Modularité et Désordre Intrinsèque : Dialogue Moléculaire de Protéines Archétypales avec leurs Partenaires Physiologiques

Julie Ledoux

université PARIS-SACLAY

Modularité et Désordre Intrinsèque : Dialogue Moléculaire de Protéines Archétypales avec leurs Partenaires Physiologiques

Modularity and Intrinsic Disorder: Molecular Dialogue Between Archetypal Proteins and their Physiological Partners

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°573, interfaces : matériaux, systèmes, usages (INTERFACES)
Spécialité de doctorat : Bioingéniérie
Graduate School: Sciences de l'ingénierie et des systèmes. Référent : ENS Paris-Saclay

Thèse préparée au **Centre Borelli** (Université Paris-Saclay, ENS Paris-Saclay, CNRS), sous la direction de **Luba TCHERTANOV**, Directrice de recherche CNRS

Thèse soutenue à Paris-Saclay, le 10 octobre 2023, par

Julie LEDOUX

Composition du Jury

Marc BAADEN Directeur de recherche CNRS, Institut de Biologie Physico-Chimique	Président
Alexandre DE BREVERN Directeur de recherche INSERM, Université Paris-Cité	Rapporteur & Examinateur
Sophie BARBE Directrice de recherche INRAE, Toulouse Biotechnology Institute	Rapporteur & Examinatrice
Alessandra CARBONE Professeure, Sorbonne Université	Examinatrice
Matthieu MONTES Professeur, Conservatoire National des Arts et Métiers	Examinateur

THESE DE DOCTORAT

NNT : 2023UPAST138

université PARIS-SACLAY | ÉCOLE DOCTORALE Interfaces: matériaux, systèmes, usages

Titre : Modularité et Désordre Intrinsèque : Dialogue Moléculaire de Protéines Archétypales avec leurs Partenaires Physiologiques

Mots clés : dynamique moléculaire ; interactome ; dynasome ; RTK KIT ; hVKORC1 ; allostérie

Résumé : La plupart des protéines sont des assemblages de modules structurellement et/ou fonctionnellement indépendants ou quasi indépendants. Les IDPs sont des protéines manifestant de telles propriétés et peuvent être soit entièrement désordonnées, soit accompagnées de régions bien ordonnées. Ces protéines jouent un rôle essentiel dans divers processus biologiques.

En utilisant des approches numériques (*in silico*), y compris la modélisation 3D et des simulations étendues de dynamique moléculaire, nous avons examiné deux archétypes d'IDP : le récepteur tyrosine kinase (RTK) KIT et le complexe 1 de la vitamine K époxyde réductase humaine (hVKORC1). Ces protéines sont des cibles thérapeutiques essentielles pour le cancer et de la coagulation sanguine, respectivement.

Par la génération de modèles complets de leurs formes sauvages et mutantes, nous avons caractérisé leurs propriétés intrinsèques et dynamiques (DYNASOME) et étudié leurs interactions avec leurs partenaires physiologiques (INTERACTOME) afin d'étudier l'initiation des voies de signalisation régulées par le RTK KIT et la réaction d'échange thiol-disulfure déclenchant l'activation de hVKORC1.

L'ensemble, cette recherche offre une base pour une exploration plus poussée de l'activation des protéines et des mécanismes de résistance, en particulier en termes de régulation allostérique. De plus, les données obtenues seront utiles pour des avancées thérapeutiques potentielles (médicaments *allo-network*). Enfin, les stratégies et les protocoles établis dans cette étude peuvent être étendus à l'étude d'autres IDPs modulaires.

Title: Modularity and Intrinsinc Disorder: Molecular Dialogue Between Archetypal Proteins and their Physiological Partners

Keywords: molecular dynamics; interactome; dynasome; RTK KIT; hVKORC1; allostery

Abstract: Most proteins are assemblies of structurally and/or functionally independent or quasi-independent modules. IDPs are proteins that manifest such properties and can be either fully disordered or accompanied by well-ordered regions. These proteins play vital roles in various biological processes.

By employing numerical *(in silico)* approaches, including 3D modeling and extended molecular dynamics simulations, we examined two archetypical IDPs: the receptor tyrosine kinase (RTK) KIT and the human vitamin K epoxide reductase complex 1 (hVKORC1). These proteins are critical therapeutic targets in cancer and blood coagulation, respectively.

By generating full-length dynamical models of their wild type and mutant forms, we characterized their intrinsic and dynamical properties (DYNASOME) and studied their interactions with their physiological partners (INTERACTOME) to investigate the initiation of signaling pathways regulated by RTK KIT and the thiol-disulfide exchange reaction leading to hVKORC1 activation.

Overall, this research offers a basis for further exploration on protein activation and resistance mechanisms, particularly in terms of allosteric regulation. Moreover, the obtained data will be useful for potential therapeutic advancements (allo-network drugs). Ultimately, the strategies and protocols established in this study can be extended to the investigation of other modular IDPs.

REMERCIEMENTS

Tout d'abord, je tiens à exprimer ma profonde gratitude envers ma directrice de thèse, Luba Tchertanov pour son accompagnement précieux tout au long de mes stages de master et plus encore ces trois dernières années. Votre expertise interdisciplinaire, votre engagement inébranlable envers l'excellence scientifique et votre dévouement à ma réussite ont été une source constante d'inspiration dans l'élaboration de ces recherches et de mon évolution académique. Je n'oublierai pas nos nombreuses et longues discussions, les innombrables idées partagées et les encouragements que vous avez toujours exprimés pour me faire sans cesse grandir en tant que chercheuse. Je suis ardemment reconnaissante du temps et à la patience que vous avez investis en moi. Je suis certaine que notre collaboration continuera à me guider dans ma carrière et ma quête de connaissances. Merci pour tout.

Je souhaite ensuite remercier les rapporteurs Alexandre de Brevern et Sophie Barbe qui ont généreusement consacré leur temps à lire et à évaluer ce manuscrit. Vos commentaires pertinents et vos suggestions avisées aideront grandement à améliorer la qualité de mes travaux à la fois présents et futurs. Pour cela, je vous en suis grandement reconnaissante.

A tous les membres du jury présents lors de la soutenance de mon doctorat – Marc Baaden, Alexandre de Brevern, Sophie Barbe, Alessandra Carbone et Matthieu Montes – j'exprime ma sincère gratitude pour leur présence et l'intérêt qu'ils ont porté à mes travaux de recherches. Je vous remercie pour vos questions et vos remarques instructives qui ont fait de cette soutenance une excellente expérience de discussions et d'apprentissage.

Je tiens à adresser mes remerciements aux membres de mon jury de mi-thèse – Konrad Hinsen, Alexey Alexandrov, Alain Trouvé, Isaure Chauvot de Beauchêne et Elodie Laine – pour leur intérêt pour mes recherches encore naissantes et dont les conseils précieux m'ont permis d'aborder sereinement la seconde moitié de mon doctorat.

Je remercie à nouveau Alain Trouvé et Elodie Maignant du Centre Borelli pour leur contribution directe à mes recherches et les discussions enrichissantes qui m'ont permis d'appréhender une autre approche de notre discipline et de ma thèse.

Mes remerciements au Centre Borelli et à Nicolas Vayatis pour m'avoir accueillie au sein du centre de recherches et aux équipes d'administration du Centre et de l'école doctorale pour leur aide précieuse tout au long de mon doctorat.

Merci aux anciens doctorants et stagiaires de l'équipe BiMoDyM – Maxim, Enki, Aurélien et Marina – dont le travail immense n'aura su qu'enrichir ces travaux de recherches.

Je remercie également les chercheurs de l'Institut de Biologie Physico-Chimique, en particulier Chantal Prévost et Charles Robert, pour l'intérêt qu'ils ont porté à mes recherches, leur gentillesse et leur accueil chaleureux lors de mes visites à l'institut.

Aux amis de longue date qui ne m'ont pas lâchée depuis 15 ans et à ceux de moins longue date mais tout aussi importants – Audrey, Grégoire, Oli, Romain, Loïc, Stéphane-Jean, Camille, Juliette, Edouard, Clémence, Solène, Paul-Henri, Mélanie, Adrien, Alice, Emeric, Céline, Raphaëlle, Jules – merci d'avoir été présents pendant ces trois ans. Aparté personnel pour Romain : (1) t'as vu, j'ai fait une *vraie* thèse de sciences, il y a 6 instances de π dans le manuscrit, alors RAFO, (2) je t'ai promis que je remercierais Brandon Sanderson, alors voilà : merci Brandon !

Maman, Papa, Clément. Je ne sais que dire. Votre amour et votre soutien a été une source de force depuis toujours. Maman, Papa, je vous suis immensément reconnaissante d'avoir tout fait pour que je puisse réaliser des études qui me passionnent. Vous m'avez apporté le courage et la ténacité, la curiosité et l'envie de découverte. Vous m'avez ouverte à tout et c'est grâce à vous que je suis ici. Clément, tu es le pilier fraternel que toute grande-sœur rêve d'avoir. Je t'adore frangin.

A tous les trois, je vous dédie ce manuscrit.

Robin. On se connait depuis un moment maintenant. Ta présence est une bouffée d'air frais. Je suis heureuse que tu sois entré dans la famille et dans ma vie. A Minette qui n'a eu de cesse de me déconcentrer, je ne dirai qu'une chose « *prrrrrr ?* » toi-même.

Cette recherche a été financée par le ministère de l'Enseignement Supérieur, de la Recherche et de l'Innovation et l'Agence Nationale de la Recherche (ANR, projet 18-CE20-0025-01) ; a été soutenue par le Centre National de la Recherche Scientifique (CNRS), l'Ecole Normale Supérieure (ENS) Paris-Saclay et l'Institut Farman. Je remercie le Centre Informatique National de l'Enseignement Supérieur (CINES), le Grand Equipement National de Calcul Intensif (GENCI) et l'Institut du Développement et des ressources en Informatique Scientifique (IDRIS) pour m'avoir donné accès à leurs ressources de calculs intensifs.

TABLE DES MATIERES

TABLE DES FIGURES

CHAPITRE 7. IDENTIFICATION ET CARACTERISATION DES PROTEINES PARTENAIRES DU RKT KIT ET DE HVKORC1 176

PARTIE 1. INTRODUCTION

Pour comprendre les fonctions ou dysfonctions d'un système biologique, sa description complète à l'échelle atomique est nécessaire. Au préalable, chacun des constituants isolés d'un système (ex. dans une cellule : protéines, ADN, ARNs, petites molécules) doit être examiné par ses propriétés structurales, dynamiques et son énergie libre en relation avec ses fonctions.

Pour les protéines en particulier, cette compréhension est requise pour leur étude en tant qu'espèce individuelle et le contrôle de leur coopération, par exemple, lors de leur assemblage en complexes supramoléculaires. L'ensemble de ces descripteurs fournit une vision précise sur les mécanismes de phénomènes vitaux. Pour illustrer, le contrôle de la réponse cellulaire à un signal extracellulaire est la fonction principale des **récepteurs à activité tyrosine kinase** (RTK). Ces récepteurs sont des régulateurs clés des processus cellulaires normaux et pathogènes, en particulier oncogènes. Leurs mutations gain de fonction sont responsables de l'indépendance de leur activation à la fixation de leur ligand spécifique. Cette indépendance induit une signalisation aberrante connue dans certains cancers et maladies inflammatoires. Un autre exemple est la **Vitamine K époxyde réductase** humaine (hVKORC1), une protéine contribuant à la physiologie sanguine et dont l'activation est dépendante d'une protéine *redox* (thiorédoxine).

Une première étape clé de cette recherche est la description des protéines à toutes leurs échelles : de leur séquence (1D) à leurs repliements locaux (structures secondaires, 2D) et globaux (structure tertiaire et quaternaire, 3D). L'ensemble des niveaux de repliements et leur dynamique (ensemble conformationnel) constitue leur **DYNASOME**[1] (**Figure 2**).

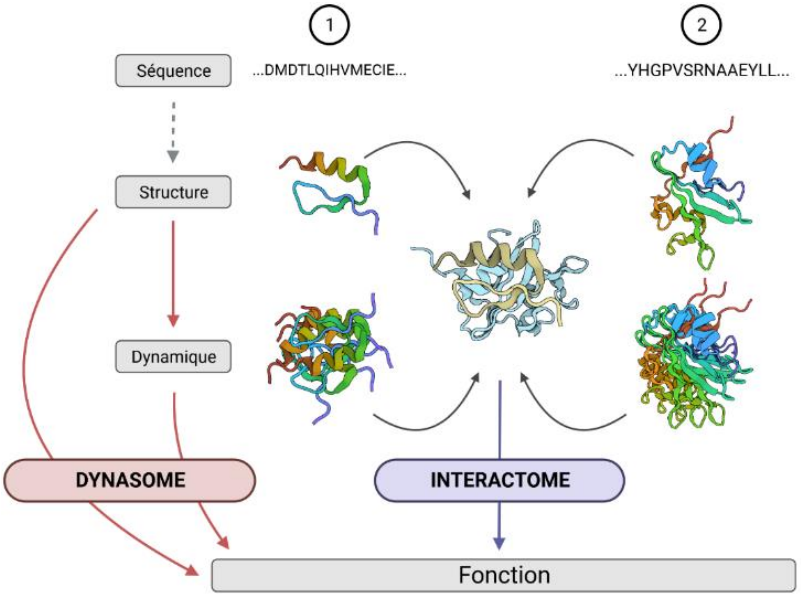


Figure 2 Niveaux de description des protéines pour comprendre leurs fonctions.

Cette description ascendante est une fondation de la recherche sur les relations du **DYNASOME** avec les données biologiques d'une protéine et la caractérisation de son **INTERACTOME** ou l'ensemble des interactions moléculaires protéines-protéines au sein d'une cellule, d'un tissu ou d'un organisme[2]. Le **DYNASOME** et l'**INTERACTOME** sont deux grandes échelles de caractérisation d'un système biologique et représentent une des plus grandes énigmes en biologie.

Les principaux enjeux et questions liés à ces deux grands axes de recherche en biologie sont présentés au CHAPITRE 1.

CHAPITRE 1. PROTEINES MODULAIRES, DESORDRE INTRINSEQUE ET ALLOSTERIE

1.1. MODULARITE DES PROTEINES

La plupart des protéines sont des structures plus ou moins complexes de régions structurellement et/ou fonctionnellement (quasi) indépendantes. Ces régions appelées « domaines » sont jointes chimiquement par des boucles flexibles de tailles variables souvent désordonnées[3]. Les domaines protéiques sont réunis en trois groupes mutuellement non exclusifs : les **domaines structuraux** dont le repliement est indépendant du reste de la protéine, et pouvant s'associer en protéines multidomaines, les **domaines fonctionnels** réalisant une activité spécifique même isolés du reste de la protéine, et les **domaines de localisation topologique** représentant la position spatiale spécifique dans l'environnement cellulaire. Les récepteurs à activité tyrosine kinase, par exemple, manifestent toutes ces qualités. Ils sont assemblés en domaines topologiques localisés dans des milieux différents : un domaine extracellulaire (ED), un domaine transmembranaire (TMD) et un domaine cytoplasmique (CD)[4]. Chaque domaine possède sa propre structure 3D : *immunoglobulin-like* ou *Ig-like* de feuillets β (ED), hélice α (TMD), une composition mixte d'hélices α et de feuillets β formant une structure globulaire (majorité du CD) ou entièrement dépliée. Chacun de ces domaines effectue ses propres fonctions comme la liaison d'un ligand spécifique et la dimérisation (ED), l'hydrolyse de l'adénosine triphosphate (ATP) par le domaine kinase (CD) et les interactions avec ses partenaires cellulaires (CD)[5]. D'autres exemples sont des modules communs à de grandes familles de protéines : les domaines catalytiques (ex. le domaine ribonucléase[6]), les domaines d'interactions (ex. les domaines *Src Homology*) ou encore les motifs de structures secondaires (ex. les motifs *EF-hand* ou *zinc finger*)[7]. Ces modules sont identifiés par leur homologie de séquence, ontologie, caractéristiques structurales communes ou encore leurs réseaux d'interactions[3,8–11].

L'architecture modulaire des protéines est particulièrement flagrante dans les protéines de haut poids moléculaire ou les protéines effectuant plusieurs fonctions.

Les **modules** consistent en un groupe de résidus hautement coopératifs participant aux fonctions d'une protéine[12]. D'une part, ils facilitent les **interactions protéines-ligands, protéines-protéines et/ou protéines-ADN/ARNs** et contribuent à leur affinité de liaison. D'autre part, ils participent, avec leurs boucles et par des mécanismes dits « allostériques », au **transfert d'une information** entre régions protéiques plus ou moins éloignées[12–15] (voir section 1.3).

Outre son rôle physiologique crucial, la modularité des protéines permet nombre d'applications en recherche, biotechnologie et pharmacologie. Parmi ces applications,

on peut citer le *screening in silico* de librairies de petites molécules ou peptides visant un seul module fonctionnel, le *design* de modulateurs intermodules et/ou inter-protéiques, des méthodes de biologie expérimentale[7] et en biologie computationnelle, pour la réduction de la taille des protéines étudiées et l'optimisation les calculs.

On a décrit brièvement les caractéristiques structurales et fonctionnelles des modules. En plus des boucles flexibles les reliant, les modules protéiques peuvent receler à la fois des régions bien structurées et « intrinsèquement désordonnées ». Cette structuration des protéines en blocs hybrides permet de réaliser de nombreuses fonctions biologiques (auto-inhibition, activation, phosphorylation, réaction enzymatique, transfert de protons, ou interactions protéines-protéines).

1.2. DESORDRE INTRINSEQUE

Le paradigme historique en biologie structurale affirme que les protéines assurent leurs fonctions grâce à une structure tridimensionnelle (3D) bien définie. Or, les protéines sont loin d'être des entités rigides et statiques. Elles forment des macromolécules hybrides se plaçant dans un *continuum* structural, aux deux extrémités se trouvant les protéines « ordonnées » (bien structurées et stables) et entièrement « désordonnées »[16].

Les protéines ou leurs régions désordonnées (IDPs/IDRs) ne possèdent pas de structure 3D conservée. Elles existent comme un ensemble de conformations hétérogènes leur permettant d'assurer leurs fonctions physiologiques ou pathologiques. Les IDPs constituent une partie très significative des protéines dans les trois règnes du vivant[17,18] et concernent, par une estimation approximative, à un minimum de 40 % du protéome eucaryote.

1.2.1. MANIFESTATION DU DÉSORDRE INTRINSÈQUE

Quelles propriétés moléculaires révèle une « protéine désordonnée » ? Dans les études empiriques, le désordre se manifeste notamment par une impossibilité de cristallisation, plusieurs positions d'occupation des atomes dans une structure ou encore des valeurs excessives de facteur bêta ou résistance à la protéolyse[19,20].

Les IDPs/IDRs s'identifient par un ensemble de caractéristiques qui leur sont propres.

Au niveau de la séquence protéique. Les comparaisons de la séquence et de la structure estimée par méthodes empiriques (ex. diffraction aux rayons X, dichroïsme circulaire, résonnance magnétique nucléaire, diffraction aux petits angles) ont montré

que la composition en acides aminés des régions ordonnées et désordonnées est dramatiquement différente[21]. Les IDRs sont enrichies en résidus de petite taille ou polaires (*A, G, R, Q, S, P* et *E*), fréquemment exposés au solvant ou présents dans les boucles[21]. Au contraire, la quantité de résidus hydrophobes à chaîne aliphatique ou cycliques (*W, C, F, I, Y, V, L* et *N*) est appauvrie. En revanche, les IDRs sont indifférentes à l'enrichissement en *H, M, T* et *D*. Le biais de composition, le manque de complexité de la séquence (motifs répétitifs) et les propriétés physico-chimiques des acides aminés permettent la discrimination entre les régions ordonnées et désordonnées. Ces éléments sont à la base du développement de nombreux outils de prédiction des IDRs[22]. En 2021, plus de 100 prédicteurs sont recensés[23]. La prédiction du désordre a été introduite dans les *Critical Assessment of Structure Prediction* (CASP) en 2002 (CASP5) et, plus récemment, la prédiction de structures protéiques par AlphaFold a permis l'identification de régions pouvant être mal interprétées comme désordonnées bien que c'en soit un bon indicateur[24].

Au niveau structural et conformationnel. Par l'enrichissement en acides aminés chargés et le déficit en acides aminés hydrophobes, les IDRs peuvent être vues comme des poly-électrolytes de grande surface accessible au solvant[25]. Ainsi, le cœur hydrophobe souvent caractéristique des protéines ordonnées est absent ou peu stabilisant. Les IDRs n'adoptent ni structures secondaires ni structure 3D stables. De plus, elles émettent une flexibilité conformationnelle remarquable, dans un désordre intrinsèque (intramodule) et extrinsèque (intermodules). Les forces d'attractions-répulsions induites par la charge nette des résidus leur permettent d'adopter des structures plus ou moins repliées en conditions physiologiques[26]. À cet égard, les IDPs existent comme un ensemble conformationnel vaste composé d'entités hybrides dans un *continuum* ordre-désordre. Ce sont des régions subissant ou non des transitions réversibles « ordre-ordre » (changement de structure secondaire), et/ou « ordre-désordre » (passage d'une structure ordonnée vers une forme dépliée), et présentant des déplacements anisotropes (linéaires et rotationnels). Contrairement aux protéines ordonnées, l'espace conformationnel visité par les IDPs est rugueux avec une surface d'énergie libre aplatie, sans minimum global ou local ni barrières énergétiques importantes[27] (**Figure 1.1**). Il est à noter que certaines protéines, notamment redox-dépendantes, sont conditionnellement désordonnées par la présence ou l'absence de ponts disulfures stabilisants[28].

Au niveau évolutif. Les séquences protéiques sont des marqueurs de l'évolution accumulée au cours du temps à travers divers mécanismes (ex. mutations, délétions, insertions par mutations *de novo*, transferts horizontaux ou latéraux de gènes). Au niveau génomique, les IDPs sont fréquemment encodées dans les régions riches en GC et dans les exons subissant un épissage alternatif. Ces régions favorisent ainsi la diversité fonctionnelle des protéines sans influence sur leur structure[29]. Au niveau évolutif, le désordre des IDPs/IDRs est catégorisé : en ***désordre flexible***, où la séquence évolue rapidement en conservant les propriétés désordonnées de la

protéine, en ***désordre contraint*** dans lequel la séquence protéique et le désordre conformationnel sont conservés, en ***désordre non conservé*** où les IDRs sont désordonnées (ou absentes) dans certaines espèces et pas d'autres[30]. Malgré ces contraintes évolutives faibles, les séquences de longues régions désordonnées peuvent être porteuses de résidus hautement conservés nécessaires à leurs fonctions[31] (voir section 1.2.3).

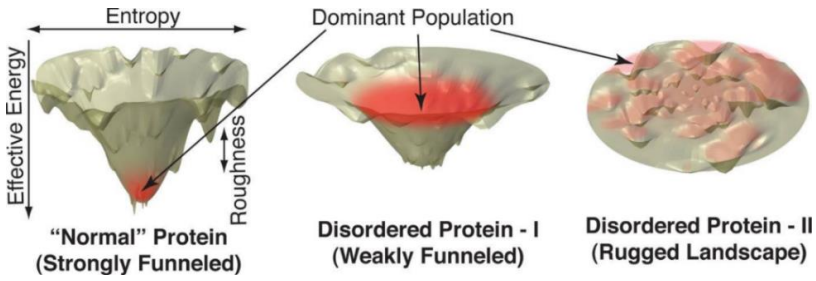


Figure 1.1 Espaces conformationnels des protéines ordonnées et intrinsèquement désordonnées représentés par des paysages énergétiques. Figure reprise de [27].

Ces propriétés conformationnelles et la diversité génétique induite par leur histoire évolutive confèrent aux IDPs/IDRs une flexibilité et une plasticité nécessaires aux fonctions physiologiques de nombreuses protéines.

1.2.2. RÔLES FONCTIONNELS DES IDPS

On n'a vu qu'au niveau évolutif, les IDPs/IDRs évoluent plus vite que les protéines ou régions ordonnées[31]. Dans l'état de la recherche actuelle, les IDPs/IDRs peuvent être regroupées en plusieurs classes basées sur la comparaison de leurs séquences[32]. La classification utilisée semble être celle de Peter Tompa[33], développée ci-après.

Les **chaînes entropiques** des protéines multidomaines sont des boucles flexibles régulant l'orientation de domaines structuraux et/ou fonctionnels les uns par rapport aux autres en leur donnant un très grand degré de liberté[34]. Elles sont généralement impliquées dans une ou plusieurs fonctions biologiques. Dans les RTKs, par exemple : la région juxtamembranaire (JMR), le *hinge* et la boucle activation (boucle A) contribuent à leur auto-inhibition et le JMR, le domaine insert kinase (KID) et queue C-terminale (C-tail) sont des plateformes de transduction du signal[35]. Les IDPs/IDRs impliquées dans la **reconnaissance moléculaire** et la modulation de l'activité des protéines sont régulées par des mécanismes allostériques intra- et/ou inter-protéine. La grande plasticité de leur surface accessible au solvant permet la liaison d'une protéine avec un ou plusieurs partenaires selon le principe *one-to-many* ou *many-to-one*[36]. Souvent, ce phénomène de reconnaissance et de liaison entraîne un changement conformationnel appelé *coupled folding and binding*. Les IDPs peuvent altérer leur fonction par les mécanismes allostériques (voir section 1.3). La double

fonction de JMR d'auto-inhibition/activation du RTK KIT et ses interactions avec ses partenaires cellulaires en est un exemple[37]. Au plus haut niveau d'organisation architecturale, les IDPs/IDRs font partie des **assembleurs moléculaires** participant à la formation de complexes supramoléculaires hautement entrelacés (ADN et ARN polymérases[38], complexes d'histones[39], ATP synthase[40]. Les **chaperonnes désordonnées** effectuent une liaison transitoire à un partenaire (ARN, protéine) et l'aident à atteindre son repliement natif et fonctionnel (ex. p50, Hsp70, GroEL)[41]. Plus généralement, elle permettra l'atteinte accélérée d'un minimum local d'énergie libre en établissant le meilleur réseau d'interactions non covalentes structurantes dans la protéine[42]. Les **modifications post-traductionnelles** (ex. phosphorylation, acétylation, méthylation, etc.) régulent l'activité des IDPs en agissant sur leur capacité à lier ou non une autre région protéique ou un ligand[43,44]. Ces résidus modifiés facilitent leur reconnaissance par des protéines effectrices provoquant une réponse protéique à des *stimuli* (ex. activation, (dé)stabilisation de complexes, modification de la localisation cellulaire). Les IDRs des **éboueurs ou *scavengers*** sont responsables de la séquestration ou la dégradation de ligands (ex. *actin-scavenger system*[45], *mRNA-decapping scavenger enzyme* DcpS[46], peroxideroine-4[47]). Les **éponges à métaux**, quant à elles, sont impliquées dans la séquestration des ions métalliques (ex. protéines contenant des motifs *EF-hand*[48] ou *zinc finger*[49,50]). Enfin, les IDPs/IDRs favorisent la création de certaines **organelles sans membrane** (assemblages protéines-protéines ou ARNs-protéines) en promouvant une séparation de phase liquide-liquide[51]. Ces composants cellulaires spécialisés (ex. corps de Cajals, *P-bodies*[52]) assistent la régulation et l'isolation contrôlée de molécules dans des compartiments séparés du reste du contenu cellulaire.

Le plus souvent, ces fonctions ne peuvent se faire que par la présence de courts motifs fonctionnels particuliers, conservés et encodés dans la séquence.

1.2.3. MOTIFS FONCTIONNELS ENCODES DANS LES IDRS

Les ***molecular recognition features*** (MoRFs) sont des motifs courts de 10 à 70 acides aminés dont la fonction principale est la reconnaissance et la liaison de partenaires spécifiques à travers une préstructuration dans l'état non lié (ex. le domaine TAD de p53[53,54]) et/ou la transition d'une région désordonnée à des structures ordonnées[55]. Un seul MoRFs peut lier plusieurs protéines partenaires pour adopter des structurations multiples après liaison. Plus courts, les ***short linear motifs*** (SLiMs) sont des segments fonctionnels hautement conservés encodés dans 3 à 10 acides aminés[56]. Les SLiMs servent de motifs d'amarrage moléculaire en augmentant la spécificité de reconnaissance. Leurs interactions sont transitoires et faibles. De plus, leur rôle d'interrupteur moléculaire permet la liaison exclusive de différents partenaires sur un même site[57,58]. Ces deux propriétés sont souvent régulées par la présence de modifications post-traductionnelles. Les **domaines intrinsèquement désordonnés**

(IDDs) sont des modules fonctionnels et structuraux de 20 acides aminés ou plus. Ils s'engagent dans des interactions plus fortes avec des partenaires que les MoRFs ou les SLiMs où ils forment des complexes beaucoup plus stables. Contrairement aux motifs fonctionnels courts, les IDDs sont capables de reconnaître et lier non seulement des régions ordonnées, mais aussi d'autres IDDs en induisant leur repliement mutuel[59].

La stabilisation par les MoRFs et les SLiMs est réversible (métastable) et ces deux éléments fonctionnels, avec les IDDs ne sont pas mutuellement exclusifs dans les IDPs/IDRs.

1.2.4. REGULATION DE L'EXPRESSION ET DE L'ACTIVITE DES IDPS

Les IDPs sont au cœur de nombreux processus cellulaires couplés, critiques et indispensables à l'exécution de nombreuses fonctions vitales. Un tel couplage nécessite une régulation fine pour assurer leur présence en quantité et qualité appropriée pendant un temps requis. Cette régulation des IDPs s'effectue selon deux voies : la régulation spatio-temporelle et l'augmentation de la complexité de la signalisation.

Pour une IDP, la **régulation spatio-temporelle** s'exerce sur : l'***expression spécifique dans un tissu ou une cellule*** de son gène et son épissage alternatif[60] (ex. le facteur de transcription FOXO[61]), l'***expression quantitative et temporelle*** de son ARNm et sa traduction en réponse à un stimulus extérieur (ex. le récepteur à l'insuline)[62], la ***compartimentalisation*** de la protéine dans les organelles sans membrane, sa ***dégradation*** plus rapide que les protéines ordonnées dans le protéasome[63].

L'**augmentation de la complexité de la signalisation** est médiée par différents facteurs : des ***protéines adaptatrices*** permettant la sélection d'une voie de signalisation plutôt qu'une autre (ex. les voies dépendantes et indépendantes du *myeloid differentiation response protein 88* (MyD88) dans la transduction du signal par le *Toll-like receptor 4*[64], une signalisation dépendante de la ***modification post-traductionnelle*** de plusieurs sites spécifiques régulant les interactions avec des partenaires et l'activation de voies de signalisation ou leur dégradation (ex. le RTK KIT[65]), le besoin d'interactions complexes impliquant des ***protéines bistables*** (ex. dimérisation des RTKs[65], les phénomènes de ***rétrocontrôles*** négatifs (ex. l'activation de Grb2/SOS par EGFR entraîne l'activation de c-Cbl inhibant l'activité de EGFR[66].

Ces processus permettent une haute limitation de la production de transcrits, une régulation de l'activité des IDPs et leur dégradation, associés au fonctionnement normal de la cellule. Cependant, l'apparition de variations génétiques et/ou une dérégulation de tels mécanismes d'inhibition sont responsables de pathologies.

1.2.5. IDPs ET PATHOLOGIES

Les nombreuses fonctions médiées par les IDPs sont cruciales pour la physiologie et la régulation de nombreux processus cellulaires. Les IDPs sont, pour la plupart, représentées dans des protéines participant à la signalisation cellulaire, associées aux cancers comme les proto-oncogènes et les suppresseurs de tumeurs[67].

Pour les processus physiologiques, l'activité des IDPs doit être parfaitement contrôlée[62]. Cependant, la sur- et sous-expression ou la dérégulation de l'activation des IDPs sont à l'origine de leur activité aberrante produisant des interactions cellulaires anormales[68]. Les IDPs sont hautement régulées de la transcription à la traduction et dans leurs modifications post-traductionnelles. Dans le génome et pendant la transcription, des évènements mutationnels sont les causes principales de cette dérégulation[67,68].

Les **mutations somatiques**, portées par les motifs linéaires[69] ou des résidus centraux des fragments hautement fonctionnels, influencent l'apparition de repliements alternatifs menant à l'activation constitutive des protéines et/ou leur résistance à l'inhibition[70]. Ces mutations *hotspot* provoquent l'altération de la variabilité conformationnelle de la protéine et à la modification des interactions avec leurs partenaires[71]. Un **épissage alternatif** altéré induit la création ou délétion de régions fonctionnelles[29]. La **translocation chromosomique** est responsable de la production de protéines de fusion (des hybrides non naturels de plusieurs protéines) préférentiellement hautement désordonnées[72] (ex. la fusion des gènes *BCR* et *ABL*[73]). Leur manque de similarité structurale avec des protéines ordonnées leur permet d'échapper à la reconnaissance de leur structure alternative et à leur dégradation par leurs protéines régulatrices. La **variabilité du nombre de copies** (CNV) affecterait également l'expression des gènes codant pour les IDPs[74].

Ces variants génétiques impactent la structure du gène, de la protéine, mais aussi sur la disponibilité exacte des IDPs dans la cellule en quantité, qualité et pour un temps limité. Une diminution de cette disponibilité est responsable de la réponse physiologique faible de la cellule. Une augmentation, quant à elle, est à l'origine de l'accroissement des interactions de la protéine avec un partenaire usuel (partenaire 1). Lorsque les mécanismes de régulation ne sont plus suffisants, l'association de l'IDP avec un partenaire 2 similaire au partenaire 1 (mais pas identique) permet l'activation de nouveaux processus. L'altération de la réponse cellulaire médiée par cette IDP modifiée est alors due à une perte de fonction du partenaire 1 ou un gain de fonction du partenaire 2. Ce principe est aussi valide aux relations protéines / acides nucléiques[62]. Ainsi, ces altérations génétiques perturbent l'équilibre fin de nombreuses fonctions cellulaires.

Les fonctions développées par ces variations génétiques aboutissent à une activité cellulaire incontrôlée (ex. prolifération accrue, suppression de l'apoptose) conduisant

au développement de formes des cancers ou d'autres maladies graves comme le diabète, maladies cardiovasculaires ou maladies neurodégénératives[68].

1.2.6.IDPs/IDRs EN TANT QUE CIBLES POUR LE DEVELOPPEMENT DE MOLECULES THERAPEUTIQUES

L'implication importante des IDPs dans les pathologies suggère qu'elles sont des cibles de choix pour le développement de nouvelles molécules thérapeutiques.

Sous l'influence du paradigme « clé-serrure », les premiers modulateurs « classiques » de l'activité des protéines (petites molécules ou peptides) étaient développés pour cibler une poche comme le site actif ou un site alternatif souvent localisé dans une région ordonnée.

Les modulateurs dits **réversibles** se distinguent par leur capacité à influencer la concentration de substrat nécessaire à l'activité d'une protéine et/ou la vitesse de la réaction qu'elle réalise[75]. Ils ne se lient pas de manière covalente à(aux) protéine(s) qu'ils reconnaissent. Les ***inhibiteurs compétitifs*** entrent en concurrence avec le substrat pour se loger dans sa poche. Ce mode d'inhibition implique une similarité structurale et chimique de l'inhibiteur avec le substrat initial. Les ***inhibiteurs non compétitifs*** s'établissent dans une poche alternative pour diminuer, à concentration de substrat égal, la vitesse de réaction et de formation de produits. Ils ne modifient pas la reconnaissance du substrat par la protéine. Les inhibiteurs compétitifs et non compétitifs se lient à la protéine seule, au contraire des ***inhibiteurs incompétitifs*** interagissant avec le complexe protéine-substrat sur le site actif ou un site alternatif révélé par l'assemblage protéine-protéine. Les **inhibiteurs allostériques** se logent sur un site alternatif du site actif pour moduler positivement ou négativement l'activité de la protéine en agissant sur son affinité au substrat par un mécanisme allostérique[75] (voir section 1.3).

Les **modulateurs irréversibles** s'engagent, eux, dans la consolidation des complexes moléculaires en empêchant leur dissociation. La majorité des modulateurs de protéines interagissent avec leur(s) cible(s) par des interactions non covalentes réversibles. Cependant, 30 % des médicaments mis sur le marché comprennent des modulateurs covalents[76].

Le *design* de modulateurs spécifiques d'un site protéique requiert l'identification empirique des poches à cibler ou la structure du complexe. Une telle recherche est dépendante de la disponibilité d'une structure 3D, au moins partielle. Les IDPs/IDRs ne possèdent pas de structure 3D stable rendant difficile, voire impossible, l'identification de ces poches et la position spatiale des résidus *hotspot* qui leur sont spécifiques (MoRFs, SLiMs). De plus, la non spécificité de certaines molécules inhibitrices, l'apparition systématique de mutations entrainant des résistances aux traitements,

leurs effets secondaires et l'absence de réponse cellulaire laissent les cliniciens avec des alternatives limitées[67,77]. Par conséquent, il est nécessaire de trouver de nouvelles stratégies pour le développement de nouvelles thérapeutiques plus spécifiques et efficaces visant ces protéines.

La modulation des IDPs/IDRs par une modification de leur hybridité ordre-désordre est une stratégie prometteuse principalement réalisée par de petites molécules imitant des domaines protéiques peptidiques ou non peptidiques (macrocycles)[78]. Une première stratégie est la **séquestration de la protéine ou du complexe dans un état non fonctionnel par réarrangement structural** *(drug-induced (mis)folding)*, c'est-à-dire basculer les conformations majoritaires de l'espace conformationnel *(population shift)* vers des conformations inaptes à la formation de complexes fonctionnels[79] ou d'agrégats toxiques[80,81] (oligomères, condensats), ou formant des agrégats non toxiques[82]. Ces molécules peuvent agir sur un site allostérique dont la modulation par une petite molécule n'est pas compétitive, les propriétés dynamiques et structurales intrinsèques des IDRs[83], les motifs conservés d'interaction protéines-protéines (MoRFs, SLiMs) avec ou sans modifications post-traductionnelles[84,85]. Les **modulations allostériques ou compétitives** inactivent non seulement la fonction principale de reconnaissance, mais aussi facilitent la dégradation de ces IDPs et les protéines de fusion aberrantes[81]. Pour les facteurs de transcription, une seconde stratégie consiste à **bloquer l'interaction dimère/ADN** ou diminuer sa spécificité de reconnaissance lorsque le dimère est responsable de l'augmentation de l'expression d'une des protéines le constituant[86].

L'allostérie a été plusieurs fois citée comme gouvernant la fonction des modules protéiques et comme potentiel levier de modulation des IDPs/IDRs. Mais à quoi fait référence ce phénomène ?

1.3. REGULATION ALLOSTERIQUE DES PROTEINES

Classiquement, l'allostérie est une régulation de la fonction des protéines par un **effecteur** (ex. la fixation d'un ligand sur un site différent du site actif).

Les effets d'un effecteur peuvent être considérés comme des perturbations locales ou globales d'un objet biologique et sont parfois décrits par une propagation du signal et qualifiés par dynamique moléculaire[87–93]. La propagation d'un signal de perturbation à travers la structure 3D d'une protéine relève du concept de couplage ou de communication allostérique[88,94,95]. Cette perturbation mécanique induite provoque une séquence de réarrangements conformationnels au sein d'une protéine et/ou un changement de flexibilité dynamique d'un ou de plusieurs sites éloignés spatialement du site de perturbation[87,96]. Ces modifications structurales et/ou comportementales des protéines sont suscitées par une variété d'effecteurs : des interactions non covalentes (fixation d'ions, ligands, protéine, ARN, ADN), l'absorption

de la lumière, des évènements covalents tels que des modifications post-traductionnelles, des mutations, des réactions chimiques (protonation/déprotonation, oxydation/réduction de ponts disulfures, etc.)[87,94].

Les effets allostériques sont de nature variée et régulent la majorité des processus biologiques (ex. signalisation, canaux ioniques, transcription, métabolisme, etc.)[96–98]. Historiquement, les modèles proposés voyaient l'allostérie comme la modification discrète des états conformationnels de protéines multimériques (modèles de Monod, Wyman et Changeux[99], et de Koshland, Nemethy et Filmer[100]. Or, non seulement les phénomènes allostériques peuvent montrer des transitions structurales des protéines, mais aussi l'altération de la dynamique de plusieurs régions avec ou sans changements conformationnels[87,95]. Ce couplage est optimal lorsque les deux régions concernées sont intrinsèquement désordonnées ou non repliées[34,96,101].

Comme mentionné précédemment (voir section 1.2), par leur flexibilité inhérente, les IDPs peuvent être vues comme un ensemble statistique de conformations. Les modifications structurales et/ou dynamiques, induites par un effecteur et traduites par régulation allostérique, redistribuent alors cet ensemble *(population shift)* entre des états fonctionnellement différents[87,95,97,102,103] (**Figure 1.2**). La modification (ir)réversible des conformations majoritaires de l'ensemble est observable sur le paysage d'énergie libre[88,104].

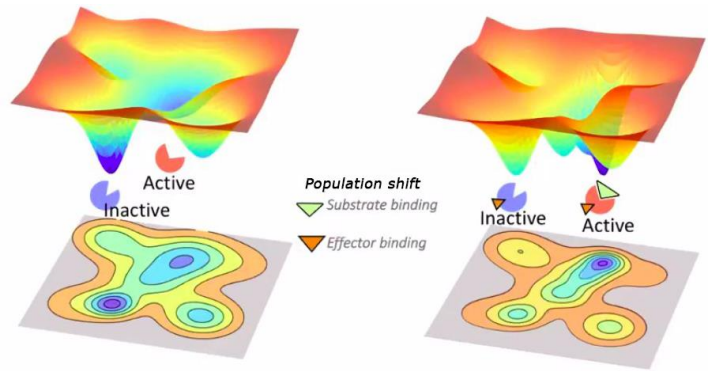


Figure 1.2 Représentation schématique de la redistribution de la population *(population shift)* d'une enzyme à l'issue de la liaison d'un substrat et d'un effecteur. Figure reprise de [104].

Par analogie avec sa définition, l'allostérie concerne des évènements de propagation d'un signal d'une région protéique à d'autres. L'analyse de l'histoire évolutive des protéines a permis de découvrir que de nombreux résidus conservés sont essentiels non seulement pour leurs fonctions, mais aussi à la communication intra- et inter-protéines. Cette communication s'effectue à travers un réseau allostérique[105,106]. Ainsi, ces réseaux sont sensibles à l'altération de ces résidus, par exemple, par mutation de type substitution[15,107,108].

La régulation allostérique est une part importante de la signalisation cellulaire[98].

On a vu dans la section précédente que les IDPs/IDRs jouent un rôle prépondérant dans la médiation des évènements biologiques par divers phénomènes et mécanismes. Les IDPs/IDRs sont impliquées dans la régulation allostérique complexe des protéines[88,109]. On a également vu que les IDPs/IDRs sont le plus souvent concernées par les altérations génétiques ou post-traductionnelles responsables de la modulation et/ou l'altération des réseaux allostériques. C'est pourquoi les IDPs/IDRs ont un rôle majeur dans la régulation allostérique[34].

L'altération de la régulation allostérique est à l'origine de la dysfonction des IDPs (et *vice-versa*) et, par conséquent, est responsable de pathologies[110]. Ainsi, l'étude des phénomènes allostériques est essentielle et pour le développement de nouvelles molécules thérapeutiques plus sélectives, spécifiques et efficaces[94,109,110] (voir section 1.2.6) pour comprendre les mécanismes cellulaires des processus physiologiques et pathogènes qui leur sont associés.

Des exemples archétypaux de ces phénomènes fondamentaux (modularité, désordre intrinsèque et régulation allostérique) sont le **RTK KIT**, une protéine clé de la régulation de la signalisation cellulaire, et la **Vitamine K époxyde réductase** humaine (hVKORC1), une protéine contribuant à la physiologie sanguine. Les fonctions et propriétés de ces deux protéines sont respectivement décrites dans les Chapitres CHAPITRE 2 et CHAPITRE 3.

CHAPITRE 2. LE RECEPTEUR TYROSINE KINASE KIT : STRUCTURES ET FONCTIONS AUX ETATS SAUVAGES ET PATHOLOGIQUES

Les récepteurs tyrosines kinases (RTKs) sont des protéines de la membrane cytoplasmique adaptées à la reconnaissance de nombreuses biomolécules. Ils contrôlent la transduction d'un signal extracellulaire vers le noyau *via* une cascade de signalisation altérant l'expression de nombreux gènes liés aux processus cellulaires normaux et pathogènes.

Notre intérêt s'est porté sur un membre représentatif des RTKs, le **RTK KIT**. KIT, ou le cluster de différenciation CD117, est une protéine membranaire à activité tyrosine kinase de la famille III à laquelle s'ajoutent les deux *platelet-derived growth factor receptors* (PDGFR) α et β, la *FMS-like tyrosine kinase 3* (FLT3) et le *colony stimulating factor 1 receptor* (CSF1R)[65]. La stimulation du KIT par le *stem-cell factor* (SCF) dans le milieu extracellulaire permet le recrutement dans le cytoplasme de partenaires initiateurs de voies de signalisation responsables de la survie, la prolifération, la différentiation et la migration cellulaire[65].

Les mutations gain de fonction du RKT KIT sont responsables de l'indépendance du KIT à la stimulation par le SCF ainsi que la surexpression de son gène. Son activation constitutive et son activité peu contrôlée entrainent une signalisation aberrante en partie induite ou responsable de cancers comme les leucémies myéloïdes aiguës (LAMs), mastocytoses, tumeurs stromales gastro-intestinales (GISTs), mélanomes, tumeurs des cellules germinales (TGMs)[111] et maladies inflammatoires[112]. Du type de suractivation du récepteur dépend sa sensibilité à des inhibiteurs d'activité kinase comme l'imatinib. Cette molécule a montré des réponses satisfaisantes dans des cancers tels que les GISTs, mais certaines mutations du KIT sont résistantes à ce traitement[113].

Malgré de nombreuses études en biologie et physiopathologie, les principes physiques détaillés des phénomènes d'activation du KIT et de ses voies de signalisation ou encore les effets des mutations sur sa structure et les mécanismes de résistance induite restent peu élucidés et les moyens de description adéquats à cette échelle restent limités.

On décrira dans ce chapitre les caractéristiques structurales et fonctionnelles du KIT aux états sauvages et pathogènes, ses mécanismes de résistance et son potentiel en tant que cible pour le développement de nouveaux inhibiteurs. On exposera les difficultés de telles études *in silico* et les résultats précédents répondant à ces questions complexes. La grande majorité de ces résultats a été produite par l'équipe BiMoDyM.

2.1. STRUCTURES DU KIT SAUVAGE ET MUTE, ET LEURS RESEAUX ALLOSTERIQUES

À l'instar des autres RTKs (excepté la famille STYK1), KIT est constitué d'un domaine extracellulaire (ED), par lequel est reconnu le SCF, une hélice α transmembranaire (TMD) et un domaine kinase cytoplasmique (CD). Le CD possède plusieurs sites tyrosines et sérines phosphorylables reconnaissables par des protéines partenaires de la signalisation cellulaire.

2.1.1. ARCHITECTURE GENERALE DE LA PROTEINE

Le KIT est une protéine de 976 acides aminés (aas) composée de trois domaines de localisation topologiques chacun constitués de nombreux domaines structuraux et fonctionnels[114]. La structure intégrale du RTK KIT n'est pas résolue pour ses formes mono- et dimériques (état inactif et actif). Cependant, les méthodes empiriques ont permis de décrire la structure de ses domaines partiels, donnant une vision très claire sur leurs rôles structuraux et fonctionnels (mécanisme d'activation, fonction catalytique), mais aussi de les utiliser en tant que cibles pour le développement d'inhibiteurs (**Figure 2.1**).

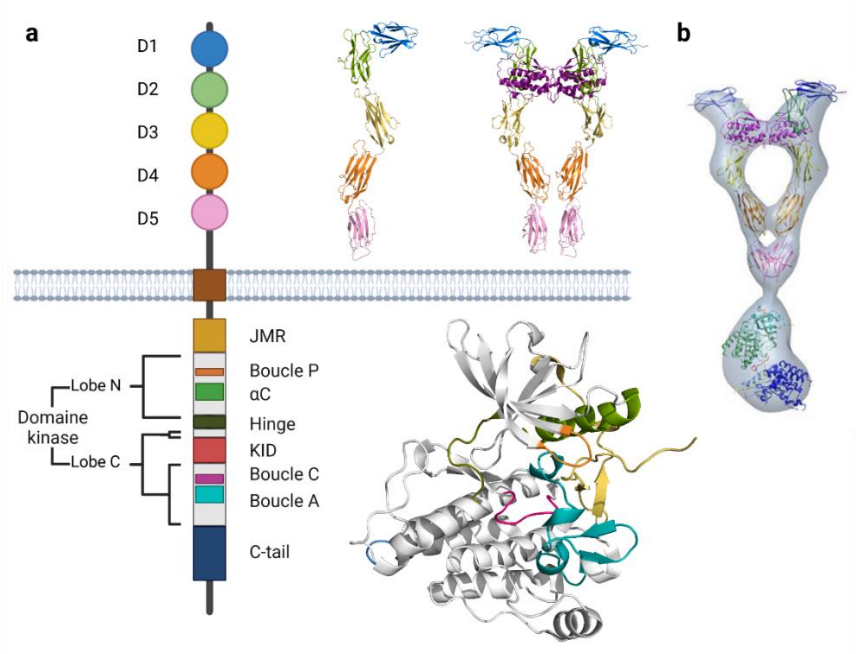


Figure 2.1 Architecture du RTK KIT (à gauche) et la structure partielle de ses domaines (à droite) (**a**) utilisés pour reconstruire la structure du dimère à partir de données de cryomicroscopie électronique (**b**).

Le **domaine extracellulaire** (ED) du récepteur KIT est constitué de domaines à motifs immunoglobulines (Ig). Ces domaines *Ig-like* sont en particulier rencontrés dans le système immunitaire où ils possèdent une architecture similaire dans les autres RTKs[115]. Structurellement, chaque domaine *Ig-like* est constitué d'un sandwich de deux feuillets β antiparallèles structurés en clé grecque. Ces feuillets β sont maintenus face à face par un pont disulfure. L'ED de KIT (Q26-P524) possède cinq de ces domaines *Ig-like* (D1 à D5) connectés par de courtes boucles et sur lesquels dix asparagines sont les sites de glycosylation par le N-acétyl-D-glucosamine (voir les structures de la *Protein Data Bank*[116], PDB ID : 2EC8, 2E9W[117]).

Le **domaine transmembranaire** (TMD) du KIT monomérique est représenté par une unique hélice α hydrophobe de dix-huit acides aminés (L525-L543), connectant l'ED avec le domaine cytoplasmique. Pour KIT, il n'existe pas de données empiriques sur la structure et l'orientation de cette hélice dans la membrane. Cependant, des conformations résolues par RMN pour la forme dimérisée du récepteur PDGFR-β (de la même famille que KIT) existent (PDB ID : 2L6W[118]).

Le **domaine cytoplasmique** (CD) de KIT (T544-V976) est constitué de quatre grands sous-domaines fonctionnels et structuraux : la région juxtamembranaire (JMR), le domaine kinase (KD), le *kinase insert domain* (KID) et le domaine C-terminal (C-ter). *Le **JMR*** est une longue boucle de 40 acides aminés (T544-K581) située entre l'hélice transmembranaire et le domaine kinase. Il est divisé en quatre sous-segments : JM-proximal (JM-P, T544-D552), JM-binding (JM-B, Y553-V559), JM-switch (JM-S, V560-I571) et JM-zipper (JM-Z, D572-K581). Cette région possède quatre tyrosines phosphorylables : Y568 et Y570 observées *in vivo*, et Y547 et Y553 observée *in vitro*[65,119]. Selon les données structurales, le ***domaine kinase*** de KIT (W582-S688, L769-S931) adopte un repliement commun aux protéines à activité tyrosine et sérine/thréonine kinase. Deux lobes N- et C-terminaux (lobes N et C), entre lesquels se situe le site actif, composent le domaine kinase. Les deux lobes sont connectés par une boucle d'environ dix acides aminés *(hinge)* permettant le déplacement relatif de ces deux lobes. Ce déplacement est critique pour la catalyse enzymatique en organisant spatialement les résidus catalytiques dans un arrangement conservé pour l'hydrolyse et le transfert du phosphate γ de l'ATP. Le ***lobe N*** (W582-L678) est composé de cinq feuillets β antiparallèles (notés β1 and β5, où β3 porte le résidu conservé K623) et d'une hélice α (αC). La boucle de liaison du phosphate du lobe N ou ***boucle P*** (G598-G601) est un coude riche en glycine et possédant le motif GxGxΦG (Φ = F/Y, F dans le cas de KIT). Elle participe à la coordination des phosphates de l'ATP/ADP dans le site catalytique. L'***hélice catalytique*** du lobe N ou αC (L631-G648), située entre le site catalytique et JMR, contient le résidu conservé E640. Le ***hinge*** (T670-L678) est une boucle de jonction entre les deux lobes et contribue à la stabilisation de l'adénine de l'ATP/ADP dans le KIT actif. Le ***lobe C*** (L679-S688, L769-S931) est principalement hélicoïdal avec six hélices α (αD-αI et αEF). Il contient les résidus catalytiques nécessaires au clivage de l'ATP et au transfert de son phosphate γ, les domaines de régulation de l'activité du récepteur et la phosphotyrosine Y900[120]. La ***boucle***

catalytique ou boucle C (H790-N797), est localisée à proximité du site actif et contient le résidu conservé D792 ainsi que la base catalytique N797. Elle permet, entre autres, l'orientation de l'ATP et de son cofacteur magnésium (Mg^{2+}) dans une position favorable au transfert du phosphate. La ***boucle d'activation*** ou boucle A est une longue boucle très flexible possédant deux tripeptides conservés (les motifs DFG et APE respectivement à ses extrémités N- et C-terminales), ainsi que la phosphotyrosine Y823. Le ***kinase insert domain*** (KID) (F689-D768) et le **domaine C-terminal** (C-tail) sont des domaines dont les structures n'ont pas été résolues expérimentalement, à la fois pour KIT mais aussi pour les RTKs les possédant. Le KID est la plateforme majeure de phosphorylation et d'interactions du KIT avec ses protéines partenaires de la signalisation cellulaire. En effet, il contient cinq sites de phosphorylation dont trois sites tyrosines (Y703, Y721, Y730) et deux sites sérines alternatifs (S741 et S746)[120]. La fonction d'une dernière tyrosine, Y747, reste à déterminer. Le C-tail (T932-V976) est une boucle non structurée porteuse de la phosphotyrosine Y936[120]. Cette description structurale détaillée du CD se repose sur de nombreuses études cristallographiques pour le CD clivé aux états inactifs et actifs (PDB IDs : 1T45[121] et 1PKG[122]). Toutefois, dans chaque structure empirique, deux tiers du JMR, le KID complet et C-tail restent non résolus.

Le RTK KIT est l'exemple typique d'une protéine topologiquement modulaire par la localisation extra- et intracellulaire de ses domaines, leurs structures et leurs fonctions. L'absence de données empiriques pour une grande partie de la protéine, mais surtout pour certaines régions du CD est un indicateur de leur désordre intrinsèque.

2.1.2. Structure du mutant D816V du KIT

Parmi les nombreux mutants du KIT, la plus étudiée et discutée dans la littérature est la mutation D816V, KIT^{D816V} (plus de 3000 publications, selon le catalogue des mutations somatiques *Cosmic v. 97*[123], au 27 août 2020). Cette mutation localisée dans la boucle A présente un intérêt clinique important par les pathologies qu'elle provoque (mastocytoses hématopoïétiques, leucémies[124]), et est souvent responsable de la résistance du récepteur KIT à l'imatinib[125]. Bien que la structure de KIT^{D816V} ne soit pas résolue empiriquement, une structure cristallographique du KIT^{D816H} a montré un réarrangement conformationnel de l'état d'auto-inhibition de JMR, suggérant son impact sur le phénomène d'auto-inhibition et l'activité kinase[126] (**Figure 2.2**).

La modélisation et l'étude par dynamique moléculaire (MD) d'une structure partielle du mutant KIT^{D816V} montrent ce même réarrangement structural et conformationnel ainsi qu'une altération locale en aval de la position mutée par le dépliement d'une hélice 3_{10} (I817-N819)[70,127] (**Figure 2.2**).

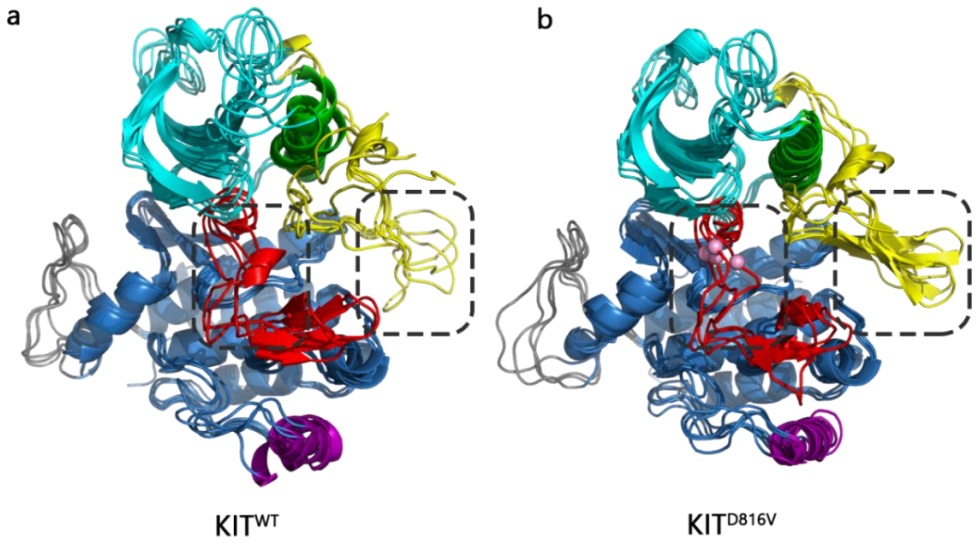


Figure 2.2 Réarrangements structuraux de JMR (jaune) et de la boucle A (rouge) dans KIT^{WT}(**a**) et KIT^{D816V} (**b**). La mutation D816V est représentée en rose sur la boucle A. Figure adaptée de [108].

2.2. MECANISMES D'ACTIVATION

La plupart des RTKs (sauf le récepteur à l'insuline) existent en tant que monomères ancrés dans la membrane plasmique et dont la fixation de leurs ligands spécifiques entraine dimérisation et activation[65]. Bien que certains de ces récepteurs peuvent retourner au noyau[128], nous nous sommes plutôt intéressés à leur localisation et mécanismes d'activations canoniques.

2.2.1. ACTIVATION DU RTK KIT SAUVAGE, KIT^{WT}

L'activation du KIT et son mécanisme reposent sur la coopération de plusieurs acteurs. A la surface de la cellule, le KIT inactif existe en forme monomérique exposant le site de liaison à son ligand spécifique, le *stem-cell factor* (SCF).

La liaison du SCF dans les boucles reliant les domaines *Ig-like* D1, D2 et D3 de l'ED de l'un de ces monomères entraine l'interaction simultanée d'un second monomère du KIT par leurs domaines Ig D1, D2 et D3. La réorientation des domaines D4 et D5 à la suite de cette interaction stabilise la formation d'un complexe homodimérique par complémentarité de surface[117]. Ainsi, KIT combine à la fois une dimérisation médiée par la fixation du SCF mais aussi par les monomères eux-mêmes.

La dimérisation du RTK KIT rapproche les TMD des deux monomères. Le positionnement relatif des deux hélices transmembranaires du dimère est inconnu pour KIT mais est caractérisé par RMN pour PDGFR-β, un autre membre de sa famille[118]. Ces données montrent un croisement en « X » des deux hélices permettant

la réorientation relative des CD de deux monomères et leur rapprochement[118,129]. Cette configuration particulière en « X » serait compatible avec une activation du domaine kinase[130].

Bien que possédant un JMR incomplet, les structures du CD du KIT sauvage aux états inactif et actif résolues par cristallographie aux rayons X[121,122] permettent d'apprécier les réarrangements structuraux observés dans le CD à l'activation, de postuler un mécanisme d'activation et le rôle de JMR dans ce processus. Dans l'état inactif, JMR entre en interactions non covalentes fortes avec les lobes du domaine kinase, forçant une configuration de la boucle A où la chaîne latérale de F811 empêche l'activité kinase en encombrant le site actif *(DFG-out)*[122]. Dans l'état actif, le détachement du JMR du domaine kinase crée les conditions nécessaires au départ de F811 de sa position bloquante *(DFG-in)*. Ainsi, JMR joue un rôle majeur dans l'auto-inhibition et l'activation du KIT.

Le domaine kinase du KIT catalyse une réaction de clivage et de transfert du phosphate γ d'une molécule d'ATP vers le groupement hydroxyle (–OH) d'une de ses propres tyrosines (autophosphorylation) et la production d'une molécule d'ADP et d'eau. Pour produire ce groupement phosphate et le transférer, le domaine kinase doit présenter un état capable de lier l'ATP et de réaliser son clivage. Un transfert optimal du phosphate requiert l'arrangement spatial précis de plusieurs résidus catalytiques conservés dans toutes les protéines kinases : K623 du motif VAVK situé sur le brin β3 du lobe N, E640 de l'hélice αC, les résidus D792 et N797 de la boucle C (motif HRDLAARN) et D810 du motif DFG, en N-terminal de la boucle A.

L'ordre dans lequel les évènements d'autophosphorylation croisée des huit phosphotyrosines portées par KIT n'est pas aléatoire et il a été montré que Y568 et Y570 situés sur JMR sont les premiers résidus phosphorylés[119]. Contrairement à sa localisation dans la boucle A, Y823 n'est probablement pas nécessaire à l'activation du KIT sauvage, mais joue un rôle clé dans la régulation de son signal pour la prolifération et la survie de la cellule[119].

La protéine ainsi active dispose des sites de reconnaissance et de recrutement de protéines partenaires des cascades de signalisation (voir section 2.3.1). La cellule pourra alors répondre en adéquation à un signal extérieur (**Figure 2.3**).

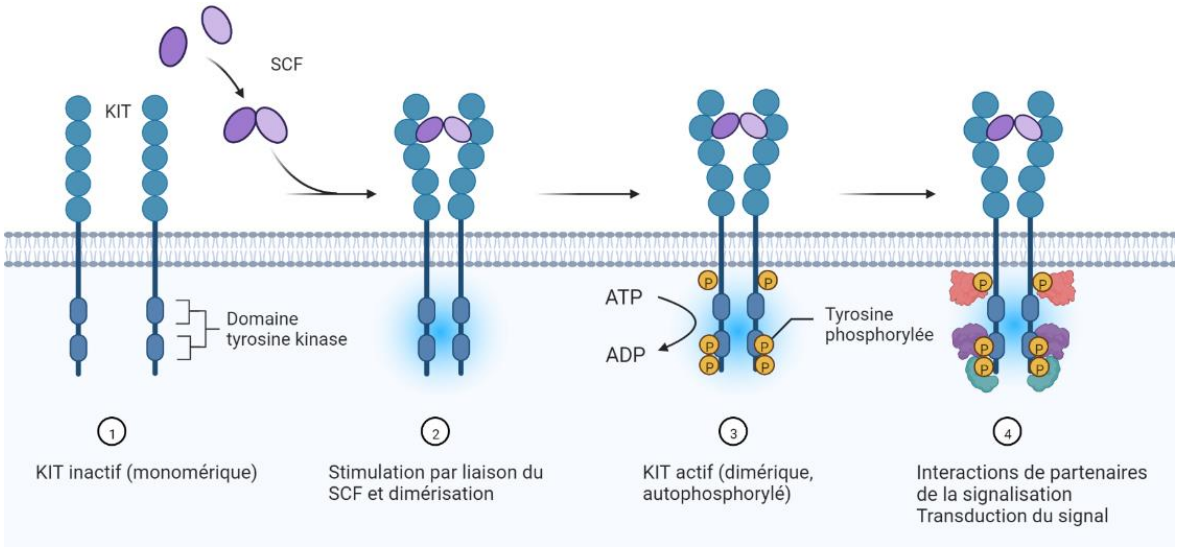


Figure 2.3 Schéma simplifié du mécanisme d'activation du RTK KIT. La fixation du SCF provoque la dimérisation du récepteur, l'activation de la fonction kinase, l'autophosphorylation de l'homodimère, le recrutement de protéines partenaires et enfin la transduction du signal.

2.2.2. ACTIVATION CONSTITUTIVE DU MUTANT ONCOGENE KIT^{D816V}

La substitution de D (chargé négativement) par V (hydrophobe) à la position 816 (boucle A), est responsable de l'activation constitutive du récepteur[113]. L'impact de cette mutation sur le CD du KIT^{D816V} au niveau structural a été étudié et comparé à sa forme sauvage par modélisation et simulation de dynamique moléculaire[70].

Cette étude a permis d'observer dans KIT^{D816V} plusieurs effets, à savoir : le réarrangement structural de JMR et le détachement du segment JM Switch du lobe C vers le solvant, la diminution de leur énergie de liaison, l'altération de la structure d'une partie de la boucle A et la disparition du couplage entre JMR et la boucle A observée dans KIT^{WT}. Ces effets ont été interprétés sous l'angle des voies de communication[70,108]. Ainsi, le paysage énergétique serait modifié vers une majorité de conformations dont l'auto-inhibition est levée et JMR, désordonné, est en contact avec le solvant, à l'instar de la forme active du KIT^{WT} [126].

Ainsi, la mutation gain de fonction D816V promouvrait le détachement spontané de JMR du reste du domaine kinase, libérant le site actif et permettant l'hydrolyse de l'ATP indépendamment de la stimulation du SCF[127]. Cette activation incontrôlée par le SCF du récepteur procèderait avant même son assignation à la membrane et sa dimérisation[131]. Dans son équivalent murin, cette mutation, D814V, permettrait au récepteur d'agir comme un dimère dont l'interface de dimérisation ne se situerait pas dans l'ED.

Le KIT^{D816V} peut alors recruter de nombreux partenaires alternatifs de l'initialisation des voies de signalisation (voir section 2.3.2).

2.3. CONTROLE DES VOIES DE SIGNALISATION PAR LE RTK KIT

2.3.1. SIGNALISATION DU KIT^{WT}

En tant que RTK, les fonctions principales du KIT sont d'assurer le contrôle de la survie, la prolifération, la croissance, la migration et l'adhésion cellulaire. En réponse à la stimulation du récepteur par son ligand spécifique, le SCF, le KIT déclenche de nombreuses voies de signalisation par le recrutement de protéines partenaires à ses sites de reconnaissance sélectifs contenant des phosphotyrosines spécifiques (voir section 2.2.1). *In fine*, le KIT activé effectue la régulation positive ou négative de l'expression de gènes particuliers. Le recrutement au KIT de protéines de l'initiation de la signalisation cellulaire peut s'effectuer de manière directe ou indirecte *via* des protéines adaptatrices.

Nous allons tout d'abord voir quelles sont les protéines partenaires du KIT, et quelles sont leurs spécificités pour les régions phosphorylées reconnues, puis répertorier les voies activées par ces protéines identifiées dans la littérature.

Activation des fonctions principales du KIT. À l'activation du KIT, ses résidus Y568 et Y570 (JMR) sont phosphorylés et deviennent des plateformes pour le recrutement des SFK (*Src family kinases* Src, Fyn, Lyn)[132–137], *Csk homologous kinase* (CHK)[138] et du complexe *Growth Factor Receptor Bound Protein 2 / Son Of Sevenless* (Grb2/SOS) seul ou complexé avec la *GRB2-associated-binding protein 2* (GAB2)[136], *Adapter Protein with a PH and SH2 domain* (APS)[139] ou Shc[140,141]. Dans le domaine insert kinase (KID), Grb2/SOS s'associe à pY703 (KID)[140], le complexe *phosphatidylinositol 3-kinase* (PI3K, composé de la sous-unité régulatrice p85 et catalytique p110) s'engage à KIT *via* p85 à pY721 (KID)[142], la *phospholipase C gamma* (PLCγ) et *phospholipase* D (PLD) sont recrutés à pY730 (KID)[136]. La protéine p85 et le complexe Crk/p85/$p120^{cbl}$ sollicitent pY900 (lobe C) après sa phosphorylation par Src phosphorylé[143]. Enfin, pY936 (C-tail) est le site de reconnaissance de Grb2/SOS, de Grb7 et Grb10[140].

KIT^{WT} contrôle 4 principales voies de signalisation régulant la survie, la prolifération, la croissance, la migration et l'adhésion cellulaire : les voies des *Mitogen-activated protein kinases* (MAPK)[144], PI3K/Akt[145], *c-Jun N-terminal kinase* (JNK)[137(pp. 3-),146] et *janus kinase / signal transducer and activator of transcription* (JAK/STAT)[147] (**Figure 2.4**, **a**).

Désactivation des fonctions du KIT. L'activité et les évènements de transduction du signal du KIT sont fermement régulés par ses effecteurs – SCF (activation), ATP et Mg^{2+} (réaction enzymatique) et phosphorylation (signalisation). Au récepteur, la régulation négative du KIT est assurée à trois niveaux : (i) son détachement de la surface cellulaire et sa dégradation intracellulaire, (ii) l'inhibition de l'activité kinase, (iii) la déphosphorylation de phosphotyrosines.

Tout d'abord, le **détachement de la surface cellulaire** est favorisé par la *protein kinase C* (PKC) et permet le clivage du CD au niveau du TMD[148,149]. Cette méthode permet la dissociation du dimère sur la membrane. Le recrutement direct ou indirect à KIT d'ubiquitines ligases (c-Clb, SLAP) favorise la **compartimentalisation et la dégradation** du récepteur : l'ubiquitine ligase c-Cbl semble jouer un rôle majeur dans la dégradation du KIT. Elle est recrutée directement à pY568 (JMR) et pY936 (C-tail) et indirectement *via* APS à pY568 et pY900 (lobe C) et pY936 (C-tail)[139], Grb2 à pY568, pY703 (KID) et pY936[150], p85 à pY721 (KID)[143], Crk à pY900[143,151]. L'ubiquitine ligase *src-like adaptator protein* SLAP est également directement recrutée à pY568[152] ainsi que la *suppressor of cytokine signaling 6* SOCS6[153]. Le recrutement de la protéine adaptatrice LnK à pY568 et pY570 favorise également une régulation négative de l'activité du KIT[154]. L'inhibition de l'activité kinase est induite par le recrutement de la protéine kinase C aux sérines S741 et S746 (KID)[155]. Enfin le **recrutement de phosphatases** comme SHP-1 et SHP-2[156,157] ou les *inositol triphosphate 5' phosphatases* SHIP1 and SHIP2[158] assurent la déphosphorylation des résidus tyrosines. La boucle A du KIT semble jouer un rôle prépondérant dans le recrutement de SHP-1 et SHP-2 (**Figure 2.4**, **b**).

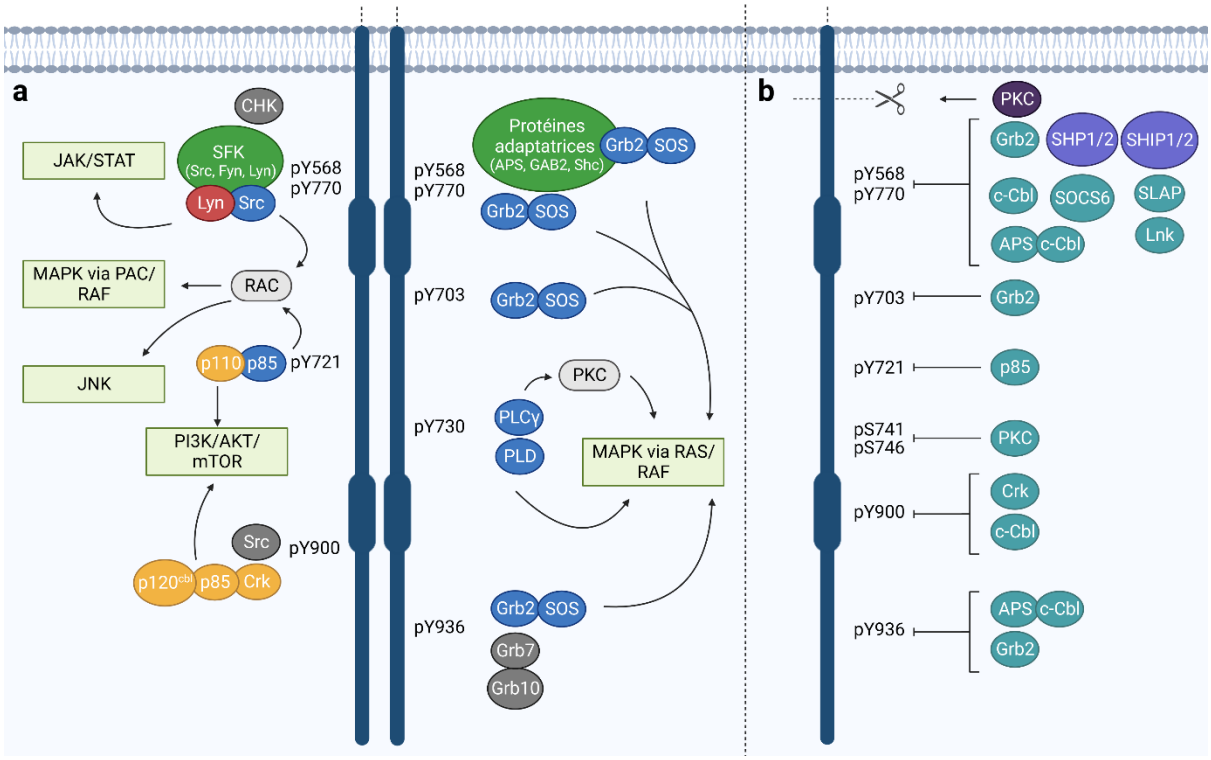


Figure 2.4 Interactome du RTK KIT^{WT}. (**a**) Voies de signalisation activées et représentées pour simplification sur l'un ou l'autre des monomères du KIT dimérisé. Les protéines sont colorées selon leur fonction et les voies qu'elles activent (vert : protéines adaptatrices ; rouge, jaune, bleu : activateurs des voies de signalisation ; gris clair : intermédiaires de voies de signalisation ; gris foncé : protéines dont l'activité est inconnue) (**b**) Voies de régulation négative de l'activité du KIT et représentées pour simplification sur un monomère du KIT dimérisé. Les protéines sont colorées selon leur fonction de régulation du KIT (violet foncé : détachement de la surface cellulaire ; violet clair : phosphatases ; turquoise : dégradation intracellulaire)

2.3.2. SIGNALISATION DES MUTANTS ONCOGENES DU KIT

Les sites de mutations oncogènes du KIT sont en majorité localisés et dans l'ED et le CD. Ils sont divisés en deux classes selon le mode d'activation de la protéine et sa réponse à la stimulation par le SCF[159].

Dans la **classe I**, les mutations du KIT ponctuelles (substitutions D419A, N505), insertions/délétions (délétion ΔY418,D419) ou encore duplications (Dup A502,Y503) sont exclusivement localisées dans le domaine extracellulaire. Les mutants sont assignés à la membrane plasmique et possèdent, à l'instar du KIT sauvage, une activité kinase, une internalisation et une dégradation dépendante du SCF même lorsque celui-ci est en faible concentration. Le niveau d'expression du récepteur à la surface de la cellule est dépendant de la mutation, mais chacune lui confère une longue demi-vie à sa surface. Dans la **classe II**, les mutants du KIT (T417I/ΔY418-419, V560D, D816V) sont localisés dans des compartiments cellulaires (réticulum endoplasmique, appareil de Golgi), sont constitutivement actifs, indépendants de la stimulation par le SCF et possèdent une demi-vie courte.

Parmi les nombreuses mutations du KIT, D816V est la plus étudiée. Parce qu'elle est responsable de pathologies graves et peut montrer une résistance à des inhibiteurs (ex. imatinib[125]).

Le mutant KIT^{D816V} diffère de sa forme sauvage tout d'abord par sa **localisation cellulaire**. Si le KIT est situé en surface de la cellule, KIT^{D816V} est en surface du réticulum endoplasmique (tumeurs stromales gastro-intestinales, GIST) ou l'appareil de Golgi (mastocytoses)[131,160]. Une faible quantité du récepteur muté est assignée à la membrane où il exerce une influence sur le chimiotactisme[161], mais est rapidement internalisé par endocytose[162]. Il a été suggéré que les caractéristiques structurales et dynamiques de KIT^{D816V} sont responsables de sa rétention dans des compartiments cellulaires, en particulier l'appareil de Golgi[163].

Activation des fonctions principales de KIT^{D816V}. Le KIT^{D816V} permet le recrutement de partenaires alternatifs à la forme sauvage du récepteur. Seules la présence ou l'absence d'interactions ont été caractérisées expérimentalement, et inévitablement, les mécanismes de reconnaissances de ces partenaires ne se sont pas encore élucidés. Tout d'abord, KIT^{D816V} contourne l'ancrage de ses SFKs sauvages (Src et Fyn) à pY568 et pY570 par les deux protéines kinases Syn et Syk[158,164] et interagit directement avec STAT3 et STAT5 à ces positions[158]. Bien que les sites d'interactions soient inconnus, KIT^{D816V} réalise une interaction directe avec la sous-unité catalytique p110 de PI3K[165], STAT1, FES et CrkL[158]. Contournant le recrutement de p85, il réalise également l'activation constitutive des protéines JNK, c-Cbl et Shc[159]. Les voies de signalisation activées par KIT^{D816V} sont quasi similaires à sa forme sauvage, notamment les voies PI3K/Akt, STAT, MAPK et JNK par CrkL[166]. Cependant, les protéines recrutées pour leur activation (partenaires directs de KIT^{D816V}) sont différentes. De plus, KIT^{D816V} active en

aval la protéine kinase FES impliquée dans la prolifération cellulaire[167]. *In fine*, cette signalisation permet l'augmentation de l'expression de gènes associés à la voie des MAPK et protéines médiatrices de la régulation de l'apoptose caspase-dépendante[168,169].

Désactivation des fonctions de KIT^{D816V}. Elle est attribuée à l'inhibition de quelques-uns de ses mécanismes de dégradation. KIT^{D816V} est responsable de la dégradation de sa phosphatase SHP-1, SHIP-1 et SHIP-2[158,170] et la phosphorylation directe de la protéine adaptatrice SLAP empêchant le recrutement de l'ubiquitine ligase c-Cbl[152]. Cette activation de SLAP serait à l'origine de la faible localisation en surface de KIT^{D816V} [152].

Agressivité oncogène de KIT^{D816V}. L'oncogénicité forte de ce mutant du KIT serait contrôlée par la phosphorylation de Y568 et Y570 (JMR)[158]. Par ailleurs, le recrutement PI3K et la tyrosine phosphorylée pY823 seraient essentiels à sa capacité oncogène[171,172].

Les nombreuses mutations pathogènes du KIT sont majoritairement localisées dans le CD et provoquent, son activité aberrante et l'activation de cascades de signalisation en partie altératives et indépendantes du SCF. De nombreuses études publiées dans la littérature ont montré que ces mutations sont responsables ou induites dans des pathologies graves. Ainsi, KIT est une cible parfaite pour le développement de modulateurs spécifiques de son activité et de ses fonctions dans la signalisation cellulaire.

2.4. KIT, UNE CIBLE POUR LE DEVELOPPEMENT D'INHIBITEURS

La présence de ces nombreuses mutations et leur pathogénicité font de KIT une cible préférentielle pour l'inhibition de son activité souvent rendue aberrante.

L'activation du KIT et son autophosphorylation déclenchent l'activité kinase du récepteur. De fait, le site actif est à l'origine la cible privilégiée pour le développement de **petites molécules inhibitrices compétitives** de l'ATP. Ces inhibiteurs peuvent reconnaitre sélectivement la forme inactive ou active de la protéine[173]. Cependant, ils sont peu sélectifs de KIT et ciblent également d'autres kinases, entrainant des effets secondaires importants[174] et/ou des résistances[175–177]. Ces résistances sont parfois contournées par la combinaison d'inhibiteurs, mais ces effets restent dépendants du patient[178]. Des **peptides inhibiteurs compétitifs** de l'interface de reconnaissance de KIT avec ses partenaires (SFK, JAK2) ainsi que des **inhibiteurs allostériques** de kinases en général sont également en cours de développement[179,180], tout comme des **inhibiteurs covalents** de l'ATP[181].

Le développement de nouveaux inhibiteurs sélectifs de KIT visant des sites

alternatifs non caractérisés structuralement ou le développement de modulateurs d'interactions du KIT avec ses partenaires de la signalisation nécessitent de posséder une structure complète du CD. Pour cela, il est indispensable de modéliser les domaines du KIT ou leurs fragments non résolus empiriquement, mais également les protéines partenaires du KIT^{WT} et le KIT^{D816V}.

2.5. PREMIERS PAS VERS LA STRUCTURE COMPLETE DU DOMAINE CYTOPLASMIQUE DU KIT

En tant que domaine d'initiation de la signalisation cellulaire, porteur principal de mutations pathogènes et en tant que cible historique de la modulation de l'activité de la protéine, le CD de KIT est d'une importante décisive. Le CD contient probablement 4 IDRs (JMR, KID, boucle A et C-tail), toutes contenant de sites de phosphorylation. De plus, JMR, KID et C-tail servent à la fois de plateforme d'initiation des cascades de signalisation, mais aussi de régions cibles à fort potentiel pour le développement de molécules thérapeutiques spécifiques du KIT. Malencontreusement, cette étude est impossible, car ces fragments sont partiellement (JMR) ou complètement absents (KID, C-tail) des structures du KIT résolues empiriquement.

Pour cela, il est indispensable de modéliser la structure complète du CD de KIT^{WT} en intégrant à la structure cristallographique du domaine kinase partie ces fragments manquant. L'absence de séquence homologue a obligé la génération et l'incorporation de modèles *ab initio* de ces fragments. Le modèle *de novo* du CD du KIT^{WT} en forme inactive a été réalisé en plusieurs étapes. Tout d'abord par la modélisation *ab initio* du KID (Isaure Chauvot de Beauchêne, 2013[182]) puis l'ajout du modèle le plus probable et de C-tail au CD (François Inizan, 2016[183], **Figure 2.5**). La caractérisation par simulation de dynamique moléculaire (DM) du CD seul en solution a permis d'observer la diversité conformationnelle du KIT^{WT} majoritairement portée par les fragments JMR (toujours partiel) KID et C-terminal, ainsi qu'une transition globale conformationnelle du KID laissant penser à un caractère désordonné de ce domaine. Malheureusement, ces simulations de DM ne reflètent pas l'environnement natif du KIT^{WT}, normalement intégré dans la membrane et il est évident que les grands déplacements du JMR (simulé en N-terminal) sont des artéfacts dus à sa liberté artificielle dans l'eau.

Par conséquent, l'étape suivante est la modélisation du CD complet du KIT par ajout du fragment manquant de JMR et l'intégration du KIT dans la membrane par le TMD. Ce modèle permettra d'également modéliser le CD de KIT^{D816V}, le caractériser et le comparer à KIT^{WT}. Une attention particulière sera portée sur les fragments supposés désordonnés du KIT (JMR, KID, boucle A et C-tail). Cette modélisation permettant étudier la modularité du récepteur, son DYNASOME, les effets des phosphorylations, et utiliser la protéine phosphorylée pour l'étude des interactions avec ses partenaires de signalisation.

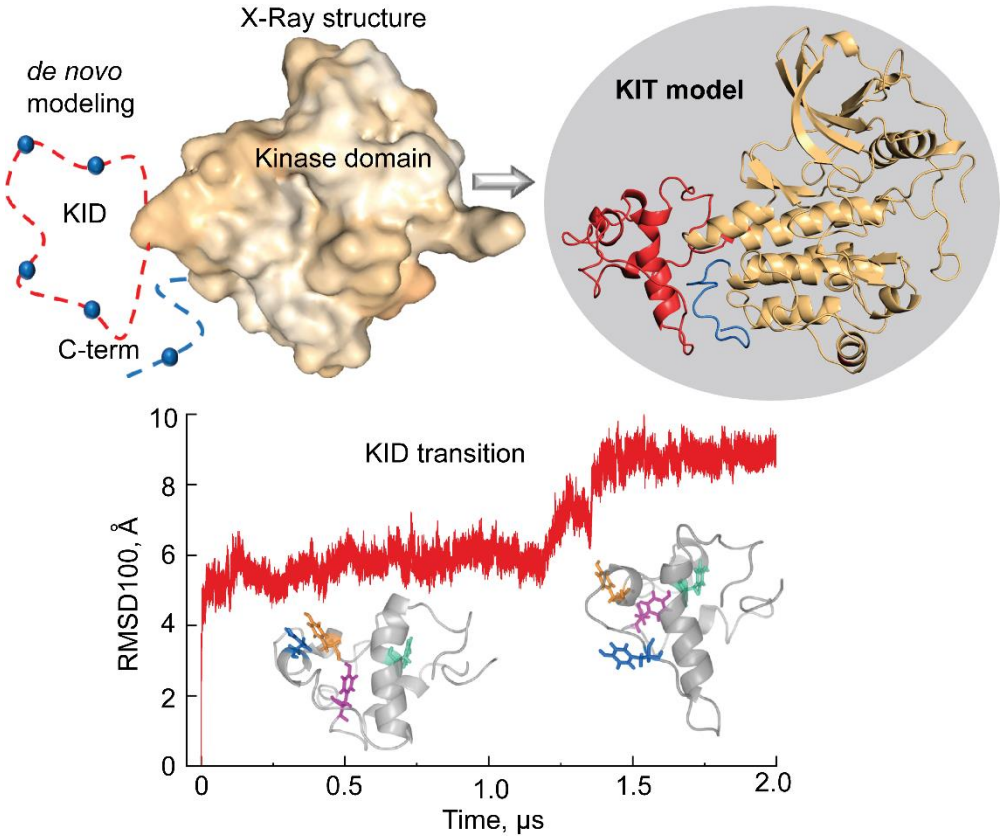


Figure 2.5 Modèle du domaine cytoplasmique du KIT étudié par simulation de dynamique moléculaire. Le modèle a été réalisée par l'ajout des modèles *de novo* du KID (en rouge) et C-tail (en bleu) à la structure cristallographique du domaine kinase de KIT (en surface) (**haut**). Transition conformationnelle vue pendant la simulation de dynamique moléculaire (**bas**). Abstract graphique de la publication [183]).

CHAPITRE 3. LA VITAMINE K EPOXYDE REDUCTASE : UNE ENZYME TRANSFORMANTE DE LA VITAMINE K

La vitamine K époxyde réductase (VKOR) est une protéine transmembranaire du réticulum endoplasmique impliquée dans la γ-carboxylation de glutamates des protéines dépendantes de la vitamine K (VKDPs)[184,185]. Ces protéines sont engagées dans de nombreux processus cellulaires comme l'hémostase, la calcification des os, l'homéostase du calcium, la transduction du signal et le stress oxydatif[186–190]. La vitamine K et les VKPDs sont également des cibles thérapeutiques dans le traitement et la prévention de cancers et maladies inflammatoires[189,191–193].

L'utilisation physiologique de la vitamine K repose sur un cycle d'oxydoréduction. Lors de ce processus, la vitamine K hydroquinone ($vitK^H$) est oxydée par la γ-glutamyl-carboxylase (GGCX), en vitamine K-2,3-époxyde ($vitK^E$). Pour servir à nouveau de substrat à la GGCX, la $vitK^E$ doit être réduite en $vitK^H$. Cette réduction est catalysée par VKOR.

Les mutations du VKOR humain (**hVKORC1**, *human VKOR complex 1*) sont responsables de la variation de la réponse des patients sous traitement aux anticoagulants de type antivitamine K (AVKs). Certaines mutations sont incriminées dans des phénomènes de résistance au traitement[194,195], dans une hypersensibilité de l'enzyme[196] et dans une modification de son activité[197]. Ainsi, la réponse physiologique aux AVKs est fortement dépendante du patient.

Malgré de nombreuses études biochimiques et biophysiques produisant des résultats fréquemment contradictoires[198–200], l'absence de structure du hVKORC1 rend difficile la compréhension détaillée de son activation et du mécanisme catalytique de réduction de la vitamine K, et par conséquent limite le développement d'inhibiteurs

Dans ce chapitre, on décrira le rôle de la vitamine K, ses isoformes et son cycle physiologique. Puis on se concentrera sur hVKORC1, sa structure probable, son cycle enzymatique et son potentiel comme cible pour le développement de nouveaux AVKs et enfin la difficulté de telles études.

3.1. LA VITAMINE K ET SON CYCLE

La vitamine K est importante pour de nombreux processus cellulaires en tant que cofacteur de la γ-glutamyl-carboxylase (GGCX), enzyme catalysant la γ-glutamyl carboxylation de résidus glutamates[201]. Cette modification post-traductionnelle est indispensable à l'activation des protéines dépendantes de la vitamine K (VKDPs). Parmi les 17 VKDPs connues[202] les protéines engagées dans la physiologie sanguine sont

les plus nombreuses, en particulier les facteurs II, VII, IX et X, la prothrombine, les protéines Z, S, C[202] impliquées dans les cascades de la coagulation. Récemment, il a été rapporté que le taux de vitamine K semble être extrêmement réduit dans les poumons des patients hospitalisés dans le cadre de l'épidémie de COVID-19, en particulier ceux sous ventilation mécanique ou décédés[203]. Leur taux de protéine S libre dans le sang serait diminué[204] et compatible avec la thrombogénicité associée à l'infection[205].

L'activation des facteurs de coagulation susmentionnés nécessite une γ-carboxylation d'un glutamate par la γ-glutamyl-carboxylase (GGCX) dont la catalyse est effectuée dans les hépatocytes, ostéoblastes ou muscles lisses[201]. La vitamine K possède une fonction primordiale dans l'hémostase : elle est indispensable aux réactions impliquant les facteurs de coagulation qui lui sont dépendantes (facteurs II, VII, IX et X)[206], elle est le cofacteur principal de la GGCX et elle est impliquée dans l'activation des protéines anticoagulantes C et S[201]. L'activation des cofacteurs dépendants de la vitamine K entraine son oxydation en vitamine K époxyde. Pour redevenir substrat de la GGCX, la vitamine K époxyde être réduite en vitamine K par hVKORC1.

La vitamine K est en réalité une famille de molécules liposolubles composée de la phylloquinone (K1), les ménaquinones (K2), ménadione (K3) et ménadiol (K4)[207]. La vitamine K1 est la forme prédominante de l'alimentation (légumes verts et plantes chlorophylliennes), les vitamines K2 sont présentes dans les aliments fermentés, les viandes et les produits animaux[208] et les vitamines K3 et K4 en sont des formes synthétiques[207]. La vitamine K apportée de l'alimentation est en majorité stockée dans le foie, lieu de synthèse des VKDPs[209].

La mise en évidence du rôle de la vitamine K1 dans l'activation de la GGCX a permis de postuler un cycle réactionnel dans lequel elle est réversiblement oxydée et réduite en des substrats fonctionnels.

Cycle de la vitamine K. La vitamine K existe naturellement sous une forme oxydée quinone (vitK^{Q}). Or, la γ-carboxylation nécessite sa forme réduite hydroquinone (vitK^{H}). Dans l'organisme, la quinone est réduite en hydroquinone par une quinone oxydoréductase (QR) dépendante du NADPH. À son tour, la vitK^{H} est convertie en vitamine K 2,3-époxyde (vitK^{E}) à la suite de la γ-carboxylation (**Figure 3.1**, **a**). Complétant le cycle, la vitK^{E} est réduite en vitK^{Q} initiale par hVKORC1 associée à des réducteurs de type groupement thiol[210] (**Figure 3.1**, **b**).

Mécanisme catalytique de réduction de la vitK^{E} en vitK^{Q} par hVKORC1. L'oxydoréduction d'une molécule implique un transfert réversible de protons en provenance d'une autre. Le mécanisme d'échange thiol-disulfure suggéré dans le cycle de la vitamine K concerne les deux cystéines d'un motif conservé CX_1X_2C[211] présent dans le site actif du hVKORC1[212,213]. Sous sa forme inactive, ces deux cystéines forment

un pont disulfure (S−S). À l'activation de l'enzyme, ce pont est réduit en deux groupements thiol (–SH) possédant une activité enzymatique (voir section 3.4). Le mécanisme d'actualité propose la protonation du groupement époxyde de vitK^{E}, l'attaque d'une cystéine de hVKORC1 et le réarrangement du proton, puis l'attaque de la seconde cystéine sur la vitK^{E}. Finalement, ce mécanisme produit la vitK^{Q} et une molécule d'eau[212]. Un mécanisme similaire est supporté par une étude théorique réalisée par mécanique quantique[213].

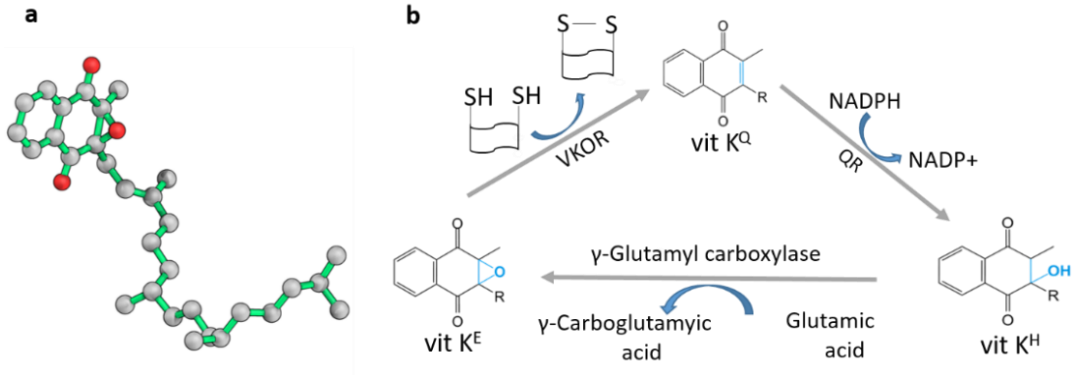


Figure 3.1 La vitamine K 2,3 époxide (**a**), et le cycle redox métabolique de la vitamine K (**b**).

3.2. STRUCTURE DU VKOR

La majorité des protéines membranaires sont difficiles à étudier en raison de leurs surfaces partiellement hydrophobes et de leur flexibilité. Pour ces motifs, la structure du VKOR humain a fait l'objet de nombreuses controverses.

3.3. ETAT DE L'ART

Le VKOR humain (hVKORC1, *human VKOR complex 1*) est une petite protéine (163 aas) transmembranaire du réticulum endoplasmique. Elle composée d'un domaine transmembranaire (TMD), d'une boucle luminale (boucle L) et de deux extrémités terminales (**Figure 3.2**). La localisation cellulaire de ces domaines permet de considérer hVKORC1 comme topologiquement modulaire. Deux domaines principaux, le TMD et la boucle L possèdent un couple de cystéines dont les chaînes latérales peuvent être soit réduites soit oxydées. Les cystéines du TMD sont contenues dans le motif catalytique CX_1X_2C (C132 et C135)[211]. En revanche, celles de la boucle L (C43 et C51) n'appartiennent à aucun motif conservé mais entrent dans le mécanisme d'activation de hVKORC1 par réaction d'échange thiol-disulfure avec sa thiorédoxine[214]. De plus les fragments et résidus R39-G46, R58-L65, L76, N77 de cette boucle sont strictement conservés dans les homologues mammifères de hVKORC1 et dont les séquences sont déposées et révisées dans Uniprot.

Malgré un nombre important de structures de protéines résolues empiriquement (61 330 structures de protéines humaines déposées dans la PDB, au 16 mars 2023), celle du hVKORC1 a longtemps été un mystère et la topologie de son TMD source de nombreux débats.

En sa qualité de protéine membranaire, l'ancrage du domaine transmembranaire (TMD) a longtemps fait l'objet de discussions. Différents modèles de topologie de ce domaine ont été avancés pour hVKORC1 : un modèle de TMD à 3 hélices (3H TM), 4 hélices (4H TM) et 5 hélices (5H TM)[198–200]. Chaque modèle proposé se base seulement sur un ou deux paramètres mesurés empiriquement.

La distribution des résidus chargés affectant l'orientation du TMD[215,216] ont initialement suggéré le modèle 3H TM[200]. L'étude biophysique de deux enzymes de mammifères, hVKORC1 et VKORC1L1 (*VKORC1 like protein 1*, un orthologue), soutient le modèle 4H TM comme le plus probable[199]. L'absence de structure empirique de hVKORC1 a favorisé le développement d'un modèle par homologie avec la structure du VKOR de *Synechococcus sp*. (bVKOR) possédant 5H TM (PDB ID : 4NV5[217]) (**Figure 3.2**). Cette structure cristallographique dispose également d'un domaine thiorédoxine (Trx-*like)* connecté de manière covalente au TMD. Une très faible similarité des séquences de bVKOR et hVKORC1 (36%) rend difficile la modélisation par homologie de hVKORC1. Plusieurs tentatives ont produit des topologies divergentes et controversées de TMD représentées par les modèles en 3H- 4H- et 5H TM, suggérés biochimiquement[200]. Par la longueur de séquence de hVKORC1 (163 aas) et le besoin de hVKORC1 d'être bien ancré dans la membrane, l'organisation 5H TM semblable à bVKOR (181 aas pour son TMD et boucle) nous a paru surréaliste.

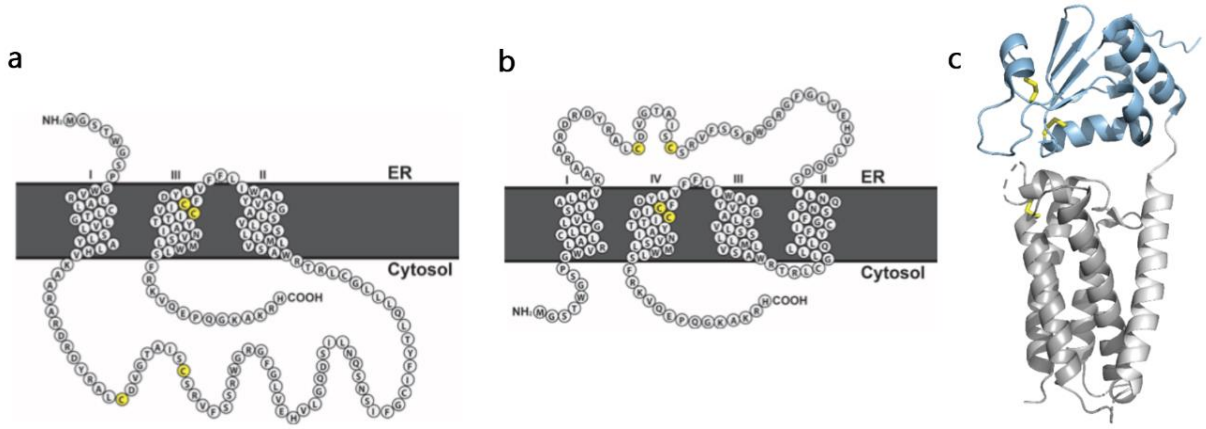


Figure 3.2 Topologies 3H- (**a**), 4H- (**b**) et 5H-TM (PDB ID : 4NV5[217]) (**c**) du domaine transmembranaire de VKOR proposés par des études biochimiques et structurales. Les figures (a) et (b) sont adaptées de [199]. Les cystéines fonctionnelles sont colorées en jaune.

3.4. MODELISATION D'UNE STRUCTURE DU HVKORC1 A QUATRE HELICES

En 2017, l'équipe BiMoDyM a réalisé la modélisation *in silico* du modèle 3D du hVKORC1 à topologie de TMD à 4 hélices[218].

Bien que l'alignement des séquences du VKOR de *Synechococcus sp.* (bVKOR, Uniprot : Q2JJF6) et de hVKORC1 (Uniprot : Q9BQB6) n'a montré qu'une faible similarité (36%), la prédiction des structures secondaires de hVKORC1 et de la topologie du TMD supportent le modèle 4H TM. Ces résultats ont permis la modélisation par homologie de ce domaine sur 4 hélices de la structure du bVKOR. En revanche, la boucle luminale (boucle L), inexistante dans bVKOR, a été modélisée *de novo* pour hVKORC1. Le modèle final du hVKORC1 est constitué d'un TMD à 4 hélices (TM1-TM4) et d'une longue boucle (boucle L) possédant également une courte hélice. L'ordre de ces régions est le suivant : TM1 (W10-A32), boucle L (R33-N77), TM2 (Q78-R98), TM3(A102-L124) et TM4 (C132-R151). Les deux cystéines C43 et C51 sont localisées dans la boucle L tandis que C_{132} et C_{135} se situent dans TM4. Ce modèle supporte l'hypothèse de la localisation luminale de la boucle L dans le réticulum endoplasmique. Les simulations de dynamique moléculaire (MD) du modèle de hVKORC1 dans son environnement natif (intégré dans une membrane) ont montré la stabilité du TMD et le désordre intrinsèque de la boucle L (**Figure 3.3**, **a**). Une étude cristallographique publiée en 2021 a confirmé l'exactitude de ce modèle *de novo*[218].

Dans le réticulum endoplasmique, hVKORC1 existe en deux états où les chaînes latérales de deux paires de cystéines, C43 et C51, C132 et C135, sont respectivement oxydées (pont disulfure, S−S, inactif) ou réduites (groupements thiol −SH, actif). En plus de ces deux états, hVKORC1 possède des états intermédiaires réactionnels (états métastables) apparaissant lors de l'activation enzymatique.

Une analyse détaillée des données de MD du modèle *de novo* a permis de prédire les différents états de hVKORC1 (inactif, actif et intermédiaires) contribuant au processus enzymatique (**Figure 3.3**, **b**).

L'affinité de la vitamine K et de quatre anticoagulants de type antivitamine K (AVKs) avec ces modèles ont été explorés *in silico* et validés empiriquement. En récapitulant ces résultats obtenus *in silico*, deux propositions du processus enzymatique de hVKORC1 ont été postulées. La première hypothèse, basée sur le mécanisme enzymatique de bVKOR, suggère l'implication directe d'un partenaire thiorédoxine dans une liaison covalente réversible avec hVKORC1[219] (**Figure 3.3**, **c**). Dans la seconde hypothèse, la thiorédoxine initie la réduction de hVKORC1 en tant que seul donneur de protons (**Figure 3.3**, **d**). Cette interprétation simple reflète le cycle catalytique complet et le plus rationnel du mécanisme enzymatique hVKORC1[220].

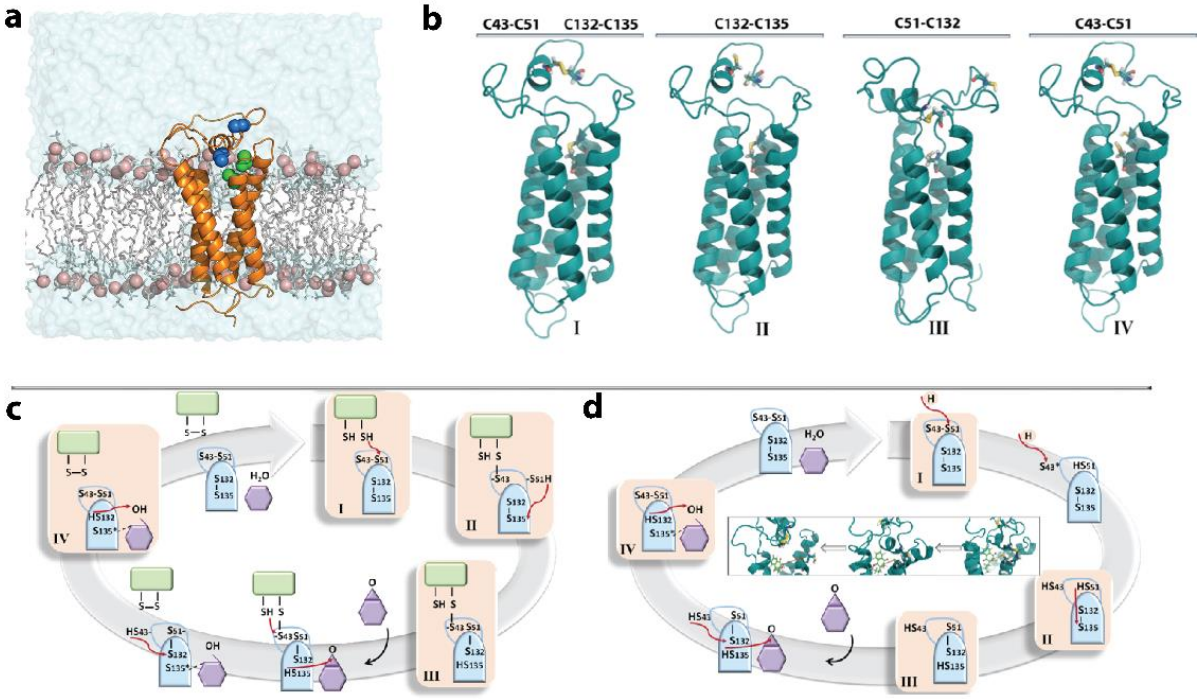


Figure 3.3 Simulation du modèle de hVKORC1 et ses états enzymatiques. Modèle de hVKORC1 intégré dans une membrane (**a**) dont la simulation a permis d'identifier quatre états enzymatiques du hVKORC1 (**b**). L'état inactif entièrement replié (I) et deux états actifs, liés à la vitK^{E} (époxyde) (II) et la vitK^{H} (hydroquinone) (III). (**c**, **d**) Cycles d'oxydoréduction de la vitamine K par hVKORC1 basés sur deux protéines (c) et une seule protéine (d) et proposés à partir des états décrits précédemment. Figure adaptée de [220].

3.5. ACTIVATION DE HVKORC1 PAR LA REACTION LA REACTION D'ECHANGE THIOL-DISULFURE : UNE ETAPE PEU ETUDIEE

Pour être activé et apte pour les réactions enzymatiques transformant la vitamine K, la réduction du pont disulfure C132-C135 du site actif du hVKORC1 est nécessaire. Pour disposer d'équivalents réducteurs, l'intervention d'une enzyme thiol-réductrice (thiorédoxine, Trx) est donc indispensable. La participation d'un partenaire dans l'activation du hVKORC1 est hautement probable bien que celui-ci soit encore inconnu. En effet, plusieurs intervenants ont été suggérés (le glutathion, le système thiorédoxine-NADPH et la protéine disulfure isomérase PDI[196,212]). Des études expérimentales ont supposé comme acteurs de l'activation de hVKORC1, TMX1 et TMX4 de la famille des *thioredoxin-related transmembrane proteins* ainsi que l'*endoplasmic resident protein 18* (Erp18). L'étude *in silico* réalisée par BiMoDyM[221], puis de récentes études empiriques[222,223] suggèrent plutôt PDI comme partenaire le plus probable du hVKORC1.

La coopération de hVKORC1 avec une protéine thiorédoxine implique un processus d'activation. Ce processus est l'une des étapes les moins étudiées dans la

physiologie de hVKORC1. Schématiquement, le mécanisme d'activation de hVKORC1 postule le transfert de protons et d'électrons de groupements thiol de la Trx vers deux cystéines oxydées dans la boucle L de hVKORC1. Puis, ces protons sont apportés au motif CX_1X_2C dans le TMD[224]. Une fois ce motif réduit, les protons sont une dernière fois transférés à la $vitK^E$ en oxydant le motif CX_1X_2C dans le site actif, réduisant la $vitK^E$ en $vitK^Q$ et libérant une molécule d'eau[211,225]. Pour réduire de manière répétée la $vitK^E$, hVKORC1 doit être régulièrement activée par un partenaire Trx délivrant des équivalents réducteurs *via* le même type de réaction d'échange thiol-disulfure (**Figure 3.4**).

Les réactions d'échanges thiol-disulfures sont au cœur du repliement oxydatif des protéines et un mécanisme clé dans presque toutes les enzymes générant et isomérisant les ponts disulfures[226]. Comprendre les mécanismes d'échanges thiol-disulfures reste un défi scientifique majeur.

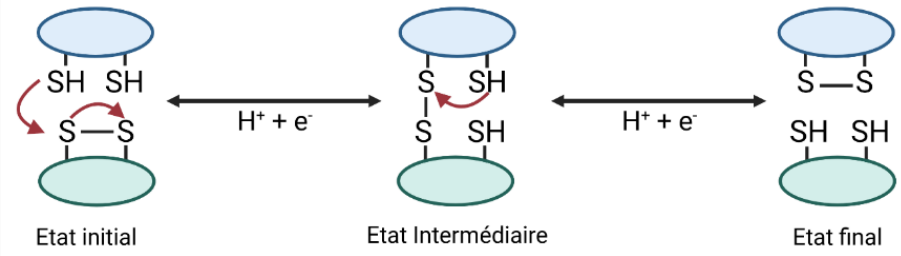


Figure 3.4 Réaction d'échange thiol-disulfure entre deux partenaires ou deux fragments intramoléculaires.

La réaction (attaque nucléophile) est initiée par le rapprochement colinéaire d'un thiolate attaquant à l'axe de la liaison disulfure (**Figure 3.4**). Cette exigence stérique est nécessaire aux interactions entre partenaires oxydoréducteurs bien structurés[227,228] et est satisfaite par des enzymes capables d'adapter leur repliement à chaque étape du procédé.

3.6. hVKORC1, UNE CIBLE POUR LE DEVELOPPEMENT D'ANTICOAGULANTS

La présence de mutations sur hVKORC1 rend possible la dérégulation de son processus enzymatique. Un polymorphisme génétique dans hVKORC1 est associé à des taux de recyclage de la vitamine K modifié provoquant, par exemple, hémorragies ou thrombose. De plus, ces variants génétiques de hVKORC1 entrainent une réponse variable des patients aux anticoagulants antivitamine K (AVKs)[194–197].

Une étude des 25 mutants hVKORC1 a montré que seuls 6 présentaient une résistance accrue aux AVKs et 10 entraînaient une perte quasi totale de l'activité *in*

vitro[197]. Ces mutations sont localisées soit dans le TMD, soit dans la boucle L. Les cystéines de la boucle L (C43 et C51) sont des résidus impliqués dans les deux cycles enzymatiques supposés de hVKORC1, rendant ce domaine particulièrement important dans le processus enzymatique et les mutations de hVKORC1 sont nombreuses et majoritairement localisées dans cette boucle[229]. En tant que protéine centrale dans le cycle de la vitamine K, hVKORC1 est une cible remarquable pour le développement de nouveaux AVKs.

Actuellement, les AVKs disponibles sont des dérivés coumariniques ou de l'indanedione[230]. Ces inhibiteurs sont compétitifs de la vitamine K dans le site actif de hVKORC1 et bloquent la réduction de la vitK^{E} en vitK^{Q} [231]. Cette inhibition diminue l'activité de la GGCX, limitant alors l'hémostase. Or, on l'a vu, les phénomènes de résistances sont souvent associés à des mutations de la boucle L[229].

Pour contourner ces résistances, la boucle L pourrait être une nouvelle cible dans le développement de nouveaux AVKs comme des modulateurs allostériques empêchant l'entrée de la vitamine K dans le site actif ou inhibant l'activation de l'enzyme par alternation de sa plasticité. De plus, l'exploration des interactions de hVKORC1 avec sa thiorédoxine pourra délivrer une autre cible : une interface d'interactions formée lors du processus de reconnaissance et d'attachement de la thiorédoxine à la boucle L. Les modulateurs de ces interactions pourront bloquer ou réguler l'activation de hVKORC1.

OBJECTIFS DE LA THESE ET CONTRIBUTIONS

En résumé, malgré de très grandes différences de fonctions, structures et propriétés biochimiques et biophysiques, le RTK KIT et hVKORC1 possèdent des traits très similaires. Ce sont des protéines membranaires, probablement composées de modules quasi indépendants contenant des domaines/fragments intrinsèquement désordonnés. Par conséquent, les problèmes liés à leur étude sont identiques et les objectifs finaux similaires, c'est-à-dire la reconstitution de leur DYNASOME et de leur INTERACTOME.

Les objectifs du travail de mon doctorat pour la thématique **RTK KIT** a été la modélisation du CD complet du KIT par ajout du fragment manquant de JMR et l'intégration du KIT dans la membrane par le TMD. Ce modèle sera une fondation pour la modélisation du CD de KIT^{D816V} et la caractérisation de deux protéines, KIT^{WT} et KIT^{D816V} avec une attention particulière sur les fragments supposés désordonnés du KIT, JMR, KID, boucle A et C-tail. Cette modélisation donnera la possibilité d'étudier la modularité du récepteur, les effets de phosphorylation, et son DYNASOME. En prenant en compte sa modularité structurale et fonctionnelle possible, les effets de phosphorylation seront étudiés premièrement sur KID possédant 3 phosphotyrosines fonctionnelles. La génération d'un complexe KIT / protéine partenaire (PP) a été planifiée pour définir le protocole optimisé pour la reconstitution de l'INTERACTOME du KIT (ensemble de ses complexes avec ces partenaires directs). Tout ce travail a été réalisé par moi-même lors de mon stage de Master 2 et mon doctorat. Par le monitorat de stage de Marina Botnari (étudiante en Master 2 *In silico Drug Design*, Université Paris-Cité, 2023), les nouvelles poches des KIT^{WT} et KIT^{D816V} ont été identifiées et caractérisées. Ce travail servira à postuler des POCKETOMES du KIT.

Les objectifs du travail sur la thématique **hVKORC1** consistent en la caractérisation de sa modularité, de son DYNASOME et l'identification de sa protéine *redox* (*Trx-like*) à partir de quatre candidats suggérés dans la littérature. Cette identification ouvrira la voie pour explorer son INTERACTOME complet composé des complexes Trx/hVKORC1 assemblés lors d'un échange thiol-disulfure : un complexe non covalent de « réactifs » ou « complexe précurseur », un complexe covalent intermédiaire caractérisant l'état transitoire, et un complexe non covalent de « produits » de l'échange thiol-disulfure ou « complexe successeur ». De tels modèles 3D des complexes Trx/hVKORC1 donneront les bases structurales à la description des mécanismes d'activation de hVKORC1.

Lors de mon doctorat, j'ai réalisé une partie du travail concernant la thématique hVKORC1 en collaboration avec Maxim Stolyarchuk (doctorant, 2019-2022), et par le monitorat de deux étudiants – Enki Bachelier (étudiant en Licence 2 Mathématiques-Sciences de la Vie, Université Paris-Saclay, 2021) et Marina Botnari. Tout particulièrement, j'ai réalisé l'identification du partenaire de hVKORC1 le plus probable

et sa caractérisation exhaustive. De plus, avec Enki Bachelier, j'ai effectué une étude approfondie des structures cristallographiques de hVKORC1 publiées en 2021 et leur exploration par simulation de dynamique moléculaire classique (cDM) et accélérée (GaDM). Maxim Stolyarchuk a modélisé les premiers complexes de ce partenaire avec hVKORC1 (complexes probables non covalents précurseurs et successeurs, et deux états intermédiaires métastables covalents) avec ma participation directe. Les nouvelles poches du hVKORC1^{WT} et quatre de ses mutants, en particulier localisées sur la boucle L, ont été identifiées et caractérisées avec Marina Botnari.

Pour les deux thématiques, RTK KIT et hVKORC1, j'ai, par ailleurs, contribué aux discussions de tous les résultats, la préparation et la réaction des articles, et la communication des résultats lors de congrès.

CHAPITRE 4. METHODES DE CARACTERISATION ET DE MODELISATION DE LA STRUCTURE ET DE LA DYNAMIQUE DU RTK KIT ET DE HVKORC1

Vu la similarité des protéines KIT et hVKORC1 et les objectifs d'étude, nous avons utilisé les mêmes informations initiales (séquence, structure ou modèles structuraux), les mêmes méthodes mathématiques d'exploration et des protocoles similaires. Pour faire preuve de concision, on ne s'intéressera qu'aux méthodes expérimentales qui ont permis la résolution des structures du RTK KIT, hVKORC1 et un partenaire du KIT, utilisés comme points de départ pour construire les modèles, et aux méthodes computationnelles pour la modélisation et l'analyse des données de dynamique moléculaire de ces systèmes.

4.1. METHODES EXPERIMENTALES

Pour l'étude de la structure des protéines, ont été appliquées pratiquement toutes les méthodes physiques et physicochimiques permettant d'obtenir des informations sur ces molécules à l'état solide et en solution. La plus grande quantité de données a été obtenue en utilisant l'analyse par diffraction des rayons X, les méthodes de détection des rayons X à faibles angles (SAXS), la microscopie électronique, méthodes spectroscopiques ou encore le dichroïsme circulaire.

Cristallographie aux rayons X. La diffraction des rayons X est la méthode la plus efficace, plus informative et plus précise pour étudier la structure des grosses molécules. Dans de nombreux cas, l'analyse par diffraction des rayons X de cristaux de protéines ou d'acides nucléiques a permis de déterminer complètement la structure tertiaire de ces molécules avec une résolution de 3 Å ou mieux (selon les statistiques de la PDB, la résolution des structures est comprise entre 0.48 et 11 Å). Cette méthode permet, en théorie, d'obtenir la structure de la protéine la plus stable correspondant à un minimum d'énergie potentielle. La densité des électrons à l'intérieur du cristal est responsable de la diffraction des rayons X (longueur d'onde entre 1 et 100 Å) qui lui sont envoyé. Les cartes de diffraction obtenues sont ensuite traitées pour rétablir la carte de densité électronique obtenue à partir du *pattern* de diffusion des rayons X par un cristal. La mise en relation avec des données biologiques comme la séquence peut alors permettre la reconstruction de la structure de la protéine[232].

Cryomicroscopie électronique (Cryo-EM). La microscopie électronique permet l'obtention d'images de basse résolution des protéines étudiées et est particulièrement adaptée pour l'étude des grandes protéines ou d'assemblage macromoléculaire. La méthode la plus performante est la cryo-EM pour laquelle l'échantillon étudié est préservé dans un état proche de l'état natif par flash-congélation dans l'hélium liquide.

Les micrographes (images) obtenus par cette méthode sont des projections en 2D de la structure 3D. Ces images sont ensuite assemblées et réorientées pour obtenir les structures 3D de la protéine sous différents angles et à des résolutions plus ou moins fines. Les structures résolues par cryo-EM et déposées dans la PDB sont minoritaires et leurs résolutions varient entre 1.15 Å à 70 Å, permettant de caractériser plutôt une forme géométrique d'une protéine ou un complexe macromoléculaire que sa structure au niveau atomique[232]. Une méthode alternative d'estimation de la forme d'une protéine est la diffraction aux petits angles, une technique analytique qui mesure l'intensité des rayons X diffusés par un échantillon (protéine en solution) en fonction de l'angle de diffusion. Les mesures sont effectuées à de très petits angles, généralement entre 0,1 et 5°.

Expérimentalement, les protéines désordonnées sont difficiles à caractériser. Dans le cas de la cristallographie aux rayons X, l'obtention d'un cristal parfait à partir d'une solution concentrée et pure de la protéine à étudier est requise. Une haute résolution signifie que les protéines cristallisées ont adopté des conformations rigoureusement identiques. L'absence de densité électronique pour certaines régions protéiques (souvent des IDRs) rend difficile la résolution d'une structure complète. Les régions peu repliées ou désordonnées, adoptent quant à elles un ensemble riche de conformations dans le cristal qui empêche leur résolution. L'obtention de la structure d'une protéine désordonnée est impossible. Similairement, dans le cas de la cryo-EM, la flexibilité des régions désordonnées les empêcherait d'être observées [233].

4.2. MODÉLISATION MOLÉCULAIRE

Les méthodes computationnelles peuvent apporter certaines réponses aux difficultés expérimentales susmentionnées et prédire une structure à partir de données empiriques mises à disposition dans des bases de données (séquence, structures de protéines similaires).

Modélisation par homologie. C'est une méthode comparative basée sur la notion « à séquence homologue, repliement similaire ». Si la structure d'une protéine A est connue et que sa séquence et celle d'une protéine B sont homologues (similarité ≥ 50 %), alors il est possible de modéliser la structure B *(target)* à partir de celle de A *(template)*. Une méthode connue s'appuie sur le principe de satisfaction des contraintes spatiales. Modeller[234] est un logiciel basé sur ce principe permettant de créer plusieurs modèles d'une protéine *target* à partir d'une structure connue d'une protéine similaire utilisée comme *template* et de réaliser l'optimisation scorée de leur géométrie. La qualité stéréochimique d'un modèle peut être visualisée par un plot de Ramachandran[235]. On note que cette méthode n'est pas universelle, car la séquence homologue pourrait avoir une structure très différente[236] et deux protéines très peu similaires en séquence peuvent avoir des structures très similaires[237].

Modélisation *de novo*. Elle consiste à prédire une structure en se basant seulement sur l'information de séquence et la structure des fragments. La méthode Rosetta est sans doute la plus connue[238]. Elle consiste à enfiler localement sur la séquence cible des fragments homologues de protéines résolues expérimentalement. Cet assemblage se fait aléatoirement par un algorithme de Monte-Carlo.

Modélisation par apprentissage (AlphaFold). AlphaFold est un modèle d'apprentissage profond prenant en entrée une séquence protéique et délivre en sortie une structure. Le réseau est basé sur l'affinage d'une information mutuelle d'évolution encodée dans un alignement multiple (MSA) et contraintes spatiales encodées dans une *pair representation* (les acides aminés qui évoluent le plus sont probablement proches sur la structure) jusqu'à prédiction finale d'un modèle 3D de la protéine[239]. La dernière version de AlphaFold a obtenu les meilleurs résultats au rendez-vous d'experts *Critical Assessment of Structure Prediction* (CASP) 14 en 2020[240].

Ces trois grands principes de la modélisation moléculaires concernent non seulement des protéines seules, mais, dans le cas de la modélisation par homologie ou AlphaFold, permettent la modélisation de complexes moléculaires protéines-protéines. Dans le cas où de tels complexes ne sont pas empiriquement résolus ou si une petite molécule (ligand) est considérée comme partenaire d'une protéine, il est nécessaire de se tourner vers des méthodes d'amarrage moléculaire.

4.3. AMARRAGE MOLECULAIRE

L'amarrage moléculaire *(docking)* est un ensemble de méthodes permettant de modéliser des complexes moléculaires entre une protéine et un partenaire (ligand ou une autre protéine). Ces méthodes sont appréciées pour la compréhension de phénomènes biologiques complexes (ex. transduction du signal, complexation, réactions enzymatiques). Elles permettent, dans le cas du *structure-based drug design*, la prédiction de régions pouvant être ciblées par des ligands, ou leur activité sur une protéine cible à moduler dans le cas de *screening*.

Les méthodes de *docking* reposent sur deux étapes : l'échantillonnage de conformations des partenaires, puis le calcul d'un score de *docking*.

L'**échantillonnage des conformations d'un partenaire** peut s'effectuer selon plusieurs méthodes : par la complémentarité de surface entre les partenaires[241], basées sur des fragments (peptides)[242], une recherche stochastique (Monte-Carlo[243], algorithme génétique[244]) ou par simulation de dynamique moléculaire[245].

Protéines et ligands sont des molécules possédant un très grand degré de liberté interne (flexibilité). Les méthodes de *docking* peuvent prendre en compte ou non cette flexibilité. Le ***docking* rigide** concerne les méthodes dans lesquelles la flexibilité des

deux molécules n'est pas prise en compte dans la recherche de l'espace conformationnel des complexes. Le ***docking* flexible** prend en compte les torsions entre liaisons par consultation de librairies de rotamères des acides aminés.

Le calcul d'une **fonction de score** permet l'évaluation des poses de *docking* pour sélectionner les plus favorables. Ces fonctions de score sont groupées en trois types : les fonctions basées sur des paramètres physiques, sur la connaissance *(knowledge-based)*, sur des données empiriques ou sur un apprentissage supervisé[246]. Celles ***basées sur les paramètres physiques*** calculent d'une part une « énergie libre de liaison » du complexe par la somme de termes représentant les interactions non covalentes entre les partenaires, d'autre part une « énergie libre » de solvatation par des modèles comme les *Generalized Born* ou *Poisson-Boltzmann Surface Areas*[247] (ex. *High Ambiguity Driven protein−protein Docking*, HADDOCK[248]). Les fonctions de scores ***basées sur des résultats empiriques*** relèvent de l'utilisation de données biochimiques ou biophysiques pour des complexes dont la structure a été résolue[249]. Enfin, les fonctions de score ***basées sur la connaissance*** se fondent sur la comparaison des paires d'atomes en interaction dans le complexe modélisé avec des structures références de complexes[250,251].

Quelques méthodes de *docking* protéine-ligand et protéines-protéines ont récemment été répertoriées dans des *reviews*[252,253].

L'obtention des modèles de structures protéiques, qu'ils soient pour des protéines seules ou en complexes n'est que la première étape pour étudier les phénomènes biologiques complexes ou la modulation de telles molécules. Ces systèmes ne sont pour la plupart pas dans une configuration de minimum d'énergie nécessaire à l'obtention d'une structure à l'équilibre. Définir une fonction d'énergie potentielle permettant d'évaluer la proximité d'un système à son état d'équilibre est donc essentiel.

4.4. CHAMPS DE FORCE

Le calcul de l'énergie potentielle est la pierre angulaire des méthodes computationnelles visant à étudier les propriétés structurales, thermodynamiques et dynamiques des particules (atomes) d'un système.

Un champ de force est l'expression des termes mathématiques de la fonction d'énergie potentielle. Le champ de force contient l'ensemble des paramètres propres à chacun des atomes et leurs interactions. Ces paramètres sont mesurés expérimentalement ou calculés par les méthodes théoriques (mécanique quantique)[254]. Le choix du champ de force dépend des propriétés du système[255].

Le calcul de l'énergie potentielle du système regroupe les six termes décrivant ces phénomènes d'interactions. Les champs de force pour la simulation calculent l'énergie potentielle conformationnelle par la somme de l'énergie potentielle de termes concernant les relations entre atomes liés par liaisons covalentes (étirement des liaisons $V_{liaisons}$, angle de valence V_{angle}, angle dièdre $V_{dièdre}$, angles de torsions impropres V_{imp}) et les interactions non covalentes (interactions de van der Waals V_{vdW} par un potentiel de Lennard-Jones 6-12 et interactions électrostatiques V_{elec} par la loi de Coulomb). La fonction d'énergie potentielle le plus populaire est décrite pour un système de N atomes (Équation 1) :

$$V = V_{liaison} + V_{angle} + V_{dièdres} + V_{imp} + V_{vdW} + V_{elec} \quad (1)$$

Ou développée (Équation 2) :

$$V = \sum_{i,j}\left(\frac{1}{2}\, k_{ij}^{r}(r_{ij} - r_{ij}^{0})^2\right) + \sum_{i,j,k}\left(\frac{1}{2}\, k_{ijk}^{\theta}\left(\theta_{ijk} - \theta_{ijk}^{0}\right)^2\right) + \sum_{i,j,k,l}\left(\frac{1}{2} k_{\Phi_{ijkl}}\left[1 + \cos\left(n\Phi_{ijkl} - \Phi_{ijkl}^{0}\right)\right]\right) + \sum_{i,j,k,l}\left(k_{ijkl}^{\xi}\left(\xi_{ijkl} - \xi_{ijkl}^{0}\right)^2\right) + \sum_{i}\sum_{j\neq i} 4\varepsilon_{ij}\left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right] + \sum_{i}\sum_{j\neq i} f \times \frac{q_i q_j}{\varepsilon_r r_{ij}} \quad (2)$$

Les liaisons, angles et de torsions sont représentés comme des potentiels harmoniques et les angles dièdres par une fonction cosinus déviant de valeurs d'équilibre (r_{ij}^0, θ_{ijk}^0, ξ_{ijkl}^0, Φ_{ijkl}^0) et dépendant de constantes de force (k_{ij}^r, k_{ijk}^θ, k_{ijkl}^ξ, $k_{\Phi_{ijkl}}$). Au potentiel des angles dièdres, la fonction cosinus prend en compte la multiplicité n (le nombre de fois où ce potentiel sera minimum si la liaison subit une rotation de 360°). Le potentiel de Lennard-Jones reflète les forces d'attractions-répulsions induites par les fluctuations des nuages électroniques de deux atomes séparés d'une distance r, avec σ_{ij} la distance où V_{vdW} est nulle et ε_{ij} la profondeur du puits de leur V_{vdW} minimum. La loi de Coulomb représente les interactions électrostatiques dépendantes des charges partielles q de deux atomes, de la constance diélectrique du milieu ε_r et de la distance r_{ij} entre les deux atomes. Le paramètre $f = \frac{1}{4\pi\varepsilon_0}$, ε_0 la permittivité du vide, est le facteur de conversion électrique.

Quelques champs de forces utilisés pour la simulation de biomolécules sont CHARMM, AMBER, GROMOS, OPLS[255].

La majorité des champs de forces pour les protéines ont été développés pour des protéines ordonnées et sont difficilement transférables pour la description des protéines désordonnées. Plusieurs stratégies visent à améliorer ces champs de forces en ajustant les paramètres des angles dièdres et leur distribution, les interactions protéine-eau[256].

Des exemples de champ de force spécifiques des IDPs sont ff14IDPSFF[257], Drude2019IDPS[258] pour les protéines, et le champ de force TIP4P-B[259] pour un modèle d'eau.

4.5. SIMULATIONS DE DYNAMIQUE MOLECULAIRE

Les fonctions des protéines sont étroitement associées aux mouvements internes se produisant dans une large gamme d'échelles de temps, des vibrations atomiques aux mouvements globaux (ex. les transitions conformationnelles à grande échelle), couvrant la femtoseconde à la milliseconde ou à une échelle de temps plus longue. Il est difficile d'apprécier expérimentalement au niveau atomique une telle dynamique et les mouvements associés aux interactions intra- et intermoléculaires. Ainsi, la simulation de cette dynamique (MD) est un outil indispensable pour l'étude *in silico* de tels phénomènes.

Protocole général de la simulation de dynamique moléculaire dite classique (cMD). Une cMD se déroule en quatre étapes. Premièrement, on établit la **topologie** du système. Elle contient toutes les informations intrinsèques le décrivant, les paramètres physicochimiques des atomes (masses, constantes physiques, charges), boîte de simulation, conditions périodiques aux bordures, fonction d'énergie potentielle. Les protéines peuvent être présentées sous une forme simplifiée (gros-grain) ou complète (tout-atome). De même, la solvatation du système peut être simulée implicitement par la constante diélectrique de l'eau ou explicitement par ces molécules. Puis, on réalise l'optimisation de la géométrie du système et la **minimisation** de son énergie potentielle totale. Dans cette étape appelée **équilibration**, les paramètres thermodynamiques du système (pression, température) sont ajustés à des valeurs de référence (1 bar, 300 ou 320 K). Ainsi, le système atteint une configuration d'équilibre sous la fonction d'énergie potentielle choisie. La **production** est la simulation cMD est en elle-même. C'est une procédure itérative où (1) les vélocités des atomes sont initialisées par leur tirage aléatoire dans la distribution de Maxwell-Boltzmann ; (2) les forces agissant sur chaque atome sont calculées conformément à la fonction d'énergie potentielle ; (3) les équations du mouvement de Newton sont intégrées pour un pas de temps Δt ; (4) les nouvelles coordonnées et vélocités des atomes sont obtenues ; (5) les nouvelles quantités physiques telles que les énergies cinétique et totale sont calculées. Les étapes 3 à 5 sont répétées jusqu'à ce que les conditions d'arrêt soient rencontrées.

Limitations de la cMD. Bien qu'outil important pour la description des protéines, la cMD possède plusieurs obstacles à une description précise et proche des phénomènes biologiques. Tout d'abord, au niveau de la précision des champs de force utilisés et leur adaptation au système simulé pour décrire les interactions et les comportements dynamiques du système. Ensuite, les échelles de temps de simulation sont parfois trop

courtes pour permettre une relaxation complète du système et/ou un échantillonnage suffisant pour décrire les phénomènes structuraux et dynamiques d'intérêt.

Simulation de dynamique moléculaire accélérée (aMD). Beaucoup de méthodes existent pour contourner les limites de la cMD. La aMD est une méthode d'amélioration de l'échantillonnage de l'espace conformationnel visant à relever les barrières d'énergie potentielle en introduisant un biais au potentiel total calculé à chaque étape de la simulation, si ce potentiel dépasse une valeur de référence[260,261]. Ainsi la probabilité que le système « tombe » dans un puits d'énergie est diminuée, lui permettant d'explorer d'autres zones de son espace conformationnel.

In fine, la MD (classique ou accélérée) produit une trajectoire dans laquelle la protéine évolue dans un espace 3D. Ainsi, on obtient un ensemble de conformations du système décrit par un jeu de coordonnées cartésiennes en fonction du nombre de pas simulés.

4.6. ANALYSES STRUCTURALES ET DE LA DYNAMIQUE ESSENTIELLE

Lors de la MD, les mouvements enregistrés concernent non seulement les atomes les uns par rapport aux autres mais également la protéine dans son milieu (déplacements globaux des conformations). La trajectoire obtenue est porteuse de nombreuses informations sur la protéine, à la fois structurales, dynamiques et sur l'ensemble conformationnel échantillonné. Le post-traitement de ces données et leur analyse permettent de combler le vide entre la description atomique numérique, la biophysique et les fonctions biologiques. Des outils de base largement utilisés et reposant sur des techniques de réduction de dimensionnalité fournissent des informations sur les systèmes.

Recalage des données. Pour comparer les conformations, et effectuer des calculs corrects, il est nécessaire de réaliser une superposition ou recalage. Cette méthode consiste à éliminer les translations et rotations entre deux structures pour minimiser leur distance au carré.

Root mean square deviation (RMSD). Il calcule la distance euclidienne moyenne entre les N atomes de deux conformations (Équation 3).

$$\mathrm{RMSD}(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\mathbf{r}_i(t) - \mathbf{r}_i^{ref}\right)^2} \qquad (3)$$

Où, pour la $i^{ème}$ paire d'atomes, $\mathbf{r}_i(t)$ représente les coordonnées d'une conformation au temps t et $\mathbf{r}_i^{ref}$ celles d'une conformation de référence.

Root mean square fluctuation (RMSF). Il mesure la déviation standard d'un atome par la distance moyenne au carré par rapport à sa position moyenne (Équation 4).

$$\mathrm{RMSF}(t) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle)^2} \quad (4)$$

Où, pour le $i^{ème}$ atome, $\mathbf{r}_i(t)$ représente ses coordonnées au temps t et $\langle \mathbf{r}_i \rangle$ ses coordonnées moyennes.

Assignation des structures secondaires. La méthode *Define Secondary Structure of Proteins* (DSSP)[262] est basée sur un processus de reconnaissance de motifs de liaisons hydrogène optimales entre les groupements –NH et –C=O du squelette peptidique et assigne à chaque résidu la structure secondaire la plus probable ou aucune.

Interactions non covalentes. Elles correspondent à des interactions de courte portée mettant en jeu des phénomènes d'attractions-répulsions. Ces interactions peuvent être décrites par des critères géométriques. Une liaison hydrogène existe si la distance entre les atomes lourds (N, O, S) donneurs (D) et accepteurs (A), est inférieure à 3.6 Å et l'angle $\widehat{DHA}$ est supérieur à 120° (parfois 90°). Si des charges formelles opposées sont mises en jeu dans cette liaison, on parlera de pont salin. Une interaction hydrophobe existe si la distance entre deux carbones appartenant à la chaîne latérale de deux résidus est inférieure à 3.6 Å.

Rayon de giration (Rg). Il mesure le niveau de compaction d'une protéine par le calcul de la distance moyenne des atomes au centre de masse (Équation 5).

$$R_g = \sqrt{\frac{\sum_{i=1}^{N} m_i r_i^2}{\sum_{i=1}^{N} m_i}} \quad (5)$$

Où m_i est la masse de l'atome i, r_i sa distance au centre de masse de la protéine.

Surface accessible au solvant (SASA). C'est une mesure géométrique de l'exposition de la protéine à son solvant. Son principe repose sur le calcul de la surface de contact des sphères S du solvant et R_i, un atome i à condition que S ne soit en contact qu'avec R_i[263].

Angles de courbure. Cette méthode permet de calculer l'angle relatif entre deux structures secondaires ou par rapport à leur position initiale (Équation 6).

$$\Theta = \arccos \frac{\mathbf{v_1} \cdot \mathbf{v}_2}{\sqrt{\|\mathbf{v}_1\|} \cdot \sqrt{\|\mathbf{v}_2\|}} \quad (6)$$

Où Θ est l'angle en radian, $\mathbf{v}_1$ et $\mathbf{v}_2$ les vecteurs de coordonnées représentatifs des

deux éléments comparés.

Analyses en Composantes Principales (ACP). L'ACP est une technique classique de la réduction de la dimension de données. En dynamique moléculaire, elle est utilisée pour la caractérisation des mouvements collectifs sous-jacents à la dynamique à l'équilibre. Les fluctuations aléatoires d'une protéine cachent la contribution de mouvements coopératifs d'importance pour sa fonction biologique. Ces modes englobent l'entièreté de la structure (modes globaux) ou ses sous-domaines (modes internes).

Une simulation de DM délivre la trajectoire d'un système dans une matrice de coordonnées cartésiennes T × 3N (où N est le nombre d'atomes et T le nombre de pas de temps de simulation). Pour effectuer l'ACP, les mouvements de translation et de rotation ont été supprimés en superposant chaque conformation de la trajectoire sur la conformation moyenne.

La corrélation entre les mouvements atomiques peut être exprimée par la matrice de covariance **C** (Équation 7).

$$C_{ij} = (\mathbf{r}_i(t) - \langle \mathbf{r}_i \rangle)\left(\mathbf{r}_j(t) - \langle \mathbf{r}_j \rangle\right) \quad (7)$$

Où pour les i et j^{èmes} atomes, **r**(t) représente leurs coordonnées à l'instant t et ⟨**r**⟩ leurs coordonnées moyennes.

Cette matrice carrée symétrique est diagonalisable par une matrice de transformation **V** telle que la variance de **C** est maximale (Équation 8).

$$\mathbf{\Lambda} = \mathbf{V}^{T}\mathbf{C}\mathbf{V} \quad (8)$$

Où **Λ** est une matrice diagonale 3N × 3N contenant les valeurs propres λ_i. La i^{ème} colonne de **V** le vecteur propre associé à λ_i.

Les valeurs propres définissent les fluctuations ou mouvements totaux du système le long des vecteurs propres qui leur sont associés.

En arrangeant les composantes principales (modes) par ordre décroissant de valeurs, le premier mode décrit les mouvements les plus coopératifs et les plus amples (fluctuations maximales). Les composantes principales suivantes expliquent les mouvements restants. Le pourcentage de mouvements expliqués par chaque mode m est exprimé par l'Équation 9.

$$\alpha(m) = \frac{\lambda_m}{\lambda_{total}} \quad (9)$$

Plus les modes sont petits, plus les mouvements expliqués relatent les modes de vibration des liaisons atomiques[264].

Le degré de collectivité k d'un domaine D dans un mode m peut être calculé directement (Équation 10) :

$$k_m^D = \frac{1}{n} \exp\left\{ -\sum_{i=1}^{n} \alpha_i(m)^2 \ln \alpha_i(m)^2 \right\} \quad (10)$$

Soit en appliquant un poids inversement proportionnel à la contribution du mode m aux mouvements totaux (Équation 11).

$$k_m^{D'} = \frac{1}{\lambda_m} k_m^D \quad (11)$$

Où n est le nombre d'atomes dans D, λ_m la valeur propre associée au mode m et $\alpha_i(m)$ la contribution de l'atome i au mouvement de D (Équation 12).

$$\alpha_i(m)^2 = \frac{x_i(m)^2 + y_i(m)^2 + z_i(m)^2}{\alpha_D(m)^2} \quad (12)$$

La contribution $\alpha_D(m)^2$ d'un domaine D au mouvement global décrit par un mode m est calculée par l'Équation 13 :

$$\alpha_D(m)^2 = \sum_{i=1}^{n} \left[x_i(m)^2 + y_i(m)^2 + z_i(m)^2 \right] \quad (13)$$

Cross-corrélations dynamiques. Elle décrit la corrélation temporelle entre chaque paire d'atomes[265] par l'Équation 14 :

$$CCM_{ij} = \frac{\langle \Delta \mathbf{r}_i \cdot \Delta \mathbf{r}_j \rangle}{\sqrt{\langle \Delta \mathbf{r}_i^2 \rangle} \cdot \sqrt{\langle \Delta \mathbf{r}_j^2 \rangle}} \quad (14)$$

Où **CCM** est la matrice de cross-corrélations, $\Delta \mathbf{r}_i$ et $\Delta \mathbf{r}_j$ sont les vecteurs de déplacements des atomes i et j.

Si $CCM_{ij}=1$, les fluctuations des atomes i et j sont complètement corrélés (même phase et même période). Si $CCM_{ij}=-1$, les fluctuations de i et j sont complètement anti-corrélées, et si $CCM_{ij}=0$, les fluctuations de i et j ne sont pas corrélées.

Analyse des modes normaux (AMN). Elle permet la description des états flexibles d'une protéine autour d'une conformation d'équilibre. L'idée générale de l'AMN est la suivante : quand un système oscillatoire à l'équilibre (un minimum énergétique pour une protéine) est légèrement perturbé, des forces s'appliquent pour la ramener à son état d'équilibre.

Une protéine de N atomes est représentée comme un réseau de N oscillateurs

harmoniques de force constante k situés à une distance de coupure d et dont les modes de vibration reproduisent le mouvement général d'un système[266].

Au point d'énergie minimal, le potentiel V du système peut être approximé par l'Équation 15 :

$$V = \frac{k}{2}\sum_{ij}^{M}(\Gamma_{ij})\left(|R_{ij}| - |R_{ij}^{0}|\right)^{2} \qquad (15)$$

Où M est le nombre de ressorts, $|R_{ij}| - |R_{ij}^{0}|$ est la distance entre les nœuds i et j par rapport à la structure équilibrée, et Γ_{ij} est un élément de la matrice laplacienne correspondant au contact inter-résidus entre les nœuds i et j. Les dérivées secondes de **V** donnent accès à la matrice hessienne **H** décrivant la contribution de chaque paire d'atomes de coordonnées cartésiennes (X, Y, Z) au potentiel V (Équation 16).

$$H_{ij,\, j\neq i} = \frac{k\Gamma_{ij}}{\left(R_{ij}^{0}\right)^{2}}\begin{bmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{bmatrix} \qquad (16)$$

La diagonalisation de **H** donne les $3N - 6$ valeurs propres (modes) non nulles λ_k représentant les fréquences de vibration et les vecteurs propres V_k décrivant les fréquences de vibration et leurs directions.

Les six premiers modes sont nuls ou négatif et correspondent aux translations et rotations de la protéine dans l'espace 3D. Ainsi, les modes de basse fréquence de vibration représenteront les mouvements collectifs les plus lents et les plus amples (réarrangements structuraux intra- et interdomaines) difficilement atteignables par simulation de dynamique moléculaire. Les modes de plus haute fréquence impliquent un faible nombre d'atomes et se rapprochent des modes vibratoires des liaisons.

Un avantage de cette méthode est qu'elle est peu complexe dans le temps. Quelques limites sont l'approximation du minimum local qui rend l'ANM pertinente que pour une conformation à l'équilibre, l'absence de prise en compte du solvant, la linéarité des mouvements calculés quand les déplacements internes d'une protéine peuvent comprendre à la fois rotations et translations. Computationnellement, les calculs sous-jacents sont complexes en temps lorsque N augmente (diagonalisation de la matrice hessienne).

4.7. CARACTERISATION DE L'ENSEMBLE CONFORMATIONNEL

En se détachant de la variable de temps, on peut voir la trajectoire de dynamique moléculaire comme un sous-ensemble de conformations échantillonnées dans l'ensemble conformationnel d'une protéine. Les analyses suivantes permettent de

rendre compte de la diversité conformationnelle de cet espace.

Apprentissage non supervisé. L'ensemble de ces méthodes d'apprentissage dites non supervisées a pour but le regroupement des conformations les plus similaires obtenues lors de la dynamique. Ces méthodes sont souvent basées sur un critère de distance entre conformations en utilisant l'information de structure (coordonnées cartésiennes) ou variables intrinsèques (ex. structures secondaires, angles dièdres, distances internes, rayon de giration… etc.). Les méthodes de partitionnement utilisées lors des analyses incluent l'*ensemble-based clustering*[267], le clustering basé sur les structures secondaires, les K-means, le *density-based spectral clustering*[268] et la classification hiérarchique ascendante. Pour certaines méthodes, les performances pour la sélection du meilleur algorithme ou des meilleurs paramètres, c'est-à-dire la capacité d'un algorithme à partitionner au mieux les données, ont été évaluées par la silhouette[269].

Paysage d'énergie libre de Gibbs. Une autre stratégie pour l'étude de l'ensemble conformationnel est sa description par l'énergie libre de Gibbs (ΔG), soit la probabilité de trouver un système (protéine) dans un état particulier. Cette représentation des conformations échantillonnées projette les données sur deux coordonnées de réaction ou variables collectives R (Équation 17).

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (17)$$

Où k_B est la constante de Boltzmann, T la température, R_1 et R_2 deux coordonnées de réaction, et P une probabilité définie sur l'histogramme bidimensionnel de R_1 et R_2. Cette méthode permet la reconstruction du paysage d'énergie libre, de comparer les différents états visités par la protéine lors de la simulation et visualiser les barrières énergétiques les séparant[270].

Pour aborder les deux thématiques, RTK KIT et hVKORC1, produire les modèles, générer les simulations et analyser les données, nous avons utilisé les mêmes méthodes mathématiques et approches bioinformatiques. De subtiles différences concernent le choix de certains critères (seuils, structure de référence), dans les données analysées (trajectoires uniques ou concaténées pour les protéines entières ou des domaines particuliers tronqués). L'ensemble de ces éléments de modélisations et analyses est présenté dans l'annexe Matériels et Méthodes.

PARTIE 2. RESULTATS

Dans l'introduction, nous avons vu l'ensemble des propriétés inhérentes aux protéines à multiples domaines topologiques et/ou structuraux pouvant comprendre une ou plusieurs régions ordonnées ou désordonnées, ainsi que les défis associés à leur description et à l'étude des complexes qu'ils forment. Le RTK KIT et hVKORC1 sont présentés comme deux exemples archétypaux de protéines modulaires possédant possiblement 1 à 4 régions désordonnées et ont été choisis comme objets d'étude de ma thèse.

Chaque protéine a été caractérisée comme une entité intégrale et par ses domaines pour mettre en évidence ses propriétés intrinsèques (spécifiques à un domaine) et extrinsèques (relations entre domaines).

Ainsi, les résultats de la recherche sur ces protéines sont regroupés dans la thèse de manière à caractériser dans cet ordre leurs propriétés intrinsèques, leur modularité, leur DYNASOME et leur INTERACTOME. La majorité des résultats présentés étant publiée, le contenu de ces publications a été adapté en plusieurs grands axes de description, allant des moins (modélisation, DYNASOME) aux plus complexes (INTERACTOME). Nous avons en parallèle mis l'accent sur les voies applicatives des études du DYNASOME et de l'INTERACTOME pour la définition de nouvelles cibles pour le développement de modulateurs protéiques (POCKETOME). Cette approche innovante, le *allo-network drug design* présente un réel intérêt thérapeutique.

CHAPITRE 5. PROTEINES INTRINSEQUEMENT DESORDONNEES : DE LA MODELISATION AUX PROPRIETES DYNAMIQUES – DYNASOME

Dans ce chapitre, nous allons présenter les caractéristiques structurales et dynamiques intrinsèques de deux protéines – le récepteur tyrosine kinase KIT et la vitamine K époxyde réductase humaine (hVKORC1) – dans le but de reconstituer leur DYNASOME ou ensemble des niveaux de repliements et leur dynamique.

On décrira tout d'abord de telles propriétés pour le RTK KIT et hVKORC1. Additionnellement, pour hVKORC1, on comparera les propriétés intrinsèques du modèle *de novo* proposé par BiMoDyM en 2017 avec les structures cristallographiques de hVKORC1 résolues en 2021.

Ce chapitre est une adaptation des articles suivants :

1. **Ledoux, J.**, Trouvé, A., & Tchertanov, L. (2022). The Inherent Coupling of Intrinsically Disordered Regions in the Multidomain Receptor Tyrosine Kinase KIT. *International Journal of Molecular Sciences*, *23*(3), 1589. https://doi.org/10.3390/ijms23031589

2. Stolyarchuk, M.[+], **Ledoux, J.**[+], Maignant, E., Trouvé, A., & Tchertanov, L. (2021). Identification of the Primary Factors Determining the Specificity of Human VKORC1 Recognition by Thioredoxin-Fold Proteins. *International Journal of Molecular Sciences*, *22*(2), 802. https://doi.org/10.3390/ijms22020802

3. **Ledoux, J.**, Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

Les données supplémentaires et les méthodes relatives à toutes ces publications sont présentées dans les annexes de la thèse.

5.1. ORDRE ET DESORDRE DU RTK KIT

Résumé. *Le récepteur tyrosine kinase (RTK) KIT régule un grand nombre de processus cellulaires par son domaine cytoplasmique (CD). Ce domaine est constitué d'un domaine kinase entouré de sous-domaines hautement flexibles : la région juxtamembranaire, le domaine insert kinase et la queue C-terminale. Ces régions sont des recruteuses clés de protéines initiatrices de la signalisation. Pour fonder une base structurale pour la caractérisation de KIT avec ses protéines de signalisation (INTERACTOME), nous avons modélisé le CD complet lié à son hélice transmembranaire. Ce modèle du KIT à l'état inactif a été étudié par simulation de dynamique moléculaire dans des conditions mimant son*

environnement naturel. Par les descriptions structurales et dynamiques de cette protéine à plusieurs domaines, nous avons expliqué les désordres intrinsèques (intradomaines) et extrinsèques (interdomaines) du KIT et représenté l'ensemble conformationnel échantillonné par simulation de MD par des paysages d'énergie libre de Gibbs. Des mouvements fortement couplés au sein de chaque domaine et entre domaines distants de KIT prouvent l'interdépendance fonctionnelle de ces régions. Ce phénomène est largement observé dans de nombreuses protéines. Enfin, nous avons suggéré que KIT dans son état inactif intègre toutes les propriétés de la protéine à l'état actif pour les évènements post-transductionnels.

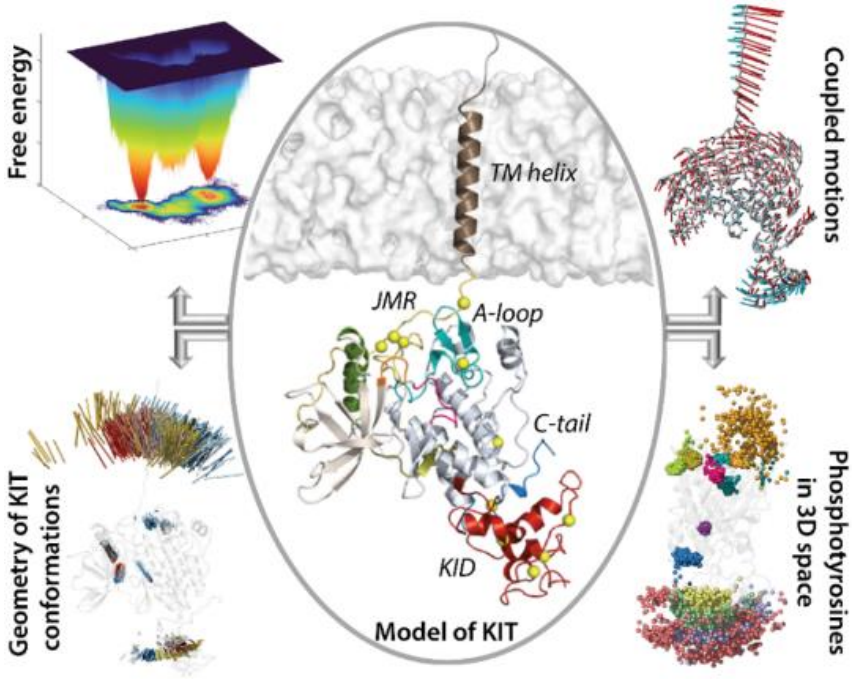


Figure 5.1 Abstract graphique de la section [271]

5.1.1. INTRODUCTION

Receptor tyrosine kinases (RTKs) are cell surface receptors with a highly selective affinity to numerous ligands—growth factors, cytokines, and hormones. Each RTK acts as a sensor for its specific extracellular ligand, whose binding triggers dimerization of the receptor, activation of its kinase function, and auto-phosphorylation of tyrosine residues in the cytoplasmic domain[65]. This mechanism leads to the recruitment and activation of multiple downstream signalling proteins, which carry the signal to the nucleus, thus altering the patterns of the gene transcriptions governing various aspects of the cell physiology. Initiation of this cascade-like process involves different regions of multidomain RTKs, each of them acting in a finely concerted behaviour, mediated by a tightly regulated allosteric mechanism controlling all their physiological processes[96,97]. The explicit elucidation of the signalling cascades represents a critical

and unsolved problem in cell biology.

Each RTK, depending on the location of its different domains, is composed of an extracellular domain (ED) and cytoplasmic domain (CD), linked by a single transmembrane helix (TM) (**Figure 5.2**, **A**). In turn, each domain of RTKs has a modular architecture consisting of several structural blocks, interconnected by coiled linkers providing high conformational plasticity. The ED, highly variable in RTKs, is formed by diversified units (Ig-like, cysteine-rich, and cadherin fragments[65] containing the highly selective ligand-binding site, while the CD architecture is similar and usually composed of the juxtamembrane region (JMR), bi-lobe tyrosine kinase (TK) domain with an ATP-binding region (N-lobe), and phosphotransferase domain (C-lobe), linked by a loop (hinge). In some RTK families, the canonical TK domain is interrupted by a kinase insert domain (KID)[272] (**Figure 5.2**, **B**).

The ligand-induced stimulation of RTKs promotes conformational changes of the ED, governing its dimerization (except IsulinR) and the signal transmission from the extracellular environment to the intracellular area. The conceptually straightforward mechanisms for the ligand-induced dimerization of RTKs are surprisingly different and encoded primarily by the sequence and structure of the extracellular domain[65,273]. As the structures of the intracellular domain of RTKs in the inactive state are distinct, the activation mechanism is specific for each receptor. As for the structures of the active state, they are rather similar, in which key regulatory elements, including the 'activation loop' (A-loop) and αC-helix in the kinase N-lobe, adopt a particular configuration in all activated TK domains that is necessary for catalysis and phosphotransfer reaction[274,275] (**Figure 5.2**, **C**).

The numerous studies focused on the regions containing tyrosine residues—JMR, A-loop, KID, and C-terminal—put in evidence their roles in different steps of RTK activation. For instance, in several RTKs, FLT3[276], KIT[121], and the EPH family[277], the inactive state of the TK domain is maintained by the auto-inhibited configuration of JMR, which is stabilised by extensive contacts with residues of the TK domain, including A-loop. The role of A-loop in kinase activation is not conserved in RTKs, e.g., the A-loop tyrosine is not necessary for EGF receptor activation[278], whereas its phosphorylation is essential for the activation of the insulin receptor[279,280].

KIT, one of 58 human RTKs, is activated by the binding of a growth factor, the stem cell factor (SCF), and regulates a variety of critical cellular processes, such as proliferation and differentiation, cell survival and metabolism, cell migration, and cell cycle control[114,129]. The aberrant activation of KIT inducing deregulation of signalling networks is associated with the progression of many cancer types, including human acute myeloid leukaemia, aggressive systemic mastocytosis, melanoma, gastrointestinal stromal tumour, and stomach cancers[281–283]. Disclosure of the KIT-activated pathways in carcinogenesis will be a crucial step towards the development of KIT-targeted therapies[284].

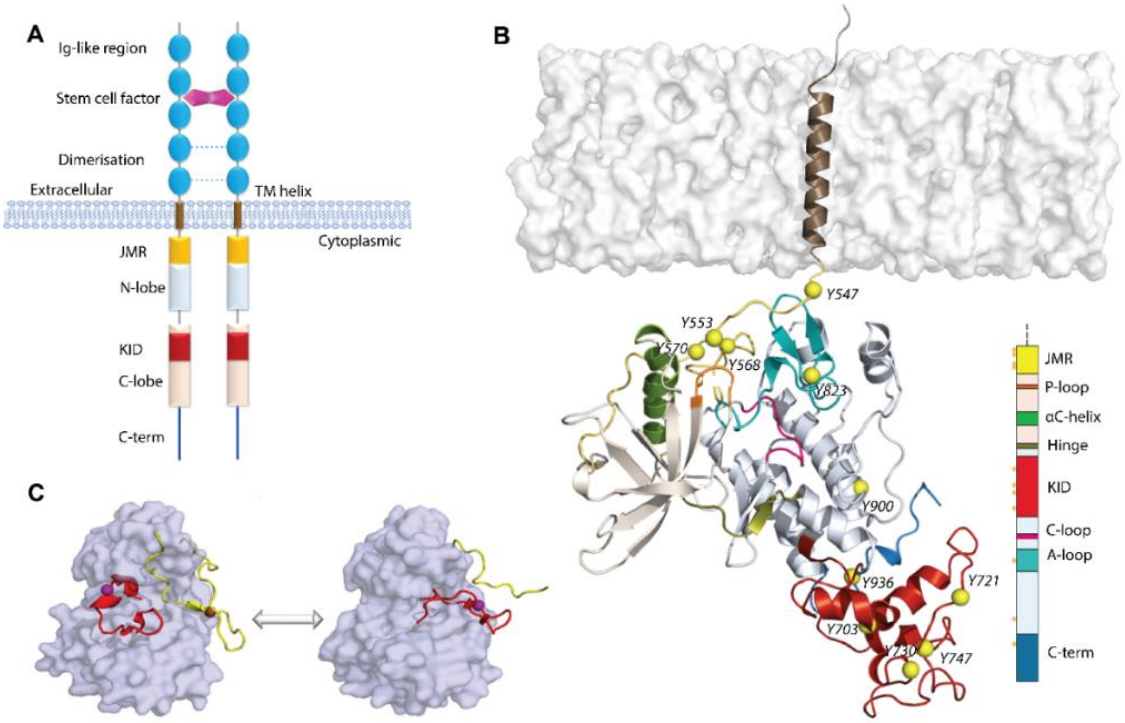


Figure 5.2 Structure of RTKs, illustrated on KIT, a member of the RTK family III. (**A**) Structural composition of KIT: an extracellular domain (ED) with five Ig-like regions, a transmembrane domain (TM helix), and a cytoplasmic domain (CD), composed of the juxtamembrane region (JMR), N- and C-lobe, spliced by a kinase insert domain (KID), and C-tail domain. The stem cell factor (SCF) extracellular binding induces dimerization and activation of KIT. (**B**) Structural model of the KIT containing CD and TM helix. The protein is shown as cartoon, membrane as grey surface, the phosphotyrosine residues (Y) as yellow balls. The KIT regions are coloured as shown on the scheme. (**C**) The inactive-to-active state transition of KIT is shown using the crystallographic structures of KIT CD in the inactive (PDB: 1T45) and active (PDB: 1PKG) states. In the inactive state (left), JMR (in yellow) is in the auto-inhibited conformation, stabilized through contacts with A-loop (in red), αC-helix, and C-loop. A-loop is packed to the TK domain. Both regions, JMR and A-loop, protect the catalytic site from ATP binding. In the active state (right), JMR and A-loop move from the TK domain to a solvent exposed position and deploy out of the active site, allowing ATP to access its binding site. The protein is shown as the solvent accessible surface, with JMR and A-loop as cartoon.

The human KIT is composed of 976 amino acids, in which residues L525–Y545, K546–K581, W582–S931, and T932-R946 represent the TMD, JMR, TK domain, and C-tail, respectively. In response to SCF binding, KIT activates several signalling pathways, using its rich set of phosphorylation sites localised on different fragments, i.e., JMR, A-loop, KID, C-lobe, and C-tail, which are the docking sites (scaffolds) of numerous proteins. In KIT, eight tyrosine phosphorylation sites have been identified *in vivo* (Y568 and Y570 in JMR; Y703, Y721, and Y730 in KID; Y823 in A-loop; Y900 in the C-lobe; and Y936 in C-tail)[155], as well as two additional sites detected *in vitro* in the activated kinase domain (Y547 and Y553 in JMR)[119].

As expected from the critical role of JMR in the regulation of KIT kinase activity, two tyrosine residues, Y568 and Y570, are involved in the maintenance of the auto-

inhibited conformation that blocks the regulatory αC-helix and ATP-binding P-loop, thereby suppressing kinase activity[37,121]. Upon SCF-stimulated activation, JMR adopts the solvent-accessible position, and these two tyrosine residues, Y568 and Y570, are the first to be phosphorylated and involved in downstream signalling[138]. Both phosphorylated tyrosine residues identified *in vivo* act as docking sites for signalling molecules with Src 2 homology domains (SH2), which, in turn, transmit the signal further through the cell[285].

The A-loop Y823 is likely to play a role of a pseudo-substrate, interacting with D792 in C-loop and, thus, maintaining the inactive conformation of KIT. It was reported that phosphorylation of Y823 is not required for KIT activation[119]; however, this residue is crucial for cell survival and proliferation[171]. The function of Y823 in post-transduction processes is still not clear, and it may interact with signalling proteins, but such interaction partners have yet to be identified[120]. The functions of JMR and A-loop appear to be strongly coordinated, not only in the inactive state of the KIT (auto-inhibition function), but also during KIT activation, when their concerted departure from the auto-inhibited positions in the inactive state contributes to other conformational changes in the protein, e.g., the removal of αC-helix and P-loop from their positions in the inactive state.

It is widely accepted that KID does not influence the kinase activity of KIT[286], and its functional role is to provide alternative binding sites for adaptors, signalling, and scaffolding proteins in the cytoplasm, through five functional phosphorylation sites, three tyrosine (Y703, Y721, and Y730), and two serine (S741 and S746)[287]. The C-terminal tail, containing phosphotyrosine residue Y936, also contributes directly to intracellular signalling[288]. Phosphorylation of KIT at Y703 and Y936 activates the mitogen-activated protein kinase (MAPK) pathway[289]. CrkII was identified to specifically bind Y900 in a phosphorylation-dependent manner, possibly via the p85 subunit of PI3-kinase[151].

