



HAL
open science

A stochastic conformational sampling method for complex protein architectures involving disordered regions

Ilinka Clerc

► **To cite this version:**

Ilinka Clerc. A stochastic conformational sampling method for complex protein architectures involving disordered regions. Networking and Internet Architecture [cs.NI]. Université de Toulouse, 2024. English. NNT : 2024TLSEI016 . tel-04871821

HAL Id: tel-04871821

<https://theses.hal.science/tel-04871821v1>

Submitted on 7 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Doctorat de l'Université de Toulouse

préparé à l'INSA Toulouse

Une méthode d'échantillonnage stochastique d'ensembles de
conformations pour des systèmes protéiques complexes
comprenant des régions désordonnées

Thèse présentée et soutenue, le 25 juillet 2024 par

Ilinka CLERC

École doctorale

EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse

Spécialité

Informatique et Télécommunications

Unité de recherche

LAAS - Laboratoire d'Analyse et d'Architecture des Systèmes

Thèse dirigée par

Juan CORTES et Pau BERNADO

Composition du jury

M. Pierre WEISS, Président, CNRS Occitanie Ouest

Mme Malene Ringkjøbing JENSEN, Rapporteuse, CNRS Alpes

M. Matthieu MONTES, Rapporteur, Conservatoire National des Arts et Métiers

Mme Jessica ANDREANI, Examinatrice, CEA Paris-Saclay

M. Juan CORTES, Directeur de thèse, CNRS Occitanie Ouest

M. Pau BERNADO, Co-directeur de thèse, INSERM Occitanie Méditerranée

Membres invités

Mme Nathalie SIBILLE, CNRS Occitanie Est

Résumé

A la différence des protéines globulaires, largement étudiées en biologie structurale, certaines protéines dites désordonnées (IDPs) et certaines régions désordonnées (IDRs) dans des architectures protéiques n'adoptent pas une forme bien définie en solution. Cette flexibilité leur confère des fonctionnalités complémentaires à celles des protéines structurées et leur étude ouvre de nouvelles perspectives prometteuses dans les domaines du médical, des biotechnologies et des biomatériaux. Ces systèmes désordonnés ne peuvent être représentés par une seule conformation, mais nécessitent des modèles d'ensembles comprenant plusieurs milliers de conformations possibles. Ces ensembles représentent la distribution d'états que la protéine peut adopter en solution. L'obtention de ces ensembles par des méthodes expérimentales seules (RMN, SAXS...) est insuffisante et doit être couplée à des approches computationnelles. Ce couplage de méthodes demeure un défi et un enjeu majeur dans ce domaine de recherche. Dans cette thèse, nous présentons MoMA-FReSa (Molecular Motion Algorithms - Flexible Region Sampler), une nouvelle méthode d'échantillonnage stochastique qui s'attaque à cette problématique. MoMA-FReSa est une méthode computationnelle polyvalente et adaptable à une large gamme de systèmes contenant des régions désordonnées, y compris les plus complexes. Grâce à l'alliance d'une méthode de décomposition hiérarchique du système, d'une stratégie d'ordonnement sophistiquée et d'un processus d'échantillonnage stochastique, ce programme permet de générer de vastes ensembles de conformations dans des temps très courts. La thèse s'articule en quatre chapitres. La première partie présente un panorama exhaustif des méthodes d'échantillonnage de régions et protéines désordonnées existantes, et positionne MoMA-FReSa comme une approche innovante et prometteuse dans ce contexte. Le chapitre 2 explicite la méthodologie employée dans MoMA-FReSa et établit ses capacités sur un ensemble de référence illustrant la diversité des systèmes pouvant être rencontrés. Le chapitre 3 poursuit ensuite en présentant divers cas pratiques et collaborations mettant en œuvre MoMA-FReSa, démontrant ainsi le potentiel de ce programme sur des applications plus innovantes. Enfin, le chapitre 4 explore des perspectives d'amélioration de la méthode en utilisant de l'apprentissage par renforcement. Une interface utilisateur est en cours de développement pour rendre cette méthode accessible à la communauté scientifique sous forme d'applications dans la suite MoMA.

Mots clés : Méthodes informatiques, Echantillonnage Stochastique, Ensembles de conformations, Protéines intrinsèquement désordonnées

Abstract

Unlike globular proteins, widely studied in structural biology, intrinsically disordered proteins (IDPs) and intrinsically disordered regions (IDRs) within proteins do not exhibit a well-defined three-dimensional structure in solution. This flexibility confers upon them complementary functionalities to those of structured proteins and their study opens up promising new opportunities in the fields of medicine, biotechnology, and biomaterials. These disordered systems cannot be represented by a single conformation but require ensemble models of several thousands of possible conformations. These ensembles represent the distribution of states that the protein can adopt in solution. Obtaining these ensembles only using experimental data (NMR, SAXS...) is not possible. Indeed, these data must be coupled with computational approaches. This coupling of methods remains a major challenge in this research field. In this thesis, we present MoMA-FReSa (Molecular Motion Algorithms - Flexible Region Sampler), a novel and promising stochastic sampling method that addresses this problem. MoMA-FReSa is a versatile computational method adaptable to a wide range of systems containing disordered regions, including the most complex ones. By combining a hierarchical decomposition of the system with a sophisticated scheduling strategy and with a stochastic sampling process, this program allows generating large ensembles of conformations in short times. The thesis is divided into four chapters. The first part presents a comprehensive overview of existing sampling methods for disordered regions and proteins, and positions MoMA-FReSa as an innovative and promising approach in this context. Chapter 2 explains the methodology employed in MoMA-FReSa and establishes its capabilities on a benchmark illustrating the diversity of systems that can be encountered. Chapter 3 continues by presenting various practical cases and collaborations using MoMA-FReSa, demonstrating the potential of this program for more innovative applications. Finally, Chapter 4 explores perspectives for improving the method using reinforcement learning. A user-friendly interface is under development to make the method accessible to the scientific community in the form of applications in the MoMA suite.

Keywords: Computational methods, Stochastic Sampling, Ensembles of conformations, Intrinsically disordered proteins

Remerciements

Merci !

A toute l'équipe RIS, de m'avoir accueillie dans une période compliquée, entre covid et confinement.

A ceux qui étaient là au début, à celle qui fut comme un modèle, une mentore. A ceux qui ont partagé le même fardeau que moi, ceux que j'ai rencontrés autours de jeux de cartes, celui que j'ai retrouvé par hasard. . .

A ceux qui se sont plongés avec moi dans les protéines et les lignes de code. A ceux avec qui j'ai partagé un petit bout de ma thèse, de mes pauses, ou de mes frites. A ceux qui sont sur le point de soutenir à leur tour, vous y êtes presque, c'est la dernière ligne droite !

A ceux qui resteront après, à celles qui ont amené le renouveau dans l'équipe, je suivrai votre parcours avec un regard attentif. A tous les éternels de l'équipe à l'image de notre sapin. . .

A Juan qui a su m'écouter, me comprendre, me guider, quand j'étais au plus haut comme au plus bas. A Pau, par tes conseils et ta gentillesse, tu as su m'aider à garder le cap et la motivation nécessaire pour continuer d'avancer. Merci à vous deux de m'avoir accompagnée avec bienveillance. Je suis si fière du chemin parcouru ensemble, merci !

A l'équipe pédagogique de l'INSA pour m'avoir permis de découvrir l'autre côté de la barrière et vivre mon rêve de découvrir l'enseignement. A tous mes élèves qui ont fait de mes premiers cours une expérience magique.

A tous mes amis.

A ceux que je me suis faits au Lycée. Ceux qui m'ont suivi à l'INSA pour une nouvelle aventure. Toi que j'ai rencontré au collège, la distance et le temps ne nous ont jamais vraiment séparées, chaque fois que je te revois j'ai l'impression de t'avoir quitté hier. . .

A ceux rencontrés à l'INSA, ceux des soirées pizzas, ceux du club Neotokyo. . .

A ceux qui sont restés, à ceux qui sont partis, loin en France, dans un autre pays. A ceux qui sont revenus, à ceux qui reviendront peut-être. . .

Aux amis faits en chemin, le groupe de Montpellier, les amis faits en ligne, de Paris, de Poitiers, certains que j'ai eu la chance de parfois rencontrer. . .

A mes deux confidentes, complices et amies, qui sont un peu avec moi tous les jours, et tous les jours pendant une thèse ce n'est pas rien. Toi dont la rencontre a été un petit bouleversement dans ma vie et toi qui fut, pour mon plus grand bonheur, mise sur mon chemin.

Car ma thèse je la dois aussi un peu à tout ça :

A ces après-midis à Florensac, aux attractions du Luna Parc. A cette soirée irréelle pour un concert, à cette marche féérique au milieu des lanternes, à ces anniversaires comme au temps de nos apparts étudiants. . .

A ces soirées tous connectés à rire, à lutter ensemble, à s'amuser. Aux soirs Top Chef ou animés, à ces visionnages et débriefs toujours de qualité. . .

A ces soirées jeux de société qui arrêtent le temps. A nos retrouvailles annuelles que j'attends avec impatience et aux retrouvailles qui se font attendre. . .

A ces laser games, ces escapes games, ces après-midis comédies musicales. . .

A toutes ces conversations hors du temps que j'entame sans un bonjour, comme si on s'était parlé la veille après des mois sans échanges.

A ma famille.

A mes parents sur qui je peux toujours compter et qui font tout pour mon bonheur. A mon frère avec qui je ne mesure pas la distance physique qui nous sépare, dont nos échanges et moments de partage participent à mon équilibre...

A mon amour, mon partenaire, avec qui j'ai avancé et muri, main dans la main durant cette longue période. J'ai hâte de voir de quoi demain sera fait avec toi...

A tous mes soutiens ronronnant de Florensac et à Ocelot et Hildi dont la chaleureuse présence quotidienne compense largement les heures de sommeil volées...

A vous tous que je vois autours de grands repas un peu trop bruyant chaque année, je chéris nos liens un peu plus chaque jour avec le temps qui passe...

A toi qui es partie, toi aussi tu étais là le jour où j'ai soutenu, merci pour tout l'amour que tu m'as donné, il me porte et me protège.

A tous ceux qui sont venus me voir pour la soutenance, à toutes les petites mains qui ont tout préparé pour faire de cette journée une vraie fête dont je me souviendrai toute ma vie. Ce jour-là, je vous avais autours de moi, mes collègues, mes amis, ma famille et ceux qui depuis le temps en font presque partie...

Merci aux fous qui ont campé à la belle étoile pour venir à ma soutenance, à celle qui est venue après sa garde sans avoir dormi. A ceux qui n'ont pas pu venir mais qui ont pensé à moi. A ceux qui ont regardé de loin et commenté, ce jour-là j'étais une star.

Merci d'avoir été là, ce jour-là ou avant.

Je suis heureuse et très fière de ce que j'ai accompli et je remercie chaque main tendue, chaque épaule qui a reçu mes larmes et mes doutes, chaque parole qui a su me guider.

A tous ceux qui comptent, à tous ceux que j'aime.

Merci à vous.

Contents

1	Introduction	1
1.1	Context: The study of Intrinsically Disordered Proteins	1
1.1.1	An overview of the structure and function of Intrinsically Disordered Proteins	1
1.1.2	An overview of computational approaches for the study of disordered proteins	3
1.2	State of the art: The molecular modeling perspective	5
1.2.1	Interactions of structured domains mediated or regulated by disordered linkers	5
1.2.2	Interactions between disordered regions and structured domains	7
1.2.3	Extreme fuzzy complexes and phase separation behavior	11
1.2.4	Emergence of novel approaches and machine learning to model protein disorder	13
1.3	Contributions and structure of the thesis	13
1.3.1	Chapter 2: An exhaustive description of MoMA-FReSA	14
1.3.2	Chapter 3: Collaborations and applications	15
1.3.3	Chapter 4: A deep learning approach to enhance MoMA-FReSA	16
1.3.4	Software availability	16
2	MoMA-FReSa: A conformational sampling method for complex disordered systems	17
2.1	Introduction	17
2.2	Method	18
2.2.1	Method overview	18
2.2.2	System representation through the pre-processing stage	21
2.2.3	Pre-processing construction steps	30
2.2.4	Sampling principle	34
2.2.5	Main sampling process	35
2.2.6	Post-processing in MoMA-FReSa and beyond	45
2.2.7	Implementation details	47
2.3	Results	48
2.3.1	Benchmark and protocol overviews	48
2.3.2	Analysis of the Results	51
2.3.3	Study of the effect of the new distance test for loops	58
2.4	Discussion and conclusion	59
3	Applications of MoMA-FReSa	61
3.1	Introduction	61
3.2	Conception of multi-modular systems	62
3.2.1	General presentation	62
3.2.2	Conception work	65
3.2.3	Discussion and perspectives	69
3.3	Estimation of the effective concentration in biomolecular interactions	69
3.3.1	General presentation	69

3.3.2	Results and analyses	72
3.3.3	Discussion and perspectives	75
3.4	Structural analysis of linkers in multi-domain proteins	76
3.4.1	General presentation	76
3.4.2	Post-processing analysis	77
3.4.3	Discussion and perspectives	80
3.5	Sampling of structural motifs	81
3.5.1	General presentation	81
3.5.2	Ensemble generation and analysis	82
3.5.3	Discussion and perspectives	87
3.6	Conclusion and perspectives	88
4	A Reinforcement Learning method for MoMA-FReSa	89
4.1	Introduction	89
4.2	Context and implementation choices	90
4.2.1	A learning approach suited to our sampling method	90
4.2.2	Learning methods employed	91
4.3	Implementation in MoMA-FReSa	93
4.3.1	Use of the learning method during the sampling procedure	93
4.3.2	Buffering and backtracking integration to the learning procedure	95
4.3.3	Global architecture	96
4.4	Ongoing work and perspectives	98
4.4.1	Ongoing work	98
4.4.2	Perspectives and areas of exploration	99
4.5	Conclusion	100
	Conclusions and perspectives	101
A	Résumé en Français	105
A.1	Chapitre 1: Introduction	105
A.2	Chapitre 2 : Une description exhaustive du MoMA-FReSA	106
A.3	Chapitre 3 : Applications dans un cadre collaboratif	107
A.4	Chapitre 4 : Une approche d'apprentissage profond pour améliorer MoMA-FReSA	107
A.5	Disponibilité du logiciel	108
	Bibliography	109

Introduction

Contents

1.1	Context: The study of Intrinsically Disordered Proteins	1
1.1.1	An overview of the structure and function of Intrinsically Disordered Proteins	1
1.1.2	An overview of computational approaches for the study of disordered proteins	3
1.2	State of the art: The molecular modeling perspective	5
1.2.1	Interactions of structured domains mediated or regulated by disordered linkers	5
1.2.2	Interactions between disordered regions and structured domains	7
1.2.3	Extreme fuzzy complexes and phase separation behavior	11
1.2.4	Emergence of novel approaches and machine learning to model protein disorder	13
1.3	Contributions and structure of the thesis	13
1.3.1	Chapter 2: An exhaustive description of MoMA-FReSA	14
1.3.2	Chapter 3: Collaborations and applications	15
1.3.3	Chapter 4: A deep learning approach to enhance MoMA-FReSA	16
1.3.4	Software availability	16

1.1 Context: The study of Intrinsically Disordered Proteins

1.1.1 An overview of the structure and function of Intrinsically Disordered Proteins

In the last few decades, Intrinsically Disordered Proteins and Regions (IDPs/IDRs) have emerged as key actors in multiple fundamental biological processes [62, 147]. Due to the lack of permanent secondary and tertiary structure, IDPs/IDRs are highly malleable molecules adapted to perform specialized functions that complement those of their globular counterparts [224, 208]. Intrinsic disorder is abundant in eukaryotic proteomes, where it contributes to the cellular complexity by participating in the vast majority of signaling and regulation events [220, 48]. Their amino acid sequence, rich in charged and non-structuring residues [175, 206], and often displaying low complexity [133], determines their lack of permanent structure. These sequence features have been widely used in bioinformatics approaches to identify disorder and function from proteomics data [216, 126, 145]. While some proteins display disorder all along the sequence (IDPs), in other cases disorder is only present in specific segments of the sequence, which are named IDRs [224, 208]. IDRs can be placed between globular domains (linkers), restricting their relative distance and

orientation, or at the N- or C-termini as disordered tails of folded domains [202]. These distinct disordered protein architectures define the types of the resulting assemblies occurring upon binding to the biological partners (see Figure 1.1).

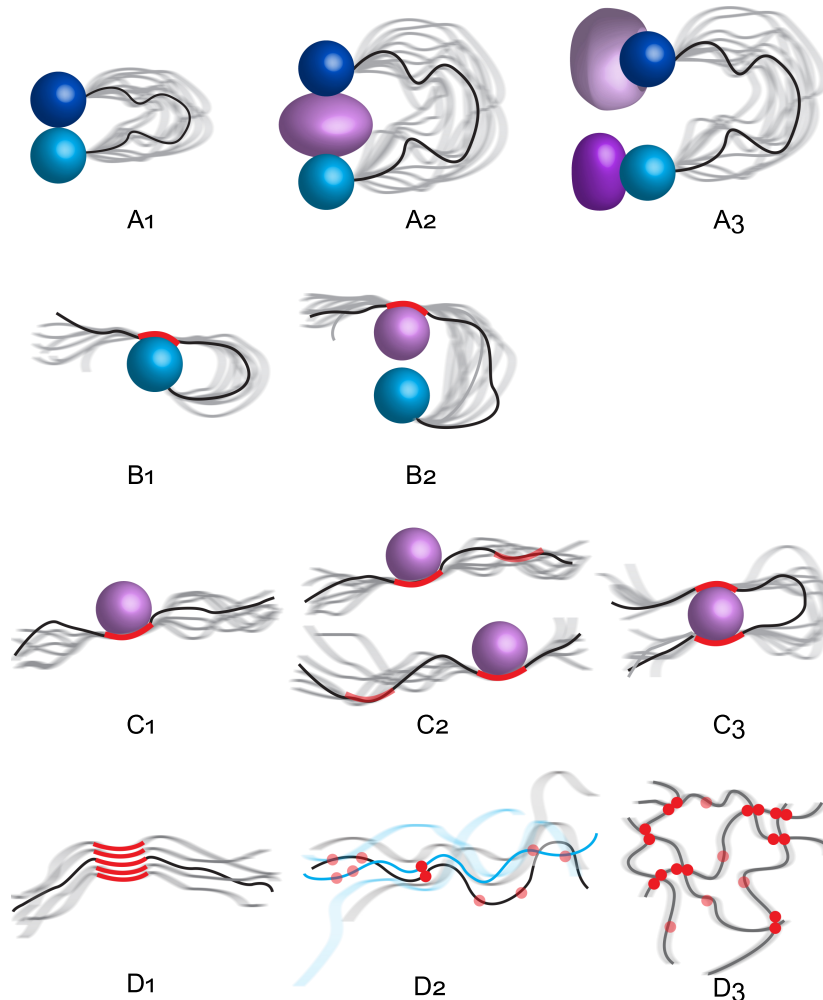


Figure 1.1: Illustration of different types of interactions involving IDPs/IDRs. They are classified depending on the ordered/disordered nature of the interacting regions. The representation is not aimed to be exhaustive, and combined interaction types do also exist. A-class cartoons refer to proteins consisting of multiple structured domains connected by flexible linkers. The domains can either interact intramolecularly (A₁) or with other biomolecules (A₂ and A₃). B-class cartoons represent interactions driven by SLiMs (in red) placed in IDRs that recognize their own globular domain (B₁) or another protein (B₂). C-class cartoons represent bimolecular interactions involving an IDP with a globular protein. While in C₁ assembly a single SLiM (in red) recognizes the globular domain, C₂ and C₃ represent scenarios where two similar SLiMs of the IDP interact with a globular protein with one (C₂) or two (C₃) binding sites. D-class cartoons represent the interaction between disordered proteins that either form amyloid-like structures (D₁), extremely fuzzy complexes (D₂) or unstructured condensates with liquid-like behavior (D₃). In D₂ and D₃, multiple low-affinity non-specific interactions (red dots) are present.

From a functional perspective, most of these disordered segments act as interaction specialists [204]. Their plasticity enables highly specific recognition by adapting their bound confor-

mation to the physicochemical nature of the partner surface. These interactions are normally performed via evolutionary-conserved short linear interaction motifs (SLiMs) inserted within the chain [53, 209]. SLiMs, which encompass from 3 to 10 contiguous amino acids, are defined according to consensus sequences that are considered as the hot-spots of the interaction [209]. The large number and sequence variability of the identified interacting segments exemplify the richness of recognition events performed by IDRs [116]. Interestingly, several proteins share the same consensus sequence, identifying the family of binding partners recognized. Differences in the remaining residues and/or flanking regions can modulate the thermodynamics and kinetics of the recognition event, as well as the capacity to discriminate between several partners (specificity). Interaction mechanisms are classified according to the flexibility adopted by the disordered fragment upon binding [76]. While some disordered segments present a well-defined rigid structure in the bound form, others display an almost complete conformational freedom with multiple, very weak fuzzy contacts with the partner [76, 148, 29]. A prominent example of these extremely fuzzy complexes is the formation of liquid-like membraneless compartments, which have emerged in the recent years as a very efficient mechanism for the spatio-temporal organization in living cells [186, 14]. Beyond the existence of these two extreme scenarios, the growing number of interactions reported suggests that there is a continuum of flexible binding modes [136]. Furthermore, post-translational modifications, often occurring in disordered segments modifying their physicochemical properties and enormously increasing the number of interaction possibilities [203], can act as switches to turn recognition events on and off [49, 7]. This large spectrum of interaction modes and regulation mechanisms explains the variability of functional outcomes and the numerous pathologies associated with the malfunction of disordered proteins and their complexes [207].

The conformational characterization of disordered proteins and their complexes still represents a challenge for biophysicists. The most suitable structural biology techniques for their study, Nuclear Magnetic Resonance (NMR) [135, 63], Small-Angle Scattering (SAS) [45, 168], and single-molecule Förster Resonance Energy Transfer (smFRET) [90], or their combination [143], provide average information that reports on the ensemble of co-existing conformations present in solution [167, 15]. Moreover, interactions mediated by IDPs/IDRs are often characterized by their low affinity, inducing an equilibrium between bound and unbound forms that further complicates their structural characterization *in vitro* [44].

In this context, computational methods, either alone or in combination with experimental data, have become pivotal for the structural and dynamic characterization of this elusive class of biomolecules. A large variety of computational methods have been specifically developed, adapting the level of description to the size of the molecular system, the question to be addressed, and the availability of experimental information [20, 178, 103, 23, 184]. The final aim of most of these approaches is the generation of conformational ensembles representing realistic pictures of biomolecular entities with the capacity to provide the structural bases of cellular mechanisms and anticipate functional properties [120].

1.1.2 An overview of computational approaches for the study of disordered proteins

Various computational methods can be applied to the structural investigation of IDPs/IDRs and their interactions [20, 178, 103, 23, 184]. The choice of the method depends on different factors: (1) availability of experimental data, (2) level of detail and timescale at which the molecular mechanism has to be investigated, (3) size of the molecular system, (4) computing power available.

When experimental (biophysical or biological) information is lacking, bioinformatics tools can be applied to identify binding motifs from IDP sequences [60, 114], and to predict interactions between these motifs and protein partners [166, 77, 157]. These predictive tools deliver relevant insights to understand functional mechanisms involving IDPs. However, they only provide a partial and qualitative picture of molecular interactions. The study of thermodynamic and kinetic aspects of protein interactions requires a more global exploration of conformational states and transitions. This exploration can be based on different types of models and algorithms.

Whereas molecular dynamics (MD) simulations using all-atom physics-based force field models are widely used for the investigation of interactions involving globular proteins, the applicability of “standard” MD approaches to IDPs/IDRs is relatively limited [20, 187]. A first limitation comes from the fact that force fields, such as Amber and CHARMM, were mostly developed having globular proteins as targets. Thus, they tend to enrich the structure with secondary structure elements (α -helices and β -strands), and to produce collapsed conformations. Recent versions of these force-fields (e.g., [91, 189, 172, 160]) have been introduced to mitigate these effects. In particular, a balanced description of protein-water interactions thanks to new water models [158] or rescaling approaches [21] have been shown to be critical for improving the description of disordered proteins [172, 212].

The other limitation of all-atom MD simulations is their computational cost, which precludes their routine use to investigate large structural rearrangements or interaction mechanisms requiring long timescales. One of the main reasons for this high computational cost is the large size of the simulation box containing the protein(s) and water molecules, due to the large radius of gyration of IDPs (with respect to folded protein) and their fluctuations.

MD protocols applied to IDPs/IDRs often rely on enhanced sampling techniques, such as replica-exchange [193, 75], metadynamics [119, 8] or combined approaches [38, 159], which are more efficient than basic MD techniques to explore multiple-basin energy landscapes (we refer the interested reader to specialized reviews [1, 226] for further information on enhanced sampling techniques). Note also that advances in software and hardware, enabling efficient parallel computing, have significantly contributed to extending the applicability of all-atom MD approaches, in particular thanks to the exploitation of graphics processing units (GPUs) [192]. Despite these methodological and technical advances, in practice, all-atom MD simulations are nowadays applicable to the investigation of relatively small systems (e.g. interactions involving protein fragments or a small number of disordered peptides) or too short timescales for larger systems.

The investigation of larger systems and/or longer timescales relies on the application of coarse-grained (CG) models. Although these models do not provide the same level of detail as the all-atom ones, they allow a much wider exploration of the conformational energy landscape. CG models can range from simple $G\bar{o}$ -like models [227, 110] to more complex ones, considering one or several beads per amino acid residue, such as AWSEM-IDP [222], PLUM [13], MARTINI [55], SIRAH [107], CALVADOS [200, 201, 199] or Mpipi [98]. ABSINTH [213] can be considered as an intermediate approach between all-atom and CG models, since it only considers dihedral angles as variables, so that small groups of bonded atoms move as rigid bodies. For their application to IDPs, special attention has been paid to the (implicit) solvation terms included in most of these models. Note that implicit solvation models can be applied using other exploration algorithms, in addition to MD. This is for instance the case of ABSINTH, which was specially developed for Monte Carlo (MC) simulations [213].

Although MD-based methods are attractive due to their accuracy (particularly for atom-

istic simulations) and capacity to provide information on the temporal evolution of the molecular system, other types of algorithms are more efficient in sampling the huge conformational space of IDPs/IDRs. In addition to MC, several methods based on stochastic sampling techniques have been proposed to generate ensemble models of IDPs/IDRs. The most popular examples of these methods are TraDES [71] and Flexible-Meccano [16, 149]. These approaches incrementally construct IDP/IDR conformations using probability distributions of the dihedral ϕ and ψ angles of amino acid residues extracted from experimentally-determined protein structures, and can include information about secondary structure propensities along the sequence. A recent variant of these methods, operating with three-residue fragments, has been shown to generate higher-quality conformational models of IDPs containing partially structured elements, which naturally emerge as they are encoded in the protein sequence [67]. The methodology developed in the present thesis inscribes in the context of these statistical sampling methods.

While modeling approaches can provide an *ab initio* description of IDP/IDR conformational ensembles based only on physics- and/or knowledge-based models, their predictive capabilities can be greatly improved by taking advantage of available experimental information. In this respect, NMR, SAS, smFRET and other experimental results can be used for correcting the model inaccuracies, either by biasing or restraining the sampling into the most relevant regions of the conformational space, or by reweighting the simulation results *a posteriori*. Numerous algorithms have been proposed for this combination of simulation methods and experimental data (e.g. [176, 27, 93, 180, 25, 26, 165, 118, 31]). The interest to consider experimental data is particularly true for fast, stochastic approaches to generate conformational ensemble models. Actually, ensembles generated by TraDES and Flexible-Meccano are usually filtered and refined based on experimental data using computational tools such as ENSEMBLE [115], ASTEROIDS [146], EOM [18, 205], the Maximum Occurrence [19, 142], and BME [32]. Integrative approaches, combining several complementary experimental and computational methods, are applied to derive more accurate structural models of IDPs/IDRs and their complexes (e.g., [30, 80]).

1.2 State of the art: The molecular modeling perspective

This section describes how different computational strategies, alone or in combination with experimental data, have been applied to study disordered biomolecular complexes. Note that the aim is not to provide an exhaustive enumeration of computational studies on IDPs/IDRs, but to briefly describe the methods and exemplify them with some relevant applications that, in some cases, are addressed in some of the applications of the methods developed along this thesis. The section is mainly organized according to the different architectures illustrated in Figure 1.1, with a final sub-section presenting the latest methods, notably based on machine learning approaches.

1.2.1 Interactions of structured domains mediated or regulated by disordered linkers

The majority of proteins in prokaryotes and eukaryotes are composed of several domains connected by linkers [4]. Domain-linker-domain (DLD), illustrated in the Figure 1.1.A, is the most common architecture, but more complex combinations of globular domains connected by flexible linkers exist. Although linkers can be very long, they typically involve from 2 up to ~ 30 residues [78, 169], displaying high levels of flexibility and absence of permanent

secondary structure. MD simulations in combination with ^{15}N NMR relaxation experiments have shown that this flexibility occurs in a broad range of timescales [212].

Linkers are not mere connectors between domains. Indeed, their length and sequence have been evolutionarily tailored to play key functional roles, being frequently involved in allosteric mechanisms [130, 152, 92]. One of the main advantages of this architecture is their capacity to enhance the effective local concentration, C_{eff} , of the linked domains, thus promoting intra- or inter-molecular interactions (Figure 1.1.A₁-A₂). They are also key components in signaling processes: linkers can propagate conformational changes in one domain, e.g. induced by ligand binding, to the other domain, which may activate or inhibit other interactions (Figure 1.1.A₃). Below, we present examples of functional roles of linkers, and discuss how they have been investigated using various computational approaches.