Consequently, the different KIT regions regulate the catalytic process and events that activate and control the signalling cascade. As such, KIT functions are dependent on more than one region, and these regions should be directly or collaterally coupled. The study of interconnections between different, distant, or adjacent KIT regions, at the structural and dynamical levels, may shed light on their cooperativity required for different KIT functions.

Such study is pertinent because the 3D model of the nearly complete cytoplasmic domain of KIT, in which the empirically (X-ray crystallography) determined kinase domain, completed by the *de novo* KID and C-tail, was reported[183]. To study the interconnections between the functional region, i.e., JMR, TK domain, KID, and C-tail, the full-length cytoplasmic domain of KIT was investigated in its native environment, with a fully reconstructed JMR, attached to the transmembrane (TM) helix and inserted into the membrane (**Figure 5.2**, **B**). The study was performed by conventional

molecular dynamics (cMD) simulation, which generates atomic-level data of intrinsically high resolution.

We suggested that such a level of description of KIT, with the fully reconstructed JMR, KID, and C-tail, will elucidate the structural and dynamical properties of its different functional regions that contain (or not) the phosphotyrosine residues. Such characterisation can more explicitly explain the role of each region in maintaining KIT inactive state and establish the relationships between the regions showing either their cooperation or autonomy.

5.1.2. RESULTS

5.1.2.1. DATA GENERATION AND PROCEEDING

The 3D model of the KIT CD possessing KID and C-tail[183] was completed by the JMR segment T544-W557 and transmembrane helix attached to JMR. The multidomain construct of 431 amino acids (I516-R946) was studied by conventional MD simulations (cMD and all-atom, with explicit water and membrane) as a membrane protein. The extended MD simulation was repeated three times (replicas 1–3, each of 2 μs, started with different randomised initial atomic velocities and performed upon strictly identical conditions) to extend conformational sampling and examine the consistency and completeness of the produced KIT conformations. Each simulation started with an equilibrated conformation, obtained after minimising the neutralised solvated model. The generated data sets were analysed for a full-length construct and per domain/region. To avoid the motion of the protein as a rigid body, all data were normalised by least-square fitting of MD conformations to the initial conformation (t = 0 μs) as a reference.

5.1.2.2. GENERAL CHARACTERISATION OF MD CONFORMATIONS

The root-mean-square deviations (RMSDs), computed for each conformation, respective to the same initial conformation, display comparable profiles in all cMD trajectories, demonstrating the good reproducibility of the generated data (**Figure S1**). As shown by the per-domain analysis, the large variability of the RMSDs, calculated for all Cα-atoms of KIT, is mainly impacted by JMR, KID, and C-terminal, while the RMSDs of the TK N- and C-lobes show high stability. Similarly, the profile of the root-mean-square fluctuations (RMSFs) curves is comparable in the three MD trajectories, with differences only in the amplitude of the RMS fluctuations of the N- and C-terminals and KID. It is also interesting to note the increased RMSD and RMSF values for A-loop in trajectory 2.

5.1.2.3. 2D Folding and 3D Structure of KIT in Inactive State

The secondary structure interpretation of KIT conformations indicates that the folding is generally well-conserved in the TK domain and corresponds perfectly to those in the crystallographic structures 1T45 (inactive state) and 1PKG (active state), while JMR, KID, and C-tail changed their folding during and between each trajectory (**Figure 5.3**). The intrinsic disorder of KID was previously characterised[183,290] and we suggest that JMR and C-tail are also intrinsically disordered regions (IDRs).

Indeed, the repeated conversion of two 3_{10}-helices (I563-N566 and D572-Q575) into turn or bend, as well as the partial instability of two β-strands, are good arguments to classify JMR as an IDR. Similarly, the alteration of α-, 3_{10}-helices, bend, turn, and coil in C-tail proves its disordered nature. A-loop also displays a noticeable transformation of its fold, evidenced as a reversible transition of a turn to a 3_{10}-helix in three distinct A-loop segments, and a full unfolding of the antiparallel sheet, as observed in trajectory 2 (**Figure 5.3**, **A**). Surprisingly, the secondary structures of αC-helix, considered early as a canonical structure that only changes its position upon the activation[121,122], was slightly perturbed in some conformations by decrease of its size.

According to these observations, the multidomain KIT comprises at least four ID regions—JMR, KID, A-loop, and C-tail—centred around the structurally stable core, the TK domain. The TK domain, together with its secondary structure stability, exhibits weak or very weak inherent dynamics, as viewed by the small values of RMSD, RMSF for the residues from this domain (**Figure S1**), and limited angular variations (≤ 10°) of αC- and αE-helix, the representative fragments from the N- and C-lobe, respectively (**Figure S2**).

The great variability of the RMSD and RMSF values of KIT apparently derived from the two distinct factors, i.e., (i) an unsteady position of the flexible KID, JMR, and C-tail, with respect to the relatively stable TK domain, (ii) the intrinsic properties KID, JMR, and C-tail, connected to their unstable (metastable and transient) folding and/or conformational changes. Each KIT region can be regarded, in a first approximation, as a pseudo-rigid body with its local centre of gravity (centroid, C). Centroids, determined on JMR (C_{JMR}), KID (C_{KID}), C-tail ($C_{C\text{-}tail}$), and the TK domain (C_{TKD}), are the nodes of a dynamical tetrahedron that reflects the displacement of each region, respective to one another (**Figure 5.4**, **A**).

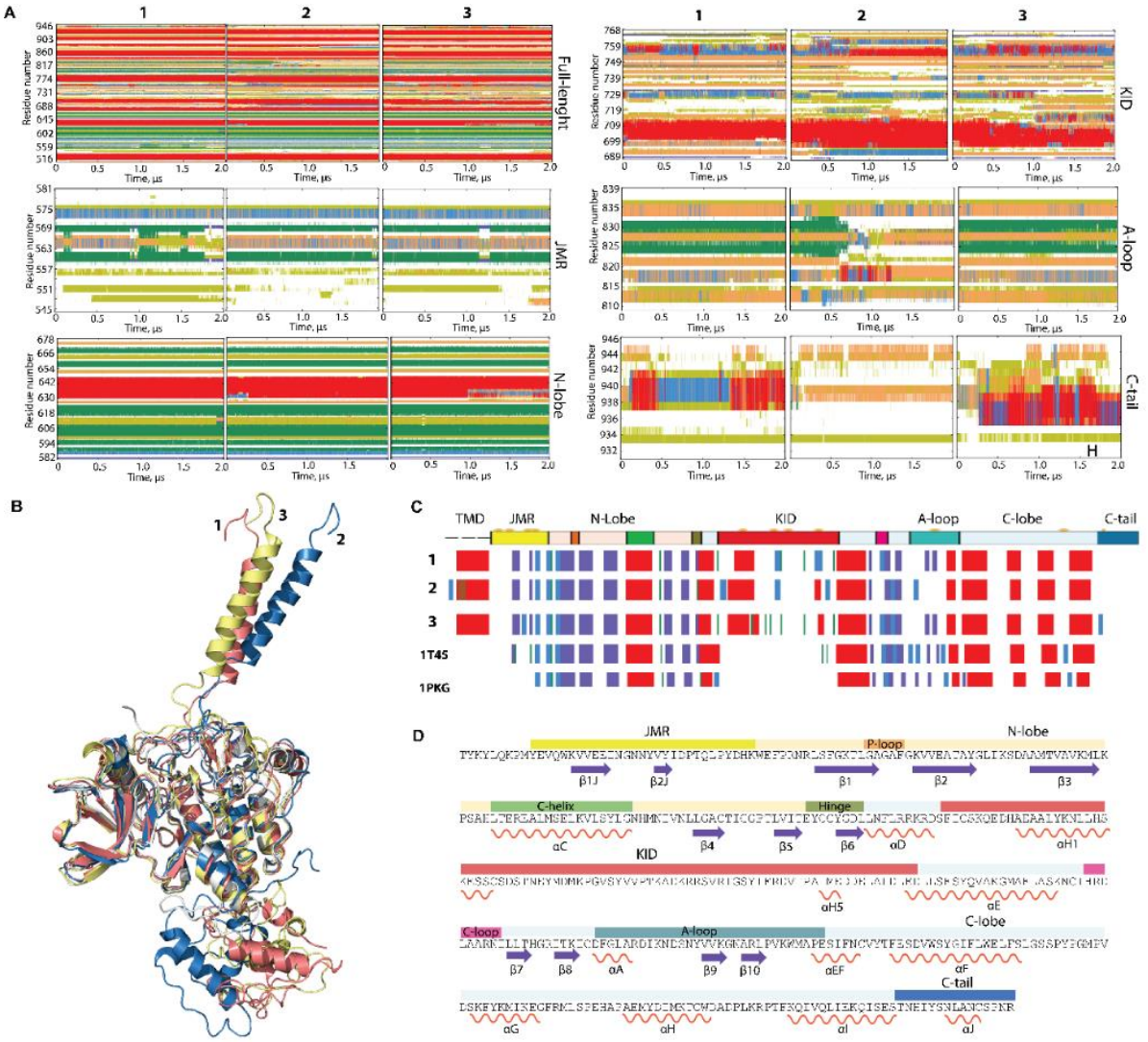


Figure 5.3 Folding of RTK KIT. (**A**) The time-related evolution of the secondary structures of the entire, full-length KIT and per domain/region, as assigned by the define secondary structure of proteins (DSSP) method[262]: α-helices in red, 3_{10}-helices in blue, parallel β strands in green, antiparallel β strands in dark blue, turns in orange, and bends in dark yellow. The three cMD replicas (1–3) were analysed individually. (**B**) The 3D structure of KIT is shown by superimposition of the final conformation of the TK domain (t = 2 μs) of each trajectory. (C) The secondary structures—αH- (red), 3_{10}-helices (light blue), and β-strands (dark blue)—assigned for a mean conformation of every MD trajectory (1–3) and the crystallographic structures 1T45 and 1PKG. (D) The secondary structures—αH- (red) and β-strands (dark blue)—assigned on the mean conformation of the concatenated trajectory are labelled as in [291].

The particularly large range of $C_{KID}\cdots C_{C\text{-}tail}$ distances with two well-resolved and considerably distant apexes at 12 and 23 Å and reflects the distinct position of the C-tail, with respect to KID, effect that was observed in [183]. Using the data generated during three 2-μs cMD simulations, we found that C-tail is positioned in proximity to KID and C-lobe, at approximately equal frequency—in 40 and 46% of conformations, respectively. The other conformations (14%) are positioned between these two border locations.

A low variance of distances between the centroids of KID and the TK domain (34–38 Å) suggests that the different positions of KID, derived mainly from its turn round (twist), relative to the TK domain and not from a linear displacement. As the αH1-helix is the most structurally stable element of the disordered KID[290], it can be assigned as a hint (reference) for the characterisation of the KID position in the KIT. The KID displacement relative to the TK domain was measured with the distance between the Cα-atoms of Y703 (αH1-helix) and Y774 (αE-helix), as well as the bending angle defined by the vectors coinciding with the αH1- and αE-helix axes. These metrics describe, in a first approximation, two main types of motion—linear (translation) and angular (rotation).

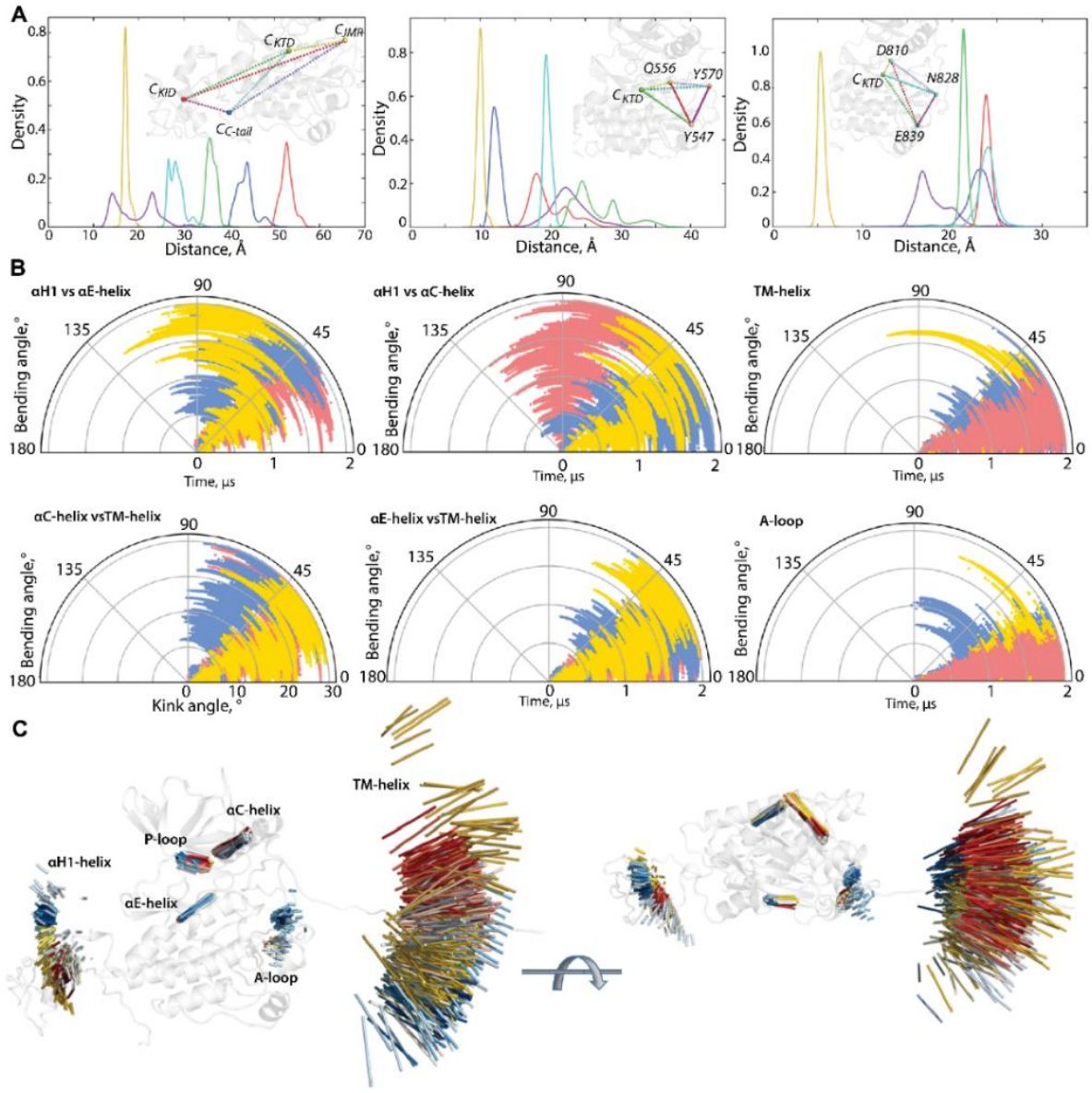


Figure 5.4 Geometry of KIT conformations from the cMD trajectory. (**A**) Geometry of the tetrahedrons with nodes designed on centroids (C) of the KIT domains or the C-atoms of residues, as shown on inserts. (**B**) Curvature between the KIT helices, TM helix, and curvature of the β-hinge of A-loop. Calculations are performed after least-square fitting of the data on the kinase domain. Conformations from different trajectories are distinguished by colour: red (1), blue (2), and yellow

(3). (**C**) Positions of hints—TM-helix, αC-helix, αE-helix, αH1-helix, P-loop, and β-hinge of A-loop—each taken 10 ns, are superimposed on the mean conformation of KIT, calculated on the concatenated data. The protein is shown as cartoon, each hint presented by an axis of helix or vector collinear with a β-strand. Two orthogonal projections are shown. All calculations are performed on cMD conformations, each taken 10 ps from the trajectories distinguished by colour—red (1), blue (2), and yellow (3) in (B,C). The colour gradient shows the evolution of a trajectory, from light (t = 0 μs) to dark (t = 2 μs).

During the first 0.5 μs of simulations, the distance between the tyrosine residues of αH1- and αE-helix varies slightly around their mean values (mv of 25–27 Å), while the bending angle between the αH1- and αE-helix shows quite different angular positions, from parallel to orthogonal (**Figure 5.4**, **B**; **Figure S2**). Obviously, the main movement of KID, relative to the TK domain, is rotation, while its linear displacement is rather limited. The measured metrics, distance, and bending angle do not correlate with each other.

The distance between the centroids of TKD and JMR ($C_{TKD}\cdots C_{JMR}$) is a very approximate metric to estimate the position of the long and flexible JMR, composed of multiple segments—the JM-Proximal (JM-P), JM-Buried (JM-B), JM-Switch (JM-S), and JM-Zipper (JM-Z)—functionally specific and, in the 3D structure of KIT, are largely distant [114].To improve the accuracy of the characterization of the JMR positions, we calculated the distance between the TK domain (CTKD) to the selected residues of each JMR segment—Y547 (JM-P), Q556 (JM-B), and Y570 (JM-S)—except for JM-Z, which is adjacent to the TK domain N-lobe, while maintaining its position (**Figure 5.4**, **B**). The distances between JM-P and CTKD, and JM-P and the other JMR fragments, show the most variations which suggest more than one JM-P positions, a typical manifestation of structural disorder. Likewise, the geometry of the tetrahedron, determined on CTKD and the residues of the extended A-loop, shows a large displacement of the β-hairpin (represented by N828), while the two ends of A-loop (calculated on D810 and E839) changed slightly. Displacement of the β-hairpin, with alternating secondary structures and a wide range of A-loop curvature, reveals heterogeneous conformations of A-loop, indicating its partial disorder.

The transmembrane (TM) helix conserves its perfect α-helical structure but displays a change in its orientation in space. To find out if pivot spaces were preferred, the distribution of the bending and kink angles[292] were obtained (**Figure 5.4**, **B**). We found that the TM helix pivots within a space defined by a cone with an apex angle of 45°. Such movement of the TM helices is frequently observed for membrane proteins in a crowded environment[293,294].

The geometry of the multidomain cytoplasmic region of KIT can be described as relative to the structurally conserved TK domain, through the bending angles between the representative fragments, taken as hints. The bending angles show the abundant diversity of the TMD orientation, in respect to the TK domain, which is richer for the

N-lobe (TM-helix versus αC-helix) than the C-lobe (TM-helix versus αE-helix) (**Figure 5.4**, **B**). As was expected, most variations of the bending angle are observed for KID.

Finally, the positions of the selected hints (TM-helix, αC-helix, αE-helix, αH1-helix, P-loop, and β-hinge of A-loop), superimposed on the mean conformation of KIT, show their relative dynamical positions during the cMD simulations, reflecting the generic geometry of KIT (**Figure 5.4**, **C**; **Figure S3**).

5.1.2.4. Inter-Domain, Non-Covalent Interactions of Kit in Inactive State

We have suggested that the positions of JMR, KID, and C-tail, observed in KIT conformations, are controlled by non-covalent interactions, primary by H-bonds. This hypothesis is based on the analysis of non-covalent contacts stabilising the crystallographic structure of the kinase domain of the inactive KIT[121], which evidenced that the abundance of H-bonds is the most important factor providing the directional interactions that underpin the protein structure (**Figure S4**). Intra-domain interactions mainly contribute to the stabilisation of the β-sheet in N-lobe (H-bonds) and coiled-coil structure in the C-lobe (principally, hydrophobic forces).

For our analysis, in each MD conformation of KIT, the H-bonds involved in the formation of regular structures (helix or sheet) and those contributing to the intra-domain framework (e.g., the β-sheet and helix motifs) were excluded. The remaining H-bond contacts represent an H-bonding pattern of multidomain KIT that displays the interaction between distinct domains (**Figure 5.5**). In terms of protein topology, the H-bond pattern represents interactions between residues that are either close-positioned in the sequence (e.g., A- and C-loop) or distant (e.g., JMR and A-loop) but are neighbours in three-dimensional structure. In terms of strength, these interactions are strong and long-lived (regular) or correspond to weak and unsteady contacts. Some contacts, both regular and rare, are multiple and frequently represent bifurcated H-bonds or involve pairs of adjacent residues forming side (parallel) interactions.

By focusing on the functionally significant regions contributing to the activation/deactivation mechanism of KIT, we found that the extended JMR forms multiple and regular H-bonds, involving its different segments interacting with distinct fragments of kinase domain—the JM-Proximal segment (JM-P, T544–D552) with the antiparallel sheet of A-loop (H-bond K546···N828), JM-Buried segment (JM-B, Y553–V559) with αC-helix (H-bond Y553···E640), JM-Zipper (JM-Z, residues D572–K581) with the αC-helix (H-bond P577···S645), and JM-Switch (JM-S, V560–I571) with C- (H-bond W557···H790 and bifurcate H-bond H790···Q556···R791) and A-loop (H-bond Q556···F811). Similarly, the catalytic (C-) and activation (A-) loops are tightly coupled by the salt bridge (D792···R815) and H-bonds D792···Y823. The position of the P-loop in the N-lobe of the kinase domain is maintained by the H-bond P600···K626.

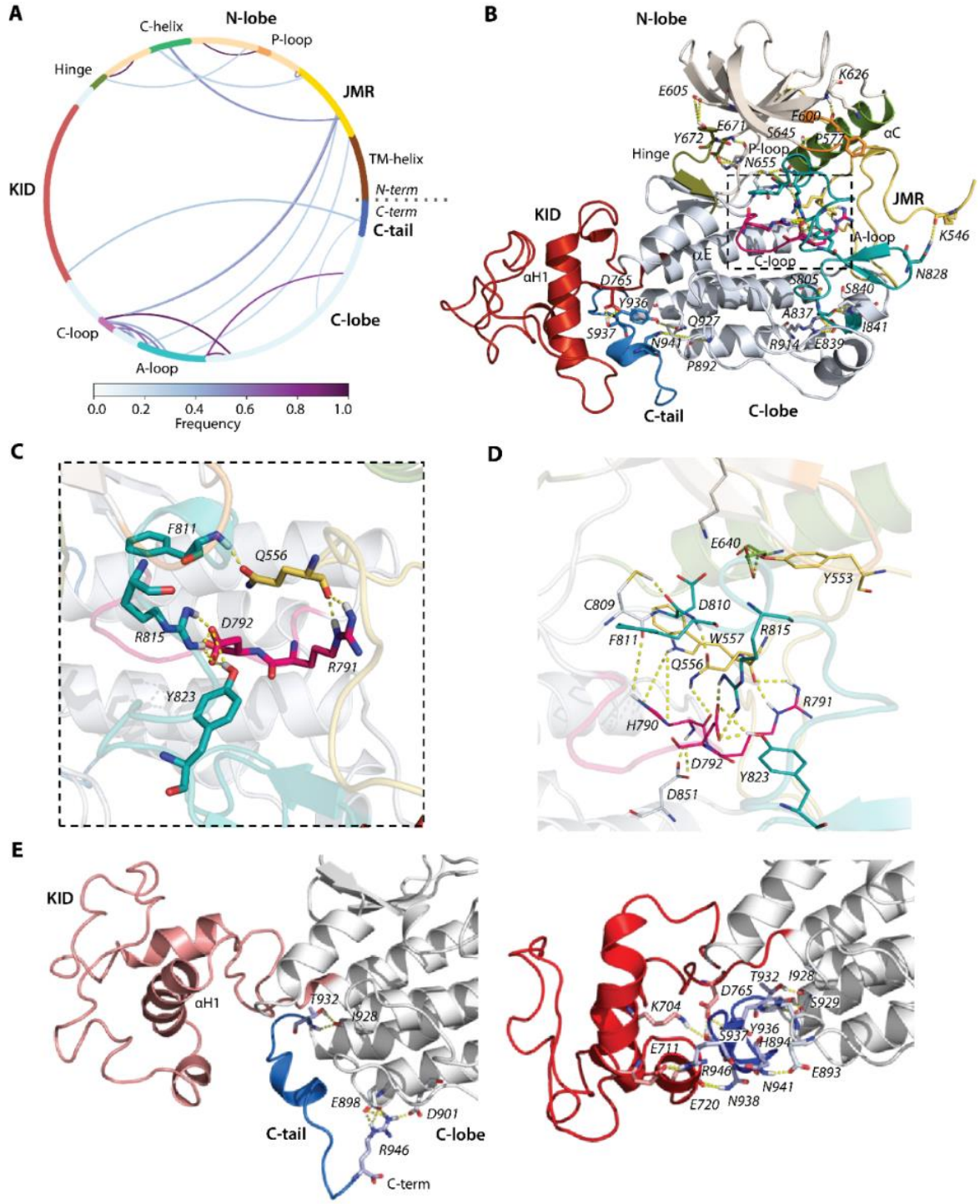


Figure 5.5 Hydrogen bonds stabilizing the inactive state of RTK KIT. (**A**) The cords diagram compiles the H-bonds of multidomain KIT, shown as curves, coloured according to the occurrences, from 0 (white) to 100% (violet). The H-bonds (yellow dashed lines), shown on 3D structure of KIT (**B**), are zoomed on the active site (**C**) and active site with neighbour residues (**D**). (**E**) H-bonds stabilizing C-tail at the TK domain (left) and at KID (right). (**A–E**) The protein is shown as cartoon, in which the domains and functionally related fragments are distinguished by colour and labelled in bold and regular font, respectively. Residues contributing to H-bonds are shown as sticks and labelled in italic; the H-bonds are shown as yellow dashed lines. Calculations are performed on the concatenated trajectories 1–3.

This extensive H-bond pattern of KIT performs crucial functions—(i) maintaining the JMR in its auto-inhibited position[121], (ii) retaining F811 in the active site (the DFG motif, D810-F811-G812)[122] and (iii) coupling of D792 with two residues of A-loop R815 and Y823 that ensures the allosteric communication between A-loop and JMR[14]. Further stabilisation of A-loop in the packed position is provided by the negatively-charged flanking residues from its ends, forming the regular H-bonds with the residues of the C-lobe—the catalytic aspartate of the DFG motif forms H-bond D810···C809, and glutamate constitutes a salt bridge with R914 (E839···R914). The hinge between N- and C-lobe is involved in multiple H-bonds with each lobe. From one side, it interacts with a twisted β-sheet of five antiparallel strands of the N-lobe (**Figure 5.5**, **B**), as well as from the other side its β-strand (β7) is linked with β-strand (β8) of C-lobe, forming the stable antiparallel inter-lobe β-sheet (not shown). Likewise, the inter-domain H-bonds contribute to C-tail stabilisation. As we observed, the cMD conformations are grouped into two large clusters, with respect to the relative positions of C-tail, KID, and C-lobe (**Figure 5.4**, **A**). Except the H-bond, I928···T932 observed in conformations of both clusters, the conformations of these clusters display two binding modes that hold C-tail, either at KID or adjacent to C-lobe (**Figure 5.5**, **C**).

In particular, the C-tail localised at proximity to KID interacts simultaneously with charged residues of KID and C-lobe, serving as a bridge between these two KIT domains (**Figure 5.5**, **E**). For the interaction with KID and C-lobe, C-tail uses alternative sets of residues. C-tail at C-lobe is maintained by the multiple H-bonds of its R946, with the charged residues E898 and D901 in C-lobe. These interactions are probably false, due to the N-terminal status of R946. Phosphotyrosine Y936 does not appear to be involved in H-bonds. Similarly, the primary phosphorylation sites Y568 and Y570 (identified *in vivo*) from JMR apparently do not contribute to the KIT interaction network and are largely exposed with their OH moieties to the solvent. One of the two additional phosphorylation sites found *in vitro* (Y553) is involved in the interaction with αC-helix through its sidechain (**Figure 5.5**, **B**, **D**). Likewise, Y823 of the A-loop interacts with the sidechain carboxyl of D792, stabilising the A-loop near the C-loop and making D792 unavailable for any catalytic process.

The position of the phosphorylation sites in the KIT structure strongly depends on their guest fragments from their conformational features and relative position in KIT. Thus, the large displacement (mainly rotational) of KID from the TK domain is reflected in the expanded distributions of the Cα-atoms position and hydroxyl groups of Y721, Y747, and Y730, located on the highly flexible fragments of the disordered KID, while the OH groups of Y703 in the stable αH1-helix form the narrower cluster (**Figure 5.6**).

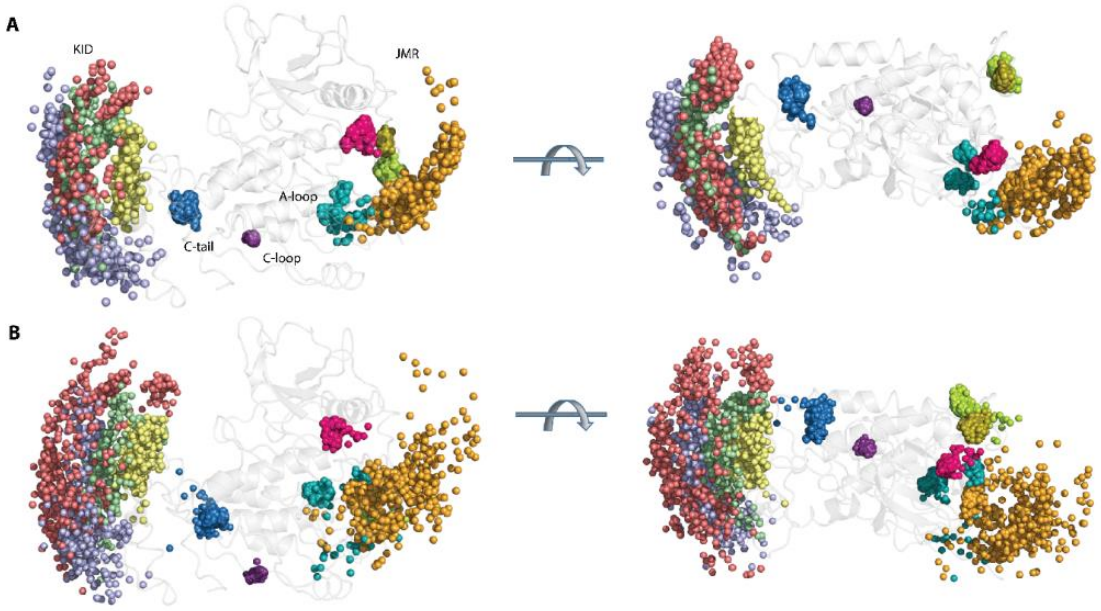


Figure 5.6 Geometry of the tyrosine residues in KIT. The spatial distribution of the Cα-atoms from the tyrosine residues (**A**) and their hydroxyl groups (OH), presented by the oxygen (O) atoms (**B**), is shown in two orthogonal projections with the coloured Cα- and O-atoms: Y547 in orange, Y553 in magenta, Y568 in smith green, Y570 in lime, Y703 in yellow, Y721 in lilac, Y730 in red, Y747 in green, Y823 in teal, Y900 in violet, and Y936 in blue. The Cα- and O-atoms positions were extracted from the MD conformations, taken each 10 ns, fitted on the TK domain of the initial structure (t = 0 ns), and superimposed on this structure (countered cartoon in grey).

Likewise, the compact distribution of the Y553, Y568, and Y570 locations, viewed by the Cα-atoms and the OH groups, reflect the stable position of JMR-B, JM-Z, and JMR-S, while the wide-ranging distribution of the Y547 location corresponds either to the multiple inherent JMR conformations or ample displacement of JM-P, with respect to the TK domain. The three clusters of the unique phosphotyrosine of A-loop, Y823, apparently reflect three different conformations of A-loop.

5.1.2.5. KIT Structure Per Domain, Internal Motion, and Intra-Domain Interactions

To characterize the inherent structural and conformational properties of the variable KIT regions, i.e., JMR, KID, A-loop, and C-tail, and estimate the contribution of these properties to the structural and dynamical relationships of KIT, each of these fragments was analysed individually. First, the conformations of each fragment were grouped by ensemble-based clustering[267] using different RMSD cut-off values, varying from 2.0 to 5.0 Å, with a step of 0.5 Å. With less than 4 Å results in many poor-populated clusters, while a cut-off value of 4 Å was sufficient to regroup the conformations of all fragments into clusters that give the best cumulative population (>95%; **Figure S5**).

The majority of JMR conformations forms the two most populated clusters, C1 (68%) and C2 (31%), composed of the conformations observed in each cMD trajectory (**Figure 5.7**). All JMR conformations show similar secondary structures, a short β-sheet in JM-S segment, and a transient 3_{10}-helix in JM-Z, regularly undergoing reversible folding–unfolding events. The representative conformations of the most populated clusters, 1 and 2, differ only in the position of the JM-P segment containing Y547, which was identified *in vitro* as a phosphotyrosine[119]. Such alternative position of JM-P leads to a large area of the Y547 location, while the other tyrosines are almost superimposed (**Figure 5.6**; **Figure 5.7**).

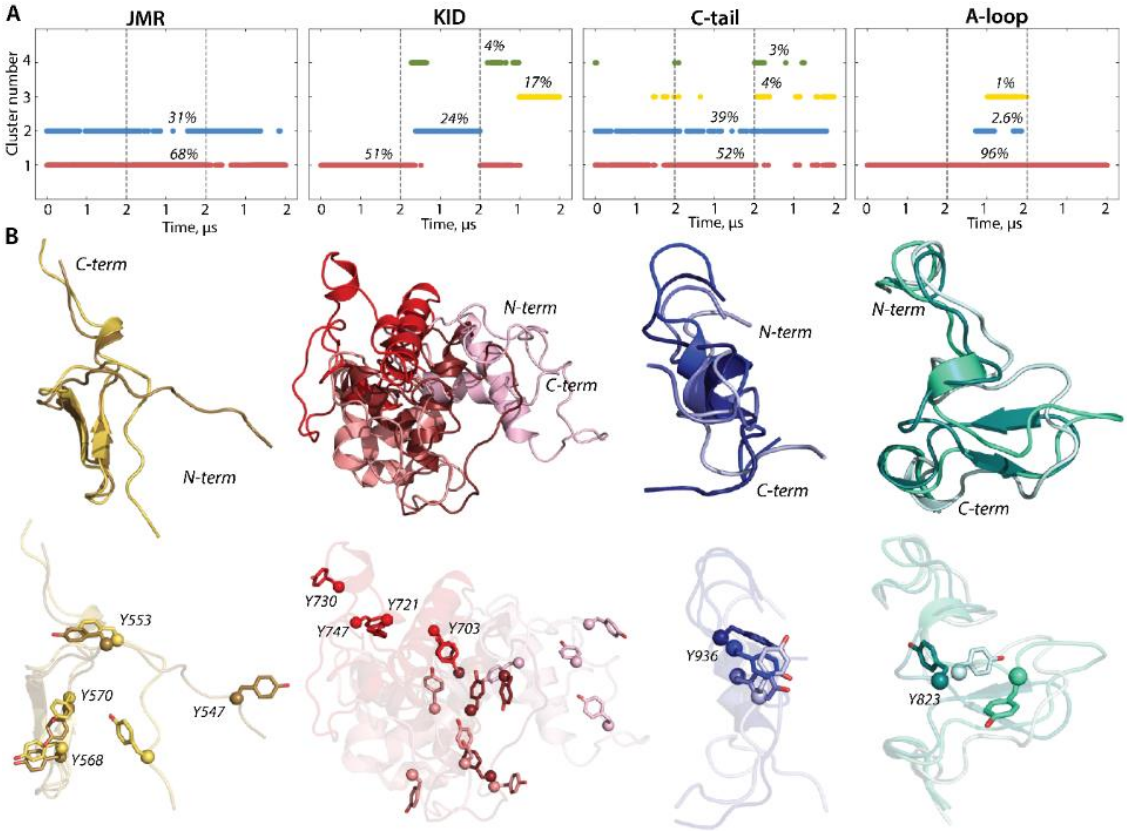


Figure 5.7 Structure and conformation of the disordered fragments of KIT—JMR, KID, A-loop, and C-tail. (**A**) The clusters of conformations, obtained by ensemble-based clustering (cut-off 4 Å) and their population. (**B**) Superimposed representative conformations from the clusters. The protein is shown as cartoon, with the tyrosine residues as sticks. The colour gradient shows the population of clusters, from dark (most populated) to light (less populated). Calculations were carried out on cMD conformations, taken every 100 ps from the concatenated trajectories after the fitting on the initial conformation of the TK domain (at t = 0 μs).

The conformations of the intrinsically disordered KID, previously characterised in [290], were grouped into four clusters. Two clusters, C2 (24%) and C3 (17%), are composed of conformations generated over the alone trajectory, 2 and 3, respectively, while the most populated cluster C1 (51%) and small cluster C4 (4%) are composed of conformations from at least two independent trajectories. The representative

conformations from the distinct clusters differ at the folding level (2D) and in 3D structure organization, reflecting a high level of intrinsic disorder in KID. The ample rotation of KID, with respect to the TK domain and its large conformational diversity, leads to dispersed location of the tyrosine residues.

The conformations of C-tail are grouped in the two most populated clusters, C1 (52%) and C2 (39%), and two lowly populated clusters, C3 (4%) and C4 (3%). The representative conformations of C1, C2, and C3 show similar secondary structures, described as an extended random coil with a small transient helix in the middle (α-helix → 3_{10}-helix), and differ mainly by the orientation of the C-terminal residues. All clusters regroup conformations generated over the three independent trajectories, and, apparently, C-tail secondary structures do not influence a cluster separation. Indeed, the small transient helix is only observed in trajectories 1 and 3, while the conformations from trajectory 2 are folded as a random coil (**Figure 5.3**). The tyrosine residue Y936 shows a close position and orientation of its OH group in most conformations (1–3 clusters) and is different in the only rare intermediate conformations.

5.1.2.6. Intrinsic Motion in KIT and its Interdependence

The intrinsic dynamics of the multidomain KIT was first analysed with the cross-correlation matrix, computed for the Cα-atom pairs of the full-length protein. As the matrices calculated for the three MD trajectories are very similar (**Figure S6**); therefore, we will discuss only one, randomly chosen, matrix map.

The Cα–Cα pairwise, cross-correlation map demonstrates highly coupled motions within each KIT domain and between the structural domains, even largely distant (**Figure 5.8**, **A**).

The patchwork pattern in N-lobe reflects the positively correlated motion of the seven stands in the crossed β-pleated sheet and their coupling with αC-helix. Similarly, the helices of C-lobe are mutually correlated (positively), forming a map of well-defined blocks, distorted by A-loop. Both TK lobes correlate negatively with the KID and TM helix and positively with JMR. C-tail correlates positively with C-lobe and negatively with N-lobe. The inter-lobe correlations are not homogeneous, and, apparently, P-loop and αC-helix of N-lobe play a specific role in the motion correlations. The motions of TM helix and JMR are contrariwise.

Such correlation patterns can be partially explained by the overall architectural features of the studied KIT, which has a strongly extended shape. Movements of one end (TM helix) are counterbalanced by movements of the opposite end (KID) to provide a stable balance of KIT around its centre of gravity. On the other hand, the highly coupled motion in KIT may reflect the allosteric effect(s) and has the functionally

related content—regulation function. In both cases, the cross-correlation pattern indicates the strongly coupled movements of the largely distant domains/regions and reflects the block-segmented movement of the multidomain KIT. The similar pattern, even more pronounced and contrasted, was observed on the cross-correlation matrix (**Figure 5.8**), calculated using the normal mode analysis (NMA), calculated on the mean KIT conformation of each trajectory with force field for Cα-atoms, developed by [295]. Moreover, the NMA cross-correlation maps are very comparable for the three mean conformations and calculated using different force fields [266,296], showing a great reproducibility of the results (**Figure S7**).

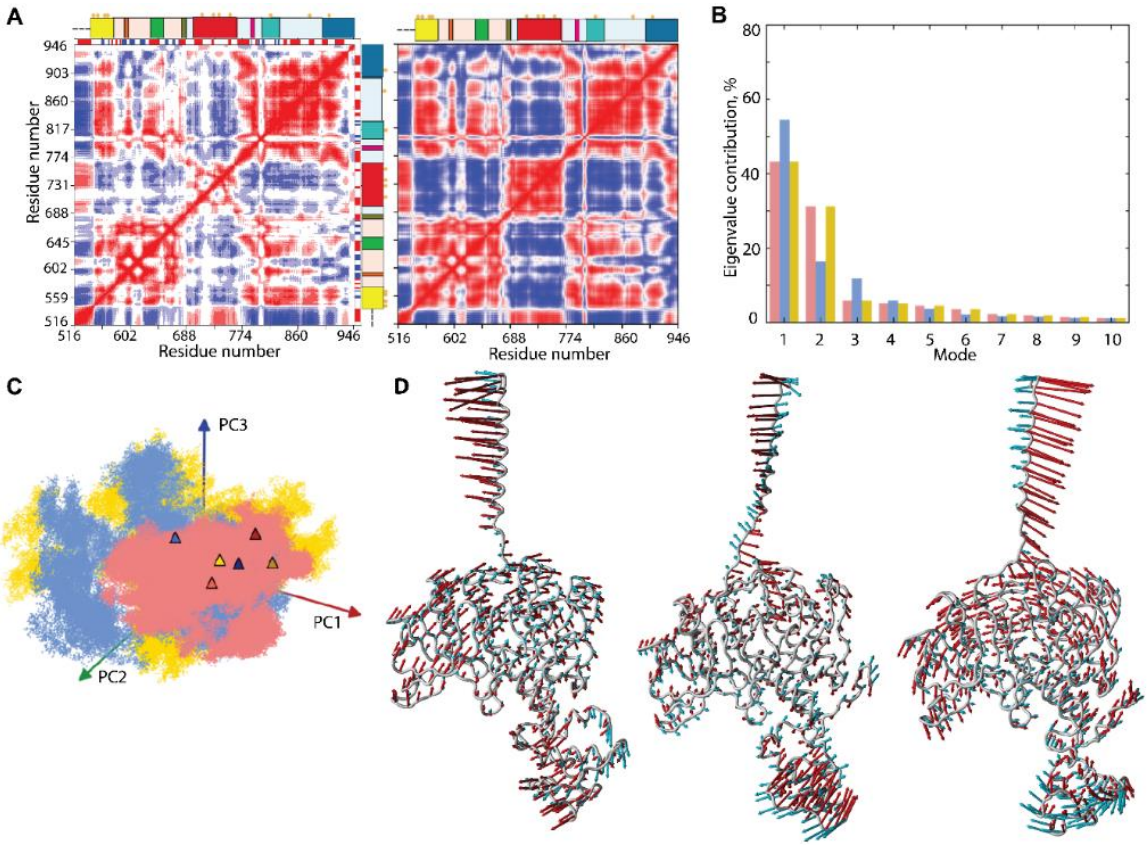


Figure 5.8 Intrinsic motion in KIT and its interdependence. (**A**) Dynamical inter-residue, cross-correlation map, computed for the Cα-atom pairs of MD conformations (left) and resulting from NMA (right) of KIT. The displayed results represent trajectory 1. Correlated (positive) and anti-correlated (negative) motions between Cα-atom pairs are shown as a red-blue gradient. (**B**) PCA modes calculated for KIT after least-square fitting of the MD conformations to the mean conformation. The bar plot gives the eigenvalue spectra, in descending order, for the first 10 modes calculated on cMD trajectories 1–3 (left). (**C**) Projection of the KIT cMD conformations onto the first two modes, calculated with principal component analysis (PCA) (right). MD trajectories 1, 2, and 3 denoted in red, blue, and yellow, respectively. Light and dark symbols display the first and the last conformations for each trajectory. (**D**) Atomic components in PCA modes 1–2 are drawn as red (1st mode) and cyan (2nd mode) arrows, projected on the cartoon of KIT. A cut-off distance of 4 Å was used.

The collective motions of KIT, characterised by PCA, showed that two ten modes describe ~70–80% of the total fluctuations of the multidomain KIT (**Figure 5.8**, **B**). The two first modes clearly reflect a highly coupled motions of the multidomain KIT. The great mobility of the TM helix is gradually increased from its C- to N-ends. The amplitude and direction of motion of the TM helix and KID differ in the three trajectories (**Figure 5.4**; **Figure 5.8**, **D**), suggesting a larger conformational space for the KIT than was observed in each trajectory, probably larger than the total space of all trajectories. The TK domain global motions in KIT demonstrate a lower amplitude, with respect to that of TM helix and KID, as that the motions of all TK residues are collective and may be described as a circular pendulum-like movement along a common virtual rotational axis. Interestingly, the virtual axis is coincident (centred) with the active site of KIT.

The projection of the cMD conformations on the first two PCA modes revealed that (i) the conformational density is extended in 1 and 2 trajectories, as well as a more compact in replicate 3; and (ii) the superimposed KIT conformations of the three cMD simulations provide a wide coverage (overlap) of PC subspaces (**Figure 5.8**, **C**).

5.1.2.7. Conformational Space of KIT and its Representation as Landscape

Since the RTK KIT contains at least four disordered fragments, showing reversible protein folding–unfolding events and a large conformational diversity, its conformational space can be represented more explicitly as a 'free energy landscape' model. Such interpretation of the intrinsically disordered multidomain protein leads to quantitatively significant results, allowing the comparison between different states. The relative Gibbs free energy, (ΔG, defined on chosen coordinates called 'reaction coordinates' or 'collective variables', describes the conformations of a protein between two or more states, measured as the probability of finding the system in those states. Such representations of protein sampling, with the use of reaction coordinates, can be the quintessential model system for barrier crossing events in proteins[297]

For the evaluation of the relative free energy (ΔG) and reconstruction of its landscape, the distant measures—radius of gyration (Rg), distance (RMSD), and the PCA components (PC1 and PC2)—were used as reaction coordinates for the description of the (ΔG landscape of KIT. The free energy landscape (FEL), as a function of RMSD and radius of gyration Rg (FEL_{Rg}^{RMSD}), as well as the PCA components (FEL_{PC2}^{PC1}), calculated for the concatenated replicas of KIT, is shown in **Figure 5.9**.

Each FEL shows a rugged landscape, reflecting a high conformational heterogeneity of the multidomain KIT. This heterogeneity adds complexity to the interpretation of the free energy map and limits the detection of spontaneous state-to-state transition, when using the conventional sampling method (cMD).

Nevertheless, both FEL of KIT show well-defined areas of minimum energy indicated in red, which represent more stable conformations (the thermodynamically more favourable state), while the reddened areas indicate conformational transitions of the protein.

The unimodal Gaussian distribution of Rg does not separate the KIT conformations on isolated clusters but delimits the most populated region; the (ΔG minima, showed by lower energy values, are rather defined by the multimodal distribution of RMSDs. The FEL_{Rg}^{RMSD} of KIT has two depth minima, separated from each other by energy barriers of various heights, shown as an additional smooth local minimum. Regarding the FEL_{PC2}^{PC1}, this local smooth minimum is absent because the free energy landscape, reconstructed on the principal components, only represents the conformations reflecting the 'essential' dynamics of protein, which constitutes two almost iso-energetic minima. As the first two PCs only represent movements on a larger spatial scale (larger spatial scale motions), so that the energy landscape, defined on these reaction coordinates, FEL_{PC2}^{PC1}, loses smaller spatial scale motions (e.g., intermediate conformations).

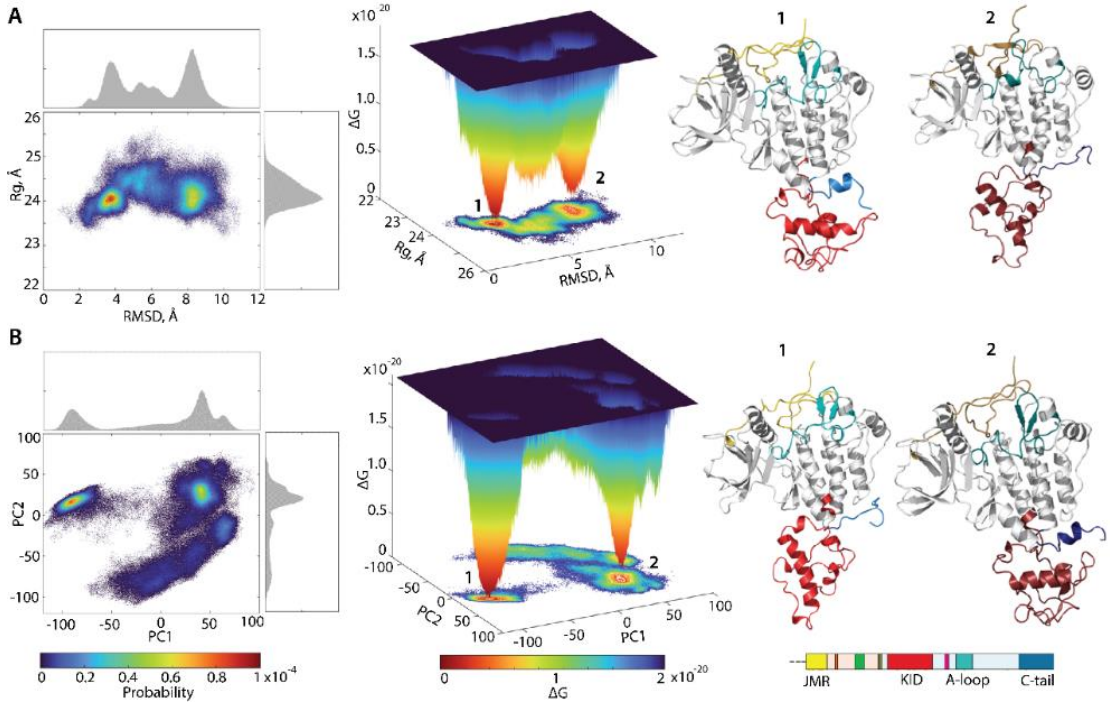


Figure 5.9 Free energy landscape (FEL) of KIT as a function of the reaction coordinates, (**A**) Rg (in Å) versus RMSD (in Å) and (**B**) two PCA components (PC1 versus PC2), was generated on the MD conformations of KIT for the conformational ensemble, sampled on the merged replicas and fitted on the initial conformation taken at t = 0 µs. (Left column) The two-dimensional representation of FEL of the KIT conformational ensembles. Density distribution of each reaction coordinate is shown at the top and right, respectively. (Middle column) The three-dimensional representation of the relative Gibbs free energy. (Left and middle column) The red colour represents high occurrence, yellow and green represent low, and blue represents lowest occurrence. The free energy surface was plotted using Matlab. (Right column) KIT conformations from wells 1 and 2.

Nevertheless, these FELs, reconstituted on two different sets of reaction coordinates, showed that the ensemble of KIT conformations can recover the two-state picture characterized by two global minima (wells); each of them is composed of similar conformations, which are very distinct between the two minima (**Figure 5.9**). The composition of these wells is permuted, due to the different reaction coordinates which lead to the permuted population of two minima: the conformations of the first and second minima on FEL_{Rg}^{RMSD} correspond to the content of the second and first minima on the FEL_{PC2}^{PC1}. The KIT conformations from the two global minima in both landscapes display large conformational alteration in KID and C-tail, which appears to be the main factor in such a two-state landscape.

The low approximate population of conformations in the two deepest wells on the free energy landscape of KIT (16 and 7% on the FEL_{Rg}^{RMSD} and 12 and 11% on the FEL_{PC2}^{PC1} indicated that the great number of KIT conformations are the intermediate between these states. We suggested that the estimation of the contribution of each domain to the total energy landscape of the multidomain KIT (the per-domain relative free energy (ΔG) will complete the reconstructed landscape of the protein. Each domain was considered individually, but as a dependent subdomain of the entire KIT; therefore, the corresponding data were fitted on the initial conformation (t = 0 μs) of the full-length of the cytoplasmic region of KIT.

As was expected, the tyrosine kinase domain showed only one highly populated (54%) and deep well completed by similar conformations, while the second well, if it still exists, is considerably reduced and contains 4% conformations (**Figure 5.10**). The second well may be composed of the KIT showing alternative conformations of its structural fragment, e.g., A-loop. Indeed, the two-state profile of A-loop, with unevenly filled contents of the wells unequally (40 and 4%), is in good agreement with the profile of the TK domain, which confirms our hypothesis.

Surprisingly, the FEL determined on the JMR showed only one highly populated deep well (47%). By comparing this observation with the ensemble-based clustering of JMR (**Figure 5.7**), its impact on the total free energy landscape can rather be attributed to its alternate position, relative to the tyrosine kinase domain, than as an effect of its internal conformational features.

The free energy landscape of KID differs from that reported in [290] and was apparently biased by the false movement of JMR. Indeed, in KIT with the highly flexible JMR with the N-terminal status, the unique global energy minimum of the KID, accompanied by the two local minima showing higher energy values, was observed[290]. The free-energy landscape of KID from KIT, composed of the cytoplasmic domain linked to TMD (present study), shows two deep wells with an almost equivalent population (15 and 14%), supplemented by a series of satellite wells with the essentially lower population (from 0.5 to 2.5%). Such a two-main state profile of KID is presumably the main factor contributing to the two-state profile of KIT.

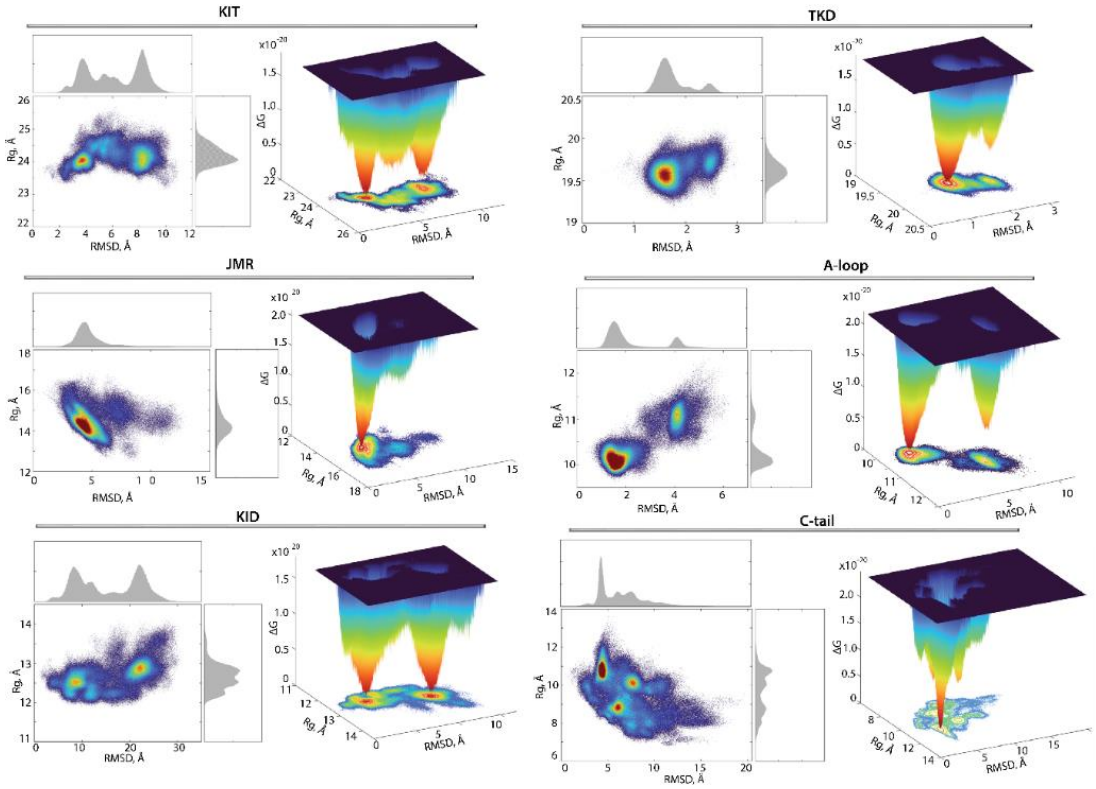


Figure 5.10 Free energy landscape (FEL) of KIT and its subdomains, as a function of the reaction coordinates, Rg (in Å) versus RMSD (in Å). The FELs were generated for each entity using the ensemble of MD conformations, sampled on the merged replicas and fitted on TK domain. (Left) The two-dimensional representation of FEL. Density distribution of each reaction coordinate is shown at the top and right, respectively. (Right) The three-dimensional representation of the relative Gibbs free energy. The red colour represents high occurrences, yellow and green represent low occurrences, and blue represents the lowest occurrences. The free energy surface was plotted using Matlab.

The FEL profile of C-tail shows unique low energy deep minimum, which is complemented by two wells with the higher energy values. As C-tail conformations have similar secondary structures and differ mainly in the orientation of the C-terminus. These intrinsic properties do not contribute significantly to the FEL, which is primarily impacted by the alternate location of the C-tail, with respect to KID and/or to TK domain.

5.1.3. Discussions

Remarkably, the multidomain modular RTK KIT consists of a quasi-stable TK domain, crowned by at least four intrinsically (inherently) disordered (ID) domains/regions—JMR, KID, A-loop, and C-tail. These disordered regions contain functional tyrosine residues and act as a ground platform for the recruitment of signalling proteins. The implication of disordered regions in post-translational

modifications is well-known[298–300]. Nevertheless, despite their fundamental role in mediating the signalling cascade, the intrinsically disordered domains of KIT (and other RTKs) are critically under-studied, leaving us with a naïve, over-simplistic, and rather schematic view of these domains as structurally impersonal chains of varying lengths, enabling conformational adaptability to ensure constitutive attachment of specific protein partners.

This paper focuses on the functionally significant domains in KIT signalling and their disorder, described at two levels, such as the intra-domain intrinsic disorder (the local disorder) and extrinsic disorder (inter-domain), observed at the KIT level.

Since KIT ID domains have no single, well-defined 3D structure, they cannot adequately be described as simple statistical coiled or helically folded chains equally populating all MD conformations allowed by their backbone torsion angles. Instead, KIT ID domains contain transient, short- and long-length structures, which display various degrees of compaction and elongations. Therefore, each ID domain of KIT is not homogeneous, but represents a very complex mixture of a broad variety of differently folded conformations, ranging from the partially folded to fully unfolded, which, in turn, are foldable. In KIT, these regions are mainly composed of polar and charged residues, while the portion of hydrophobic residues is reduced, presenting an archetypal sequence composition of the intrinsically disordered proteins[301,302].

The ID regions of KIT belongs to two types—the very elongated (extended) and poorly folded regions (JMR, A-loop, and C-tail), as well as the globule-like (collapsed) KID, with a high level of the helical structures.

The high content of unfolded residues (random coil) in JMR and A-loop implies a potentially high degree of disorder in these regions. However, since they, in the inactive state of KIT, are attached to the TK domain by many H-bonds, anchoring each end of A-loop and bind the JMR along most of its length, their disorder is limited. Only short segments of these regions, not stabilized by H-bonds, show local disorder, which is evidenced by the alternating positions of JM-P on JMR and unfolding of the β-hinge in A-loop. Since the disordered segment of JMR contains the functional tyrosine Y547, this residue is highly dislocated in 3D space over a wide range and, therefore, is accessible for phosphorylation events, even in inactive KIT or its intermediate state. Y823 in A-loop is also irregularly positioned in space, located in at least in three different positions, but its side chain is still oriented towards the active site of KIT.

We suggest that, upon activation of KIT, induced by SCF binding, JMR displaced from its packed auto-inhibited position will achieve the most extended and flexible conformation, so that its level of disorder will increase, and, therefore, the level of its adaptability required for scaffolding (docking sites) and recruitment of different protein binding partners will also increase.

The position of Y936 and orientation of its phosphorylable sidechain is highly conserved in all conformations of the extended C-tail, regardless of the location of its disordered C-end, either near KID or close to the TK domain. It may well be that these results do not fully reflect the role of the C-end; in our model, it was shortened, with respect to the native length.

The higher portion of charged and polar residues in KID, compared to JMR and A-loop, is reflected in the enhanced disorder of KID, which involves almost all residues of the domain, except of a single helix (αH1), which shows stable folding (secondary structures) but varies greatly in its orientation, respective to the TK domain. The structure of the αH1-helix is conserved, not only in the KID fused to the tyrosine kinase domain but also in KID simulated as a cleaved polypeptide[290].

The KID disorder is manifested first as transient folding (secondary structures) and presented as a collection of highly variable helices, permanently converting between α- and 3_{10} helices, as well as between a helix and non regular structure (turn, bend, and coil), whose length and type change. Second, as KID helices are connected by flexible linkers of varying length, depending on the order of the helical folding, this promotes many relative orientations of the helices, leading to a large set of very heterogeneous conformations. These two factors, transient folding and high conformational flexibility, are the main items characterizing the intrinsic intra-domain (local) disorder of KID. Nevertheless, as we have reported [290], the intrinsically disordered KID acquires a globular shape, stabilized by non-covalent interactions—H-bonds and hydrophobic forces. Hydrophobic contacts are compiled, as a well-organised hydrophobic core maintaining KID compactness.

Moreover, this compact globule-like domain is displaced (linear and angular displacement) as a 'pseudo-rigid' body, with respect to the TK domain, such representing the inter-domain extrinsic disorder. This movement is also highly heterogeneous, both in space and time, with a variable contribution of the linear and angular components, manifested in the form of small dislocations and large-scale rearrangements. This all-scale movement can involve distinct segments of KID, again showing the highly anisotropic nature of the intrinsic disordered in this domain.

The two-level disorder (intrinsic and extrinsic) provides the high conformational variability and adaptability of JMR, KID, and C-tail required for the scaffolding (docking sites) and recruitment of different protein partners of KIT and facilitates the regulation of cellular processes. The overall structure of KIT represents a continuous spectrum of conformations, with a different degree and depth of disorder, as was reported for the other functional proteins[301,302], thereby generating a complex protein structural space that defines a structure–disorder continuum, with no clear boundary between ordered and disordered proteins/regions. The disorder of at least four regions of the multidomain KIT is reflected in its free energy landscape, which lacks a unique global deep minimum that can be found in ordered proteins and appears as two local minima

joined/separated by a 'flattened plateau' containing the intermediate conformations.

Classically, multiple binding events in ID KIT, regulated by phosphorylation, can be characterized as binary on/off switches. However, it has been reported that phosphorylation can generate more complex responses[303], and multi-site phosphorylation can additionally generate sensitive threshold responses, as well as graduated responses. Therefore, successive phosphorylation events can additively modify (increase or decrease) the binding affinity, allowing for graduated responses, or they can modulate the conformational set, with an impact on signalling output.

Such simplified and flattened energy landscape has shown that KIT is extremely sensitive to different environmental changes (e.g., phosphorylation) that can alter its free energy landscape in different ways, e.g., lowering some energy barriers, while raising certain energetic minima. This explains the conformational plasticity of ID regions, allowing them to evolve faster than protein domains that adopt defined stable structures, and its ability to interact with several different partners and, therefore, fold in different ways.

Since ID domains are multiple in RTK KIT, we asked, does the disorder/order of one domain depend on the disorder/order of other remote regions? In other words, we wish to understand whether the folding–unfolding process of the intrinsically disordered domains in inactive KIT is orchestrated or not. Does the allosteric regulation of KIT involve the disordered domains/regions?

Highly coupled motions between distant sites of KIT, as evidenced by the cross-correlation maps, suggests its association with the functional dependence of these regions, which is classified as allosteric regulation, the phenomenon largely observed in many proteins[88,95–97]. In particular, the coupling motions in each lobe, N- and C-lobe, of the TK domain and between the lobes reflect the allosteric regulation of the kinase function, which is well-described for different non receptor and receptor tyrosine kinases[96,105,107]. The coupled motions of two activating regions of KIT, A-loop, and JMR were characterized, in terms of their allosteric communication in the wild-type KIT, which was disrupted in oncogenic mutants[14,108].

The mechanism of regulation of the RTK regions directly involved in cell signalling is still a matter that is little studied, due to the absence of structural data characterizing these regions.

To the best of our knowledge, we have presented, for the first time, a model of a full-length cytoplasmic region of an RTK KIT attached to a transmembrane helix and its molecular dynamics simulations, under conditions that mimic the natural environment of the KIT. We demonstrated the tight coupling between the KIT remote regions populated by phosphotyrosines (JMR, KID, and C-tail), which serves as a scaffold for the recruitment and activation of signalling proteins.

The classical molecular models, the Koshland-Nemethy-Filmer[100] and Monod-Wyman-Changeux[99] paradigms describing the allostery, 'a second secrete of life'[94,304] take, as a physical basis, the conformational changes between well-defined structural states of proteins but do not consider other factors, such as conformational dynamics, monomeric-oligomeric states, intrinsic disorders, and negligible conformational changes[102]. Recent empirical observations, demonstrating that allostery, can be facilitated by dynamic and intrinsically disordered proteins[97,103,109] and have resulted in a new paradigm to understand allosteric mechanisms, which focuses on the conformational ensemble and statistical nature of the interactions responsible for the transmission of information[87,103].

The manifestation of allostery in intrinsically disordered proteins (IDP) is one of the most sophisticated phenomena observed in the last decade[96,302,305–308]. The conformational dynamics of folded structures and large-scale disorders are important for allostery[87,101], but the quantitative understanding of this phenomenon remains a great challenge. It is not easy to understand and describe the allosteric phenomenon with a common point of view, encompassing both highly structured and disordered systems. Moreover, it was suggested that the intrinsic disorder of the RTKs renders not only infer flexibility and high accessibility of binding sites, but certain chain dimensions and spatial organizations may influence the organization of the signalling complexes, orchestration of protein interactions and, in the end, signalling outcomes[298].

Following this concept, instead of considering the ID domains of KIT as passive scaffolds for its protein partners, we put forward a more complex view of active orchestration via organizational and operational features left uncovered within their disorder. Moreover, we suggest that all properties of the activated RTK KIT and post-transduction events, initiated by the active KIT, are encoded in its inactive state.

5.2. ORDRE ET DESORDRE DE HVKORC1

Résumé. *La vitamine K époxyde réductase humaine (hVKORC1) est une protéine transmembranaire du réticulum endoplasmique convertissant la vitamine K sous sa forme 2,3-époxyde en vitamine K hydroquinone impliquée dans l'activation des protéines dépendantes de la vitamine K. Ce mécanisme est initié par hVKORC1 activé. Cette activation est dépendante de la réduction de deux ponts disulfures : entre deux cystéines de la boucle luminale (boucle L) par une protéine redox partenaire, puis entre les cystéines du motif CX_1X_2C du site actif par les cystéines précédemment réduites. L'activation initiale, par la protéine redox, est l'une des étapes la moins étudiée dans le cycle du hVKORC1, et cette réaction spontanée est dépendante de l'adaptation de leurs configurations. Par simulations de dynamique moléculaire, nous avons étudié le repliement et la plasticité conformationnelle de hVKORC1 à l'état inactif*

(entièrement oxydé) en utilisant les structures cristallographiques et le modèle de novo *disponibles. La boucle L possède un repliement composé d'hélices α et* 3_{10} *transitoires et une forme principalement « fermée ». Par cette description, nous avons montré que le hVKORC1 oxydé est une enzyme composée d'un domaine transmembranaire stable et d'une boucle luminale désordonnée. Nous avons suggéré que, même si le niveau de désordre de la boucle L est limité par un pont disulfure, la boucle L possède une plasticité suffisante pour sa reconnaissance par une protéine redox partenaire.*

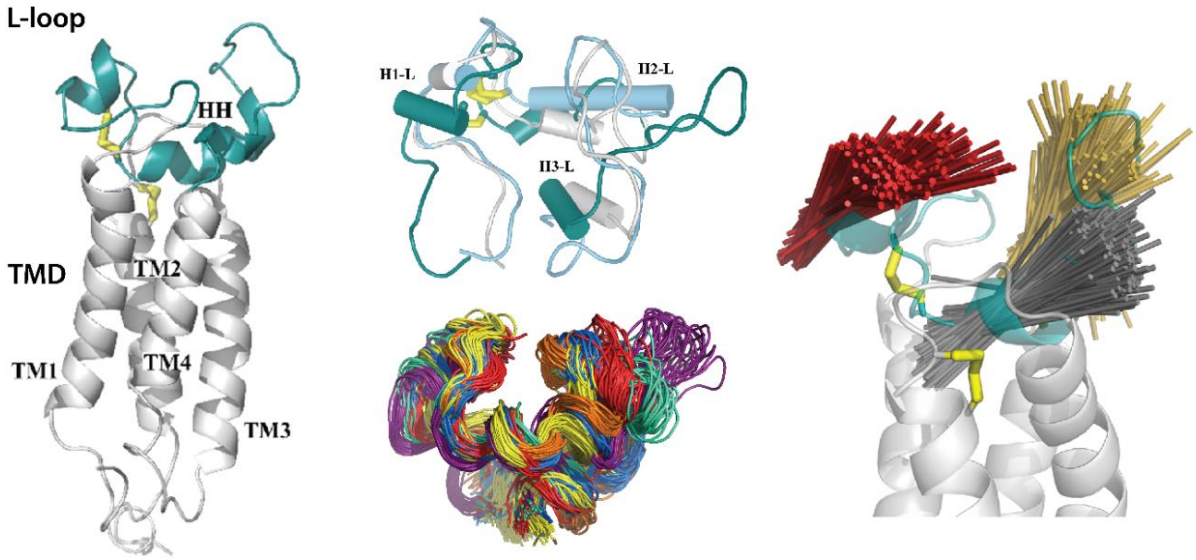


Figure 5.11 Abstract graphique de la section.

5.2.1. INTRODUCTION

In the present study, we focused the vitamin K epoxide reductase complex 1 (VKORC1), an endoplasmic reticulum-resident transmembrane protein that is responsible for the activation of vitamin K-dependent proteins, and it is involved in several vital physiological and homeostasis processes[211]. Vitamin K (K from *koagulation*, German), a natural K-vitamer generated by plants and preferentially transported to the human liver[309], is required for post-synthesis modification of proteins involved in blood coagulation (blood pro- and anti-clotting enzymes, e.g., endothelial anticoagulant protein S) as well as proteins outside the coagulation cascade[310]. Recently it was reported that vitamin K levels appear to be extremely reduced in the lungs of hospitalised COVID-19 patients and particularly in those who need mechanical ventilation in intensive care unit and/or died[203]. It seems therefore that activation of endothelial protein S in these patients is more severely compromised than activation of hepatic procoagulant factors[311] and is compatible with enhanced thrombogenicity in COVID-19[205].

Within cells, vitamin K participates in a cyclic process initiated in the endoplasmic reticulum (ER) lumen through reduction from the inactive vitamin K 2,3-epoxide to active vitamin K quinone by enzyme vitamin K epoxide reductase (VKOR)[225,312]. After each vitamin K-dependent protein activation, vitamin K is recycled to the initial inactive state.

Although different hypotheses based on biochemical, biophysical, *in silico* and crystallographic studies have been proposed, explicit mechanisms of vitamin K reduction in mammals, in particular human VKOR (hVKORC1), are not currently well defined. Each suggested mechanism postulates that redox proteins transfer electrons to two conserved cysteines in a luminal loop (L-loop) of hVKORC1. Further, these electrons are transferred to a CXXC motif in the enzyme transmembrane domain (TMD)[224]. Finally, the reduced CXXC motif of hVKORC1 transfers electrons to vitamin K[211,225]. Each step of vitamin K reduction is tightly coupled to the motif CXXC oxidation in hVKORC1 active site. To repeatedly reduce vitamin K, hVKORC1 must be regularly activated by a redox partner delivering reducing equivalents through thiol-disulphide exchange reaction. Cooperation of hVKORC1 with a redox protein implies an activation process that represents a less studied step in hVKORC1 vital cycle.

VKOR is the target of oral anticoagulants (VKAs) like warfarin, which dampens coagulation by limiting the supply of vitamin K. Its functional role is a catalyst in the reduction of vitamin K, requiring cooperation with a redox partner that delivers reducing equivalents. A particularly interesting problem is the enzymatic activation of hVKORC1 by the thiol–disulphide exchange. This process involves molecular recognition at the highest level required for proton-transfer reactions.

Recently, 3D models of human VKORC1 (hVKORC1) have been reported along with functionally related enzymatic states[220]. The models that were generated for the metastable states of hVKORC1 and their validation through *in silico* and *in vitro* screening have led to a conceptually plausible mechanism for enzymatic reactions based on a sequence array of hVKORC1-activated states involved in vitamin K transformation.

Consequently, the study of hVKORC1 is pertinent as hVKORC1 has dual interest – as a clinical target for the development of vitamin K reduction modulators, and as an enzyme activated by its redox protein through thiol-disulphide exchange reaction, a biological fundamental process. In the present work, we analysed the structural and conformational properties of hVKORC1 to provide an accurate target model for fundamental research and pharmacological applications, focusing on the exploration of hVKORC1 ability to recognize its redox protein.

5.2.2. RESULTS

5.2.2.1. GENERAL CHARACTERISATION

Human VKORC1 is composed of two domains: the extended luminal loop (L-loop), which contains the cysteine residues that participate in the electron exchange between the redox enzyme and hVKORC1, and the transmembrane domain (TMD), which includes two other cysteine amino acids from the highly conserved CXXC active site that is essential for vitamin K quinone reduction[313,314]. Based on studies of bacterial VKOR homologues, it was proposed that the loop cysteines of hVKORC1 allow protons to be shuttled to the active-site cysteines[224,226].

Earlier, a four-helix transmembrane domain structural model of human VKORC1 in its four functional states was reported[220]. Here, the focus is on the inactive (oxidised) state of hVKORC1, in which two pairs of cysteine residues, C43–C51 and C132–C135, are covalently linked to form disulphide bridges S···S. hVKORC1 was studied by MD simulations of the model that mimics the protein in its natural environment, i.e., hVKORC1 embedded in the membrane and surrounded by water molecules (**Figure 5.12**, **A**). While the extended L-loop (R33-N77 aas) has demonstrated high conformational variability in the protonated forms of hVKORC1[220], the inactive state of hVKORC1 was studied by repeated 500 ns MD simulations (replicas 1–3) using random initial velocities.

The RMSDs computed for the positions of all Cα-atoms relative to the initial structure (t = 0 ns) showed comparable behaviour over the three MD trajectories, with a mean value (mv) of 5 Å (**Figure 5.12**, **B**). The per-domain RMSDs showed that the N- and C-terminals are the fragments that contribute most to large RMSD values (up to 13 Å), while the TMD curves demonstrate a highly stable profile with the smallest RMSDs (2 Å). RMSDs computed for the Cα-atoms of the L-loop, after fitting to its initial conformation, showed alternated values, small or large, that were maintained over a large time scale (50–100 ns). The altered RMSD values, viewed as a set of well-defined slopes, indicate the possible conformational transitions in the L-loop. To check the suggested conformational transitions, MD conformations picked before and after each sudden RMSD change were compared. Three conformations of the L-loop that were chosen from replica 3 at t = 150, 250 and 375 ns showed significant differences in the folding and orientation of the helices and the loops, which revealed structural and conformational transitions (**Figure 5.12**, **E**).

The larger RMSD values computed for Cα-atoms of the L-loop, after rigid alignment based on the initial conformation of the TMD compared to the RMSDs computed after rigid alignment based on the initial conformation of the L-loop, suggest the displacement of the L-loop from the TMD as a pseudo-rigid body (**Figure 5.12**, **C**). The profile of the RMSF curves is similar in the three MD trajectories, with

differences only in the amplitude of the RMS fluctuations of the highly flexible regions of hVKORC1, the N- and C-terminals and the extended L-loop (**Figure 5.12**, **D**). The 2D and 3D structures of hVKORC1 is generally conserved over the MD trajectories and shows a fully helical fold of the protein, with the four long-living extended (of 15–19 residues) transmembrane α-helices, TM1–TM4, observed in the reduced forms of hVKORC1[220] and the three short helices on the L-loop (**Figure 5.12**, **E**, **F**).

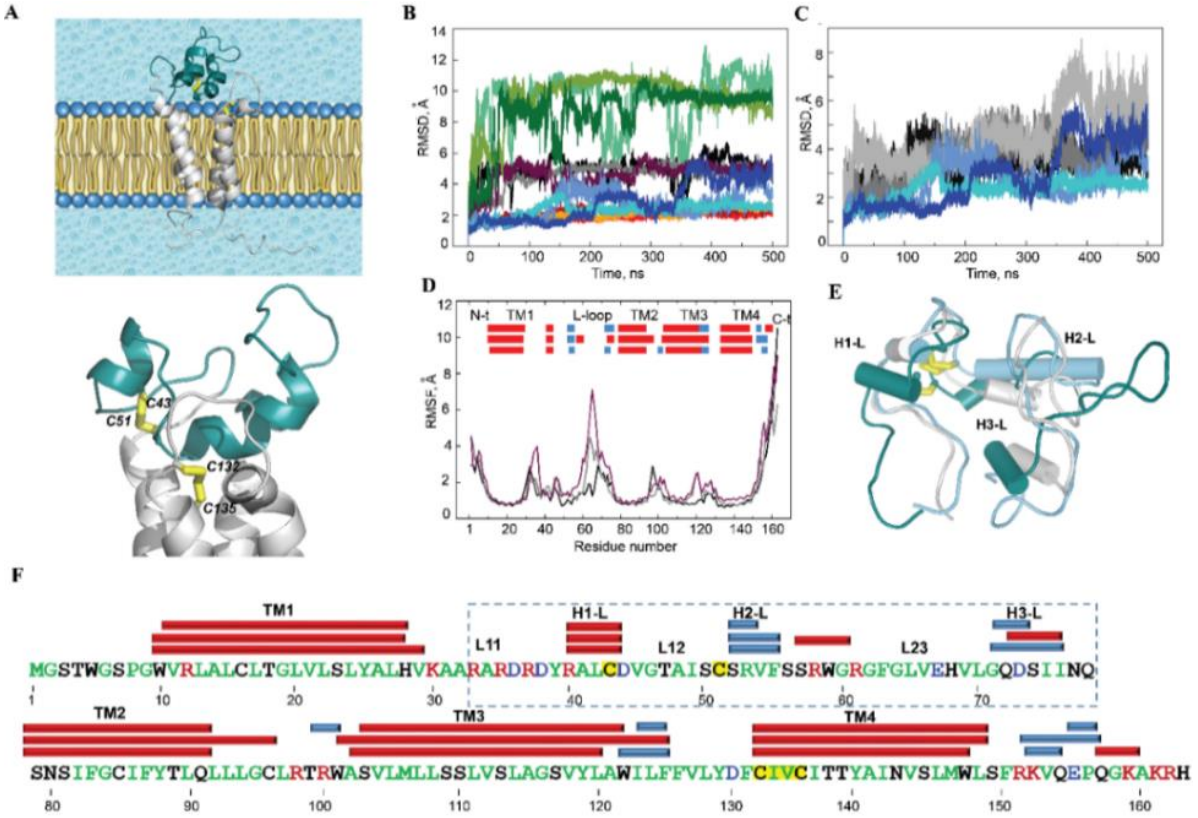


Figure 5.12 hVKORC1 in its inactive state and its conventional MD simulations. (**A**) 3D model of hVKORC1 in its inactive state; it was inserted into the membrane (top) and zoomed in on the L-loop (bottom). The L-loop is highlighted by the colour teal; disulphide bridges formed by cysteine residues C43-C51 and C132-C135 are drawn as yellow sticks. Transmembrane helices (TMs) are numbered as in [220]. (**B**,**C**) RMSDs computed for each MD trajectory (replicas 1–3) from initial coordinates (at t = 0 ns, the same for all replicas) on the Cα-atoms of full-length hVKORC1 (in black, grey and rose brown), of the transmembrane domain (in orange, red and grenadine), of the L-loop (in clear aqua, bleu and navy) and of the N- and C-terminals (in teal, green and deep green) after fitting to the initial conformation of the respective fragment (B); of the L-loop (i) after fitting to its initial conformation (clear aqua, blue and navy blue) and (ii) after fitting of the protein coordinates to the initial conformation of the TMD (black, grey and silver) (C). (**D**) RMSFs computed for Cα-atoms of the MD conformations (replicas 1–3) after fitting to the initial conformation (at t = 0 ns, the same for all replicas; in black, grey, and rose brown). In the insert, the folded secondary structures, αH- (red) and 3_{10}-helices (blue), were assigned for a mean conformation of each MD trajectory. (**E**) Superimposition of the L-loop conformations picked from replica 3 at 150 (grey), 250 (light blue) and 375 ns (deep teal). (**F**) The hVKORC1 sequence (Q9BQB6) and the secondary structure assignment for a mean conformation over each MD trajectory. Residues are coloured according to their properties: positively and negatively charged residues are in red and blue, respectively; hydrophobic residues are in green; polar and amphipathic residues are in black;

residues C43, C51 and the CX_1X_2C motif are highlighted by a yellow background. α- and 3_{10}-helices are shown as red and blue batons, respectively. Secondary structure labelling is shown above the hVKORC1 sequence. The L-loop sequence is surrounded by dashed lines.

5.2.2.2. The Luminal Loop of hVKORC1: Structure and Dynamics

Since the luminal loop (L-loop) is the fragment targeted by a Trx-fold protein, our focus is mainly on its intrinsic structural and dynamical properties and their connection with those of the transmembrane domain of hVKORC1.

L-loop folding, encompassing 30%, 36% and 22% of all residues in replicas 1–3, respectively, is presented by three small (3–4 residues) transient helices, H1-L, H2-L and H3-L, which are partially converted between the αH- and 3_{10}-helices (**Figure 5.12**, **E**; **Figure S8**). Despite the transient structure of helices, their positions on the sequence are well conserved. The L-loop helices are interconnected by coiled linkers, which, together with the linker joining the L-loop to TM1 from the transmembrane domain of hVKORC1, display RMSF values that suggest the high mobility of these loops (**Figure 5.12**, **D**). H1-L, mainly folded as a regular α-helix, contains C43 at its C-cap, which is linked covalently to C51, an N-cap residue of H2-L helix, forming the S···S bridge between two cysteines. Such covalent bonding significantly restricts the conformational mobility of this fragment. The large coiled linker connecting H2-L and H3-L helices is composed of hydrophobic residues, with the inserted charged and polar amino acids in the proximity of each helix (**Figure 5.12**, **F**).

The intrinsic dynamics of hVKORC1 was first analysed with the cross-correlation matrix computed for the Cα-atom pairs of the full-length protein and the L-loop. The Cα–Cα distance pairwise patterns demonstrate the coupled motions within each hVKORC1 domain, the L-loop and the TMD and between two structural domains (**Figure 5.13**, **A**; **Figure S9**). The regular pattern in the TMD reflects the correlated motion of the TM helices that is mainly associated with their collective drift, observed earlier in all metastable states of hVKORC1[220]. The motion of the L-loop correlates with the movement of the linkers that connect the TM-helices and join the L-loop to the TMD.

The cross-correlations computed on only L-loop atoms display different maps in the three replicas, with either a fine-grained pattern (replicas 1 and 2) or a pattern composed of well-defined blocks of nearly equal size (replica 3), reflecting the highly coupled motion of the L-loop fragments consisting of 10–12 residues from the L-loop helices and their adjacent linkers. The difference in patterns is associated with the disparity of L-loop motion—small or medium in replicas 1 and 2 and broad in replica 3, as evidenced by RMSFs and PCA.

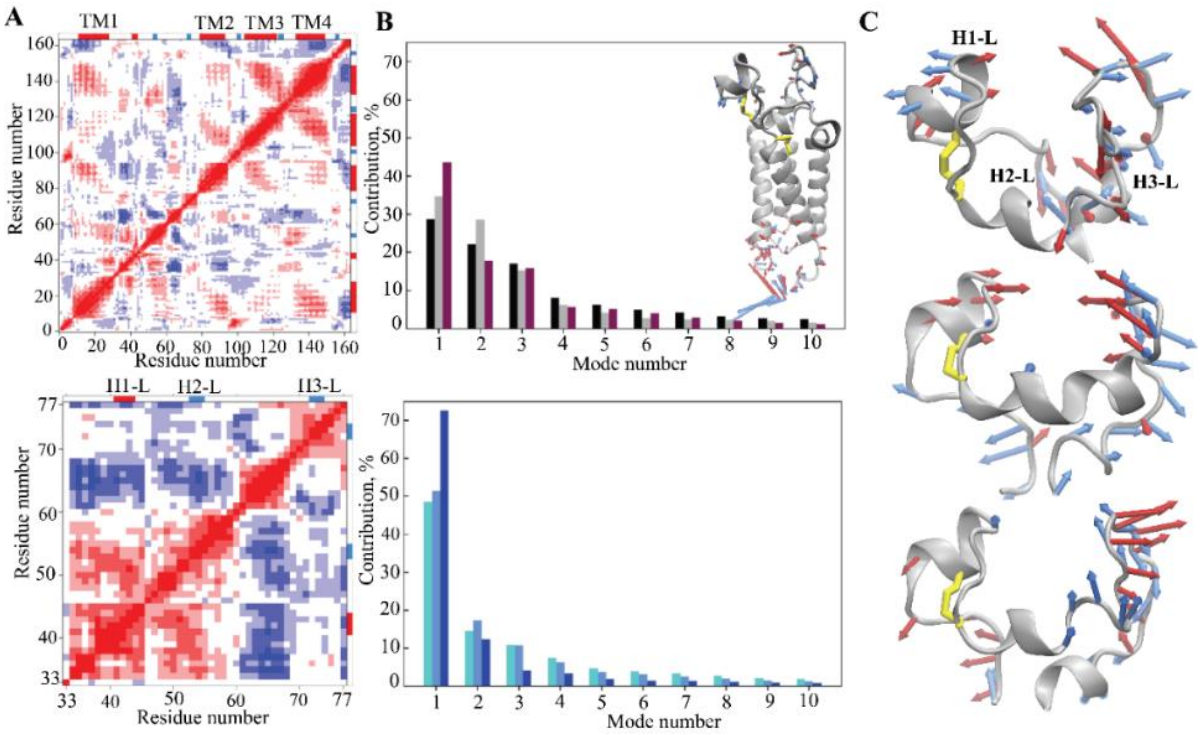


Figure 5.13 Intrinsic motion of hVKORC1 and its L-loop. (**A**) The inter-residue cross-correlation map computed for the Cα-atom pairs after fitting to the respective first conformation (t = 0 ns) of the full-length hVKORC1 (top) and the L-loop (bottom) is shown for the three replicas. Correlated (positive) and anticorrelated (negative) motions between the Cα-atom pairs are shown as a red–blue gradient. (**B**) The PCA modes of the full-length hVKORC1 (top) and the L-loop (bottom), calculated for each MD trajectory after least-square fitting of the MD conformations to the average conformation of the respective domain as a reference. The bar plot gives the eigenvalue spectra in descending order for the first 10 modes. The data for replicas 1–3 are coloured black, grey and rose brown, respectively, while, for the full-length hVKORC1, the colouring is clear aqua, blue and navy blue for the L-loop. (**C**) Atomic components in the first PCA modes of the L-loop are drawn as red (1st mode) and blue (2nd mode) arrows projected onto the respective average structure from replicas 1 (top), 2 (middle) and 3 (bottom). Only motion with an amplitude ≥ 2Å is shown. The S-S bridge of hVKORC1 is shown using yellow sticks.

The collective motions of hVKORC1, characterised by PCA, showed that ten modes describe ~80–90% of the total fluctuations of both the full-length hVKORC1 and the L-loop (**Figure 5.13**, **B**). Similar to the RMSF values, the first two PCA modes denote the great mobility of the terminal residues (N- and C-terminus) and the L-loop (**Figure 5.13**, **B**, **insert**). PCA analysis performed on only the Cα-atoms of the L-loop showed that two first modes describe most of the L-loop motion that displays the large-amplitude collective movements of the L-loop fragments—helices and adjacent-coiled linkers. The amplitude and direction of motion of the L-loop fragments differ in the three trajectories (**Figure 5.13**, **C**), suggesting a larger conformational space for the L-loop than was observed in each trajectory, probably larger than the total space of all trajectories. The first two modes in replicas 1 and 2 showed a highly coupled motion of the H1-L helix and the L23 linker in a scissors-like manner, while the collective

motion in 3 mainly displays a displacement of L23, which is horizontal with respect to the rest of the L-loop and vertical with respect to the TMD (**Figure 5.13**, **C**; **Figure S9**).

To characterise the conformational changes of the L-loop that are associated with a great deal of flexibility and mobility, the most emblematic residues, in view of their fluctuations (RMSFs), were first selected. Two sets of residues — (1) C43, V54 and S74, located on the L-loop helices (the midpoint residues of H1-L, H2-L and H3-L) and showing the minimal values of RMSFs, and (2) R35, G46 and G64, positioned on the L-loop linkers L11, L12 and L13, respectively, and displaying the greatest RMSF values— were chosen (**Figure 5.14**, **A**). Each set of residues was completed by residue C135 from the TMD and was then used to define two tetrahedrons, T1 and T2, designed on the Cα-atoms. It is suggested that light may be shed on the conformational features of the L-loop by analysis of the six straight edges corresponding to the distances between each pair of residues.

Analysis of T1 geometry showed (i) the great stability of Cα–Cα distances (d) between C43 (H1-L helix), V54 (H2-L helix) and C135 over nearly all the simulated time and in all the replicas; (ii) high conservation of Cα–Cα distances between each of the three residues and S74 (H3-L) over a substantial time period (200–300 ns or more), followed by (iii) a synchronic change of these distances (Δ of 6–8 Å), indicating the displacement of the H3-L helix with respect to the other helices, H1-L and H2-L (**Figure 5.14**, **B**). As was expected, T2, which is determined using the most fluctuating residues, showed less conserved geometry, displaying synchronic changes in all or at least 3–4 distances (Δ of 8–15 Å).

Comparison of T1 and T2 metrics showed an absence of coupling between their geometries. Similarly, no evident relation was found between the secondary structure of the L-loop and the T1 or T2 geometries, suggesting that the relative positions of the residues from the L-loop helices and from the linkers connecting these helices are disconnected from the folding–unfolding effects in the L-loop (**Figure 5.14**, **B**, **C**).

This analysis revealed (i) the high stability of the H1-L helix in terms of its secondary structure, as well as in its relative position with respect to TMD, (ii) the quasistable spatial position of the transient H2-L helix relative to H1-L and TMD, (iii) the large displacement of the transient H3-L helix from the anchored structural motif formed by HL-1 and H2-L and of the coiled linkers L11–L13.

To illustrate the relative orientation of the L-loop helices, their structural drift was analysed. The axis of each helix was defined for the conformations from trajectories 1–3 (sampled every 100 ps, concatenated data), superposed and projected on a randomly chosen conformation of the L-loop (**Figure 5.14**, **D**). The superimposed axes (elongated by 50% to better represent their position and direction) form a reap-like distribution for all helices. The axes of the three helices differ in length and their spatial orientation within each reap-like distribution and between the helices.

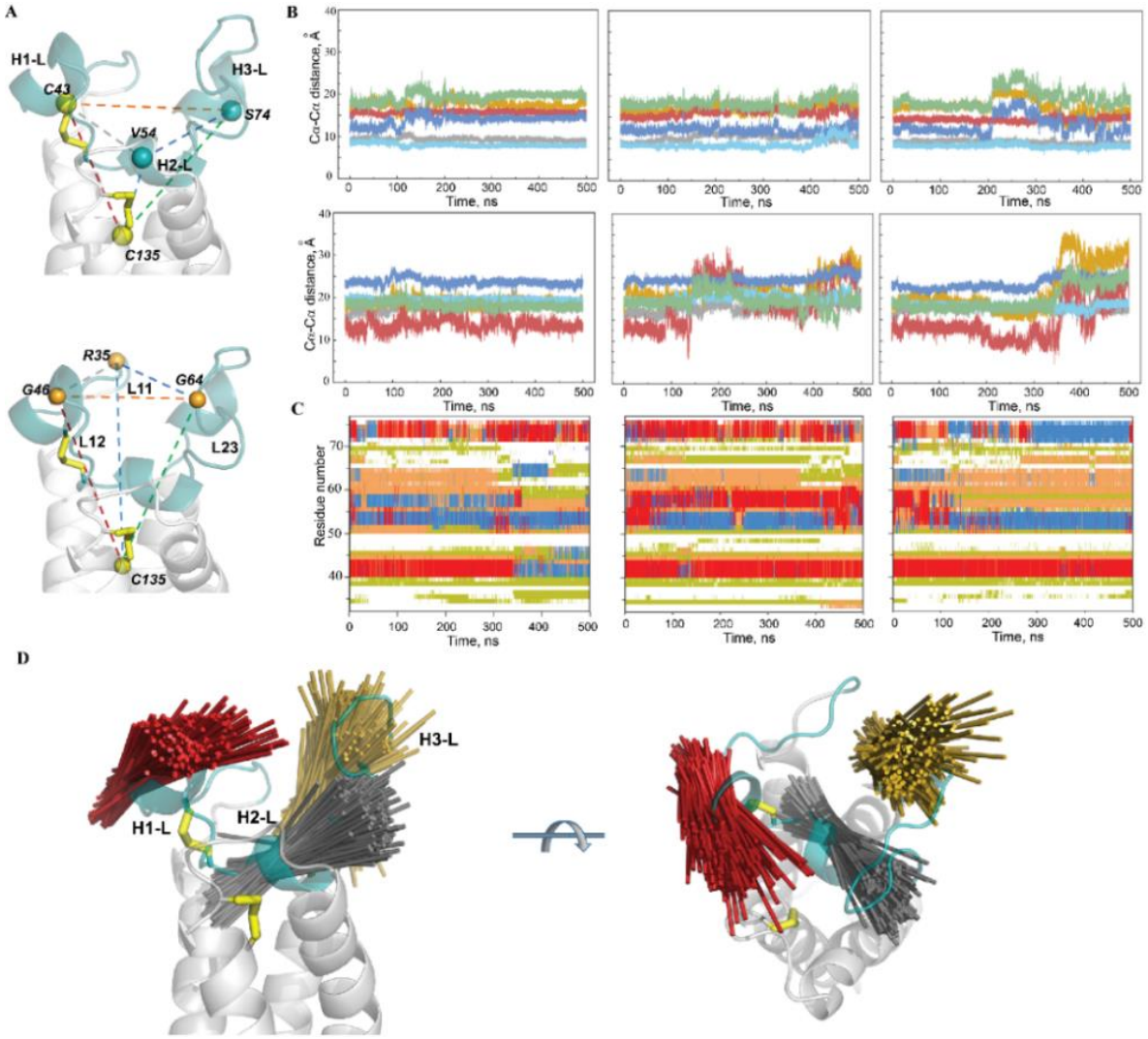


Figure 5.14 Geometry and folding of the L-loop from hVKORC1 in its inactive state. (**A**) Two tetrahedrons, T1—defined for the Cα-atom of C135 and for the midpoint residues of each L-loop helix, and T2—defined for the Cα-atom of C135 and for the most fluctuating residues (with the greatest RMSF values), from the L-loop linkers. (**B**) Distances between each pair of Cα-atoms from the tetrahedrons T1 (top) and T2 (bottom) over each MD trajectory. The distance curves and the edges of a tetrahedron are coloured similarly. (**C**) The time-dependent evolution of the secondary structure of each residue, as assigned by the Define Secondary Structure of Proteins (DSSP) method: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow. (**D**) Drift of the L-loop helices observed over the MD simulations (concatenated trajectory, sampled every 100 ps). Superimposed axes of helices from the L-loop are covered on the randomly chosen conformation of hVKORC1 in two orthogonal projections. The axis of each helix is defined as a line connecting the two centroids assigned for the first and the last residues.

5.2.2.3. Conformational Variability of the hVKORC1 L-Loop

To depict the conformational space explored over the MD simulations using the L-loop of hVKORC1 in its inactive state and to distinguish the most probable conformations, the generated conformations were analysed using ensemble-based

clustering[267]. Conformations of each MD trajectory were grouped with different RMSD cut-off values that varied from 1.6 to 3.0 Å, with a step of 0.2 Å. Using of cut-off value ≥ 2.2 Å results in a poor number of clusters, while more restricted cut-off values of 1.8 and 2 Å were sufficient to regroup the L-loop conformations into clusters that give the best cumulative population (> 90%; **Figure 5.15**, **A**). Interestingly, clustering with these cut-off values produces an equal number (six) of clusters, with a nonzero population in all replicas (**Figure S10**). Taking a cut-off of 2.0 Å as the criterion, the population of each cluster obtained for each trajectory was compared. In replica 1, most conformations form the two most-populated clusters, C1 (48%) and C2 (32%); the other conformations are regrouped in clusters with a low population of 0.5–9 %. In each of the other replicas (2/3), the MD conformations are regrouped in the three most-populated clusters, with a comparable density between the replicas: C1 (41/33%), C2 (22/23%) and C3 (20/15%). The MD conformations that form the most populated clusters, C1 and C2, are individually regrouped within the narrow time ranges in trajectory 3 only, while in two other simulations, they are observed over a long period for each trajectory as coexisting with the conformations from the other clusters (**Figure 5.15**, **B**). The conformations from the lowly populated clusters are usually observed in time ranges where the RMSD varies significantly and may show the transient states of the L-loop.

The representative conformations from different clusters of the same replica are divergent at the folding level (2D) and in 3D-structure organisation (**Figure S10**). An archetypical example is the considerable disparity between the conformations from clusters C2 and C3 of trajectory 3 that represents the L-loop before and after the transition, which is evidenced by the RMSD curve (**Figure 5.12**, **B**). In contrast, some representative conformations of the clusters from different replicas showed a convenient similarity, for instance, C2 and C1 from replicas 1 and 3, respectively.

It is supposed that the L-loop conformational spaces generated by the three independent trajectories are partially overlapped. To verify this hypothesis, a clustering analysis was performed on the merged trajectory composed of the L-loop conformations from three replicas. Accordingly, the number of clusters obtained with the same RMSD cut-off values is significantly lower for the merged data than the sum of clusters obtained individually for each replica (**Figure 5.15**, **A**), which confirms the overlapping of the conformational spaces of the L-loop covered over the three replicas of MD simulation of hVKORC1 in its inactive state.

The first three clusters of the concatenated trajectory (cut-off 2.0 Å) contain 31%, 14% and 12% of all conformations, while the other conformations form the poorly populated clusters. The cumulative population of the clusters, with a density > 4% on the merged data, is reduced (72%) with respect to individual trajectories but is still meaningful and statistically rich for the characterization of the most frequent L-loop conformations. Regarding the composition of the clusters, it was found that the dense clusters of the merged trajectory, $C1^m$ and $C3^m$, are composed of conformations from

different trajectories (C1^{m} and C3^{m} are comprised of conformations from replicas 1/2/3, with proportions of 83/4/12% and 28/12/58%, respectively), while the other clusters are composed of conformations from the unique trajectory—2 (C2^{m} and C5^{m}) and 3 (C4^{m} and C6^{m}), respectively (**Figure 5.15**, **C**).

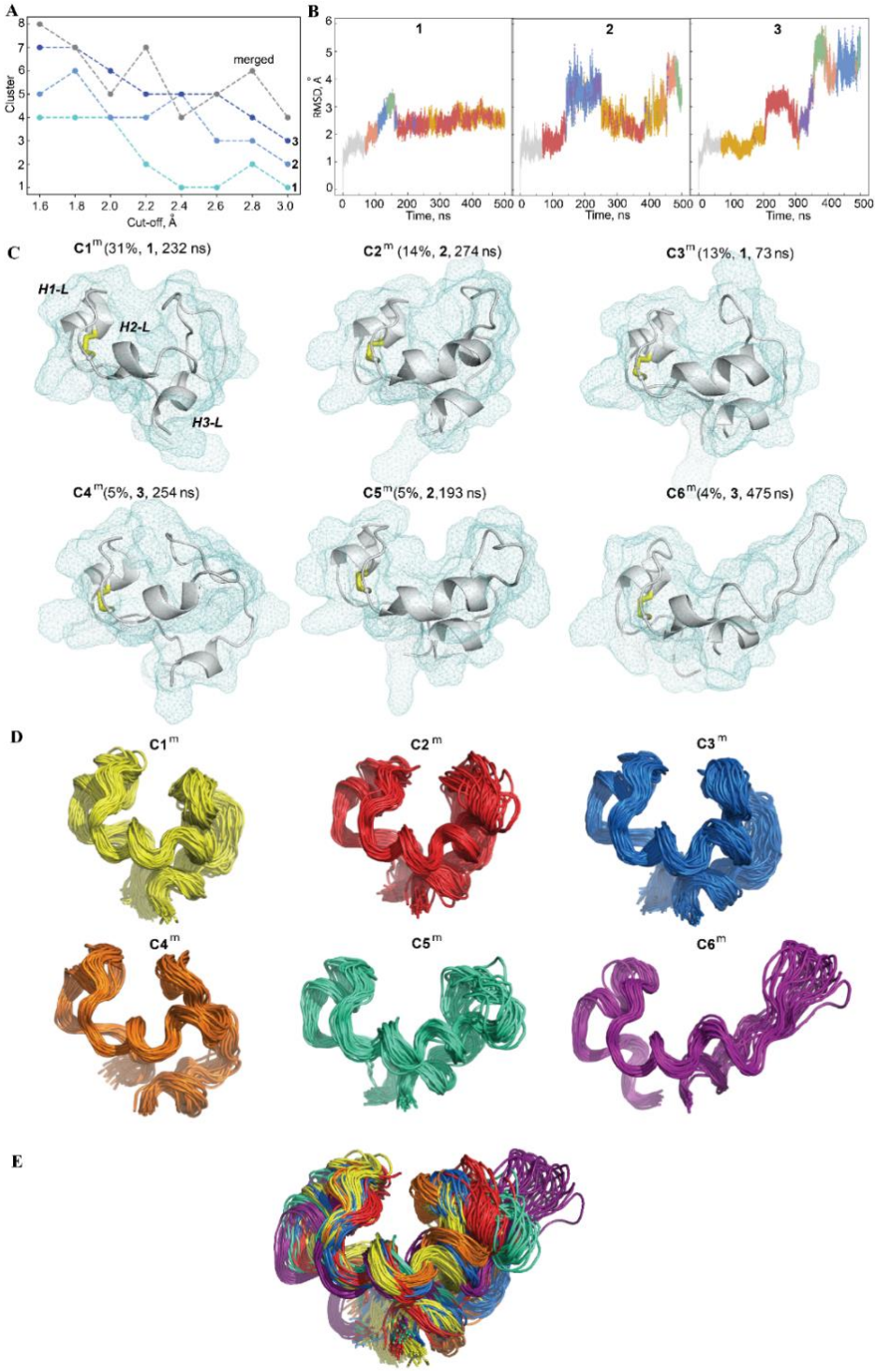


Figure 5.15 Ensemble-based clustering of L-loop MD conformations. (**A**) Number of clusters obtained for each MD trajectory (1, 2 and 3) and the concatenated trajectory. The first 70 ns of every trajectory were omitted from the computation. Clustering was performed on each 10-ps frame of every trajectory using cut-off values that varied from 1.6 to 3.0 Å, with a step of 0.2 Å. (**B**)

Location of the MD conformations grouped in clusters, with a cut-off of 2.0 Å for the RMSD curves of trajectories 1–3. Clusters C1–C6 are arbitrarily distinguished by colours in each trajectory: orange (C1), red (C2), blue (C3), rose (C4), green (C5) and violet (C6). (**C**) Representative conformations of the L-loop from clusters (C^m) with population ≥ 4%, obtained with a cut-off of 2.0 Å for the merged trajectory. The L-loop is shown as ribbons with a meshed surface, with disulphide bridges C43–C51 drawn as yellow sticks. The L-loop surface is displayed as meshed contours. The population of each cluster is given in brackets (in %), together with the replica number (in the bold) and the time (in ns) over which the representative conformation was recorded within a replica. (**D**) Conformations of the L-loop (taken every 100 frames) of each cluster (C^m) of the merged trajectory, and (**E**) superposed conformations from the $C1^m$–$C6^m$ clusters. In (D, E), the L-loop is drawn as a tube.

The representative conformation of each cluster, generated for the concatenated trajectory, showed that the principal factors leading to the conformational difference of the L-loop consist of (i) a variable length of the H2-L helix; a decrease of that promotes (ii) an elongation of linker L23, which, in turn, encourages (iii) the repositioning of the H3-L helix with respect to the H1-L and H2-L helices (**Figure 5.15**, **C**, **D**). In contrast to H2-L, the length of the H1-L and H3-L helices is better conserved. The whole shape of the conformations from different clusters well-reflects the "scissor-like" motion of the H1-L helix and the L23 loop that is observed in the PCA modes. The compact shape of the L-loop corresponds to the "closed" position of the H1-L helix and the L23 loop, which is a typical feature of most L-loop conformations (see the highly populated clusters, $C1^m$–$C4^m$). The conformations grouped in cluster $C6^m$ show an elongated shape, with an "open" position of the H1-L helix and the L23 loop. Cluster C5m is composed of intermediate conformations between the "open" and "closed" forms.

The clustering enabled (i) the splitting of MD conformations of the L-loop into groups composed of similar geometry and shape (within a cut-off), (ii) the assembly of a great majority of conformations into a limited number of clusters, and (iii) a distinction between dense clusters with a statistically reasonable population.

5.2.2.4. INTRA-L-LOOP INTERACTIONS

To establish the forces that stabilise L-loop conformations, contact maps were computed for each representative conformation from the most populated clusters (> 4%) found on the concatenated trajectory. The contact maps show the multiple intra-L-loop interactions between the linkers, between the linkers and helices and between the helices (**Figure S11**). Nevertheless, the patterns of such contacts differed in clusters $C1^m$–$C6^m$. The most common pattern found in the maps describes the contact of L11 with H2-L and H3-L helices and of L23 with H3-L, which are systematically observed in clusters $C1^m$–$C5^m$.

Analysis of the H-bonds showed that the L-loop conformations are stabilised by

mutual H-bonds that form extensive networks (**Figure 5.16**; **Table S1**). Comparing these H-bond networks in "closed" conformations (clusters $C1^m$–$C5^m$), it is noted that D36, D38, D44, R53, R61 and E67 are the key residues that form the salt bridges. In the "open" conformation (cluster $C6^m$), the set of interacting residues that form the salt bridges is composed of R35, D38, D44 and R58.

The salt bridge that is stabilised by the pairing of charged residues when a combination of two noncovalent interactions is formed, H-bonding and ionic-bonding, is the most observed contribution to the stability of the entropically unfavourable folded conformation of proteins[315]. Indeed, in the highly compact "closed" L-loop conformations from $C1^m$ and $C2^m$, R53 interacts with D36 and D38, forming the R53-based "salt bridge pattern" that stabilises the proximal position of H2-L and the L11 linker. In the conformations from cluster $C3^m$, the "salt bridge pattern" is formed by R61 interacting with D36 and E67, which stabilises the tight location of two distant linkers, L11 and L23.

These interactions in $C3^m$ are completed by the contact of R53 (H2-L) with D36 (L11), causing an overlap of the two "salt bridge patterns", namely, R61- and R53-based patterns. Additionally, in $C3^m$, the other "salt bridge pattern" is formed by R37 contacting with D44, which stabilises H1-L and the L12 loop in a tight spatial position. In $C4^m$, the R53- and R61-based "salt bridge patterns" are clearly separated, while each positively charged residue interacts with different subsets of the negatively charged residues, i.e., R61 with D36 and E67 and R53 with D38. These two "salt bridge patterns" gather two neighbouring helices, H1-L and H2-L, and two distant linkers, L11 and L23.

In $C4^m$, like $C3^m$, the "salt bridge pattern" formed with R37 and D44 is clearly separated from the R61- and R53-based "salt bridge patterns". Such a spatial separation of two "salt bridge patterns" is observed in the "open" conformations of the L-loop from $C5^m$ and C6m clusters, in which two "salt bridge patterns" are formed by R53 (H2-L) interacting with D36 and D38 ($C5^m$) or with D38 and D44 ($C6^m$), and either by R37 (L11) interacting with R40 and D44 ($C5^m$) or by R35 bound to R35 and D38 ($C6^m$).

Besides the salt-bridge interactions, the charged residues also contribute to H-bonds by interaction with the different polar and hydrophobic residues, which either act as H-donors or H-acceptors for the atoms in their main or side chains. All these ionic and H-bond interactions between the charged residues and between the charged and polar residues contribute to the tight spatial L-loop arrangement, in which the helices and linkers from the remote sequence segments are localised at proximity. It is interesting to note that R40, D44, R53 and R61 interact in any conformation of the L-loop, independent of the L-loop's shape, by forming either the salt bridges or the H-bonds.

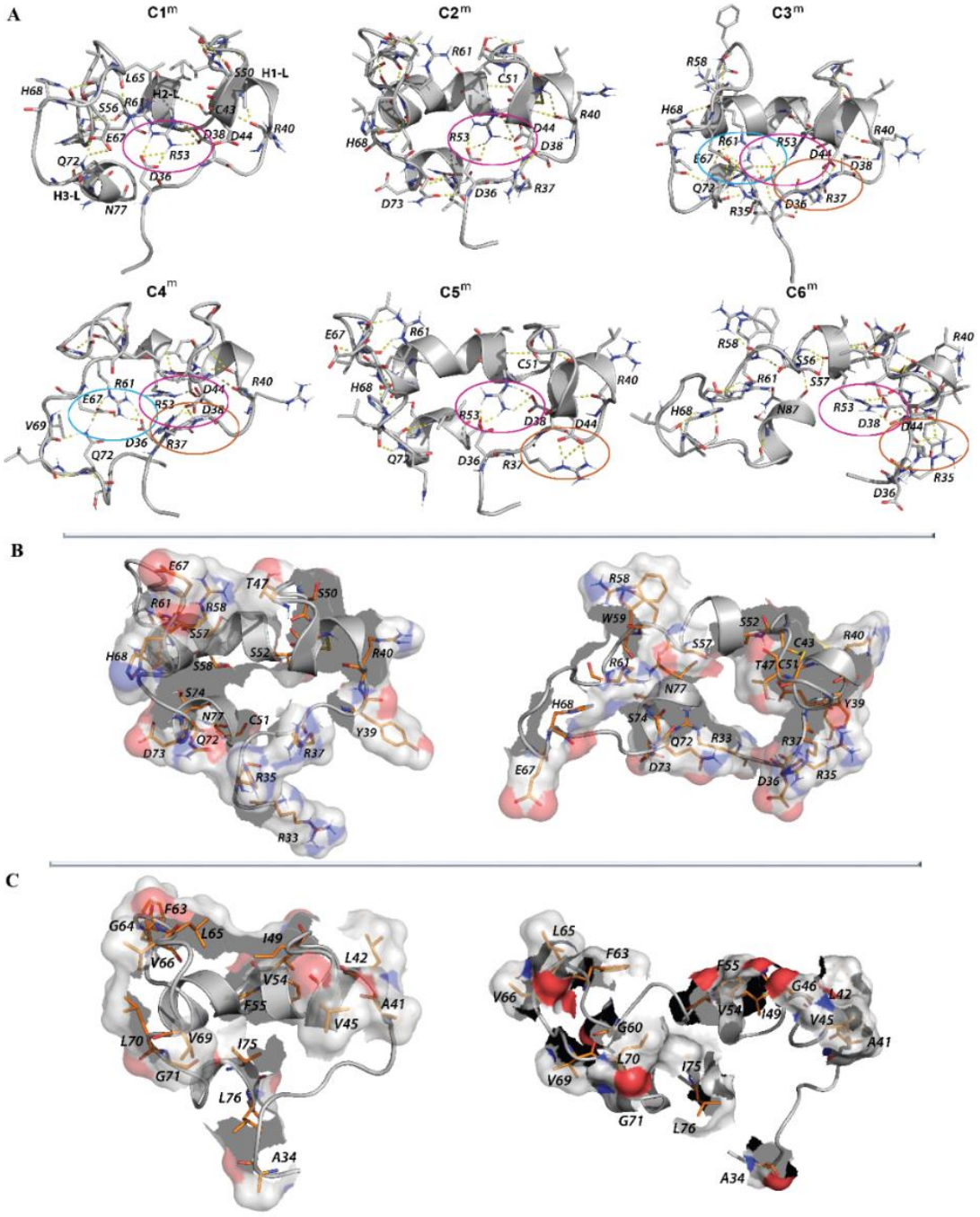


Figure 5.16 Interacting residues in L-loop conformations. (**A**) Intraloop H-bond interactions in the L-loop conformations from clusters C1m–C6m. H-bonds D-H···A (D···A < 3.6 Å, ∠DHA ≥ 120°), where D and A are H-donor and H-acceptor (O/N) atoms, were analysed in a representative conformation from each cluster of the merged trajectories. Interactions that stabilised the helices were not considered. The L-loop is shown as ribbons, with the interacting residues as sticks and H-bond traces as dashed lines. Common H-bonding motifs are encircled by magenta (at R53), blue (at R61) and orange (at R37). The most characteristic donor and acceptor groups are labelled. N, O and C atoms are in blue, red and grey, respectively. (**B**) Charged and polar residues protruding from the L-loop. (**C**) Hydrophobic residues protruding from of the L-loop. The L-loop is shown as ribbons, with the residues exposed to the solvent displayed as sticks with a space-filling encounter. In (B,C), the N, O and C atoms are in blue, red and orange, respectively.

Nevertheless, many charged and polar residues that are not involved or are partially involved in intra-L-loop interactions protrude from the L-loop, as illustrated by the "closed" and "open" conformations of the L-loop (**Figure 5.16**, **B**). Considering the spatial position of the solvent-exposed residues with respect to L-loop cysteine residues (C43 and C51) that participate in the thiol–disulphide exchange reaction, the residues from sequence S52 to E67 are most likely involved in interactions with a redox protein.

As the L-loop also contains many hydrophobic residues, their contribution to intra- and intermolecular interactions was evaluated. Although hydrophobic forces are known to be relatively weak interactions, such interactions can add up to make an important contribution to the overall stability of a conformer or molecular complex[316].

Multiple contacts between the A41, G46, A48, I49, V54, L70 and L76 hydrophobic residues were observed in "closed" conformations, while in the "open" conformations, such contacts involved V45, F55, F63, L70 and L76 (**Figure S12**; **Table S1**). These hydrophobic contacts may reflect the stabilising interactions that complete the H-bond contribution as well as the repulsive forces that equilibrate the strong salt-bridge interactions.

Similar to the charged and polar residues, some hydrophobic side chains are oriented toward the exterior of the L-loop, putting them in positions accessible to the solvent, such that the number of such residues is significantly higher in the "closed" conformations than in the "open" ones (**Figure 5.16**, **C**; **Table S1**). One part of these residues (F55, G56, F63, L65, V66) belongs to the sequence S52–E67, which was postulated to be involved in interactions with a redox protein.

5.2.3. Discussions

Vitamin K epoxide reductase is a membrane protein that reduces vitamin K using a membrane-embedded cysteine-containing redox centre. Such activity requires the cooperation of hVKORC1, with a redox partner that delivers reducing equivalents. The physiological redox partner of hVKORC1 remains uncertain; nevertheless, four proteins—PDI, ERp18, Tmx1 and Tmx4—were suggested as the most likely H-donors of hVKORC1[222,224]. Deciphering the molecular origins of hVKORC1 recognition by an unknown redox protein is not a trivial task.

We suggested that a careful *in silico* study of the isolated proteins would provide useful information. In particular, quantitative metrics and qualitative estimations can shed new light on the target (hVKORC1) features and the peculiarities of redox proteins. Such information may help in predicting (i) the protein fragments participating in hVKORC1 recognition by a Trx and (ii) the most probable partner of hVKORC1.

What was learned from studying hVKORC1 protein?

The L-loop is known to bind to and accept reducing equivalents from species-specific partner oxidoreductases essential for hVKORC1 enzymatic function *in vivo*[214], so this domain was carefully characterised. We found that the L-loop in the inactive (oxidised) state of hVKORC1 is noticeably less flexible compared to the reduced states of hVKORC1[220] and more folded, showing three helices connected by coiled linkers. This three-helix fold of the L-loop was generally maintained over the MD simulations, while the length and spatial positions of the helices were highly variable. This variation is reflected in many L-loop conformations, varying from a compact "closed" conformation, which is prevalent, to an extended "open" conformation.

It was established that the H2-L helix is the fundamental actor that controls the conformational features of the L-loop. This transient helix converts between the αH- and the 3_{10}-folds, adapting in length from short to elongated. The shortened H2-L helix, in which the S56-R61 segment is unfolded, promotes the elongation of the coiled linker L23, connecting the H2-L and H3-L helices. The extended linker L23 shows (i) great mobility with respect to the H1-L helix, which can be described as a "scissor-like" motion, and (ii) a large vertical displacement with respect to the TMD. Moreover, the extended linker L23 delivers increasing mobility to H3-L, evidenced by its displacement with respect to H2-L.

At the sequence level, the L-loop has been reported to be conserved between VKORs from different species[317]. Particularly high conservation was found for the S56-G63 segment, which in hVKORC1 is followed by 5-residue hydrophobic insert GFGLV, which is completed by glutamic acid (E67) and histidine (H68). Sequence conservation, along with observed structural and dynamical properties of the H2-L helix and its adjacent linker L23, suggests their possible functional role. From the analysis of the H-bonding patterns in the L-loop, regular exposition of the charged (R58, R61, E67 and D73) and polar residues (S56, S57, W59, H68 and N77) to the outer side of the L-loop was observed in positions favourable for contact with a solvent or protein. Therefore, we postulated that the S56-R61 segment, a part of the more extended S53-N77 segment, is a platform for the recognition of a protein partner.

Charged residues have been shown to be instrumental in the definition of binding specificity, while sometimes contributing little binding energy to the interactions themselves[318,319]. In other cases, charged residues were found to promote high-affinity binding[320,321]. They are also the main players in "electrostatic steering", which is a long-range mechanism in which electrostatic forces can steer a ligand protein to a binding site on the receptor protein; this drastically increases the association rate[322,323]. Often, charged residues that are important for protein–protein interactions are conserved across families of evolutionarily related proteins and protein complexes[324–326].

Moreover, the tryptophan residue (W59) from the S56-R61 segment, following 5-residue hydrophobic insert GFGLV, may act as an anchoring residue that binds the two proteins. Tryptophan residues have been shown to exhibit a strong tendency to remain within the interfacial region[327]. The role of the hydrophobic effect as a driving force in protein folding and assembly is well described[328].

5.3. COMPARAISON DU MODELE *DE NOVO* DE HVKORC1 AVEC SES STRUCTURES CRISTALLOGRAPHIQUES

Résumé. *En 2021, des structures cristallographiques du VKOR de* Takifugu rubripes *et de hVKORC1 ont été résolues par cristallographie aux rayons X. Ces structures empiriques ont confirmé la topologie à quatre hélices prédites par la modélisation* in silico *et ayant permis la modélisation du modèle* de novo *proposé par l'équipe BiMoDyM en 2017. Outre une position dans la membrane de la boucle L incohérente avec sa fonction de reconnaissance par sa protéine redox, la simulation de dynamique moléculaire de modèles issus des structures cristallographiques a montré une rigidité aberrante de la boucle L induite par la présence d'une liaison hydrogène robuste. La destruction de cette liaison a permis de restaurer la flexibilité de la boucle et les propriétés structurales et dynamiques des modèles relaxés se sont approchées de celles du modèle* de novo. *Par ces résultats, nous avons conclu que l'utilisation des structures cristallographiques n'est pas raisonnable et qu'une conformation à boucle L « fermée » du modèle* de novo *est la plus appropriée pour les études de reconnaissance et d'activation du hVKORC1 par sa protéine redox.*

5.3.1. INTRODUCTION

To advance our knowledge on hVKORC1 as a target for its redox protein, in the section we explored all available structural information – the recently reported crystallographic structures of VKOR[218] and compared them with *de novo* model (**Figure 5.17**).

As we focused on a fully oxidised state enzyme, in which two pairs of cysteine residues form disulphide bridges, we concentrated our analysis on structures reported two forms of the oxidised state, which were obtained at different crystallisation conditions and referenced in [218] as 'open' and 'closed' (**Figure 5.17**, **b**). These terms are not related to L-loop conformation, and describe the forms obtained either by co-crystallisation of hVKORC1 with ligands, the holo form, or ligand-free enzyme, the apo form. Since crystallographic data revealed structurally different forms of the same enzymatic state with a common topology/connectivity, it perfectly confirms the large flexibility of L-loop suggested *in silico*[220].

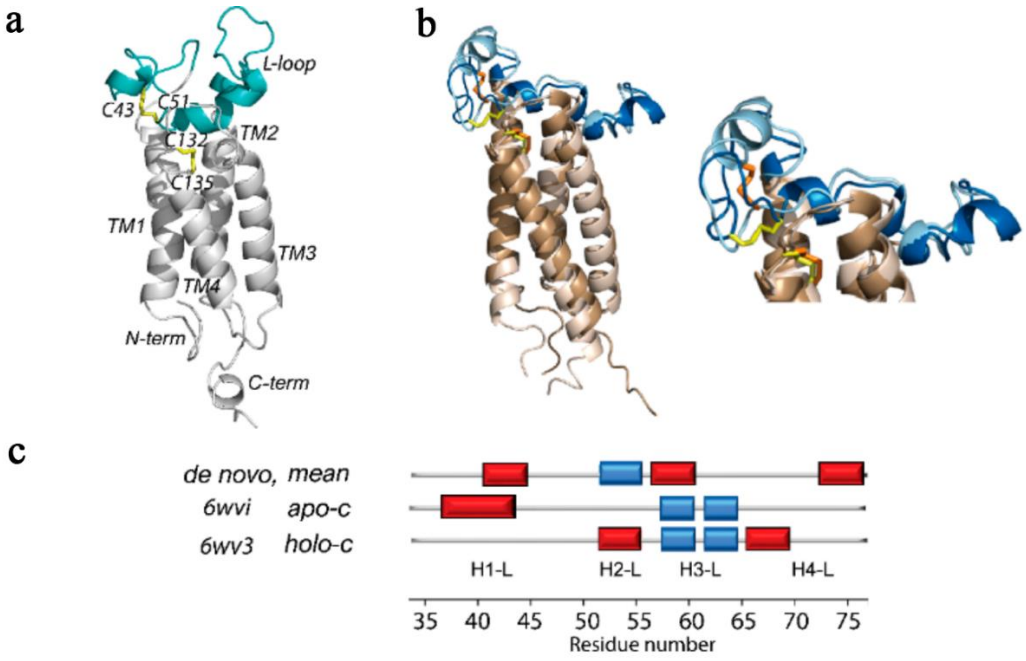


Figure 5.17 The hVKORC1 structure. (**a**) 3D *de novo* models of hVKORC1 in the oxidised inactive state (left, [220]). (**b**) Superimposed crystallographic structures characterising the oxidised state of VKOR-like protein from *Takifugu rubripes* (free-ligand) (PDB ID: 6wvi) and hVKORC1 with ligands (warfarin and glycerol monooleate) (PDB ID: 6wv3), refer enced as apo-c (light blue) and holo-c (dark blue) forms (middle), with a zoomed view on the L-loop (right). Ligands are not shown for clarity. (a, c) Protein and its fragment are shown as ribbons; disulphide bridges denoted in yellow and orange sticks. TMD helices coloured in grey and brown. (**c**) Interpretation of L-loop secondary structure in the *de novo* model (mean conformation over cMD trajectory) and two crystallographic forms of oxidised VKOR. α- and 3_{10}-helices are shown in red and blue respectively. Helices numbering is adapted from [211].

Focusing on hVKORC1 as the target of its redox protein PDI, we first address the central question – is hVKORC1 L-loop a properly folded two-state domain that provides a reversible open-to-closed conformational transition (on-off switch), postulated in [218], or is it an inherently disordered (ID) region possessing great plasticity/adaptability required to perform various steps during the hVKORC1 activation process?

To explore this question as much as possible in a more explicit and pragmatic way, we characterised structural and conformational properties of hVKORC1 in the fully oxidised (inactive) state using two crystallographic structures (apo and holo forms) and conformational spaces generated by molecular dynamics (MD) simulations of their homology models, *de novo* model with an intrinsically disordered L-loop [220], and L-loop extracted from these models and simulated as an isolated polypeptide.

We first postulated that two crystallographic forms must be easily reversible during MD simulations under identical conditions (unbound protein in water solution). MD conformational sets derived from two crystallographic forms and *de novo* model were

compared to reach an expected consensus on their common conformational spaces.

This computational study aimed comprehensively investigation the structural and conformational properties of enzymes.

5.3.2. RESULTS

5.3.2.1. TWO CRYSTALLOGRAPHIC FORMS OF HVKORC1 FULLY OXIDISED STATE: ANALYSES AND HYPOTHESIS

This paragraph summarises a comparative analysis of VKOR crystallographic structures recently reported in [218]. This analysis was carried out to improve our knowledge of hVKORC1 regarded as a target of its redox protein, an aspect not yet discussed to date at structural level.

The crystallographic structures represent two alternative forms of VKOR oxidised state obtained upon different crystallisation conditions: by co-crystallisation of human VKORC1 with vitamin K antagonists (VKAs) and additive molecules (glycerol monooleate) located in the active site pocket, and the ligand-free VKOR-like protein from *Takifugu rubripes* and ligand-bound VKOR in which the vitamin K quinone (KQ) or vitamin K epoxide (KO) is in non catalytic site. These two forms were called by the authors open and closed respectively for the ligand-free and ligand-bound active site forms. To compare these two crystallographic forms, the respective structures were retrieved from the Protein Data Bank (PDB)[116] and their original atomic coordinates were extracted from well-resolved structures. To avoid inaccuracy in the used terminology for various conformations of VKOR, the forms identified by crystallography will be referred as apo-c (PDB ID: 6wvi) and holo-c form (PDB ID: 6wv3), while the MD conformations of hVKORC1 with varying L-loop conformations from compact globular-like to elongated arrangement, will be referred as the open and closed.

The Root Mean Square Deviation (RMSD) calculated on Cα-atoms (the N- and C-terminal residues were excluded) showed that two forms differ mainly in L-loop (4.0 Å) while TMD is similar (0.7 Å) (**Figure 5.17**; **Figure 5.18**).

L-loop showed a helical folding composed of two helices (H1-L and H3-L) in apo-c form and three helices (H2-L, H3-L, and H4-L) in holo-c form (**Figure 5.17**; **Figure 5.18**). Noteworthy, in both forms, is that H3-L is made up of two adjacent small 3_{10}-helices separated by R61 apparently acting as a breaker of regular structures. The double 3_{10}-helix is formed by the same residues in both structures, and its position in 3D space is equivalent. In addition to folding differences, the L-loop shows divergent conformations, resulting in a distinct localisation of the L-loop disulphide bridge

compared to the one in the active site, where they are distant in apo-c and neighbouring in holo-c.

We first hypothesised that ligands may play a decisive role in stabilising the holo-c form. Surprisingly, there are no H-bonds between ligands and residues of L-loop. Both ligands, warfarin and glycerol monooleate, form non-covalent interactions only with residues from TM1, TM2 and TM4 of TMD, with exception of a unique hydrophobic contact described as interaction C-H···π, between V54 (H2-L) and warfarin (**Figure 5.18**, **b**). Therefore, if L-loop residues do not make a significant contribution to ligands binding, the structural and conformational difference of L-loop in two crystallographic forms depends on other factors. To identify such factors, non-covalent contacts stabilising each crystallographic form were calculated for two sub-regions of L-loop, defined in [218] as cap (R33-F55) and anchor (S56-N77), the second sub-region, a very flexible in *de novo* model, will be further called as hinge. Both sub-regions are stabilised by multiple H-bonds differing greatly between two crystallographic forms in the cap (so-called the form-specific contacts), while most H-bonds in the hinge are observed in both forms (**Figure 5.18**, **d**;). In general, distances characterising H-bonds in two forms are shorter in apo-c compared to holo-c.

Remarkably, residue N80 acts as a trifurcated binding centre in both crystallographic forms to maintain H2-L (holo-c) or coil (apo-c) and H3-L at proximity to transmembrane helix TM2 through simultaneous interactions with W59, G60 and F63. Additionally, N80 provides affinity to warfarin as a donor in H-bond (in holo-c form). Similarly, residue Q78 is involved in bifurcated H-bonding with G62 and I75 in both forms. The form-specific weak H-bonds and hydrophobic contacts involving either intra−L-loop residues alone or with TMD contribute to incremental stabilisation of L-loop conformation in each form.

This analysis of crystallographic structures showed that (i) L-loop of hVKORC1 in the fully oxidised state is potentially able to display large structural and conformational rearrangements; (ii) L-loop residues do not contribute significantly to ligands binding; (iii) L-loop in both crystallographic forms is mainly stabilized by form-specific H-bonds in the cap while the hinge is stabilised by H-bonds common in both forms. We naturally suggested that (i) L-loop structural and conformational difference in holo-c form versus apo-c is driven primary by steric factors associated with accommodation of ligands or their absence, and (ii) removing the ligands from hVKORC1 would relax the target promoting closed-to-open transition referred in [218].

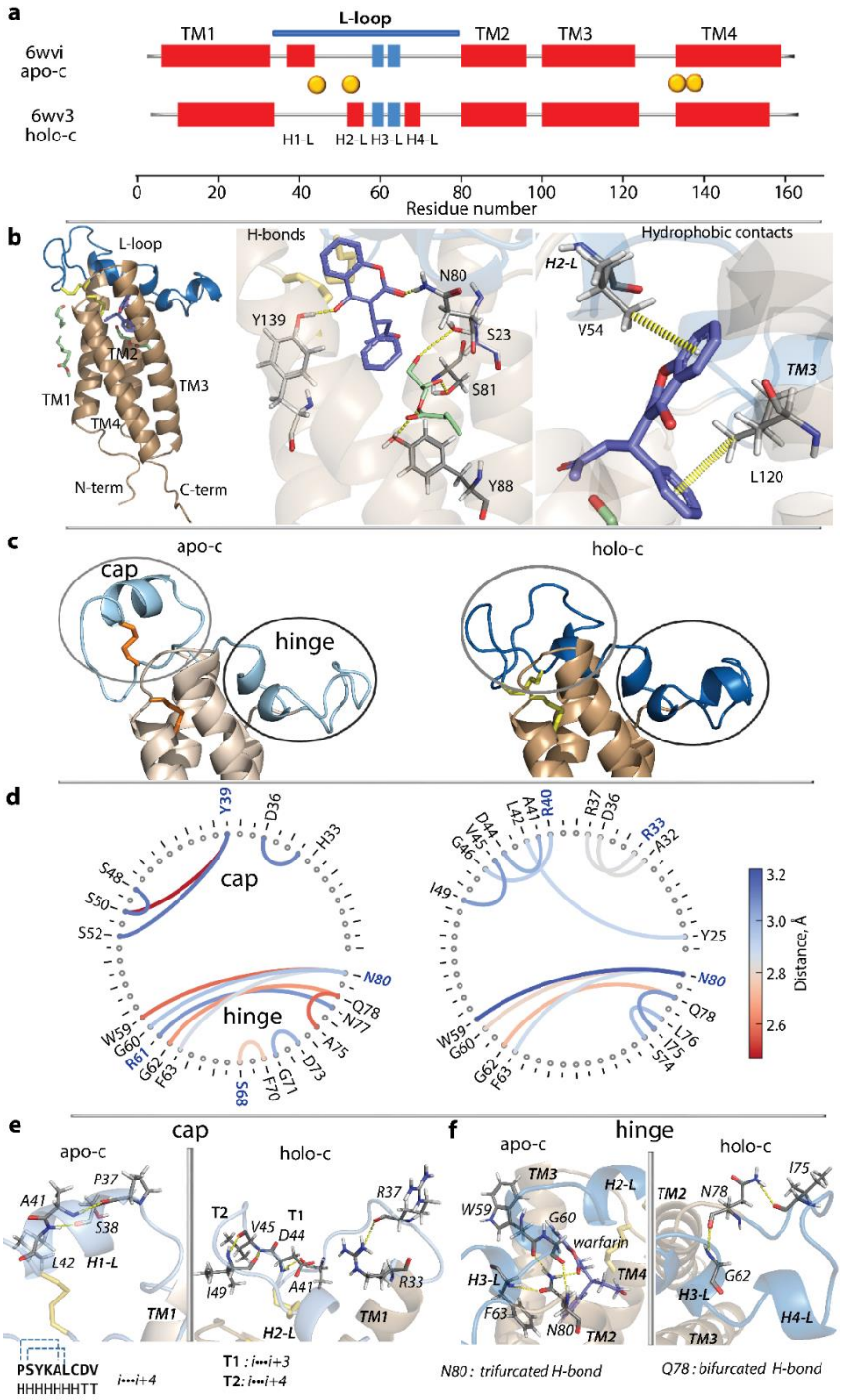


Figure 5.18 Structure of two crystallographic forms of the oxidised state of VKOR. (**a**) Secondary structures interpretation (DSSP) of crystallographic structures of VKOR-like protein from *Takifugu rubripes* (free ligand, PDB ID: 6wvi) and human VKOR (co-crystallized with warfarin and glycerol monooleate, PDB ID: 6wv3). α- and 3_{10}-helices are shown in red and blue respectively and numbered as in [221]. L-loop segment is delimited by a blue bar; positions of cysteine residues indicated by orange balls. (**b**) Structure of human VKORC1 co-crystalized with warfarin and glycerol monooleate (PDB ID: 6wv3) (left). H-bonds (middle) and hydrophobic interactions (right) stabilising warfarin and glycerol monooleate in structure 6wv3. (**c**) L-loop conformation from crystallographic structures 6wvi (apo-c) and 6wv3 (holo-c). (**d**) H-bonds stabilising the cap and hinge in L-loop in two crystallographic forms showed as a string diagram. Residues contributing to H-bonding by

their side chains are labelled in blue bold. (e-f) H-bonds stabilising α-helix and turns in L-loop cap in two crystallographic forms (**e**), and bifurcated and trifurcated H-bonds of residues N80 and Q78 (**f**) in both crystallographic structures, illustrated for holo-c form. (b-c, e) Protein is shown as cartoon with L-loop in blue (PDB ID: 6wvi) and light blue (PDB ID: 6wv3), TMD domain in brown (PDB ID: 6wvi) and light brown (PDB ID: 6wv3), and disulphide bridges as yellow (6wvi) and orange (PDB ID: 6wv3) sticks; non-covalent interactions (H-bonds and hydrophobic contacts) are shown by yellow dashed lines. Figures were prepared from atomic coordinates. The numbering of residues corresponds to the respective sequences of crystallographic data.

5.3.2.2. WHY ARE TWO CRYSTALLOGRAPHIC FORMS OF THE SAME HVKORC1 STATE NOT REVERSIBLE DURING MD SIMULATIONS?

The structure of human VKORC1 crystallised in holo-c form (PDB ID: 6wv3) was studied by conventional molecular dynamics (cMD) simulation (200-ns trajectory) without ligands. Surprisingly, removal of ligands did not produce expected effects on cMD hVKORC1 conformations (**Figure 5.19**). Indeed, although cMD data revealed transient folding and flexibility of L-loop, evidenced by (i) increased RMSD values, (ii) instability of the regular fold showing the reversible transition of helices to coil, and (iii) flexibility of L-loop inter-helices linkers, the observed effects did not provide significant plasticity that would lead the transition of holo-c to apo-c form. Upon cMD simulation, the cap H-bonds are significantly weakened or vanished, while almost all H-bonds in hinge are maintained despite highly transient folding.

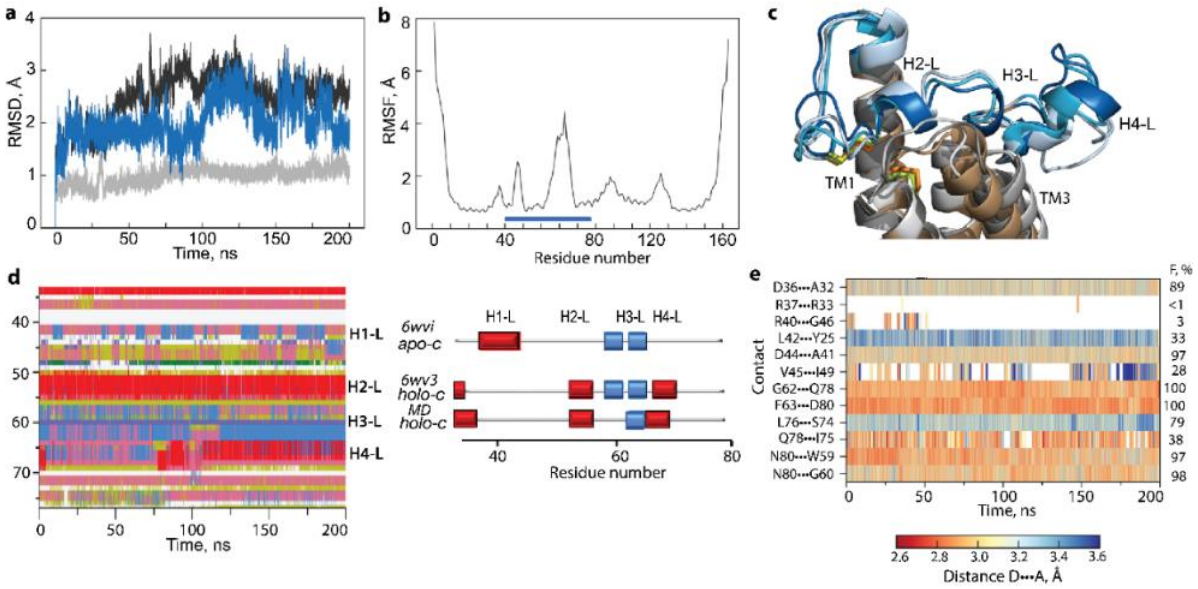


Figure 5.19 Analysis of cMD simulation of hVKORC1 holo-c form without warfarin. (**a**) RMSDs from initial coordinates computed for the whole protein Cα-atoms (black), TMD (grey) and L-loop (blue) after fitting on the crystallographic structure related domain. (**b**) RMSFs computed for Cα-atoms after fitting on the crystallographic structure. Blue horizontal bar delimits L-loop residues. (**c**) Superimposition of L-loop conformations randomly picked at t= 165 and 200 ns on crystallographic structure (PDB ID: 6wv3). Protein is shown as cartoon with disulphide bridges as sticks. VKOR conformations and their structural elements are distinguished by colour: TMD helices,

L-loop and disulphide bridges in sand, dark blue and yellow (PDB ID: 6wv3); in dark grey, sky blue and lemon (cMD frame picked at t = 165 ns); and in light grey, light blue and orange (cMD frame picked at t = 200 ns). (**d**) (Left) Time-dependent evolution of each residue secondary structures as assigned by the Define Secondary Structure of Proteins (DSSP) method for L-loop: α-helices are in red, 3_{10}-helices in blue, turn in orange and bend in dark yellow. (Right) Interpretation of L-loop secondary structures in two crystallographic forms (two first lines) and for the average MD conformation (third line) of holo-c form. α- and 3_{10}-helices are shown in red and blue respectively and numbered as in [221]. (**e**) Frequency (F, %) and length of H-bonds observed upon cMD simulation of holo-c form.

We hypothesised that (i) the 200-ns cMD simulation was insufficient to observe the expected holo-c to apo-c transition, and (ii) the extended conformational sampling of both forms would help efficiently detect their conformational spaces overlap and observation of the expected transition. To generate comparable datasets for two forms, human full-length hVKORC1 homology models, apo-h and holo-h, were built using the crystallographic structures, apo-c and holo-c, as templates. Then, to improve conformational sampling, each form of the human protein was investigated using the robust Gaussian accelerated MD (GaMD) methodology[261].

Analysis of the 500 ns GaMD simulation data (RMSD, RMSF, secondary structures) showed that the well-conserved TM helices of hVKORC1 vary slightly only at their ends, while L-loop helices exhibit an unstable fold in each form and a tendency to equalise the fold between two forms (**Figure 5.20**). Segment P37-C43, either folded as an α-helix (H1-L) in apo-h or as a random coil in holo-h, preserves its structure observed in the crystallographic template. In contrast, other L-loop segments show unstable fold in both forms. Particularly, fragment S52-F55, a coil in apo-c, appears as a transient 3_{10}-helix, further stabilised as an α-helix, and corresponds well to H2-L helix observed in holo-h (and holo-c) form. Moreover, segment V66-V69, stabilised as H4-L helix in holo-c, is essentially unfolded in the second half of GaMD simulation, similarly to apo-c. Despite these obvious changes in L-loop secondary structures, the spatial positions of the cap and hinge regions relative to TMD are preserved in both forms.

Curious fixedness of L-loop observed during GaMD simulations of models derived from the crystallographic structures of apo and holo forms, evokes the question of L-loop role in hVKORC1 activation. L-loop rigidity conflicts with its primary functions – recognition and binding of redox protein by hVKORC1 – for which L-loop must be easily adaptable, therefore highly flexible. Moreover, the secondary structures variation observed in L-loop, particularly transitions from the helix to coil or vice versa, leads to the coil elongation associated with the increase of L-loop flexibility.

To understand the factors contributing to L-loop fixedness, we analysed the H-bonds stabilising each form. This analysis revealed that residue Q78 participates in a unique stable H-bond, Q78···G62, well preserved during GaMD simulations in both forms (occurrence of 100 %), while its bonding to I75 is maintained in only 25 and 24% of conformations of apo-h and holo-h form respectively (**Figure 5.20**, **f**). The

trifurcated H-bonding of N80, observed in the crystallographic structures, has fully disappeared in both forms during GaMD simulations. Consequently, there is no stable H-bond interaction between L-loop and TMD that would hold them together in both forms. Possibly intra−L-loop Q78···G62 H-bond stabilises the flattened conformation of hinge and holds this rigidified hinge close to TM2, limiting its displacement.

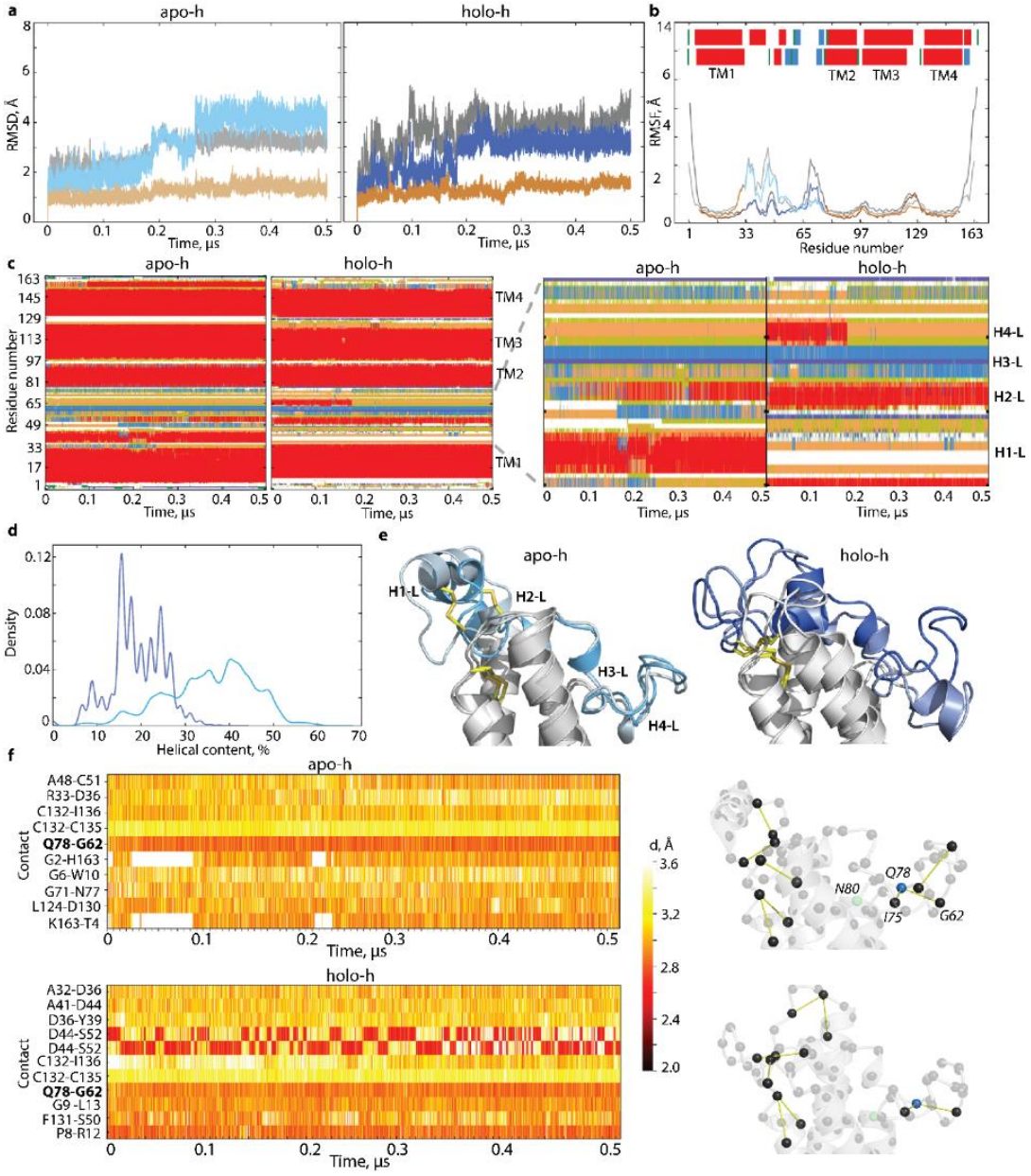


Figure 5.20 Accelerated MD simulations (GaMD) of human VKORC1 homology models, derived from crystallographic structures. (**a**) RMSDs computed for Cα-atoms of the full-length protein, TMD, and L-loop after fitting on the initial conformation (t = 0 ns) of the related domain. (**b**) RMSFs computed for the Cα atoms of two forms GaMD conformations after fitting on the respective average conformations. (a-b) Apo-h and holo-h forms are in light and dark colours respectively: L-loop in blue, TMD in brown, total hVKORC1 in grey. (**c**) Time-dependent evolution of each residue secondary structures as assigned by DSSP method for L-loop: α-helices are in red, 3_{10}-helices in blue, turn in orange and bend in dark yellow. (**d**) Helical content of each form. (**e**) Superimposition of hVKORC1 conformations picked at 0 and 500 ns. Protein is shown as cartoon with disulphide

bridges-forming cysteine residues in yellow sticks. TMD helices are in grey, L-loop conformations are distinguished by colour: t = 0 and 500 ns are in rey and blue for apo, and dark blue and violet for holo-c form (**f**) (Left) Non-covalent contacts (time series of H-bond events for H-bonds observed with frequency ≥ 0.8) in two forms of the oxidised hVKORC1. The contacts strength is shown by colour: from the strongest (2.7 Å, in red) to the weakest (3.6 Å, in white). (Right) Graphs of non-covalent contacts zoomed on L-loop. Vertices represent residues and a link between two residues reflects H-bonding.

5.3.2.3. ROLE OF H-BOND Q78···G62 ON L-LOOP INHERENT DYNAMICS

We assumed that (i) breaking Q78···G62 H-bond would give L-loop more mobility and (ii) if this contact is functionally crucial for L-loop, it will be restored upon MD simulation. To examine this hypothesis, two datasets, relaxed apo-h and relaxed holo-h, were produced in 500 ns cMD simulations of apo-h and holo-h forms so that during the first 100 ns of cMD H-bond Q78···G62 was prevented; then, both forms were simulated with the fully lifted restriction.

By comparing the relaxed models with the models stabilised by H-bond Q78···G62, we observed that the H-bond alternation (its presence or absence) is connected to L-loop folding (**Figure 5.21**).

Despite the cap secondary structures conservation in both models (H1-L helix in apo-h and a random coil with two turns in holo-h), (i) H2-L helix formed in apo-h over GaMD simulation, was preserved in the relaxed apo-h model only during the first 100 ns (with prevented H-bond), and later (ii) (with the fully lifted restriction on the L-loop ends), was transformed into short-lived 3_{10}-helix which was further partially denatured in a coil; (iii) two adjacent 3_{10}-helices (H3-L) were converted into a unique helix folded as an α-helix in the relaxed apo-h and a 3_{10}-helix in the relaxed holo-h ; (iv) H4-L helix was also unstable in two forms, reversibly transiting from α-helix to coil. As a result, the number of folded structures (α- and 3_{10}-helices) in L-loop of each relaxed model was not identical to the forms stabilised by H-bond Q78···G62 (**Figure 5.20**, **c**, **d**; **Figure 5.21**, **b**, **c**). Each relaxed form shows a clear evolution of its folding, increased in apo-h and reduced in holo-h, demonstrating a tendency to be comparable in both forms.

Curiously the expected restoring of H-bond Q78···G62 upon the fully lifted restriction was not observed during the simulation of any hVKORC1 forms studied.

In addition to evident change in L-loop folding, the relaxed forms demonstrate significant conformational mobility, viewed by the displacement of their helices with respect to TMD and the decrease of the distance between centroids defined on cap and hinge of L-loop especially noticeable in relaxed apo-h model (**Figure 5.21**, **e**). Conformations of the relaxed forms showed that a temporary restriction of H-bond Q78···G62 promotes the displacement of two L-loop regions, hinge to cap, and that this effect is more pronounced in relaxed apo-h form. The proximal position of cap

and hinge was observed early in the predominant L-loop conformation of the *de novo* model[220].

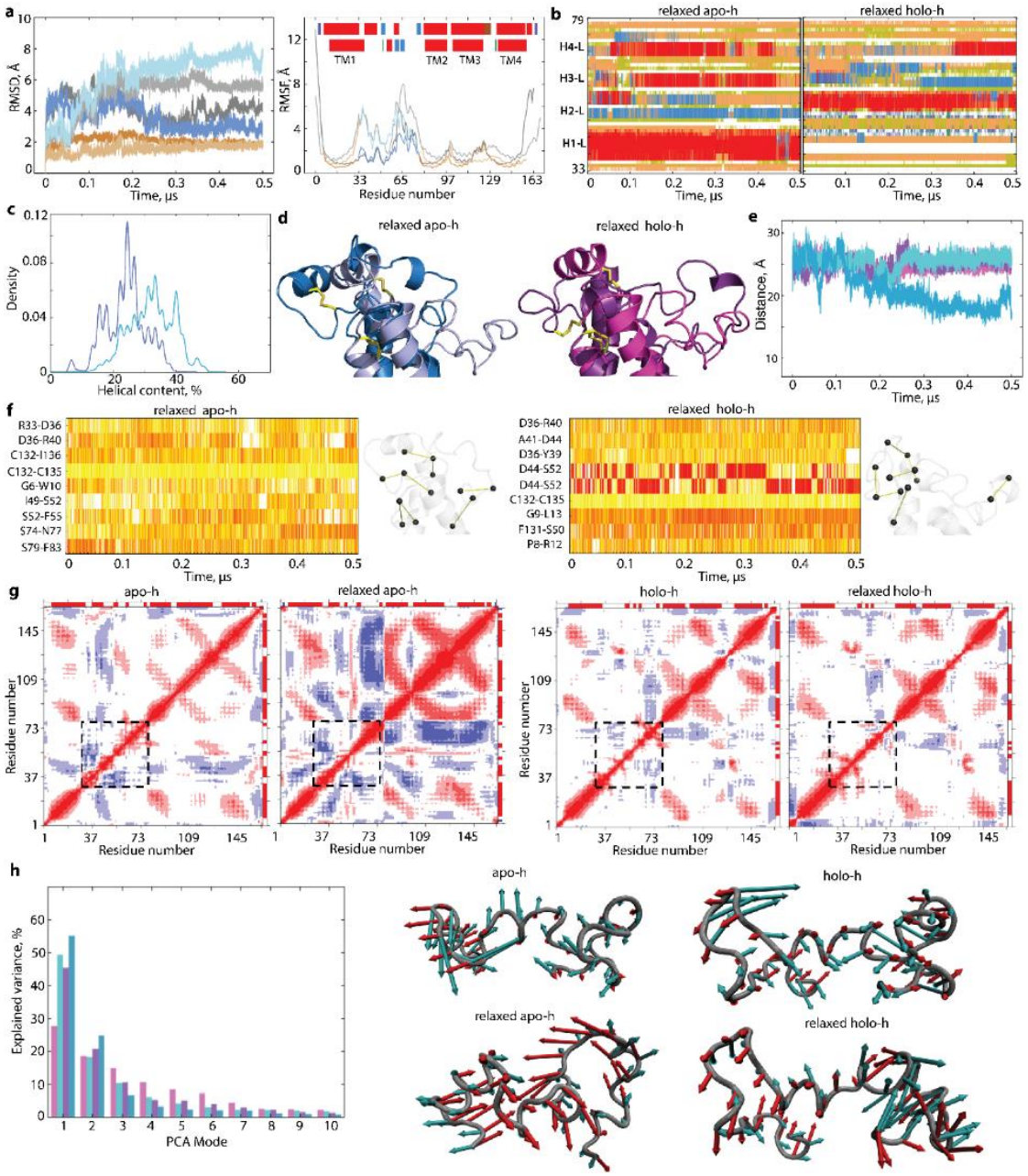


Figure 5.21 Conventional MD simulations of the homology models of hVKORC1 with conditioned H-bond Q78···G62. (**a**) RMSDs from the initial coordinates computed for Cα-atoms of the full-length protein (grey), TMD (brown) and L-loop (blue) after fitting to the initial conformation (t = 0 ns) of each related domain (left). RMSFs computed for the Cα atoms of each form after fitting on the initial conformation (right). (Insert) Secondary structures interpretation of each form average structure, relaxed apo-h (top) and relaxed holo-h (bottom). α- and 3_{10}-helices are shown in red and blue respectively. Conformations are distinguished by colour: relaxed holo-h and relaxed apo-h are in dark and light respectively. (**b**) Time-dependent evolution of each residue secondary structures as assigned by DSSP for L-loop: α-helices are in red, 3_{10}-helices in blue, turn in orange and bend in dark yellow. (**c**) Helical content of each form. (**d**) Superimposition of randomly chosen conformations of relaxed apo-h and holo-h forms (in dark colours) into the conformations taken

at t = 0 ns (in light colours). Protein is shown as cartoon with disulphide bridges-forming cysteine residues in yellow sticks. (**e**) Distance between the N- and C- ends of L-loop. (**f**) Non-covalent contacts (occurrence ≥ 80%) in two forms of the oxidised hVKORC1. The strength of contacts is shown by colour: from the strongest (2.7 Å, in red) to the weakest (3.6 Å, in white). Graphs of each form non-covalent contacts are shown on the right. Vertices represent residues and a link between two residues reflects H-bonding. (**g**) Intrinsic motion of hVKORC1 and its L-loop represented by the inter-residue cross-correlation maps computed for Cα-atom pairs after fitting on the initial conformation of apo-h and holo-h forms, and their relaxed versions. Secondary structures are projected onto the matrices borders (α-helix/β-strand in red/blue). L-loop is delimited with dashed lines. Correlated (positive) and anticorrelated (negative) motions between the Cα-atom pairs are shown as a red–blue gradient. (**h**) L-loop PCA modes calculated for each MD trajectory after least-square fitting of the MD conformations to the average conformation. The bar plot gives the eigenvalue spectra in descending order for the first 10 modes (left) Conformations are distinguished by colour: apo-h is in blue light, relaxed apo-h is in rose, holo-h is in dark blue and relaxed holo-h is in purple (middle and right). Atomic components in the first PCA modes of the L-loop in apo (middle) and holo (right) form are drawn as red (1st mode) and blue (2nd mode) arrows projected onto the respective average structure. Only motion with an amplitude ≥ 4 Å is shown.

Together with the significant structural and conformational reorganisation, increasing dynamical coupling between TMD and L-loop was observed only in apo-h form with the relaxed H-bond Q78···G62 as showed by the cross-correlation maps (**Figure 5.21**, **g**). The strong positive correlations inside TMD, early attributed to the collective drift of TM helices[220], were observed in all models.

Similar to the increased RMSF values, the first two PCA modes denote essential collective motions of L-loop that are greater in the relaxed forms. The two first modes characterize most of L-loop motion displaying large-amplitude collective movements of helices and adjacent coiled linkers (**Figure 5.21**, **h**). The amplitude and direction of L-loop motion in relaxed models of each form, especially apo-h, are increased compared to the forms stabilized by H-bond Q78···G62, suggesting a larger conformational space for the relaxed L-loop.

5.3.3. Discussions

As was reported earlier, the native inactive hVKORC1 *de novo* model represents a transmembrane four-helix bundle crowned by intrinsically disordered luminal loop, protruding in the endoplasmic reticulum lumen[220]. hVKORC1 published X-Ray structures with terminals restrained by green fluorescent protein[218] mostly confirm the *de novo* model correctness. First, they delivered a solid empirical affirmation of hVKORC1 four-helix TM domain, initially predicted by *de novo* modelling. Second, the conformational transition reported in [218], is a strong argument of L-loop flexibility observed in *de novo* model. Nevertheless, a careful analysis of MD simulation data obtained for the human VKORC1 homology models built from crystallographic structures reported two forms of the enzyme oxidised state, demonstrated the L-loop

curious fixedness and raises the questions of L-loop role in hVKORC1 activation. L-loop rigidity conflicts with its primary functions − recognition and binding of redox protein − for which L-loop must be easily adaptable and highly flexible. Moreover, L-loop expected conformational transformation, explained by the authors in terms of open-to-closed transition[218], was not observed under MD simulations of two forms in identical conditions (unbound protein in water solution).

Our search for sterical and physical conditions required for L-loop transition identified intra L-loop H-bond, Q78···G62, as a main factor leading to L-loop rigidity. Even a short-lived constraint preventing this H-bond formation increased L-loop flexibility in both hVKORC1 forms delivered from crystallography. Moreover, extended cMD simulations of isolated L-loop, cleaved from crystallographic forms showed per se rupture of this H-bond. Apparently, Q78···G62 H-bond stabilises certain L-loop conformations under particular circumstances (e.g., protein crystallisation conditions), but generally hVKORC1 L-loop poses great flexibility. This L-loop quality is mandatory, and determined by its functional role to easily adapt its conformation in response to an external stimulus (redox protein) or biochemical (e.g., thiol-disulphide reaction) factors.

It has been established that increased intrinsic plasticity represents an important prerequisite for effective molecular recognition[300,329–331] and long-range conformational changes mediates enzymatic reactions[332].

As hVKORC1 is an enzyme using thiol-disulphide reaction for its activation by a redox protein followed by protons transfer to the active site for vitamin K processing, hVKORC1 intrinsically disordered L-loop is the excellent and optimal platform to ensure this complicated multi-step reaction. This biochemical process requires transformations between the functional groups –SS–, –SH and –S• of two proteins, hVKORC1 and its redox protein, during the transfer of two protons/electrons. The kinetics and mechanisms of thiol–disulphide substitution and redox reactions having a pivotal role in biology are well-described for different small molecular and enzymatic systems[333,334]. In particular, computational study of thiol-disulphide exchange reactions reactivity in thioredoxins and in other proteins concluded that these reactions are critically fine-tuned by the active site atomistic details[335]. A prime example is the hydrophobic pocket around the thioredoxin family CXXC motif, the geometry, dynamics and electrostatic environment of which decide on the redox potential and kinetics[336]. Given the multitude of possible thiol-disulphide exchange reactions, an important but hitherto unresolved question is how specificity is achieved. Nevertheless, the highly dynamic disordered nature of regions playing pivotal role in such reactions was systematically reported [337].

All data generated by MD simulation of two different crystallographic forms, their models derived from these structures and the *de novo* model, confirmed that L-loop intrinsic disorder consists of two reversible processes − transient folding and high

conformational flexibility – leading to L-loop enormous conformational diversity ranging its conformations from closed compact globule-like shape (the most prevalent) to rare open elongated boat-shape through limitless number of intermediaries. All structurally organised L-loop segments are involved in the reversible folding-unfolding process (structural transitions), and L-loop exhibits great conformational flexibility supplied by linear and rotational displacements, and their combination in either folded or coiled segments.

L-loop, as a highly disordered region, possesses large conformational plasticity supplying great capacity for multiple structural and conformational arrangements required for different steps of the thiol-disulphide exchange reaction leading to hVKORC1 activation. Such interpretation of hVKORC1 structure does not contradict an increased propensity of L-loop to be disordered. L-loop sequence contains a large part of polar and charged residues (54%) while the hydrophobic residues portion is reduced to 44%, a typical composition of ID proteins[302,331]. It was shown that L-loop flexibility depends on the oxidation/reduction state of hVKORC1, and L-loop of its fully oxidized (inactive state) is considerably less flexible and more folded compared to the reduced states[220]. Nevertheless, 'oxidized L-loop' demonstrates a remarkable structural and conformational plasticity evidenced by variation of its helices in length and spatial position, giving rise to myriad L-loop conformations. As L-loop transient folding (at the secondary structures level) was observed in the quasi-rigid (apo-h and holo-h) and flexible (relaxed apo-h and *de novo*) species, this process is possibly disassociated with L-loop conformational flexibility.

5.4. Conclusion sur le desordre du RTK KIT et de hVKORC1

La modélisation et la simulation de dynamique moléculaire de deux protéines, le RTK KIT et hVKORC1, a mis en évidence leur caractère hybride. Ces protéines sont composées de domaines ordonnés (domaine kinase du KIT et domaine transmembranaire de hVKORC1), et de régions désordonnées (JMR, KID, boucle A et C-tail pour KID, boucle L pour hVKORC1).

Ces IDRs possèdent de nombreuses structures secondaires transitoires, mais réversibles durant les simulations, et une plasticité conformationnelle leur conférant une haute ou moindre flexibilité. Ce désordre intrinsèque diffère d'une IDR à l'autre mais également d'une protéine à l'autre. De plus, la variabilité de la position relative des domaines les uns par rapport aux autres ainsi que leur couplage dynamique fort caractérise leur désordre extrinsèque.

Ainsi, nous avons pu caractériser le DYNASOME de ces deux protéines, et en particulier leurs régions désordonnées délivrant des cibles soigneusement étudiées pour l'exploration de leurs interactions avec leurs protéines partenaires et le développement de nouvelles molécules thérapeutiques.

CHAPITRE 6. MODULARITE DES PROTEINES RKT KIT ET HVKORC1

Nous avons vu que les modules sont des régions protéiques (quasi) indépendantes liées par des boucles flexibles souvent désordonnées. La question que nous pouvons nous poser est la suivante : est-ce que ces régions fonctionnelles désordonnées peuvent être considérées comme des modules de leur protéine respective ? Quel est le degré d'indépendance de ces modules ? Quelle est leur application pratique ?

Dans ce chapitre, nous tenterons de comprendre la modularité du RTK KIT et de hVKORC1 et comment celle-ci peut être utile pour l'étude des relations intermodules et des interfaces entre domaines protéiques.

Ce chapitre est une adaptation des articles suivants :

1. **Ledoux, J.**, Trouvé, A., & Tchertanov, L. (2021). Folding and Intrinsic Disorder of the Receptor Tyrosine Kinase KIT Insert Domain Seen by Conventional Molecular Dynamics Simulations. *International Journal of Molecular Sciences*, *22*(14), 7375. https://doi.org/10.3390/ijms22147375

2. **Ledoux, J.**, & Tchertanov, L. (2022). Does Generic Cyclic Kinase Insert Domain of Receptor Tyrosine Kinase KIT Clone Its Native Homologue? *International Journal of Molecular Sciences*, *23*(21), 12898. https://doi.org/10.3390/ijms232112898

3. **Ledoux, J.**, Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

Les données supplémentaires et les méthodes relatives à toutes ces publications sont présentées dans les annexes de la thèse.

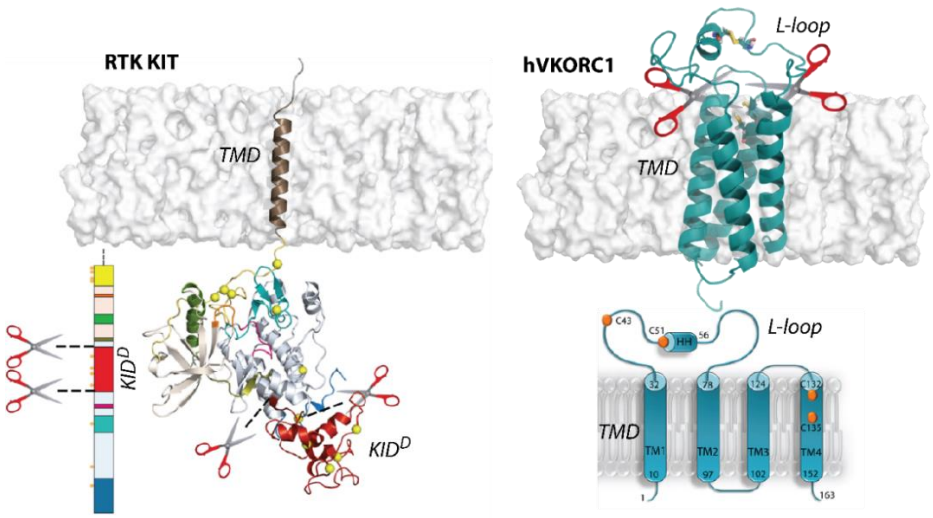


Figure 6.1 Abstract graphique du chapitre.

6.1. LE DOMAINE INSERT KINASE (KID) DU RTK KIT

Résumé. *L'une des régions clés de recrutement des protéines partenaires de signalisation du RTK KIT est le KID. Cette région intrinsèquement désordonnée a été étudiée par simulation de dynamique moléculaire en tant que polypeptide clivé du reste du CD de KIT. Ce KID clivé a été simulé soit avec restriction de ses extrémités terminales pour imiter la présence du domaine kinase, soit libre dans le solvant. Nous avons montré que le KID clivé préserve des propriétés structurales et dynamiques analogues à sa forme jointe au KIT : un repliement hélicoïdal composé d'hélices α et 3_{10} transitoires, une forme globulaire maintenue par des interactions non covalentes majoritairement spécifiques. Construits par différentes coordonnées de réactions, les paysages d'énergie libre montrent les nombreux minima locaux des ensembles conformationnels hétérogènes. Toutes ces propriétés sont associées au désordre intrinsèque de KID qu'il soit clivé ou non du reste du CD de KIT. Par l'ensemble de ces résultats, nous avons conclu que le KID est un domaine ou module intrinsèquement désordonné du KIT structurellement quasi indépendant du CD. Cette comparaison nous a permis de constater que le KID clivé simulé avec des contraintes reflète mieux les propriétés structurales et dynamique du KID joint au domaine kinase, et pourra être utilisé pour l'étude* in silico *des effets de phosphorylation sur le KID.*

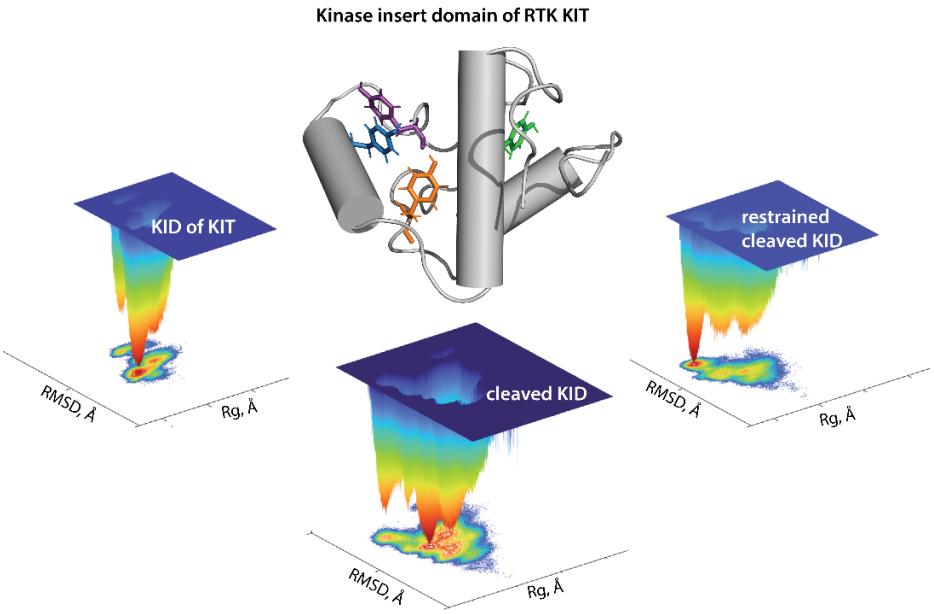


Figure 6.2 Abstract graphique de la section[290].

6.1.1. INTRODUCTION

Receptor tyrosine kinases (RTKs) act as sensors for extracellular ligands, the binding of which trigger dimerisation of RTK, activation of its kinase function and auto-phosphorylation of specific tyrosine residues in the cytoplasmic domain[65,66]. This leads to the recruitment and activation of multiple downstream signalling proteins, which carry the signal to the nucleus, where it alters patterns of gene transcription such as those governing various aspects of the cell physiology. Initiation of this cascade-like process involves different regions of the multidomain RTKs, each of them performing specific actions that are finely concerted by a tightly regulated allosteric mechanism controlling all functional processes of RTK[96]. Explicit elucidation of the signalling cascade represents an important and unsolved problem in cell biology.

The modular extracellular domain (ED) of RTK, containing characteristic structural motifs involved in specific ligand binding, is formed by various motifs (Ig-like, cysteine-rich, cadherin fragments, etc.) interconnected by coiled linkers, providing high conformational plasticity (**Figure 6.3**, **A**). The binding of ligands affects the monomer–dimer equilibrium of RTKs by stabilising the dimeric state through a global conformational change in the ED[338]. The signal induced by ligand-binding propagates across the transmembrane (TM) domain to the cytoplasmic domain (CD) and promotes its activation. The cytoplasmic domain of RTKs also has a modular structure composed of the juxtamembrane region (JMR), the split tyrosine kinase (TK) domain with the proximal (N-) and distal (C-) lobes linked by a kinase insert domain (KID), and the C-terminal tail (C-term).

We concentrate our study on the RTK KIT from the PDGFR (III-type) family. The physiological actions of KIT controlling cell survival, proliferation, differentiation, and migration depend on the activation of specific or overlapping pathways[339], which endows the activity of the SCF/KIT system (where SCF is the stem cell factor that regulates the KIT activation and therefore, triggers the initiation of multiple signal transduction pathways) of great complexity. Aberrant regulation of KIT signalling networks is associated with the progression of many cancer types, including human acute myeloid leukaemia, aggressive systemic mastocytosis, melanoma, gastrointestinal stromal tumour, and stomach cancers[281,282]. Disclosure of the KIT activated pathways in carcinogenesis will be a crucial step towards the development of KIT targeted therapies[284].

Functions of the TK domain of KIT, similarly to other RTKs, are mainly attributed to catalytic activity and trans-phosphorylation, while the post-transduction processes and the binding of intracellular proteins are associated with JMR, KID and C-terminal, which are the regions possessing multiple phosphorylation sites[65] (**Figure 6.3**, **B**). The overall structural feature of these flexible fragments is the intrinsic disorder which is a vital condition for the creation of the dynamic networks of interacting proteins[340].

A first structural model of the full-length cytoplasmic domain of KIT has been published[183] (**Figure 6.3**, **C**). The primary analysis of the structural and dynamical properties of this model showed that the conformational variability of KIT is provided mainly by JMR, KID and C-terminal, which were interpreted rather as intrinsically disordered (ID) regions demonstrating significant structural and conformational plasticity.

Because ID proteins (IDPs) lack stable secondary structures (2D) under physiological conditions, they exist as heterogeneous conformational (3D) ensembles, and are capable of rapidly changing conformation upon influence by an effector (e.g., binding of ligand/cofactor/protein). Principally, the determination of a single structure of the ID regions or ID proteins has no physical relevance, as such structures present only an isolated element (a lone conformation) from a huge conformational space. A more pertinent approach is the description of such proteins in terms of the probabilistic population of the different regions, and the correlation of these probabilities with the protein function. Most experimental techniques employed to study IDPs[341] suffer from a conundrum: the empirical observables, which make it possible to assess the protein disorder, represent an average over the conformational ensemble, but the ensemble itself cannot be unequivocally inferred from the experiments. Therefore, computational methods provide an advantageous approach for analysing IDPs. Molecular dynamics (MD) simulation provides great insights into this challenge[342], as an approach that can, in principle, produce the structural ensemble of biomolecules.

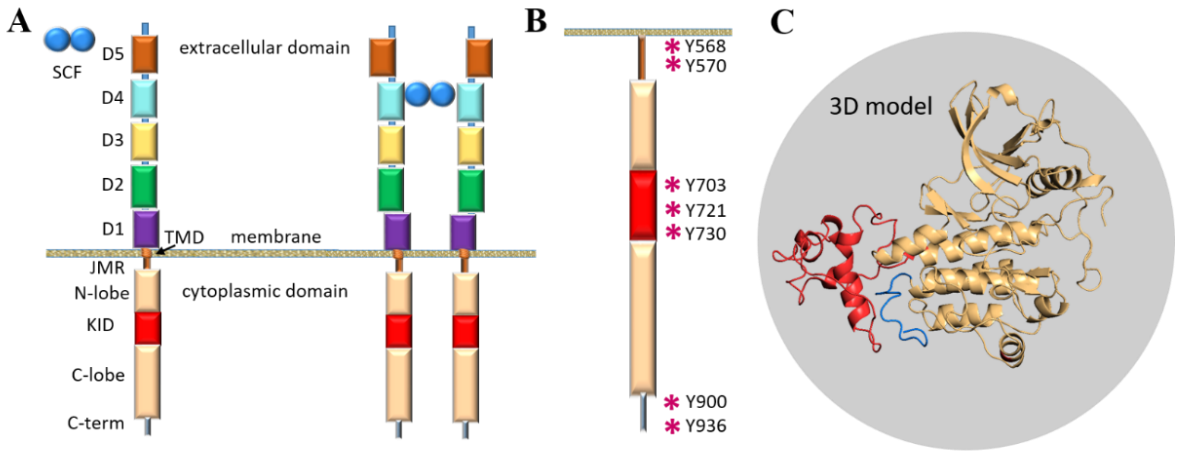


Figure 6.3 The modular structure of RTKs illustrated with KIT, a member of the RTK family III. (**A**) Structural composition of KIT: an extracellular domain (ECD) with five Ig-like regions (D1–D5), a transmembrane domain (TMD) and a cytoplasmic domain (CD) comprising a juxtamembrane region (JMR), an ATP-binding region (N-lobe), the phosphotransferase domain (C-lobe) with a kinase insert domain (KID) and a C-terminal. The stem cell factor (SCF) extracellular binding induces dimerisation and activation of KIT. (**B**) The tyrosine residues (Y, showed by asterisk) of JMR, KID and C-terminal identified as the phosphorylation sites involved in the recognition of cellular partners. (**C**) 3D structural model of the KIT CD with KID (red) and C-term (blue)[183].

In the present study, the focus is on the KID of KIT, a binding hub assuming exquisite specificity of the receptor. As has been reported that five functional phosphorylation sites of KID from KIT, three tyrosine (Y703, Y721, Y730), and two serine (S741 and S746), provide the alternative binding sites for the adaptors, signalling and scaffolding proteins in the cytoplasm[287]. Phosphorylation of Y703 supplies the binding site for the SH2 domain of Grb2, an adaptor protein initiating the Ras/MAP kinase signalling pathway. Phosphorylated Y721 and Y730 are the recognition sites of PI3K and phospholipase C (PLCγ), respectively. The function of Y747 has not yet been described. Phosphorylated serine residues, S741 and S746, bind PKC (protein kinase C) and contribute to re-control of PKC activity under the receptor stimulation. The functional importance of KID is also emphasised by newly identified mutations of K704, N705 and S725, which were reported as activating in gastrointestinal stromal tumours[343]. Consequently, a study of KID structure-dynamics features and their relation to KID function is still crucial but/and obviously challenging.

The conformational dynamics of KID from KIT was probed by conventional molecular dynamics (cMD) simulations. This method generates the atomic-level data required for a detailed analysis of the conformational space and the identification of folding intermediates related to function and/or characterisation of functionally important phenomena related to allosteric regulation[344]. Allostery is often a fine adaptive mechanism of protein modulation in the cellular environment (e.g., adaptation of binding partners to form biomolecular assemblies) during signal transduction, catalysis, and gene regulation[95,96].

One approach for studying the assembly of multidomain proteins and their folding is to use the modular domain of the protein, which preserves binding capabilities even when the domain is removed from the context of the full-length protein[7]. The ability of modular protein domains to independently fold and bind both *in vivo* and *in vitro* has been taken advantage of by a significant portion of proteomics studies that have used modular domains to assess the protein–protein interactions required for a diverse set of cellular processes, including signal transduction and subcellular localisation.

The use of KID as a cleaved polypeptide represents a promising strategy for the exploration (empirically and numerically) of KIT signal transduction and the modulation of protein function by controlled interference with the underlying molecular interactions. Such use will be fully justified if we can prove that KID of KIT is a modular domain that preserves its structural and dynamic properties. As we seek draw conclusions with respect to the usefulness or the disadvantage of using cleaved KID as a reduced model to study the RTK KIT 'interactome'[345], a comparative analysis of this model with the KID of KIT was carried out. We also suggest a possible dependence of KID on the kinase domain of KIT, which can either be induced locally by geometric restriction on the KID terminus or globally promoted by long-range allosteric effects.

Our present study of KID from KIT focuses on its folding features and intrinsic disorder with the aim of defining a KID species suitable for the exploration of the KIT 'interactome'.

6.1.2. Results

6.1.2.1. Data Generation

The kinase insert domain of KIT, composed of the 80 amino acids (aas) (F689–D768), was examined by conventional MD simulations (all-atom, with explicit water) as the cleaved polypeptide (KID^{C}) and as the subdomain of the kinase domain (KID^{D} from KIT). The cleaved KID was simulated as a fully unconstrained entity (KID^{C}), and as a polypeptide with restrained distance between two Cα-atoms of terminal residues F689 and D768 (cleaved restrained KID, KID^{CR}), which is conserved in crystallographic structures of KIT[183]. The MD simulation run (of 1.8 µs for KID^{CR} and 2 µs for KIT and KID^{C}) was repeated four (KID^{C}) and two (KIT and KID^{CR}) times with different randomised initial atomic velocities to extend conformational sampling of each studied entity and to examine the consistency and completeness of the produced KID conformations. The other parameters of the MD simulations of KID and KIT were strictly identical. Each simulation was started with an equilibrated structure obtained after minimising the neutralised solvated protein which is either a 3D *de novo* model of KID or of KIT.

6.1.2.2. General Characterisation of MD Simulation Data

The data analysis was performed on each trajectory after the least-square fitting of MD conformations to the initial KID conformation, to avoid the motion of the domain as a rigid body.

First, the global stability of KID throughout the MD simulations was estimated using the root mean square deviations (RMSDs) computed on the Cα atoms relative to the initial KID model (at t = 0 µs) as a frame of reference. For the cleaved KID, simulated as the unconstrained polypeptide (KID^{C}), the RMSD profiles differ among trajectories 1–4, and their values vary along each trajectory, but the ranges of RMSD variations are comparable between trajectories (**Figure 6.4**, **A**, **top panel**). Well-resolved slopes on RMSD curves are either a single event showing a rapid increase of RMSD values, or a two-event process showing an alternating increase/decrease in RMSD. Such sudden changes in RMSD, observed at t = 0.40–0.55, 0.77–1.79 and 1.55–1.60 µs in trajectories 1, 2 and 4, respectively, reflect regular conformational transitions in KID suggested in [183] based on a single simulation. The RMSDs cover a large range of values (4–9 Å) arranged in two main peaks per trajectory with different (most

trajectories) or comparable (one trajectory) populations (**Figure 6.4**, **B**, **top panel**). Two main peaks in each replica are separated by 1–3 Å, while the most populated peaks, composed of similar conformations, from replicas 1–4 are shifted only by ~1 Å. While RMSD of MD conformations were calculated from the initial coordinates, which are identical in all trajectories, we can suggest (1) at least two groups of highly different KID conformations within each trajectory; (2) partial overlapping of conformations from different trajectories.

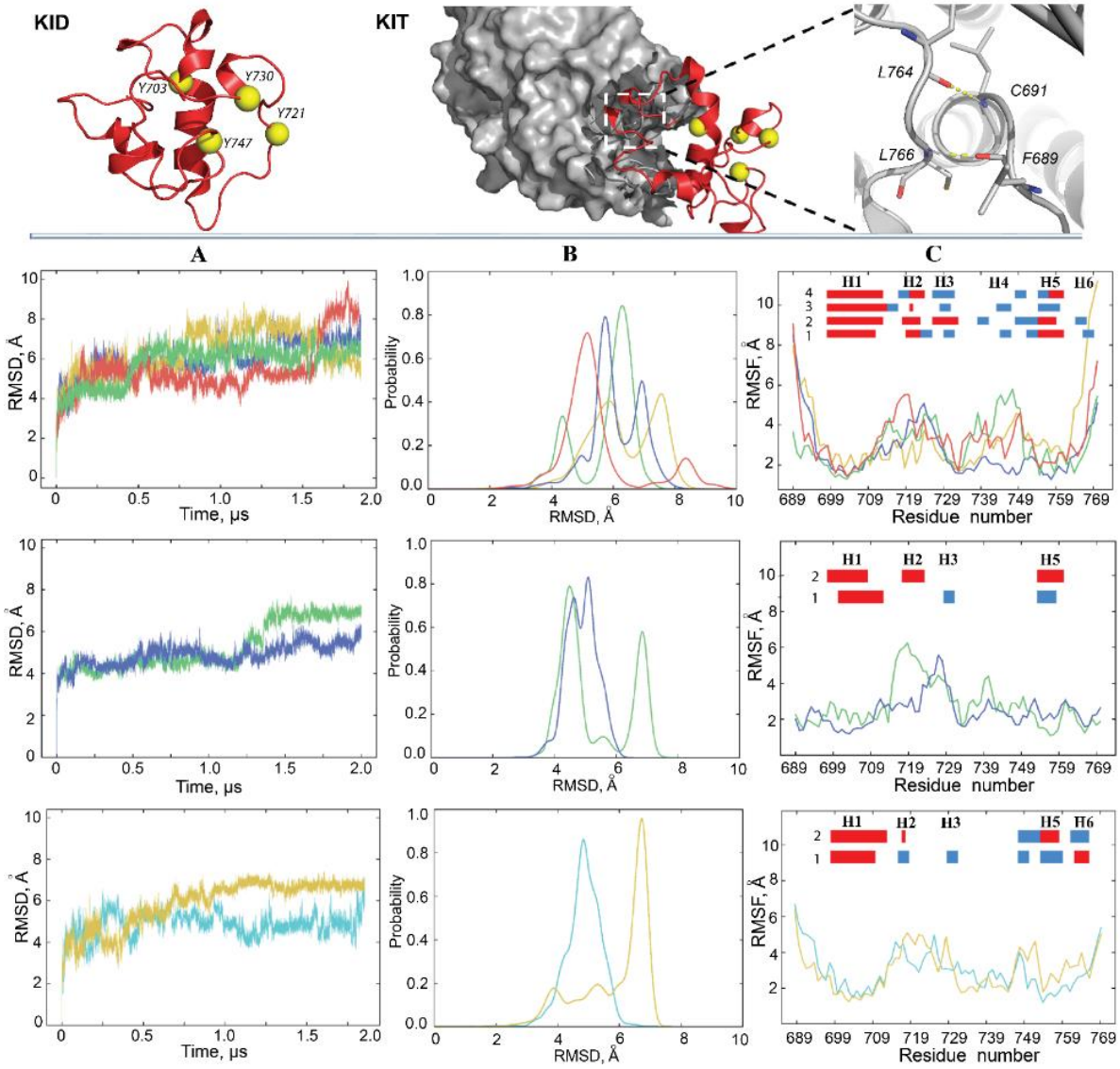


Figure 6.4 Conventional MD simulations of KID. (**Top panel**) Structural models of the cleaved KID (KID^{C}) (left) and KID fused to KIT (KID^{D}) (middle) in which the N- and C-ends are stabilised by pairs of H-bonds (right). Proteins are displayed as ribbons (KID), surface-filled model (kinase domain), yellow balls (the tyrosine residues) and sticks (residues formed H-bonds). Columns (A–C) show the statistical descriptors of KID^{C} (top), KID^{D} (middle) and KID^{CR} (bottom) computed on the all Cα-atoms of KID for MD conformations of each trajectory after fitting on initial conformation. (**A**) RMSDs, (**B**) probability distributions of the RMSDs, and (**C**) RMSFs computed on the Cα atoms after

fitting on initial conformation. In the insert, the folded secondary structures, αH- (red) and 3_{10}-helices (blue), labelled as H1–H6, were assigned for an average conformation of each MD trajectory. In (A–C), MD replicas are distinguished by colour: 1–4 of KID^C are in green, yellow, blue, and red; 1–2 of KID^D are in green and blue; 1–2 of KID^{CR} are in blue light and yellow respectively.

Like the cleaved KID, the RMSD curves of KID^D show either a rather dramatic increase or a series of small reversal changes (**Figure 6.4**, **A**, **middle panel**). The probability distribution of such RMSD variations appears as two distinct well-separated peaks or two overlapped distributions in a slightly narrow range with respect to KID^C (**Figure 6.4**, **B**, **middle panel**). RMSD of the cleaved KID simulated with the restrained distance between its N- and C-end vary within a range like KID^D; nevertheless, the probability distribution shows a single peak for each replica (**Figure 6.4**, **A**, **B**, **lower panel**). Second, the root mean square fluctuations (RMSFs) were calculated for each residue of KID. As expected, the highest RMSFs are mainly observed for residues at KID^C extremities, while fluctuations of these residues in KID^D and KID^{CR} are limited. Inter-residual contacts in KID^D led to the identification of two stable (occurrence of 97–99%) hydrogen bonds N–H···O liking two pairs of residues, C691–L764, and L766–F689, maintaining KID N- and C-end in very proximal position (**Figure 6.4**, **right**; **Figure S13**). Apparently, these strong H-bonds restrain the mobility of KID extremities as reflected in small RMSFs. Like KID^D, the N- and C-ends of cleaved KID are stabilised by H-bonds, but the other pairs of residues are involved.

Curiously, in all KID entities three hydrophobic residues, V731, V732 and P733, positioned on the random coil show small RMSFs in respect to preceded and followed residues, therefore, the RMSF curves in the different entities of KID display a comparable profile, which is described as the 'camel double-humped' contour (KID^C) or close to this contour (KID^D and KID^{CR}) (**Figure 6.4**, **C**).

The two descriptors, RMSD and RMSF, indicate the highly heterogeneous conformational composition of each data set obtained by cMD simulation of each KID entity. This data reveals the intrinsically disordered nature of KID from KIT, previously suggested in [183] based on a single trajectory of simulation. It is well known that a high content of polar and charged residues increases the propensity of a protein to be disordered[302,346,347]. KIT, composed of 58% such residues, is a good candidate for being an intrinsically disordered region, and this feature sequence-dependent and disconnected from KID context, either as a cleaved polypeptide or a domain of KIT.

Heterogeneous KID conformations were analysed with the ensemble-based clustering using the RMSD criterion[267]. To better grasp the conformational diversity of each studied KID entity, the conformations were clustered with the RMSD threshold r = 3, 4 and 5 Å. With Å, we obtained a reasonably limited number of clusters regrouping almost all the conformations generated on each trajectory (**Figure S14**–**Figure S16**). Since some clusters' representative conformations of different replicas showed striking similarity, we suggest that the conformational spaces generated by

the independent trajectories of each KID entity partially overlap. Clustering analysis of the merged (concatenated) trajectory with the same RMSD threshold (r = 4 Å) produced the number of clusters which is less than a sum of clusters obtained for each replica. Such results do not contradict the supposed overlap of conformational spaces sampled by replicated trajectories.

6.1.2.3. Folding and Compactness of KID

The secondary structure interpretation (DSSP) indicates that the helical fold of KID is constituted of α- and 3_{10}-helices (**Figure 6.4**, **C**; **Figure S17**). These helices are varied in number, length and $\alpha/3_{10}$ ratio. These variations are observed within a trajectory, between the trajectories, and for different KID species, but the position of some helices is curiously conserved.

KID^C is composed of six helices, H1–H6i is made up of 44–56% of amino acids. H1-helix, the largest (15–16 aas) of all KID helices, is a long-leaved α-helix well-conserved in all replicas. Other helices of varying lengths are rather transient, switching between α- and 3_{10}-helices (H2, H3 and H5) or between 3_{10}-helix and random coil, which is partially folded as reverse turn or bend (H4 and H6). Similarly, KID^D shows a helical fold, but the number of helices (H1, H2 and H3 or H5 on average structure) and the portion of residues (25–30%) constituting these helices are significantly diminished. In KID^D, like KID^C, H1-helix is the largest compared to other helices; however, its length is reduced to 11–12 aas. The H1-helix is long-lived, while the other helices are fully transient.

The highly diminished folding in KID^D prompts two hypotheses:

Hypothesis 1. The KID folding depends on the KID ends, which are stabilised by strong H-bonds in KID fused to KIT and highly flexible in the cleaved KID.

Hypothesis 2. The KID folding depends on the status of KID as an entity, which is either autonomous polypeptide or collateral subdomain influenced by the kinase domain.

The cleaved KID, simulated with the constrained distance between the terminal residues (KID^{CR}) showed that 30–35% of amino acids are involved in regular structures, the extended and long-lived α-helix H1 (15–17 aas) and three to five transient helices, indicating that the KID^{CR} folding (at the 2D level) is more ordered than in KID^D and less than in KID^C (**Figure S17**). It seems that the constrained polypeptide in general better represents the folding of the native KID than the unconstrained, but the expected equivalence was not observed. The mid-domain of KID is organised similarly in KID^{CR} and KID^D, while folding of C-terminal residues is quasi-identical in KID^{CR} and KIDC. It is probable that the applied constraints in KID^{CR} were rather soft compared to the

restricted intra-protein geometry in KID^D. The increased flexibility of border residues in KID^{CR} compared to KID^D, as seen by RMSFs, and the greater similarity of 2D folding of residues from C-end (H6) in KID^{CR} and KID^D, support this hypothesis. On the other hand, we can suggest a certain bias of such comparison derived from a different number of independent MD trajectories that is 4 for KID^C and 2 for KID^D and KID^{CR}. Interestingly, in all studied KID entities, residues forming H1 and H5 helices show the smallest RMSF values compared to the other helices.

The three-dimensional (3D) structure of KID in all studied entities represents a compact array of αH- and 3_{10}-helices linked by short or extended loops (random coils or turns) that play a principal role in the conformational diversity of KID. In the most of conformations, KID retains its collapsed globule-like shape that is slightly elongated towards the KID connected to the kinase domain, and therefore, is best described as a flattened (oblate) ellipsoid with an opening given by the distance between its ends (**Figure 6.4**). This opening in KID^D is controlled by hydrogen bonds linking residues from N- and C-extremities. The enlarged displacement of the border residues in the cleaved KID promotes a diminishing or a full destabilisation of such H-bonding.

To characterise the size and compactness of KID, the mass-weighted radius of gyration (Rg) was calculated for each entity of KID on MD conformations. We compared the distribution of Rg vs. RMSD between each ensemble of KID conformations (on every trajectory on each KID entity) (**Figure S18**). All distributions show at least two heavily populated regions separated by an area with a lower number of intermediates. The mean values of Rg for each peak are comparable in the two replicas on KID^D (and on KID^{CR}) but are slightly different between entities. In KID^D, the Rg of each peak (mv of 11.80(1) and 12.22(1) Å) shows a slightly increased value in comparison with KID^{CR} (mv of 11.33(1) and 12.00(2) Å). In KID^C, the Rg mean values of the heavily populated regions vary between replicas, but these variations spanned (cover) the Rg range observed in KID^D and KID^{CR} (mv of 11.64(1) and 12.52(2) Å).

Since strong variations in Rg (and RMSD) are observed along the same trajectory on cleaved KID, this indicates a higher conformational variability of KID^C, rather than more exhaustive sampling over the four replicates. Additionally, since the cluster analysis of the individual and concatenated trajectories showed at least a partial overlap of the conformational spaces generated on the replicas for each KID entity (**Figure S14**–**Figure S16**), the more in-depth analysis of each KID entity was performed on the concatenated data.

6.1.2.4. Folding and Intra-Domain Interactions of KID

The propensity of globular proteins to be compact is the key reason that their folded states achieve high packing density. We suggest that the contact map can characterise KID collapsibility. The high instability of folded structures, which are,

except for H1, transient, converting between αH- and 3_{10}-helices or between helix and random coil, together with their great mobility, producing an expected irregularity of contacts, resulted in the smeared pattern on the contacts maps (**Figure 6.5**, **A**).

The contact maps are sufficiently different between replicas of the same entity of KID (e.g., replicas 1–4 of KID^{C}), but strikingly similar between distinct KID entities (e.g., replicas 3 on KID^{C} and 1 on KID^{CR}, or replicas 4 on KID^{C} and 2 on KID^{CR}). Such patterns in the contact map can be linked, on the one hand, to a large difference in folding of the same KID entity observed in the MD replicas, and on the other hand, to a partial similarity of the conformational and structural characteristics of the different entities of KID.

According to the contact maps, the residues of N- and C-ends of the cleaved KID, despite their high (in KID^{C}) or moderate (in KID^{CR}) flexibility, are still involved in intramolecular contacts. Apparently, the change in conformation in the cleaved KID leads to the appearance of alternative hydrogen bonds, which are formed by the same residues as in KID^{D}, and the other neighbour residues (C691 ··· D765, S692 ··· D765, L766 ··· F689, E767 ··· I690).

The high-occurrence of contacts between multiple residues from the folded and randomly coiled segments of the mid-domain of KID^{C} indicate diverse non-covalent interactions between helices (H1, H2 and H5), helices and coiled regions, and randomly coiled regions. This large number of intra-molecular contacts suggests that multiple regions contribute to maintaining the inherent (intrinsic) 3D structure of KID in all studied entities, independently of the fold.

While a considerable portion of KID undergoes both large conformational changes and high fluctuations, we focused on residues showing low fluctuations (RMSF)—the 'pseudo-rigid' segments E699-S709, P754-L764 and V731-P733—and analysed their contacts with all the other residues (**Figure 6.5**, **B**, **C**). It is interesting to note that these 'pseudo-rigid' fragments show a different fold, a long-leaved helix (H1), a random coil linking H3 and H4 helices, and a transient helix (H5), respectively.

The most fascinating observations from the analysis of contacts formed by the 'pseudo-rigid' fragments are as follows: (i) a large number of KID residues are involved in intramolecular contacts, (ii) the charged and polar 'contacting' residues are the most abundant compared to hydrophobic residues, (iii) residues V728–D737 constituting the linker connecting H3 to H4 or H5 are regularly involved in intramolecular contacts, and (iv) residues of the N- and C-ends interact with the 'pseudo-rigid' segments. The quantitative estimation of the number of contacts stabilising KID is minimal in KID^{D}, while in the cleaved KID their number is increased to 46% (KID^{C}) and 52% (KID^{CR}). Interestingly, there are a fairly limited number of residues that do not contribute to intramolecular interactions in KIT.

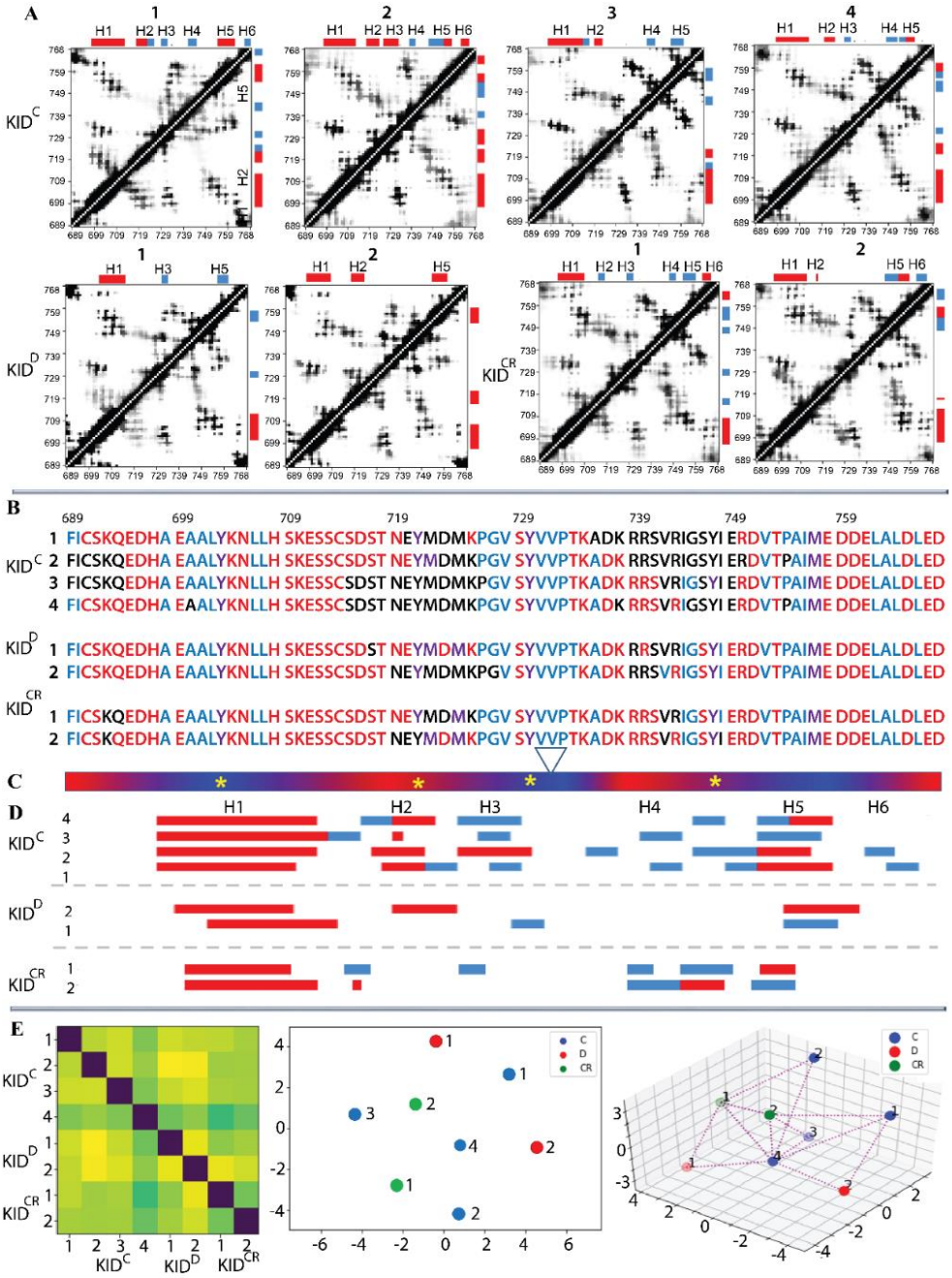


Figure 6.5 Estimation of intramolecular contacts in KID. (**A**) Dynamic contact maps of pairwise distances Cα−Cα < 10 Å computed for each MD trajectory. Black–white gradient shows a frequency (from 1 to 0) of contact during a trajectory. Secondary structure bars are shown at the top and on the left of the contact map. (**B**) Residues participating in steady contacts (in colour) were identified from (A) and analysed visually (PyMOL). Residues from the 'pseudo-rigid' segments E699–S709, P754–L764 and V731–P733 were used as an origin for the computing of contacts to all residues of KID. Only the contacts with an occurrence greater than 80% were considered. Polar, hydrophobic, and amphiphilic residues are denoted in red, blue and violet respectively. (**C**) Red–blue gradient shows the RMSF values, from large (>4 Å, in red) to small (<2 Å, in blue). The position of tyrosine residues is shown by the yellow asterisks. (A,**D**) The αH- (red) and 3_{10}-helices (blue), were assigned by DSSP on average conformation from each MD trajectory and labelled from H1 to H6. (**E**) The per-residue modelling by finite-state Markov models of the secondary structure dynamics of KID. Transition probability matrix from one folding state to another (from one letter to another) was

obtained on data encoding the secondary structure for each residue of each replica. (Left) The Fisher-Rao matrix (8 × 8 of size), where the first 4 replicas are the cleaved KID^{C} (group C), the next two the KID^{D} (group D), and the last two the cleaved restricted KID^{CR} (group CR). (Right) Multi-Dimensional Scaling (MDS) represents an 'as isometric as possible' embedding of the data in 2D and 3D (i.e., a representation by placing points on a plane and in Cartesian coordinates while preserving the calculated inter-distances as well as possible).

Analysis of the contribution of 'contacting' residues in the intramolecular H-bonds (including the salt bridges) and the hydrophobic interactions, inspected separately, showed that both types of interactions form an extended and dense network of contacts that stabilises a compact globular shape of all studied KID entities (**Figure 6.6**, **A**). To illustrate that these complex intramolecular contacts held a 'globule-like' shape of KID, we depicted the H-bonds and hydrophobic interactions on a randomly chosen conformation of each entity (**Figure 6.6**, **B**).

In particular, the helices H1 and H5 interact mainly through the multiple hydrophobic interactions stabilising their proximal position. Residues from H1 and H5 form H-bonds and hydrophobic interactions with residues from the loop connecting H3 and H4, and/or from H4. These abundant interactions stabilise a closed location of the structurally disordered regions of KID mid-domain to helices H1 and H5. H1 interacts with H2 and H3 mainly via the H-bonds formed by charged and polar residues. Finally, the N- and C-end residues interact with each other and with H1-helix. The charged (E758, E595, E699, E711 and E761) and polar (N705, N719) residues form the H-bonds by the interaction of their sidechains with the main-chain atoms of the other residues.

The observed patterns of non-covalent intramolecular contacts in KID display a key role of the αH1-helix which, like a drop of glue, attaches all the structural fragments of KID around them. The results obtained indicate that intramolecular interactions, van der Waals and electrostatic forces, are a dominant factor in the stabilisation of the compact 'globule-shaped' (collapsed) KID and that the collapse induced by the intramolecular force conquers the solvent-induced expansion.

Since the non-covalent contacts of KID are folding-dependent, we further focused on the time-connected dynamics of the KID secondary structure. To compare the folding dynamics more finely in KID entities, we carried out a study based on per-residue modelling by finite-state Markov models, which was carried out on data encoding the secondary structure for each KID residue of each replica in an estimated transition probability matrix from one folding type to another. The eight-category classification (eight-letter code) of secondary structures was used. We thus obtain for each replica a sequence of transition matrices (one per residue). A suitable distance (Fisher-Rao distance) between these families is then calculated for all the pairs of replicas. We obtain the Fisher-Rao matrix (8 × 8 of size), where the first four replicas are the cleaved KID^{C} (group C), the next two the KID^{D} (group D), and the last two the

cleaved restricted KID^{CR} (group CR) (**Figure 6.5**, **F**). Multi-Dimensional Scaling (MDS) was performed to get an 'as isometric as possible' embedding of the data in 2D (i.e., a representation by placing points on a plane while preserving the calculated inter-distances as well as possible). Interestingly, the three groups do not form separate clusters, as shown in the 2D representation. Group C occupies the space quite well with replica number 4 at the central position, which is also central for all KID trajectories, while the two points of group D occupy extreme positions. These observations corroborate well with the secondary structures (**Figure 6.7**, **C**, **E**) and the inter-distance matrix (**Figure 6.5**, **A**).

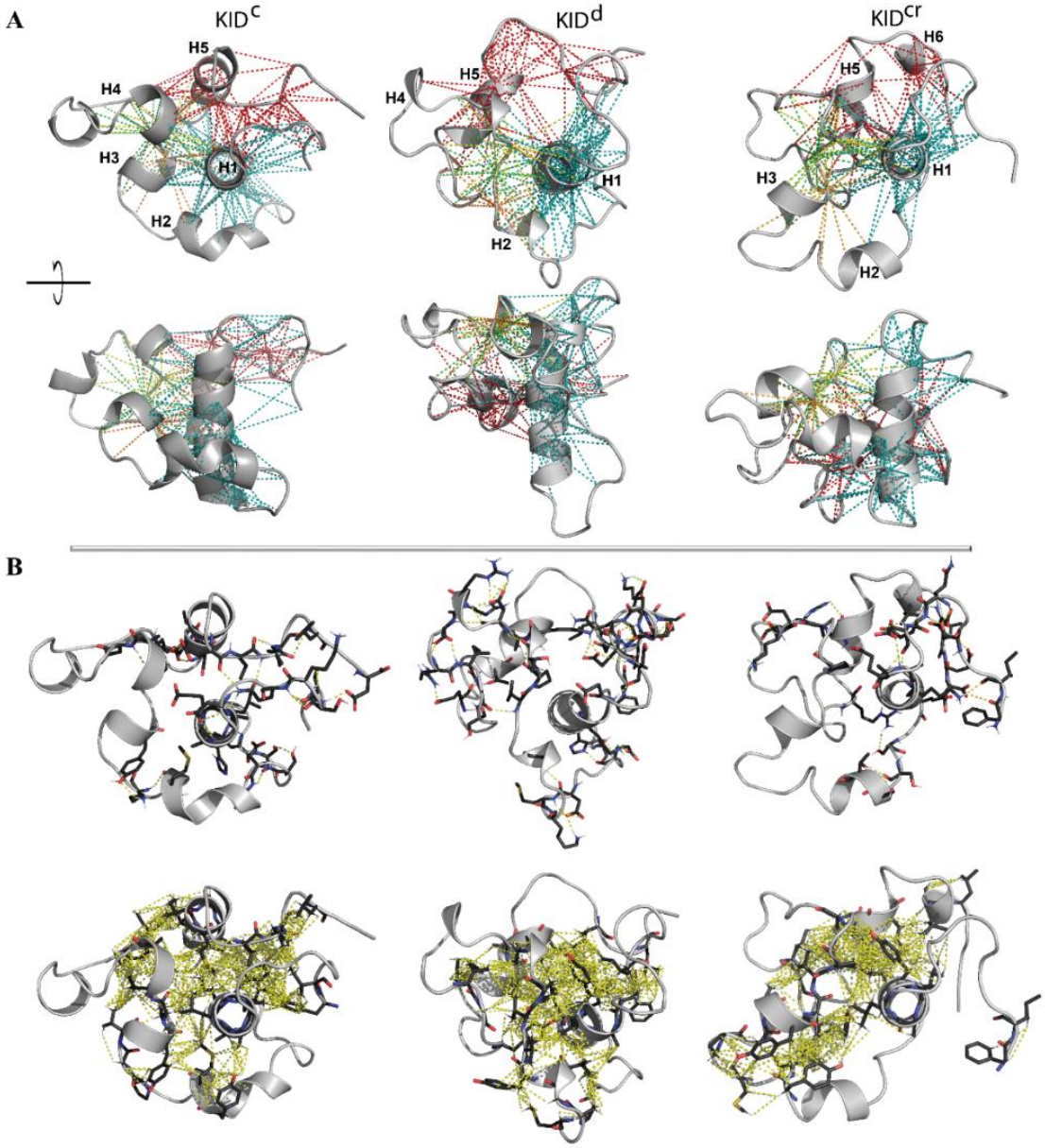


Figure 6.6 Non-covalent interactions maintaining the inherent (intrinsic) 3D structure of KID. (**A**) The intramolecular contacts, H-bonds, and hydrophobic interactions (dashed lines), formed by residues E699–709 (H1) (in teal), P754-L764 (H5) (in red) and V731/V732/P733 (orange/green/yellow) with all the residues of KID at least in one trajectory, are superimposed on

a randomly chosen KID conformation shown in two orthogonal projections. Protein is shown in grey ribbons, helices H1–H6 are labelled for KIDC. (**B**) H-bonds (D-H···A ≤ 3.6 Å, where D (D = O/N/S) is a donor atom and A (A = O/N/S) is an acceptor atom) (top) and van der Waals contacts (C-H···A = O/N/S ≤ 4 Å) between all side chain atoms (bottom) are shown on a randomly chosen conformation. (A,B) The interactions stabilising the regular structures (helices) are not considered.

6.1.2.5. GEOMETRY OF THE TYROSINE RESIDUES IN KID

As KID contains four tyrosine residues, three of which are known to be phosphorylation sites (Y703, Y721, and Y730) and one (Y747) that has an unclear functional role[348], we focused on their structural features. These residues do not belong to the helices H1–H6 identified in KID, except Y703, which is positioned on a highly conserved αH1-helix. The other phosphotyrosine residues are located on transient fragments converting between α-, 3_{10}-helix and random coil. This observation suggests that the higher solvent accessibility of the phosphorylation sites afforded by the absence of secondary structure facilitates the post-transduction (translational) modifications required for recognition and recruitment of downstream molecules that are adaptors or signalling proteins.

We suggest that the geometry of the phosphotyrosine residues, key elements for substrate binding, may reflect (recover) the structural or conformational features of KID. Geometry was described by using a tetrahedron designed on the Cα-atoms of tyrosine residues regarded as nodes connected by edges (**Figure 6.7**). In the studied KID entities, the tetrahedron geometry varies greatly during the MD trajectories. Even if the edges of the tetrahedron are maintained over an extended period (30–50 ns), their lengths are further altered, either for all edges (synchronous transform) or only for certain edges (asynchronous transform).

Changes of the tetrahedron geometry during an MD trajectory are viewed either as instantaneous events or as stepwise processes. The long-time preservation of inter-tyrosine distances is apparently associated with the maintenance of the secondary structure in KID, while the synchronous/asynchronous change reflects the important transformation at the level of the helices. Consequently, the inter-tyrosine geometry is closely related to the folding–unfolding process in KID. Nevertheless, the other structural factors, such as relative orientation of the helices and flexibility of the coiled linkers, can contribute significantly to the high variability of tyrosine residue geometry. A coherent consequence of such variability of geometry of the tyrosine residues is the great dispersion of the hydroxyl groups (**Figure 6.7**, **F**).

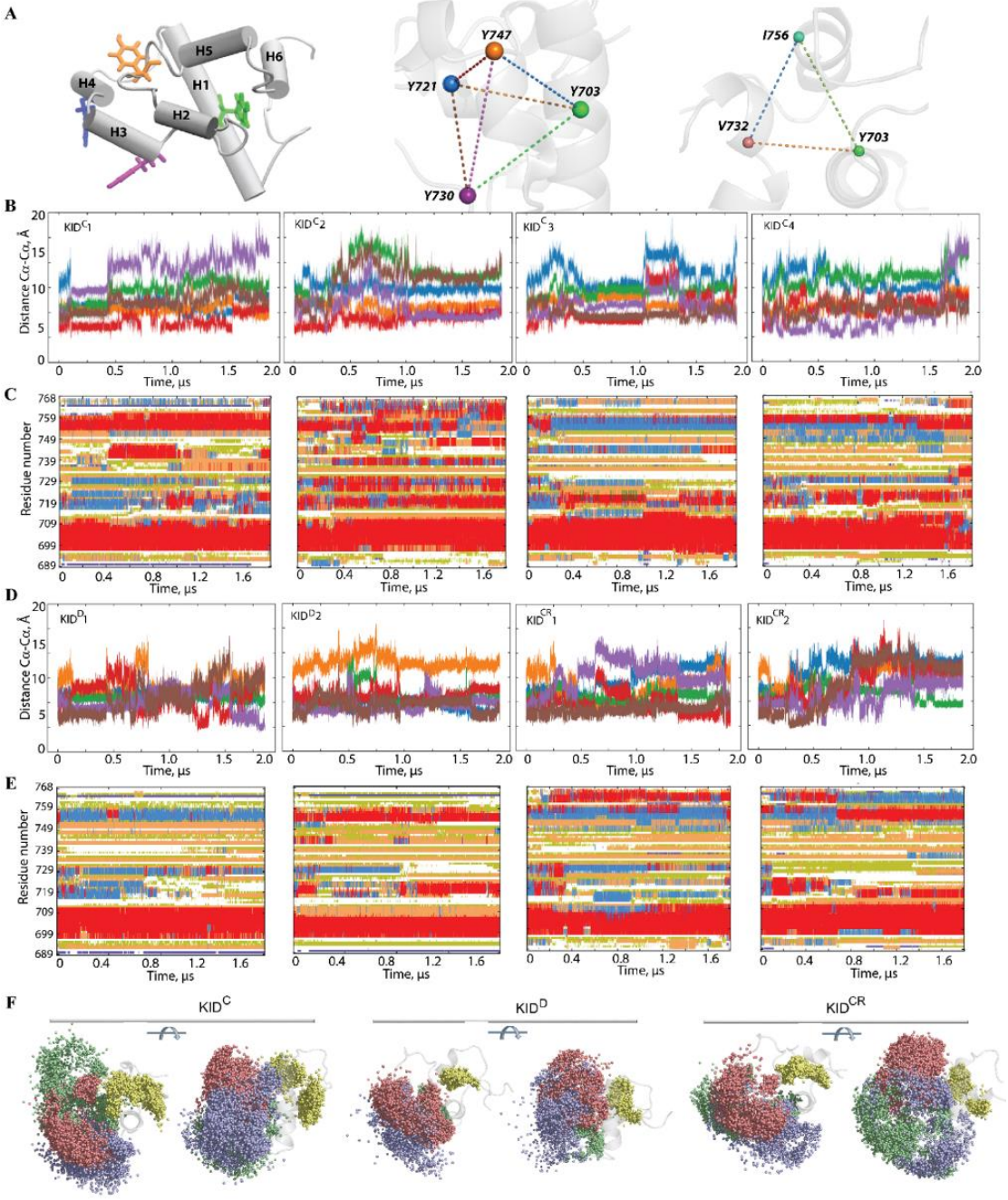


Figure 6.7 The inter-residue geometry of tyrosines in the isolated unconstrained KID and its relationship with folding. (**A**) KID structure with helices shown as solid cylinders and tyrosine residues as sticks (left), tetrahedron delimited on the Cα-atoms of tyrosine residues (middle) and triangle designed on the most stable among the MD simulation residues (with the smallest RMSF values) (right). (**B**,**D**) Distances between each pair of tyrosine residues in each MD trajectory, coloured as the edges of the tetrahedron. (**C**,**E**) The time-related evolution of the secondary structures of each residue as assigned by DSSP with the type-coded secondary structure bar. (**D**) Distances between the most 'stable' (minimal RMSF values) residues Y703, V732 and I756 over each cMD trajectory, coloured as the edges of the triangle in (A). (**F**) The spatial distribution of hydroxyl groups presented by the oxygen atoms of the tyrosine residues in KID is shown in two orthogonal projections with the oxygen atoms of Y703, Y721, Y730 and Y747, respectively coloured in yellow, blue, red and green.

Furthermore, we focus on finding geometrically conserved elements of KID and their spatial relationships with incongruent structures. We suggest that residues Y703 (αH1-helix), V732 (H5 helix) and I756 (linker connecting H4 and H5 helices), with the smallest RMSF values, are a 'rigid subset' if the inter-residue distance is conserved in MD conformations. As expected, the inter-residue distances display essentially small variations compared to the distances between the tyrosine residues (**Figure S19**). Nevertheless, the almost invariant geometry of the most 'rigid' residues over the large periods of simulation time (≥ 1 μs) is followed either by an instantaneous change or by staggering alternations of their values. There is no obvious correlation between the geometry of a 'rigid subset' formed by Y703, V732 and I756, and the geometry of a tetrahedron designed on the tyrosine residues, Y703, Y721, Y730 and Y747; nonetheless, their changes without a doubt are connected.

6.1.2.6. Free Energy Landscape as a Quantitative Measure of Folding and Disorder in KID

A promising strategy for the in-depth analysis of a protein conformational space is to consider the 'free energy landscape' along specifically chosen coordinates called 'reaction coordinates' or 'collective variables', which describe the conformation of a protein[349–351]. Such interpretation leads to quantitatively significant results that allow comparison between different states of a protein. The relative Gibbs free energy (ΔG between two or more states is a measure of the probability of finding the system in those states. Such representations of protein sampling with use of reaction coordinates can be the quintessential model system for barrier crossing events in proteins[297]. These can be estimated from incomplete sampling of the states, as long as it is an unbiased sampling.

The statistical quantities—radius of gyration, RMSD, helical folding, contacts, surface exposed to solvent—usually used for the description of the conformational properties of proteins, were regarded as reaction coordinates (collective variables) for the evaluation of the relative free energy ΔG) and reconstruction of its landscape (**Figure 6.8**).

First, the probability (P) of Rg, P(Rg), and of RMSD, P(RMSD), were used as the reaction coordinates for the evaluation of the relative free energy ΔG). The free energy landscape (FEL) as a function of RMSD and radius of gyration Rg (FEL_{Rg}^{RMSD}) for the concatenated replicas of KID is shown for each KID entity.

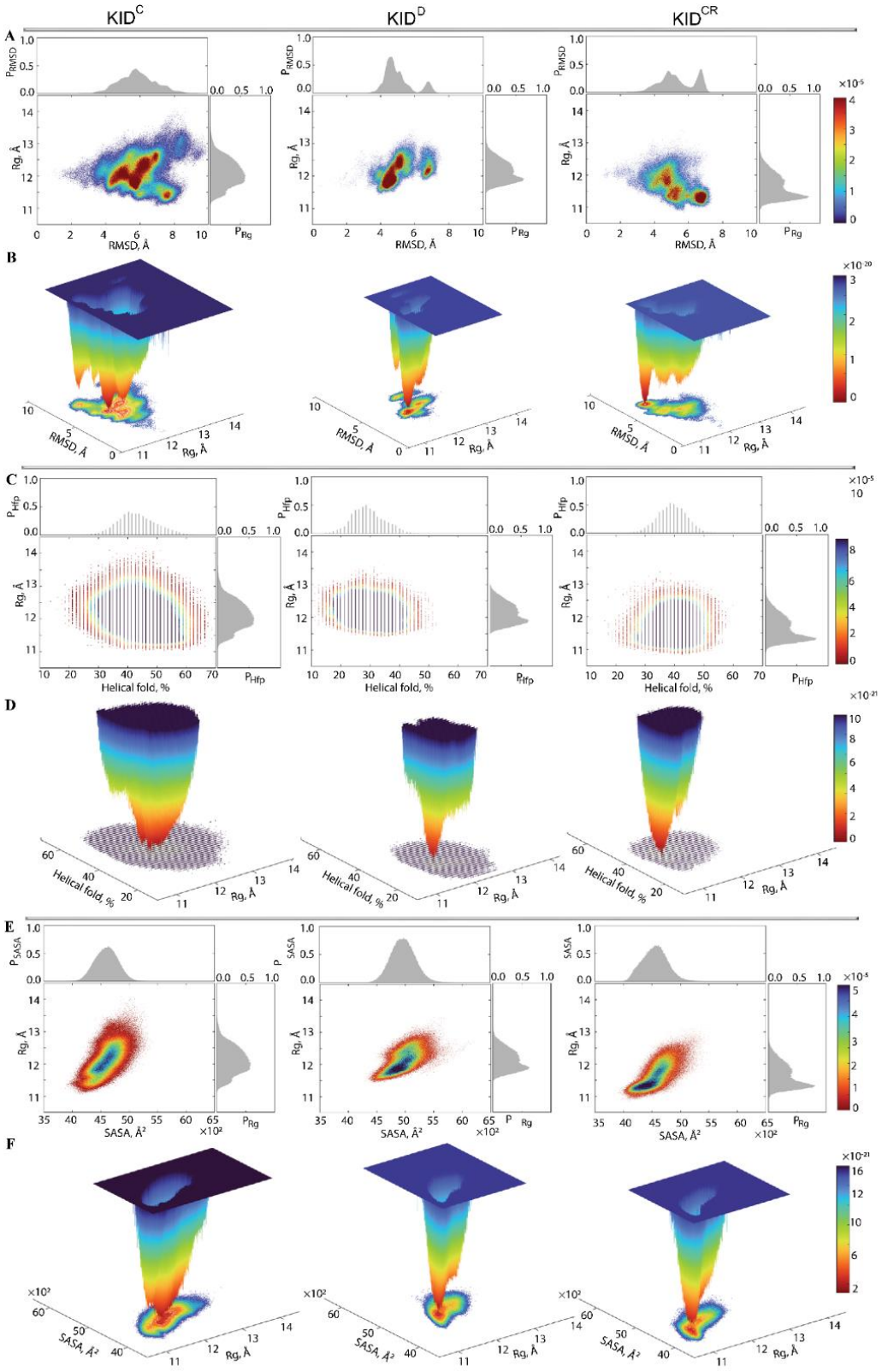


Figure 6.8 Free energy landscape (FEL) of KID as a function of the reaction coordinates. Each FEL generated on the MD conformations of different KID entities (KID^{C}, KID^{D} and KID^{CR}) is displayed for the conformational ensemble sampled on the merged replicas. Calculations of the radii of gyration are performed on the Cα-atoms. The 2-dimensional representation of FEL of the KID conformational ensembles plotted as a function of Rg (in Å) versus (A) RMSD (in Å); (C) H_{fp}, which

describes the fraction of helices (as calculated by the DSSP) (in %); (E) SASA, which describes the solvent-accessible surface area (as calculated with Cpptraj) (in Å2). Probability distribution of each reaction coordinate is shown at the top and right, respectively. The red colour represents high occurrence, yellow and green low, and blue represents the lowest occurrences. The 3-dimensional representation of the relative Gibbs free energy as a function of Rg and (B) RMSD; (D) H_{fp}; (F) SASA. The blue colour represents the high energy state, green and yellow low and blue represents the lowest metastable state. The free energy surface was plotted using Matlab.

Each FEL_{Rg}^{RMSD} shows a rugged landscape indicating a high conformational heterogeneity, which is best represented in the cleaved KID (**Figure 6.8**, **A**). The heterogeneity likely builds up the entropy barrier and adds complexity to the free energy map, and thus limits spontaneous state-to-state transition when using conventional advanced sampling methods. Nevertheless, the FEL_{Rg}^{RMSD} of each KID entity shows areas of minimum energy indicated by the red colour. The red areas represent more stability, while the reddened areas indicate transitions in the conformation of the protein followed by the thermodynamically more favourable state. The $FEL_{Rg}^{RMSD}$of different entities of KID have a unique global energy minimum (in dark red), which is clearly different from the other local minima showed by higher energy values (four minima in KID^{C} and two in KID^{D} and KID^{CR}). Regarding the FEL_{Rg}^{RMSD} ranges, it becomes the narrowest in KID^{D} and the largest in KIDC.

To characterise the KID conformations in terms of compactness, despite very small discrepancies in the values of Rg, the KID conformations were defined, based on mean values, as compact (11.2–11.6 Å), semi-compact (11.7–12.2 Å) and loose (12.3–12.8 Å). In KID^{D}, the lowest energy well is constituted of KID conformations having a semi-compact structure (Rg of 11.83(1) Å) (**Figure 6.8**, **B**, **Table S3**). Two other wells are composed of loose conformations showing very close degrees of compaction (Rg of 12.2 and 12.3 Å).

In KID^{CR}, the narrow well of the lowest energy includes the KID conformations with a compact structure (Rg of 11.33(1) Å). Two other wells contain compact and semi-compact conformations (Rg of 11.2 and 11.8 Å, respectively). The low energy well on the large-ranged FEL_{Rg}^{RMSD} map of KID^{C} is composed of compact conformations (Rg of 11.80(1) Å); the other wells are composed of conformations having different compactness, with Rg varying from 11.4 Å (compact) to 12.8 Å (loose).

Second, the relative free energy (ΔG was evaluated while using as reaction coordinates the helical folding order parameter (H_{fp}), describing the fraction of helices (as calculated by DSSP), and the radius of gyration (Rg). The unimodal Gaussian distribution of H_{fp} does not separate the KID conformations on isolated clusters but describes the range of variation of the helical content and delimits the most populated region (**Figure 6.8**, **C**). The H_{fp} and Rg values are apparently decorrelated. The free energy landscape, $FEL_{Rg}^{H_{fp}}$, reconstructed on these coordinates, shows a unique well for

each KID entity, but their profile and position is differed between the studied KID (**Figure 6.8**, **D**). This unique deep well, large in KID^{C} and narrow in KID^{D} and KID^{CR}, is composed of conformations that show a different order of helical folding that is greater in cleaved KID (KID^{C} and KID^{CR}) compared to the KID domain of KIT (KID^{D}).

Third, for evaluation of the relative free energy ΔG) of KID, the solvent-accessible surface area (SASA) and the radius of gyration (Rg) were used as reaction coordinates. Statistically, similarly to the helical content, the SASA is also represented by the unimodal Gaussian distribution, which is highly symmetric in KID^{C} and KID^{D}, and slightly asymmetric in KID^{CR}. The SASA and Rg are highly correlated metrics, showing the dependence of the solvent-accessible surface from the KID size. The free energy landscape FEL_{Rg}^{SASA}, reconstructed on these reaction coordinates, shows deep and shallow wells which are adjacent in each KID entity. In KID^{C} and KID^{D} these wells are composed of semi-compact KID conformations which are more exposed to solvent in KID^{D} and rather buried in KIDC. In KID^{CR}, the deeper well is populated by the compact and buried conformations.

The free energy landscape of KID, represented as a function of the radius of gyration Rg and of an alternate metric (RMSD, helical folding parameter, SASA) as the second reaction coordinate, groups the KID conformations in a statistically found collection of multi-minima, or super-wells. On each FEL we can distinguish the most probable states of different KID entities and compare their functionally related properties. We can also understand how these statistical properties vary in different parts of the energy landscape.

As emerges from the observations mentioned above, the representative conformations of the deepest well (the lowest energy minimum) located on each FEL do not show close similarity in their folding (secondary structure), tertiary organisation (3D structure) or the spatial position of tyrosine residues for the same KID entity (**Figure 6.9**; **Table S3**). Nevertheless, in KID^{D} these randomly chosen conformations cognate well (RMSD of 2.0–3.1 Å), while in the cleaved KID, KID^{C} and KID^{CR}, they are largely different (**Figure S20** Representative conformations of KID from the deep wells on the free energy landscape (FEL) plotted as a function of Rg *versus* RMSD. KID is displayed as a red cartoon contoured with a surface-filled model. Position of the tyrosine residues (Cα-atoms) are shown as balls coloured in yellow, blue, red and green for Y703, Y721, Y730 and Y747 respectively.**Figure S20**; **Table S3**).

The conformational sub-set of the deepest well on each FEL indicated the homogenous composition for each KID entity, as shown by similar first-principles metrics, RMSD and Rg, varying a little. It seems interesting to compare the similarity of the conformational subset from the deepest well of each KID entity with those obtained by ensemble-based clustering. The Rg mean values of the same KID entity computed on two conformational subsets are equivalent. This observation is valid for all the KID species studied. Comparing the most populated conformational subsets

generated by ensemble-based clustering with those of the deepest wells, we suggest their agreement, although very approximate. Obviously, each deepest well on FEL^{RMSD}_{Rg} is formed by the most similar KID conformations of close size then the most populated cluster contained the more heterogeneous conformations.

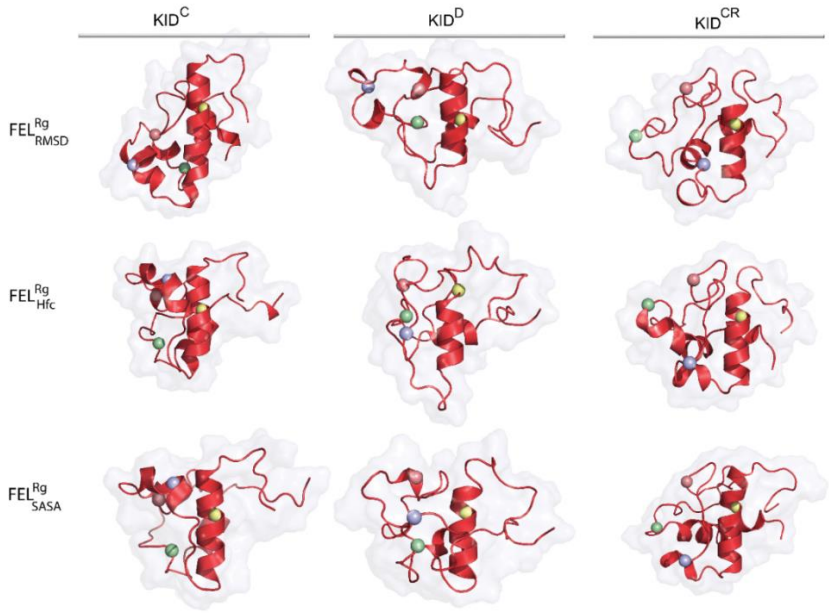


Figure 6.9 Representative conformations of the deepest well on the free energy landscape (FEL) of KID. KID displayed as red ribbons contoured with a surface-filled model. Position of the tyrosine residues (Cα-atoms) is shown as balls coloured in yellow, blue, red and green for Y703, Y721, Y730 and Y747, respectively.

6.1.3. Discussions

In the present study, we focused on the kinase insert domain of RTK KIT, the key recruitment region for a host of downstream signalling proteins—enzymes and adapter/scaffolding proteins[35]. Using the MD simulation-based study, we showed that the KIT kinase insertion domain is intrinsically disordered, and this property is disconnected from the KID context, as a cleaved polypeptide or as a domain of the receptor. Intrinsic disorder is a central feature in the function of numerous proteins, enabling them to serve as molecular switches and hubs in biological networks[308,331]. The intrinsically disordered regions (IDRs) of these proteins are increasingly recognised for their prevalence and their critical roles in regulatory intermolecular interactions. It has been estimated that IDRs in the human proteome contain ~132,000 binding motifs[352]. Disordered proteins are believed to account for a large fraction of all cellular proteins, playing roles in cell-cycle control, signal transduction, transcriptional and translational regulation, and large macromolecular complexes[353].

Principally, the RTK KIT contains three regions, JMR, KID and C-terminal, which retain the phosphotyrosine residues regulating signalling in multiple ways. The

activation loop (A-loop) of KIT holds also two tyrosine residues, Y812 and Y823, the functions of which are still under discussion[119,171]; however, it was demonstrated that the mutation of Y823 causes aberrant downstream signalling and a possible binding of proteins by the phosphorylated Y823 was suggested. Consequently, these four phosphotyrosine-containing regions of KIT are the principal regulatory subunits, maintaining the spatiotemporal control of KIT signalling.

In KIT, these regions are mainly composed of polar and charged residues, reaching up to 58% in KID and C-terminal, 46% in JMR and 43% in A-loop, while the portion of hydrophobic residues is reduced to 34% in JMR and KID, and 36% in C-terminal, while in A-loop their amount reach up to 47%. High content of polar and charged residues and a low population of hydrophobic residues is an archetypal signature of the intrinsically disordered proteins[301,302]. The lower portion of charged and polar residues in JMR and A-loop compared to KID and C-term, is reflected in the enhanced stability of their 2D and 3D structure in each receptor's state, inactive or active, as shown in crystallographic structures[121,122]. At the same time, A-loop and JMR, compared to the other KIT regions, show the largest difference in their position and structure in the inactive and active states (RMSD of 17 and 10 Å for A-loop and 5 and 2 Å for JMR), suggesting their intrinsic disorder. While KID is longer in sequence and has a higher fraction of disordered residues compared to the other phosphotyrosine-containing KIT regions, its intrinsic disorder is more abundant.

The intrinsic disorder of KID makes it difficult to study by experimental techniques[354], and the 3D numerical model is the unique description of KID from KIT[183] By taking these 3D data as the initial structure for the study of its evolution over time, we obtained an accurate and detailed atomistic physical model which presents the intrinsically disordered kinase insertion domain of KIT as a set of heterogeneous conformations. The ensembles of conformations generated for each studied KID entity were characterised by using statistical descriptors which are tightly related with either the geometrical (structural) features or the physical properties.

Our results show that accurate physical models of flexible biomolecular systems, such as proteins consisting of several intrinsically disordered regions, are well founded and can serve to pave the way to establishing the relationship between conformational flexibility and biological function.

The KID folding displays a collection of highly variable helices altered in length and type (α- or 3_{10}-helices). Except for H1, a unique helix constantly folded as an α-helix and varying only in its length, the other KID helices are transient, converting between the α- and 3_{10}-helices or a helix and non regular secondary structures—turns, bends, and coils. Apparently, the helical fold is a proper feature of KID, disconnected from KID status (as a cleaved species or a domain fused to KIT) and of a context of MD simulations (unconstrained or restrained distance F689-D768). Nevertheless, a quantity of such folding depends partially not only on the restriction applied to KID extremities,

but also on the kind of such restriction. First, a greater degree of the 2D folding was detected in the cleaved KID (KID^{C} and KID^{CR}), evidenced by the arrangement of transient helices H4 and H6, not observing in KID fused to KIT (KID^{D}). Moreover, comparing 2D structures of the cleaved KID, we observed that H2 and H3 helices are more extended in KID simulated with restrained N- and C-ends (KID^{CR}).

The conserved α-helical architecture of the H1 helix, reported for KID embedded to KIT[183], is also observed in each studied KID entity, independently from its context—KID fused to KIT or cleaved polypeptide. Such structural consistency, together with conservation of its spatial position, suggests that this helix, immediately adjacent to the kinase domain, is the inner ordered motif of KID, which is critical for KIT function. Indeed, despite the changes in the helix length along the same trajectory or in different trajectories of distinct KID entities, H1, like a drop of glue, which attaches the KID structural fragments around them by manifold contacts and stabilises a 'globule-like' shape of the intrinsically disordered KID. The 'organising role' in stabilising the KID structure was previously attributed to tyrosine Y747 located on the helix H4[183]. We suggest that the Y747 and H1-helix functions are complementary and can be mutually dependent.

The reduced portion of folded secondary structure in KID fused to KIT (KID^{D}) suggests that its folding is controlled by the kinase domain. We postulate that the KID folding is principally sequence dependent, but partly allosterically regulated by the kinase domain of KIT. Allostery within or mediated by intrinsically disordered proteins ensures robust and efficient signal integration through mechanisms that would be extremely unfavourable or even impossible for ordered globular protein upon its interaction with the partners[43,101].

The other inherent structural feature of KID is related to flexible linkers connecting these helices, which favour many relative orientations of helices, leading to a vast set of highly heterogeneous conformations. Apparently, the conformational flexibility may be more important than the specific secondary structures before binding. The length of KID linkers is variable and strongly depends on the order of helical folding. Our study revealed that despite great flexibility, once in either a cleaved state or fused to KIT, KID acquires a compact globule-like shape, with a helical fold fraction ranging from the poor (in KID^{D}) to rich (KID^{C}). A helical structure is, in general, not stable by itself—short helices taken out from stable globular proteins are found to be unstable when isolated[355]. Additional stabilising interactions of the helical structure must, therefore, be present in globular proteins.

In KID, the bulky hydrophobic residues of the weakly fluctuating regions localised on ordered structures (H1 and H5) or random coil (linker between H3 and H4/H5), interact strongly between themselves, forming a ternary hydrophobic core which stabilises the entire tertiary structure of the helically folded KID. The stabilisation of the N- and C-termini of KID is achieved by switchable H-bonds formed by terminal

residues in various combinations in different KID entities.

The intrinsically disordered KID contains three known phosphorylation sites (residues Y703, Y721 and Y730) that control KIT signalling. Tyrosine Y703 is located on the αH1-helix, the most conserved structure compared to the rest of KID, although its length varies greatly in each KID entity. The other phosphorylation sites are located on fully transient structures. These observations raised the possibility that, when targeting signalling proteins, KID could display a rich landscape of overlooked folding/dynamic properties. We suggest that (i) transient helices are structures preconfigured to specifically localise signalling proteins, which are selective for alternative phosphotyrosine sites of KID, and may facilitate phosphorylation-mediated regulation of signalling cascade, and (ii) KID has evolved an unusual structural flexibility to stabilise long-lived folding intermediates and, thus, maximise their 'recognisability' to signalling proteins. The specificity of intermolecular interactions of KID with signalling proteins is apparently determined by sequence- and structure-based selectivity, which are the two determining factors in 'molecular recognition'. Furthermore, the transient helices provide inherent stability near phosphorylation sites in KID, consistent with the finding of short and transient helical structures near the phosphorylation sites of the (unphosphorylated) disordered region of the Src homology 2 protein domains (SH2) of adaptor or signalling molecules—Grb2, PI3K, PLCγ[339].

Several questions/issues naturally arise regarding fundamental aspects of the folding/unfolding process in KID, such as the following: (i) explicitly characterising the free energy landscape of folding, which, as we found, is coupled to the KID status and naturally to its binding properties; (ii) to choose topographic characteristics as a criterion for referencing a cleaved KID as the appropriate species for the study of post-transduction process. Furthermore, it is of significant interest to explore to what extent the funnel landscape perspective also applies to intrinsically disordered proteins.

The first principles describing the protein properties—the size (the radius of gyration, Rg), the RMSD (the measure of the average distance between the conformations), the amount of the organised secondary structure (the helical folding order parameter, H_{fp}) and the solvent accessible surface area (SASA)—were used as reaction coordinates (collective variables) for the evaluation of the relative free energy (ΔG) of KID. Similar measurements have been used to compute interesting protein physics[356–358]. Furthermore, such first-principles-based descriptions of the free energy landscape of a disordered protein could be a part of machine learning algorithms that attempt to find energy functions that will finely represent the protein landscape.

The native structure of a protein, as stated in free energy theory[359], shows the lowest free energy in the large conformational space. In this case, the free energy landscape of protein folding is funnel shaped and oriented towards a single well (basin) corresponding to a native structure. The 'trapping' conformations observed during the folding process show a shallower well depth than the overall bias towards the native

structure, which ensures that the native structure is both thermodynamically favourable and kinetically accessible. This is valid in the case of a well-ordered protein, whereas the IDP we suggest has several 'quasi-native' states. Despite some conventionality in using RMSD as a reaction coordinate, the free energy landscape defined on Rg and RMSD showed a series of well-resolved wells, which can be interpreted as 'quasi-native' states of intrinsically disordered KID. Landscapes defined on the basis of structurally or physically related measures (SASA, Rg or order of helicity), used as reaction coordinates, displayed a single deep and enlarged well.

We show that the cleaved unconstrained KID explores multiple conformational substates, which are represented by conformations ranging from compact to lose, regardless of their helical content, and showed that the solvent-accessible surface area is highly dependent on the radius of gyration. KID with ends naturally restricted in KID^{D} or mimicking a native restriction in KID^{CR} exhibit a detectable reduction in helical structure and reduced conformational variability (diminished number of substates). Most of the conformations forming the deepest well on the FELs of KID^{CR} are characterised by high compactness and reduced solvent accessible surface area, while in KID^{D}, this well is completed by semi-compact conformations that have a larger solvent-accessible surface area than in KID^{CR} and comparable with those in KIDC. In general, KID has been found to occupy numerous conformational substates in the weakly funnelled free energy landscape, which distinguishes it from the highly directed, time-dependent free energy landscape of regularly folded globular proteins[360–362].

The free energy landscapes constructed from first principles describe KID conformations through a collection of minima (so-called basins) and provide a direct 'polymetric' evaluation of conformational ensemble each KID species and a comparison between the conformational ensembles of different KID entities. This description is much more accurate than grouping conformations using ensemble-based clustering and allows direct comparison between proteins based on structural metrics or physically related parameters. By comparing the respective species, we found that the profiles of the free energy landscapes of KID^{D} and KID^{CR} are approximately identical, with only one deep well, while in the unconstrained cleaved KID the free energy landscape is highly disrupted, which leads to slightly different wells with clearly pronounced minima. This indicates that the conformational ensemble of the cleaved KID simulated using restricted N- and C-termini (KID^{CR}), better reproduces the KID from the KIT.

Therefore, for studying post-transduction events of KIT, KID^{CR} is the most suitable entity. We suggest that the N- and C-termini of the cleaved KID can be stabilised by a linker, which is a chemical or polypeptide fragment 10 Å in length. Such a cyclic polypeptide would be a bona fide generic molecule mimicking the native KID for both computational and empirical studies of KID phosphorylation and recruitment of a multitude of signal transducers to its docking sites. Nevertheless, it seems to us that

the use of the cleaved KID cannot be universal. Even for the study of individual downstream signalling pathways mediated, for example, by KIT possessing 'oncogenic driver' mutations in the tyrosine kinase domain, a full-length cytoplasmic domain would be required.

6.2. LE DOMAINE INSERT KINASE CYCLISE DU RTK KIT MIME-T-IL SON HOMOLOGUE NATIF ?

Résumé. *L'étude du KID du RTK KIT en tant que polypeptide clivé et la comparaison de ses propriétés structurales et dynamiques avec le KID inséré dans le KIT a montré que le KID clivé simulé avec une restriction de distance à ses extrémités terminales reproduit mieux les propriétés du KID natif. L'application de telles contraintes n'est valable que pour des études* in silico. *Nous avons suggéré que le module KID clivé et cyclisé serait une forme de KID plus adaptée pour l'étude des évènements post-transductionnels du KIT via le KID. Ainsi, nous avons modélisé le macrocycle du KID en pontant les extrémités terminales par quatre glycines pour imiter les contraintes imposées* in silico. *La simulation de dynamique moléculaire du KID cyclique et sa comparaison croisée avec le KID clivé simulé sous contraintes et le KID joint au KIT montrent la conservation de son désordre intrinsèque et de ses propriétés dynamiques. Ainsi, nous avons prouvé que cette forme cyclique du KID est un analogue adapté et approprié aux études empiriques de la signalisation cellulaire initiée par le KID.*

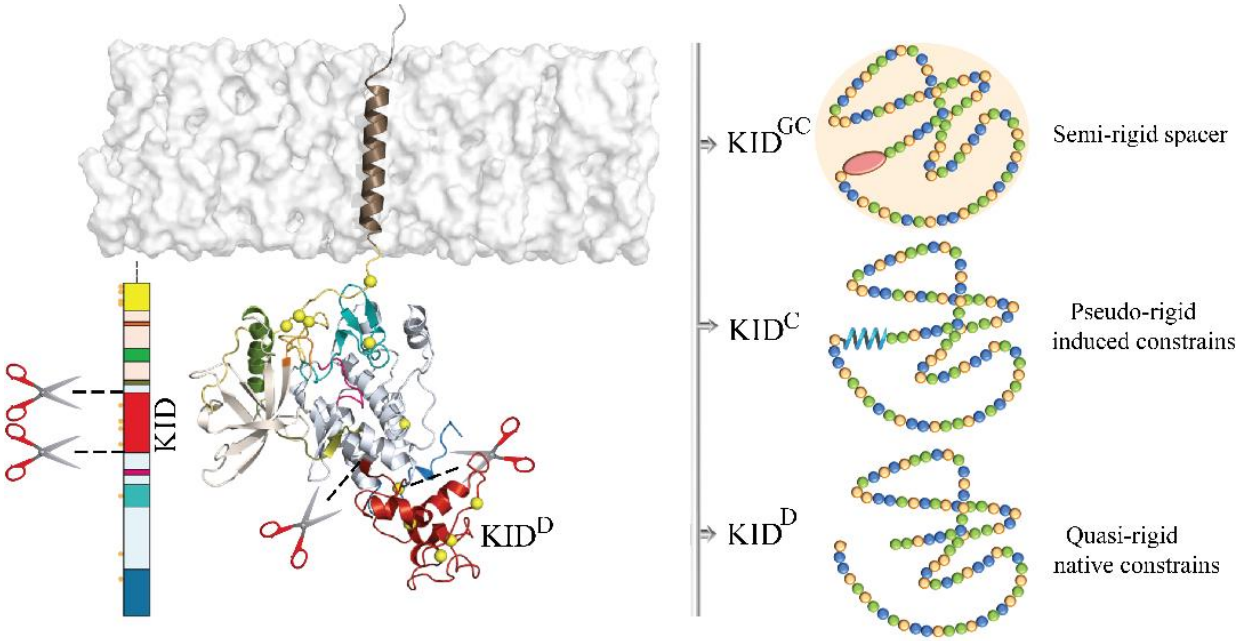


Figure 6.10 Abstract graphique de la section[363].

6.2.1. INTRODUCTION

Most proteins of the human proteome are assembled from several structural (or functional) units. These so-called protein domains form a modular architecture of a macromolecule[364,365]. It is generally accepted that protein structural domains can fold, function, select and be selected independently of the rest of the protein[366]. In many cases, protein domains in isolation can be successfully expressed and tend to fold spontaneously into their native 3D structure while retaining their overall functions. Due to the complexity of using the standard empirical and/or computational biophysical methods for studying large proteins, their splitting on isolated domains is a very attractive strategy. This approach provides a promising route for using cleaved domains as independent units/blocks in research and biotechnology of large multidomain proteins (MDPs), particularly for studying protein–protein interactions[12]. In these cases, general rules are then derived by looking at the individual domains in detail, assuming that large proteins behave like beads on a string, i.e., the function of a large protein can be understood by summarising the functions of its domains.

In many cases, modular proteins have a hybrid composition. They include structurally well-organized domains connected by intrinsically disordered regions (IDRs), which appear to be crucial elements in the cooperative folding of MDPs and the modulation of protein functions[302,352]. As the inherent flexibility of IDRs delivers dynamic cross-domain communication between remote domains, enabling cooperative allosteric regulation in an MDP, the use of cleaved IDRs as stand-alone entities is still questioned[367,368].

Intrinsically disordered proteins (IDPs), including one or several IDRs, are involved in regulatory pathways and cell signalling and sample an extensive range of conformations[43,300,302,305]. Investigation of structural ensembles of IDPs is difficult for both experiment and simulation. On the other hand, if an IDP has a modular architecture in its structures, this property yields a more efficient functional activity performance. It was recognised that modules consist of groups of highly cooperative residues[365,369], which may possess certain functional independence. Usually, protein modules are interconnected through amino acids that maintain the shortest pathways between all amino acids and are, thus, crucial for signal transmission, leading to robust and efficient communication networks[7,12,367,370,371]. This modular organisation is advantageous and, as such, has been conserved in its improved version. Many kinds of structural disorganisations can lead to deteriorated processes prompting dysfunction of normal physiological functions and causing severe diseases[308,372].

Studying the physiologically or pathologically related processes, in particular post-transduction effects, is frequently limited or impossible due to the poor solubility and stability of the associated proteins. From a practical point of view, considering the large size of such proteins and the technological and methodological problems of studying

IDPs, certain specific proteins can be analysed per domain.

Receptor tyrosine kinases (RTKs) are the archetypical modular membrane proteins possessing both well-folded and disordered domains acting together in ligand-induced activation and regulation of a post-transduction process that tightly couples extracellular and cytoplasmic events. They ensure the fine-turning control signal transmission from the outside of the cell inward through the cell to the genes by signal transduction[338,345,373].

Deregulation of RTK KIT, including overexpression and gain of function mutations, has been detected in several human cancers. The mutation-induced disorder is directly linked to leukaemia[374,375], in almost all cases of systemic mastocytosis[282] and other hematopoietic cancers; gastrointestinal stromal tumour (GIST)[376], melanoma[377] and others[378].

Similar to all RTKs, KIT contains a tyrosine kinase domain (TKD) crowned by several IDRs–juxtamembrane region (JMR), kinase insert domain (KID), and C-tail[114], which are inherently coupled[271]. In turn, the TKD of KIT is also composed of two sub-domains–N- and C-lobes—enriched by an IDR called activation (A-) loop, tightly collaborating in the activation/deactivation process[379]. Therefore, using these IDRs as independent isolated units instead of their natively fused to TKD states requires careful consideration and further investigation[380,381].

Each KIT IDR contains functional phosphotyrosine residues that act as critical regulatory elements that contribute to KIT activation and/or mediating protein–protein interactions. JMR is the bi-functional segment playing a regulatory role in the activation/deactivation process and the recruitment of signalling proteins. At the same time, KID only participates in the selective recognition and binding of adaptors, signalling and scaffolding proteins[114,345,382]. Multiple functional phosphorylation sites of KID from KIT, three tyrosine (Y703, Y721, Y730), and two serine (S741 and S746) provide alternative binding sites for the intracellular proteins[287]. Phosphorylation of Y703 supplies the binding site for the SH2 domain of Grb2, an adaptor protein initiating the Ras/MAP kinase signalling pathway. Phosphorylated Y721 and Y730 are PI3K and phospholipase C (PLCγ) recognition sites, respectively. The function of Y747 has not yet been described. Phosphorylated serine residues, S741 and S746, bind PKC (protein kinase C) and contribute to the negative feedback of PKC activity under receptor stimulation.

The phosphorylation and binding of KIT domains having multiple functional phosphorylation sites is a great challenge[65,66]. Given its many phosphotyrosines, nothing is known about how such processes occur. Is single-site tyrosine phosphorylation sufficient for a protein to bind (a one-to-one process) to such a domain? Or is protein binding a more collective event, described as multithreaded processes—one to many, or many to one, or many to many—in which, for example,

protein binding induces conditions for the phosphorylation of another tyrosine followed by binding of another protein, or partner binding requires phosphorylation of two or more tyrosine sites at the target? To answer these questions, it is necessary to consider many cases of phosphorylation/binding events described by a factorial function. Only for KID with three functional tyrosine residues, the number of combinations analysed is seven. However, if we take into account that in RTK KIT, eight tyrosine phosphorylation sites have been identified *in vivo* (Y568 and Y570 in JMR; Y703, Y721, and Y730 in KID; Y823 in A-loop; Y900 in the C-lobe; and Y936 in C-tail)[155] as well as two additional sites having been detected *in vitro* in the activated kinase domain (Y547 and Y553 in JMR)[119], the number of combinations is drastically increased. The modularity of KIT yields a more efficient performance of the functional activity study.

To begin such exploration, even by *in silico* methods, a single modular domain should be carefully determined and optimised before studying phosphorylation effects. Our recent *in silico* study (3D *de novo* modelling and molecular dynamics (MD) simulations) suggested that the cleaved KID (isolated protein) better reproduces the natively fused KID if simulated with locally N- and C-ends to mimic the native steric condition[290]. To deliver the usable species as an initial template for the empirical studies, a generic cyclic KID of RTK KIT, composed of the 80-amino-acid cleaved polypeptide (F689–D768) cyclised by insertion of four Gly residues acting as a physical connector or spacer between its N- and C-KID termini, was proposed as an entity that would be best suited for future studies on the KIT post-transduction effects involving KID. We suggested that this generic cyclic KID (KID^{GC}) is also an intrinsically disordered protein (IDP). As the characterisation of IDPs is not a trivial problem, in this paper, we report the structural description of the KID and resort to the recapitulation of the available KIT KID data obtained for the different KID species to compare them in terms of Gibbs free energy.

The characterisation in terms of structural and biophysically related metrics of the conformational spaces generated by the large-scale MD simulations of KID which was considered to be (I) a generic macrocycle (KID^{GC}), (II) a cleaved isolated polypeptide (KID^{C}), and (III) a natively TKD-fused domain (KID^{D})[183,271] **Figure 6.11**), inspired us to examine the crucial question: how does a KID evolve when studied in isolation compared to more complex architectures? While the allostery of a multidomain protein and the role of quaternary structure in modulating affinity is well established in many proteins[96], we asked whether the folding of the KID^{GC} and its binding sites are in correct position for its functioning.

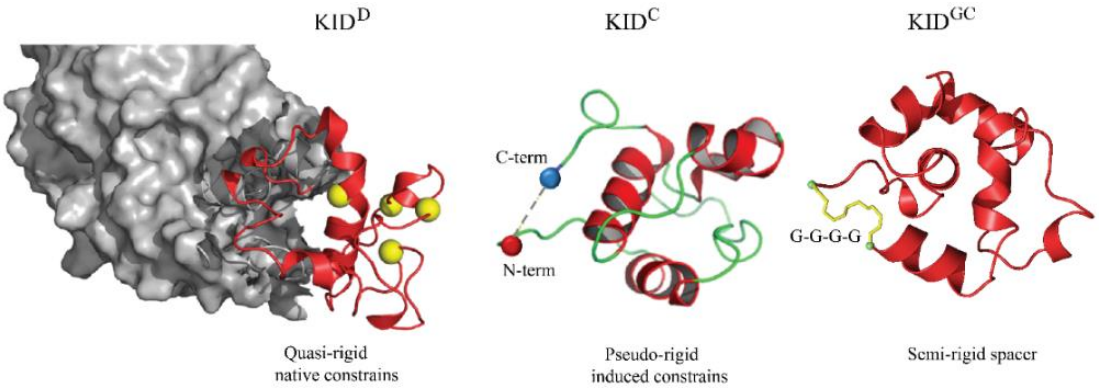


Figure 6.11 Kinase insert domain (KID) of RTK KIT as a domain fused to TKD (KID^D), a cleaved isolated polypeptide (KID^C) and a generic cyclic entity (KID^{GC}). 3D structures are taken as randomly chosen conformations from the MD simulation of each species, with KID shown as a red cartoon, TKD as the grey surface, and the motif representing GGGG being yellow sticks. TKD is shown partially with a focus on the region adjacent to KID.

6.2.2. Results

A 3D model of the KID generic macrocycle (KID^{GC}) was obtained from the randomly chosen MD conformation of the KID polypeptide from a restrained isolated KID. Composed of 80 amino acids (F689–D768) by integration of a short spacer constituted of four glycine residues (GGGG motif). An optimised and well-equilibrated model of KID^{GC} was studied by extended (2-µs) MD simulation running twice in strictly identical conditions. As a newly produced species, KID^{GC} was first characterised in terms of conventional descriptors using generated MD replicas and further compared with native KID's fused to TKD of KIT (KID^D) and cleaved KID simulated as an isolated polypeptide with easy restrained end-to-end distance (induced pseudo-rigid constraints) (KID^C).

6.2.2.1. General Characterisation of KID^{GC}

The root-mean-square deviations (RMSDs) calculated on KID^{GC} conformations from two replicas in the same initial structure show good convergence (**Figure 6.12**, **A**). Compared with the great and fast variations of RMSDs in KID^D and KID^C, associated with the significant conformational transitions[290], the RMSD curves of KID^{GC} are significantly smoothers and vary within 2–3 Å.

Similar to the other KIDs, according to DSSP[262], KID^{GC} shows an essential portion of the helical fold (45–51%), which is composed of 5–6 transient helices frequently transforming into the other structural motifs (α-helix↔3_{10}-helix↔turn/bend) (**Figure 6.12**, **C**, **D**). A clear predominance of α-helix, two times more frequent than 3_{10}-helix, is evident. Nevertheless, the overall occurrence of each helix computed on the

concatenated trajectories is dissimilar: 80% for H1 and H6 and only 50–70% for other helices—H2–H5 (**Figure 6.12**, **E**). Comparing the KIDs' helicity, we noted that the GGGG spacer significantly increased the helical content in KID^{GC} concerning KID^{D} and KID^{C} having the portion of the helically folded residues of 25–30 and 30–35%, respectively[290].

The elastic GGGG motif retains the dynamical ability of the N- and C-terminals residues of KID^{GC} characterised by an inter-distance from 5 to 15 Å, likely in $KID^{D}KID^{D}$ and KID^{C} [290] and mean value (m.v.) of 10 Å, corresponding precisely to the value observed in all KIT crystallographic structures[116,183]. The root-mean-square fluctuation (RMSFs) curves subdivide the KID^{GC} residues into two groups, characterised by high and small RMDF values. Additionally, the groups of residues showed the extreme amplitude of fluctuations, either the highest or lowest, which are conserved in different KID. We found early on that the weakly fluctuating residues are involved in the multiple non-covalent interactions stabilising the globule-like shape of KID[290]. Here, we focus on the KID's highly fluctuating residues. We suggest that these residues may be the main factors that influenced the conformational diversity of KID. In addition to the highly fluctuating N- and C-terminal residues interconnected in KID^{GC} by the elastic spacer GGGG, three other segments, C714-M722 (1), R739-V742 (2) and E758 (3), systematically show the enlarged RMSF values during MD simulations (**Figure 6.12**, **E**). These residues are either from the unregular (random coil) or partially folded transient structures.

To estimate the residues fluctuations concerning a stable αH1-helix, taken as reference structure, and visualise the variance, the position of each mid-point residue (Cα-atom) of the maximally fluctuating KID^{GC} residue was aligned on αH1-helix (A701-N705) and projected into KID 3D structure. First, the maximally fluctuating residues are nearly equidistant from the KID^{GC} H1-helix (RMSF value of ≈ 6.5 Å), and their spatial position is described as the elongated surface distributions with an apparent shape of the oblate spherical sector comparable for all maximal fluctuation residues both in length and area occupied differed in spatial position.

Most instances of C691 are distributed along the y–z plane. R740, E758, and E767 are distributed mainly on the x–y plane (**Figure 6.12**, **F**). As was expected, distributions formed by N- and C-terminals (C691 and E767) linked by the GGGG motif are closely positioned. Distributions of the highly fluctuating residues from segments 1, 2, and 3 are mutually orthogonal and, together with the N- and C-terminal arrays, represent a spiral galaxy form, as viewed at the top.

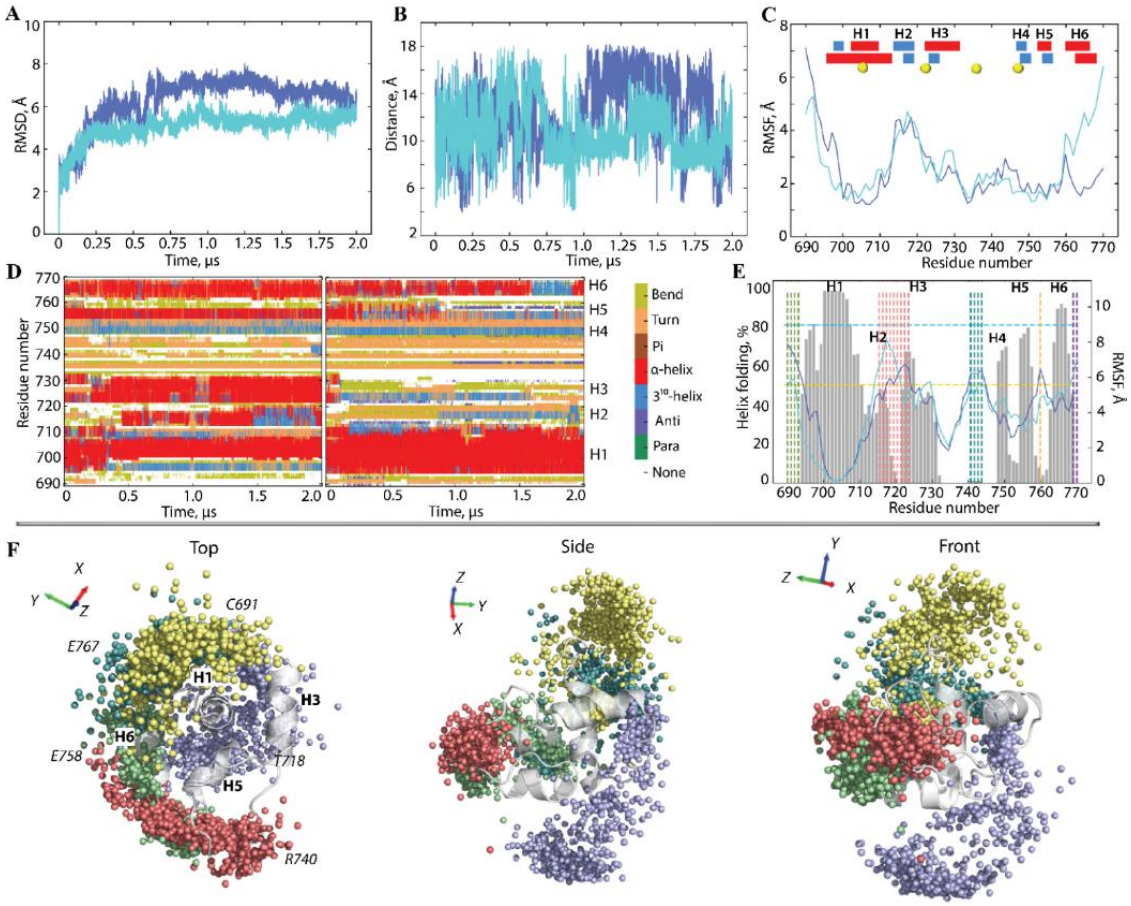


Figure 6.12 Conventional MD simulations of KID^{GC}. (**A**) RMSDs, computed on the Cα atoms after fitting on initial conformation (at t = 0 ns), (**B**) distances between the Cα-atoms from F689 and D768 residues, and (**C**) RMSFs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns). In (C), the insert shows the interpretation of the folded secondary structures, αH- (red) and 3_{10}-helices (blue), labelled as H1-H6, assigned for a mean conformation of each MD trajectory. Yellow balls show the tyrosine residue position. In (A–C), MD replicas 1–2 are distinguished by colour, light and dark blue. (**D**) The secondary structure time-related evolution of each KID residue as assigned by DSSP with type-coded secondary structure bar. (**E**) Superimposition of the helical structure content (in % of the total simulation time) calculated for each residue of KID^{GC} of the concatenated trajectories and showed as grey histograms (left axis) into the RMSFs calculated after the alignment on H1-helix (A700-L706) (right axis). Coloured dashed lines trace groups of residues with higher RMSF values (>6 Å): F689-S691 in green, C714-M722 in red, R739-V742 in blue, E758 in orange, and E767-D768 in violet. (**F**) The spatial position of the highly fluctuating residues (the Cα-atoms of median residue) is distinguished by colour (C691 in yellow, T718 (1) in violet, R740 (2) in red, E758 (3) in green and E767 (in teal) projected on KID^{GC} structure after alignment on H1-helix (A701-N705). Three views—top, side and front—concerning the αH1-helix axis are shown.

The tyrosines—key KID residues—show highly variable spatial positions. Distances connecting the apexes of a tetrahedron designed on the Cα-atoms of four tyrosine residues display very fluctuating values (**Figure 6.13**, **A**, **B**). Distances Y730-Y747, Y703-Y730, Y721-Y730, and Y703-Y721 represent the asymmetric bimodal skew-normal distributions of quasi-equivalent probability, with a minimal contribution (< 0.1) of the second components.

The main features of these distribution differed only in their maxima values position—at 8, 11, 14 and 16 Å, respectively. Two other tetrahedron distances, Y721-Y747 and Y703-Y747, are described as multimodal distributions—bimodal (Y721-Y747) with the maxima at 8 and 15 Å, and three-modal (Y703-Y747) with the utmost at 11, 13, and 17 Å. The tyrosine residues (Cα-atoms) projected into KID 3D structure after the alignment on Y703 at αH1 helix display different spatial distributions—the compact for Y730 (red), more enlarged for Y747 (green), and broad and subdivided on the separated clusters for Y721 (lilac) (**Figure 6.13**, **C**). Similar to the highly fluctuating residues, the tyrosine residues are distributed mainly on a semi-sphere around αH1-helix with Y721 and Y747 locations mostly on the y–z and x–y planes, respectively.

The tyrosine residues (Cα-atoms) projected into KID 3D structure after the alignment on Y703 at αH1 helix display different spatial distributions—the compact for Y730 (red), more enlarged for Y747 (green), and broad and subdivided on the separated clusters for Y721 (lilac) (**Figure 6.13**, **C**). Similar to the highly fluctuating residues, the tyrosine residues are distributed mainly on a semi-sphere around αH1-helix with Y721 and Y747 locations mostly on the y–z and x–y planes, respectively.

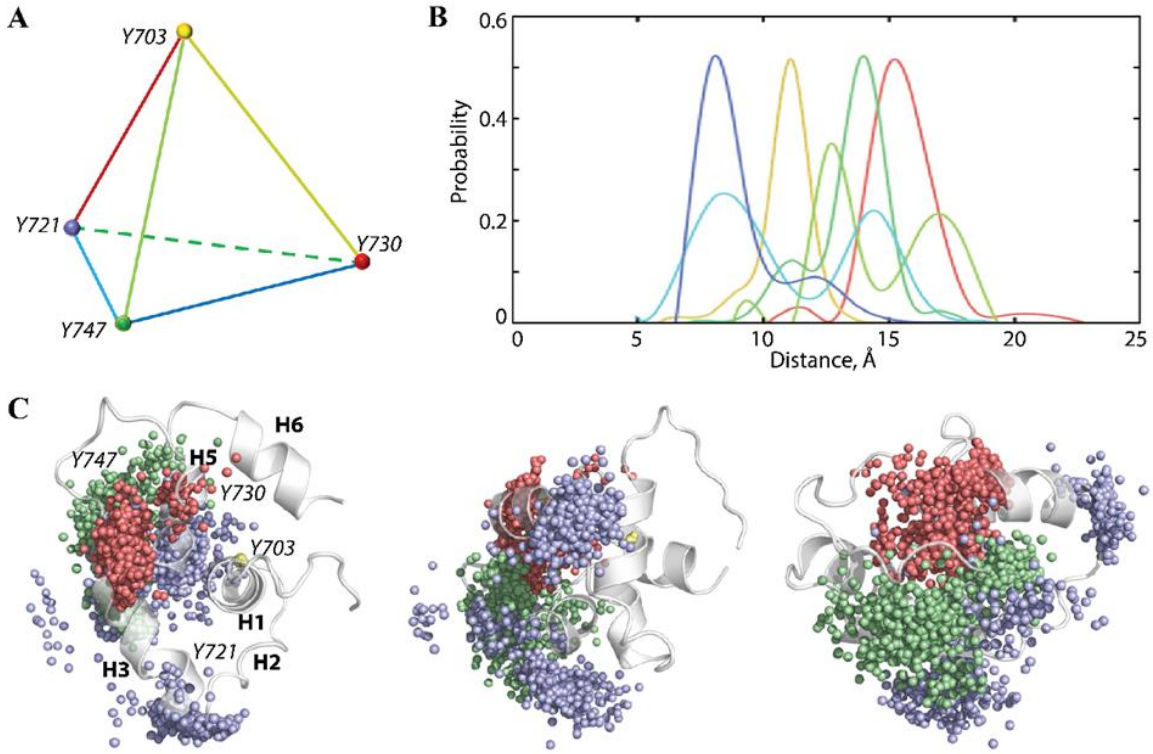


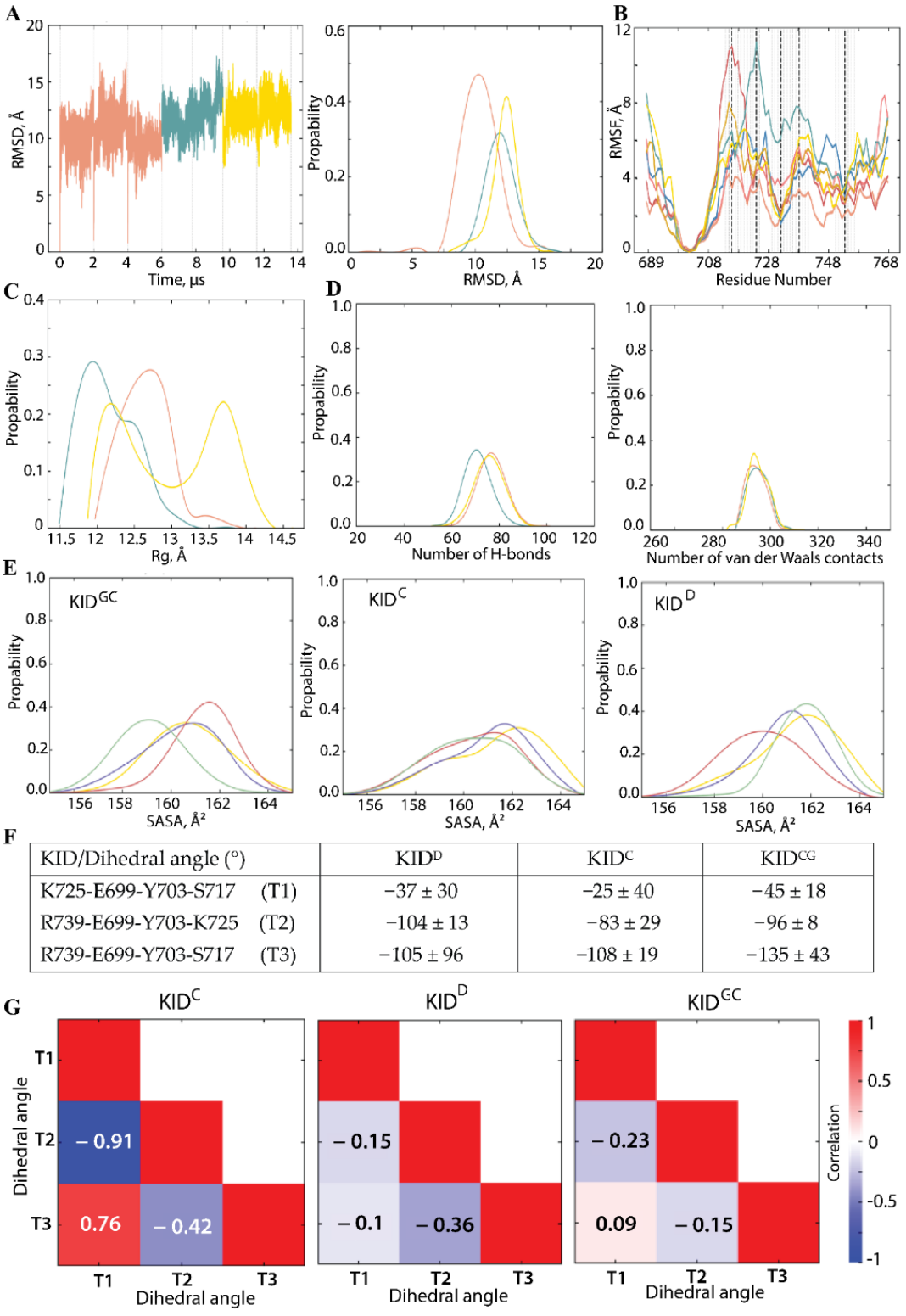
Figure 6.13 The cumulative spatial position of the tyrosine residues Y721, Y730, and Y747 relative to Y703 from the conserved αH1-helix. (**A**) The tetrahedron is defined on the Cα-atoms (vertex) of each tyrosine. (**B**) Probability of distances between each pair of tyrosine residues (tetrahedron edges). Curves are coloured as tetrahedron edges (A). (**C**) Position of the Cα-atoms of each tyrosine residue–Y703 (yellow), Y721 (lilac), Y730 (red), and Y747 (green), projected into the KID^{GC} 3D structure (grey cartoon) after alignment on αH1-helix (A701-N705) and shown at three orientations: top (left), side (middle), and front (right), with respect to the αH1 axe. Each point (MD frame) took 5 ns of the 14-μs MD concatenated trajectory.

6.2.2.2. Comparative Analysis of KID^{GC}, KID^{D}, and KID^{C}

To designate or not the KID^{GC} as a species having the qualities comparable to those of the native KID fused to the tyrosine kinase domain of KIT (KID^{D}) or the cleaved KID (KID^{C}), and therefore to postulate the appropriate model for the study of posttranslational processes of RTK KIT, we compared the conformational and structural proprieties of these three entities. As the MD trajectories of all KID species were generated upon strictly identical conditions that differed only in the mode of preserving its end-to-end distance, we postulated that these data might be analysed together as a unique concatenated trajectory (dataset) describing the same object—the intrinsically disordered polypeptide. Before the data analysis, all data were normalised by a fitting on the most stable structural element of KID—αH1-helix (A701-N705) taken from the KID^{D} conformation at t = 0 ns and further analysed either as all datasets or using their selected components—a unique replica or the merged replicas for a given entity (**Figure 6.14**).

The normalised RMSDs of each analysed KID show a frequent alteration (increase/decrease) in value which are comparable between the species so that the concatenated trajectory can be viewed as a continuous 14 μs trajectory of IDP KID. The RMSD probability curves for distinct KID represent a Gaussian distribution that is partially superimposed, showing very close mean values, from 10 to 12 Å. The RMSF curves show that the minimally and maximally fluctuating residues are the same in all studied KID or at least the nearest ones. Indeed, V732 and P754 systematically display minimal RMSF values, while S717, K725, and R739 exhibit the highest.

We previously found that KID^{D} and KID^{C} display a compact globule-like shape, stabilised by a dense network of non-covalent contacts[290]. Some residues, mainly having minimal fluctuations, are more likely to be close to each other than others. Still, it is in good agreement with a broad conformational ensemble without apparent specificity between KID^{D} and KID^{C}. The radius of gyration (Rg), characterising a protein shape, shows a slightly asymmetric normal distribution for KID^{D} and KID^{C} with maxima at 12.7 and 12.0 Å, respectively. In contrast, Rg of KID^{GC} shows a multimodal distribution with two equally weighted means of the bell-shaped normal distributions with maxima at 12.2 and 13.7 Å. The total number of H-bonds stabilising each KID entity (including the intra-helix contacts) is strictly identical in KID^{GC} (m.v. of 76) and KID^{D} (m.v. of 76.7). In contrast, their number is slightly reduced (3%) in KID^{C}. It should be noted that the total number of van der Waals interactions is precisely the same in all the KIDs studied and that they are almost four times more numerous than hydrogen bonds.



KID/Dihedral angle (°)		KID^{D}	KID^{C}	KID^{CG}
K725-E699-Y703-S717	(T1)	−37 ± 30	−25 ± 40	−45 ± 18
R739-E699-Y703-K725	(T2)	−104 ± 13	−83 ± 29	−96 ± 8
R739-E699-Y703-S717	(T3)	−105 ± 96	−108 ± 19	−135 ± 43

Figure 6.14 Comparative characterisation of KID using the conventional molecular dynamics (cMD) simulations of different KID entities. (**A**) RMSD was computed on the Cα-atoms (left) as well as its probability (right). (**B**) RMSF was calculated on the Cα atoms of KID^{GC} (yellow, sand), KID^{D} (red, pink, coral) and KID^{C} (teal, blue). Vertical dashed lines mark the minimally (V732 and P754) and maximally (S717, K725, and R739) fluctuating residues. (**C**) The radius of gyration (Rg) of KID entities. (**D**) Non-covalent interactions, hydrogen (H-) bonds (left) and van der Waals contacts (right), and stabilising KID entities were computed for the MD frames taken each 100 ns. Contacts with donor–acceptor

(D-A) distance between heavy atoms (D and A = N, O, S) ≤ 3.6 Å, and angle at H atom (DHA) ≥ 120° were interpreted as H-bonds; distances between C-atoms ≤ 3.6 Å were attributed to van der Waals contacts. Only contacts with occurrence ≥40% were taken into consideration. (**E**) The colour denotes the solvent-accessible surface areas (SASA) of each tyrosine residue—Y703 in yellow, Y721 in violet, Y730 in red, and Y747 in green. (**F**) The dihedral (pseudo-torsion) angle is defined for the most fluctuating residues in each KID entity relative to the αH1-helix and (**G**) Correlations between the dihedral angles. (A–G) Analysis was performed on KID^{GC} (yellow), KID^{D} (red), and KID^{C} (teal) per trajectory, per entity (merged replicas), and using the concatenated data for all entities. All calculations were performed after fitting all conformations on the most stable structural element of KID—H1 (A701-N705) from the KID^{D} conformation taken at t = 0 ns.

The other metrics characterising geometry, the highly fluctuating residues—distance, pseudo-valent and pseudo-torsion (dihedral) angles—showed a substantial variance in their values (**Table S4**). The values of pseudo-torsion angles defined the angle between two hyperplanes formed by the highly fluctuating residues—K725-E699-Y703-S717 (T1, synclinal), R739-E699-Y703-K725 (T2, anticlinal), and R739-E699-Y703-S717 (T3, anticlinal), are particularly interesting. The value range of the same dihedral angle is compatible for each KID entity (**Figure S21**).

6.2.2.3. Intrinsic geometry of tyrosine residues in KID^{GC}, KID^{D} and KID^{C}

Focusing on the KID key residues—tyrosines—we analysed and compared the metrics related to their properties. The solvent-accessible surface area (SASA) of tyrosine residues is comparable in all studied species and between the functional phosphotyrosines—Y703, Y721, and Y730—that control KIT signalling and Y747 with the non identified empirically function[348].

Based on our previous *in silico* calculations, we have assigned to Y747, located on the H4 helix, the "organising role" in the assembly of KID structure at the tertiary and quaternary level and suggested that the Y747 and αH1-helix functions are complementary and can be mutually dependent[183]. As a single Y703 is localised in the stable αH1-helix, i.e., the most conserved structural element of KID varying only in length, the other phosphotyrosines are localised on the fully transient structures, so we have supplemented for each tyrosine residue the Ramachandran plots providing an additional view on the secondary structure in each KID and their tyrosines backbone configuration (**Figure 6.15**).

The Ramachandran plot shows the statistical distribution of the combinations of the backbone dihedral angles φ and ψ and visualises energetically allowed and forbidden regions for the dihedral angles[235]. Typically, the permitted areas and folding of the secondary structure are residue dependent. For all non glycine and non proline residues of KID, the α-helices are found at m.v. of −64 ± 2 (ψ) and −41 ± 2◦ (φ), while the 3_{10} helices are in the upper part of the α-helices region, at −60 (ψ) and −25◦ (φ)[383].

For unphosphorylated tyrosines, the parallel and antiparallel β-sheets are localised in ranges of −119 ± 17 to 131 ± 16◦ (ψ) and −126 ± 18 to 142 ± 16◦ (φ), respectively. Left-handed helices are found at 60◦ (ψ) and 50◦ (φ).

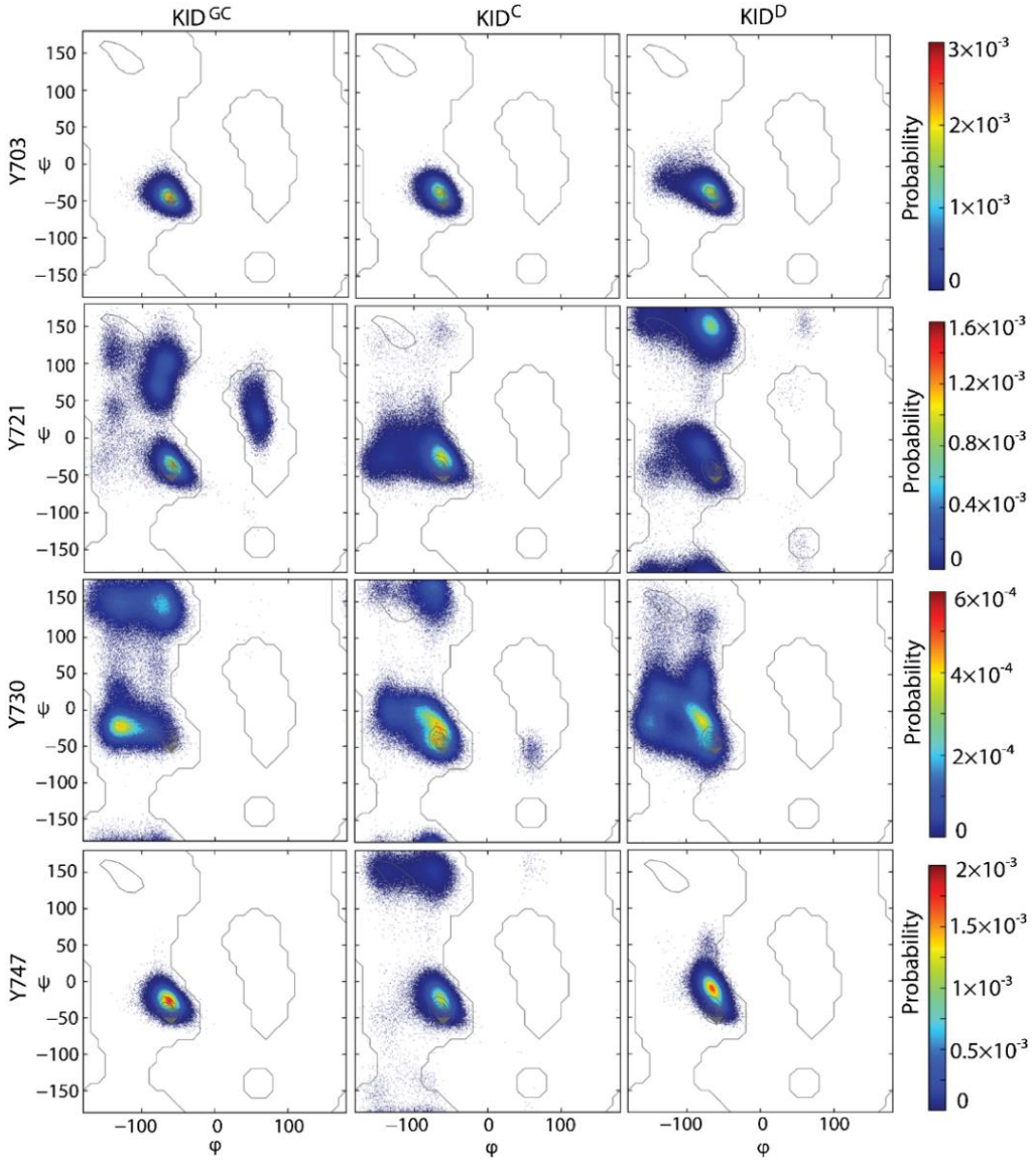


Figure 6.15 Ramachandran plots of the tyrosine residues in each KID entity. The colour scale shows the population density. Contours represent the unphosphorylated tyrosine backbone-dependent authorized regions[384]. The red colour represents high occurrence, yellow and green represent low, and blue represents the lowest occurrence.

In general, this analysis supports the secondary structure interpretation in KID assessed in the previous works by the DSSP, which was described as a helical fold composed of the α- and 3_{10}-helices[183,271,290]. Identifying helicity is in good agreement between the two methods, DSSP and Ramachandran plot. Nevertheless, the last method signalises the β-strand structures, which were not sampled by the DSSP program. Both methods confirm that the KID of RTK KIT is an archetypical intrinsically

disordered entity, regardless of the context studied, either as a domain of RTK KIT or as a cleaved isolated protein, and this inherent property is manifested primarily at the secondary structure level. Each sequence segment is folded as a partially unstable (transient) structure or represents an irregular coil. Curiously, all conformational ensembles generated from different KID entities evidence that KID polypeptide tends mostly to a disordered state with a great propensity to exhibit structured transient regions.

A unique but widely diffused distribution of Y721 in KID^{C} characterises its organisation into α- and 3_{10} helices. In contrast, a unique diffused distribution of Y730 in KID^{D} and KID^{GC} corresponds to an unfolded coil. Several well-resolved maxima characterise the Ramachandran plots of Y721 in the areas corresponding to a coil transiting to helix and β-strand (KID^{D}), the coil transiting to helix (KID^{C}), and the coil transiting to α-helix and left-handed helix (KID^{GC}). Interestingly, the β-strand area in KID^{GC} is presented by at least three distinct clusters that correspond to different types of secondary structures—parallel and antiparallel β-sheets and type II turn. The Y747 residue, considered in the literature as a rather nonfunctional tyrosine (non phosphotyrosine), showed a single sharp distribution in KID^{GC} and KID^{D} corresponding to 3_{10}-helix. In contrast, in KID^{C} an additional diffuse distribution around β-sheets is observed.

The Ramachandran plot of KID tyrosine residues showed the distributions of all accessible φ and ψ values. Still, the character of these distributions is very different for tyrosines within the same KID entity and between the same tyrosine in other KID. In all KID entities, only Y703 forms a single dense maximum corresponding to the α-helical structure with a small contribution of 3_{10}-helices. The unique and thick maximum observed for Y747 in KID^{GC} and KID^{D} corresponds to the area of the 3_{10}-helix rather than the α-helix.

All KID tyrosine residues, physiologically or structurally pivotal, are located in sequence regions characterised by different degrees of disorder. Only tyrosine Y703 is localised in the structurally conserved αH1-helix, varying in length, in all KID entities; Y747 is positioned on a sequence segment that is folded as regular 3_{10}-helix (H4) in KID^{D} and KID^{GC} or as a partially transient structure (α-helix ↔ 3_{10}-helix) in KID^{C}. Two other tyrosine residues, Y721 and Y730, are located on a fully transient backbone (α-helix ↔ 3_{10}-helix ↔ β-strand ↔ β-turn) in all studied KID.

The tyrosine residues are exposed to the solvent by their side chain and involved either in backbone–backbone H-bond interactions or entirely unlinked non-covalently from their environment (**Figure 6.16**).

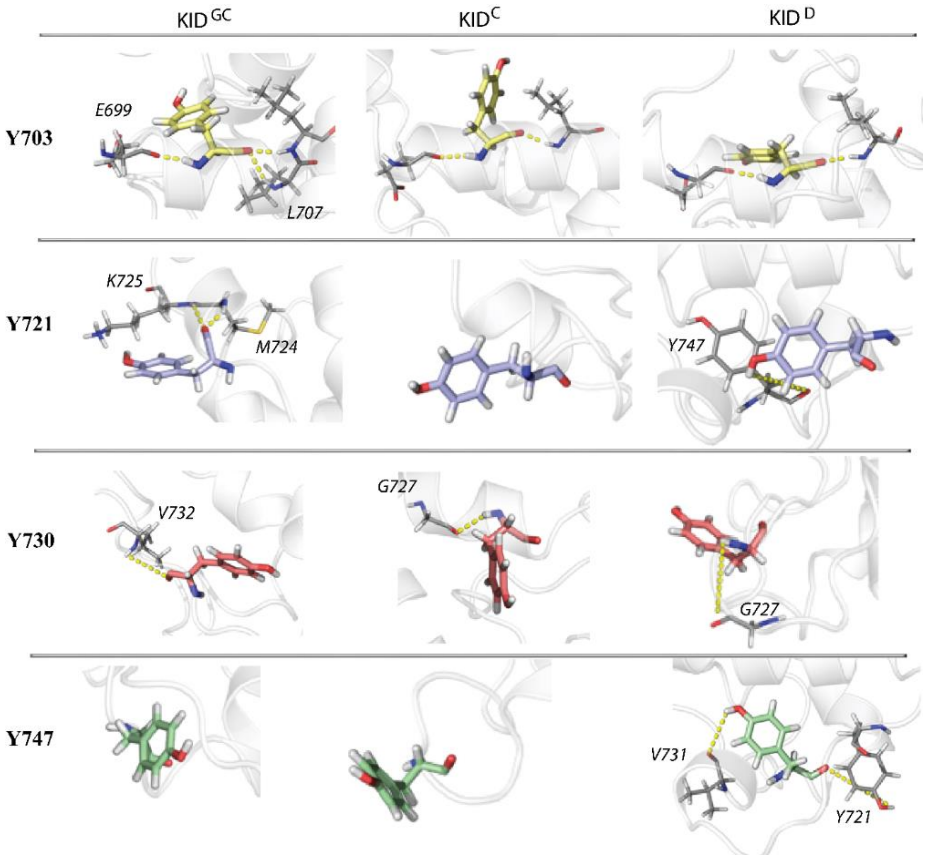


Figure 6.16 Snapshots of the KID fragments with tyrosine residues were taken at the top of each distribution (**Figure 6.13**, **B**). Oxygen and nitrogen atoms are in red and blue; respectively, carbon atoms are coloured differently for each tyrosine residue (Y703 in yellow, Y721 in lilac, Y730 in coral, and Y747 in green) and similarly (in grey) for other KID residues. Protein is shown as a grey cartoon. The yellow dashed lines represent the hydrogen bonds between atoms of the named residues.

Tyrosine Y703 showed equivalent H-bond interactions in all KID entities, either thanks to its helical folding, or similar orientation of its sidechain relative to its environment. Only the Y747 in KID^D is a single exception forming an H-bond by its sidechain.

6.2.2.4. Multiparameter Clustering of KID Conformations

KID in any studied entity is an IDP possessing transient helices interconnected with flexible loops, increasing the difficulty of regrouping structurally similar conformations based on criteria such as the RMSD. A set of 31 features (metric space) related mainly to the intrinsic polypeptide geometrical properties were selected for clustering to deliver the independence of KID-generated conformations from any referencing structure. Those algorithms require data pre-processing such as scaling and the important features selection step to improve the clustering results by discarding redundant embedded information.

First, the data were scaled between 0 and 1 for each KID entity and each metric (feature) individually. Next, the feature selection was performed by looking for high correlations/anti-correlations between feature pairs. Finally, the data dimensionality was reduced by Principal Component Analysis (PCA), keeping the first k components explaining up to 80% of the variance.

The correlation matrix constructed on these metrics revealed several correlations (**Figure 6.17**, **A**). Focusing on features with correlation coefficients (c.c.) ≥ 0.6 or ≤ −0.6, we first observed that the S717-K725-R739 triangle area positively correlates to the distance S717-K725 (c.c. of 0.8). The rest of the considered correlation values depend solely on features involving tyrosine residues (pairwise distances and dihedral angles).

The size (volume) of the tetrahedron formed by the tyrosine residues is positively correlated with the inter-tyrosine distances Y721-Y730 and Y721-Y747 (c.c. 0.74 and 0.67, respectively). Such dependence is mainly delivered by the spatial mobility of Y721, located between the highly fluctuating residues S717 and K725, whereas Y730 and Y747 are positioned near the low fluctuating residues V732 and P754, respectively. Other correlated tyrosine features are the backbone dihedral angles. Tyrosine Y721 ψ angle is positively correlated with Y730 φ angle with a coefficient of 0.6. This indicates that the terminal Cα-atoms of fragment Y721-Y730 twist in the same direction during MD simulation. On the contrary, Y747 φ and ψ angles are anti-correlated (c.c. −0.6), suggesting a twist in the opposite direction of Y747 NH- Cα and Cα-CO planes. We identified two KID metrics (features) with high correlation/anti-correlation (≥0.8 or ≤ −0.8): the distance S717-K725 and S717-K725-R739 triangle area. Keeping both does not add robust discriminating information for clustering. For further analysis, the latter was removed from the dataset.

The PCA dimensionality reduction on the remaining 30 features showed that the first two or three principal components (PCs) explain only minimal variance, 37 and 10%, respectively (**Figure 6.17**, **B**). The most portion of variance (80%) is described by the twelve first PC, which were selected as a final dataset.

However, among more than 30 features used for clustering, five metrics representing the KID shape (the radius of gyration, Rg), the distance between the most fluctuating residues (S717 and K725), the distance between Y747 with any other functional tyrosine, and two parameters characterising the internal geometry of tyrosine (ψ angles of Y721 and Y730) clearly distinguished the clusters formed by similar conformations (**Table S4**).

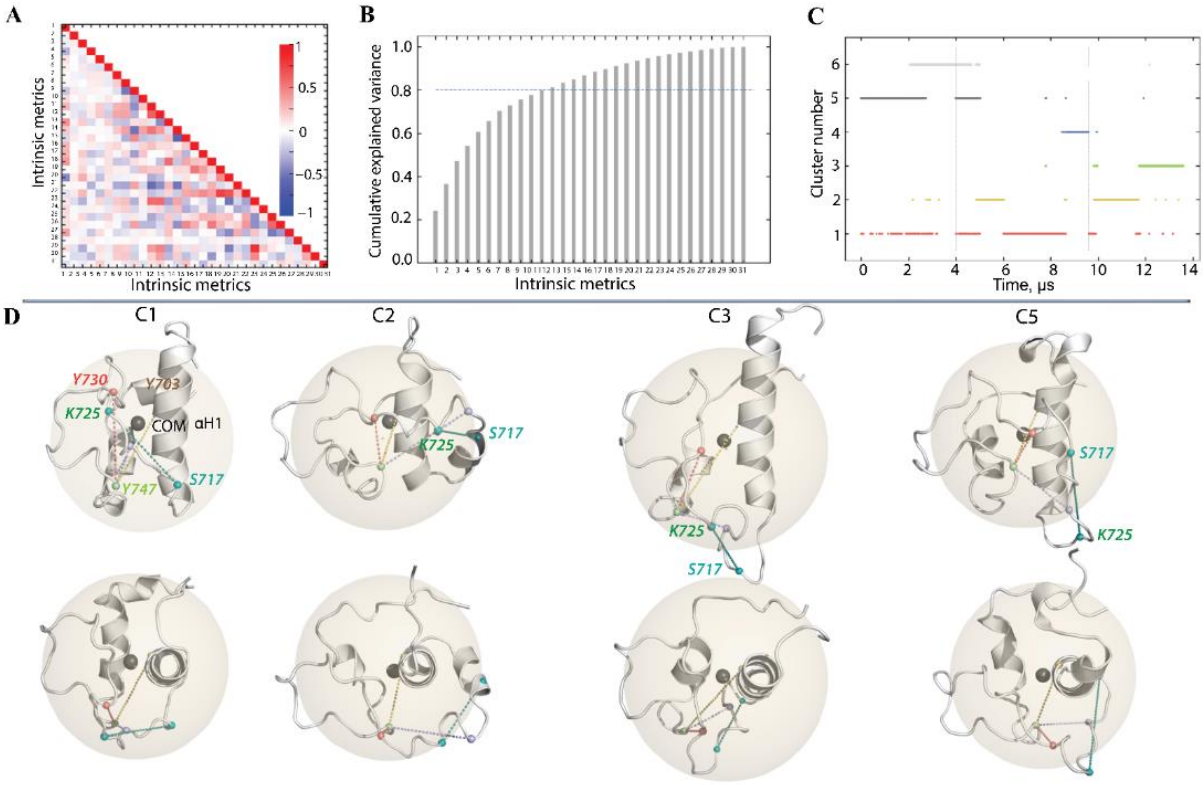


Figure 6.17 Clustering of KID conformations was performed on the concatenated 14 μs trajectory. (**A**) A correlation matrix showing correlation coefficients between 31 features. The features are (1) Rg, (2) number of hydrophobic contacts, (3) number of H-bonds, (4–7) SASA for each tyrosine residue, (8–15) backbone dihedral angles ϕ and ψ, (16–21) distances between the tyrosine residues, (22) tyrosine tetrahedron volume, (23–25) distances between the maximally fluctuating residues, (26–28) dihedral angle between the maximally fluctuating residues, (29) triangle area between the maximally fluctuating residues, (30) distance between the minimally fluctuating residues, (31) dihedral angle between the minimally fluctuating residues. (**B**) PCA analysis performed on the pre-processed data. (**C**) Clusters obtained from the concatenated 14 μs trajectory and their composition. Each colour represents different cluster. (**D**) The representative conformations from the most populated clusters are shown in three views—side and top—concerning the αH1-axis. Protein is shown as a cartoon. The sphere corresponds to the gyration radii (Rg).

Thereby, C1, C2, and C5 represent le most compact clusters based on the radius of gyration, whereas C3 is the most extended. However, the tyrosine geometry and most fluctuating residues in C1, C2, and C3 show apparent differences: among the most populated cluster, C1 and C3 have a similar shape as KID ante and post-transition conformations, as observed in [385]. Despite its extended shape, C3 possesses a tighter residue-wise rectangular geometry.

Further, all generated KID conformations were classified according to their similarities using different clustering methods—DBSCAN[268], K-means algorithm and hierarchical agglomerative. The clustering performance was evaluated with the Silhouette score[269] and Calinski–Harabasz score[386].

A first run of the data in each algorithm on a set of hyperparameters was conducted, and their performance was calculated to find the most suitable method. The K-means method showed the best scores, followed by hierarchical agglomerative clustering and DBSCAN (**Figure S22**). The best agreement between scores was obtained for K-means with k = 5 and k = 6. The clustering with k = 5 gave the best Silhouette (0.35 versus 0.33), whereas k = 6 yielded the best Calinksi–Harabasz score (37,906 versus 39,536). Given the difficulty of distinguishing the optimal number of clusters based on relative performance score values, the final clustering was performed for k = 5 and k = 6.

The scores for both types of clustering performance are similar, 0.33 and 0.30 with a Silhouette and 39,536 and 37,906 with a Calinski–Harabasz score for k = 5 and k = 6, respectively. A contingency table to identify strong clusters showed that both clusterings agree well (**Table S5**). The results show strong agreement for 65% of the total clustered conformations.

In particular, the cluster population strongly agrees in terms of the similarity of conformations in between two clusters found for k = 5 or 6 for $C5_{k=5}/C1_{k=6}$ and $C3_{k=5}/C6_{k=6}$ (21%), and for $C1_{k=5}/C5_{k=6}$ and $C4_{k=5}/C2_{k=6}$ (44%) (**Figure S23**). However, the difference between the results obtained by the two clusterings is observed in only 35% of the total conformations (C2 for k = 5 or C3 and C4 for k = 6). To avoid this ambiguity, we chose to keep the last clusters as C5 and C6, respectively. Finally, the strong population size is 22.9, 21.2, 13.8 and 7.1% for clusters C1–C4, respectively. The more ambiguous clusters, C5–C6, encompass 22.5 and 11.9% of the total clustered KID conformations.

The composition of the cluster population shows that the MD conformations of each KID object are contained in all clusters, albeit in different proportions. Cluster C1 is composed of conformations from all simulated KID: KID^{D} (1%), KID^{C} (20%) and KID^{GC} (3%); C2 comprises a mix of KID^{C} (8%) and KID^{GC} (13%); C3 and C4 are composed only of conformations issued from a lone KID entity—KID^{GC} (14%) and KID^{C} (7%), respectively. Finally, C5 and C6 are composed of a mixture of KID^{D} and KID^{C} with a prevalence for KID^{D} (17 and 11%) when the KID^{C} population represents only 6 and 1% in the respective clusters. The remaining population of conformations not regrouped into the clusters represents less than 1%. The representative conformations from the most populated clusters—C1, C2, C3, and C5—composed of the MD conformations of all KID entities are shown in **Figure 6.17**, **D** with their gyration radii (Rg).

6.2.2.5. The Gibbs Free Energy (ΔG) Landscape of KID Conformations

The conformational diversity of IDP KID can also be assessed via the Gibbs free energy (ΔG) landscape as it was applied in [29,38] for KID^{D} and KID^{C}. The ΔG representation provides a statistical overview of the KID conformational ensemble as

a function of two reaction coordinates. It is essential to use a statistical thermodynamic treatment to analyse the data rather than assuming a two-state transition. Such treatment could be simple, but it should consider conformational entropy explicitly in terms of ensembles of microstates. Molecular simulations can test the physical significance of the choice of model used to analyse the generated data.

In our case, using rich data of the concatenated trajectory obtained by merging all trajectories of KID^{GC}, KID^{D}, and KID^{C} presents a rare opportunity to compile the MD conformations obtained from different KID entities simulated under similar conditions.

First, to generate the free energy landscape (FEL) of IDP KID, we used the first two principal components (PC1 and PC2) of a PCA as reaction coordinates (**Figure 6.18**, **A)**. The FEL PC1 vs. PC2 shows a rugged landscape revealing KID high conformational diversity with well-defined minima indicating the multimodal distribution of both PC1 and PC2 (**Figure 6.18**, **B**, **C**). The deepest well, W5, together with the adjacent low minimum W6, forms a conformational space (area 1) separated from the other (area 2) by a very high energy barrier. This splitting was created due to the bimodal profile of the PC1 component separating these two regions on a three-dimensional relief. Area 2 is complex, and it displays a series of minima represented by the lowest combined well (W1–W4) and distant minima W1, W3, and W4, separated due to the multimodal distribution of the PC2 component.

Interestingly, the observed minima correspond perfectly to the multi-parameter clustering results: such clusters, C1, C3–C6 (**Figure 6.18**, **A**), are identifiable with the deepest wells on the FEL of KID (**Figure 6.18**, **B**, **C**). The wells W1 and W3 are deep but more spread out. The remaining centred extended well (W1–W4) includes conformations from C1–C4, but those clusters are well-defined with K-means.

Further, we explored the credibility of using as the reaction coordinates for FEL generation the metrics characterising the highly fluctuating residues, the pseudo-torsion angle (T), and the respective distance D between two highly fluctuating atoms of KID. As the residues S717, K725, and R739 systematically exhibit the highest RMSF values in all studied KID species (**Figure 6.14**, **B**), and the pseudo-torsion angle characterising their relative position are correlated (**Figure 6.14**, **F**), 2D and 3D FELs were generated using three pairs of these metrics (T versus D) (**Figure 6.18**, **D**). The analysed residues are regularly positioned at the KID sequence, separated by 13–14 residues. In the 3D structure, these residues are also located on highly remote structural segments, but their positions are not equidistant in three orthogonal directions of KID (**Figure S21**). Nevertheless, we suggested that the FELs constructed using these metrics can supplement the description of the KID free energy.

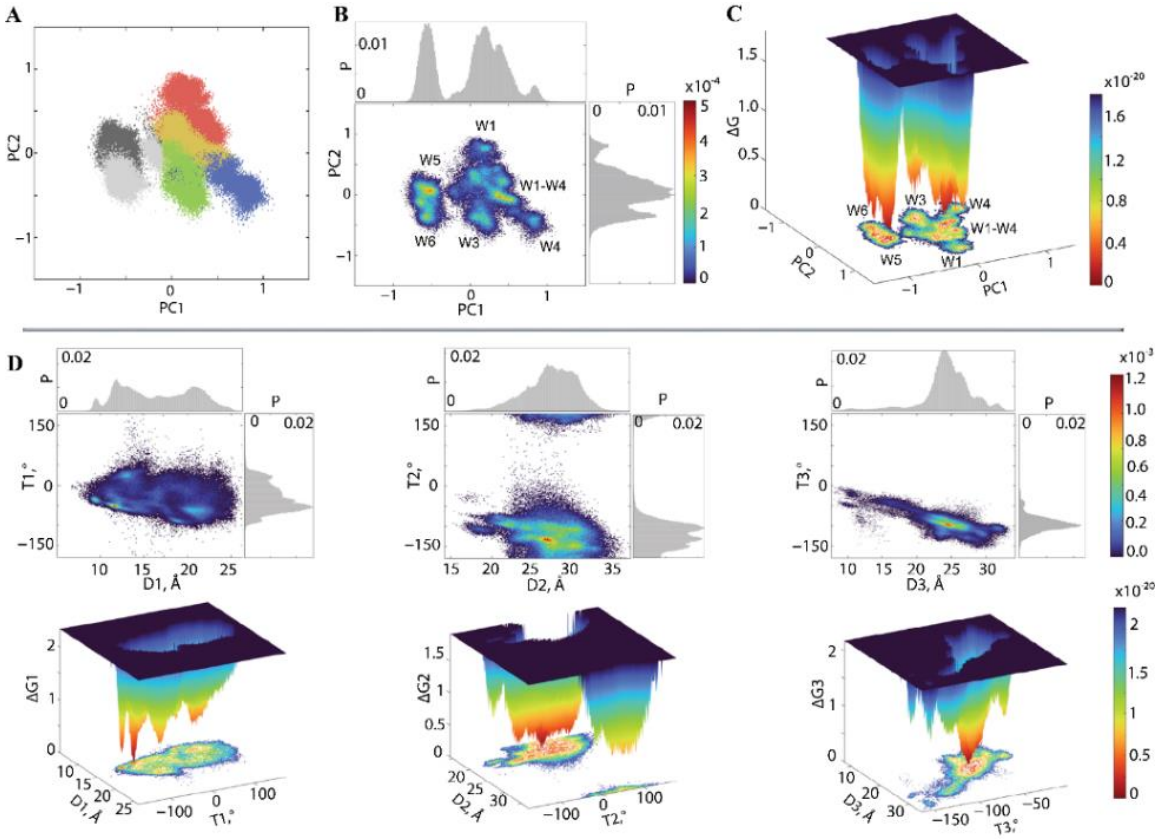


Figure 6.18 Free energy landscape (FEL) of KID in the 2- and 3-dimensional representations as a function of the reaction coordinates. (**A**) Projection of KID conformations on two principal components, PC1 and PC2. Clusters are coloured as in **Figure 6.17**, **C**: C1 in red, C2 in yellow, C3 in green, C4 in blue, C5 in dark grey, and C6 in light grey. (**B**, **C**) FEL of KID defined on PC1 and PC2 as the reaction coordinates. (**D**) FELs of KID determined by using as the reaction coordinates the most fluctuating metrics in KID—pseudo-torsion (dihedral) angles T1 (K725-E699-Y703-S717), T2 (R739-E699-Y703-K725), and T3 (R739-E699-Y703-S717) and the respective distances D1–D3 between two highly fluctuating atoms. FELs were generated on the 14 μs concatenated trajectory composed of MD conformations of all KID entities studied—KID^{C}, KID^{D}, and KID^{GC}. Blue represents the high energy state, green and yellow low and red represents the lowest stable state. The free energy surface was plotted using Matlab.

All FELs are highly different, showing either multiple low minima (T1 vs. D1 and T2 vs. D2) or only the one surrounded by significantly lower wells (T3 vs. D3). The FELs with multiple minima have different profiles reflecting the anisotropic positions of the reference residues in the 3D space. High energy barriers separate the wells on each FEL.

The Gibbs free energy landscape dissected from the first principal components and geometrical metrics of the highly fluctuating residues show multiple minima, as expected for intrinsically disordered proteins.

6.2.3. Discussions

In this paper, we analysed KID, a crucial domain for the RTK KIT transduction process, represented by three entities: (i) a generic macrocycle (KID^{GC}), (ii) a cleaved isolated polypeptide (KID^{C}), and (iii) a natively fused TKD domain (KID^{D}). Obtained results lead us to the postulate that these KID entities have similar structural and dynamic characteristics indicating the intrinsically disordered nature of this domain. This finding means that both polypeptides, cyclic KID^{GC} and linear KID^{C}, are valid models of KID integrated into the RTK KIT and will be helpful for further computational and empirical studies of post-transduction KIT events.

Previous studies showed that KID, either as an isolated polypeptide or integrated into KIT, has a helical folding and globular shape stabilised by multiple non-covalent interactions[271,290]. The newly constructed generic cyclic KID^{GC} displayed a similar, but more compact, globular shape and was characterised by increased helical content.

The functional tyrosines Y703, Y721, and Y730 are located on a sphere of varying radii and are fully accessible to solvent. The size of the segments is different—more compact for the Y730 as for the stabiliser Y747 and very dispersed for Y721 spreading over the whole hemisphere. Moreover, in all KID species, phosphotyrosines are not stabilised by specific non-covalent bonds (with side chain contribution) that allow their full solvent availability and accessibility to protein–protein interactions and phosphate transfer.

Analysis of the tyrosine residues' dihedral angles (Ramachandran plots) delivered additional information regarding their secondary structure interpreted by DSSP. In particular, Y703 is undoubtedly located in the same region of φ vs. ψ distribution in all studies species, corresponding to a helical folding. At the same time, the geometry of Y721 and Y730 diffuses between regions corresponding to a helical folding, coil, and β-strand with various frequencies between all species. The dihedral area of Y747 in KID^{GC} and KID^{D} is identical.

Like KID^{D} and KID^{C}, KID^{GC} displays high flexibility providing its conformational diversity, as seen by the residue fluctuation profiles and clustering.

Hence, the apparent disorder of KID arises from the competition between intra-KID non-covalent interactions and high flexibility, which causes KID to dynamically alternate between sub-ensembles with different unstable fold architectures. This behaviour contrasts with the usual disorder interpretation indicative of absent tertiary interactions.

It is likely that the intrinsic disorder permits KID to bind partners via either

conformational selection (fold first and then bind)[371] or induced-fit (bind first and fold while bound)[387] processes or alternating between conformational selection and induced fit[388]. Moreover, to fold upon binding as a conformational switch, KID sequences must fully encode all the structures they form in complex with diverse partners.

By grouping the generated MD conformations of all KID species according to their intrinsic characteristics, we observed the partial overlap of their conformational spaces. Therefore, whether isolated linear polypeptide, cyclised macrocycle or the domain integrated into the native KIT, KID explores similar conformations. The representative conformations of the most populated clusters show that KID has mainly a compact globular spherical shape and, less frequently, an ellipsoidal surface. These shape differences are reflected in the configuration of the tyrosine residues and the distance between the most fluctuating residues.

The Gibbs free energy landscape generated from the first principal components and geometrical metrics of the highly fluctuating residues, which form a set of KID intrinsic dynamical and geometrical features, show multiple minima as expected for intrinsically disordered protein.

We do not expect a similarity between the FEL constructed by using as reaction coordinates two principally distinct metrics—the first principal components (PC1 vs. PC2) and two geometrical measures (the inter-residue distances and pseudo-torsion angle). In computational structural biology, classical PCA reduces the big data dimensionality of extensive MD concatenated trajectories. PC1 and PC2 are the product of the eigenvectors and eigenvalues of the covariance matrix and characterise two orthogonal directions in space along which projections have the most significant variance, interpreted as the amplest atomic displacements in each MD conformation, mainly contributing to the essential dynamics[389]. Our analysis used the normalised and feature-selected dataset of intrinsic features separated from the dynamics. In contrast, the distance and pseudo-torsion (dihedral) angle describe only a subset of this dataset as the systematically measured relative geometry of the three chosen residues with the highest RMSF values.

Clustering based on over thirty features independent from any referencing structure and free-energy landscape construction on the features dataset projection on two first principal components should be best-suited to study the conformational diversity of KID of RTK KIT.

Our results demonstrate that KID^{GC} and KID^{D} display similar structural, conformational, and dynamic properties and energy-related characteristics; KID^{GC} can be used for empirical studies of KID phosphorylation and binding with its specific signalling proteins. However, since the kinase domain is a central hub of both receptor

activation and communication between distant functional regions—JMR, A-loop, KID, and C-terminal, investigation of signal transduction mechanisms or the mechanisms of allosteric regulation of KIT in the native or mutated state would require full-length KIT or, at least, its full-length cytoplasmic domain.

6.3. MODULARITÉ DU HVKORC1

Résumé. *La région de recrutement de la protéine redox responsable de la première étape d'activation du hVKORC1 est la boucle luminale (boucle L). Les simulations de dynamique moléculaire en mécanique classique et accélérée de la boucle L clivée du domaine transmembranaire avec ou sans restriction aux résidus terminaux ont montré que la boucle L conserve ses propriétés désordonnées et varie d'une forme arrondie (conformation « fermée ») à aplatie (conformation « ouverte »). Les paysages d'énergie libre construits sur deux coordonnées de réaction pour l'ensemble des conformations issues des simulations de la boucle L insérée dans hVKORC1 et la boucle L clivée montrent de nombreux minima locaux composés de conformations hétérogènes. Par cette comparaison, nous avons suggéré que hVKORC1 est une protéine modulaire dont la boucle L désordonnée, clivée ou non du reste de la protéine, possède toutes les propriétés requises pour la reconnaissance et l'initiation de l'activation du hVKORC1 par une protéine partenaire redox. Les simulations de dynamique moléculaire de hVKORC1 complet et de la boucle L clivée du modèle* de novo *ou des structures cristallographiques ont permis de délivrer des cibles adéquates de hVKORC1 pour la modélisation de son INTERACTOME.*

6.3.1. INTRODUCTION

hVKORC1 main functions – reversible activation/deactivation over the thiol-disulphide exchange with its redox protein, and recurrent reduction of vitamin K – involve L-loop and TMD, respectively. Our previous studies[220,221] and results reported above in this work, have shown that (i) transient secondary structures and high conformational variability qualify L-loop as an intrinsically disordered region, (ii) the TMD structure is strictly ordered, and apparently do not dependent on L-loop disorder, and (iii) there are no long-lived H-bond interactions between TMD and L-loop. These hVKORC1 properties observed in all studied models, except for limited conformational flexibility of the L-loop in the crystallographic structures, suggest that hVKORC1 is a modular protein and that L-loop and TMD are two of its structural and functional subdomains.

As two kinds of L-loop disorder – transient folding and conformational flexibility – do not appear to affect hVKORC1 TMD structure at a given (oxidised) state, we

examined if hVKORC1 can be considered a modular protein composed of the transmembrane domain and L-loop, as two sub-domains of the enzyme having their distinct structural properties and fulfilling their own functions? Then, in the context of enzyme activation, we asked what conformation of hVKORC1 L-loop is an authentic target for redox protein PDI that may be used for automatic docking experiments?

The ability of modular protein domains to independently fold and bind proteins has allowed a significant number of protein-protein interaction studies *in silico*, *in vivo* and *in vitro* performed on isolated modules used as more accessible items[7].

6.3.2. RESULTS

6.3.2.1. IS HVKORC1 A MODULAR PROTEIN?

To investigate this domain as a suitable promoter in the activation/deactivation process of hVKORC1 by its redox protein, we studied L-loop as an isolated polypeptide cleaved from TMD. The slightly extended L-loop (R33-N80) was cleaved from (i) the homology models of hVKORC1, holo-h and apo-h, generated from two crystallographic forms of the oxidised state, (ii) their derivatives with the relaxed H-bond Q78···G62, and (iii) the *de novo* model of the same enzymatic state. The initial models of L-loop differ considerably as shown by the RMSD values, which are smaller between the apo-h and holo-h forms (of 4.0 Å) than between each form and *de novo* model (of 8.4 Å) (**Figure 6.19**, **a**).

First, the cleaved polypeptides were studied by the conventional (cMD) and accelerated (GaMD) simulations with the soft restraints (elastic retaints) to maintain the flexible ends of each polypeptide at the distance observed in hVKORC1. Restriction on the L-loop ends distance mimics the steric conditions imposed by the enzyme transmembrane domain. Although using constraints upon the MD simulation with constant pressure is a questionable subject, the applied restriction on L-loop ends allowed the distance preservation between R33 and N80 residues in the range of 15-20 Å, comparable with the distance of 15 Å in the empirical structures[218] and of 16 Å in the *de novo* model[220].

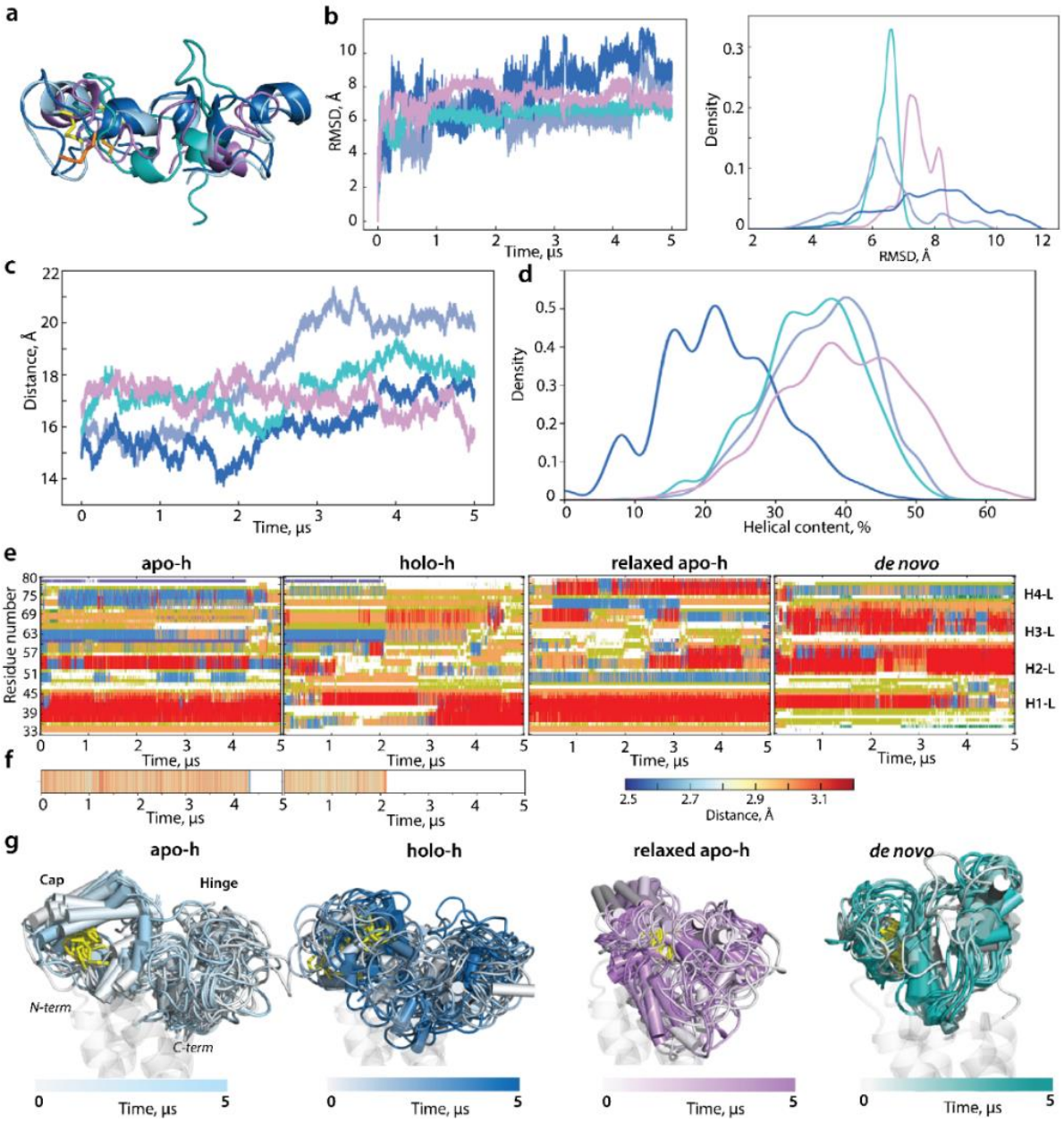


Figure 6.19 Structure and conformation of hVKORC1 L-loop studied as isolated polypeptide. (**a**) L-loop was extracted from apo-h and holo-h, relaxed apo-h and *de novo* models, and simulated (cMD) over 5 μs. (**b**) RMSDs of each form after Cα-atoms fitting to their respective initial conformations (left), and RMSD probability density (right). (**c**) Distance between L-loop ends (Cα–Cα atoms). (d) RMSFs computed for the Cα atoms of each form after fitting on their respective initial conformation. (**d**) Helical content of each model. (**e**) Time-dependent evolution of secondary structure in each L-loop form, (from left to right) apo-h, holo-h, relaxed apo, and *de novo*, as assigned by DSSP: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow. (**f**) Time-dependent evolution of H-bond Q78···G62 in L-loop cleaved from apo-h and holo-h forms. (**g**) Superimposition of 20 conformations taken each 25 ns of cleaved L-loop cMD simulations. The colour gradient shows the time-dependent conformations, from light (t = 0) to dark (t = 5μs). (a, g) Protein is shown as cartoon with disulphide bridges in yellow sticks. L-loop cleaved from different hVKORC1 forms is distinguished by colour: blue light (apo-h), dark blue (holo-h), teal (*de novo*) and lilac (relaxed apo-h).

Remarkably, even without any condition, H-bond Q78···G62 stabilising L-loop hinge has broken at 4.3 µs (apo-h) and 2.2 µs (holo-h) of cMD simulation of the cleaved L-loop. Curiously, the H-bond rupture was spontaneous and apparently unrelated to the variation in distance between the N- and C-ends of the L-loop. The H-bond disruption is accompanied by several structural effects displayed as (i) a change in the folding of residues R61 and N80, (ii) a complete unfolding of the H3-L double helix, (iii) a tendency of H2-L to fold as a 3_{10}-helix and (iv) stabilization of the long H1-L helix.

The extended (5-µs) cMD simulation of cleaved L-loop has generated more similar conformations between apo-h, especially in its relaxed version, and *de novo* model, than between two forms derived from crystallographic structures (**Figure 6.19**). Furthermore, L-loop cleaved from apo-h and relaxed apo-h forms and *de novo* model show (i) close RMSD profiles and values, and (ii) very similar portion of ordered (helical) structures (40, 38 and 38% in apo-h, relaxed apo-h and *de novo* respectively) which is twice higher than in holo-h form (21%). Despite the equal number of helices (four helices), the principal difference between apo-h and *de novo* model is a structural organisation of cap showed a large H1-L helix in apo-h and relaxed apo-h, and poorly folded or unfolded segment in *de novo* model; the other helices rather different in type (α- and 3_{10}-helices) and length. Nevertheless, in all studied cleaved L-loop models, all helices are transient, reversibly conversed between α- and 3_{10}-helix (α-helix → 3_{10}-helix), and between 3_{10}-helix and turn or bend (3_{10}-helix → turn/bend). These recurring transitions of L-loop secondary structures over the trajectory promote helices with great length variation. Proportions of folded structures (α- and 3_{10}-helices) in cleaved L-loop are identical to L-loop in the corresponding model of hVKORC1 (**Figure 5.20**, **d**; **Figure 5.21**, **c**; **Figure 6.19**).

These transformations in cleaved L-loop lead to an overall change in its shape, which tends to be more compact in both shapes and approximates L-loop prevalent shape in the *de novo* model (**Figure 6.19**, **g**). Similar results were observed over a GaMD simulation of the same length (data not shown).

Secondly, we investigated whether L-loop structure and conformations depend on steric conditions imposed on N- and C-ends, either naturally occurring in hVKORC1 or mimicked by soft constraints on cleaved L-loop terminal residues. Cleaved L-loop (apo-h, holo-h, relaxed apo-h, and *de novo*) was studied by 0.5 µs cMD simulations without any constraints.

RMSD and RMSF values of each fully liberated polypeptide vary within ranges observed for L-loop either fused to TMD or simulated as cleaved polypeptide under constraints (**Figure 6.20**). Similar to restrained L-loop, released L-loop folding (secondary structures) is transient in each form, and folding order comparable with end-restrained L-loop. In particular, the total of folded structures in cleaved L-loop simulated with constrains or not, is equivalent in apo-h (40/40%), holo-h (22/21%) and relaxed apo-h (38/37%) but decreased in *de novo* model (38/25%). Curiously, some

structural effects, the alternative structural content of residues R61 and N80 and full unfolding of double H3-L helix, were observed in apo-h L-loop simulated under both conditions on the terminal residues, either as the end-restrained sample or as the released polypeptide.

The distance between the L-loop ends in apo-h form was 21 Å during the first 250 ns of simulations, and further gradually decreased to 13 Å and maintained until the end of simulation. Similar variations of this distance were observed in apo-h form and *de novo* model simulated without hindrance. It should be noted that in many conformations of apo-h and *de novo*, the N-and C-ends distance values in unrestrained L-loop agree well with L-loop fused to TMD. In holo-h form this distance varies reversely in a very large range, from 5 to 35 Å. The shape of unconstrained L-loop is similar to related constrained forms (**Figure 6.19**; **Figure 6.20**).

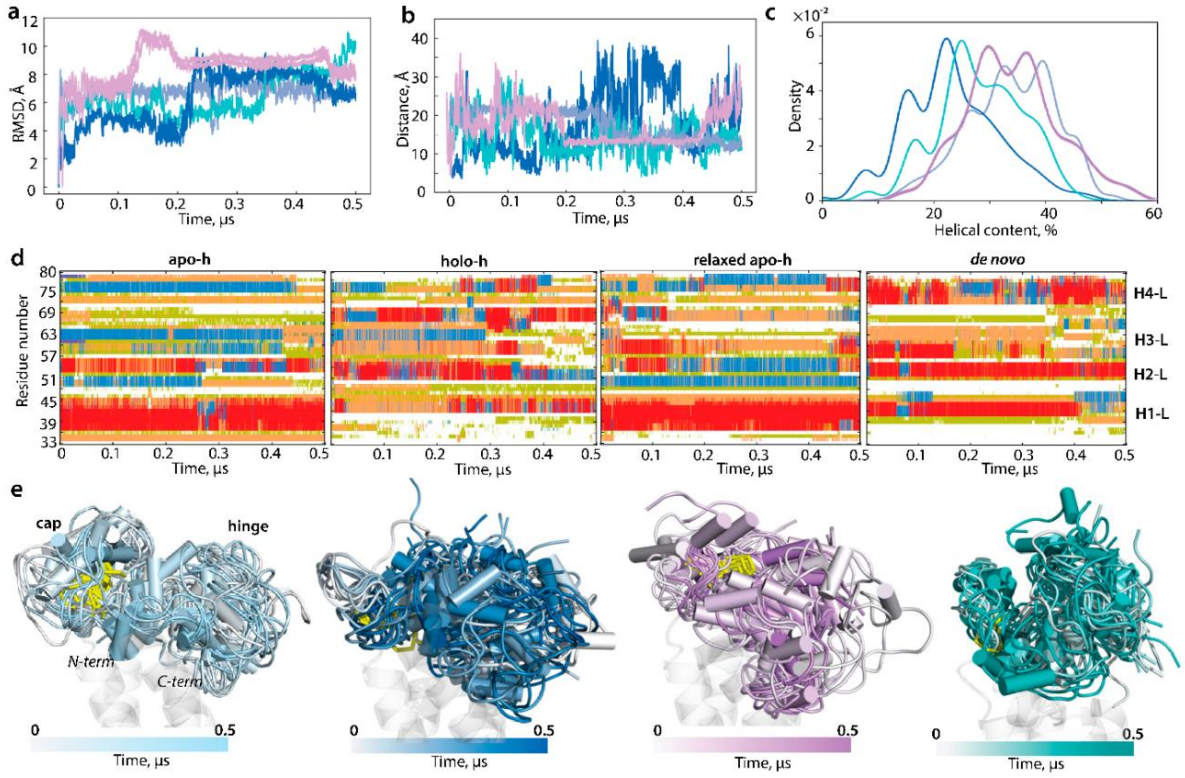


Figure 6.20 Structure of cleaved L-loop studied as fully released polypeptide. (**a**) RMSDs of each form after Cα-atoms fitting to their respective initial conformation. (**b**) Distance between L-loop ends (Cα−Cα atoms). (**c**) Projection of L-loop cMD conformations onto the first three modes calculated from principal components (PCs). (**d**) Time-dependent evolution of each L-loop form secondary structures, (from left to right) apo-h, holo-h, relaxed apo, and *de novo*, as assigned by DSSP: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow. (**e**) Overall helical content (left), the content of α- (middle) and 3_{10}-helices (right) observed in cMD conformations of each form. (f) Superimposition of 20 conformations taken each 25 ns of cMD simulations of L-loop relaxed apo-h and holo-h forms. The colour gradient shows the evolution of the trajectory, from light (t = 0) to dark (t = 5μs). Protein is shown as cartoon and disulphide bridges in yellow sticks. The studied entities of L-loop are distinguished by colour: apo-h is in light blue, h olo-h is in dark blue, relaxed apo-h is in lilac and *de novo* model is in teal.