Linkers in bi-specific antibodies

The role of linkers to enhance C_{eff} , as well as their effects on stability, affinity and activity, have been of particular interest in the context of bi-specific antibodies conceived from the combination of different antibodies or antibody fragments [35]. These engineered molecules have a great potential for diagnostic and therapeutic applications. The simplest and most common architecture, called single-chain variable domain (scFv) format, consists of antigen-binding sites of two antibodies connected through a linker. Theoretical methods based on simple worm-like models have been proposed to investigate the binding affinity of these systems [231], allowing to establish a relationship between the linker length and C_{eff} . However, predictions provided by such simple models can be inaccurate since they do not consider sequence-dependent structural properties of the linker and disregard possible interactions with the domains. Both, linker sequence and interactions, have been shown to be important for the conformational preferences of multi-domain proteins [108, 139]. Therefore, more detailed models are required for their investigation. In their study, Mittal *et al.* [139] performed simulations using the ABSINTH together with an MC-based method called Hamiltonian Switch Metropolis Monte Carlo (HS-MMC) [140] specially developed to enhance sampling of IDRs connected to a folded domain. Although only relatively small artificial constructs involving SH3 and WW domains were used in this study, the approach and the conclusions can be generalized to other systems, including scFvs. For this type of systems, perturbation-response methods are a valuable tool to investigate the dynamical coupling between the two complementarity-determining regions (CDR), as well as the role of the linker in this mechanism. As an interesting example of such methods, Ettayapuram-Ramaprasad *et al.* [69] proposed an implementation based on an effective Hessian matrix computed from all-atom MD simulations. This Hessian matrix represents an ensemble-based elastic network that captures collective motions, from which the effect of local perturbations can be exhaustively investigated.

Linkers in multi-domain enzymes

Multi-domain enzymes are another type of proteins for which the study of the functional roles of linkers has attracted interest over the past two decades. MD simulations have been widely used for this purpose. For instance, standard all-atom MD protocols with simulation times of 20 ns were used to investigate the role of the linker in cullin-RING E3 ubiquitin ligases [125], unveiling that allosterically controlled linker motions modulate the distance between the domains, and therefore the ubiquitin transfer reactions. Nevertheless, these types of “basi” techniques cannot be applied to investigate thermodynamic and kinetic properties

that would require extremely long simulations. CG models and enhanced sampling methods are the natural alternatives in this case. As an example, Li *et al.* [122] assessed the essential role of disordered linkers in allosteric regulation processes using a G $\bar{5}$ -like model and umbrella sampling combined with a theoretical thermodynamic analysis. Their results suggested that the influence of the linker can be characterized by a C_{eff} that depends on the linker length and flexibility.

The case of bimodular cellulases

Numerous studies of multi-domain proteins involving flexible linkers are based on a combination of experimental and computational methods. Bimodular cellulases composed of covalently bound catalytic and cellulose-binding modules can be considered as a typical example. For instance, structural properties of a long disordered linker, containing 88 residues, in an artificial protein conceived from two natural cellulases were investigated by SAXS combined with molecular modeling tools [215]. More precisely, high-temperature MD simulations were applied as a conformational sampling technique, and a subset of the resulting models was selected to collectively fit the experimental data. Results of this study showed that the linker does not behave like a pure random coil, and suggest that the structural properties of the linker are essential for the function of these bimodular enzymes. Similar results have been observed in other studies combining SAXS and theoretical approaches [179, 17]. Moreover, bioinformatics analyses showed that sequence features are conserved in different families of bimodular cellulase enzymes, and suggest that the linker length has been evolutionarily optimized based on the type of the connected domains [181]. In this study, the authors also applied all-atom replica-exchange MD simulations together with circular dichroism to investigate the effects of glycosylation in the linker. Results of their analysis showed that the linkers are not rigidified by the addition of mono- or disaccharides, although they tend to adopt more extended conformations. Overall, this work demonstrated that linker length and composition is important for the activity of these enzymes, but a more clear description of functional roles remained to be elucidated. One of these roles was revealed by μs -scale all-atom MD simulations, showing that glycosylated linkers bind dynamically and non-specifically to the cellulose surface [155]. The predicted enhancement of binding affinity due to the linker was confirmed experimentally. The importance of the linker for the processivity in cellulases, as well as in other DLD enzymes, has been investigated using bioinformatics tools and a statistical kinetic model [195]. Results of this theoretical work suggested that processivity may result from the kinetic bias of binding due to spatial constraints imposed by the linker, which favors rebinding over full release of the substrate. They also show that the linker length and flexibility have been finely tuned through evolution to optimize this process. In Subsection 1.3.2, we describe the contribution of our methodology in a broader project aiming at optimizing the linker properties to increase the processivity and activity of an enzyme.

1.2.2 Interactions between disordered regions and structured domains

The interaction between IDPs/IDRs and their globular partners is very often mediated by SLiMs inserted into disordered chains [53] (see Section 1.1 for additional details about SLiMs). In the unbound form, SLiMs can be pre-structured, reducing the entropic cost of the interaction and, as a consequence, tuning its thermodynamics [51, 72]. The inherent flexibility enables a single SLiM to recognize multiple partners with different structures and affinities (Figure 1.2), with p53 being the most notorious example of this promiscuity [208]. Several

proteins contain successive SLiMs and can be perceived as molecular platforms that bring to proximity different proteins involved in the same metabolic or signaling pathway to form high-order molecular assemblies [221]. For instance, this capacity is exploited by nuclear receptor co-regulators to assemble a large number of proteins to trigger gene transcription (see below), or by viruses to hijack the eukaryotic translational machinery [52]. In this section, we will describe how computational methods have helped to understand SLiM recognition events. Then, we describe the architectures emerging when several adjacent SLiMs recognize one or multiple sites in the globular partner.

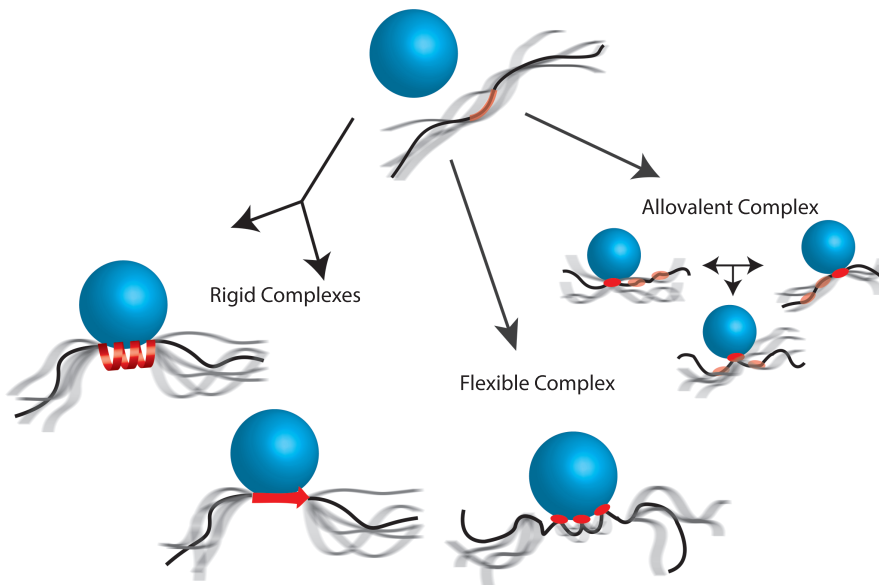


Figure 1.2: Representation of the different scenarios when an IDP interacts with a globular domain. Upon binding the interacting SLiM can adopt a rigid structure in a folding-upon-binding process. The final structure of this rigid segment, depending on the properties of the receptor, can be a canonical secondary structure (α -helix or β -strand), or adopt a coil conformation. Note that the interaction region of the IDP can be pre-structured in the unbound form, tuning the thermodynamic stability of the complex. The complex can also be dynamic, displaying multiple weak specific interactions that bind and unbind continuously while maintaining the overall architecture of the complex. Allovalent complexes occur when several SLiMs adjacently positioned in the chain can interact with a single receptor site and the bound conformation is continuously exchanging.

Modeling partner recognition by short linear motifs

MD simulations have emerged as a powerful tool to study binding modes of IDPs. MD simulations are especially well-suited when the recognition and binding to the partner is achieved by SLiMs since in these cases the computational effort can be reduced by simulating only a small fragment of the IDP. In many the cases, high-resolution structures of the bound form are available from X-ray crystallography or NMR. Alternatively, experimentally-assisted computational docking with programs such as FlexPepDock [166], HADDOCK [77, 39] or IDPLZerD [157] can be used to model the SLiM in the bound form. More recently, Machine-Learning (ML)-based approach AlphaFold2 [99] and its variants to model stable complexes [70] have been successfully used to model rigid complexes of IDPs with their partners [34]. MD studies of partner recognition by IDPs have primarily centered on discriminating between two

mechanistically different binding modes: *conformational selection*, when the preformed bound conformation is a requirement for binding, and *induced fit*, when the optimal conformation is only adopted upon binding. For instance, the structural ensembles of Gab2 in the unbound state as well as in complex with Grb2 were generated using MD simulation with NMR-derived backbone chemical shifts as restraints [113]. Interestingly, it was observed that the secondary structure elements involved in recognition and binding of the partner were already present in the unbound state as well, albeit transiently. Disruption of these secondary structure elements resulted in an affinity reduction, establishing Gab2-Grb2 interaction as a typical example of *conformational selection*. On the other hand, umbrella-sampling all-atom and coarse-grained MD simulations to study the binding between c-myc and KIX revealed a different scenario [95]. It was observed that the probability of crossing the transition state and the time required to do so did not depend on the structuration of c-myc at the beginning of the simulations, indicating that both unstructured and structured c-myc were capable of binding KIX with comparable rates. It was also noted that the transition state ensemble was heterogeneous with a wide diversity of c-myc conformations. A yet different mode of binding was observed for p53-MDM2 binding using very long unbiased MD simulations and Markov State Models (MSMs) [230]. In this case, binding almost always preceded folding, providing a classic example of ‘fly-casting’ followed by induced fit.

Another example of simultaneous binding and folding was described by Robustelli *et al.* for the interaction between the α -helical molecular recognition element (α -MoRE) of the intrinsically disordered C-terminal domain of the measles virus nucleoprotein (NTAIL) and the X domain (XD) of the phosphoprotein of the same virus using unbiased MD simulations [173]. As in the case of c-myc-KIX complex, the transition state was found to be highly heterogeneous. An interesting observation, however, was that if the α -MoRE formed long helices in the beginning of the binding event, it actually unfolded before forming the additional intermolecular contacts of the native conformation. This is in contrast to the conformational selection phenomenon observed for Gab2-Grb2. It was concluded that there was no clear temporal separation between binding and folding events as observed in other cases [154].

The above-described interactions can also occur intramolecularly if a disordered tail recognizes the globular domain to which it is attached (see Figure 1.1.B₁) [10, 131, 6]. Due to the concomitant increase of the C_{eff} , this architecture enables interactions that would have a very low affinity in an intermolecular scenario. Furthermore, the inherent flexibility of the resulting loop-like fragment between the the domain and the binding motif decreases the entropic cost of the interaction [5]. In Subsection 1.3.2, we will discuss the application of my ensemble generation approach to explicitly quantify C_{eff} in a flexible biomolecular system. An example of such intramolecular interaction is the auto-inhibition of DNA binding activity of Ets1 by its disordered C-terminal IDR having a Serine rich region (SRR) that can be phosphorylated [102]. Kasahara *et al.* used high-temperature canonical MD simulations to generate a wide range of structures which were then used to seed multi-canonical MD simulations for enhanced sampling. The simulations showed an increased number of contacts between the phosphorylated SRR and a helix in the core of the protein which is responsible for DNA binding, compared to non-phosphorylated SRR indicating a direct competitive mode of inhibition. Free energy surface analyses based on Principal Component Analysis (PCA) followed by clustering of conformations showed that these auto-inhibitory states existed in the non-phosphorylated state as well but their population was significantly increased upon phosphorylation due to alteration of the free energy landscape of Ets1.

Modeling allovalent complexes

Allovalent interactions occur when multiple similar (or equivalent) SLiMs are adjacently found in the same protein and interact with a partner with a single interaction site [148, 76] (see Figure 1.1.C₂). These polyvalent proteins enable a special type of fuzzy complex in which the different SLiMs alternatively recognize the partner and dynamically exchange their position from unbound to bound forms (Figure 1.2). This competition of weak interactions for the same binding site increases the overall stability of the complex through cooperative effects that cannot be accounted for by traditional thermodynamic models [109, 128]. The continuous binding-dissociation-rebinding processes are very difficult to model, hampering the deep understanding of the structural and kinetic signatures of allovalency.

The interaction of phosphorylated Sic1 (pSic1) with *cdc4* is the prototypical example of allovalent complex. Sic1 contains nine similar CDK phosphorylation sites spread along the chain that can interact with the Cdc4 [197]. Interestingly, the increase of the affinity is not linear with the number phosphorylated sites, and the K_d reaches the submicromolar range only in the presence of at least 6 of them [138]. This non-linear cooperative mechanism makes Sic1 extremely sensitive to the cellular level of the Cdk kinase [28]. Structural ensembles of Sic1 and pSic1 have been determined by combining NMR and SAXS data, which were integrated using the program ENSEMBLE [137]. A simplistic model of the allovalent complex was built by docking the ensemble of the unbound pSic1 to Cdc4 using the site-specific fraction of bound form determined by NMR and the crystallographic structure of Cdc4 with a model peptide. Although this model provides some insights into the binding mode, the thermodynamic and kinetic features of the complex remain elusive, requiring more advanced computational tools. MD simulations were performed to understand the allovalent recognition of a fragment of the nuclear pore complex (NPC) protein Nup135 and importin- β [134]. Like many other NPC proteins, Nup135 contains multiple FG dipeptides inserted in the sequence that, by weakly interacting with specific proteins, facilitate their translocation to the nucleus. Individual conformations of Nup135 derived from unbiased MD simulations were collected, mixed with importin- β and submitted to a 2 μ s MD simulation. The specific association of the two proteins was repeatedly observed along the trajectory, with the FG-repeats docking into previously identified binding pockets on the surface of importin- β [11]. Although the structural details of the FG recognition could be observed, the limited sampling hampered the extraction of the site-exchange kinetics and the evaluation of the differences between the alternative sites.

Modeling partner recognition by multiple different short linear motifs

A different scenario occurs when multiple adjacent SLiMs can recognize the same globular partner through different anchoring points. In these circumstances, the disordered chain forms a long flexible loop-like structure that connects the bound segments (see Figure 1.1.C₃). This recognition mechanism is often associated to cooperative binding through the increase of the C_{eff} of other SLiM(s) when one or more SLiM(s) are already bound. Note that this mechanism is similar to the case of disordered linkers connecting globular domains explained in Subsection 1.2.1. Organisms have developed these complex regulation mechanisms in order to modulate biological outputs. There are many examples of interactions that involve multisites, but very few of them have been structurally characterized. Thus, the complexity of the whole system, including the interplay of the different interacting regions, often remains undescribed.

Complexes involving disordered co-regulators and homo- or heterodimeric nuclear re-

ceptor (NR) that regulate gene transcription are prototypical examples of the C_3 scenario. The interaction motifs of co-activators and co-repressors, called NR-boxes, share LxxLL and LxxI/HLxxI/L consensus sequence, respectively. Intriguingly, co-regulators contain a different number of consecutive NR-boxes depending on the organism and can potentially recognize the two binding sites of the NR dimers. For this multisite binding, the balance between an asymmetric model, where a single NR anchoring point is occupied, and a deck model, where both anchoring points are engaged, will depend on the affinity of the individual NR-boxes and the effective concentration dictated by the number and distance between the SLiMs. For the specific case of NRs, local affinities are modulated by endogenous ligands. The complex between the co-repressor N-CoR_{NRID} with the RXR/RAR NR heterodimer could not be fully characterized at the residue level by NMR due to chemical exchange observed in the interacting regions [46]. In order to have a global picture of the complex, all-atom models of N-CoR_{NRID} were generated using Flexible-Meccano [16, 149] and docked to one site of the heterodimer using the crystallographic structure as a template, representing the asymmetric model. To represent the deck model, with the two NR-boxes simultaneously bound to the heterodimer, steered MD simulations were performed on some conformations of the asymmetric ensemble forcing the second NR site to dock on the other face of the NR. By comparing the averaged SAXS profiles computed from both ensembles with the experimental one, the relative populations of the two binding modes in the apo form and in the presence of NR ligands were determined [46]. For the case of co-activators, no detailed model of the complexes has been proposed, although the presence of simultaneous binding has been demonstrated [56, 174, 182]. Interestingly, for TIF2_{NRID} co-activator, NMR experiments highlighted the involvement of TIF2_{NRID} NR-box2 flanking region in its interaction with RXR/RAR heterodimer. The specific fragment encompassing NR-box2 and its flanking ordered region was co-crystallized with RAR bound to an agonist, and revealed an interacting helix turn helix motif of the TIF2_{NRID} fragment on the RAR surface [182]. The exact role of this flanking region in the recognition mechanism and the effects on the overall arrangement of the complex remain to be deciphered. Again, computational approaches should play a pivotal role to address these questions.

Another example concerns the interaction of a 60-residue long fragment of the tumor-suppressor p53¹⁻⁶⁰ with the metastasis-associated S100A4 protein through three anchoring points [61]. This study combined NMR data with MD simulations to determine the structure and dynamics of this fuzzy complex. The fact that the linkers between the three interaction motifs are short makes the modeling of the system less complicated. Indeed, the conformational sampling of long flexible loops connecting simultaneously bound SLiMs is one of the remaining challenges in the field. Although numerous methods have been reported for loop modeling in folded proteins [185, 152, 117], existing approaches mainly aim at predicting the most likely loop conformation(s) rather than exhaustively sampling the conformational space of the loop. Moreover, only a few of these methods remain computationally efficient when the loop length exceeds 15 residues. One of them is a robotics-inspired method that exploits a large structural database of three-residue fragments [9]. First tests with this method applied to IDPs show its ability to rapidly generate conformational ensemble models of loops involving around 100 residues (unpublished work).

1.2.3 Extreme fuzzy complexes and phase separation behavior

Several IDPs can also interact with each other. The association can give rise to highly disordered complexes [219] (illustrated in Figure 1.1.D₂ and D₃) or to rigid particles, such as amyloids (Figure 1.1.D₁). In this last case, large aggregates are formed by the perfect

arrangement of chains stabilized by a dense network of hydrogen bonds. This case will not be described here, and the reader is referred to other publications [84, 111, 94]. At the other extreme of flexibility, recent publications describe the formation of high-affinity complexes between two IDPs that retain their flexibility upon binding [29, 223]. For the case of Borgia *et al.* [29], this new kind of biomolecular interaction can be explained by the large opposite electrostatic charges of the two proteins, histone H1 and its nuclear chaperone prothymosin- α . The integration of NMR and smFRET data into one-bead-per-residue CG simulations unveiled that the complex was maintained by multiple long-range electrostatic interactions without the need for defined binding sites and specific interactions. Interestingly, ternary complexes displaying a high exchange rate are formed at high concentrations [190]. The lack of specificity in the interactions causes this phenomenon and triggers the formation of large oligomers, a phenomenon that is reminiscent of liquid-liquid phase separation (LLPS).

Multiple pieces of evidence indicate that dynamical, multivalent interactions between IDRs/IDPs are major drivers of cellular LLPS processes and provide the structural scaffold for the so-called membrane-less organelles [14, 33]. Remarkably, the structural and functional characterization of these condensates is attracting ever-growing attention since they are currently recognized to play a major role in organizing cellular biochemistry [186, 24]. Computational approaches have the potential to play a key role in this challenge, given the difficulties in tackling the daunting complexity of these biomolecular assemblies with standard structural biology techniques and/or polymer physics theories [33, 178]. In particular, molecular simulations can provide access to elusive structural details of the condensates and complement theoretical and experimental investigations of the molecular grammar governing LLPS [123, 218, 132] with the final aim of establishing sequence-structure-function relations.

Not surprisingly, the length and time scales associated with cellular LLPS, which are collective processes involving intermolecular interactions among a large number of large-sized biomolecules, have favored the development and applications of suitable CG molecular models. In this respect, CG models based on one-bead-per-residue description have been shown to provide a reasonable compromise of accuracy and computational efficiency and are a popular choice for simulating the LLPS equilibria of flexible proteins [20, 178, 103, 23, 184]. Most accurate versions of these models explicitly take into account the protein sequence and rely on inter-residue energy functions that include implicit-solvent Debye-Hückel electrostatics and contact potentials accounting for excluded volume and short-range attraction. The latter terms are defined according to hydrophobicity scales or statistical potentials and tuned to reproduce experimental structural data or affinities [104, 57]. While the quantitative predictive capabilities of one-bead-per-residue potential should not be overstated [50], this approach has been successfully applied to shed light on how the protein sequence determines the phase behavior of IDPs, as well as the structural and dynamical properties of condensed phase [57, 132, 89]. Furthermore, CG simulations at this level of resolution can be easily extended to include folded domains [43], post-translational modifications [141], thermoresponsive behavior [58] and interactions with RNA molecules [170]. Moving to higher-resolution models, a recent study indicated that the popular MARTINI CG force-field, which relies on a four-atoms to one-bead mapping and an explicit solvation model, can accurately describe the condensation of FUS prion-like domain, upon a fine tuning of its energy function against experimental transfer free-energies [12]. Conversely, ultra-coarse grained simulations, where a single bead may represent a protein domain or an entire biomolecule, have been successfully applied to get some insight into the internal organization of multi-component mixtures that mimic more closely the complexity of cellular condensates [87, 42, 144, 65].

So far, the role of atomistic MD in this field has been rather limited due to the demanding computational requirements of this approach, which make the direct simulation of phase separation processes extremely challenging with present-day computational resources [86]. Nevertheless, recent studies have indicated novel strategies to take advantage of all-atom, explicit-solvent MD simulations based on accurate last-generation force fields in the characterization of biomolecular LLPS. Notably, MD simulations of protein fragments at high-concentration were used to dissect the molecular interactions driving the LLPS with a “divide-and-conquer” strategy and they provided results in good agreement with NMR and mutagenesis data with a limited computational cost [150]. Furthermore, a high-resolution picture of protein dynamics in the condensed phase was obtained by generating an initial CG configuration of phase-separated proteins, which was then mapped back to all-atom resolution and simulated in the microsecond timescale thanks to a specialized supercomputer [229].

1.2.4 Emergence of novel approaches and machine learning to model protein disorder

In the last few years, innovative and promising methods have emerged to generate conformational ensembles of biomolecular systems. Some strategies combine molecular dynamics (MD) simulations with stochastic methods, to overcome inherent limitations of MD simulations [162]. Concretely, these strategies exhaustively sample small fragments of IDPs using MD simulations and, in a second step, stochastically selected fragments are connected similarly to Flexible-Meccano, to generate conformations of the whole system [191, 161]. These conformations can be used to initialize MD simulations aiming to refine the exploration in some areas of the space.

IDPConformerGenerator [198, 228] also relies on fragment-by-fragment construction using dihedral angles to generate conformational ensembles of IDPs. More recently this method evolved with a new module of Local Disordered Region Sampling (LDRS), which attempts to tackle the challenge of generating IDR ensembles [127]. While this addition represents an advancement in the capabilities of the method, it still faces limitations in terms of the diversity of systems it can handle and the computational efficiency. The inherent complexity of IDRs generally necessitates specialized methods for optimal results, tailored for specific systems, such as those focusing on multi-loop regions [196].

One of the most impactful advancements in structural bioinformatics in recent years is the emergence of AlphaFold2 [100, 99]. This freely available tool, developed by Google DeepMind and powered by deep learning techniques, has achieved remarkable success in the field of protein structure prediction [188, 34]. The growing importance of machine learning for solving structural biology problems is also implying a major change in the way IDP/IDR modeling is approached [124]. Notably, several recent methods use machine learning and deep learning to predict IDP conformational ensembles, such as idpSam [97], IDPFold [234], and Phanto-IDPs [233]. Machine learning seems particularly well-suited for loop modeling, with diverse methods exploring this type of approaches during the last years [151] [9]. In Chapter 4 of this manuscript, I will present a Reinforcement Learning (RL) based method to enhance conformational sampling of disordered regions.

1.3 Contributions and structure of the thesis

The biological relevance of IDPs/IDRs underlines the importance of having detailed structural models of this class of proteins and their complexes. These models guarantee a molecular

perspective of key cellular processes and eventual rational interventions with pharmacological aims [3]. The co-existence of an astronomical number of conformations and the averaged nature of the experimental data that can be recorded for IDPs make the use of computational methods unavoidable. The immense challenges in the field are exemplified in the study of liquid-like droplets, which have attracted the interest of a large community from diverse scientific domains [24, 2]. These highly concentrated protein condensates are inherently disordered and display multivalent, weak intermolecular interactions that are modulated by external parameters such as pH, temperature or phosphorylation states [156]. Therefore, they present multiple challenges for computational modeling.

The growing interest of the structural bioinformatics community to overcome challenges posed by IDPs/IDRs is encouraging. The improvement in the force fields, for both all-atom and CG simulations, to adapt them to disordered states, the development of enhanced sampling strategies, as well as the generalization of parallelized software and the use of GPUs are the most prominent hints of these developments. The increase in the number of experimental studies focusing on IDPs/IDRs is also crucial as they continue identifying novel biological mechanisms. Moreover, databases and repositories assembling experimental and omics data improve our structural and functional knowledge of these proteins, and provide new opportunities to develop and validate the theoretical methods [88, 120]. This new data is rich in information and can be used, for instance, to improve current force fields, or can be exploited to conceive more accurate conformational sampling methods [67]. The use of data mining and machine learning methods to analyze and exploit relevant information from these databases is a very promising avenue for the improvement of predictive molecular modeling approaches and for the development of new tools to tackle the challenging questions posed by disordered proteins and their complexes.

In this context, this thesis introduces MoMA-FReSa (Molecular Motion Algorithms - Flexible Regions Sampler), an innovative method to explore the conformational states of disordered biomolecular entities. As highlighted in Section 1.2, existing methods often cater to specific types of systems and/or are very computationally expensive. MoMA-FReSa, however, aspires to be a more general and computationally accessible tool, capable of exploring the vast conformational space of a wide range of systems and to fill the critical gap in the field of IDP/IDR modeling. In addition, we aim to have the capacity to model the most complex biomolecular systems, which can not be studied by MD. Globally, our primary objective is to provide researchers with a versatile tool that transcends the limitations of current methods.

By providing a detailed roadmap for the manuscript, the following subsections outline the structure of this manuscript and highlight the main results and contributions of this thesis.

1.3.1 Chapter 2: An exhaustive description of MoMA-FReSA

MoMA-FReSa is a novel all-atom stochastic sampling method designed for the exploration of conformational space of IDPs and IDRs. It builds upon established methods such as Flexible-Meccano [149] [16] by utilizing dihedral angles of amino acids to generate ensemble models. To develop a more versatile aspect, MoMA-FReSa broads some aspect of these traditional methods. For example, contrary to Flexible-Meccano [16], the three backbone dihedral angles are considered (not only ϕ and ψ) and the two sampling directions are considered according to the systems (not only N-to-C).

Our sampling approach aligns with recent methods also based on the construction of the backbone through the selection of dihedral angles as IDPConformerGenerator [198]. While IDPConformerGenerator also introduced the Local Disordered Region Sampling (LDRS) mod-

ule for generating IDR ensembles [127], its limitations in the diversity of systems that can be addressed and its computational efficiency highlight the need for a more general and optimized method.

MoMA-FReSa aligns with the philosophy of the MoMA suite (<https://moma.laas.fr/>) developed at the *Laboratoire d'Analyse et d'Architecture des Systèmes of the Centre National de la Recherche Scientifique* (LAAS-CNRS). This approach particularly follows the work of Alejandro Estaña [67] on three-residue fragment approaches, Amélie Barozet [9] on loop modeling. The selection of dihedral angles relies on a database of three-residue fragments [68] extracted from experimentally determined protein structures [73].

The first objective of MoMA-FReSa is to establish a unified algorithm capable of handling various biomolecular systems with disordered parts. To achieve this, the method employs a pre-processing step that hierarchically decomposes the system into sub-regions of amino acid residues. This initial stage not only grants MoMA-FReSa versatility but also optimizes the subsequent sampling process. During this sampling phase, MoMA-FReSa acts as a state-space search algorithm aiming to find a feasible conformation by a nondeterministic problem-solving approach. In effect, to optimally manage the exploration of the wide space of the possible conformations, MoMA-FReSa employs a hierarchical stochastic process.

Chapter 2 presents the method of MoMA-FReSa and provides a preliminary assessment of its capabilities. The chapter demonstrates the versatility of MoMA-FReSa by applying it to a diverse benchmark of systems. The results in this chapter provide convincing evidence of the potential of MoMA-FReSa.

1.3.2 Chapter 3: Collaborations and applications

Following the establishment of the methodology and capabilities of MoMA-FReSa in the previous chapter, Chapter 3, we apply the program to multiple cases. This chapter showcases the versatility of MoMA-FReSa beyond traditional sampling through various collaborative projects:

- **Conception of Multi-Modular Systems through the iGEM Competition:** MoMA-FReSa was used for the design of a multi-modular protein in the context of the iGEM (International Genetically Engineered Machine) synthetic biology competition. A team from *Toulouse Institut National des Sciences Appliquées - Université Paul Sabatier* (INSA-UPS) required MoMA-FReSa in the development of CALIPSO, a targeted drug delivery system designed for cancer treatment (<https://2023.igem.wiki/toulouse-insa-ups/home>).
- **Estimating Effective Concentrations in Molecular Interactions:** A collaboration was initiated with the group of Dr. Lucia Chemes at the University of San Martin (Argentina) to estimate effective concentrations in intramolecular and intermolecular interactions with MoMA-FReSa. This initial collaboration focused on a particularly challenging system [164].
- **Structural Analysis within the CORNFLEX Project:** The capabilities of MoMA-FReSa for structural analysis were explored in the context of the CORNFLEX project, a collaborative research initiative involving multiple French institutions (<https://anr.fr/Project-ANR-22-CE45-0003>). This collaboration also provided a valuable opportunity to illustrate some of the post-processing analysis capabilities of MoMA-FReSa.