Comparing each form of cleaved L-loop, simulated either under constrained N- and C-ends or fully released species, with the L-loop fused to hVKORC1, we affirm that in all cases L-loop is an intrinsically disordered region having (i) a helical fold with quasi-equal content of folded structures, (ii) a similar sequence position of transient helices (H1-H4), and (iii) a comparable overall 3D shape. Consequently, L-loop structural properties — folding, conformations and degree of the intrinsic disorder — are independent from its physicochemical context as the entity. We postulate that L-loop folding is not controlled by hVKORC1 transmembrane domain. At the same time, TMD restricts L-loop ends distance, probably optimising L-loop geometry.

Consequently, this analysis clearly confirms hVKORC1 modular structure composed of the quasi-rigid and stable transmembrane domain and intrinsically disordered L-loop. The TMD stability is maintained by an extended network of non-covalent interactions (**Figure S24**) organising four TM helices in a structural motif described as alpha-helical coiled-coil[390,391]. This motif is conserved in hVKORC1 inserted into membrane or placed in an aqueous solution. Indeed, during hVKORC1 MD simulations in different environments, the TMD helices do not vary in structure and are held together by similar or strictly identical non-covalent interactions supplying the TMD a tight packing in each model.

6.3.2.2. How are L-loop conformations similar?

Comparing time-related folding, we observe that transient events in different segments of L-loop sequence may be mutually related or not. Thus, ordering of one helix favours an unfolding of the other, as seen in the *de novo* model or holo-h form of L-loop simulated with restrained N- and C-ends (**Figure 6.19**, **e**). On the other side, holo-h form conformations display H2-L unfolding (from ~1 μs) followed by H3-L transition to coil (from 2 μs), synchronous with H4-L unfolding. The significant decrease in holo-h L-loop folding is partially compensated by H1-L increase in length and transient H3-L folding as α-helix. Nevertheless, the helical content of this form is obviously lower than in other studied L-loop entities. Folding of distinct segments in apo-h and relaxed apo-h forms shows different folding-unfolding effects relationships – interrelated, disconnected and their combination. Therefore, interrelations between folding-unfolding effects in disordered L-loop is apparently more sophisticated than their description at the secondary structures level, and probably also depend on global and local flexibility.

To study two structural processes in L-loop, folding and flexibility, and their interrelations, we tried to regroup similar conformations of cleaved L-loop simulated with constrained N- and C-ends as they represent the richest data (5μs trajectories). A search for similar MD conformations of L-loop from different form of VKOR was performed on the concatenated trajectories (apo-h, holo-h, relaxed apo-h and *de novo*). Primary, we found that L-loop conformations with minimal RMSD values (4-5 Å)

show shape similarity, while their secondary structures are highly different (**Figure 6.21**). Secondly, the secondary structures-based clustering[324] regrouped MD conformations (0.65 was chosen as the most appropriate value from tested cut-off: 0.5, 0.6, 0.65, 0.7 and 0.8) in seven clusters (C1-C7). Clusters C5 and C6 contain conformations derived from only one L-loop form (respectively apo-h and apo-h relaxed); clusters C1-C4 and C7 comprise conformations from several L-loop forms (**Figure 6.21**, **b**). Only clusters C2 and C4 are constituted by conformations generated over all four trajectories.

We observed that (i) the secondary structures of conformations regrouped in a cluster are only partially similar, usually showing resemblance for the one or two L-loop segments, while others differ significantly; (ii) conformations of each cluster C2, C4 or C7 are either similar in shape (C4 and C7) or largely distant (C2); (iii) independent from shape similarity, conformations within each cluster are highly different (RMSD of 6.5-10.0 Å) (**Figure 6.21**, **c**, **d**).

The results produced by RMSD and secondary structures-based clustering showed that both approaches are not suitable for analysis of intrinsically disordered L-loop. We searched more intuitively/manually L-loop conformations with both minimum RMSD and close structural similarities that resulted in finding more comparable conformations than those obtained using automatic clustering based on a single criterion.

Thirdly, concatenated data combining cMD trajectories of L-loop cleaved from apo-h, holo-h, relaxed apo-h forms, and *de novo* model were analysed with principal component analysis (PCA). Projection of the generated conformations on the principal components PC1-PC3, shows partially overlapped subspaces (**Figure 6.21**, **d**). These overlapped areas relay successively all subspaces and form compact generic ensemble of L-loop conformations. Evidently, the generic ensemble combined from subspaces does not represent a full conformational space of disordered L-loop but reflects more exhaustively its conformational properties than a unique subspace.

Conformational space of the cleaved L-loop simulated with unrestrained N- and C- ends is less compact and significantly less explored (by factor 10). Nevertheless, similarly, to constrained L-loop, each subspace overlaps with another, and subspace formed by relaxed apo-h conformations showed an overlapping with apo-h and holo-h forms.

Clearly, cMD simulation of each cleaved L-loop generates only a limited part of the intrinsically disordered L-loop overall conformational space. We suggested that combined representation of these limited portions using the free energy landscape model will allow a better characterization of the generated set, if it is not yet complete. We are aware that such a representation does not provide the free energy quantitative characteristics but can be useful for the comparison of different datasets.

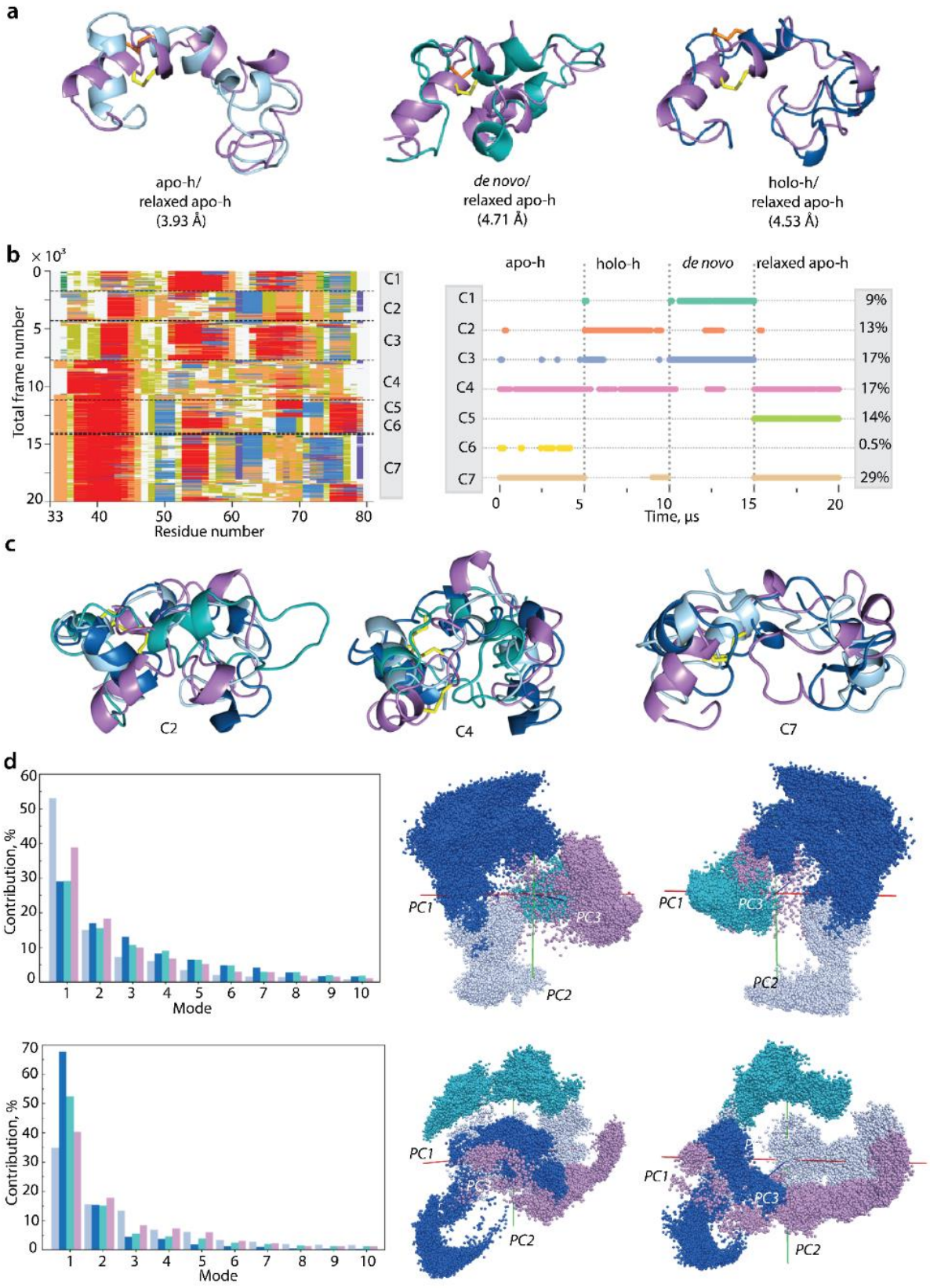


Figure 6.21 Clustering of MD conformations of L-loop simulated as an isolated polypeptide cleaved from each form – apo-h holo-h, relaxed apo-h and *de novo* model. (**a**) Conformations with the lowest RMSD values (< 4.0 and 5.0 Å, shown in parentheses) are superimposed. (**b**) Secondary structures-based clustering[392] with threshold 0.65 (left) performed on the time-depended

evolution of each L-loop form secondary structures as assigned by DSSP: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow. Population and composition of clusters (right). The simulation time is displayed as cumulative for four trajectories. (**c**) Superimposition of L-loop from the most populated clusters C2, C4 and C7, composed of conformations generated from at least three different forms (b). (**d**) PCA analysis of concatenated cMD trajectories of cleaved L-loop simulated with constrained N- and C-ends (top panel) and without constraints (bottom panel). The bar plot gives the eigenvalue spectra in descending order for the first 10 modes calculated on each cMD trajectory (left). Projection of L-loop cMD conformations onto the first three principal components (PC) (right). The concatenated trajectories were least-square fitted on the mean conformation to remove rigid-body motions. (a-d) L-loop from apo-h (blue light), holo-h (blue dark), relaxed apo-h (lilac) and *de novo* (teal) is shown as cartoon with helices and disulphide bridges in yellow sticks.

The relative Gibbs free energy, ΔG, defined on chosen coordinates called reaction coordinates, describes a protein conformations between two or more states, measured as the probability of finding the system in those states[297]. For the relative free energy (ΔG) evaluation and reconstruction of L-loop conformational ensemble landscape, primary measures — radius of gyration (Rg) and distance (RMSD) — were used as reaction coordinates. The free energy landscape (FEL) as a function of RMSD and radius of gyration Rg (FEL^{Rg}_{RMSD}) was characterised for the concatenated data of cMD trajectories of L-loop cleaved from each form of hVKORC1 (holo-h, apo-h, relaxed apo-h and *de novo* model). The FEL^{Rg}_{RMSD} determined on normalised conformations shows two very closely positioned narrow deepest potential wells, W1 and W2, complemented by W3, an adjacent well satellite (**Figure 6.22**).

All these very proximal wells, separated by InfleCS* method[393] are composed of compact globule-like conformations generated from relaxed apo-h form and *de novo* model (W1), relaxed apo-h and apo-h forms (W2) and a mixture of apo-h, relaxed apo-h and *de novo* (W3). Conformations from these wells are similar principally by size (Rg values ~10-10.5 Å) and shape (closed form of L-loop), and apparently are enabled to reversible transition as viewed by a low ΔG barrier on FEL^{Rg}_{RMSD}.

Shallow flattened wells W4 and W5 include L-loop conformations derived from all analysed samples. These wells L-loop conformations differ mainly in shape and size – compact globular-like L-loop (closed conformation) in W4, and elongated L-loop (open conformation) in W5. Despite the conformations similar shape in W4 and W1-W3, higher RMSD values in W4 are a discriminating factor leading to the separation of W4.

Both analytical methods, PCA and reconstruction of the relative free energy landscape using the primary measures — radius of gyration (Rg) and distance (RMSD) — as two reaction coordinates, showed that sampling of L-loop by multiple independent MD simulations from largely different initial molecular conformation originated either from empirically determined structures or theoretically predicted

models, converge to similar L-loop conformations. Study of cleaved L-loop accesses this convergence more rapidly than for the entire protein, even sampled by GaMD simulation.

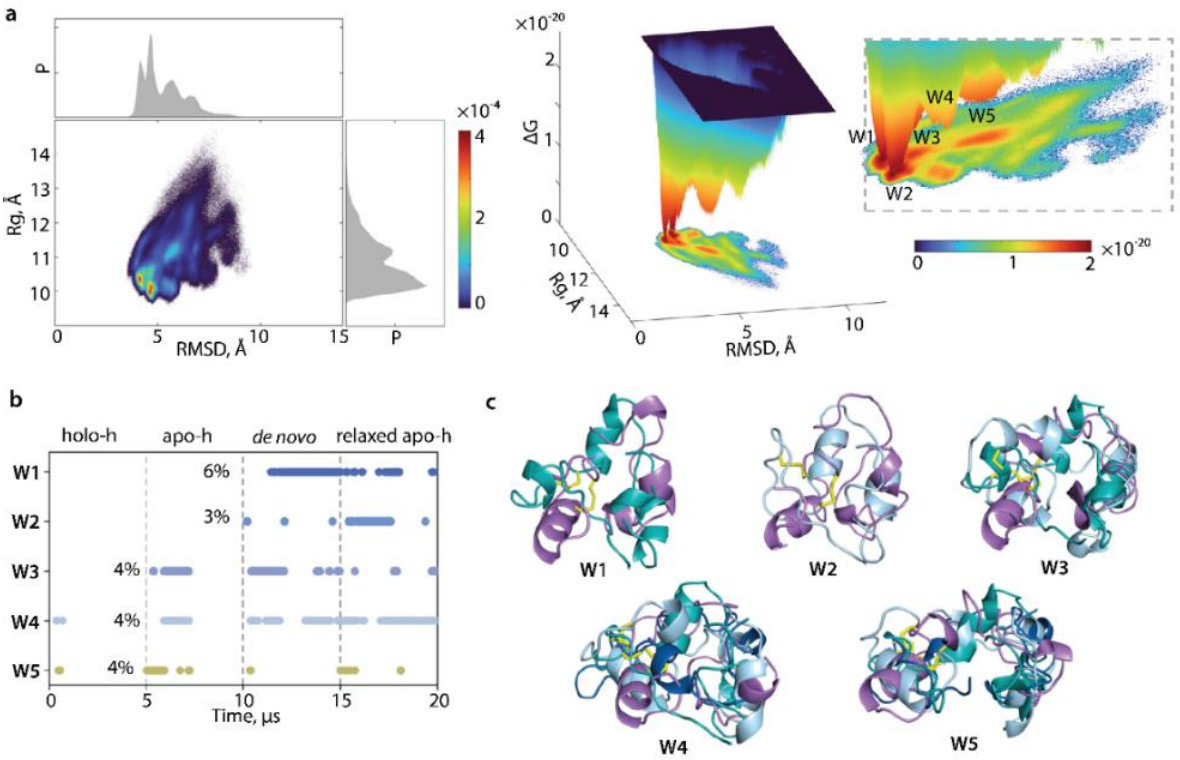


Figure 6.22 Free energy landscape (FEL) of cleaved L-loop as a function of reaction coordinates, Rg versus RMSD. (**a**) 2D- (left) and 3D dimensional (middle) representation of the L-loop conformational ensembles relative Gibbs free energy, zoomed on the principal wells (right). Probability density (P) of each reaction coordinate is shown on the 2D map top and right. (**b**-**c**) Content of each well (W1-W5) illustrated as the superimposed representative conformations. All conformations were generated by cMD simulations of different forms of L-loop and fitted on the average conformation calculated on the merged data. The red colour on 2D and 3D-diagrams represents the highest occurrence, yellow and green low, and blue represents the lowest occurrence. The free energy surface was plotted using Matlab. L-loop is shown as a cartoon distinguished by colour: the holo-h is in light blue, the apo-h is in dark blue, relaxed apo-h is in lilac and *de novo* model is in teal. Content of wells for the merged data is shown over the concatenated time range.

6.3.3. Discussions

According to the obtained results, hVKORC1 is composed of the stable transmembrane domain and intrinsically disordered L-loop. The TMD stability is due to multiple inter-helices non-covalent interactions – H-bonds and van-der-Waals contacts –, which maintained four transmembrane helices together, forming a coiled coil[390]. This structural motif is perfectly preserved in hVKORC1 inserted into a membrane or placed in an aqueous solution, as was shown by classical or accelerated

MD simulations. Location of hVKORC1 on the endoplasmic reticulum membrane is crucial for its proper folding[394], and once it is already folded, we hypothesised that the membrane does not affect hVKORC1 structure maintained by preformed interactions within the TMD. Moreover, TM helices position is well-maintained on a membrane, showing only collective drift[220].

The absence of stable non-covalent contacts between well-ordered highly stable TMD and fully disordered L-loop, and their limited dynamical coupling, are good arguments to conceptualise hVKORC1 as a modular protein. Moreover, cleaved L-loop and L-loop fused to TMD showed very comparable structural and conformational properties, as was confirmed by different simulation methods (cMD and GaMD) and conditions of simulation (in membrane or water solution only). In all studied cases, the fully oxidised hVKORC1 L-loop adopts a compact globular-like shape (closed conformation), the most prevalent and apparently the most energetically favourable conformation of L-loop in solution. Therefore, the role of L-loop hinge, which, according to [218], is partially embedded in the membrane and acts as an anchor, requires detailed study.

One approach for studying the assembly of multidomain proteins and their folding is to use the protein modular domain, which preserves binding capabilities even when the domain is removed from the full-length protein context[7]. Dynamic modularity displays clear sub-domain architectures that gives protein module enhanced flexibility and might influence on its ability to respond to the redox protein selection[364,395]. Modular protein domain's ability to independently fold and bind both *in vivo* and *in vitro* has been taken advantage of by a significant portion of proteomics studies to assess protein–protein interactions required for a diverse set of biological processes. L-loop application (empirically and numerically) as cleaved polypeptide represents a promising strategy for hVKORC1 thiol-disulphide exchange reactions study and modulation of protein function by the controlled interference of underlying molecular interactions. Such use is fully justified as we proved that hVKORC1 L-loop is a modular domain that preserves perfectly structural and dynamic properties upon its fused and cleaved status.

These considerations are valid for an enzyme given state performing its specific function(s). hVKORC1 principal function in fully oxidised state consists in recognition and binding of its redox protein leading to inter-protein thiol-disulphide exchange reactions. Most likely, this function is performed exclusively by L-loop. The next step, intra-hVKORC1 thiol-disulphide exchange reactions resulted in activation of the CXXC motif forming hVKORC1 active site, requires a tight cooperation between L-loop and transmembrane domain on all levels – structural, dynamical (structural/conformational transitions), physical (electron transfer) and biochemical (bond cleavage/formation). Vitamin K transformation in the active site can be realised without L-loop participation, while the last step of hVKORC1 catalytic cycle – restitution of the initial oxidase state – requires again L-loop contribution as a principal actor. Consequently, hVKORC1

structure can be conceptualised as a context-aware self-organising system providing global feedback to modulate and coordinate vitamin K transformation.

L-loop rich conformational space promoted by two different processes - transient folding and high flexibility - creates a serious problem in grouping L-loop conformations into clusters containing similar conformations. RMSD-based clustering methods, typically used to define prevalent protein structure, are dependent on proper cut-off between cluster groups to be effective [45]. This sensitivity to input parameters and low separation dimensionality, RMSD-based methods are not the best choice for analysis of disordered proteins that sample a large conformational space. Moreover, RMSD similarity calculated in Cartesian coordinate space, is sensitive to how the structures are aligned beforehand, and ignores real deviations within the backbone. Similarly, the secondary structure based hierarchical clustering local alignment algorithm[392,396] is not the best solution for disordered protein. Application of these two different approaches to L-loop clustering is not crowned with any success. The difficulty is to regroup disordered L-loop transient structures in a way that is reductive enough to provide required simplification while being flexible enough to accommodate a wide range of irregular structural configurations[397].

To avoid or reduce the problem, we decided to represent L-loop conformational space by reconstruction of its free energy landscape using as reaction coordinates the primarily descriptors – radius of gyration and RMSD. As molecular dynamics simulation of L-loop cleaved from distinct forms of hVKORC1 produced conformational subspaces which are partially overlapped, we suggested that their combining would be an appropriate approach to represent the generic L-loop conformational space, even if it is not yet complete. The free energy landscape, modelled on cumulative conformational space, specifies qualitatively the shape, folding, and dynamics of disordered L-loop and allows comparison between its different configurations. L-loop free energy landscape showed some local minima populated by ensembles of quasi-isoenergetic L-loop conformations typically seen in disordered proteins[398].

As we analysed hVKORC1 as a target of its redox protein, in particularly focusing on its ability to form protein–protein interactions mandatory for thiol-based redox switches, we concentrated on the central question – what conformation of hVKORC1 oxidised state is an authentic target of redox protein PDI?

Mechanisms of folding coupled to binding is poorly understood, but it has been hypothesized on theoretical grounds that binding kinetics may be enhanced by a 'fly-casting' effect, where the disordered protein binds weakly and non specifically to its target and folds when approaching the cognate binding site[399]. Consequently, ID protein capable of adapting to binding surfaces through coupled folding and binding[400] and the same binding region may have the capacity to bind several different partners with very similar affinities[401].

6.4. CONCLUSION SUR LA MODULARITE DU RTKI KIT ET DE HVKORC1

En réponse à notre question, « les régions fonctionnelles désordonnées peuvent-elles être considérées comme des modules de leur protéine respective ? », nous avons analysé les caractéristiques structurales et dynamiques de régions désordonnées du RTK KIT (KID) et de hVKORC1 (boucle L) en tant que polypeptides clivés de leur protéine respective libres dans le solvant ou, dans le cas du KID, également cyclisé. La comparaison de ces espèces clivées avec leur équivalent natif intégré dans leur protéine a montré leurs grandes similarités.

Par leur petite taille, l'utilisation des modules clivés, dans notre cas, de ces deux domaines intrinsèquement désordonnés, est une approche prometteuse pour les études *in silico* et *in vitro*. Ces cibles sont par conséquent plus faciles à modéliser pour la reconstruction des INTERACTOMES de KIT par le KID et de hVKORC1 par sa boucle L.

CHAPITRE 7. IDENTIFICATION ET CARACTERISATION DES PROTEINES PARTENAIRES DU RKT KIT ET DE HVKORC1

L'INTERACTOME d'une protéine est l'ensemble des complexes qu'elle forme avec des protéines partenaires. Les régions de recrutement d'une protéine par une autre encode à la fois un motif ou des motifs courts de reconnaissance pouvant comprendre des résidus altérés par une modification post-traductionnelle.

De fait, pour reconstruire un INTERACTOME, il est nécessaire d'obtenir des cibles de chacun des acteurs du complexe. Pour le RTK KIT, il s'agit du KID phosphorylé et du domaine SH2 de ses partenaires de signalisation. Pour hVKORC1, il s'agit de sa forme inactive (oxydée) et de sa protéine redox partenaire en forme réduite.

Dans un premier temps, on utilisera le module KID pour l'étude des effets de phosphorylation sur son désordre intrinsèque. Puis on caractérisera le domaine SH2 de l'un de ses partenaires physiologiques, la phosphatidylinositol 3-kinase (PI3K). Enfin, pour hVKORC1, nous allons identifier et caractériser la protéine redox la plus probable parmi quatre candidates décrites dans la littérature.

Ce chapitre est une adaptation des articles suivants :

1. **Ledoux, J.**, & Tchertanov, L. (2023). Site-Specific Phosphorylation of RTK KIT Kinase Insert Domain: Interactome Landscape Perspectives. *Kinases and Phosphatases*, *1*(1), 39–71. https://doi.org/10.3390/kinasesphosphatases1010005

2. Stolyarchuk, M.[+], **Ledoux, J.**[+], Maignant, E., Trouvé, A., & Tchertanov, L. (2021). Identification of the Primary Factors Determining the Specificity of Human VKORC1 Recognition by Thioredoxin-Fold Proteins. *International Journal of Molecular Sciences*, *22*(2), 802. https://doi.org/10.3390/ijms22020802

Les données supplémentaires et les méthodes relatives à toutes ces publications sont présentées dans les annexes de la thèse.

7.1. LE KID DU RTK KIT, UNE CIBLE DU DOMAINE SH2 DE LA PROTEINE DE SIGNALISATION PI3K

Résumé. *Le KID est un module intrinsèquement désordonné du RTK KIT et une région de recrutement clé de protéines partenaires (PPs) de la signalisation cellulaire médiée par ce récepteur. Phosphorylé à des résidus tyrosines spécifiques, il fournit des sites de reconnaissance pour les domaines SH2 de ses PPs et permet leur activation par le transfert de ce phosphate d'une protéine à l'autre. La simulation de dynamique moléculaire et l'étude comparative des*

effets de toutes les combinaisons de phosphorylation sur le KID clivé ont montré que chaque espèce maintient un désordre intrinsèque général, mais une dynamique spécifique à chaque combinaison de phosphorylation, que ce soit dans la corrélation des mouvements ou la distribution spatiale des tyrosines phosphorylées. La modélisation et la simulation de dynamique moléculaire de chacune de ces espèces du KID phosphorylés ont permis de générer un nombre conséquent de cibles du KID pour la description de l'INTERACTOME du KIT par ce domaine.

7.1.1. INTRODUCTION

Human cell-to-cell communication is monitored by signals from its environment, to which the cell must respond appropriately. Cell membrane receptors with tyrosine kinase activity (RTKs) frequently promote the external-internal exchange of information[65,338,382]. Once activated by its specific molecular stimuli, such as growth factors, these RTKs phosphorylate their downstream cytoplasmic substrates - adaptors, signalling and scaffolding proteins. Reversible protein phosphorylation provides a central regulatory mechanism in cells and represents a crucial step of post-transduction processes (PTPs)[402]. PTP involves many intrinsically disordered (ID) proteins that most contain phosphorylation sites[306,308,330,403].

KIT is a champion among RTKs regarding the number of ID regions and site-specific phosphotyrosines. As was reported early, the cytoplasmic domain of RTK KIT contains a tyrosine kinase domain (TKD) crowned by at least four ID regions/domain – juxtamembrane region (JMR), kinase insert domain (KID), activation (A-) loop and C-terminal tail, which are inherently coupled[271].

Each KIT ID region is a key regulatory element contributing to KIT activation and/or mediating protein-protein (PP) interactions through its functional phosphotyrosine residues. Two tyrosines, Y568 and Y570, identified *in vivo*, and two additional sites, Y547 and Y553, detected *in vitro*[119], provide bi-functional activity of JMR playing a regulatory role in the KIT activation/deactivation process as well as in the recruitment of adaptors, signalling and scaffolding proteins[37,404]. The KID of KIT likely only participates in such recruitment through protein partners' (PP) selective recognition and binding[272]. Multiple functional phosphorylation sites of KID, three tyrosine (Y703, Y721, Y730) and two serine (S741 and S746) residues provide alternative binding sites for intracellular PPs[287,405]. Phosphorylation of Y703 supplies the binding site for the SH2 (Src Homology 2) domain of Grb2 (Growth Factor Receptor-bound protein 2), an adaptor protein initiating the MAPK (mitogen-activated protein kinase) pathway[144,406]. Phosphorylated Y721 and Y730 are the recognition sites of phosphatidylinositol 3-kinase (PI3K) and phospholipase C (PLCγ), respectively[149]. The function of Y747 has not yet been described empirically. Considering its abundant intra-KID contacts, the 'organising role' of tyrosine Y747 in stabilising the KID structure was previously

attributed[385]. Phosphorylated serine residues, S741 and S746, bind protein kinase C (PKC) and contribute to the retro-control of PKC activity under receptor stimulation. The other phosphotyrosines in KIT are Y823 and Y900 from the TKD C-lobe, and Y936 from the C-terminal tail[155,407].

A high population of phosphorylation sites in the cytoplasmic domain of RTK KIT (in particular, 8 tyrosines) and their location in ID regions furnish extraordinarily sophisticated problems to study its post-transduction processes. To our great satisfaction, we found that RTK KIT is a modular protein, similar to many proteins of the human proteome[364]. Consequently, one part of post-transduction processes may be studied by using the individual domains of KIT, regarded in certain approximations as independent from the rest of the protein. This approach provides a promising route for using such domains as independent units in research and biotechnology of large multidomain proteins, particularly for studying PP interactions[12].

Many issues remain despite the apparent simplification of multidomain protein studies when using a per-domain approach for research (e.g., reduced molecular size).

The phosphorylation and binding to signalling proteins of each KIT domain having multiple functional phosphorylation sites is a great challenge[65,66]. First, the most archetypical question is: how does phosphorylation influence the structure and conformation of intrinsically disordered (ID) proteins? It was reported that PTPs induced significant changes in the structural and dynamical properties of ID proteins by affecting their energy landscapes[44,298,299]. PTPs cause a broad spectrum of effects, from local stabilisation or destabilisation of the secondary structure to global disorder-to-order transformations. Such structural and conformational events cause a complete change of a protein folding varied from an intrinsically disordered to well-folded structure or even a sporadic switch between monomeric and oligomeric states[388].

Secondly, for ID regions containing more than one phosphorylation site, the question arises about the required number of phosphorylated sites for a protein(s) binding. Focusing on the multi-site KIT KID, we are still unsure if single-site tyrosine phosphorylation is sufficient to create a signalling protein scaffold (a one-to-one process). It can be suggested that protein binding to multisite KID is a more collective multithreaded pre-process one to many, or many to one, or many to many. For example, a protein binding induces the required conditions for another KID tyrosine phosphorylation followed by binding of another protein specific to the newly created scaffold or a partner binding requires phosphorylation of two or more tyrosine sites at the KID. To inspect these hypotheses, it is necessary to consider many cases of phosphorylation events, which are described by the factorial function. Only for KID possessing three phosphotyrosines, the number of combinations analysed is 7.

The most sophisticated enigma, a bona fide Pandora's box, is understanding the phosphate binding order as a time-dependent sequence of the PTP events. Decoding

the intrinsic kinetics of multisite phosphorylation is key to understanding how multiple phosphorylated sites within a domain can collectively shape the transcriptional response. Characterisation of such phosphorylation kinetic should be based on biochemical and/or biophysical techniques, which may be applied to complex systems, including multiple phosphorylation and kinase cascades[408,409].

In our paper, to enrich the reliable knowledge of the KIT KID post-transduction processes, we first apply a systematic *in silico* approach to map the phosphorylation effects on the multisite KID possessing three functional tyrosines – Y703, Y721 and Y730. Such study should help to understand the phosphorylation-induced effects on KID and deliver the phosphorylated KID structures as suitable targets to probe signalling proteins binding. The structural and conformational changes induced by phosphorylation of the tyrosine residues (all possible combinations are considered) in the 80-residue KID of RTK KIT were investigated using 3D modelling and the extended molecular dynamics (MD) simulation.

7.1.2. Results

7.1.2.1. Modelling of the phosphorylated KIT KID

To model the alternatively phosphorylated KID entities, we used the cleaved KID from RTK KIT, previously studied as an unphosphorylated (native and inactive) species[271,290]. Looking ahead a bit, we have shown that KID from the inactive state of KIT is very similar to that in the active state (unpublished data, partially presented in Supplementary Information **Figure S25**). This similarity approves the use of the KID extracted from the RTK KIT in an inactive state, published in [271,385], as an initial model for probing phosphorylation of KID tyrosine sites. Seven 3D models of the phosphorylated KID (**Figure 7.1**) were generated from the KID conformation taken at t = 2 μs of MD simulation of the full-length cytoplasmic domain of KIT in the inactive state completed by its transmembrane helix inserted into a membrane[271].

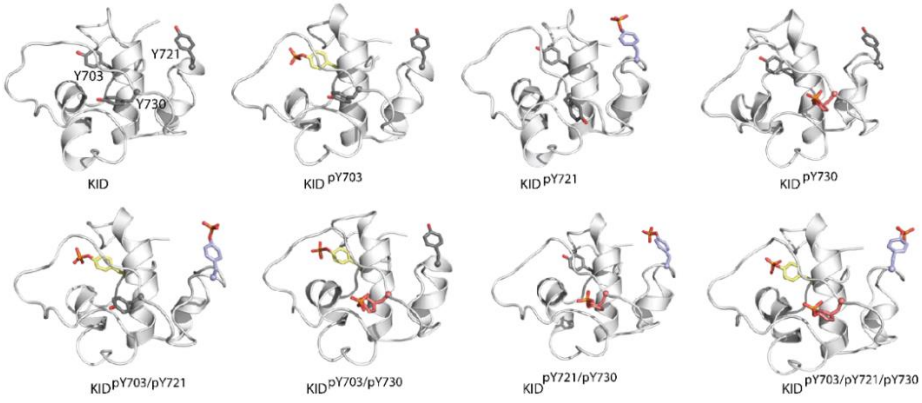


Figure 7.1 3D models of the phosphorylated KID. All models were generated from the KID

conformation taken at t = 2 μs of MD simulation of the full-length cytoplasmic domain of KIT in an inactive state[271] by phosphorylation of tyrosine residues with phosphate group -O-PO_3^{2-}. Protein is shown as a cartoon, and phosphorylated tyrosine residues Y703, Y721 and Y730 are shown as yellow, light blue and pink sticks, respectively.

7.1.2.2. General characterisation of the KIDs MD simulations

Each optimised and well-equilibrated phosphorylated KID (p-KID) model was studied by extended (2-μs) MD simulation running three times in strictly identical conditions using random initial velocity. First, the generated MD trajectories were characterised by conventional methods using a commonly used descriptors and compared between them and with those of the native KID.

The root means standard deviations (RMSDs) calculated on MD conformations from three replicas respective to the same initial structure of each p-KID show their good convergence (**Figure S26**). Compared to the ample variations of RMSDs in the native KID, the RMSD of p-KID entities is slightly reduced. Nevertheless, their profiles, except smoothers curves of KID^{pY721}, indicate significant conformational transitions, as was observed in unphosphorylated KID[290].

7.1.2.3. Structural and dynamical features of p-KIDs

Assignment of secondary structures (SS) with DSSP reveals a helical fold in all p-KIDs, like the native KID species (**Figure 7.2**, **A**; **Figure S27**). Except for H1-helix that in all studied KIDs is constantly folded as the α-helix varying only in length, the other helices are transient, reversibly converted between the folded and unfolded structures (α-helix ↔ 3_{10}-helix ↔ turn/bend), a phenomenon typical for IDPs. Depending on the unfolded residues fraction, the number and length of transient helices vary significantly within a MD trajectory of the same KID species and between them. Whether phosphorylated or not, KID contains from two to six helices formed by 20 - 35% amino acids (aas). The most folded helical structure is observed in KID^{pY721} (35%) and $KID^{pY721/pY730}$ (32%), while the less folded in KID^{pY703} (20%). The other phosphorylated KIDs showed a folding in the 23 – 25% range, which differs from the native KID (29%). The ratio of α/3_{10}-helices, varying from 2 to 5 times in the KIDs, except KID^{pY703} (by a factor of 21), indicates the apparent predominance of α-helices in all the KID studied. Furthermore, the doubly phosphorylated $KID^{pY721/pY730}$ shows two β-strands (D723 - V728 aas) appearing in 45% of the simulation time instead of its 30% lifetime in mono-phosphorylated KID^{pY721}. In another doubly phosphorylated KID, $KID^{pY703/pY730}$, a pair of short strands is observed in the D737-R740 fragment for 20% of the simulation time.

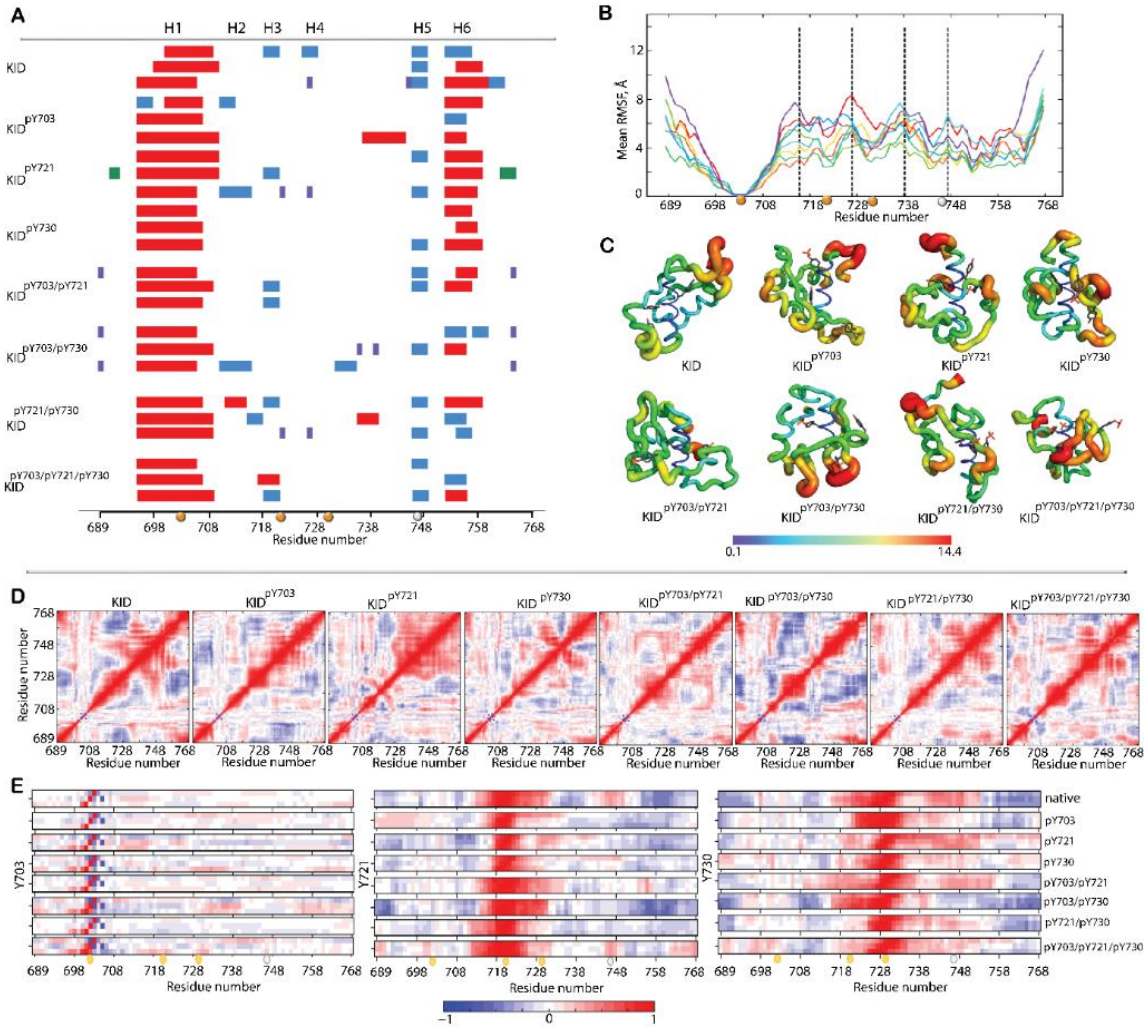


Figure 7.2 Structural and dynamical properties of unphosphorylated and phosphorylated KID. (**A**) The αH- (red), 3_{10}-helices (blue), and β-strands (violet and green), were assigned by DSSP on average conformation from each MD trajectory for every KID. The helical structures are labelled from H1 to H6. (**B**) RMSFs were computed on the Cα atoms of MD conformations after fitting on initial conformation (at t = 0 ns) and alignment on the best-conserved portion of αH1-helix (Y703 - L706). Different KID entities are distinguished by colour: KID (lilac), KID^{pY703} (red), KID^{pY721} (orange), KID^{pY730} (yellow), KID$^{pY703/pY721}$ (lime), KID$^{pY703/pY730}$ (green), KID$^{pY721/pY730}$ (turquoise), and KID$^{pY703/pY721/pY730}$ (blue). (**C**) The RMSFs of the average conformations for KIDs are presented as tubes. The tube size is proportional to the residue atomic fluctuations computed on the backbone atoms. Red–blue gradient shows the RMSF values, from large (> 14 Å, in red) to small. (**D**) Dynamical inter-residue cross-correlation maps computed for all Cα-atom pairs of MD conformations of concatenated trajectories of each KID. (**E**) Cross-correlations zoomed on the phosphotyrosines and their immediate environment after fitting on the αH1-helix. The blue-red gradient shows the correlations from -1 (blue) to 1 (red). The tyrosine positions are shown as balls: phosphotyrosines Y703, Y721 and Y730 in yellow, and Y747 in grey.

These observations illustrate the most apparent phosphorylation-induced effects on KID folding (secondary structures) in KIT entities possessing (i) the phosphorylation pY721 either as the single (KID^{pY721}) or double (KID$^{pY721/pY730}$) sites; and (ii) the phosphorylation pY730 in the doubly phosphorylated KID, KID$^{pY703/pY730}$ and

$KID^{pY721/pY730}$. In both cases, phosphorylation either increases the overall KID folding, as evidenced in (i) or promotes additional folding in only its immediate environment, manifested in (ii), while the overall folding is diminished.

While a significant portion of p-KID residues is not folded in regular structures, they all exhibit high flexibility, as observed in unphosphorylated KID [10, 20, 34]. To rationally compare the phosphorylation-induced effects on the KID flexibility, the root means square fluctuations (RMSFs) were calculated on MD conformations from the concatenated replicas of each p-KID respective to its initial structure (at t = 0 μs) and fitted on the best-conserved portion (Y703 - L706) of αH1-helix. This approach characterises the p-KID RMSF variations respective to the most conserved KID structural element – αH1-helix.

The RMSF curves and tubes defined on the average conformations of KID and p-KIDs demonstrate (i) reduced RMSF values of N- and C-terminals residues in all p-KIDs relative to KID; (ii) partial re-distribution of the most and least fluctuating fragments along the KID sequence (**Figure 7.2**, **B**, **C**; **Table S6**).

For example, the highly fluctuating fragment near D716 in the native KID shows reduced RMSF values in all p-KIDs. In contrast, the moderate fluctuations of KID segment G727-V728 significantly increased in KID^{p703}, showing RMSF values up to 8 Å. The minimally fluctuating regions vary little between different p-KIDs and relative to the native KID in terms of their position at the KID sequence and RMSF values.

In general, KID^{p703} and $KID^{pY721/pY730}$ are the most flexible p-KIDs, while $KID^{pY703/pY721}$ and $KID^{pY703/pY730}$ are more 'rigid' entities. After a detailed comparison of all studies KIDs, we noted that the minimally fluctuating fragments (RMSF value ≤ 3 Å) are M722-M724 (in $KID^{pY703/pY721}$, $KID^{pY703/pY730}$ and $KID^{pY721/pY730}$), V732 (in $KID^{pY703/pY730}$), R743-G745 (in KID^{pY730} and $KID^{pY703/pY721}$), and T753-P754 (in KID, KID^{pY721}, KID^{pY730} and $KID^{pY703/pY721/pY730}$). These minimally fluctuating segments of p-KID regions are involved in non-covalent intramolecular interactions contributing to stabilising their tertiary structures, as observed in the native KID[290].

The maximally fluctuating fragments (RMSF value ≥ 6 Å) are D716 (in KID), P726-V728 (in KID^{pY703}, KID^{pY730} and $KID^{pY703/pY721/pY730}$), K738-R739 (in KID, KID^{pY721}, KID^{pY730}, $KID^{pY721/pY730}$, $KID^{pY703/pY721/pY730}$) and I748 (in $KID^{pY721/pY730}$). The fluctuations of those regions are increased by at least 2 Å compared to the native KID, suggesting an enhancement of conformational flexibility upon phosphorylation.

These results evidenced the different degrees of flexibility of the p-KIDs, containing either minimally or maximally fluctuating fragments relative to αH1-helix, taken as a reference. Curiously, these fragments do not contain tyrosine residues.

The inherent dynamics of the intrinsically disordered KIDs were analysed with the cross-correlation matrix computed for the all Cα-atom pairs of each KID. The Cα–Cα

pairwise cross-correlation maps demonstrate highly coupled motions between the KID structural fragments, even vastly distant (**Figure 7.2**, **D**). The maps are highly different for the KIDs studied, showing either the block-like patterns composed of clearly delimited correlated segments (KID, KID^{pY703} and $KID^{pY703/pY730}$) or very smooth patterns (KID^{pY721}, KID^{pY730}, KID $^{pY721/pY730}$ and $KID^{pY703/p721/pY730}$). Focusing on the phosphotyrosines, we note that the motion of pY703 is correlated with only a minimal number of residues (2-3 amino acids, aas) close to Y703 at the KID sequence. These correlations are positive and negative with preceding and following residues, respectively. This suggests that Y703 acts as a pivotal kerner of αH1 helix. Apart from this region, Y703 doesn't show any strong correlation with the rest of KID residues.

In contrast, Y721 and Y730 show strong correlations with many residues close to these sites (strong positive correlations) or distant residues (strong or moderate negative correlations). Surprisingly, the number of neighbouring (sequence position) correlated residues with Y721 is decreased on mono-phosphorylated KID or moderately shifted when phosphorylated with pY730. Concerning Y730, the correlated region is larger than Y703 and Y721 (up to 50% of residues). In KID^{pY703}, Y721 correlates positively with well-defined proximal residues in terms of sequence position when positive correlations are more diffused in other KIDs. This pattern is clearly reduced in KIDs phosphorylated at Y721.

7.1.2.4. SHAPE OF P-KIDS AND THEIR STABILISATION BY H-BONDS

The p-KIDs size described by the radius of gyration (Rg) is nearly similar in most studied proteins (**Figure 7.3**, **A**). Statistically, the Rg shows the Gaussian unimodal distribution except for the bimodal curve for KID with three phosphorylated sites – $KID^{pY703/p721/pY730}$. The most probable Rg values vary in the narrow range, from 11.7 to 13.5 Å, with a mean value (mv) of 12.5 Å for all p-KIDs. Only mv of Rg for KID^{p703} and $KID^{pY703/p721/pY730}$ (main peak) is slightly diminished (12.2-12.3 Å). The Rg values of p-KID correspond well to those observed in the native KID. Also, like native KID, the p-KID conformations are stabilised by many H-bonds that maintain their globular-like shape (**Figure 7.3**, **B**, **C**), similar to that in unphosphorylated KID[290].

In all p-KIDs, the phosphotyrosines are located on the protein surface, and their phosphorylated side chain is well exposed to the solvent or proteins. However, pY703 is systematically involved in H-bonds by its phosphate oxygen atoms acting as acceptors. Indeed, two oxygen atoms of pY703 form the salt bridge with R740 in mono-phosphorylated (KID^{pY703}), bi-phosphorylated ($KID^{pY703/pY721}$ and $KID^{pY703/pY721}$) and three-phosphorylated entities (**Figure 7.3**, **D**). A geometrically favourable contact was considered an H-bond if its occurrence was more than 30%, it suggests that many p-KID conformations having pY703 may have an abundant number of alternative orientations disfavouring pY703 intra-KID H-bonds. Other phosphotyrosines sidechains of holds their oxygen atoms available for post-transduction events.

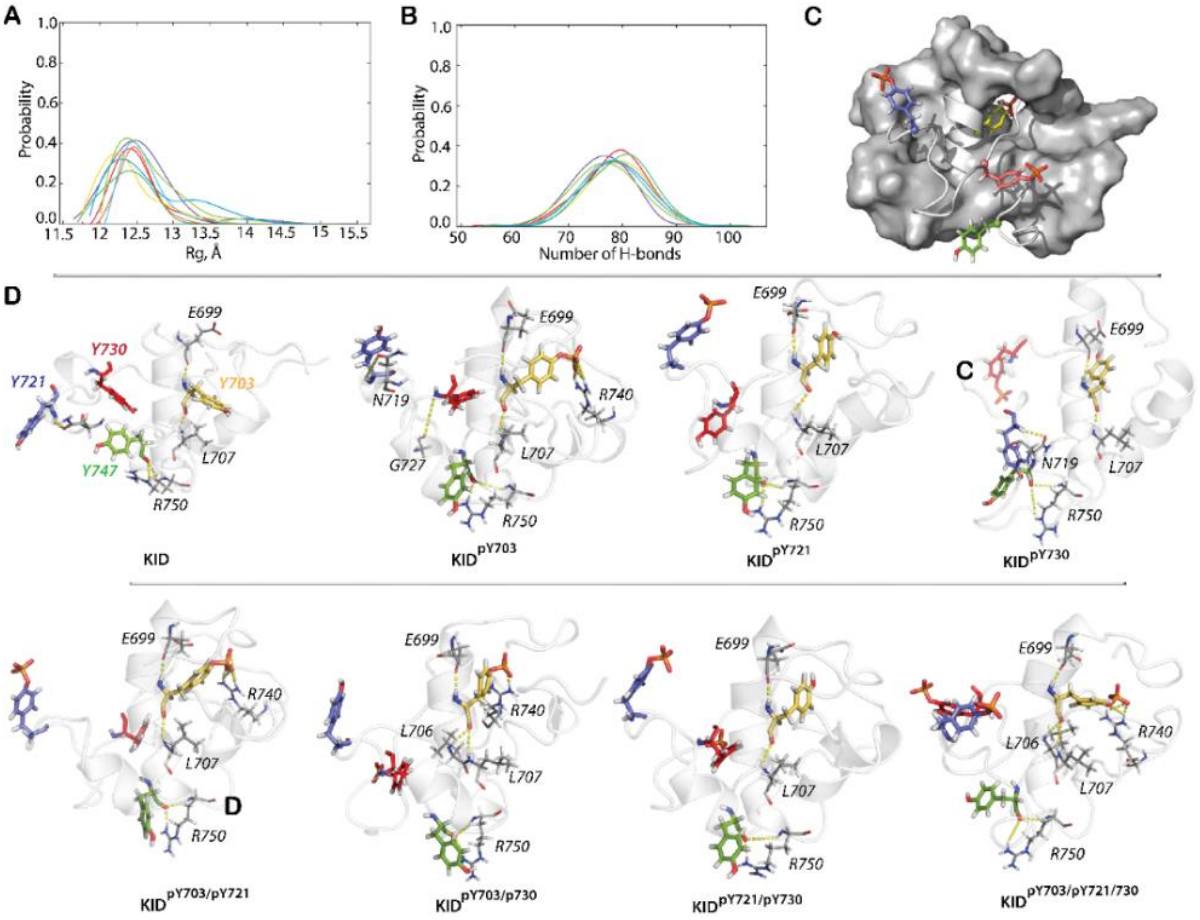


Figure 7.3 The KIDs shape and its stabilisation by H-bonds. Radii of gyration (Rg) (**A**) and number of H-bonds stabilising KIDs conformations (**B**) that maintain their globular-like shape (C). Different KID entities are distinguished by colour: KID (lilac), KID^{pY703} (red), KID^{pY721} (orange), KID^{pY730} (yellow), KID$^{pY703/pY721}$ (lime), KID$^{pY703/pY730}$ (green), KID$^{pY721/pY730}$ (turquoise), and KID$^{pY703/pY721/pY730}$ (blue). In (**C**) protein is shown as a surface cartoon, and tyrosine residues as coloured sticks: Y703 in yellow, Y721 in blue, Y730 in red and Y747 in green. (**D**) H-bonds involving tyrosine residues. H-bonds were calculated for the contacts D-H···A with distance D···A ≤ 3.6 Å and pseudo-valent angle DHA ≥ 120° observed with an occurrence ≥ 30%. In (C, D) Proteins are shown as cartoons, tyrosine residues as large coloured sticks and contacting residues as think grey sticks. Dashed yellow lines display H-bonds. The numbering of tyrosine residues is shown on the KID.

7.1.2.5. THE P-KIDS SOLVENT ACCESSIBILITY

Analysis of the total Solvent Accessibility Surface (SASA) of the unphosphorylated KID and its phosphorylated derivatives shows the unimodal distribution for all studied entities, which differed slightly only by the peak height (85-100%) (**Figure 7.4**, **A**). The SASA calculated for each phosphotyrosine individually in each p-KID entity systematically showed two peaks at proximity to 160 and 200 Å^2. The Y703 smaller value of SASA (at 162 Å^2) is observed in KID^{pY721}, KID^{pY730} and KID$^{pY721/pY730}$ while the second peak (at 205 Å^2) in KID^{pY703}, KID$^{pY703/pY721}$, KID$^{pY703/pY730}$ and KID$^{pY703/pY721/pY730}$. The Y721 smaller SASA value (at 160-162 Å^2) is observed in KID^{pY703}, KID^{pY730} and

$KID^{pY703/pY730}$, the second peak (at 203 $Å^2$) in KID^{pY721}, $KID^{pY703/pY721}$, $KID^{pY721/pY730}$ and $KID^{pY703/pY721/pY730}$. For Y730 the smaller SASA values is observed in KID^{pY703}, KID^{pY721} and $KID^{pY703/pY721}$, while the second peak (at 200 $Å^2$) is observed in KID^{pY730}, $KID^{pY703/pY730}$, $KID^{pY721/pY730}$ and $KID^{pY703/pY721/pY730}$.

The main conclusions of these observations are as follows: (i) the global SASA is almost identical in all the p-KIDs studied with slight variations in the distribution height, a minimum (82%) in $KID^{pY703/pY721}$ and a maximum (98 %) in $KID^{pY703/pY721/pY730}$; (ii) SASA of each tyrosines shows two peaks; (iii) each smaller peak, corresponding to unphosphorylated tyrosines, is equivalent to the native KID's, suggesting unchanged availability to the solvent of those tyrosine; (iv) the second peak of each tyrosine and with the most significant SASA value is composed only of its phosphorylated derivatives; (v) the 40 $Å^2$ increment between two SASA peaks is consistent for all p-KIDs studied, and independent of the number of phosphorylated sites (mono-, bi- and tri-phosphorylated); (vi) SASA of unphosphorylated Y747 shows a near-equivalent unimodal distribution in all proteins studied.

The tyrosines' representative atoms' spatial distributions – the oxygen atom from the tyrosine OH group from the native KID, and the phosphorus atom in p-KID, superposed on the respective average conformations of the native KID and its phosphorylated derivatives, are very extended and show an oblate spherical sector shape, frequently subdivided into two regions for the same tyrosine (**Figure 7.4**, **C**). Such distributions, often overlapping between the different tyrosines, demonstrate clearly that the SASA of tyrosine residues of the proteins studied is (i) divergent in terms of surface area, spatial position, and orientation and (ii) suggests a great diversity of MD conformations in which the phosphotyrosine residues are either well-exposed to the solvent or this access is limited by the steric constraints induced by the phosphotyrosine neighbouring residues' environment.

To illustrate such cases, we represented the accessible surfaces for the tyrosine residues traced out by the van-der-Waals surfaces of the water molecules atoms when in contact with the protein. This alternative approach to estimating the possible contact surface is coherent with the SASA method. Traced on each phosphorylated tyrosine, the contact surface displays the systematic accessibility of the tyrosine residues sidechain to the solvent, surprisingly whether phosphorylated or not, except for unphosphorylated Y703 in the native KID and in doubly phosphorylated $KID^{pY721/yY730}$, where its accessibility is constrained, and correlating with its small SASA value (at 162 $Å^2$) (**Figure 7.4**, **A**, **B**). Once Y703 is phosphorylated, its surface is accessible to the solvent, similarly to the other KID phosphotyrosines. Such availability of each phosphotyrosine is necessary for the recognition of KID linear motifs by protein partners, phosphate transfer and signal transduction events.

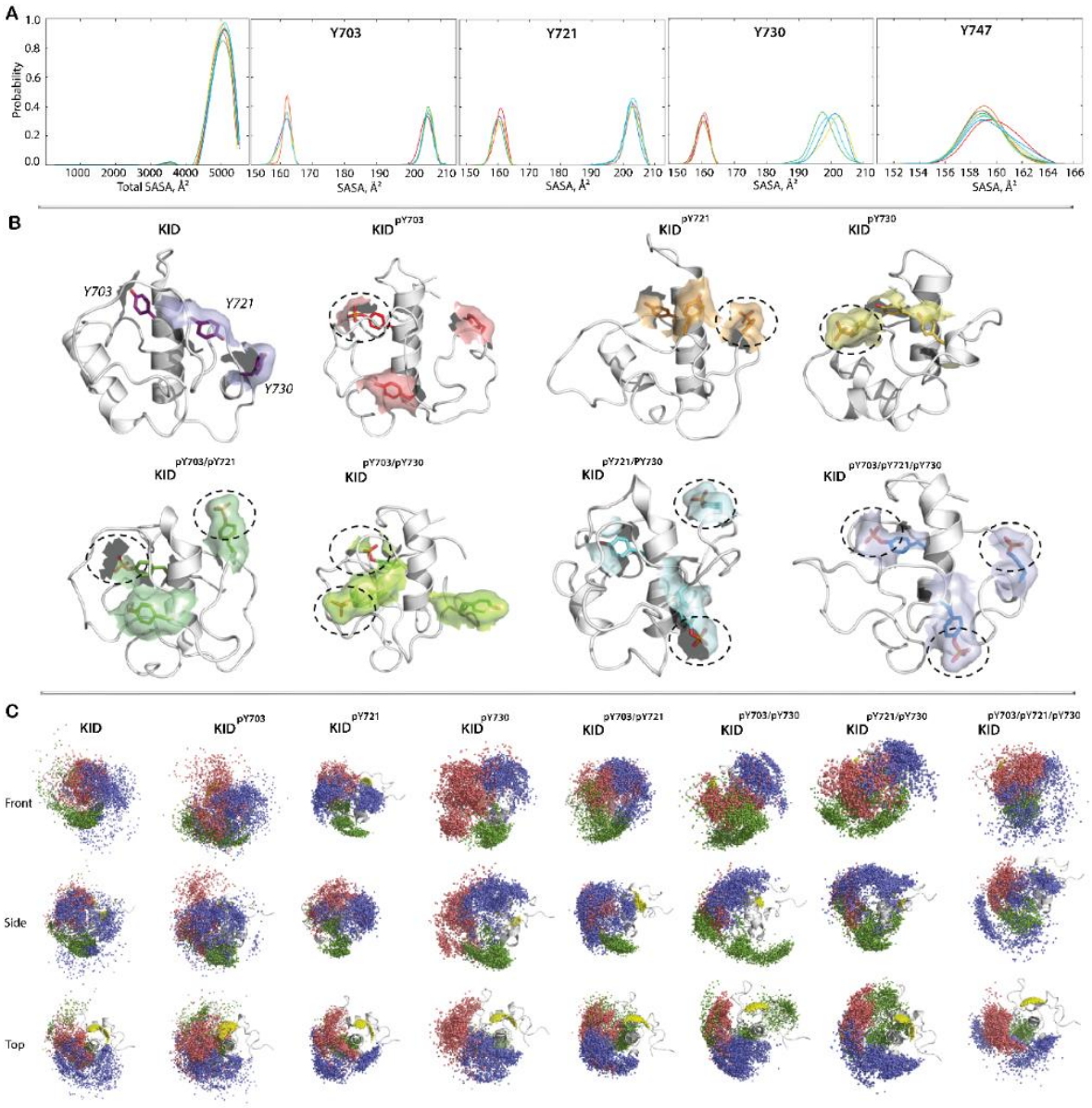


Figure 7.4 The solvent accessibility surface of phosphotyrosines and their spatial distribution. (**A**) The SASA probability distribution of total KID and each of its tyrosine residues. Different KID entities are distinguished by colour: KID (lilac), KID^{pY703} (red), KID^{pY721} (orange), KID^{pY730} (yellow), KID$^{pY703/pY721}$ (lime), KID$^{pY703/pY730}$ (green), KID$^{pY721/pY730}$ (turquoise), and KID$^{pY703/pY721/pY730}$ (blue). (**B**) The solvent/protein contact surface is traced by the van de Waals surfaces of water molecules in contact with the protein. Dashed ovals contour the contact surfaces of the phosphotyrosine in each p-KID. The numbering of tyrosine residues is shown in KID. (**C**) Spatial distribution of KID tyrosines (after a fitting of conformations taken each 2 ns from the concatenated trajectories on Y703-L706, t = 0 μs) are presented by the oxygen atom of hydroxyl-group or phosphorus atoms for unphosphorylated and phosphorylated tyrosine residues respectively. Three orthogonal projections with atoms of KID, Y703, Y721, pY730 and Y747, respectively, coloured in yellow, blue, red and green.

7.1.3. DISCUSSIONS

As the phosphate is an effector, its covalent binding to KID may promote significant effects on this intrinsically disordered domain. It was earlier reported that multisite phosphorylation gradually tunes the affinity of the inhibitor SIC1 for the disordered cyclin-dependent kinase CDC4[410], and a similar mechanism had earlier been suggested for the beta-catenin/E-cadherin complex[411].

Our systematic study of phosphorylation-induced effects on KIT KID showed that the one-site, two-site, and three-site phosphorylation considerably affected the KID structure. We evidenced it as (i) an increase in the overall p-KID folding or additional folding in the phosphorylated tyrosines' immediate environment, and (ii) an alternation of p-KID's flexibility. However, regardless of the phosphorylation site position and the number of phosphorylated tyrosine residues, all p-KIDs are intrinsically disordered entities holding the inherent coupled dynamics specific for each p-KID. Focusing on the phosphorylation sites, their motions are also site-specific. In particular, Y703 motion is correlated with only a minimal number of the sequence adjacent residues, positively and negatively with preceding and following residues, respectively. In contrast, Y721 and Y730 show strong correlations with many residues located close to Y721 or Y730 (strong positive correlations) or distant residues (strong or moderate negative correlations). Despite evident structural and dynamical alterations in p-KIDs, their globular-like shape, maintained by an extended H-bonds network pattern, is universally conserved. The solvent-exposed phosphotyrosine residues' position in all p-KIDs—a primary determinant of local residue flexibility[412]—and their sidechains extended spatial distribution allow each phosphate group availability for post-transduction events with high probability.

This modelling of the phosphorylated KID generates an exhaustive number of well characterised targets, which opened a route for describing KID INTERACTOME.

The tyrosine phosphorylation of RTK KIT generates the necessary conditions to recruit and activate downstream signalling proteins through pY binding to their SH2 domains. Consequently, we focused on phosphatidylinositol-3 kinase (PI3K) N-terminal SH2 domain, which preferentially binds to KIT KID via the phosphorylated tyrosine pY721[413,414].

7.2. PI3K, UN PARTENAIRE DE SIGNALISATION DU RTK KIT

Résumé. *La sous-unité régulatrice p85 de la phosphatidylinositol 3-kinase (PI3K) est composée, entre autres, de deux domaines SH2 N- et C-terminaux spécialisés dans la reconnaissance et la liaison de tyrosines phosphorylées. Le module N-terminal est spécifique, dans le cas de KIT, de KID phosphorylé en*

Y721. Le transfert de phosphate du KID^{pY721} à p85 entraine une réponse protéique de PI3K et l'activation de la voie de signalisation du même nom. La simulation de dynamique moléculaire de SH2 libre, issu d'une structure cristallographique de ce domaine cristallisé avec un peptide du KID^{pY721}, a montré un repliement de feuillets β entre deux hélices α stables, et la présence de trois régions désordonnées hautement flexibles. Nous avons constaté que deux de ces régions (boucles) portent les résidus conservés de reconnaissance du groupement phosphate et suggèrent une grande adaptabilité de SH2 lors de la reconnaissance de KID^{pY721}. L'ensemble des données générées a permis de délivrer les cibles de p85 pour la génération du complexe KID^{pY721}/SH2, le premier prototype des complexes moléculaire composant l'INTERACTOME du KIT via KID^{pY721}.

7.2.1. INTRODUCTION

Tyrosine phosphorylation controls RTK KIT cell signalling through the recruitment and activation of proteins involved in downstream signalling pathways, mediated through pY binding of the SH2 (Src Homology 2) and/or PTB (phosphotyrosine-binding) domains of signalling proteins[120,382,415]. SH2 and PTB domains typically recognise a pY residue within a specific amino acid sequence context.

In particular, the SH2 domain of human phosphatidylinositol 3 kinase (PI3K) has been reported to bind preferentially to pYφXφ (where φ is a residue with a hydrophobic side chain and X is any residue)[416]. Therefore, the specificity of SH2 domains is conferred to some extent by binding various pY-containing motifs.

7.2.2. RESULTS

7.2.2.1. AVAILABLE CRYSTALLOGRAPHIC STRUCTURE RELATED TO THE KID BINDING: ANALYSES AND HYPOTHESIS

The Protein Data Base (PDB) was searched for the KIT signalling proteins focusing on KID-specific partners to have an empirically determined benchmark structure. We identified a co-crystallized molecular complex (PDB ID: 2IUH, 2.0 Å resolution)[417] composed of a fragment of the signalling protein PI3K and a phosphopeptide *TNEYMDMK* (p-pep) of the KIT KID (residues T718-K725), including the phosphorylated tyrosine pY721. To develop the first PP complex, we choose PI3K, which is an intracellular signal transducer enzyme specific to KID of KIT[120]. It contains a catalytic subunit of 110 kDa (p110) and a small regulatory subunit of 85 kDa (p85). The PI3K p85 subunit contains two SH2 domains, an N- and a C-terminal SH2 domains, that

bind to two closely spaced pYXXM motifs (pY is the phosphorylated tyrosine; X is any amino acid; M is Met). The p85 is phosphorylated during the binding of PI3K to the RTK KIT and detached from p110, which then phosphorylates the phosphatidylinositol 4,5-biphosphate (PIP_2) of the plasma membrane[418].

The structure of molecular complex 2IUH includes the p85 (N-term) sequence G321-D440 with a single point mutation Q330N. Further this domain will be referenced as SH2 for simplicity. It represents an archetypical structure of the SH2 domain consisting of a central antiparallel β-sheet formed by three or four β-strands (β1-β4) flanked by two α-helices (αH1 and αH2) (**Figure 7.5**, **A**, **B**).

The KID p-pep is located on surface of the SH2 binding pocket formed by the central β-sheet and α-helices residues (**Figure 7.5**, **B**, **E**). The positively charged and polar residues from the β-sheet combine two quasi-symmetrical pocket surfaces in respect to the p-pep main axis, while the polar, positively and negatively charged residues from α-helices surround the N- and C-terminals of p-pep. Such SH2 surface is highly favourable to the multiple non-covalent interactions with each p-pep residue without exception, encouraging the sandwich-like position of the p-pep and SH2 domain (**Figure 7.5**, **C**, **D**).

First, the pY721 of p-pep is engaged in hydrogen (H-) bonds and ionic bonding with R340, R358 and S361 of the SH2 domain, forming salt bridges. Second, its nearest negatively charged residues, E720 and D723, form H-bonds with K379 and N417. The polar N719 contributes to the H-bond pattern with its homologs, N344 and N378, acting as an acceptor and donor, respectively. The H-bond pattern of p-pep is completed by the main chain interaction involving T718. Finally, the bifurcate van-der-Waals contact formed by the SH2 M722 adds to the tight p-pep binding to the domain. The peptide position in the SH2 binding pocket and the H-bond pattern maintaining it correspond to the canonical mode of SH2 binding[419].

The overall binding pocket of SH2, defined with Fpocket[420] is more significant than its surface occupied by p-pep (**Figure 7.5**, **F**) and can accommodate the larger KID fragment or retain p-pep in alternative orientation concerning the X-Ray pose. Following these suggestions, we emphasize the p-pep features. It should be noted that the short p-pep (TNEYMDMK) contains three residues preceding pY721 and four residues following it. These residues show similar biophysical properties making the p-pep almost symmetric concerning its recognition properties. Moreover, in a solvent, the polypeptide p-pep *per se* is a highly flexible entity. While the SH2 binding pocket also demonstrates symmetry in its surface residues which form two pairs of areas showing similar biophysical properties (**Figure 7.5**, **E**), we suggested that (i) the p-pep/SH2 complex in solution may differ from its conformation in the solid state; (ii) the p-pep may be having an alternative orientation in the SH2 binding pocket in respect to X-Ray pose, and (iii) KIT KID docking into PI3K SH2 may also have the alternative solutions. Such suggestions prompt us to study the p-pep/SH2 complex in

solution.

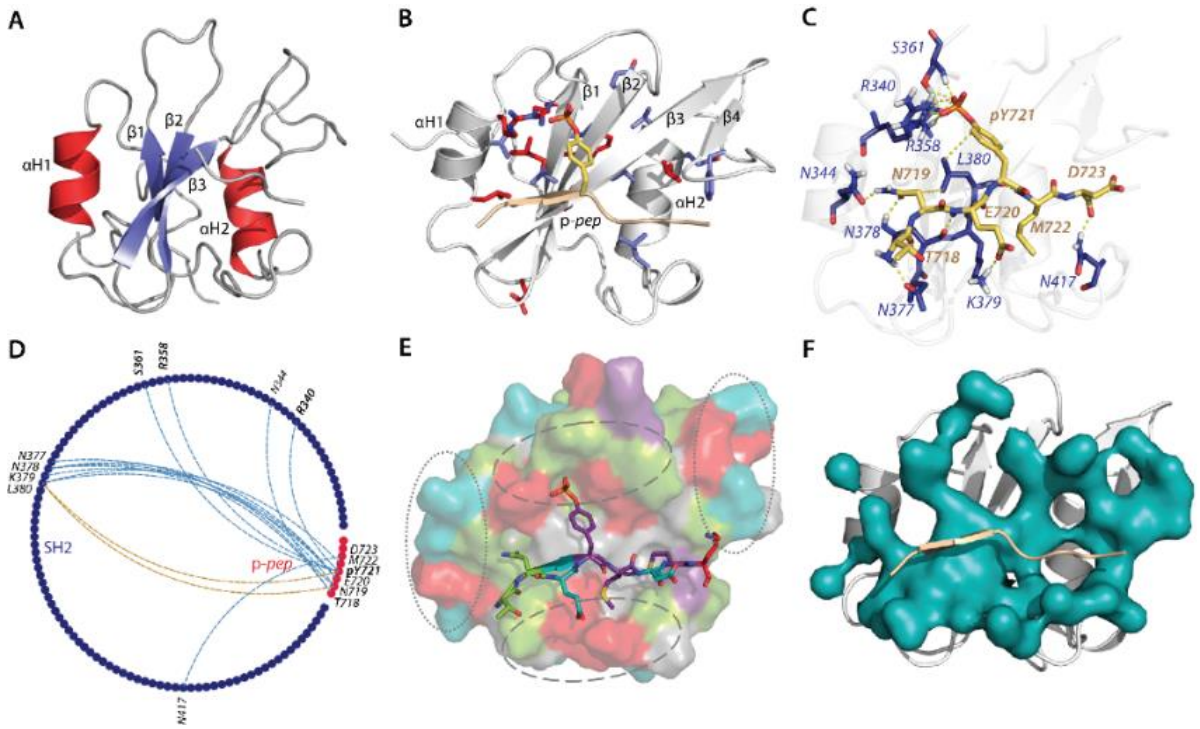


Figure 7.5 The crystallographic structure of SH2 domain and its characterisation. (**A**) The general view on structure of free-ligand SH2 domain of p85 (PDB ID: 1PIC). Protein is shown as a cartoon: α-helices, β-strands and coils are displayed in red, blue and grey respectively. (**B**) Structure of co-crystallized molecular complex composed of the SH2 domain fragment of PI3K and a phosphopeptide TNEYMDMK (p-pep) of the KIT KID. SH2 and p-pep are shown as a cartoon in grey and beige, respectively. The highly conserved (identical) residues and residues similar by physicochemical properties among the SH2 domains are in red and blue, respectively. (**C**) The non-covalent interactions in the molecular complex p-pep/SH2. (**D**) H-bonds (in blue) and hydrophobic contact (in beige) stabilising the p-pep (in red) and SH2 domain of PI3K (in darkblue) showed as a string diagram. (**E**) Surface representation of the p-pep/SH2 binding pocket with the p-pep (in sticks), stained by physicochemical property of amino acids: the positively and negatively charged are in red and blue respectively, polar in green, amphiphilic in purple, and hydrophilic in grey. Two pairs of areas delimited by pointed and dashed lines are formed by residues displaying similar physicochemical properties. (**F**) Definition of the SH2 binding pocket by Fpocket.

7.2.2.2. MOLECULAR DYNAMICS SIMULATIONS OF THE P-PEP/SH2 MOLECULAR COMPLEX

To study the dynamical properties of the p-pep/SH2 complex, we used the extended (2 μs) MD simulation. As the three replicas of simulations showed the remarkable similarity of RMSD and RMSF profiles and values (**Figure 7.6**, **B**-**D**), further analysis was performed on the concatenated trajectories for the time-dependent or time-independent metrics.

The stable and low RMSD values (1.5 - 2.0 Å) observed in the SH2 domain during all replicas are typical for excellent simulation convergence. The higher RMSF values are kept for the residues located in the loop between β1 and β2 strands, which is a part of the SH2 binding pocket accommodating pY721 from p-pep in structure 2IUH. The p-pep residues from N- and C-terminals fluctuate significantly. Such increased flexibility of the SH2 and p-pep facilitates the mutual adaptation of the two entities.

The folding (2D) and tertiary (3D) structure of the p-pep/SH2 complex is very similar between the MD replicas and during a replica (**Figure 7.6**, **A**, **E**). Congruent to the empirical structure 2IUH, the MD conformations of the SH2 domain consist of the central 'core' formed by three β-strands, β1 (G353-R358), β2 (Y368-R373) and β3 (N378-F384), organised in an anti-parallel sheet, and two highly conserved α-helices, αH1 and αH2. The central axis of these α-helices is nearly parallel. As it was observed in the crystallographic structure 2IUH, the MD conformations of the SH2 domain contain the second antiparallel sheet formed by two short β-strands, β4 and β5, well conserved during the MD simulations.

Apart from these well-organised and highly conserved structural units during the MD simulation, we localised in the SH2 domain two intrinsically disordered regions (IDRs) – F1 and F2. These IDRs, one positioned between β1 and β2 stands (F1) and the other between αH2-helix and β5 (F2), manifest a structural disorder, evidenced by transient structures reversibly converting between the folded and unfolded states (3_{10} helices ↔ β-strand ↔ turn ↔ coil).

The cross-correlation matrix computed on the concatenated trajectories of SH2 shows a very smooth pattern which partially reflects the expected coupling motion between the β-stands within a 'core' structure, and the relative rigidity of SH2 during the simulation (**Figure 7.6**, **I**).

The principal component analysis (PCA) demonstrates that the first six modes describe 80% of the motions' variance of the p-pep complex, with the 1st and 2nd modes explaining only 30 and 20% of the movement, respectively (**Figure 7.6**, **F**). The major contributor to these modes is the p-pep showing a great degree of dynamical disorder that is displayed as the largest displacements of its N- and C-terminals in mutually perpendicular directions, (**Figure 7.6**, **G**). Nevertheless, the central part of p-pep conserved its position in the binding pocket of SH2 and demonstrated only a slight motion. Such p-pep location is stabilised by highly conserved salt bridges formed by pY721 with R340, R358 and S361 of the SH2 domain, as was observed in the X-ray structure 2IUH. The other multiple H-bonds and van-der-Waals contacts, stabilising the p-pep in 2IUH, disappeared during MD simulation and only bifurcated H-bond E720···L380···M722 is maintained with the occurrences ≥ 50% (**Figure 7.6**, **H**), but interactions with M724 of the pYXXM recognition motif were only transitory. The second significant contributors to the PCA modes are the IDRs of SH2, F1 and F2, which showed a pendulum-like movement with a relatively large amplitude in the direction

orthogonal to the axis of α-helices.

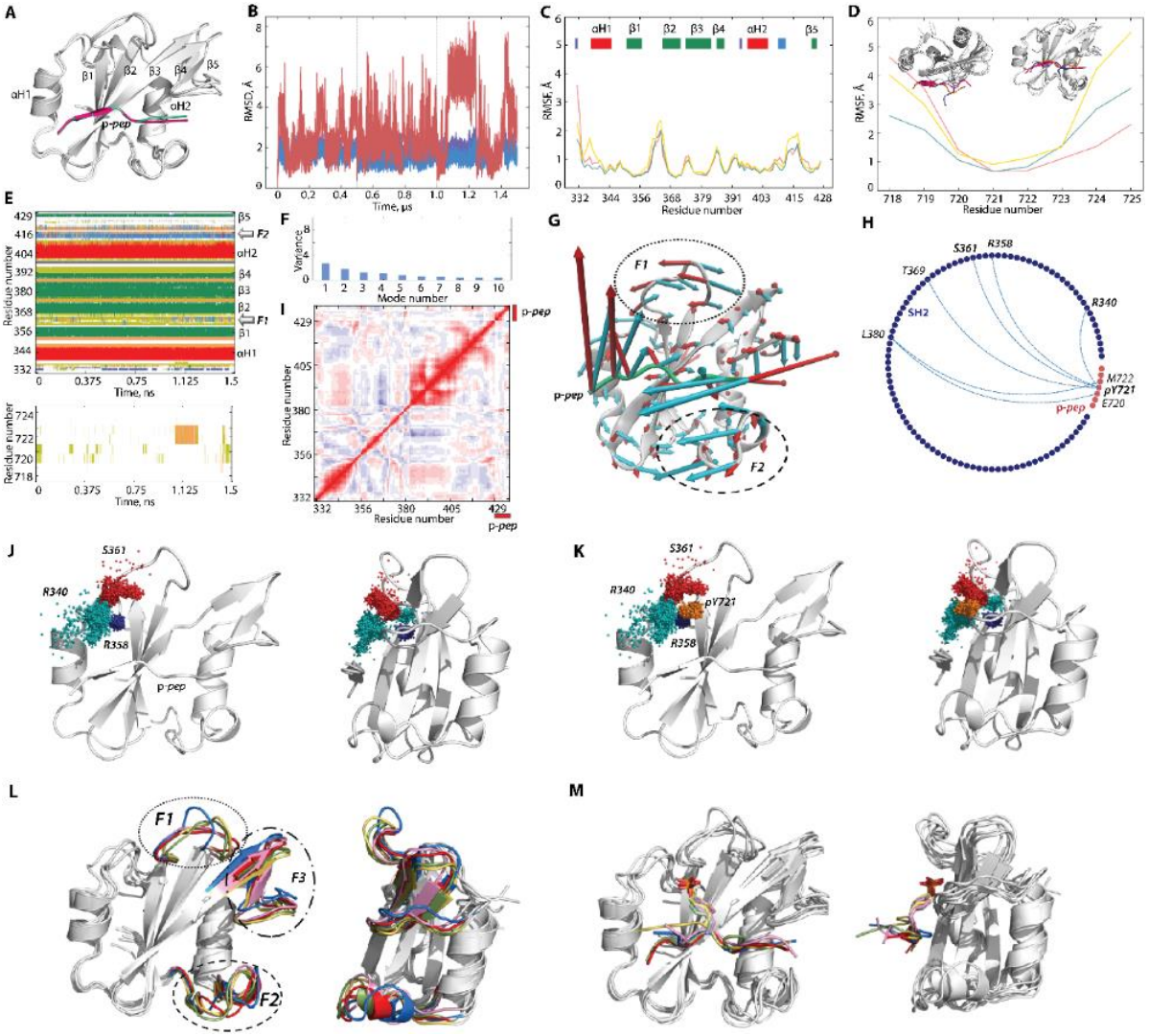


Figure 7.6 Structural and dynamical properties of the p-pep/SH2 complex. (**A**) Superimposition of the equilibrated p-pep/SH2 complex at t = 0 μs (p-pep in pink) on the X-Ray structure 2IUH (p-pep in cyan). (**B**) RMSDs are computed on the Cα atoms after fitting on the 'core' β-sheet (G353-R358, Y368-R373, N378-F384) of the initial conformation (at t = 0 ns). The whole complex is coloured in purple, p-pep in red, and the SH2 domain in blue. (**C**) Each MD trajectory RMSFs were computed on the Cα atoms of the SH2 domain after least-square fitting on the 'core' β-sheet of the initial conformation (at t = 0 ns) identical for each trajectory. The folding of SH2 mean conformation is shown in the insert. (**D**) RMSFs computed on the Cα atoms of p-pep after fitting on the initial conformation (at t = 0 ns) identical for each trajectory. The insert shows two conformations illustrating the large RMSF values for its N- and C-terminals. (B-D) MD replicas 1-3 are distinguished by colour (red, yellow, and blue). (**E**) Time-dependent evolution of each residue secondary structure as assigned by DSSP method for SH2 domain (top) and p-pep (bottom): α-helices are in red, 3_{10}-helices in blue, β-strand in green, turn in orange, and bend in dark yellow (replica 2). (F) SH2 PCA modes calculated for the concatenated MD trajectory (replicas 1-3) after least-square fitting of the MD conformations on the 'core' β-sheet (αC-atoms) of the initial conformation (t= 0 ns). The bar plot gives the eigenvalue spectra in descending order for the first 10 modes. (**G**) Atomic components in the two first PCA modes of the SH2 domain are drawn as red (1st mode) and blue (2nd mode) arrows projected onto the average structure. Only motion with an

amplitude ≥ 4 Å is shown. SH2 is in grey, and p-pep is in green. (**H**) H-bonds (blue dashed lines) stabilising the p-pep (in red) and the SH2 domain of PI3K (in blue) showed as a string diagram. Only the contacts with an occurrence ≥ 50% were taken into consideration. (**I**) Dynamical inter-residue cross-correlation maps computed for all Cα-atom pairs of SH2 MD conformations of concatenated trajectories after fitting on the 'core' structure β-sheet (αC-atoms) of the initial conformation (t = 0 ns). The blue-red gradient shows the correlations from -1 (blue) to 1 (red). (**J-K**) Distribution of the representative atoms from residues R340, R358 and S361 which act as H-bond and/or salt bridge donor/acceptor centres in non-covalent interaction with pY721 of p-pep (data were taken every 500 ps). Representative atoms are OG, the oxygen atom of the S361 side chain; CZ, the carbon atom at two amino groups of R340 and R358. OG and CZ atoms are defined on the structural formula of serine and arginine; P is the phosphorus atom of pY721. (**L**) Superimposition of the SH2 domain's representative conformations of each cluster shown in two orthogonal projections. Clustering was based on the RMSD values (cut-off 0.75 Å) after least-squares fitting on the β-sheet core (G353-R358, Y368-R373, N378-F384). Protein is displayed as a grey cartoon with three IDRs, F1, F2 and F3, distinguished by colour corresponding to the respective cluster and delimited by pointed (F1), dashed (F2) and dotted (F3) lines. (**M**) Conformation of the p-pep in the representative conformations from C1-C4 clusters the p-pep/SH2 complex and its orientation in respect to the SH2 binding pocket. Two orthogonal projections are shown.

Focusing on the SH2 domain residues contributing to the pY721 binding in structure 2IUH and during MD simulation, we generated the spatial distribution of their representative atoms. The representative atoms from residues R340, R358 and S361, acting as H-bond and salt bridge donor/acceptor centres in non-covalent interaction with pY721 of p-pep atoms are OG, the oxygen atom of S361 side chain; CZ, the carbon atom at two amino groups of R340 and R358; (defined on structural formula of serine and arginin; P is the phosphorus atom of pY721).

The oxygen (OG) atom of the appreciably fluctuating S361 residue and the carbon (CZ) atoms from the low-fluctuating R340 showed similar enlarged distributions around the pY721 well-packed areas (**Figure 7.6**, **J**, **K**). From the other side, the CZ carbon atoms of the lowly-fluctuating R358 manifest compact spatial distributions. We suggested that in favourable steric conditions the long side chain of arginine broadly explores its conformational space, generating many different rotamers due to its outstanding conformational flexibility originating from multiple torsional degrees of freedom.

As we aimed to define the principal factors governing the p-pep recognition and binding by SH2 of PI3K, we first studied the conformational features of R340, R358, S361 and pY721.

The commonly used descriptors characterising rotation around each bond, showed that the arginine residue conceivably exhibits a significant number of rotamers (**Figure S28**, **A**). The arginine rotamers are described in the IUPAC conformational terms trans (t) and gauche (g) (https://goldbook.iupac.org/) and a population. According to the torsion angles values, in the p-pep/SH2 complex, the *g-ttt* rotamer represents 56% of all R340 conformations. The other R340 conformations are fully

heterogeneous, showing a reciprocal similarity with the occurrence of less than 10%. The R358 rotamers are regrouped into two clusters, C1, containing 85% of all MD conformations in g+ttg+ configuration, and C2, which comprises 14% of g+g+tt rotamers. Residue S361 having only one torsional degree of freedom, shows two significant conformers, g- (60%) and g+ (37%). The pY721 MD conformers are regrouped in three clusters, C1 (36%), C2 (35%) and C3 (21%), containing g-g-g+t, g-g-g+g- and g-g-g+g+, respectively. Such results are explained by the local environment surrounding R340, R358 and S361. Despite salt bridges formed with pY721 phosphate group: the majority of R340 neighbours are polar or identically charged residues that create repulsion forces favouring the flexibility of its sidechain; R358 neighbours have small hydrophobic or cyclic sidechains, consequently limiting the explored rotamers by attractive forces; S361's neighbours are polar or charged residues with donor sidechains for the majority. Stabilisation and destabilisation of non-covalent interactions between S361 and its surroundings may explain the resulting rotamers.

This conformational (or local) disorder of R340, R358 and S361 is an additional contributor to the inherent disorder of the SH2 domain. As S361 and R358 are located on the IDR F1, they appear three types of disorder – (i) the backbone folding/unfolding (structural events), (ii) displacement/rotation (dynamical events), and (iii) inherent (local) rotational disorder. The high population of R358 rotamer (85%, g+ttg+) indicates a small degree of dynamical and local disorders, also evidenced by the compact spatial distribution of the CZ carbon atoms of the lowly-fluctuating R358 (**Figure 7.6**, **J**, **K**). The highly fluctuating residue S361 is involved in structural, dynamical, and local disorders, resulting in its sparse distribution around the pY721 well-packed area. The lowly fluctuating residue R340 contributes to only local rotational disorder demonstrating a large conformational space.

Secondly, the SH2 MD conformations of the p-pep/SH2 complex were clustered based on the RMSD values (cut-off 0.75 Å), calculated after the least square fitting on the β-sheet 'core'. Five clusters, C1-C5, reveal different populations, the higher (C1, 37%), low (C2, 17%) and relatively lesser (C3-C4, 8% and C5, 4%) (**Figure S29**). The representative conformations of these clusters display the noticeable structural and conformational disparity detected in (i) both IDRs, F1 and F2, deriving from their unstable folding and large movements, and (ii) the β-sheet formed with β4 and β5 strands ending by a coiled C-term (**Figure 7.6**, **L**). Consequently, the RMSD-based clustering of the SH2 conformations from the p-pep/SH2 complex is identified as the third SH2 IDR, F3.

7.2.2.3. CHARACTERISATION OF THE FREE-LIGAND SH2 DOMAIN OF PI3K

To determine the impact of the p-pep binding into the PI3K SH2 domain and define the correct target for KIT KID docking, we examined the free-ligand (without p-

pep) SH2 domain using extended (2 μs) MD simulation. Like the p-pep/SH2 complex, three replicas of the free-ligand SH2 domain showed remarkable similarity of their RMSD and RMSF profiles and values (**Figure 7.7**, **C**; F**Figure S30**). Further analysis was performed either on the one trajectory for characterisation of the time-dependent metrics or concatenated trajectories for computing of time-independent statistical measures.

The stable and low RMSD values (1.5 - 2.5 Å) observed during all replicas are typical for excellent simulation convergence. The RMSF profile of free-ligand SH2 is similar to pep/SH2 complex, but the highest RMSD values observed for residues located in the loop between β1 and β2 strands, are twice as much (**Figure 7.7**, **C**, **D**).

The folding (2D) and tertiary (3D) structure of the free-ligand SH2 domain is similar to p-pep/SH2 complex (**Figure 7.7**, **A**; **Figure 7.6**). Identical to the p-pep/SH2 complex, we localised in the free-ligand SH2 domain three intrinsically disordered regions (IDRs) – F1, F2 and F3 –, which manifest (i) a structural disorder, evidenced by transient structures reversibly converting between 3_{10}-helix ↔ β-strand ↔ turn ↔coil, (ii) a dynamical disorder apparent as the highest fluctuations, at that (iii) degree of such SH2 disorder is increased compared to the p-pep/SH2 complex (**Figure 7.7**, **A**-**D**; **Figure 7.6**, **C**, **E**).

The cross-correlation matrix computed on the concatenated trajectories of the free-ligand SH2 shows more contrast and better interpretable pattern than in the p-pep/SH2 complex. It reflects clearly (i) the coupling motion between the β-strands within a 'core' β-sheet structure; (ii) the weak anti-correlation between two IDRs, F1 and F2; (iii) the per block correlation (positive or negative) of each IDR, F1, F2 and F3, with the other SH2 structural fragments (**Figure 7.7**, **E**).

PCA demonstrates that the first four modes describe a great majority of the variance of the SH2 motion (80%), with the 1st mode explaining 60% (**Figure 7.7**, **F**). The major contributor to the 1st mode is the IDR F1 showed a significant amplitude movement in the direction of the binding pocket of SH2 (**Figure 7.7**, **G**).

The 2nd mode, which showed a nearly perpendicular direction to the 1st vector, confirms the F1 reciprocating motion concerning the pocket is accommodating the KID p-pep in the p-pep/SH2 complex. Such flexibility facilitates binding pocket evolution upon the recognition process of p-pep. As seen from the PCA, the collective motion of the free-ligand SH2 domain differs markedly from that of the p-pep/SH2 complex. In the complex we observed large amplitude motions of all IDSs, F1, F2 and F3, while in free-ligand SH2 F2 and F3 motions are significantly reduced, while F1 exhibit ample motions toward SH2 binding pocket. Direction of F1 movement in the free-ligand SH2 is completely different from p-pep/SH2 complex's.

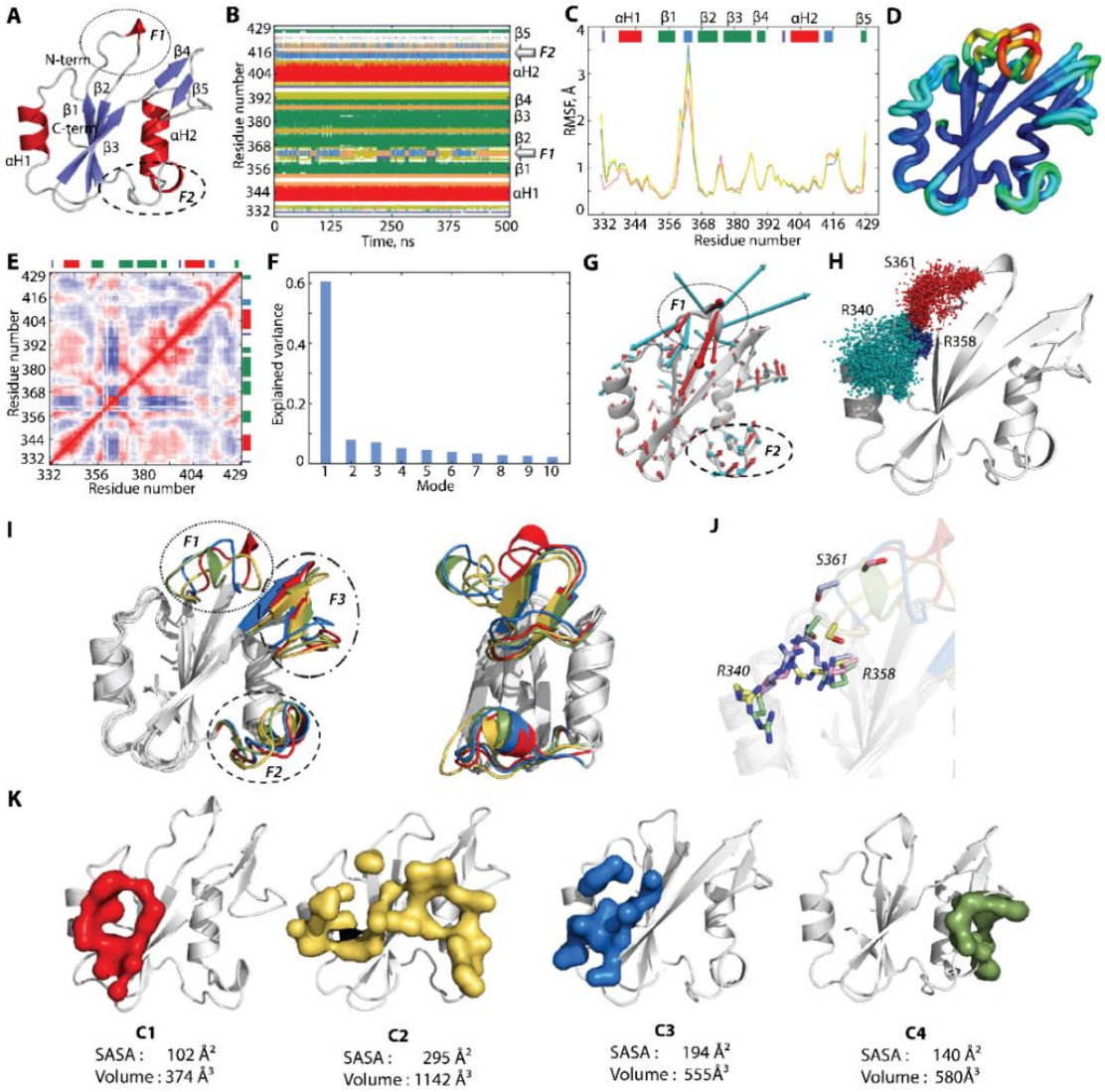


Figure 7.7 Structural and dynamical properties of the free-ligand SH2 domain from p85α PI3K. (**A**) The mean conformations were calculated on the concatenated trajectories 1-3. Protein is shown as a cartoon: α-helices, β-strands and coils are displayed in red, blue and grey, respectively. Areas F1 and F2, the most variable in 2D folding are delimited by pointed and dashed ellipses. (**B**) The secondary structure time-dependent evolution of each SH2 residue as assigned by DSSP method: α-helices are in red, 3_{10}-helices in blue, β-strand in green, turn in orange, and bend in dark yellow (replica 2). (**C**) RMSFs computed for the Cα atoms of the free-ligand SH2 domain conformations of each replica after fitting on the initial conformation (t=0 ns), identical for trajectories 1-3. The 2D folding of SH2 mean conformation is shown in the insert. (**D**) The average conformation of the SH2 domain is presented as tubes. The tube size is proportional to the residue atomic fluctuations computed on the backbone atoms. The red–blue gradient shows the RMSF values, from large (> 3.7 Å, in red) to small. (**E**) Dynamical inter-residue cross-correlation maps, computed for all Cα-atom pairs of MD conformations of concatenated trajectories of SH2 after fitting on the 'core' β-sheet structure of the initial conformation (t= 0 ns). Blue-red gradient shows the correlations from -1 (blue) to 1 (red). (**F**) PCA modes calculated for the concatenated MD trajectories (1-3) of free-ligand SH2 after least-square fitting on the 'core' β-sheet (αC-atoms) of the initial conformation (t = 0 ns). The bar plot gives the eigenvalue spectra in descending order for the first 10 modes. (**G**) Atomic components in the two first PCA modes are drawn as red (1st mode) and blue (2nd mode) arrows projected onto the average structure of the SH2 domain. Only motion with an amplitude ≥

4 Å is shown. Two areas, F1 and F2, manifested the most significant displacement are delimited by pointed and dashed ellipses. (**H**) Distribution of the representative atoms from residues R340, R358 and S361 which act as H-bond and/or salt bridge donor/acceptor centres in non-covalent interaction with pY721 of p-pep (data were taken every 500 ps). Representative atoms are: OG, the oxygen atom of S361 side chain; CZ, the carbon atom at two amino groups of R340 and R358. OG and CZ atoms are defined on the structural formula of serine and arginine. (**I**) Superimposition of the representative conformations from the free-ligand SH2 clusters are shown in two orthogonal projections. Protein is displayed as a grey cartoon with three IDRs, F1, F2 and F3, distinguished by colour corresponding to the respective cluster and delimited by pointed (F1), dashed (F2) and dotted (F3) lines. (**J**) Orientation of the side chains of R340, R358 and S361 in the representative conformations from C1-C4 clusters of the free-ligand SH2 domain. (**K**) The SH2 binding pockets (Fpocket) defined for each representative conformation from clusters C1-C4. Only pockets around the binding pocket observed in X-ray structure 2IUH are shown with its SASA and volume values.

To estimate the other p-pep binding effects, we generated the spatial distributions of the representative atoms from residues R340, R358 and S361 (for details, see 2.2.2) of the free-ligand SH2 domain. The oxygen (OG) atoms of the highly fluctuating S361 residue showed a large distribution with a mushroom cap shape. In contrast, the CZ carbon atoms from the low-fluctuating R358 are concentrated in a minimal area (**Figure 7.7**, **H**). On another side, the CZ carbon atoms of R340, belonging to the lowly-fluctuating αH1-helix, manifest the most extensive spatial distribution.

Considering the capacity of arginine and serine to rotational conformational flexibility originating from their torsional degrees of freedom, we used the descriptors defined above (see 7.2.2.2). According to these descriptors, the two most populated clusters of R340 from the free-ligand SH2 represent only 45% of all conformations from that 34, and 11% are the ttg+t and tttt rotamers respectively (**Figure S28**, **B**). The other conformations of R340 are fully heterogeneous, showing occurrences of less than 10%. The R358 rotamers are regrouped into four main clusters, each having a population of more than 10%, and comprise 83% of all MD conformations. The most populated clusters are composed of g+ttg+ (36%), tttt (20%) and g+tg-t (16%) rotamers. Residue S361 having only one torsional degree of freedom, shows two major conformers, g- (45%) and g+ (41%), while g- presents only 14%.

This analysis displayed that the rotational conformational flexibility of residues R340, R358 and S361 is increased in the free-ligand SH2 domain concerning the p-pep/SH2 complex. However, the residue-related specificity is conserved. Indeed, R340 is more locally disordered in both cases, as well as only 56% (in the complex) and 45% (in the free-ligand SH2) of its rotamers regrouped in the cluster(s). In comparison, 99 and 83% of the R358 rotamers, and 97 and 100% of the S361 rotamers form the well-defined clusters in the p-pep/SH2 complex and the free-ligand SH2, respectively. This suggests that pY721 phosphate group may be the main factor for the limitations of rotamer diversity of R340, R358 and S361.

The free-ligand SH2 conformations regrouped using the RMSD criteria (cut-off

0.75 Å) reveal four clusters of differed populations, the highest (C1, 74%), low (C2, 18%) and relatively minor (C3 and C4, 3%) (**Figure 7.6**, **B**). Superimposition of the representative conformations from each cluster exhibits that the main differences between these conformations are detected in (i) two IDRs, F1 and F2, derived from their reversibly transiting fold and significant conformational motion, and (ii) the β-sheet formed by β4- and β5-strands exhibiting variability in length completed by the collective movement of these β-strands and coiled C-term (**Figure 7.7**, **G**, **I**). Again, the RMSD-based clustering of the free-ligand SH2 conformations is identified as the third IDR, F3, likely to the p-pep/SH2 complex.

As S361 and R358 are located on the IDR F1, they appear with two types of disorder – (i) the backbone reversible folding/unfolding co-occurring with displacement/rotation motion, and (ii) the side chain rotational disorder (local), while R340 contributes to only local rotational disorder demonstrating a large conformational space (**Figure 7.7**, **H**, **J**).

Finally, the pockets found in the representative conformations of C1-C4 clusters (only pockets consistent with the binding pocket observed in the X-ray structure were considered) differ in their (i) location in the SH2 domain, (ii) volume and (ii) SASA values (**Figure 7.7**, **K**). The largest (SASA of 295 $Å^2$) and voluminous (1142 $Å^3$) pocket is observed in the conformation from C2. Its position on the SH2 surface is highly coherent with the X-ray structure's 2IUH. The SASA value (295 $Å^2$) is also close to that observed in 2IUH (215 $Å^2$). Nevertheless, the volume (1142 $Å^3$) is three times smaller than in 2IUH (3312 $Å^3$). The pocket of conformations from cluster C2 comprises two functional residues, R340 and R358, on its surface. Pockets found on the representative conformations from clusters C1 and C3 are localised close to αH1-helix and characterised by a significantly reduced SASA and volume (102 A^2 and 374 A^3 in cluster C1, and 194 A^2 and 555 A^3 in C3). Moreover, both reduced pockets include only R340. The pocket defined on conformation from cluster C4 is located in proximity with αH2-helix. Its metrics, SASA (140 A^2) and volume (580 A^3), are also diminished, and pocket's surface does not contain any functional residues.

Of SH2 eight residues interacting with the p-pep of KID in the crystal, which portion is part of the surface pockets found in the representative conformation of each cluster? Despite the significant difference between the pockets located on the conformations of clusters C1 and C2, seven out of eight residues are localised on their pocket surface. The pocket defined on the representative conformation of cluster C3 contains only four residues from eight and zero in C4.

Considering the pockets size, the residues forming these pockets, and the juxtaposition of the SH2 residues with those interacting with the p-pep of KID, it seems that the representative conformation of cluster C2 is an appropriate target for KID recognition/binding. Nevertheless, the large population (74%) of conformations in cluster C1 and the IDR F1 close to the binding site may favour a better adaptation of

the SH2 domain conformation to accommodate the KIT KID. The SH2 binding pocket adaptability evidenced by differences between the peptide complexes formed by p-pep with KIT and PDGFR was reported in [417].

7.2.3. Discussions

To deliver the appropriate molecular entity that specifically binds KID^{pY721}, the available crystallographic structure 2IUH— containing the PI3K N-terminal SH2 domain co-crystallized with the KIT KID phosphorylated peptide TNEYMDMK (p-pep)—was used as an initial cornerstone.

The archetypal structure of the SH2 domain is well conserved in a solid state and an aqueous solution in both SH2 forms, bound to p-pep and free-ligand entity. P-pep localisation, corresponding to the canonical mode of SH2 bonding[419], is maintained in the crystal by multiple non-covalent interactions involving all p-pep residues without exception. In an aqueous solution, the number of p-pep contacts with SH2 is significantly decreased. However, in both cases, pY721 of p-pep is engaged in multi-branching hydrogen and ionic bonding with R340, R358, and S361 of the SH2 domain, forming salt bridges. In a solvent, the formation of salt bridges is mainly due to entropy, usually accompanied by unfavourable ΔH contributions due to the desolvation of interacting ions upon association[421]. Due to many ionisable side chains of amino acids present in a protein, the pH in which it is placed is crucial for its stability. However, interactions between SH2 and KID M724 of the recognition linear motif pYXXM were not conserved during simulation.

This decrease of non-covalent contacts in solution is related to (i) the three intrinsically disordered regions (IDRs F1, F2, and F3) of SH2, which manifest either a structural disorder, evidenced by transient structures reversibly converting between 3_{10}-helix; β-strand; turn; coil, and/or a dynamical (conformational) disorder, and (ii) the high flexibility of p-pep N- and C-terminals. In the free-ligand SH2, the degree of such disorder is considerably increased compared to the p-pep/SH2 complex. We also established that, in this state SH2 coupled motion with the other SH2 structural fragments is increased: between (i) the β-strands of a 'core' β-sheet structure, (ii) between IDRs, and (iii) between each IDR. In free-ligand SH2, residues R340, R358, and S361, contributing to p-pep binding in the complex, appear in different types of disorder: R358 and S361 as located on the IDR showing backbone reversible folding/unfolding co-occurring with displacement/rotation motion and the sidechain rotational disorder (local), while R340 contributes to only local rotational disorder demonstrating a large conformational space.

These detailed characterisations of free-ligand SH2 and co-crystallised complex of a KID^{pY721} peptide with SH2, make a solid foundation for building of full molecular complex of KID^{pY721} with SH2.

7.3. IDENTIFICATION DE LA PROTEINE REDOX DU HVKORC1

Résumé. *Les réactions d'oxydoréductions sont des évènements majeurs de nombreux processus biologiques. Elles impliquent le transfert d'électrons de résidus cystéines par une protéine donneuse vers les cystéines oxydées d'une protéine accepteuse induisant le clivage d'un pont disulfure. Une telle réaction est responsable de l'activation et de la fonction de hVKORC1. Parmi de nombreuses protéines redox, quatre protéines candidates (PDI, ERp18, Tmx1, and Tmx4) ont été recensées dans la littérature comme partenaires probables de hVKORC1. L'analyse comparative par simulation de dynamique moléculaire de chacune de ces protéines a montré leurs similarités structurales mais surtout leurs différences majeures en séquence, en dynamique, en flexibilité et en surface accessible du site d'interaction. Selon cette analyse, seule l'une de ces protéines, la* Protein Disulfide Isomerase *(PDI), possède les propriétés structurales et dynamiques cohérentes avec une reconnaissance de la boucle L de hVKORC1 par deux fragments. L'identification de PDI comme partenaire redox de hVKORC1 a également été confirmée par des études* in vitro. *Ainsi, la simulation de dynamique moléculaire de PDI a permis de délivrer des conformations cibles de PDI pour la génération du complexe moléculaire hVKORC1/PDI, le premier représentant de l'INTERACTOME de hVKORC1 formé au cours de son activation par réaction d'échange thiol-disulfure.*

7.3.1. INTRODUCTION

Thioredoxins (Trxs) are disulphide reductases that are responsible for maintaining proteins in their reduced state inside cells. Trxs are involved in a wide variety of fundamental biological functions ([226] and references herein) and, therefore, are vital for all living cells, from archaebacteria to mammals. The wide variety of Trx reactions is based on their broad substrate specificity and potent capacity to reduce multiple cellular proteins[422]. This broad specificity for thioredoxin and related proteins has made it difficult to distinguish the true physiological partners for the protein from *in vitro* artefacts.

All membrane associated Trx proteins possess an active site made up of two vicinal cysteine (C) residues embedded in a conserved CX1X2C motif. These two cysteines, separated by two residues, play a key role in the transfer of two hydrogen atoms to the oxidised target and the breaking of the Trx–disulphide bond (**Figure 7.8**, **A**). This disulphide-relay pathway is accompanied by an electron transfer in the opposite direction. An intermediate state during the electron transfer is a mixed disulphide bond formed by a pair of cysteine residues from two proteins, which can be resolved by the nucleophilic attack of a thiol group from one of the flanking cysteine residues. Through

this mechanism, the disulphide is exchanged within one thiol oxidoreductase or between a disulphide donor and a target protein[423]. Thiol–disulphide exchange reactions occur between redox-sensitive biomolecules if donors and acceptors can interact in the appropriate orientations when attacking and leaving groups[228].

The thioredoxin fold is the most common structure found in thiol oxidoreductases; it has been carefully described in [424]. It is illustrated with the crystallographic structure of human protein disulphide isomerase (PDI; **Figure 7.8**, **B**;), which is the best-characterised enzyme that assists in the process of oxidative folding[425,426]. The PDI structure consists of a central five-stranded propeller with four flanking α-helices, an architecture that contains extra regions compared to the classical thioredoxin fold (a four-stranded β-sheet with three α-helices formed by about 80 residues).

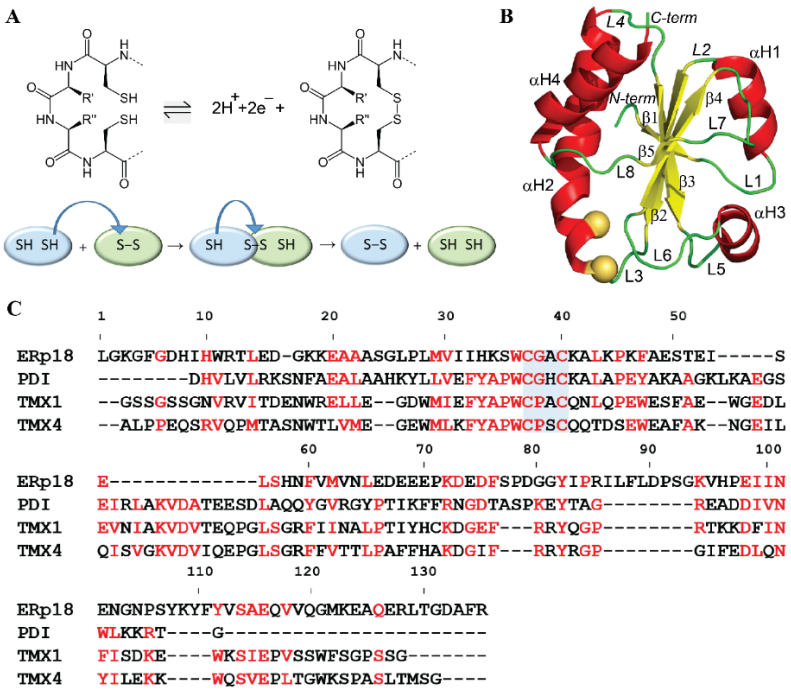


Figure 7.8 Thioredoxin-fold protein as a physiological reductant of human vitamin K epoxide reductase complex 1 (hVKORC1). (**A**) Oxidation of the two cysteine residues in the CX_1X_2C motif of Trx-fold proteins forms a disulphide bond, a process associated with the loss of two hydrogen atoms and, hence, two electrons (top). Mechanism of disulphide exchange between Trx and a target (bottom). The *H*-donor enzyme and a target are coloured in blue and green, respectively. (**B**) The Trx-fold is illustrated using the X-ray structure of human PDI deposed in the PDB (PDB ID: 4ekz). The protein is shown as red ribbons, with two cysteine residues from the CX_1X_2C motif as yellow balls. The four α-helices (in red), five β-strands (in yellow) and eight loops (in green) are numbered. (**C**) Comparison of the sequences of Trx-fold proteins ERp18, PDI, Tmx1 and Tmx4. Sequences were aligned on ERp18, having the most elongated sequence with ESPript3 (http://espript.ibcp.fr/). The solution with the best score is shown. The residues are coloured according to the consensus values: red indicates strict identity or similarity, while non conserved residues are in black. Blue highlights the CX_1X_2C motif.

The dithiol/disulphide group in the CX_1X_2C motif, which is located at the head of

the αH2 helix, protrudes from the protein surface and is exposed to a solvent. Such a spatial arrangement of the CX_1X_2C motif is probably to ensure the full accessibility of the first cysteine, which is required to react with the cysteine residue of a target to accomplish redox processes. It has been reported that the reactive thiolate of this first cysteine can be stabilised by the positive dipole at the head of the αH2 helix and by a network of hydrogen (H) bonds that are formed between the thiolate and neighbouring residues presented by the helix-turn structure[427].

In the present study, the focus is on the Trx's function as a physiological reductant (H donor) of vitamin K epoxide reductase complex 1 (hVKORC1).

Since the physiological reductant of hVKORC1 has not yet been identified, initial exploration was made of four human redoxin proteins, namely, protein disulphide isomerase (PDI), endoplasmic reticulum oxidoreductase (ERp18), thioredoxin-related transmembrane protein 1 (Tmx1) and thioredoxin-related transmembrane protein 4 (Tmx4), reported as the most probable *H*-donors of hVKORC1[222,224]. These proteins have distinct compositions for the active site CX_1X_2C—CGHC in PDI, CGAC in ERp18, CPAC in Tmx1 and CPSC in Tmx4—and they show broad but distinct substrate specificity. The nature of this specificity is the focus of this work. To evaluate the one most likely to reduce hVKORC1, a detailed comparison of these redoxins was first provided at different levels of the protein's organisation—sequence, secondary and tertiary structure, intrinsic dynamics, intraprotein interactions governing structural and conformation properties, and composition of the surface exposed to the targets.

This study principally leans on molecular dynamics (MD) simulation of the chosen Trx-fold proteins in the reduced state. It is suggested that a careful analysis of the simulation data will deliver quantitative and qualitative metrics to shed light on the following questions: (i) Are the 1D, 2D and 3D properties and the dynamic features good indicators for the prediction of the protein fragments participating in hVKORC1 recognition by a Trx? (ii) From the in-silico study of proteins, is it possible to predict which of them is the most likely partner of hVKORC1?

7.3.2. RESULTS

7.3.2.1. SEQUENCES AND STRUCTURAL DATA

Structures of PDI (PDB ID: 4ekz[426]), ERp18 (PDB ID: 1sen[428]) and Tmx1 (PDB ID: 1x5e; to be published) were used to extract the coordinates of a domain containing the CX_1X_2C motif (**Figure 7.8**, **C**; **Table S7**). This domain was chosen for the study of all proteins because ERp18, Tmx1 and Tmx4 proteins only constituted of one Trx-fold domain (a). The sequences of the four selected Trx proteins show a low identity/similarity (**Figure 7.8**, **C**; **Table S8**) along with the best scores for Tmx1 and

Tmx4 (47/68%). The ERp18 sequence differs most from those of PDI, Tmx1 and Tmx4 (23/38%, 15/23% and 15/23%, respectively). A 3D model of Tmx4 was built from Q9H1E5 (https://www.uniprot.org/uniprot/), with the Tmx1 structure as a template.

The ERp18, PDI and Tmx1 empirical structures and the Tmx4 homology model were optimised (when necessary) to obtain a CX_1X_2C motif in the reduced state. These were then used for the conventional MD simulations (two 500-ns trajectories for each protein), running under strictly identical conditions.

7.3.2.2. General Characterisation of Trx-Fold Proteins Using MD Simulations

The global stability of each Trx-fold protein over the course of a simulation was estimated using the root mean square deviation (RMSD) that showed (i) similar behaviour for the same protein among both MD replicas and (ii) significant disparity between the different proteins (**Figure 7.9**, **A**). Comparable RMSDs for PDI over each replica and between replicas characterise a highly stable protein structure during the simulation. Similar to PDI, the RMSD values for ERp18, Tmx1 and Tmx4 varied within a narrow range after elimination of the largest amplitude N/C-terminal residues. This demonstrates the good structural stability of each Trx, which is a quality that is typical of well-organised folded regular proteins.

Indeed, in all studied Trx-fold proteins, the properly ordered secondary structures (SS or 2D structure) were shown to be long-lived α-helices and β-strands. These ordered structures are interconnected by coiled linkers to form a stable globular 3D arrangement that is described as a four- or five-stranded antiparallel β-sheet sandwiched between four α-helix-bundle structures, which is an archetypical fold of the Trx family of proteins (**Figure 7.9**, **B**). Similar to the RMSDs, the root means square fluctuations (RMSFs) agree well between the pair of replicas for protein (**Figure 7.9**, **C**; **Figure S31**). The most pronounced difference in RMSFs between the two replicas is only observed in ERp18, in which β5 is partially unfolded and the L7 and L8 loops are joined together, resulting in large fluctuations.

Further characterisation of each protein and a comparison between the proteins is frequently completed by the observations obtained for a randomly chosen single trajectory or concatenated data. This is because the RMSDs and RMSFs in both replicas of each protein display comparable profiles and a similar range of values, and the 2D and 3D structures of each protein are perfectly matched (the RMSD values between the average structures of replicas 1 and 2 are less than 0.5 Å; **Figure 7.9**). The exception is PDI, in which the αH2-helix showed a different length over two replicas that was caused by the distant fold of its N-terminal.

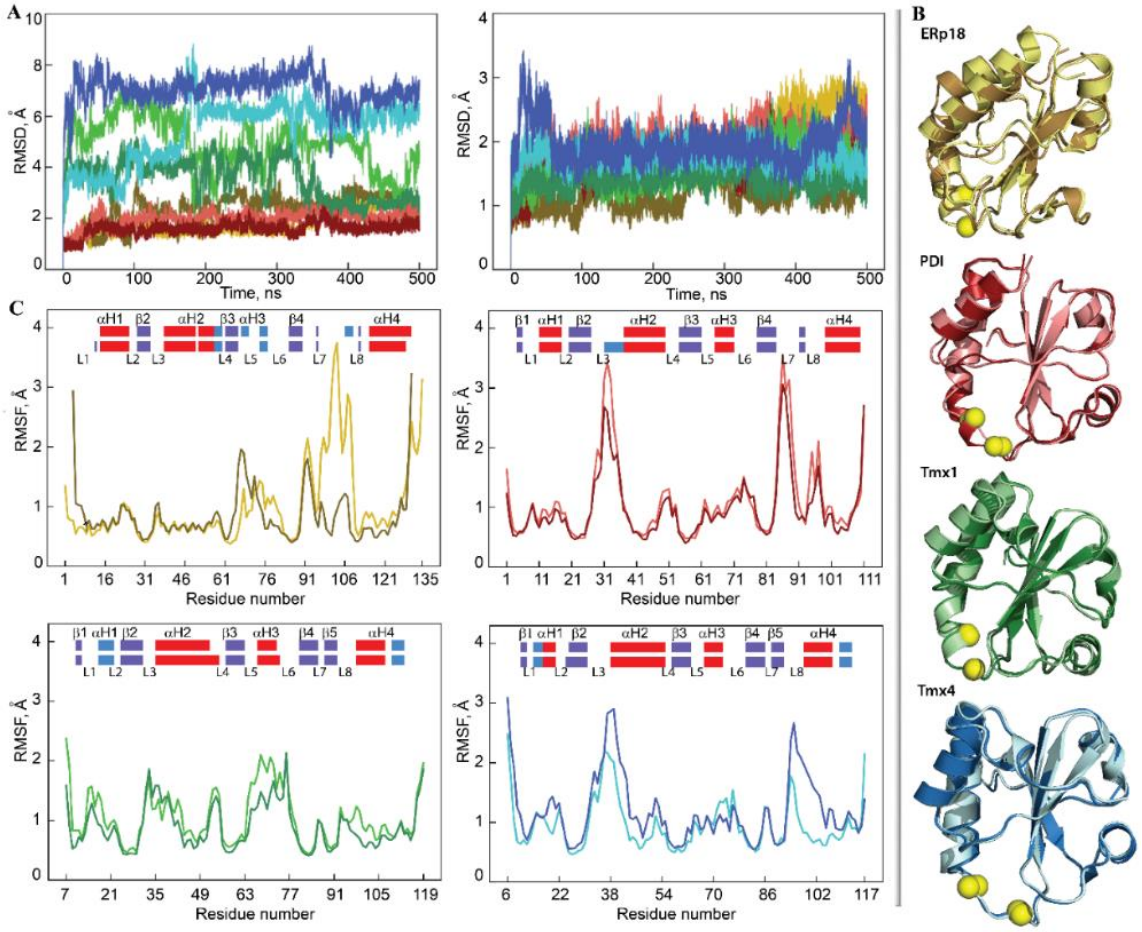


Figure 7.9 Characterisation of the MD simulations for the four Trx-fold proteins ERp18, PDI, Tmx1 and Tmx4. (**A**) RMSDs from the initial coordinates computed for all Cα-atoms (right) in each protein after fitting to initial conformation. (**B**) The superimposed average structures of each protein over replicas 1 and 2. Cysteine residues are shown as yellow balls. RMSD values of 0.5, 0.4, 0.3 and 0.4 Å in Erp18, PDI, Tmx1 and Tmx4, respectively. (**C**) RMSFs computed for the Cα-atoms using RMSF amplitude values less than 4 Å for the MD conformation of each protein after fitting to the initial conformation. Highly fluctuating residues (3, 6 and 5 in ERP18, Tmx1 and Tmx4, respectively) were excluded from the RMSD computation. In the insert, the secondary structures—αH- (red), 3_{10}-helices (light blue) and β-strands (dark blue)—were assigned for a mean conformation of every MD trajectory, 1 (top) and 2 (bottom) of each protein and were labelled as in the crystallographic structure of human PDI. (A–C) Proteins are distinguished by colour (first/second replicas): ERp18 (yellow/brown), PDI (light/dark red), Tmx1 (light/dark green) and Tmx4 (light/dark blue). The numbering of the residues in each Trx-fold protein is arbitrary and starts from the first amino acid in the 3D model.

How different are the 2D and 3D structures for the four proteins? The organised secondary structures, α- and 3_{10}-helices and β-strands, involve 55%, 60%, 60% and 56% of the residues in ERp18, PDI, Tmx1 and Tmx4, respectively, where the helical and β-strand fold portions vary from 36% to 42% and from 13% to 22% of total folding, respectively. Although all ordered 2D structures (helices and strands) are generally conserved across the studied proteins, their positions, lengths and qualities (e.g., α- or 3_{10}-helix) are slightly different (**Figure 7.9**; **Figure S31**).

The helical fold of each protein is represented by α-helices of different lengths (of 7–18 residues) and by 3_{10}-helices that consist of 3–4 residues. H1, which is a long-lived α-helix in ERp18 and PDI, is transient and converts between α- and 3_{10}-helices in Tmx1 and Tmx4. H2, which is the longest α-helix (14–18 residues) that contains the CX_1X_2C motif at its N-extremity, is generally conserved in all proteins; however, it may be partially split into two helices (ERp18) or reduced in size (PDI). The folding of the CX_1X_2C motif is different in the four proteins, and this represents a part of the regular α-helix (ERp18 and Tmx1), a transient helix fluctuating between α- and 3_{10}-helices or/and a turn (PDI) and a coiled structure (Tmx4). In ERp18, H3 consists of a pair of short 3_{10}-helices, while in the other proteins, it is a single and stable α-helix. H4 is a long and stable α-helix in ERP18 and PDI, while in Tmx1 and Tmx4, it is folded as a shorter α-helix and is joined to a 3_{10}-helix.

This analysis illustrates that although the studied proteins share a similar structure, their folding is noticeably different; this reflects their sequence-dependent character.

Additionally, the atomistic RMS fluctuations of the studied proteins show (i) minimal RMSF values for all β-strands forming the antiparallel β-sheet in all proteins, while the helices may have discernable fluctuations (e.g., αH2 and αH3 in Tmx1, and αH2 in Tmx4), and, as was expected, (ii) strong differences in the fluctuations of the coiled linkers (**Figure 7.9**, **C**). These linkers, which interconnect the core β-stands and the surrounding α-helices, are the most variable elements in the studied proteins in terms of sequence composition, length, and conformation. It is also noted that moderate (in the order of 1.5–2.5 Å) but systematically observed fluctuations of fragment L5-αH3-L6 arose in all studied proteins. This fragment is structurally adjacent to the CX_1X_2C motif and may play a role in thiol–disulphide exchange reactions.

7.3.2.3. Intrinsic Motion and Its Interdependence on Trx-Folded Proteins

Since a protein's dynamics influence its functional *properties,* intrinsic motions of Trx-fold proteins were compared. First, a cross-correlation map was computed for all Cα-atom pairs of each protein (**Figure 7.10**, **A**). The positively correlated motion of β2-, β3- and β4-strands, which was observed in each studied protein, reflects their concerted movement in the β-barrel. To equilibrate structural stability, the other fragments in Trx-fold proteins display a motion that tends to correlate negatively. As such, in ERp18, in addition to the β-barrel coupled motion, the structural moieties with the strongest correlation are L7 and αH4. In PDI, a regular fractal-like pattern shows the correlated motion of αH1 with αH2 and L7 and αH3 with L7 and L8. In Tmx1, the coupled motion is observed between the αH2-helix and the αH4-helix and between the β2-stand and the αH3-helix. Tmx4 demonstrates correlated motions between the αH2-helix and the β3-strand and between the αH3-helix and β4/β5-strands.

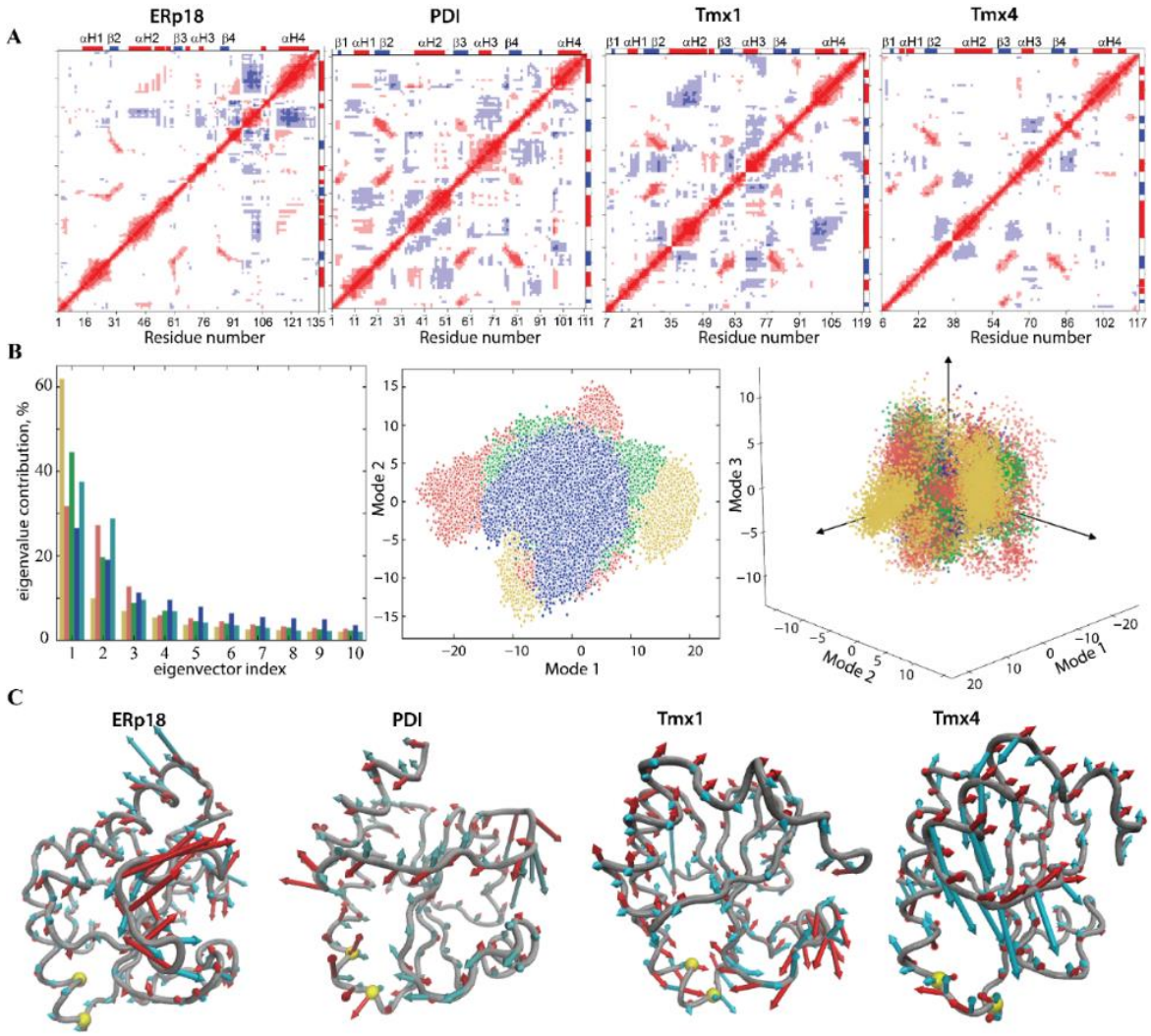


Figure 7.10 Intrinsic motion in the Trx-folded proteins and its interdependence. (**A**) Inter-residue cross-correlation maps computed for the Cα-atom pairs of ERp18, PDI, Tmx1 and Tmx4 after the fitting procedure. Secondary structure projected onto the protein sequences (α-helix/β-strand in red/blue) is shown at the border of matrices. Correlated (positive) and anticorrelated (negative) motions between Cα-atom pairs are shown as a red–blue gradient. (**B**) The PCA modes calculated for each protein after least-square fitting of the MD conformations to the average conformation as a reference. The bar chart gives the eigenvalue spectra in descending order for the first 10 modes (left). Projection of ERp18, PDI, Tmx1 and Tmx4 MD conformations with the principal component (PC) in 2D (middle) and 3D subspaces (right). MD conformations were taken every 100 ps (2D) and 10 ps (3D). The protein data is referenced by colour—ERp18 (dark yellow), PDI (brown), Tmx1 (green) and Tmx4 (dark blue and light blue for two replicas). (**C**) Collective motions characterised by the first two PCA modes. Atomic components in PCA modes 1–2 are drawn as red (1st mode) and cyan (2nd mode) arrows projected on a tube representation of each protein. For clarity, only motion with an amplitude ≥ 2 Å is represented. Cysteine residues are shown as yellow balls. All computations were performed on the Cα-atoms with RMSF fluctuations less than 4 Å for each protein after fitting on the initial conformation.

The collective motion of Trx-fold proteins and its impact on their conformational properties was studied using a principal component analysis (PCA). The principal

components (PCs) were determined, and the MD conformations for each protein were projected onto the PC subspace formed by the first two and first three eigenvectors. This indicated that Tmx1 (green) and Tmx4 (blue) conformations were grouped in a unique compact region for each protein, and these regions were perfectly superimposed for both proteins, while the conformations of PDI (red) and ERp18 (yellow) were trapped in two or three separate regions that were located in a slightly enlarged space (**Figure 7.10**, **B**). Randomly selected conformations from the distinct regions in the projection of the first two PCA modes showed that their conformational difference is mainly associated with a motion that leads to a slight skew of the H5-helix and displacement of the H3-helix in ERp18 and a disparity in the H2-helix length in PDI.

From the ten calculated PCA modes describing ~95% of total backbone fluctuations of each Trx-fold protein, the first two most dominant modes were used to illustrate ample collective movements qualitatively (**Figure 7.10**, **C**). The PCA modes of the Trx-fold proteins reveal the essential mobility of their fragments, which is either similar in the four proteins or has different features for a given protein. For instance, in ERp18, the greatest mobility is observed for L7 and L8 loops that are joined together due to the unfolding of the β5-strand. In PDI and Tmx4, the L7 and L8 loops are well separated by the β5-strand, but each of them shows the coupled motion of a large amplitude. Uniquely, in Tmx1, the αH3-helix and its joint L5 loop display a high amplitude motion. In PDI, Trx1 and Trx4, the collective motion of the H2-helix and joint L3 loop is comparable in amplitude but differs in direction.

7.3.2.4. Focus on the Region of Trx-Fold Proteins Potentially Involved in Target Recognition and/or Electron Transfer Reaction

To compare the four Trx-like proteins regarded as probable functional effectors of hVKORC1, the focus was on two fragments that may be involved in target recognition and/or electron transfer reaction. The first fragment, F1, comprises L3 and an N-extremity of αH2-helix that includes the CX_1X_2C motif and the second, F2, which is structurally adjacent to the CX_1X_2C motif, is composed of L5-αH3-L6. Both fragments form a frontal region that is exposed to the solvent in each Trx-fold protein, which may interact directly with a target during the electron-exchange process, similar to a bacterial protein containing a Trx-fold domain that is covalently bonded to VKOR (PDB ID: 4N5V[217]). The delimiting of these two regions is very approximate because the sequences and 2D structures of the studied proteins show significant differences. To have segments of a comparable length in different proteins, the boundaries of fragments were chosen so that their lengths were equal (17 residues; **Figure 7.11**, **A**).

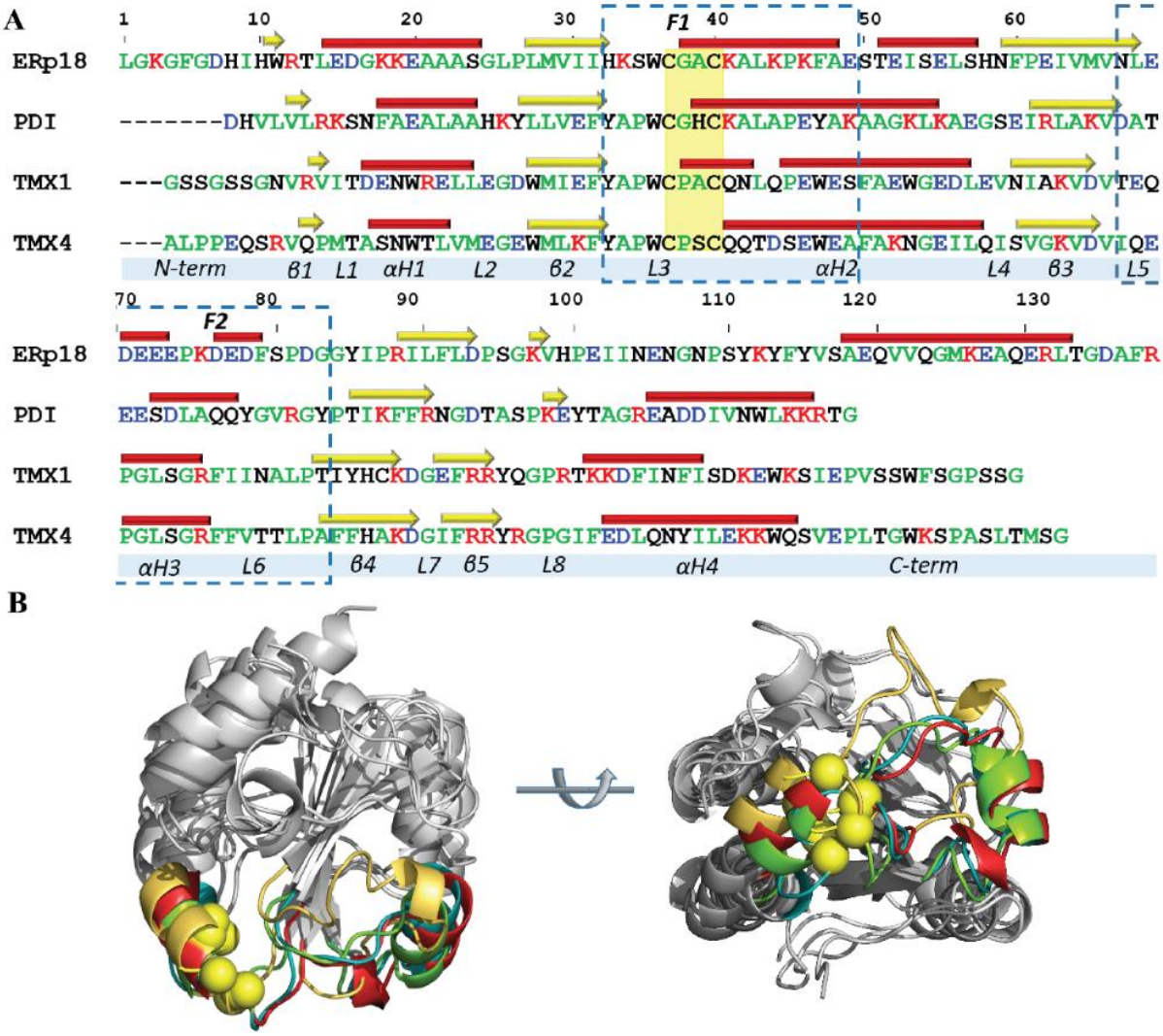


Figure 7.11 Sequence and folding of Trx-like proteins. (**A**) Alignment of the sequences and the secondary structure assigned to a mean conformation of the concatenated trajectory of each studied protein. Residues are coloured according to their properties—the positively and negatively charged residues are in red and blue, respectively; the hydrophobic residues are in green; the polar and amphipathic residues are in black; the CX_1X_2C motif is highlighted by a yellow background. The α-helices and β-strands are shown as red batons and yellow arrows, respectively. Secondary structure labelling is shown below the Tmx4 sequence. (**B**) The superimposed 3D structures of the Trx-fold proteins are shown in two orthogonal projections. The proteins are drawn as ribbons, with the cysteine residue as yellow balls. The F1 and F2 regions (and secondary structure labels) that are potentially involved in target recognition and/or the electron transfer reaction are outlined by dashed lines in (A) and differentiated by colour in (B) to distinguish between the proteins: ERp18 (dark yellow), PDI (red), Tmx1 (green) and Tmx4 (dark blue).

The F1 region in PDI, Tmx1 and Tmx4 is similarly initiated by tyrosine, which is the residue reported to be a breaker of secondary structures, while in ERp18, the role of "a breaker" is given to histidine, followed by lysine, which are amino acids that are more likely to be present in disordered regions[331,429]. The following residues of the L3-loop, a pair of hydrophobic residues (APs), are perfectly conserved in PDI, Tmx1 and Tmx4,

while in ERp18, these positions are occupied by positively charged and polar residues (KSs). Furthermore, the specific CX_1X_2C motif for each studied protein is preceded by tryptophan (W), which is a highly conserved residue in the four proteins. Tryptophan is an amphipathic residue that, similar to tyrosine, is often found at the surface of proteins and is sometimes classified as polar.

It is suggested that the F1 region of Trx-folded proteins, which contains the CX_1X_2C motif, contributes to redox reactions rather than target recognition. Nevertheless, a double action of the F1 fragment as both redox agent and recognition platform for a target has not been excluded.

The second surface region of Trx-fold proteins, F2, which is in proximity of the CX_1X_2C motif, consists of the αH3-helix and its two adjacent loops, L5 and L6. This fragment shows a negligible or no similarity/identity between the four proteins and, thus, may convey the highest degree of specificity in the discrimination/recognition of a partner. The most critical difference consists of the sequence composition of the L5 loop and the αH3-helix and the length of the αH3-helix. In ERp18, a set of five negatively charged amino acids (EDEEEs), which are positioned on the L5 loop and the αH3-helix, are separated by proline (P) and lysine (K) from the other three negatively charged amino acids (DEDs). This promotes a breakup of the H3-helix into two smaller 3_{10}-helices. In the other proteins, the number of negatively charged residues in this region is diminished to four in PDI and one in Tmx1 and Tmx4. The two last proteins, Tmx1 and Tmx4, have the same αH3-helix content and differ only in the combination of amino acids in L5. Despite a great difference in the αH3-helix composition of PDI compared to that of Tmx1 and Tmx2, the length of the helix in the three proteins is equivalent (6 aas). In all studied proteins, the short loop L5 contains at least one negatively charged residue and one polar residue, while the extended L6 loop is mainly composed of hydrophobic residues enriched by one or two polar residues with an inserted charged amino acid (the negative in ERp18 and the positive in PDI).

As the αH3-helix is moving considerably in Tmx1 and moderately in the other proteins (**Figure 7.10**, **B**), we suggest that the αH3-helix can adapt its orientation to get the best position with respect to the target and, together with its joint loops, L5 and L6, is able to build the recognition (docking) site(s) for target accommodation. The F2 region is the most dissimilar fragment in the studied proteins, and it has a sequence composed of hydrophobic stretches folded into a polar lipid environment. F2 also contains polar and charged residues required for stretches of sequence that are exposed to a solvent in cytosolic or extracellular environments[430]. Therefore, F2, which is positioned in the proximity of the CX_1X_2C motif, is a fragment of a Trx-fold protein that can contribute to hVKORC1 recognition.

7.3.2.5. GEOMETRY OF THE CX_1X_2C MOTIF

Focusing on the CX_1X_2C motif, a key agent in thiol–disulphide exchange reactions, its geometry was characterised in each Trx-fold protein. It was observed that structurally, the CX_1X_2C motif constitutes either a part of the αH2-helix (in ERp18 and Trx1), which is transient in PDI, or an extension of the L3 loop (in Tmx4). Both cysteine residues that are located on a coil are largely exposed to the solvent, whereas only one cysteine is exposed in the folded CX_1X_2C, while the other cysteine is buried in the protein chain.

Surprisingly, the folding of the CX_1X_2C (CGAC) motif in the calculated conformations (MD simulation) of ERp18 is coherent with those observed in the experimentally determined structure (PDB ID: 1sen), despite the different protein states, namely, reduced (MD simulation), with two protonated thiol groups, and oxidised (X-ray analysis), in which two deprotonated thiol groups form a disulphide bridge. In both protein states studied by the two different methods, the first cysteine from the CGAC motif is the N-cap residue (the last nonhelical residue) of the α-H2 helix.

The second unexpected observation is connected to the different folding of the CX_1X_2C (CGHC) motif in the calculated (MD simulation) and empirical structures (X-ray analysis) of PDI when studied in the same state (reduced). Indeed, the CX_1X_2C motif in the crystallographic structure of PDI was reported as folded, with the C37 positioned at the cap of the α-H2 helix (PDB ID: 4ekz), while in the MD conformations, the structure of this motif is transient and alternated between the helical fold (α- or 3_{10}-helices) and the turn/coiled structure, demonstrating high conformational plasticity.

The folding of the CX_1X_2C motif in Tmx1 (CPAC) in the MD conformations and the NMR structures (PDP ID: 1X5E; both are in a reduced state) is equivalent, with the first cysteine as an N-cap residue of the downstream α-H2 helix, similar to ERp18. In Tmx4, a protein with the most similar sequence to Tmx1, the CX_1X_2C motif (CPSC) demonstrates a coiled structure. In these two proteins, the conserved proline constitutes the characteristic CPX_2C motif, and the observed structural differences may be connected either to the X_2 residue or to the long-distance structural effects.

The geometry of the CX_1X_2C motif was described by two metrics: a distance S···S′ between the protonated sulphur atoms and a dihedral angle S−Cα−Cα′−S′ (**Figure 7.12**, **A**; **Figure S32**). In proteins ERp18 and Tmx1, the mean value (mv) of these parameters (4 Å and 60°, respectively) describe a synclinal configuration (Prelog–Klyne nomenclature) of the sulphur atoms that is well-conserved over the MD simulations. Nevertheless, a rare but not-negligible number of Tmx1 conformations revealed a syn-periplanar or anticlinal orientation of sulphur atoms that promoted a slight increase in the S···S distance. Such restrained geometry of the CX_1X_2C motif in Erp18 and Tmx1 is apparently related to its location on the well-folded αH2-helix. By contrast, the CX_1X_2C

motif located on a coiled L3 loop in Tmx4 stimulates a highly divergent orientation of sulphur atoms, running from syn-periplanar configuration to an antiperiplanar configuration, as was evidenced by a large variation in the dihedral angle S−Cα−Cα−S. The measured metrics, distance S···S and dihedral angle in PDI had values close to those in Erp18 and Tmx1. Nevertheless, a large number of conformations displayed a strongly variant geometry, which is similar to Tmx4. Such richness in PDI conformations corresponds to the transient structure of the N-terminal of the H2-helix, conversed between the helical fold (α- and 3_{10}-helices) and turn structure.

To better characterise the dynamical behaviour of the CX_1X_2C motif over two trajectories for each protein and to compare the different proteins, 3D skeletal shape trajectories of the motif's atoms were described in Kendall's shape space[431]. For a given integer k, Kendall's shape space is the manifold of dimension 3k − 7 dimension of all possible configurations of k atoms in R^3 considered up to a rigid transformation (translation, rotation and scaling). It has a Riemannian structure with a computable geodesic distance. The framework allows the use of geometric statistics and dimension reduction methods like multidimensional scaling (MDS) to analyse the shape trajectories[432]. These methods offer various ways of visualising all the data in a common space, summarizing them with a reduced number of variables and comparing them to each other. A tetrahedron, defined for the S- and Cα-atoms of two cysteine residues, C37 and C40, was extracted from conformations over MD simulations (**Figure 7.12**, **C**). The four proteins can be condensed in two major groups that are weakly overlapping (clearly visible in the 3D view): ERp18 and TMX1 on the one hand, PDI and TMX4 on the other hand, the latter group displaying a larger shape variation.

This analysis is illustrated by the superposition of the thiol groups (Cα-S-H) from the CX_1X_2C motif of the MD conformations for each protein (**Figure 7.12**, **B**). The orientation of the thiol groups favours H-bond interaction (S−H···S) only in ERp18 and some Tmx1 conformations. In PDI, both the thiol groups are shown to have the most variant orientation within a group and between groups, which reflects their high mobility.

The H-bond between the sulphur atoms of each cysteine is characterised for two cases: (1) the S-atom from C37 is the H-donor to the S-atom of C40, and (2) the S-atom from C40 is the H-donor to the S-atom of C37 (**Figure 7.12**, **D**; **Figure S33**). Monitoring of the geometry of S-H···S (1) showed a very low probability (0.1–0.9%) of such an interaction in all proteins. Contact (2) has a probability of 72% in Erp18 and 27% in Tmx1. Analysis of the contact metrics (distance S···S and angle at H-atom) indicated that a typical S-H···S H-bond is slightly stronger in Tmx1 than ERp18. Such an H-bond was not observed in the other studied proteins.

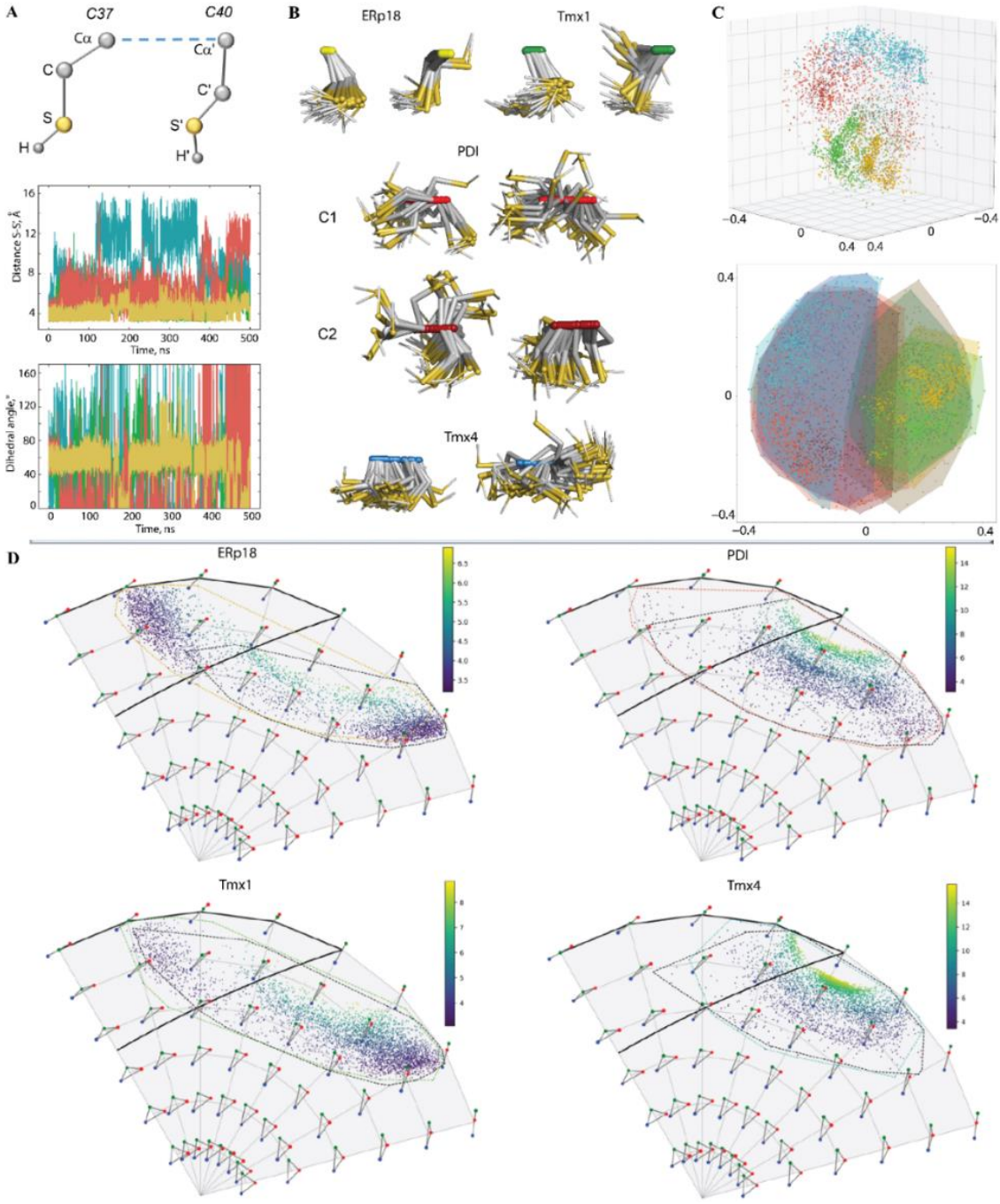


Figure 7.12 The CX_1X_2C motif geometries for ERp18, PDI, Tmx1 and Tmx4. (**A**) Geometry of the CX_1X_2C motif (left) is described by distance S···S′ (middle) and dihedral angle (right), determined as an absolute value of the pseudo torsion angle S−Cα(C37)−Cα′(C40)−S′. Only one replica 2 is shown. (**B**) Superposition of the thiol groups (Cα-C-S-H) from the CX_1X_2C motif of each protein is shown for either only one MD trajectory (ERP18, Tmx1 and Tmx4) or for both (PDI). Samples were taken for each 100-ns frame. (**C**) Multidimensional scaling (MDS) in 2D and 3D on the set of S-C-C-S tetrahedrons. Embedded points have been coloured according to the partner and replica they belong to. (**D**) Evolution of the shape of the triangles S-H···S on Kendall's disk of 3D triangles; each data point is coloured according to the S···S distance. Representative triangles are regularly sampled on the disk. The thick black line delimits the area of conformations favouring H-bond interaction. The dashed areas are contouring subpopulations according to the S-atom being the H-donor.

As expected, the S-H···S H-bond does not influence the folding of the CX_1X_2C motif (**Figure S34**). For instance, this H-bond is observed in conformations from clusters C1, C2 and C4 of ERp18, and it is absent in the others (C3 and C5), although the CX_1X_2C motif is well folded in both cases. Interestingly, both thiol groups do not contribute to H-bond interaction in the most prevalent PDI conformation with an unfolded CX_1X_2C motif, but K41, which is next to the C40 residue, is H-bound to H39 and L43. In the folded CX_1X_2C motif of PDI, C37 is in contact with P35 through the H-bond formed by the main chain atoms. Apparently, this interaction contributes to the stabilisation of the PDI conformation in the folded state, but it is not the unique factor that leads to such a structure. Similar but not-equivalent H-bonds are observed in the well-structured Tmx1 motif and the fully unfolded Tmx4 motif.

Structure organisation of the CX_1X_2C motif strongly influences their reactivity, affecting such properties as their accessibility and protonation state (i.e., pK_a)[433]. Functional analyses of each cysteine in the consensus CX_1X_2C motif demonstrated that N-terminal cysteine is important for the formation of a transient S−S bond with the substrate, whereas C-terminal cysteine is involved in substrate release[434]. In proteins, specific hydrogen-bond donors and an electropositive local environment tend to lower the pK_a by stabilising thiolate, and a hydrophobic environment or an electronegative local environment tends to raise the pK_a by destabilising a negatively charged, as opposed to a neutral form, side chain[433,435,436].

7.3.3. Discussions

Regarding the probable redox partners of hVKORC1 (Erp18, PDI, Tmx1 and Tmx4), it was observed that despite their similar architecture, each protein is characterised by its own sequence-dependent structural and dynamical features. In particular, it was observed that the CX1X2C motif's different folds are connected to the divergent configuration of the thiol groups—either as part of the well-folded αH2-helix (Erp18 and Tmx1), with the restrained cis-geometry of sulphur atoms, or as a part of a coiled structure, with the alternating orientation of sulphur atoms that runs from a syn-periplanar configuration to an antiperiplanar configuration (Tmx4), or as part of a transient structure (PDI) reversed between the helical fold (α- and 3_{10}-helices) and turn-coil structure, leading to a large number of thiol group configurations.

Focusing on the F1 region, suggested to be able to form intermolecular interactions with a target, it is noted that only F1 of PDI and the targeted S56–R61 segment of hVKORC1 have similar structural properties, or rather, a structural disorder that describes an intrinsically disordered region (IDR). Indeed, two IDRs, which are the transient N-terminal of the α-helix H2 in PDI and the transient H2-L helix comprising the S56–R61 segment from the L-loop of hVKORC1, show similar structural heterogeneity and plasticity that is consistent with an affinity that is sensitive to changes in local frustration distribution and thermodynamics.

Numerous publications have reported that many protein–protein interactions (PPIs) are mediated by protein regions that are not confined to a single folded conformation prior to binding, namely, IDRs that participate in PPIs (interacting IDRs)[55,437,438]. IDRs are increasingly recognised for their prevalence and their critical roles in regulatory intermolecular interactions[308]. It has been hypothesised that some traits make IDRs particularly suitable for interactions that involve signalling and regulation, complementing globular domains that more often perform catalytic functions. It has been estimated that IDRs in the human proteome contain ~132,000 binding motifs[352]. Disordered proteins are believed to account for a large fraction of all cellular proteins, playing roles in cell-cycle control, signal transduction, transcriptional and translational regulation, and large macromolecular complexes[353]. Nevertheless, even if fragment F1 is considered the most probable fragment to form intermolecular interactions with a target, the mobility of linker L5 and the αH3 helix from F2 of PDI means that F2 has strong compatibility with the highly mobile S56–R61 segment of hVKORC1. Moreover, F2 shows the most dissimilar sequence in the studied proteins, and it also has a great number of hydrophobic, polar and charged residues that are exposed to solvents. Consequently, F2 is also potentially able to contribute to stabilising a supramolecular complex. These two fragments are very close to the CX1X2C motif, which is either joined in a sequence (F1; sequence vicinity) or adjacent in a 3D structural space (F2; spatial vicinity).

This makes clear that we can begin to construct models of the molecular complex formed by hVKORC1 and PDI, where PDI is the most probable redox partner of hVKORC1.

7.4. Conclusion sur l'identification et la caracterisation des proteines partenaires du RKT KIT et du hVKORC1

Un point de départ pour étudier les interactions du RTK KIT est la structure déterminée empiriquement (rayons X) référencée 2IUH dans la PDB. Cette structure représente un complexe du domaine SH2 N-terminal de la sous-unité p85 du transducteur de signal intracellulaire PI3K, lié au phosphopeptide TNEYMDMK (p-pep, résidus T718-K725) du KIT KID avec la tyrosine phosphorylée pY721. L'étude des propriétés structurales et dynamiques (par simulation MD étendue) du complexe moléculaire p-pep/SH2 et de son dérivé SH2 libre de son ligand, a mis en évidence l'impact de la liaison de p-pep sur le domaine PI3K SH2 et a fourni le partenaire protéique optimisé pour l'amarrage sur KID phosphorylé à la même position que p-pep (KID^{pY721}).

Après une étude *in silico* des quatre Trxs, suggérées comme des partenaires les plus probables du hVKORC1 par des études empiriques, nous avons fourni une caractérisation détaillée et comparative de leurs séquences, structures secondaires et

tertiaires, leurs dynamiques, leurs interactions intraprotéiques et la composition des surfaces exposées aux cibles. Nous avons identifié sur chaque protéine des sites potentiellement impliqués dans la reconnaissance de hVKORC1. Similairement, sur la protéine hVKORC1 sous forme inactive (oxydé), nous avons (1) montré qu'un faible repliement de la boucle d'activation (boucle L) favorise une flexibilité conformationnelle riche qui permet la formation de sites d'arrimage appropriés pour la reconnaissance et la liaison d'un partenaire redox (Trx) et (2) identifié le fragment dans la boucle L potentiellement impliqué dans la reconnaissance d'une protéine redox. Une telle analyse permet de prédire les sites de reconnaissance/liaison putatifs sur chaque protéine isolée, et la protéine disulfite isomérase (PDI) est suggérée comme le partenaire hVKORC1 le plus probable.

Nous avons identifié et caractérisé les protéines partenaires pour chaque cible, caractériser leurs propriétés structurales et dynamiques et tenter de générer les complexes moléculaires de RTK KIT et hVKORC1 avec ces protéines partenaires.

CHAPITRE 8. AMARRAGE PROTEINE-PROTEINE : MODELISATION DES COMPLEXES MOLECULAIRES OUVRANT LA VOIE VERS LA GENERATION DES INTERACTOMES DE RTK KIT ET HVKORC1

Nous avons précédemment caractérisé les propriétés structurales et dynamiques, ainsi que les régions d'interactions probables entre différents partenaires du RTK KIT et de hVKORC1 : KID^{pY721} et le domaine SH2 de p85, et la boucle L de hVKORC1 et PDI. Ainsi, nous avons préparé chaque protéine pour étudier leur reconnaissance et leurs interactions.

Dans ce chapitre, nous présenterons la génération de complexes précurseurs entre ces deux couples de protéines par deux méthodes : une méthode appelée *user-guided*, basée sur une connaissance préalable des propriétés intrinsèques obtenues *in silico* pour chacun des partenaires, les critères biophysiques nécessaires à la formation des complexes (ex. distances optimales entre résidus spécifiques) ; une méthode d'amarrage classique (HADDOCK). Chaque modèle généré a ensuite été affiné par simulation de dynamique moléculaire

Ce chapitre est une adaptation des articles suivants :

1. **Ledoux, J.**, & Tchertanov, L. (2023). Site-Specific Phosphorylation of RTK KIT Kinase Insert Domain: Interactome Landscape Perspectives. *Kinases and Phosphatases*, *1*(1), 39–71. https://doi.org/10.3390/kinasesphosphatases1010005

2. Stolyarchuk, M.[+], **Ledoux, J.**[+], Maignant, E., Trouvé, A., & Tchertanov, L. (2021). Identification of the Primary Factors Determining the Specificity of Human VKORC1 Recognition by Thioredoxin-Fold Proteins. *International Journal of Molecular Sciences*, *22*(2), 802. https://doi.org/10.3390/ijms22020802

3. **Ledoux, J.**, Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

Les données supplémentaires et les méthodes relatives à toutes ces publications sont présentées dans les annexes de la thèse.

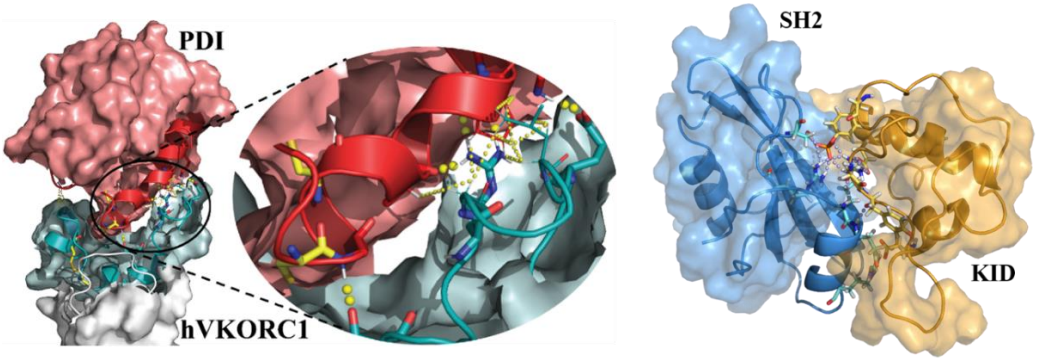


Figure 8.1 Abstract graphique du chapitre.

8.1. Reconnaissance et liaison de KIT et PI3K par le complexe KID^{pY721}/SH2

Résumé. *La structure du domaine SH2 de la sous-unité régulatrice p85 de la phosphatidylinositol 3-kinase (PI3K) cristallisée avec un peptide du KID phosphorylé en Y721 (PDB ID : 2IUH) est une information de valeur pour générer le complexe KID^{pY721}/SH2. L'étude directe de la structure cristallographique montre que le peptide est stabilisé dans une poche polaire de symétries horizontales et verticales, et que le phosphate de pY721 est maintenu en contact avec SH2 par un réseau complexe parmi lequel nous avons observé des liaisons hydrogène avec trois résidus conservés et hautement stables pendant la simulation de dynamique moléculaire. L'amarrage automatique du peptide cristallisé avec SH2 mais également du KID suggère deux orientations possibles, et pour le peptide et pour KID^{pY721}/SH2, l'un par rapport à l'autre. Pour prendre en compte le désordre intrinsèque et l'adaptabilité des régions de reconnaissance des deux partenaires, nous avons modélisé et simulés ces deux complexes non covalents par leur rapprochement progressif. Ainsi, nous avons pu postuler un premier précurseur du complexe KID^{pY721}/SH2 pour la modélisation de l'INTERACTOME du KIT.*

8.1.1. Introduction

SH2 (Src Homology 2) domains are well-characterized and most studied modules contributing to protein-protein interaction (PPI) and capable of recognizing sequences/docking sites containing a phosphorylated tyrosine. This post-transduction event is a key regulator of physiological molecular activation in the cell, making SH2 domains a key player with a fundamental role in cell signalling. Therefore, several pathologies result from the deregulation of these PPIs mediated by SH2 domains.

In response to extracellular stimuli (SCF), RTK KIT is activated and initiates signalling by phosphorylation of their phosphotyrosine-containing cytoplasmic platforms (KID, JMR and C-tail) and then, downstream adaptors, signalling or scaffolding proteins that are recruited to these phosphotyrosine (pY) primarily via SH2) and pY binding domains (PTB)[120,382,415].

The highly selective interaction of signalling proteins with pY binding motifs specifically activates signalling pathways, such as canonical signalling via Ras mitogen-activated protein kinase (MAPK), phosphatidylinositol 3-kinase (PI3K) and phospholipase C-gamma (PLC-γ)[416].

It is well known that the SH2 affinity for pY depends on the amino acid sequence at the pY site. In Src-family kinases (SFKs) pYEEI motif is preferentially bound, while the SH2 domains of PI3K or PLC-γ preferentially bind to pYφXφ (where φ is a residue with a hydrophobic side chain)[416].

In this study, we analysed the empirically determined (X-ray) structure retrieved from the Protein Data Bank (PDB) and referenced it as 2IUH. This structure represents a complex of intracellular signal transducer PI3K p85 N-terminal SH2 domain (SH2) bound to the phosphopeptide TNEyMDMK (p-pep, residues T718–K725) of KIT KID with the phosphorylated tyrosine pY721[417]. Study of the structural and dynamical properties (by extended MD simulation) of the molecular complex p-pep/SH2 and its free-ligand SH2 derivative, highlighted the impact of p-pep binding on the PI3K SH2 domain and provided the optimised protein partner for KID docking.

Docking of the phosphorylated KID into the PI3K SH2 domain was performed, and the resulting *de novo* models of the molecular complex were compared to the unique empirically determined structure 2IUH. The *de novo* KID^{pY721}/SH2 model is postulated as the first comprehensive precursor of RTK KIT signalling complex. This model adds new information about the well-studied RTK KIT, offers insights into its less studied post-transduction events, and highlights RTK KIT interactions and signalling pathways through KID. RTK KIT interactions with its downstream proteins represent the prominent perspective for developing novel therapeutic agents.

8.1.2. Results

8.1.2.1. Can the Crystallographic Structure of Complex Formed by p-pep of KID and PI3K SH2 Domain be Reproduced by Docking?

Prior to studies of the PI3K SH2 domain recognition/binding by KIT KID, we performed a bench test to investigate if docking can reproduce the empirically determined p-pep/SH2 complex. First, the docking trials were conducted using the

structure 2IUH as a benchmark set. The p-pep/SH2 complex was separated into unbound protein SH2 (X-ray Model, M1) and free p-pep peptide, and these isolated entities were docked with High Ambiguity Driven DOCKing (HADDOCK)[248]. Unlike other docking approaches, based on the combination of energetics and shape complementarity of studied proteins, HADDOCK uses biophysical interactions data—in our case, the H-bond contacts between p-pep and SH2 domain, to drive the docking process. The p-pep residues were considered as the 'active centres' while the 'passive' residues were defined as all SH2 residues positioned the distance ≤ 4 Å from residues interacting with the p-pep in structure 2IUH.

From 1 000 decoys generated by rigid-body docking, 200 solutions were refined (semi-rigid docking). Analysis of these solutions by using the fraction of common contacts (FCC) clustering[439] produced four clusters, C1-C4, with different populations (**Figure 8.2**, **A**). Each cluster, the most populated C1 (57%) and the sparsely populated C2 (10%), C3 (8%) and C4 (5%), contains the p-pep complex models showing differently oriented p-pep. Nevertheless, the great majority of them are occupying the SH2 binding pocket. In the binding pocket, the major p-pep position is similarly observed in structure 2IUH. However, the p-pep orientation shows that its N-terminal is located close to either αH1-helix, similarly to the crystallographic structure 2IUH or in the opposite direction, in proximity to αH2-helix. The p-pep is rarely oriented along the β-strands direction (perpendicular to the major p-pep position). Curiously, the p-pep orientation with its N-term located at the αH1-helix is observed in the four best solutions for C1, C3 and C4 clusters, and only in C2 its N-term is localised at the αH2 helix (**Figure 8.2**, **B**).

The p-pep in two major orientations in the SH2 binding pocket is stabilised by the different networks formed by non-covalent interactions – H-bonds, salt bridges and van der Waals contacts (**Figure 8.2**, **C**, **D**). The non-covalent interaction pattern stabilising the p-pep/SH2 complex is much more crowded for the best docking solutions which are composed cluster C1 in respect to that of C2. Interesting that specific interactions, the salt bridges formed by pY721 with R340, R358 and S361, and H-bonds N719···N365, D723···N378 and K725···N344, are maintaining the complex with p-pep in two opposite orientations and consequently, are conserved for the major p-pep location. The non-covalent interaction patterns observed in the representative conformation of cluster C1 and X-Ray structure 2IUH are quasi-identical (recognition of pYXXM linear motif); the difference consists of some additional contacts in the 'docked' complex.

Consequently, docking trials showed that HADDOCK reproduces the benchmark structure. The used protocol was further applied in docking of the *de novo* model of KIT KID^{pY721} (ligand) into the PI3K SH2 domain (target) issued (i) from the empiric structure 2IUH (Model 1, M1) and (ii) the representative conformation of the free-ligand SH2 from cluster C1 (Model 2, M2).

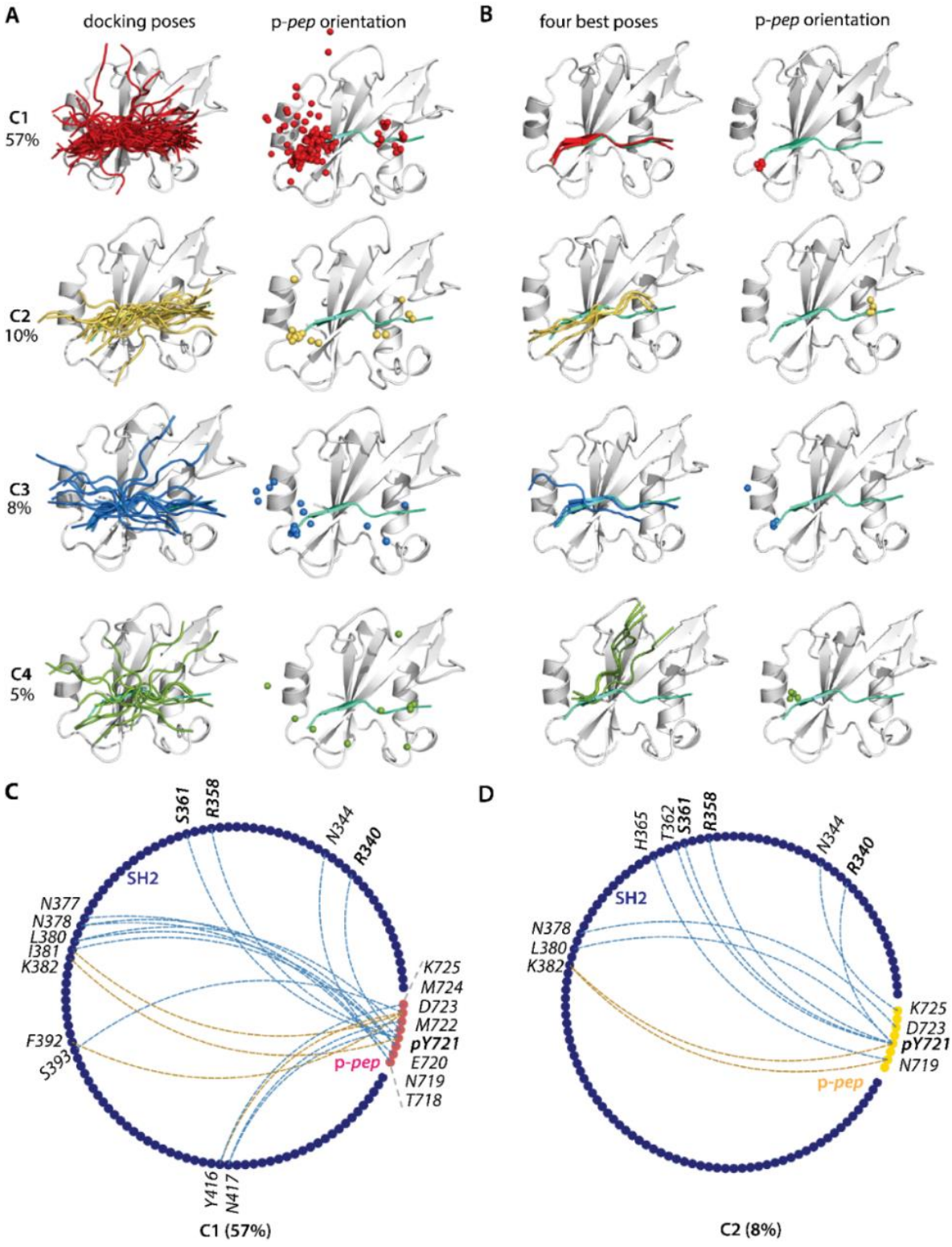


Figure 8.2 Computational docking of p-pep (ligand) onto SH2 (target) performed with HADDOCK using an information-driven method (benchmark). (**A**) Docking poses distributed into four clusters (left) showed the p-pep orientation (right). (**B**) Superimposition of the top 4 solutions (left) and the p-pep orientation (right). (A-B) The target is a grey cartoon; the ligand is coloured per cluster. The N-terminal of p-pep is presented as a ball. (**C**-**D**) H-bonds (in blue) and hydrophobic contact (in beige) stabilising the p-pep (in red/yellow) and SH2 domain of PI3K (in blue) showed as a string diagram for the representative complex from cluster C1 (C) and C2 (D).

8.1.2.2. DOCKING OF KID INTO PI3K SH2 DOMAIN

The encouraging outcomes from the bench test prompt us first to dock the KIT KID^{pY721} into the SH2 domain from 2IUH (M1). The KID^{pY721} docking into Model 1 (1 000 decoys generated by rigid-body docking) produced the KID/SH2 (M1) complexes. The majority of retained 200 solutions (60%) were regrouped using the RMSD criteria (cut-off 0.6 Å) onto seven clusters (C1-C7) with the relatively low population varied from 22 to 5%, in total regrouping 70% (cut-off 5%) of all solutions (**Figure S35**, **A**).

The docking poses show different positions of KID^{pY721} (and its p-pep) concerning the SH2 binding pocket, but the p-pep major position is similar to what is observed in structure 2IUH. Nevertheless, the KID^{pY721} orientation viewed by its p-pep is quite divergent even within the same cluster, showing a circular-distributed orientation of the KID p-pep around the SH2 (M1) binding site. After refinement of the 200 docking solutions, the KID p-pep position in the SH2 (M1) binding cavity corresponds better to the X-ray structure, however its orientation corresponds either to the structure 2IUH (C4) or appears an opposite arrangement (C1) (**Figure S35**, **B**).

The docking solutions obtained upon the docking of KID^{pY721} into the SH2 domain from 2IUH (M1) are discouraging. These ambiguous results may be due either to an incorrect initial structure of the one or two anchored partners, SH2 (M1) or/and KID^{pY721}, or to the docking algorithm.

First, we suggested that either (i) the use of the SH2 structure from 2IUH (M1) as a target is not a good idea, or (ii) the p-pep orientation in the SH2 binding site from the X-Ray structure may correspond to a wrong solution due to p-pep small length and binding site symmetry of their residues showing the similar biophysical properties.

Trying to explore our hypothesis, primarily a 'bad target choice', we performed a docking (always with HADDOCK) of KID^{pY721} into the SH2, which is the representative conformation of the most populated cluster C1 of the free-ligand SH2 (M2). Surprisingly, the KID^{pY721} docking into SH2 (M2) produced similar results as obtained for the KID^{pY721} docking into M1 (**Figure S36**).

The non-covalent interaction patterns in the KID^{pY721}/SH2 complex models obtained by the KID docking into the two models of SH2 (M1 and M2) differ (**Figure 8.3**). Moreover, this difference is the most meaningful compared to the p-pep/SH2 complex in the solid state or water solution. Only the salt bridges formed by pY721 with residues R340, R358 and S361 is the most conserved motif of the non-covalent bonding, which was observed in clusters C1 and C2 (benchmark) and cluster C2 of KID^{pY721}/SH2(M2) obtained by the docking trials.

This motif is observed in the p-pep/SH2 complex in the solid state (structure 2IUH), in solution (MD conformations) and in the benchmark docking results, as well as in

KID^{pY721}/SH2 complex, obtained by the KID^{pY721} docking into MD conformation of the SH2 domain. This highly conserved motif of the non-covalent bonding stabilising the KID^{pY721} p-pep and full-length KID^{pY721} would be helpful for constrained docking or dynamics simulation. Contrary to the benchmark (p-pep docking to 2IUH SH2 domain), the recognition of KID complete linear motif pYXXM, with stabilisation of M724 with SH2 is missing in the docking solutions of both M1 and M2.

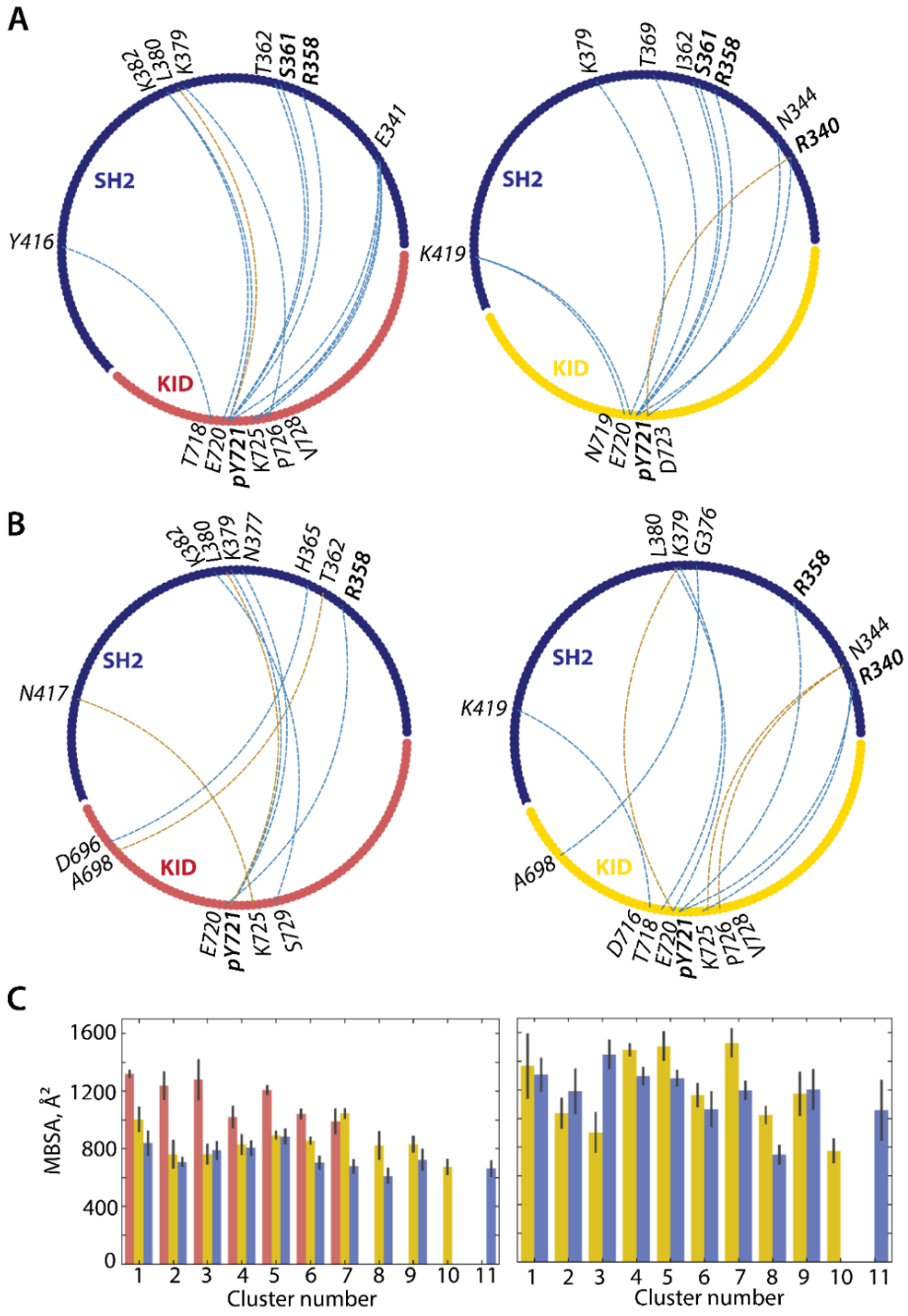


Figure 8.3 Analysis of the KIT KID^{pY721}/PI3K SH2 models obtained by computational protein-protein docking. (**A**, **B**) H-bonds (in blue) and hydrophobic contacts (in beige) stabilising the KID (red/yellow area) and PI3K SH2 domain (dark-blue area) in KID^{pY721}/SH2(M1) (A), and KID^{pY721}/SH2(M2) (B) docking solutions for the representative conformations from clusters C1 (left) and C2 (right) are showed as a string diagram. (**C**) The p-pep buried surface area calculated for

complexes p-pep/SH2 (left) and KID^{pY721}/SH2 (right). Models of p-pep/SH2 are in red, KID^{pY721}/SH2 (M1) are in yellow, and KID^{pY721}/SH2 (M2) are in blue. Protein-protein docking was performed with HADOCK using an information-driven method.

The buried surface area, which measures the size of the interface in a protein-protein complex [43], was calculated for complexes p-pep/SH2 and KID^{pY721}/SH2. Comparing only the p-pep/SH2, the mean buried surface area (MBSA) of the complex obtained by re-Docking of the X-Ray structure 2IUH (benchmark trials) is globally greater in the first six clusters (1-6). In contrast, the other clusters' MBSA values are almost identical within the standard deviations (**Figure 8.3**, **C**). The MBSA values of the KID^{pY721}/SH2 conformations from the first two clusters (1-2) are nearly similar between the models KID^{pY721}/SH2 (M1) and KID^{pY721}/SH2 (M2), while in cluster 3 its value is more significantly different. Indeed, the MBSA of KID^{pY721}/SH2 (M2) is greater than of KID^{pY721}/SH2 (M1). This result is not unexpected because the total amount of surface area buried within its fold is tightly coupled to the overall flexibility of a protein[412]. Moreover, as KID is the intrinsically disordered protein, its vast conformational landscape makes difficult the identification of favourable states for SH2 recognition and docking.

As the protein-protein binding ambiguous results may be also produced by the method applied, it is worthwhile to note that HADDOCK quantitative measures of binding modes–number of clusters, score and population–do not allow comparisons between docking solutions reflecting conformational and orientational characteristics, even though a simple superposition of the found solutions showed that a reference (X-ray) pose was observed.

8.1.2.3. Intuitive User-Guided Modelling of Molecular Complex KID^{pY721}/SH2

To resolve the problem of the KID^{pY721}/SH2 complex building and to take into account KID inherent flexibility, we applied an alternative approach which we previously used for the construction of a protein-protein complex for the pair of partners, the vitamin K epoxide reductase (VKOR) and its redox protein, the Protein Disulphide Isomerase (PDI)[221]. The 3D models of the KID^{pY721}/SH2 complex were built using the crystallographic structure of the KID p-pep/SH2 as a reference for the initial positioning of KID^{pY721} relative to the SH2 domain.

To be most objective in the modelling of the KID^{pY721}/SH2 complex, the structure 2IUH was not used as a template because (i) of suggested alternative KID^{pY721} position/orientation in the SH2 binding pocket, (ii) of a high structural/conformational variability of KID and free-ligand SH2, and (iii) of a considerable difference between the p-pep and KID^{pY721} size.

For modelling the KID^{pY721}/SH2 complex, a conformation of KID^{pY721} having an excellent similarity of its TNEYMDMK fragment with the p-pep from the empirical structure 2IUH (minimal RMSD values) was chosen as the initial structure to bring the two proteins as close as possible. As for the initial SH2 model, the conformation from the most populated cluster C1 of the free-ligand SH2 domain was chosen and positioned at KID^{pY721} so that (i) the distance between the phosphorus atom of KID^{pY721} and the CZ atoms of R340, R358 and OG atom of S361 from the SH2 was at least 10 Å; (ii) KID^{pY721} was alternatively placed above the middle of the SH2 binding pocket, with its TNEYMDMK fragment oriented (a) similarly to p-pep in structure 2IUH and (b) in the opposite direction, as evidenced by HADDOCK docking.

The obtained proto-models of KID^{pY721}/SH2 complex, CM1 and CM2, were explored using the Gaussian accelerated Molecular Dynamics (GaMD) simulation[260,261] with restrains applied to the distances between the phosphorus atom (KID^{pY721}) and nitrogen/oxygen atoms of R340, R358 and S361 (SH2). Restraint distances were gradually diminished during a stepped 350-ns GaMD simulation then removed entirely (**Figure 8.4**, **A**).

After constraints relaxation, the inter-protein distances between the phosphorus atom from KID pY721 and SH2 nitrogen atoms from R340 and R358 are well conserved until the end of the MD simulation (500 ns) in both models. However, the contact between the KID pY721 phosphorus atom and the SH2 S361 oxygen atom shows large variations in two models. Curiously, this distance fell regularly to the value observed before all constraints were relaxed, and even lower.

The RMSD curves and values of CM1 and CM2 models, as well as each protein that composes these models, display similar profiles (**Figure S37**, **A**). The RMSD of SH2 is stable in both models, and the principal contributor to the increase of RMSD values is KID^{pY721}. As expected, KID^{pY721} showed an overall decrease in fluctuations, except for the N/C-ends, compared to free KID^{pY721} (**Figure S37**, **C**).

The SH2 folding in complex KID^{pY721}/SH2 is globally well-conserved in both models, except for a coiled β1 and β2 stands linker, recognised previously as IDR F1. However, the character of the F1 reversible folding in the complex most resembles the p-pep/SH2 complex compared to the free-ligand SH2 (**Figure 7.6**, **E**; **Figure 7.7**, **B**; **Figure S37**, **B**). KID^{pY721} folding in both models, CM1 and CM2, shows intrinsic disorder in a large portion of protein, except a stable αH1-helix, like free KID^{pY721}. Nevertheless, some distinguishable differences are likely complex-specific. In that manner, KID^{pY721} showed that (i) in CM1, the small β-sheet (T718-V728) is well conserved during simulation with and without constraints, while in CM2, it was transformed in the systematically unfolded state (turn; coil) in the unconstrained GaMD simulation; (ii) in CM1, the reversible H5 (310-helix; turn; coil) is folded as a well-stable αH5-helix in the range of 230–500 ns, while in CM2, this fragment shows the highly reversible structure. Focusing on the secondary structures' evolution of KID p-pep during GaMD, we noted

that after removing the restraints, its folding is drastically decreased in CM1 (random coil) and increased in CM2.

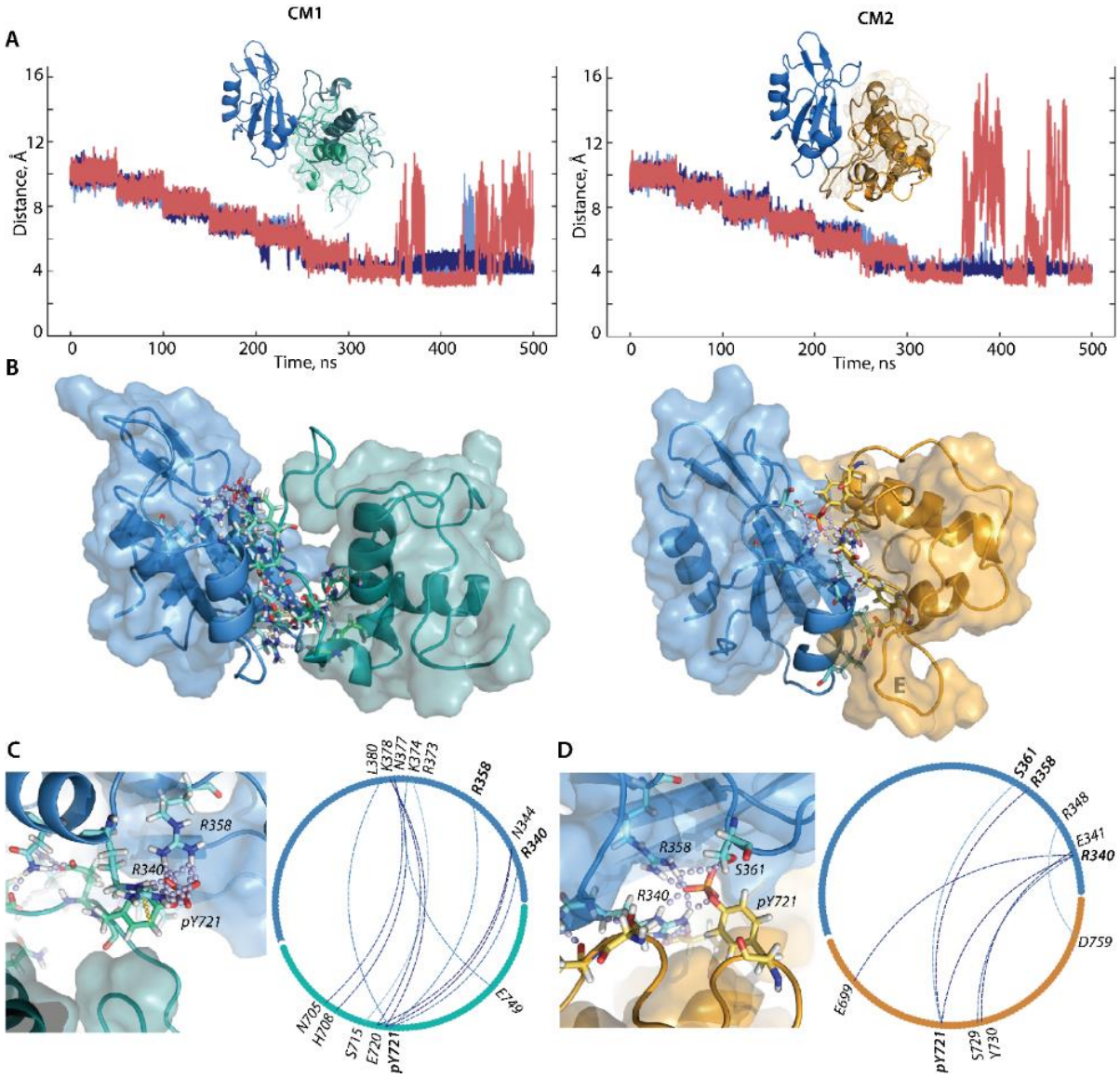


Figure 8.4 Intuitive modelling and GaMD simulation of molecular complex KID^{pY721}/SH2, represented by two alternative models in which the KID TNEYMDMK peptide is oriented similarly to p-pep in structure 2IUH (CM1, left) and in the opposite direction (CM2, right). (**A**) Distance variations between the phosphorus atom (KID^{pY721}) and nitrogen/oxygen atoms of R340 (blue light), R358 (dark blue) and S361 (SH2) (red). Each model of non-covalent complex and its time-dependent evolution (snapshots taken at t = 0, 75, 125, 175, 225, 275, 325 and 500 ns) is shown in insert. (**B**) 3D models of complex KID^{pY721}/SH2 at 500 ns of GaMD simulation. (**C**-**D**) Inter-protein H-bonds stabilising the KID (teal) and PI3K SH2 domain (blue) in CM1 (C) and CM2 (D) are showed as 3D structure (left) and a string diagram (right) displaying the high (≥ 80%, in dark blue) and low (≤ 50%, in blue light) probability of these contacts. Proteins are shown as surfaced cartoon, KID in teal (CM1) and orange (CM2), and SH2 in blue. The interface residues are shown as sticks, non-covalent contacts as dashed lines, lilac for H-bond and yellow for van der Waals contacts.

Intermolecular contacts formed during the unconstrained simulations showed that KID^{pY721} and SH2 in the CM1 model are linked by multiple H-bonds (11 contacts) of high probability (with occurrence ≥ 80%, 6 H-bonds), or low probability (with occurrence between 30 and 50%, 5 H bonds), which together form a very large interface of protein-protein interaction (**Figure 8.4**, **B**, **C**).

As expected, the major contributor to the CM1 interface contact network is two salt bridges formed by pY721 (KID^{pY721}) with R340 and R358 (SH2). These salt bridge interactions pulled together the disordered region of KID^{pY721} and two SH2 regions—the dynamically disordered linker F1 (through the interaction with R358), and the SH2 αH1helix (by interaction with R340). Surprisingly, these salt bridges are low probability events, and contact pY721···S361 is nearly disappeared (probability ≤ 30%). The salt bridge interactions are completed by H-bonding of neighbour residues to pY721, E720, and T718, each acting as a bifurcated centre. As shown, E720 interacts with N344 and K379, and T718 bounds N344 and N377. The H-bonds formed by KID^{pY721} residues N705 and H708 (from the stable αH1-helix) with N377 and K379 (from the β-sheet core of SH2), produce the second interface area of contacts, spatially separated from those formed by pY721 and its closest residues.

The CM2 KID^{pY721}/SH2 interface is formed by a limited number of H-bonds (seven contacts) and substantially restricted the number of residues contributing to binding (R340R361 from SH2 and E699-Y730 from KID) as well as structural elements involved (**Figure 8.4**, **B**, **D**). First, KID pY721 interacts with R340, R358 (the high probability contacts), and S361 (the low probability contact) of the SH2 domain. This principal binding interactions motif between KID^{pY721} and SH2 is completed by H-bonds between (i) R340 (αH1-helix of SH2) with E699 (αH1-helix of KID^{pY721}) and S729 (disordered H3 of KID pY721), making R340 a trifurcate centre, and (ii) E341 (αH1-helix of SH2) with Y730 (disordered H3 of KID^{pY721}). These contacts showed that in CM2, the recognition between two proteins, KID^{pY721} and SH2, is maintained by strong and stable interactions formed by two salts bridges and crosswise H-bonds involving the limited protein regions.

We also note the reduced size, estimated by the radius of gyration (Rg), of the KID^{pY721}/SH2 complex represented by the CM2 model with respect to CM1 (**Figure S37**, **D**).

8.1.3. Discussions

We attempted to describe the RTK KIT INTERACTOME by modelling a set of macromolecular complexes formed by KIT with its cellular proteins partners (PPs) involved in signal transduction. To initiate this modelling, we built a 3D model of the first molecular complex formed by the most phosphotyrosine-rich kinase insert domain (KID) of RTK KIT with a phosphatidylinositol 3 kinase (PI3K) SH2 domain

binding preferentially to KID.

Many interacting protein partners, RTK KIT and PI3K, are intrinsically disordered proteins (IDPs) with a modular structure. Indeed, as we previously reported, the multidomain RTK KIT comprises the sub-domains (JMR, TKD, KID, and C-term), whether structurally well-ordered or intrinsically and extrinsically disordered[271,290]. Similarly, PI3K from the non receptor tyrosine kinases Src-family, constitutes a typical example of modular architecture: an amino-terminal SH3 and SH2 domains, flanking a kinase domain by intra-molecular SH3-binding and SH2-binding sites[12,367].

The modular architecture of protein structures is advantageous for a more efficient execution of their functional activity (e.g., allosteric regulation of protein–protein interactions, involved in cell signalling). Some parts of such interaction interfaces participate in the information transfer (inter-protein communication), while other interaction regions appear to contribute only binding affinity (switching)[380,380,381].

It is well known that ~40–60% of the human proteome appears to be composed of protein domains/regions that are intrinsically disordered[43,308,437]. IDPs are paradigmatic challenges because: they are disordered in their inactive state[43,437,440] can fold partially or fully upon the biological effectors binding[308,441] can bind selectively diverse partners[442,443] and exhibit allosteric regulation without a well-defined quaternary or even tertiary structure[101,103].

These three fundamental properties—modularity, intrinsic disorder, and allostery—provide proteins with a finely regulated molecular mechanism that illustrates how nature can govern cellular signalling[44,101,300,308,403,444].

Each module of multidomain proteins studied, KIT and PI3K, may possess structural, dynamic, and functional independence, as was evidenced for KIT KID[290] (see also sections 0, 6.2 and 0) and SH2[445] (see section 7.2).

To construct KIT INTERACTOME through KID^{pY721}/SH2, and prior to the docking of these domains, the bench test performed on the empirically determined structure 2IUH (co-crystalized complex of a KID^{pY721} peptide and PI3K SH2 domain) indicates a good reproducibility of the crystallographic solution. Such result was obtained with High Ambiguity Driven protein–protein DOCKing (HADDOCK)[248], which uses biophysical interactions data—in our case, the H-bond contacts between p-pep and SH2 domain—to drive the docking process.

The benchmark structure successful reproducibility encouraged the HADDOCK docking of KID into the SH2 domain, represented by the SH2 structure from 2IUH and the most probable MD conformation of the free-ligand SH2.

Independently from the SH2 structure, the obtained docking solutions show different positions of KID^{pY721} (and its p-pep) with respect to the SH2 binding pocket.

Although the major p-pep docking position in the SH2 binding pocket matches its position in the structure 2IUH, KID^{pY721} shows a circular-distributed orientation of its p-pep around the SH2 binding site. These discouraging docking solutions may be attributed to (i) HADDOCK, which uses the quantitative measures of binding modes, such as the number of clusters, score, and population that do not allow comparisons between solutions reflecting conformational and orientational characteristics, (ii) the incorrectness of the X-ray solution, or (iii) the basic difference in the binding of peptide and protein to a target.

We suggest that p-pep can be located and oriented differently compared to its crystallographic structure pose. On one side, the p-pep pseudo symmetry concerning pY721, is issued from the biophysical properties' similarity and this fragment's high inherent conformational flexibility in a solution. On the other side, the double pseudo symmetry of the SH2 binding pocket is due to the similar biophysical properties of the pocket surface residues. Moreover, it is well known that peptides exhibiting either extended conformation or adopting β-turn or α-helix as a motif for target recognition can be completely buried in cavities, making multiple high-affinity interactions with a target[446], while the interaction interface between the two proteins is limited and involves significantly fewer amino acids from each partner, usually defined as 'hot-spot' residues and making the largest contributions to complex formation[447,448].

We suggested that to build the molecular complex composed of two intrinsically disordered proteins, which couple folding and binding possesses [72], an alternative strategy may be used. The empirically resolved structure 2IUH, and the high conservation of the pY721 binding with residues R340, R358, and S361 in the solid state and water solution can be used for the intuitive user-guided building of the p-KID/SH2 complex as reference supports. Considering the high conformational variability (structural instability and flexibility) of both proteins, KID and SH2, in solution, the MD conformations of KID^{pY721} and free-ligand SH2 are the most appropriate starting structures in such a study.

A direct use of the X-ray 2IUH structure is not a dogma for modelling the KID^{pY721}/SH2 complex, but it can serve as a reference for the initial positioning of KID^{pY721} concerning SH2.

Two models of the KID^{pY721}/SH2 complex with KID^{pY721} positioned in front of the SH2 cavity, but alternatively oriented, either coherent to the p-pep orientation in structure 2IUH (CM1) or oppositely oriented (CM2), were further explored by accelerated (GaMD) simulations. The preliminary results showed that in both probed models, the proteins are bonded by a combination of salt bridges and hydrogen bonds, when approaching different domains from both proteins. These inter-domain interactions create a small binding cleft, including a few residues in model CM2, while in CM1 the inter-protein interface represents an area twice larger on each protein and spans widely spaced amino acids in protein sequences and structures.

Although a very compact and regular CM2 interface model as well as an increased helical folding of the KID T718–K725 fragment and the reduced size of the KID/SH2 complex are very attractive arguments for the choice of this model as functionally related, there are still doubts in such a conclusion.

For an objective assignment of the functionally related model, we are now engaged in extended unconstrained MD simulations (2–3 μs) of both models to generate necessary and sufficient data for detailed comparative analysis of their structural, dynamical, and recognition properties as well as an accurate estimation of the binding energy.

Returning to the interacting proteins, RTK KIT and PI3K, regarded as modular disordered proteins, it seems that molecular modelling and molecular dynamics simulations provide powerful tools for the exploration of such proteins and their complexes. Such a study will be most effective when analysed in close conjunction with experiments on a protein function, which would play an essential role in validating and improving the modelling and simulations.

8.2. MODELISATION MOLECULAIRE DE HVKORC1 ET PDI POUR L'INITIATION DE LA REACTION D'ECHANGE THIOL-DISULFURE

Résumé. *Pour effectuer sa fonction de recyclage de la vitamine K, hVKORC1 doit être activé par une protéine redox partenaire dont la plus probable est PDI selon notre prédiction et les récentes études empiriques. Cette activation nécessite la reconnaissance et la liaison de ces deux protéines. Sans structure empirique du complexe, et connaissant les contraintes stériques nécessaires à une réaction spontanée de transfert d'électrons, nous avons proposé deux modèles du complexe hVKORC1/PDI différant par l'orientation des fragments putatif F1 et F2 de PDI (identifiés* in silico*) par rapport à l'axe principal de hVKORC1. Pour prendre en compte le désordre de la boucle L et des fragments de PDI, nous avons modélisé ces deux complexes non covalents (M1 et M2) par simulation de dynamique moléculaire avec l'application d'un gradient descendant de distance S···S entre les deux protéines. Ces complexes ont également été reproduits par amarrage protéine-protéine sur une conformation complète du hVKORC1 ou sur une conformation de la boucle L seule, confirmant nos hypothèses sur la modularité structurale et fonctionnelle de hVKORC1 et la justesse de nos modèles construits par la méthode* user-guided*. La simulation de dynamique moléculaire plus poussée de ces deux complexes a montré que seul le modèle M1 répond aux critères de stabilité et de surface d'interface nécessaires à un contact prolongé entre les deux protéines. Ainsi, nous avons proposé ce modèle comme le complexe précurseur non covalent pour l'étude des réactions d'échange thiol-disulfure entre hVKORC1 et PDI, un premier pas vers*

la modélisation des états intermédiaires et finaux de l'INTERACTOME de hVKORC1.

8.2.1. INTRODUCTION

Thiol-disulphide exchange reactions are central to oxidative protein folding and a key mechanism in almost all enzymes generating and isomerizing disulphide bonds[226]. Understanding the mechanisms of thiol-disulphide exchange still remains a significant intellectual challenge 50 years after the classic studies of Anfinsen and colleagues on refolding of reduced ribonuclease A (RNase)[446]. The reaction is initiated by the nucleophilic attack of a thiolate on a disulphide (**Figure 8.5**). An attacking thiolate approaches along the disulphide axis and this requirement for collinearity establishes the orientation necessary for interactions between well-structured redox partners[227,228]. Thus, disulphide exchange reactions have significant steric requirements that must be met by enzymes capable of adapting their folding at each step of the process.

By comparing four thioredoxin proteins as promising redox partners of hVKORC1, we suggested that human Protein Disulphide Isomerase (PDI) is the most probable hVKORC1 redox protein, and *de novo* model of hVKORC1 in the oxidised inactive state[220] may be used as its target. Moreover, we identified the hVKORC1 fragment as the most probable PDI binding site, and two possible PDI fragments, F1 and F2, which may recognize this binding site. Further, we aimed to evaluate these *in silico* predictions to answer: How do the predicted results correspond to a model of the complex formed by hVKORC1 and its possible partner?

A central goal of this study is to understand, at the atomistic level, the recognition mechanisms between Trx and hVKORC1 (a process preceding the electrons' transfer reaction) and, thereby, identify shared vulnerable sites that can be targeted with anti-hVKORC1 or anti-Trx therapeutics.

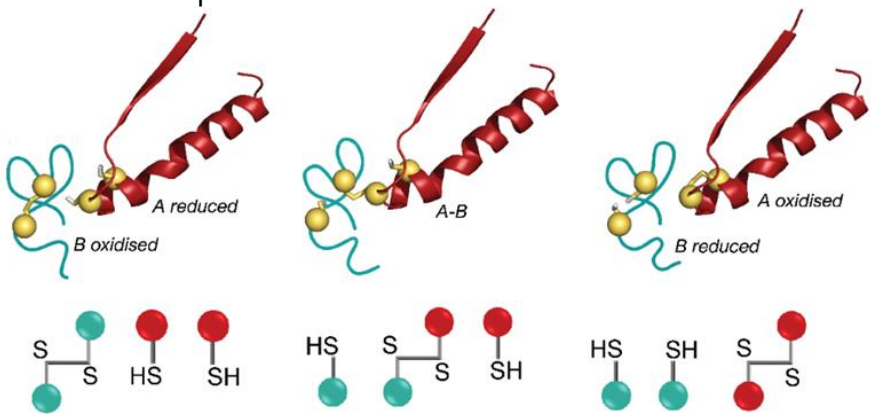


Figure 8.5 Thiol-disulphide exchange reactions between PDI and hVKORC1. Thiol-disulphide exchange reactions involve reduced (proton-coupled) PDI (A, in red) and oxidised hVKORC1 (B, in deep teal). Generation of mixed-disulphide intermediate A-B involves an in-line transition state.

Mechanistic models (top panel) and simplified scheme (bottom panel) of states A, A-B and B show a reduction of pair of cysteine residues of oxidised protein (A, hVKORC1) by a CX1X2C dithiol motif (B, PDI).

8.2.2. RESULTS

8.2.2.1. USER-GUIDED MODELLING OF HVKORC1/PDI MOLECULAR COMPLEX

The molecular complex of hVKORC1 was constructed with PDI to probe our hypothesis on the identification of a hVKORC1 redox partner (see the Discussion section); 3D models of the complex were constructed using the crystallographic structure of the VKOR from bacteria (bVKOR; PDB ID: 4nv5[200]) as a reference for the initial positioning of PDI relative to hVKORC1.

To be most objective in the modelling of the human PDI–VKORC1 complex, the structure of bVKOR was not used as a template because of (i) the suggested alternative VKOR activation mechanisms in bacteria and in eukaryotes, that is, in their respective native environments, which employ significantly different mechanisms for electron transfer[449] (ii) a high structural difference between the Trx- and L-loop domains in bVKOR and human proteins (RMSD values are 4.5 and 4 Å between bVKOR and the "closed" and "open" conformations of hVKORC1, respectively), (iii) very low sequence identity/similarity (15/20%), and (iv) a very large distance between the cysteine residues from the Trx-like and VKOR-like domains (the minimal S···S distance of 16 Å) in bVKORC1 (**Figure S38**).

For modelling the human PDI–VKORC1 complex, a conformation of hVKORC1 with the most extended "open" L-loop (the least probable conformation) was chosen as the initial target structure in order to bring the two proteins as close as possible. As for the initial PDI model, the conformation with a well-ordered and long αH2-helix that is similar to the X-ray structure of PDI[426] was chosen and positioned above hVKORC1 so that (i) the distance between the sulphur atoms from C37 of PDI and C43 of hVKOR1 was as short as possible (12.5 Å) and (ii) each PDI fragment that was suggested to be a fragment able to form the intermolecular interactions with a target, namely, F1 and F2, was alternatively placed above the middle of the L-loop surface. The obtained proto-models, Model 1 and Model 2, were explored using MD simulation for conditions (see the Methods section), where restraints apply to the distance S···S between C37 (PDI) and C43 (hVKORC1). The restraints were gradually diminished during a stepped 80-ns MD simulation run (**Figure 8.6**).

For both models, structural rearrangement occurred inside each protein and between the proteins, with diminishing S···S distance. In Model 1, the extended "open" conformation of the hVKORC1 L-loop was observed at an S···S distance of 12 Å, which then adopted a "closed" conformation at a shortened S···S distance (of 10 and 8 Å),

with the αH1-L helix and the L23 linker located in a proximal position, which is the most probable conformation of the L-loop in isolated hVKORC1. The initially well-ordered and long αH2-helix of PDI is rotated by 30° (at an S···S distance of 10 Å), followed by the bending of the helix, and then (at the S···S distance of 8 Å) by depletion of two helices, a small 310-helix in the proximity of the CX1X2C motif and a shortened αH-helix, which demonstrates a folding–unfolding effect observed in MD simulations of PDI in an isolated state.

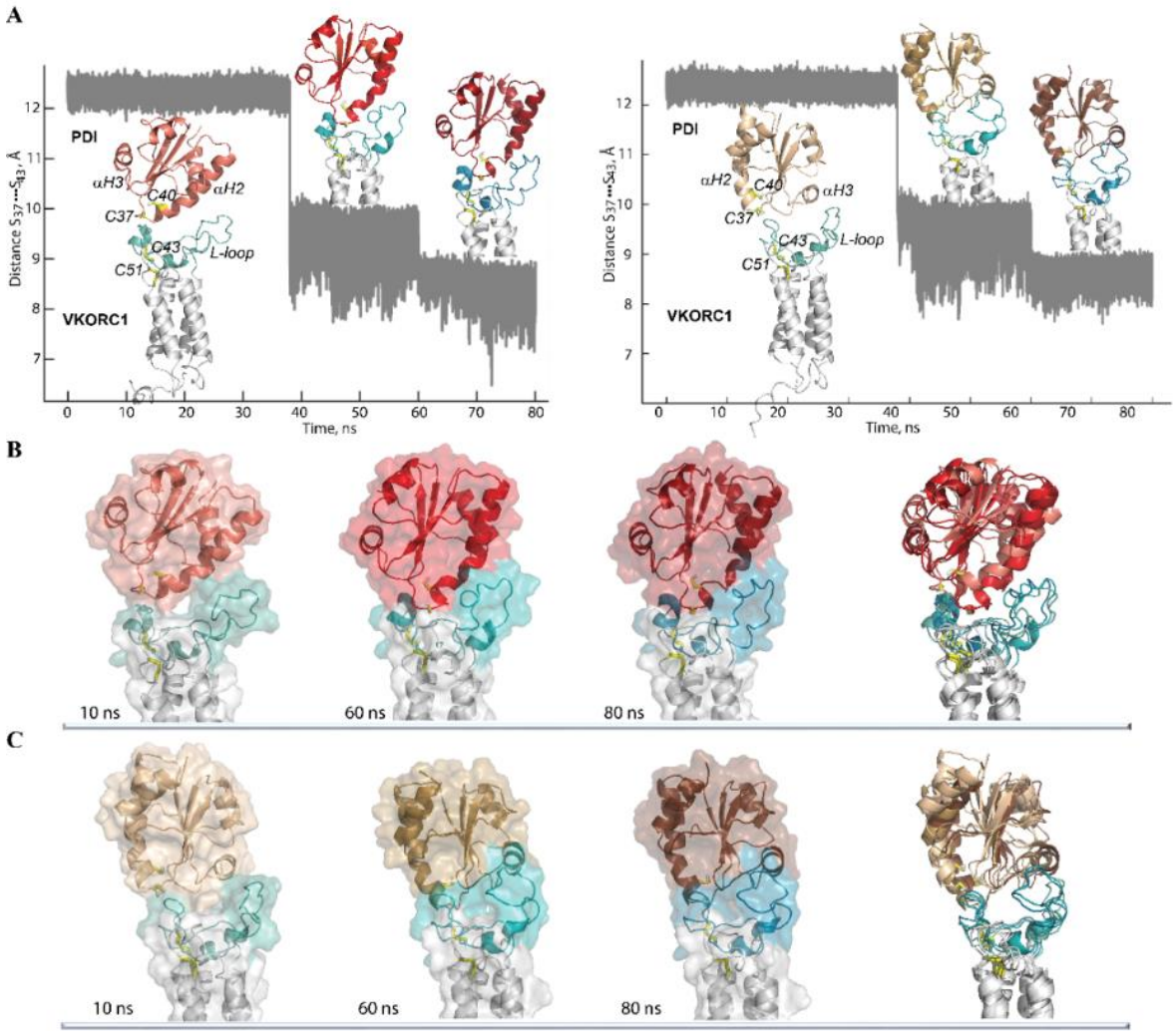


Figure 8.6 Modelling of human PDI–hVKORC1 complex. (**A**) MD simulations of 3D model PDI–VKORC1 complexes were performed, with gradually diminished distance (from 12.5 to 8.0 Å) between the sulphur (S) atoms of C37 from PDI and of C43 from the L-loop of hVKORC1. PDI has two orientations with respect to hVKORC1, with F1 (Model 1, left) and F2 (Model 2, right) positioned above the middle of the L-loop surface. Both models of the PDI–hVKORC1 complex are shown as snapshots taken at t = 10, 60 and 80 ns, with different S···S distances. The reference residues and fragments are labelled. (**B**, **C**) Conformations of the PDI–hVKORC1 complex, with two different PDI orientations, chosen at t = 10, 60 and 80 ns, and their superposition at all three times. In (A–C), the proteins are depicted as ribbons or as ribbons and surfaces and are distinguished by colour: a red palette was used for PDI and a cyan palette for hVKORC1, both nuanced by the tonality from light to dark to distinguish the conformations chosen at t = 10, 60 and 80 ns.

Similarly, in Model 2, a gradually diminishing S···S distance from 12 to 8 Å promotes a change in the L-loop conformation from "open" to "closed" in hVKORC1, while in PDI, a departure of the αH3-helix from its initial position to the location most exposed to the solvent (a 4.5–5.0 Å parallel displacement of the helix) was observed. The conformational changes observed during the simulations of the two PDI–hVKORC1 complex models are reflected in the folding of "interacting" proteins. The extended "open" conformation of the hVKORC1 L-loop, taken as the initial structure for complex modelling, showed increased folding (by 50%) in Model 1, with a decrease in the S···S distance from 12 to 8 Å, while in Model 2, its helical fold was reduced by 40% (**Table S9**). As for PDI, the folded content of its initial and final conformations was the same for both models.

Analysis of the intermolecular contacts at the interface between PDI and hVKORC1 (in the conformation taken at t = 80 ns) showed that these two proteins in Model 1 are linked through two salt bridges formed by R61 (hVKORC1) and E46 (PDI) and by D67 (hVKORC1) and K49 (PDI) (**Figure 8.7**, **A**). Hydrophobic contacts were also observed between two pairs of residues: A42 (PDI) and G62 (hVKORC1) and P45 (PDI) and L65 (VKORC1). Moreover, G62 (hVKORC1) interacts with P45 (PDI). The PDI–hVKORC1 interface interactions are completed by an H-bond between the side chain of S57 (hVKORC1) and the main chain of G38 (PDI), the amino acid in the proximity of the CX1X2C motif, and by the hydrophobic interaction between V45 (hVKORC1) and G82 (PDI). All distances between the interacting D···A atoms were ranged from 2.5 to 3.2 Å, which characterise strong interactions.

Spatially, two sets of interactions stabilising the PDI–hVKORC1 complex were observed. The first set, which is composed of S57(hVKORC1)···G38(PDI) and V45(hVKORC1)···G82(PDI), is localised in the proximity of the active sites, the CGHC motif of PDI and disulphide bridge C43–C51 of hVKORC1 and probably stabilises their close location, which is induced, in part, by a steric requirement imposed on the sulphur atoms from C37 and C43 to be in the closed position. The second set, which is composed of multiple contacts between the residues from short sequence segments, A42-K49 from PDI and R61-E67 from hVKORC1, forms a very compact regular interaction pattern that describes the highly specific recognition between two molecules that are maintained by two salt bridges and by crosswise hydrophobic interactions. This pattern of interactions stabilises the αH2-helix of PDI and the L23 linker from hVKORC1 in a close position that is independent of any interaction with Set 1 and, consequently, may present a first step in the PDI–hVKORC1 recognition process.

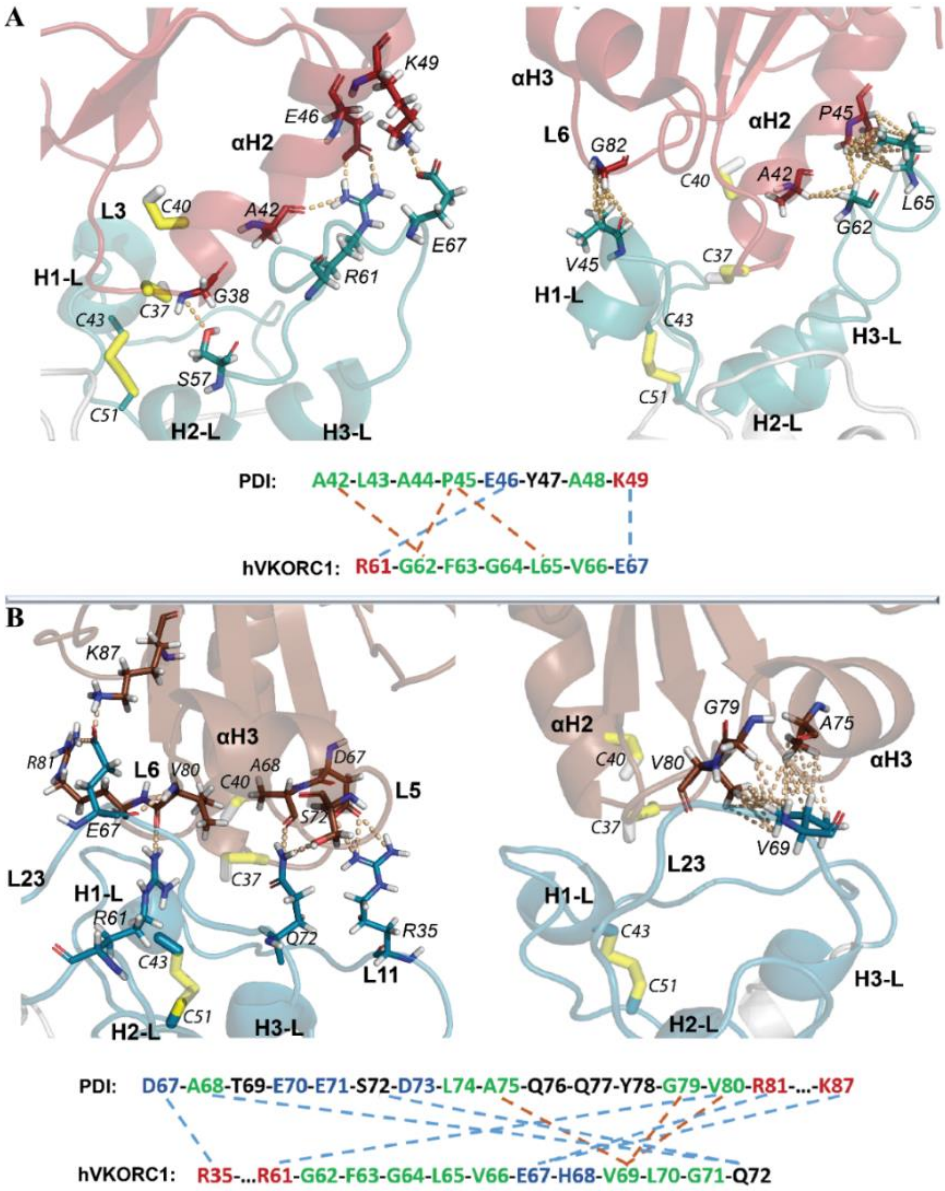


Figure 8.7 Intermolecular contacts at the interface between PDI and hVKORC1 in two models of the PDI–hVKORC1 complex. The intermolecular H-bonds and hydrophobic contacts between PDI and hVKORC1 in Model 1 (**A**, top) and Model 2 (**B**, top). (A, B) The proteins are shown as coloured ribbons: PDI in red and brown and hVKORC1 in cyan (L-loop), with the interacting residues and thiol groups as sticks. The contacts are indicated by dashed lines: H-bonds in yellow and hydrophobic contacts in salmon. The structural fragments and residues participating in the contacts are labelled. Analysis of the intermolecular contacts was performed on conformations taken at t = 80 ns. (A, B) A pattern of H-bond (in blue) and hydrophobic (in orange) contacts between the PDI and hVKORC1 residues (bottom). Residues are coloured according to their properties: the positively and negatively charged residues are in red and blue, respectively; the hydrophobic residues are in green; the polar and amphipathic residues are in black.

The residues of hVKORC1 that form salt-bridges and H-bonds are located on the transient H2-L helix and on the L23 linker, which is composed of a segment that was predicted to be the most putative recognition region in an isolated hVKORC1. Similarly,

PDI residues participating in hVKORC1 recognition belong to the F1 fragment were regarded as a possible putative recognition site. Surprisingly, a hydrophobic interaction with V45 (hVKORC1) is formed by G82 (PDI), which is a residue from the F2 fragment that is also predicted to be a fragment that contains possible recognition sites.

In Model 2, the interaction interface between PDI and hVROR1 is also formed by two salt bridges generated by R35 (hVKORC1) and D67 (PDI) and by E67 (VKORC1) interacting with R81 and K87 from PDI. Other electrostatic interactions are presented by the H-bonds of Q72 (hVKORC1) with A68 and S72 from PDI and of R61 (hVKORC1) with V80 of PDI. Hydrophobic contacts are observed as a three-furcate interaction of the three PDI amino acids (A75, G79 and V80) attached to a unique amino acid (V69) of hVKORC1. Unlike the compact interface contact network in Model 1, the interacting residues in both proteins of Model 2 are distributed over large sequence segments, from D67 to K87 in PDI and from R35 to Q75 in hVKORC1. This highly enlarged interface interaction network seems less probable because of the small probability of a synchronised approach of two space-separated binding sites to the target.

It is interesting that two amino acids, R61 and E67, of hVKORC1 form salt bridges in both models, Model 1 and Model 2, but by selecting different PDI residues. Remarkably, both amino acids belong to a hVKORC1 segment that is predicted to be the putative recognition site by analysis of the isolated protein.

Although very compact and regular, the interface interaction pattern formed by the closely localised residues in both proteins from Model 1, together with the increased helical folding of the L-loop by 50%, is a very attractive argument for the choice of this model for being functionally related, though there is still doubt in such a conclusion.

Other characteristics are considered to better justify or challenge our hypothesis. First, from a superimposition of each model on the experimentally defined structure of bVKORC1, the best fit at the level of Trx-like domain orientation with respect to hVKORC1 is observed for Model 1 (**Figure S39**), but analysis of the interaction between the Trx-like and VKOR domains in the bacterial protein showed only a single short contact (between Q40 from the αH2-helix of Trx and L46 of the L-loop), an observation that largely mismatches the interaction patterns observed in both models.

Finally, to check the stability of the interactions between the two proteins in Model 1 and Model 2, the models were simulated at t = 80 ns under more relaxed (“soft”) conditions, which gives more tolerant restrains on the distance S···S between C37 (PDI) and C43 (hVKORC1).

In the two MD simulations of Model 1, which have different “soft” constraints (a time range of 80–100 ns), the distance S···S either varied within an enlarged range (7–11 Å) or, surprisingly, showed a tendency to decrease (6–10 Å) with respect to the

simulation with a more "hard" restriction (a time range of 60–80 ns) (**Figure S40**, **A**). The MD conformations of Model 1, generated using different "soft" constraints, showed very similar structures of PDI–hVKORC1 that differed only in the folding of the H2-L helix from the L-loop of hVKORC1 and the αH2 helix of PDI. Each of these structural effects was observed in isolated proteins. The interface interactions between the residues from the αH2 helix of PDI and L23 from the L-loop of hVKORC1 were very similar for conformations taken at t = 100 ns and t = 80 ns. With respect to the conformation chosen at t = 80 ns, some novel contacts involving residues from H2-L (hVKORC1) and the L3 loop and of PDI are observed in the conformation taken at t = 100 ns (**Figure S41**).

These results show that the highly specific recognition between the two molecules is maintained by the strong and stable interactions formed by two salt bridges and by crosswise hydrophobic interactions preserved in Model 1.

The MD conformations of Model 2, generated using "soft" and "hard" constraints, showed similar structures of PDI–hVKORC1, which differed only in the position of the αH3 helix of PDI and the L-loop of hVKORC1 (**Figure S40**, **B**). The interactions observed at the interface between the two proteins are not preserved, except for a single salt bridge between E67 (hVKORC1) and R81 (PDI) (**Figure S42**).

8.2.2.2. PROTEIN-PROTEIN DOCKING OF HVKORC1 ONTO PDI

To evaluate hVKORC1 conformations as putative targets for Protein Disulphide Isomerase (PDI) suggested as redox partner[221], we used High Ambiguity Driven protein-protein DOCKing (HADDOCK)[248]. Unlike other protein-protein docking approaches, based on combination of energetics and shape complementarity, HADDOCK uses biophysical interactions data, in our case, a short distance between sulfur atoms from cysteine residues of two interacting protein, PDI and hVKORC1, to drive the docking process.

In the docking analysis, PDI per se is an invariable component taken from [221], while hVKORC1 is a variable item that can be any randomly chosen conformation generated by cMD simulation of the theoretical model *(de novo)* or crystallographic structure. We suggested that using different target conformations will help discriminate an authentic conformation specific to its ligand. This study was carried out with the aim of answering the essential question: what conformation of hVKORC1 is an optimal target for PDI?

Prior to docking studies, we performed a bench test to investigate if docking with HADDOCK can reproduce the predicted *de novo* PDI-hVKORC1 complex? Docking trials were performed using the published *de novo* structural model (Model 1) of PDI-hVKORC1 complex[221] as benchmark set. The theoretical model (*de novo* 3D model)

application as a reference is imputable to the absence of empirical structural data for PDI-hVKORC1 complex. PDI-hVKORC1 complex (Model 1) was separated into unbound proteins and docked with HADDOCK and Ambiguous Interaction Restraints (AIRs)[248], using a pair of cysteine residues thiol groups, C37 from PDI and C43 from hVKORC1, as active centres to drive the docking process. For objectivity, each protein was considered as a target and as a ligand.

Docking of PDI as a ligand into hVKORC1 as a target (scenario i), showed two clusters, C6 and C9 (numbered by HADDOCK) formed with models of PDI-hVKORC1 complexes similar to a benchmark structure (**Figure S43**). Docking of hVKORC1 as a ligand into PDI as a target (scenario ii), does not lead to a benchmark solution. In both scenarios for a ligand-target pair, (i) and (iii), both docked proteins structures and conformations are well conserved with maximal RMSD values of 0.4 (PDI) and 1.0 Å (hVKORC1).

Curiously, two alternative scenarios in a docking of ligand-target pair give rise to largely different solutions. Focusing on solutions produced for case (i), we noted that PDI is located above L-loop and occupies an approximately similar spatial position for HADDOCK solutions and benchmark, changing only its orientation resulted from rotation of PDI around the central hVKORC1 axis. Only exception is cluster C10, where PDI is located on one side of L-loop in a position perpendicular to hVKORC1 central axis. In case (ii), the position of PDI on one side of L-loop is observed in most docking solutions (6 clusters out of 10). Since such solutions are not compatible with the membrane position, they have not been considered. A HADDOCK solution was classified as interpretable if PDI position matched the Trx domain position in hVKORC1bacterial homolog, and human VKORC1 uses the same electron transfer pathway as its bacterial homologues[224]. Docking trials showed that HADDOCK reproduces the benchmark model (**Figure 8.8**).

Comparing the obtained results for two docking scenarios, we found scenario (i) is the right choice leading to the benchmark solution. Therefore, this scenario was further used in docking examinations.

Two strongly different hVKORC1 conformations with L-loop compact (closed, most probable) and elongated (open, least probable) shapes were randomly chosen from data generated by cMD simulations of *de novo* model of hVRORC1[221] and used as targets for PDI docking. The docking results show that both conformations of hVKORC1, with elongated (open) and compact (closed) L-loop produced solutions where PDI is positioned in the same space volume as in the benchmark complex, however, PDI orientation with respect to the target is highly divergent.

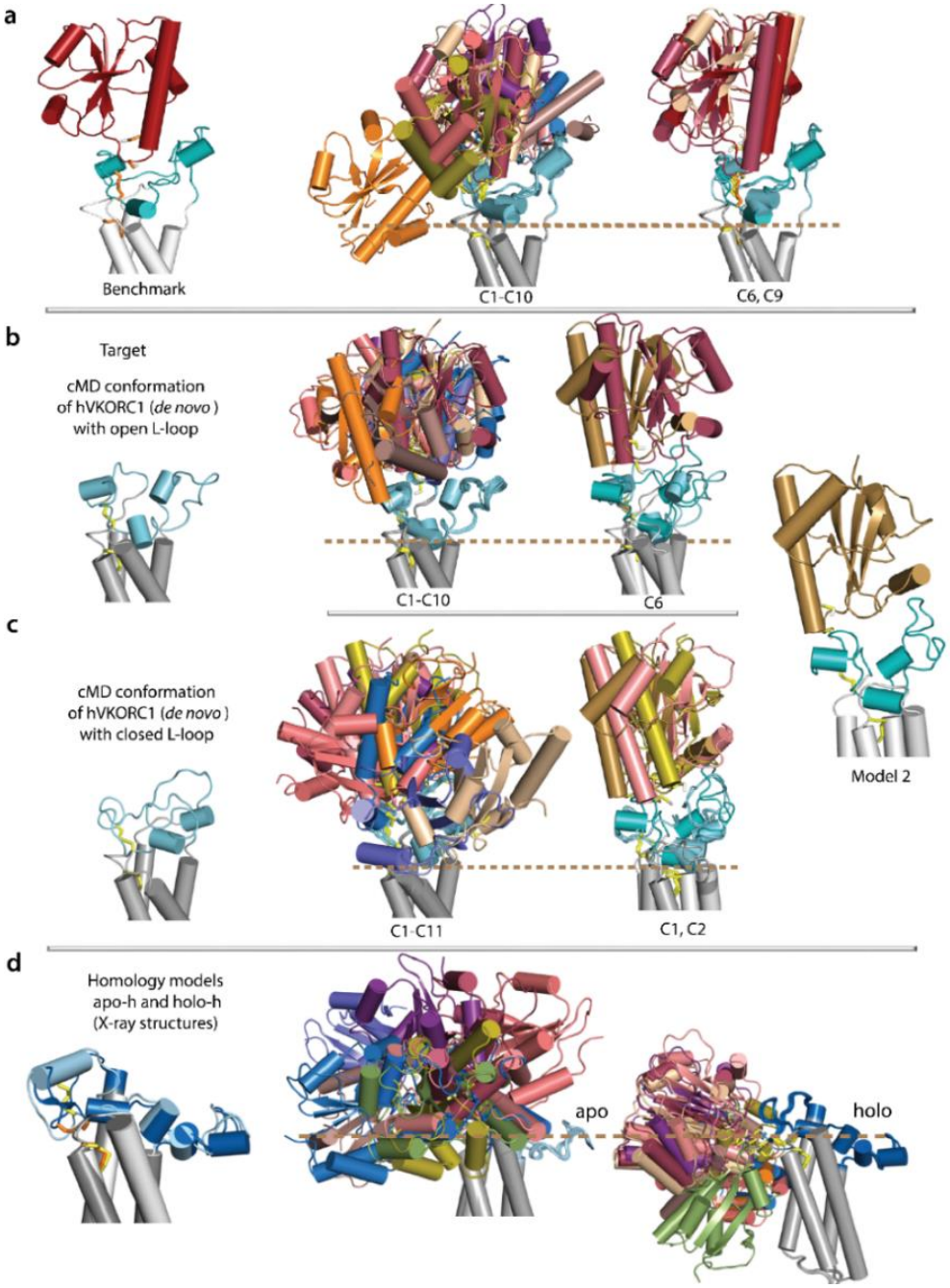


Figure 8.8 Computational protein-protein docking of PDI (ligand) onto hVKORC1 (target) performed with HADOCK using an information–driven method. (**a**) Benchmark complex from [221] (left); Superimposition of the top 10 solutions (middle) and best solutions, clusters C6 and C9, (right) on benchmark. (**b**-**c**) (Left) hVKORC1 conformations with different L-loop shape, elongated (open, b) and compact (closed, c) conformations *de novo* model[220]. (Middle) Superimposition of the top solutions obtained for hVKORC1 with closed (b) and open (c) L-loop conformations. (Right) Model 2 suggested as possible in [221] used for superimposition of HADDOCK solutions, clusters C6, and C1, C2 obtained for hVKORC1 with open (b) and closed (c) L-loop conformations. Two orthogonal projections are shown. (**d**) hVKORC1 homology models, apo-h and holo-h, quasi-identical to X-ray structures (PDB IDs: 6wv3 and 6wvi) (left). Superimposition of the top of 10 HADDOCK solutions for PDI docked onto apo-h (middle) and holo-h (right) forms of hVKORC1. (a-d) Protein is shown as a cartoon with helices as cylinders and disulphide bridges in yellow sticks. A possible boundary of the membrane is denoted as dashed line.

The HADDOCK quantitative metrics of binding modes – number of clusters, score, and population – don't allow comparison between solutions for open and closed conformations, even if a simple superimposition of found solutions showed no benchmark solution was observed. Surprisingly, HADDOCK solutions for hVKORC1 forms showed close similarity to the alternative PDI-hVKORC1 complex Model 2 proposed in [221] as a possible solution (**Figure 8.8**). Moreover, the number of such interpretable solutions is greater for hVKORC1 with L-loop in closed conformation.

PDI docking onto hVKOR models apo-h and holo-h derived directly from crystallographic structures[218] did not produce the expected benchmark solution, nor solutions corresponding to the alternate model of the PDI-hVKORC1 complex. Moreover, many HADDOCK solutions have low compatibility with bacterial homologue of hVKORC1, the unique empirical structure in which VKOR and Trx-like domain are covalently bound[224].

Finally, to test if cleaved L-loop is a valid target for PDI docking, PDI was docked onto open- and closed L-loop cleaved from the respective *de novo* models. These docking experiments found again HADDOCK solutions corresponding to Model 2 and showed that (i) solutions are very similar to PDI docking onto the full-length hVKORC1, and (ii) closed cleaved L-loop is the best target of PDI (**Figure 8.9**). Similarly, PDI docking into L-loop cleaved from hVKORC1 models apo-h and holo-h derived directly from the crystallographic structures[218], did not produce any reasonable solution.

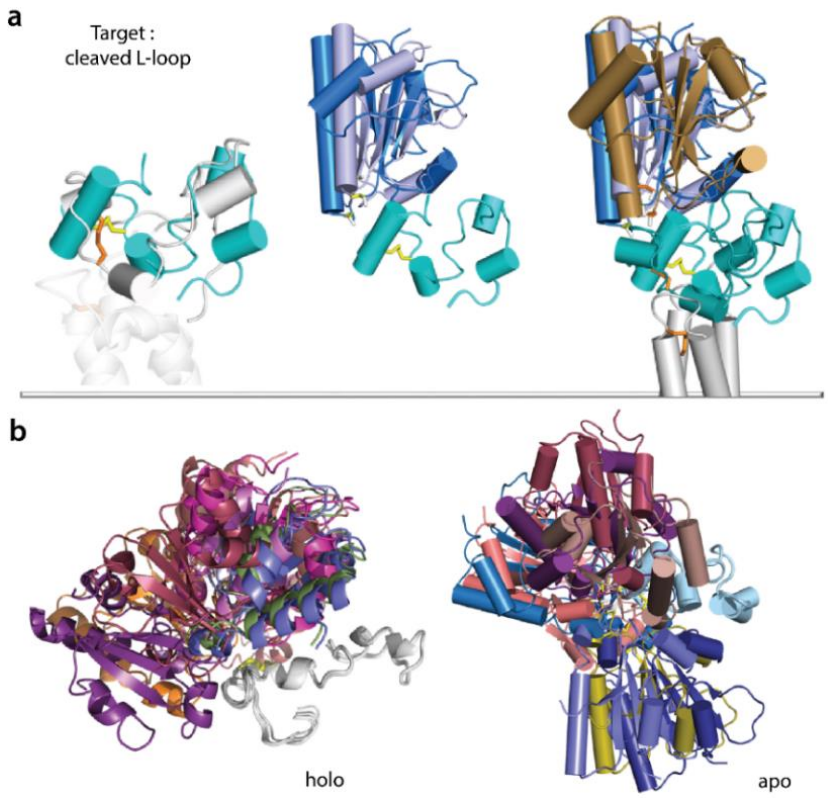


Figure 8.9 Protein-protein computational docking of PDI (ligand) constant conformation and cleaved L-loop (target) represented by different conformations. Docking was performed by

HADDOCK using an information–driven method. (**a**) (left) Superimposition of L-loop cleaved from Model 1 of PDI-hVKORC1 complex at t = 441 ns of cMD simulation (grey)[221] and taken at 4.62 µs of cMD simulation of cleaved L-loop from apo-h form simulated as an isolated species (teal). (Middle and right) The two best docking solutions obtained for PDI docking onto L-loop conformation with compact (closed) shape superimposed on L-loop from Model 2 (middle) and on Model 2 (right) suggested as possible in [221]. (**b**) Superimposition of the best solutions obtained for PDI docking onto cleaved L-loop from holo- and apo- forms (X-ray). (a,b) Protein is shown as a cartoon with helices as cylinders or spiral and disulphide bridges as yellow sticks.

8.2.3. DISCUSSIONS

As vitamin K is involved in various crucial biological processes[450,451], in particular in blood coagulation, an explicit understanding of each step leading to its transformation by hVKORC1 is required to control such processes in the context of its deregulation activity leading to severe diseases[194,452]. At a molecular level, the formation of a hVKORC1-redox protein molecular complex is a fundamental process leading to thiol-based redox switches, occurring primarily at an inter-protein level between cysteine residues of redox proteins and the L-loop and, further, inside hVKORC1, between cysteine residues of the L-loop and the highly conserved CXXC motif, to transform vitamin K epoxide to its reduced form, hydroquinone[312].

Molecular modelling and molecular dynamics simulations provide powerful tools for the exploration of proteins and their complexes.

8.2.3.1. ON THE USER-GUIDED MODELLING OF HVKORC1/PDI COMPLEX

Direct use of the X-ray structure of VKOR from bacteria (a protein with covalently bound Trx-like and VKOR-like domains, which has low sequence and structure similarity compared to the human proteins PDI and hVKORC1) is not appropriate for the modelling of the human complex but can be a reference for the initial positioning of PDI with respect to hVRORC1. Using conventional MD simulations, two models of the PDI–hVKORC1 complex, with the PDI in two alternative positions, which were either exposed by F1 (Model 1) or F2 (Model 2) in front of the L-loop of hVKORC1, were studied. In both probed models, proteins bind to each other using a combination of hydrogen bonds, salt bridges and hydrophobic contacts formed by residues from the different protein domains. These domains are small binding clefts and include a few peptides in Model 1, while in Model 2, the molecular interface represents large areas on each protein and spans widely spaced amino acids in protein sequences.

How do the "interacting" residues predicted by analysis of isolated proteins correspond to the contacts in the complex formed by hVKORC1 and PDI?

In Model 1, the interface contact network is composed of two salt bridges formed

by two pairs of charged residues, R61 and E67, from hVKORC1, which, together with S57, G62 and L65, also contribute to the stabilisation of the two proteins. These residues are amino acids from the L-loop segment that was predicted as a platform for recognition of a protein partner by hVKORC1. In Model 2, the interaction interface between PDI and hVROR1 is also completed by two salt bridges formed by R35 and E67 from hVKORC1 interacting with D67, R81 and K87 from PDI and by H-bonds formed by Q72 and R61 of hVKORC1 with A68, S72 and V80 of PDI. In both models, two amino acids of hVKORC1, R61 and E67, participate in strong electrostatic interactions, salt bridges or H-bonds but with different PDI residues. It is remarkable that both amino acids belong to a hVKORC1 segment that is predicted to be the putative recognition site from analysis of the isolated protein. The contacting PDI residues are mainly predetermined by PDI orientation with respect to the L-loop.

Based on limited data from the stepped finite-time simulations, is it possible to conclude which model is the correct one?

In both models, the optimised (enhanced) orientation of PDI with respect to hVKORC1 is maintained by the multiple interactions between the two molecules.

In Model 1, intermolecular contacts are observed between the two short length peptides, R61-E67 from hVKORC1 and A42-K49 from PDI, which form two salt bridges and three crosswise hydrophobic interactions. Such a compact regular interaction pattern may describe highly specific recognition between the two molecules, maintaining the αH2-helix of PDI and the extended L23 linker of hVKORC1 in a close position and, consequently, may present the first step in the PDI–hVKORC1 recognition process. The other set of interactions, S57(hVKORC1)···G38(PDI) and V45(hVKORC1)···G82(PDI), is located in close vicinity to the CGHC motif of PDI and disulphide bridge C43-C51 of hVKORC1. This is induced by a steric requirement imposed on the sulphur atoms from C37 and C43 that holds them in a closed position. Moreover, as these contacts are formed by the main chain atoms, they are rather nonspecific.

In Model 2, the interaction interface between PDI and hVROR1 represents a large area for each protein and spans long-spaced amino acids of the protein sequences (D67-K87 in PDI and R61-Q72, completed by R35 in hVKORC1). The two salt bridges, which are formed by R35 (L11 from hVKORC1) and D67 (L5 from PDI) and by E67 (L23 from hVKORC1) interacting with R81 and K87 from L6 of PDI, involve two regions on each protein that are separated by large distances in the sequence and the 3D structure. The other H-bonds involve the residues located between the two remote salt bridges. The dense cluster of hydrophobic contacts is realised as a three-furcate interaction of three PDI amino acids (A75, G79, and V80) attached to a single amino acid (V69) of hVKORC1.

In both models of the PDI–hVKORC1 complex, interacting hydrophobic motifs

from both proteins form “interacting hydrophobic cores”, which may be the key factors in the recognition process. The total number of noncovalent contacts between PDI and hVKORC1 in Model 2 is 9, while in Model 1, it is only 5. It was reported that the number of connections between each pair of proteins is a strong predictor of how tightly the proteins connect to each other[453].

Nevertheless, despite the large number of H-bonds and the dense cluster of hydrophobic contacts, it appears that the enlarged interface interaction network observed in Model 2 is less likely, due to the low probability of a synchronised approach of the two space-separated binding sites on PDI to the two space-separated binding sites on the target.

Moreover, based on the stepped simulations of Model 1, the diminishing distance between the two proteins promoted an increase in the helical folding of the L-loop by 50%, while in Model 2, its helical fold was reduced by 40%. While proteins become disordered on their own, their native conformation is stabilised upon binding[447,448]. The folded content of the initial and final PDI conformations is the same in both models; nevertheless, its conformation is adapted in both models by the folding–unfolding of the αH2-helix in Model 1 and by the removal of the αH3 helix in Model 2.

The specificity of intermolecular interactions in PDI–hVKORC1 is apparently determined by sequence- and structure-based selectivity, which are the two determining factors in “molecular recognition”. A natural implication of the conformational selection model is the particular range of surface shapes visited by each protein and their collective complementarity, which is adjusted throughout the binding process. It was recognised that cooperativity derives from the hydrophobic effect, the driving force in single-chain protein folding[454]. The hydrophobic folding units that are observed at the interfaces of two-state complexes similarly suggest the cooperative nature of two-chain protein folding, which is also the outcome of the hydrophobic effect[307,455,456]. Nevertheless, although the hydrophobic effect plays a dominant role in protein–protein binding, it is not as strong as that observed in the interior of protein monomers; its extent is variable. The binding site is not necessarily at the largest patch of the hydrophobic surface. There are high portions of buried charged and polar residues at the interface, suggesting that hydrogen bonds and ion pairs contribute more to the stability of protein-binding than to that of protein-folding. Protein-binding sites have neither the largest total buried surface area nor the most extensive nonpolar buried surface area. They cannot be uniquely distinguished by their electrostatic characteristics, as observed by parameters such as unsatisfied buried charges or the number of hydrogen bonds.

The question is then to test if electrostatic and hydrophobic interactions in the PDI–hVKORC1 complex can be conserved qualitatively. The MD simulations of Model 1, performed upon different “soft” constraints that supplied an increased degree of

freedom for proteins and allowed them to be removed, proved the stability of the interactions formed by salt bridges and by the crosswise hydrophobic contacts. As Model 1 of the PDI–hVKORC1 complex showed stable interface interactions under such conditions, it was proposed as the first precursor to probe thiol–disulphide exchange reactions between PDI and hVKORC1.

8.2.3.2. On the protein-protein docking of hVKORC1 onto PDI

Like the user-guided modelling, we have taken PDI as the hVKOR-interacting redox partner, although this question remains open to discussion and still awaits empirical identification[222,224,457]. Docking of PDI (ligand) onto hVKORC1 (target), performed with HADDOCK preliminary tested on the PDI–hVKORC1 complex[221] as a benchmark, clearly indicated that the most interpretable solutions were found for the L-loop closed form used as a target only. Note, that a HADDOCK solution was classified as interpretable if the PDI position matched hVKORC1 bacterial homologue thioredoxin-like domain position, human VKORC1 was shown to use the same electron transfer pathway as its bacterial homologues[224]. Surprisingly, HADDOCK interpretable solutions did not correspond to the benchmark complex, but the other alternative model (Model 2) of PDI-hVKORC1 complex reported in [221]. Given two probable solutions for PDI-VKORC1 complex (Model 1 and Model 2), this would seem to be an area with substantial potential for further development.

As modularity provides biological systems with a convenient way to present binding sites on stable protein scaffolds, in the right position for function, and allows regulation by module rearrangement[369], we investigated whether cleaved L-loop separated from TMD will retain its fused context scaffolding properties. PDI docking onto L-loop as a target produced solutions like those obtained by PDI docking onto hVKORC1. We found that (i) only L-loop closed conformation allows the recognise and binding of PDI, and (ii) PDI docking onto L-loop produced again the most interpretable solutions corresponding to Model 2. Considering cleaved and fused L-loop similar properties in folding and conformational plasticity, and also capacity to recognise PDI, we suggested that cleaved L-loop is a convenient entity in studies of hVKORC1 recognition/activation by its redox protein.

Also, application of hVKORC1 (the membrane protein) in aqueous solution, as shown here, is likely to prove to be very useful in practice either *in silico* studies or *in vitro* experiments. Although, for today, there are no empirical data available for a complex of hVKORC1 with its redox protein, our results can be useful to engender working hypothesis for such studies. Therefore, we are now waiting for needed experimental validation (currently being undertaken by biologist colleagues) of the predictions given in this article. Experimental validation of the model of the PDI–hVKORC1 complex is essential for the continuation of this research, which will allow a better understanding of the redox chemistry underlying vital cell processes.

8.3. CONCLUSION SUR LA MODELISATION DES INTERACTOMES DE KIT ET DE HVKORC1

Pour chaque cible, RTK KIT, hVKORC1 et leurs protéines partenaires, nous avons modélisé deux complexes alternatifs et proposé, pour chaque, le complexe le plus probable.

Les complexes non covalents de RTK KIT avec ses protéines en aval, modélisés ici par le complexe KID^{pY721}/SH2 représentent une perspective de premier plan pour l'étude de l'initiation des voies de signalisations et le transfert de phosphate. Le complexe de hVKORC1/PDI nous permettra d'étudier en profondeur les mécanismes d'activation de hVKORC1 et les réactions d'échange thiol-disulfure. Enfin, pour ces deux complexes, nous avons pu délivrer de nouvelles cibles, des poches allostériques intramoléculaires, mais surtout des interfaces d'interactions pour le développement de nouvelles molécules thérapeutiques visant à contourner les résistances spécifiques du RTK KIT et de hVKORC1 à leurs traitements actuels respectifs.

CHAPITRE 9. EFFETS DE MUTATIONS SUR LE RTK KIT ET HVKORC1

Le RTK KIT et hVKORC1 sont deux protéines entrant dans des processus physiologiques indispensables à l'exécution de nombreuses fonctions vitales, la transduction du signal (RTK KIT) et le recyclage de la vitamine K (hVKORC1). La présence de mutations dans des régions fonctionnelles critiques de ces protéines est à l'origine de la dérégulation de ces processus finement contrôlés.

Dans ce chapitre, nous étudierons et présenterons les effets induits par des mutations ponctuelles pathogènes localisées dans les régions fonctionnelles de ces deux protéines.

Pour le RTK KIT, nous nous sommes concentrés sur le mutant le plus étudié, porteur de la substitution D816V. Cette mutation faux-sens de la boucle A est responsable de l'activation constitutive du récepteur, en partie responsable de pathologies graves, et de la résistance aux inhibiteurs utilisés pour les soigner. Pour hVKORC1, nous avons étudié les effets de quatre mutations de la boucle L associés aux phénomènes de résistances aux anticoagulants antivitamine K (A41S, H68Y) ou à la modification de l'activité de l'enzyme (S52W et W59R).

Le DYNASOME de chacune de ces formes mutées a été étudié en détails à partir de données de simulation de dynamique moléculaire et comparé avec les résultats obtenus pour les formes sauvages du RTK KIT et de hVKORC1.

Ce chapitre présente les résultats faisant l'objet de manuscrits en cours de rédaction et qui seront soumis fin juillet (Ledoux *et al.*) et septembre (Botnari *et al.*) 2023.

1. **Ledoux, J.**, Botnari, M., & Tchertanov, L. (2023). Receptor Tyrosine Kinase KIT: Mutation-Induced Conformational Shift Promotes Alternating Allosteric Pockets.

2. Botnari, M., **Ledoux, J.**, & Tchertanov, L. (2023). Synergy of Mutation-Induced Effects in Human Vitamin K Epoxide Reductase: Perspectives and Challenges for the Design of Allo-Network Modulators.

Les données supplémentaires et les méthodes relatives à ces résultats sont présentées dans les annexes de la thèse.

9.1. MUTATION D816V DU RTK KIT

Résumé. *Le récepteur tyrosine kinase (RTK) KIT est un régulateur clé des processus cellulaires normaux et joue un rôle critique dans le développement et la progression de nombreuses maladies. Les effets induits par les mutations*

conduisent à l'activation constitutive du domaine cytoplasmique favorisant un contrôle aberrant de la signalisation intracellulaire. Nous communiquons le premier modèle structural du domaine cytoplasmique complet d'un mutant oncogène du KIT (KIT^{D816V}) étudié par simulations de dynamique moléculaire. La comparaison des propriétés structurales et dynamiques de KIT^{D816V} avec celles du KIT sauvage (KIT^{WT}) a permis d'évaluer les effets induits par la mutation sur le couplage intramoléculaire entre chaque domaine de la protéine et entre les domaines, en particulier les régions fonctionnelles intrinsèquement désordonnées. Par ailleurs, la représentation d'un paysage d'énergie libre de Gibbs, communs aux deux protéines, a permis de montrer leur superposition partielles et des minima locaux de composition hétérogène. L'ensemble de ces données a délivré des conformations du KIT muté pour son étude en tant que cible pour la reconstruction de son INTERACTOME avec ses protéines partenaires alternatives mais également pour la recherche de poches allostériques pour le développement de nouvelles molécules thérapeutiques inhibant son activité aberrante.

9.1.1. INTRODUCTION

Receptor tyrosine kinases (RTKs) are the cytoplasmic proteins that control the signal transduction of extracellular signals to the nucleus through tightly coupled signalling cascades which alter the expression pattern of numerous genes[338,382]. RTK KIT, also known as the CD117 differentiation cluster, is a family III membrane protein consisting of 976 amino acids (aas). It is targeted by a highly specific cytokine, the Stem-Cell Factor (SCF). Stimulation by SCF in the extracellular medium enables KIT to recruit protein partners into the cytoplasm. KIT initiate and propagate critical signalling pathways through specific phosphotyrosine binding of their downstream proteins contain Src homology (SH2) or phosphotyrosine-binding (PTB) domains[339,415].

The RTK KIT activated signalling pathways that control many important cellular processes such as proliferation, survival, migration, development, and functions of many cell types, including germ cells and immature haematopoietic cells[120]. Under physiologic conditions, KIT gene expression, protein activity and activated signalling processes are quantitatively and temporally perfectly controlled. Dysregulation of KIT activity underpins abnormal cell development leading to tumorigenesis[170]. In particular, constitutive activation may confer oncogenic properties upon normal cells and triggers RTK KIT-induced signalling independently of SCF stimulation[170]. KIT overexpression and gain-of-function mutations have been reported in different types of cancer, such as gastrointestinal stromal tumours (GISTs; in 70–80% of cases), acute myeloid leukaemia, melanoma, systemic mastocytosis, and others [124,282,375].

KIT physiological functions are highly related to its modular architecture, plasticity and transmembrane location which provide tight cooperation of KIT's extracellular and

cytoplasmic domains (ED and CD) through the transmembrane linker (TMD) (**Figure 9.1**, **a**).

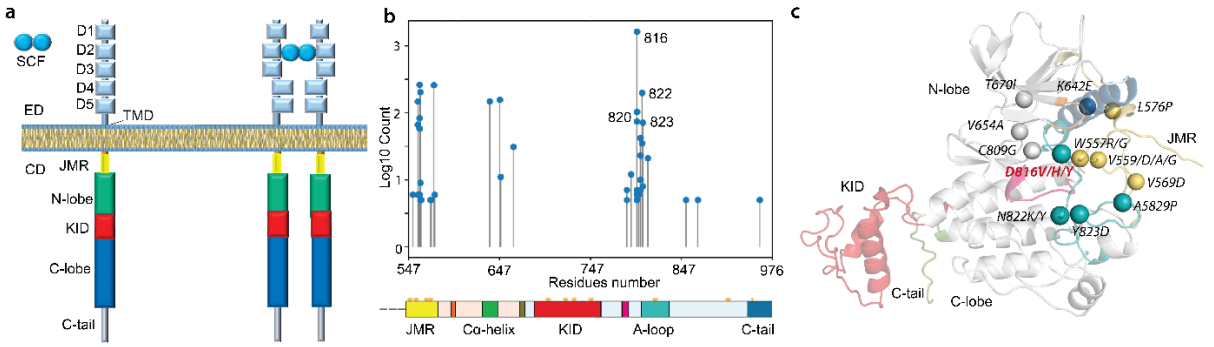


Figure 9.1 RTK KIT SCF-induced activation and its cytoplasmic missense mutations. (**a**) The modular architecture of RTK KIT in monomeric (SCF-unbound) and dimeric (SCF-bound) states. The extracellular domain (ED) composed of five Ig-like domains (D1-D5) is linked by the transmembrane domain (TMD) to the cytoplasmic domain (CD) consistent of a juxtamembrane region (JMR), tyrosine kinase bi-lobe domain (TKD containing N- and C-lobes), kinase domain insert (KID) and C-terminal tail. SCF extracellular binding induces KIT dimerisation and activation. (**b**-**c**) Most studied KIT CD somatic mutations in cancer (count of the number of mentioning papers).

SCF extracellular binding to KIT ED promotes the ligand-induced receptor dimerisation and CD activation. The departure of crucial functional regions – juxtamembrane region (JMR), catalytic Cα-helix and activation (A-) loop – from their autoinhibited positions (structural effects), and phosphorylation of specific tyrosine residues (biochemical/biophysical effects) lead to time-limited signalling, adaptor or scaffold proteins binding to KIT followed by initiation of intracellular signalling cascade[65]. This large set of protein-protein complexes constitutes KIT INTERACTOME.

Wild-type KIT (KIT^{WT}) activates sophisticated signalling pathways through multiple protein partners, either directly or indirectly linked to KIT. Proteins such as Src family kinases (SFK) or phospholipase C (PLCγ) activate the Mitogen-Activated Protein Kinase (MAPK) signalling pathway through the RAS/RAF, phosphatidylinositol 3-kinase (PI3K) pathway as well as, via RAC, the MAPK and c-jun N-terminal kinase (JNK) pathways[407].

The most common perturbations of KIT signalling pathways result from KIT somatic and germline mutations. These gain-of-function mutations typically affect residues from regions involved in the inactive-to-active conformational transition leading to the constitutive (SCF-binding independent) activation of the receptor. Specifically, JMR, A-loop and Cα-helix, which play crucial roles in KIT kinase activation and the transition from an inactive to an active state, are primary locations of gain-of-function mutations (**Figure 9.1**, **b**, **c**). Mutations at position 816 in the A-loop of KIT are frequently observed in cancers, with the substitution D816V being widely implicated in haematological malignancies such as mastocytosis and leukaemia[282,283,374,377,458].

The constitutively active KIT^{D816V} aberrant signalling is associated with structural transformations in JMR and A-loop which disrupt the coupling between these regions as well as their communication pathway[70]. Recently, experimental evidence revealed that KIT^{D816V} mutant does not dimerise like KIT^{WT} and exhibits a decreased stability in the tyrosine kinase domain[459,460]. Signalling amplification compared to the SCF-activated KIT^{WT} has been confirmed[459]. Comparative studies of downstream signalling pathways activated by oncogenic KIT^{D816V} and KIT^{WT}, conducted at qualitative and quantitative levels, have shown differences in signalling potential and alterations in downstream proteins involved in KIT signalling pathways[158,164,166].

In terms of specific pathology-related involvement, KIT^{D816V} is considered the primary driver in many diseases. The availability of tyrosine kinase inhibitors has suggested KIT^{D816V} as a therapeutic target. However, *in vitro* studies investigating the efficacy of imatinib have revealed that although the drug effectively inhibits KIT^{WT}, it does not inhibit KIT^{D816V} [461]. The resistance of KIT^{D816V} to imatinib has prompted *in silico* investigations to explain its mechanisms[462] and numerous *in vitro* studies to evaluate the efficacy of new tyrosine kinase inhibitors. Dasatinib, a potent multi-target kinase inhibitor targeting ABL, SRC, KIT, PDGFR and other tyrosine kinases, demonstrates significant inhibitory activity against KIT^{WT} and KIT^{D816V} [463]. However, anticancer drugs targeting tyrosine kinases not only affect tumours but also have the potential for serious side effects, such as pulmonary or cardiovascular toxicities[464].

Therefore, it is crucial to provide a synergistic description of KIT^{D816V} that can offer prognostic information and open new avenues for research exploring alternative targeted therapeutic strategies.

In the present work we explored the effects of the D816V mutation on RTK KIT at the atomistic level using 3-D model of the full-length cytoplasmic domain of KIT^{D816V} bound to TM helix embedded in membrane. Based on recent empirical data, we studied a monomeric KIT^{D816V} through extended molecular dynamics (MD) simulations. By conducting a detailed analysis of simulation data to compare KIT^{D816V} to KIT^{WT} [271] we aimed to identify the effects caused by the D816V mutation.

The multidomain modular native KIT consists of a quasi-stable TKD, which is surrounded by four intrinsically disordered (ID) regions – JMR, KID, A-loop and C-tail – each containing functional tyrosine residues. Those ID regions act as a platform ground for the recruitment of signalling proteins (JMR, KID and C-tail) or as main promoters of the KIT activation mechanism (JMR and A-loop). With this information in mind, we focused on the following essential questions: (i) Does the kinase domain retain its stability in KIT^{D816V} mutant? (ii) Does the carcinogenic mutation D816V influence the intrinsic disorder of modular KIT? (iii) Which intrinsic disorder events – transient folding, conformational variability or coupling – are the main factors promoting the constitutively active state of KIT^{D816V}? (iv) As tight coupling between JMR, KID and C-tail was observed in KIT^{WT}, we asked whether this coupling is

maintained in KIT^{D816V} mutant? (v) What factors are decisive in forming the constitutively active state of KIT^{D816V}?

The answers to these questions will help establish the causes prompted the KIT^{D816V} signalling deregulation. Thus, clarifying the inherent dynamics of KIT (DYNASOME[1]), describing the allosteric pockets (POCKETOME[465]) and identifying KIT interactions with protein partners (INTERACTOME[2]) might aid in the development of highly effective KIT-specific inhibitors that act simultaneously on intra-molecular targets and the interface between interacting proteins – known as allo-network drugs[466].

The first step in such a study is a detailed description of KIT^{D816V} and a comparison with inactive KIT^{WT} to identify D816V-induced effects on KIT's biophysical properties and establish connections with gain-of-function empirical data. To the best of our knowledge, we report an exhaustive comparison between inactive KITWT and one of its oncogenic mutants for the first time.

9.1.2. Results

9.1.2.3. Data generation and proceeding

The KIT^{D816V} 3-D model (sequence I516-R946) was derived by homology modelling from KIT^{WT} full-length CD model with its transmembrane helix (TMD)[271] (**Figure S44**) and studied by all-atom molecular dynamics (MD) simulation in it natural environement (embedded into the membrane through the TMD and submerged in water). Three independent 2-µs MD simulation replicates were generated to enhance conformational sampling and examine consistency and completeness of the KIT^{D816V} conformations produced under strictly identical conditions. The generated MD trajectories were analysed for the full-length construct and per domain/region using unique and concatenated trajectories. To avoid rigid body motions, KIT^{D816V} trajectories were normalised by least-square fitting on the initial structure (t = 0 µs). For comparative analysis of KIT^{D816V} data with the previously published of KIT^{WT} [271] an additional normalisation was performed by least-square fitting of all MD conformations to the same initial structure (KIT^{WT}, t = 0 µs).

9.1.2.4. General characterisation of MD trajectories

The root-mean-square deviations (RMSDs), computed for each KIT^{D816V} MD conformation display comparable profiles between replicated trajectories, demonstrating a good reproducibility of the generated data (**Figure S45**). As was observed in KIT^{WT}, KIT^{D816V} RMSD values are mainly impacted by KID (up to 22 Å), JMR (up to 15 Å), and C-tail (up to 18 Å), while the TKD lobes showed significantly smaller

RMSD values (< 4 Å). KIT^{D816V} root-mean-square fluctuations (RMSFs) profiles are comparable showing only differences in the highly fluctuating KID, as was observed in KIT^{WT}. The RMSDs and RMSFs calculated after fitting on each domain showed a systematic decreasing of their values, indicating greater inter-domain effects in respect to intra-domain.

9.1.2.5. KIT FOLDING IN INACTIVE (WILD TYPE) AND CONSTITUTIVELY ACTIVE (MUTANT) STATES

The KIT^{D816V} TKD folding (2D structure) remains generally well-conserved throughout MD simulation (**Figure 9.2**). The average structural folding of the N- and C-lobes corresponds closely to the empirically (X-ray) characterised structures of inactive (auto-inhibited) KIT^{WT} (PDP ID: 1T45) and active KIT^{WT} (PDB ID: 1PKG) [121,122].

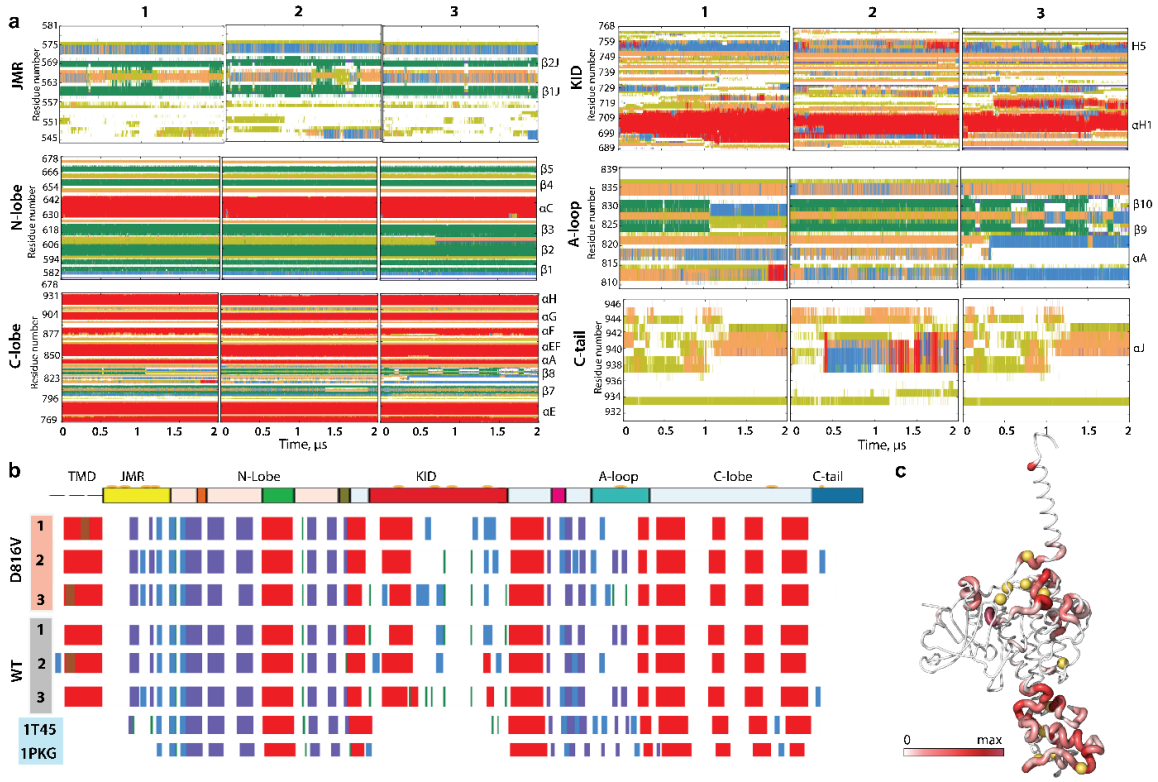


Figure 9.2 KIT^{D816V} folding. (**a**) The time-related evolution of the secondary structures of KIT^{D816V} per domain/region, as assigned by the Define Secondary Structure of Proteins (DSSP) method: α-helices in red, 3_{10}-helices in blue, parallel β strands in green, antiparallel β strands in dark blue, turns in orange, and bends in dark yellow. The three MD replicas (1–3) were analysed individually. (**b**) The secondary structures—αH- (red), 3_{10}-helices (light blue), and β-strands (dark blue)—assigned for a mean conformation of each MD trajectory (1–3) of KIT^{D816V}, KIT^{WT} and the crystallographic structures of inactive KIT^{WT} (PDP ID: 1T45) and active KITWT (PDB ID: 1PKG). (**c**) Increasing of helical folding in KIT^{D816V} in respect to KIT^{WT} illustrated by the thickness of the colored ribbons as a function of the difference in probability of folding per couple of residues i of KIT^{WT} and KIT^{D816V} (large ones are red, small ones are white).

The folding of the four KIT^{D816V} intrinsically disordered regions (IDRs), absent in the crystallographic structures – JMR, KID, A-loop and C-tail –, along the MD trajectories show conserved secondary structures or structures transitioning between folded, alternatively folded or unfolded (αH ↔ 3_{10}-helix ↔ β-strand ↔ turn ↔ bend ↔ coil). The folding content is quite similar to the KIT^{WT} (**Figure S46**). The effects of the D816V mutation on their folding are consistent with earlier observations in a partial KIT^{D816V} construct, studied by MD simulation[70].

Regarding the A-loop, its helical folding significantly increased in KIT^{D816V} (13%) compared to KIT^{WT} (3%). This increase occurs either through a folding up- or downstream of the point mutation, or by disruption of A-loop's two β-strands (**Table S1**). As a result, the effects of the D816V mutation on KIT folding predominantly influence the A-loop. However, when estimating the difference in the probability of secondary structures formation for residue pair i between KIT^{WT} and KIT^{D816V}, it becomes apparent that the effects of the D816V mutation on KIT folding are more pronounced in the IDRs.

9.1.2.6. KIT PLASTICITY: MUTATION-INDUCED EFFECTS ON THE CONFORMATIONAL SPACE

Besides the reversible transition in the folding of IDRs, the conformational plasticity of KIT^{D816V} arises from linear and rotational displacements of local structures within each ID region and between regions/domains, ordered or disordered. To capture the conformational heterogeneity of KIT^{D816V}, and in particular to compare KIT^{D816V} and KIT^{WT} proteins, we analysed the conformational ensemble of each KIT subdomain by Principal Component Analysis (PCA), a multivariate statistical technique used for reducing the dimensionality of large datasets, increasing interpretability and at the same time minimising information loss. This method systematically reduces the number of dimensions needed to describe the protein dynamics through a decomposition process that filters the observed motions from the largest to the smallest spatial scales.

Prior to comparative KIT^{D816V} vs KIT^{WT} analysis, we note that the three first modes describe ≈ 85-90% of all collective motion observed in each of three KIT^{D816V} MD trajectories (**Figure 9.3**, **a**). By projecting the KIT^{D816V} MD conformations from the three independent replicates onto the first two or three principal components, we observe a significant overlap, indicating similar essential motions and conformational features among the replicates.

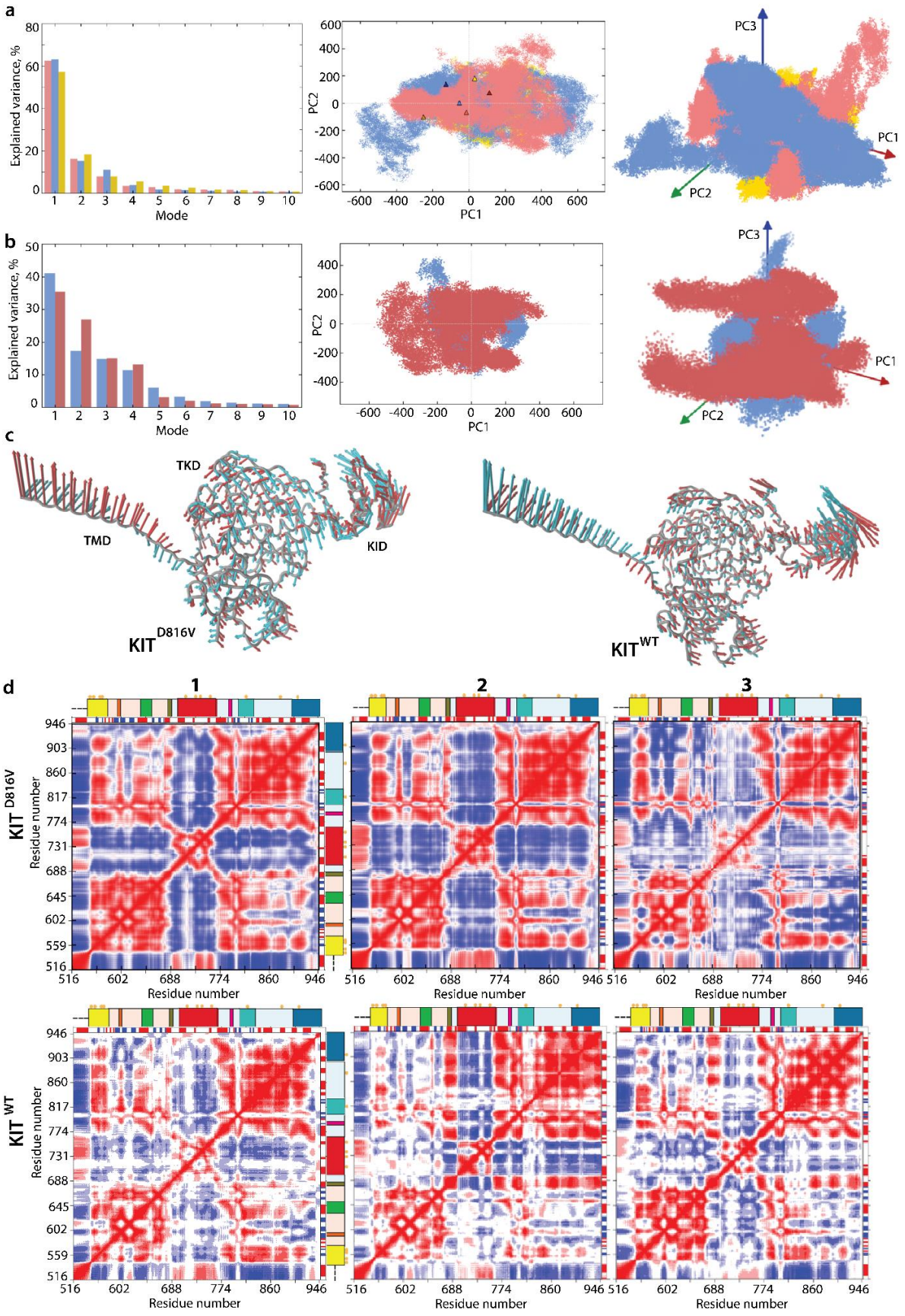


Figure 9.3 PCA of the MD conformations of KIT. (**a**) PCA modes calculated for each trajectory of KIT^{D816V} after least-square fitting of the MD conformations to the initial conformation (t = 0μs). The

bar plot gives the eigenvalue spectra, in descending order, for the first 10 modes calculated on MD trajectories 1–3 (left). Projection of the KIT^{D816V} MD conformations onto the first two (middle), and three PCA modes (right). MD trajectories 1, 2, and 3 denoted in red, blue, and yellow, respectively. Light and dark symbols display the first and the last conformations for each trajectory. (**b**) PCA modes calculated for concatenated trajectory of KIT^{D816V} (red) and KIT^{WT} (blue) after least-square fitting of the MD conformations to the initial conformation (t = 0 μs). The bar plot gives the eigenvalue spectra, in descending order, for the first 10 modes calculated on each concatenated trajectories (left). Projection of the KIT^{D816V} (red) and KIT^{WT} (blue) MD conformations onto the first two (middle) and three PCA modes (right). (**c**) Atomic components in PCA modes 1–2 are drawn as red (1st mode) and cyan (2nd mode) arrows, projected on the cartoon of KIT^{D816V} (left) and KIT^{WT} (right). A cut-off distance of 4 Å was used. Intrinsic motion in KIT and its interdependence. (**d**) Dynamical inter-residue, cross-correlation map, computed for the Cα-atom pairs of MD conformations of KIT^{D816V} (top) and KIT^{WT} (bottom) from each individual trajectory. Correlated (positive) and anti-correlated (negative) motions between Cα-atom pairs are shown as a red-blue gradient.

Overlapping areas are formed by the conformations that are perfectly reproduced during replicates representing the statistically rich 'cloned' data. Non-overlapping areas, on the other hand, consist of newly observed conformations that complement the 'cloned' data and contribute to the generic sets. While the generic conformational space from multiple subspaces does not represent the complete conformational space of disordered KIT, it more comprehensively reflects its conformational properties compared to a single subspace.

The significant overlap observed in the PCA analysis of three KIT^{D816V} MD replicates justifies their merging, allowing for further calculations on the concatenated trajectory. Additionally, since the KIT^{WT} MD trajectories were also highly comparable[271], we used these two conformational generic ensembles characterizing two different proteins, KIT^{D816V} and KIT^{WT}, for comparison. Surprisingly, the conformational spaces of the two proteins exhibit a large overlap, suggesting, a good similarity of their conformational subsets in the initial approximation (**Figure 9.3**, **b**).

The two first modes clearly reflect highly coupled motions of the each multidomain KIT (**Figure 9.3**, **c**). The TM helix shows gradually increased mobility from its C- to N-ends. In both KIT proteins, the global motions of the TK domain exhibit lower amplitudes compared to the TM helix and KID. The motions of all TK residues are collective and can be described as a circular pendulum-like movement along a common virtual rotational axis. Interestingly, this virtual axis is coincident with the active site of each KIT. Notably, the collective motions of TKD calculated from the concatenated trajectories exhibit only slight differences between KIT^{D816V} and KIT^{WT}, indicating better circularity in KIT^{WT} compared to KIT^{D816V}. Analysis of the individual trajectories further highlights these differences in the TKD motion directions of two proteins (**Figure S47**).

The intrinsic dynamics of both multidomain KIT proteins were analysed using the cross-correlation matrix, computed for the Cα-atom pairs of the full-length protein. While the matrices calculated for the three MD trajectories of each protein are very similar, they differ between the proteins (**Figure 9.3**, **d)**. The Cα–Cα pairwise, cross-correlation map demonstrates highly coupled motions within each KIT domain and between the structural domains, even if they are widely separated. However, the cross-correlation pattern in KIT^{D816V} has more pronounced contracts compared to KIT^{WT}, reflecting a higher degree of correlated motion.

In N-lobe, the pattern of each protein reflects the positively correlated motion of the seven stands in the crossed β-pleated sheet, along with their coupling with the αC-helix. Similarly, the helices in the C-lobe show positive mutual correlations, forming well-defined blocks that are distorted by the A-loop. Both TK lobes positively correlate with each other, negatively with the KID and TM helix and positively with JMR, with these correlations being significantly higher in KIT^{D816V}.

The intra- and inter-lobe correlations differ between in KIT^{D816V} and KIT^{WT}. The N-lobe of KIT^{D816V} is a more homogeneous block, reflecting highly collective motions of the P-loop and αC-with helix JMR. The motions of the TM helix and JMR are reversed in both proteins. These correlation patterns can be partially explained by the overall architectural features of the studied KIT proteins, which have a strongly extended shape. Movements of one end (TM helix) are counterbalanced by movements of the opposite end (KID) to provide a stable balance of KIT around its centre of gravity. On the other hand, the highly coupled motion in KIT may reflect the allosteric effect(s) and is functionally relevant, particularly in KIT^{D816V}, which exhibits more active regulatory functions.

9.1.2.7. Impact of D816V mutation on the Inter-Domain Non-Covalent Interactions stabilising KIT

To examine the impact of D816V mutation on the non-covalent interactions stabilising the KIT 3D structure, we compared H-bond patterns observed in KIT^{D816V} and KIT^{WT} (**Figure 9.4**, **a**). A large number of H-bonds were observed in both proteins; however, two new strong H-bonds connecting the A-loop extremities with αC-helix and JMR were observed in KIT^{D816V}. In contrast the strong H-bond that stabilises the A-loop extremity with C-lobe in KIT^{WT} was absent in KIT^{D816V}.

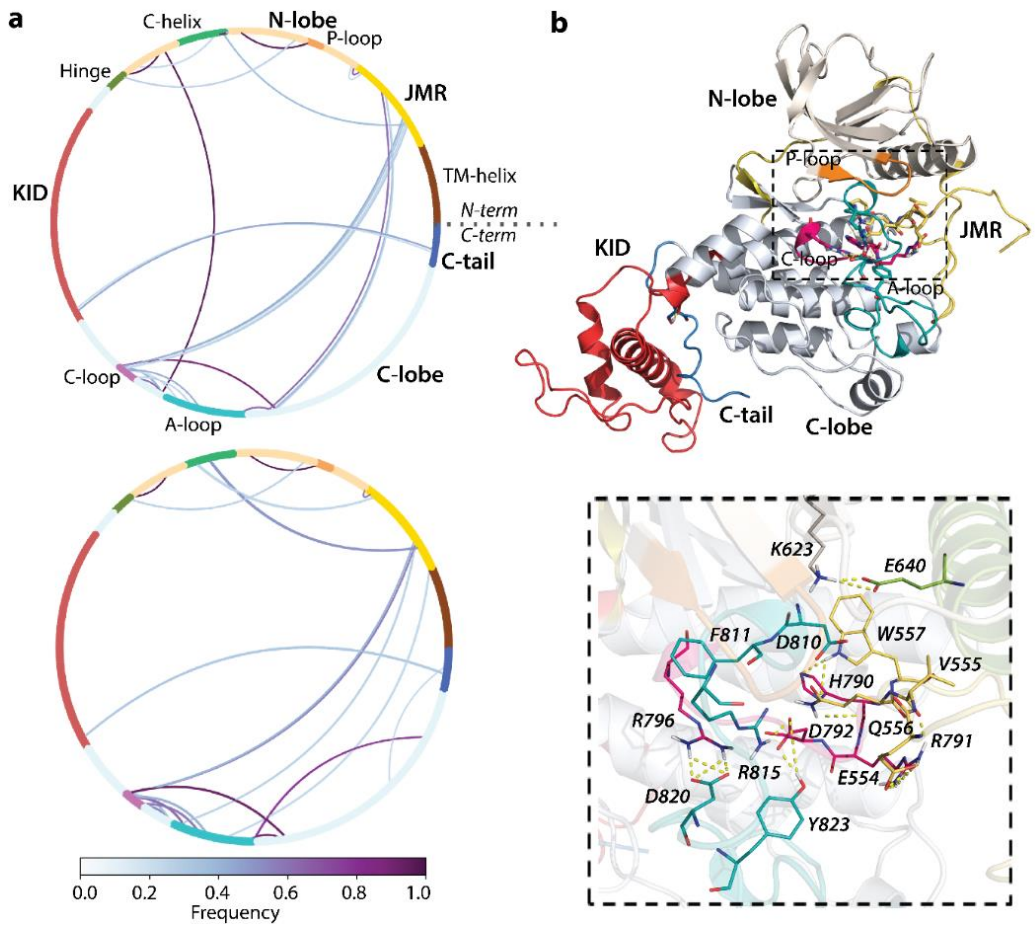


Figure 9.4 Hydrogen bonds stabilising RTK KIT. (**a**) The cords diagram compiles the H-bonds of multidomain KIT^{D816V} (top) and KIT^{WT} (bottom). H-bonds are shown as curves, coloured according to the occurrences, from 0 (white) to 100% (violet). (**b**) The H-bonds (yellow dashed lines), shown on 3D structure of KIT^{D816V} (top), are zoomed on the active site with neighbour residues (bottom). The protein is shown as cartoon, in which the domains and functionally related fragments are distinguished by colour and labelled in bold and regular font, respectively. Residues contributing to H-bonds are shown as sticks and labelled in italic. Calculations are performed on the concatenated trajectory.

Focusing on the active site and residues near the active site, we observe significant changes in the H-bonds pattern in KIT^{D816V} compared to KITWT (**Figure 9.4**, **b**). The D820 residues in the A-loop forms a salt bridge with R796 in C-loop, and Y823 in the A-loop interacts through a H-bond with R815 stabilising two sub-fragments of the A-loop. Additionally, Y823 and D792 from C-loop, form a bifurcate bond. K623 from N-lobe β-sheet interacts with E640 from αC-helix, forming a salt bridge. Interestingly, the JMR W557 forms a unique H-bond with the π-system of H790 from C-loop using its N-H donor group. It is worth noting that the residues D810, F811 and G812 which form the highly conserved catalytic (DFG) motif, do not participate in H-bonding in either protein.

The distinctive H-bond patterns observed in KIT^{D816V}, which differ significantly from KIT^{WT}, appear to play a crucial role in the constitutive activation of the oncogenic

mutant. Surprisingly, the H-bond between the A-loop residue Y823 and C-loop E792, previously interpreted as a key interaction maintaining the allosteric pathway between A-loop and JMR in the inactive native protein[14] is observed in the constitutively active KIT^{D816V}.

On one hand, the TKD, a core structure of KIT, shows the remarkable stability in a given state of protein, whether it is the wild-type inactive (KIT^{WT}) or constitutively active (KIT^{D816V}), with nearly similar H-bond patterns. On the other hand, the observed collective motions in each protein result in many significantly different KIT conformations. This conformational heterogeneity primarily affects the functional KIT regions that possess tyrosine residues – JMR, A-loop, KID and C-tail. Similar to KIT^{WT}, the position of phosphorylation sites in KIT^{D816V} strongly depend on their host fragments, which vary in conformational properties and relative position within protein.

For instance, the extended displacement (linear and rotational) of KID from the TK domain is reflected in the expanded distributions of the Cα-atoms position and hydroxyl groups of Y721, Y747, and Y730, which are located on the highly flexible fragments of the disordered KID. In contrast, the OH group of Y703 in the stable αH1-helix form a narrower cluster, mainly due to the KID global rotational displacement relative TKD (**Figure 9.5**).

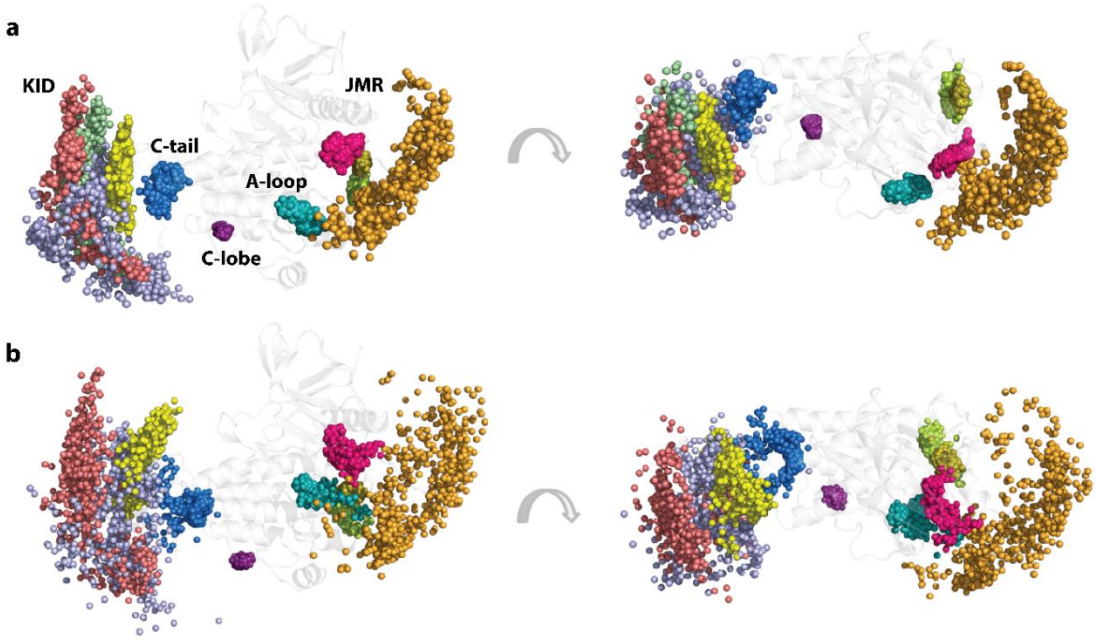


Figure 9.5 Geometry of the tyrosine residues in KIT^{D816V}. The spatial distributions of the Cα-atoms from the tyrosine residues (**a**) and their hydroxyl groups (OH), presented by the oxygen (O) atoms (**b**), are shown in two orthogonal projections with the coloured Cα- and O-atoms: Y547 in orange, Y553 in magenta, Y568 in smith green, Y570 in lime, Y703 in yellow, Y721 in lilac, Y730 in red, Y747 in green, Y823 in teal, Y900 in violet, and Y936 in blue. The Cα- and O-atoms positions were extracted from the MD conformations, taken each 10 ns, fitted on the TKD of the initial structure (t = 0 ns), and superimposed on this structure (countered cartoon in grey).

Similarly, the compact distribution of the locations of Y553, Y568, and Y570, viewed by the Cα-atoms and the OH groups, reflects the stable position of JMR-B, JM-Z, and JMR-S, while the wide-ranging distribution of the Y547 location corresponds to either the multiple inherent JMR conformations or ample displacement of JM-P, with respect to the TK domain. In KIT^{D816V} the compact single cluster representing the unique phosphotyrosine Y823 in the A-loop differs from that in KIT^{WT}, which is subdivided into three different clusters. The JMR Y547 exhibits an enlarged distribution in the two orthogonal directions compared to the unidirectional movement in KIT^{WT}. This distribution of Y547 is likely derived from spatial linear and rotational displacements with respect to the TKD, similar to the KID tyrosine residues Y721, Y747, and Y730.

The other tyrosine residues, Y900 and Y936, located in C-lobe and C-tail respectively, form a unique cluster, either very compact (Y900) or enlarged (Y936). Their positions and compactness are similar to those observed in KIT^{WT}.

9.1.2.8. PER DOMAIN CLUSTERING OF KIT^{D816V} CONFORMATIONS

To characterize the inherent structural and conformational properties of the highly variable KIT^{D816V} regions, namely JMR, KID, A-loop, and C-tail, and estimate their contribution to the global dynamics of protein, each of these fragments was individually analysed. First, the conformations of each fragment were grouped by ensemble-based clustering[267] with varying RMSD cut-off values, ranging from 2.0 to 4.0 Å in increments of 0.5 Å. A cut-off value of 4 Å for JMR, KID and C-tail, while a cut-off value of 2 Å for A-loop were found to be sufficient to group the conformations into clusters that accounted for the majority of the population (> 95%).

For JMR, the majority of conformations were grouped into three highly populated clusters: C1 (52%), C2 (19%) and C3 (12%), which consisted of conformations observed in each MD trajectory (**Figure 9.6**). All JMR conformations showed different secondary structures including a short β-sheet in JM-S segment and a transient 3_{10}-helix in JM-Z, regularly undergoing reversible folding–unfolding events. The representative conformations of the clusters primarily differed in the position of the JM-P segment which contains Y547 identified *in vitro* as a phosphotyrosine[119]. The alternative positions of JM-P resulted in a wide range of Y547 location, while the other tyrosines were almost superimposed (**Figure 9.5**; **Figure 9.6**).

The conformations of the intrinsically disordered KID were grouped into five clusters. Only cluster C1 (63%) comprised conformations generated over three trajectories, while the remaining clusters consisted of conformations from individual trajectory. The representative conformations from thesesdistinct clusters vary in folding (2D) and 3D structure organisation, highlighting the high degree intrinsic disorder in KID. The extensive rotational and linear displacements within KID's structural units contributed to its diverse conformational landscape, resulting in dispersed locations of tyrosine

residues, as illustrated in **Figure 9.5**.

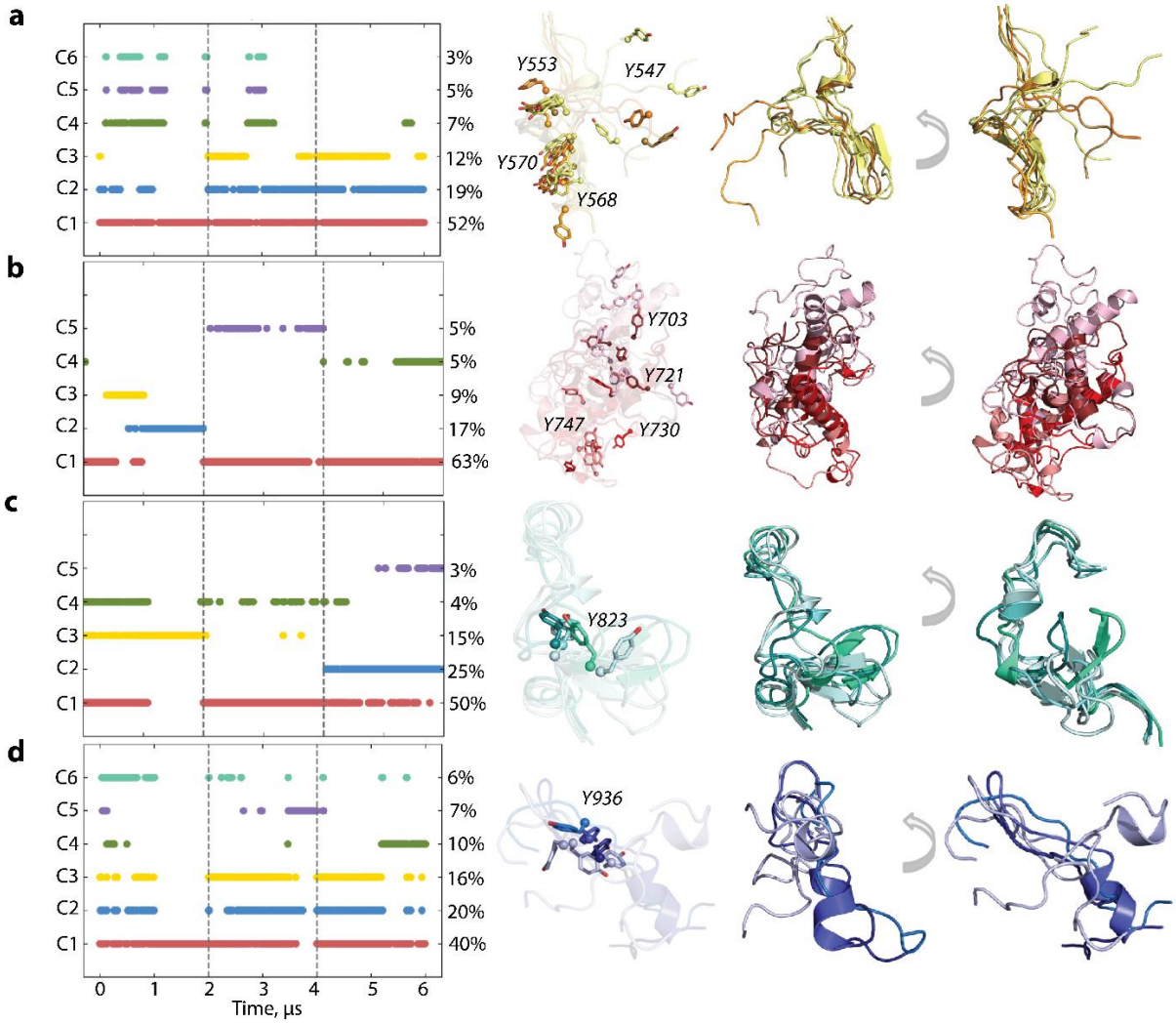


Figure 9.6 Structure and conformation of the disordered fragments of KIT^{D816V}—JMR, KID, A-loop, and C-tail. (**a**-**d**) The clusters obtained by ensemble-based clustering (cut-off of 4 Å for JMR, KID and C-tail, and 2 Å for A-loop) and their population (left column). Superimposed representative conformations from the clusters (right column). The protein is shown as cartoon, the tyrosine residues as sticks. The colour gradient shows the population of clusters, from dark (most populated) to light (less populated). Tyrosine residue numbering is shown for the most populated cluster. Calculations were carried out on the MD conformations of concatenated trajectory, taken every 100 ps after the fitting on the initial conformation of the TKD (at t = 0 μs).

When focusing on the A-loop, it was observed that the most populated cluster (C1, 50%) comprised conformations from all trajectories and contained a well conserved β-hairpin and small 3_{10}-helix transient helix. The conformations in the less populated clusters showed decreased folding (3_{10}-helix ↔ turn) and displaced β-hairpin.

The conformations of the C-tail are grouped into three most populated clusters, C1 (40%), C2 (20%) and C3 (16%) and the three lowly populated clusters, C4-C6 (10-6%). The representative conformations showed similar secondary structures, characterized

as an extended random coil with a small transient helix in the middle (α-helix ↔ 3_{10}-helix ↔ coil) and differ mainly in the orientation of the C-terminal residues. All clusters contained conformations generated from the three independent trajectories, and the separation of clusters did not appear to be influenced by the C-tail secondary structures. The tyrosine residue Y936 showed a close position and orientation of its OH group in most conformations (1–3 clusters) while it differed in the rare intermediate conformations (**Figure 9.5**).

This cluster analysis of the individual regions in KIT^{D816V} demonstrated their inherent structural and conformational disorder, which was found to be greater compared to KIT^{WT} [271]. This highlights the mutation-induced effects on each functional fragment of KIT – JMR, A-loop, KID and C-tail.

However, it is important to note, that this analysis only reflects the intrinsic features of each analysed fragment in both proteins, while the extrinsic disorder, which describes the global protein plasticity (inter-domain/region relationships) has not yet been revealed.

9.1.2.9. WHAT ARE WE LEARNING FROM THE CUMULATIVE FREE ENERGY LANDSCAPE OF KIT^{D816V} AND KIT^{WT}?

The inherent structural and conformational properties of both KIT proteins, KIT^{WT} and KIT^{D816V}, reveal that these multidomain proteins consist of a relatively stable TK domain surrounded by four intrinsically disordered regions – JMR, KID, A-loop, and C-tail.

To characterize and compare the conformational spaces of these disordered proteins, an explicit 'free energy landscape' model was utilized. This model interprets the reversible local folding-unfolding and global conformational diversity of intrinsically disordered proteins by defining the relative Gibbs free energy (ΔG) on selected coordinates known as 'reaction coordinates' or 'collective variables'. These reaction coordinates describe the conformations of the protein between different states and can be used to measure the probability of finding the system in those states. This approach provides a quantitative representation of the sampling of disordered proteins and can be considered a fundamental model system for studying barrier crossing events in such proteins[297].

Since the only difference between KIT^{D816V} and KIT^{WT} is the point mutation (D816V) which can be regarded as an effector promoting irreversible structural and conformational events in the protein, we hypothesized that the cumulative representation of two conformational spaces as a ‘free energy landscape’ model, may shed of light on key differences/similarities between these proteins.

For the evaluation of the relative free energy (ΔG) and the reconstruction of its landscape, the PCA components (PC1 and PC2) were used as reaction coordinates. The free energy cumulative landscape (FEL) as a function the PCA components was calculated on the concatenated trajectory combining the MD conformations of both KIT.

Each 2D FEL exhibits a rugged landscape – reminiscent of the artistic technique known as 'pointillisme' which employs small, juxtaposed areas of colour – reflecting a high conformational heterogeneity of the analysed conformations (**Figure 9.7**).

The observed heterogeneity in conformational landscapes of multidomain proteins like KIT, which consist of at least four intrinsically and extrinsically disordered regions, arises from various factors and the overlap of two distinct conformation spaces. This complexity poses challenges in interpreting the free energy map. However, analysis of the three-dimensional free energy landscape (3D FEL) reveals the presence of smooth or well-defined areas with minimum energy, depicted in red. These regions correspond to more stable conformations, representing thermodynamically favourable states. On the other hand, the reddened areas indicate conformational transitions occurring within the protein.

To gain a more detailed understanding of the effects of the D816V mutation on KIT in each domain separately, it was necessary to analyse each sub-domain individually. The principal component analysis (PCA) of the full-length cytoplasmic domain (CD) considered the folding and positions of all regions relative to the kinase domain together. However, this global analysis does not allow us to specifically assess the mutation effects on each domain. Therefore, to eliminate the influences of other parts of the protein and focus on individual sub-domains, principal components of an PCA were calculated for each sub-domain by cleaving it from the structure and aligning it with the TK domain. This approach enables a more in-depth examination of the effects of the D816V mutation within each specific sub-domain of KIT.

The FEL of the full-length cytoplasmic domain of both KIT reveals the presence of multiple minima or wells W1-W5, separated by energy barriers of varying heights (**Figure 9.7**, **a**).

Wells W1 and W5 have a bi-component non-equivalent content. W1 is composed of 73% KIT^{WT} conformations and 27% KIT^{D816V} conformations, while W5 consists of 78% KIT^{WT} conformations and 22% KIT^{D816V} conformations. W2 is entirely composed of KIT^{WT} conformations (100%) indicating a distinct conformational state exclusive to KIT^{WT}. On the other hand, W3 and W4 predominantly include KIT^{D816V} conformations, with composition 87 and 96% respectively. The KIT^{WT} conformations from W1 and W2 exhibit significant differences in various structural elements. Hence, differences are observed in the (i) orientation of KID relative to TK domain, (ii) folding of A-loop, C-loop and C-tail, and (iii) conformation of JMR.

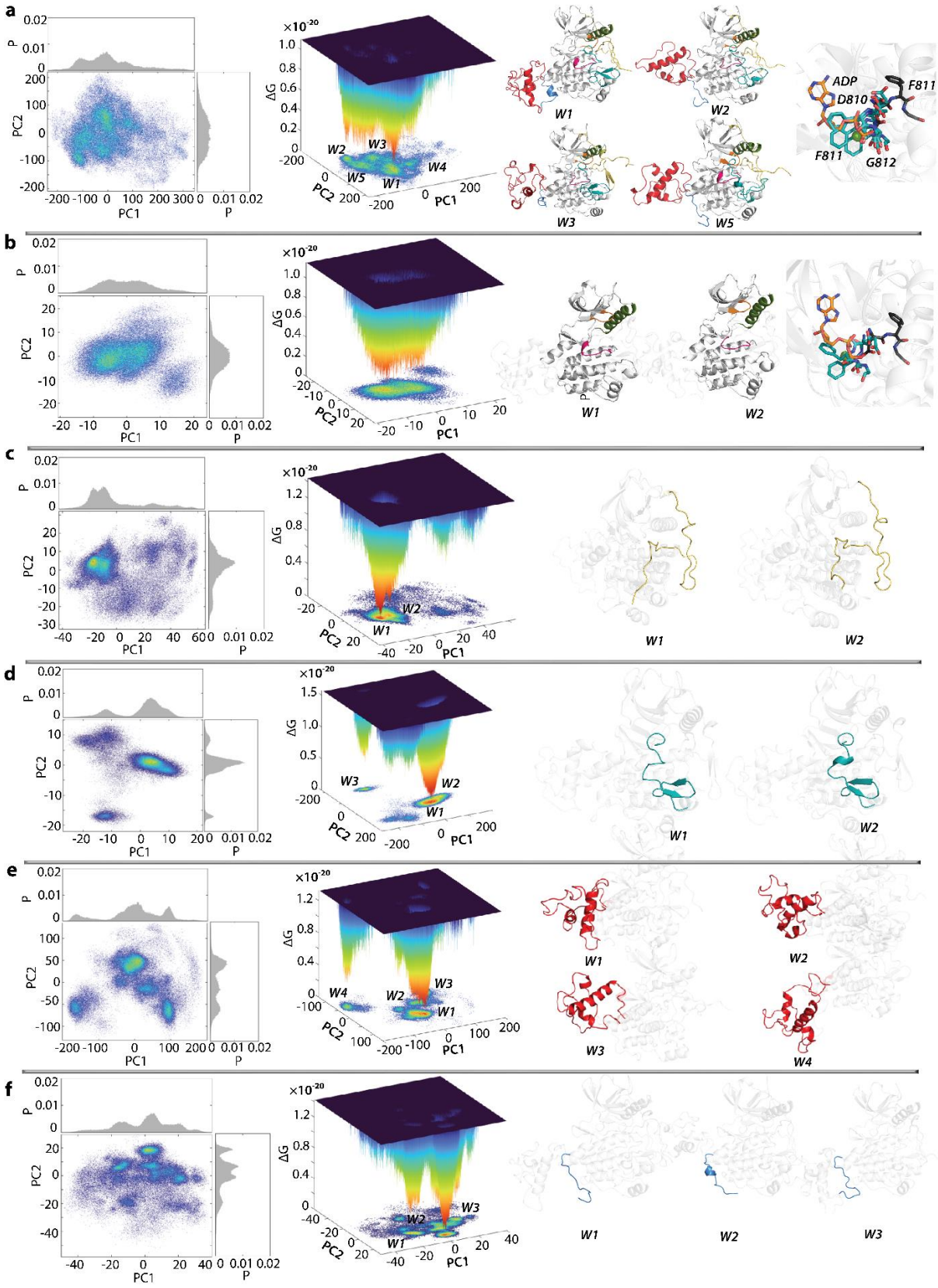


Figure 9.7 Free energy landscape (FEL) of two KIT proteins as a function of the reaction coordinates, taken as two PCA components (PC1 vs PC2). FELs were generated on the concatanated trajectory consisting of MD conformations of KIT^{WT} and KIT^{D816V} mutant, normalised to the initial conformation of KIT^{WT} (t = 0 μs). (**a-f**) The 2D representation of FEL of the KIT conformational

ensembles focusing the full-length cytoplasmic domain (**a**); the full-length cytoplasmic domain without A-loop (**b**); the JMR (**d**); the A-loop (**d**); the KID (**e**) and C-tail (**f**) (left colomn). The 3D representation of the relative Gibbs free energy (middle column). The red colour represents high occurrence, yellow and green represent low, and blue represents lowest occurrence. Each identified well contains at least 1% of the total conformations and $\Delta G \leq 0.45 \times 10^{-20}$ K_bT. The major representation conformation (in cartoon) from the most populated minima (right column) is highlighted by colour: white (TKD or not relevant domain), yellow (JMR), orange (P-loop), green (αC helix), red (KID), pink (C-loop), A-loop (teal), C-tail (blue).

W1 and W4 distinguish themselves from W2 and W3 by a helix folding in the region encompassing catalytic residue D792. A-loop, the mutated sub-domain in KIT^{D816V}, retains a fair interaction with TK domain and JMR but with the latter other regions. In W1, W3 and W4, the same A-loop fragment interacts with JM-B whereas it interacts with JM-switch (V560-I571) turn between the β-strands bearing the functional phosphotyrosines Y568 and Y570. A-loop anti-parallel β-strands remain in W1-W2 but are partially or totally unfolded / folded into helix in W3 and W4 respectively. Additionally, two helices appear in W4 up- and downstream from position 816. Finally, C-tail remains unfolded in W2-W4 varying only by the position relative to TK domain and KID with the similar direction in W3 and W4 and opposite in W2. In W1, C-tail is folded in a unique short helix in contact with TK domain with its C-terminal in interaction with KID.

The major W5 components are KIT^{WT} conformations, comprising fragments that are either similar to W1 (C-loop) or to W2 (A-loop and C-tail), while the KID conformation is W5 differs in both W1 and W2. In contract, the KIT^{D816V} conformations, which are major in W3 and W4, are differed mostly from KIT^{WT} conformations in terms of KID and JMR folding as well as in JMR and C-tail conformations. However, other structural/functional fragments in KIT^{D816V} are similar to KIT^{WT} from either from W1 (A-loop) or W2 (C-loop) or W5 (JMR conformation). In any case, A-loop proximity to JMR seems to indicate non-bonded interactions between both domains as found in JMR auto-inhibition position in inactive KIT^{WT} crystallographic structure (PD ID: 1T45). As for P-loop, folding and positioning do not change considerably between wells. With αC-helix, only its position changes in W4, bending closer to P-loop but not as much as in active KIT^{WT} crystallographic structure (PDB ID: 1PKG).

Interestingly, in all conformations within the lowest energy wells (W1-W5), the conserved catalytic (DFG) motif is in DFG-in conformation. This is noteworthy because in the active state of KIT^{WT}, the DFG-out configuration is typically observed (PDB ID: 1PKG), indicating a distinct conformational change associated with catalytic activity.

When constructing the FEL of the full-length cytoplasmic domain without A-loop, two local minima, W1 and W2, are observed, each with a two-component composition (**Figure 9.7**, **b**). W1 is composed of KIT^{WT} (60%) and KIT^{D816V} (34%) making a 2:1 population ratio. On the other hand, W2 contains KIT^{D816V} (90%) as the major

component. The two major conformations from W1 and W2 are different in the folding of C-loop. For the TK domain, the mutation (D816V) appears to have affected the orientation of the αC-helix, and unfolded C-loop N-terminal containing the catalytic residue D792. We note that the conserved *catalytic* (DFG) motif is in DFG-in conformation in both major conformations.

The FEL of JMR showed two adjacent minima, W1 and W2. Both are predominantly composed of KIT^{WT} (84% in both W1 and W2) with the similar folding and conformational properties (**Figure 9.7**, **c**). KIT^{D816V} is a minor component, representing only 16% of the population. In both representative conformations, JMR is unfolded and maintained in an auto-inhibition position, differing only by the position of JM-proximal (T544-D552) and JM-B relative to the TK domain. It appears that JMR is not a domain affected by the mutation. However, clustering analysis reveals that the KIT^{D816V} JMR exhibits a larger conformational variability compared to KIT^{WT}. The lowest energy conformations In both protein differ in the folding and position relative to the TK domain of all functional sub-domain: JMR is either unfolded or folded in anti-parallel strands, and its JM-binding (JM-B, Y553-V559) region either is sticking to TK domain (W1) and is relatively free in the solvent (W2).

The KID FEL is 263characterized by a series of the largely distant minima (W1-W4) (**Figure 9.7**, **e**). Two of them, W1 and W2, are bi-component wells showing either a quasi-equal composition (W1 contains 53 % of KIT^{WT} and 47% of KIT^{D816V}), or a clear prevalence of only one (W2 contains 81%of KIT^{D816V} and 19 % of KIT^{WT}, or a fully one-component composition (W3 and W4 contain 100% of KIT^{WT} and KIT^{D816V} respectively). When comparing KID conformations from KIT^{WT} forming W1 and W3 wells, the main difference lies in the αH1-helix orientation and the KID conformation. Similarly, the KID of KIT^{D816V} from W2 and W4 is differed by its helix's orientation.

Despite all differences, all wells exhibit a compact globular shape in the KID domain, whether in KIT^{WT} or KIT^{D816V}, with all αH1, 310-H3 and αH5 similarly oriented relative to each other. When comparing KID from two proteins, the same difference appeared at the folding level. However, the mutation appears to have affected the overall domain position relative to TK domain and the helices length and but not their number.

The KID, the longest functional region, exhibits the highest disorder in terms of folding and position relative to TK domain. Across all wells, KID maintains certain characteristics: stable αH1 and αH5 helices that differ mainly in helix content and orientation relative to the TK domain, and overall compactness while retaining a globular shape. In W1, KID is unfolded except for those two helices which are both perpendicular to each other, with αH1 being also perpendicular to TK domain. In W2-W4, KID includes additional 3_{10} helices H2 and H4. Both wells substantially differ in αH1 and αH5 positions: perpendicular to each other in W2 with αH1 pivoting by 90° compared to W1; both parallel to TK domain in W3; αH1 C-terminal oriented about

45° farther from TK domain but nevertheless retaining a cross-like position relative to TK domain and perpendicular to αH5.

The A-loop being affected by the mutant at position 816, remains the most attractive sub-domain to evaluate the effect of the mutation. Similar to JMR, the FEL of A-loop showed two largely separated bi-component minima, W1 and W2. W1 contains both KIT^{WT} (76%) and KIT^{D816V} (24%) while W2 is composed of KIT^{D816V} (85%) and KIT^{WT} (15%) (**Figure 9.7**, **d**). The differences between both wells content consists in A-loop folding rather than its orientation relative to the TK domain. In W1 and W2, A-loop is mostly unfolded except for two anti-parallel β-strands at the same position. However, A-loop in W2 shows a 3_{10}-helix positioned just below the mutation position. Such folding was observed as an intermittent event in KIT^{WT} (Ledoux *et al.*, 2022) but shows much more stability in KIT^{D816V}.

The FEL of the highly flexible C-tail is described by a series of differently populated clusters, three of them are the most populated (**Figure 9.7**, **f**). All clusters are bi-component, but with various protein populations; W1 is composed of 91% of KIT^{WT}, W2 contains quasi-equal population of the both proteins (52% of KIT^{WT} and 48% of KIT^{D816V}), while W3 is completed by 75% of KIT^{D816V} and 25 % of KIT^{WT}. C-tail of KIT^{WT} shows different structure in W1 and W2, while C-tail of KIT^{D816V} from W3 is similar to KIT^{WT} from W1, where it is unfolded and differs only by its position relative to TK domain. Either it is oriented on one side of TK domain (W1) or on the opposite (W3). In the ambiguous W2, C-tail is partially folded in a small helix and its C-terminal oriented as in W1.

However, we must note that such observations are done based on the small wells' populations in respect to the overall sampled conformations during the MD simulations. As such, it is still challenging to observe clear structural differences between KIT^{WT} and KIT^{D816V} even with FEL representation on the two first principal components of a PCA. Either the reaction coordinates are not sufficiently suitable for such study or, the multiple disorder levels and number of disordered fragments or the length/number of replicates are insufficient to better capture the conformational spaces of both KIT species.

As all FELs were reconstructed on the principal components, they only represent the conformations reflecting the 'essential' dynamics of protein. As the first two PCs only represent movements on a larger spatial scale (larger spatial scale motions) and describe only 60-20% of variance, so that the energy landscape, defined on these reaction coordinates, loses smaller spatial scale motions (e.g., intermediate conformations).

In summary, the study of the conformational landscapes of KIT^{WT} and $KITD^{816V}$ reveals several key observations. While there are numerous wells that encompass conformations of both KIT species and demonstrate overlapping folding patterns and

disordered positions of the fragments relative to the TK domain, there is a distinct set of wells that exclusively contain either KIT^{WT} or KIT^{D816V} subdomains. This suggests that the mutation at position D816V has significant effects on the dynamic behavior and internal movements of KIT, ultimately leading to its constitutively activated state.

9.2. MUTATIONS DE LA BOUCLE L DE HVKORC1

Résumé. *La vitamine K époxyde réductase humaine (hVKORC1) est une enzyme clé transformant la vitamine K en une forme fonctionnelle pour la coagulation sanguine. Son activation nécessite les équivalents réducteurs fournis par un partenaire redox. L'actrice principale de cette activation est la boucle L, une région souvent porteuse de mutation faux-sens. Les effets de quatre de ces mutations – A41S et H68Y associées aux phénotypes de résistances aux anticoagulants AVKs, S52W et W59R impliqués dans la perte quasi totale de l'activité de l'enzyme – ont été étudiés par es méthodes* in silico. *Nous avons montré que, même mutée, la boucle L conserve ses propriétés désordonnées. Les mutations affectent particulièrement le repliement et la dynamique des hélices transmembranaires par une forte corrélation de leurs régions adjacentes avec la boucle L et l'alternance de corrélations croisées locales et globales. En représentant par un paysage d'énergie libre de Gibbs les ensembles conformationnels générés, nous avons pu observer leur superposition partielle, avec la forme sauvage de hVKORC1 mais aussi entre mutants, au travers de puits de conformations hétérogènes composés de plusieurs espèces. L'ensemble des données générées a permis de délivrer ces mutants de hVKORC1 comme cibles dans la reconstruction de leur INTERACTOME avec PDI mais aussi pour la recherche de nouvelles poches allostériques pour le développement de nouveaux anticoagulants antivitamine K.*

9.2.1. INTRODUCTION

The human vitamin K epoxide reductase (hVKORC1), a small (163 amino acids) endoplasmic reticulum (ER) protein, contributes to the reduction of inactive vitamin K 2,3-epoxide to active vitamin K quinone, required for blood coagulation[312]. Four functional cysteine residues of hVKORC1 control all steps of its repeatedly reproduced enzymatic cycle. Two cysteine residues, C43 and C52, are located in the ER luminal loop (L-loop), and two others, C132 and C135, in the luminal end the transmembrane domain (TMD) helix, forming $C_{132}XXC_{135}$ motif of the active site. To regain catalytic activity, these cysteines must be reduced by an external redox protein through a sequential electron-transfer process mediated by C43 and C51, which prior to intra-protein electron-transfer to $C_{132}XXC_{135}$ motif involved in the vitamin K transformation. It was predicted that the protein disulphide isomerase (PDI) is hVKORC1 redox partner

initiating its activation by delivering the reducing equivalents[221]. This *in silico* prediction was later confirmed *in vitro*[197].

The perfectly controlled multi-step enzymatic process of the native hVKORC1 may be deregulated by missense mutations. A genetic polymorphism in hVKORC1 is associated with low or accelerated vitamin K recycling rates[196,197] causing serious diseases such as hemorrhages and thrombosis, including the enhanced thrombogenicity in severe Covid-19 cases[203]. Moreover, hVKORC1 polymorphisms affect anti-vitamin K anticoagulant drugs (AVKs) dose response promoting resistance to treatment[194,195]. Currently available AVKs are vitamin K competitive coumarin or indanedione derivatives[230]. If hVKORC1 mutations are frequently associated to resistance phenomena, AVKs physiological response to AVKs is also highly patient-dependent. A study of the 25 hVKORC1 mutants[197] showed that only 6 increased hVKORC1 resistance to AVKs and 10 led to loss of activity *in vitro*. These mutations are located either in the TMD, or in the luminal loop (R33-N77). Out of 45 L-loop residues, 19 show different genetic variations.

hVKORC1 is a small multifaceted protein composed of structurally diverse domains performing their divergent functions: the well-folded TMD and intrinsically disordered L-loop[221,467]. Consequently, its study is of significant interest for (i) fundamental research – as an enzyme activated by its protein redox through thiol–disulphide exchange reactions[217,337,457], a crucial process in biology as having primary roles in defense mechanisms against oxidative stress or redox regulation of cell signaling[190]; (ii) clinically-related investigation to understand the mechanisms of mutation-induced resistance to AVKs; and (iii) pharmacologically-relevant exploration and development of strategies alternative to the classical active site inhibition. Such research can be based on a pioneering *de novo* 3D model of hVKORC1[220], the correctness of which was confirmed by a crystallographic study[218]. Furthermore, hVKORC1 in inactive state (fully oxidised) has been considered as a modular protein consisting of the structurally stable TMD and intrinsically disordered (ID) L-loop[221,467]. This hVKORC1 *de novo* model and its redox protein PDI, were carefully studied by *in silico* methods and used for the generation of a first PDI/hVKOC1 precursor molecular complex[221,467].

hVKORC1 mutants represent an obvious importance for both fundamental and applied research. In this work we will subsequently focus on L-loop mutants. Using *in silico* techniques – 3D modelling and molecular dynamics (MD) simulation – we studied by the mutation-induced effects on the inherent structural and dynamical properties focusing on four mutants in fully oxidised state: two mutations – A41S and H68Y – associated with the resistance phenotype observed in patients, and two others – S52W and W59R – leading to loss of hVKORC1 activity[197] (**Figure 9.8**, **A**). The choice of mutations was not influenced by the level of AVKs responses (resistance) or deactivation reported in the paper [197], but was rather guided by the differences in physico-chemical properties between the native and mutated residues as well as their position on the L-loop sequence. Point mutations A41S and S52W are in the 'cap'

stabilised by disulphide bridge and the remaining W59R and H68Y positioned in the 'hinge', a fully relaxed L-loop sub-region, involved in PDI recognition[221].

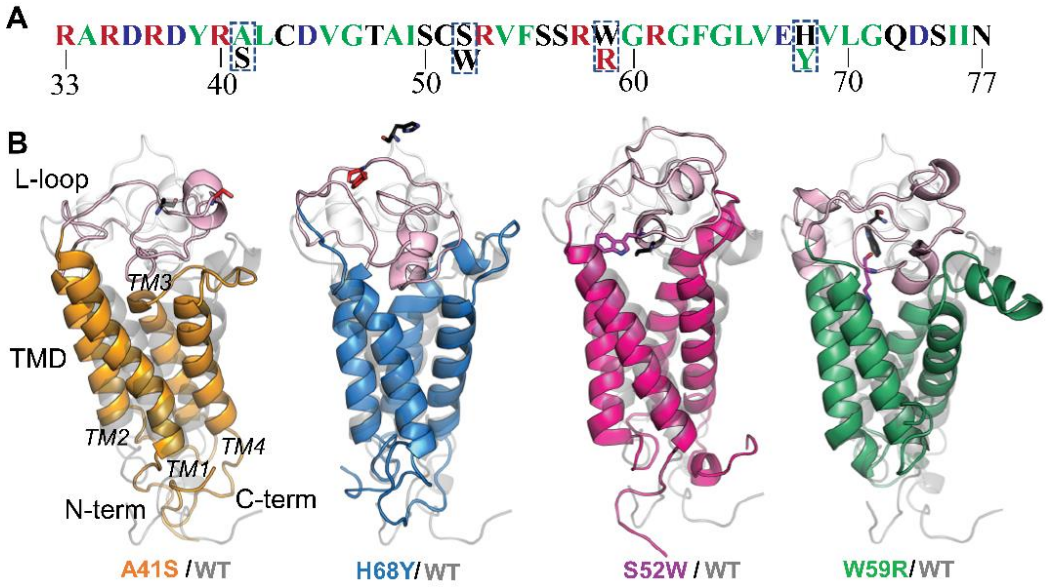


Figure 9.8 hVKORC1 mutants with missense mutations located in L-loop. (**A**) The L-loop sequence (R33-N77) with residues coloured according to their properties: positively and negatively charged residues in red and blue, respectively; hydrophobic residues in green; polar and amphipathic residues in black. Mutated residues are contoured. (**B**) The 3D structural models of hVKORC1^{A41S}, hVKORC1^{H68Y}, hVKORC1^{S52W} and hVKORC1^{W59R} mutants are superimposed with hVKORC1^{WT}. All models represent the last conformation of a randomly chosen trajectory. Proteins are displayed as cartoons, native and mutated residues as black and coloured sticks.

The inherent hVKORC1 mutation-induced effects were compared and interpreted in structural, dynamical and free energy-related terms.

9.2.2. Results

9.2.2.1. Modelling and data proceeding

The 3D models of hVKORC1 mutants were produced by homology modelling using the native enzyme validated *de novo model* as a template, then refined by conventional MD simulation (cMD, all-atom, with explicit water) (**Figure 9.8**, **B**). For each hVKORC1 mutant, three independent MD trajectories (replicas 1–3, each of 0.5 µs) were generated upon strictly identical conditions to extend conformational sampling and examine their consistency and completeness. The generated data sets were analysed for a full-length construct and per domain/region individually using either a single trajectory or concatenated data. To avoid the protein motion as a rigid body, all data were normalised by least-square fitting of MD conformations to the initial

conformation (t = 0 μs). Further, all data generated for the native protein and its mutant were merged and normalised by fitting onto the hVKORC1^{WT} initial conformation (t=0 μs) either on all, or on TMD domain Cα atoms. Such standardised data were used for comparing between them.

9.2.2.2. General characterisation of the MD trajectories

The root means squares deviations (RMSDs) computed for each conformation on the individual trajectories display comparable profiles for each mutant, demonstrating the good reproducibility of the generated data. Nevertheless, the RMSD values show high variation and significantly increasing values (up to 8-10 Å) in some trajectories of hVKORC1^{W59R} and hVKORC1^{H68Y} while other replicas vary slightly, likely to hVKORC1^{WT} (**Figure S48**). hVKORC1^{S52W} and hVKORC1^{A41S} RMSD values and their variations in all trajectories are lower (from 2 to 6.5 Å). Similarly, the root-mean-square fluctuations (RMSFs) curves are comparable in the three MD trajectories for each mutant, differing only in highly fluctuating N- and C-terminals, L-loop and linkers connecting TMD helices and L-loop to TM1. These RMSF profiles are typical for hVKORC1, displaying a saddle form for L-loop residues for which a hollow separates two maxima. The RMSF maxima differ largely for trajectories of the same mutant and between mutants.

To improve mutants and hVKORC1^{WT} RMSFs comparison, their values were recalculated for concatenated trajectories of each protein and standardised on same initial conformation (t=0 μs for hVKORC1^{WT}). Comparing the RMSDFs curves, the following effects may be noted: (i) the L-loop saddle is much more pronounced respective to hVKORC1^{WT}, (ii) the TM linkers in mutants fluctuate more compared to hVKORC1^{WT}; (iii) the maximally fluctuating fragments show different values in hVKORC1 mutants (**Figure 9.9**, **A**).

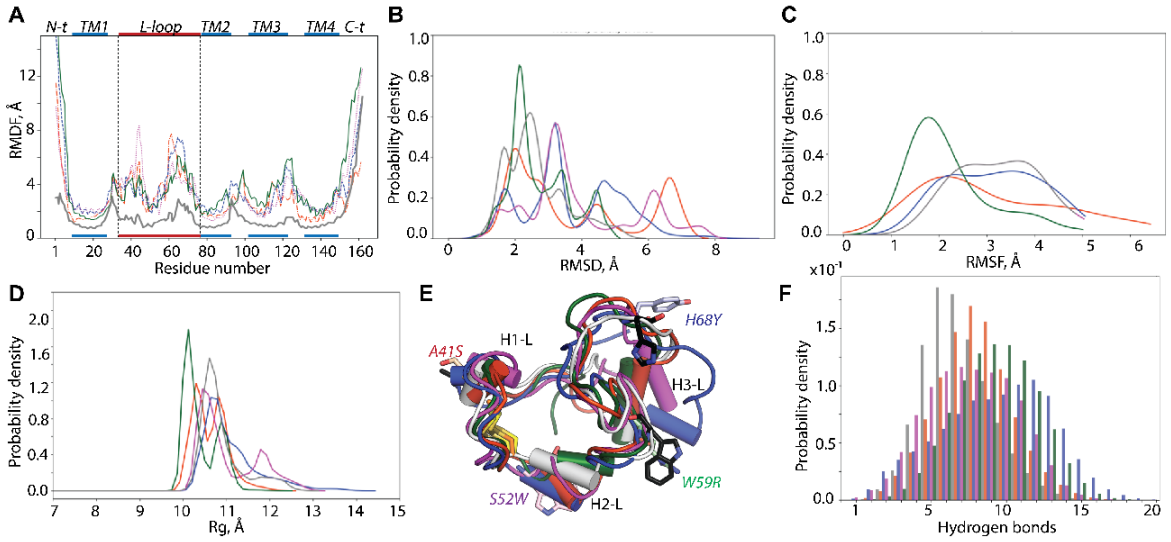


Figure 9.9 Conventional MD simulationsof hVKORC1 mutants in inactive state. (**A**) RMSDs calculated for the concatenated data on the Cα of each mutant and standardised to the initial

coordinates of hVKORC1^{WT} (t = 0 µs). RMSDs (**B**), RMSFs (**C**) and radius of gyration (Rg) (**D**) probability densities. (**E**) Superposition of the most probable conformations (RMSDs peaks) of each mutant studied. (**F**) Number of hydrogen bonds stabilizing L-loop of each mutant compared to hVKORC1. (**A**-**F**) hVKORC1 mutants are coloured in red (A41S), purple (S52W), blue (H68Y) and green (W59R). hVKORC1^{WT} is in gray.

As L-loop is a key moiety in hVKORC1 activation, and bears many missense mutations, we primarily focused our analysis on this region. L-loop RMSD distributions show the multimodal Gaussian curves, while the RMSF distributions are monomodal with very large variance (**Figure 9.9**, **B**). Each RMSD curve shows a unique highest peak corresponding to the RMSD value specific for each protein. The lower peaks are either adjacent to these highest peaks (hVKORC1^{WT}, hVKORC1^{W59R} and hVKORC1^{H68Y}) or largely distant (hVKORC1^{A41S}, hVKORC1^{S52W}) on the RMSD axis. We note that the RMSD curves are most similar in hVKORC1^{W59R} and hVKORC1^{WT} by their profile, the values range (from 1.5 to 5 Å), the highest peak position (at 2.1 and 2.3 Å respectively) and its probability. Other mutants RMSD probability curves demonstrate the enlarged value range (up to 7-8 Å). The RMSF probability curves represent monomodal flattened distributions with very large variance in all mutant, except hVKORC1^{W59R} that shows the RMSF distinct peak ranged from 1.5 to 2.5 Å for a great majority of its MD conformations (**Figure 9.9, C**).

9.2.2.3. L-LOOP SHAPE AND CONFORMATIONAL FEATURES

Despite RMSD and RMSF Gaussian distributions high heterogeneity, L-loop size evaluated by the radius of gyration (Rg), showed that the majority of L-loop in all mutants possesses a compact globular shape of comparable size (Rg between 10-11 Å). However, hVKORC1^{A41S} and hVKORC1^{W59R} display two L-loop sizes, one of them is slightly reduced and the other increased relative to hVKORC1^{WT} (**Figure 9.9**, **D**). Few hVKORC1 mutants L-loop conformations exhibits an enlarged size (up to 13-14 Å).

Suggesting the dependence between L-loop conformation and its shape, we characterised the conformational space explored each hVKORC1 mutant L-loop over the MD simulations with a RMSD-based clustering[267]. First, the L-loop conformations of each mutant were grouped with different RMSD cut-off values that varied from 2.4 to 7.0 Å, with a step of 0.2 Å (**Figure S49**). With a 4.0 Å cut-off value, up to 98-100 % of conformation were clustered.

Comparing the L-loop population of each cluster obtained for every hVKORC1 mutant, we observed that in hVKORC1^{A41S}, most conformations are included in two high populated clusters, C1 (64%) and twice and half smaller C2 (26%); the rest of conformations forms a low populated cluster (9%) (**Figure 9.10**, **A**; **Figure S49**). L-loop hVKORC1^{H68Y} conformations are regrouped mainly in cluster, C1 (69%), with remaining distributed in three clusters with a comparable population: C2 (12%); C3 (9%) and

C4(8%). hVKORC1^{S52W} L-loop conformations form two nearly equally populated clusters, C1 (37%) and C2 (32%), while the other clusters are less populated: C3(17%), C4(9%) and C5 (5%). hVKORC1^{W59R} L-loop are grouped into two clusters, C1(62%) and C2(36%).

For each mutant, all cluster representative conformation is divergent in folding (2D) and 3D-structure. A similar observation can be made by comparing the representative conformations of the most populated clusters of different mutants. Similar, the representative conformations of the most populated clusters of different mutants are structurally distinct. Nevertheless, the most populated cluster C1 in all mutants is composed of compact globular-like 'closed' L-loop conformations, as was observed in hVKORC1^{WT} [221,467]. The poorly populated clusters contain extended L-loop conformations ('open' shape) except hVKORC1^{W59R}. For this mutant, both clusters are populated by 'closed' L-loop conformations differing slightly by the globularity degree. Similar to hVKORC1^{WT}, between the mutants 'closed' and 'open', conformations with intermediate 'opening' are observed in mutants.