- **Sampling of Structural Motifs by studying Ethylene Receptor 2 in *Vitis vinifera*:** In the context of a collaboration with students at the *École Nationale Supérieure Agronomique de Toulouse* (INP-ENSAT), we applied MoMA-FReSa to study a flexible region of the second ethylene receptor (ETR2) of *Vitis vinifera* [41]. Through this study, we addressed the challenge of modelling structural motifs. This work also served to demonstrate the diverse conformational sampling methods offered by MoMA-FReSa.

1.3.3 Chapter 4: A deep learning approach to enhance MoMA-FReSa

Finally, Chapter 4 explores an improved version of MoMA-FReSa, integrating a Reinforcement Learning (RL) model-free method [163, 183]. Artificial intelligence (AI) techniques have become increasingly powerful tools in protein modeling and design. Numerous sampling methods now incorporate machine learning elements or are fully based on deep-learning approaches [228, 234, 97, 233]. Following this promising trend, we worked of a machine learning approach to improve the capabilities of MoMA-FReSa. Among all the machine learning approaches, RL was well adapted because this method does not require experimental data for learning. Instead, learning is performed simultaneously to conformational sampling.

The primary target was to improve sampling of systems containing loops. In effect, modeling loops is challenging. Even if modeling method for multi-loop have emerged [196], most of the actual loop sampling methods focus on the modeling of single-loop proteins. MoMA-FReSa has the capacity to sample systems involving one or multiple loops as well as other flexible regions, as described in Chapter 2. However systems involving loops are the most time-consuming ones for MoMA-FReSa. Due to the significant challenges they represent for traditional methods [151], loops are particularly well-suited for machine learning approaches. In the continuity of precedent methods [9], we implemented a custom RL method specifically tailored to our approach. However, the objective of this implementation extends beyond loops ; we aim at improving the performance of MoMA-FReSa across all types of system types, especially those with high complexity.

Our approach uses a Curiosity-Driven Reinforcement Learning strategy [74]. Our implementation focuses on dihedral angles and, more specifically, secondary structure propensities. To do this, MoMA-FReSa benefits from a special sampling approach, which considers the structural type of the amino acid residue (α , β , or γ) during the sampling process. While the initial results presented in Chapter 4 are not yet conclusive, we remain optimistic about the potential of the implemented model. We believe that further development holds promise in transforming this addition into a valuable tool. This work is ongoing, and several possibilities for improvement are currently being explored.

1.3.4 Software availability

MoMA-FReSa was developed with the goal of becoming a widely accessible tool for the scientific community. While a fully functional version has already been used in collaborative research in Chapter 3, no operative version is currently available for public download.

Development is ongoing and the creation of an intuitive user interface built in Python is the current primary focus. This user interface will serve as the foundation for a web interface, ultimately leading to a web application freely accessible to researchers. Additionally, plans are in place to offer a freely available Apptainer container for users who prefer a local installation. We will discuss this ongoing work in more detail in the Conclusions of the thesis.

MoMA-FReSa: A conformational sampling method for complex disordered systems

Contents

2.1	Introduction	17
2.2	Method	18
2.2.1	Method overview	18
2.2.2	System representation through the pre-processing stage	21
2.2.3	Pre-processing construction steps	30
2.2.4	Sampling principle	34
2.2.5	Main sampling process	35
2.2.6	Post-processing in MoMA-FReSa and beyond	45
2.2.7	Implementation details	47
2.3	Results	48
2.3.1	Benchmark and protocol overviews	48
2.3.2	Analysis of the Results	51
2.3.3	Study of the effect of the new distance test for loops	58
2.4	Discussion and conclusion	59

2.1 Introduction

In Chapter 1, we introduced MoMA-FReSa, a novel computational method designed to address the limitations of current structural modeling approaches for proteins with disordered regions. While biophysical techniques like X-ray crystallography, Nuclear Magnetic Resonance (NMR), and cryo-electron microscopy (cryo-EM) provide precise structural models for folded proteins, Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs) pose a significant challenge. Experimental methods such as NMR, Small-angle X-ray scattering (SAXS) or Förster resonance energy transfer (FRET) offer valuable insights into the ensemble of conformations adopted by IDPs/IDRs in solution, but these measurements are not sufficient to provide unambiguous atomistic representations without the coupling with computational methods.

Given the critical biological roles played by disordered regions, understanding their structural behavior is a current and challenging objective. MoMA-FReSa is a valuable tool in this context, as it can generate large ensembles of conformations of complex biomolecular systems, which can be subsequently refined using experimental data. Notably, the key strength

of MoMA-FReSa lies in its ability to model a wide range of molecular systems containing disordered regions, outperforming previous methods in this regard.

In this chapter, we describe the method of MoMA-FReSa in Section 2.2, highlighting how the innovative combination of a pre-processing strategy and a sampling process leads to an efficient exploration of the conformational space across a vast range of systems. Section 2.3 shows the performance of MoMA-FReSa using a carefully chosen benchmark set. Finally, we discuss the obtained results and mention potential future directions for MoMA-FReSa in Section 2.4.

2.2 Method

MoMA-FReSa is an all-atom conformational sampling method specifically designed for complex biomolecular structures. To effectively handle the diverse challenges presented by disordered biomolecular systems, MoMA-FReSa employs a structure decomposition and hierarchization procedure. This pre-processing strategy decomposes each system into smaller sub-regions of amino-acids, making the overall process more manageable.

Key to the operation of MoMA-FReSa is this region-oriented approach. The core of the method lies in defining and scheduling these regions in a pre-processing stage, a process referred to as hierarchical scheduling. Then, established all-atom sampling methods are effectively utilized within individual regions. This hierarchical approach allows MoMA-FReSa to efficiently sample various types of biomolecular assemblies involving disordered regions, since the pre-processing construction effectively mitigates the inherent complexity of such systems.

The following subsections explain in detail the different components of MoMA-FReSa. First, Subsection 2.2.1 gives an overview of the method, from the pre-processing to the post-processing, through the core method. A given example called *Illustrative Example* is defined to illustrate the following subsections. Sections 2.2.2 and 2.2.3 detail the pre-processing procedure, starting with a theoretical foundation and then expressing its practical application. In Subsection 2.2.4, we give an overview of the key elements of the sampling strategy, and present MoMA-FReSa as a nondeterministic problem-solving approach. Subsection 2.2.5, goes deeper in this process and provides a step-by-step description of the core functionalities of the sampling method. Then, Subsection 2.2.6 focuses on the post-processing analysis steps. Finally, Subsection 2.2.7 introduces the various fundamental tools employed by the sampling procedure.

2.2.1 Method overview

MoMA-FReSa, a novel technique for exploring the structural landscapes of proteins involving highly-flexible (disordered) regions, employs a hierarchical strategy. This subsection details the objectives, target systems, and core principles of MoMA-FReSa. It also introduces an *Illustrative Example*, used as a guide throughout this chapter. Then, the subsection provides an overview of the innovative pre-processing decomposition of the method, fundamental residue-centered sampling techniques, and post-processing capabilities.

Objective and system definition

MoMA-FReSa is designed as a comprehensive tool for sampling IDPs and proteins containing IDRs. It aims to generate a large and unbiased ensemble of conformations, covering the maximum of the accessible conformational space, within a reasonable time frame. This method

is applicable to a wide range of systems, with one or multiple polypeptide chains and with diverse disorder arrangements.

The core principle of MoMA-FReSa lies in the segmentation of the protein system into distinct amino-acid regions. These regions are classified as either flexible (corresponding to IDRs) or rigid (ordered/globular regions). The sampling then focuses on the flexible regions. Indeed, only disordered regions, lacking a predefined structure, require sampling.

The primary goal of MoMA-FReSa can be redefined as follows: to efficiently generate a conformation for the complete system respecting the constraints imposed by both rigid and flexible regions. Ensembles can then be generated by applying this method multiple times. Compared to conventional methods, MoMA-FReSa addresses two key challenges: (i) an efficient sampling to generate large ensembles of conformations in a reasonable time and (ii) a general method applicable to a large variety of disordered biomolecular systems.

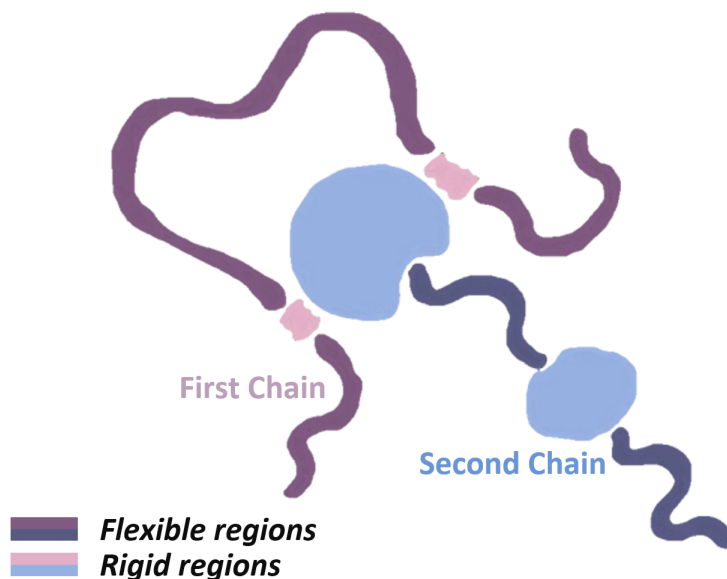


Figure 2.1: *Illustrative Example* system. One chain (pink) bound by two short motifs to another multi-domain chain (blue).

Figure 2.1 introduces a complex system composed of two interacting protein chains. These chains contain multiple disordered regions and are connected through two specific binding motifs. The diverse nature of the flexible regions, the geometric constraints imposed by the interaction, and the requirement to maintain specific interactions during sampling present significant challenges for conventional sampling methods. This system, called *Illustrative Example*, is used in the following subsections as a guiding case study to demonstrate the effectiveness of the proposed pre-processing procedure in simplifying complex sampling problems.

Pre-processing stage

Effective management of sampling all flexible regions in MoMA-FReSa needs a crucial pre-processing stage. This involves two key components:

- **System Decomposition:** The protein system is segmented in distinct regions based on biological principles. This decomposition ensures an accurate representation of all system-specific structural features.

- **Hierarchical Scheduling:** The obtained flexible regions are prioritized for sampling. The obtained flexible regions are sampled in parallel as much as possible, sometimes respecting a priority order between them. The implementation of a predefined sampling priority order when necessary ensures effective and broad applicability of MoMA-FReSa to various protein systems of growing complexity.

These decomposition and hierarchization processes are the two fundamental aspects of the pre-processing stage, and the two key steps of the sampling procedure, which are described in Subsections 2.2.2 and 2.2.3, respectively.

A core sampling method applied to disordered residues

The core sampling procedure is employed regardless of the specific protein structure, as all the structural specificities are captured by the decomposition and the scheduling pre-processing work. Thus, the core sampling procedure should be as general as possible, able to be applied to all type of systems with a given decomposition in regions and the associated scheduling obtained thanks to the pre-processing stage. The general approach for this global sampling process is detailed in Subsection 2.2.4, followed by a deeper description of the method in Subsection 2.2.5.

Contrary to the pre-processing stage, which establishes a foundation specific to the chosen system and only needs to be performed once, the core sampling procedure is iterative. This means the main step is executed multiple times until a desired number of conformations is generated.

Sampling a region essentially means sampling the conformations of all its residues. Among all types of molecular representations explored in Chapter 1, MoMA-FReSa adopts a basic one where bond lengths and angles are fixed, requiring only three dihedral angles (ϕ , ψ , and ω) for each residue, as shown in Figure 2.2. In this sampling procedure, only the backbone placement is studied and the placement of the side chains is an object of the post-processing stage.

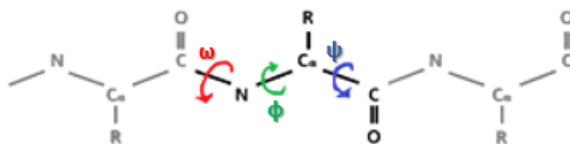


Figure 2.2: Illustration of the three dihedral angles of a residue in a polypeptide chain.

The core objective of the algorithm is to assign a specific value to each of the three dihedral angles for every flexible residue. This selection defines the conformation of each residue and ultimately leads to the complete conformation of the entire biomolecule.

Post-processing operations

Beyond its pre-processing and sampling methods, MoMA-FReSa offers a comprehensive suite of post-processing tools. These tools cover fundamental tasks like side-chain placement and energy calculations or fill more complex protein sampling analysis needs, such as contact assessment. MoMA-FReSa can also generate various reports for deeper analyses of the investigated system, such as reports on contact assessments to evaluate user-specified bound

pairs or energy computations. All these capacities are detailed in Subsection 2.2.6. In addition, some illustrations of the alliances of MoMA-FReSa with these post-processing tools are explored deeper in Chapter 3.

2.2.2 System representation through the pre-processing stage

MoMA-FReSa applies a pre-processing stage that decomposes the system into regions based on the intrinsic flexibility of consecutive residues. This decomposition is a relatively straightforward process. However, the challenge lies in two aspects: (i) accurately classifying regions based on their flexibility and capturing their properties, and (ii) defining an optimal sampling order between these regions for an efficient conformational exploration of the complete system. This subsection develops these challenges using a designated *Illustrative Example* as a reference. The result of the pre-processing is a signed adjacency matrix that encodes the intricate relationships between the regions.

A region-oriented decomposition

The elements of the biomolecular system S are broken down to generate a specific tree, shown in Figure 2.3. The initial division follows established biochemical principles: protein polypeptide chains (abbreviated chains) composed by amino-acid residues. For a system S with l chains the set of chains is noted $\{C_i, \forall i \in \llbracket 1, l \rrbracket\}$. MoMA-FReSa is versatile and can handle a wide range of systems, including single-chain proteins, multi-chain proteins, and protein complexes. Regardless of the complexity of the system or the origin of the polypeptide chains (whether from a single protein or different proteins within a complex), MoMA-FReSa treats all chains on the same level of the tree.

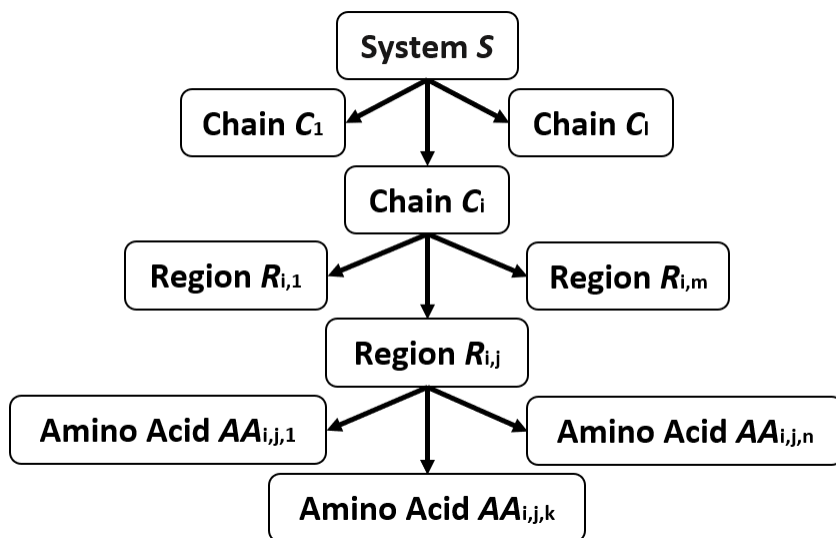


Figure 2.3: Structure tree in MoMA-FReSa developed for one example of each level

While protein chains are defined by their amino-acid sequences (called residues here), MoMA-FReSa introduces an intermediate level between chains and residues. The model splits each chain into consecutive segments of residues called regions. For a chain C_i of m regions, the ensemble of regions is noted $\{R_{i,j}, \forall j \in \llbracket 1, m \rrbracket\}$. Each region $R_{i,j}$ is a fragment containing n residues noted $\{AA_{i,j,k}, \forall k \in \llbracket 1, n \rrbracket\}$. The purpose of this decomposition is to

group consecutive disordered residues (noted $\tilde{A}A_{i,j,k}$) into distinct disordered/flexible regions (noted $\tilde{R}_{i,j}$).



Figure 2.4: Illustration of how a chain C_i is split in different regions where R represents globular regions and \tilde{R} the disordered ones.

As shown in Figure 2.4, MoMA-FReSa identifies flexible regions containing at least three consecutive disordered residues. The remaining residues are grouped into rigid regions. This ensures complete dissociation: neighbours of flexible regions are always rigid regions, and *vice versa*. In this decomposition, all regions $R_{i,j}, \forall j \in \llbracket 1, m \rrbracket$ can have up to two flanking regions: the N-terminal one $R_{i,j-1}$ if $j > 1$ and the C-terminal one $R_{i,j+1}$ if $j < m$. $R_{i,1}$ is the N-terminal region of C_i and $R_{i,m}$ is the C-terminal one. For a given (i, j) with $1 < j < m$, $AA_{i,j-1,n_{j-1}}$, the last residue of the n_{j-1} -residue long region $R_{i,j-1}$, is the N-terminal neighbour of $AA_{i,j,1}$, the first residue of $R_{i,j}$; $AA_{i,j+1,1}$, the first residue of the region $R_{i,j+1}$, is the C-Terminal neighbour of AA_{i,j,n_j} , the last residue of the n_j -residue long region $R_{i,j}$.

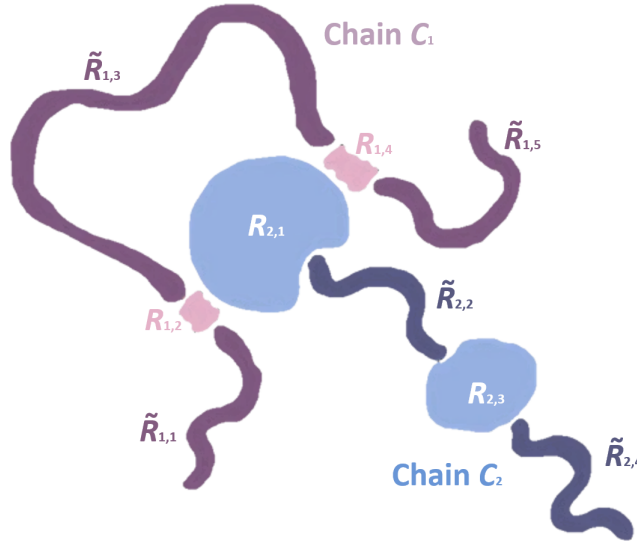


Figure 2.5: *Illustrative Example* system update with the regions split. One chain with five regions bound to another chain with four regions. R represents globular regions and \tilde{R} the disordered ones.

Figure 2.5 shows the *Illustrative Example* defined in Subsection 2.2.1 with the new notation. The first polypeptide chain is composed by a central disordered region $\tilde{R}_{1,3}$, two terminal flexible regions $\tilde{R}_{1,1}$ and $\tilde{R}_{1,5}$ and the two rigid interacting domains $R_{1,2}$ and $R_{1,4}$. The other chain is constituted by two rigid domains $R_{2,1}$ and $R_{2,3}$ connected by a disordered region $\tilde{R}_{2,2}$, and a C-terminal flexible region $\tilde{R}_{2,4}$. In a first step, MoMA-FReSa assigns each of the regions and builds the tree for this complex, displayed in Figure 2.6.

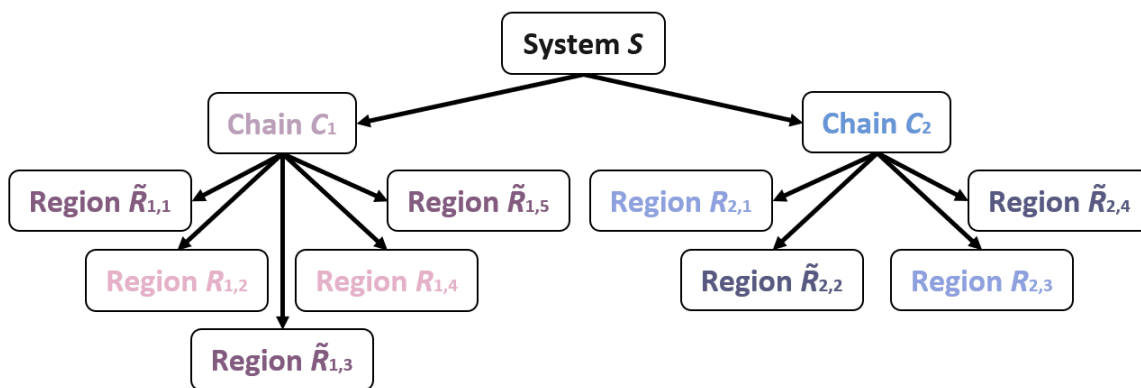


Figure 2.6: Structure tree of the *Illustrative Example* without the lower level where R represents globular regions and \tilde{R} the disordered ones

Region organization and graph construction

As seen previously, MoMA-FReSa can split each system S in regions alternating flexible and rigid regions. The system S can be expressed as a simple graph of regions where each region acts as a vertex within this graph, connected to its neighbouring regions by edges. Figure 2.7 illustrates this concept for the *Illustrative Example*. This visual representation employs distinct vertex shapes to differentiate between flexible and rigid regions, providing a clear distinction between their roles.

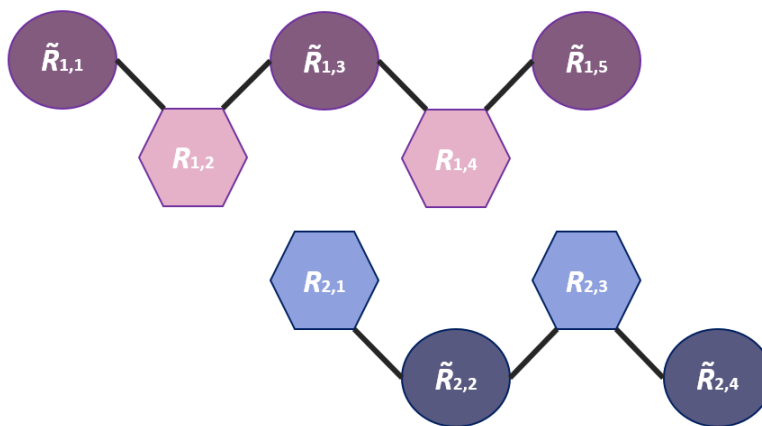


Figure 2.7: Primary simple graph of the *Illustrative Example* structure where R represents globular regions and \tilde{R} the disordered ones.

Rigid regions interactions and graph update

While the simple graph representation effectively captures the connectivity between regions, it fails in depicting all potential interactions within the system. Rigid regions can interact with each other, forming constraints that are not reflected in this basic representation. To address this limitation, MoMA-FReSa introduces the concept of a bound relation \mathfrak{B} between two rigid regions, denoted as $R_1 \mathfrak{B} R_2$. This relation implies a geometric constraint where the relative positions of R_1 and R_2 are fixed due to intra-system interactions. The \mathfrak{B} relation adheres to the following properties:

- **Rigid Region Specificity:** The \mathfrak{B} relation applies exclusively to rigid regions: if $R_1 \mathfrak{B} R_2$, R_1 and R_2 must be rigid.
- **Symmetry:** The \mathfrak{B} relation is symmetrical: $R_1 \mathfrak{B} R_2 \iff R_2 \mathfrak{B} R_1$.
- **Transitivity:** The \mathfrak{B} relation is transitive: if $R_1 \mathfrak{B} R_2$ and $R_2 \mathfrak{B} R_3$ so $R_1 \mathfrak{B} R_3$.
- **Irreflexivity:** While the combination of symmetry and transitivity might suggest elements in relation with themselves (i.e., $R_i \mathfrak{B} R_i$), these valid relationships are ignored in the context of this method.

By incorporating the \mathfrak{B} relation, MoMA-FReSa enhances the graph representation, enabling the modeling of geometric constraints arising from interactions between rigid regions. The method considers all different \mathfrak{B} relationships between pairs of distinct rigid regions. This enriched model provides a more comprehensive picture of the structure of the system.

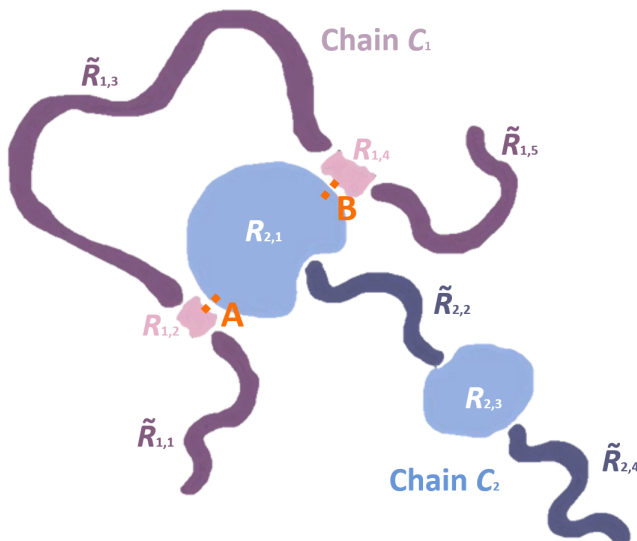


Figure 2.8: *Illustrative Example* system updated with apparent bonds.

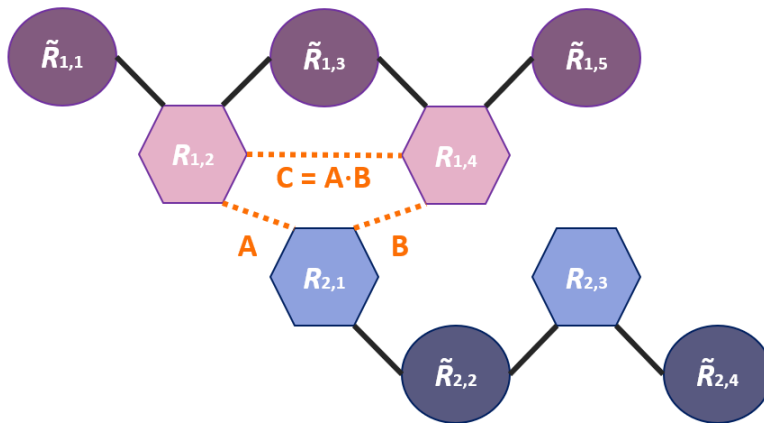


Figure 2.9: *Non-oriented System Graph* of the *Illustrative Example* where the relationships generated by the relation \mathfrak{B} are displayed.

The chain C1 of the *Illustrative Example*, as described in Subsection 2.2.1, has two binding sites interacting to the same rigid region, $R_{2,1}$, as illustrated in Figure 2.8. These two relationships are named $A = R_{1,2}\mathfrak{B}R_{2,1}$ and $B = R_{2,1}\mathfrak{B}R_{1,4}$. By the property of transitivity, a third resulting geometry relationship is also considered: $C = R_{1,2}\mathfrak{B}R_{1,4} = A \cdot B$.

As illustrated in Figure 2.9, a second type of edge is added to the graph to represent the relation \mathfrak{B} between rigid regions of the system. In fact, the resulting graph, named *Non-oriented System Graph*, is the addition of two simple graphs, one based on the split of the chains in regions and one based on the interaction of rigid regions.

Flexible regions and graph orientation

With all types of relationships properly defined in the *Non-oriented System Graph*, the next step involves establishing an order of sampling among these regions. While the method prioritizes parallel sampling of flexible regions for efficiency, ensuring their correct spatial positioning is crucial. If flexible regions were sampled without the proper context, their movement could lead to collisions and render previous sampling steps obsolete. Consequently, some flexible regions must wait for the prior sampling of others, potentially leading to a change in their spatial arrangement.

Each flexible region can have one or two starting residues (first residues potentially sampled within the region). If a starting residue has a neighbouring rigid residue, its position is defined with respect to the position of this rigid region. In MoMA-FReSa, the global position of a flexible region is determined by the combined global positions of the rigid neighbours of its starting residues. To optimize the overall process, MoMA-FReSa restricts the sampling of flexible regions to those for which rigid neighbours of their starting residues have already been positioned.

Based on the *Non-oriented System Graph*, another graph named *Oriented System Graph* is created. This graph uses directed edges between consecutive regions of the chains to express the order of sampling. The orientation definitions are:

- $R_1 \rightarrow \tilde{R}_2$: The rigid region R_1 needs to be fixed at a given position before sampling \tilde{R}_2 .
- $\tilde{R}_1 \rightarrow R_2$: The flexible region \tilde{R}_1 needs to be successfully sampled before placing R_2 .

It is important to note that the *Oriented System Graph* remains non-oriented when it comes to rigid region interactions due to the symmetrical nature of the \mathfrak{B} relation. In terms of hierarchical organization, the \mathfrak{B} relationship forces the simultaneous and definitive spatial placement of all connected rigid regions.

Rigid regions are most of the time considered only as “obstacles” placed at the beginning of the procedure or during the sampling, adding structural constraints during the sampling process. Flexible regions, which are the core of the problem and whose sampling is the final goal of MoMA-FReSa, can be classified in the following types:

- **Tail:** Tails are flexible regions placed at one of the termini of polypeptide chains. Their sampling starts from its flanking rigid region. As shown in Figure 2.10, their representation in the *Oriented System Graph* follows the rule: $R \rightarrow \tilde{R}_T$.
- **Loop:** Loops are flexible regions positioned between two rigid regions. The positions of these two rigid domains are fixed during the sampling of the loop. This represents a sampling step between two already-placed rigid regions. As shown in Figure 2.10, the rule is $R \rightarrow \tilde{R}_L \leftarrow R$.