L-loop predominant globular-like shape is stabilised by the extended H-bond networks. The number of H-bonds in the majority of L-loop conformations is typically between 5 and 12 contacts for all mutants, are higher in hVKORC1^{WT}, hVKORC1^{H68Y} and hVKORC1^{S52W}, and lower in hVKORC1^{A41S} and hVKORC1^{W59R} (**Figure 9.9**, **F**).

To understand the factors contributing to L-loop fixedness, we analysed the time-dependent occurrence of H-bonds stabilising each protein. We observed that some H-bonds are conserved in all mutants. For instance, the salt-bridge interaction D38···R53, observed in hVKORC1^{WT}, is conserved in the 'closed' conformations and is stronger in hVKORC1^{H68Y}, hVKORC1^{S52W} and hVKORC1^{W59R} (**Figure 9.10**, **B**). The loss of such interactions favors the 'opening' of L-loop conformations represented in the lowly populated clusters (e.g., conformation from C4 in hVKORC1^{S52W}). Such conformations are typically observed in time ranges where the RMSD varies significantly and may correspond to the transient states of L-loop. The backbone-backbone H-bonds L42···A48 and T47···I49 are conserved in all mutants and along all MD trajectories. As the S−S covalent bridge, the salt-bridge interactions and backbone-backbone H-bonds L42···A48 and T47···I49 contribute to the L-loop 'cap' stabilisation. The observe H-bonds (**Figure 9.10**, **C**) explain L-loop 'cap' and 'hinge' cohesion in each mutant, likely in hVKORC1^{WT} [221,467]. L-loop 'closed' form is apparently maintained by non-covalent interactions between hydrophobic residues from these two L-loop modules, 'cap' and 'hinge'. The observed patterns of van der Waals intramolecular contacts in hVKORC1 L-loop display their key role in maintain two L-loop fragments at proximity. These results indicate that intramolecular interactions, van der Waals and electrostatic forces, are dominant factors in the stabilisation of the compact 'globule-shaped' (collapsed) L-loop. Such intermolecular forces, causing the L-loop globularity, conquers with its expansion ('opening') caused by the solvent.

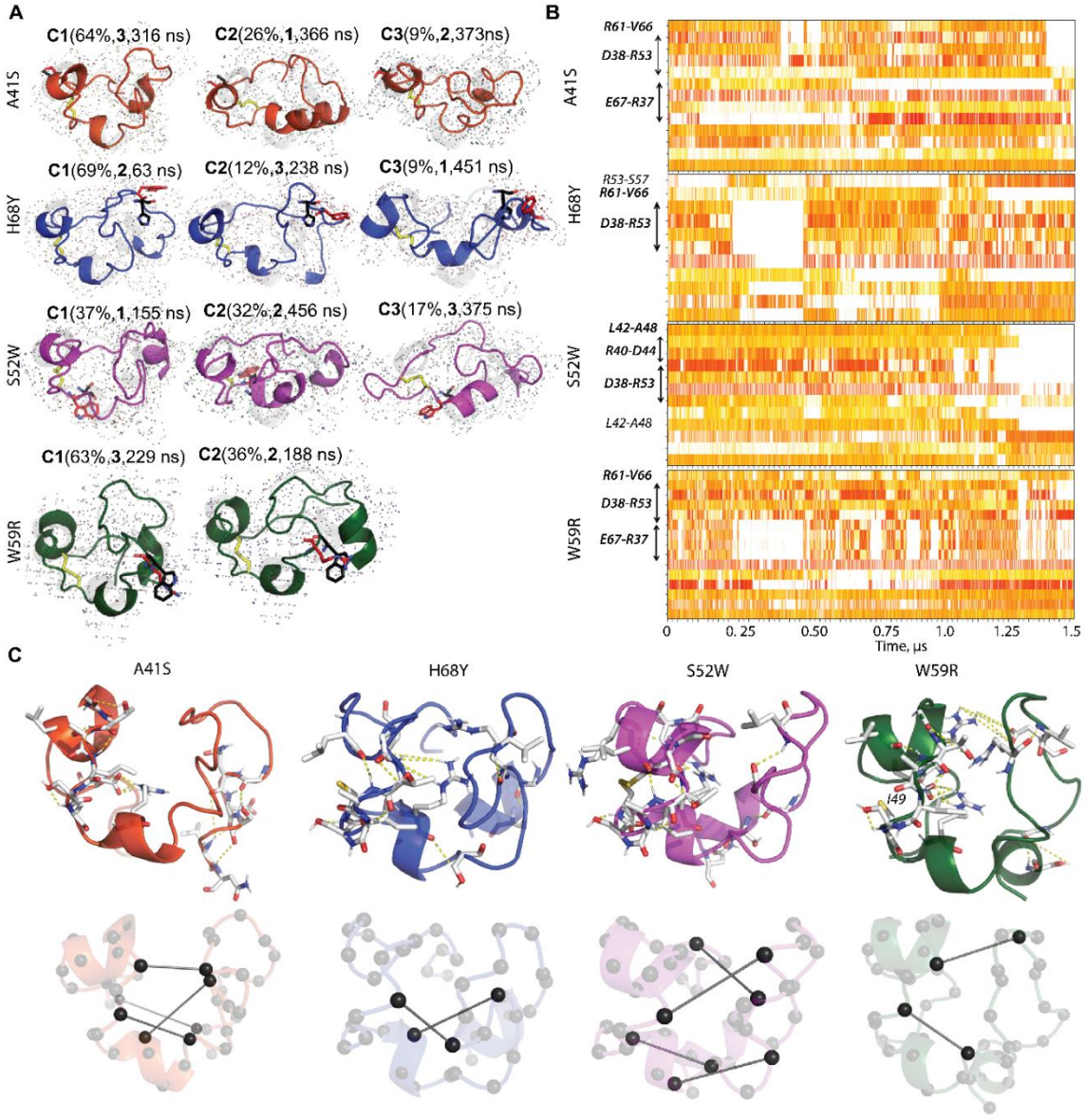


Figure 9.10 L-loop conformations characterisation. (**A**) Ensemble-based clustering of L-loop MD conformations from hVKORC1 mutants obtained for concatenated trajectories after fitting on their respective TMD domain. Clustering was performed on each 100-ps frame for the merged trajectory of each mutant using cut-off value of 4.0 Å. Representative conformations of the L-loop from clusters (C^m) with population ≥ 9%. The population of each cluster is given in brackets (in %), together with the replica number (bold) and the time (in ns) over which the representative conformation was recorded. (**B**) Non-covalent contacts (time series of H-bond events for H-bonds observed with frequency ≥ 0.6) in each mutant of hVKORC1 were calculated for the merged data. The contact's strength is shown by colour: from the strongest (2.6 Å, in red) to the weakest (3.6 Å, in white). (**C**) H-bonds (yellow dashed lines) patterns stabilising the major conformation in each mutant. H-bonds (D-H···A ≤ 3.6 Å, where D (D = O/N/S) is a donor atom and A (A = O/N/S) is an acceptor atom) are shown on a representative conformation of C1 cluster for each mutant. Interactions stabilising regular structures (helices) are not considered. (**A**, **C**) Mutated proteins (in cartoon) are distinguished by colour: $hVKORC1^{A41S}$ (orange red), $hVKORC1^{H68Y}$ (dark blue), $hVKORC1^{S52W}$ (fuchsia) and $hVKORC1^{W59R}$ (dark green); $hVKORC1^{WT}$ (gray). The L-loop of mutants is shown superposed with L-loop of $hVKORC1^{WT}$ (grey). Disulphide bridges C43–C51 drawn as yellow sticks, the mutated and native residues are shown as red and black sticks respectively.

9.2.2.4. L-LOOP FOLDING AND PLASTICITY

Similar to the native enzyme, L-loop helical folding in mutants involves 30%, 34%, 27% and 27% of all residues (mean values) in hVKORC1^{A41S}, hVKORC1^{H68Y}, hVKORC1^{S52W} and hVKORC1^{W59R}, respectively. Such fold is presented by three small (3–4 residues) transient helices, H1-L, H2-L and H3-L, converting partially between αH- and 3_{10}-helices (**Figure S50**). αH-helices are influenced by mutations and their total content varies from 13 (in hVKORC1^{W59R}) to 17% (in hVKORC1^{A41S}) (**Figure 9.11**, **A**).

Such fold is presented by three small (3–4 residues) transient helices – H1-L, H2-L and H3-L–, converting partially between αH- and 3_{10}-helices (**Figure 9.11**, **B**; **Figure S50**). αH-helices are more susceptible to mutations and their total content varies from 13 (in hVKORC1^{W59R}) to 17% (in hVKORC1^{A41S}).

In most cases, the helical content in mutants is the same as in hVKORC1^{WT}. Nevertheless, it presents different probability in each studied protein. L-loop conformations of hVKORC1^{H68Y} showed the lower (18%) and higher (28%) helical content observed. hVKORC1^{A41S} also exhibits two main helical content probabilities at 28 and 35%. In most of hVKORC1^{W59R} L-loop conformations, 28-30% helical folding is observed and only a little part of them is either poorly or highly folded.

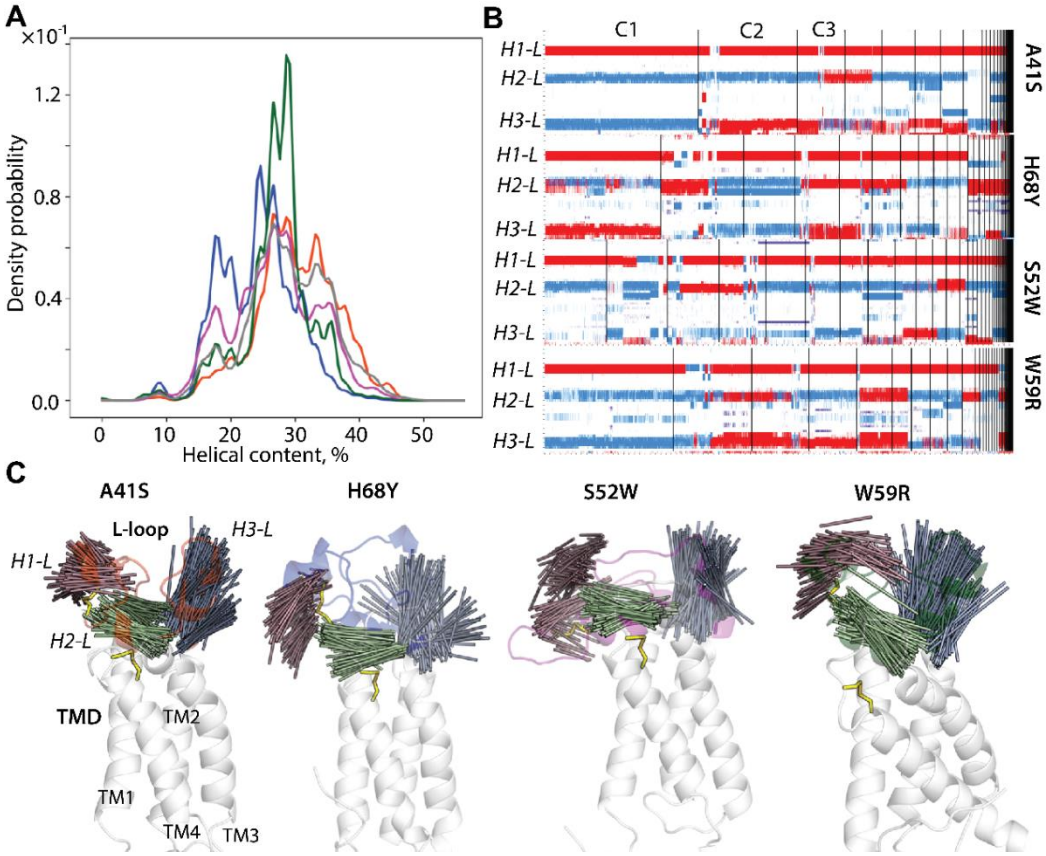


Figure 9.11 Folding of L-loop in hVKORC1 mutants. (**A**) Helical content (helices α, 3_{10} and π) probability (in %) estimated on the concatenated trajectories. (**B**) Clustering of L-loop MD

conformations using their secondary structure content. The α- and 3_{10}-helice are in red and blue respectively. (**C**) Drift of the L-loop helices observed over the MD simulations (concatenated trajectory, sampled every 100 ps) of each mutant. Superimposed axes of helices from the L-loop are covered on a randomly chosen conformation. The axis of each helix is defined as a line connecting the two Cα assigned for the first and the last residues.

Clustering of MD conformations based on the secondary structures showed that H1-L is the most conserved α-helix in all mutants (**Figure 9.11**, **B**). In only a few conformations, α-helix H1-L transits to a small 3_{10}-helix. H2-L helix is a fully transient (α-helix ↔ 3_{10}-helix), at that a fraction of a small 3_{10}-helix is predominant. This highly transient 3_{10}-helix is observed as a short or elongated, a single or double, with a fraction, varying in order hVKORC1^{A41S} > hVKORC1^{S52W} > hVKORC1^{W59R} ≈ hVKORC1^{H68Y}. Similarly, H3-L is fully transient, exhibiting distinct fractions of two types of helices in different mutants. In hVKORC1^{S52W}, H3-L transits also to coil (α-helix ↔ 3_{10}-helix ↔ coil). Combination of divergent fractions of two helices, is fundamental criterion clustering of MD conformations.

Despite the helices transient structure, their positions on the sequence are well conserved, likely to the native enzyme. The L-loop helices are interconnected by coiled linkers, which, together with the linker joining the L-loop to TM1 from the transmembrane domain of hVKORC1, display RMSF values that suggest a higher mobility of these loops compared to hVKORC1^{WT}.

H1-L, mainly folded as a regular α-helix, contains C43 linked covalently to C51 located on the coil connecting H1-L and H2-L helices. Such covalent bonding significantly restricts the conformational mobility of this fragment. The large, coiled linker connecting H2-L and H3-L helices favors the relative displacement of these helices. To illustrate the relative orientation of L-loop helices, their dynamics drift was analysed. The axis of each helix was defined for conformations from the merged trajectories (sampled every 100 ps) of each mutant, superposed and projected on a randomly chosen conformation (**Figure 9.10**, **C**). The superimposed axes (elongated by 50% to better represent their position and direction) form a dense distribution for helices in each mutant. In contrast to the nice reap-like distribution in hVKORC1$^{WT\ [221]}$, each helix axes in all mutants differs largely in its spatial orientation either within each helix distribution and/or between the same helix in distinct mutants.

This observation, at first glance, seems to be inconsistent with the favored 'closed' L-loop conformation. We note that in several mutants, the part of TMD helices adjacent to L-loop showed the mutant-induced partial unfolding (**Figure S51**). In particular, the TM2 and TM3 upper segments adjacent to L-loop are partially disordered (α-helix ↔ 310-helix or/and α-helix ↔ 3_{10}-helix↔coil) in all mutants. This disorder diminishes the α-helical fold of TMD and increases flexibility of these helices, that may influence L-loop plasticity, evidenced as the enlarged L-loop helices drift. We suggested that L-loop helices drift and the TM2 and TM3 upper segments motion may be correlated.

9.2.2.5. hVKORC1 MUTANTS INHERENT DYNAMICS

To characterise L-loop plasticity, the inherent dynamics of hVKORC1 mutants was first analysed by estimating the collective motions by the Principal Component Analysis (PCA) performed on the only Cα-atoms. Similar to the RMSF values, the first two PCA modes denote the terminal residues (N- and C-terminus) great mobility (**Figure S51**). As N- and C-term (M1-G9 and K152-H163 aas) are not the subject of our study, they were excluded from the further computation.

Ten modes describe ~90–95% of the total fluctuations of both hVKORC1 containing TMD and L-loop, or only L-loop, and the two first modes characterise most motion (**Figure 9.12**, **A**). Projection of the MD conformations on the first PCA modes revealed that the collective motion distribution of the two domains, TMD and L-loop, and only the one L-loop, shows that (i) the sampled conformational space is largely extended in all mutants in respect to a more compact area in hVKORC1^{WT}; (ii) the superimposed hVKORC1 conformations provide a poor overlap of hVKORC1^{WT} and its mutants subspaces, especially for L-loop (**Figure 9.12**, **B**).

The first principal component (PC1) is the direction in space along which projections have the largest variance, and the second (PC2) is the direction which maximises variance among all directions orthogonal to the first. These two PCs characterise adequately the degree of anharmonicity in the molecular dynamics of each protein studied. Atomic displacements in the two first PCA modes were projected onto the respective average structures to visualise the direction and amplitude of the principal motion (lowest frequency, most collective) (**Figure 9.12**, **C**, **D**).

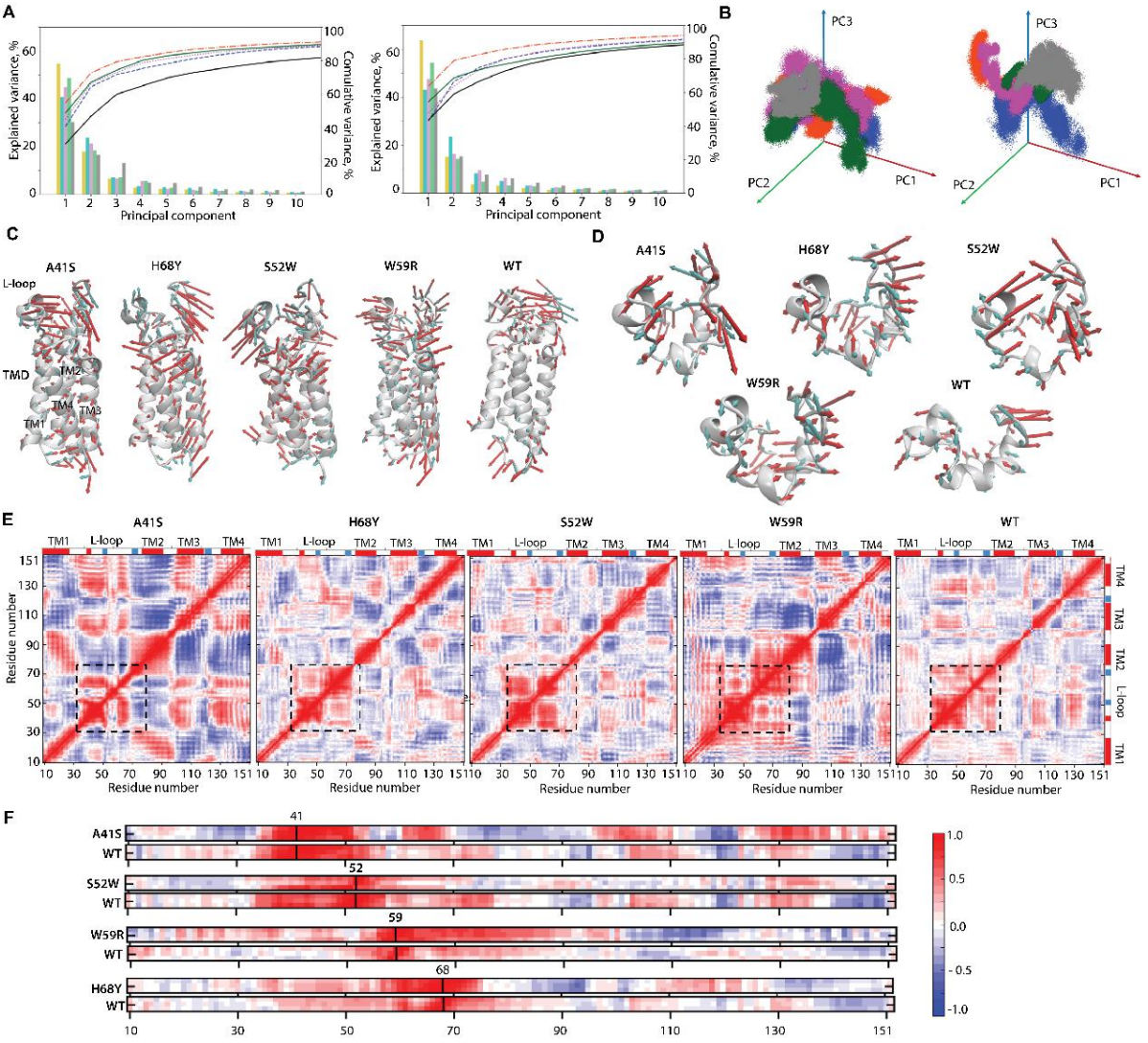


Figure 9.12 Inherent motion of hVKORC1 and its L-loop. (**A**) The PCA modes of the hVKORC1 including TMD and L-loop (left) and the L-loop (right), calculated for concatenated MD trajectory after least-square fitting of the MD conformations to the average conformation of the respective domain as a reference. The bar plot gives the eigenvalue spectra in descending order for the first 10 modes. (**B**) Projection of the hVKORC1 including TMD and L-loop (left) and the L-loop (right) onto the first three modes, calculated with principal component analysis (PCA). (**A-B**) The data for mutated proteins are distinguished by colour: hVKORC1^{A41S} (orange red), hVKORC1^{H68Y} (dark blue), hVKORC1^{S52W} (fuchsia) and hVKORC1^{W59R} (dark green); hVKORC1^{WT} (gray). (**C**, **D**) Atomic components in the two first PCA modes of the analysed hVKORC1 domains are drawn as red (1st mode) and cyan (2nd mode) arrows projected onto the respective average structure presented in grey cartoon. Only motion with an amplitude ≥2Å is shown. (**E**) The inter-residue cross-correlation map computed for the Cα-atom pairs after fitting to the respective first conformation (t = 0 ns) of each hVKORC1 individually (concatenated data of each protein, including TMD and L-loop. The L-loop region is delimited by dashed lines. (**F**) Correlation of each point mutation with the other hVRORC1 residues. (**E, F**) Correlated (positive) and anticorrelated (negative) motions between the Cα-atom pairs are shown as a red–blue gradient. Correlation matrices were calculated after fitting on each hVKORC1 respective TMD domain (t = 0).

The ample motion of L-loop is observed in all mutants, similar to the native protein.

The L-loop large-amplitude collective movements involves all its structural elements – helices and coiled linkers – in all proteins studied. However, the L-loop motion in mutants differs from the scissor-like movements observed in hVKORC1^{WT}, and also alters between mutants. Despite the divergence of the collective motion of the mutants, the L-loop fragments – a 'cap' stabilised by a disulfide bridge and non-covalent interactions, and a relaxed 'hinge'– demonstrate high amplitude movements of these fragments, which are oriented either in the opposite way or in the orthogonal direction with respect to each other. Moreover, the amplitude of these movements is different for different mutants.

In addition, in some mutants we detected a large collective motion of the TM helices, the effect is not observed in hVKORC1^{WT}. These nearly static TMD helices in hVKORC1^{WT}, affected only by the collective drift in membrane[220], in all mutants show the ample motion of their regions adjacent to L-loop.

To follow L-loop and TMD motions, the cross-correlation matrices were computed for the Cα-atom pairs of hVKORC1 mutants composed of TMD and L-loop. The Cα–Cα distance pairwise patterns demonstrate the coupled motions within each hVKORC1 domain, L-loop and TMD, and between two topological domains (**Figure 9.12**, **E**). We noted that (1) the fine-grained muted patterns of hVKORC1^{H68Y}, hVKORC1^{S52W} and hVKORC1^{WT} are composed of well-defined blocks, typically corresponding to structural elements (helices or coils) and showed positive or negative correlations; (2) the hVKORC1^{A41S} and hVKORC1^{W59R} patterns are noticeably brighter and composed of grain-enlarged blocks that frequently overlap between structural units (helices, coils) and/or domains (TMD and L-loop). Focusing on L-loop, its inherent pattern is also different between the studied proteins. In particular, we observed that (i) the well-defined red block bounded by L-loop residues and subdivided into two sub-blocks, demonstrates positive correlations of its structural units – 'cap' and 'hinge' only in hVKORC1^{H68Y} and hVKORC1^{WT}; (iii) the L-loop red block is significantly reduced in hVKORC1^{A41S} and hVKORC1^{S52W} and increased in hVKORC1^{W59R}.

Consistency of L-loop 'hinge' and TM2 motions promotes reduction of L-loop fraction showing independent dynamics, while TM1 contribution in motion consistent with L-loop movement leads to the increase of the dynamical coupling of L-loop and TMD, observed early in the metastable states of hVKORC1^{WT} occuring during the enzymatic cycle[220].

Focusing on the point mutations correlation with the other proteins residues and comparing them to the native protein, we note the important mutation-induced long-range (in the sequence terms) effects (**Figure 9.12**, **F**). In L-loop 'cap' region of hVKORC1^{A41S}, is positively correlated with the 'hinge', TM3 and TM4, and negatively with TM1, TM3 and the loop linking TM3 and TM4. Such correlations are observed in hVKORC1^{WT} with a lesser strength. Mutation H68Y and S52W does not change protein dynamics except for the positive correlation that is increase in the mutated position

direct environment. Moreover, in hVKORC1^{H68Y}, moderate anti-correlation with TM2 and positive correlation with TM3 and its linker to TM4 appeared, whereas in hVKORC1^{S52W}, such correlations disappeared. Finally, in hVKORC1^{W59R}, modest positive correlations are observed with TM1 and L-loop 'cap' region. Additionally, this mutation triggers strong positive and negative correlations with L-loop 'hinge', TM2 and TM3 respectively when it does not influence the dynamic of TM4. Thereby, mutation W59R is strongest influencer of hVKORC1 dynamics (half the protein), where other mutations affected only a quarter of the protein.

9.2.2.6. FREE ENERGY LANDSCAPE AS A QUANTITATIVE MEASURE OF THE MUTATION-INDUCED EFFECTS IN HVKORC1

To compare hVKORC1 mutants conformational space, we adopted a promising strategy for the in-depth analysis – generation of the 'free energy landscape' along specifically chosen coordinates called 'reaction coordinates' or 'collective variables', which describe the conformation of a protein[349–351]. The intrinsic conformation energy landscape can be quantified by the density of states, a statistical energy distribution that may be quantified by transforming the canonical ensemble representation to microcanonical representation. Such interpretation leads to quantitatively significant results that allow comparison between different protein forms, states or mutants. The relative Gibbs free energy ΔG is a measure of the probability of finding the system in a given state. Such representations of protein sampling with use of reaction coordinates can be the quintessential model system for assessing barrier crossing events in proteins[297]. This can be estimated from incomplete sampling of the states, as long as it is an unbiased sampling.

In our case, using rich data obtained by merging of all generated trajectories for hVKORC1WT and its four mutants, we were able to compare MD conformations of different hVKORC1 proteins that differ in sequence by only one residue. Since all proteins were simulated under identical conditions, these data, after normalisation to the same conformation, can also be interpreted in terms of free energy.

To generate the free energy landscape (FEL) of hVKORC1 and its mutants, we used the first two PCA principal components (PC1 and PC2) as reaction coordinates. The FEL constructed on PC1 *versus* PC2 shows a rugged landscape revealing the L-loop high conformational diversity with well-defined minima indicating the multimodal distribution of both PC1 and PC2 (**Figure 9.13**, **A**, **B**).

The deepest well, W1, together with the adjacent low minimum, forms a conformational space separated from the other by a very high energy barrier. This FEL profile is derived, on the one hand, from the bimodal distribution of the PC1 component showing a well-defined sharp probability maximum supplemented by a flat and extended peak, and on the other hand, from a series of maxima on the

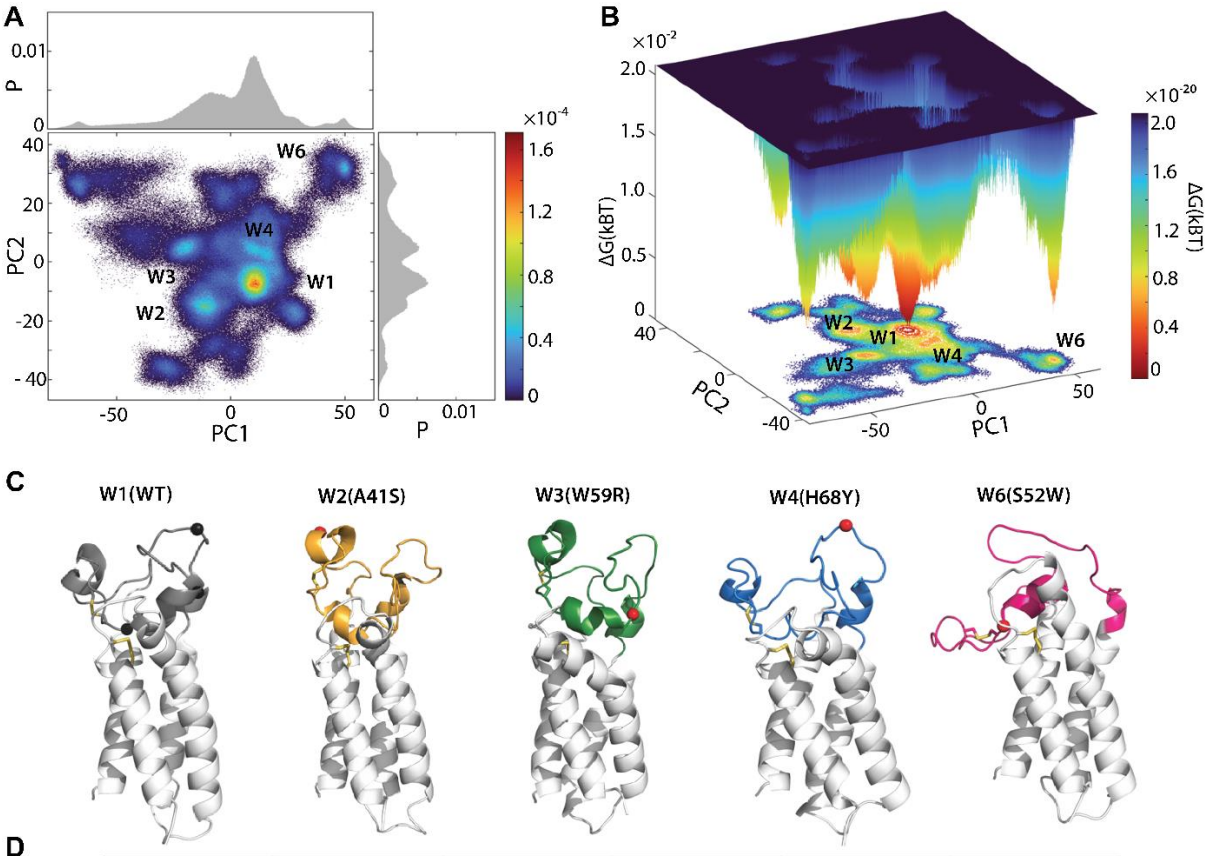
PC2multimodal distribution with a single distinct higher peak. In the results, the deepest W1 on the hVKORC1 FEL is completed by a series of minima represented by lower adjacent wells (W2–W4) and a distant low minimum W6. High energy barriers strongly separate the wells on the FEL.

The Gibbs free energy landscape exhibiting multiple minima was first searched for their protein content. Each well on the FEL includes a different content, consisting of MD conformations of two or three proteins (W1, W2 and W4) or a single protein (W4 and W6) (**Figure 9.13**, **C**).

Further, we explored the content of each well regarding the proteins composition. The wells W1, W2 and W4 are the multi-protein sub-ensembles compiled the MD conformations of different proteins. In particular, the hVKORC1^{WT} MD conformations are the major component (56%) of W1, coexisting with hVKORC1^{H68Y} mutant (23%); the other mutants are also presented in W1 but in a minor quantity (9-4%). Similarly, W2 is principally composed of conformations of two proteins, hVKORC1^{A41S} (56%) and hVKORC1^{W59R} (35%); the hVKORC1^{WT} MD conformations (5%) complete this W2 sub-ensemble. The three-component W4 contains hVKORC1^{H68Y}, hVKORC1^{WT} (37%) and hVKORC1^{S52W} (15%) MD conformations. In difference, W3 and W6 are the mono-protein sub-ensembles composed of only hVKORC1^{W59R} (95%) and hVKORC1^{S52W} (100%) MD conformations, respectively.

From this analysis, some general conclusions may be postulated: (1) the minimum of intrinsic energy landscape spectrum corresponds to conformational ensemble composed of the native (hVKORC1^{WT}) or native-like L-loop by its shape and size ('closed' with Rg of 10.7 Å), observed in hVKORC1^{W59R} and a little portion of the other mutants; (2) the second well (W2) proximal to W1, is composed mainly of two mutants, hVKORC1^{A41S} and hVKORC1^{W59R}, having a most compact globular L-loop (Rg of 10.2 Å) with highly conserved H1-L helix; (3) W1 and W2 contain only 'closed' L-loop conformations different mainly at level of their compactness ; (4) W3 and W6 minima are composed of the 'closed-open' and 'open' L-loop conformations from hVKORC1^{W59R} hVKORC1^{S52W} with Rg of 11 and 12 Å, respectively; (4) W4 compiles all 'open–closed' L-loop conformations of all proteins, except hVKORC1^{A41S} in which L-loop conformations always are 'closed'.

Our results show that the free energy landscape minima reflect rather intrinsic local L-loop conformational features and intrinsic local folding. The conformations of native protein and its mutants havong the similar shape and size dependent of their plasticity, are overlaping.



	W1	W2	W3	W4	W6
Total population	21	6	3	5	1
WT	**56**	5	0	37	0
A41S	7	**60**	0	0	0
H68Y	23	0	4	**48**	0
S52W	9	0	1	15	**100**
W59R	4	35	**95**	0	0

Figure 9.13 Free energy landscape (FEL) of hVKORC1^{WT} and its four mutants. The 2- (**A**) and 3-dimensional (**B**) representations of FEL defined on PC1 and PC2 as the reaction coordinates. FELs were generated on the 7.5 μs concatenated trajectory composed of MD conformations from all hVKORC1 proteins studied—hVKORC1^{A41S}, hVKORC1^{H68Y}, hVKORC1^{S52W}, hVKORC1^{W59R} and hVKORC1^{WT}. N- and C-terminal were excluded from computation. Blue represents the high energy state, green and yellow low and red represents the lowest stable state. (**C**) Representative conformations taken at minima of each well. Protein is shown as carton with L-loop distinguished by colour: hVKORC1^{A41S} (orange), hVKORC1^{H68Y} (blue), hVKORC1^{S52W} (fuchsia) and hVKORC1^{W59R} (green), hVKORC1^{WT} (grey). The point mutation position is shown as red (in mutant) and black (in hVKORC1^{WT}) balls. (**D**). Population (%) and content of each well on the FEL. The major population is in bold. Each hVKORC1 population is the percentage of conformation in a given well relative to the total well population. The total population of each minimum was estimated for conformations in wells with Gibbs free energy ≤ 0.8 × 10^{-20} k_BT.

9.3. CONCLUSION SUR LES EFFETS DE MUTATIONS SUR LE RTK KIT ET HVKORC1

Pour chaque protéine, RTK KIT et hVKORC1, nous avons généré les premiers modèles, étudié par simulation de dynamique moléculaire, et comparé à leur forme sauvage, une sélection de mutants provoquant activation constitutive et résistance aux traitements médicamenteux et/ou modifiant l'activité enzymatique.

L'ensemble des données générées et les propriétés respectives de KIT et de hVKORC1 donne un point de départ à une recherche plus fondamentale pour comprendre les mécanismes sous-jacents aux phénomènes de résistances mais également pour une recherche plus applicative, à savoir l'indentification et la caractérisation de poches alternatives à celles actuellement ciblées et pouvant servir pour le développement de nouveaux modulateurs allostériques dits « allo-network ».

Ces résultats sont encore exploratoires et une analyse plus poussée des données générées sera nécessaire pour une meilleure compréhension des phénomènes entourant les mutations du RTK KIT et de hVKORC1.

CHAPITRE 10. POCKETOMES DES FORMES SAUVAGES ET MUTEES DU RTK KIT ET DE HVKORC1

La caractérisation des mutants du RTK KIT et de hVKORC1 est une base pour une recherche plus applicative, notamment d'intérêt pharmacologique. Un domaine de la recherche en *drug design* est l'identification et la caractérisation de poches à la surface des protéines pouvant être ciblées par des molécules.

Le RTK KIT et hVKORC1 font l'objet de nombreuses mutations entrainant des résistances ou des effets secondaires graves lors du traitement des patients. Ces médicaments ciblent en majorité les sites actifs de ces protéines et ne sont généralement pas spécifiques de la protéine à cibler.

De fait, il est nécessaire de procéder à l'identification de nouvelles poches alternatives pouvant influencer les mécanismes allostériques gouvernant la structure et la dynamique du RTK KIT et de hVKORC1 et qui pourront permettre le développement de nouvelles molécules dites *allo-network*.

Ce chapitre présente les résultats faisant l'objet de manuscrits en cours de rédaction et qui seront soumis fin juillet (Ledoux *et al.*) et septembre (Botnari *et al.*) 2023.

1. **Ledoux, J.**, Botnari, M., & Tchertanov, L. (2023). Receptor Tyrosine Kinase KIT: Mutation-Induced Conformational Shift Promotes Alternating Allosteric Pockets.

2. Botnari, M., **Ledoux, J.**, & Tchertanov, L. (2023). Synergy of Mutation-Induced Effects in Human Vitamin K Epoxide Reductase: Perspectives and Challenges for the Design of Allo-Network Modulators.

Les données supplémentaires et les méthodes relatives à ces résultats sont présentées dans les annexes de la thèse.

10.1. POCKETOMES DE KIT^{WT} ET KIT^{D816V}

10.1.1. INTRODUCTION

The inactive KIT has been widely studied for the design of the ATP competitive inhibitors (type 2) as imatinib or sunitinib [113,462]. Inactive conformations are referred to as "DFG-out" conformations because the Mg-binding DFG motif, which commonly makes conformation-specific molecular interactions with TKIs, is oriented out of the active site. The concept of designing conformational control inhibitors to target the

activated form of kinases and enable the inhibition of a wide range of kinase mutants has been proposed. Indeed, a class of inhibitors targeting the 'switch control pocket' or allosteric inhibitors were reported[468]. A structure of KIT co-crystallised with allosteric inhibitor 3G8 (PDB ID: 6HH1; to be published) shows its position in the allosteric pocket adjacent to αC-helix.

In all cases, this research was performed using the partial structural data in which the JMR, KID and C-tail regions are not defined or defined only partially. Our full-length model of cytoplasmic domain of KIT would largely extend the searching of the novel pockets. We used our long-timescale unbiased MD simulations of both, proteins KIT^{WT} or KIT^{D816V}, to explore their pockets, characterise their properties and select the most perspective for development of selective allosteric modulators.

10.1.2. RESULTS

We simulated the mutation-induced impact on the conformational space of proteins by considering pockets identified by the Fpocket algorithm[420]. Our protocol of allosteric pockets prediction is comprised of three steps: (i) Search of the optimal criteria for Pocket hunting. This step is carried out using the isovalue from 0.35 to 0.50 with a step of 0.5 for both proteins. (ii) Pockets in the input protein structure were identified by Fpocket protein cavity detections, which uses Voronoi tessellation and alpha shapes to identify each pocket. Two isovalues (0.35 and 0.50) were used. (ii) Monitoring of the pockets volume variation during the concatenated trajectory of each protein. (iii) The pockets were ranked according to the computed volume together with their local hydrophobic densities (a feature from Fpocket).

Different allosteric pockets were localised KIT (**Figure 10.1**). Regardless of the isovalue criterion used, we found known pockets that are already targets of inhibitors[469]. The P1 pocket corresponds to the ATP binding pocket and is the preferred target of competitive inhibitors of this molecule such as imatinib. The P2 pocket is an allosteric pocket located between JMR, the αC helix and the β-strands of the A-loop also targeted by imatinib and other kinase inhibitors such as sunitinib or sorafenib. These results allow us to validate the application of such criteria to the search for new alternative pockets on KIT^{WT} and KIT^{D816V}.

The results of our search for pockets enabled us to observe that KIT^{D816V} has both more pockets (between 6 and 8 for KIT^{WT}, and 10 for KIT^{D816V}) but also pockets of smaller volume, located in more or less deep cavities on the surface of the receptor, but with volumes still acceptable to consider them as drug targets.

Outside the known pockets P1 and P2, we observed similar pockets between the two proteins. With an isovalue of 0.35, all KIT^{D816V}'s P2-P8 and P3-P10 pockets appear to correspond to KIT^{WT}'s P2 and P3 pockets respectively. Furthermore, with an isovalue

of 0.5, KIT^{D816V}'s P10 pocket appears to partially represent KIT^{WT}'s P3 pocket. Also, the sum of KIT^{WT}'s P3 and P8 pockets partially corresponds to its own P3 pocket found with an isovalue of 0.35. Furthermore, both proteins show individual pockets at similar positions and comprising the same residues: the P4 pocket located between JMR, A-loop and C-lobe, and the P5 and P6 pockets on the surface of N-lobe.

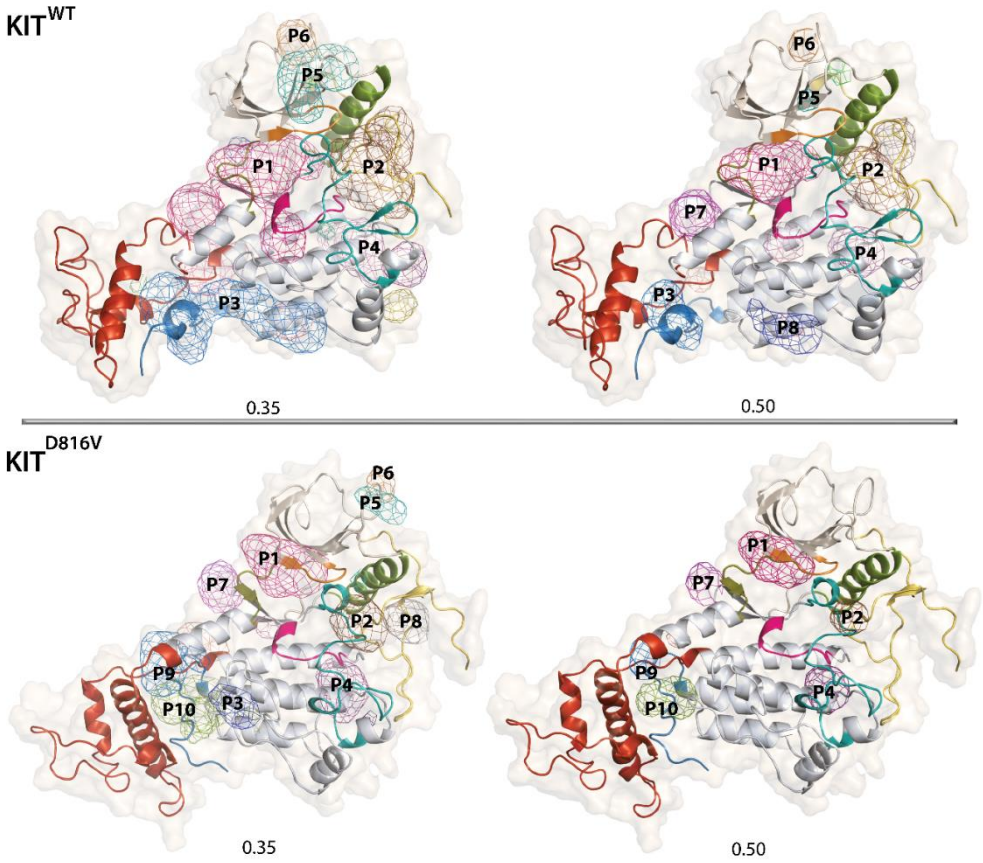


Figure 10.1 RTK KIT^{WT} and KIT^{D816V} POCKETOME found with Fpocket with isovalue 0.35 and 0.5. Pockets are ranked starting from 1 according to their volume and represented on a randomly chosen structure as coloured meshes.

Despite this, we found at least one pocket that appeared to be dependent on the wild-type or mutated protein, regardless of the isovalue criterion used. We might have thought that such a pocket would be located in a site close to the point of mutation but this P9 pocket, located between KID, the N-terminus of C-tail and the last helix of C-lobe, is only present in KIT^{D816V}.

The mutation could therefore have an influence on the size of the common pockets or their division into several sub-pockets, but the results show only one pocket which appears to be exclusive to the mutated receptor (with such isovalue criterion).

We have seen that the disorder and dynamics of KIT IDRs are increased in the mutated receptor, and the very strong anti-correlation between KID and C-lobe could be at the origin of the prolonged presence of such a pocket.

10.2. POCKETOMES DE HVKORC1^{WT} ET SES MUTANTS

10.2.1. INTRODUCTION

After characterising hVKORC1 mutant induced-effects (see section 9.2), we focused on hVKORC1 pockets as possible targets in drug design, in particular, allosteric L-loop pockets. Despite a clear advantage of allosteric drugs, allosteric designs in many cases do not always work[470–472]. It was proposed the way to novel generation of drugs which harness inherent protein allosteric regulation and the protein network-level data (cellular level), named '*allo-network drugs*'[466]. We do not intend to develop allo-network AVKs but aim to design and characterise the perspective allosteric target sites, in native hVKORC1 and each of its studied mutant. For the first time, we communicate the intra-protein allosteric sites observed in hVKORC1 and its four mutants.

10.2.2. RESULTS

Pocket search in hVKORC1^{WT} (with MDpocket[473]) along with the well-known active site pocket was achieved by identification of the potential allosteric binding site on the L-loop surface (**Figure 10.2**, **A**, **B**). Binding sites were numbered according to their enclosed volume (V) – a large P1 (150 -500 Å^3) located at L-loop, and a tiny P2 (50 to 200 Å^3), corresponding to active site cavity of hVKORC1 in the oxidised state. Focusing on these two pockets, P1 and P2, we localised three arginine (R) residues contributing to their formation. We analysed the volume of pockets as a function of arginine quantity at the pocket surface (**Figure 10.2**, **D**). The P1 surface contains from one to three arginine residues – R35, R53, and R61 – and the pocket volume apparently depends on their orientation. Indeed, in smaller P1, the side chain of R35 is oriented outside of the pocket cavity, while the side chains of R53 and R61 are positioned inside the cavity and contribute to stabilising the 'open' conformation of the L-loop. In contrast, if the side chains of all arginine residues are outside of the pocket cavity, the volume of P1 is considerably increased. The surface of the small P2 pocket comprises one to two arginine residues, R35 and R61. The smallest P2 is observed in the 'closed' L-loop conformations with R35 and R61 oriented with side chains outside and inside the pocket, respectively. The 'open' L-loop conformations with a similar orientation of R35 and R61 exhibit a larger P2.

Orientation of the arginine side chain is not a single factor impacting pockets volume. A similar effect occurs from the orientation of the other residues' side chains, for instance, D36, L65 and ER7. Surprisingly, a pocket search in mutants clearly demonstrates the mutation-induced effects on the appearance of pockets, their localisation and size. In difference to hVKORC1^{WT}, in hVKORC1^{H68Y} and hVKORC1^{S52W}, only P2 appeared in MD conformations. The volume of this pocket in hVKORC1^{H68Y} is

comparable with that in hVKORC1^{WT} but is twice larger in hVKORC1^{S52W}. hVKORC1^{A41S} and hVKORC1^{W59R} show three pockets – P2 that is slightly larger than in hVKORC1^{WT}, a tiny pocket P3 adjacent to P2, and P1 with a diminished volume, slightly in hVKORC1^{A41S} and considerably in hVKORC1^{W59R}, compared to hVKORC1^{WT}.

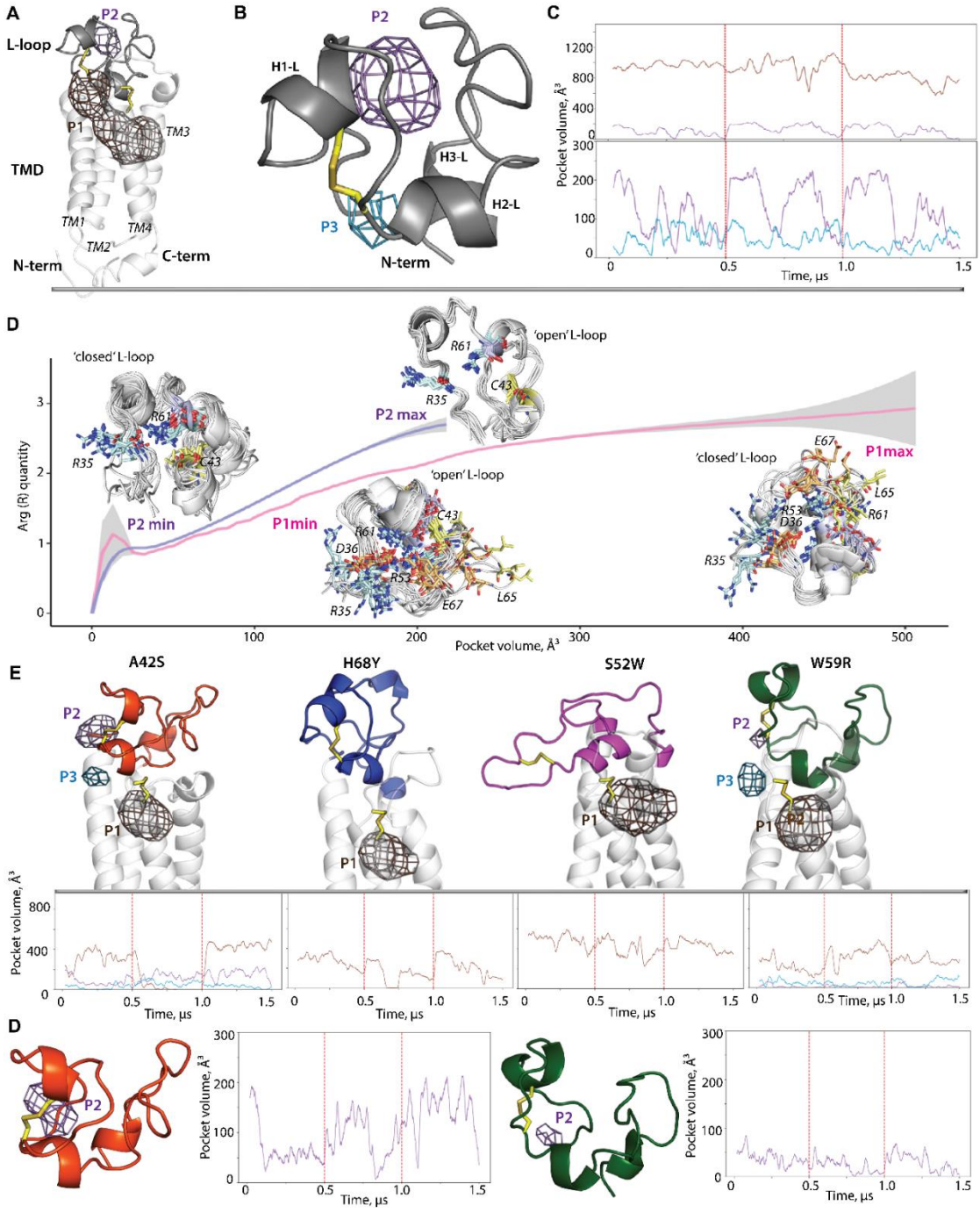


Figure 10.2 Pockets observed in hVKORC1^{WT} and its four mutants. (**A-B**) Two pockets localised in hVKORC1^{WT}. (**C**) Variations of the P1 and P2 volume in hVKORC1^{WT}. (**D**) Volume of pockets as a function of arginine quantity at the pocket surface. Conformations corresponding to the minimal and maximal values are shown together with the orientation of the side chains of residues forming the pockets surface. (**E**) Pockets observed in four mutants (top) and variation of their volume along the MD simulation (bottom).

10.3. CONCLUSION SUR LA DESCRIPTION DES POCKETOMES DE RTK KIT ET DE HVKORC1 DANS LEURS FORMES SAUVAGES ET MUTEES

La recherche de poches a permis de localiser non seulement les poches connues sur chacune des protéines sauvages et mutées mais a aussi fait la lumière sur de nouvelles poches allostériques. La localisation, le volume et le nombre de poches diffèrent entre KIT^{D816V} et KIT^{WT} mais aussi entre $hVKORC1^{WT}$ et ses mutants ou entre ses mutants. Ces poches, de volumes acceptables pour les considérer comme cibles, ouvrent une voie pour le développement de modulateurs allostériques sélectifs.

Tout comme l'étude de l'influence des mutations sur le RTK KIT et sur hVKORC1, l'exploration des poches, leurs propriétés physicochimiques, géométriques, dynamiques et leur druggabilité est encore en cours et sera approfondie promptement.

CONCLUSION GENERALE

Le DYNASOME et l'INTERACTOME sont deux grands axes de description des protéines comparable avec les observations empiriques et permettant de comprendre leurs fonctions et leurs dysfonctions. Ces axes mettent en jeu plusieurs concepts fondamentaux de la biologie.

En premier lieu, nous avons exploré diverses propriétés intrinsèques et extrinsèques des protéines. En particulier, nous avons décrit leur composition en termes de modules à la fois structuraux et fonctionnels. Puis nous nous sommes concentré sur les propriétés des boucles liant ou composant ces modules et leur désordre intrinsèque variable, comment ce désordre se manifeste, leurs fonctions mais aussi leur association quasi systématique à certaines pathologies graves. Nous avons également vu que ces protéines ou régions désordonnées (IDPs ou IDRs) sont des leviers à considérer pour le développement de nouvelles molécules thérapeutiques, spécifiques de la protéine à cibler ou de ses interactions avec des partenaires.

La question de l'étude *in silico* des IDPs n'est pas triviale et le champ de développement de méthodes d'analyses est en plein essor dans le monde scientifique.

Nous avons choisi deux protéines archétypales manifestant toutes ces propriétés.

Tout d'abord le récepteur tyrosine kinase KIT (RTK KIT), une protéine membranaire clé de la signalisation cellulaire possédant un domaine kinase stable entouré de plusieurs régions absentes dans les structures cristallographiques (le domaine juxtamembranaire JMR, le *kinase insert domain* KID et la queue C-terminal C-tail). Ces quatre régions possèdent tous les sites phosphotyrosines responsables à la fois de l'activation du récepteur et/ou de sa reconnaissance par les domaines SH2 de ses protéines partenaires de la signalisation cellulaire. Les mutations gain de fonctions que ce récepteur peut porter sont responsables en majorité de cancers. Une mutation en particulier, la substitution D816V, est responsable de maladies sanguines telles que les leucémies myéloïdes et est la plus étudiée dans le monde médical et scientifique en général. Ce mutant du KIT est encore le sujet de nombreuse énigmes concernant ses propriétés intrinsèques mais aussi les mécanismes de son activation constitutive et ceux responsables de sa résistance aux traitements médicamenteux actuels.

Ensuite, la vitamine K époxyde réductase complexe 1 (hVKORC1) est une enzyme responsable de la transformation de la vitamine K vers une forme utilisable par les protéines qui lui sont dépendantes. Nous avons vu les difficultés de l'étude de hVKORC1, par l'absence de sa structure résolue par méthode expérimentale mais aussi par le mystère entourant l'identité de la protéine redox responsable de l'initiation de son activation et sa reconnaissance par la boucle luminale (boucle L) de hVKORC1. Cette enzyme est une cible privilégiée dans les traitements par anticoagulants

antivitamine K (AVKs) dont la réponse reste dépendante du patient. Une telle individualité de réponse est le résultat de mutations, notamment de la boucle L, entrainant résistances ou modifications de l'activité de l'enzyme.

En parallèle, nous avons exploré certaines méthodes expérimentales utilisées pour établir la structure des protéines, ainsi que les méthodes *in silico* ayant servi à la génération des modèles de ces deux protéines, leur étude par simulation de dynamique moléculaire et l'application de différentes méthodes mathématiques nécessaires à l'extraction d'informations à partir de ces données massives, requises à la fois sur la structure des protéines mais aussi leur dynamique.

Malgré les différences de structures et de fonctions, les problèmes liés à l'étude du RTK KIT et de hVKORC1 sont quasi identiques. L'objectif de la thèse était de caractériser, les DYNASOMES de RTK KIT et de hVKORC1 inactif (sauvages et mutés) ainsi que le DYNASOME des protéines partenaires respectives de ces protéines.

Ainsi nous obtiendrons des jeux de cibles à la fois des protéines seules en formes sauvages, utiles et pour générer les complexes non covalents de RTK KIT et de hVKORC1 avec leur protéine partenaire, et pour leurs applications en tant que les cibles thérapeutiques. Pour la génération de complexes macromoléculaires, nous avons tenté, et avec succès, d'établir un protocole permettant de modéliser les complexes précurseurs pour chaque protéine cible (RTK KIT et hVKORC1) valide pour reconstruire leur INTERACTOME comme un ensemble de complexes de ces protéines avec leurs partenaires. En plus de l'identification et de la caractérisation des poches allostériques présentes sur ces protéines, nous obtiendrons des jeux de cibles présente leurs interfaces d'interactions. Ces ensembles de cibles intramoléculaires (poches allostériques) et intermoléculaires (interfaces d'interactions) pourront par la suite être utilise pour le développement de nouvelles molécules thérapeutiques spécifiques *(allo-network drug design)* contournant les résistances et effets secondaires associées à chacune de ces protéines.

Concernant le RTK KIT, nous avons complété le modèle du domaine cytoplasmique (CD) développé à partir de la structure cristallographique du CD du KIT sauvage inactif, et inséré dans une membrane par son hélice transmembranaire. Nous avons pu identifier et caractériser par simulation de dynamique moléculaire (DM), quatre IDRs entournant le domaine kinase stable : JMR, KID, la boucle d'activation (boucle A) et C-tail. Ainsi, nous avons étudié avec précision le DYNASOME du KIT, en particulier leur couplage entre ces IDRs.

Contrairement au RTK KIT, le hVKORC1 complet a été modélisé par l'équipe BiMoDyM et a permis de caractériser les différents états rencontrés par la protéine au cours du cycle enzymatique à l'exception de l'état initial, à savoir l'activation de hVKORC1 par sa protéine redox. Dans le but de décrire l'INTERACTOME du hVKORC1 sauvage, nous nous sommes concentrés sur son état inactif (oxydé) reconnu par sa

protéine partenaire. Nous avons donc caractérisé, en simulation de DM, le DYNASOME de ce modèle *de novo* hVKORC1 inactif (publié en 2017) et les avons comparée aux données de DM des structures cristallographiques du hVKORC1 disponibles depuis 2021. Nous avons identifié dans le modèle *de novo* que la boucle L est désordonnée. En revanche, l'analyse et la comparaison des propriétés structurales et dynamiques de ce modèle avec les données de DM de modèles obtenus par homologie à partir des structures cristallographiques du VKOR de *Takifugu rubripes*, ont montré que seule la boucle L du modèle *de novo* possède les propriétés adéquates et requises pour la reconnaissance de sa protéine redox et par conséquent reconstruire le l'INTERACTOME du hVKORC1.

Nous nous sommes ensuite demandé si les domaines désordonnés identifiés chez le RTK KIT (KID), et chez hVKORC1 (boucle L) sont des modules (quasi) indépendants du reste de leur protéine respective. L'analyse de leurs propriétés en DM a montré qu'isolés en solution et libre ou avec contraintes artificielles, le KID et la boucle L maintiennent un désordre aux propriétés similaires à leur équivalent intégré dans leur protéine respective. De plus, nous avons établi ces mêmes conclusions pour un KID cyclisé par un linker de glycines. Ainsi, nous avons prouvé que le KID et la boucle L sont des modules du RTK KIT et de hVKORC1 et que ces espèces clivées de leur protéine sont des cibles adaptées pour les études respectives d'évènements de transduction du signal par le KID ou l'activation du hVKORC1 par sa protéine redox.

Pour reconstruire l'INTERACTOME du RTK KIT sauvage et de hVKORC1 sauvage inactif, il est nécessaire d'obtenir des ensembles de conformations de tous les acteurs en jeu et identifier les conformations les plus pertinentes pour étudier leur reconnaissance par leurs protéines partenaires.

Cette question n'est également pas triviale et trouver les bonnes interfaces d'interactions entre deux protéines ainsi que leur évaluation font l'objet du développement de nombreuses méthodes d'amarrage moléculaire.

Nous avons déjà caractérisé la boucle L du hVKORC1 en tant que cible. En revanche, pour la reconnaissance et l'interaction avec ses protéines partenaires de signalisation, RTK KIT doit être phosphorylé à des sites phosphotyrosines précis. Comme nous ne connaissons ni l'ordre de phosphorylation ni le nombre de tyrosine phosphorylées à chaque instant, en se concentrant sur le domaine KID, nous avons étudié toutes les combinaisons de phosphorylation possibles sur ses trois phosphotyrosines et leur influence sur la structure et la dynamique du KID. Nous avons montré que la phosphorylation ne modifie pas significativement le repliement du KID mais possède une influence sur sa dynamique et la distribution spatiale des phosphotyrosines, différente pour chaque combinaison de phosphorylation. Ainsi, nous avons obtenu des ensembles de conformations du KID phosphorylé pour la reconstruction de l'INTERACTOME du RTK KIT par le KID.

Nous nous sommes ensuite concentré sur les partenaires du KID et de hVKORC1.

La structure cristallographique d'un complexe du domaine SH2 N-terminal de la sous-unité p85α de la phosphatidylinositol 3-kinase (PI3K) avec un peptide de KID phosphorylé en Y721 nous a permis une étude approfondie à la fois du complexe mais aussi du domaine SH2 seul. Nous avons montré (1) que la poche de liaison du peptide est une poche polaire possédant une symétrie dans deux directions et stabilisant fortement le peptide en cours de la simulation de DM en particulier par de nombreuses interactions de la phosphotyrosine avec le domaine SH2, (2) que le domaine SH2 simulé seul en solution est composé d'une région stable et de trois fragments de désordre variable porteurs de résidus conservés de reconnaissance et de liaison de phosphotyrosines. Ainsi, nous avons obtenu un jeu de conformations cibles du domaine SH2 du partenaire de KID phosphorylé en Y721 qui nous permettra la reconstruction de son INTERACTOME par ce site.

Contrairement à ce que nous venons de voir, la difficulté de caractériser le partenaire du hVKORC1 réside dans son identification. Sur quatre protéines candidates recensées dans la littérature, notre étude *in silico* a montré que seuls deux fragments nommés F1 et F2 de la *Protein Disulfide Isomerase* (PDI) possèdent les qualités structurales et dynamiques nécessaires à sa reconnaissance par la boucle L de hVKORC1. L'identification de PDI que nous avons faite a été plus tard confirmée par étude *in vitro* et nous a conforté dans l'utilisation adéquate de cette protéine pour la reconstruction du complexe d'initiation de l'activation (ou complexe précurseur) de hVKORC1 par PDI.

Ainsi nous avons caractérisé les DYNASOMES de tous les partenaires inclus dans les INTERACTOMES du RTK KIT sauvage et du hVKORC1 sauvage inactif. L'étape suivante est de modéliser les complexes KID^{pY721} et le domaine SH2 de p85α (KID^{pY721}/SH2), et la boucle L de hVKORC1 inactif et PDI (hVKORC1/PDI).

Pour étudier la reconnaissance de deux paires de protéines, nous avons utilisé deux approches d'amarrage *(docking)* moléculaire : une basée sur le *docking* automatique des deux partenaires, et une que nous avons appelé *user-guided* basé sur une connaissance ou intuition préalable des résidus en interactions entre chaque partenaire. Pour ces deux complexes, les résultats de l'application de cette méthode ainsi que ceux d'un *docking* automatique a révélé deux orientations possibles des partenaires l'un par rapport à l'autre. Pour le complexe KID^{pY721}/SH2, le fragment de KID^{pY721} équivalent au peptide cristallisé dans le complexe résolu par rayon X était orienté soit dans le même sens que dans le complexe cristallographique soit dans le sens opposé. Le complexe hVKORC1/PDI, quant à lui, n'est pas résolu empiriquement. Pendant la reconstruction *de novo* de ce complexe, nous avons observé que l'orientation de PDI par rapport à hVKORC1 diffère par sa position relative à la boucle L, interagissant soit par le fragment F1 soit par le fragment F2 identifiés lors de la caractérisation de PDI seule. Pour prendre en compte le désordre intrinsèque de tous

les fragments des partenaires interagissant entre eux, nous avons reconstruit les deux modèles de KID^{pY721}/SH2 et hVKORC1/PDI par la simulation de DM ou approche *user-guided*. Le rapprochement graduel des deux partenaires a permis d'observer une adaptation mutuelle de la surface d'interactions, l'identification des résidus la stabilisant et le repliement complémentaire des deux partenaires.

Par ces résultats, nous avons non seulement réalisé un premier pas pour la reconstruction des INTERACTOMES complets du RTK KIT et hVKORC1 sauvages mais également obtenu des surfaces d'interactions qui pourront être utilisées lors du développement de nouvelles molécules thérapeutiques spécifiques respectivement de ces protéines.

Or, les pathologies associées à ces deux protéines, RTK KIT et hVKORC1, sont développées lorsque certaines de leurs régions (JMR, boucle A pour KIT ; boucle L pour hVKORC1) acquièrent des mutations, en particulier dans les domaines d'interactions avec des protéines partenaires.

L'ensemble des résultats suivants sont exploratoires et font encore l'objet de recherches en cours.

Nous nous sommes intéressés à deux jeux de mutations de type substitutions d'acides aminés. Pour le RTK KIT, une mutation de la boucle d'activation D816V responsable de cancers hématologiques ; pour hVKORC1, quatre mutations de la boucle L responsables soit de la résistance du récepteur aux AVKs (A41S, H68Y) ou de la modification de l'activité de hVKORC1 (S52W, W59R).

Les modèles des formes sauvages de ces deux protéines ont servi de point de départ à la génération du premier modèle du CD complet du KIT muté (KIT^{D816V}) et des premiers modèles du hVORC1 muté dans sa boucle L.

Nous avons caractérisé l'ensemble de ces mutants et avons comparé à leurs formes sauvages leurs propriétés structurales et dynamiques, ainsi que les espaces conformationnels explorés leur de la simulation de dynamique moléculaire.

Nous avons montré que la mutation D816V n'influence que peu la structure du domaine kinase, mais a un effet non seulement sur le désordre intrinsèque et extrinsèque des fragments désordonnés du KIT en l'augmentant par rapport à sa forme sauvage, mais également une partie de son repliement à des régions fonctionnelles et stratégiques comme JMR et la boucle A. Le récepteur KIT^{D816V} est plus dynamiquement organisés et la mutation D816V provoque l'augmentation forte du couplage entre domaines et un inversement de la direction des mouvements collectifs par rapport à la forme sauvage de KIT. Pour les mutants de hVKORC1, le repliement global est majoritairement conservé. La flexibilité de la boucle L est maintenue bien que la position et l'orientation spatiale des hélices transitoires soient variables parmi les mutants. De plus, nous avons observé une déstabilisation de l'orientation des quatre

hélices transmembranaires qui peut être due à la mutation et/ou par les conditions de simulations (sans membrane). Enfin, au niveau dynamique, chaque mutation influence non seulement le couplage local autour de la position mutée mais également sur une majorité la structure de chaque hVKORC1 étudié. En revanche, KIT^{D816V} et les quatre mutants de hVKORC1 présentent des similitudes avec leurs formes sauvages respectives (et entre mutants pour hVKORC1) avec qui ils partagent une partie de leurs espaces conformationnels.

L'ensemble de ces résultats offre tout d'abord des modèles dynamiques explorables pour la recherche fondamentale sur les mécanismes d'activation et de résistance des protéines, ainsi que leur régulation allostérique. Également, ils sont une base pour la reconstruction des INTERACTOMES des mutants du RTK KIT et de hVKORC1. Enfin, pour une application dans le développement de stratégies alternatives de développement de modulateurs sélectifs et spécifiques contrôlant l'activation excessive et la formation de leurs complexes avec leurs protéines partenaires. Ces médicaments dits *allo-network* permettront le contrôle spécifique de la signalisation du KIT et de l'activation de hVKORC1 par une modulation allostérique. Ainsi, ils pourraient contourner les effets secondaires résultants de leur inhibition par des molécules non spécifiques agissant sur des protéines homologues.

Les modèles des formes sauvages et mutées du RTK KIT et de hVKORC1 ont servi de point de départ pour une recherche plus applicatives. En se concentrant sur les protéines seules, nous avons recherché de nouvelles poches pouvant être ciblées dans le développement de nouvelles molécules contournant les phénomènes de résistances induits par les mutations.

Cette étude très préliminaire nous a permis d'identifier des poches jusqu'alors inconnues, communes entre les formes sauvages et mutées ou propres à chacune des espèces. L'investigation concernant les propriétés physicochimiques, géométriques et la druggabilité potentielle de ces proches est encore en cours.

L'ensemble des données générées et des résultats de ma thèse sera le commencement d'exploration plus approfondie des mécanismes d'activation, de résistance et les mécanismes allostériques entourant le RTK KIT et hVKORC1 sauvages et mutés. E plus, cette recherche permettra des avancées thérapeutiques potentielles concernant ces deux protéines (DYNASOME, INTERACTOME). Enfin, les stratégies et protocoles développés lors de ce travail pourront être appliqués à l'étude d'autres IDPs modulaires et leurs interactions mutuelles.

Dans le cas du RTK KIT, la reconstruction de l'INTERACTOME du KIT complet et le développement de tels médicaments *allo-network* spécifiques du KIT seront abordés, je l'espère, dans le cadre du doctorat de Marina Botnari.

PRODUCTIONS SCIENTIFIQUES

L'ensemble de ces résultats a fait l'objet de 6 publications longues, 2 courtes dans des revues à comité de lecture et 2 abstracts longs. La recherche exposée a également été communiquée dans 4 congrès internationaux et 2 conférences où j'ai été invitée. Enfin, une partie des résultats présentés dans la thèse sont le sujet de 2 manuscrits de publications longues en cours de préparation.

Le détail de ces productions scientifiques est présenté ci-après.

En gras, les publications où je fais partie des auteurs et les conférences où j'ai été présentatrice. Le symbole $^{+}$ indique une contribution égale à la recherche publiée.

PUBLICATIONS

Publications longues dans des revues à comité de lecture

1. **Ledoux, J.**, & Tchertanov, L. (2023). Site-Specific Phosphorylation of RTK KIT Kinase Insert Domain: Interactome Landscape Perspectives. *Kinases and Phosphatases*, *1*(1), 39–71. https://doi.org/10.3390/kinasesphosphatases1010005

Special issue: Research on Protein Phosphorylation in Genetic Diseases

2. **Ledoux, J.**, & Tchertanov, L. (2022). Does Generic Cyclic Kinase Insert Domain of Receptor Tyrosine Kinase KIT Clone Its Native Homologue? *International Journal of Molecular Sciences*, *23*(21), 12898. https://doi.org/10.3390/ijms232112898

Special issue: Intrinsically Disordered Proteins (IDPs) 2.0; Section: Molecular Biophysics

3. **Ledoux, J.**, Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

Special issue: Intrinsically Disordered Proteins (IDPs) 2.0; Section: Molecular Biophysics

4. **Ledoux, J.**, Trouvé, A., & Tchertanov, L. (2022). The Inherent Coupling of Intrinsically Disordered Regions in the Multidomain Receptor Tyrosine Kinase KIT. *International Journal of Molecular Sciences*, *23*(3), 1589. https://doi.org/10.3390/ijms23031589

Special issue: Biophysical characterization and Molecular Engineering of Multidomain Proteins 2.0; Section: Molecular Biophysics

5. **Ledoux, J.**, Trouvé, A., & Tchertanov, L. (2021). Folding and Intrinsic Disorder of the Receptor Tyrosine Kinase KIT Insert Domain Seen by Conventional Molecular Dynamics Simulations. *International Journal of Molecular Sciences*, *22*(14), 7375. https://doi.org/10.3390/ijms22147375

Special issue: Structural, Functional and Folding Strategies of Oligomeric Proteins; Section: Molecular Biophysics

6. Stolyarchuk, M.[+], **Ledoux, J.**[+], Maignant, E., Trouvé, A., & Tchertanov, L. (2021). Identification of the Primary Factors Determining the Specificity of Human VKORC1 Recognition by Thioredoxin-Fold Proteins. *International Journal of Molecular Sciences*, *22*(2), 802. https://doi.org/10.3390/ijms22020802

Special issue: Molecular Recognition in Biological and Bioengineered Systems; Section: Molecular Biophysics

Publications courtes dans des revues à comité de lecture

1. **Ledoux, J.**, & Tchertanov, L. (2022). Receptor Tyrosine Kinase KIT: A New Look for an Old Receptor. In I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, & F. Ortuño (Eds.), *Bioinformatics and Biomedical Engineering* (Vol. 13347, pp. 133–137). Springer International Publishing. https://doi.org/10.1007/978-3-031-07802-6_11

2. **Ledoux, J.**, Stolyarchuk, M., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. In I. Rojas, O. Valenzuela, F. Rojas, L. J. Herrera, & F. Ortuño (Eds.), *Bioinformatics and Biomedical Engineering* (Vol. 13347, pp. 138–141). Springer International Publishing. https://doi.org/10.1007/978-3-031-07802-6_12

Manuscrits en préparation

1. **Ledoux, J.**, Botnari, M., & Tchertanov, L. (2023). Receptor Tyrosine Kinase KIT: Mutation-Induced Conformational Shift Promotes Alternating Allosteric Pockets.

2. Botnari, M., **Ledoux, J.**, & Tchertanov, L. (2023). Synergy of Mutation-Induced Effects in Human Vitamin K Epoxide Reductase: Perspectives and Challenges for the Design of Allo-Network Modulators.

ABSTRACTS DE CONFERENCES

1. Tchertanov, L., & **Ledoux, J.** (2022). A first comprehensive look at the order-disorder nature of RTK KIT native and carcinogenic targets. *Global Journal of Cancer Therapy*, *8*(1), 036–039. https://doi.org/10.17352/2581-5407.000046

2. Tchertanov, L., & **Ledoux, J.** (2021). Receptor Tyrosine Kinase KIT: A New Look for an Old Receptor. Book of Abstracts, 21–22.

COMMUNICATIONS ORALES

1. **Ledoux, J.**, & Tchertanov, L. (2022, June 27). *Receptor Tyrosine Kinase KIT: a new Look for an old receptor*. 9th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Gran Canaria, Spain.

2. **Ledoux, J.**, Stolyarchuk, M., & Tchertanov, L. (2022, June 27). *Human Vitamin K Epoxide Reductase as a Target of its Redox Protein*. 9th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO), Gran Canaria, Spain.

3. **Ledoux, J.**, & Tchertanov, L. (2022, April 6). *Receptor tyrosine kinase KIT: functional order-disorder basis and intrinsic coupling pattern*. Modeling of Biological Molecules Workshop (AMMIB), Ecole Polytechnique, Palaiseau, France.

4. **Ledoux, J.**, & Tchertanov, L. (2022, March 30). *Computational design of GluN3A ligand-binding domain to generate the first double-fluorescence biocaptor genetically coded*. DA-Farman Day, Ecole Normale Supérieure Paris-Saclay, Gif-sur-Yvette, France.

CONFÉRENCES INVITÉES

1. **Ledoux, J.** (2023, April 27). *Modularity and Intrinsic Disorder Modularity and Intrinsic Disorder: Molecular conversation between archetypal proteins and their physiological partners.* Institute of Physico-chemical Biology (IBPC). Paris, France.

Invitation : Chantal Prevost and Charles H. Robert

2. **Ledoux, J.,** & Tchertanov, L. (2023, March 20). *A First Comprehensive View on Intrinsic Disorder in Carcinogenic RTK KIT: delivery of a multifaceted target for the development of allosteric modulators for its constitutive activation and post-transduction events* [Webinar]. World Pharma 2023. 2nd Global Virtual Summit in Pharmaceutical and Novel Delivery Systems., London, United Kingdom.

RÉFÉRENCES

1. Hensen, U., Meyer, T., Haas, J., Rex, R., Vriend, G., & Grubmüller, H. (2012). Exploring Protein Dynamics Space: The Dynasome as the Missing Link between Protein Structure and Function. *PLoS ONE*, *7*(5), e33931. https://doi.org/10.1371/journal.pone.0033931

2. Vidal, M. (2005). Interactome modeling. *FEBS Letters*, *579*(8), 1834–1838. https://doi.org/10.1016/j.febslet.2005.02.030

3. Campbell, I. D., & Baron, M. A. (1991). The structure and function of protein modules. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *332*(1263), 165–170. https://doi.org/10.1098/rstb.1991.0045

4. Pawson, T., & Nash, P. (2003). Assembly of Cell Regulatory Systems Through Protein Interaction Domains. *Science*, *300*(5618), 445–452. https://doi.org/10.1126/science.1083653

5. Trenker, R., & Jura, N. (2020). Receptor tyrosine kinase activation: From the ligand perspective. *Current Opinion in Cell Biology*, *63*, 174–185. https://doi.org/10.1016/j.ceb.2020.01.016

6. Cerritelli, S. M., & Crouch, R. J. (2009). Ribonuclease H: The enzymes in eukaryotes: Ribonucleases H of eukaryotes. *FEBS Journal*, *276*(6), 1494–1505. https://doi.org/10.1111/j.1742-4658.2009.06908.x

7. Jadwin, J. A., Ogiue-Ikeda, M., & Machida, K. (2012). The application of modular protein domains in proteomics. *FEBS Letters*, *586*(17), 2586–2596. https://doi.org/10.1016/j.febslet.2012.04.019

8. Cho, Y.-R., & Zhang, A. (2010). Identification of functional hubs and modules by converting interactome networks into hierarchical ordering of proteins. *BMC Bioinformatics*, *11*(S3), S3. https://doi.org/10.1186/1471-2105-11-S3-S3

9. Wang, Y., & Qian, X. (2014). Functional module identification in protein interaction networks by interaction patterns. *Bioinformatics*, *30*(1), 81–93. https://doi.org/10.1093/bioinformatics/btt569

10. Yu, Y., Liu, J., Feng, N., Song, B., & Zheng, Z. (2017). Combining sequence and Gene Ontology for protein module detection in the Weighted Network. *Journal of Theoretical Biology*, *412*, 107–112. https://doi.org/10.1016/j.jtbi.2016.10.010

11. Tasdighian, S., Di Paola, L., De Ruvo, M., Paci, P., Santoni, D., Palumbo, P., Mei, G., Di Venere, A., & Giuliani, A. (2014). Modules Identification in Protein Structures: The Topological and Geometrical Solutions. *Journal of Chemical Information and Modeling*, *54*(1), 159–168. https://doi.org/10.1021/ci400218v

12. Del Sol, A., Araúzo-Bravo, M. J., Amoros, D., & Nussinov, R. (2007). Modular architecture of protein structures and allosteric communications: Potential implications for signaling proteins and regulatory linkages. *Genome Biology*, *8*(5), R92. https://doi.org/10.1186/gb-2007-8-5-r92

13. Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E*, *69*(2), 026113. https://doi.org/10.1103/PhysRevE.69.026113

14. Laine, E., Auclair, C., & Tchertanov, L. (2012). Allosteric Communication across the Native and Mutated KIT Receptor Tyrosine Kinase. *PLoS Computational Biology*, *8*(8), e1002661. https://doi.org/10.1371/journal.pcbi.1002661

15. Allain, A., Chauvot De Beauchêne, I., Langenfeld, F., Guarracino, Y., Laine, E., & Tchertanov, L. (2014). Allosteric pathway identification through network analysis: From molecular dynamics simulations to interactive 2D and 3D graphs. *Faraday Discuss.*, *169*, 303–321. https://doi.org/10.1039/C4FD00024B

16. Uversky, V. N. (2021). The protein disorder cycle. *Biophysical Reviews*, *13*(6), 1155–1162. https://doi.org/10.1007/s12551-021-00853-2

17. Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology*, *337*(3), 635–645. https://doi.org/10.1016/j.jmb.2004.02.002

18. Hu, G., Wang, K., Song, J., Uversky, V. N., & Kurgan, L. (2018). Taxonomic Landscape of the Dark Proteomes: Whole-Proteome Scale Interplay Between Structural Darkness, Intrinsic Disorder, and Crystallization Propensity. *PROTEOMICS*, *18*(21–22), 1800243. https://doi.org/10.1002/pmic.201800243

19. Oldfield, C. J., Xue, B., Van, Y.-Y., Ulrich, E. L., Markley, J. L., Dunker, A. K., & Uversky, V. N. (2013). Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1834*(2), 487–498. https://doi.org/10.1016/j.bbapap.2012.12.003

20. Suskiewicz, M. J., Sussman, J. L., Silman, I., & Shaul, Y. (2011). Context-dependent resistance to proteolysis of intrinsically disordered proteins. *Protein Science*, *20*(8), 1285–1297. https://doi.org/10.1002/pro.657

21. Williams, R. M., Obradovic, Z., Mathura, V., Braun, W., Garner, E. C., Young, J., Takayama, S., Brown, C. J., & Dunker, A. K. (2000). The protein non-folding problem: Amino acid determinants of intrinsic order and disorder. *Biocomputing 2001*, 89–100. https://doi.org/10.1142/9789814447362_0010

22. He, B., Wang, K., Liu, Y., Xue, B., Uversky, V. N., & Dunker, A. K. (2009). Predicting intrinsic disorder in proteins: An overview. *Cell Research*, *19*(8), 929–949. https://doi.org/10.1038/cr.2009.87

23. Zhao, B., & Kurgan, L. (2021). Surveying over 100 predictors of intrinsic disorder in proteins. *Expert Review of Proteomics*, *18*(12), 1019–1029. https://doi.org/10.1080/14789450.2021.2018304

24. Ruff, K. M., & Pappu, R. V. (2021). AlphaFold and Implications for Intrinsically Disordered Proteins. *Journal of Molecular Biology*, *433*(20), 167208. https://doi.org/10.1016/j.jmb.2021.167208

25. Sottini, A., Borgia, A., Borgia, M. B., Bugge, K., Nettels, D., Chowdhury, A., Heidarsson, P. O., Zosel, F., Best, R. B., Kragelund, B. B., & Schuler, B. (2020). Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nature Communications*, *11*(1), 5736. https://doi.org/10.1038/s41467-020-18859-x

26. Holehouse, A. S., Das, R. K., Ahad, J. N., Richardson, M. O. G., & Pappu, R. V. (2017). CIDER: Resources to Analyze Sequence-Ensemble Relationships of Intrinsically Disordered Proteins. *Biophysical Journal*, *112*(1), 16–21. https://doi.org/10.1016/j.bpj.2016.11.3200

27. Papoian, G. A. (2008). Proteins with weakly funneled energy landscapes challenge the classical structure–function paradigm. *Proceedings of the National Academy of Sciences*, *105*(38), 14237–14238. https://doi.org/10.1073/pnas.0807977105

28. Fraga, H., Pujols, J., Gil-Garcia, M., Roque, A., Bernardo-Seisdedos, G., Santambrogio, C., Bech-Serra, J.-J., Canals, F., Bernadó, P., Grandori, R., Millet, O., & Ventura, S. (2017). Disulfide driven folding for a conditionally disordered protein. *Scientific Reports*, *7*(1), 16994. https://doi.org/10.1038/s41598-017-17259-4

29. Romero, P. R., Zaidi, S., Fang, Y. Y., Uversky, V. N., Radivojac, P., Oldfield, C. J., Cortese, M. S., Sickmeier, M., LeGall, T., Obradovic, Z., & Dunker, A. K. (2006). Alternative splicing in concert with protein intrinsic disorder enables increased functional diversity in multicellular organisms. *Proceedings of the National Academy of Sciences*, *103*(22), 8390–8395. https://doi.org/10.1073/pnas.0507916103

30. Pancsa, R., & Tompa, P. (2016). Coding Regions of Intrinsic Disorder Accommodate Parallel Functions. *Trends in Biochemical Sciences*, *41*(11), 898–906. https://doi.org/10.1016/j.tibs.2016.08.009

31. Brown, C. J., Takayama, S., Campen, A. M., Vise, P., Marshall, T. W., Oldfield, C. J., Williams, C. J., & Keith Dunker, A. (2002). Evolutionary Rate Heterogeneity in Proteins with Long Disordered Regions. *Journal of Molecular Evolution*, *55*(1), 104–110. https://doi.org/10.1007/s00239-001-2309-6

32. Sickmeier, M., Hamilton, J. A., LeGall, T., Vacic, V., Cortese, M. S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V. N., Obradovic, Z., & Dunker, A. K. (2007). DisProt: The Database of Disordered Proteins. *Nucleic Acids Research*, *35*(Database), D786–D793. https://doi.org/10.1093/nar/gkl893

33. Tompa, P. (2005). The interplay between structure and function in intrinsically unstructured proteins. *FEBS Letters*, *579*(15), 3346–3354. https://doi.org/10.1016/j.febslet.2005.03.072

34. Papaleo, E., Saladino, G., Lambrughi, M., Lindorff-Larsen, K., Gervasio, F. L., & Nussinov, R. (2016). The Role of Protein Loops and Linkers in Conformational Dynamics and Allostery. *Chemical Reviews*, *116*(11), 6391–6423. https://doi.org/10.1021/acs.chemrev.5b00623

35. Hubbard, S. R., & Miller, W. T. (2007). Receptor tyrosine kinases: Mechanisms of activation and signaling. *Current Opinion in Cell Biology*, *19*(2), 117–123. https://doi.org/10.1016/j.ceb.2007.02.010

36. Uversky, V. N. (2013). Intrinsic Disorder-based Protein Interactions and their Modulators. *Current Pharmaceutical Design*, *19*(23), 4191–4213. https://doi.org/10.2174/1381612811319230005

37. Chan, P. M., Ilangumaran, S., La Rose, J., Chakrabartty, A., & Rottapel, R. (2003). Autoinhibition of the kit receptor tyrosine kinase by the cytosolic juxtamembrane region. *Molecular and Cellular*

Biology, *23*(9), 3067–3078. https://doi.org/10.1128/MCB.23.9.3067-3078.2003

38. Armache, K.-J., Mitterweger, S., Meinhart, A., & Cramer, P. (2005). Structures of Complete RNA Polymerase II and Its Subcomplex, Rpb4/7. *Journal of Biological Chemistry*, *280*(8), 7131–7134. https://doi.org/10.1074/jbc.M413038200

39. Peng, Z., Mizianty, M. J., Xue, B., Kurgan, L., & Uversky, V. N. (2012). More than just tails: Intrinsic disorder in histone proteins. *Molecular BioSystems*, *8*(7), 1886. https://doi.org/10.1039/c2mb25102g

40. Boreikaite, V., Wicky, B. I. M., Watt, I. N., Clarke, J., & Walker, J. E. (2019). Extrinsic conditions influence the self-association and structure of IF1, the regulatory protein of mitochondrial ATP synthase. *Proceedings of the National Academy of Sciences*, *116*(21), 10354–10359. https://doi.org/10.1073/pnas.1903535116

41. Tompa, P., & Csermely, P. (2004). The role of structural disorder in the function of RNA and protein chaperones. *The FASEB Journal*, *18*(11), 1169–1175. https://doi.org/10.1096/fj.04-1584rev

42. Sugase, K., Dyson, H. J., & Wright, P. E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. *Nature*, *447*(7147), 1021–1025. https://doi.org/10.1038/nature05858

43. Berlow, R. B., Dyson, H. J., & Wright, P. E. (2018). Expanding the Paradigm: Intrinsically Disordered Proteins and Allosteric Regulation. *Journal of Molecular Biology*, *430*(16), 2309–2320. https://doi.org/10.1016/j.jmb.2018.04.003

44. Bah, A., & Forman-Kay, J. D. (2016). Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *Journal of Biological Chemistry*, *291*(13), 6696–6705. https://doi.org/10.1074/jbc.R115.695056

45. Otterbein, L. R., Cosio, C., Graceffa, P., & Dominguez, R. (2002). Crystal structures of the vitamin D-binding protein and its complex with actin: Structural basis of the actin-scavenger system. *Proceedings of the National Academy of Sciences*, *99*(12), 8003–8008. https://doi.org/10.1073/pnas.122126299

46. Meziane, O., Piquet, S., Bossé, G. D., Gagné, D., Paquet, E., Robert, C., Tones, M. A., & Simard, M. J. (2015). The human decapping scavenger enzyme DcpS modulates microRNA turnover. *Scientific Reports*, *5*(1), 16688. https://doi.org/10.1038/srep16688

47. Siemens, H., Jackstadt, R., Kaller, M., & Hermeking, H. (2013). Repression of c-Kit by p53 is mediated by miR-34 and is associated with reduced chemoresistance, migration and stemness. *Oncotarget*, *4*(9), 1399–1415. https://doi.org/10.18632/oncotarget.1202

48. Grzybowska, E. (2018). Calcium-Binding Proteins with Disordered Structure and Their Role in Secretion, Storage, and Cellular Signaling. *Biomolecules*, *8*(2), 42. https://doi.org/10.3390/biom8020042

49. Amartely, H., David, A., Shamir, M., Lebendiker, M., Izraeli, S., & Friedler, A. (2016). Differential effects of zinc binding on structured and disordered regions in the multidomain STIL protein.

Chemical Science, *7*(7), 4140–4147. https://doi.org/10.1039/C6SC00115G

50. Hayes, P. L., Lytle, B. L., Volkman, B. F., & Peterson, F. C. (2008). The solution structure of ZNF593 from Homo sapiens reveals a zinc finger in a predominately unstructured protein. *Protein Science*, *17*(3), 571–576. https://doi.org/10.1110/ps.073290408

51. Darling, A. L., Liu, Y., Oldfield, C. J., & Uversky, V. N. (2018). Intrinsically Disordered Proteome of Human Membrane-Less Organelles. *PROTEOMICS*, *18*(5–6), 1700193. https://doi.org/10.1002/pmic.201700193

52. Spector, D. L. (2006). SnapShot: Cellular Bodies. *Cell*, *127*(5), 1071.e1-1071.e2. https://doi.org/10.1016/j.cell.2006.11.026

53. Lee, H., Mok, K. H., Muhandiram, R., Park, K.-H., Suk, J.-E., Kim, D.-H., Chang, J., Sung, Y. C., Choi, K. Y., & Han, K.-H. (2000). Local Structural Elements in the Mostly Unstructured Transcriptional Activation Domain of Human p53. *Journal of Biological Chemistry*, *275*(38), 29426–29432. https://doi.org/10.1074/jbc.M003107200

54. Joerger, A. C., & Fersht, A. R. (2010). The Tumor Suppressor p53: From Structures to Drug Discovery. *Cold Spring Harbor Perspectives in Biology*, *2*(6), a000919–a000919. https://doi.org/10.1101/cshperspect.a000919

55. Mohan, A., Oldfield, C. J., Radivojac, P., Vacic, V., Cortese, M. S., Dunker, A. K., & Uversky, V. N. (2006). Analysis of Molecular Recognition Features (MoRFs). *Journal of Molecular Biology*, *362*(5), 1043–1059. https://doi.org/10.1016/j.jmb.2006.07.087

56. Davey, N. E., Van Roey, K., Weatheritt, R. J., Toedt, G., Uyar, B., Altenberg, B., Budd, A., Diella, F., Dinkel, H., & Gibson, T. J. (2012). Attributes of short linear motifs. *Mol. BioSyst.*, *8*(1), 268–281. https://doi.org/10.1039/C1MB05231D

57. Zanzoni, A., Ribeiro, D. M., & Brun, C. (2019). Understanding protein multifunctionality: From short linear motifs to cellular functions. *Cellular and Molecular Life Sciences*, *76*(22), 4407–4412. https://doi.org/10.1007/s00018-019-03273-4

58. Neduva, V., & Russell, R. B. (2005). Linear motifs: Evolutionary interaction switches. *FEBS Letters*, *579*(15), 3342–3345. https://doi.org/10.1016/j.febslet.2005.04.005

59. Tompa, P., Fuxreiter, M., Oldfield, C. J., Simon, I., Dunker, A. K., & Uversky, V. N. (2009). Close encounters of the third kind: Disordered domains and the interactions of proteins. *BioEssays*, *31*(3), 328–335. https://doi.org/10.1002/bies.200800151

60. Buljan, M., Chalancon, G., Eustermann, S., Wagner, G. P., Fuxreiter, M., Bateman, A., & Babu, M. M. (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, *46*(6), 871–883. https://doi.org/10.1016/j.molcel.2012.05.039

61. Kodani, N., & Nakae, J. (2020). Tissue-Specific Metabolic Regulation of FOXO-Binding Protein: FOXO Does Not Act Alone. *Cells*, *9*(3), 702. https://doi.org/10.3390/cells9030702

62. Gsponer, J., Futschik, M. E., Teichmann, S. A., & Babu, M. M. (2008). Tight Regulation of Unstructured Proteins: From Transcript Synthesis to Protein Degradation. *Science*, *322*(5906), 1365–1368. https://doi.org/10.1126/science.1163581

63. Prakash, S., Tian, L., Ratliff, K. S., Lehotzky, R. E., & Matouschek, A. (2004). An unstructured initiation site is required for efficient proteasome-mediated degradation. *Nature Structural & Molecular Biology*, *11*(9), 830–837. https://doi.org/10.1038/nsmb814

64. Wang, J. Q., Jeelall, Y. S., Ferguson, L. L., & Horikawa, K. (2014). Toll-Like Receptors and Cancer: MYD88 Mutation and Inflammation. *Frontiers in Immunology*, *5*. https://doi.org/10.3389/fimmu.2014.00367

65. Lemmon, M. A., & Schlessinger, J. (2010). Cell Signaling by Receptor Tyrosine Kinases. *Cell*, *141*(7), 1117–1134. https://doi.org/10.1016/j.cell.2010.06.011

66. Volinsky, N., & Kholodenko, B. N. (2013). Complexity of Receptor Tyrosine Kinase Signal Processing. *Cold Spring Harbor Perspectives in Biology*, *5*(8), a009043–a009043. https://doi.org/10.1101/cshperspect.a009043

67. Mészáros, B., Hajdu-Soltész, B., Zeke, A., & Dosztányi, Z. (2021). Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules*, *11*(3), 381. https://doi.org/10.3390/biom11030381

68. Uversky, V. N., Oldfield, C. J., & Dunker, A. K. (2008). Intrinsically Disordered Proteins in Human Diseases: Introducing the D2 Concept. *Annual Review of Biophysics*, *37*(1), 215–246. https://doi.org/10.1146/annurev.biophys.37.032807.125924

69. Uyar, B., Weatheritt, R. J., Dinkel, H., Davey, N. E., & Gibson, T. J. (2014). Proteome-wide analysis of human disease mutations in short linear motifs: Neglected players in cancer? *Mol. BioSyst.*, *10*(10), 2626–2642. https://doi.org/10.1039/C4MB00290C

70. Laine, E., Chauvot De Beauchêne, I., Perahia, D., Auclair, C., & Tchertanov, L. (2011). Mutation D816V Alters the Internal Structure and Dynamics of c-KIT Receptor Cytoplasmic Region: Implications for Dimerization and Activation Mechanisms. *PLoS Computational Biology*, *7*(6), e1002068. https://doi.org/10.1371/journal.pcbi.1002068

71. Seera, S., & Nagarajaram, H. A. (2022). Effect of Disease Causing Missense Mutations on Intrinsically DisorderedRegions in Proteins. *Protein & Peptide Letters*, *29*(3), 254–267. https://doi.org/10.2174/0929866528666211126161200

72. Hegyi, H., Buday, L., & Tompa, P. (2009). Intrinsic Structural Disorder Confers Cellular Viability on Oncogenic Fusion Proteins. *PLoS Computational Biology*, *5*(10), e1000552. https://doi.org/10.1371/journal.pcbi.1000552

73. Dash, D. P., Trap-stamborski, V., Czajkowski, C., Rauscher, D., & Pillai, R. (2018). Identification of a Novel BCR-ABL1 Fusion Transcript in a Chronic Myeloid Leukemia (CML) Patient By Unique Molecular Methods and Subsequent Monitoring of the Patient's Response to Tyrosine Kinase Inhibitor (TKI) Therapy. *Blood*, *132*(Supplement 1), 5453–5453. https://doi.org/10.1182/blood-2018-99-119219

74. Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., De Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E., & Dermitzakis, E. T. (2007). Relative Impact of Nucleotide and Copy Number Variation on Gene Expression Phenotypes. *Science*, *315*(5813), 848–853. https://doi.org/10.1126/science.1136678

75. Smith, H. J., & Simons, C. (Eds.). (2004). *Enzymes and Their Inhibitors: Drug Development* (0 ed.). CRC Press. https://doi.org/10.1201/9780203414583

76. Sutanto, F., Konstantinidou, M., & Dömling, A. (2020). Covalent inhibitors: A rational approach to drug discovery. *RSC Medicinal Chemistry*, *11*(8), 876–884. https://doi.org/10.1039/D0MD00154F

77. Nussinov, R., Tsai, C.-J., & Jang, H. (2021). Anticancer drug resistance: An update and perspective. *Drug Resistance Updates*, *59*, 100796. https://doi.org/10.1016/j.drup.2021.100796

78. Modell, A. E., Blosser, S. L., & Arora, P. S. (2016). Systematic Targeting of Protein–Protein Interactions. *Trends in Pharmacological Sciences*, *37*(8), 702–713. https://doi.org/10.1016/j.tips.2016.05.008

79. Uversky, V. N., Davé, V., Iakoucheva, L. M., Malaney, P., Metallo, S. J., Pathak, R. R., & Joerger, A. C. (2014). Pathological Unfoldomics of Uncontrolled Chaos: Intrinsically Disordered Proteins and Human Diseases. *Chemical Reviews*, *114*(13), 6844–6879. https://doi.org/10.1021/cr400713r

80. Armiento, V., Spanopoulou, A., & Kapurniotu, A. (2020). Peptide-Based Molecular Strategies To Interfere with Protein Misfolding, Aggregation, and Cell Degeneration. *Angewandte Chemie International Edition*, *59*(9), 3372–3384. https://doi.org/10.1002/anie.201906908

81. Zheng, Y., Qu, J., Xue, F., Zheng, Y., Yang, B., Chang, Y., Yang, H., & Zhang, J. (2018). Novel DNA Aptamers for Parkinson's Disease Treatment Inhibit α-Synuclein Aggregation and Facilitate its Degradation. *Molecular Therapy - Nucleic Acids*, *11*, 228–242. https://doi.org/10.1016/j.omtn.2018.02.011

82. Di Giovanni, S., Eleuteri, S., Paleologou, K. E., Yin, G., Zweckstetter, M., Carrupt, P.-A., & Lashuel, H. A. (2010). Entacapone and Tolcapone, Two Catechol O-Methyltransferase Inhibitors, Block Fibril Formation of α-Synuclein and β-Amyloid and Protect against Amyloid-induced Toxicity. *Journal of Biological Chemistry*, *285*(20), 14941–14954. https://doi.org/10.1074/jbc.M109.080390

83. Akoury, E., Gajda, M., Pickhardt, M., Biernat, J., Soraya, P., Griesinger, C., Mandelkow, E., & Zweckstetter, M. (2013). Inhibition of Tau Filament Formation by Conformational Modulation. *Journal of the American Chemical Society*, *135*(7), 2853–2862. https://doi.org/10.1021/ja312471h

84. Oldfield, C. J., Cheng, Y., Cortese, M. S., Romero, P., Uversky, V. N., & Dunker, A. K. (2005). Coupled Folding and Binding with α-Helix-Forming Molecular Recognition Elements. *Biochemistry*, *44*(37), 12454–12470. https://doi.org/10.1021/bi050736e

85. Jung, K.-Y., Wang, H., Teriete, P., Yap, J. L., Chen, L., Lanning, M. E., Hu, A., Lambert, L. J., Holien, T., Sundan, A., Cosford, N. D. P., Prochownik, E. V., & Fletcher, S. (2015). Perturbation of the c-Myc–Max Protein–Protein Interaction via Synthetic α-Helix Mimetics. *Journal of Medicinal Chemistry*, *58*(7), 3002–3024. https://doi.org/10.1021/jm501440q

86. Kiessling, A., Sperl, B., Hollis, A., Eick, D., & Berg, T. (2006). Selective Inhibition of c-Myc/Max Dimerization and DNA Binding by Small Molecules. *Chemistry & Biology*, *13*(7), 745–751. https://doi.org/10.1016/j.chembiol.2006.05.011

87. Cui, Q., & Karplus, M. (2008). Allostery and cooperativity revisited. *Protein Science*, *17*(8), 1295–1307. https://doi.org/10.1110/ps.03259908

88. Wodak, S. J., Paci, E., Dokholyan, N. V., Berezovsky, I. N., Horovitz, A., Li, J., Hilser, V. J., Bahar, I., Karanicolas, J., Stock, G., Hamm, P., Stote, R. H., Eberhardt, J., Chebaro, Y., Dejaegere, A., Cecchini, M., Changeux, J.-P., Bolhuis, P. G., Vreede, J., … McLeish, T. (2019). Allostery in Its Many Disguises: From Theory to Applications. *Structure*, *27*(4), 566–578. https://doi.org/10.1016/j.str.2019.01.003

89. Chennubhotla, C., & Bahar, I. (2006). Markov propagation of allosteric effects in biomolecular systems: Application to GroEL–GroES. *Molecular Systems Biology*, *2*(1), 36. https://doi.org/10.1038/msb4100075

90. Chennubhotla, C., & Bahar, I. (2007). Signal Propagation in Proteins and Relation to Equilibrium Fluctuations. *PLoS Computational Biology*, *3*(9), e172. https://doi.org/10.1371/journal.pcbi.0030172

91. Chennubhotla, C., Yang, Z., & Bahar, I. (2008). Coupling between global dynamics and signal transduction pathways: A mechanism of allostery for chaperonin GroEL. *Molecular BioSystems*, *4*(4), 287. https://doi.org/10.1039/b717819k

92. Piazza, F., & Sanejouand, Y.-H. (2009). Long-range energy transfer in proteins. *Physical Biology*, *6*(4), 046014. https://doi.org/10.1088/1478-3975/6/4/046014

93. Pandini, A., Fornili, A., Fraternali, F., & Kleinjung, J. (2012). Detection of allosteric signal transmission by information-theoretic analysis of protein dynamics. *The FASEB Journal*, *26*(2), 868–881. https://doi.org/10.1096/fj.11-190868

94. Fenton, A. W. (2008). Allostery: An illustrated definition for the 'second secret of life.' *Trends in Biochemical Sciences*, *33*(9), 420–425. https://doi.org/10.1016/j.tibs.2008.05.009

95. Tsai, C.-J., Del Sol, A., & Nussinov, R. (2009). Protein allostery, signal transmission and dynamics: A classification scheme of allosteric mechanisms. *Molecular BioSystems*, *5*(3), 207. https://doi.org/10.1039/b819720b

96. Changeux, J.-P., & Christopoulos, A. (2017). Allosteric modulation as a unifying mechanism for receptor function and regulation. *Diabetes, Obesity and Metabolism*, *19*, 4–21. https://doi.org/10.1111/dom.12959

97. Gunasekaran, K., Ma, B., & Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins: Structure, Function, and Bioinformatics*, *57*(3), 433–443. https://doi.org/10.1002/prot.20232

98. Nussinov, R., Tsai, C.-J., & Liu, J. (2014). Principles of Allosteric Interactions in Cell Signaling. *Journal of the American Chemical Society*, *136*(51), 17692–17701. https://doi.org/10.1021/ja510028c

99. Monod, J., Wyman, J., & Changeux, J.-P. (1965). On the nature of allosteric transitions: A plausible model. *Journal of Molecular Biology*, *12*(1), 88–118. https://doi.org/10.1016/S0022-2836(65)80285-6

100. Koshland, D. E., Némethy, G., & Filmer, D. (1966). Comparison of Experimental Binding Data and Theoretical Models in Proteins Containing Subunits *. *Biochemistry*, *5*(1), 365–385. https://doi.org/10.1021/bi00865a047

101. Hilser, V. J., & Thompson, E. B. (2007). Intrinsic disorder as a mechanism to optimize allosteric coupling in proteins. *Proceedings of the National Academy of Sciences*, *104*(20), 8311–8315. https://doi.org/10.1073/pnas.0700329104

102. Changeux, J.-P. (2013). 50 years of allosteric interactions: The twists and turns of the models. *Nature Reviews Molecular Cell Biology*, *14*(12), 819–829. https://doi.org/10.1038/nrm3695

103. Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature*, *508*(7496), 331–339. https://doi.org/10.1038/nature13001

104. Maria-Solano, M. A., Serrano-Hervás, E., Romero-Rivera, A., Iglesias-Fernández, J., & Osuna, S. (2018). Role of conformational dynamics in the evolution of novel enzyme function. *Chemical Communications*, *54*(50), 6622–6634. https://doi.org/10.1039/C8CC02426J

105. Ahuja, L. G., Aoto, P. C., Kornev, A. P., Veglia, G., & Taylor, S. S. (2019). Dynamic allostery-based molecular workings of kinase:peptide complexes. *Proceedings of the National Academy of Sciences*, *116*(30), 15052–15061. https://doi.org/10.1073/pnas.1900163116

106. Agarwal, P. K., Billeter, S. R., Rajagopalan, P. T. R., Benkovic, S. J., & Hammes-Schiffer, S. (2002). Network of coupled promoting motions in enzyme catalysis. *Proceedings of the National Academy of Sciences*, *99*(5), 2794–2799. https://doi.org/10.1073/pnas.052005999

107. Marsiglia, W. M., Katigbak, J., Zheng, S., Mohammadi, M., Zhang, Y., & Traaseth, N. J. (2019). A Conserved Allosteric Pathway in Tyrosine Kinase Regulation. *Structure*, *27*(8), 1308-1315.e3. https://doi.org/10.1016/j.str.2019.05.002

108. Chauvot De Beauchêne, I., Allain, A., Panel, N., Laine, E., Trouvé, A., Dubreuil, P., & Tchertanov, L. (2014). Hotspot Mutations in KIT Receptor Differentially Modulate Its Allosterically Coupled Conformational Dynamics: Impact on Activation and Drug Sensitivity. *PLoS Computational Biology*, *10*(7), e1003749. https://doi.org/10.1371/journal.pcbi.1003749

109. Ferreon, A. C. M., Ferreon, J. C., Wright, P. E., & Deniz, A. A. (2013). Modulation of allostery by protein intrinsic disorder. *Nature*, *498*(7454), 390–394. https://doi.org/10.1038/nature12294

110. Nussinov, R., & Tsai, C.-J. (2013). Allostery in Disease and in Drug Discovery. *Cell*, *153*(2), 293–305. https://doi.org/10.1016/j.cell.2013.03.034

111. Sheikh, E., Tran, T., Vranic, S., Levy, A., & Bonfil, R. D. (2022). Role and significance of c-KIT receptor tyrosine kinase in cancer: A review. *Bosnian Journal of Basic Medical Sciences*, *22*(5), 683–698. https://doi.org/10.17305/bjbms.2021.7399

112. Reber, L., Da Silva, C. A., & Frossard, N. (2006). Stem cell factor and its receptor c-Kit as targets for inflammatory diseases. *European Journal of Pharmacology*, *533*(1–3), 327–340. https://doi.org/10.1016/j.ejphar.2005.12.067

113. Foster, R., Griffith, R., Ferrao, P., & Ashman, L. (2004). Molecular basis of the constitutive activity and STI571 resistance of Asp816Val mutant KIT receptor tyrosine kinase. *Journal of Molecular Graphics and Modelling*, *23*(2), 139–152. https://doi.org/10.1016/j.jmgm.2004.04.003

114. Roskoski, R. (2005). Structure and regulation of Kit protein-tyrosine kinase—The stem cell factor receptor. *Biochemical and Biophysical Research Communications*, *338*(3), 1307–1315. https://doi.org/10.1016/j.bbrc.2005.09.150

115. Gelfand, I. M., Chothia, C., & Kister, A. E. (2001). Immunoglobulin Fold: Structures of Proteins in the Immunoglobulin Superfamily. In John Wiley & Sons, Ltd (Ed.), *ELS* (1st ed.). Wiley. https://doi.org/10.1038/npg.els.0003051

116. Berman, H. M. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. https://doi.org/10.1093/nar/28.1.235

117. Yuzawa, S., Opatowsky, Y., Zhang, Z., Mandiyan, V., Lax, I., & Schlessinger, J. (2007). Structural Basis for Activation of the Receptor Tyrosine Kinase KIT by Stem Cell Factor. *Cell*, *130*(2), 323–334. https://doi.org/10.1016/j.cell.2007.05.055

118. Muhle-Goll, C., Hoffmann, S., Afonin, S., Grage, S. L., Polyansky, A. A., Windisch, D., Zeitler, M., Bürck, J., & Ulrich, A. S. (2012). Hydrophobic Matching Controls the Tilt and Stability of the Dimeric Platelet-derived Growth Factor Receptor (PDGFR) β Transmembrane Segment. *Journal of Biological Chemistry*, *287*(31), 26178–26186. https://doi.org/10.1074/jbc.M111.325555

119. DiNitto, J. P., Deshmukh, G. D., Zhang, Y., Jacques, S. L., Coli, R., Worrall, J. W., Diehl, W., English, J. M., & Wu, J. C. (2010). Function of activation loop tyrosine phosphorylation in the mechanism of c-Kit auto-activation and its implication in sunitinib resistance. *Journal of Biochemistry*, *147*(4), 601–609. https://doi.org/10.1093/jb/mvq015

120. Lennartsson, J., & Rönnstrand, L. (2012). Stem Cell Factor Receptor/c-Kit: From Basic Science to Clinical Implications. *Physiological Reviews*, *92*(4), 1619–1649. https://doi.org/10.1152/physrev.00046.2011

121. Mol, C. D., Dougan, D. R., Schneider, T. R., Skene, R. J., Kraus, M. L., Scheibe, D. N., Snell, G. P., Zou, H., Sang, B.-C., & Wilson, K. P. (2004). Structural Basis for the Autoinhibition and STI-571 Inhibition of c-Kit Tyrosine Kinase. *Journal of Biological Chemistry*, *279*(30), 31655–31663. https://doi.org/10.1074/jbc.M403319200

122. Mol, C. D., Lim, K. B., Sridhar, V., Zou, H., Chien, E. Y. T., Sang, B.-C., Nowakowski, J., Kassel, D. B., Cronin, C. N., & McRee, D. E. (2003). Structure of a c-Kit Product Complex Reveals the Basis for Kinase Transactivation. *Journal of Biological Chemistry*, *278*(34), 31461–31464. https://doi.org/10.1074/jbc.C300186200

123. Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., Boutselakis, H., Cole, C. G., Creatore, C., Dawson, E., Fish, P., Harsha, B., Hathaway, C., Jupe, S. C., Kok, C. Y., Noble, K., Ponting,

L., Ramshaw, C. C., Rye, C. E., … Forbes, S. A. (2019). COSMIC: The Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research*, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

124. Reiter, A., George, T. I., & Gotlib, J. (2020). New developments in diagnosis, prognostication, and treatment of advanced systemic mastocytosis. *Blood*, *135*(16), 1365–1376. https://doi.org/10.1182/blood.2019000932

125. Crespo, A., & Fernández, A. (2008). Induced Disorder in Protein–Ligand Complexes as a Drug-Design Strategy. *Molecular Pharmaceutics*, *5*(3), 430–437. https://doi.org/10.1021/mp700148h

126. Gajiwala, K. S., Wu, J. C., Christensen, J., Deshmukh, G. D., Diehl, W., DiNitto, J. P., English, J. M., Greig, M. J., He, Y.-A., Jacques, S. L., Lunney, E. A., McTigue, M., Molina, D., Quenzer, T., Wells, P. A., Yu, X., Zhang, Y., Zou, A., Emmett, M. R., … Demetri, G. D. (2009). KIT kinase mutants show unique mechanisms of drug resistance to imatinib and sunitinib in gastrointestinal stromal tumor patients. *Proceedings of the National Academy of Sciences*, *106*(5), 1542–1547. https://doi.org/10.1073/pnas.0812413106

127. Vendôme, J., Letard, S., Martin, F., Svinarchuk, F., Dubreuil, P., Auclair, C., & Le Bret, M. (2005). Molecular Modeling of Wild-Type and D816V c-Kit Inhibition Based on ATP-Competitive Binding of Ellipticine Derivatives to Tyrosine Kinases. *Journal of Medicinal Chemistry*, *48*(20), 6194–6201. https://doi.org/10.1021/jm050231m

128. Rovida, E., & Dello Sbarba, P. (2014). Possible mechanisms and function of nuclear trafficking of the colony-stimulating factor-1 receptor. *Cellular and Molecular Life Sciences*, *71*(19), 3627–3631. https://doi.org/10.1007/s00018-014-1668-2

129. Opatowsky, Y., Lax, I., Tomé, F., Bleichert, F., Unger, V. M., & Schlessinger, J. (2014). Structure, domain organization, and different conformational states of stem cell factor-induced intact KIT dimers. *Proceedings of the National Academy of Sciences*, *111*(5), 1772–1777. https://doi.org/10.1073/pnas.1323254111

130. Bell, C. A., Tynan, J. A., Hart, K. C., Meyer, A. N., Robertson, S. C., & Donoghue, D. J. (2000). Rotational Coupling of the Transmembrane and Kinase Domains of the Neu Receptor Tyrosine Kinase. *Molecular Biology of the Cell*, *11*(10), 3589–3599. https://doi.org/10.1091/mbc.11.10.3589

131. Bougherara, H., Subra, F., Crépin, R., Tauc, P., Auclair, C., & Poul, M.-A. (2009). The Aberrant Localization of Oncogenic Kit Tyrosine Kinase Receptor Mutants Is Reversed on Specific Inhibitory Treatment. *Molecular Cancer Research*, *7*(9), 1525–1533. https://doi.org/10.1158/1541-7786.MCR-09-0138

132. Samayawardhena, L., Hu, J., Stein, P., & Craig, A. (2006). Fyn kinase acts upstream of Shp2 and p38 mitogen-activated protein kinase to promote chemotaxis of mast cells towards stem cell factor. *Cellular Signalling*, *18*(9), 1447–1454. https://doi.org/10.1016/j.cellsig.2005.11.005

133. O'Laughlin-Bunner, B. (2001). Lyn is required for normal stem cell factor-induced proliferation and chemotaxis of primary hematopoietic cells. *Blood*, *98*(2), 343–350. https://doi.org/10.1182/blood.V98.2.343

134. Linnekin, D., DeBerry, C. S., & Mou, S. (1997). Lyn Associates with the Juxtamembrane Region

of c-Kit and Is Activated by Stem Cell Factor in Hematopoietic Cell Lines and Normal Progenitor Cells. *Journal of Biological Chemistry*, *272*(43), 27450–27455. https://doi.org/10.1074/jbc.272.43.27450

135. Shivakrupa, R., & Linnekin, D. (2005). Lyn contributes to regulation of multiple Kit-dependent signaling pathways in murine bone marrow mast cells. *Cellular Signalling*, *17*(1), 103–109. https://doi.org/10.1016/j.cellsig.2004.06.004

136. Gommerman, J. L., Sittaro, D., Klebasz, N. Z., Williams, D. A., & Berger, S. A. (2000). Differential stimulation of c-Kit mutants by membrane-bound and soluble Steel Factor correlates with leukemic potential. *Blood*, *96*(12), 3734–3742.

137. Timokhina, I., Kissel, H., Stella, G., & Besmer, P. (1998). Kit signaling through PI 3-kinase and Src kinase pathways: An essential role for Rac1 and JNK activation in mast cell proliferation. *The EMBO Journal*, *17*(21), 6250–6262. https://doi.org/10.1093/emboj/17.21.6250

138. Price, D. J., Rivnay, B., Fu, Y., Jiang, S., Avraham, S., & Avraham, H. (1997). Direct Association of Csk Homologous Kinase (CHK) with the Diphosphorylated Site Tyr568/570 of the Activated c-KIT in Megakaryocytes. *Journal of Biological Chemistry*, *272*(9), 5915–5920. https://doi.org/10.1074/jbc.272.9.5915

139. Wollberg, P., Lennartsson, J., Gottfridsson, E., Yoshimura, A., & Rönnstrand, L. (2003). The adapter protein APS associates with the multifunctional docking sites Tyr-568 and Tyr-936 in c-Kit. *Biochemical Journal*, *370*(3), 1033–1038. https://doi.org/10.1042/bj20020716

140. Thömmes, K., Lennartsson, J., Carlberg, M., & Rönnstrand, L. (1999). Identification of Tyr-703 and Tyr-936 as the primary association sites for Grb2 and Grb7 in the c-Kit/stem cell factor receptor. *The Biochemical Journal*, *341 (Pt 1)*(Pt 1), 211–216.

141. Ahmed, S. B. M., & Prigent, S. A. (2017). Insights into the Shc Family of Adaptor Proteins. *Journal of Molecular Signaling*, *12*, 2. https://doi.org/10.5334/1750-2187-12-2

142. Serve, H., Hsu, Y. C., & Besmer, P. (1994). Tyrosine residue 719 of the c-kit receptor is essential for binding of the P85 subunit of phosphatidylinositol (PI) 3-kinase and for c-kit-associated PI 3-kinase activity in COS-1 cells. *Journal of Biological Chemistry*, *269*(8), 6026–6030. https://doi.org/10.1016/S0021-9258(17)37564-6

143. Sattler, M., Salgia, R., Shrikhande, G., Verma, S., Pisick, E., Prasad, K. V. S., & Griffin, J. D. (1997). Steel Factor Induces Tyrosine Phosphorylation of CRKL and Binding of CRKL to a Complex Containing c-Kit, Phosphatidylinositol 3-Kinase, and p120CBL. *Journal of Biological Chemistry*, *272*(15), 10248–10253. https://doi.org/10.1074/jbc.272.15.10248

144. Thatcher, J. D. (2010). The Ras-MAPK Signal Transduction Pathway. *Science Signaling*, *3*(119). https://doi.org/10.1126/scisignal.3119tr1

145. Vidal, S., Bouzaher, Y. H., El Motiam, A., Seoane, R., & Rivas, C. (2022). Overview of the regulation of the class IA PI3K/AKT pathway by SUMO. *Seminars in Cell & Developmental Biology*, *132*, 51–61. https://doi.org/10.1016/j.semcdb.2021.10.012

146. Dehbashi, M., Kamali, E., & Vallian, S. (2017). Comparative genomics of human stem cell factor (SCF). *Molecular Biology Research Communications*, *6*(1). https://doi.org/10.22099/mbrc.2017.3919

147. Linnekin, D., Weiler, S. R., Mou, S., DeBerry, C. S., Keller, J. R., Ruscetti, F. W., Ferris, D. K., & Longo, D. L. (1996). JAK2 Is Constitutively Associated with c-Kit and Is Phosphorylated in Response to Stem Cell Factor. *Acta Haematologica*, *95*(3–4), 224–228. https://doi.org/10.1159/000203882

148. Yee, N. S., Langen, H., & Besmer, P. (1993). Mechanism of kit ligand, phorbol ester, and calcium-induced down-regulation of c-kit receptors in mast cells. *The Journal of Biological Chemistry*, *268*(19), 14189–14201.

149. Yee, N. S., Hsiau, C. W., Serve, H., Vosseller, K., & Besmer, P. (1994). Mechanism of down-regulation of c-kit receptor. Roles of receptor tyrosine kinase, phosphatidylinositol 3′-kinase, and protein kinase C. *The Journal of Biological Chemistry*, *269*(50), 31991–31998.

150. Sun, J., Pedersen, M., & Rönnstrand, L. (2008). Gab2 Is Involved in Differential Phosphoinositide 3-Kinase Signaling by Two Splice Forms of c-Kit. *Journal of Biological Chemistry*, *283*(41), 27444–27451. https://doi.org/10.1074/jbc.M709703200

151. Lennartsson, J. (2003). Identification of Tyr900 in the kinase domain of c-Kit as a Src-dependent phosphorylation site mediating interaction with c-Crk. *Experimental Cell Research*, *288*(1), 110–118. https://doi.org/10.1016/S0014-4827(03)00206-4

152. Kazi, J. U., Agarwal, S., Sun, J., Bracco, E., & Rönnstrand, L. (2013). Src-Like Adaptor Protein (SLAP) differentially regulates normal and oncogenic c-Kit signaling. *Journal of Cell Science*, jcs.140590. https://doi.org/10.1242/jcs.140590

153. Bayle, J., Letard, S., Frank, R., Dubreuil, P., & De Sepulveda, P. (2004). Suppressor of Cytokine Signaling 6 Associates with KIT and Regulates KIT Receptor Signaling. *Journal of Biological Chemistry*, *279*(13), 12249–12259. https://doi.org/10.1074/jbc.M313381200

154. Simon, C., Dondi, E., Chaix, A., De Sepulveda, P., Kubiseski, T. J., Varin-Blank, N., & Velazquez, L. (2008). Lnk adaptor protein down-regulates specific Kit-induced signaling pathways in primary mast cells. *Blood*, *112*(10), 4039–4047. https://doi.org/10.1182/blood-2008-05-154849

155. Blume-Jensen, P., Wernstedt, C., Heldin, C.-H., & Rönnstrand, L. (1995). Identification of the Major Phosphorylation Sites for Protein Kinase C in Kit/Stem Cell Factor Receptor in Vitro and in Intact Cells. *Journal of Biological Chemistry*, *270*(23), 14192–14200. https://doi.org/10.1074/jbc.270.23.14192

156. Paulson, R. F., Vesely, S., Siminovitch, K. A., & Bernstein, A. (1996). Signalling by the W/Kit receptor tyrosine kinase is negatively regulated in vivo by the protein tyrosine phosphatase Shp1. *Nature Genetics*, *13*(3), 309–315. https://doi.org/10.1038/ng0796-309

157. Pati, S., Gurudutta, G. U., Kalra, O. P., & Mukhopadhyay, A. (2010). The structural insights of stem cell factor receptor (c-Kit) interaction with tyrosine phosphatase-2 (Shp-2): An in silico analysis. *BMC Research Notes*, *3*(1), 14. https://doi.org/10.1186/1756-0500-3-14

158. Chaix, A., Arcangeli, M.-L., Lopez, S., Voisset, E., Yang, Y., Vita, M., Letard, S., Audebert, S., Finetti,

P., Birnbaum, D., Bertucci, F., Aurrand-Lions, M., Dubreuil, P., & De Sepulveda, P. (2014). KIT-D816V oncogenic activity is controlled by the juxtamembrane docking site Y568-Y570. *Oncogene*, *33*(7), 872–881. https://doi.org/10.1038/onc.2013.12

159. Shi, X., Sousa, L. P., Mandel-Bausch, E. M., Tome, F., Reshetnyak, A. V., Hadari, Y., Schlessinger, J., & Lax, I. (2016). Distinct cellular properties of oncogenic KIT receptor tyrosine kinase mutants enable alternative courses of cancer cell inhibition. *Proceedings of the National Academy of Sciences*, *113*(33). https://doi.org/10.1073/pnas.1610179113

160. Xiang, Z., Kreisel, F., Cain, J., Colson, A., & Tomasson, M. H. (2007). Neoplasia Driven by Mutant c- *KIT* Is Mediated by Intracellular, Not Plasma Membrane, Receptor Signaling. *Molecular and Cellular Biology*, *27*(1), 267–282. https://doi.org/10.1128/MCB.01153-06

161. Taylor, M. L., Dastych, J., Sehgal, D., Sundstrom, M., Nilsson, G., Akin, C., Mage, R. G., & Metcalfe, D. D. (2001). The Kit-activating mutation D816V enhances stem cell factor–dependent chemotaxis. *Blood*, *98*(4), 1195–1199. https://doi.org/10.1182/blood.V98.4.1195

162. Tabone-Eglinger, S., Subra, F., El Sayadi, H., Alberti, L., Tabone, E., Michot, J.-P., Théou-Anton, N., Lemoine, A., Blay, J.-Y., & Emile, J.-F. (2008). KIT Mutations Induce Intracellular Retention and Activation of an Immature Form of the KIT Protein in Gastrointestinal Stromal Tumors. *Clinical Cancer Research*, *14*(8), 2285–2294. https://doi.org/10.1158/1078-0432.CCR-07-4102

163. Obata, Y., Horikawa, K., Takahashi, T., Akieda, Y., Tsujimoto, M., Fletcher, J. A., Esumi, H., Nishida, T., & Abe, R. (2017). Oncogenic signaling by Kit tyrosine kinase occurs selectively on the Golgi apparatus in gastrointestinal stromal tumors. *Oncogene*, *36*(26), 3661–3672. https://doi.org/10.1038/onc.2016.519

164. Sun, J., Pedersen, M., & Rönnstrand, L. (2009). The D816V Mutation of c-Kit Circumvents a Requirement for Src Family Kinases in c-Kit Signal Transduction. *Journal of Biological Chemistry*, *284*(17), 11039–11047. https://doi.org/10.1074/jbc.M808058200

165. Sun, J., Mohlin, S., Lundby, A., Kazi, J. U., Hellman, U., Påhlman, S., Olsen, J. V., & Rönnstrand, L. (2014). The PI3-kinase isoform p110δ is essential for cell transformation induced by the D816V mutant of c-Kit in a lipid-kinase-independent manner. *Oncogene*, *33*(46), 5360–5369. https://doi.org/10.1038/onc.2013.479

166. Chaix, A., Lopez, S., Voisset, E., Gros, L., Dubreuil, P., & De Sepulveda, P. (2011). Mechanisms of STAT Protein Activation by Oncogenic KIT Mutants in Neoplastic Mast Cells. *Journal of Biological Chemistry*, *286*(8), 5956–5966. https://doi.org/10.1074/jbc.M110.182642

167. Voisset, E., Lopez, S., Dubreuil, P., & De Sepulveda, P. (2007). The tyrosine kinase FES is an essential effector of KITD816V proliferation signal. *Blood*, *110*(7), 2593–2599. https://doi.org/10.1182/blood-2007-02-076471

168. Foley, C. J., Freedman, H., Choo, S. L., Onyskiw, C., Fu, N. Y., Yu, V. C., Tuszynski, J., Pratt, J. C., & Baksh, S. (2008). Dynamics of RASSF1A/MOAP-1 Association with Death Receptors. *Molecular and Cellular Biology*, *28*(14), 4520–4535. https://doi.org/10.1128/MCB.02011-07

169. Sharma, S., & Gangenahalli, G. (2016). Gene Expression Profiling of Human c-Kit Mutant

D816V. *Journal of Cancer Therapy*, *07*(06), 439–454. https://doi.org/10.4236/jct.2016.76046

170. Piao, X., Paulson, R., van der Geer, P., Pawson, T., & Bernstein, A. (1996). Oncogenic mutation in the Kit receptor tyrosine kinase alters substrate specificity and induces degradation of the protein tyrosine phosphatase SHP-1. *Proceedings of the National Academy of Sciences*, *93*(25), 14665–14669. https://doi.org/10.1073/pnas.93.25.14665

171. Agarwal, S., Kazi, J. U., Mohlin, S., Påhlman, S., & Rönnstrand, L. (2015). The activation loop tyrosine 823 is essential for the transforming capacity of the c-Kit oncogenic mutant D816V. *Oncogene*, *34*(35), 4581–4590. https://doi.org/10.1038/onc.2014.383

172. Hashimoto, K., Matsumura, I., Tsujimura, T., Kim, D.-K., Ogihara, H., Ikeda, H., Ueda, S., Mizuki, M., Sugahara, H., Shibayama, H., Kitamura, Y., & Kanakura, Y. (2003). Necessity of tyrosine 719 and phosphatidylinositol 3′-kinase–mediated signal pathway in constitutive activation and oncogenic potential of c-kit receptor tyrosine kinase with the Asp814Val mutation. *Blood*, *101*(3), 1094–1102. https://doi.org/10.1182/blood-2002-01-0177

173. Zhang, J., Yang, P. L., & Gray, N. S. (2009). Targeting cancer with small molecule kinase inhibitors. *Nature Reviews Cancer*, *9*(1), 28–39. https://doi.org/10.1038/nrc2559

174. Amitay-Laish, I., Stemmer, S. M., & Lacouture, M. E. (2011). Adverse cutaneous reactions secondary to tyrosine kinase inhibitors including imatinib mesylate, nilotinib, and dasatinib: Practical approach. *Dermatologic Therapy*, *24*(4), 386–395. https://doi.org/10.1111/j.1529-8019.2011.01431.x

175. Roberts, K. G., Odell, A. F., Byrnes, E. M., Baleato, R. M., Griffith, R., Lyons, A. B., & Ashman, L. K. (2007). Resistance to c-KIT kinase inhibitors conferred by V654A mutation. *Molecular Cancer Therapeutics*, *6*(3), 1159–1166. https://doi.org/10.1158/1535-7163.MCT-06-0641

176. Zheng, S., Pan, Y.-L., Tao, D.-Y., Wang, J.-L., & Huang, K.-E. (2009). Secondary C-kit mutation is a cause of acquired resistance to imatinib in gastrointestinal stromal tumor. *Scandinavian Journal of Gastroenterology*, *44*(6), 760–763. https://doi.org/10.1080/00365520802647459

177. Guo, T., Agaram, N. P., Wong, G. C., Hom, G., D'Adamo, D., Maki, R. G., Schwartz, G. K., Veach, D., Clarkson, B. D., Singer, S., DeMatteo, R. P., Besmer, P., & Antonescu, C. R. (2007). Sorafenib Inhibits the Imatinib-Resistant *KIT T670I* Gatekeeper Mutation in Gastrointestinal Stromal Tumor. *Clinical Cancer Research*, *13*(16), 4874–4881. https://doi.org/10.1158/1078-0432.CCR-07-0484

178. Vincenzi, B., Nannini, M., Badalamenti, G., Grignani, G., Fumagalli, E., Gasperoni, S., D'Ambrosio, L., Incorvaia, L., Stellato, M., Spalato Ceruso, M., Napolitano, A., Valeri, S., Santini, D., Tonini, G., Casali, P. G., Dei Tos, A. P., & Pantaleo, M. A. (2018). Imatinib rechallenge in patients with advanced gastrointestinal stromal tumors following progression with imatinib, sunitinib and regorafenib. *Therapeutic Advances in Medical Oncology*, *10*, 175883591879462. https://doi.org/10.1177/1758835918794623

179. Niv, M. Y., Rubin, H., Cohen, J., Tsirulnikov, L., Licht, T., Peretzman-Shemer, A., Cna'an, E., Tartakovsky, A., Stein, I., Albeck, S., Weinstein, I., Goldenberg-Furmanov, M., Tobi, D., Cohen, E., Laster, M., Ben-Sasson, S. A., & Reuveni, H. (2004). Sequence-based Design of Kinase Inhibitors Applicable for Therapeutics and Target Identification. *Journal of Biological Chemistry*, *279*(2), 1242–

1255. https://doi.org/10.1074/jbc.M306723200

180. Schoepfer, J., Jahnke, W., Berellini, G., Buonamici, S., Cotesta, S., Cowan-Jacob, S. W., Dodd, S., Drueckes, P., Fabbro, D., Gabriel, T., Groell, J.-M., Grotzfeld, R. M., Hassan, A. Q., Henry, C., Iyer, V., Jones, D., Lombardo, F., Loo, A., Manley, P. W., … Furet, P. (2018). Discovery of Asciminib (ABL001), an Allosteric Inhibitor of the Tyrosine Kinase Activity of BCR-ABL1. *Journal of Medicinal Chemistry*, *61*(18), 8120–8135. https://doi.org/10.1021/acs.jmedchem.8b01040

181. Ohren, J. F., Chen, H., Pavlovsky, A., Whitehead, C., Zhang, E., Kuffa, P., Yan, C., McConnell, P., Spessard, C., Banotai, C., Mueller, W. T., Delaney, A., Omer, C., Sebolt-Leopold, J., Dudley, D. T., Leung, I. K., Flamme, C., Warmus, J., Kaufman, M., … Hasemann, C. A. (2004). Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nature Structural & Molecular Biology*, *11*(12), 1192–1197. https://doi.org/10.1038/nsmb859

182. Chauvot De Beauchêne, I. (2013). *Structural modeling of activation and resistance mechanisms of a tyrosine kinase, by molecular dynamics and ligand docking.* Ecole Normale Supérieure Cachan.

183. Inizan, F., Hanna, M., Stolyarchuk, M., Chauvot De Beauchêne, I., & Tchertanov, L. (2020). The First 3D Model of the Full-Length KIT Cytoplasmic Domain Reveals a New Look for an Old Receptor. *Scientific Reports*, *10*(1), 5401. https://doi.org/10.1038/s41598-020-62460-7

184. Stenflo, J., Fernlund, P., Egan, W., & Roepstorff, P. (1974). Vitamin K Dependent Modifications of Glutamic Acid Residues in Prothrombin. *Proceedings of the National Academy of Sciences*, *71*(7), 2730–2733. https://doi.org/10.1073/pnas.71.7.2730

185. Willems, B. A. G., Vermeer, C., Reutelingsperger, C. P. M., & Schurgers, L. J. (2014). The realm of vitamin K dependent proteins: Shifting from coagulation toward calcification. *Molecular Nutrition & Food Research*, *58*(8), 1620–1635. https://doi.org/10.1002/mnfr.201300743

186. Fridell, Y.-W. C., Villa, J., Attar, E. C., & Liu, E. T. (1998). GAS6 Induces Axl-mediated Chemotaxis of Vascular Smooth Muscle Cells. *Journal of Biological Chemistry*, *273*(12), 7123–7126. https://doi.org/10.1074/jbc.273.12.7123

187. Goruppi, S., Ruaro, E., Varnum, B., & Schneider, C. (1997). Requirement of Phosphatidylinositol 3-Kinase-Dependent Pathway and Src for Gas6-Axl Mitogenic and Survival Activities in NIH 3T3 Fibroblasts. *Molecular and Cellular Biology*, *17*(8), 4442–4453. https://doi.org/10.1128/MCB.17.8.4442

188. Berkner, K. L., & Runge, K. W. (2004). The physiology of vitamin K nutriture and vitamin K-dependent protein function in atherosclerosis. *Journal of Thrombosis and Haemostasis*, *2*(12), 2118–2132. https://doi.org/10.1111/j.1538-7836.2004.00968.x

189. Jadhav, N., Ajgaonkar, S., Saha, P., Gurav, P., Pandey, A., Basudkar, V., Gada, Y., Panda, S., Jadhav, S., Mehta, D., & Nair, S. (2022). Molecular Pathways and Roles for Vitamin K2-7 as a Health-Beneficial Nutraceutical: Challenges and Opportunities. *Frontiers in Pharmacology*, *13*, 896920. https://doi.org/10.3389/fphar.2022.896920

190. Bjørklund, G., & Chirumbolo, S. (2017). Role of oxidative stress and antioxidants in daily nutrition and human health. *Nutrition*, *33*, 311–321. https://doi.org/10.1016/j.nut.2016.07.018

191. Tanaka, M., & Siemann, D. W. (2021). Therapeutic Targeting of the Gas6/Axl Signaling Pathway in Cancer. *International Journal of Molecular Sciences*, *22*(18), 9953. https://doi.org/10.3390/ijms22189953

192. Gul, S., Maqbool, M. F., Maryam, A., Khan, M., Shakir, H. A., Irfan, M., Ara, C., Li, Y., & Ma, T. (2022). Vitamin K: A novel cancer chemosensitizer. *Biotechnology and Applied Biochemistry*, *69*(6), 2641–2657. https://doi.org/10.1002/bab.2312

193. Markowska, A., Antoszczak, M., Markowska, J., & Huczyński, A. (2022). Role of Vitamin K in Selected Malignant Neoplasms in Women. *Nutrients*, *14*(16), 3401. https://doi.org/10.3390/nu14163401

194. Rost, S., Fregin, A., Ivaskevicius, V., Conzelmann, E., Hörtnagel, K., Pelz, H.-J., Lappegard, K., Seifried, E., Scharrer, I., Tuddenham, E. G. D., Müller, C. R., Strom, T. M., & Oldenburg, J. (2004). Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. *Nature*, *427*(6974), 537–541. https://doi.org/10.1038/nature02214

195. Oldenburg, J., Watzka, M., Rost, S., & Müller, C. R. (2007). VKORC1: Molecular target of coumarins. *Journal of Thrombosis and Haemostasis*, *5*, 1–6. https://doi.org/10.1111/j.1538-7836.2007.02549.x

196. Wallin, R., Wajih, N., & Hutson, S. M. (2008). VKORC1: A Warfarin-Sensitive Enzyme in Vitamin K Metabolism and Biosynthesis of Vitamin K-Dependent Blood Coagulation Factors. In *Vitamins & Hormones* (Vol. 78, pp. 227–246). Elsevier. https://doi.org/10.1016/S0083-6729(07)00011-8

197. Hodroge, A., Matagrin, B., Moreau, C., Fourel, I., Hammed, A., Benoit, E., & Lattard, V. (2012). VKORC1 mutations detected in patients resistant to vitamin K antagonists are not all associated with a resistant VKOR activity. *Journal of Thrombosis and Haemostasis*, *10*(12), 2535–2543. https://doi.org/10.1111/jth.12019

198. Wang, X., Dutton, R. J., Beckwith, J., & Boyd, D. (2011). Membrane Topology and Mutational Analysis of Mycobacterium tuberculosis VKOR, a Protein Involved in Disulfide Bond Formation and a Homologue of Human Vitamin K Epoxide Reductase. *Antioxidants & Redox Signaling*, *14*(8), 1413–1420. https://doi.org/10.1089/ars.2010.3558

199. Cao, Z., Van Lith, M., Mitchell, L. J., Pringle, M. A., Inaba, K., & Bulleid, N. J. (2016). The membrane topology of vitamin K epoxide reductase is conserved between human isoforms and the bacterial enzyme. *Biochemical Journal*, *473*(7), 851–858. https://doi.org/10.1042/BJ20151223

200. Tie, J.-K., Jin, D.-Y., & Stafford, D. W. (2012). Human Vitamin K Epoxide Reductase and Its Bacterial Homologue Have Different Membrane Topologies and Reaction Mechanisms. *Journal of Biological Chemistry*, *287*(41), 33945–33955. https://doi.org/10.1074/jbc.M112.402941

201. Stafford, D. W. (2005). The vitamin K cycle. *Journal of Thrombosis and Haemostasis*, *3*(8), 1873–1878. https://doi.org/10.1111/j.1538-7836.2005.01419.x

202. Wen, L., Chen, J., Duan, L., & Li, S. (2018). Vitamin K-dependent proteins involved in bone and cardiovascular health (Review). *Molecular Medicine Reports*. https://doi.org/10.3892/mmr.2018.8940

203. Janssen, R., Visser, M. P. J., Dofferhoff, A. S. M., Vermeer, C., Janssens, W., & Walk, J. (2021). Vitamin K metabolism as the potential missing link between lung damage and thromboembolism in Coronavirus disease 2019. *British Journal of Nutrition*, *126*(2), 191–198. https://doi.org/10.1017/S0007114520003979

204. Sim, M. M. S., & Wood, J. P. (2022). Dysregulation of Protein S in COVID-19. *Best Practice & Research Clinical Haematology*, *35*(3), 101376. https://doi.org/10.1016/j.beha.2022.101376

205. Klok, F. A., Kruip, M. J. H. A., Van Der Meer, N. J. M., Arbous, M. S., Gommers, D. A. M. P. J., Kant, K. M., Kaptein, F. H. J., Van Paassen, J., Stals, M. A. M., Huisman, M. V., & Endeman, H. (2020). Incidence of thrombotic complications in critically ill ICU patients with COVID-19. *Thrombosis Research*, *191*, 145–147. https://doi.org/10.1016/j.thromres.2020.04.013

206. Brenner, B., Kuperman, A., Watzka, M., & Oldenburg, J. (2009). Vitamin K–Dependent Coagulation Factors Deficiency. *Seminars in Thrombosis and Hemostasis*, *35*(04), 439–446. https://doi.org/10.1055/s-0029-1225766

207. Mladěnka, P., Macáková, K., Kujovská Krčmová, L., Javorská, L., Mrštná, K., Carazo, A., Protti, M., Remião, F., Nováková, L., & the OEMONOM researchers and collaborators. (2022). Vitamin K – sources, physiological role, kinetics, deficiency, detection, therapeutic use, and toxicity. *Nutrition Reviews*, *80*(4), 677–698. https://doi.org/10.1093/nutrit/nuab061

208. Halder, M., Petsophonsakul, P., Akbulut, A., Pavlic, A., Bohan, F., Anderson, E., Maresz, K., Kramann, R., & Schurgers, L. (2019). Vitamin K: Double Bonds beyond Coagulation Insights into Differences between Vitamin K1 and K2 in Health and Disease. *International Journal of Molecular Sciences*, *20*(4), 896. https://doi.org/10.3390/ijms20040896

209. Benton, M. E., Price, P. A., & Suttie, J. W. (1995). Multi-Site-Specificity of the Vitamin K-Dependent Carboxylase: In Vitro Carboxylation of Des-.gamma.-carboxylated Bone Gla Protein and Des-.gamma.-carboxylated Pro Bone Gla Protein. *Biochemistry*, *34*(29), 9541–9551. https://doi.org/10.1021/bi00029a031

210. Tie, J.-K., Jin, D.-Y., Straight, D. L., & Stafford, D. W. (2011). Functional study of the vitamin K cycle in mammalian cells. *Blood*, *117*(10), 2967–2974. https://doi.org/10.1182/blood-2010-08-304303

211. Goodstadt, L. (2004). Vitamin K epoxide reductase: Homology, active site and catalytic mechanism. *Trends in Biochemical Sciences*, *29*(6), 289–292. https://doi.org/10.1016/j.tibs.2004.04.004

212. Silverman, R. B., & Nandi, D. L. (1988). Reduced thioredoxin: A possible physiological cofactor for vitamin k epoxide reductase. further support for an active site disulfide. *Biochemical and Biophysical Research Communications*, *155*(3), 1248–1254. https://doi.org/10.1016/S0006-291X(88)81274-9

213. Davis, C. H., Deerfield, D., Wymore, T., Stafford, D. W., & Pedersen, L. G. (2007). A quantum chemical study of the mechanism of action of Vitamin K epoxide reductase (VKOR). *Journal of Molecular Graphics and Modelling*, *26*(2), 401–408. https://doi.org/10.1016/j.jmgm.2006.10.005

214. Rishavy, M. A., Usubalieva, A., Hallgren, K. W., & Berkner, K. L. (2011). Novel Insight into the Mechanism of the Vitamin K Oxidoreductase (VKOR). *Journal of Biological Chemistry*, *286*(9), 7267–7278. https://doi.org/10.1074/jbc.M110.172213

215. Hartmann, E., Rapoport, T. A., & Lodish, H. F. (1989). Predicting the orientation of eukaryotic membrane-spanning proteins. *Proceedings of the National Academy of Sciences*, *86*(15), 5786–5790. https://doi.org/10.1073/pnas.86.15.5786

216. vonHeijne, G. (1989). Control of topology and mode of assembly of a polytopic membrane protein by positively charged residues. *Nature*, *341*(6241), 456–458. https://doi.org/10.1038/341456a0

217. Liu, S., Cheng, W., Fowle Grider, R., Shen, G., & Li, W. (2014). Structures of an intramembrane vitamin K epoxide reductase homolog reveal control mechanisms for electron transfer. *Nature Communications*, *5*(1), 3110. https://doi.org/10.1038/ncomms4110

218. Chatron, N., Chalmond, B., Trouvé, A., Benoît, E., Caruel, H., Lattard, V., & Tchertanov, L. (2017). Identification of the functional states of human vitamin K epoxide reductase from molecular dynamics simulations. *RSC Advances*, *7*(82), 52071–52090. https://doi.org/10.1039/C7RA07463H

219. Liu, S., Li, S., Shen, G., Sukumar, N., Krezel, A. M., & Li, W. (2021). Structural basis of antagonizing the vitamin K catalytic cycle for anticoagulation. *Science*, *371*(6524), eabc5667. https://doi.org/10.1126/science.abc5667

220. Jin, D.-Y., Tie, J.-K., & Stafford, D. W. (2007). The Conversion of Vitamin K Epoxide to Vitamin K Quinone and Vitamin K Quinone to Vitamin K Hydroquinone Uses the Same Active Site Cysteines. *Biochemistry*, *46*(24), 7279–7283. https://doi.org/10.1021/bi700527j

221. Stolyarchuk, M., Ledoux, J., Maignant, E., Trouvé, A., & Tchertanov, L. (2021). Identification of the Primary Factors Determining the Specificity of Human VKORC1 Recognition by Thioredoxin-Fold Proteins. *International Journal of Molecular Sciences*, *22*(2), 802. https://doi.org/10.3390/ijms22020802

222. Wajih, N., Hutson, S. M., & Wallin, R. (2007). Disulfide-dependent Protein Folding Is Linked to Operation of the Vitamin K Cycle in the Endoplasmic Reticulum. *Journal of Biological Chemistry*, *282*(4), 2626–2635. https://doi.org/10.1074/jbc.M608954200

223. Chetot, T., Benoit, E., Lambert, V., & Lattard, V. (2022). Overexpression of protein disulfide isomerase enhances vitamin K epoxide reductase activity. *Biochemistry and Cell Biology*, *100*(2), 152–161. https://doi.org/10.1139/bcb-2021-0441

224. Schulman, S., Wang, B., Li, W., & Rapoport, T. A. (2010). Vitamin K epoxide reductase prefers ER membrane-anchored thioredoxin-like redox partners. *Proceedings of the National Academy of Sciences*, *107*(34), 15027–15032. https://doi.org/10.1073/pnas.1009972107

225. Tie, J., & Stafford, D. W. (2008). Structure and Function of Vitamin K Epoxide Reductase. In *Vitamins & Hormones* (Vol. 78, pp. 103–130). Elsevier. https://doi.org/10.1016/S0083-6729(07)00006-4

226. Hatahet, F., & Ruddock, L. W. (2009). Protein Disulfide Isomerase: A Critical Evaluation of Its Function in Disulfide Bond Formation. *Antioxidants & Redox Signaling*, *11*(11), 2807–2850. https://doi.org/10.1089/ars.2009.2466

227. Bach, R. D., Dmitrenko, O., & Thorpe, C. (2008). Mechanism of Thiolate−Disulfide Interchange Reactions in Biochemistry. *The Journal of Organic Chemistry*, *73*(1), 12–21. https://doi.org/10.1021/jo702051f

228. Winther, J. R., & Thorpe, C. (2014). Quantification of thiols and disulfides. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1840*(2), 838–846. https://doi.org/10.1016/j.bbagen.2013.03.031

229. Watzka, M., Geisen, C., Bevans, C. G., Sittinger, K., Spohn, G., Rost, S., Seifried, E., Müller, C. R., & Oldenburg, J. (2011). Thirteen novel VKORC1 mutations associated with oral anticoagulant resistance: Insights into improved patient diagnosis and treatment. *Journal of Thrombosis and Haemostasis*, *9*(1), 109–118. https://doi.org/10.1111/j.1538-7836.2010.04095.x

230. Pengo, V., & Denas, G. (2018). Optimizing quality care for the oral vitamin K antagonists (VKAs). *Hematology*, *2018*(1), 332–338. https://doi.org/10.1182/asheducation-2018.1.332

231. Heestermans, M., Poenou, G., Hamzeh-Cognasse, H., Cognasse, F., & Bertoletti, L. (2022). Anticoagulants: A Short History, Their Mechanism of Action, Pharmacology, and Indications. *Cells*, *11*(20), 3214. https://doi.org/10.3390/cells11203214

232. Misra, G. (Ed.). (2017). *Introduction to Biomolecular Structure and Biophysics*. Springer Singapore. https://doi.org/10.1007/978-981-10-4968-2

233. Nwanochie, E., & Uversky, V. N. (2019). Structure Determination by Single-Particle Cryo-Electron Microscopy: Only the Sky (and Intrinsic Disorder) is the Limit. *International Journal of Molecular Sciences*, *20*(17), 4186. https://doi.org/10.3390/ijms20174186

234. Webb, B., & Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics*, *54*(1). https://doi.org/10.1002/cpbi.3

235. Ramachandran, G. N., Ramakrishnan, C., & Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *Journal of Molecular Biology*, *7*(1), 95–99. https://doi.org/10.1016/S0022-2836(63)80023-6

236. He, Y., Chen, Y., Alexander, P., Bryan, P. N., & Orban, J. (2008). NMR structures of two designed proteins with high sequence identity but different fold and function. *Proceedings of the National Academy of Sciences*, *105*(38), 14412–14417. https://doi.org/10.1073/pnas.0805857105

237. Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, *2*, S19–S24. https://doi.org/10.1016/S1359-0278(97)00059-X

238. Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein Structure Prediction Using Rosetta. In *Methods in Enzymology* (Vol. 383, pp. 66–93). Elsevier. https://doi.org/10.1016/S0076-6879(04)83004-0

239. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

240. *Groups Analysis: Zscores—CASP14*. (n.d.). Retrieved June 26, 2023, from https://predictioncenter.org/casp14/zscores_final.cgi

241. Yan, Y., & Huang, S.-Y. (2019). Pushing the accuracy limit of shape complementarity for protein-protein docking. *BMC Bioinformatics*, *20*(Suppl 25), 696. https://doi.org/10.1186/s12859-019-3270-y

242. Murray, C. W., & Rees, D. C. (2009). The rise of fragment-based drug discovery. *Nature Chemistry*, *1*(3), 187–192. https://doi.org/10.1038/nchem.217

243. Hart, T. N., & Read, R. J. (1992). A multiple-start Monte Carlo docking method. *Proteins: Structure, Function, and Genetics*, *13*(3), 206–222. https://doi.org/10.1002/prot.340130304

244. Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, *61*(8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

245. Huang, N., Kalyanaraman, C., Bernacki, K., & Jacobson, M. P. (2006). Molecular mechanics methods for predicting protein–ligand binding. *Phys. Chem. Chem. Phys.*, *8*(44), 5166–5177. https://doi.org/10.1039/B608269F

246. Li, J., Fu, A., & Zhang, L. (2019). An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences*, *11*(2), 320–328. https://doi.org/10.1007/s12539-019-00327-w

247. Feig, M., Onufriev, A., Lee, M. S., Im, W., Case, D. A., & Brooks, C. L. (2004). Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. *Journal of Computational Chemistry*, *25*(2), 265–284. https://doi.org/10.1002/jcc.10378

248. Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, *125*(7), 1731–1737. https://doi.org/10.1021/ja026939x

249. Guedes, I. A., Pereira, F. S. S., & Dardenne, L. E. (2018). Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges. *Frontiers in Pharmacology*, *9*, 1089. https://doi.org/10.3389/fphar.2018.01089

250. Muegge, I. (2000). A knowledge-based scoring function for protein-ligand interactions: Probing the reference state. *Perspectives in Drug Discovery and Design*, *20*(1), 99–114. https://doi.org/10.1023/A:1008729005958

251. Chen, P., Ke, Y., Lu, Y., Du, Y., Li, J., Yan, H., Zhao, H., Zhou, Y., & Yang, Y. (2019). DLIGAND2: An

improved knowledge-based energy function for protein–ligand interactions using the distance-scaled, finite, ideal-gas reference state. *Journal of Cheminformatics*, *11*(1), 52. https://doi.org/10.1186/s13321-019-0373-4

252. Torres, P. H. M., Sodero, A. C. R., Jofily, P., & Silva-Jr, F. P. (2019). Key Topics in Molecular Docking for Drug Design. *International Journal of Molecular Sciences*, *20*(18), 4574. https://doi.org/10.3390/ijms20184574

253. Rosell, M., & Fernández-Recio, J. (2020). Docking approaches for modeling multi-molecular assemblies. *Current Opinion in Structural Biology*, *64*, 59–65. https://doi.org/10.1016/j.sbi.2020.05.016

254. Chmiela, S., Sauceda, H. E., Müller, K.-R., & Tkatchenko, A. (2018). Towards exact molecular dynamics simulations with machine-learned force fields. *Nature Communications*, *9*(1), 3887. https://doi.org/10.1038/s41467-018-06169-2

255. Mackerell, A. D. (2004). Empirical force fields for biological macromolecules: Overview and issues. *Journal of Computational Chemistry*, *25*(13), 1584–1604. https://doi.org/10.1002/jcc.20082

256. Mu, J., Liu, H., Zhang, J., Luo, R., & Chen, H.-F. (2021). Recent Force Field Strategies for Intrinsically Disordered Proteins. *Journal of Chemical Information and Modeling*, *61*(3), 1037–1047. https://doi.org/10.1021/acs.jcim.0c01175

257. Song, D., Luo, R., & Chen, H.-F. (2017). The IDP-Specific Force Field *ff14IDPSFF* Improves the Conformer Sampling of Intrinsically Disordered Proteins. *Journal of Chemical Information and Modeling*, *57*(5), 1166–1178. https://doi.org/10.1021/acs.jcim.7b00135

258. Cui, X., Liu, H., & Chen, H.-F. (2022). Polarizable Force Field of Intrinsically Disordered Proteins with CMAP and Reweighting Optimization. *Journal of Chemical Information and Modeling*, *62*(20), 4970–4982. https://doi.org/10.1021/acs.jcim.2c00835

259. Mu, J., Pan, Z., & Chen, H.-F. (2021). Balanced Solvent Model for Intrinsically Disordered and Ordered Proteins. *Journal of Chemical Information and Modeling*, *61*(10), 5141–5151. https://doi.org/10.1021/acs.jcim.1c00407

260. Miao, Y., Feher, V. A., & McCammon, J. A. (2015). Gaussian Accelerated Molecular Dynamics: Unconstrained Enhanced Sampling and Free Energy Calculation. *Journal of Chemical Theory and Computation*, *11*(8), 3584–3595. https://doi.org/10.1021/acs.jctc.5b00436

261. Miao, Y., & McCammon, J. A. (2017). Gaussian Accelerated Molecular Dynamics: Theory, Implementation, and Applications. In *Annual Reports in Computational Chemistry* (Vol. 13, pp. 231–278). Elsevier. https://doi.org/10.1016/bs.arcc.2017.06.005

262. Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, *22*(12), 2577–2637. https://doi.org/10.1002/bip.360221211

263. Lee, B., & Richards, F. M. (1971). The interpretation of protein structures: Estimation of static accessibility. *Journal of Molecular Biology*, *55*(3), 379-IN4. https://doi.org/10.1016/0022-

2836(71)90324-X

264. Amadei, A., Linssen, A. B. M., & Berendsen, H. J. C. (1993). Essential dynamics of proteins. *Proteins: Structure, Function, and Genetics*, *17*(4), 412–425. https://doi.org/10.1002/prot.340170408

265. Ichiye, T., & Karplus, M. (1991). Collective motions in proteins: A covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Genetics*, *11*(3), 205–217. https://doi.org/10.1002/prot.340110305

266. Bauer, J. A., Pavlović, J., & Bauerová-Hlinková, V. (2019). Normal Mode Analysis as a Routine Part of a Structural Investigation. *Molecules*, *24*(18), 3293. https://doi.org/10.3390/molecules24183293

267. Lyman, E., & Zuckerman, D. M. (2006). Ensemble-Based Convergence Analysis of Biomolecular Trajectories. *Biophysical Journal*, *91*(1), 164–172. https://doi.org/10.1529/biophysj.106.082941

268. Ester, M., Kriegel, H. P., & Sander, J. (n.d.). A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, *96*(34), 226–331.

269. Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

270. Gapsys, V., Michielssens, S., Peters, J. H., De Groot, B. L., & Leonov, H. (2015). Calculation of Binding Free Energies. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (Vol. 1215, pp. 173–209). Springer New York. https://doi.org/10.1007/978-1-4939-1465-4_9

271. Ledoux, J., Trouvé, A., & Tchertanov, L. (2022). The Inherent Coupling of Intrinsically Disordered Regions in the Multidomain Receptor Tyrosine Kinase KIT. *International Journal of Molecular Sciences*, *23*(3), 1589. https://doi.org/10.3390/ijms23031589

272. Locascio, L. E., & Donoghue, D. J. (2013). KIDs rule: Regulatory phosphorylation of RTKs. *Trends in Biochemical Sciences*, *38*(2), 75–84. https://doi.org/10.1016/j.tibs.2012.12.001

273. Hubbard, S. R. (1999). Structural analysis of receptor tyrosine kinases. *Progress in Biophysics and Molecular Biology*, *71*(3–4), 343–358. https://doi.org/10.1016/S0079-6107(98)00047-9

274. Huse, M., & Kuriyan, J. (2002). The Conformational Plasticity of Protein Kinases. *Cell*, *109*(3), 275–282. https://doi.org/10.1016/S0092-8674(02)00741-9

275. Nolen, B., Taylor, S., & Ghosh, G. (2004). Regulation of Protein Kinases. *Molecular Cell*, *15*(5), 661–675. https://doi.org/10.1016/j.molcel.2004.08.024

276. Griffith, J., Black, J., Faerman, C., Swenson, L., Wynn, M., Lu, F., Lippke, J., & Saxena, K. (2004). The Structural Basis for Autoinhibition of FLT3 by the Juxtamembrane Domain. *Molecular Cell*, *13*(2), 169–178. https://doi.org/10.1016/S1097-2765(03)00505-7

277. Wybenga-Groot, L. E., Baskin, B., Ong, S. H., Tong, J., Pawson, T., & Sicheri, F. (2001). Structural

Basis for Autoinhibition of the EphB2 Receptor Tyrosine Kinase by the Unphosphorylated Juxtamembrane Region. *Cell*, *106*(6), 745–757. https://doi.org/10.1016/S0092-8674(01)00496-2

278. Hubbard, S. R. (2006). EGF Receptor Activation: Push Comes to Shove. *Cell*, *125*(6), 1029–1031. https://doi.org/10.1016/j.cell.2006.05.028

279. Blundell, T. L. (2021). Using a synthetic switch to regulate insulin receptor activation. *Proceedings of the National Academy of Sciences*, *118*(33), e2111313118. https://doi.org/10.1073/pnas.2111313118

280. Choi, E., Kikuchi, S., Gao, H., Brodzik, K., Nassour, I., Yopp, A., Singal, A. G., Zhu, H., & Yu, H. (2019). Mitotic regulators and the SHP2-MAPK pathway promote IR endocytosis and feedback regulation of insulin signaling. *Nature Communications*, *10*(1), 1473. https://doi.org/10.1038/s41467-019-09318-3

281. Pham, D. D. M., Guhan, S., & Tsao, H. (2020). KIT and Melanoma: Biological Insights and Clinical Implications. *Yonsei Medical Journal*, *61*(7), 562. https://doi.org/10.3349/ymj.2020.61.7.562

282. Gilreath, J., Tchertanov, L., & Deininger, M. (2019). Novel approaches to treating advanced systemic mastocytosis. *Clinical Pharmacology: Advances and Applications*, *Volume 11*, 77–92. https://doi.org/10.2147/CPAA.S206615

283. Kitamura, Y., & Hirotab, S. (2004). Oncogenic protein tyrosine kinases: Kit as a human oncogenic tyrosine kinase. *Cellular and Molecular Life Sciences*, *61*(23), 2924–2931. https://doi.org/10.1007/s00018-004-4273-y

284. Ghosh, S., Marrocco, I., & Yarden, Y. (2020). Roles for receptor tyrosine kinases in tumor progression and implications for cancer treatment. In *Advances in Cancer Research* (Vol. 147, pp. 1–57). Elsevier. https://doi.org/10.1016/bs.acr.2020.04.002

285. Heldin, C.-H., Östman, A., & Rönnstrand, L. (1998). Signal transduction via platelet-derived growth factor receptors. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, *1378*(1), F79–F113. https://doi.org/10.1016/S0304-419X(98)00015-8

286. Zhang, H.-M., Yu, X., Greig, M. J., Gajiwala, K. S., Wu, J. C., Diehl, W., Lunney, E. A., Emmett, M. R., & Marshall, A. G. (2010). Drug binding and resistance mechanism of KIT tyrosine kinase revealed by hydrogen/deuterium exchange FTICR mass spectrometry: Conformational Basis for KIT Drug Inhibition. *Protein Science*, *19*(4), 703–715. https://doi.org/10.1002/pro.347

287. Amit, I., Wides, R., & Yarden, Y. (2007). Evolvable signaling networks of receptor tyrosine kinases: Relevance of robustness to malignancy and to cancer therapy. *Molecular Systems Biology*, *3*(1), 151. https://doi.org/10.1038/msb4100195

288. Masson, K., Heiss, E., Band, H., & Rönnstrand, L. (2006). Direct binding of Cbl to Tyr568 and Tyr936 of the stem cell factor receptor/c-Kit is required for ligand-induced ubiquitination, internalization and degradation. *Biochemical Journal*, *399*(1), 59–67. https://doi.org/10.1042/BJ20060464

289. Sun, J., Pedersen, M., Bengtsson, S., & Rönnstrand, L. (2007). Grb2 mediates negative regulation

of stem cell factor receptor/c-Kit signaling by recruitment of Cbl. *Experimental Cell Research*, *313*(18), 3935–3942. https://doi.org/10.1016/j.yexcr.2007.08.021

290. Ledoux, J., Trouvé, A., & Tchertanov, L. (2021). Folding and Intrinsic Disorder of the Receptor Tyrosine Kinase KIT Insert Domain Seen by Conventional Molecular Dynamics Simulations. *International Journal of Molecular Sciences*, *22*(14), 7375. https://doi.org/10.3390/ijms22147375

291. Hubbard, S. R. (1997). Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. *The EMBO Journal*, *16*(18), 5572–5581. https://doi.org/10.1093/emboj/16.18.5572

292. Cordes, F. S., Bright, J. N., & Sansom, M. S. P. (2002). Proline-induced Distortions of Transmembrane Helices. *Journal of Molecular Biology*, *323*(5), 951–960. https://doi.org/10.1016/S0022-2836(02)01006-9

293. D'rozario, R. S. G., & Sansom, M. S. P. (2008). Helix dynamics in a membrane transport protein: Comparative simulations of the glycerol-3-phosphate transporter and its constituent helices. *Molecular Membrane Biology*, *25*(6–7), 571–583. https://doi.org/10.1080/09687680802549113

294. Dunton, T. A., Goose, J. E., Gavaghan, D. J., Sansom, M. S. P., & Osborne, J. M. (2014). The Free Energy Landscape of Dimerization of a Membrane Protein, NanC. *PLoS Computational Biology*, *10*(1), e1003417. https://doi.org/10.1371/journal.pcbi.1003417

295. Hinsen, K. (1998). Analysis of domain motions by approximate normal mode calculations. *Proteins: Structure, Function, and Genetics*, *33*(3), 417–429. https://doi.org/10.1002/(SICI)1097-0134(19981115)33:3<417::AID-PROT10>3.0.CO;2-8

296. Bahar, I., Lezon, T. R., Bakan, A., & Shrivastava, I. H. (2010). Normal Mode Analysis of Biomolecular Structures: Functional Mechanisms of Membrane Proteins. *Chemical Reviews*, *110*(3), 1463–1497. https://doi.org/10.1021/cr900095e

297. Appadurai, R., Nagesh, J., & Srivastava, A. (2021). High resolution ensemble description of metamorphic and intrinsically disordered proteins using an efficient hybrid parallel tempering scheme. *Nature Communications*, *12*(1), 958. https://doi.org/10.1038/s41467-021-21105-7

298. Iakoucheva, L. M. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research*, *32*(3), 1037–1049. https://doi.org/10.1093/nar/gkh253

299. Collins, M. O., Yu, L., Campuzano, I., Grant, S. G. N., & Choudhary, J. S. (2008). Phosphoproteomic Analysis of the Mouse Brain Cytosol Reveals a Predominance of Protein Phosphorylation in Regions of Intrinsic Sequence Disorder. *Molecular & Cellular Proteomics*, *7*(7), 1331–1348. https://doi.org/10.1074/mcp.M700564-MCP200

300. Dunker, A. K., Babu, M. M., Barbar, E., Blackledge, M., Bondos, S. E., Dosztányi, Z., Dyson, H. J., Forman-Kay, J., Fuxreiter, M., Gsponer, J., Han, K.-H., Jones, D. T., Longhi, S., Metallo, S. J., Nishikawa, K., Nussinov, R., Obradovic, Z., Pappu, R. V., Rost, B., … Uversky, V. N. (2013). What's in a name? Why these proteins are intrinsically disordered: Why these proteins are intrinsically disordered. *Intrinsically Disordered Proteins*, *1*(1), e24157. https://doi.org/10.4161/idp.24157

301. Uversky, V. N. (2011). Intrinsically disordered proteins from A to Z. *The International Journal of Biochemistry & Cell Biology*, *43*(8), 1090–1103. https://doi.org/10.1016/j.biocel.2011.04.001

302. Uversky, V. N. (2014). Introduction to Intrinsically Disordered Proteins (IDPs). *Chemical Reviews*, *114*(13), 6557–6560. https://doi.org/10.1021/cr500288y

303. Ferrell, J. E., Jr, & Ha, S. H. (2014). Ultrasensitivity part II: Multisite phosphorylation, stoichiometric inhibitors, and positive feedback. *Trends in Biochemical Sciences*, *39*(11), 556–569. https://doi.org/10.1016/j.tibs.2014.09.003

304. Monod, J. (1971). *Chance and necessity: An essay on the natural philosophy of modern biology* (1st American ed.). Knopf.

305. Uversky, V. N. (2013). Unusual biophysics of intrinsically disordered proteins. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1834*(5), 932–951. https://doi.org/10.1016/j.bbapap.2012.12.008

306. Dunker, A. K., Cortese, M. S., Romero, P., Iakoucheva, L. M., & Uversky, V. N. (2005). Flexible nets. The roles of intrinsic disorder in protein interaction networks. *FEBS Journal*, *272*(20), 5129–5148. https://doi.org/10.1111/j.1742-4658.2005.04948.x

307. Tsai, C.-J., Ma, B., Sham, Y. Y., Kumar, S., & Nussinov, R. (2001). Structured disorder and conformational selection. *Proteins: Structure, Function, and Genetics*, *44*(4), 418–427. https://doi.org/10.1002/prot.1107

308. Wright, P. E., & Dyson, H. J. (2015). Intrinsically disordered proteins in cellular signalling and regulation. *Nature Reviews Molecular Cell Biology*, *16*(1), 18–29. https://doi.org/10.1038/nrm3920

309. Schurgers, L. J., & Vermeer, C. (2002). Differential lipoprotein transport pathways of K-vitamins in healthy subjects. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1570*(1), 27–32. https://doi.org/10.1016/S0304-4165(02)00147-2

310. Berkner, K. L. (2008). Vitamin K-Dependent Carboxylation. In *Vitamins & Hormones* (Vol. 78, pp. 131–156). Elsevier. https://doi.org/10.1016/S0083-6729(07)00007-6

311. McCann, J. C., & Ames, B. N. (2009). Vitamin K, an example of triage theory: Is micronutrient inadequacy linked to diseases of aging? *The American Journal of Clinical Nutrition*, *90*(4), 889–907. https://doi.org/10.3945/ajcn.2009.27930

312. Garcia, A. A., & Reitsma, P. H. (2008). VKORC1 and the Vitamin K Cycle. In *Vitamins & Hormones* (Vol. 78, pp. 23–33). Elsevier. https://doi.org/10.1016/S0083-6729(07)00002-7

313. Tie, J. -K., & Stafford, D. W. (2016). Structural and functional insights into enzymes of the vitamin K cycle. *Journal of Thrombosis and Haemostasis*, *14*(2), 236–247. https://doi.org/10.1111/jth.13217

314. Tie, J.-K., & Stafford, D. W. (2017). Functional Study of the Vitamin K Cycle Enzymes in Live Cells. In *Methods in Enzymology* (Vol. 584, pp. 349–394). Elsevier. https://doi.org/10.1016/bs.mie.2016.10.015

315. Pylaeva, S., Brehm, M., & Sebastiani, D. (2018). Salt Bridge in Aqueous Solution: Strong Structural Motifs but Weak Enthalpic Effect. *Scientific Reports*, *8*(1), 13626. https://doi.org/10.1038/s41598-018-31935-z

316. Pace, C. N., Fu, H., Fryar, K. L., Landua, J., Trevino, S. R., Shirley, B. A., Hendricks, M. M., Iimura, S., Gajiwala, K., Scholtz, J. M., & Grimsley, G. R. (2011). Contribution of Hydrophobic Interactions to Protein Stability. *Journal of Molecular Biology*, *408*(3), 514–528. https://doi.org/10.1016/j.jmb.2011.02.053

317. Bevans, C., Krettler, C., Reinhart, C., Watzka, M., & Oldenburg, J. (2015). Phylogeny of the Vitamin K 2,3-Epoxide Reductase (VKOR) Family and Evolutionary Relationship to the Disulfide Bond Formation Protein B (DsbB) Family. *Nutrients*, *7*(8), 6224–6249. https://doi.org/10.3390/nu7085281

318. Davis, S. J., Davies, E. A., Tucknott, M. G., Jones, E. Y., & Van Der Merwe, P. A. (1998). The role of charged residues mediating low affinity protein–protein recognition at the cell surface by CD2. *Proceedings of the National Academy of Sciences*, *95*(10), 5490–5494. https://doi.org/10.1073/pnas.95.10.5490

319. Slagle, S. P., Kozack, R. E., & Subramaniam, S. (1994). Role of Electrostatics in Antibody-Antigen Association: Anti-Hen Egg Lysozyme/Lysozyme Complex (HyHEL-5/HEL). *Journal of Biomolecular Structure and Dynamics*, *12*(2), 439–456. https://doi.org/10.1080/07391102.1994.10508750

320. Nelson, C. A., Viner, N. J., Young, S. P., Petzold, S. J., & Unanue, E. R. (1996). A negatively charged anchor residue promotes high affinity binding to the MHC class II molecule I-Ak. *The Journal of Immunology*, *157*(2), 755–762. https://doi.org/10.4049/jimmunol.157.2.755

321. Stenlund, P., Lindberg, M. J., & Tibell, L. A. E. (2002). Structural Requirements for High-Affinity Heparin Binding: Alanine Scanning Analysis of Charged Residues in the C-Terminal Domain of Human Extracellular Superoxide Dismutase. *Biochemistry*, *41*(9), 3168–3175. https://doi.org/10.1021/bi011454r

322. Schreiber, G. (2002). Kinetic studies of protein–protein interactions. *Current Opinion in Structural Biology*, *12*(1), 41–47. https://doi.org/10.1016/S0959-440X(02)00287-7

323. Wade, R. C., Gabdoulline, R. R., Lüdemann, S. K., & Lounnas, V. (1998). Electrostatic steering and ionic tethering in enzyme–ligand binding: Insights from simulations. *Proceedings of the National Academy of Sciences*, *95*(11), 5942–5949. https://doi.org/10.1073/pnas.95.11.5942

324. Haberland, J., & Gerke, V. (1999). Conserved charged residues in the leucine-rich repeat domain of the Ran GTPase activating protein are required for Ran binding and GTPase activation. *Biochemical Journal*, *343*(3), 653–662. https://doi.org/10.1042/bj3430653

325. Unkles, S. E., Rouch, D. A., Wang, Y., Siddiqi, M. Y., Glass, A. D. M., & Kinghorn, J. R. (2004). Two perfectly conserved arginine residues are required for substrate binding in a high-affinity nitrate transporter. *Proceedings of the National Academy of Sciences*, *101*(50), 17549–17554. https://doi.org/10.1073/pnas.0405054101

326. Zhao, N., Pang, B., Shyu, C.-R., & Korkin, D. (2011). Charged residues at protein interaction

interfaces: Unexpected conservation and orchestrated divergence. *Protein Science*, *20*(7), 1275–1284. https://doi.org/10.1002/pro.655

327. De Planque, M. R. R., Bonev, B. B., Demmers, J. A. A., Greathouse, D. V., Koeppe, R. E., Separovic, F., Watts, A., & Killian, J. A. (2003). Interfacial Anchor Properties of Tryptophan Residues in Transmembrane Peptides Can Dominate over Hydrophobic Matching Effects in Peptide−Lipid Interactions. *Biochemistry*, *42*(18), 5341–5348. https://doi.org/10.1021/bi027000r

328. Chandler, D. (2005). Interfaces and the driving force of hydrophobic assembly. *Nature*, *437*(7059), 640–647. https://doi.org/10.1038/nature04162

329. Uversky, V. N. (2002). Natively unfolded proteins: A point where biology waits for physics. *Protein Science*, *11*(4), 739–756. https://doi.org/10.1110/ps.4210102

330. Dunker, A. K., Brown, C. J., Lawson, J. D., Iakoucheva, L. M., & Obradović, Z. (2002). Intrinsic Disorder and Protein Function. *Biochemistry*, *41*(21), 6573–6582. https://doi.org/10.1021/bi012159+

331. Dunker, A. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., Oldfield, C. J., Campen, A. M., Ratliff, C. M., Hipps, K. W., Ausio, J., Nissen, M. S., Reeves, R., Kang, C., Kissinger, C. R., Bailey, R. W., Griswold, M. D., Chiu, W., Garner, E. C., & Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. https://doi.org/10.1016/S1093-3263(00)00138-8

332. Kamerlin, S. C. L., & Warshel, A. (2010). At the dawn of the 21st century: Is dynamics the missing link for understanding enzyme catalysis?: Critically Examining the Dynamical Proposal. *Proteins: Structure, Function, and Bioinformatics*, *78*(6), 1339–1375. https://doi.org/10.1002/prot.22654

333. Nagy, P. (2013). Kinetics and Mechanisms of Thiol–Disulfide Exchange Covering Direct Substitution and Thiol Oxidation-Mediated Pathways. *Antioxidants & Redox Signaling*, *18*(13), 1623–1641. https://doi.org/10.1089/ars.2012.4973

334. Jensen, K. S., Hansen, R. E., & Winther, J. R. (2009). Kinetic and Thermodynamic Aspects of Cellular Thiol–Disulfide Redox Regulation. *Antioxidants & Redox Signaling*, *11*(5), 1047–1058. https://doi.org/10.1089/ars.2008.2297

335. Carvalho, A. T. P., Swart, M., Van Stralen, J. N. P., Fernandes, P. A., Ramos, M. J., & Bickelhaupt, F. M. (2008). Mechanism of Thioredoxin-Catalyzed Disulfide Reduction. Activation of the Buried Thiol and Role of the Variable Active-Site Residues. *The Journal of Physical Chemistry B*, *112*(8), 2511–2523. https://doi.org/10.1021/jp7104665

336. Li, W., Baldus, I. B., & Gräter, F. (2015). Redox Potentials of Protein Disulfide Bonds from Free-Energy Calculations. *The Journal of Physical Chemistry B*, *119*(17), 5386–5391. https://doi.org/10.1021/acs.jpcb.5b01051

337. Kolšek, K., Aponte-Santamaría, C., & Gräter, F. (2017). Accessibility explains preferred thiol-disulfide isomerization in a protein domain. *Scientific Reports*, *7*(1), 9858. https://doi.org/10.1038/s41598-017-07501-4

338. Schlessinger, J. (2000). Cell Signaling by Receptor Tyrosine Kinases. *Cell*, *103*(2), 211–225. https://doi.org/10.1016/S0092-8674(00)00114-8

339. Rönnstrand, L. (2004). Signal transduction via the stem cell factor receptor/c-Kit. *Cellular and Molecular Life Sciences*, *61*(19–20), 2535–2548. https://doi.org/10.1007/s00018-004-4189-6

340. Oved, S., & Yarden, Y. (2002). Molecular ticket to enter cells. *Nature*, *416*(6877), 133–136. https://doi.org/10.1038/416133a

341. Schramm, A., Bignon, C., Brocca, S., Grandori, R., Santambrogio, C., & Longhi, S. (2019). An arsenal of methods for the experimental characterization of intrinsically disordered proteins – How to choose and combine them? *Archives of Biochemistry and Biophysics*, *676*, 108055. https://doi.org/10.1016/j.abb.2019.07.020

342. Kasahara, K., Terazawa, H., Takahashi, T., & Higo, J. (2019). Studies on Molecular Dynamics of Intrinsically Disordered Proteins and Their Fuzzy Complexes: A Mini-Review. *Computational and Structural Biotechnology Journal*, *17*, 712–720. https://doi.org/10.1016/j.csbj.2019.06.009

343. The AACR Project GENIE Consortium, The AACR Project GENIE Consortium, André, F., Arnedos, M., Baras, A. S., Baselga, J., Bedard, P. L., Berger, M. F., Bierkens, M., Calvo, F., Cerami, E., Chakravarty, D., Dang, K. K., Davidson, N. E., Del Vecchio Fitz, C., Dogan, S., DuBois, R. N., Ducar, M. D., Futreal, P. A., … Zhang, H. (2017). AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discovery*, *7*(8), 818–831. https://doi.org/10.1158/2159-8290.CD-17-0151

344. Bu, Z., & Callaway, D. J. E. (2011). Proteins MOVE! Protein dynamics and long-range allostery in cell signaling. In *Advances in Protein Chemistry and Structural Biology* (Vol. 83, pp. 163–221). Elsevier. https://doi.org/10.1016/B978-0-12-381262-9.00005-7

345. Paul, M. D., & Hristova, K. (2019). The RTK Interactome: Overview and Perspective on RTK Heterointeractions. *Chemical Reviews*, *119*(9), 5881–5921. https://doi.org/10.1021/acs.chemrev.8b00467

346. Habchi, J., Tompa, P., Longhi, S., & Uversky, V. N. (2014). Introducing Protein Intrinsic Disorder. *Chemical Reviews*, *114*(13), 6561–6588. https://doi.org/10.1021/cr400514h

347. Oldfield, C. J., & Dunker, A. K. (2014). Intrinsically Disordered Proteins and Intrinsically Disordered Protein Regions. *Annual Review of Biochemistry*, *83*(1), 553–584. https://doi.org/10.1146/annurev-biochem-072711-164947

348. Mabe, S., Nagamune, T., & Kawahara, M. (2014). Detecting protein–protein interactions based on kinase-mediated growth induction of mammalian cells. *Scientific Reports*, *4*(1), 6127. https://doi.org/10.1038/srep06127

349. Banavali, N. K., & Roux, B. (2005). Free Energy Landscape of A-DNA to B-DNA Conversion in Aqueous Solution. *Journal of the American Chemical Society*, *127*(18), 6866–6876. https://doi.org/10.1021/ja050482k

350. Pietrucci, F. (2017). Strategies for the exploration of free energy landscapes: Unity in diversity

and challenges ahead. *Reviews in Physics*, *2*, 32–45. https://doi.org/10.1016/j.revip.2017.05.001

351. Hénin, J., Fiorin, G., Chipot, C., & Klein, M. L. (2010). Exploring Multidimensional Free Energy Landscapes Using Time-Dependent Biases on Collective Variables. *Journal of Chemical Theory and Computation*, *6*(1), 35–47. https://doi.org/10.1021/ct9004432

352. Wong, E. T. C., So, V., Guron, M., Kuechler, E. R., Malhis, N., Bui, J. M., & Gsponer, J. (2020). Protein–Protein Interactions Mediated by Intrinsically Disordered Protein Regions Are Enriched in Missense Mutations. *Biomolecules*, *10*(8), 1097. https://doi.org/10.3390/biom10081097

353. Dosztányi, Z., Chen, J., Dunker, A. K., Simon, I., & Tompa, P. (2006). Disorder and Sequence Repeats in Hub Proteins and Their Implications for Network Evolution. *Journal of Proteome Research*, *5*(11), 2985–2995. https://doi.org/10.1021/pr060171o

354. Necci, M., Piovesan, D., CAID Predictors, DisProt Curators, & Tosatto, S. C. E. (2020). *Critical Assessment of Protein Intrinsic Disorder Prediction* [Preprint]. Bioinformatics. https://doi.org/10.1101/2020.08.11.245852

355. Dill, K. A. (1990). Dominant forces in protein folding. *Biochemistry*, *29*(31), 7133–7155. https://doi.org/10.1021/bi00483a001

356. Onuchic, J. N., Wolynes, P. G., Luthey-Schulten, Z., & Socci, N. D. (1995). Toward an outline of the topography of a realistic protein-folding funnel. *Proceedings of the National Academy of Sciences*, *92*(8), 3626–3630. https://doi.org/10.1073/pnas.92.8.3626

357. Onuchic, J. N., Socci, N. D., Luthey-Schulten, Z., & Wolynes, P. G. (1996). Protein folding funnels: The nature of the transition state ensemble. *Folding and Design*, *1*(6), 441–450. https://doi.org/10.1016/S1359-0278(96)00060-0

358. Koretke, K. K., Russell, R. B., & Lupas, A. N. (2002). Fold recognition without folds. *Protein Science*, *11*(6), 1575–1579. https://doi.org/10.1110/ps.3590102

359. Onuchic, J. N., & Wolynes, P. G. (2004). Theory of protein folding. *Current Opinion in Structural Biology*, *14*(1), 70–75. https://doi.org/10.1016/j.sbi.2004.01.009

360. Weinkam, P., Zimmermann, J., Romesberg, F. E., & Wolynes, P. G. (2010). The Folding Energy Landscape and Free Energy Excitations of Cytochrome *c*. *Accounts of Chemical Research*, *43*(5), 652–660. https://doi.org/10.1021/ar9002703

361. Wolynes, P. G. (2015). Evolution, energy landscapes and the paradoxes of protein folding. *Biochimie*, *119*, 218–230. https://doi.org/10.1016/j.biochi.2014.12.007

362. Chu, W.-T., & Wang, J. (2018). Quantifying the Intrinsic Conformation Energy Landscape Topography of Proteins with Large-Scale Open–Closed Transition. *ACS Central Science*, *4*(8), 1015–1022. https://doi.org/10.1021/acscentsci.8b00274

363. Ledoux, J., & Tchertanov, L. (2022). Does Generic Cyclic Kinase Insert Domain of Receptor Tyrosine Kinase KIT Clone Its Native Homologue? *International Journal of Molecular Sciences*, *23*(21), 12898. https://doi.org/10.3390/ijms232112898

364. Sergio Hleap, J., & Blouin, C. (2015). The Semantics of the Modular Architecture of Protein Structures. *Current Protein & Peptide Science*, *17*(1), 62–71. https://doi.org/10.2174/1389203716666150923104720

365. Dohmen, E., Klasberg, S., Bornberg-Bauer, E., Perrey, S., & Kemena, C. (2020). The modular nature of protein evolution: Domain rearrangement rates across eukaryotic life. *BMC Evolutionary Biology*, *20*(1), 30. https://doi.org/10.1186/s12862-020-1591-0

366. Wang, Y., Zhang, H., Zhong, H., & Xue, Z. (2021). Protein domain identification methods and online resources. *Computational and Structural Biotechnology Journal*, *19*, 1145–1153. https://doi.org/10.1016/j.csbj.2021.01.041

367. Lim, W. A. (2002). The modular logic of signaling proteins: Building allosteric switches from simple binding domains. *Current Opinion in Structural Biology*, *12*(1), 61–68. https://doi.org/10.1016/S0959-440X(02)00290-7

368. Dueber, J. E., Yeh, B. J., Chak, K., & Lim, W. A. (2003). Reprogramming Control of an Allosteric Signaling Switch Through Modular Recombination. *Science*, *301*(5641), 1904–1908. https://doi.org/10.1126/science.1085945

369. Kuriyan, J. (1993). Modular Protein Structures. *Current Science*, *64*(2), 85–95.

370. Del Sol, A., & Carbonell, P. (2007). The Modular Organization of Domain Structures: Insights into Protein–Protein Binding. *PLoS Computational Biology*, *3*(12), e239. https://doi.org/10.1371/journal.pcbi.0030239

371. Luong, T. D. N., Nagpal, S., Sadqi, M., & Muñoz, V. (2022). A modular approach to map out the conformational landscapes of unbound intrinsically disordered proteins. *Proceedings of the National Academy of Sciences*, *119*(23), e2113572119. https://doi.org/10.1073/pnas.2113572119

372. Iakoucheva, L. M., Brown, C. J., Lawson, J. D., Obradović, Z., & Dunker, A. K. (2002). Intrinsic Disorder in Cell-signaling and Cancer-associated Proteins. *Journal of Molecular Biology*, *323*(3), 573–584. https://doi.org/10.1016/S0022-2836(02)00969-5

373. Weiss, F. U., Daub, H., & Ullrich, A. (1997). Novel mechanisms of RTK signal generation. *Current Opinion in Genetics & Development*, *7*(1), 80–86. https://doi.org/10.1016/S0959-437X(97)80113-X

374. Furitsu, T., Tsujimura, T., Tono, T., Ikeda, H., Kitayama, H., Koshimizu, U., Sugahara, H., Butterfield, J. H., Ashman, L. K., & Kanayama, Y. (1993). Identification of mutations in the coding sequence of the proto-oncogene c-kit in a human mast cell leukemia cell line causing ligand-independent activation of c-kit product. *Journal of Clinical Investigation*, *92*(4), 1736–1744. https://doi.org/10.1172/JCI116761

375. Malaise, M., Steinbach, D., & Corbacioglu, S. (2009). Clinical implications of c-Kit mutations in acute myelogenous leukemia. *Current Hematologic Malignancy Reports*, *4*(2), 77–82. https://doi.org/10.1007/s11899-009-0011-8

376. De Silva, M. C., & Reid, R. (2003). Gastrointestinal stromal tumors (GIST): C-kit mutations, CD117 expression, differential diagnosis and targeted cancer therapy with imatinib. *Pathology*

Oncology Research, *9*(1), 13–19. https://doi.org/10.1007/BF03033708

377. Carvajal, R. D. (2011). KIT as a Therapeutic Target in Metastatic Melanoma. *JAMA*, *305*(22), 2327. https://doi.org/10.1001/jama.2011.746

378. Longley, B. J., Reguera, M. J., & Ma, Y. (2001). Classes of c-KIT activating mutations: Proposed mechanisms of action and implications for disease classification and therapy. *Leukemia Research*, *25*(7), 571–576. https://doi.org/10.1016/S0145-2126(01)00028-5

379. Reshetnyak, A. V., Opatowsky, Y., Boggon, T. J., Folta-Stogniew, E., Tome, F., Lax, I., & Schlessinger, J. (2015). The Strength and Cooperativity of KIT Ectodomain Contacts Determine Normal Ligand-Dependent Stimulation or Oncogenic Activation in Cancer. *Molecular Cell*, *57*(1), 191–201. https://doi.org/10.1016/j.molcel.2014.11.021

380. Dueber, J. E., Yeh, B. J., Bhattacharyya, R. P., & Lim, W. A. (2004). Rewiring cell signaling: The logic and plasticity of eukaryotic protein circuitry. *Current Opinion in Structural Biology*, *14*(6), 690–699. https://doi.org/10.1016/j.sbi.2004.10.004

381. Buck, E., & Iyengar, R. (2003). Organization and Functions of Interacting Domains for Signaling by Protein-Protein Interactions. *Science's STKE*, *2003*(209). https://doi.org/10.1126/stke.2092003re14

382. Schlessinger, J. (2014). Receptor Tyrosine Kinases: Legacy of the First Two Decades. *Cold Spring Harbor Perspectives in Biology*, *6*(3), a008912–a008912. https://doi.org/10.1101/cshperspect.a008912

383. Hovmöller, S., Zhou, T., & Ohlson, T. (2002). Conformations of amino acids in proteins. *Acta Crystallographica Section D Biological Crystallography*, *58*(5), 768–776. https://doi.org/10.1107/S0907444902003359

384. Shapovalov, M. V., & Dunbrack, R. L. (2011). A Smoothed Backbone-Dependent Rotamer Library for Proteins Derived from Adaptive Kernel Density Estimates and Regressions. *Structure*, *19*(6), 844–858. https://doi.org/10.1016/j.str.2011.03.019

385. Inizan, F., Hanna, M., Stolyarchuk, M., Chauvot de Beauchêne, I., & Tchertanov, L. (2020). The First 3D Model of the Full-Length KIT Cytoplasmic Domain Reveals a New Look for an Old Receptor. *Scientific Reports*, *10*(1), 5401. https://doi.org/10.1038/s41598-020-62460-7

386. Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, *3*(1), 1–27. https://doi.org/10.1080/03610927408827101

387. Lätzer, J., Papoian, G. A., Prentiss, M. C., Komives, E. A., & Wolynes, P. G. (2007). Induced Fit, Folding, and Recognition of the NF-κB-Nuclear Localization Signals by IκBα and IκBβ. *Journal of Molecular Biology*, *367*(1), 262–274. https://doi.org/10.1016/j.jmb.2006.12.006

388. Sen, S., & Udgaonkar, J. B. (2019). Binding-induced folding under unfolding conditions: Switching between induced fit and conformational selection mechanisms. *Journal of Biological Chemistry*, *294*(45), 16942–16952. https://doi.org/10.1074/jbc.RA119.009742

389. Balsera, M. A., Wriggers, W., Oono, Y., & Schulten, K. (1996). Principal Component Analysis and Long Time Protein Dynamics. *The Journal of Physical Chemistry*, *100*(7), 2567–2572. https://doi.org/10.1021/jp9536920

390. Woolfson, D. N. (2005). The Design of Coiled-Coil Structures and Assemblies. In *Advances in Protein Chemistry* (Vol. 70, pp. 79–112). Elsevier. https://doi.org/10.1016/S0065-3233(05)70004-8

391. Burkhard, P., Stetefeld, J., & Strelkov, S. V. (2001). Coiled coils: A highly versatile protein folding motif. *Trends in Cell Biology*, *11*(2), 82–88. https://doi.org/10.1016/S0962-8924(00)01898-5

392. Volkert, L. G., & Stoffer, D. A. (2004). A comparison of sequence alignment algorithms for measuring secondary structure similarity. *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing (IEEE Cat. No.04CH37612)*, 182–189. https://doi.org/10.1109/CIBCB.2004.1393952

393. Westerlund, A. M., & Delemotte, L. (2019). InfleCS: Clustering Free Energy Landscapes with Gaussian Mixtures. *Journal of Chemical Theory and Computation*, *15*(12), 6752–6759. https://doi.org/10.1021/acs.jctc.9b00454

394. Braakman, I., & Hebert, D. N. (2013). Protein Folding in the Endoplasmic Reticulum. *Cold Spring Harbor Perspectives in Biology*, *5*(5), a013201–a013201. https://doi.org/10.1101/cshperspect.a013201

395. Hleap, J. S., Susko, E., & Blouin, C. (2013). Defining structural and evolutionary modules in proteins: A community detection approach to explore sub-domain architecture. *BMC Structural Biology*, *13*, 20. https://doi.org/10.1186/1472-6807-13-20

396. Gower, J. C., & Ross, G. J. S. (1969). Minimum Spanning Trees and Single Linkage Cluster Analysis. *Applied Statistics*, *18*(1), 54. https://doi.org/10.2307/2346439

397. Ezerski, J. C., & Cheung, M. S. (2018). CATS: A Tool for Clustering the Ensemble of Intrinsically Disordered Peptides on a Flat Energy Landscape. *The Journal of Physical Chemistry B*, *122*(49), 11807–11816. https://doi.org/10.1021/acs.jpcb.8b08852

398. Teilum, K., Olsen, J. G., & Kragelund, B. B. (2021). On the specificity of protein–protein interactions in the context of disorder. *Biochemical Journal*, *478*(11), 2035–2050. https://doi.org/10.1042/BCJ20200828

399. Tsai, C.-J., Kumar, S., Ma, B., & Nussinov, R. (1999). Folding funnels, binding funnels, and protein function. *Protein Science*, *8*(6), 1181–1190. https://doi.org/10.1110/ps.8.6.1181

400. Iešmantavičius, V., Dogan, J., Jemth, P., Teilum, K., & Kjaergaard, M. (2014). Helical Propensity in an Intrinsically Disordered Protein Accelerates Ligand Binding. *Angewandte Chemie International Edition*, *53*(6), 1548–1551. https://doi.org/10.1002/anie.201307712

401. Levy, Y., Cho, S. S., Onuchic, J. N., & Wolynes, P. G. (2005). A Survey of Flexible Protein Binding Mechanisms and their Transition States Using Native Topology Based Energy Landscapes. *Journal of Molecular Biology*, *346*(4), 1121–1145. https://doi.org/10.1016/j.jmb.2004.12.021

402. Olsen, J. V., Blagoev, B., Gnad, F., Macek, B., Kumar, C., Mortensen, P., & Mann, M. (2006).

Global, In Vivo, and Site-Specific Phosphorylation Dynamics in Signaling Networks. *Cell*, *127*(3), 635–648. https://doi.org/10.1016/j.cell.2006.09.026

403. Bondos, S. E., Dunker, A. K., & Uversky, V. N. (2022). Intrinsically disordered proteins play diverse roles in cell signaling. *Cell Communication and Signaling*, *20*(1), 20. https://doi.org/10.1186/s12964-022-00821-7

404. Binns, K. L., Taylor, P. P., Sicheri, F., Pawson, T., & Holland, S. J. (2000). Phosphorylation of tyrosine residues in the kinase domain and juxtamembrane region regulates the biological and catalytic activities of Eph receptors. *Molecular and Cellular Biology*, *20*(13), 4791–4805. https://doi.org/10.1128/MCB.20.13.4791-4805.2000

405. Edling, C. E., & Hallberg, B. (2007). c-Kit—A hematopoietic cell essential receptor tyrosine kinase. *The International Journal of Biochemistry & Cell Biology*, *39*(11), 1995–1998. https://doi.org/10.1016/j.biocel.2006.12.005

406. Shapiro, P. (2002). Ras-MAP Kinase Signaling Pathways and Control of Cell Proliferation: Relevance to Cancer Therapy. *Critical Reviews in Clinical Laboratory Sciences*, *39*(4–5), 285–330. https://doi.org/10.1080/10408360290795538

407. Lennartsson, J., Jelacic, T., Linnekin, D., & Shivakrupa, R. (2005). Normal and Oncogenic Forms of the Receptor Tyrosine Kinase Kit. *STEM CELLS*, *23*(1), 16–43. https://doi.org/10.1634/stemcells.2004-0117

408. Waudby, C. A., Alvarez-Teijeiro, S., Josue Ruiz, E., Suppinger, S., Pinotsis, N., Brown, P. R., Behrens, A., Christodoulou, J., & Mylona, A. (2022). An intrinsic temporal order of c-JUN N-terminal phosphorylation regulates its activity by orchestrating co-factor recruitment. *Nature Communications*, *13*(1), 6133. https://doi.org/10.1038/s41467-022-33866-w

409. Salazar, C., & Höfer, T. (2006). Kinetic models of phosphorylation cycles: A systematic approach using the rapid-equilibrium approximation for protein–protein interactions. *Biosystems*, *83*(2–3), 195–206. https://doi.org/10.1016/j.biosystems.2005.05.015

410. Mittag, T., Orlicky, S., Choy, W.-Y., Tang, X., Lin, H., Sicheri, F., Kay, L. E., Tyers, M., & Forman-Kay, J. D. (2008). Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proceedings of the National Academy of Sciences*, *105*(46), 17772–17777. https://doi.org/10.1073/pnas.0809222105

411. Huber, A. H., & Weis, W. I. (2001). The Structure of the β-Catenin/E-Cadherin Complex and the Molecular Basis of Diverse Ligand Recognition by β-Catenin. *Cell*, *105*(3), 391–402. https://doi.org/10.1016/S0092-8674(01)00330-0

412. Marsh, J. A. (2013). Buried and Accessible Surface Area Control Intrinsic Protein Flexibility. *Journal of Molecular Biology*, *425*(17), 3250–3263. https://doi.org/10.1016/j.jmb.2013.06.019

413. Lev, S., Givol, D., & Yarden, Y. (1992). Interkinase domain of kit contains the binding site for phosphatidylinositol 3′ kinase. *Proceedings of the National Academy of Sciences*, *89*(2), 678–682. https://doi.org/10.1073/pnas.89.2.678

414. Vajravelu, B. N., Hong, K. U., Al-Maqtari, T., Cao, P., Keith, M. C. L., Wysoczynski, M., Zhao, J., Moore Iv, J. B., & Bolli, R. (2015). C-Kit Promotes Growth and Migration of Human Cardiac Progenitor Cells via the PI3K-AKT and MEK-ERK Pathways. *PLOS ONE*, *10*(10), e0140798. https://doi.org/10.1371/journal.pone.0140798

415. Schlessinger, J., & Lemmon, M. A. (2003). SH2 and PTB Domains in Tyrosine Kinase Signaling. *Science's STKE*, *2003*(191). https://doi.org/10.1126/stke.2003.191.re12

416. Zhou, S. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell*, *72*(5), 767–778. https://doi.org/10.1016/0092-8674(93)90404-E

417. Nolte, R. T., Eck, M. J., Schlessinger, J., Shoelson, S. E., & Harrison, S. C. (1996). Crystal structure of the PI 3-kinase p85 amino-terminal SH2 domain and its phosphopeptide complexes. *Nature Structural & Molecular Biology*, *3*(4), 364–374. https://doi.org/10.1038/nsb0496-364

418. Backer, J. M., Myers, M. G., Shoelson, S. E., Chin, D. J., Sun, X. J., Miralpeix, M., Hu, P., Margolis, B., Skolnik, E. Y., & Schlessinger, J. (1992). Phosphatidylinositol 3′-kinase is activated by association with IRS-1 during insulin stimulation. *The EMBO Journal*, *11*(9), 3469–3479. https://doi.org/10.1002/j.1460-2075.1992.tb05426.x

419. Wagner, M. J., Stacey, M. M., Liu, B. A., & Pawson, T. (2013). Molecular Mechanisms of SH2- and PTB-Domain-Containing Proteins in Receptor Tyrosine Kinase Signaling. *Cold Spring Harbor Perspectives in Biology*, *5*(12), a008987–a008987. https://doi.org/10.1101/cshperspect.a008987

420. Le Guilloux, V., Schmidtke, P., & Tuffery, P. (2009). Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, *10*(1), 168. https://doi.org/10.1186/1471-2105-10-168

421. Marcus, Y., & Hefter, G. (2006). Ion Pairing. *Chemical Reviews*, *106*(11), 4585–4621. https://doi.org/10.1021/cr040087x

422. Lee, S., Kim, S. M., & Lee, R. T. (2013). Thioredoxin and Thioredoxin Target Proteins: From Molecular Mechanisms to Functional Significance. *Antioxidants & Redox Signaling*, *18*(10), 1165–1207. https://doi.org/10.1089/ars.2011.4322

423. Hudson, D. A., Gannon, S. A., & Thorpe, C. (2015). Oxidative protein folding: From thiol–disulfide exchange reactions to the redox poise of the endoplasmic reticulum. *Free Radical Biology and Medicine*, *80*, 171–182. https://doi.org/10.1016/j.freeradbiomed.2014.07.037

424. Martin, J. L. (1995). Thioredoxin—A fold for all reasons. *Structure*, *3*(3), 245–250. https://doi.org/10.1016/S0969-2126(01)00154-X

425. Dobson, C. M., & Karplus, M. (1999). The fundamentals of protein folding: Bringing together theory and experiment. *Current Opinion in Structural Biology*, *9*(1), 92–101. https://doi.org/10.1016/S0959-440X(99)80012-8

426. Wang, C., Li, W., Ren, J., Fang, J., Ke, H., Gong, W., Feng, W., & Wang, C. (2013). Structural Insights into the Redox-Regulated Dynamic Conformations of Human Protein Disulfide Isomerase. *Antioxidants & Redox Signaling*, *19*(1), 36–45. https://doi.org/10.1089/ars.2012.4630

427. Guddat, L. W., Bardwell, J. C., & Martin, J. L. (1998). Crystal structures of reduced and oxidized DsbA: Investigation of domain motion and thiolate stabilization. *Structure*, *6*(6), 757–767. https://doi.org/10.1016/S0969-2126(98)00077-X

428. Rowe, M. L., Ruddock, L. W., Kelly, G., Schmidt, J. M., Williamson, R. A., & Howard, M. J. (2009). Solution Structure and Dynamics of ERp18, a Small Endoplasmic Reticulum Resident Oxidoreductase,. *Biochemistry*, *48*(21), 4596–4606. https://doi.org/10.1021/bi9003342

429. Imai, K., & Mitaku, S. (2005). Mechanisms of secondary structure breakers in soluble proteins. *BIOPHYSICS*, *1*, 55–65. https://doi.org/10.2142/biophysics.1.55

430. Schwartz, R. (2006). Frequencies of hydrophobic and hydrophilic runs and alternations in proteins of known structure. *Protein Science*, *15*(1), 102–112. https://doi.org/10.1110/ps.051741806

431. Kendall, D. G. (1984). Shape Manifolds, Procrustean Metrics, and Complex Projective Spaces. *Bulletin of the London Mathematical Society*, *16*(2), 81–121. https://doi.org/10.1112/blms/16.2.81

432. Dryden, I. L., & Mardia, K. V. (2016). *Statistical shape analysis with applications in R* (Second edition). John Wiley & Sons.

433. Kortemme, T., & Creighton, T. E. (1995). Ionisation of Cysteine Residues at the Termini of Model α-Helical Peptides. Relevance to Unusual Thiol pKaValues in Proteins of the Thioredoxin Family. *Journal of Molecular Biology*, *253*(5), 799–812. https://doi.org/10.1006/jmbi.1995.0592

434. Xu, S., Sankar, S., & Neamati, N. (2014). Protein disulfide isomerase: A promising target for cancer therapy. *Drug Discovery Today*, *19*(3), 222–240. https://doi.org/10.1016/j.drudis.2013.10.017

435. Dyson, H. J., Jeng, M.-F., Tennant, L. L., Slaby, I., Lindell, M., Cui, D.-S., Kuprin, S., & Holmgren, A. (1997). Effects of Buried Charged Groups on Cysteine Thiol Ionization and Reactivity in *Escherichia coli* Thioredoxin: Structural and Functional Characterization of Mutants of Asp 26 and Lys 57. *Biochemistry*, *36*(9), 2622–2636. https://doi.org/10.1021/bi961801a

436. Pinitglang, S., Noble, M., Verma, C., Thomas, E. W., & Brocklehurst, K. (1996). Studies on the enhancement of the reactivity of the (Cys-25)-S-/(His159)-Im+H ion-pair of papain by deprotonation across p *K* a 4. *Biochemical Society Transactions*, *24*(3), 468S-468S. https://doi.org/10.1042/bst024468s

437. Tompa, P., Davey, N. E., Gibson, T. J., & Babu, M. M. (2014). A Million Peptide Motifs for the Molecular Biologist. *Molecular Cell*, *55*(2), 161–169. https://doi.org/10.1016/j.molcel.2014.05.032

438. Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R. J., Daughdrill, G. W., Dunker, A. K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D. T., Kim, P. M., Kriwacki, R. W., Oldfield, C. J., Pappu, R. V., Tompa, P., Uversky, V. N., Wright, P. E., & Babu, M. M. (2014). Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, *114*(13), 6589–6631. https://doi.org/10.1021/cr400525m

439. Takemura, K., & Kitao, A. (2019). More efficient screening of protein-protein complex model structures for reducing the number of candidates. *Biophysics and Physicobiology*, *16*(0), 295–303. https://doi.org/10.2142/biophysico.16.0_295

440. Uversky, V. N. (2013). A decade and a half of protein intrinsic disorder: Biology still waits for physics. *Protein Science: A Publication of the Protein Society*, *22*(6), 693–724. https://doi.org/10.1002/pro.2261

441. Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, *293*(2), 321–331. https://doi.org/10.1006/jmbi.1999.3110

442. Demarest, S. J., Martinez-Yamout, M., Chung, J., Chen, H., Xu, W., Dyson, H. J., Evans, R. M., & Wright, P. E. (2002). Mutual synergistic folding in recruitment of CBP/p300 by p160 nuclear receptor coactivators. *Nature*, *415*(6871), 549–553. https://doi.org/10.1038/415549a

443. Waters, L., Yue, B., Veverka, V., Renshaw, P., Bramham, J., Matsuda, S., Frenkiel, T., Kelly, G., Muskett, F., Carr, M., & Heery, D. M. (2006). Structural Diversity in p160/CREB-binding Protein Coactivator Complexes. *Journal of Biological Chemistry*, *281*(21), 14787–14795. https://doi.org/10.1074/jbc.M600237200

444. Panda, A., & Tuller, T. (2020). Exploring Potential Signals of Selection for Disordered Residues in Prokaryotic and Eukaryotic Proteins. *Genomics, Proteomics & Bioinformatics*, *18*(5), 549–564. https://doi.org/10.1016/j.gpb.2020.06.005

445. Wandless, T. J. (1996). SH2 domains: A question of independence. *Current Biology*, *6*(2), 125–127. https://doi.org/10.1016/S0960-9822(02)00440-2

446. Stanfield, R. L., & Wilson, I. A. (1995). Protein-peptide interactions. *Current Opinion in Structural Biology*, *5*(1), 103–113. https://doi.org/10.1016/0959-440X(95)80015-S

447. Keskin, O., Ma, B., & Nussinov, R. (2005). Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues. *Journal of Molecular Biology*, *345*(5), 1281–1294. https://doi.org/10.1016/j.jmb.2004.10.077

448. Keskin, O., Gursoy, A., Ma, B., & Nussinov, R. (2008). Principles of Protein−Protein Interactions: What are the Preferred Ways For Proteins To Interact? *Chemical Reviews*, *108*(4), 1225–1244. https://doi.org/10.1021/cr040409x

449. Tie, J.-K., Jin, D.-Y., & Stafford, D. W. (2012). Mycobacterium tuberculosis Vitamin K Epoxide Reductase Homologue Supports Vitamin K–Dependent Carboxylation in Mammalian Cells. *Antioxidants & Redox Signaling*, *16*(4), 329–338. https://doi.org/10.1089/ars.2011.4043

450. Bus, K., & Szterk, A. (2021). Relationship between Structure and Biological Activity of Various Vitamin K Forms. *Foods*, *10*(12), 3136. https://doi.org/10.3390/foods10123136

451. Ferland, G. (2012). The Discovery of Vitamin K and Its Clinical Applications. *Annals of Nutrition and Metabolism*, *61*(3), 213–218. https://doi.org/10.1159/000343108

452. Marchetti, G., Caruso, P., Lunghi, B., Pinotti, M., Lapecorella, M., Napolitano, M., Canella, A., Mariani, G., & Bernardi, F. (2008). Vitamin K-induced modification of coagulation phenotype in VKORC1 homozygous deficiency. *Journal of Thrombosis and Haemostasis*, *6*(5), 797–803. https://doi.org/10.1111/j.1538-7836.2008.02934.x

453. Vangone, A., & Bonvin, A. M. (2015). Contacts-based prediction of binding affinity in protein–protein complexes. *ELife*, *4*, e07454. https://doi.org/10.7554/eLife.07454

454. Dill, K. A., Bromberg, S., Yue, K., Chan, H. S., Ftebig, K. M., Yee, D. P., & Thomas, P. D. (2008). Principles of protein folding—A perspective from simple exact models. *Protein Science*, *4*(4), 561–602. https://doi.org/10.1002/pro.5560040401

455. Tsai, C.-J., Lin, S. L., Wolfson, H. J., & Nussinov, R. (1997). Studies of protein-protein interfaces: A statistical analysis of the hydrophobic effect: Protein-protein interfaces: The hydrophobic effect. *Protein Science*, *6*(1), 53–64. https://doi.org/10.1002/pro.5560060106

456. Tsai, C.-J., & Nussinov, R. (1997). Hydrophobic folding units at protein-protein interfaces: Implications to protein folding and to protein-protein association. *Protein Science*, *6*(7), 1426–1437. https://doi.org/10.1002/pro.5560060707

457. Soute, B. A. M., Groenen-van Dooren, M. M. C. L., Holmgren, A., Lundström, J., & Vermeer, C. (1992). Stimulation of the dithiol-dependent reductases in the vitamin K cycle by the thioredoxin system. Strong synergistic effects with protein disulphide-isomerase. *Biochemical Journal*, *281*(1), 255–259. https://doi.org/10.1042/bj2810255

458. Ashman, L. K. (1999). The biology of stem cell factor and its receptor C-kit. *The International Journal of Biochemistry & Cell Biology*, *31*(10), 1037–1051. https://doi.org/10.1016/S1357-2725(99)00076-X

459. Rajan, V., Prykhozhij, S. V., Pandey, A., Cohen, A. M., Rainey, J. K., & Berman, J. N. (2022). KIT D816V is dimerization-independent and activates downstream pathways frequently perturbed in mastocytosis. *British Journal of Haematology*, bjh.18116. https://doi.org/10.1111/bjh.18116

460. Arock, M., Sotlar, K., Akin, C., Broesby-Olsen, S., Hoermann, G., Escribano, L., Kristensen, T. K., Kluin-Nelemans, H. C., Hermine, O., Dubreuil, P., Sperr, W. R., Hartmann, K., Gotlib, J., Cross, N. C. P., Haferlach, T., Garcia-Montero, A., Orfao, A., Schwaab, J., Triggiani, M., … Valent, P. (2015). KIT mutation analysis in mast cell neoplasms: Recommendations of the European Competence Network on Mastocytosis. *Leukemia*, *29*(6), 1223–1232. https://doi.org/10.1038/leu.2015.24

461. Yoshida, C., Yamaguchi, H., Doki, N., Murai, K., Iino, M., Hatta, Y., Onizuka, M., Yokose, N., Fujimaki, K., Hagihara, M., Oshikawa, G., Murayama, K., Kumagai, T., Kimura, S., Najima, Y., Iriyama, N., Tsutsumi, I., Oba, K., Kojima, H., … the Kanto CML Study Group. (2023). Importance of TKI treatment duration in treatment-free remission of chronic myeloid leukemia: Results of the D-FREE study. *International Journal of Hematology*, *117*(5), 694–705. https://doi.org/10.1007/s12185-023-03549-3

462. Da Silva Figueiredo Celestino Gomes, P., Chauvot De Beauchêne, I., Panel, N., Lopez, S., De Sepulveda, P., Geraldo Pascutti, P., Solary, E., & Tchertanov, L. (2016). Insight on Mutation-Induced Resistance from Molecular Dynamics Simulations of the Native and Mutated CSF-1R and KIT. *PLOS ONE*, *11*(7), e0160165. https://doi.org/10.1371/journal.pone.0160165

463. Shah, N. P., Lee, F. Y., Luo, R., Jiang, Y., Donker, M., & Akin, C. (2006). Dasatinib (BMS-354825) inhibits KITD816V, an imatinib-resistant activating mutation that triggers neoplastic growth in most patients with systemic mastocytosis. *Blood*, *108*(1), 286–291. https://doi.org/10.1182/blood-2005-

10-3969

464. Cheng, F., Xu, Q., Li, Q., Cui, Z., Li, W., & Zeng, F. (2023). Adverse reactions after treatment with dasatinib in chronic myeloid leukemia: Characteristics, potential mechanisms, and clinical management strategies. *Frontiers in Oncology*, *13*, 1113462. https://doi.org/10.3389/fonc.2023.1113462

465. Kufareva, I., Ilatovskiy, A. V., & Abagyan, R. (2012). Pocketome: An encyclopedia of small-molecule binding sites in 4D. *Nucleic Acids Research*, *40*(D1), D535–D540. https://doi.org/10.1093/nar/gkr825

466. Nussinov, R., Tsai, C.-J., & Csermely, P. (2011). Allo-network drugs: Harnessing allostery in cellular networks. *Trends in Pharmacological Sciences*, *32*(12), 686–693. https://doi.org/10.1016/j.tips.2011.08.004

467. Ledoux, J., Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

468. Bandara, G., Bai, Y., Chan, E. C., Maric, I., Simakova, O., Wise, S. C., Flynn, D., Metcalfe, D. D., Gilfillan, A. M., & Wilson, T. M. (2011). Targeting the KIT Activating Switch Control Pocket: A Novel Mechanism to Inhibit Mast Cell Activation and KIT D816V Neoplastic Mast Cell Proliferation. *Blood*, *118*(21), 1740–1740. https://doi.org/10.1182/blood.V118.21.1740.1740

469. Zhao, Z., & Bourne, P. E. (2020). Overview of Current Type I/II Kinase Inhibitors. In P. Shapiro (Ed.), *Next Generation Kinase Inhibitors* (pp. 13–28). Springer International Publishing. https://doi.org/10.1007/978-3-030-48283-1_2

470. Okuzumi, T., Fiedler, D., Zhang, C., Gray, D. C., Aizenstein, B., Hoffman, R., & Shokat, K. M. (2009). Inhibitor hijacking of Akt activation. *Nature Chemical Biology*, *5*(7), 484–493. https://doi.org/10.1038/nchembio.183

471. Cameron, A. J. M., Escribano, C., Saurin, A. T., Kostelecky, B., & Parker, P. J. (2009). PKC maturation is promoted by nucleotide pocket occupation independently of intrinsic kinase activity. *Nature Structural & Molecular Biology*, *16*(6), 624–630. https://doi.org/10.1038/nsmb.1606

472. Gopi, H., Umashankara, M., Pirrone, V., LaLonde, J., Madani, N., Tuzer, F., Baxter, S., Zentner, I., Cocklin, S., Jawanda, N., Miller, S. R., Schön, A., Klein, J. C., Freire, E., Krebs, F. C., Smith, A. B., Sodroski, J., & Chaiken, I. (2008). Structural determinants for affinity enhancement of a dual antagonist peptide entry inhibitor of human immunodeficiency virus type-1. *Journal of Medicinal Chemistry*, *51*(9), 2638–2647. https://doi.org/10.1021/jm070814r

473. Schmidtke, P., Bidon-Chanal, A., Luque, F. J., & Barril, X. (2011). MDpocket: Open-source cavity detection and characterization on molecular dynamics trajectories. *Bioinformatics*, *27*(23), 3276–3285. https://doi.org/10.1093/bioinformatics/btr550

474. Shen, M., & Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. *Protein Science*, *15*(11), 2507–2524. https://doi.org/10.1110/ps.062416606

475. Laskowski, R. A., MacArthur, M. W., Moss, D. S., & Thornton, J. M. (1993). PROCHECK: A program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*, *26*(2), 283–291. https://doi.org/10.1107/S0021889892009944

476. Grant, B. J., Rodrigues, A. P. C., ElSawy, K. M., McCammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), 2695–2696. https://doi.org/10.1093/bioinformatics/btl461

477. Jo, S., Kim, T., Iyer, V. G., & Im, W. (2008). CHARMM-GUI: A web-based graphical user interface for CHARMM. *Journal of Computational Chemistry*, *29*(11), 1859–1865. https://doi.org/10.1002/jcc.20945

478. Lomize, A. L., Pogozheva, I. D., Lomize, M. A., & Mosberg, H. I. (2006). Positioning of proteins in membranes: A computational approach. *Protein Science*, *15*(6), 1318–1333. https://doi.org/10.1110/ps.062126106

479. Schott-Verdugo, S., & Gohlke, H. (2019). PACKMOL-Memgen: A Simple-To-Use, Generalized Workflow for Membrane-Protein–Lipid-Bilayer System Building. *Journal of Chemical Information and Modeling*, *59*(6), 2522–2528. https://doi.org/10.1021/acs.jcim.9b00269

480. Case, D. A., Cheatham, T. E., Darden, T., Gohlke, H., Luo, R., Merz, K. M., Onufriev, A., Simmerling, C., Wang, B., & Woods, R. J. (2005). The Amber biomolecular simulation programs. *Journal of Computational Chemistry*, *26*(16), 1668–1688. https://doi.org/10.1002/jcc.20290

481. Maier, J. A., Martinez, C., Kasavajhala, K., Wickstrom, L., Hauser, K. E., & Simmerling, C. (2015). ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *Journal of Chemical Theory and Computation*, *11*(8), 3696–3713. https://doi.org/10.1021/acs.jctc.5b00255

482. Berendsen, H. J. C., Postma, J. P. M., Van Gunsteren, W. F., DiNola, A., & Haak, J. R. (1984). Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics*, *81*(8), 3684–3690. https://doi.org/10.1063/1.448118

483. Duane, S., Kennedy, A. D., Pendleton, B. J., & Roweth, D. (1987). Hybrid Monte Carlo. *Physics Letters B*, *195*(2), 216–222. https://doi.org/10.1016/0370-2693(87)91197-X

484. Andersen, H. C. (1983). Rattle: A "velocity" version of the shake algorithm for molecular dynamics calculations. *Journal of Computational Physics*, *52*(1), 24–34. https://doi.org/10.1016/0021-9991(83)90014-1

485. Roe, D. R., & Cheatham, T. E. (2013). PTRAJ and CPPTRAJ: Software for Processing and Analysis of Molecular Dynamics Trajectory Data. *Journal of Chemical Theory and Computation*, *9*(7), 3084–3095. https://doi.org/10.1021/ct400341p

486. Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: Visual molecular dynamics. *Journal of Molecular Graphics*, *14*(1), 33–38. https://doi.org/10.1016/0263-7855(96)00018-5

487. Sagui, C., Pedersen, L. G., & Darden, T. A. (2004). Towards an accurate representation of electrostatics in classical force fields: Efficient implementation of multipolar interactions in

biomolecular simulations. *The Journal of Chemical Physics*, *120*(1), 73–87. https://doi.org/10.1063/1.1630791

488. Konagurthu, A. S., Lesk, A. M., & Allison, L. (2012). Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics*, *28*(12), i97–i105. https://doi.org/10.1093/bioinformatics/bts223

489. Van Gunsteren, W. F., & Berendsen, H. J. C. (1988). A Leap-frog Algorithm for Stochastic Dynamics. *Molecular Simulation*, *1*(3), 173–185. https://doi.org/10.1080/08927028808080941

490. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Lee, H., & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *The Journal of Chemical Physics*, *103*(19), 8577–8593. https://doi.org/10.1063/1.470117

491. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

492. DeLano, W. L. (2002). Unraveling hot spots in binding interfaces: Progress and challenges. *Current Opinion in Structural Biology*, *12*(1), 14–20. https://doi.org/10.1016/S0959-440X(02)00283-X

493. Trott, O., & Olson, A. J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, *31*(2), 455–461. https://doi.org/10.1002/jcc.21334

494. Tian, C., Kasavajhala, K., Belfon, K. A. A., Raguette, L., Huang, H., Migues, A. N., Bickel, J., Wang, Y., Pincay, J., Wu, Q., & Simmerling, C. (2020). ff19SB: Amino-Acid-Specific Protein Backbone Parameters Trained against Quantum Mechanics Energy Surfaces in Solution. *Journal of Chemical Theory and Computation*, *16*(1), 528–552. https://doi.org/10.1021/acs.jctc.9b00591

495. Meagher, K. L., Redman, L. T., & Carlson, H. A. (2003). Development of polyphosphate parameters for use with the AMBER force field. *Journal of Computational Chemistry*, *24*(9), 1016–1025. https://doi.org/10.1002/jcc.10262

496. Dickson, C. J., Madej, B. D., Skjevik, Å. A., Betz, R. M., Teigen, K., Gould, I. R., & Walker, R. C. (2014). Lipid14: The Amber Lipid Force Field. *Journal of Chemical Theory and Computation*, *10*(2), 865–879. https://doi.org/10.1021/ct4010307

497. Peters, E. A. J. F., Goga, N., & Berendsen, H. J. C. (2014). Stochastic Dynamics with Correct Sampling for Constrained Systems. *Journal of Chemical Theory and Computation*, *10*(10), 4208–4220. https://doi.org/10.1021/ct500380x

498. Evans, D. J., & Holian, B. L. (1985). The Nose–Hoover thermostat. *The Journal of Chemical Physics*, *83*(8), 4069–4074. https://doi.org/10.1063/1.449071

499. Hünenberger, P. H., Mark, A. E., & Van Gunsteren, W. F. (1995). Fluctuation and Cross-correlation Analysis of Protein Motions Observed in Nanosecond Molecular Dynamics Simulations.

Journal of Molecular Biology, *252*(4), 492–503. https://doi.org/10.1006/jmbi.1995.0514

500. Kemmink, J., Darby, N. J., Dijkstra, K., Nilges, M., & Creighton, T. E. (1996). Structure Determination of the N-Terminal Thioredoxin-like Domain of Protein Disulfide Isomerase Using Multidimensional Heteronuclear ^{13}C/ ^{15}N NMR Spectroscopy. *Biochemistry*, *35*(24), 7684–7691. https://doi.org/10.1021/bi960335m

501. Wang, C., Yu, J., Huo, L., Wang, L., Feng, W., & Wang, C. (2012). Human Protein-disulfide Isomerase Is a Redox-regulated Chaperone Activated by Oxidation of Domain a′. *Journal of Biological Chemistry*, *287*(2), 1139–1149. https://doi.org/10.1074/jbc.M111.303149

502. Wang, L., Zhang, L., Niu, Y., Sitia, R., & Wang, C. (2014). Glutathione Peroxidase 7 Utilizes Hydrogen Peroxide Generated by Ero1α to Promote Oxidative Protein Folding. *Antioxidants & Redox Signaling*, *20*(4), 545–556. https://doi.org/10.1089/ars.2013.5236

503. Biterova, E. I., Isupov, M. N., Keegan, R. M., Lebedev, A. A., Sohail, A. A., Liaqat, I., Alanen, H. I., & Ruddock, L. W. (2019). The crystal structure of human microsomal triglyceride transfer protein. *Proceedings of the National Academy of Sciences*, *116*(35), 17251–17260. https://doi.org/10.1073/pnas.1903029116

ANNEXES

A. MATERIELS ET METHODES

A.1. METHODES UTILISEES DANS LA THEMATIQUE RTK KIT

A.1.1. LEDOUX, J., TROUVE, A., & TCHERTANOV, L. (2022). THE INHERENT COUPLING OF INTRINSICALLY DISORDERED REGIONS IN THE MULTIDOMAIN RECEPTOR TYROSINE KINASE KIT. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES*, *23*(3), 1589. HTTPS://DOI.ORG/10.3390/IJMS23031589

A.1.1.1. 3D modelling

The full-length cytoplasmic domain of KIT in the inactive state

The 3D structure of the inactive RTK KIT (K558-R946) was taken at 2 µs of molecular dynamics (MD) simulation, as reported in [385]. Structure of the missing JMR fragment (T544-W557) was retrieved from the PDB structure 3G0E (resolution: 1.6 Å)[126]. Five thousand (5000) models of the full-length cytoplasmic domain of KIT (T544-R946), in the inactive state, were generated for the human sequence P10721 (https://www.uniprot.org/uniprot/ accessed 17 June 2021) using Modeller 10.1[234] and the available structural data. The best model was chosen according to its DOPE score[474] and stereochemical quality (Procheck)[475].

As the generated N-terminal of KIT was not accessible for its linking to the transmembrane domain, the alternative conformations were calculated by the normal mode analysis (NMA) from the R library bio3D[476]. Fifteen (15) modes were calculated at 310 K, with all available force fields. The seventh mode, calculated with the All-atom Elastic Network 2 (aaenm2) force field, showed the greatest fluctuations. As the fragment T544-W557 is rigidified by the strong H-bond T544···K547, a short (1 ns) MD simulation (in water solution) of the inactive KIT was performed, upon constrains to remove this H-bond. The model of KIT, with relaxed N-terminal end ($KIT^{T544-R946}$), allowed the latter modelling steps.

The full-length cytoplasmic domain of KIT in the inactive state linked to the transmembrane helix

The 3D model of the transmembrane α-helix (L521-L543), created with the Builder module of PyMOL 1.9 (http://www.pymol.org/pymol), and last conformation of $KIT^{T544-R946}$, from the 1-ns MD simulation, were used as templates for construction of the full-length cytoplasmic domain of KIT, in the inactive state, with its transmembrane helix

($KIT^{L521\text{-}R946}$).

Five thousand (5000) models of $KIT^{L521\text{-}R946}$ were generated with Modeller 10.1[234] using the human sequence (P10721). To avoid the sticking of the transmembrane domain to the cytoplasmic domain during the minimisation procedure, refinement was performed only on H517-L521 and M541-Y545. The best model was chosen according to its DOPE score[474] and stereochemical quality, assessed with Procheck[475].

The full-length cytoplasmic domain of KIT with transmembrane helix inserted into the membrane

A phosphatidylcholine (POPC) lipid bilayer was generated using Charmm-Gui[477]. As a single transmembrane α-helix was not found in the Orientation of Protein in Membrane (OPM) database[478], the orientation of the double-helix of PDGFR-β, a cousin of KIT, was considered. Suggesting that the single helix in monomer of KIT can have an alternative orientation, $KIT^{L521\text{-}R946}$ was oriented manually, so that its transmembrane helix was positioned perpendicularly to the bilayer with polar residues, next to the phospholipids' polar heads and apolar residues among their tails. Finally, to reduce the number of residues located in the extracellular area, the N-terminal extremity of KIT was reduced to I516-R946 ($KIT^{I516\text{-}R946}$).

A.1.1.2. Molecular dynamics simulations

Systems set-up. Each system, $KIT^{T544\text{-}R946}$ and $KIT^{I516\text{-}R946}$, wrapped in a 23 Å-width leaflet lipid bilayer of POPC (2), were solvated with TIP3P water model in a rectangular box, with the PACKMOL-Memgen[479] and LEaP modules of AmberTools20[480] (http://ambermd.org/AmberTools.php), using the ff14SB all-atom force field[481] for protein and Lipid17 for membrane: (i) hydrogen atoms were added; (ii) protonation states of amino-acids at physiological pH were assigned, as well as the histidine residues protonated on their ε-nitrogen atoms; (iii) no counter-ions were added, as the system is already of neutral charge. The systems, $KIT^{T544\text{-}R946}$, in the water solution (1), and $KIT^{I516\text{-}R946}$, wrapped in a 23 Å-width leaflet lipid bilayer of POPC (2), contained 71,062 atoms in total, with 6415 atoms of protein and 64,647 atoms of water (1), as well as 172,368 atoms in total with 6869 atoms of protein, 42,074 atoms of the lipid membrane, and 123,423 atoms of water (2).

Minimisation, equilibration and data generation. The systems were equilibrated using the Sander module of AmberTools20. For system (1), 30,000 minimization steps were performed, i.e., (i) 10,000 on the all-atom fixed protein to relax the water, (ii) 10,000 with fixed Cα atoms to allow the relaxation of sidechains, and (iii) 10,000 without any constraints on the system. For (2), the positional constraints of various forces were applied and subsequently decreased at each minimisation/equilibration step to allow a smooth equilibration. The values were for: all protein atoms—10, 10, 2.5, 1, 0.5, and 0.1; the phosphate atom of POPC—2.5, 2.5, 1, 0.5, and 0.1 kcal/mol; the dihedral angle around the double bond of the oleoyl chain of POPC (restricted to 0°)

and dihedral angle formed by the glycerol carbon and the oleoyl ester oxygen atoms of POPC (restricted to 120°)—250, 250, 100, 50, and 25 kcal/mol. The system (2) was minimised during 5000 steps (2500 steps of steepest descent then 2500 steps of conjugated gradient).

For both systems, the following steps were executed. A 100 ps thermalisation step was performed, where the temperature (atoms velocity) was gradually increased from 0 to 310 K using the Berendsen thermostat[482]. Then, a 100 ps equilibration step with a constant volume and 100 ps equilibration step with constant pressure (1 bar) were performed. Periodic boundaries conditions and isotropic position scaling were imposed with the Berendsen barostat[482]. For these two steps, temperature regulation was performed with Langevin dynamics with friction coefficient $\gamma = 1$. Finally, a 100 ps molecular dynamics was completed at 310 K (Langevin dynamics), constant volume, and constant pressure with a hybrid Monte-Carlo barostat[483]. In (2), the membrane surface tension was set to 0 on the xy plane. Lastly, a mini (100 ps) molecular dynamics simulation of the previous conditions was completed, without any constraint on the system.

All equilibration steps and molecular dynamics simulation were carried out with an integration step of 2 fs. Non bonded interactions were calculated with the particle mesh Ewald summation (PME), with a cut-off of 10 Å, and bonds involving hydrogen atoms were constrained with SHAKE algorithm[484]. The initial velocities were reassigned, according to the Maxwell-Boltzmann distribution, and the same parameters (simulation conditions) as the mini dynamics were applied. Coordinates were recorded every 1 ps. The system was simulated with AMBER18[480] (http://ambermd.org/AmberMD.php) using the PMED Cuda module, running of the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/). For system (1), a unique short trajectory of 1 ns and, for system (2), three extended trajectories of 2 µs were performed.

A.1.1.3. Data analysis

All standard analyses were performed using the CPPTRAJ 4.25.6 program[485] of AmberTools20. Analysis of MD conformations (every 10 ps) was realized after least-square fitting on residues of the TK domain (W582-S688, L769-S931) or on residue-truncated trajectories of each fragment to remove rigid-body motions.

(1) The RMSD and RMSF values were calculated for the Cα-atoms, using the initial model (at t = 0 ns) as a reference. The RMSD, RMSF, and cross-correlations were calculated for the Cα-atoms on the initial conformation (t = 0 µs) as a reference;
(2) Secondary structural propensities for all residues were calculated using the define secondary structure of proteins (DSSP) method[262]. The secondary structure types were assigned for residues, based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 ps for the individual and

concatenated trajectories, respectively;

(3) Clustering analysis was performed on the productive simulation time of each MD trajectory, using an ensemble-based approach[267]. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots, according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cut-off r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD conformations are grouped into n reference clusters, based on their RMSDs from each reference structure. The cut-off was varied from 3 to 5 Å. The analysis was performed every 100 ps;

(4) The H-bonds between donor (D) and acceptor (A) atoms N, O, and S were monitored, according to the following geometrical parameters: $d_{(D-A)} \leq 3.6$ Å, $\widehat{DHA} \geq 120°$. Hydrophobic contacts were considered for all hydrophobic residues, with side chains within 4 Å of each other;

(5) The principal components analysis (PCA) modes were calculated for the backbone atoms (N, H, Cα, C, and O) after least-square fitting on the average conformation calculated on the concatenated data. The eigenvectors were visualized with NMWiz module for VMD[486];

(6) The normal modes and cross-correlation matrices of the average conformation of each replicate were calculated with the R library bio3D[476], at 310 K, with all available force fields;

(7) Curvature angles of selected secondary structures (helices or β-strand), relative to others or to their initial position (t = 0 μs), were calculated with Equation (1):

$$\Theta = \arccos \frac{v_1 \cdot v_2}{\sqrt{\|v_1\|} \cdot \sqrt{\|v_2\|}} \quad (1)$$

Where Θ is the angle in radian and $\mathbf{v}_1$ and $\mathbf{v}_2$ the secondary structures representative vectors coordinates. The vectors where delimited in the N- to C-terminal directions on the Cα atoms of the most structurally stable residues, according the DSSP: P524−C537 (TMD), T594−A597 (β_1 in P-loop), L637−G648 (αC helix), A701−N705 (αH1 in KID), S771−L783 (αE in C-lobe), and V824−K826 (β_9 in A-loop). The kink angle of the TMD helix (P524−C537) was calculated according to [292], after least-square fitting of the TMD Cα on their initial conformation. Using the previous formula, the kink angle corresponds to the angle between the vector from the mid-point of hinge (V530) in the N-terminal direction and vector from V530 in the C-terminal direction;

(8) The radius of gyration Rg was calculated from the atomic coordinates for all atoms but hydrogens.

(9) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates with Equation (3)[270]:

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (3)$$

Where k_B represents the Boltzmann constant, T is the temperature, $P_{(R_1,R_2)}$ denotes the probability density of states along the two reaction coordinates, calculated using their joint probability, and P_{max} denotes the maximum probability. The population of each well was roughly estimated using a square defined with R_1 and R_2 value intervals and containing red to orange ΔG colors.

A.1.2. Ledoux, J., Trouve, A., & Tchertanov, L. (2021). Folding and Intrinsic Disorder of the Receptor Tyrosine Kinase KIT Insert Domain Seen by Conventional Molecular Dynamics Simulations. *International Journal of Molecular Sciences*, *22*(14), 7375. https://doi.org/10.3390/ijms22147375

A.1.2.1. 3D modelling of cleaved KID

The 3D model of the full-length cytoplasmic domain of KIT was taken from [385]. The coordinates of kinase insert domain (KID, sequence F689–D768 aas) were extracted and used as a model of the cleaved KID.

A.1.2.2. Molecular dynamics simulation

Systems set-up. For MD simulation, models of KIT and KID (inactive unphosphorylated state) were prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBER)[480] using the ff99SB all-atom force field parameter set: (i) hydrogen atoms were added, (ii) covalent bond orders were assigned, (iii) protonation states of amino acids were assigned based on their solution for pK values at neutral pH, (iv) histidine residues were considered neutral and protonated on ε-nitrogen atoms, and (v) Na^+ counter-ions were added to neutralise the protein charge.

Each protein was solvated with explicit TIP3P water molecules in a periodic rectangular box with at least 12 Å distance between the proteins and the boundary of the water box. The total number of atoms in the systems (protein, water molecules and counter ion) was 69,089 and 19,537 for the KIT and KID, respectively.

The setup of the systems was performed with the Simulated Annealing with NMR-Derived Energy Restraints (SANDER) module of AMBER. First, each system was minimised successively using the steepest descent and conjugate gradient algorithms as follows: (i) 10,000 minimisation steps where the water molecules have fixed protein atoms, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein side chains to relax, and (iii) 10,000 minimisation steps without any constraint

on the system. After relaxation, each system was gradually heated from 10 to 310 K at constant volume using the Berendsen thermostat[482] while restraining the solute Cα atoms by 10 kcal/mol/Å^2. Thereafter, the system was equilibrated for 100 ps at constant volume (NVT) and for a further 100 ps at constant pressure (NPT) maintained by the Monte Carlo method [483]. Final system equilibration was achieved by a 100 ps NPT run to assure that the water box of the simulated system had reached the appropriate density. The electrostatic interactions were calculated using the PME (particle mesh Ewald summation)[487] with a cut-off of 10.0 Å for long-range interactions. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution.

Minimisation, equilibration and data generation. All trajectories were produced using the AMBER ff99SB force field with the PMEMD module of AMBER 16 and AMBER 18[480] (GPU-accelerated versions) running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2× CPU Intel Xeon E5-2680 and Nvidia GTX 780ti) and the supercomputer JEAN ZAY at IDRIS.

The multiple extended trajectories were generated for each equilibrated system: two 2-μs trajectories for KIT with KID, four 1.8-μs replicas for cleaved KID (KID^C) and two1.8-μs replicas for cleaved KID with the restrained distance (10 Å) between the Cα-atoms of terminal residues, F689 and D768 (KID^{CR}).

A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog method. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step. The van der Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution. Coordinates were recorded every 1 ps.

A.1.2.3. Data analysis

Unless otherwise stated, all recorded MD trajectories, individual and merged, were analysed (RMSFs, RMSDs, DSSP, clustering) with the standard routines of the CPPTRAJ 4.15.0 program[485] of AMBER 20 Suite. All analysis was performed on the MD conformations (every 10 ps) by considering either all simulations or the production part of the simulation, which was generated after the removal of non well-equilibrated conformations (0–70 ns), as was shown by the RMSDs, or on residues with a fluctuation of less than 4 Å, as shown by the RMSFs, and after least-square fitting of the MD conformations for a region of interest, thus removing rigid-body motion from the analysis.

- ***Conventional analysis***

(1) The RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference;

(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method[262]. The secondary structure types were assigned for residues based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 and 20 ps for the individual and concatenated trajectories, respectively;
(3) Clustering analysis was performed on the productive simulation time of each MD trajectory using an ensemble-based approach[267]. The analysis was performed every 100 ps. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cut-off r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD con-formations are grouped into n reference clusters based on their RMSDs from each reference structure. The cut-off was varied from 3 to 5 Å;
(4) H-bonds between heavy atoms (N, O, and S) as potential donors/acceptors were calculated with the following geometric criteria: donor/acceptor distance cut-off was set to 3.6 Å, and the bond angle cut-off was set to 120°. Hydrophobic contacts were considered for all hydrophobic residues with side chains within a 4 Å of each other;
(5) Contact search was performed with Cpptraj. Contact present if Cα–Cα distance < 10 Å. Map normalised over the number of frames such as the value represents the contact frequency in the considered trajectory;
(6) The mass-weighted radius of gyration (Rg) was calculated for all atoms but hydrogens.
(7) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates with Equation (1)[270]:

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (1)$$

Where k_B represents the Boltzmann constant,T is the temperature. $P_{(R1, R2)}$ denotes the probability of states along the two reaction coordinates, which is calculated using a k-nearest neighbour scheme and P_{max} denotes the maximum probability. The 3-dimensional representations of the free energy surface were plotted using Matlab (US, © 1994-2021 The MathWorks, Inc.).

- ***Advanced data analysis***

Secondary structure for each KID residue of each replica (DSSP) was classified by an 8-letter code[488] and used for calculation of an estimated transition probability matrix from one folding state to another. A suitable distance (Fisher-Rao distance) between these families was calculated for all the pairs of replicas. Multi-Dimensional

Scaling (MDS) was performed to get an 'as isometric as possible' embedding of the data in 2D (i.e., a representation by placing points on a plane while preserving the calculated inter-distances as well as possible).

A.1.3. Ledoux, J., & Tchertanov, L. (2022). Does Generic Cyclic Kinase Insert Domain of Receptor Tyrosine Kinase KIT Clone Its Native Homologue? *International Journal of Molecular Sciences*, *23*(21), 12898. https://doi.org/10.3390/ijms232112898

A.1.3.1. 3D modelling of cyclised KID

The initial 3D model of KID (sequence F689–D768) was taken at 2 µs of restrained cleaved KID molecular dynamics (MD) simulation reported in [271]. Five thousand KID models completed with four glycine residues in the C-terminal were generated with Modeller 10.1[234]. Only the GGGG motif loop was refined. The best model was chosen according to the DOPE score[474] and stereochemical quality (Procheck, Cambridge, UK)[475]. An additional five thousand models of cyclised KID^{GC} were generated from the previous loop refined KID^{GC} model using the LINK patch and assessed with the DOPE score and Procheck.

A.1.3.2. Molecular dynamics simulations

System set-up For MD simulation, the model of KID^{GC} was prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBER)[480] using the ff14SB all-atom force field parameter set: (i) hydrogen atoms were added, (ii) covalent bond orders were assigned, (iii) protonation states of amino acids were assigned based on their solution for pK values at neutral pH, (iv) histidine residues were considered neutral and protonated on ε-nitrogen atoms, and (v) Na+ counter-ions were added to neutralise the protein charge.

KID^{GC} was solvated with explicit TIP3P water molecules in a periodic octahedron box with at least a 12 Å distance between the protein and the boundary of the water box. The number of atoms in the system was 20,934, 1277 and 7 for water, protein, and counter ions, respectively.

The setup of the systems was performed with the Simulated Annealing with NMR-Derived Energy Restraints (SANDER) module of AMBER. First, the KID^{GC} system was minimised successively using the steepest descent and conjugate gradient algorithms as follows: (i) 10,000 minimisation steps with all protein atoms fixed to relax water molecules and counter ions, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein side chains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. After relaxation, the KID^{GC} system was gradually heated from 10 to 310 K at constant volume using the Berendsen thermostat[482] while

restraining the solute Cα-atoms by 10 kcal/mol/Å^2. After that, the system was equilibrated for 100 ps at constant volume (NVT) and a further 100 ps at constant pressure (NPT) maintained by the Monte Carlo method[483]. A 100 ps NPT run achieved final system equilibration to assure that the water box of the simulated system had reached the appropriate density. The electrostatic interactions were calculated using the PME (particle mesh Ewald summation)[487] with a cut-off of 10.0 Å for long-range interactions. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution.

Minimisation, equilibration and data generation. All MD trajectories were produced using the AMBER ff14SB force field with the PMEMD module of AMBER 18[480] (GPU-accelerated versions) running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2× CPU Intel Xeon E5-2680 and Nvidia GTX 780ti, Canonical Ltd., London, UK) and the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/).

The 2 μs trajectories (replicas) were generated for KID^{GC} equilibrated system. A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog method[489]. Coordinate files were recorded every 1 ps. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step[490]. The van der Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution. Coordinates were recorded every 1 ps.

A.1.3.3. Data analysis

Unless otherwise stated, all recorded MD trajectories, individual and merged, were analysed with the standard routines of the CPPTRAJ 4.15.0 program[485] of AMBER 20 Suite and Python library Scikit-Learn[491]. All data analysis was performed on the MD conformations (every 10 ps) by considering all concatenated data or the production part of the simulation after least-squares fitting of the MD conformations for a region of interest, thus removing rigid-body motion from the analysis.

- ***Conventional analysis***

(1) The RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference;
(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method[262]. The secondary structure types were assigned for residues based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 ps for the individual and concatenated trajectories;
(3) H-bonds between heavy atoms (N, O, and S) as potential donors or acceptors were calculated with the following geometric criteria: donor–acceptor distance cut-off was set to 3.6 Å, and the bond angle at H-atom cut-off to 120°; van der Waals contacts were considered for residues with side-chains C atoms within a 3.6 Å of

each other;
(4) The mass-weighted radius of gyration (Rg) and Solvent Accessible Surface Area (SASA) were calculated for all atoms except hydrogens;
(5) The tyrosine dihedral angles (ϕ, ψ) distributions (Ramachandran plots[235]) were compared to the unphosphorylated tyrosine backbone-dependent allowed area from the Dunbrack database[384];
(6) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates (Equation (1))[270] :

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (1)$$

Where k_B represents the Boltzmann constant, T is the temperature. $P_{(R1, R2)}$ denotes the probability of states along the two reaction coordinates, calculated using a k-nearest neighbour scheme, and P_{max}—the maximum probability.

- ***Clustering***

Thirty-one non reference-dependent (intrinsic) features were calculated for the individually concatenated data of KID^C, KID^D and KID^{GC} every 100 ps to create a starting dataset for clustering.

Pre-Processing and features selection. First, the data were scaled between 0 and 1 for each KID entity concatenated data and each feature.

To reduce the data dimensionality, a two-step process was applied for features selection: (i) for each pair of correlated features (coefficient of correlation should be ≥ 0.8 or ≤ −0.8), one feature was removed from the dataset, then (ii) the first k PC of a PCA explaining up to 80% of the variance were kept for further analysis.

Clustering. First, the hyperparameters of each algorithm were tweaked to maximise the clustering performances. (i) The Density-Based Spectral Clustering (DBSCAN)[393] for a minimum sample size of 5% of the data and ε distance between 0.1 and 1 each 0.05; (ii) K-means for *k* between 2 and 15 with 1000 iterations each; (iii) Hierarchical Agglomerative clustering using Ward distance metrics and cutting the tree at *k* between 2 and 15. The best method and associated hyperparameters are then finally applied to the dataset.

The clustering performances (parameters tweaking and final clustering) were assessed with the Silhouette[269] and Calinski–Harabasz scores[386].

A.1.4. Ledoux, J., & Tchertanov, L. (2023). Site-Specific Phosphorylation of RTK KIT Kinase Insert Domain: Interactome Landscape

A.1.4.1. 3D modelling

P-KID models

The initial 3D model of KID (sequence F699-D768, Uniprot P10721) was taken at 2 μs of MD simulation of the inactive KIT model reported in [271]. Tyrosine residues were phosphorylated by adding the phosphate group -O-PO_3^{2-} with PyMOL Builder module[492].

The active KIT$^{p568/p570}$/2Mg^{2+}/ATP

The initial 3D model of the activated KIT (sequence I516-R946, Uniprot P10721) was constructed using the crystallographic structure 1PKG (resolution 2.9 Å)[122] after 2 Mg^{2+}/ ATP docking with Autodock Vina[493] into the kinase domain, and model of the inactive KIT completed with transmembrane domain and C-terminal tail as reported in [271].

The PI3K regulatory subunit p85α N-terminal SH2 (N-SH2) domain

The template structure of p85α (N-)SH2 domain complexed with a KID^{p721} peptide[116] was retrieved from the PDB (PDB: 2IUH, resolution 2.0 Å) and used to model the free-ligand (N-)SH2 domain (sequence E332-S429, Uniprot P27986), further referenced as SH2 for a simplicity.

Five thousand models of (i) p-KID – KID^{pY703}, KID^{pY721}, KID^{pY730}, KID$^{pY703/pY721}$, KID$^{pY703/pY730}$, KID$^{pY721/pY730}$, KID$^{pY703/pY721/pY730}$; (ii) active KIT$^{p568/p570}$/2Mg^{2+}/ATP; (iii) SH2 were generated with Modeller 10.1[234]. The best models were assessed using the DOPE score[474] and Procheck[475].

KID^{pY721}/SH2 complex

Molecular complex of KID^{pY721} (model) with SH2 (model, cluster C1 representative conformation) was modelled using the structure of SH2 domain complexed with a KID^{pY721} peptide[417] as a reference for the initial KID positioning with respect to SH2 domain. A conformation of KID^{pY721} having an excellent similarity of its TNEYMDMK fragment with p-pep from the structure 2IUH was chosen as the initial structure of KID ^{pY721}. KID ^{pY721} was placed in two orientations of its TNEYMDMK fragment with respect to SH2, with p-pep N-terminal at αH1- and αH2 helix, respectively. In both cases, KID^{pY721} was positioned in front of SH2 binding pocket. The initial distance between p-atom from pY721 of KID and CZ and OG atoms of R340, R358 and S361 of SH2 in each built complex was at least 10 Å.

A.1.4.2. Molecular dynamic simulation

Systems set-up. The systems were prepared with the LEAP module of AMBER 20 (http://ambermd.org/AmberTools.php; accessed on 17 June 2021), using the ff19SB (phosaa19SB and phosaa19SB) all-atom force fields[480,494] for phosphorylated KID and active KIT. The latter was inserted in a phosphatidylcholine (POPC) lipid bilayer using Charm-Gui membrane and prepared with the additional lipid17 and ATP force fields[495]. Then, (i) hydrogen (H) atoms were added, (ii) covalent bond orders were assigned, (iii) protonation states of amino acids were assigned based on their solution for pKa values at neutral pH, (iv) histidine residues were considered neutral and protonated on ε-nitrogen atoms, and (v) Na+ counter-ions were added to neutralise the protein charge. All systems studied were solvated with explicit TIP3P water molecules in a periodic octahedron box with at least a 12 Å distance between the protein and the boundary of the water box.

Minimisation, equilibration and data generation. The setup of the systems was performed with the SANDER module of AMBER 20. First, p-KID and free-ligand SH2 were minimised using the steepest descent and conjugate gradient algorithms as follows: (i) 10,000 minimisation steps where the water molecules have fixed protein atoms, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein side chains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. After relaxation, the systems were gradually heated from 10 to 310 K at constant volume using the Berendsen thermostat[482] while restraining the solute Cα-atoms by 10 kcal/mol/Å^2. After that, the systems were equilibrated for 100 ps at constant volume (NVT), and a further 100 ps at constant pressure (NPT) maintained by the Monte Carlo method[483]. A 100-ps NPT run achieved final system equilibration to assure that the water box of the simulated system had reached the appropriate density. Active KIT was equilibrated using the same protocol as in [271].

All MD trajectories were produced with the PMEMD module of AMBER 20 (GPU-accelerated versions) and the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/).

Classical MD simulations. Three trajectory replicas of (i) 2 μs for each p-KID equilibrated system and active KIT and (ii) 500 ns for SH2 were generated with an integration time step of 2 fs. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step and bonds involving hydrogen atoms were constrained with the SHAKE algorithm[484]. The van der Waals interactions were modelled using a 6-12 Lennard-Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution. Coordinates were recorded every 1 ps.

In all p-KID equilibration and production steps, the N- and C-terminal Cα-atoms were restrained to a 10 ± 0.2 Å interatomic distance with a weight 20 kcal/mol to mimic

their native measure typical for a full-length KIT kinase domain[385].

GaMD simulations. To estimate the parameters needed for the Gaussian accelerated Molecular Dynamics (GaMD) simulation[260,261], 50-ns of cMD trajectories (one for each model CM1 and CM2) were used with the following distance constraints: 10 ± 0.2 Å with a weight 20 kcal/mol. Then, the 500-ns GaMD trajectories of the relaxed systems were generated, using as starting conformations the cMD conformations of the respective forms taken at t = 50 ns. Every 50 ns, the interatomic distance constraints between P of KID^{pY721}, CZ and OG of R340, R358 and S361 of SH2 were reduced by 1 Å (with the same standard deviation) from 10 to 4 Å. The boosting was applied of both total and dihedral potential energies. The boosting energy threshold was set as the maximal total potential energy calculated during the cMD. The coordinates were recorded every 1 ps.

A.1.4.3. Data analysis

All standard analyses of all protein trajectories were performed using the CPPTRAJ 4.25.6 program[485] of AmberTools20. Analysis of the protein MD conformations (every 10 ps) was realised after least-square fitting on a reference structure to remove rigid-body motions.

(1) The RMSD and RMSF values and cross-correlations were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference;
(2) Secondary structural propensities for all residues were calculated using the define secondary structure of proteins (DSSP) method[262];
(3) SH2 clustering analysis was performed on the productive simulation time of each MD trajectory using an ensemble-based approach[267]. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cut-off r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD conformations are grouped into n reference clusters, based on their RMSDs from each reference structure. The tested cut-offs were 0.5, 0.75, 1, 1.5, and 2 Å. The analysis was performed every 10 ps after SH2 fitting on the β-core (G353-R358, Y7368-R373, N378-F384) Cα-atoms at t = 0 ns;
(4) The mass-weighted radius of gyration (Rg) was calculated for all atoms but hydrogens.
(5) The H-bonds between the donor (D) and acceptor (A) atoms N, O, and S, satisfying the following geometrical parameters: $d_{(D-A)} \leq 3.6$ Å, DHA angle ≥ 120°, were monitored. Van der Waals contacts were considered for all residues, with the side chains carbon atoms within 3.6 Å of each other;
(6) The principal components analysis (PCA) modes were calculated for all non

hydrogen atoms after least-square fitting on the average conformation calculated on the concatenated data. The eigenvectors were visualised with the NMWiz module for VMD[486];

(7) SH2 dynamic correlations were calculated after fitting of MD conformations on its β-core structure (G353-R358, Y7368-R373, N378-F384) Cα-atoms at t = 0 ns.

(8) Binding pockets were found and analysed using the Fpocket program[420];

(9) Molecular docking of KID^{pY721} into SH2 was performed with HADDOCK 2.2 webserver[248] using (i) active residues R340, N344, R358, S361, N377-L380, and N417 for SH2, and (ii) T718-K725 for KID^{pY721}. Passive residues were assigned within 4 Å of the active residues. One thousand complexes were generated, and the 200 best docking solutions clustered using the FCC method (fraction of common contacts) and a cut-off of 0.6 Å. Clusters with a minimum size of 4 were kept for the refining step.

A.1.5. LEDOUX, J., BOTNARI, M., & TCHERTANOV, L. (2023). RECEPTOR TYROSINE KINASE KIT: MUTATION-INDUCED CONFORMATIONAL SHIFT PROMOTES ALTERNATING ALLOSTERIC POCKETS. – EN PRÉPARATION

A.1.5.1. 3D modelling

The coordinates of full-length KIT^{WT} (sequence I515-R946) in the inactive state (conformation at 2 µs from replicate 3) was taken from [271]. To model KIT mutant KIT^{D816V}, D816 was mutated to V with the Builder module of PyMOL (http://www.pymol.org/pymol).

A.1.5.2. Molecular dynamics simulations

Orientation in membrane. KIT^{D816V} transmembrane helix was positioned according to the orientation of KIT^{WT} conformation taken for its modelling.

System set-up. KIT^{D816V}, wrapped in a 23 Å-width leaflet lipid bilayer of POPC (2), were solvated with TIP3P water model in a rectangular box, with the PACKMOL-Memgen[479] and LEaP modules of AmberTools20 (http://ambermd.org/AmberTools.php), using the ff14SB all-atom force field[481] for protein and Lipid17 for membrane: (i) hydrogen atoms were added; (ii) protonation states of amino-acids at physiological pH were assigned, as well as the histidine residues protonated on their ε-nitrogen atoms; (iii) no counter-ions were added, as the system is already of neutral charge.

Minimisation, equilibration and data generation. The systems were equilibrated using the Sander module of AmberTools20. For system (1), 30,000 minimization steps were performed, i.e., (i) 10,000 on the all-atom fixed protein to relax the water, (ii) 10,000 with fixed Cα atoms to allow the relaxation of sidechains, and (iii) 10,000 without any constraints on the system. For (2), the positional constraints of various

forces were applied and subsequently decreased at each minimisation/equilibration step to allow a smooth equilibration. The values were for: all protein atoms—10, 10, 2.5, 1, 0.5, and 0.1; the phosphate atom of POPC—2.5, 2.5, 1, 0.5, and 0.1 kcal/mol; the dihedral angle around the double bond of the oleoyl chain of POPC (restricted to 0°) and dihedral angle formed by the glycerol carbon and the oleoyl ester oxygen atoms of POPC (restricted to 120°)—250, 250, 100, 50, and 25 kcal/mol. The system (2) was minimised during 5000 steps (2500 steps of steepest descent then 2500 steps of conjugated gradient).

Then, the following steps were executed. A 100 ps thermalisation step was performed, where the temperature (atoms velocity) was gradually increased from 0 to 310 K using the Berendsen thermostat[482]. Then, a 100 ps equilibration step with a constant volume and 100 ps equilibration step with constant pressure (1 bar) were performed. Periodic boundaries conditions and isotropic position scaling were imposed with the Berendsen barostat[482]. For these two steps, temperature regulation was performed with Langevin dynamics with friction coefficient $\gamma = 1$. Finally, a 100 ps molecular dynamics was completed at 310 K (Langevin dynamics), constant volume, and constant pressure with a hybrid Monte-Carlo barostat[483]. In (2), the membrane surface tension was set to 0 on the xy plane. Lastly, a mini (100 ps) molecular dynamics simulation of the previous conditions was completed, without any constraint on the system.

All equilibration steps and molecular dynamics simulations were carried out with an integration step of 2 fs. Non bonded interactions were calculated with the particle mesh Ewald summation (PME), with a cut-off of 10 Å, and bonds involving hydrogen atoms were constrained with SHAKE algorithm[484]. The initial velocities were reassigned, according to the Maxwell-Boltzmann distribution, and the same parameters (simulation conditions) as the mini dynamics were applied. Coordinates were recorded every 1 ps. The system was simulated with AMBER18 (http://ambermd.org/AmberMD.php) using the PMED Cuda module, running of the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/). For system (1), a unique short trajectory of 1 ns and, for system (2), three extended trajectories of 2 µs were performed.

A.1.5.3. Data analysis

All standard analyses were performed using the CPPTRAJ 4.25.6 program[485] of AmberTools20. Analysis of MD conformations (every 10 ps) was realized after least-square fitting on residues of the TK domain (W582-S688, L769-S931), residue-truncated trajectories or after fitting on inactive KIT^{WT} to remove rigid-body motions.

(1) The RMSD and RMSF values were calculated for the Cα-atoms, using the initial model (at t = 0 ns) as a reference. The RMSD, RMSF, and cross-correlations were calculated for the Cα-atoms on the initial conformation (t = 0 µs) as a reference;

(2) Secondary structural propensities for all residues were calculated using the define secondary structure of proteins (DSSP) method[262]. The secondary structure types were assigned for residues, based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 ps for the individual and concatenated trajectories, respectively;

(3) Clustering analysis was performed on the productive simulation time of each MD trajectory, using an ensemble-based approach[267]. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots, according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cut-off r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD conformations are grouped into n reference clusters, based on their RMSDs from each reference structure. The cut-off was varied from 3 to 5 Å. The analysis was performed every 100 ps;

(4) The H-bonds between donor (D) and acceptor (A) atoms N, O, and S were monitored, according to the following geometrical parameters: $d_{(D-A)} \leq 3.6$ Å, $\widehat{DHA} \geq 120°$.

(5) The principal components analysis (PCA) modes were calculated for the backbone atoms (N, H, Cα, C, and O) after least-square fitting on the average conformation calculated on the concatenated data. The eigenvectors were visualized with NMWiz module for VMD[486];

(6) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates with Equation (1):

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (1)$$

Where k_B represents the Boltzmann constant, T is the temperature, $P_{(R_1,R_2)}$ denotes the probability density of states along the two reaction coordinates, calculated using their joint probability, and P_{max} denotes the maximum probability. The population of each well was roughly estimated using a square defined with R_1 and R_2 value intervals and containing red to orange ΔG colors.

(7) Pockets investigations were done with MDpocket with isovalue 0.5 on the concatenated data of each KIT species[473]. Well-defined grid points were exclusively extracted using VMD and PyMol. Conformational analysis was performed on nine conformations with a maximum P1 volume surpassing an arbitrary volume threshold, eight conformations with a maximum P2 volume surpassing an arbitrary volume threshold, and sixteen conformations characterized by minimal P1 and P2 pocket volumes.

A.2. METHODES UTILISEES DANS LA THEMATIQUE HVKORC1

A.2.1. STOLYARCHUK, M.+, LEDOUX, J.+, MAIGNANT, E., TROUVE, A., & TCHERTANOV, L. (2021). IDENTIFICATION OF THE PRIMARY FACTORS DETERMINING THE SPECIFICITY OF HUMAN VKORC1 RECOGNITION BY THIOREDOXIN-FOLD PROTEINS. *INTERNATIONAL JOURNAL OF MOLECULAR SCIENCES*, *22*(2), 802. HTTPS://DOI.ORG/10.3390/IJMS22020802

A.2.1.1. 3D modelling

Trx-fold proteins

Structures of PDI (PDB ID: 4ekz), ERp18 (PDB ID: 1sen) and TMX1 (PDB ID: 1X5e) were retrieved from the PDB database and atomic coordinates of domain a, which contains the CX_1X_2C motif and is present in all available structures, were extracted. The 3D homology model of hTMX4 was generated from human sequence Q9H1E5 (https://www.uniprot.org/uniprot/) using the Modeller program[234] and the empirical structure of TMX1 (PDP ID: 1X5e) that was used as a template. The 3D model of the ERp18 protein was optimised (the cysteine residues were saturated with hydrogen atoms) to obtain a reduced state of the CX_1X_2C motif.

hVKORC1

The coordinates of full-length hVKORC1 (sequence M1–H163) in the inactive state was taken from [220].

hVKORC1/Trx complex

Each complex of the PDI protein with hVKORC1 (PDI–hVKORC1) was modelled using the structure of bacterial VKOR (bVKOR; PDB ID: 4nv5) as a reference for the initial PDI positioning with respect to hVKORC1. The structures of the human Trx-fold protein and hVKORC1 were carefully superimposed with the respective domains of bVKOR. To eliminate a small intersection between part of the L-loop of hVKORC1 and PDI, the extended conformation of the L-loop was chosen. The PDI protein was placed in two orientations with respect to hVKORC1, with (i) L3 and the αH2-helix (F1) and (ii) L5 and the αH2-helix (F2) positioned in front of the predicted "binding fragment" of the L-loop from hVKORC1. The initial distance S···S between the sulphur atoms from C37 of PDI and C43 of hVKORC1 in each built complex was 16 Å.

The stereochemical quality of all 3D models was assessed by Procheck[475], which revealed that more than 95% of nonglycine/nonproline residues have dihedral angles in the most favoured and permitted regions of the Ramachandran plot, as is expected for good models.

A.2.1.2. Molecular dynamics simulations

Systems set-up. For MD simulations, all models of the isolated proteins (PDI, ERp18, Tmx1, Tmx4), hVKORC1, and the two models of the PDI–hVKORC1 complex in two orientations (PDI_{F1}–VKORC1 and PDI_{F2}–VKORC1) were prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBER)[480] using the ff14SB all-atom force field parameter set[481]: (i) hydrogen atoms were added; (ii) covalent bond orders were assigned; (iii) protonation states of amino acids were assigned based on their solution for pK values at neutral pH and histidine residues were considered neutral and were protonated for ε-nitrogen atoms; (vi) the Na^+ counter-ion was added to neutralise the protein charge.

Each membrane protein, hVKORC1 and the two models of complex PDI–VKORC1 (PDI_{F1}–VKORC1 and PDI_{F2}–VKORC1) were embedded in the equilibrated and hydrated membrane composed of 200 1,2-dilauroyl-sn-glycero-3-phosphocholine (DLPC) lipids using the replacement method available in the CHARMM-GUI membrane builder (http://www.charmm-gui.org/input/membrane)[477]. This lipid bilayer was completed with 17293 (hVKORC1), 22047 (PDI_{F1}–VKORC1) and 22567 (PDI_{F2}–VKORC1) water molecules (TIP3P), pre-equilibrated during 1.5 ns of MD using the Lipid14 tool[496] from the AMBER package.

Each protein or protein complex inserted into a membrane was solvated with explicit TIP3P water molecules in a periodic rectangular box with a distance of at least 12 Å between the proteins and the boundary of the water box. Cl^- ions were randomly placed to neutralise the system.

The total number of atoms in the isolated Trx-fold proteins (protein, water molecules and counter ion) varied from 16,065–26,386. The total number of atoms in the membrane systems (hVKORC1 and its complexes with PDI, including proteins, DLPC lipids, water molecules and counter-ions, was 72683 (hVKORC1), 92570 (PDI_{F1}–VKORC1), and 93325 (PDI_{F2}–VKORC1)). The box size varied in the range of 84 × 84 × 108–141 $Å^3$.

The set-up of the systems was performed with the Simulated Annealing with NMR-Derived Energy Restraints (SANDER module of AMBER18[480]). First, each system was minimised successively using the steepest descent and conjugate gradient algorithms, as follows: (i) 10,000 minimisation steps where water molecules have fixed, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein side-chains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. The equilibration was performed on the solvent, keeping the solute atoms (except H-atoms) restrained for 100 ps at 310 K and a constant volume (NVT). Protein, membrane and solvent (water and ions) temperatures were separately coupled to the velocity rescale thermostat, which was a modified Berendsen thermostat[497] with a relaxation time of 0.1 ps. Each system was equilibrated for 1 ns (NPT), with all nonhydrogen atoms

of the protein and the DLPC membrane harmonically restrained. Semi-isotropic coordinate scaling and Parrinello–Rahman pressure coupling were used to maintain the pressure at 1 bar, with a relaxation time of 5 ps. The Nose–Hoover thermostat[498]was applied to the protein, lipids and solvent (water and ions) separately, with a relaxation time of 0.5 ps to keep the temperature constant at 310 K. Water and ions were allowed to move freely during equilibration.

Minimisation, equilibration and data generation. All trajectories were performed using the AMBER ff14SB force field with the PMEMD module of AMBER 16 and AMBER 18 (GPU-accelerated versions) running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2x CPU Intel Xeon E5-2680 and Nvidia GTX 780ti) and the supercomputer JEAN ZAY at IDRIS.

The 500-ns MD trajectories of each fully relaxed isolated protein were generated (2 replicas for Trx-fold proteins and 3 replicas for hVRORC1) in its natural environment—the water solution for the Trx-fold protein and the solvated bilayer lipid membrane for h-VKORC1. Each PDI–hVKORC1 complex that was inserted into the solvated bilayer lipid membrane was simulated for an alternating value of distance S···S from PDI and hVKORC1 (see the next subsection for details). MD simulation of the Trx–VKORC1 complex was first performed for 38 ns, with a constrained S···S distance of 12.8 Å, which was further reduced to 10.2 Å and followed by simulation for 20 ns, and finally to 8.2 Å, followed by the last 20-ns of the simulation.

A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog algorithm[489]. The Particle Mesh Ewald (PME) method[490], with a cutoff of 9.0 Å, was used to treat long-range electrostatic interactions at every time step. The van der Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution.

Stepped MD simulations of the PDI–hVKORC1 Complex. In Model 1 and Model 2, to prevent the separation of the PDI protein from hVKORC1 and to bring them together, a restrained harmonic distance was introduced to the S···S atom pair (the sulfur atoms from C37 of PDI and C43 of hVKORC1), which was varied in a stepwise manner (**Figure S40**). Specifically, the 80-ns simulation was divided into three steps, each with different applied restraints (d): from 0 to 38 ns with d equal to 12.8–11.8 Å (Step A), from 38 to 60 ns with d equal to 10.2–9.6 Å (Step B), and 60 to 80 ns with d equal to 9.2–8.2 Å (Step C). To probe the stability of the PDI–hVKORC1 complex, the simulations of Model 1 and Model 2) were continued from 80 to 100 ns with two different "soft" restraints applied to distance S···S. While the lower limit value remained at 8.2 Å, as in the previous simulation steps (A–C), the upper limit in Step D was increased to 10.2 Å (as in the 60–80 ns step) and 12.8 Å (as in the 0–38 ns step).

A.2.1.3 Data Analysis

Unless otherwise stated, all recorded MD trajectories were analysed (RMSFs,

RMSDs, DSSP, clustering) with the standard routines of the CPPTRAJ 4.15.0 program[485] of AMBER 18 Suite.

- ***Conventional analysis***

(1) The RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference. All analysis was performed on the MD conformations (every 10 ps) by considering either all simulations or the production part of the simulation, which was generated after the removal of non well-equilibrated conformations (0–70 ns), as was shown by the RMSDs, or on residues with a fluctuation of less than 4 Å, as shown by the RMSFs. For hVKORC1, the RMSDs were individually calculated for each domain after least-square fittings of the MD conformations to the initial conformation of a domain, thus removing rigid-body motion from the analysis;
(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method[262]. The secondary structure types were assigned for residues based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 and 20 ps for the individual and concatenated trajectories, respectively;
(3) The dynamic cross-correlation (DCC) between all atoms within a molecule quantifies the correlation coefficients of motions between atoms, *i.e.,* the degree to which the atoms move together[499]. Calculations were performed on backbone Cα-atoms on the productive simulation time of each MD trajectory using an ensemble-based approach[267]. The correlation values vary between −1 and 1, where 1 illustrates a complete correlation, −1 a complete anti-correlation and 0 no correlation;
(4) The collective motions of proteins were investigated by principal component analysis (PCA). For an *N*-atom system, a trajectory matrix contains, in each column, Cartesian coordinates for a given atom at each time step x(t) fitting the coordinate data to a reference structure;
(5) The extent to which the fluctuations of a system are correlated depends on the magnitude of the cross-correlation coefficient[265];
(6) Clustering analysis was performed on the productive simulation time of each MD trajectory using an ensemble-based approach[267]. The first 70 ns were omitted from the analysis of Trx-fold proteins. The analysis was performed every 100 ps. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cutoff r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD conformations are grouped into *n* reference clusters based on their RMSDs from each reference structure. The cut-off was set to 2 Å for both clustered proteins or domains (Trx-fold and L-loop) to allow the

comparison;

(7) Drift analysis of helices was performed on the L-loop from h-VKORC1 using the centroids (C_i) defined for the main-chain atoms for amino acids (aas) at the top and bottom of each helix. Positions of these centroids were monitored over the MD simulations, and their coordinates were projected on the *x*–*z* and *y*–*z* planes. The geometry of the CX_1X_2C motif from the Trx-fold proteins was described by the distance S···S between two sulphur atoms from cysteine residues C37 and C40 and the dihedral angle determined as an absolute value of pseudotorsion angle S−Cα37− C40α−S;

(8) H-bonds between heavy atoms (N, O, and S) as potential donors/acceptors were calculated with the following geometric criteria: donor/acceptor distance cut-off was set to 3.6 Å, and the bond angle cut-off was set to 120°. Hydrophobic contacts were considered for all hydrophobic residues with side chains within a distance of 4 Å of each other.

- ***Advanced analysis***

(1) Metric multidimensional scaling (MDS) is an algorithm for dimension reduction and visualization; it computes an embedding of a set of points (a shape trajectory in our case) in a lower dimension space with respect to the pairwise distances (Kendall's ones in our case) in the original set[431];

(2) The algorithm consists of a minimisation of the cost (see Equation (1)):

$$\sum_{i \neq j} \left(d_{ij} - \|x_i - x_j\|\right)^2 \quad (1)$$

Where $D=\left(d_{ij}\right)$ is the pairwise distance matrix, and $\{x_i\}_{\mathbf{i}}$ are the embedded points. It can be implemented using the manifold.MDS class in Python's scikit learn library;

(3) The Fréchet mean of a set is a point minimising the sum of squared distances to each point of the set. As an example, the Fréchet mean $\overline{T}$ of one set $\{T_i\}_i$ of tetrahedrons is defined as in Equation (2):

$$\overline{T} \in \underset{T}{\operatorname{argmin}} \sum_i d(T,T_i)^2 \quad (2)$$

Where the distance is the Euclidean distance, the Fréchet mean is no other than the classical mean we know;

(4) Kendall's shape space of 3D triangles is isometric to the northern hemisphere of a 3D sphere of radius ½, where the equilateral triangle is at the north pole[432]. We use a planar representation of the half-sphere as a disk with the equilateral triangle at the centre by the transformation $(\varphi,\theta)\rightarrow(r=\sin(\theta),\varphi)$ from the spherical coordinates to the polar coordinates. Each 3D triangle, up to translation rotation and scaling, is represented by a unique point of the disk.

A.2.2. Ledoux, J., Stolyarchuk, M., Bachelier, E., Trouvé, A., & Tchertanov, L. (2022). Human Vitamin K Epoxide Reductase as a Target of Its Redox Protein. *International Journal of Molecular Sciences*, *23*(7), 3899. https://doi.org/10.3390/ijms23073899

A.2.2.1. Available X-ray data and homology models

VKOR X-Ray data

The original atomic coordinates of two crystallographic forms of VKOR in the inactive state, the holo-c form, obtained by co-crystallisation of human protein with inhibitor (warfarin) and glycerol monooleate (PDB ID: 6wv3), and the apo-c form that was received for the ligand-free VKOR-like protein from *Takifugu rubripes* (PDB ID: 6wvi) were retrieved from the Protein Database (PDB). These data were used (i) for comparative analysis of protein structures (holo-c versus apo-c), (ii) as template for homology modelling of the human protein (apo-h), (iii) for optimisation of the holo-c form (repair the missing residues) to obtain the full-length protein (holo-h), and (iv) as the starting conformation for MD simulation of the holo-c.

Homology models

3D homology models of the full-length human VKOR (1-163 aas), apo-h and holo-h, were generated. The apo-h model was obtained with Modeller[234] from the human sequence Q9BQB6 (https://www.uniprot.org/uniprot/) and the empirical structure 6wvi (VKOR-like protein from *Takifugu rubripes*), used as a template. The similarity/identity of two sequences is 69/46 % for the total protein, and 77/65 % for the L-loop. The holo-h model was obtained from the holo-c form by adding the missing residues/atoms.

Relaxed homology models

The relaxed hVKOR models, relaxed apo-h and relaxed holo-h, in which H-bonds G62···Q78 and G60···N80 were removed using the translations of the neighbouring residues: R58-W59 (x, y + 4, z-2), G60-F63 (x, y + 3, z-2) and Q78-S79 (x, y – 2, z), were prepared. The coordinates of TM3 helix (W101-V127) were also translated (x, y + 2, z) to avoid steric clashes.

Models of the isolated L-loop

The coordinates of L-loop, slightly extended in sequence at its C-end (R33–N80), were extracted from the holo-h, apo-h and *de novo* models and used as the starting conformations of the cleaved isolated L-loop (holo-h L-loop, apo-h L-loop and *de novo* L-loop).

The stereochemical quality of 3D model was assessed by Procheck[475]; which revealed that more than 96% of nonglycine/nonproline residues have dihedral angles in the most favoured and permitted regions of the Ramachandran plot, as is expected for good models.

A.2.2.2. Molecular dynamics simulations

Systems set-up. Each system, structure of the holo-c form, the homology models apo-h and holo-h, their relaxed models, relaxed apo-h and relaxed holo-h, and models of the cleaved L-loop, was prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBERTools 20) (http://ambermd.org/AmberTools.php)[480] using the ff14SB all-atom force field parameter set[481] and TIP3P water models : (i) hydrogen atoms were added; (ii) covalent bond orders were assigned; (iii) protonation states of amino-acids were assigned based on their solution for pK values at neutral pH, and the histidine residues were protonated on their ε-nitrogen atoms; (vi) counter-ions, Cl^-, were added to neutralise the charge of each protein; (v) each protein was placed in an octahedron water box. Each final system contains 2595 atoms of VKOR and 45987/57222 atoms of water for the apo/holo forms respectively, and 741 atoms of L-loop and 19302/18876 atoms of water for the cleaved L-loop.

Minimisation, equilibration and data generation. Each system was minimised and equilibrated using the Sander module of AmberTools20 (http://ambermd.org/AmberTools.php) using the steepest descent and conjugate gradient algorithms through the 30,000 minimisation steps as follows: (i) 10,000 minimisation steps where water molecules have fixed, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein sidechains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. A 100 ps thermalisation step was performed, where the temperature (atoms velocity) is gradually increased from 0 to 310 K using the Berendsen thermostat with imposed periodic boundaries conditions and isotropic position scaling[482]. Then, a 100 ps equilibration with constant volume (NVT) and a 100 ps equilibration with constant pressure (1 bar) (NPT) were performed. For these two steps, temperature regulation was performed with Langevin dynamics with friction coefficient $\gamma = 1$. Finally, a 100 ps molecular dynamics was completed at 310 K (Langevin dynamics), constant volume and constant pressure (hybrid Monte-Carlo barostat[483]. All equilibration steps were carried out with an integration step of 2 fs. Non bonded interactions were calculated with the Particle-Mesh Ewald summation (PME) with a cut-off of 10 Å and bonds involving hydrogen atoms were constrained with SHAKE algorithm[484].

The conventional Molecular Dynamics (cMD) trajectories of the holo-c structure (0. 2 µs), the homology models apo-h and holo-h (0.5 µs), their relaxed models (relaxed holo-h and relaxed apo-h) (0.5 µs), and models of the cleaved L-loop with restrains (5 µs) and fully relaxed (0.5 µs), were generated using the AMBER ff14SB force field with the PMEMD module of AMBER 16 and AMBER 18 (GPU-accelerated versions)[480]

running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2x CPU Intel Xeon E5-2680 and Nvidia GTX 780ti) and the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/).

Classical MD simulations. The initial velocities were reassigned according to the Maxwell-Boltzmann distribution. A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog algorithm[489]. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step. The van-der-Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution. For the relaxed models, relaxed apo-h and relaxed holo-h, to avoid the H-bonds formation, the additional constraints were applied: the distances between the Cα-atoms from G62-Q78 and G60-N80 were maintained ≥ 9.9 Å to ensure that no H-bonds could form between the backbone of the glycine residues and their respective potential donor. During the first 100 ns, the constraints between G62-Q78 and G60-N80 were maintained, then removed for the next 400 ns to fully relax the systems. The cleaved L-loop was simulated for 0.5 µs as a fully unconstrained entity, and for 5 µs as a polypeptide with the 100 kcal/mol Cartesian positional restrains of the backbone atoms (peptide N-Cα-C-O) from terminal residues R33 and N80 (cleaved restrained L-loop), to preserve R33···N88 distance that was observed in crystallographic structures[218] and *de novo* model[220].

GaMD simulations. To estimate the parameters needed for the Gaussian accelerated Molecular Dynamics (GaMD) simulation[260,261], 50-ns of cMD trajectories of hVKORC1 were used. Then, the 500-ns GaMD trajectories of the relaxed holo-h and relaxed apo-h were generated using as starting conformations the cMD conformations of the respective forms taken at t=50 ns. The boosting was applied of both total and dihedral potential energies. The boosting energy threshold was set as the maximal total potential energy calculated during the cMD. The coordinates were recorded every 1 ps.

A.2.2.3. Protein-protein docking

Protein-protein docking was performed with the HADDOCK2.4 web server (https://wenmr.science.uu.nl/haddock2.4/). HADDOCK (High Ambiguity Driven protein-protein DOCKing)[248] is a protein−protein docking approach based on available biochemical or biophysical information to drive the docking process. The docking protocol consists of several steps with user-defined input parameters. First, the topologies and coordinates files are generated for each molecule separately, and merged to generate the starting models of the complex. Second, 10,000 structures were randomly sampled and subjected to the rigid body energy minimisation (it0). Third, the best 200 structures were selected and a semi-flexible simulated annealing in torsion angle space was performed on it (it1). And finally, the obtained structures after the previous step were refined in Cartesian space with explicit solvent (TIP3P) — a short

molecular dynamics stage. After, the water-refined structures were clustered using a 7.5 Å RMSD cut-off and sorted according to the HADDOCK score. The maximum number of clusters was set to 10 and the minimal cluster size was set to 4. All other input parameters were kept default. To guide the docking, a set of ambiguous interactions restraints (AIRs), a pair of cysteine residues, C43 of hVKOR and C37 of PDI, was provided. Docking simulations were run with the same conformation of PDI (ligand protein) and with a set of different conformations of hVKOR (target protein), except of benchmark trials for which each protein, PDI and hVKORC1, was considered as ligand and target.

A.2.2.4. Data analysis

Unless stated otherwise, the data analysis was performed using CPPTRAJ 4.25.6 program[485] of AmberTools20 (http://ambermd.org/AmberTools.php) for MD conformations taken every 10 ps of simulation after least-square fitting on the initial conformation (t = 0 ns) of a region of interest, thus removing rigid-body motion from the calculations.

- ***Conventional data analysis***

(1) RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference;
(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method[262]. The secondary structure types were assigned for residues based on backbone -NH and -CO atom positions. Secondary structures were assigned every 10 and 20 ps for the individual and concatenated trajectories, respectively;
(3) The trajectories of the cleaved L-loop of different forms were compared by means of best-fit RMSD values of all C-alpha atoms in a pairwise manner between two trajectories frame by frame. The resulting two-dimensional pairwise RMS2D matrix allows finding the pairs of conformations with the minimal values of RMSD (less than the selected threshold values of 4 and 5 Å);
(4) H-bonds between heavy atoms (N, O, and S) as potential donors/acceptors were calculated with the following geometric criteria: donor/acceptor distance cut-off was set to 3.6 Å, and the bond angle cut-off was set to 120°. Hydrophobic contacts were considered for all hydrophobic residues with side chains within a 4 Å of each other;
(5) The trajectories of the cleaved L-loop (holo-h, apo-h forms and *de novo* model) were compared by means of best-fit RMSD values of all C-alpha atoms in a pairwise manner between two trajectories frame by frame. The resulting two-dimensional pairwise RMSD matrix allows finding the pairs of conformations with the minimal values of RMSD (less than the selected threshold values of 4 and 5 Å);
(6) The mass-weighted radius of gyration (Rg) was calculated for all atoms but hydrogens;

The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates (Equation (1))[270]:

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (1)$$

Where k_B represents the Boltzmann constant, T is the temperature. $P_{(R1, R2)}$ denotes the probability of states along the two reaction coordinates, calculated using a k-nearest neighbour scheme, and P_{max}—the maximum probability.

The conformational landscape was reconstructed on the trajectories of the unconstrained (0.5 μs) and constrained (5 μs)—with the constraints applied on the N- and C- ends —simulations of the cleaved L-loop from holo-h, apo-h forms and *de novo* model, using the pairs of metrics – RMSD, Rg, helical content and principal components (PC1 and PC2) obtained by the Principal Component Analysis (PCA. The 3-dimensional representations of the free energy surface were plotted using Matlab (US, © 1994-2021 The MathWorks, Inc.).

- ***Advanced Data Analysis***

To group the conformations with similar secondary structure patterns of the concatenated trajectory combined with four 5-μs cMD simulations of the cleaved L-loop (holo-h, apo-h, relaxed-apo-h and *de novo* models), the algorithm of measuring the secondary structure similarity as described in [392] was used. The scoring matrix (secondary structure elements similarity matrix) supporting the 8-DSSP state alphabet was utilised. Then, the single-link hierarchical clustering method[396] with threshold of 0.65 was carried out using SciPy (http.//https://scipy.org) to finally group similar conformations.

A.2.3. Botnari, M., Ledoux, J., & Tchertanov, L. (2023). Synergy of Mutation-Induced Effects in Human Vitamin K Epoxide Reductase: Perspectives and Challenges for the Design of Allo-Network Modulators. – en préparation

A.2.3.1. 3D modelling

3D homology models of the full-length human VKORC1 (1-163 aas) mutants, $hVKORC1^{A41S}$, $hVKORC1^{H68Y}$, $hVKORC1^{S52W}$ and $hVKORC1^{W59R}$, Modeller[234] using the *de novo* model of hVKORC1 in a fully oxidised state[220] as template. The stereochemical quality of 3D model was assessed by Procheck[475], which revealed that more than 96% of nonglycine/nonproline residues have dihedral angles in the most favoured and permitted regions of the Ramachandran plot, as is expected for good models.

A.2.3.2. Molecular dynamics simulations

Systems set-up. Each system, structure of the holo-c form, the homology models apo-h and holo-h, their relaxed models, relaxed apo-h and relaxed holo-h, and models of the cleaved L-loop, was prepared with the LEAP module of Assisted Model Building with Energy Refinement (AMBERTools 20) (http://ambermd.org/AmberTools.php)[480] using the ff14SB all-atom force field parameter set [481] and TIP3P water models : (i) hydrogen atoms were added; (ii) covalent bond orders were assigned; (iii) protonation states of amino-acids were assigned based on their solution for pK values at neutral pH, and the histidine residues were protonated on their ε-nitrogen atoms; (vi) counter-ions, were added to neutralise the charge of each protein.

Minimisation, equilibration and data generation. Each system was minimised and equilibrated using the Sander module of AmberTools20 (http://ambermd.org/AmberTools.php) using the steepest descent and conjugate gradient algorithms through the 30,000 minimisation steps as follows: (i) 10,000 minimisation steps where water molecules have fixed, (ii) 10,000 minimisation steps where the protein backbone is fixed to allow protein sidechains to relax, and (iii) 10,000 minimisation steps without any constraint on the system. A 100 ps thermalisation step was performed, where the temperature (atoms velocity) is gradually increased from 0 to 310 K using the Berendsen thermostat with imposed periodic boundaries conditions and isotropic position scaling[482]. Then, a 100 ps equilibration with constant volume (NVT) and a 100 ps equilibration with constant pressure (1 bar) (NPT) were performed. For these two steps, temperature regulation was performed with Langevin dynamics with friction coefficient $\gamma = 1$. Finally, a 100 ps molecular dynamics was completed at 310 K (Langevin dynamics), constant volume and constant pressure (hybrid Monte-Carlo barostat[483]. All equilibration steps were carried out with an integration step of 2 fs. Non bonded interactions were calculated with the Particle-Mesh Ewald summation (PME) with a cut-off of 10 Å and bonds involving hydrogen atoms were constrained with SHAKE algorithm[484].

The molecular dynamics trajectories of hVKORC1 mutants (0.5 µs) were generated using the AMBER ff14SB force field with the PMEMD module of AMBER 18 and AMBER 18 (GPU-accelerated versions)[480] running on a local hybrid server (Ubuntu, LTS 14.04, 252 GB RAM, 2x CPU Intel Xeon E5-2680 and Nvidia GTX 780ti) and the supercomputer JEAN ZAY at IDRIS (http://www.idris.fr/jean-zay/).

The initial velocities were reassigned according to the Maxwell-Boltzmann distribution. A time step of 2 fs was used to integrate the equations of motion based on the Leap-Frog algorithm[489]. The Particle Mesh Ewald (PME) method, with a cut-off of 10 Å, was used to treat long-range electrostatic interactions at every time step. The van-der-Waals interactions were modelled using a 6–12 Lennard–Jones potential. The initial velocities were reassigned according to the Maxwell–Boltzmann distribution.

A.3.3.3. Data analysis

Unless stated otherwise, the data analysis was performed using CPPTRAJ 4.25.6 program[485] of AmberTools20 (http://ambermd.org/AmberTools.php) for MD conformations taken every 10 ps of simulation after least-square fitting on the initial conformation (t = 0 ns) of a region of interest, thus removing rigid-body motion from the calculations.

(1) RMSD and RMSF values were calculated for the Cα-atoms using the initial model (at t = 0 ns) as a reference;
(2) Secondary structural propensities for all residues were calculated using the Define Secondary Structure of Proteins (DSSP) method[262];
(3) H-bonds between heavy atoms (N, O, and S) as potential donors/acceptors were calculated with the following geometric criteria: (D···A ≤ 3.6 Å, pseudo-valent angle at H-atom ∠ D-H···A > 120°, where D (D = O/N/S) is a donor atom and A (A = O/N/S) is an acceptor atom). Hydrophobic contacts were considered for all hydrophobic residues with side chains within a 4 Å of each other;
(4) The mass-weighted radius of gyration (Rg) and Solvent Accessible Surface Area (SASA) were calculated for all atoms except hydrogens;
(5) Clustering analysis was performed on the productive simulation time of each MD trajectory using an ensemble-based approach[267]. The analysis was performed every 100 ps. The algorithm extracts representative MD conformations from a trajectory by clustering the recorded snapshots according to their Cα-atom RMSDs. The procedure for each trajectory can be described as follows: (i) a reference structure is randomly chosen in the MD conformational ensemble, and all conformations within an arbitrary cutoff r are removed from the ensemble; this step is repeated until no conformation remains in the ensemble, providing a set of reference structures at a distance of at least r; (ii) the MD conformations are grouped into *n* reference clusters based on their RMSDs from each reference structure. The optimal cut-off was set to 4 Å for both clustered proteins or domains to allow comparison;
(6) To group the conformations with similar secondary structure patterns of each concatenated trajectory. The 8-letter alphabet defining secondary structure according to DSSP was simplified to a 5-letter alphabet were the secondary structure assigned to a residue is in the ensemble {none; α-helix; 3_{10}-helix; π-helix, strand}. The Jaccard distance was used to measure the pairwise dissimilarity between each conformation generated through MD simulation. Then, a complete linkage hierarchical clustering method [46] was performed for tree pruning distances from 0.05 to 1 each 0.05. And each clustering, the performance was assessed with the silhouette score[269]. The pruning distance leading to the best silhouette score clustering was further analysed;
(7) The principal components analysis (PCA) modes were calculated for the backbone atoms (N, H, Cα, C, and O) after least-square fitting on the average conformation calculated on the concatenated data;

(8) The relative Gibbs free energy of the canonical ensemble was computed as a function of two reaction coordinates with Equation (1)[270]:

$$\Delta G = -k_B T \ln \frac{P_{(R_1,R_2)}}{P_{max}} \quad (1)$$

Where k_B represents the Boltzmann constant, T is the temperature, $P_{(R_1,R_2)}$ denotes the probability density of states along the two reaction coordinates, calculated using their joint probability, and P_{max} denotes the maximum probability. The population of each well was roughly estimated using a square defined with R_1 and R_2 value intervals and containing red to orange ΔG colors.

Each minimum was estimated for conformations in wells with Gibbs free energy $\leq 0.8 \times 10^{-20}$ $k_B T$.

(9) Pockets investigations were done with MDpocket with isovalue 0.5 on the concatenated data of each KIT species[473]. Well-defined grid points were exclusively extracted using VMD[486] and PyMol. Conformational analysis was performed on nine conformations with a maximum P1 volume surpassing an arbitrary volume threshold, eight conformations with a maximum P2 volume surpassing an arbitrary volume threshold, and sixteen conformations characterized by minimal P1 and P2 pocket volumes.

B. FIGURES SUPPLÉMENTAIRES

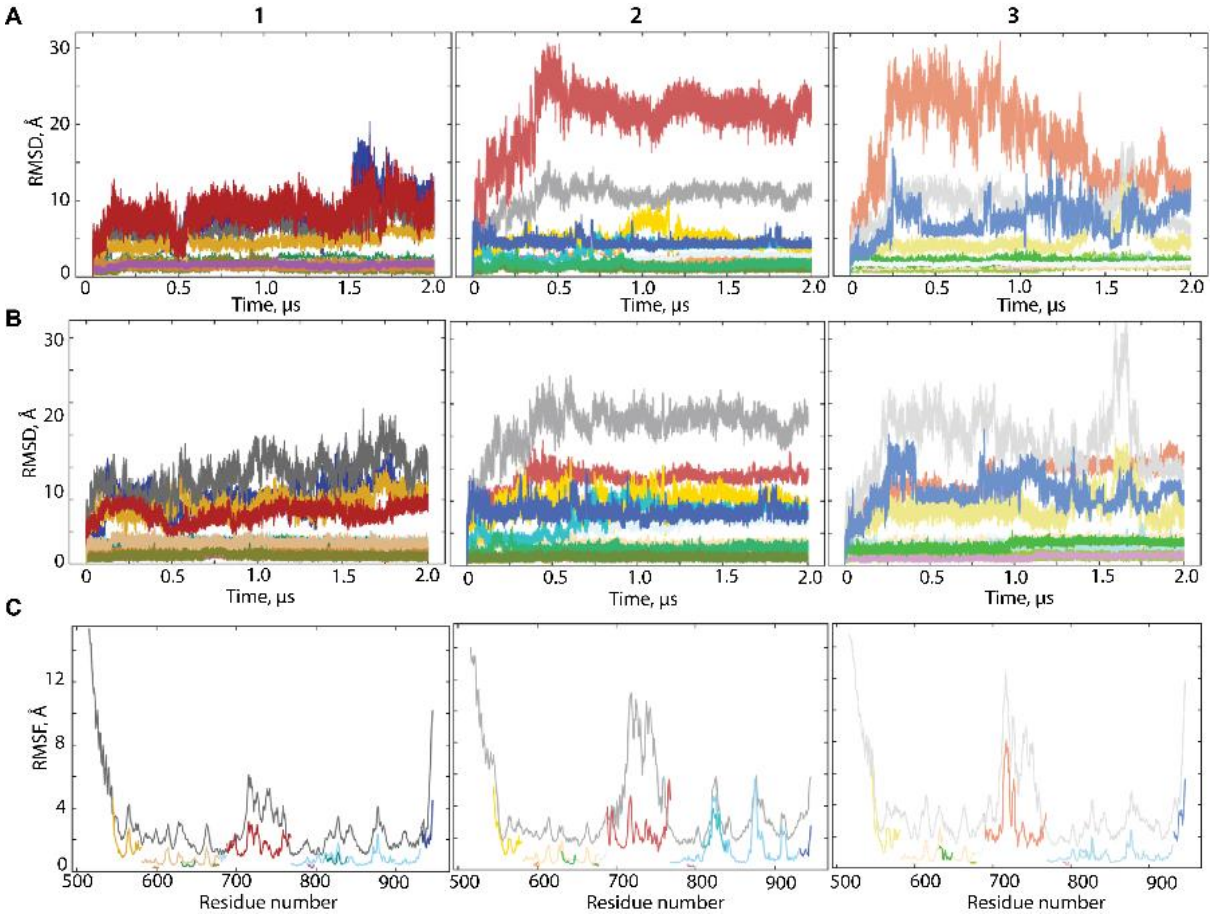


Figure S1 Molecular dynamics simulations of 3D model of the full-length cytoplasmic domain of RTK KIT in inactive state. (**A**,**B**) RMSDs from the initial coordinates (t=0 μs) computed for the Cα-atoms of the overall structure and individually for the Cα-atoms of KIT domain/regions after least-square fitting of the MD conformations of the TK domain on the initial conformations (**A**), and on the initial conformations of the respective domain (B). (**C**) RMSFs computed on the Cα atoms for MD conformations after the least-square fitting on the initial conformation of KIT or the respective domain. (**A–C**) KIT is in grey, N-lobe in beige, C-lobe in blue, JMR in yellow, P-loop in orange, αC-helix in green, hinge in olive, KID in red, C-loop in rose, A-loop in teal, C-tail in dark blue for the 1-3 trajectories (replicas) of MD simulations.

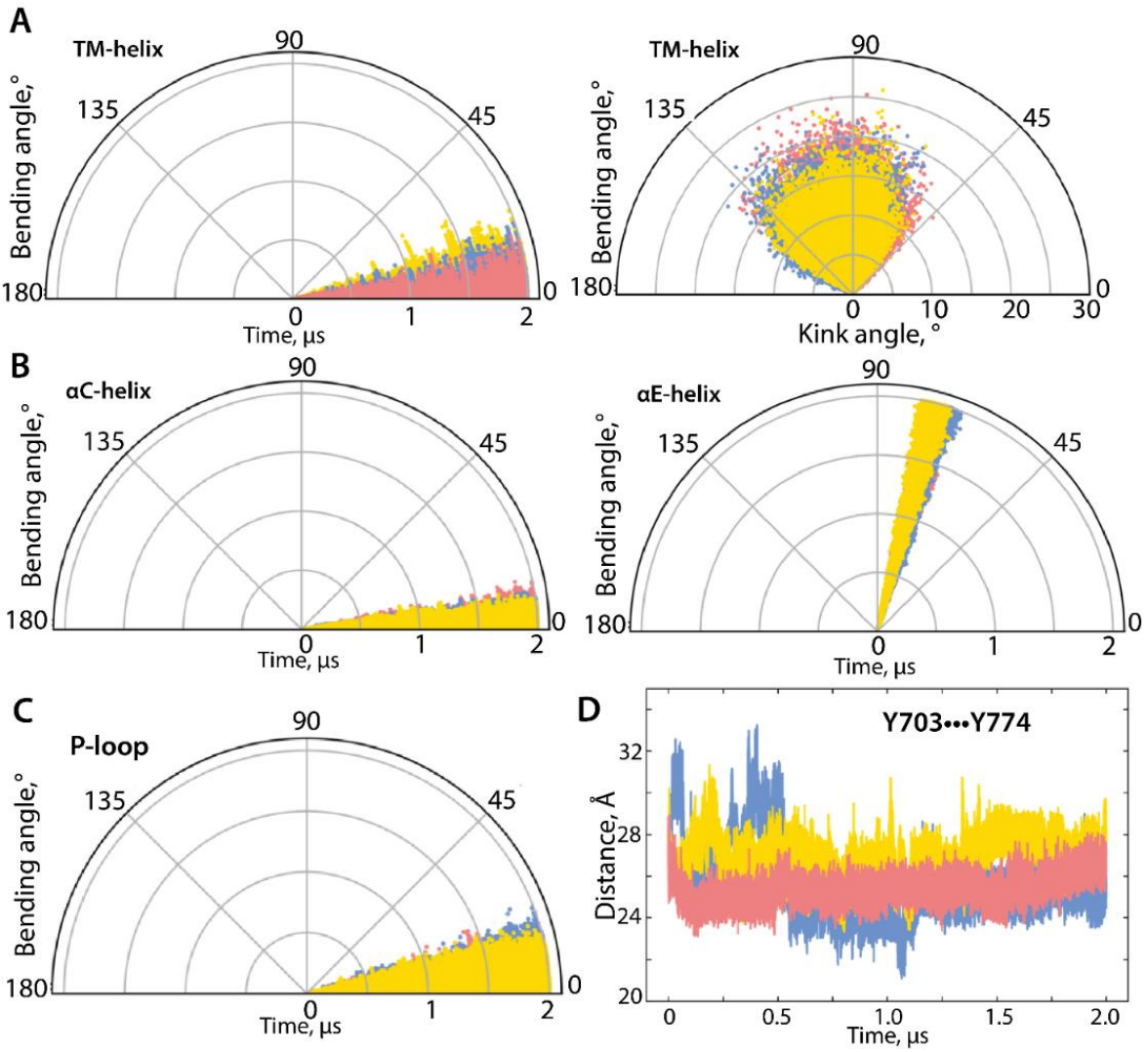


Figure S2 Geometry of the KIT fragments. (**A**) The angular variations of the TM helix, the bending angle (left) and distribution of the bending and kink angles (right). (**B**) Bending angle of the αC-helix (in respect to its staring conformation) and αE-helix (in respect to the membrane surface). (**C**) Bending angle of the P-loop. (**D**) Distance between the Cα-atoms of Y703 (αH1-helix of KID) and Y774 (αE-helix of the C-lobe). Calculations are performed after least-square fitting of the data on the TK domain. Conformations from different trajectories are distinguished by colour: red (1), blue (2) and yellow (3).

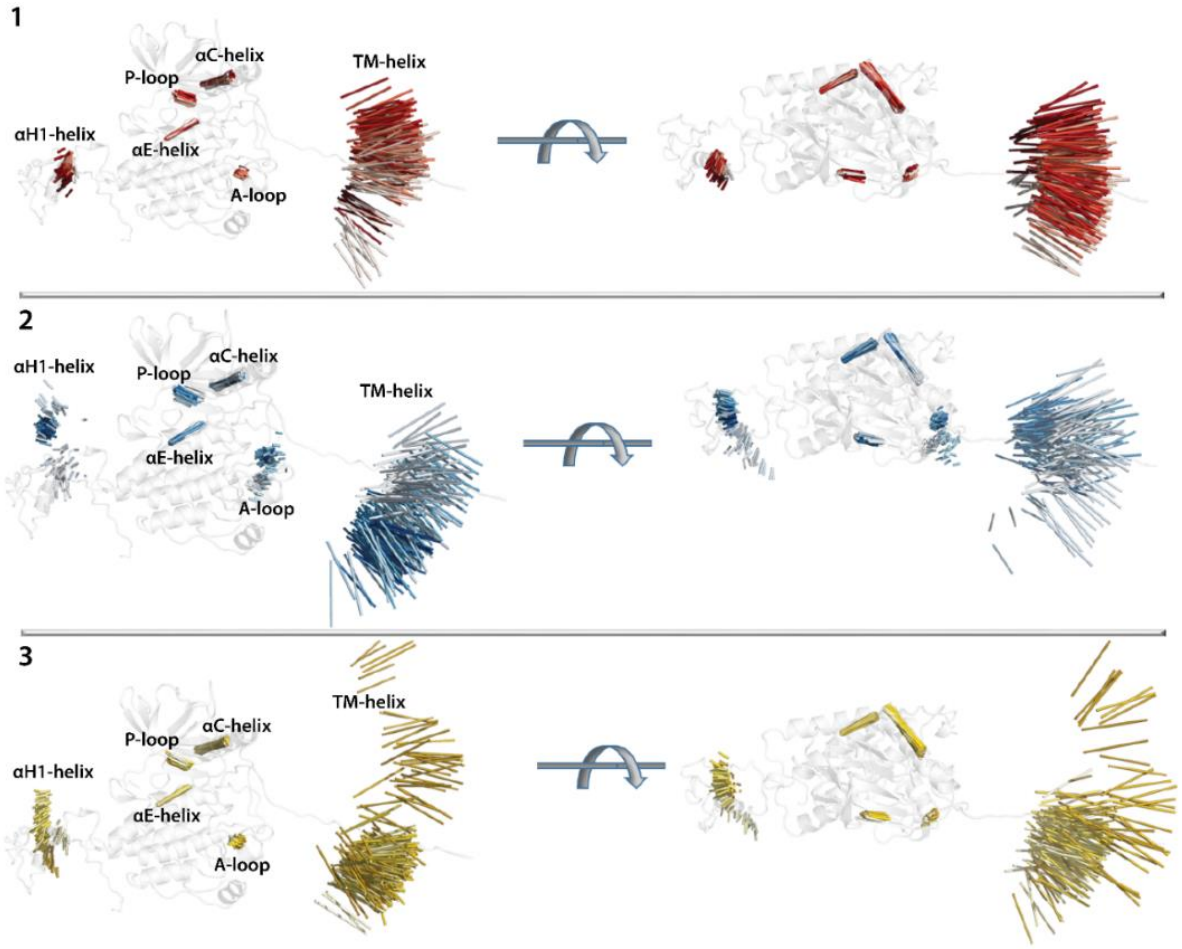


Figure S3 Conformational features of KIT. Positions of hints – TM-helix, αC-helix, αE-helix, αH1-helix, P-loop and β-hinge of A-loop – are superimposed on a mean conformation of KIT. Protein is shown as cartoon, each hint is presented by an axis of helix or by a vector colinear with a strand. Two orthogonal projections are shown. Calculations were performed on cMD conformations taken each 10 ns from the individual trajectories distinguished by colour – red (**1**), blue (**2**) and yellow (**3**). The colour gradient shows the evolution of a trajectory, from light (t = 0) to dark (t = 2 μs).

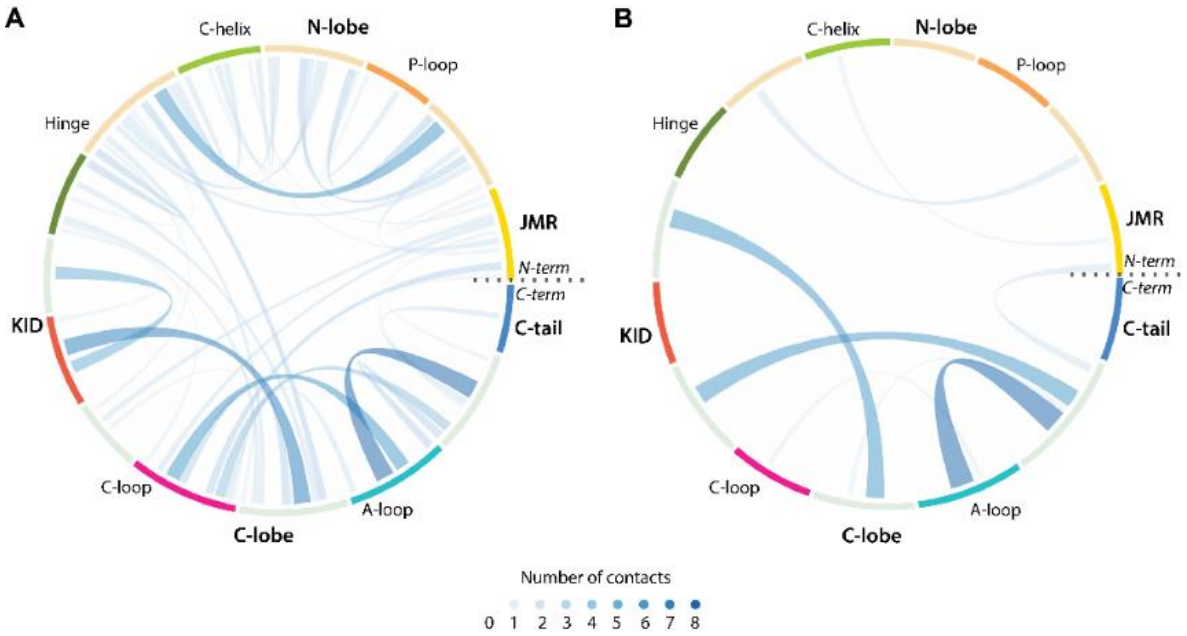


Figure S4 The non-covalent contacts stabilizing the crystallographic structure (PDB ID: 1T45) of RTK KIT in the inactive state. The string diagram compiles the H-bonds (**A**) and hydrophobic interactions (**B**) which are shown as curves coloured according to the occurrences of contacts, from 0 (white) to 8 (blue). The KIT domains and the functionnally related fragments are distinguished by colours and labelled in bold and regular characters respectively. Contacts involved in the formation of regular structures (H-bonds forming helix or sheet) and intra-domain framework except the functionally related regions were excluded from consideration.

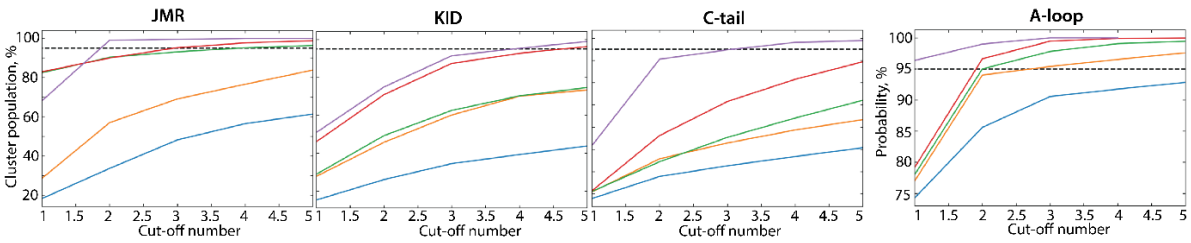


Figure S5 Ensemble-based clustering of cMD conformations of the highly varied KIT regions – JMR, KID, A-loop and C-tail. Fraction of the population of clusters obtained for the concatenated trajectory. Clustering was performed on each 100-ps frame of each trajectory using cut-off values varying from 1.0 to 4.0 Å, with a step of 0.5 Å. The lines in blue, orange, green, red, and violet correspond to cut-off of 2.0, 2.5, 3.0, 3.5 and 4.0 Å respectively.

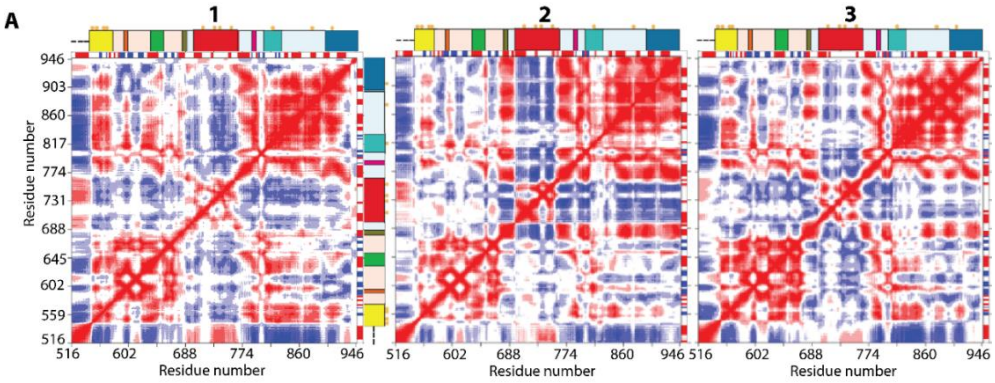


Figure S6 Intrisinc motion in KIT and its interdependence. Inter-residue cross-correlation map computed for the Cα-atom pairs of KIT computed for each cMD trajectory after least-square fitting on the initial conformation. Correlated (positive) and anti-correlated (negative) motions between Cα-atom pairs are shown as a red-blue gradient.

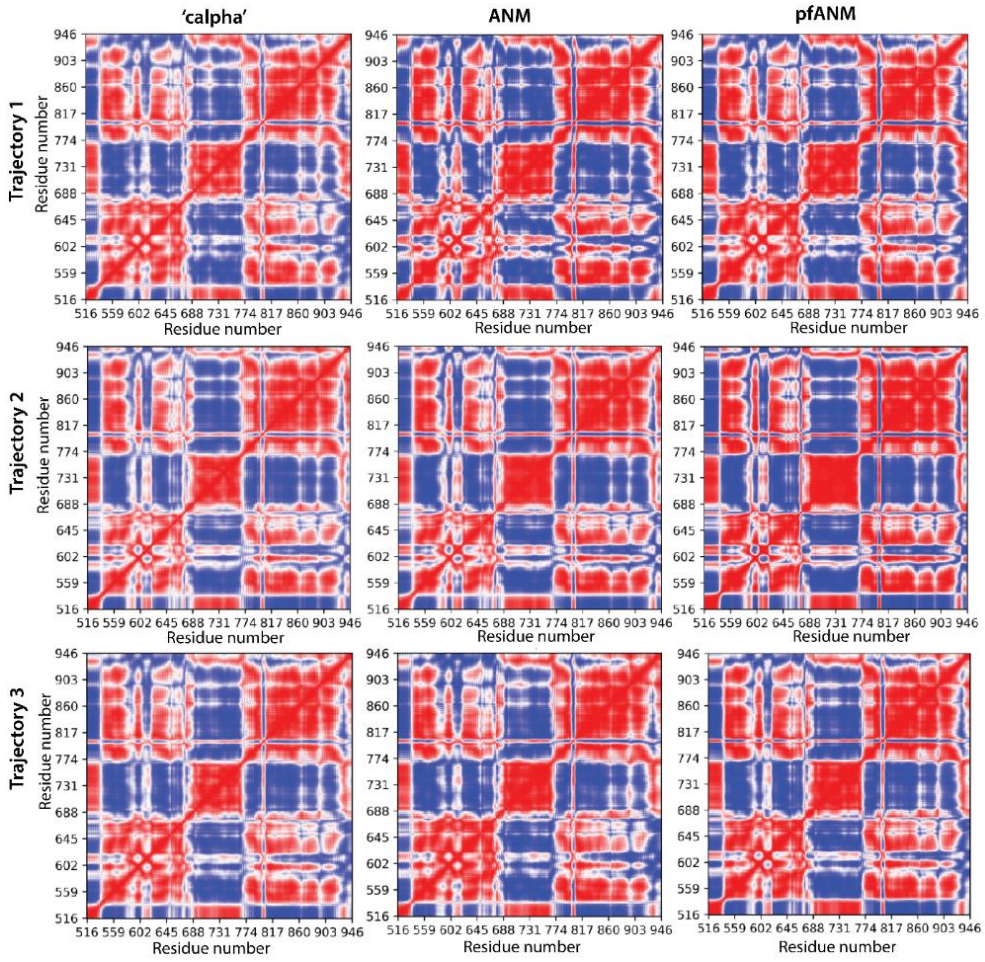


Figure S7 Intrisinc motion in KIT and its interdependence. Dynamical inter-residue cross-correlation maps resulting from NMA of the mean conformation of KIT in each cMD trajectory (1-3), calculated by using three different force fields – the 'capha', the Anisotropic Normal Model (ANM) and Elastic Network model (pfANM). Correlated (positive) and anti-correlated (negative) motions between Cα-atom pairs are shown as a red-blue gradient.

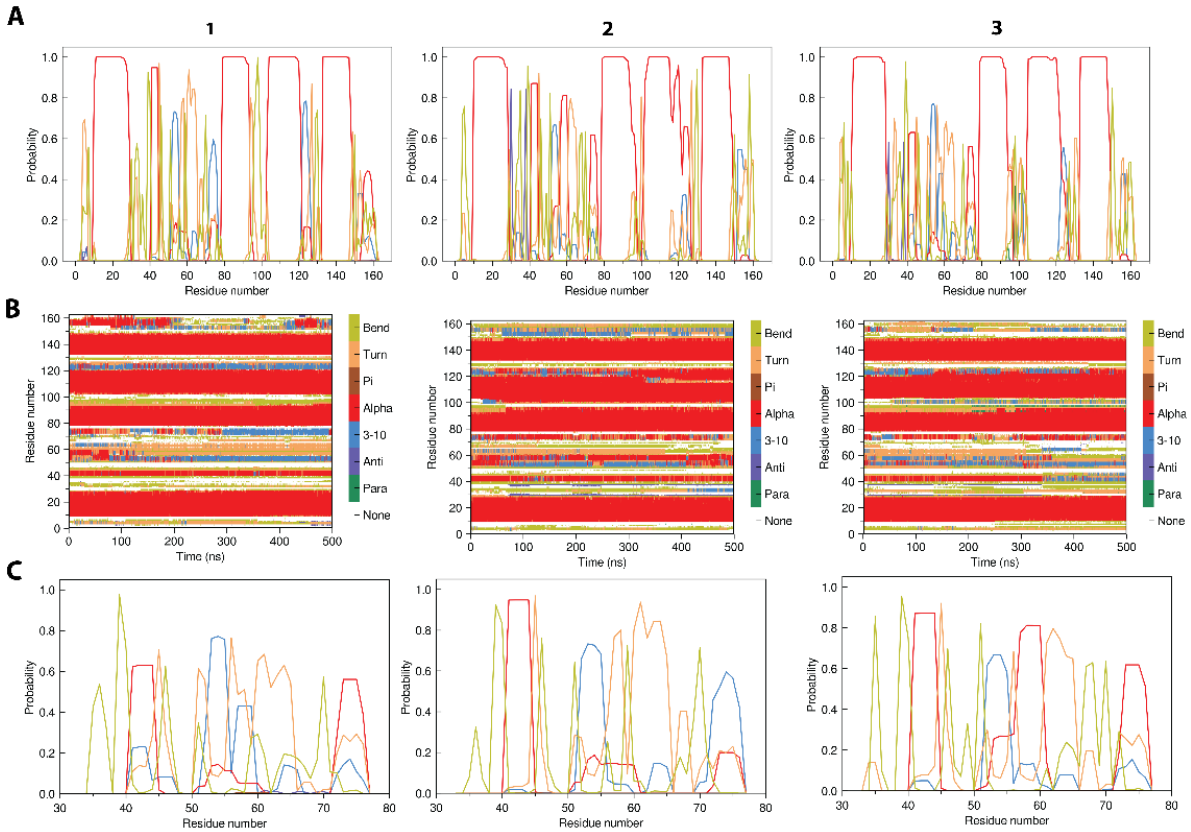


Figure S8 Folding of hVKORC1 over the MD simulations. Secondary structure of each residue of hVKORC1 (**A**) and of L-loop (**C**) assigned by DSSP. Assignment of the secondary structures to colors is given as follows: α-helix is in red, 3_{10}-helix is in blue, the parallel and antiparallel strands are in green and violet respectively; turn is in orange and bend is in dark yellow. Proportion of every secondary structure type for each residue is given as a probability. (**B**) The time-related evolution of the secondary structures of each residue as assigned by DSSP with the type-coded secondary structure bar.

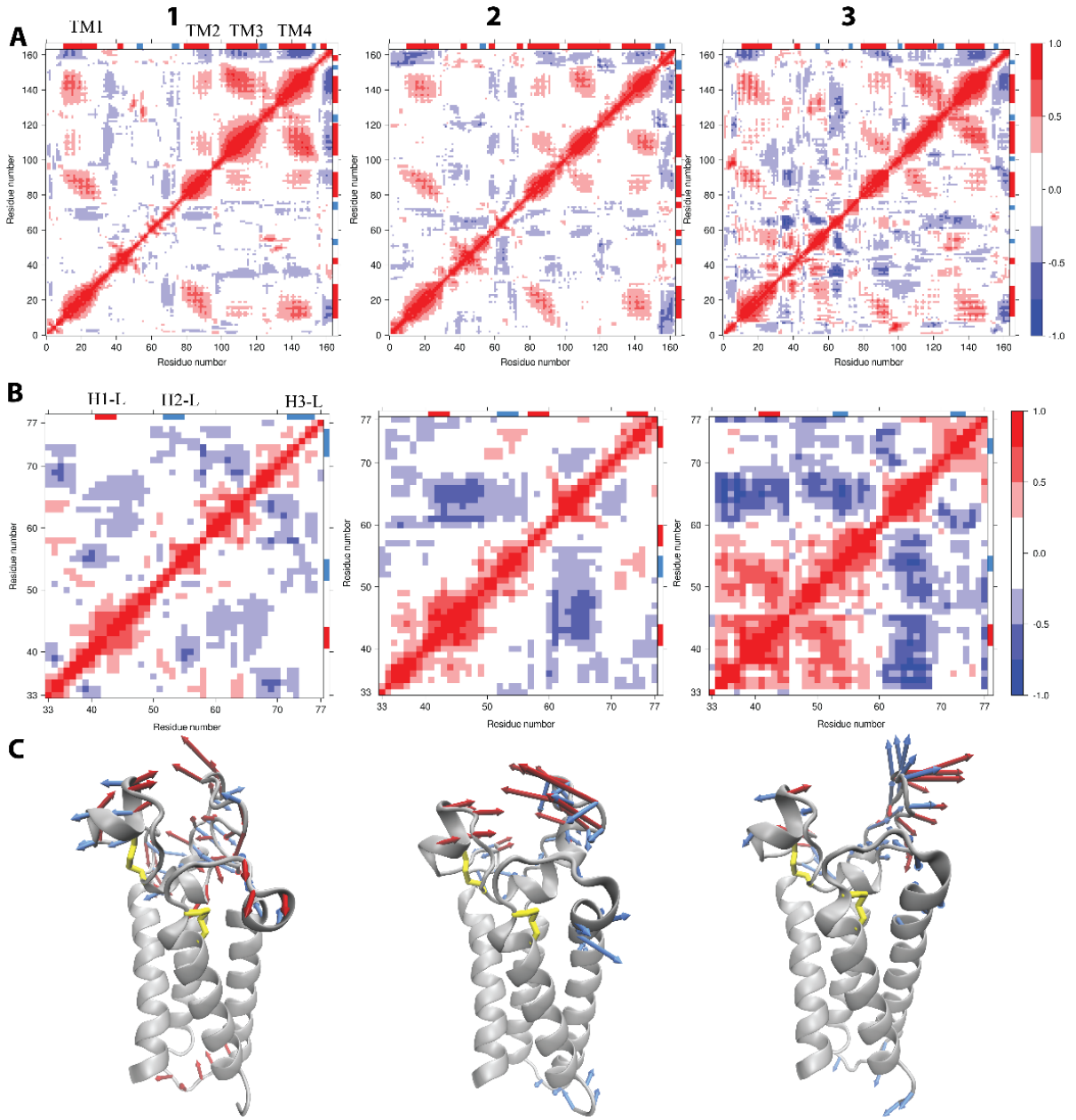


Figure S9 Intrinsic motion of hVKORC1 and its L-loop. Inter-residue cross-correlation map computed for the Cα-atom pairs after fitting on the respective first conformation (t=0ns) of the full-length hVKORC1 (**A**) and of the L-loop (**B**) over each replica 1-3. Correlated (positive) and anti-correlated (negative) motion between the Cα-atom pairs are shown as a red-blue gradient. (**C**) Atomic components in the first PCA modes of hVKORC1 (after omitting the highly fluctuating residues from the N- and C-terminals) are drawn as red (1st mode) and blue (2nd mode) arrows projected onto the respective average structure. Only the motion with an amplitude ≥ of 2 Å was represented. The protein is shown as ribbons diagrams with the S-S bridge as yellow sticks. All computation was performed on the Cα-atoms with the RMSF fluctuations less than 4 Å of each protein after fitting on initial conformation.

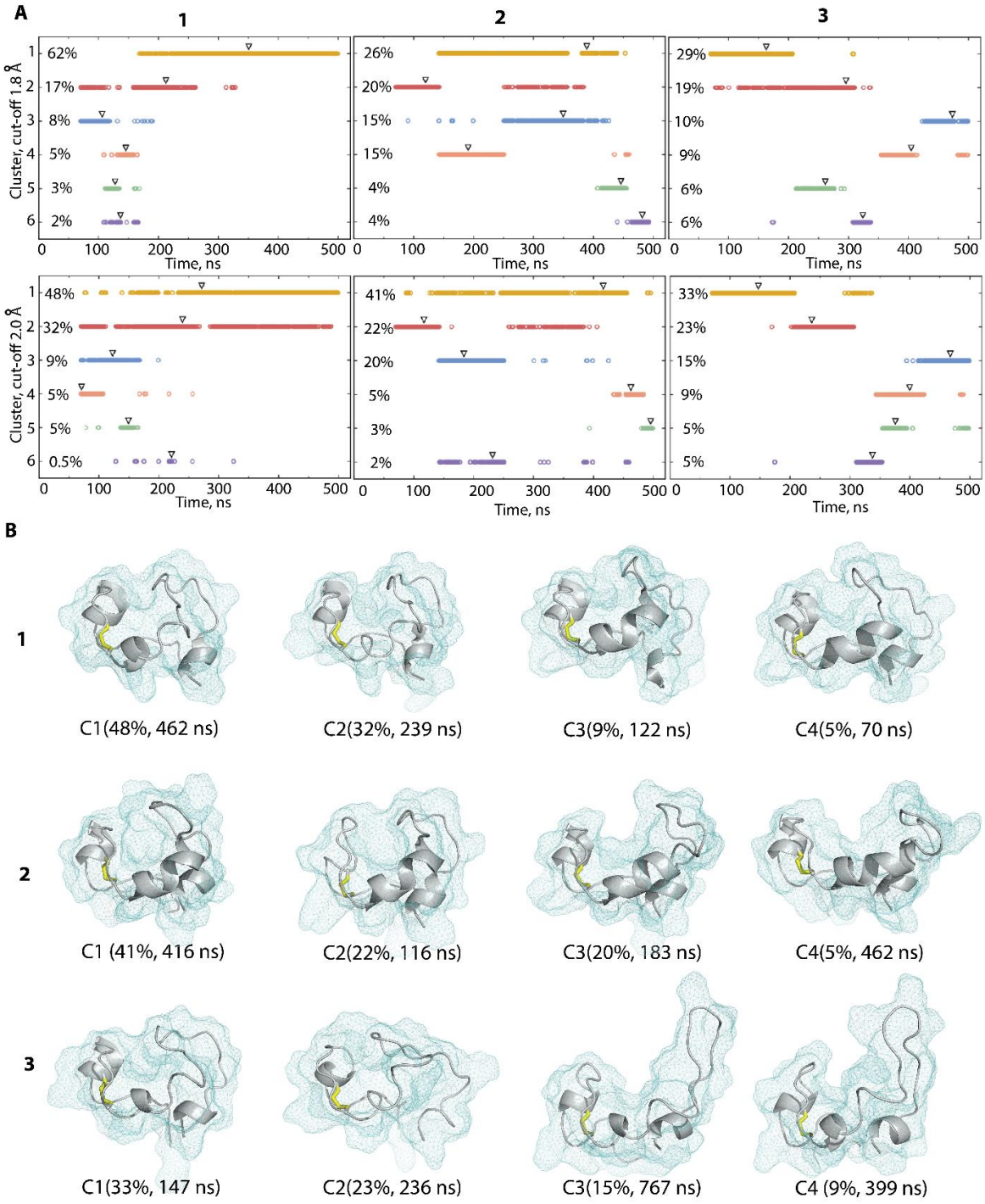


Figure S10 Clustering analysis of the L-loop conformations by using the ensemble-based clustering. (**A**) Regrouping of clusters over each trajectory of MD simulation. Calculation was performed on every 10-ps frames after omitting of the first 70 ns with using the cut-off values of 1.8 (top panel) and 2.0 Å (bottom panel). Clusters classified from the most to the less populated (C1-C6) and affiliated to the time of MD trajectory. Triangle symbol indicated the frames used as the representative conformations. (**B**) The representative conformation of the L-loop (shown as ribbon with a meshed surface and the S-S bridge as sticks) from each cluster with population >4% (cut-off of 2.0 Å). Population of each cluster is given in brackets together with the time of observation of the representative conformation.

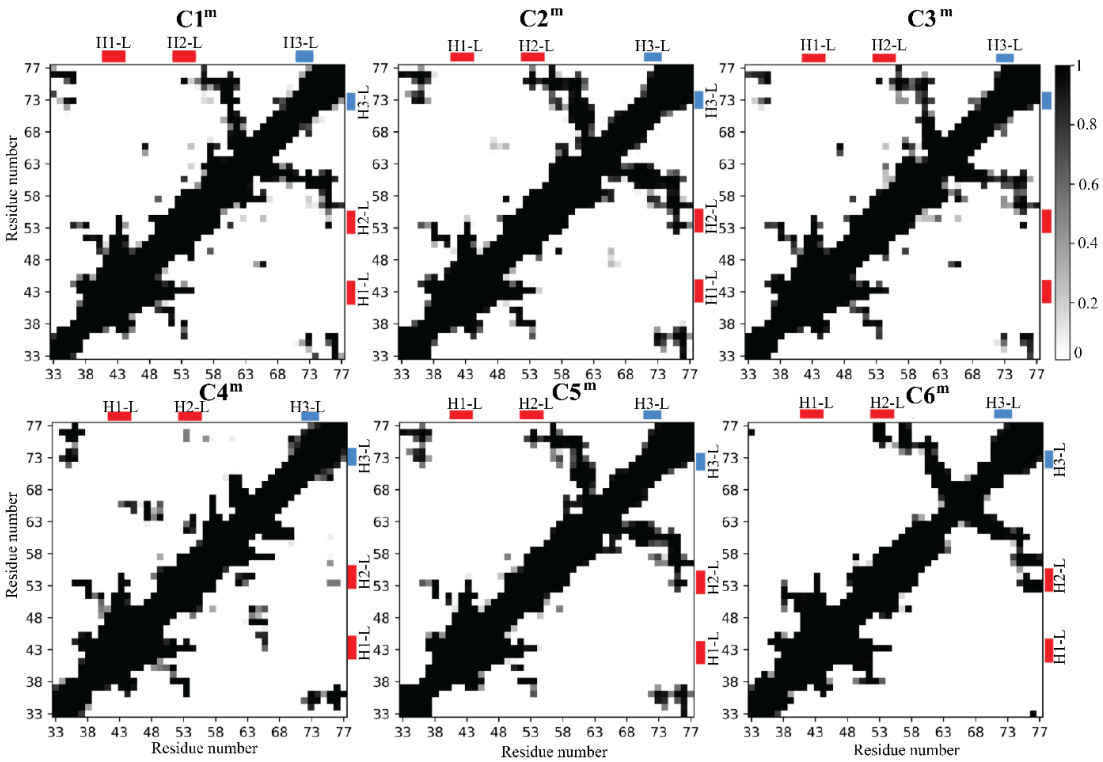


Figure S11 The contact maps of pairwise distances Cα−Cα (< 10 Å) computed for each conformation from the mostly populated clusters (> 4%) found on the concatenated trajectory. Gradient from white (0) to black (1) shows a frequency of the contact during the simulation.