- **Linker:** Linkers are flexible regions located between two rigid domains whose relative position is not fixed before the conformational sampling. One of the rigid regions of the chain is fixed before the sampling, providing a direction for building the linker (from the fixed region to the other rigid neighbour). This direction is an arbitrary notion of the program and have no consequences on the possible conformations generated by MoMA-FReSa. The linker places its following rigid region (in this given direction) after a successful sampling of all its flexible residues. The direction is generally chosen in an optimal way to reduce sampling cost and duration. The orientation for linkers is $R \rightarrow \tilde{R}_K \rightarrow R$, as shown in Figure 2.10.
- **Pure IDP:** In intrinsically disordered proteins (IDPs) the entire chain is a single flexible region that needs to be sampled. There are no flanking regions and consequently no directed edges in the graph.

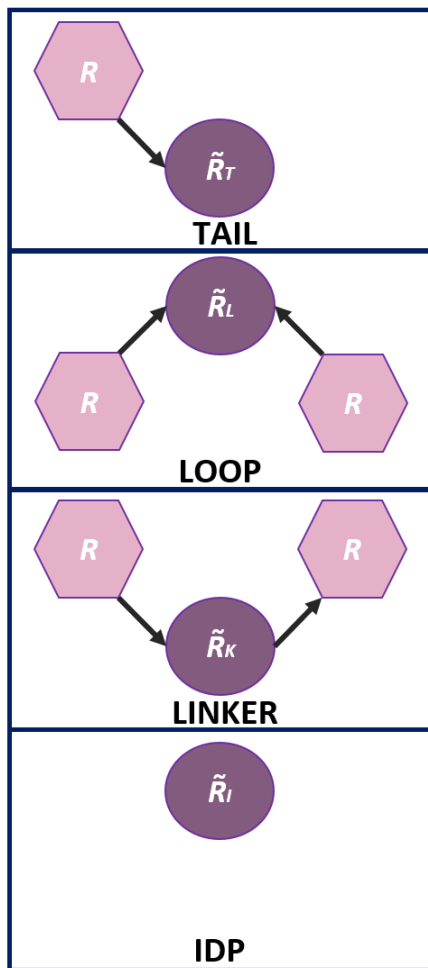


Figure 2.10: Illustration of the oriented representation for all types of flexible regions.

Figure 2.11 shows the *Illustrative Example* with a designated definition of its flexible regions. $\tilde{R}_{1,3}$ is a central loop, while $\tilde{R}_{1,1}$ and $\tilde{R}_{1,5}$ are tails. In the second chain, $\tilde{R}_{2,4}$ is a tail and $\tilde{R}_{2,2}$ is a linker that can have two possible directions (N-to-C or C-to-N) in MoMA-FReSa. The resulting *Oriented System Graph* for both linker directions is shown in Figure 2.12.

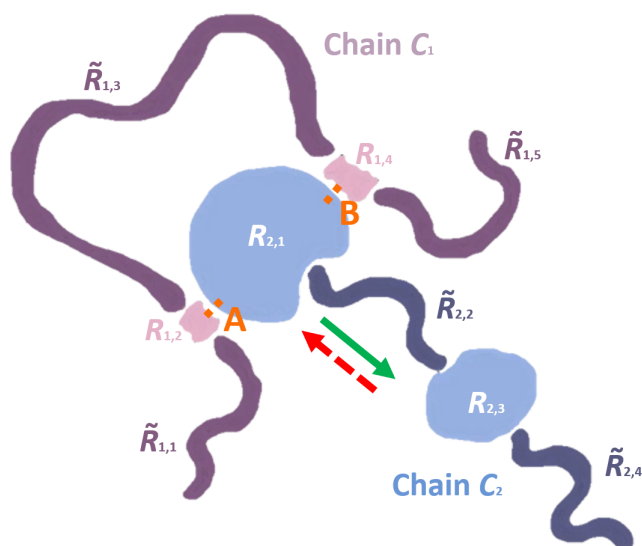


Figure 2.11: *Illustrative Example* system updated with apparent types of flexible regions. The linker can be considered as a N-to-C linker (full green arrow) or a C-to-N linker (dashed red arrow).

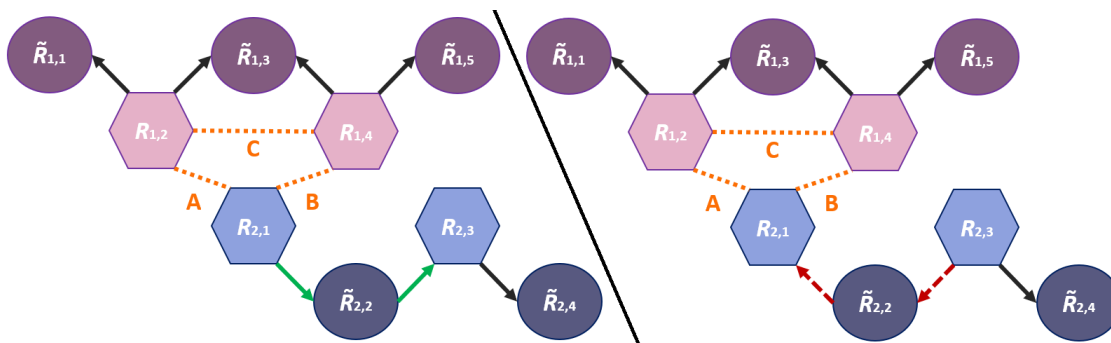


Figure 2.12: Two possible *Oriented System Graphs* of the *Illustrative Example*. On the left, example for a N-to-C linker with full green arrow ; on the right, example for a C-to-N linker dashed red arrow

Construction of signed adjacency matrices

The final step in representing order within MoMA-FReSa involves translating the *Oriented System Graph* into a signed adjacency matrix. The goal of this matrix is to efficiently capture the hierarchical relationships between all regions in the system. Essentially, it acts as a roadmap to guide the sampling process. To do this, a first signed adjacency matrix M is defined by translating the orientation in a matrix. For a system S of m_T regions, M is a $m_T \times m_T$ matrix. To create M , each region is assigned a unique numerical index. MoMA-FReSa prioritizes chain index over region index when sorting. This means regions within a chain are ordered consecutively (e.g., $R_{1,1}$ comes before $R_{1,2}$ in chain 1). Following this order, regions are assigned consecutive natural numbers for their indices in the M matrix. For instance, $\tilde{R}_{1,1}$, $R_{1,2}$, $\tilde{R}_{1,3}$, $R_{1,4}$, $\tilde{R}_{1,5}$, $R_{2,1}$, $\tilde{R}_{2,2}$, $R_{2,3}$ and $\tilde{R}_{2,4}$ are respectively re-indexed as \tilde{R}_1 , R_2 , \tilde{R}_3 , R_4 , \tilde{R}_5 , R_6 , \tilde{R}_7 , R_8 and \tilde{R}_9 .

The value at any position $M[r, c]$ represents the directed relationship between region R_r and region R_c . Here is how the orientation is encoded:

- If $R_r \rightarrow R_c$, $M[r, c] = 1$.
- If $R_r \leftarrow R_c$, $M[r, c] = -1$.

Because of this directional encoding, M is anti-symmetric. Figure 2.13 illustrates this concept of translating oriented edges into the M matrix.

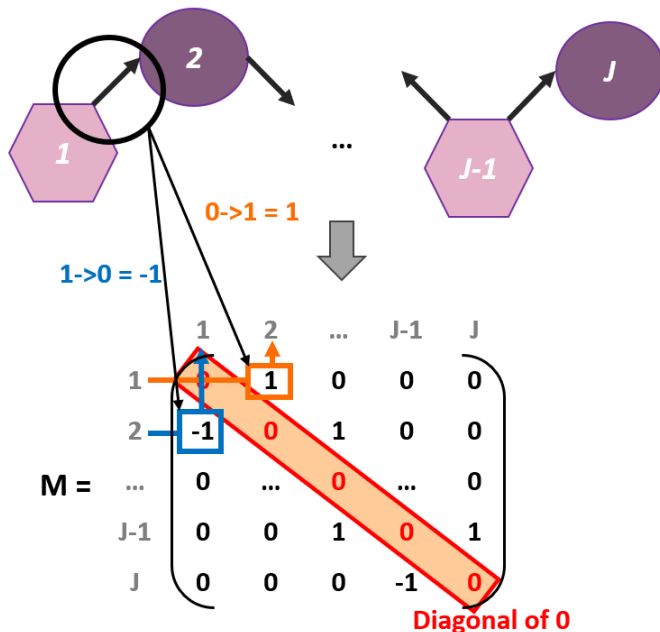


Figure 2.13: Illustration of the creation of the signed adjacency matrix M for a given *Oriented System Graph*.

Based on the left *Oriented System Graph* in Figure 2.12, a signed adjacency matrix M can be created for the *Illustrative Example* from smaller adjacency matrices for each chain. The approach is shown in Figure 2.14.

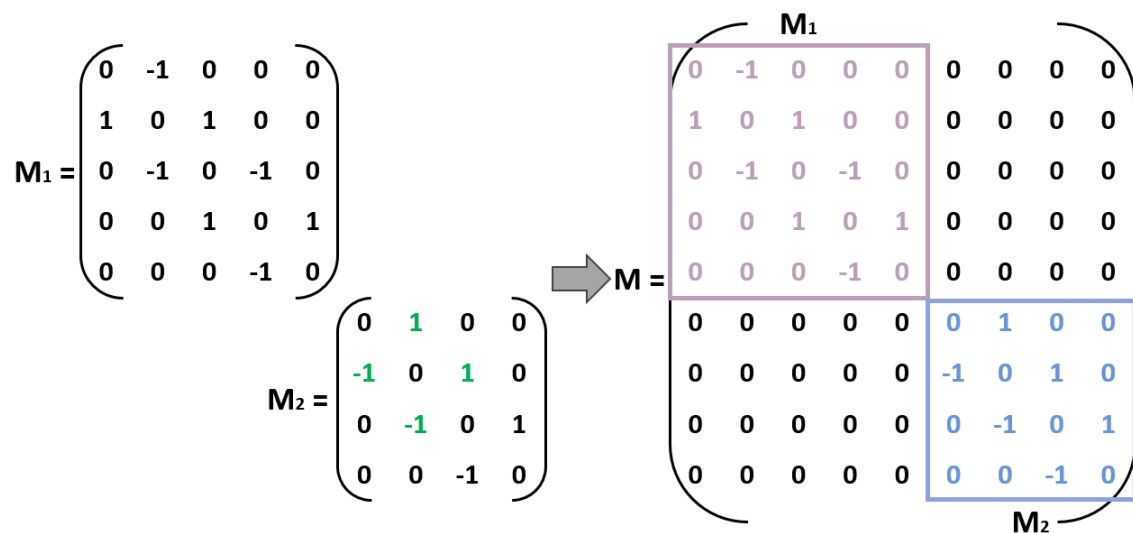


Figure 2.14: Illustration of the construction of the signed adjacency matrix M of the *Illustrative Example* with N-to-C linker.

While M captures order from edges, it does not explicitly consider \mathfrak{B} relationships, where connected rigid regions need simultaneous placement. To address this, M is refined into a new matrix, M' . M' has the same flexible lines as M : for each flexible region \tilde{R}_i , $M'[i, :] = M[i, :]$. For each rigid region R_i let us define H_i the subset of indices of the rigid regions spatially fixed to R_i position, such as $\forall j \in H_i, R_i \mathfrak{B} R_j$; then $M'[i, :] = M[i, :] + \sum_{j \in H_i} M[j, :]$. This summation mathematically encodes the requirement for simultaneous placement of bound rigid regions. This process is illustrated by Figure 2.15.

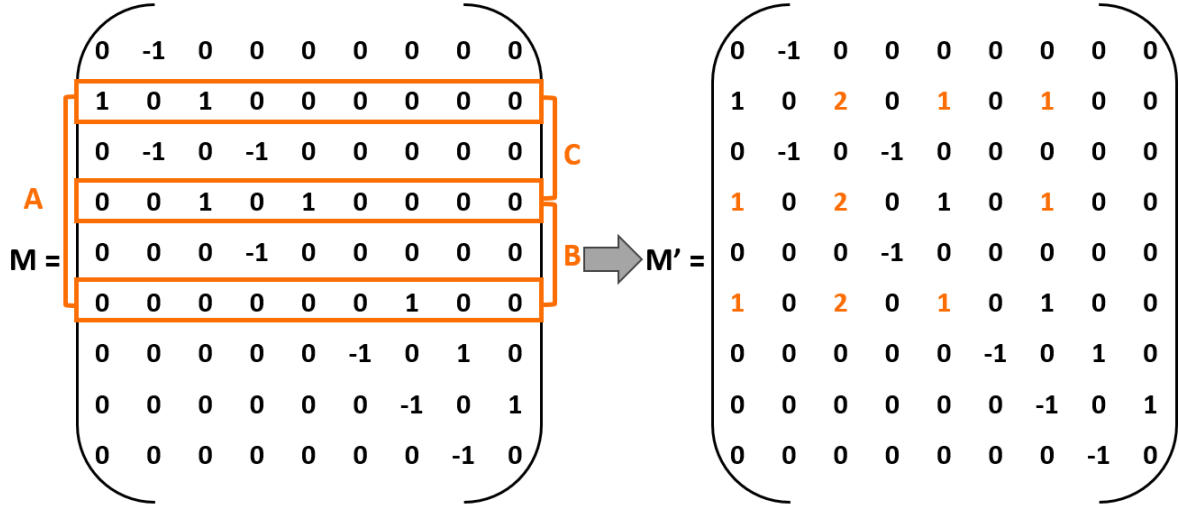


Figure 2.15: Illustration of the construction of the signed adjacency matrix M' of the *Illustrative Example* with N-to-C linker.

The same procedure can be applied to the right *Oriented System Graph* of Figure 2.12 to obtain its M' matrix, as shown in Figure 2.16.

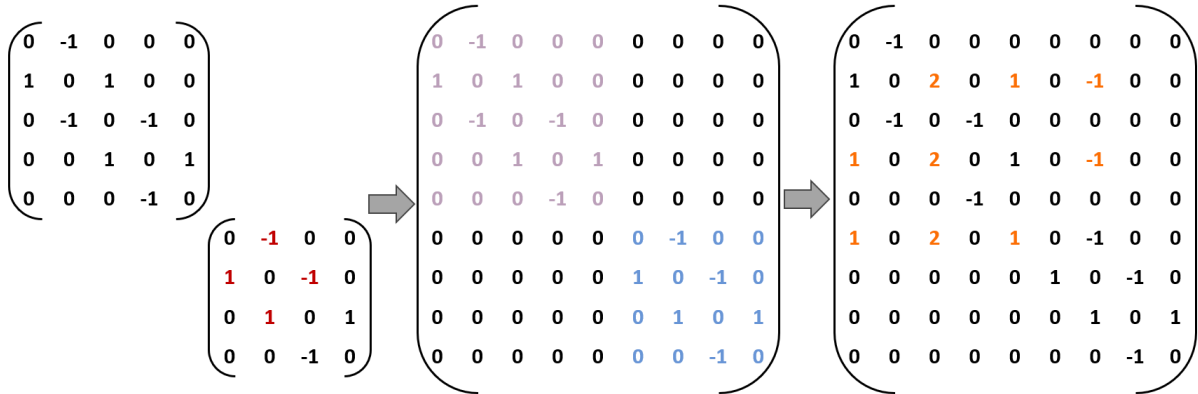


Figure 2.16: Detailed illustration of the construction of the signed adjacency matrix M' of the *Illustrative Example* with C-to-N linker.

The specific values (weights) in M' are not crucial for understanding the sampling order; it is the sign (positive or negative) that determines the hierarchy between regions. With M' in place, MoMA-FReSa possesses a concise representation of all hierarchical relationships for the sampling process. The use of this matrix during the sampling is illustrated later in Subsection 2.2.5.

2.2.3 Pre-processing construction steps

The preceding subsection outlined the general decomposition methodology employed by MoMA-FReSa. This process begins with the construction of a simple graph based on the decomposition in regions and culminates in the encapsulation of the regional order relationships within a signed adjacency matrix M' . However, achieving this outcome necessitates the consideration of several choices in accordance with the specific structure of the investigated molecular system and the defined requirements of the user. This subsection develops in more detail the pre-processing construction phase of MoMA-FReSa and the associated heuristics.

Rules of construction

During the entire decomposition process, two fundamental rules are consistently applied: one governing flexible regions and another one for rigid regions. As illustrated in Figure 2.10, all types of flexible regions adhere to the same rule:

Rule 1: A flexible region can have a maximum of one positive weight value (outgoing arrow) in the corresponding row of matrices M and M' (M and M' have the same flexible rows). Additionally, the number of positive weight values on a flexible row in M or M' must never exceed the number of negative weight values (incoming arrows).

A rigid region can only be positioned by a single flanking flexible region. Attempting to place a rigid region relative to two moving flanking regions would inevitably lead to a spatial conflict. This property extends to rigid regions connected by the \mathfrak{B} relation, signifying that these regions are spatially fixed together and require at most a single governing flexible region for conflict-free placement of the entire "block". This principle leads to the second rule:

Rule 2: A rigid row in matrix M' can have a maximum of one negative value representing an incoming arrow. This means that a group of linked rigid regions can only be placed by a single incoming flexible region.

To prevent evident circular dependencies, a third rule is introduced:

Rule 3: The addition of opposite values during M' creation is prohibited. In other words, if row i of M has a value of 1 in column c ($M[i, c] = 1$), this rule forbids its summation with a row j that has a value of -1 in the same column c ($M[j, c] = -1$). If the two associated regions are bound, the entire process is aborted.

System analysis process

The MoMA-FReSa decomposition process begins with a system represented solely by a *Non-oriented System Graph*. To transform this into an *Oriented System Graph*, a preliminary analysis is conducted based on previously established rules:

- **Chain Ends and Rule 1:** All flexible regions situated at chain extremities are designated as tails, adhering to *Rule 1*.
- **Bound Regions and Rule 3:** According to *Rule 3*, any flexible region surrounded by two bound rigid regions possesses two incoming arrows to avoid circular dependencies, classifying them as loops.

Following this initial assignment, the orientations of remaining edges remain undetermined. Applying these principles to the *Illustrative Example* leads to the graph depicted in Figure 2.17.

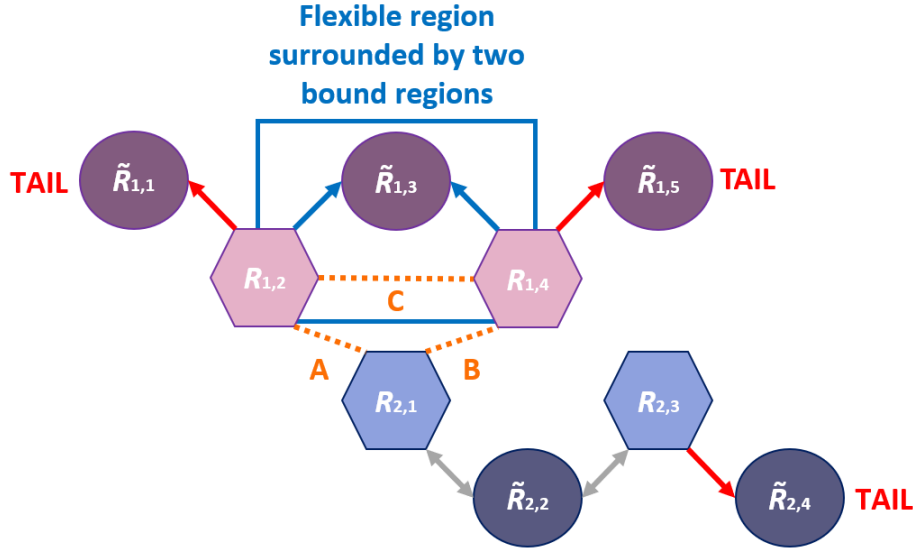


Figure 2.17: Example of construction of a first version of the *Oriented System Graph* of the *Illustrative Example* by scanning the structure of the system.

This procedure leads to the creation of a preliminary M matrix. Here, non-deterministically oriented edges are represented by unknown values denoted as $U = \pm 1$ within the corresponding rows. These unknowns are further distinguished as U_B if positioned below the diagonal of M and U_A if positioned above. This differentiation results in a prototype M' matrix. An illustrative example for the *Illustrative Example* is provided in Figure 2.18.

$$M = \begin{pmatrix}
 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & U_A & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & U_B & 0 & U_A & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & U_B & 0 & 1 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0
 \end{pmatrix} \rightarrow M' = \begin{pmatrix}
 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 2 & 0 & 1 & 0 & U_A & 0 & 0 & 0 \\
 0 & -1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 2 & 0 & 1 & 0 & U_A & 0 & 0 & 0 \\
 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 2 & 0 & 1 & 0 & U_A & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & U_B & 0 & U_A & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & U_B & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & 0
 \end{pmatrix}$$

Figure 2.18: Example of a first version of the signed adjacency matrices M and M' of the *Illustrative Example* from the first version of the *Oriented System Graph*.

Construction procedure

The next step is to assign appropriate values to the unknown weight entries within the M' matrix. Let F be the subset of indices of the flexible regions of the system S , for each $i \in F$,

a 3×3 sub-matrix M'_i centered around $M'[i, i]$ is defined. The U_A of the upper part of the matrix M'_i are the opposite to the U_B of the lower part due to the anti-symmetrical property of M . This is illustrated on the *Illustrative Example* in Figure 2.19.

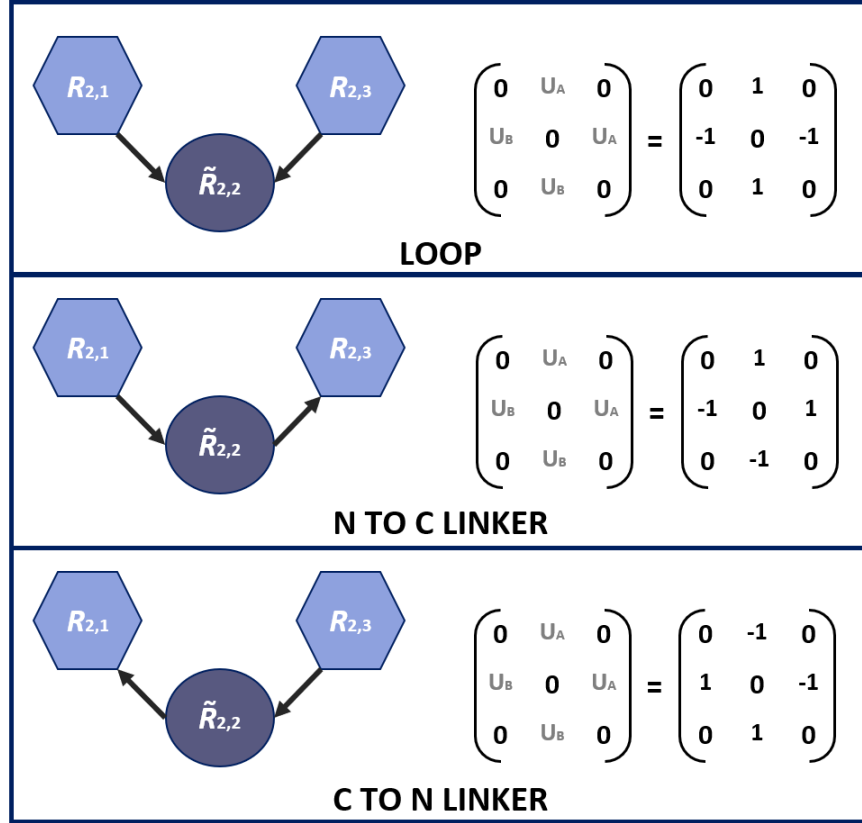


Figure 2.19: Example of effects on the sub-matrix by choosing the type of flexible regions for $\tilde{R}_{2,2}$ of the *Illustrative Example*.

Let $G \subset F$ the subset of flexible regions with at least one unknown value on the associated row. The matrix M' is updated by following these steps:

1. **Selection of a Flexible Region:** An index $i \in G$ is chosen based on a pre-defined heuristic (e.g., prioritizing the longest region within the set).
2. **Unknown Weight Value Assignment:** If unknown values exist on row i of M' , there are three potential choices: loop, N-to-C linker, or C-to-N linker. If a specific unknown weight value has already been fixed due to previous choices, only two choices remain: loop or linker (with a fixed direction). The choice is made based on another heuristic that prioritizes linkers whenever possible. If there is a choice for the direction, the linker is oriented from the larger surrounding rigid region to the smaller one.
3. **Sub-matrix and Anti-symmetry Update:** Following the chosen unknown weight value assignment and the anti-symmetrical property, the corresponding upper triangular elements U_A and lower triangular elements U_B within sub-matrix M'_i are updated.
4. **Bound property Update:** Consequently, all upper U_A and lower U_B elements of the flexible column i of M' are similarly updated accordingly the bound properties of the M' matrix.

5. **Enforcing Rule 2:** For all rows j , where $M'[j, i]$ was updated in the previous step (these are rigid rows only, the update resulting from the \mathfrak{B} relation between rigid regions), if the updated value is -1 (indicating an incoming arrow), *Rule 2* is enforced. This imposes all remaining unknown weight values within row j to be 1.
6. **Anti-symmetrical Completion (Flexible Columns):** For all columns k updated during the previous step (these are flexible columns only), the following updates are performed based on the anti-symmetry of sub-matrices M'_k : if a U_A was enforced to 1, then $M'[k - 1, k] = 1$ and $M[k, k - 1] = -1$; if a U_B was enforced to 1, then $M'[k + 1, k] = 1$ and $M[k, k + 1] = -1$.
7. **Iterative Application and Termination:** Steps 3, 4, and 5 are iteratively applied to all newly modified sub-matrices M'_k until no further updates are required within M' to comply with *Rule 2*, the \mathfrak{B} relationship and the anti-symmetry properties.
8. **Circularity Test:** A test for circular dependencies within the resulting M' matrix is performed.
9. **Updating G :** The set G is updated by removing rows that have become well-defined (i.e., no remaining unknown weight values) due to the refinement process.

MoMA-FReSa allows for user intervention during this pre-processing construction phase. A version of the final system is provided to the user, who can check if the system is properly defined or suggest corrections. For example, the user can delete geometric constraints, to generate linkers instead of loops. On the contrary, the user can enforce loops instead of linkers even in the absence of a favorable geometry. At any point where user input is prompted, the option to backtrack and modify previous selections is possible if the enforced choices become unsatisfactory.

The iterative process terminates when G becomes empty, meaning that the M' matrix has reached a fully defined final state, enabling the subsequent sampling stage. The region splitting and this signed adjacency matrix M' defined during this pre-processing phase is unique for a given system with given user needs.

Circularity test

During the MoMA-FReSa pre-processing construction, a critical step is incorporated after each decision point to ensure the constructed *Oriented System Graph* remains free of circular dependencies. This concept aligns with *Rule 3*, aiming to prevent a scenario where a region R_1 requires placement before R_2 , while R_2 necessitates placement before R_1 , essentially creating a deadlock. This undesirable situation can arise during linker assignment. To identify potential circular dependencies, a straightforward test is applied to flexible columns within the M' matrix. This test verifies the following property:

Circular property: Let S be a system of m_T regions and its $m_T \times m_T$ matrix M' , where F is the set of indices of flexible regions of the system in M' . The system encounters a circular problem, if and only if there is a subset $E \subset F$ such as $\forall i \in E, M'[:, i] \neq 0_m$ and $\sum_{i \in E} M'[:, i] = 0_m$.

An illustrative example of a circular problem is provided in Figure 2.20.

The circularity test is exclusively performed after selecting a linker and only on columns that are fully defined (i.e., no remaining unknown U values). If a circular problem is detected,

the pre-processing procedure backtracks to the beginning of step 2, it discards the problematic choice and abandons the current linker direction. If both linker directions lead to circular problems, or if the alternative direction is unavailable due to previous selections, the procedure enforces a loop for this specific flexible region.

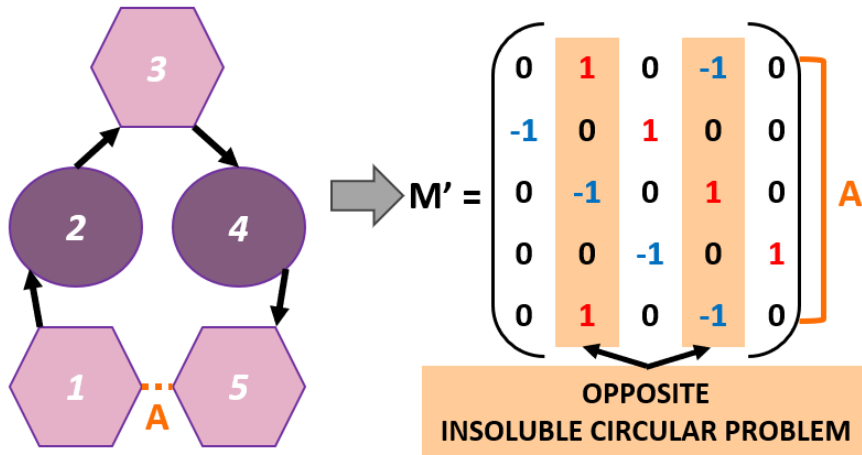


Figure 2.20: Example of circular problem appearing in a matrix M' .

2.2.4 Sampling principle

This subsection explains the core principles of the sampling method and its associated residue-centered approach, and incorporates new notations. MoMA-FReSa can be viewed as a state-space search algorithm, aimed to efficiently sampling the vast conformational space while adhering to specified constraints.

Conformational space definition

As defined in Subsection 2.2.1, our sampling approach is applied to the residues inside flexible regions. The order of sampling between the flexible regions of the system based on the pre-processing hierarchization and the order of sampling between residues of a same region are both developed in detail in next Subsection 2.2.5.

The conformation of a residue is defined by only three dihedral angles (ϕ , ψ , and ω) shown in Figure 2.2. Therefore, a conformation vector $Q_{i,j,k} = (\phi_{i,j,k}, \psi_{i,j,k}, \omega_{i,j,k})$ can be defined for each residue $\tilde{A}_{i,j,k}$.

The dihedral angle of the peptide bond, ω , behaves differently from the other two (ϕ and ψ). While ϕ and ψ can vary widely according to the Ramachandran plot (Figure 2.24), ω typically adopts only two values: 180° (*trans* conformation, preferred conformation for most residues) and 0° (*cis* conformation, less common but can occur, especially in Proline residues). Due to the limited variability of ω , some methods, such as Flexible-Meccano [149], simplify the representation of the conformation of a residue by excluding ω and using only ϕ and ψ . The conformation vector model is then reduced to $Q_{i,j,k} = (\phi_{i,j,k}, \psi_{i,j,k})$. However, this simplification is not always ideal. The presence of *cis* conformations, particularly in Proline, necessitates retaining all three angles for a more accurate description.

The algorithm ultimately aims at selecting a conformation vector for each flexible residue, resulting in the complete conformation of the biomolecule. The state for each flexible region

$\tilde{R}_{i,j}$ of n residues can be defined as a $3 \times n$ -size vector $Q_{i,j} = \bigcup_{1 \leq k \leq n} Q_{i,j,k}$, which concatenates the conformations $Q_{i,j,k}$ of each residue of the region.

Finally, for a given system S , the state is the concatenation of its m_F flexible region. For a system of l chains $C_i, i \in \llbracket 1, l \rrbracket$, of respectively $m_F(i)$ flexible regions each ($\sum_{1 \leq i \leq l} m_F(i) = m_F$), the state is written $Q = \bigcup_{\substack{1 \leq j \leq m_F(i) \\ 1 \leq i \leq l}} Q_{i,j}$. If n_F represents the total number of flexible residues in the system, the size of this vector is $3 \times n_F$.

A planning problem

MoMA-FReSa can be categorized as a state-space search algorithm. In this context, the algorithm aims at maximizing the exploration of the set of feasible conformations. \mathcal{Q} is the ensemble of all states (i.e., all conformations Q , satisfying constraints or not) and \mathcal{Q}_a is the ensemble of goal states (i.e., admissible conformations) included in \mathcal{Q} . The search-space tree is composed by the states of \mathcal{Q} and the objective is to reach states belonging to \mathcal{Q}_a through our method. In this tree, the number of transitions between a given initial state and a terminal one is basically n_F (the number of flexible residues of the system).

The conformations of each flexible residue are picked from a database of three-residue fragments detailed in Subsection 2.2.7. For a conformation vector $Q_{i,j,k}$, the p th set of angle values is extracted from the database among all the possible set for $AA_{i,j,k}$ following given instructions. This set is then applied to obtain the new conformation $Q_{i,j,k}(p)$. Since the database contains discrete values for the angles but \mathcal{Q} is continuous, Gaussian noise is applied to the data, as explained below.

The number of conformations stored within the database typically varies depending on the specific amino-acid type and the local sequence. To illustrate the vastness of the conformational space \mathcal{Q} , we can consider a hypothetical scenario with a constant number of possible conformations, denoted by P , for each residue regardless the amino-acid type. For a system containing n_F flexible residues and assuming P available conformations per residue, the total number of potential states Q is P^{n_F} . Thus, the number of states can grow quickly even for relatively small systems. For instance, a simple IDR with 10 residues and 100 conformations per residue would possess 100^{10} so 10^{20} possible states. This number is the theoretical size of \mathcal{Q} before the addition of continuity through stochastic noise. Enumerating all these states would be monstrous and testing all of them would be impossible. The complexity is even increased for systems with more residues, multiple chains and a larger number of constraints.

The aim of MoMA-FReSa as a state-space search algorithm is to navigate this huge and complex conformational landscape while respecting the relevant constraints and reach an admissible conformation in \mathcal{Q}_a . To effectively explore this continuous and vast ensemble of states, with a continuous search-space tree, MoMA-FReSa employs a sophisticated sampling algorithm involving multiple stochastic processes. Therefore, it can be considered as a non-deterministic problem solving method [79].

2.2.5 Main sampling process

This subsection provides a comprehensive description of the MoMA-FReSa sampling method. It adopts a hierarchical approach, starting with a high-level overview of the functionalities of the method. Afterwards, dedicated parts further develop each individual component of the algorithm, providing a detailed breakdown of the processes, to finally explain error management throughout these various levels, ensuring a robust and reliable sampling procedure.

Definition of the main steps

The sampling process utilizes a collision grid to ensure admissible conformations. Each system component is added to the grid step-by-step, with collision tests performed to guarantee no collision. This ensures a valid placement for all components, leading to an admissible conformation. Further details about the collision grid and the checking process are provided in Subsection 2.2.7. Prior to initiating any sampling step, all defined fixed rigid domains within the system are positioned within the collision grid. These fixed regions act as obstacles during sampling, imposing spatial constraints on the movable elements. From a matrix perspective, these fixed rigid regions can be identified in the signed adjacency matrix M' by rows without negative weights. In simpler terms, this signifies the absence of any incoming arrows on the corresponding block of spatially linked rigid regions within the *Oriented System Graph*.

The MoMA-FReSa planning approach can be decomposed into three distinct phases, illustrated in Figure 2.21. Thus, each cycle includes the three following iterative steps in : (i) selection of a flexible region, (ii) selection of a residue inside the identified region, and (iii) definition of a conformation for this residue. This process is repeated until there are no more regions to sample. At this time, an admissible state in \mathcal{Q}_a is reached in our search-space tree. During these three steps, which correspond to the three levels of the hierarchical decomposition of the system, it is advantageous to integrate stochastic variability to explore a diverse range of solutions and define the largest possible conformational ensemble. This is explained in the corresponding parts below.

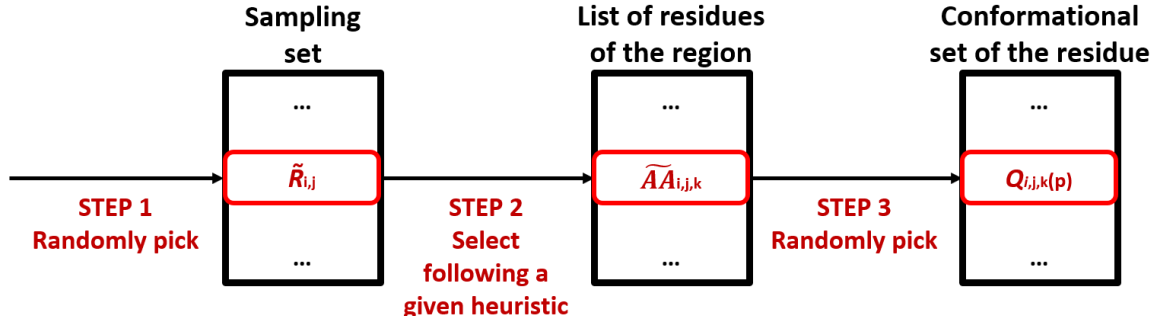


Figure 2.21: The decomposition of planning problem in three steps. Step 1: Pick a random flexible region inside a shortlist of regions - Step 2: In the chosen region, pick the good residue following the sampling order, for some types of regions chose randomly one among two - Step 3: Randomly pick a conformation in a database and slightly modify the angles.

During the sampling process, the algorithm employs a depth-first search strategy. This implies that the algorithm explores each branch of the search tree as deeply as possible until a successful conformation is identified (i.e., conformation of \mathcal{Q}_a) or a dead-end is reached (i.e., a branch of the tree that can not lead to an admissible state in \mathcal{Q}_a). This last scenario arises when a particular sampling step fails to identify a conformation that satisfies all the implemented tests after a pre-determined number of attempts. When a failure happens, the algorithm triggers a backtracking mechanism to back up on a upper search node, allowing for the exploration of alternative branches. This backtracking capability is a critical component of the ability of the algorithm to recover from dead-ends and efficiently explore the conformational space. If the number of backtracks surpasses a defined threshold, the entire sampling process is aborted, encountering a failure condition.

The main method `Generate-One-Conformation` of Algorithm 1 and the sampling

method `Sample` outlined in Algorithm 2 exhibit how the three steps upper defined are combined and executed within the sampling process. These three phases are developed in the following sub-parts of the subsections. The `Backtrack` mechanism employed to recover from potential failures during the sampling process is detailed at the end of this subsection.

Algorithm 1 Generate One Conformation

GENERATE-ONE-CONFORMATION(*max_backtracks_allowed*):

```

1: Place all the fixed rigid regions
2: Reset the Sampling Set
3: while the Sampling Set is not empty and the number of backtracks does not reached
   max_backtracks_allowed do
4:   Randomly choose a flexible region  $\tilde{R}_{i,j}$  among the Sampling Set (Step 1)
5:   Sample( $\tilde{R}_{i,j}$ ), one sampling step on the given region
6:   if Sample( $\tilde{R}_{i,j}$ ) is a success then
7:     if  $\tilde{R}_{i,j}$  does not need to be sampled anymore then
8:       Remove  $\tilde{R}_{i,j}$  of the Sampling Set
9:       Add  $\tilde{R}_{i,j}$ 's children to the Sampling Set
10:    end if
11:  else
12:    Backtrack( $\tilde{R}_{i,j}$ )
13:  end if
14: end while
15: if the Sampling Set is empty then
16:   Generate-One-Conformation is a success
17: else
18:   Generate-One-Conformation is a failure
19: end if

```

Algorithm 2 Sample

SAMPLE($\tilde{R}_{i,j}$):

```

1: Choose a residue  $\tilde{AA}_{i,j,k}$  in  $\tilde{R}_{i,j}$  (Step 2)
2: for a certain number of tries do
3:   Define a conformation  $Q_{i,j,k}$  for  $\tilde{AA}_{i,j,k}$  (Step 3)
4:   Sample is a success if the conformation passes the tests
5: end for
6: If no conformation passes the tests in the number of tries, Sample is a failure

```

Step 1 - Selection of a flexible region in the system

The initial action involves selecting one flexible region $\tilde{R}_{i,j}$ among the subset of flexible regions of S referred to as the *Sampling Set*. This *Sampling Set* guides the order in which flexible regions are explored during sampling and all regions in the set are sampled in parallel. To define this subset, MoMA-FReSa constructs a hierarchical structure called the *Draft Prioritization Tree* based on the adjacency matrix M' previously defined. This tree of regions serves to prioritize flexible regions based on their dependencies within the system.

Let F be the subset of flexible region indices in the matrix, $\forall i \in F$, the row i is scanned, if there is one positive value at the column j ($M[i, j] > 0$). This means that the region R_j is a child of the region R_i in the *Draft Prioritization Tree*. If it exists, the rigid child is

scanned in the same way to obtain its own children. Through this iterative process, the *Draft Prioritization Tree* captures and organizes all flexible regions within the system, excluding fixed rigid regions (those pre-positioned in the grid before sampling) and redundant rigid region (at most one rigid region per block of spatially linked rigid regions in \mathfrak{B} relation together). By this construction method, the children of flexible regions are rigid and the children of rigid regions are flexible. Notably, the roots of this tree are exclusively flexible regions.

A tree of flexible regions called the *Prioritization Tree* is naturally obtained from this *Draft Prioritization Tree*. The tree is refined by removing all rigid regions: the direct flexible children of a previously removed rigid child becomes the direct children of its grandparent (parent of its parent) within the *Prioritization Tree*. It is important to note that only linkers have children in this tree.

The initial *Sampling Set* is established based on the roots of these trees. This set contains all the regions that can be currently sampled in parallel and is dynamically updated throughout the process. At each cycle, a region is randomly picked in the *Sampling Set* to add a nondeterminism aspect at this step. The following rules govern how the *Sampling Set* is updated throughout the sampling process:

- **Successful Region Removal:** Each region fully and successfully sampled is removed from this subset.
- **Linker Children Addition:** If the region is a linker with flexible dependencies, its removal adds its children nodes in the *Prioritization Tree* to the *Sampling Set*.
- **Return to the Sampling Set:** The `Backtrack` procedure, defined later, can fill this subset on certain conditions.
- **Sampling Completion:** The sampling process is considered successful and over when the *Sampling Set* becomes empty.

To exemplify this concept, let us revisit the *Illustrative Example* introduced in Subsection 2.2.1. As established in Subsection 2.2.2, the *Illustrative Example* incorporates a linker $\tilde{R}_{2,2}$ that introduces dependencies. The direction of this linker impacts the initial *Sampling Set* and the structure of the *Prioritization Tree*:

- **Scenario 1: Linker from $R_{2,1}$ to $R_{2,3}$ and $\tilde{R}_{2,4}$:** In this scenario, $R_{2,1}$ (and so $R_{1,2}$ and $R_{1,4}$) is fixed and the tail $\tilde{R}_{2,4}$ is directly dependent on the linker. Figure 2.22 visually depicts the dependency relationships between flexible regions and illustrates how is extracted the initial *Sampling Set* with the first prior flexible regions. If the linker $\tilde{R}_{2,2}$ is successfully sampled, it is removed from the *Sampling Set*, and its unique direct child, $\tilde{R}_{2,4}$ is incorporated into the set instead.
- **Scenario 2: Linker from $R_{2,3}$ to $R_{2,1}$ and $\tilde{R}_{2,4}$:** In this scenario, $R_{2,3}$ is fixed and the tail $\tilde{R}_{2,4}$ is no longer directly dependent on the linker. However, the rigid regions belonging to C_1 and spatially linked to $R_{2,1}$ become entirely dependent on the successful sampling of the linker, consequently affecting the surrounding flexible regions. If the linker $\tilde{R}_{2,2}$ is successfully sampled, the following rigid regions $R_{2,1}$, $R_{1,2}$ and $R_{1,4}$ are positioned. As a result, the three flexible regions within C_1 replace $\tilde{R}_{2,2}$ within the current *Sampling Set*. Figure 2.23 summarizes the construction of the *Prioritization Tree* and the initial *Sampling Set* for this specific case.

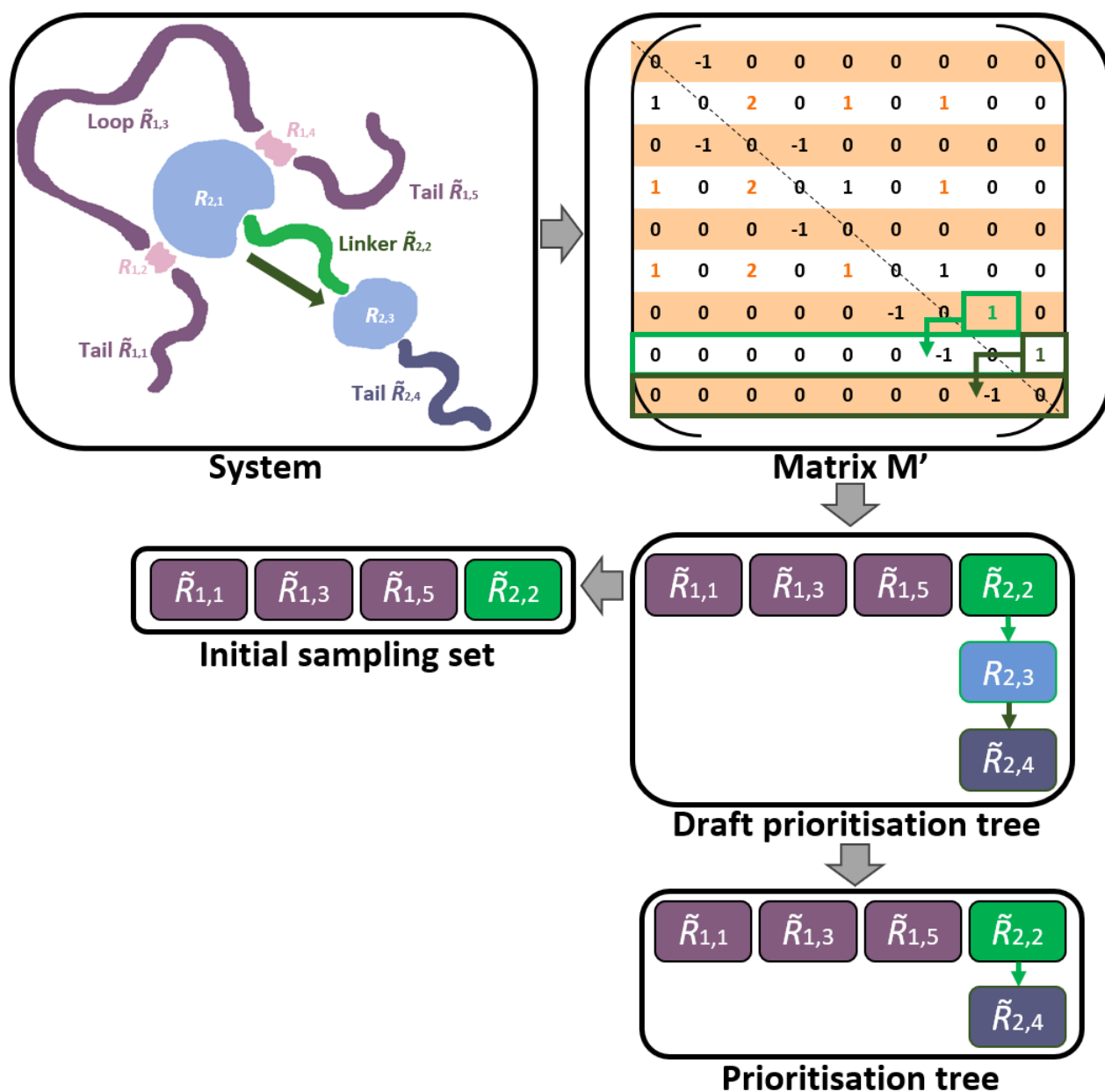


Figure 2.22: Example of the construction of the *Prioritization Tree* and the initial *Sampling Set of Illustrative Example* with a N-to-C linker.

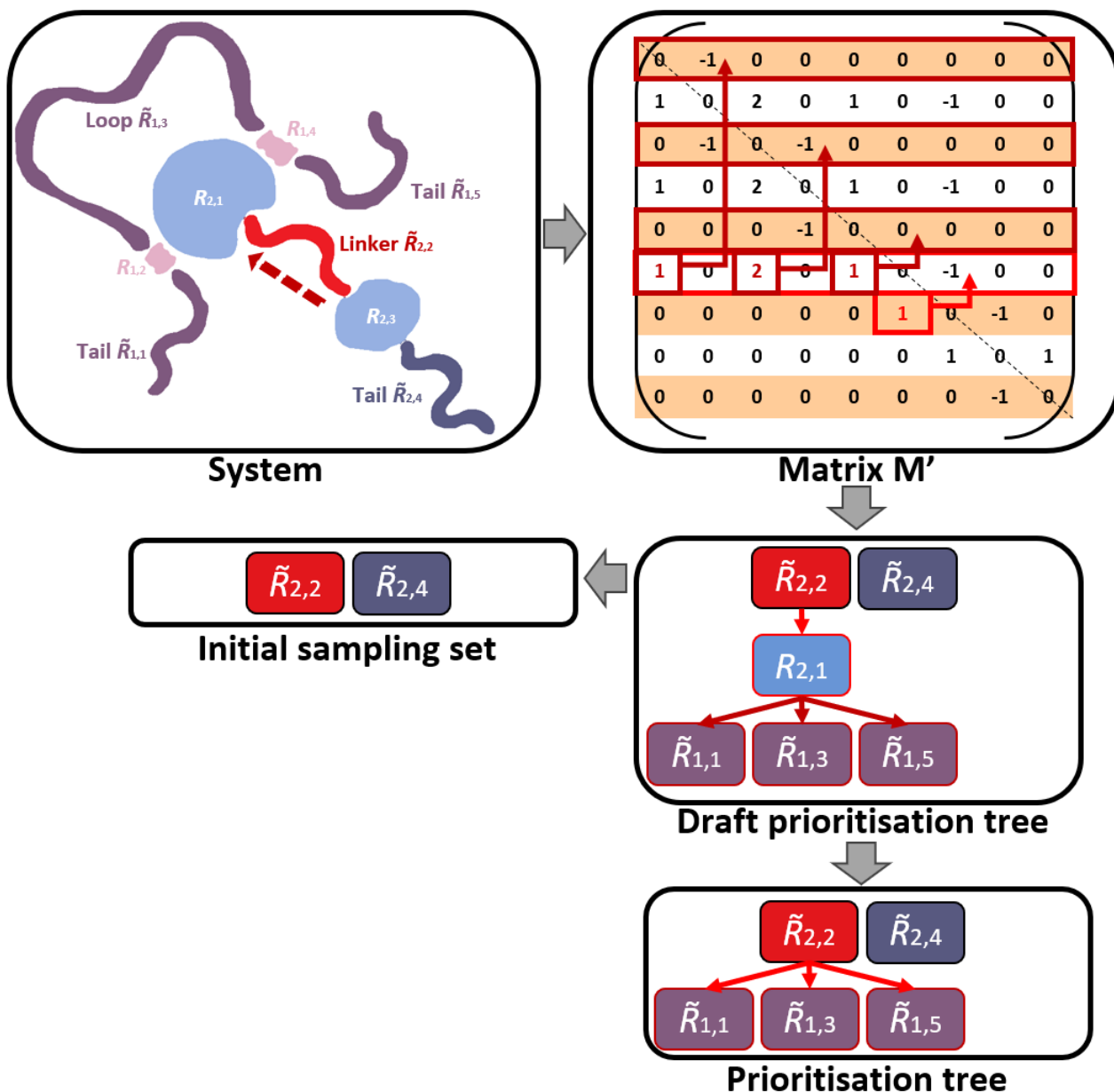


Figure 2.23: Example of the construction of the *Prioritization Tree* and the initial *Sampling Set* of *Illustrative Example* with a C-to-N linker.

Step 2 - Selection of a residue in the region

Following the selection of a specific region $\tilde{R}_{i,j}$ a pre-defined heuristic is employed to select a residue $\tilde{A}_{i,j,k}$ to be sampled within that region. Once a residue is successfully sampled, it is fixed in the space and incorporated into the grid for collision detection purposes.

The sampling order adheres to a logical order, following either an N-to-C, an C-to-N, or both directions, depending on the specific region type. Additionally, the chosen residue in a given direction is always a neighbour of the most recently sampled residue within the region considering this direction. When possible (loops, IDPs), the sampling is bidirectional, potentially leading to two eligible residues in each cycle. This choice reduces a possible bias. In such cases, the algorithm randomly picks one of these residues. The sampling approach for

each region type follows these rules:

- **Tails:** For tail regions, the sampling direction follows the free extremity. In a chain C_i of m regions, if the tail is the m^{th} region, sampling progresses from N-to-C. Conversely, if the tail region is the first region within the chain, sampling proceeds from C-to-N.
- **Linkers:** Linker regions are sampled unidirectionally, adhering to the information encoded within the signed adjacency matrix M' .
- **Pure IDPs:** In pure IDPs, sampling occurs in both directions (N-to-C and C-to-N). In a region $\tilde{R}_{i,j}$ of n residues, a starting residue $\tilde{A}_{i,j,k}$ is selected randomly ; on the C-terminal side sample from N-to-C from the residue $\tilde{A}_{i,j,k}$ to the C-terminal $\tilde{A}_{i,j,n}$: $\tilde{A}_{i,j,k} \rightarrow \tilde{A}_{i,j,n}$; on the other side sample from C-to-N from the other starting residue $\tilde{A}_{i,j,k-1}$ to the N-terminal $\tilde{A}_{i,j,0}$: $\tilde{A}_{i,j,1} \leftarrow \tilde{A}_{i,j,k-1}$. A random residue is picked between the two eligible ones except if one side has already been entirely sampled, in this case the choice on the opposite side becomes automatic.
- **Loops:** Loop regions also involve sampling in both directions. In a region $\tilde{R}_{i,j}$ of n residues, a random last triplet of residues (tripeptide) centered at a randomly sampled residue $\tilde{A}_{i,j,k}$ is selected ; on the N-terminal side sample from N-to-C from the N-terminal anchor point and first starting residue $\tilde{A}_{i,j,1}$ to the last tripeptide excluded: $\tilde{A}_{i,j,1} \rightarrow \tilde{A}_{i,j,k-2}$; on the other side sample from C-to-N from the C-terminal anchor point and other starting residue $\tilde{A}_{i,j,n}$ to the last tripeptide excluded: $\tilde{A}_{i,j,k+2} \leftarrow \tilde{A}_{i,j,n}$. Similar to IDPs, a random selection is made between the two eligible residues unless one side has been entirely sampled, at which point the choice on the opposite side becomes fixed. If both sides have been completely sampled, the final three residues are sampled collectively in a single step to close the kinematic chain. This technique is detailed in Subsection 2.2.7.

When all residues within a region have been sampled successfully, the sampling process for that region is complete. The region is then removed from the *Sampling Set*, considering its dependencies within the overall strategy.

Step 3 - Selection of a conformation for the residue

The final step involves defining a conformation for the chosen residue $\tilde{A}_{i,j,k}$. MoMA-FReSa employs a database of three-residue fragments for conformation selection, as mentioned in Subsection 2.2.4 and detailed in Subsection 2.2.7. With the exception of the final three residues within loops due to their specific closing requirement, all other residues are assigned conformations by randomly selecting an entry from this database. The selection process can be done using two strategies:

- **Single-Residue-based Sampling (SRS) Strategy:** This strategy focuses solely on the central residue of the chosen fragment when selecting a conformation from the database. For example, if the sampling residue is a Threonine, a fragment is randomly selected among all the fragments with a central Threonine, regardless of the two flanking residues of the fragment.
- **Three-Residue-based Sampling (TRS) Strategy:** This strategy considers also the neighbouring residues to select a conformation. The fragment selected in the database must have the same consecutive three amino-acids and the same secondary structural

region (according to the Ramachandran plot in Figure 2.24) for the already sampled residue. For example, if the sampling residue is a Threonine flanked by an Alanine (N-terminal neighbour) and a Glycine (C-terminal neighbour) and the Alanine has been already sampled in α -region (N-to-C directional sampling), a fragment is randomly picked among all the "(ALA,THR,GLY)" with an α conformation for the Alanine.

These strategies and their detailed implementation are elaborated upon a previous work [67]. Their utilization by MoMA-FReSa is studied in more details for chosen examples in Chapter 3.

For a given residue $\tilde{A}_{i,j,k}$, a conformation vector $Q_{i,j,k}$ represents its current conformational state. The sampling process extracts a set of potential dihedral angles from a random database entry p to define a specific conformation $Q_{i,j,k}(p)$. To introduce continuity from a discrete set and to account for uncertainties within crystallographic data, a slight random distortion is applied to the three dihedral angles within $Q_{i,j,k}(p)$. The current distortion uses a Gaussian perturbation centered on the angle value from the database with a small standard deviation (0.01 radians). This could be improved in future work by making the standard deviation dependent on the number of entries in the database for the sampled residue. For example, higher dispersion and standard-deviation could be considered for tripeptides with fewer entries in the database. Note that, at this stage, side-chain placement is not yet performed. Only the backbones atoms and a pseudo- C_β is incorporated into the collision grid as explained in Subsections 2.2.6 and 2.2.7.

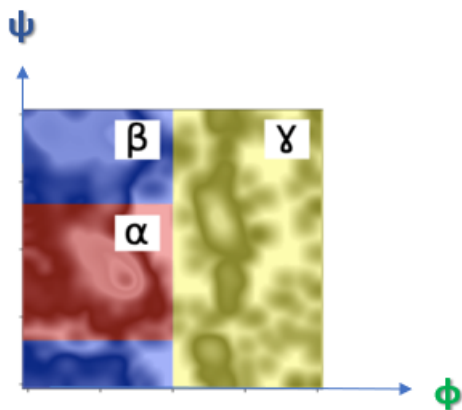


Figure 2.24: Ramachandran plot where the main secondary structures are plotted as a function of ϕ and ψ .

The user of MoMA-FReSa can choose several options according to the previous knowledge of the system under investigation:

- **Strategy Selection (SRS/TRS):** As mentioned previously, the sampling can be conducted through the Single Residue Strategy (SRS) or the Three Residue Strategy (TRS) to select the conformation of a given residue. This choice can be let to the user or chosen by default after studying the flexible region properties and can vary among the residues of a same system. When nothing is known about the system, SRS is recommended.
- **Residue Structural Region Selection:** The selection of conformations can be further narrowed down based on the structural region of the residue (corresponding to α -helix, β -strand, or just γ). In this case, the conformation is chosen randomly among the conformations inside the chosen structural region for the middle residue, so with a

specific restriction on ϕ and ψ angles according to the the Ramachandran diagram (Figure 2.24). This option is particularly relevant for future machine-learning-enhanced implementations, as discussed in Chapter 4.

- **Target-Residue Conformation Model:** The user can also define a model for the conformation of each residue of the flexible region (a target conformation) to guide the selection of new dihedral angles. For example, the user can define a target loop model, and all the residues of the loop are sampled within a given angular threshold distance to this target conformation.

Multiple conformations for the given residue are evaluated through a battery of tests until a suitable one is identified:

- **Collision Test:** The primary test is a rigorous all-atom collision test conducted using a grid-based approach (described in Subsection 2.2.7). This test ensures that the proposed conformation avoids steric clashes with other atoms within the system.
- **Specific Distance Test for Loops:** For loop regions, an additional forward-checking test is implemented at each step to assess loop closure feasibility. This test calculates the distance (denoted as $dist$) between the most recently sampled residues on each side of the loop. It also computes the maximum feasible distance ($max_feasible$) that could be achieved if all remaining residues within the loop were positioned linearly. If $max_feasible < dist$, the conformation is rejected to avoid to reach dead-end nodes of the search-space tree. To introduce some level of randomness and potentially improve loop closure success rates, the algorithm does not necessarily accept conformations that satisfy the distance check. Instead, a measure $s = \frac{max_feasible - dist}{max_feasible}$ is computed. This value renders the proportion of the maximum feasible distance that remains available for closure. The closer to 1 is s , the greater is the probability of successful loop closure. On the contrary, conformations with a s close to 0 are avoided. After some calibration tests, we made the choice to automatically accept conformations with $s > 0.33$. For conformations with lower s values, a random number u is drawn from a uniform distribution between 0 and 1, the conformation is only accepted if $u < s$. The details concerning these loop closure improvements are provided in Subsection 2.3.3.
- **TRS Additional Test:** During TRS, a test can be performed to verify if the newly formed tripeptide aligns with the database to ensure consistency with known structures. When the sampled residue is $\tilde{A}A_{i,j,k}$, the considered tripeptide is $(\tilde{A}A_{i,j,k-2}, \tilde{A}A_{i,j,k-1}, \tilde{A}A_{i,j,k})$ in N-to-C direction and $(\tilde{A}A_{i,j,k}, \tilde{A}A_{i,j,k+1}, \tilde{A}A_{i,j,k+2})$ in C-to-N. The test searches in the database for an entry matching both the tripeptide sequence and the structure type for each individual residue within the tripeptide. If no entry is found for this association, the conformation is rejected.
- **Terminal Residue Tests and Additional Considerations:** Depending on the region type, specific tests are applied to the final residue(s). Pure IDPs and tails just apply the basic tests on their last residue(s). Loops have a particular closure test. Linkers have a specific procedure for the last residue: if the collision test successes, the linker positions the following rigid regions, the program adds them into the grid and initiates a new global collision test that scans the entire grid for potential clashes. In addition, if the children of the linker in the *Prioritization Tree* include loops, a test of loop feasibility is attempted, similar to classic distance test described above.