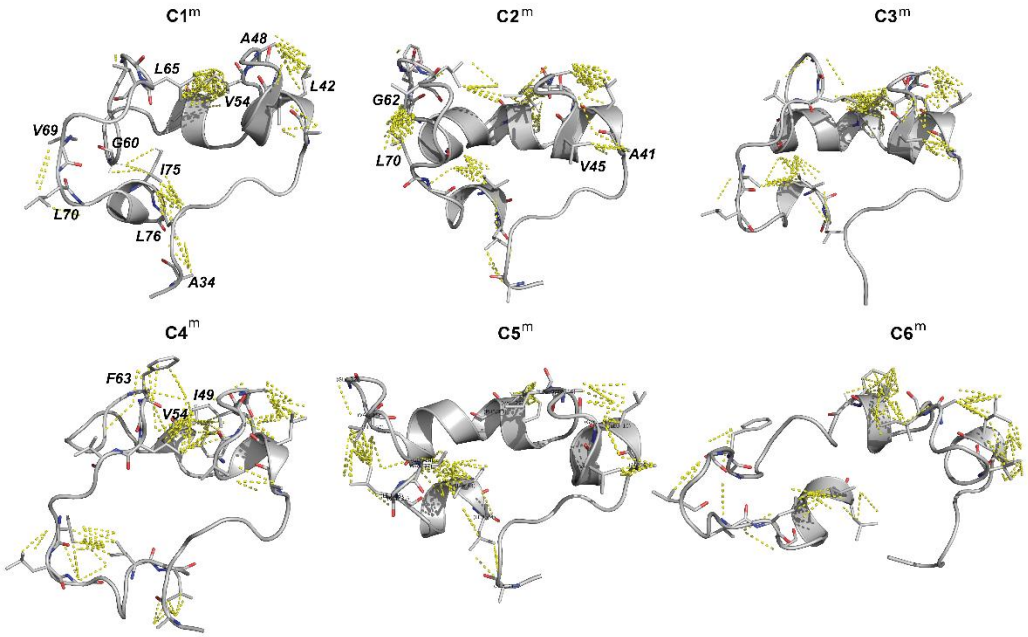


Figure S12 The hydrophobic contacts (yellow dashed lines) in L-loop of hVKORC1. The labels of residues are shown on conformation from the cluster $C1^m$; the other labels were added if required.

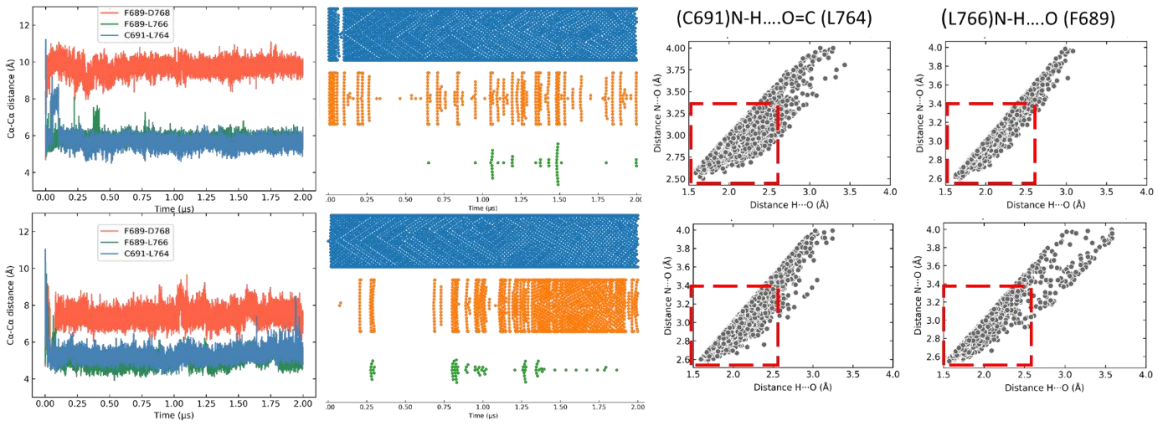


Figure S13 Stabilising contacts between residues from N- and C- extremities of KID fused to KIT. Distance between the pairs of Cα-Cα atoms (first columns at the left) and their occurrences over the MD simulations (second column). Distributions of distances between a donor (D) and an acceptor (A) and a hydrogen (H) and acceptor (A) characterising the strength of H-bonds (C691)N-H···O=C(L764) (third column) and (L766)N-H···O(F689) (fourth column).

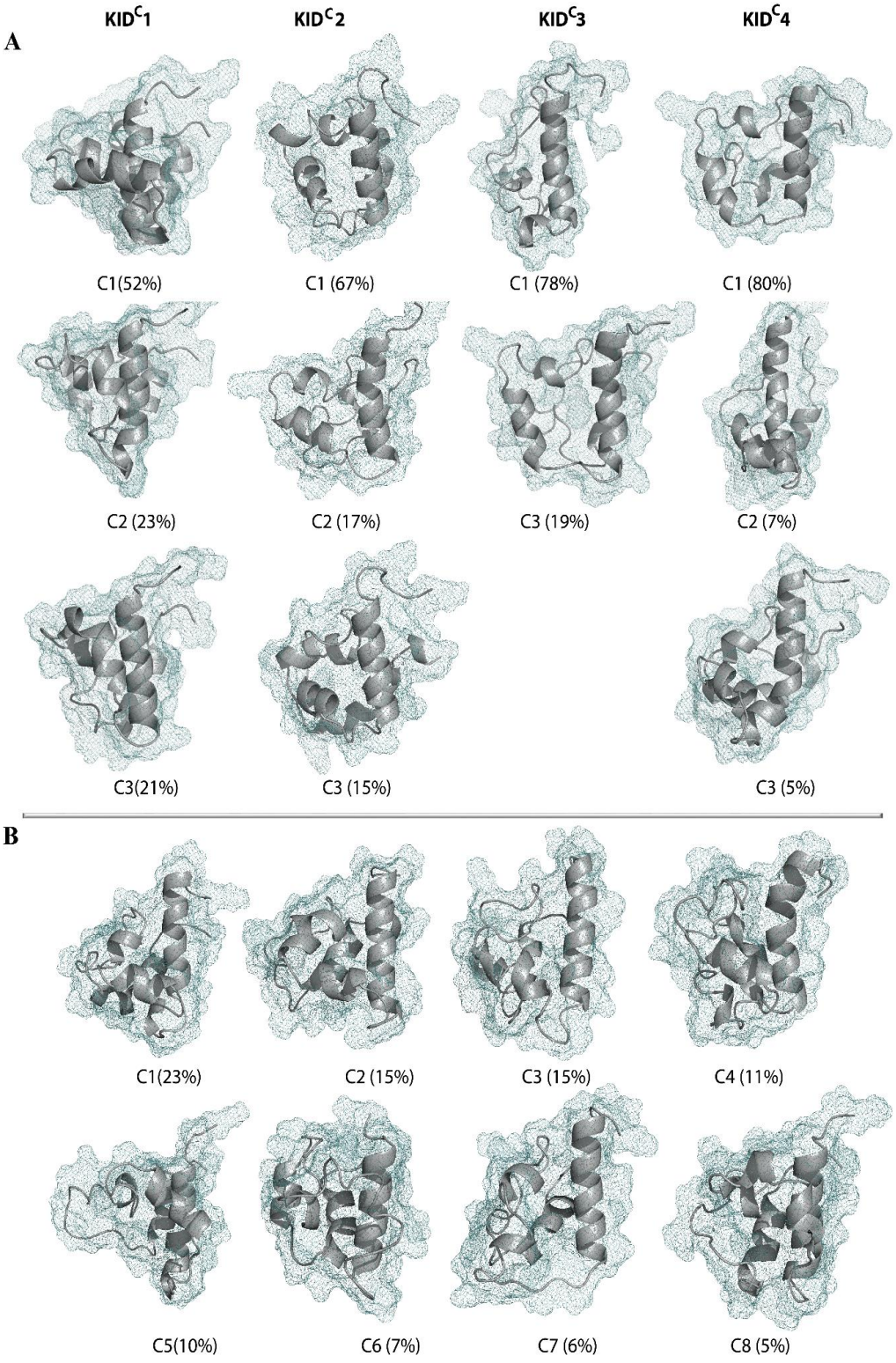


Figure S14 Clusters and their representative conformations of cleaved KID simulated as unconstrained species (KID^C). Representative conformations of clusters obtained on each trajectory (**A**) and concatenated trajectories (1-4) (**B**) of the MD simulation of KID. Conformations were clustered with the ensemble-based clustering using the RMSD cut-off of 4 Å. The first 100 ns as well as the most fluctuating residues from KID extremities with the RMSF values exceeding 6 Å were omitted from the analysis. KID is shown as a grey cartoon with a meshed surface. Population of each cluster is given in brackets.

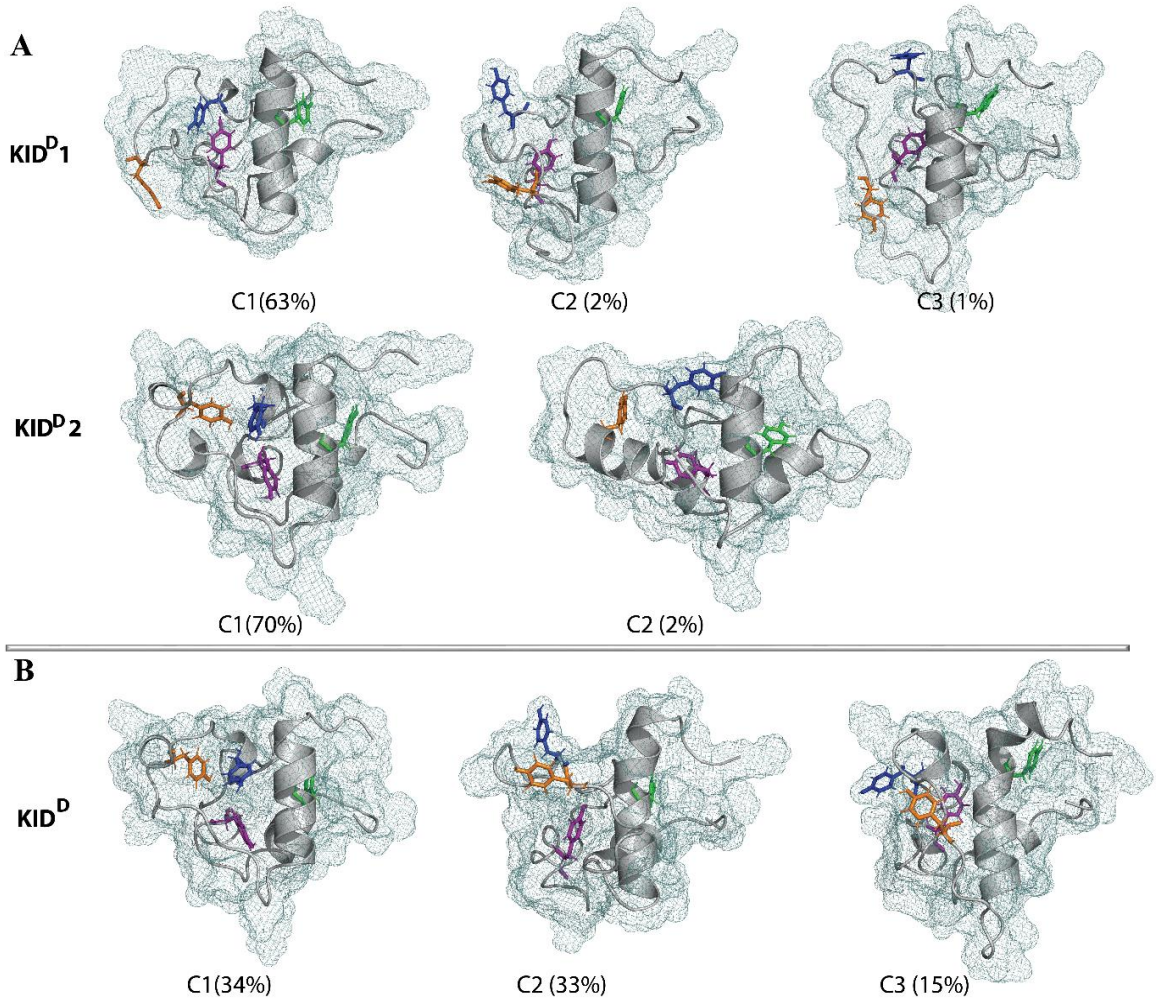


Figure S15 Clusters and their representative conformations of KID fused to KIT (KID^D). Representative conformations of clusters obtained on each trajectory (A) and concatenated trajectories (1-2) (**B**) of the MD simulation of KID. Conformations were clustered with the ensemble-based clustering using the RMSD cut-off of 4 Å. The first 100 ns as well as the most fluctuating residues from KID extremities with the RMSF values exceeding 6 Å were omitted from the analysis. KID is shown as a grey cartoon with the tyrosine residues (coloured differently) as sticks and a meshed surface. Population of each cluster is given in brackets.

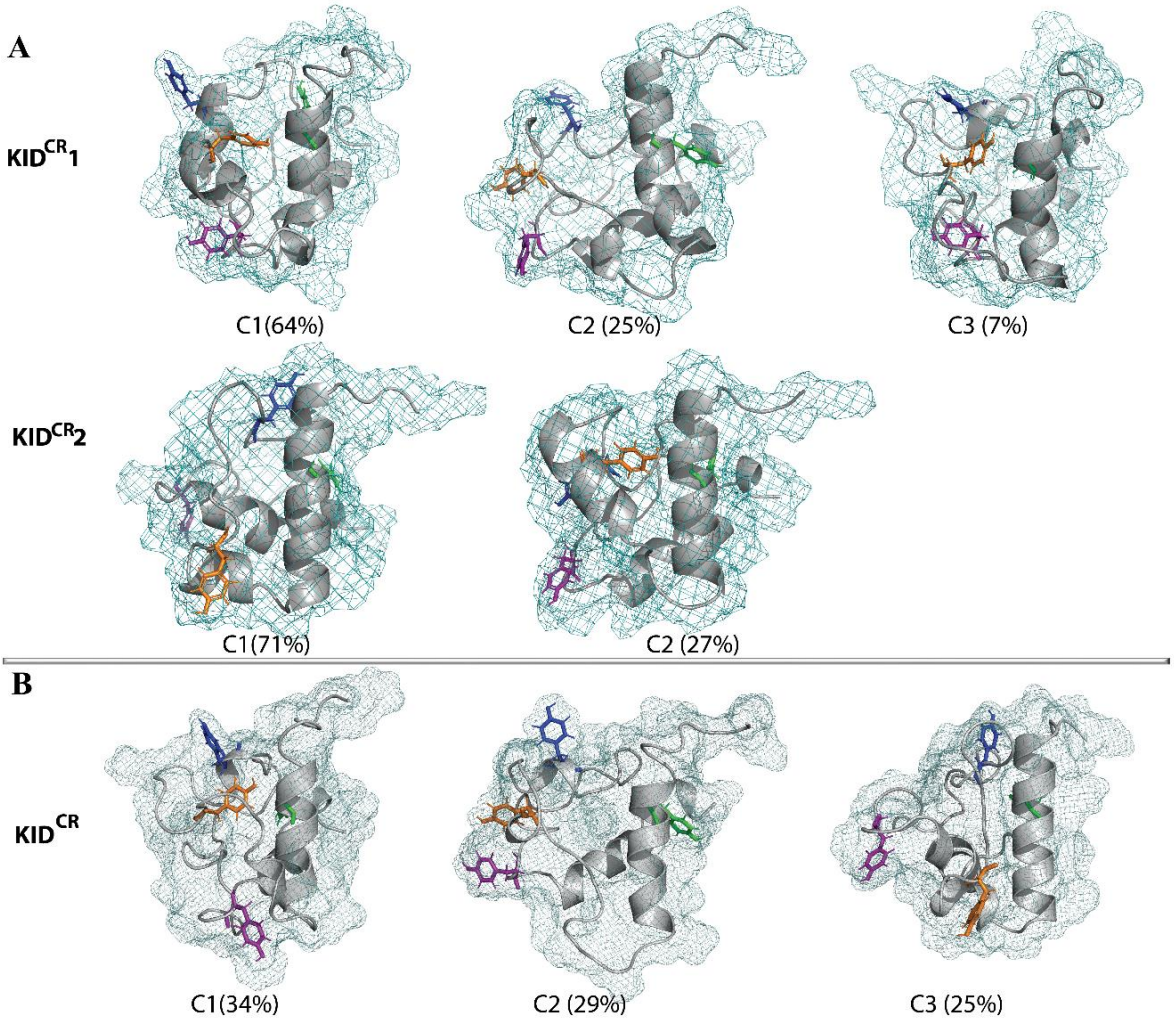


Figure S16 Clusters and their representative conformations of cleaved KID simulated with the constrained distance between its N- and C-ends (KID^{CR}). Representative conformations of clusters obtained on each trajectory (**A**) and on concatenated trajectories (1-2) (**B**) of the MD simulation of KID. Conformations were clustered with the ensemble-based clustering using the RMSD cut-off of 4 Å. The first 100 ns as well as the most fluctuating residues from KID extremities with the RMSF values exceeding 6 Å were omitted from the analysis. KID is shown as a grey cartoon in grey with the tyrosine residues (coloured differently) as sticks and a meshed surface. Population of each cluster is given in brackets.

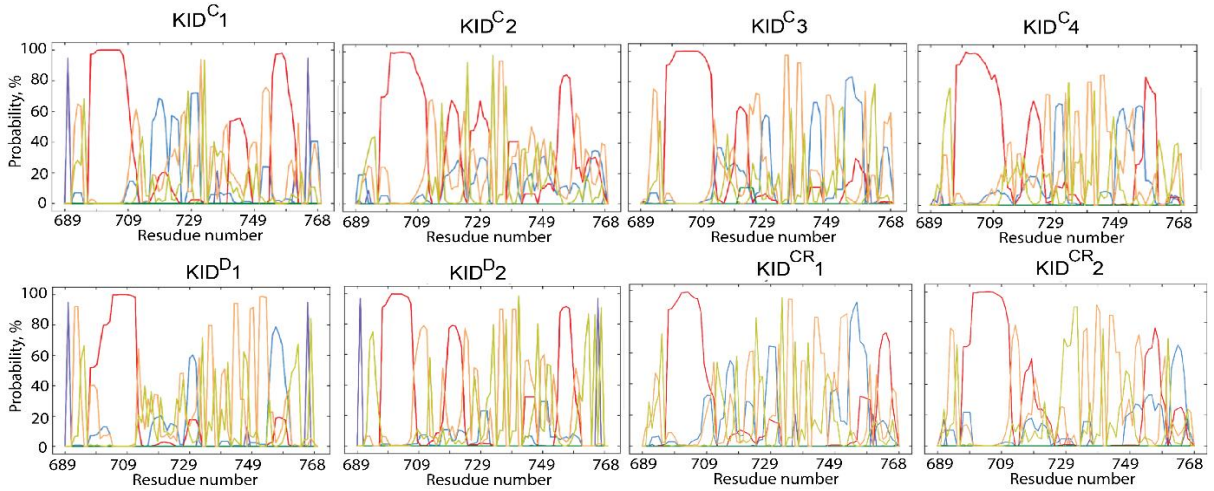


Figure S17 Secondary structure assignment (by DSSP) of KID during MD simulations. The proportion of every secondary structure type for each residue is given as a probability. Secondary structure is coded by colour: αH-, 3_{10}- and π-helices are in red, blue and green respectively; parallel and antiparallel strands are in rose and violet; turn and bend are in orange and pear.

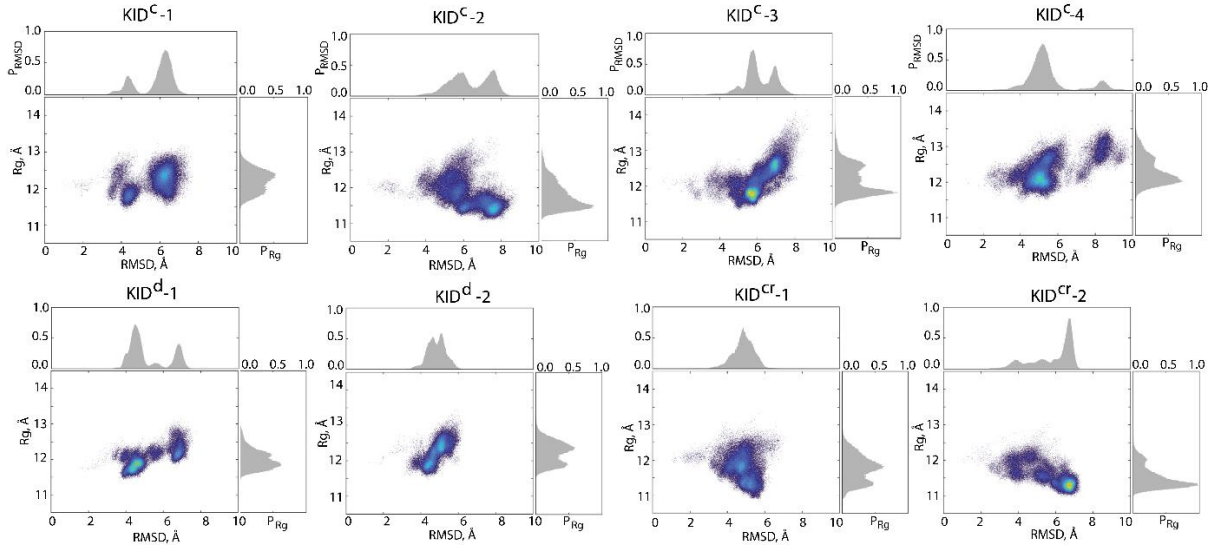


Figure S18 Compaction of KID estimated on conformational ensembles generated over each cMD trajectory for every KID entity studied. 2D representation of the free energy landscape of KID (KIDc, KIDd and KIDcr) conformational ensembles plotted as a function of Rg (in Å) versus RMSD (in Å). Probability distribution of Rg (right) and RMSD (top) calculated from each generated ensemble of KID studied as cleaved entity and as domain of KID. All calculations are performed on the Ca-atoms. The red colour represents the high energy state, yellow and green low and blue represents the lowest stable state.

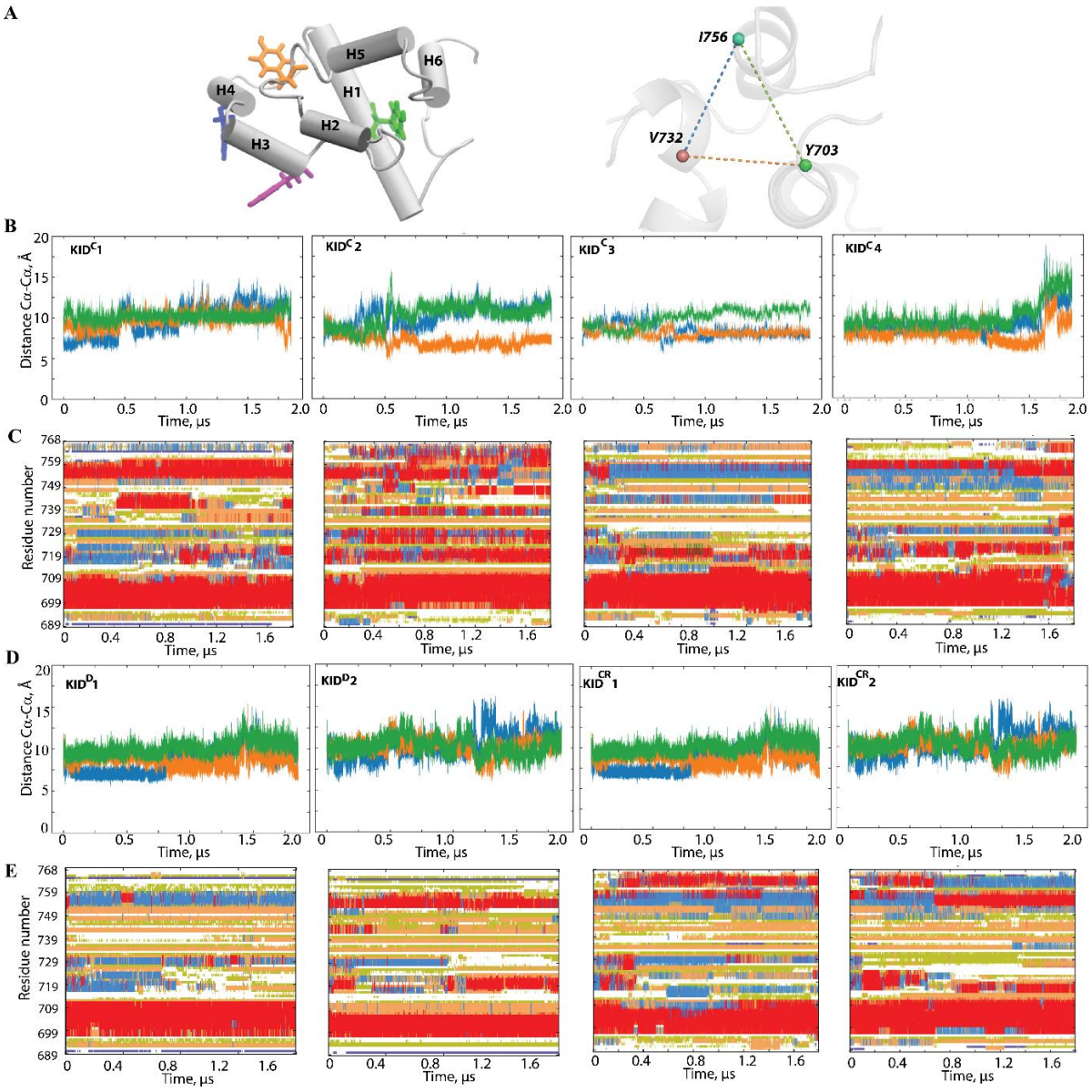


Figure S19 The inter-residue geometry of tyrosines in the isolated unconstrained KID and its relation with folding. (**A**) KID structure with helices shown as solid cylinders and tyrosine residues as sticks (left) and the triangle designed on the most stable over the MD simulations residues (with the smallest RMSF values) (right). (**B** and **D**) Distances between each pair of the tyrosine residues in each MD trajectory, coloured as the edges of triangle. (**C** and **E**) The time-related evolution of the secondary structures of each residue as assigned by DSSP with the type-coded secondary structure bar. (**D**) Distances between the most 'stable' (minimal RMSF values) residues Y703, V732 and I756, over each cMD trajectory, coloured as the edges of triangle in (**A**).

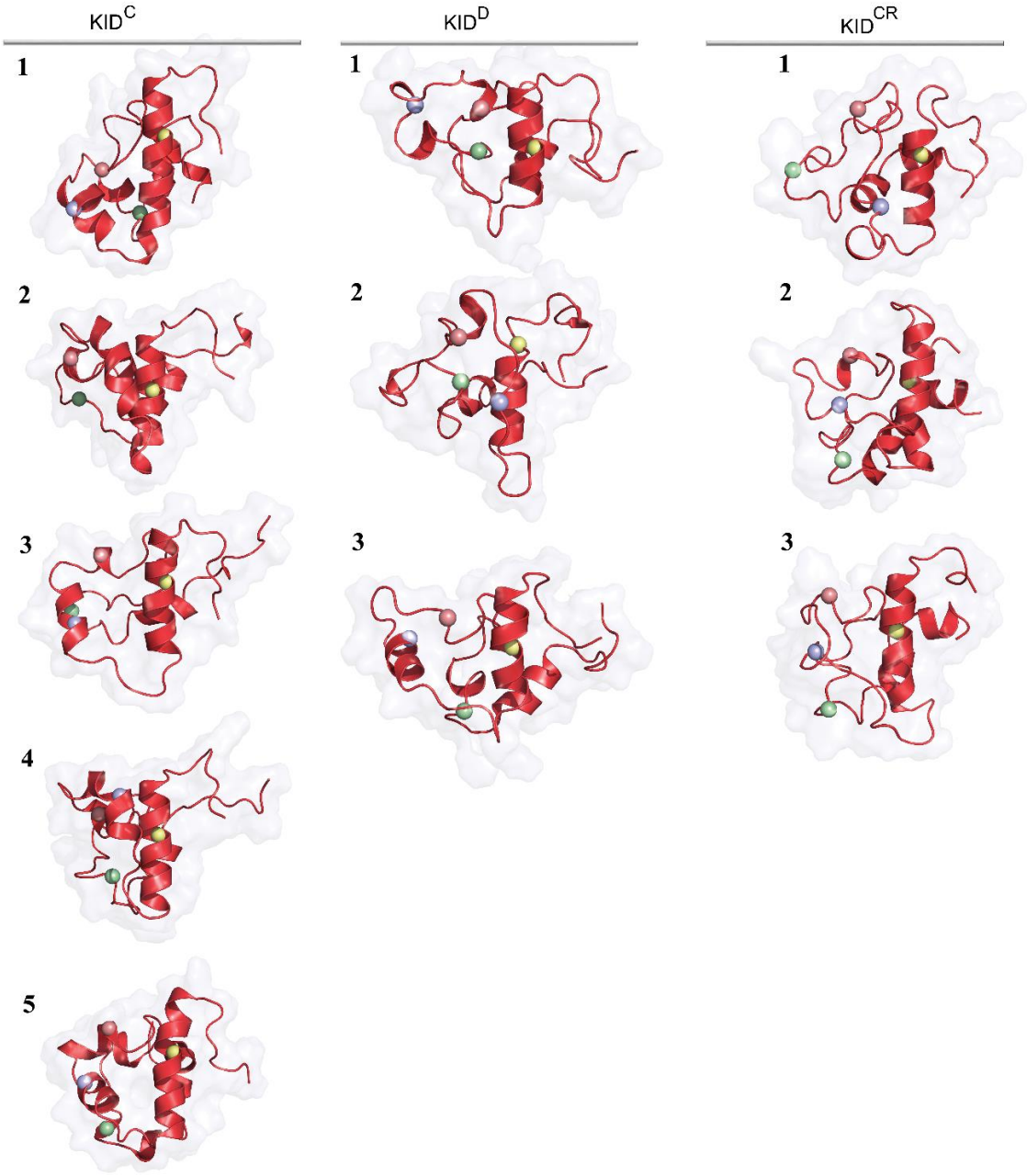


Figure S20 Representative conformations of KID from the deep wells on the free energy landscape (FEL) plotted as a function of Rg *versus* RMSD. KID is displayed as a red cartoon contoured with a surface-filled model. Position of the tyrosine residues (Cα-atoms) are shown as balls coloured in yellow, blue, red and green for Y703, Y721, Y730 and Y747 respectively.

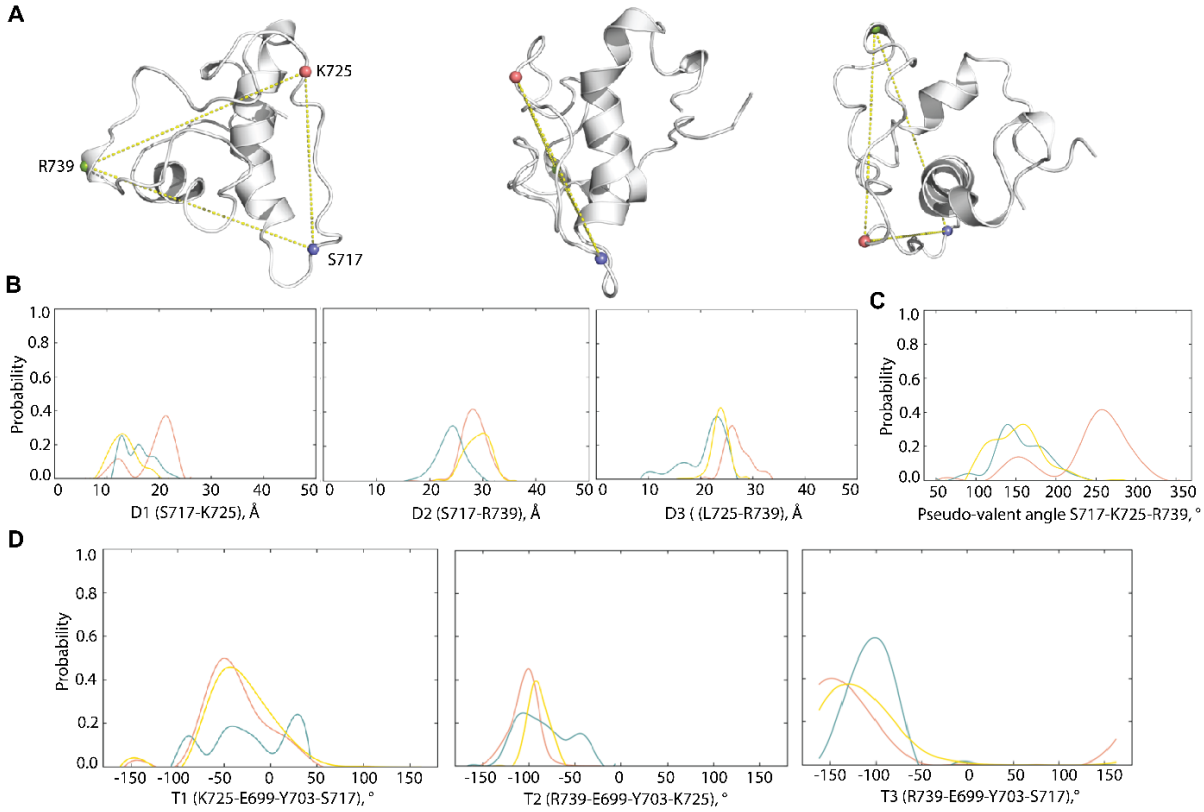


Figure S21 Geometry metrics of the most fluctuating residues. (**A**) 3D structure. Protein is shown as cartoon; the most fluctuating residues are displayed as balls. (**B**) The distance values probability. (**C**) The pseudo-valent angle probability. (**D**) Pseudo-torsion (dihedral) angle probability. (**B-C**) Metrics were calculated for each KID entity: KID^{GC} in yellow, KID^{D} in red and KID^{C} in teal.

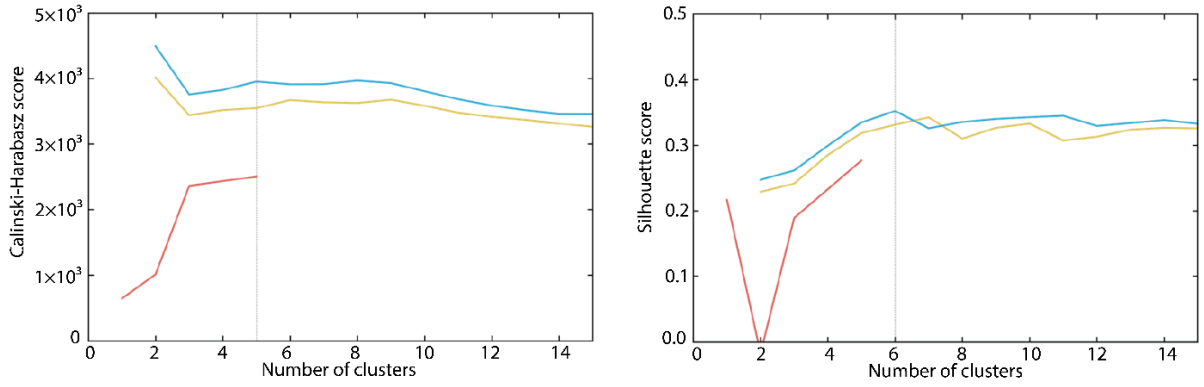


Figure S22 Performance scores were obtained during hyperparameters tweaking to find the better clustering method. Number of clusters obtained by DBSCAN (in red), K-means (in blue) and Ward's methods (in yellow). Scores were computed with Calinski-Harabasz (left) and Silhouette (right).

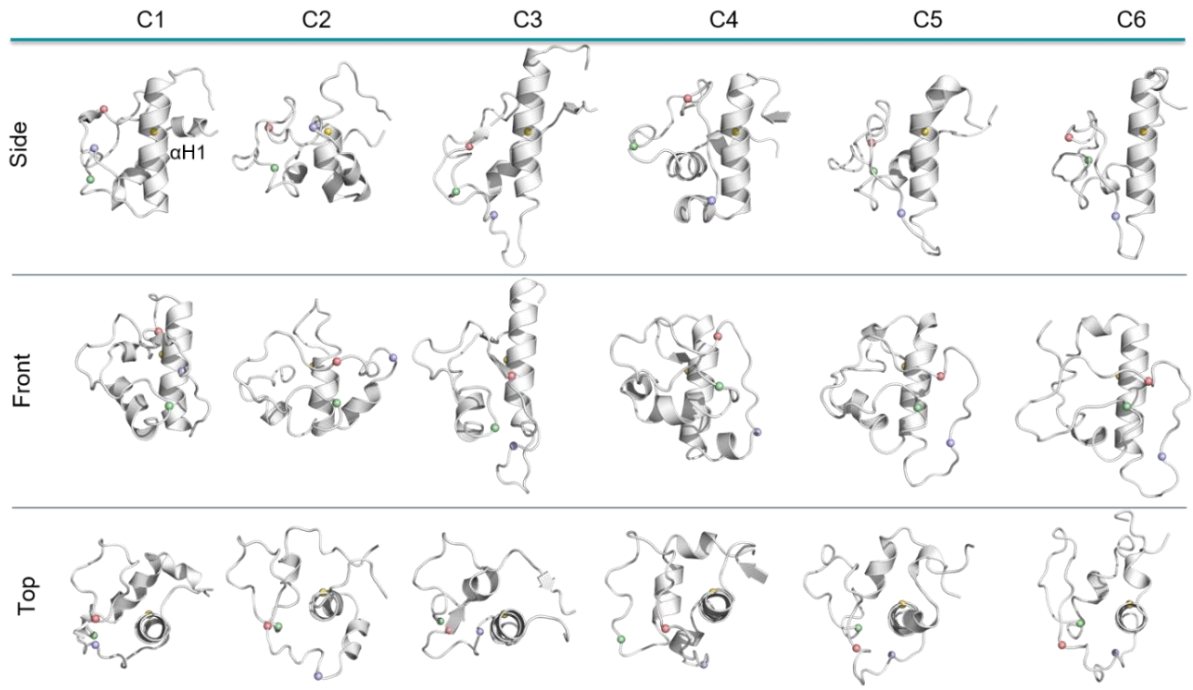


Figure S23 Representative conformations from each cluster (C1 – C6). The calculation was performed on the concatenated trajectories after fitting all data on αH1-helix taken from KID conformation at t=0. Three views–side, front and top–concerning αH1-helix are shown. Protein is displayed as cartoon, and the position of each tyrosine residue is shown by ball: Y703 in yellow, Y721 in lilac, Y730 in coral and Y747 in green.

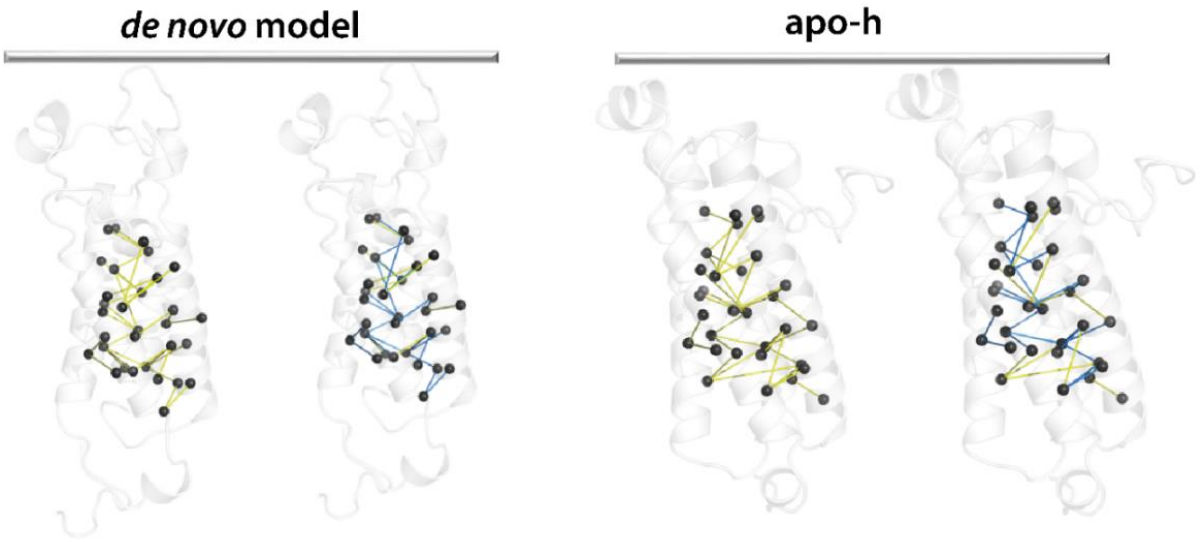


Figure S24 Non-covalent interactions stabilising hVKORC1 TMD helices in *de novo* model embedded in phospholipid bilayers mimicking the ER membrane in aqueous solution, and apo-h form simulated in aqueous solution. Non-covalent interactions, denoted as an undirected graph whose vertices represent a residue and a link between two residues reflects the presence of at least one type of non-covalent interaction. Van-der-Waals contacts (≤ 4.0 Å, events with frequency ≥ 0.8) were calculated for all heavy atoms (O, N, C and S). All observed contacts in each model (left) and blue links discriminated contacts common between two models (right).

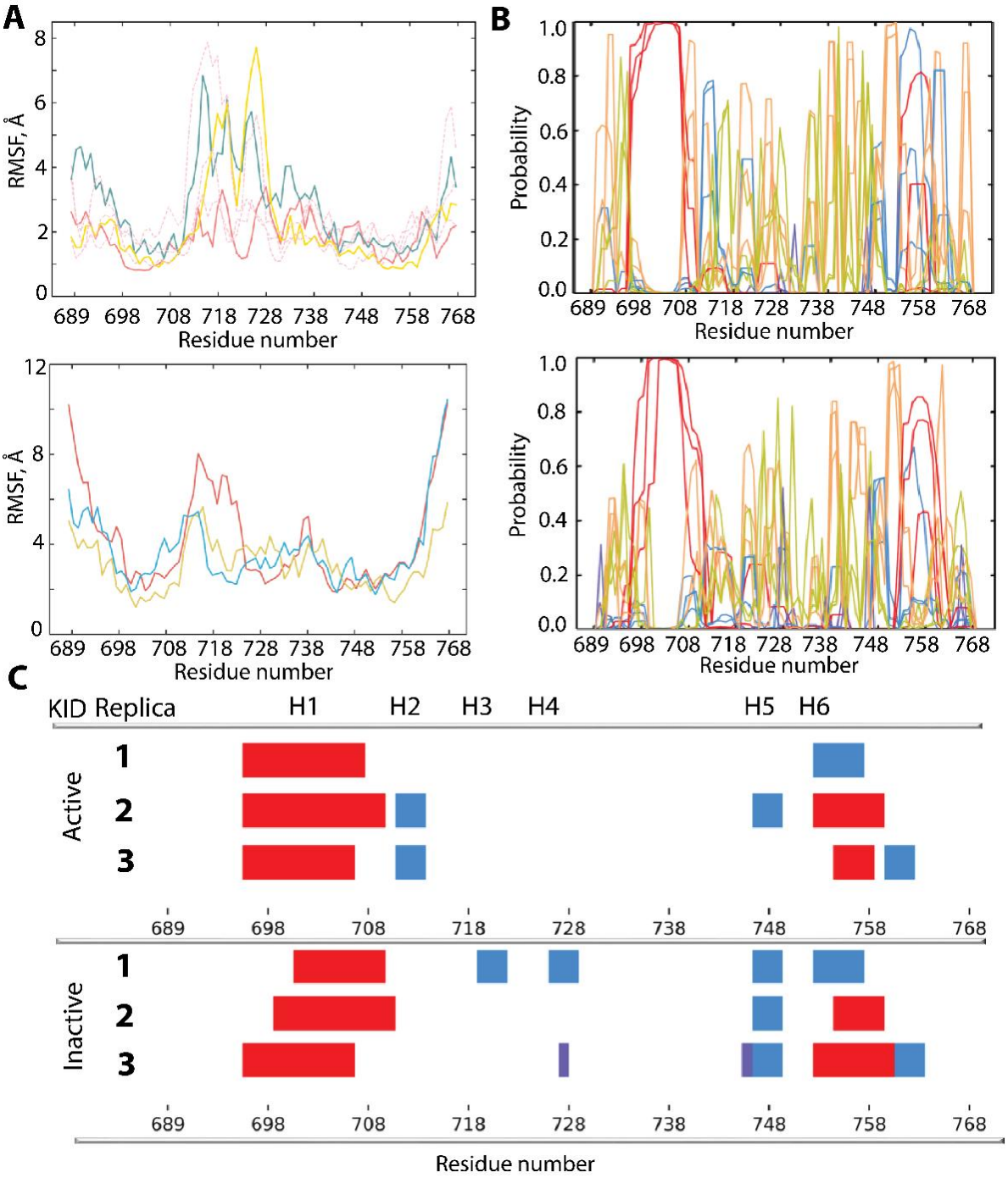


Figure S25 KID of RTK KIT in its active and inactive states. (**A**) RMSFs computed on the Cα atoms for MD conformations (1-3 replicates) of KID from KIT in active (top) and inactive (bottom) states. (**B**) Secondary structure (SS) assignments (DSSP) for KID from KIT in active (top) and inactive (bottom) states during MD simulations. For each residue, the proportion of SS type is given as a percentage of the total simulation time and shown with coloured lines: α-helix in violet, 3_{10}-helix in red, parallel and antiparallel β-sheet in blue and cyan, turn in green. (**C**) The secondary structures, αH- (red), 3_{10}-helices (light blue) and β-strands (violet), assigned for a mean conformation of each MD trajectory (1-3) of KID from KIT in active (top) and inactive (bottom) states.

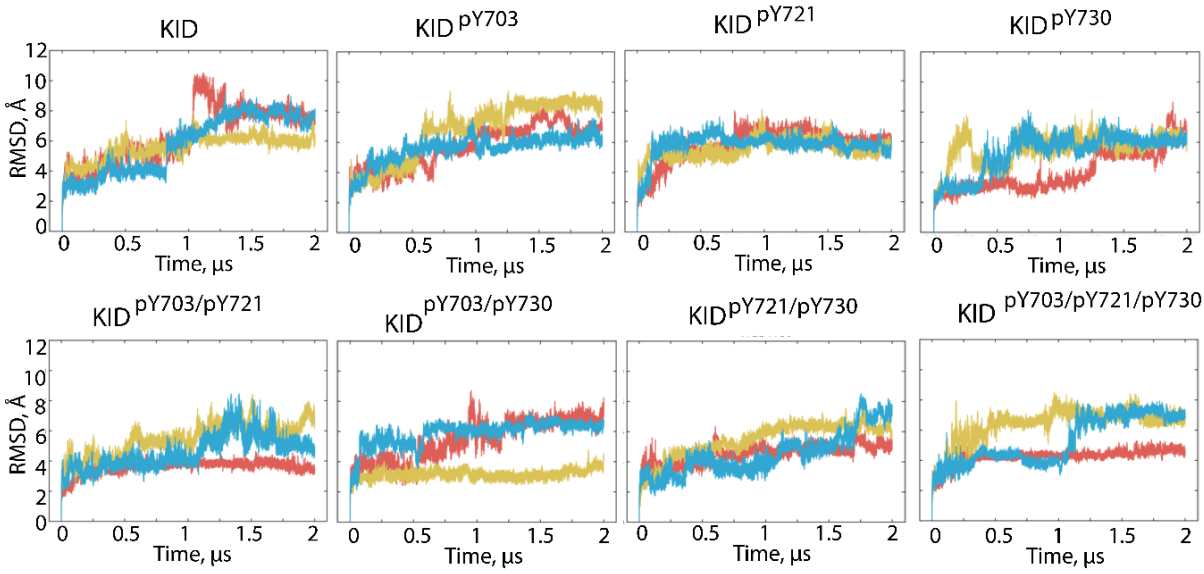


Figure S26 Conventional MD simulations of pKID. RMSDs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns). MD replicas 1-3 are distinguished by colour, light blue, red and yellow.

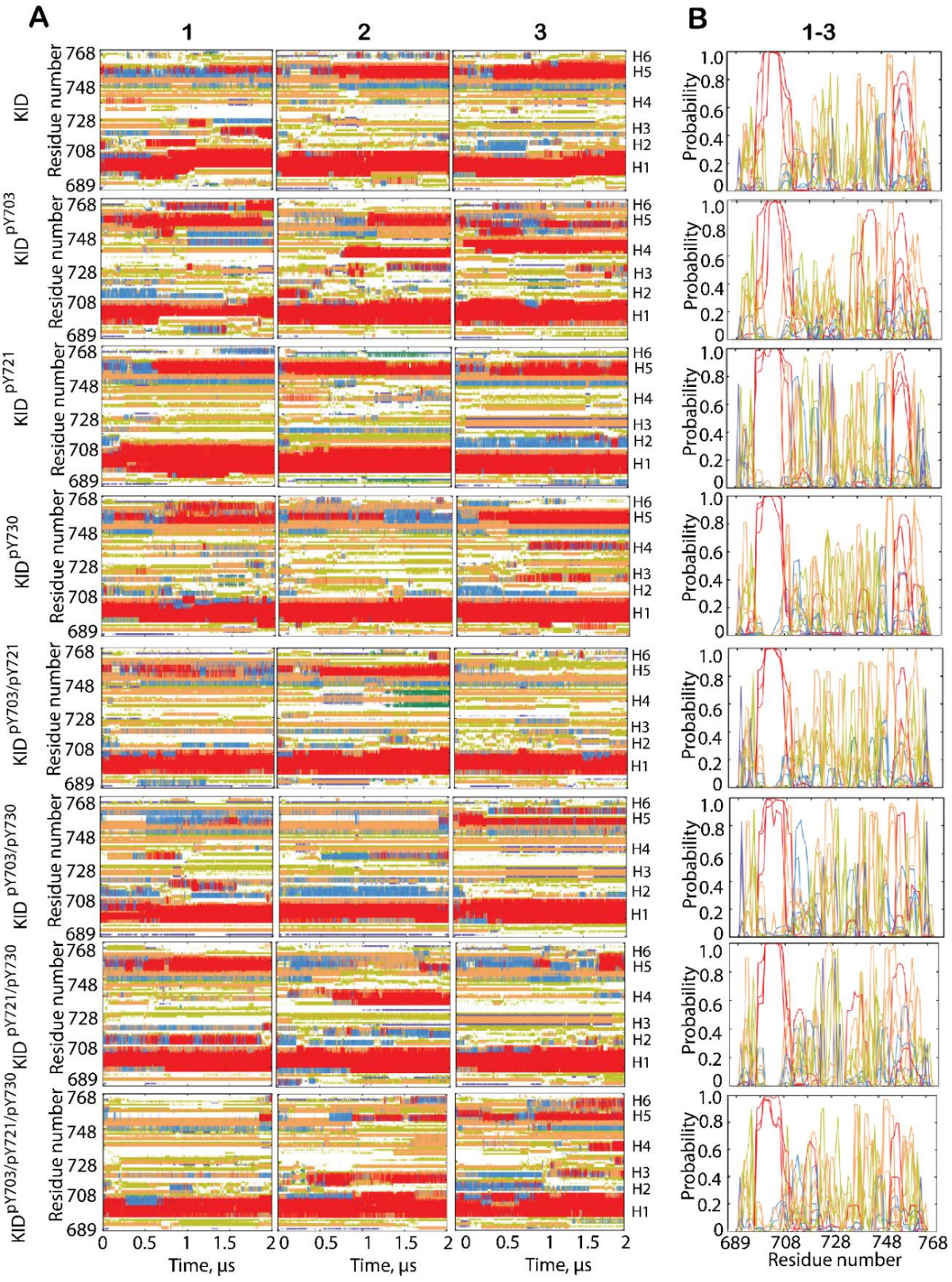


Figure S27 Secondary structure (SS) of unphosphorylated and phosphorylated KID during MD simulation (1-3 replicas), as assigned by DSSP. (**A**) The time-related evolution of SS of each residue with the type-coded SS. (**B**) For each residue, the proportion of SS type is given as a percentage of the total simulation time. (**A**-**B**) Each SS type is colour-coded: α-helix in violet, 310-helix in red, parallel, and antiparallel β-sheet in blue and cyan, turn in green.

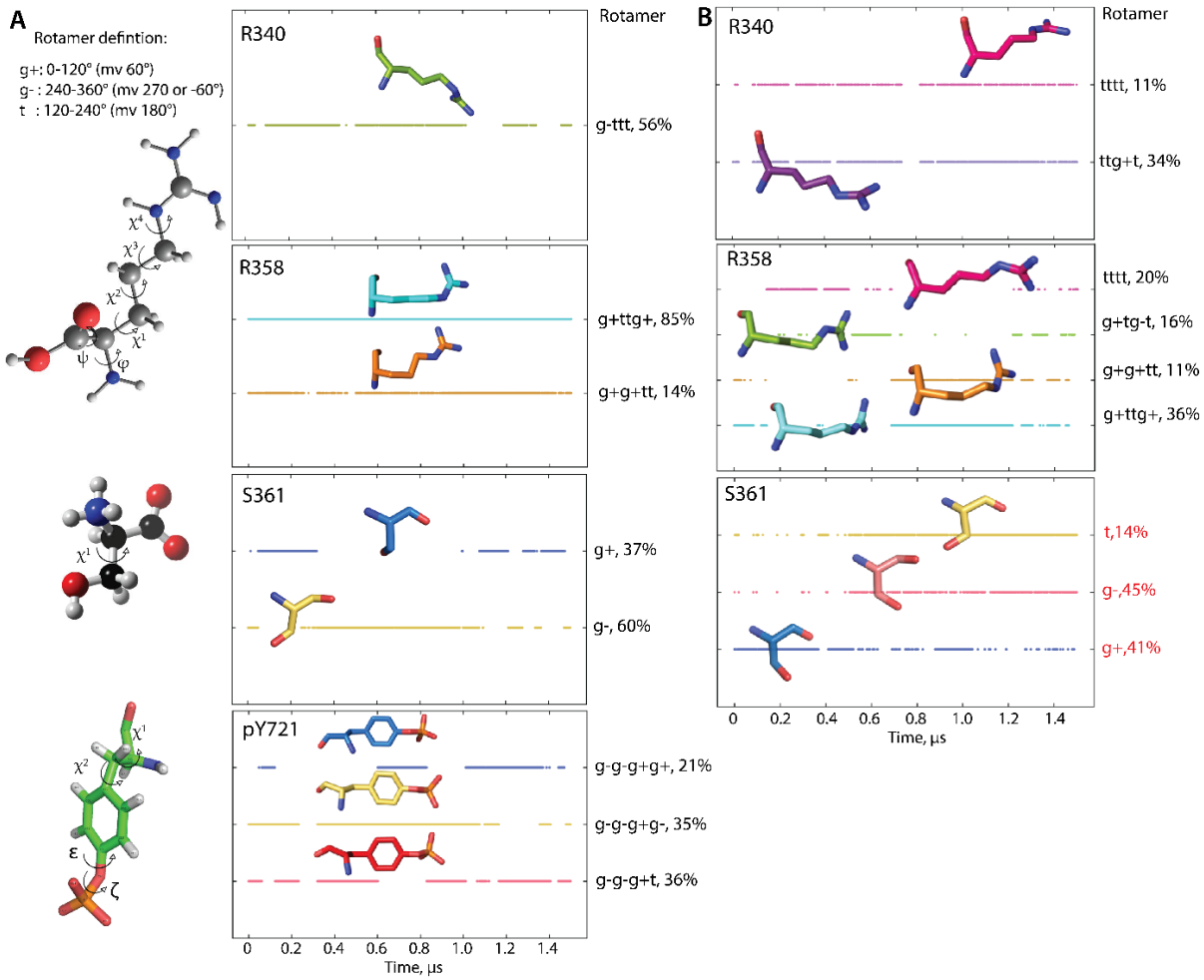


Figure S28 Conformational features of the PI3K SH2 domain in p-*pep*/SH2 complex (**A**) and free-ligand SH2 (**B**). Conformations of R340, R358, S361 and pY721 are described in the IUPAC conformational terms trans (t) and gauche (g), and a population calculated on the concatenated trajectories 1-3.

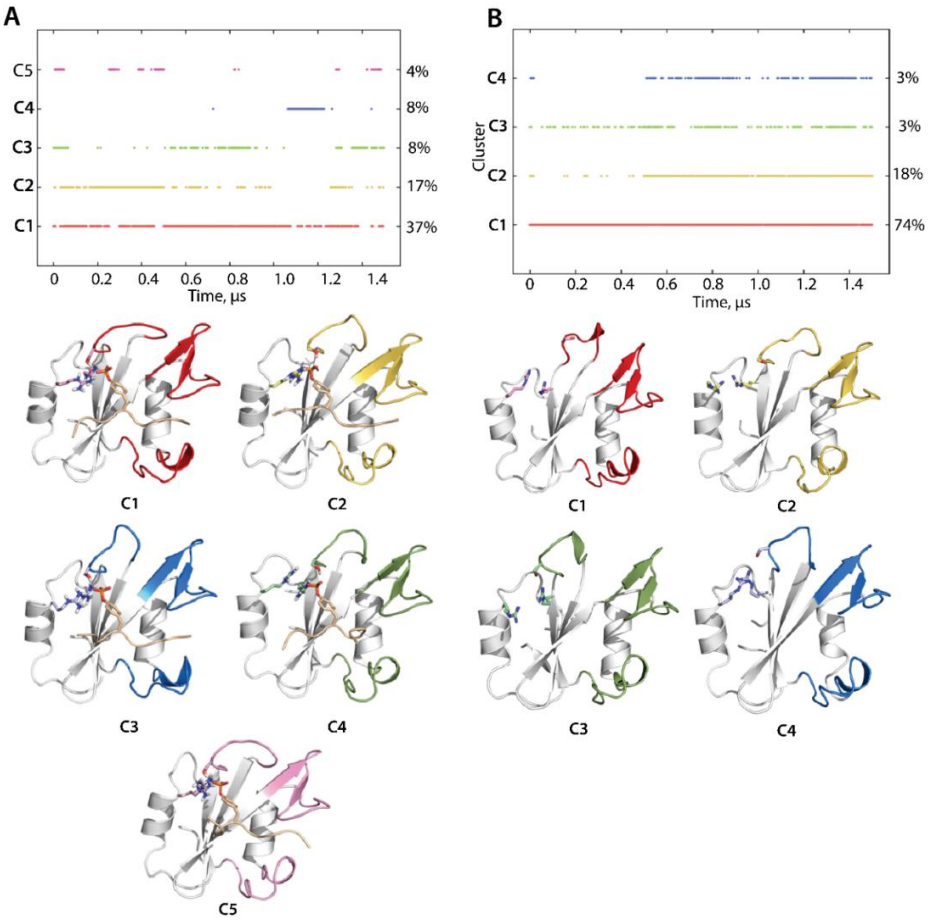


Figure S29 Conformational features of the PI3K SH2 domain in p-pep/SH2 complex (**A**) and free-ligand SH2 (**B**). The SH2 domain clusters and their population. Clustering was based on the RMSD values (cutoff 0.75 Å) after least-squares fitting on the β-sheet core (G353-R358, Y368-R373, N378-F384). Protein is displayed as a grey cartoon with three IDRs, F1, F2 and F3, distinguished by colour corresponding to the respective cluster. Residues R340, R358, S361 and pY721 are shown as sticks.

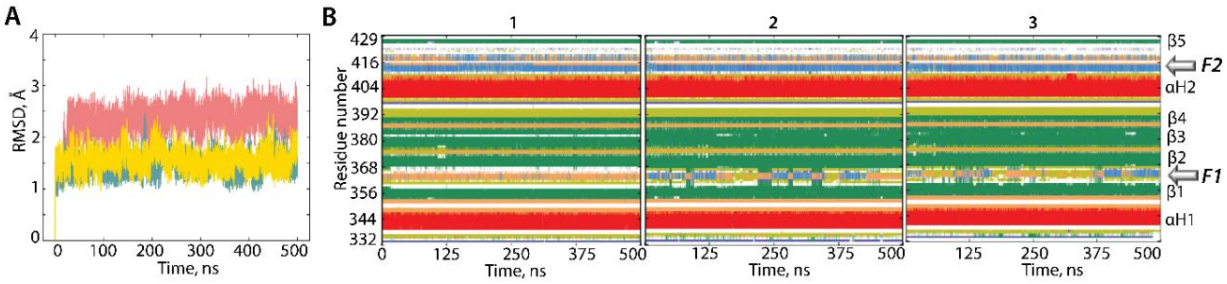


Figure S30 Conventional MD simulations of the free-ligand SH2 domain. (**A**) RMSDs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns). MD replicas 1-3 are distinguished by colour, light blue, red and yellow. (**B**) The time-related evolution of SS of each residue with the type-coded SS: α-helix in violet, 3_{10}-helix in red, parallel, and antiparallel β-sheet in blue and cyan, turn in green.

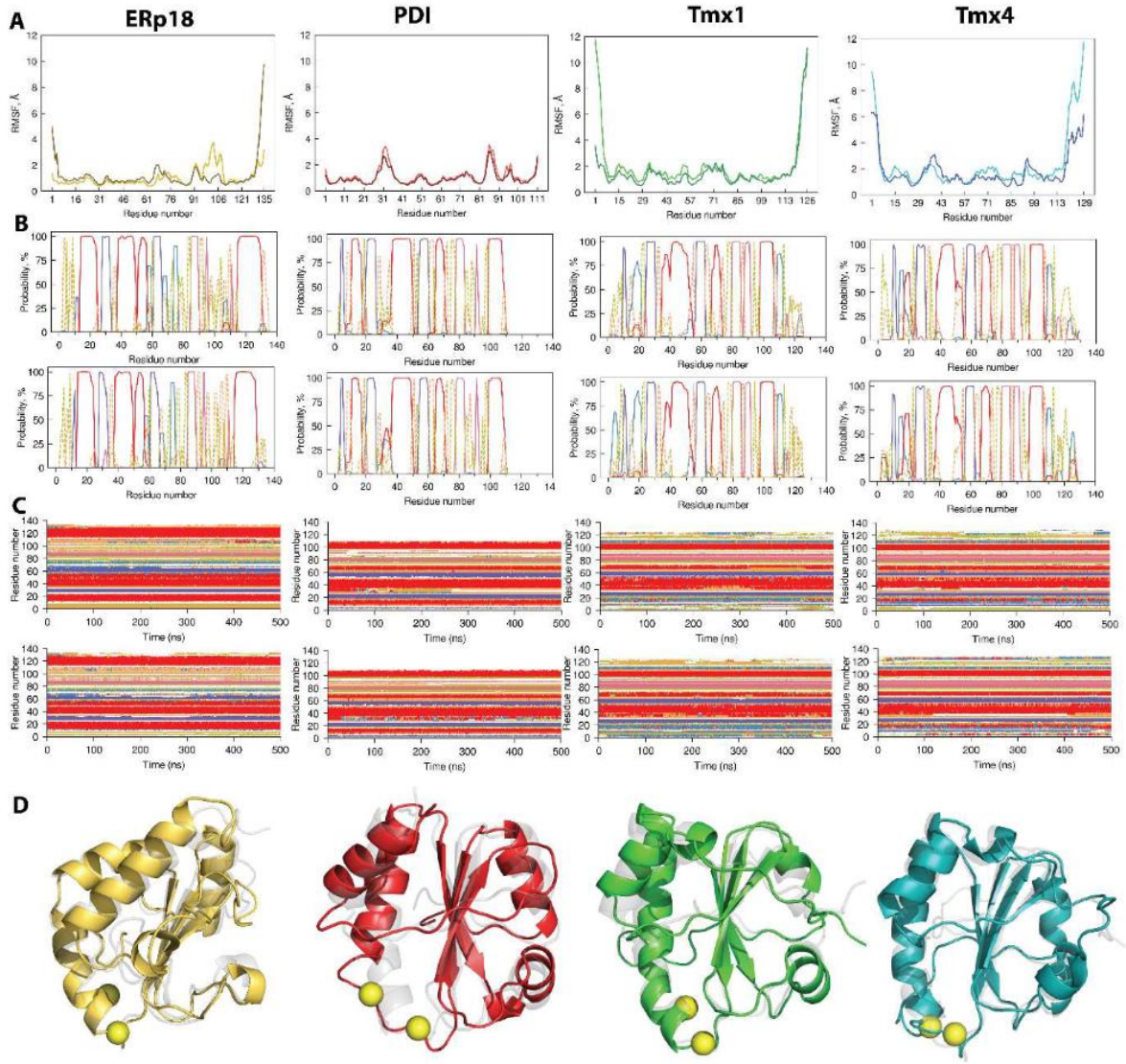


Figure S31 MD simulations of ERp18, PDI, Tmx1 and Tmx4. (**A**) RMSFs computed on the all Cα-atoms for two replica of MD simulations of each protein after fitting on initial conformation. (**B**) Proportion (given as a probability) of every secondary structure type for each residue, as assigned by DSSP. Assignment of the secondary structure type to colors is given as follows: α-helix is in red, 310-helix is in blue, the parallel and antiparallel strands are in green and violet respectively; turn is in orange and bend is in dark yellow. (**C**) The time-depended evolution of the secondary structure of each residue as assigned by DSSP: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow. In (A-C) the numbering of residue in each Trx protein is arbitrary and started from the first amino acid in a model. (**D**) The mean conformation of each protein, calculated for MD trajectory 1, is superimposed on its experimentally determined structures of ERP18, PDI and Tmx1 (in grey), and on the homology model of Tmx4. MD conformations of each protein are shown as colored ribbons – ERp18 in yellow, PDI in red, Tmx1 in green and Tmx4 in blue – with cysteine residue as yellow balls.

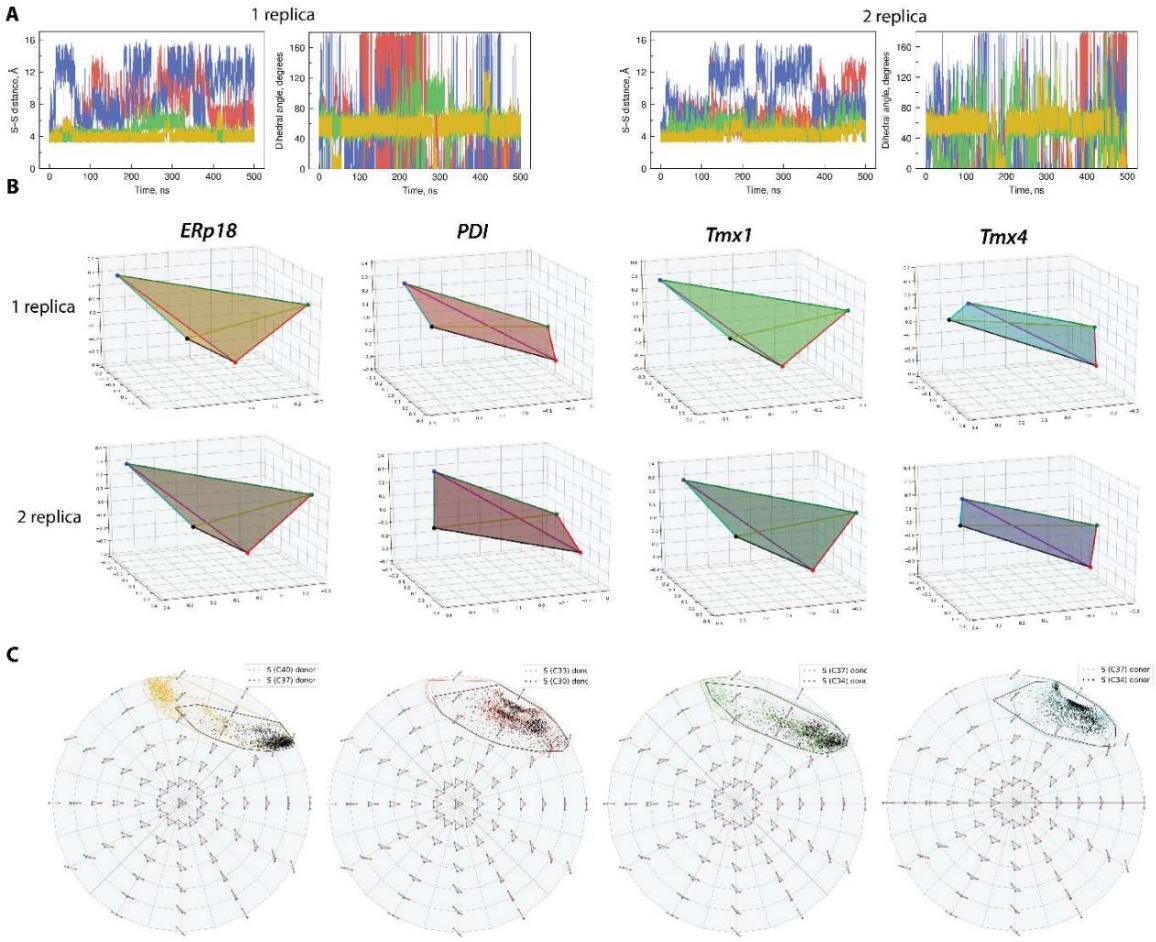


Figure S32 Geometry of CX_1X_2C motif for ERp18, PDI, Tmx1 and Tmx4. (**A**) Geometry of CX1X2C motif in each replica is described by distance S···S′ (left) and dihedral angle (right) determined as an absolute value of the pseudo torsion angle S−Cα(C37)−Cα′(C40)−S′. (**B**) Frechet mean of each replica, computed in Kendall framework. (**C**) Projection of S donor (red) - H(green)···S(blue) triangles on a planar disk, where the S-donor is alternatively on the first (black) or second carbon (color). (A-C) Proteins are distinguished by colour − ERp18 (yellow), PDI (red), Tmx1 (green) and Tmx4 (blue).

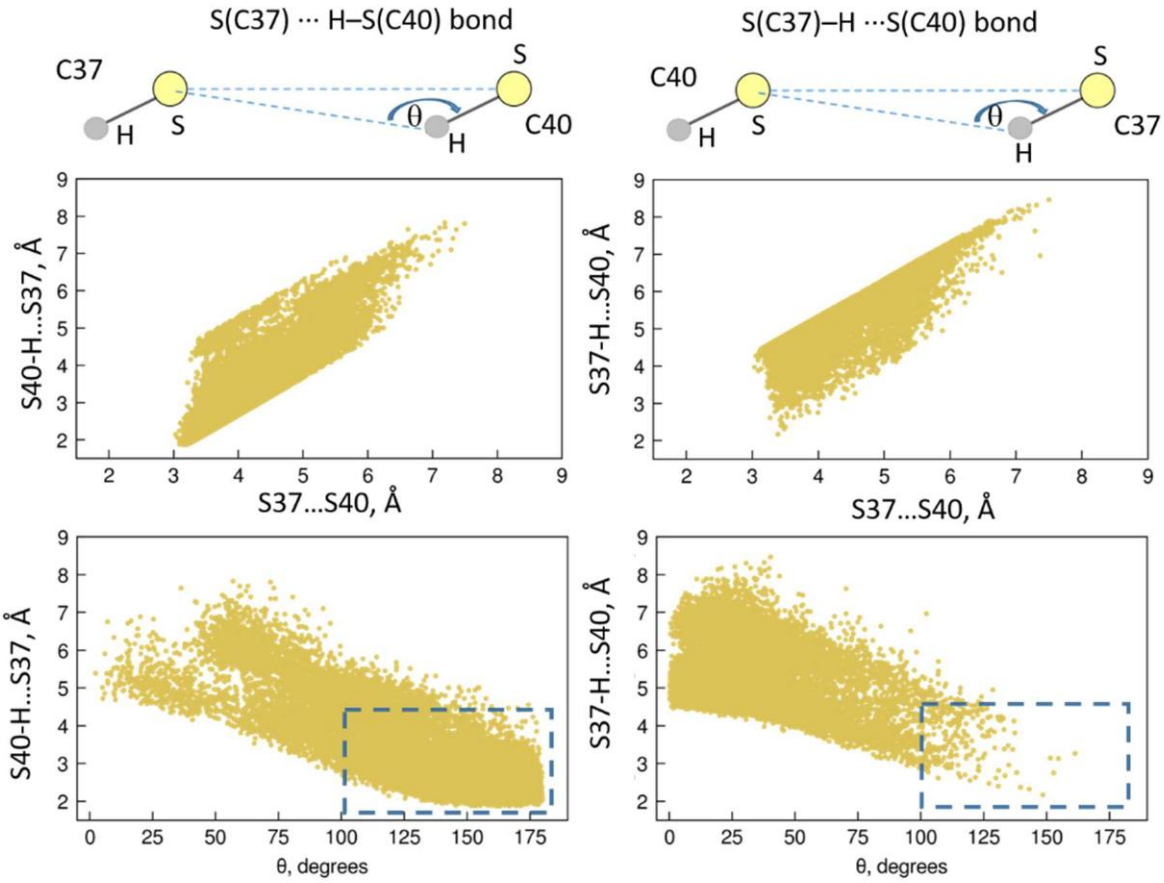


Figure S33 Hydrogen bond in CX_1X_2C fragment from ERp18. Two thiol groups from C40 and C37 areassociated by H-bond in which the sulphur atoms from each cysteine residue are the donor (C40) and acceptor (C37) groups respectively. The H-bond is characterised by the mutually correlated parameters, the interatomic distances S···S and the pseudo-covalent angle at H-atom (SH···S). The regions delimited by dashed lines correspond to the H-bond interaction.

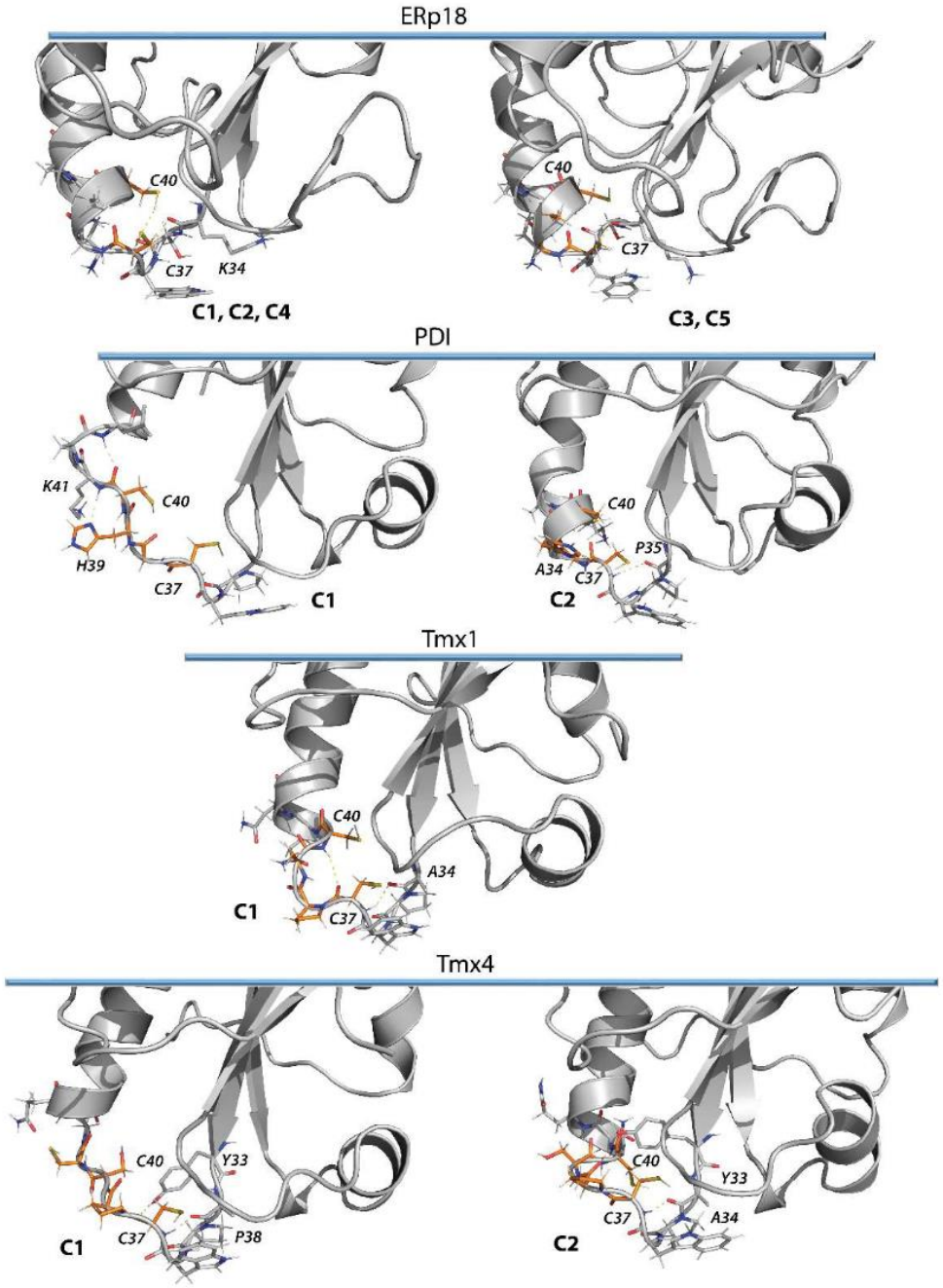


Figure S34 The intra-molecular interactions in the CX_1X_2C motif with neighbour residues was characterised for the conformations regrouped into clusters (cut-off of 2.0 Å) from the concatenated MD trajectory. To regroup the most similar MD conformations and to measure the structural differences between them, after removing the residues with the largest fluctuations from the N- and C-terminals (if its needed), ensemble-based clustering[267] was applied to the concatenated trajectory of each protein. Using the same cut-off value (2.0 Å), results in only one unique cluster encompassing 99% of the MD conformations for Tmx1, two clusters with populations of 94 and 4% in Tmx4, three clusters populated with 73, 14 and 12% of the conformations in PDI, and a large number of clusters with lower populations (38, 17, 15, 13, 5 %) in ERp18. The content of each statistically significant cluster was used for analysis of intramolecular interactions that stabilised the CX_1X_2C motif. Proteins are shown as grey ribbons along with the CX_1X_2C motif and neighbouring residues as sticks.

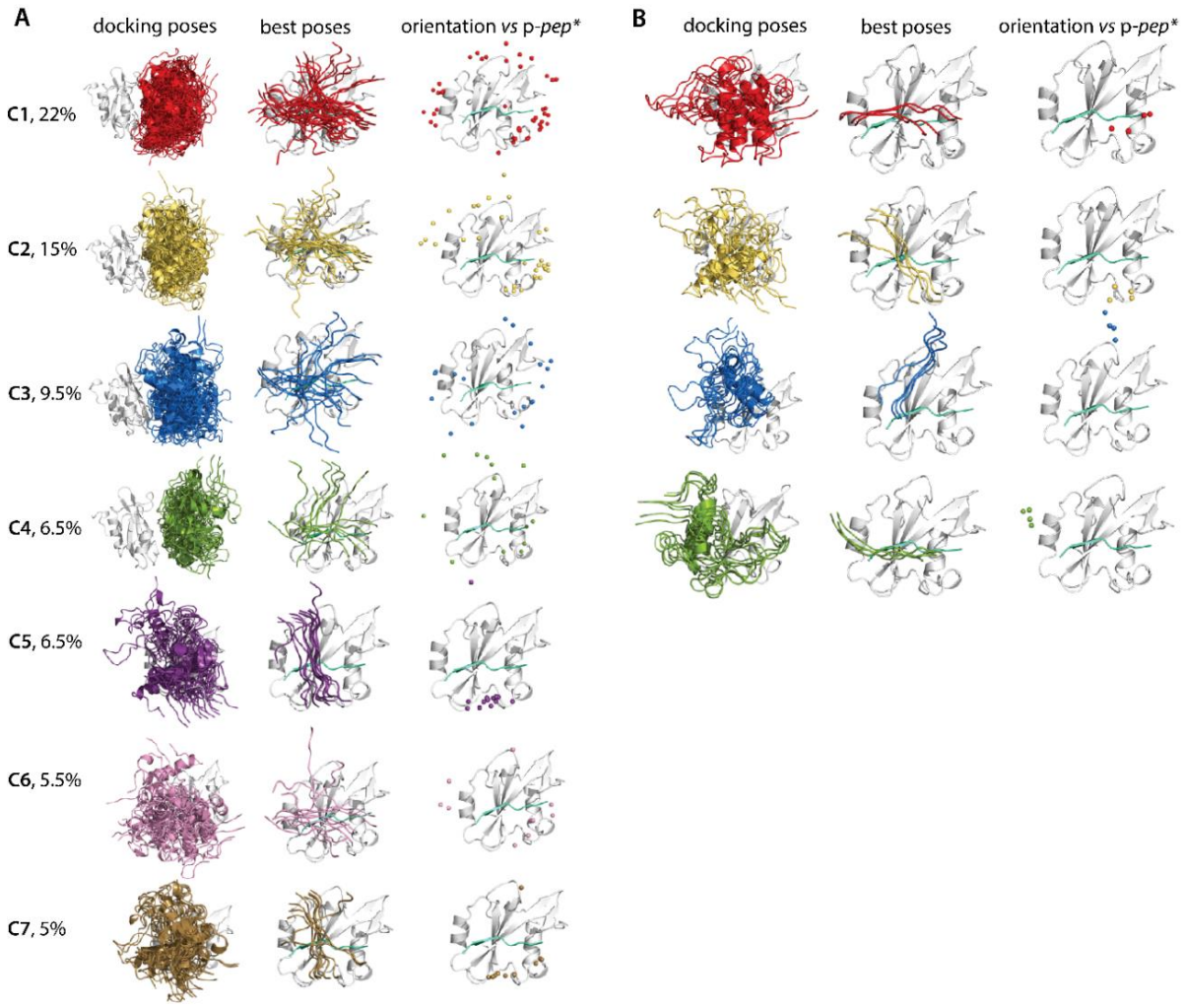


Figure S35 The protein-protein docking of KID onto SH2 performed with High Ambiguity Driven DOCKing (HADDOCK) using an information-driven method. KID is a model (conformation having an excellent structural/conformational similarity of its TNEYMDMK fragment with the p-pep from the empirical structure 2IUH), and SH2 is the crystallographic structure (M1). (**A**) Docking poses selected for refinement are distributed into seven clusters (left), showing the best cluster poses (middle) and orientation of the KID p-pep concerning the p-pep (in cyan) from structure 2IUH (right). (**B**) Docking poses after refinement (left); superimposition of the top 4 solutions (middle) and the p-pep orientation (right). (**A**-**B**) The target is a grey cartoon; the ligand (KID, left column) is coloured per cluster. To simplify the poses visualisation, KID was presented only by its p-pep in the best solution (middle column), and the N-terminal of p-pep is presented as a ball to distinguish the p-pep orientation (right column).

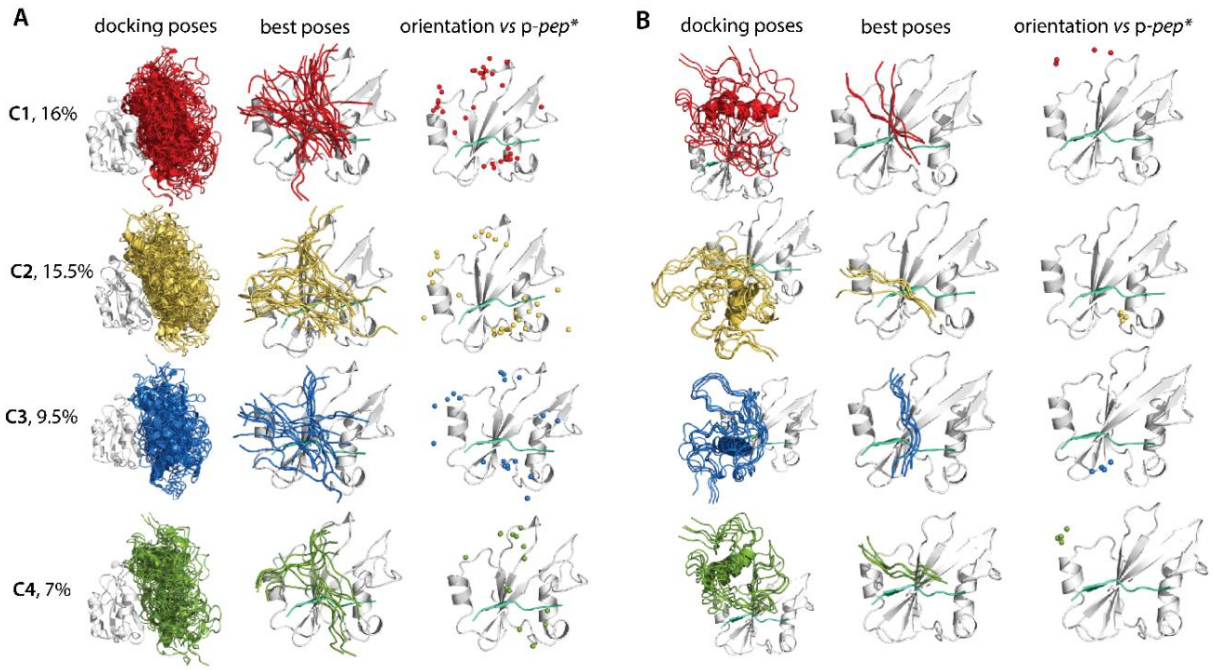


Figure S36 The protein-protein docking of KID onto SH2 performed with HADDOCK using an information-driven method. KID is a model (conformation having an excellent structural/conformational similarity of its TNEYMDMK fragment with the p-pep from the empirical structure 2IUH), and SH2 is a model (conformation taken from cluster C1, M2). (**A**) Docking poses selected for refinement are distributed into four clusters (left), showing the best cluster poses (middle) and orientation of the KID p-pep concerning the p-pep (in cyan) from structure 2IUH (right). (**B**) Docking poses after refinement (left); superimposition of the top 4 solutions (middle) and the p-pep orientation (right). (**A**-**B**) The target is a grey cartoon; the ligand (KID, left column) is coloured per cluster. To simplify the poses visualisation, KID was presented only by its p-pep in the best solution (middle column), and the N-terminal of p-pep is presented as a ball to distinguish the p-pep orientation (right column).

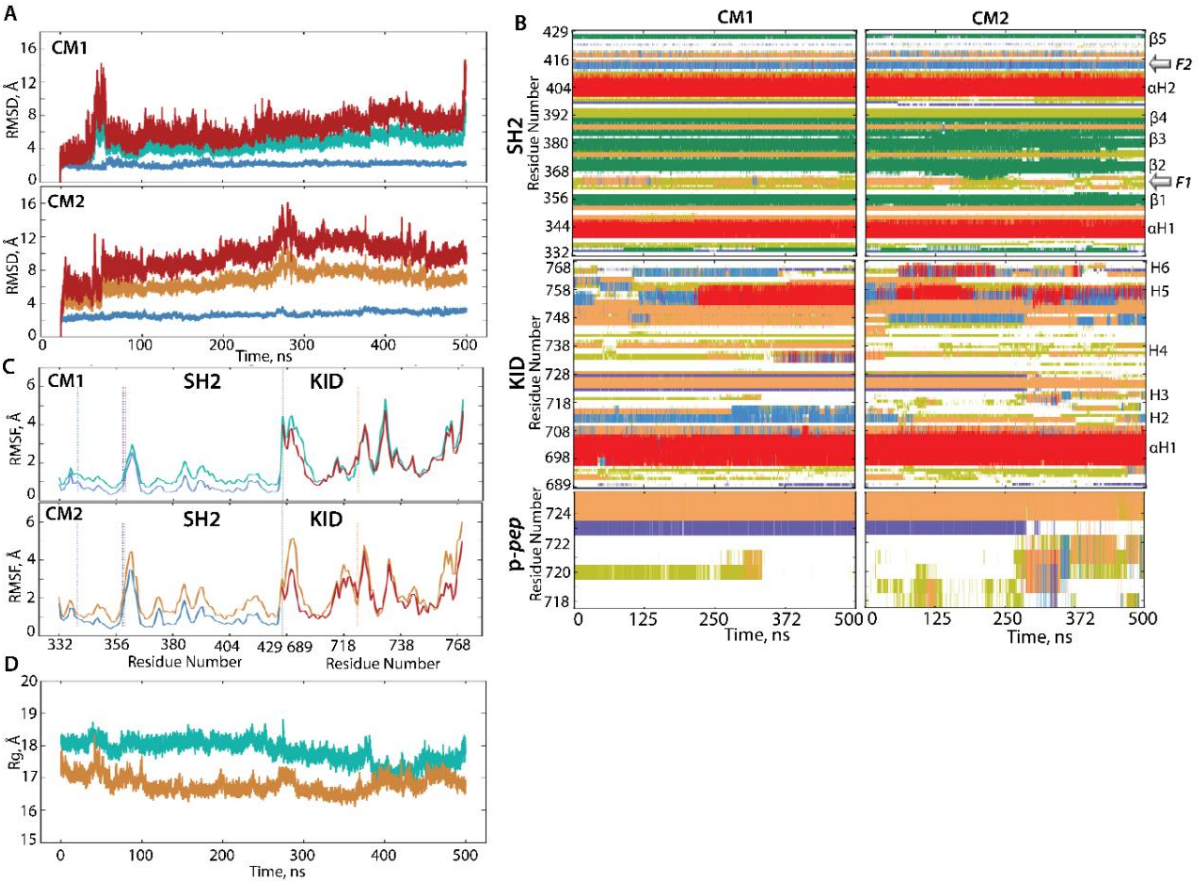


Figure S37 GaMD simulations of CM1 and CM2 models of KID/SH2 molecular complex. (**A**) RMSDs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns) of the SH2 β-sheet core structure. Proteins are distinguished by colour, KID in red (CM1 and CM2), SH2 in blue (CM1 and CM2), and complex in teal (CM1) and orange (CM2). (**B**) The time-related evolution of secondary structure (SS) of each residue in SH2 (top), KID (middle) and p-pep fragment of KID (bottom) with the type-coded SS: α-helix in violet, 310-helix in red, parallel, and antiparallel β-sheet in blue and cyan, turn in green. (**C**) RMSFs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns) of each domain separately. (**D**) Radius of gyration, Rg, computed for each model, CM1 and CM2. Proteins are distinguished by colour, KID in red (CM1 and CM2), SH2 in blue (CM1 and CM2), and complex in teal (CM1) and orange (CM2).

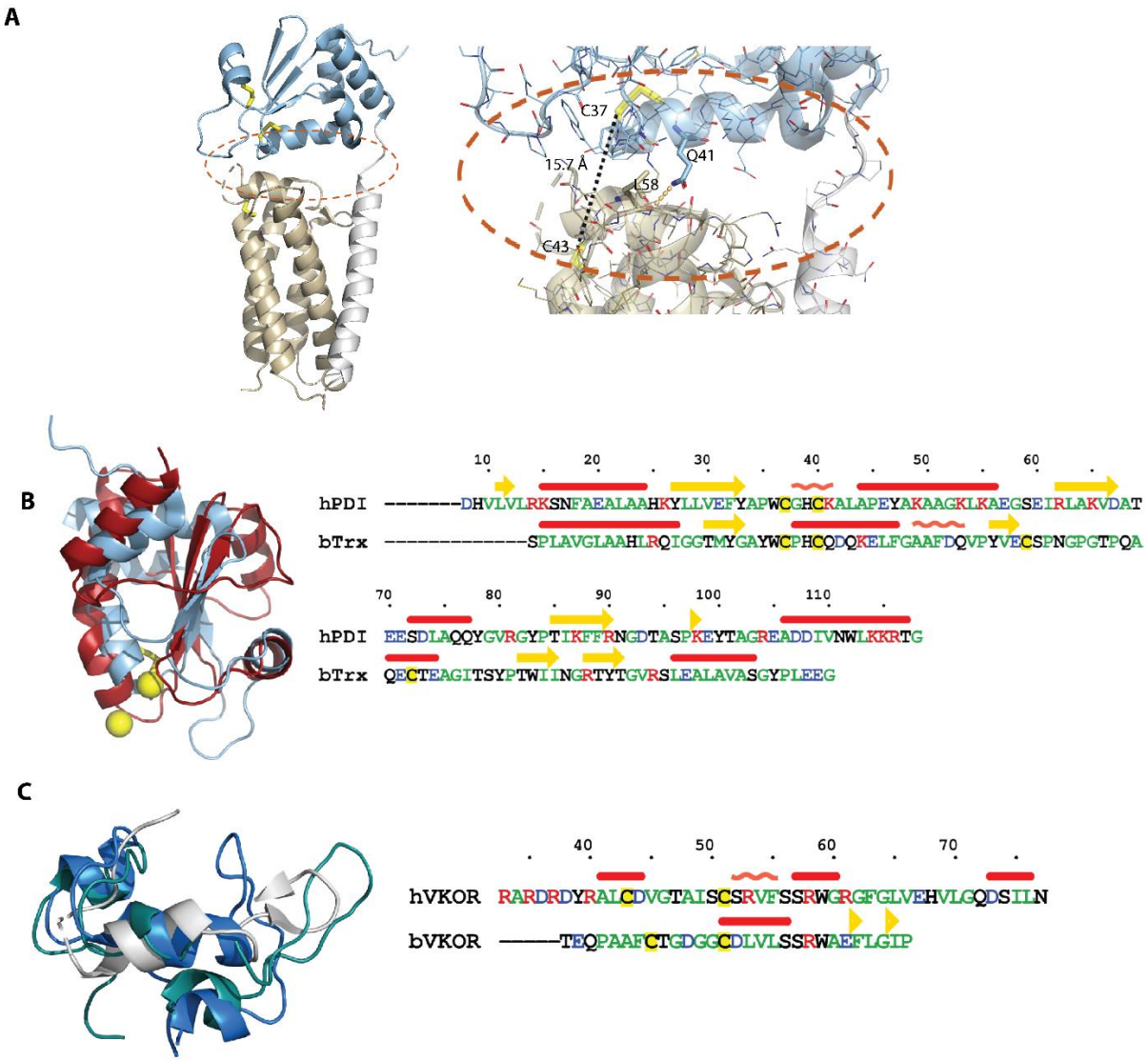


Figure S38 Comparison of the Trx-domain and L-loop from bVKOR with the PDI and L-loop from hVKORC1. (**A**) Structure of bVKOR (PDB ID: 4NV5) and zooming on the interface region (encircled). Distance between the sulphur atoms from C37 (Trx) and from C43 (L-loop) and the H-bond contact between two domains are shown as dashed lines. (**B**) Superposition of the Trx-like domain from bVKOR (blue) and of the human PDI (red) (RMSD value of 6 Å) (left). The pairwise alignment (NEEDLE program) of sequences of the Trx-like domain from bVKOR and of the human PDI (identity/similarity of 15/20%) (right). (**C**) Superposition of the L-loop from bVKOR (grey) and from hVKORC1 with the 'closed' (blue) and 'open' (cyan) conformations. RMSD values between the L-loop from bVROR and from hVKORC1 are 4.5 and 4 Å for the 'closed' and 'open' conformations (left). The pairwise alignment (NEEDLE program) of the L-loop sequences from bVKOR and from the human PDI (identity and similarity values of 15/20% (right).

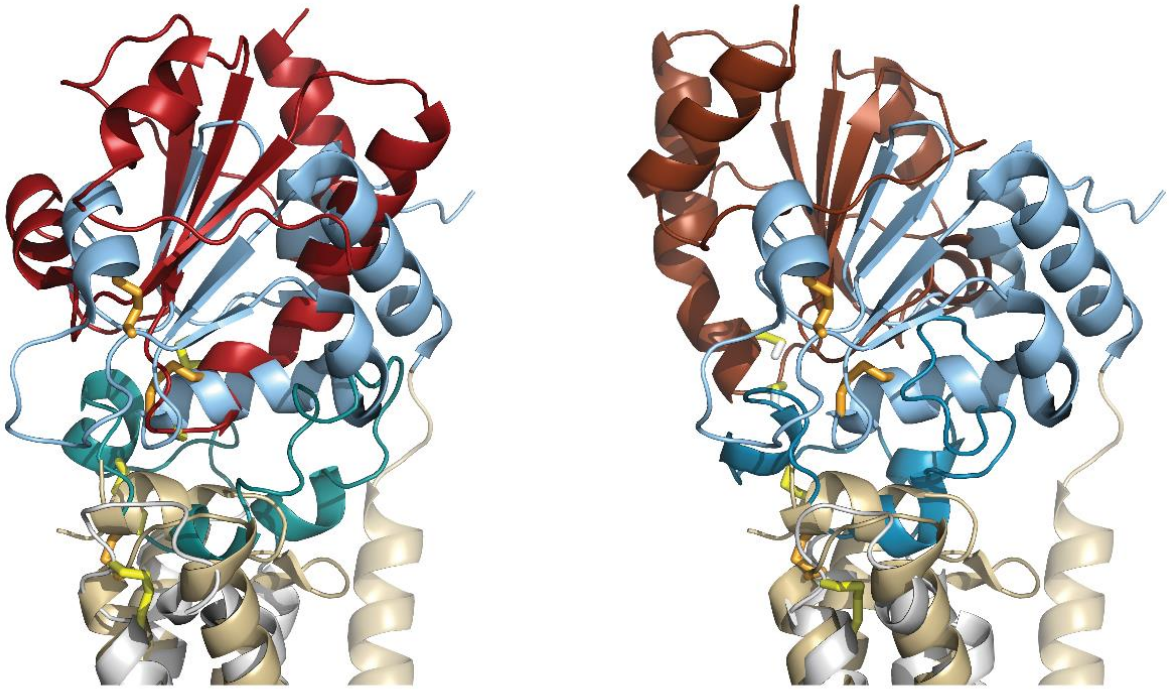


Figure S39 Superposition of Model 1 (left) and of Model 2 (right) of the PDI-hVKORC1 complex into the structure of bVKOR (PDB ID: 4NV5). Model 1 and Model 2 are represented by conformations taken at t = 80 ns of the stepped finite-time MD simulations. Proteins are shown as coloured ribbons. In the models: PDI in red (Model 1) and in brown (Model 2), hVKORC1 in grey with L-loop in cyan and blue. In the X-ray structure of bVKORC1: The Trx-like and VKOR-like domains are shown in light blue and in beige respectively. The cysteine residues are shown as sticks, orange in bVKOR and yellow in Models.

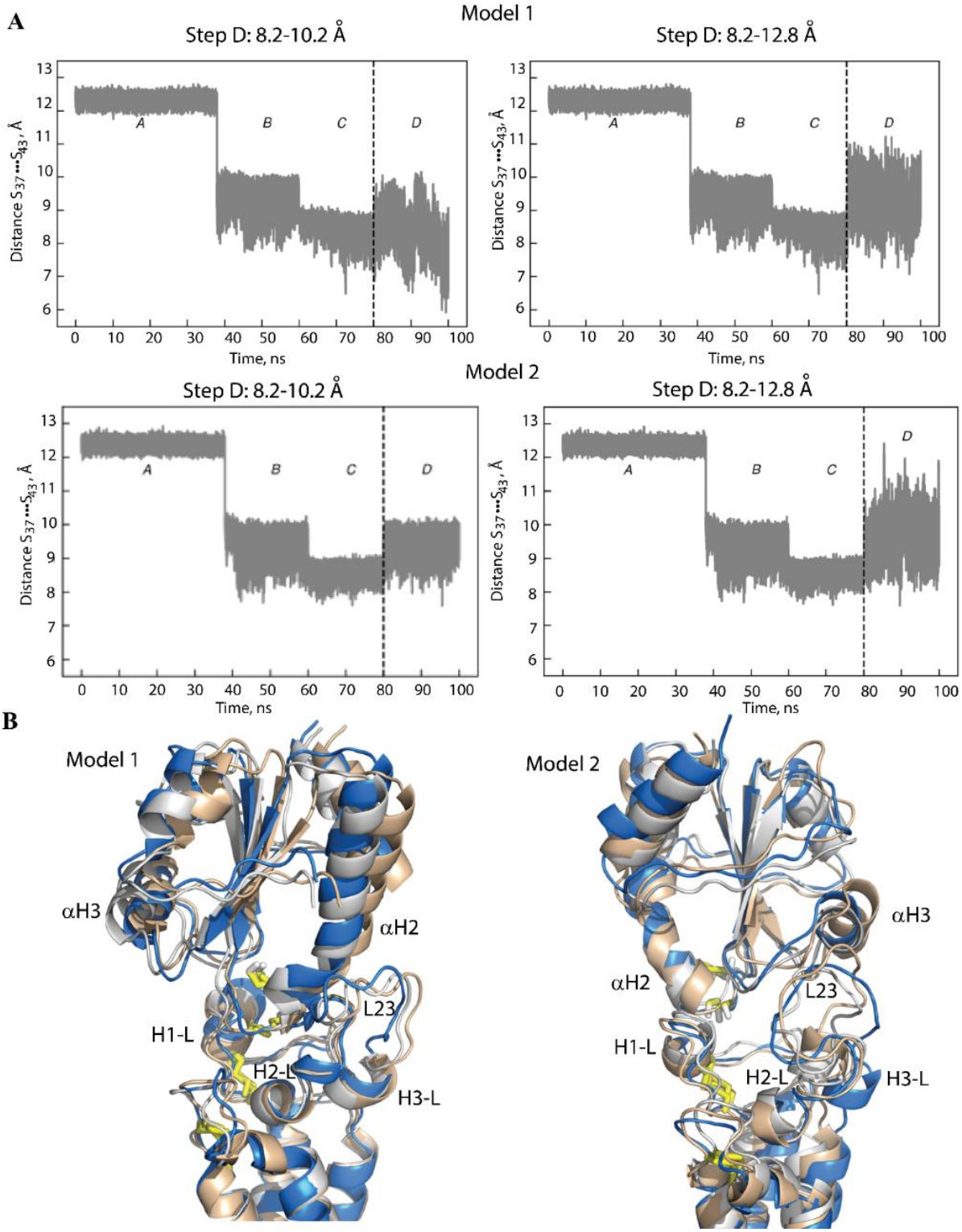


Figure S40 Modelling of the human PDI-VKORC1 complex and their MD simulations. (**A**) The MD simulations of 3D models PDI-VKORC1 complexes, Model 1 and Model 2, were performed with a gradually diminished distance (from 12 to 8 Å) (steps A-C) between the sulphur (S) atoms of C37 from PDI and of C43 from L-loop of hVKORC1, and further (step D) this distance was enlarged to 8.2-12.2 Å and to 8.2-12.8 Å. (**B**) Superimposition of confirmations of Model 1 (left) and Model 2 (right) picked at t=80 (grey) and at 100 ns with S···S distance of 8.2-10.2 Å (beige) and of 8.2-12.8 Å (blue). Proteins are depicted as ribbons. The reference fragments are labelled.

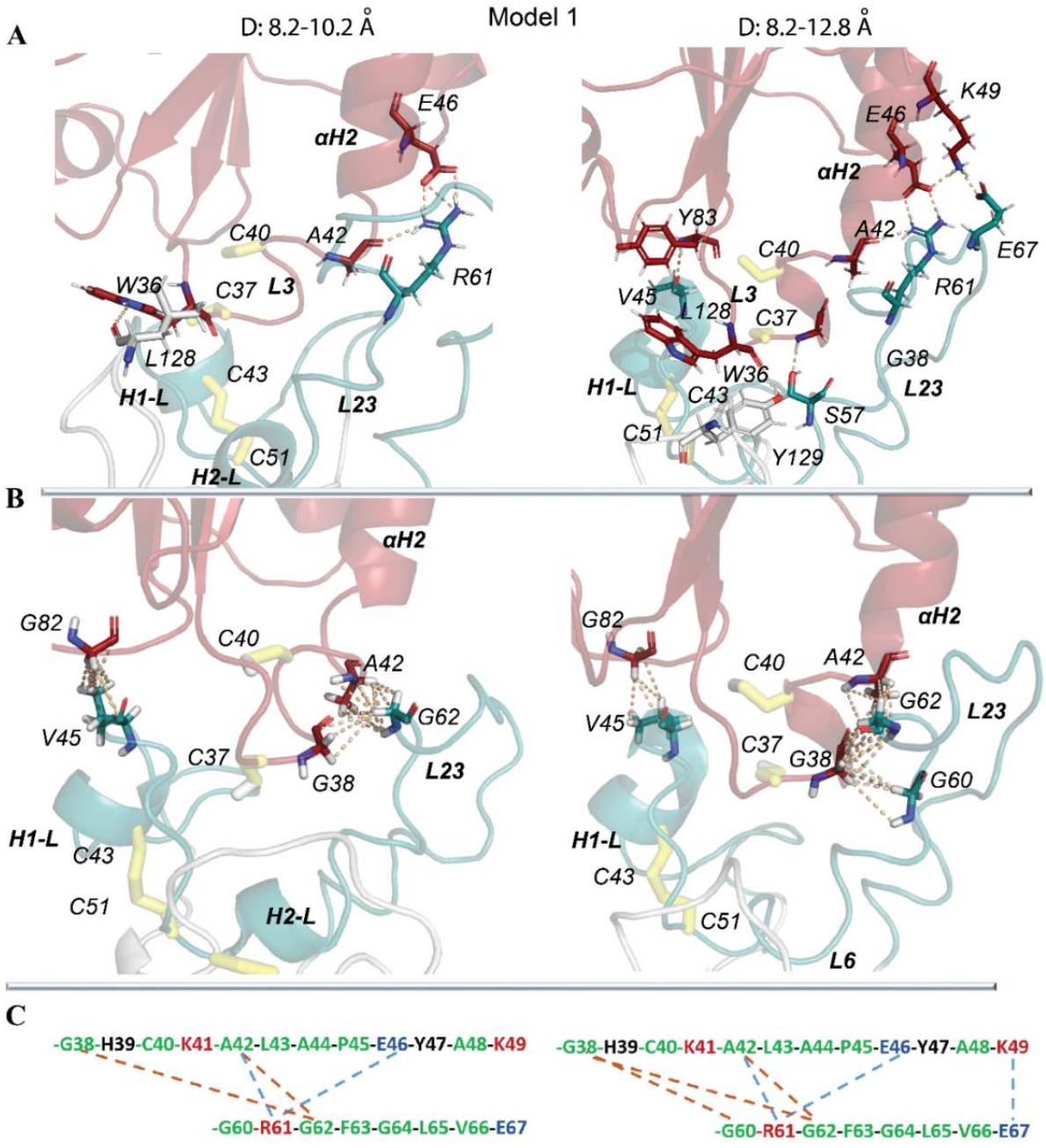


Figure S41 Intermolecular contacts at the interface between PDI and hVKORC1 in the Model 1 of the PDI-hVKORC1 complex. (**A**) The intermolecular H-bonds and (**B**) hydrophobic contacts between PDI and hVKORC1, observed in conformations generated using different 'soft' constraints. The proteins a shown as coloured ribbons, PDI in red and brown, and hVKORC1 in cyan (L-loop) with the interacting residues and thiol groups as sticks. The contacts are indicated by dashed lines, H-bonds in yellow and hydrophobic in salmon. The structural fragments and residues participating in the contacts are labelled. Analysis of intermolecular contacts was performed on conformations taken at t=80 ns. (**C**) A pattern of H-bond (in blue) and hydrophobic (in orange) contacts between the PDI and hVKORC1 residues. Residues are coloured according to their properties – the positively and negatively charged residues are in red and blue respectively, the hydrophobic residues are in green, the polar and amphipathic residues are in black.

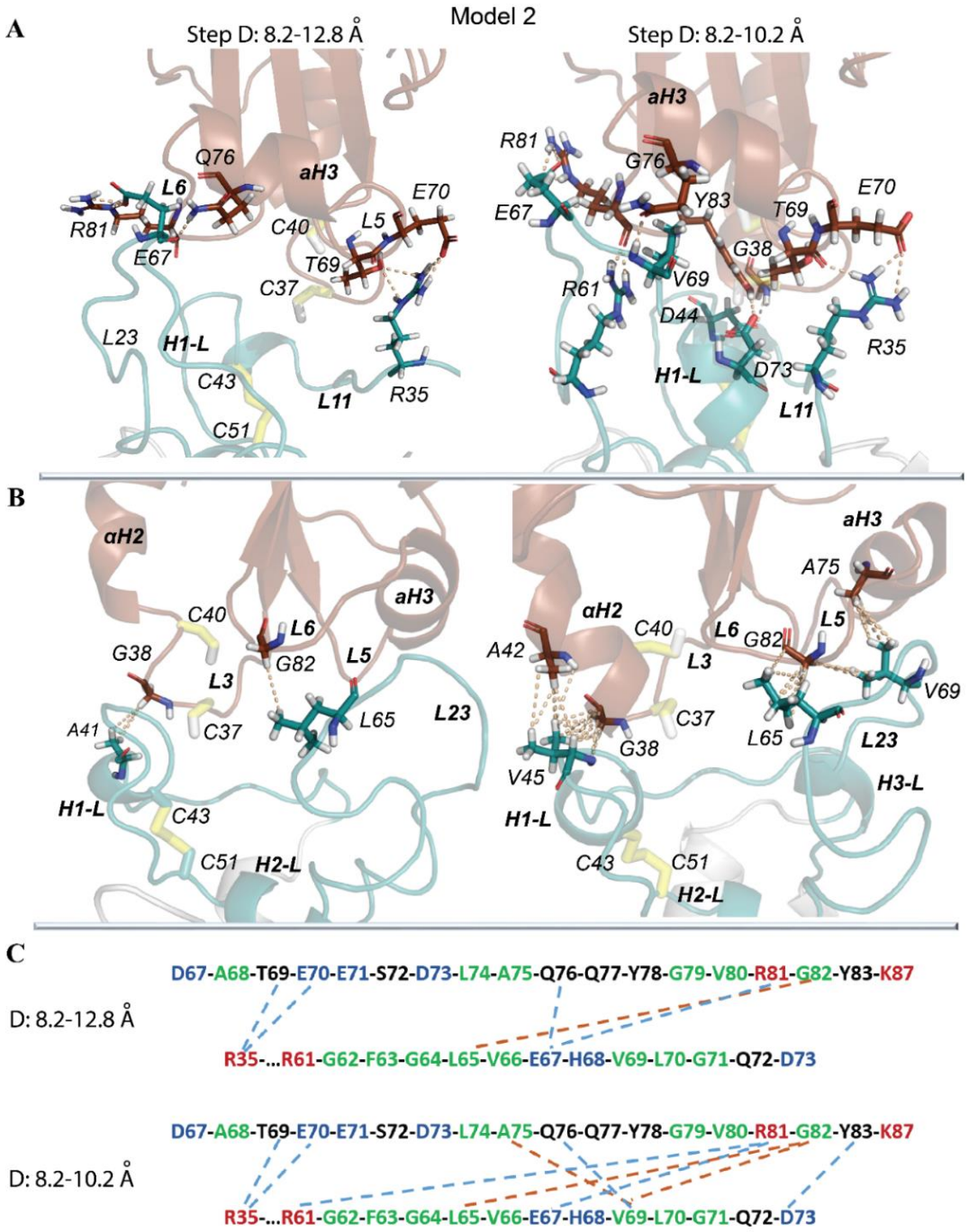


Figure S42 Intermolecular contacts at the interface between PDI and hVKORC1 in the Model 2 of the PDI-hVKORC1 complex. (**A**) The intermolecular H-bonds and (**B**) hydrophobic contacts between PDI and hVKORC1, observed in conformations generated using different 'soft' constraints. The proteins a shown as coloured ribbons, PDI in red and brown, and hVKORC1 in cyan (L-loop) with the interacting residues and thiol groups as sticks. The contacts are indicated by dashed lines, H-bonds in yellow and hydrophobic in salmon. The structural fragments and residues participating in the contacts are labelled. Analysis of intermolecular contacts was performed on conformations taken at t=80 ns. (**C**) A pattern of H-bond (in blue) and hydrophobic (in orange) contacts between the PDI and hVKORC1 residues. Residues are coloured according to their properties – the positively and negatively charged residues are in red and blue respectively, the hydrophobic residues are in green, the polar and amphipathic residues are in black.

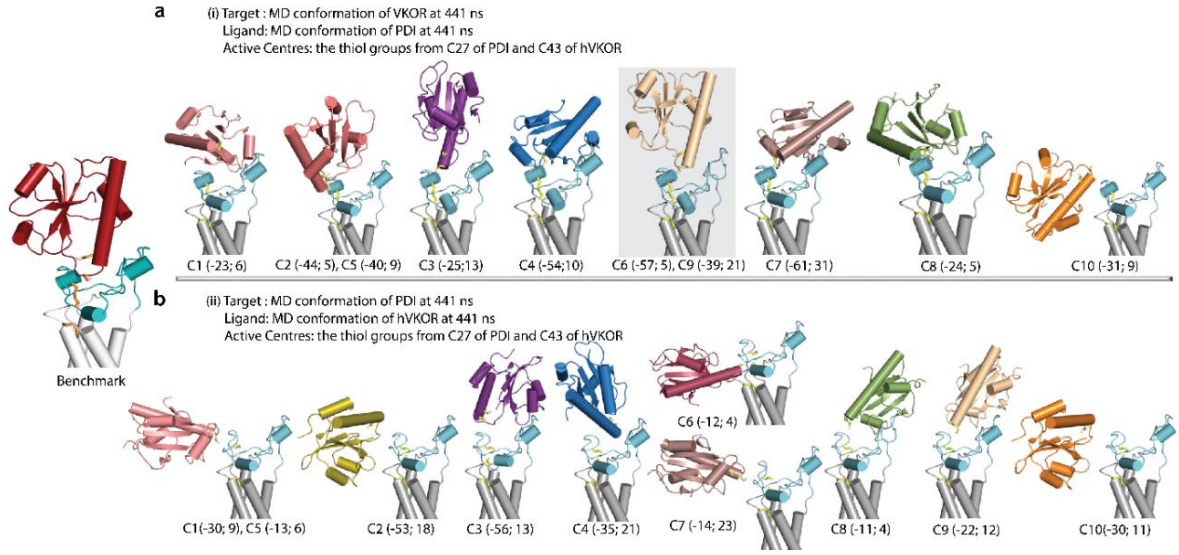


Figure S43 Protein-protein computational docking using an information–driven method on benchmark complex PDI-hVKORC1. Top 10 models produced by HADDOCK for the reference structure (benchmark) using two scenarios for a ligand-target pair, (i) (**a**) and (ii) (**b**). For each cluster, the representative conformations with the HADDOCK score (a.u.) and a cluster population (number of observation) are shown. Protein is shown as a cartoon with helices as cylinders and disulphide bridges in yellow sticks.

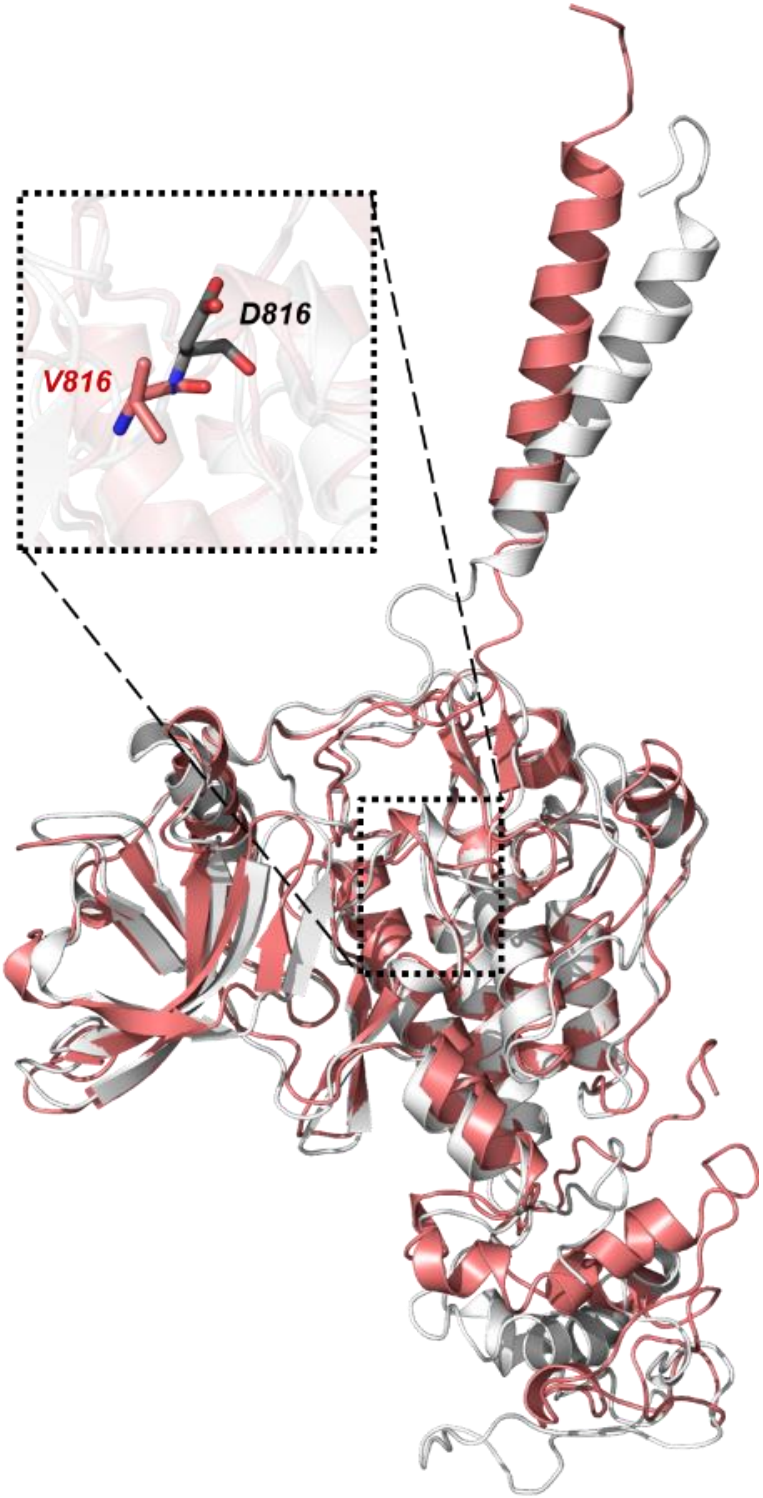


Figure S44 Homology modelling of KITD816V from the KITWT full-length cytoplasmic domain model completed by the transmembrane helix (sequence I516-R946). Superimposition of KIT^{WT} (in grey) taken at t=2μs of cMD simulation, and KIT^{D816V} (in pink) taken at t = 0 μs.

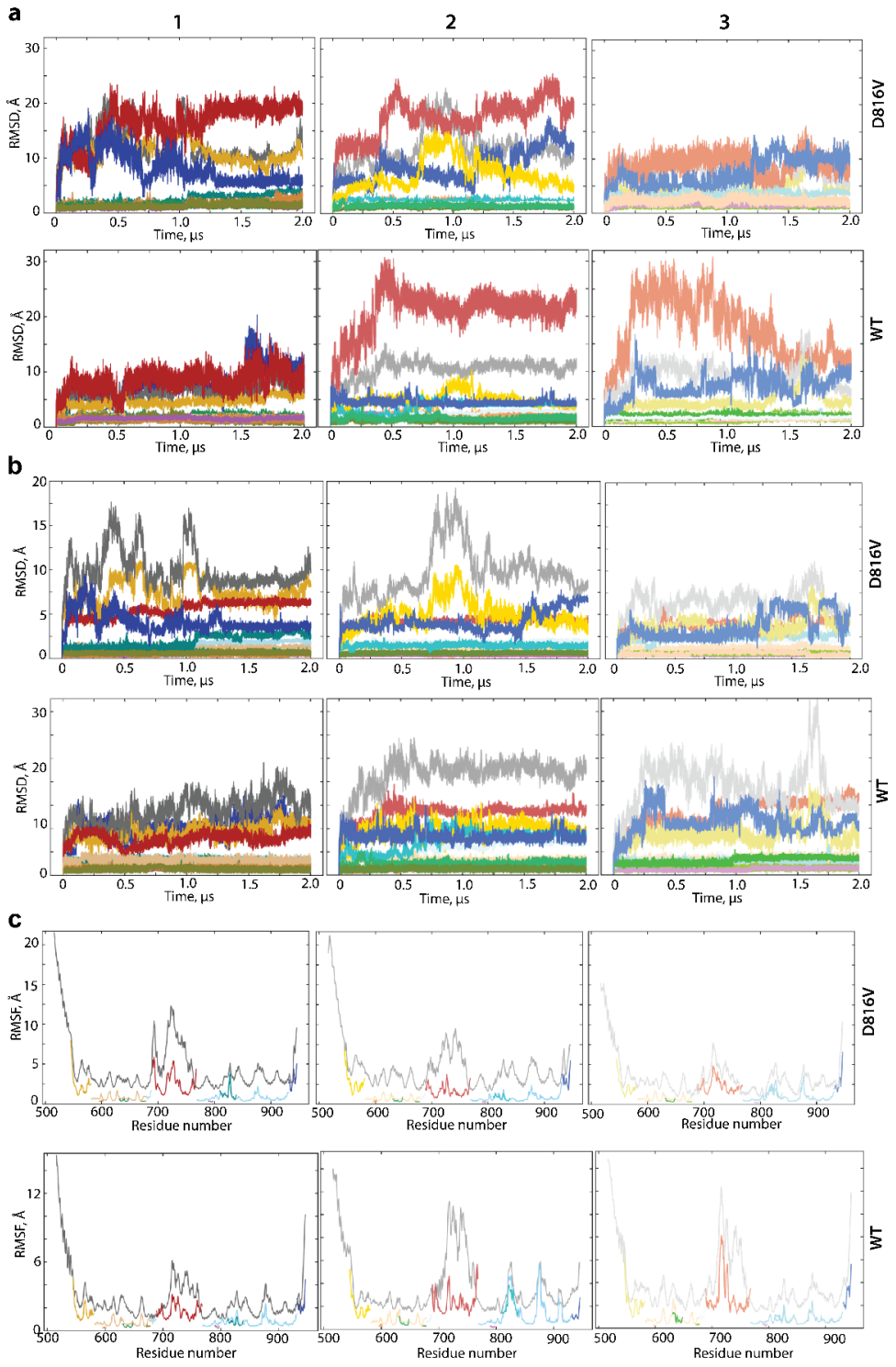


Figure S45 Molecular Dynamics (MD) simulations of the full-length cytoplasmic domain of KIT^{D816V} and KIT^{WT}. (**a**) RMSDs computed on the Cα atoms after fitting on initial conformation (at t = 0 ns) (1-3 replicas, each of 2 µs) of kinase domain and each domain/regions of KIT^{D816V} (a) and KITWT (**b**). (**c**) RMSFs computed on the Cα atoms for cMD conformations after the least-square fitting on the kinase domain initial conformation of KIT^{D816V} (top panel) and KIT^{WT}. (a-c) KIT is in grey, N-lobe in beige, C-lobe in blue, JMR in yellow, P-loop in orange, αC-helix in green, hinge in olive, KID in red, C-loop in magenta, A-loop in teal, C-tail in dark blue for the 1-3 trajectories (replicas) of MD simulations. cMD replicas 1-3 are distinguished by colour tonality, dark, lighter, and light.

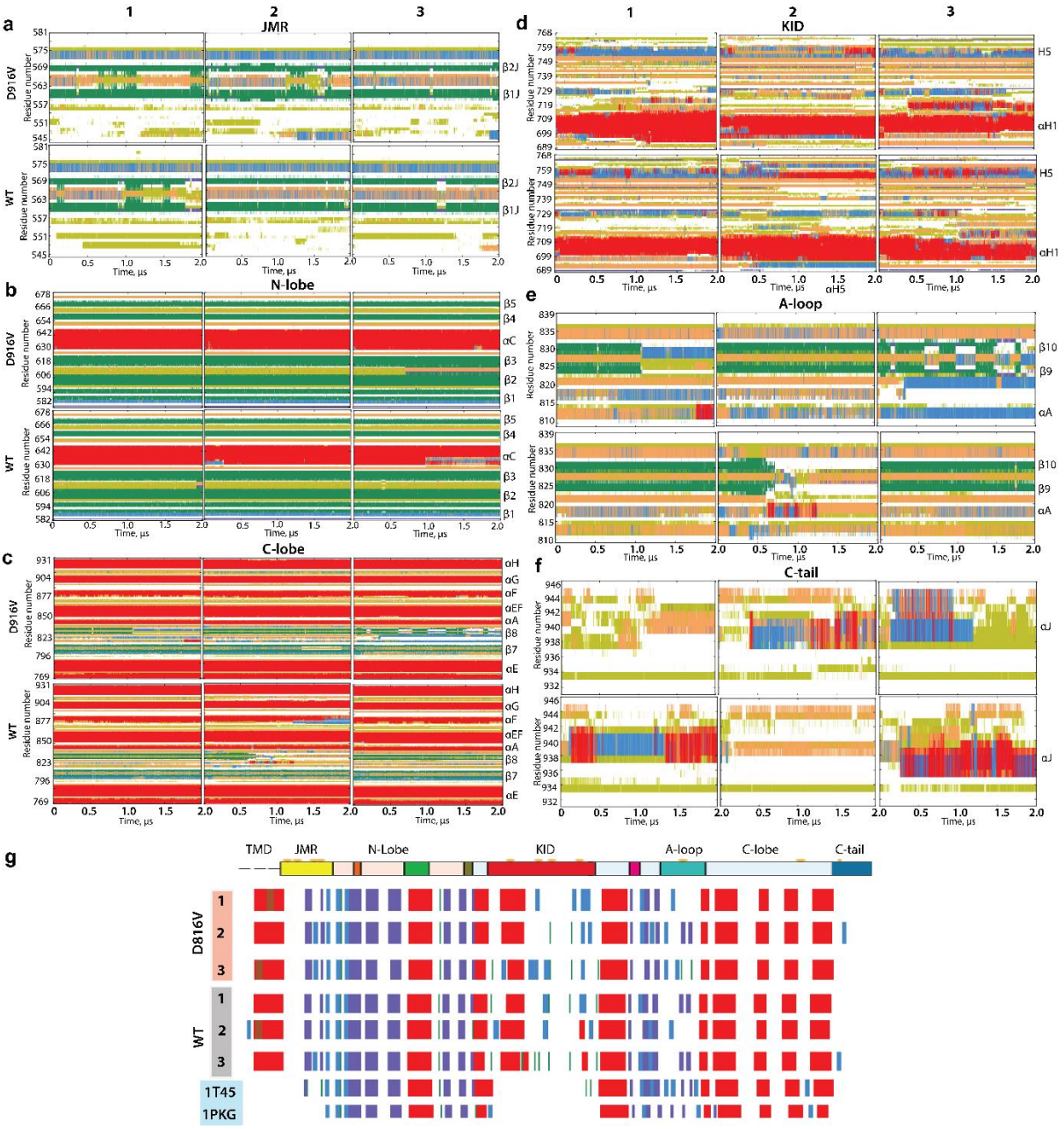


Figure S46 Folding of KITD816V (top panels) and KITWT (bottom panels). (a-f) The time-related evolution of the secondary structures of the entire, full-length KIT and per domain/region, as assigned by the define secondary structure of proteins (DSSP) method: α-helices in red, 3_{10}-helices in blue, parallel β strands in green, antiparallel β strands in dark blue, turns in orange, and bends in dark yellow. The three cMD replicas (1–3) were analysed individually. (**g**) The secondary structures—αH- (red), 3_{10}-helices (light blue), and β-strands (dark blue)—assigned for a mean conformation of every MD trajectory (1–3) of KIT^{D816V}, KIT^{WT} and the crystallographic structures 1T45 (inactive KIT) and 1PKG (active KIT). (**D**) The secondary structures—αH- (red) and β-strands (dark blue)—assigned on the mean conformation of the concateneted trajectory.

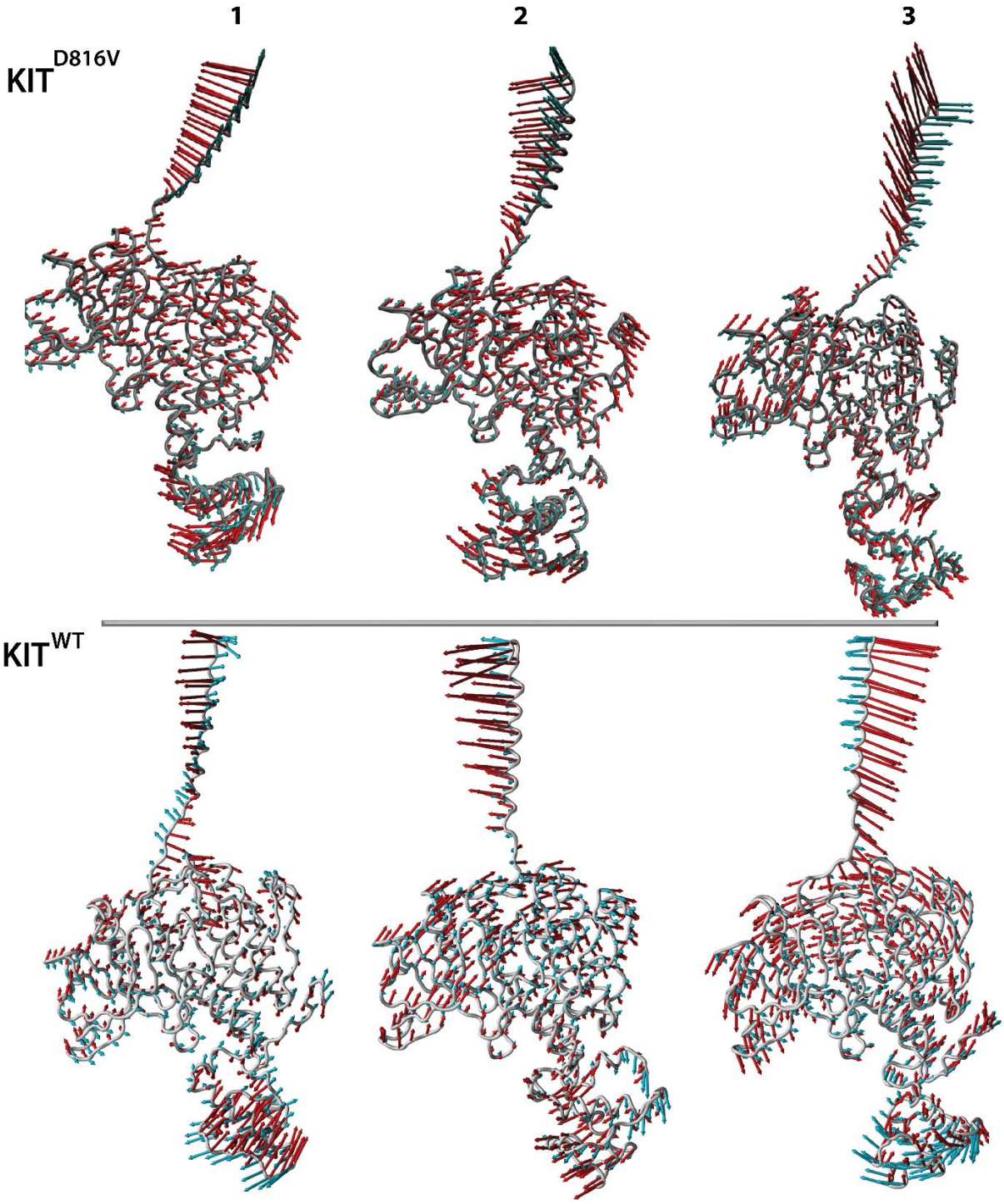


Figure S47 Principal Component Analysis modes calculated after least-square fitting of the MD conformations to the mean conformation for each respective KIT species. Atomic components in PCA modes 1–2 are drawn as red (1st mode) and cyan (2nd mode) arrows, projected on the cartoon of KIT. A cut-off distance of 4 Å was used.

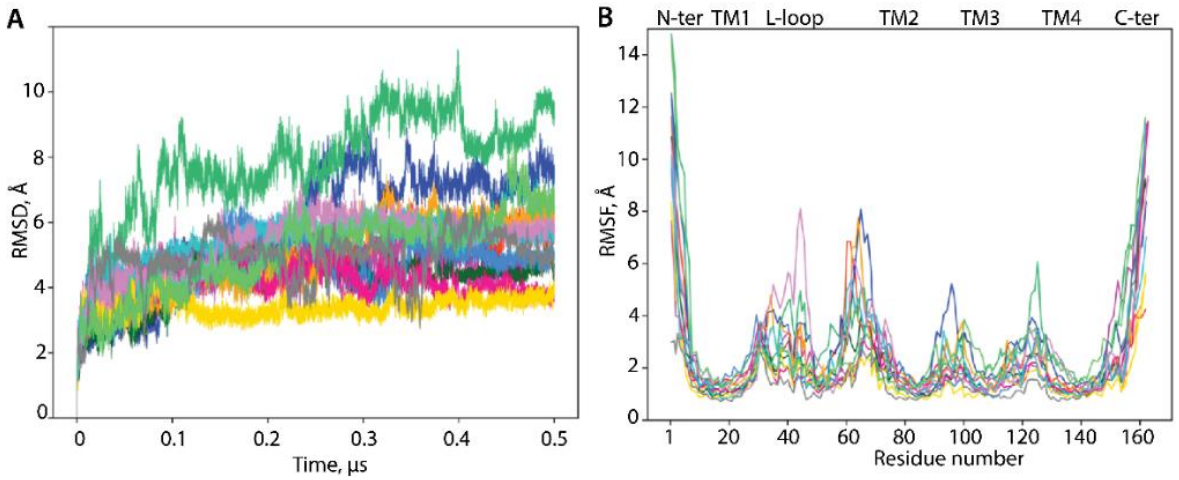


Figure S48 Characterisation of the MD simulations of four hVKORC1 mutants comparing to the native enzyme. (**A**) RMSDs from the initial coordinates computed for all Cα-atoms (right) in each protein after fitting to initial conformation. (**B**) RMSFs computed for the Cα-atoms for the MD conformation of each protein after fitting to the initial conformation (reference structure at t=0). Mutated proteins are distinguished by colour (1/2/3 replicas): hVKORC1^{A41S}(R) (orange red/orange/gold), hVKORC1^{H68Y}(R) (dark blue/blue/turquoise), hVKORC1^{S52W} (D) (fuchsia/pink/violet) and hVKORC1^{W59R} (D) (dark green/sea green/lime); hVKORC1^{WT} (gray).

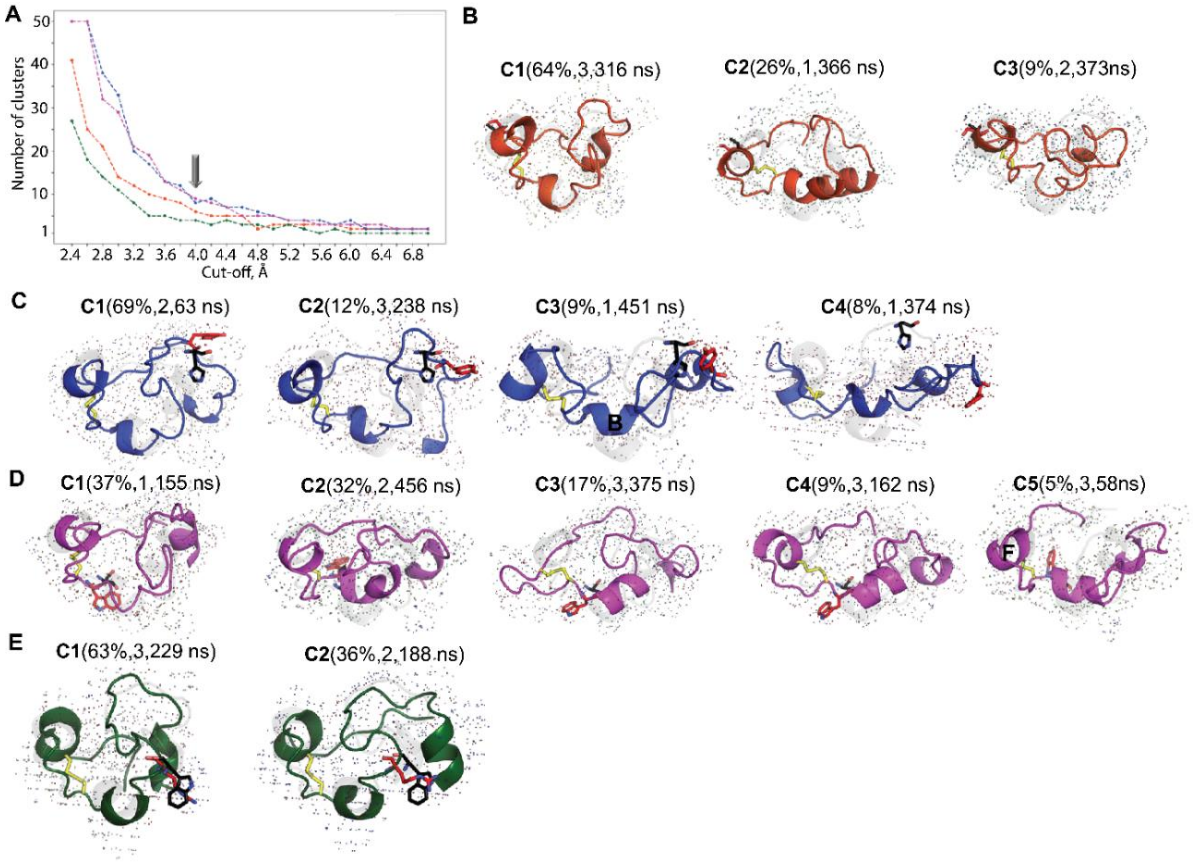


Figure S49 Ensemble-based clustering of MD conformations of the L-loop from hVKORC1 mutants. (**A**) Number of clusters obtained for each mutant using concatenated trajectories. Clustering was performed on each 100-ps frame of every trajectory using cut-off values that varied from 2.4 to 7.0 Å, with a step of 0.2 Å. The cut-off value of 4.0 Å was estimated as the optimal. (**B-E**) Representative conformations of the L-loop from clusters (**C**) with population ≥ 5%. The population of each cluster is given in brackets (in %), together with the replica number (in the bold) and the time (in ns) over which the representative conformation was recorded. Mutated proteins are distinguished by colour: hVKORC1^{A41S} (orange red), hVKORC1^{H68Y} (dark blue), hVKORC1^{S52W} (fuchsia) and hVKORC1^{W59R} (dark green); hVKORC1^{WT}(gray). The L-loop of mutants is shown as coloured cartoons superposed with L-loop of hVKORC1^{WT} presented in grey cartoon with a meshed surface. Disulphide bridges C43–C51 drawn as yellow sticks, the mutated and native residues are shown as red and black sticks respectively.

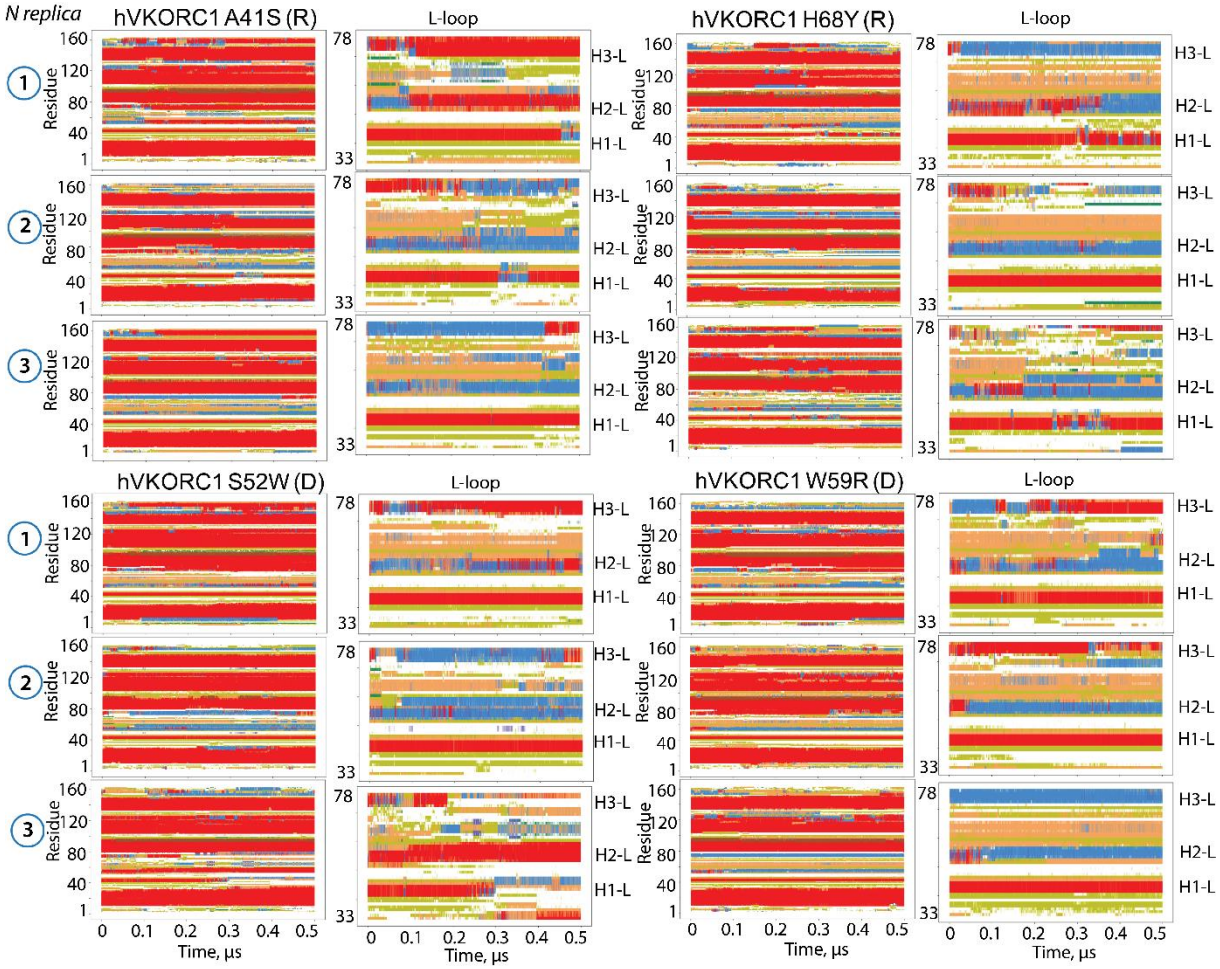


Figure S50 Folding of L-loop from hVKORC1 mutants in inactive state. The time-dependent evolution of the secondary structure of each residue, as assigned by the Define Secondary Structure of Proteins (DSSP) method: α-helix is in red, 3_{10}-helix is in blue, turn is in orange and bend is in dark yellow is shown for the full-length proteins (1st and 3rd columns) and L-loop (2nd and 4th columns). L-loop helixes (H1-L, H2-L, H3-L) are denoted. Number of replica (1-3) is encircled.

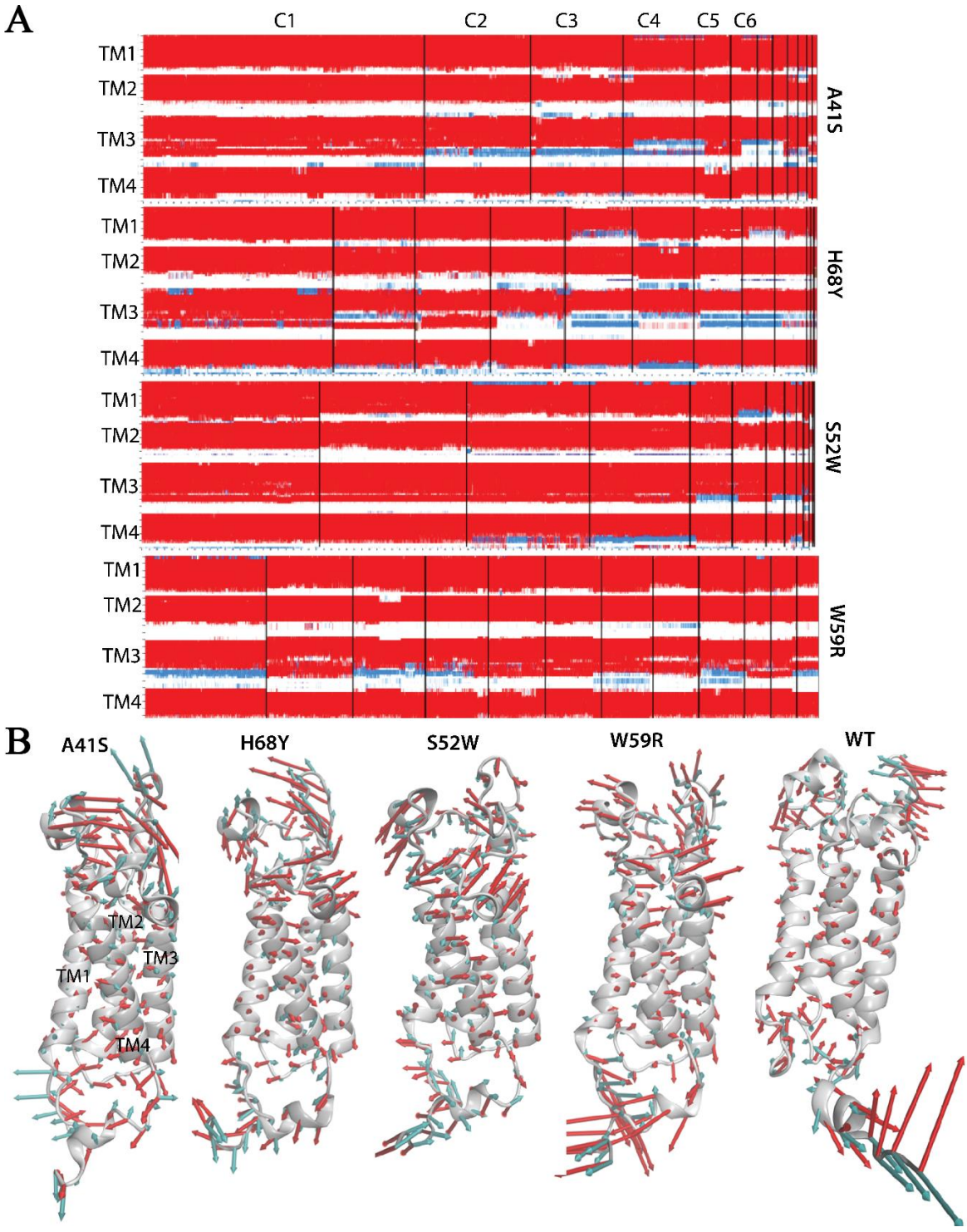


Figure S51 (**A**) Clustering of the TMD MD conformations using their secondary structure content. The α- and 3^{10}-helice are in red and blue respectively. (**B**) PCA modes calculated for concatenated MD trajectory of each hVKORC1 full-length mutant (1-163 residues) after least-square fitting of the MD conformations to the average conformation. Atomic components in the first PCA modes are drawn as red (1st mode), and blue (2nd mode) arrows projected onto the respective average structure shown as cartoon. Only motion with an amplitude ≥ 4 Å is considered.

C. TABLEAUX SUPPLÉMENTAIRES

Table S1 L-loop of hVKORC1: Residues involved in non-covalent intramolecular interactions and those available for intermolecular interactions.

Interactions and parameters:

- Hbonds criteria :
  - Distance D-A < 3.6 Å; angle at H (DHA) > 120
  - Considered donor/acceptor (D/A) atoms: N, O, S
- Contacts stabilising the secondary structure (helix) are not considered
- Hydrophobic contacts: residues: F, A, M, I, L, V, P, F, G; distance < 4 Å

Color code: positively charged – negatively charged – polar not charged -hydrophobic

INTRAMOLECULAT INTERACTIONS

Object of study	H-bonds	Hydrophobic Contacts
The 'closed' conformations, clusters $C1^m$-$C5^m$	D36, D38, R40, D44, R53, R61, E67	A41, A48, G46, I49, V54, L70, L76
The 'open' conformations, clusters $C6^m$	R35, D36, R37, D38, Y39, S50, R53, S56, S57, R58, W59, R61, N77	V45, F55, F63, L70, L76

AVAILABILITY TO INTER-MOLECULAR INTERACTION

Object of study	H-bonds	Hydrophobic Contacts
The 'closed' conformations, clusters $C1^m$-$C5^m$	R33, R35, R37, Y42, R43, C46, T50, C51, S52, S56, S57, R58, W59, H68, D73, N77	A34, L42, V45, F55, F63, L65, V66, V69, L70, L76
The 'open' conformations, clusters $C6^m$	D36, E67	G56, I49

Table S2 Hydrogen bonds involving L-loop residues in crystallographic structures 6wvi (apo-c) and 6wv3 (holo-c) of VKOR in the oxidized state. Distances (D···A and H···A, Å) between donor (D) and acceptor (A) atoms, and hydrogen (H) and acceptor (A) atoms, and the pseudo-valent angle at H (°). Amino acids forming H-bond by backbone and side chains are denoted in black and blue respectively. H-bonds observed in both forms are highlighted in blue.

H-bond	**Distance D…A/H…A/∠A–H–D**	**H-bond**	**Distance D…A/H…A/∠A–H–D**
apo-c		**holo-c**	
D36(N)…H33(O)	3.09/2.16/157.8	D36(N)…A32(O)	2.83/1.87/157.6
Y39(O)…S50(O)	2.47 /1.54/167.4	R33(N)…R37(O)	2.84/1.88/158.8
S50(N)…S48(O)	3.12/2.24/163.3	R40(N)…G46(O)	2.93/1.97/156.3
S52(N)…Y39(O)	3.12/2.23/149.9	L42(N)…Y25(O;TM1)	2.90/1.96/153.7
R61(N)…N77(O)	3.06/2.14/154.4	D44(N)…A41(O)	3.01/2.08/153.2
		I49 (O2)… V45(N)	3.05/2.20/140.5
G62(N)…Q78(O)	2.65/1.72/156.0	G62(N)…Q78(O2)	2.72/1.84/143.4
F63(N)…N80(O;TM2)	2.85/2.02/139.8	F63(N)…N80(O;TM2)	2.87/1.91/156.7
S68(N)…F70(O)	2.77/2.07/126.4	L76(N)…S74(O)	2.97/2.06/148.3
D73(N)…G71(O)	3.00/2.25/131.5		
Q78(N)…A75(O)	2.57/1.62/158.4	Q78(N)…I75(O)	3.05/2.13/150.5
N80(N;TM2)…W59(O)	2.58/1.69/148.0	N80(N;TM2)…W59(O)	2.79/1.81/164.4
N80(N;TM2)…G60(O)	2.96/2.06/150.0	N80(N;TM2)…G60(O)	3.18/2.19/166.8

Table S3 Comparison of the representative conformations of KID. The RMSD (Å) was calculated on the concatenated trajectories after least-square fitting (6 sub-tables).

1. Conformations taken from the deep wells found on FELs

KID	FEL	Well	Well	RMSD (Å)
KID^{C}	FEL_{Rg}^{RMSD}	1	2	8.4
			3	6.6
			4	7.0
			5	6.5
	FEL_{SASA}^{Rg}	1	2	4.8
KID^{CR}	FEL_{Rg}^{RMSD}	1	2	5.8
			3	5.1
	FEL_{SASA}^{Rg}	1	2	5.2

KID	FEL	Well	Well	RMSD (Å)
KID^{D}	FEL_{Rg}^{RMSD}	1	2	7.1
			3	3.9

2. Conformations taken from the deepest wells found on FELs

KID	FEL / FEL	RMSD (Å)
KID^{C}	FEL_{Rg}^{RMSD}/ FEL_{Hfp}^{Rg}	6.6
	FEL_{Rg}^{RMSD}/ FEL_{SASA}^{Rg}	6.7
	FEL_{Hfp}^{Rg}/ FEL_{SASA}^{Rg}	1.9
KID^{CR}	FEL_{Rg}^{RMSD}/ FEL_{Hfp}^{Rg}	6.9
	FEL_{Rg}^{RMSD}/ FEL_{SASA}^{Rg}	5.6
	FEL_{Hfp}^{Rg}/ FEL_{SASA}^{Rg}	4.7

KID	FEL / FEL	RMSD (Å)
KID^{D}	FEL_{Rg}^{RMSD}/ FEL_{Hfp}^{Rg}	2.0
	FEL_{Rg}^{RMSD}/ FEL_{SASA}^{Rg}	3.1
	FEL_{Hfp}^{Rg}/ FEL_{SASA}^{Rg}	2.8

3. Characterisation of conformational sub-sets from the deepest well on FEL_{Rg}^{RMSD}

KID	Population, %	mv RMSD (Å)	mv Rg (Å)
KID^{C}	11	5.70 (2)	11.80 (1)
KID^{D}	37	4.49 (2)	11.83 (1)
KID^{CR}	19	6.69 (2)	11.33 (1)

Note: The mean values are shown with the standard error.

4. Characterisation of conformational sub-sets from the most populated cluster C1 (cut-off 4 Å) obtained by ensemble-based clustering

KID	Population, %	mv RMSD (Å)	mv Rg (Å)
KID^{C}	23	3.0 (1)	12.0 (1)
KID^{D}	34	3.9 (0.3)	11.83 (3)
KID^{CR}	34	5 (1)	11.15 (4)

Note: The mean values are shown with the standard error.

5. Conformations taken from the most populated cluster C1 obtained by ensemble-based clustering

KID	RMSD (Å)
KID^{C} / KID^{D}	6.54
KID^{C} / KID^{CR}	8.21
KID^{D} / KID^{CR}	7.03

6. Conformations taken from the deepest wells found on FELs and from the most populated clusters (C1) obtained by the ensemble-based clustering

KID	FEL	RMSD (Å)
KID^{C}	FEL_{Rg}^{RMSD}	7.9
	FEL_{Hfp}^{Rg}	4.0
	FEL_{SASA}^{Rg}	3.7
KID^{CR}	FEL_{Rg}^{RMSD}	3.4
	FEL_{Hfp}^{Rg}	3.8
	FEL_{SASA}^{Rg}	2.6

KID	FEL	RMSD (Å)
KID^{D}	FEL_{Rg}^{RMSD}	4.2
	FEL_{Hfp}^{Rg}	3.9
	FEL_{SASA}^{Rg}	3.5

Table S4 Non overlapping metrics (features) value intervals in each most populated KID cluster show their differences in geometric properties.

Cluster		Metric (Feature)	Value range	
C_{N1}	C_{N2}		C_{N1}	C_{N2}
C1	**C2**	Distance S717-K725	17.1 ± 2.7 Å	12.4 ± 2.0 Å
		Distance Y721-Y747	9.6 ± 2.0 Å	14.4 ± 1.7 Å
	C3	Radius of gyration	12.4 ± 0.4 Å	13.7 ± 0.2 Å
		ψ Y721	-30.7 ± 34.9 °	73.9 ± 30.9 °
		ψ Y730	-18.0 ± 24.1 °	140.8 ± 29.3 °
	C5	Distance S717-K725	17.1 ± 2.7 Å	21.5 ± 1.7 Å
		Distance Y730-Y747	13.1 ± 3.9 Å	6.5 ± 0.8 Å
		ψ Y721	-30.7 ± 34.9 °	152.0 ± 22.0 °
C2	**C3**	Distance Y703-Y747	12.0 ± 1.4 Å	16.8 ± 1.2 Å
		Distance Y721-Y747	14.4 ± 1.7 Å	8.7 ± 1.0 Å
		Radius of gyration	12.4 ± 0.3 Å	13.7 ± 0.2 Å
		ψ Y721	-28.2 ± 18.6 °	73.9 ± 30.9 °
		ψ Y730	-11.6 ± 35.1 °	140.8 ± 29.3 °
	C5	Distance S717-K725	12.4 ± 2.0 Å	21.5 ± 1.7 Å
		Distance Y721-Y747	14.4 ± 1.7 Å	10.0 ± 1.1 Å
		ψ Y721	-28.2 ± 18.6 °	152.0 ± 22.0 °
C3	**C5**	Distance S717-K725	13.0 ± 2.0 Å	21.5 ± 1.7 Å
		Distance Y703-Y747	16.8 ± 1.2 Å	11.3 ± 0.8 Å
		Distance Y730-Y747	9.0 ± 1.6 Å	6.5 ± 0.8 Å
		Radius of gyration	13.7 ± 0.2 Å	12.6 ± 0.3 Å
		ψ Y721	73.9 ± 30.9 °	152.0 ± 22.0 °
		ψ Y730	140.8 ± 29.3 °	-6.9 ± 33.6 °

Table S5 Contingency table of agreement (%) on cluster population between K-means run with k = 5 and k = 6. The best population sizes are in bold.

		k = 6						**Total**
		C1	C2	C3	C4	C5	C6	
k = 5	C1	0.0	0.0	0.0	**0.2**	**21.2**	0.0	**21.4**
	C2	0.0	0.0	**22.5**	**11.9**	0.1	0.0	**34.5**
	C3	0.0	0.0	0.0	0.0	0.0	**7.1**	**7.1**
	C4	0.0	**22.9**	0.0	**0.2**	0.0	0.0	**23.1**
	C5	**13.8**	0.0	0.0	0.0	0.0	0.0	**13.8**
Total		**13.8**	**22.9**	**22.5**	**12.3**	**21.3**	**7.1**	**100**

Table S6 KID conformational flexibility described by the RMSF mean, minimal and maximal values (over three replicas and for residues S713-E761) relative to the most conserved H1 helix segment (Y703 - L706) for each KID entity.

KIDs	**Mean RMSF, Å**	**Min RMSF, Å**	**Max RMSF, Å**
KID	5.3	3.9 (V732-P733) 2.8 (P754)	7.7 (D716) 7 (K738-R739)
KID^{pY703}	5.8	4.8 (D723) 4 (T753-P754)	8 (G727-V728)
KID^{pY721}	3.9	3.2 (V731) 2.4 (T753-P754)	5.4 (V728) 6.1 (K738-R739)
KID^{pY730}	4.3	2.1 (G745) 2.0 (T753-P754)	5.8 (G727-Y730) 7.1 (R739)
KID$^{pY703/pY721}$	3.7	2.6 (M724) 2.3 (I744-G745) 2.0 (R743)	5.5 (S715-D716) 4.0 (I748)
KID$^{pY703/pY730}$	3.5	2.4 (M722-M724) 2.8 (V732) 3.1 (I744)	3.6 (E720) 5.4 (V728) 5.4 (R739) 4.5 (Y747-I748)
KID$^{pY721/pY730}$	5.0	3.0 (D723-M724) 4.0 (Y730-V732) 4.1 (K745) 3.5 (M757)	4.7 (720-721) 5.9 (V728) 7.7 (K738) 6.7 (I748)
KID$^{pY703/pY721/pY730}$	4.6	4.4 (V732-T734) 2.1 (T753)	6.7 (P726-G727) 6.0 (K738)

Table S7 Structure of the thioredoxine-fold proteins (from human) deposited in the PDB. The PDB identification code, method, protein domain studied, protein state, active site motif and reference are presented. Structures used in this study are highlighted in blue.

Protein	PDB code	Method	Domain	State	CX_1X_2C motif	Reference
PDI	1x5c	NMR	a'	Reduced	CGHC	To be published
PDI	1mek	NMR	a	Oxidized	CGHC	[500]
PDI	3uem	X-ray, 2.29 Å	bb'a'	Reduced	CGHC	[501]
PDI	4ekz	X-ray, 2.51 Å	abb'xa'	Reduced	CGHC	[502]
PDI	4e11	X-ray, 2.88 Å	abb'xa'	Oxidized	CGHC	[502]
PDI	6i7s	X-ray, 2.5 Å	ab'xa'	Reduced	CGHC	[503]
Erp18	1sen	X-ray 1.20 Å	a	Oxidized	CGAC	To be published
Erp18	2k8v	NMR	a	Oxidized	CGAC	[428]
TMX1	1x5e	NMR	a	reduced	CPAC	To be published

Table S8 Sequence identity / similarity in ERp18, PDI, Tmx1 and Tmx4. The sequences of domain a from the experimentally determined structures of ERp18, PDI, and Tmx1, as well as the 129 amino acid fragment from the Q9H1E5 sequence of Tmx4 (https://www.uniprot.org/uniprot/) were aligned with ERp18, that was used as a reference for alignment and numbering. The identity / similarity after aligning the sequences of domain a (full-length), fragments F1 (33-50) and F2 (67-84).

	PDI	**Tmx1**	**Tmx4**
Full-length			
ERp18	23/38	15/23	15/23
PDI		26/42	23/38
Tmx1			47/68
F1 (17 aas); F2 (17 aas)			
ERp18	60/70; 22/22	40/60; 11/22	30/60; 11/7
PDI		60/70; 0/0	60/70; 0/0
Tmx1			80/90; 50/67

Table S9 Helical fold (top) of L-loop in isolated hVKORC1 (**A**), in Model 1 and Model 2 (**B**); of the PDI in Model 1 and Model 2 (**C**).

A. Isolated L-loop

Cluster	Helical folding, %
1	29
2	27
3	34
4	25
5	34
6	20

B. L-loop of hVKORC1 (complex)

Conformation	Helical folding, %
Model 1	
t = 10 ns	25
t = 60 ns	36
t = 80 ns	38
Model 2	
t = 10 ns	25
t = 60 ns	31
t = 80 ns	15

C. PDI folding (complex)

Conformation	Folding proportion, % (total / helix / sheet)
PDB 4ekz, domain a	60 / 39 / 21
Model 1	
t = 10 ns	56 / 33 / 23
t = 60 ns	57 / 37 / 20

Table S10 Folding of the intrinsically disordered regions in KIT^{WT} and KIT^{D816V}

	KIT^{WT} mean folding (%)			**KIT^{D816V} mean folding (%)**		
	Helix	**Strand**	**Total**	**Helix**	**Strand**	**Total**
JMR	13 ± 4	14 ± 0	27 ± 4	10 ± 4	16 ± 2	29 ± 5
KID	30 ± 8	3 ± 2	33 ± 5	32 ± 5	2 ± 2	33 ± 7
A-loop	**3 ± 3**	13 ± 9	17 ± 5	**13 ± 5**	9 ± 8	22 ± 9
C-tail	13 ± 9	0	13 ± 9	15 ± 11	0	15 ± 11