- **Additional Potential Tests:** The framework can be extended to incorporate other tests within this stage, such as evaluations based on energy/scoring functions for partial conformations during the construction, or the satisfaction of other constraints derived from experiments.

A conformation for a given residue is considered successful if it passes all the designated tests. If a predetermined number of attempts are unsuccessful in satisfying feasibility conformation, this step is considered as a failure. Such a scenario indicates a dead-end within the search-space tree. A backtracking procedure is then applied, as detailed in the following part.

Failure management

MoMA-FReSa employs a backtracking process to manage potential dead-ends encountered throughout the sampling steps, particularly in complex systems where finding a feasible state might require numerous attempts. If the number of backtracks surpasses a predefined threshold, this conformation is considered as trapped and the sampling process is aborted to avoid excessive exploration of unproductive branches within the search-space tree that do not lead to states in \mathcal{Q}_a .

To facilitate backtracking, MoMA-FReSa maintains a record of failures at the lowest level. These failure reports document the cause of the failure (e.g., collision) and a list of residues implicated in the issue (e.g., a list of all colliding residues). Once a successful sampling is achieved at the same level, the corresponding failure list is erased. When a failure occurs within a region $\tilde{R}_{i,j}$, a `Backtrack` procedure is initiated as detailed in the Algorithm 3. This process randomly selects a failure report from the list within $\tilde{R}_{i,j}$ and extracts the cause and responsible residues. This random selection ensures unbiased exploration of different failure scenarios.

Algorithm 3 Backtrack

BACKTRACK($\tilde{R}_{i,j}$):

- 1: Pick a random failure report fr_x in $\tilde{R}_{i,j}$
 - 2: **for** $AA_{u,v,w}$ in the fr_x responsible list **do**
 - 3: $R_{u,v}$ is defined as the region of $AA_{u,v,w}$
 - 4: **if** $R_{u,v}$ is flexible **then**
 - 5: Select a random number of `Deconstruction` steps according to the properties of $\tilde{R}_{u,v}$
 - 6: Initialize a Boolean `apply_side` according to the failure cause of fr_x
 - 7: **for** a certain number of `Deconstruction` steps **do**
 - 8: `Deconstruction` on $\tilde{R}_{u,v}$ on $AA_{u,v,w}$ side (if `apply_side`) or not
 - 9: **end for**
 - 10: **end if**
 - 11: **end for**
-

The `Backtrack` process described in Algorithm 3 can initiate a `Deconstruction` procedure for each responsible residue $AA_{u,v,w}$ identified within a failure report. However, this procedure is only applied if the region, $R_{u,v}$ to which the responsible residue $AA_{u,v,w}$ belongs to is flexible.

The number of `Deconstruction` steps for a specific region is determined based on its size and characteristics. This value is drawn from a normal distribution with a mean defined as the nearest integer below (floor function) $1 + \ln(\text{sampling_size})$, where `sampling_size` is

a function of the size and the nature of the region. The standard deviation of the distribution begins at 1 and increases incrementally along with the mean. This approach ensures that larger or more complex regions could undergo a higher number of deconstruction attempts on average.

The `Deconstruction` procedure always targets the last sampled residue within a region. However, for two-sided regions like loops and pure IDPs, the process can be directional or non-directional depending the failure cause:

- **Directional Deconstruction:** If the cause of the failure is directional (e.g., collision), the deconstruction focuses on the side of the region containing the responsible residue, $AA_{u,v,w}$.
- **Non-directional Deconstruction:** If the failure cause is non-directional (e.g., distance test failure in loops), a random selection is made between the last N-terminal and C-terminal residues that were sampled.

The `Deconstruction` procedure essentially reverses the sampling steps for the designated residues:

- **Step 3 - Residue Unset:** The sampling operation for the chosen residue is nullified, effectively removing it from the grid.
- **Step 2 - Region Step Backtracking:** The sampling progress within the region is reversed by one step in the appropriate direction.
- **Step 1 - Re-inclusion into Sampling Set:** If the region had been entirely sampled and consequently removed from the *Sampling Set*, it is re-introduced. As a particular case, when a linker comes back into the *Sampling Set*, it is treated as a loop with a specific closure target centered on the first residue that was deconstructed. Additionally, the linker no longer possesses children in the *Prioritization Tree* to avoid conflicts with potential advancements made in the sampling of these children earlier in the process. This transformation from linker to loop is made to avoid the computational cost associated with the removal and subsequent addition in the collision grid of the rigid region following the linker, as well as possible subsequent steric conflicts.

2.2.6 Post-processing in MoMA-FReSa and beyond

Following each successful main sampling step, MoMA-FReSa offers a post-processing suite of tools to characterize more deeply the obtained conformation and generate informative reports. Specific examples of practical applications of these tools are presented in Chapter 3. This subsection provides explanations on the various components within this post-processing step, handled by MoMA-FReSa or using external programs.

Side-chain placement

After successfully completing the sampling process, MoMA-FReSa can refine the protein structure by placing the side-chains on the backbone. These side-chains were previously represented by pseudo- C_β atoms to reserve space during sampling. Note that the diameter of the pseudo- C_β is different for the different types of amino-acid, inspired by previous coarse-grained protein model [121].

The construction of the side-chains is based on the rotamer library from Lovell [129], which encodes the most frequent values of the side-chain dihedral angles for each amino-acid type. The side-chains are randomly extracted from this database, slightly disturbed and tested for collisions. Side-chain placement can be optional depending on the intended use of the conformation. However, if placement is requested and collision clashes occur despite the reserved space and can not be solved before the end of the process, the entire conformation is discarded.

Further measures by MoMA-FReSa

At this stage, the algorithm can assess user-specified bound pairs within the molecule. In this case, successful conformation not only requires a favorable overall structure but must also satisfy spatial constraints imposed by these designated interactions. MoMA-FReSa can generate several reports on contact assessments when an interaction constraint is added, along with other data in specific cases, such as the spatial distribution of rigid regions following linker placement.

During this final phase, MoMA-FReSa can also apply energy functions to calculate the total energy of a complete conformation. This energy value serves as an indicator of the stability of the conformation. Lower energy typically corresponds to a more favorable structure. This indicator can be used to refine or sort the set of conformations ensemble generated. In the current program a version of a hydrophathy scale (HPS) [57] [171] is used evaluate potential energy. Chapter 3 provides more details on this energy computation, as well as on interaction assessment and rigid distribution procedures.

Beyond MoMA-FReSa: Filtering and analysis of conformational ensembles

Experimental data, such as Small-Angle X-ray Scattering (SAXS) profiles, chemical shifts (CS), or Residual Dipolar Couplings (RDCs) measured by NMR, can be used to validate or refine conformational ensembles. By comparing the backcalculated data from the ensemble generated with MoMA-FReSa with experimental data, the user can filter out less compatible conformations or reweight their populations, resulting in a more accurate representation of the structure of the molecule. These potential post-processing tools are not currently implemented in MoMA-FReSa.

Conformational ensembles generated by MoMA-FReSa can be further analysed and compared using statistical tools. Notably, two tools developed in our group, WASCO and WARIO, offer valuable functionalities:

- **WASCO - a Wasserstein-based Statistical Tool to Compare Conformational Ensembles of Intrinsically Disordered Proteins:** This tool represents ensembles as ordered sets of probability distributions and offers a method to assess local and global differences between two conformational ensembles at the residue level [83].
- **WARIO - Weighted Families of Contact Maps to Characterize Conformational Ensembles of (Highly-)Flexible Proteins:** This tool is specialized can define cluster and contact maps from a conformational ensemble, as well as computing features based on these clusters [82]. These features include the average secondary structure propensities per cluster (using DSSP), together with the average radius of gyration across cluster conformations.

Chapter 3 demonstrates the complementary use of these tools alongside MoMA-FReSa.

2.2.7 Implementation details

The sampling stage of MoMA-FReSa requires the execution of fundamental steps that govern the generation of valid system conformations. To achieve this, MoMA-FReSa employs a variety of tools, some of which are explained in this subsection.

Three-residue fragment database

MoMA-FReSa utilizes a database of three-residue fragments. This database serves as a foundational element for generating realistic protein conformations during the sampling stage. This database is a dynamic element in MoMA-FReSa.

The actual fragment database is constructed from a subset of experimentally-determined high-resolution protein structures: SCOPe [73] 2.06 release. Multiple operations of filtering and identification were executed to obtain the resulting database [66]. First, to remove redundant sequences, a filter was applied on the initial set to obtain a subset containing sequences with less than 95% pairwise identity. This filtering resulted into a subset of 28,011 protein domain structures stored in PDB format.

Secondary structure information, crucial for our analysis, was assigned to each residue in these domains using DSSP [101]. Then, each structure file was processed by sliding a window of size 3 (to obtain tripeptides) along the amino-acid sequence. Only tripeptides where none of the residues participated in an α -helix, π -helix, or β -strand (DSSP codes *H*, *I* and *E*, respectively) were added to a database. This filtering obeys to our aim to model flexible regions. An additional filtering step was necessary for structures derived from NMR, which often contain multiple models. Here, a distance filter was employed on corresponding tripeptides from each model to eliminate structural redundancy.

Following this filtering process, the final database contained 2,972,319 unique tripeptides. Therefore, the three-residue fragment database covers a wide range of conformations classified in the 27 structural classes [66]. With 20 natural amino-acids, $20^3 = 8000$ tripeptides must be represented and not all of them were equally sampled. Concretely, the (*G, G, G*) tripeptide is the most represented one in the database with 2,560 conformations, while the (*C, I, W*) is the least populated one, with 1 single instance. Note that for the least represented tripeptides, due to statistical uncertainty, the code enforces SRS to keep sampling diversity.

Collision detection

Throughout the MoMA-FReSa process, collision detection between atoms is performed efficiently using a grid-based data structure [54]. Atoms are stored in a grid adjusted to the size of the largest atom. For a given atom, the collision test examines the primary grid cell of the atom and all its immediate neighbouring cells. If another atoms occupy any of these examined grid cells, inter-atomic distances are computed to evaluate collisions.

Initially, all fixed rigid regions are definitively positioned within the grid. As the sampling process progresses, sampled flexible residues and mobile rigid regions (those successfully placed after linker sampling) are progressively added to the grid. In essence, only atoms that are expected to be correctly placed are included within the grid, minimizing unnecessary computations. By the end of the process, a successful sampling run ensures that the entire system is positioned within the grid without collisions. This approach offers significant advantages in terms of computational efficiency for collision detection, usually time and memory expensive.

In order to enhance efficiency, side-chains of amino-acids can be represented by pseudo carbon β atoms (pseudo- C_β) within the grid. These pseudo- C_β serve as proxies for the entire

side-chain. The radius assigned to each pseudo- C_β is carefully chosen based on the specific amino-acid type [121]. This simplification can be applied solely to flexible regions or extended to both flexible and globular regions of the system.

Kinematic loop closure

MoMA-FReSa incorporates a robotics-inspired approach for efficiently closing loops during the sampling process. As detailed in Subsection 2.2.5, a random triplet of residues is defined at the beginning of the loop sampling and designated as its closing residues. Loops are then constructed with these closing residues in mind, incorporating distance tests, defined in the same subsection, to avoid non-resealable loops. When only these final three residues remain, MoMA-FReSa employs a single particular step to close the kinematic chain, determining the conformations of all three residues simultaneously. The specific details of this method have been described in detail in previous work [9].

2.3 Results

This section assesses the performance of MoMA-FReSa on a benchmark set consisting of various typical use cases. The objective is studying the effectiveness of MoMA-FReSa for sampling a large ensemble of conformations focusing on both basic and more complex systems.

The benchmark set is described in detail in Subsection 2.3.1. This subsection then provides explanations on the testing protocol. Subsection 2.3.2 details the results obtained on the benchmark set, highlighting the performance of algorithm on both basic and complex systems. Finally, Subsection 2.3.3 presents the improvements achieved in results for systems containing loops through the implementation of our upgraded loop sampling method.

2.3.1 Benchmark and protocol overviews

In this subsection, the procedure of test is detailed: first by a description of the benchmark and then by a description of the protocol and the studied metrics.

Benchmark definition

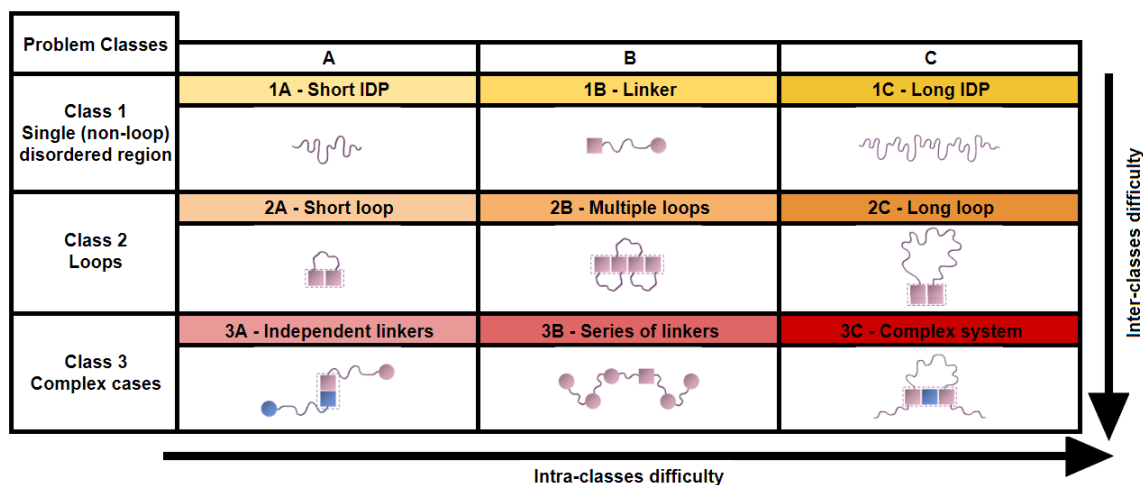


Figure 2.25: Benchmark set arranged as a matrix to highlight intra- and inter-complexity.

To evaluate the performance of MoMA-FReSa across typical use cases, a benchmark set was designed. The benchmark set categorized problems into three classes with increasing difficulty:

- **Class 1 - Single (non-loop) disordered region:** This class represented standard sampling problems tackled by existing methods, such as Flexible-Meccano [149] or IDP-ConformerGenerator [198]. It involved a single disordered region or IDP with non-loop constraints.
- **Class 2 - Loops:** Loops are a particular type of disordered regions with a higher degree of difficulty due to the loop-closure constraint. Most classical methods cannot efficiently build loops.
- **Class 3 - Complex cases:** This class combined elements from the first two classes, presenting significant sampling challenges. Examples included concatenations of a large number of regions of different types, or systems with multiple chains.

Within each class, an additional intra-class difficulty level was assigned: A, B, C (with A the lowest level of difficulty and C the hardest). For each combination of class and intra-class difficulty level, a specific problem type was defined along with a corresponding example. This resulted in a benchmark set of nine cases, as illustrated in Table 2.1:

- **1A - Short IDP:** Alpha-synuclein (human neuronal protein, UniProt ID: P37840), 140 residues.
- **1B - Linker:** A 37-residue linker of a chimeric multimodular protein involving a CBM domain (PDB ID: 1XDB) and a GH11 domain (PDB ID: 2C1F).
- **1C - Long IDP:** A silk protein (UniProt ID: P07856), 1186 residues.
- **2A - Short loop:** Short loop of 9 residues in human cyclophilin (PDB ID: 2CPL) from the benchmark set of Jacobson *et al.* [96], also used by Barozet *et al.* [9].
- **2B - Multiple loops:** Three CDR loops of 7, 9, and 16 residues in a nanobody (PDB ID: 7d4b).
- **2C - Long loop:** A long loop of 75 residues in a beef liver catalase (PDB ID: 7CAT) [47].
- **3A - Independent linkers:** Ribosomal protein L12 (PDB ID: 1RQU), a multidomain protein with a 22-residue long linker that dimerizes through its N-terminal folded domain [17].
- **3B - Series of linkers:** A section of the titin protein (PDB ID: 3B43) composed by five short linkers of 9, 8, 8, 5, and 6 residues [214].
- **3C - Complex system:** A viral IDP E1A (UniProt ID: P03255) bound to the retinoblastoma protein (PDB ID: 2R7G) by two structural motifs, forming a complex with a 71-residue loop and two tails of 7 and 17 residues [81].

By incorporating these two difficulty levels (intra-class and inter-class), a difficulty matrix was constructed to categorize the benchmark set in Figure 2.25. This benchmark set allowed us to evaluate the performance of MoMA-FReSa on a representative range of scenarios that one can encounter in practical applications. In this way, we could analyse both the average performance for common systems and the capabilities of the algorithm when facing more intricate problems.

1A - Short IDP	1B - Linker	1C - Long IDP
2A - Short loop	2B - Multiple loops	2C - Long loop
3A - Independent linkers	3B - Series of linkers	3C - Complex system
LEGEND		
	Flexible region: IDP/Tail/Linker/Loop	
	Rigid Domain fixed during the simulation	
	Moving Rigid Domain attached to a linker	
	Set of Rigid Domains fixed relative to each other	
	First Chain	
	Second Chain	

Table 2.1: Overview of the nine cases of the benchmark set, with schematic and molecular representations.

Protocol

For each example of the benchmark set, we applied MoMA-FReSa to generate $c = 1000$ conformations. The maximum number of backtrack attempts was set to $b_{max} = 300$. For all the examples, we sampled conformations in SRS strategy and without placing the side-chains, and no other post-processing was applied. MoMA-FReSa was run using a parallelized implementation, using $x = 10$ threads. After running MoMA-FReSa on each example of the benchmark, we analysed its performance using three main categories of metrics:

- **Temporal information:** This metric took two forms. The first one was the duration of the sampling process t in seconds. This duration was given for $x = 10$ threads. It is important to note that a single-threaded execution time would be slightly less than $x \times t$ due to non-parallelizable operations. The second one was the rate at which conformations were generated v (i.e., the number of conformations per second), calculated as $v = c/t$. A lower sampling duration t and a higher value of conformers per second v indicate faster performance.
- **Ratio of success:** This metric reflected the success rate of conformation generation, expressed as a percentage. It is calculated as $r = c/a \times 100$, where c was the number of successfully generated conformations and a was the total number of attempted conformations. A lower ratio of success (r) indicates a more challenging case for MoMA-FReSa.
- **Backtrack analysis:** For each of the c acceptable conformations, we picked up the number of backtrack processes $b_i, \forall i \in \llbracket 1; c \rrbracket$. At the end of the run, we calculated b_{mean} the average number of backtrack processes employed for each successfully generated conformation and the standard deviation b_{std} , which captured the variation in the number of backtracks used across successful conformations. Both the ratio of success r and backtrack analysis metrics (b_{mean}, b_{std}) provide valuable information about the difficulty a case presents for MoMA-FReSa. They are complementary and they can be interpreted together: A high ratio of success r with a high mean backtrack b_{mean} suggests a challenging case where MoMA-FReSa frequently uses backtracking but still samples successfully.

These three categories of metrics should be read together to gain a comprehensive understanding of the performance of MoMA-FReSa on each benchmark case. For instance, the sampling duration t is influenced by the size of the flexible regions (i.e. the number of flexible residues inside the system), but it is also strongly correlated to r and b_{mean} : A high number of backtracks or failed conformations can significantly impact the overall sampling time.

2.3.2 Analysis of the Results

This subsection analyses the performance MoMA-FReSa on the benchmark set, class by class. Then, a global analysis over the benchmark is presented.

Summary of the results

The results obtained by following the protocol are presented in Table 2.2. In this table, n_F and n_T are the number of flexible residues and the total number of residues, respectively.










Problem case			n_F/n_T	t	v	r	$b_{mean} \pm b_{std}$
1A	Short IDP		140/140	31.0 s	32.5 conf/s	100%	12 ± 5
1B	Linker		37/343	18.2 s	55.4 conf/s	100%	5 ± 6
1C	Long IDP		1186/1186	254.4 s	4.0 conf/s	100%	84 ± 17
2A	Short loop		9/164	123.9 s	8.1 conf/s	81.4%	96 ± 72
2B	Multiple loops		32/126	191.1 s	5.3 conf/s	66.8%	159 ± 70
2C	Long loop		75/498	1851.0 s	0.5 conf/s	15.2%	178 ± 74
3A	Independent linkers		44/240	56.2 s	18.0 conf/s	97.6%	18 ± 19
3B	Series of linkers		36/569	244.8 s	4.1 conf/s	75.6%	64 ± 57
3C	Complex system		95/450	3027.0 s	0.3 conf/s	11.9%	180 ± 75

Table 2.2: Benchmark results summary.

Class 1

All three cases in Class 1 achieved a success rate r of 100%. As expected, the computational efficiency of the method here strongly correlates with the size of the flexible region.

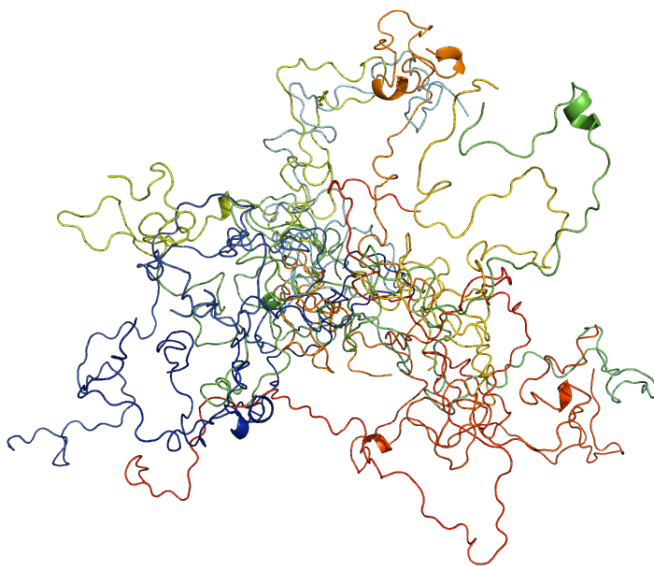


Figure 2.26: Case 1A - Short IDP - Alpha-synuclein - Illustration of multiple conformations.

Case 1A, entirely flexible, exhibited a low mean backtrack b_{mean} of 12 and a high rate of conformations per second $v = 32.5$, demonstrating the efficiency of MoMA-FReSa.

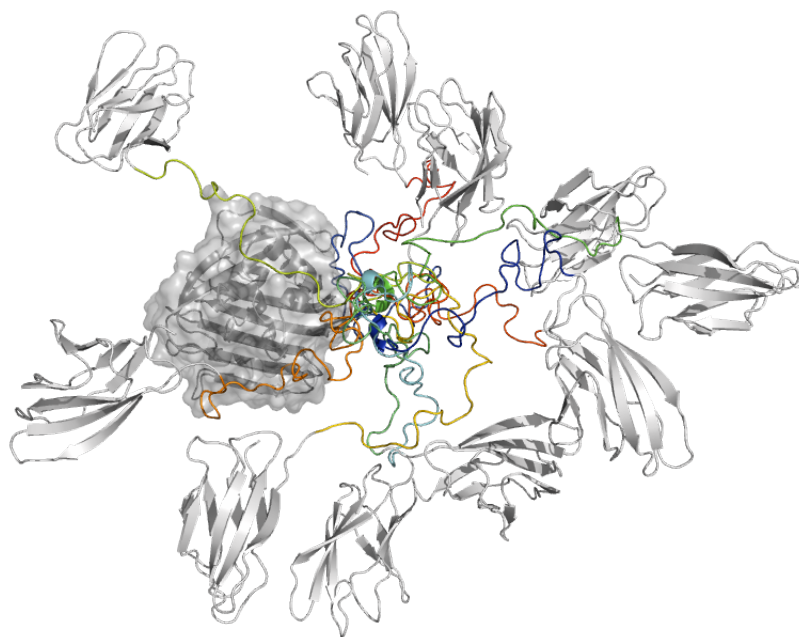


Figure 2.27: Case 1B - Linker - Chimeric linker between a CBM and a GH11 domains - Illustration of multiple conformations.

Similar to Case 1A, Case 1B has a low b_{mean} and a very high v of 55.4 conformations per second, indicating efficient handling of single-linker systems.

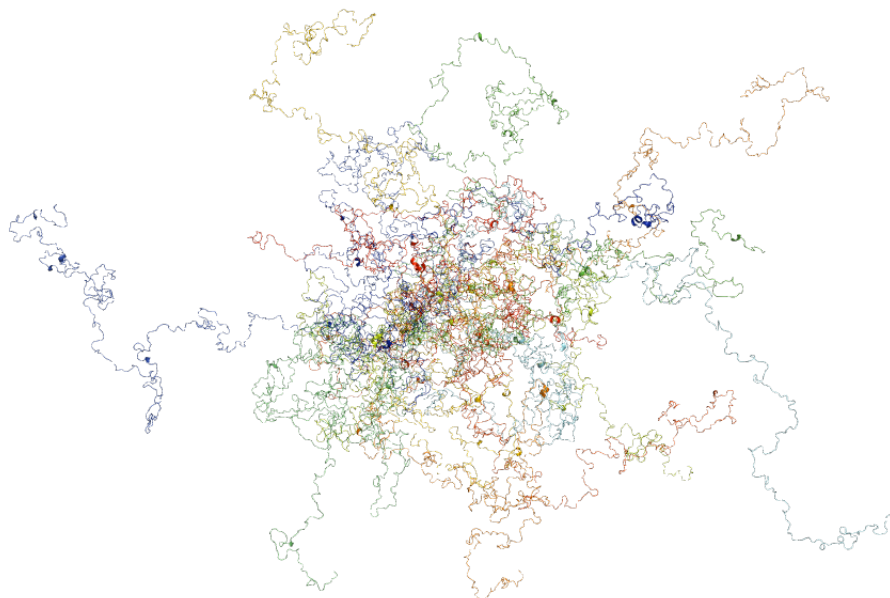


Figure 2.28: Case 1C - Long IDP - Silk protein - Illustration of multiple conformations.

Compared to Case 1A, Case 1C involved a significantly longer disordered region (roughly ten times larger). This translated to a tenfold increase in sampling duration t and a higher b_{mean} of 84. Despite the complexity of a very long IDP, MoMA-FReSa remained efficient with an v of 4 conformations per second.

Class 2

Loops inherently pose a greater challenge for sampling, reflected in consistently high b_{mean} and b_{std} values for all cases in Class 2.

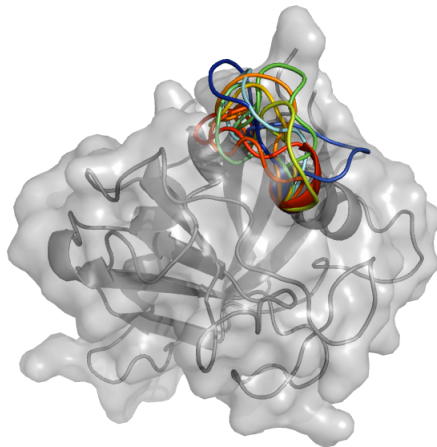


Figure 2.29: Case 2A - Short loop - Human cyclophilin - Illustration of multiple conformations.

Despite being a short loop of only nine residues, Case 2A exhibited a b_{mean} of 96 and a long duration d of 123.9 seconds. This highlighted the inherent difficulty of loop sampling, even for short ones. However, the success ratio remained high at $r = 81.4\%$.

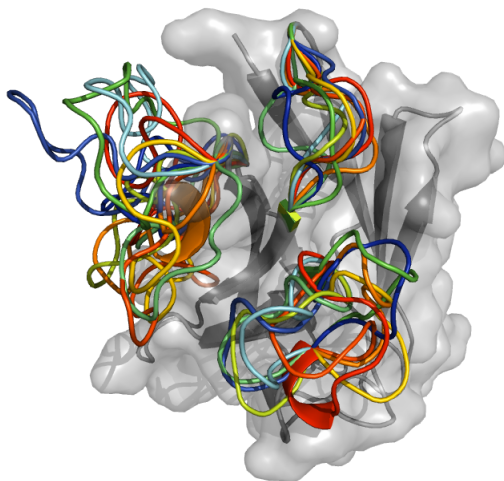


Figure 2.30: Case 2B - Multiple loops - CDR loops of a nanobody - Illustration of multiple conformations.

Case 2B involved three loops (short and medium length of 7, 9 and 16 residues). This complexity reduced the success ratio to $r = 66.8\%$ and increased the b_{mean} to 159. Nevertheless, MoMA-FReSa still achieved a respectable sampling rate r of over 5 conformations per second.

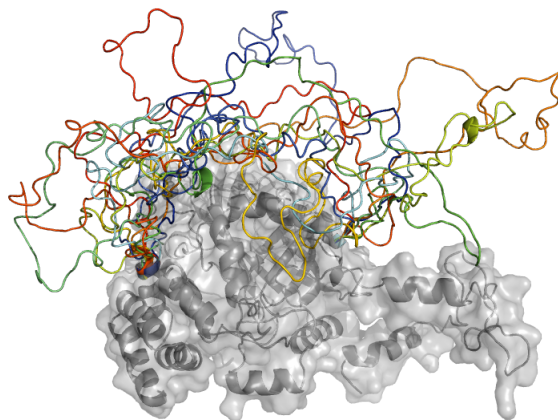


Figure 2.31: Case 2C - Long loop - Beef liver catalase - Illustration of multiple conformations.

As a long loop of 75 residues, Case 2C presented one of the most challenging scenarios for MoMA-FReSa. The number of conformations per second dropped significantly to $v = 0.5$, and the success ratio reduced to $r = 15.2\%$. As expected, both b_{mean} and b_{std} were high. Despite the lower efficiency, MoMA-FReSa was still capable of handling this case and produce a large ensemble in a reasonable time.

Class 3

Class 3 encompasses diverse cases with a wide range of results.

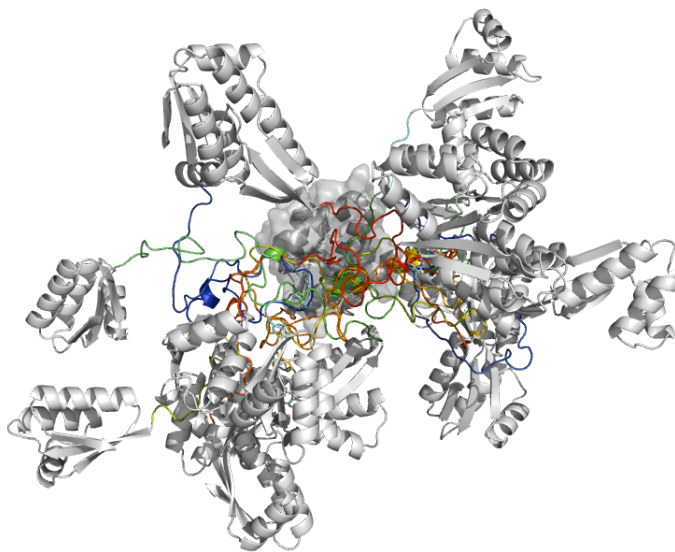


Figure 2.32: Case 3A - Independent linkers - Ribosomal protein L12 - Illustration of multiple conformations.

Case 3A, consisting of a dimer with two 22-residue long linkers, maintained excellent performance with a near-perfect success ratio, low backtrack values, and a capacity of $v = 18$ conformations per second. This outcome aligned with the strong performance observed in Case 1B.

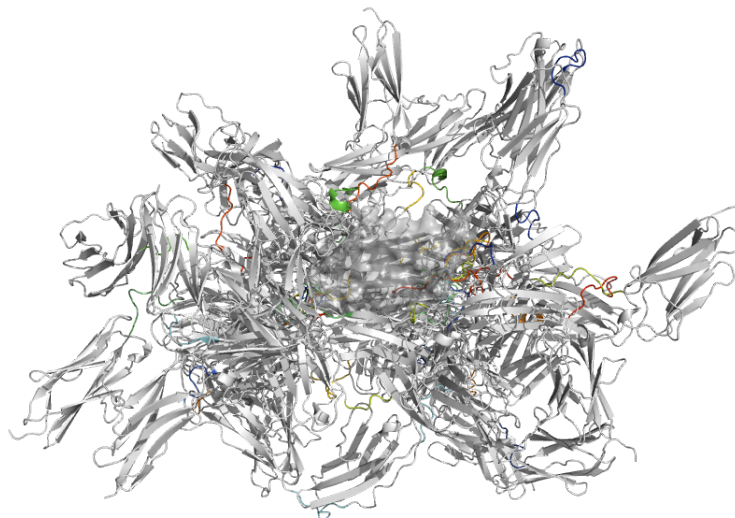


Figure 2.33: Case 3B - Series of linkers - Titin protein - Illustration of multiple conformations.

Similar to Case 3A, Case 3B involved several linkers. However, the concatenation of short linkers presented a more intricate challenge compared to the independent linkers in the previous case. MoMA-FReSa still demonstrated good performance with a sampling rate of $v = 4.1$ conformations per second and a success ratio of $r = 75.6\%$. While the b_{mean} of 54 is relatively high for linker cases, the high success ratio suggested the method effectively avoided getting trapped.

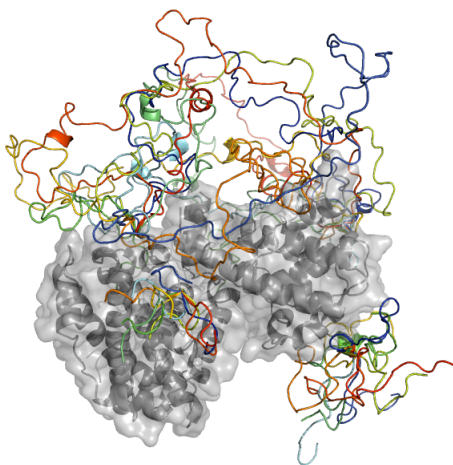


Figure 2.34: Case 3C - Complex system - Retinoblastoma protein domain in complex with adenovirus E1A - Illustration of multiple conformations.

The final case, Case 3C, incorporated a long loop of 71 residues (nearly as long as the loop in Case 2C), along with two tails and a large rigid obstacle. This complexity translated to lower performance compared to Case 2C: a success ratio of $r = 11.9\%$ and a sampling rate of $v = 0.3$ conformations per second. The backtrack values were similar to those observed in Case 2B.

Benchmark summary and discussion

This analysis highlighted the strengths and weaknesses of MoMA-FReSa across different protein system complexities. Finally, we generalize some patterns between the types of systems:

- **IDPs and Linkers:** These systems generally exhibit high success rates ($r \approx 100\%$) except for concatenated linkers, which significantly increase sampling complexity. However, even in this case, the success rate remains relatively high. The duration t for IDPs is particularly favorable considering their sizes. Interestingly, we observe $b_{mean} \approx b_{std}$, suggesting a high degree of variability in the backtracking procedures for these cases.
- **Loops:** Loops present the most challenging scenario, reflected in lower success ratios r and significantly longer duration t compared to non-loop cases of similar size. The loop length seems to be a more critical factor than the number of loops in determining the difficulty. In other words, a longer loop, represents a greater challenge than multiple shorter loops for MoMA-FReSa.

These observations provide valuable insights into the performance of MoMA-FReSa on various protein structures. The method demonstrates very good efficiency for handling simpler systems like IDPs and linkers. While it can still manage loops and more complex scenarios, the efficiency and success rates decrease as the difficulty level increases. By focusing on the number of flexible residues n_F and duration t , we observed slightly differences between the three types of regions:

- **IDPs:** IDPs exhibit a nearly linear relationship between n_F and t . This implied similar sampling times by residue for short and long IDPs, with a ratio $\frac{n_F}{t} \approx 4.6$ in our benchmark. MoMA-FReSa efficiently handles these cases due to the absence of complex geometric constraints.
- **Linkers:** While linkers initially show a linear trend similar to IDPs in single-linker studies, the presence of multiple linkers significantly alters this behavior. Independent linkers like Case 3A introduce geometric constraints, impacting sampling time. This effect is even more pronounced in series of linkers like Case 3B. Notably, cases 1B and 3B had a similar number of flexible residues, but Case 3B required 13 times longer to sample due to these geometric constraints. Overcoming these constraints in linkers remains a challenge.
- **Loops:** Loops present the most significant challenge for MoMA-FReSa. Unlike linkers, the number of loops itself is less influential than their size and complexity. Case 2C exemplifies this, where the $\frac{n_F}{t}$ ratio was 1.8 times lower compared to Case 2A due to a larger loop structure. Additionally, geometric constraints in loops, which are difficult to quantify with figures, further complicate the sampling process. Even though the loop in Case 3C was slightly smaller than Case 2C, its sampling duration t was considerably longer, indicating a more intricate structure and highlighting the substantial challenges posed by loops.

While MoMA-FReSa demonstrated overall effectiveness across various system types, certain areas require further exploration for performance optimization. Linkers and especially loops present complexities that warrant further research and potential improvements. Subsection 2.3.3 details an example of such improvements specifically designed for loops.

2.3.3 Study of the effect of the new distance test for loops

This subsection explores the effectiveness of a new distance test designed to enhance loop sampling. Previously, our primary method for assessing loop closure feasibility relied on a binary check: attempts were rejected if the maximum feasible distance ($max_feasible$) was less than the actual distance ($dist$), $max_feasible < dist$, and accepted otherwise. To improve exploration and mitigate potential traps during sampling, we introduced a stochastic test, as detailed in Subsection 2.2.5. In this new version, the test succeeds if $max_feasible \geq \frac{3}{2}dist$, and it is subject to stochastic choice when $dist \leq max_feasible < \frac{3}{2}dist$.

The primary motivation for this update was to improve performance for complex loop cases, particularly long loops. Here, we compare the results obtained with and without the upgraded test on Case 3C. The results are presented in Table 2.3.

Problem case	t	v	r	$b_{mean} \pm b_{std}$
3C before test upgrade	5481.0 s	0.2 conf/s	5.8%	177 \pm 77
3C after test upgrade	3027.0 s	0.3 conf/s	11.9%	180 \pm 75

Table 2.3: Comparison of results regarding the test upgrade on Case 3C.

These results show that the updated test leads to significant improvements in both time and success rate. We observe a reduction in sampling duration by a factor of 1.8 ($t_{old} = 1.8 \times t_{current}$) and an increase in success rate by a factor of 2 ($r_{current} = 2 \times r_{old}$).

Interestingly, the backtrack information seems relatively unchanged between the two methods. This suggests that the previously successful conformations were likely those that benefited from efficient backtracking strategies. The new method does not alter these successful conformations but rather reduces the number of unsuccessful attempts that get trapped in unproductive backtracking cycles.

Problem case	t	v	r	$b_{mean} \pm b_{std}$
2A before test upgrade	149.8 s	6.7 conf/s	76.1%	109 \pm 79
2A after test upgrade	123.9 s	8.1 conf/s	81.4%	96 \pm 72
2B before test upgrade	317.5 s	3.2 conf/s	47.1%	175 \pm 72
2B after test upgrade	191.1 s	5.3 conf/s	66.8%	159 \pm 70
2C before test upgrade	3176 s	0.3 conf/s	7.9%	185 \pm 74
2C after test upgrade	1851.0 s	0.5 conf/s	15.2%	178 \pm 74

Table 2.4: Comparison of results regarding the test upgrade on cases of class 2.

While initially intended to enhance performance for long loops, we observed similar trends across all systems with loops (including short and medium loops) as shown in Table 2.4: shorter sampling times t and higher success rates r with the upgraded test. We can see that the higher the complexity the more effective is the new approach. Notably, the improvement in Case 2C, which involves a similar long loop as Case 3C, seems quantitatively identical to this case (time reduction by 1.7 and success rate augmentation by a factor 1.9). Additionally, similar to Case 3C, the impact on backtracks for successful conformations seems negligible on all these cases. These findings demonstrate that the new distance test not only achieves its primary objective of improving long loop sampling but also provides general advantages for loop-containing systems, regardless of their length.

2.4 Discussion and conclusion

Our research demonstrated the effectiveness of MoMA-FReSa as a general sampling method. It excels in its adaptability to a broad range of systems, including very complex ones. MoMA-FReSa consistently delivered robust results across various test cases. Even in more intricate cases like very long IDPs or concatenated linkers, the results remain very satisfying. While further research on linkers could be beneficial, the current capabilities of MoMA-FReSa are excellent. Notably, it can generate conformational ensembles for challenging loops within reasonable time, especially after the implementation of the current loop distance calculation method.

To continue to overcome inherent loop sampling difficulties, we can consider machine learning methods to improve performances. Chapter 4 details a Reinforcement Learning approach specifically designed for MoMA-FReSa. This method aims at learning favorable residue conformations across Ramachandran space in the context of a specific biomolecular system. While the main interest of this approach is to improve loops, linkers can also benefit of this improvement. We also envisage to implement additional energy functions to evaluate and rank the sampled conformational states, allowing them to be weighted according to their Boltzmann probability [235].

Despite potential for further improvements, MoMA-FReSa already samples diverse systems efficiently. Chapter 3 explores its applicability in complex systems in the context of collaborations, further highlighting its vast potential beyond regular sampling needs.

Applications of MoMA-FReSa

Contents

3.1	Introduction	61
3.2	Conception of multi-modular systems	62
3.2.1	General presentation	62
3.2.2	Conception work	65
3.2.3	Discussion and perspectives	69
3.3	Estimation of the effective concentration in biomolecular interactions	69
3.3.1	General presentation	69
3.3.2	Results and analyses	72
3.3.3	Discussion and perspectives	75
3.4	Structural analysis of linkers in multi-domain proteins	76
3.4.1	General presentation	76
3.4.2	Post-processing analysis	77
3.4.3	Discussion and perspectives	80
3.5	Sampling of structural motifs	81
3.5.1	General presentation	81
3.5.2	Ensemble generation and analysis	82
3.5.3	Discussion and perspectives	87
3.6	Conclusion and perspectives	88

3.1 Introduction

As introduced in Chapter 1, MoMA-FReSa fulfills a critical need in the field of disordered proteins by offering a general and computationally efficient conformational sampling method applicable to diverse types of biomolecular disordered systems. Chapter 2 presented the method and demonstrated its efficiency in generating large conformational ensembles using a benchmark set containing diverse types of protein architectures involving disordered regions.

However, the potential of MoMA-FReSa extends far beyond its initial function. This chapter explores a broad range of diverse and original applications. We aim to demonstrate how MoMA-FReSa transcends its role as a general sampling algorithm and transforms into a multifaceted bioinformatics tool. The following sections address a spectrum of potential uses for MoMA-FReSa and offers a view on future possibilities. These applications have been developed in a collaborative way, emphasizing the ability of MoMA-FReSa to address user-specific inquiries within concrete contexts.

This chapter is structured into four distinct applications. Section 3.2 provides insights into the role of MoMA-FReSa in the design of multi-modular proteins. Section 3.3 shows its

potential in estimating effective concentrations in intramolecular or intermolecular interactions. Section 3.4 explores deeper how MoMA-FReSa can be employed for structural studies of multi-domain proteins and gives examples of post-processing analyses. Section 3.5 focuses on the application of MoMA-FReSa in the analysis of structural motifs and presents the diversity of conformational sampling methods provided by MoMA-FReSa. We conclude this chapter in Section 3.6 by discussing the perspectives highlighted by these diverse applications.

3.2 Conception of multi-modular systems

3.2.1 General presentation

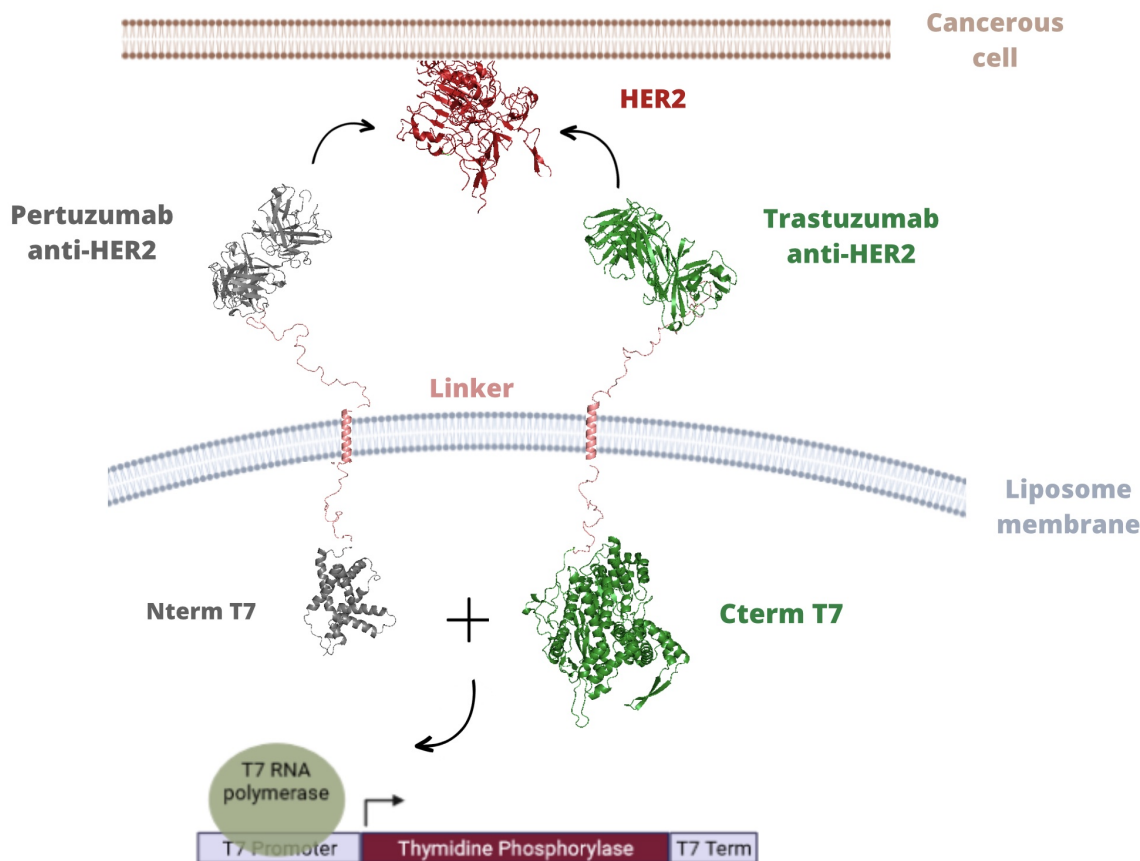


Figure 3.1: Presentation of the iGEM biosensing signal transduction strategy. Recognition of HER2 extracellular domain induces functional assembly of the split T7 RNA polymerase, which enables gene expression of target gene under control of a T7 promoter. Extracted from: <https://2023.igem.wiki/toulouse-insa-ups/structural-optimization>.

In the context of the International Genetically Engineered Machine (iGEM) synthetic biology competition, a team from *Toulouse Institut National des Sciences Appliquées - Université Paul Sabatier* (INSA-UPS) developed CALIPSO, a targeted drug delivery system designed for cancer treatment (<https://2023.igem.wiki/toulouse-insa-ups/home>). CALIPSO leverages synthetic biology to achieve localized drug production within liposomes, carrier vesicles specifically targeted to cancer cells. A critical aspect of CALIPSO resides in the design of a protein complex that can recognize cancer cells and trigger drug production on-site. However, this

protein complex presents a structural challenge as ensuring proper interaction between its various components requires meticulous optimization of linker sequences.

The core functionality of this protein complex is based on a split T7 RNA polymerase, divided into two fragments, which is inside of the liposome. Each subunit is linked to an anti-HER2 antibody through a transmembrane segment, as seen in Figure 3.1. When these antibodies bind to HER2, on the surface of a cancer cell, they bring the two T7 RNA polymerase fragments closer together. This allows them to assemble into a functional enzyme capable of initiating the transcription of a target gene located inside the liposome. Note that the two transmembrane helices are supposed to dimerize during this process forming an anti-parallel dimer.

A structural design problem

The success of this strategy resides in the design of optimal linker sequences connecting the antibodies and the T7 RNA polymerase fragments. These linkers need to be carefully designed to ensure proper antibody-HER2 binding and efficient T7 polymerase fragment assembly. Indeed, on the one side, the linkers must not hinder the ability of the antibodies to bind tightly to HER2, ensuring specific recognition of cancer cells. On the other side, the linker lengths need to be strategically chosen to allow the T7 RNA polymerase fragments to come close enough to assemble, while allowing some mobility with respect to the lipid membrane.

This section focuses on the flexible, undefined regions of the linker connecting the functional domains of the complex. While the central transmembrane helical region (TMH), with a known amino-acid sequence (RLILIIVGAIALLVHGF, oriented to the exterior), is assumed to adopt a well-defined secondary structure, the remaining linker regions lack inherent structure and must be simply optimized in terms of length.

The iGEM team required our collaboration to define the flexible parts of their synthetic system using MoMA-FReSa. This section presents this application of MoMA-FReSa on this practical case and the potential of the method in the design of multi-modular proteins for bioengineering uses.

Problem definition

The multi-modular system was considered in a pre-defined configuration where the antibodies are bound to the HER2 receptor and the T7 RNA polymerase fragments are already assembled. The distances between the different parts of the system were fixed in collaboration with the iGEM team. Notably, the distance between the antibodies and the membrane was chosen to be a good compromise between a short distance to allow the optimal formation of the complex and a reasonable distance to the membrane, in order to have enough flexibility during the assembly process. The distance of the gap between the two transmembrane helical motifs was also fixed to around 14Å. During the next steps of this study, we considered the model of the system defined in Figure 3.2, with strong geometric constraints on the placement of the elements.

The system was composed by two chains, each of them containing two unknown flexible regions. Therefore, a total of four regions had to be modelled with MoMA-FReSa in order to correctly define the system in terms of both size and amino-acid composition. In Figure 3.2, these four regions are named:

- \tilde{R}_S : The shorter internal part of the left linker connecting the N-terminus of the T7 RNA polymerase fragment to the TMH.

- \tilde{R}_L : The longer external part of the left linker connecting the transmembrane segment to the Pertuzumab anti-HER2 antibody.
- \tilde{R}_{S_r} : Counterpart of \tilde{R}_S on the right linker, connecting the C-terminus of the T7 RNA polymerase fragment to the TMH.
- \tilde{R}_{L_r} : Counterpart of \tilde{R}_L on the right linker, connecting the Trastuzumab anti-HER2 antibody to the TMH.

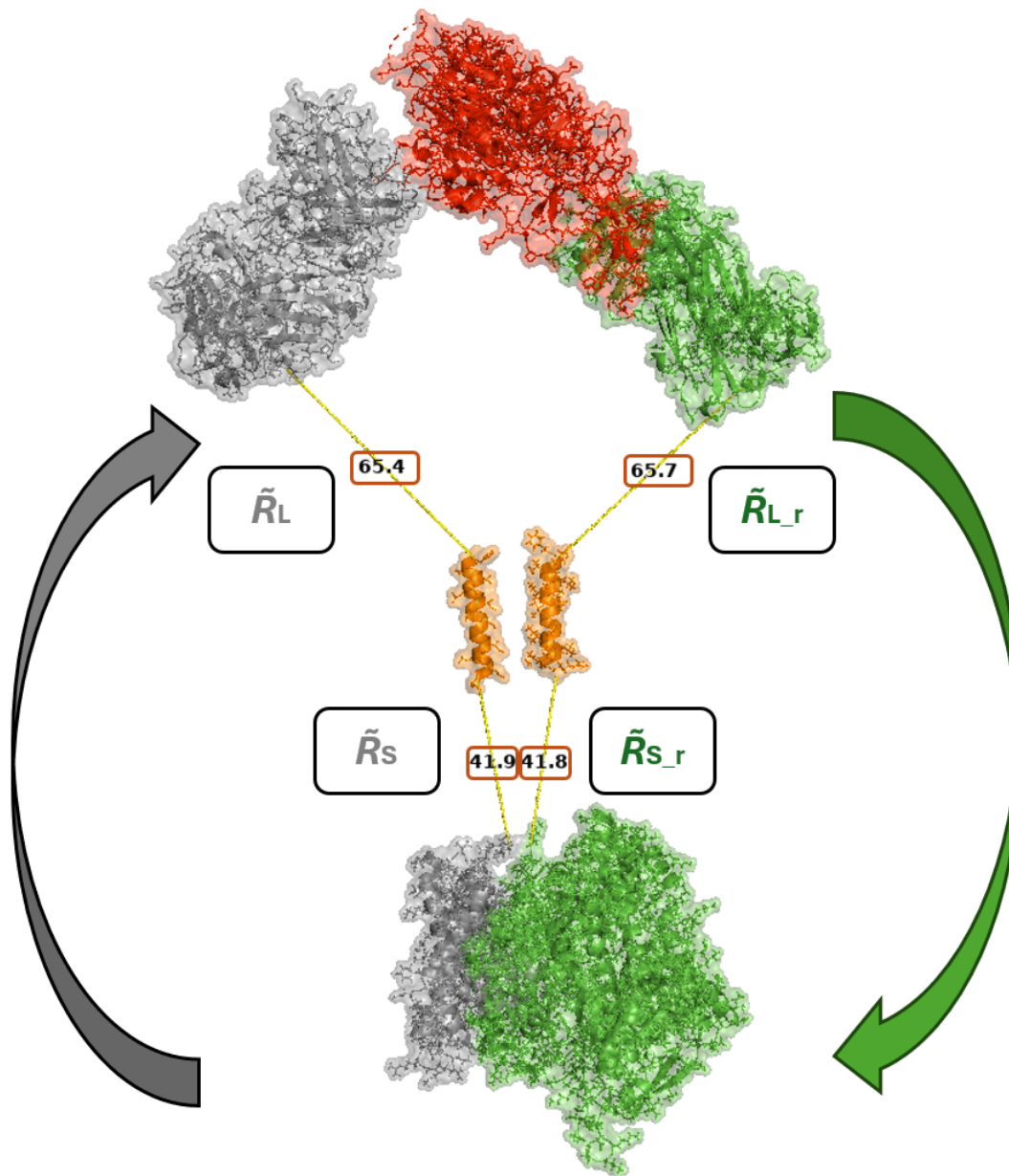


Figure 3.2: Presentation of the model studied in collaboration with the iGEM INSA-UPS team. Grey construct (left): N-terminus T7 RNA polymerase fragment and Pertuzumab anti-HER2 antibody. Green construct (right): C-terminus T7 RNA polymerase fragment and Trastuzumab anti-HER2 antibody. The red molecule is HER2. The arrows show the N-to-C direction of the green and grey constructs. Distances between the anchoring points of the disordered regions of the linkers are given in Å.

In the context of the collaboration, we made two hypothesis to orient our study:

- **Similar Sequence Pattern for Long and Short regions:** We searched a similar amino-acid type of sequences for both long and short regions. We decided to build these sequences on a repetition of a disordered sequence pattern.
- **Opposite Sequences for Reversed Regions:** Due to the opposite orientation of the right and left linker chains, we supposed that regular and reverse regions were mirrors in terms of sequences. In simpler words, \tilde{R}_{S_r} and \tilde{R}_{L_r} sequences were the reversed with respect to \tilde{R}_S and \tilde{R}_L .

We used MoMA-FReSa as a simple proxy to evaluate potential amino-acid sequences for the both types of regions to ensure proper functionality of the entire CALIPSO complex. Note that this was a highly exploratory work to help the iGEM team to subsequently perform more in-depth analyses.

Size analysis protocol

In accordance with the iGEM team, we decided to choose a standard sequence based on a known disordered motif GGGS. Note that this motif is commonly used for synthetic linkers [210]. In effect, this sequence is enriched with Glycine (G) and Serine (S) residues, which are known to promote conformational flexibility.

In these conditions, a $(GGGS)^n$ linker would allow mobility of the connected functional domains. The objective is to find an appropriate length $4 \times n_S$ (n_S repetitions of the GGGS motif) for the region \tilde{R}_S and $4 \times n_L$ (n_L repetitions of the GGGS motif) for the region \tilde{R}_L . The counterpart linker sequences (\tilde{R}_{S_r} and \tilde{R}_{L_r}) were defined to have the same amino-acid sequence as their respective partners (\tilde{R}_S and \tilde{R}_L), but in reverse order: $(SGGG)^{n_S}$ and $(SGGG)^{n_L}$.

In this context, where all the elements were fixed with respect to each other, the four flexible regions could be considered as loops, and we assumed the following hypothesis: the sequence is appropriate (in terms of length) if the probability to generate loop conformations for this sequence is high. If the probability was too low, we would had considered that the geometric constraints are too strong with this sequence to allow the wished functionality of the system. On the contrary, a high success rate could imply a more suitable length to achieve the desired end-to-end distance.

Considering the minimum required distances fixed for this system configuration, as shown in Figure 3.2, we would need at least 11 and 18 residues for loops corresponding to 42Å and 66Å, respectively. These numbers were obtained considering that the approximate length of an amino-acid residue is 3.8Å, and assuming that fully-extended conformations are feasible, which is not the case due to steric constraints. Therefore, we began testing with lengths exceeding our initial estimations (16 and 24 residues for R_S and R_L , respectively), due to the lack of successful poses with shorter sequences. To reduce computational cost, we conducted separate tests for short and long disordered regions.

3.2.2 Conception work

Optimization of the internal regions

The results for short internal regions \tilde{R}_S and \tilde{R}_{S_r} are summarized in Table 3.1 and Figure 3.3.

n_S	Size of the region	Success rate (%)
4	16	4.1
5	20	33.5
6	24	22.3
7	28	31.5
8	32	41.9
9	36	41.7
10	40	34.6
11	44	27.9
12	48	18.9

Table 3.1: Table of success rate as a function of the number of motifs, n_S .

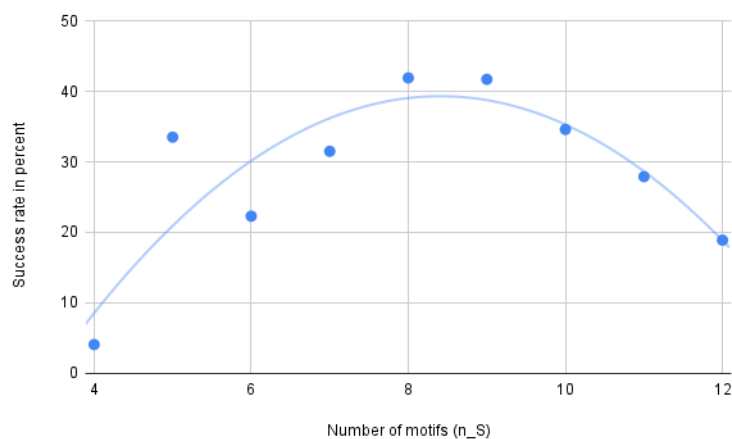


Figure 3.3: Percentage success rate as a function of the number of motifs, n_S .

Our analysis identified a size of 32 residues (4×8) as the optimal length for the short regions \tilde{S}_L and \tilde{S}_{L_r} , since the maximum of the success rate is reached at this size, although a similar rate was obtained with a n_S of 9.

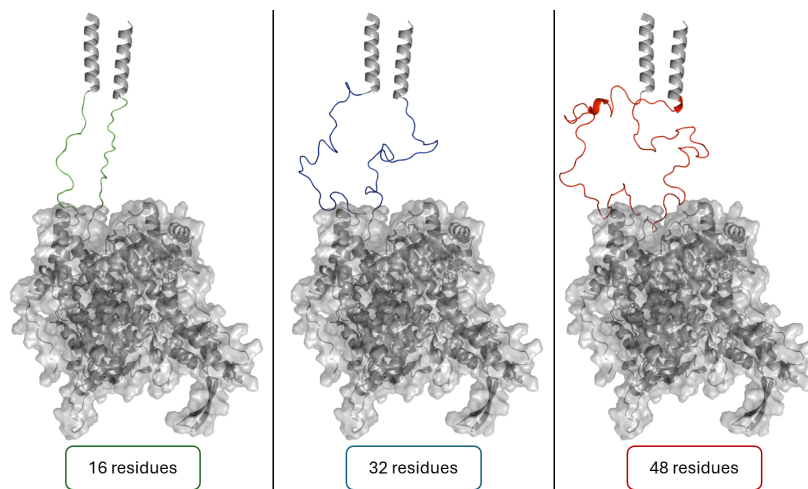


Figure 3.4: Comparison between a 16, 32 and 48 residue-long conformations for \tilde{R}_S and \tilde{R}_{S_r} .

For comparison, Figure 3.4 also shows results for 16- and 48-residue long sequences. Shorter loops likely lacked sufficient conformational freedom, while longer ones might become bulky and introduce clashes, leading in the two cases to lower success rates. According to our calculations, the 32-residue long sequence represents a good balance.

Optimization of the external regions

The results for long external regions \tilde{R}_L and \tilde{R}_{L_r} are summarized in Table 3.2 and Figure 3.5.

n_L	Size of the region	Success Rate (%)
8	32	2.4
9	36	2.2
10	40	5.8
11	44	9.5
12	48	11.4
13	52	13.7
14	56	14.3
15	60	17.2

Table 3.2: Table of success rate as a function of the number of motif repeats, n_L .

For the longer regions connecting the external components of the CALIPSO design, the success rate continued to linearly increase with n_L and we supposed that, at some point, the plot would describe a concave polynomial curve as in Figure 3.3. However, reaching the maximum of this curve was not necessarily the best choice for this configuration. Indeed, a size of 52 residues (4×13) provided a sufficient success rate for our purposes.

Regarding the initial problem, longer sizes, although potentially improving success rates further, would keep the liposome away from the cell and threaten the correct functioning of the system. As we fixed the distance of HER2 in our configuration, we could not illustrate this issue. However, these longer sequences could lead to linker conformations folded towards the membrane in our constructions. Figure 3.6 illustrates this potential issue with a 60-residue long region.

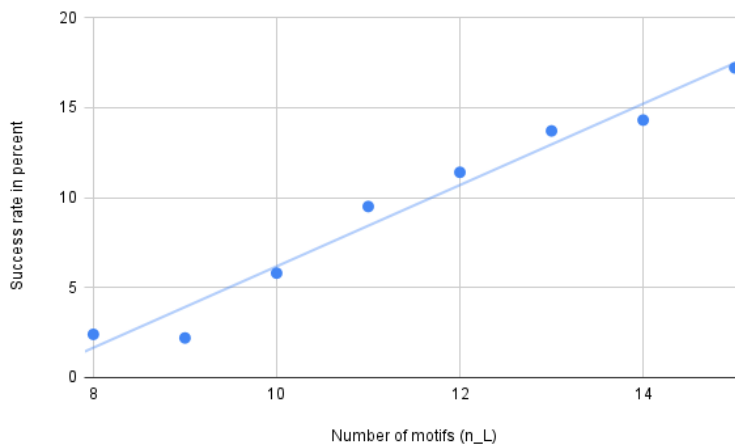


Figure 3.5: Percentage success rate as a function of the number of motifs, n_S .

Similar to the \tilde{R}_S and \tilde{R}_{S_r} regions, shorter sizes for long linkers \tilde{R}_L and \tilde{R}_{L_r} were also not ideal. Shorter sequences may not consistently reach the required distance between functional domains, leading to lower success rates.

In conclusion, we kept the 52-residues size as an optimal length. This size offers a balance between achieving the necessary distance, maintaining a high success rate, and avoiding potential issues associated with excessively long linkers.

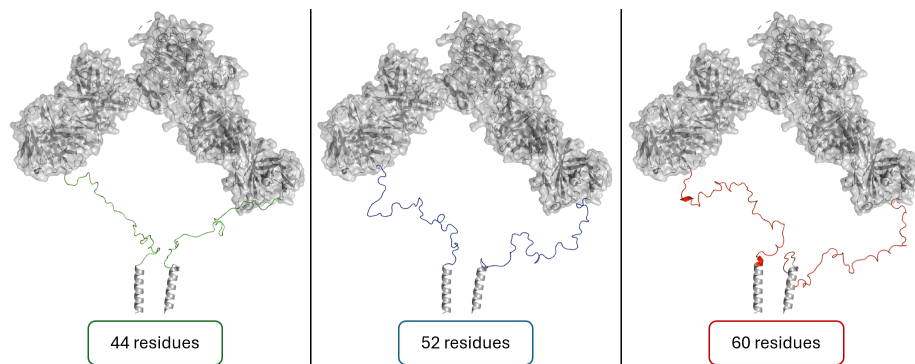


Figure 3.6: Comparison between a 44, 52 and 60 residue-long conformations for \tilde{R}_L and \tilde{R}_{L_r} .

Assessment of the complete design

To validate our analysis, we generated global conformations with short and long regions simultaneously: a repetition of the GGGs amino-acid, 8 times for \tilde{R}_S and 13 times for \tilde{R}_L , and similarly with SGGG for \tilde{R}_{S_r} and \tilde{R}_{L_r} . MoMA-FReSa handled to generate 1000 conformations with a success rate of 11.4%, a reasonable value for this type of complex cases involving long loops. Some results are presented in Figure 3.7.

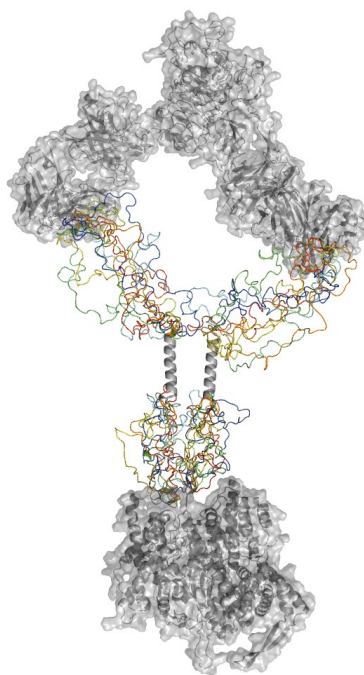


Figure 3.7: Illustration of 10 conformations of the complete system.

The model built for the CALIPSO protein with MoMa-FRESA has been fundamental for subsequent experimental design performed by the iGEM team.

3.2.3 Discussion and perspectives

In this study, we explored the potential of MoMA-FReSa for designing multi-domain proteins, using a concrete example. Our initial objective was to determine the optimal size for synthetic linkers in a given system configuration with multiple constraints. MoMA-FReSa offered several advantages in this process. First, it significantly reduced research time by allowing us to study the internal and external parts of the system independently. Additionally, the flexibility of the method allowed us to treat the regions as loops, fixing the distance between the different components of the system.

Through this example, we showed the potential of MoMA-FReSa for multi-domain protein design and how the program can take advantage of its versatility to adapt to user constraints and hypotheses. By our final generation of a conformational ensemble, we successfully obtained suitable structures and validated the method. In all this process, the optimized sampling method of MoMA-FReSa facilitated rapid prototyping of multi-modular constructs and the testing of hypotheses related to linker selection. These results are particularly significant and encouraging in the context of loop problems that are notoriously complex and time-consuming. Even if this case was highly exploratory, synthetic biology is a challenging domain and this first approach in this field by MoMA-FReSa is promising.

3.3 Estimation of the effective concentration in biomolecular interactions

3.3.1 General presentation

In collaboration with the group of Dr. Lucia Chemes, at the University of San Martin (Argentina), we explored the application of MoMA-FReSa to estimate the effective concentration of tethered ligands. The effective concentration (C_{eff}) represents the free ligand concentration required to achieve an encounter rate with the binding site equivalent to that of the tethered ligand. This parameter plays a crucial role to quantitatively understand the interaction of biomolecules containing two or more binding sites connected through flexible regions. Indeed, the C_{eff} depends on the length and the structural properties of the linker [106].

The formula for effective concentration (C_{eff}) is given by:

$$C_{\text{eff}}(r) = \frac{p(r)}{4\pi r^2} \frac{10^{27} \text{\AA}^3 L^{-1}}{N_A}$$

where r is the distance between both binding sites in the complex, and $p(r)$ is the distance distribution function between the two binding sites of the flexible protein [105]. The first term represents the probability distribution function divided by the surface area of a sphere with radius r . The second term converts the result into molar concentration.

Historically, C_{eff} estimations have been derived from analytical models inspired from polymer science [232]. In this section, exploiting the power of MoMA-FReSa, we explored how to estimate C_{eff} by generating large ensembles of the linker conformations and calculating the proportion of conformations where the ligand was bound to the target site [105].

System definition

The system under consideration involves the p107 AB domain (RBL1_HUMAN, UniProt ID: P28749) of a Rb tumor suppressor protein, named RbAB from now on [164]. The viral E1A protein competes with host factors to bind RbAB, subverting cell cycle regulation. This interaction is of higher affinity than the human one thanks to the two RbAB binding motifs present in E1A, which are named the LxCxE and the E2F. These two motifs interact with two clefts of RbAB, which we will name LxCxE-cleft and E2F-cleft, for simplicity. These two motifs are separated by a disordered linker that, when both motifs interact with RbAB, behaves as a loop [81].

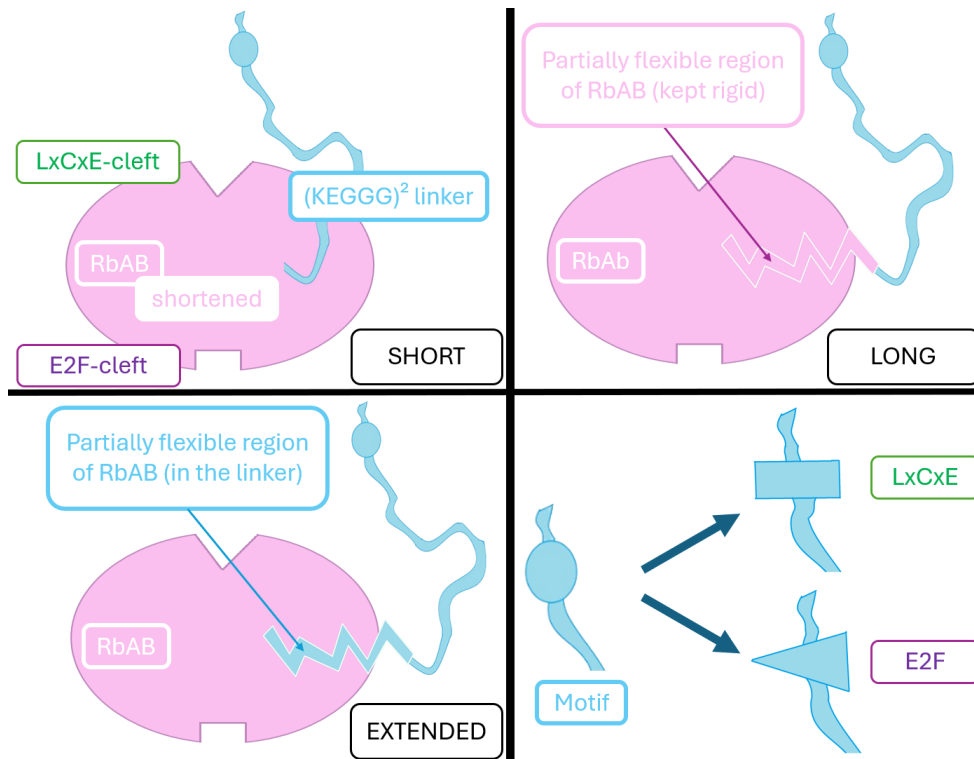


Figure 3.8: Definition of the system constructions: Short (with an ablation of a partially flexible region in RbAB) - Long (with the partially flexible region in RbAB considered to be rigid) - Extended (with the partially flexible region considered as part of the linker). The binding motif of the linker can be LxCxE or E2F.

In the group of L. Chemes, they have built two synthetic proteins based on the E1A-RbAB complex to study the effect of the linker length on the C_{eff} . First, a $(KEGGG)^2$ synthetic linker of 10 residues was fused to the RbAB, ending by one of the two E1A motifs, which would recognize one of the two clefts on the RbAB surface. The second one is a similar system, with the deletion of 24 residues from the partially flexible/disordered region of RbAB, as detailed below. Independently of the specific E1A motif attached, three different system constructions were considered from the computational perspective:

- **Short Construction (344-residues rigid domain, 10-residues linker):** A shortened version of the RbAB with a 24-residue deletion at its C-terminus, corresponding to a partially flexible/disordered region. The $(KEGGG)^2$ linker was directly attached to this new C-terminus.

- **Long Construction (368-residues rigid domain, 10-residues linker):** The original construction with no deletions. The 10-residues (KEGGG)² was attached to the original C-terminus.
- **Extended Construction (344-residues rigid domain, 34-residues linker):** Similar to the long construction with no deletions of the 24 residues at the C-terminus, but this time these residues were considered to belonging to the linker instead of to the rigid domain.

Figure 3.8 summarizes the different constructions of the system.

Contact assessment method

The explicit exploration of conformations to derive C_{eff} requires a definition of the bound state for the second site. This remains a subjective point as the identification of conformations successfully attained both binding sites requires a definition of the bound state, which has to be flexible enough to represent encounter complexes.

To address this point, we developed a post-processing tool able to check binding in a resulting conformation. MoMA-FReSa allows users to define a list of pairs of residues, a distance threshold d_{bind} and a metric to evaluate successful binding. Therefore, in order to classify a given conformation as successful, the post-processing procedure checks if there is binding according to the previously mentioned list of parameters.

Two residues are considered to interact if the distance between their C_{β} is less than the given threshold d_{bind} . With this in mind, MoMA-FReSa offers three different metrics to evaluate successful binding from a given conformation:

- **One Point:** The binding is successful if at least one of the pairs of residues interacts.
- **All Points:** The binding is successful if all the pairs of residues interact simultaneously.
- **Average:** The binding is successful if, on average, all pairs are in interaction. To check this, the mean distance between the residue pairs d_m is computed. If $d_m \leq d_{\text{bind}}$ the binding is considered successful.

It is worth noting that these distinct metrics offer more or less relaxed definitions of a successful interaction and, therefore different resulting C_{eff} . It is also important to note that these three metrics become equivalent when only one pair of residues is tested.

MoMA-FReSa reports the percentage of valid conformations that meet the binding criteria and, for the "One Point" metric specifically, it identifies the first interacting residue pair (if any). Using this tool, an estimation of the effective concentration could be achieved in a more precise and in a system-specific manner than previous numerical approaches [105]. We define a new formula for C_{eff} with MoMA-FReSa:

$$C_{\text{eff}} = \frac{R_{\text{eff}}}{V_{\text{eff}}} \frac{10^{27} \text{\AA}^3 L^{-1}}{N_A}$$

with R_{eff} the effective ratio of bound conformations (valid confirmations that pass the binding test) over all valid conformations and V_{eff} the effective volume. The V_{eff} can be computed in different ways depending on the exact definition of the metrics used.

To assess the capabilities of this tool, we performed a series of studies on the previously defined system. The impact of the chosen metric was studied in a preliminary work (not

presented in this manuscript), so we only used the "One Point" metric in the following analysis to facilitate the comparison on other factors. Note that we did not directly tackle the problem of computing the effective concentration, since we have not addressed yet the computation of V_{eff} . Therefore, only R_{eff} is reported.

3.3.2 Results and analyses

First case study: Construction and distance comparisons

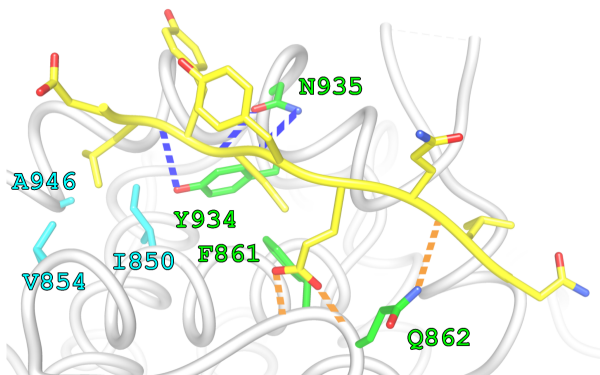


Figure 3.9: E1A-RbAB interactions. E1A and RbAB are depicted as yellow and white ribbons, respectively. RbAB residues that are involved in H-bonds are depicted as green sticks. RbAB residues that interact through hydrophobic interactions with the L amino-acid of the LxCxE-cleft are depicted as cyan sticks.

In our first application, we investigated the binding preference of RbAB in complex with E7 peptide (PDB ID: 4YOZ). This study focused on the LxCxE-cleft, which offered multiple potential pairs of residues to define a binding event (Figure 3.9). We selected four pairs based on reported distances.

We conducted two initial studies to assess the capabilities of MoMA-FReSa in this context. The first test aimed to identify the optimal system construction (short, long or extended) for the LxCxE-cleft. These results from MoMA-FReSa calculations will be then compared with experimental data (work in progress). We performed a preliminary study aiming at generating 10 bound conformations and compared the rates of R_{eff} . These evaluations were conducted using a distance threshold of $d_{\text{bind}} = 10\text{\AA}$. Table 3.3 summarizes these results.

Construction	Short	Long	Extended
R_{eff}	1.58%	0.11%	0.14%

Table 3.3: Table of the rate R_{eff} in percent for the different constructions.

The results indicated a clear preference for the short construction, where the linker extremity was positioned near the cleft. Between long and extended, extended displayed a slightly better rate. The addition of flexibility of the extended scenario slightly improved the percentage of successful conformations.

Our second study focused on the effect of the distance threshold d_{bind} on R_{eff} . We generated a large ensemble of 50000 valid conformations using the preferred short construction. We then

analysed the evolution of the number of bound conformations and R_{eff} rate throughout the generation process. The results are presented in Figures 3.10 and 3.11.

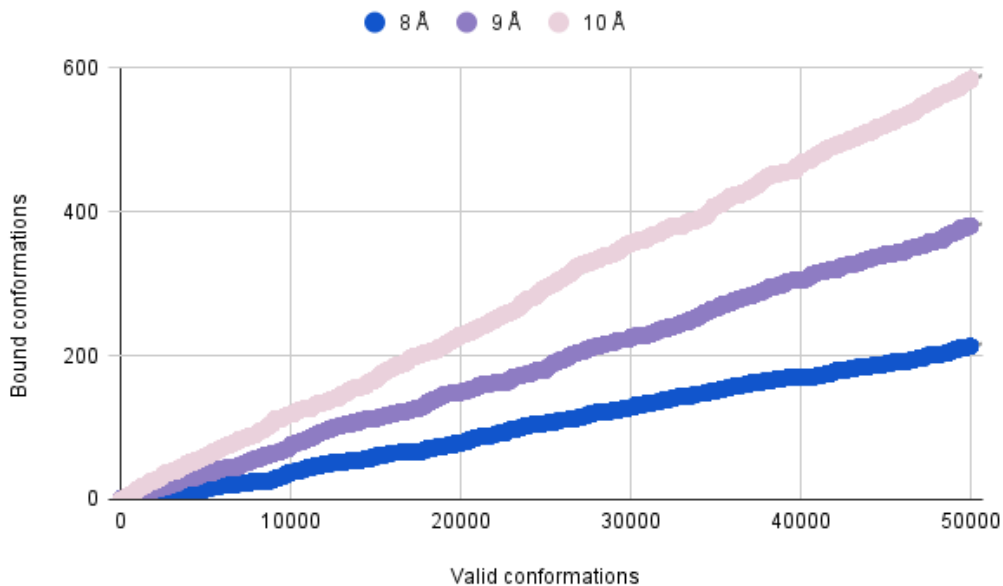


Figure 3.10: Evolution of the number of bound conformations during sampling.

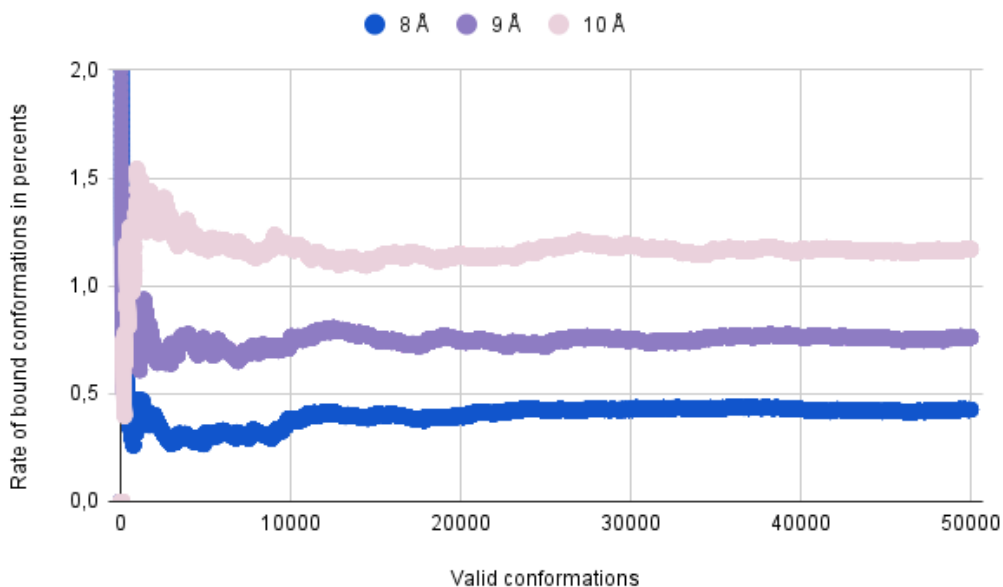


Figure 3.11: Evolution of the rate of bound conformations R_{eff} (percentage) during sampling.

As anticipated, the reduction of the distance threshold led to a decrease of R_{eff} . Interestingly, the bound rate appeared to stabilize around 20000 valid conformations: 0.43% for 8Å, 0.76% for 9Å and 1.17% for 10Å.

Second case study: Binding site comparison and choice of the pairs

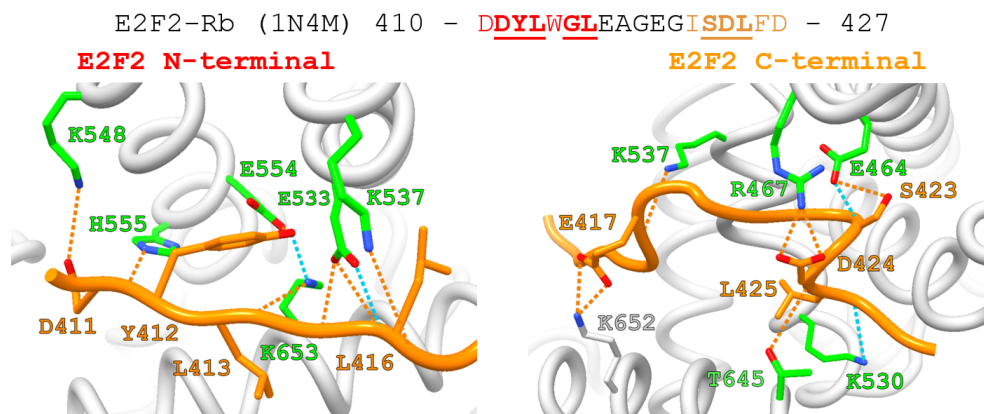


Figure 3.12: Structure of the E2F2-RbAB complex. Intermolecular hydrogen bonds in the N-terminal (left) and C-terminal (right) segments of the E2F2 peptide (orange) with the RbAB groove (PDB: 1N4M). Blue dashed lines: H-bonds fulfilling optimal criteria. Orange dashed lines: H-bonds detected by using a relaxed criterion.

Our second case study explored the binding preference of RbAB in complex with the E2F2 peptide (PDB ID: 1N4M). The E2F2 peptide exhibits two distinct secondary structures: the N-terminal residues (410–416) form an extended β -strand conformation (red sequence, to the left, in Figure 3.12, while the C-terminal residues (422–427) adopt a more twisted structure (orange sequence, to the right, in Figure 3.12).

This study aimed at determining whether the N-terminal or the C-terminal binding motif of E2F2 is preferred to bind on the E2F-cleft. We selected three amino-acid pairs on each side (N-terminal and C-terminal) for a total of six pairs, based on reported distances. For each system, we generated ensembles of one million valid conformations with $d_{bind} = 10\text{\AA}$ and analysed the bound rate for both N-terminal and C-terminal binding motifs. The results are presented in Figure 3.13.

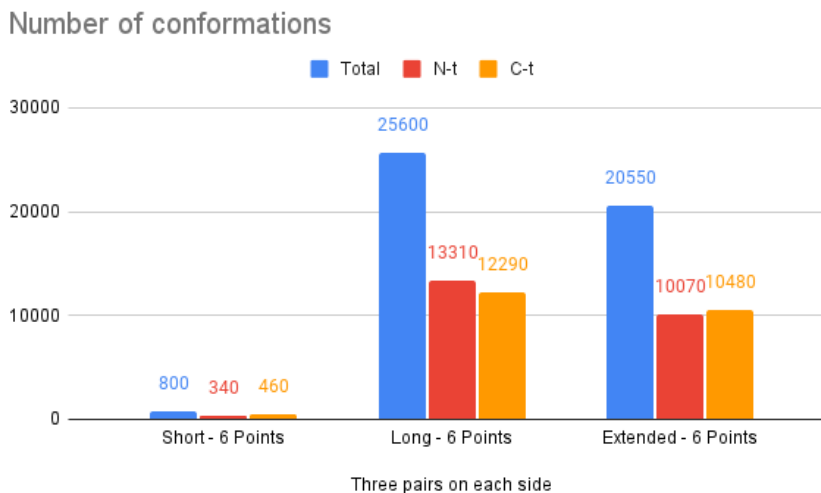


Figure 3.13: Number of bound conformations among the one million generated for each site and each construction, with six pairs of interacting residues used to define a successful binding.

In contrast to our first case study, the results indicated a preference for the long construction in this scenario. Introducing additional flexibility through the extended construction appeared unfavorable, and the short construction reached very low binding rates. We observed a roughly equal distribution of bound conformations and so R_{eff} between the N-terminal and C-terminal motifs (42%, 52% and 49% of N-terminal side binding, for short, long and extended constructions, respectively).

To further investigate these observations, we isolated two representative residue pairs (one for each side) from the initial set of six, aiming to enforce specific orientation of the binding motif on each side. The results for these two pairs are shown in Figure 3.14. This time, the distribution of bound conformations became significantly biased towards the C-terminal motif (3%, 26% and 23% of N-terminal side binding, for short, long and extended construction, respectively). This finding highlighted the importance of careful selection of binding pairs for accurate assessment of binding preferences.

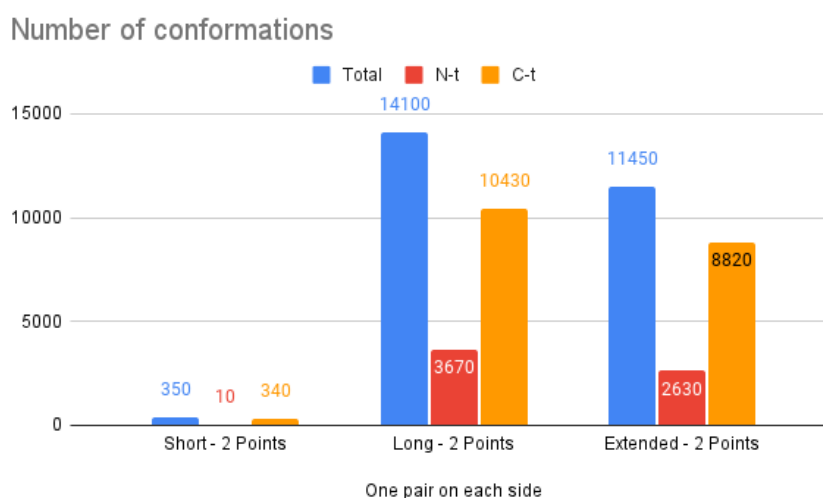


Figure 3.14: Number of bound conformations among the one million generated for each site and each construction, with two pairs of binding residues used to define a successful binding.

3.3.3 Discussion and perspectives

This work presents a preliminary exploration of the potential of MoMA-FReSa in estimating effective concentrations in multi-site binding processes. We developed a promising post-processing procedure to access binding and we conducted a series of studies on a model system provided by our Argentinian collaborators. These studies highlighted the importance of parameter settings in this post-processing tool, notably the good designation of the binding pairs and the selection of an appropriate distance threshold.

We successfully generated estimations of probability distribution R_{eff} for multiple configurations and constructions of the system. These results are promising considering the C_{eff} of these cases are difficult to measure experimentally, computational approaches like MoMA-FReSa can become particularly relevant for these studies.

The next step will be the development of an adequate complete formula to estimate effective concentration with MoMA-FReSa. This will require the calculation of the effective volume of binding V_{eff} , which is not an easy parameter when multiple distance pairs are used to define effective binding. Moreover, our method to estimate R_{eff} could also be accelerated

by using distance calculations, similar to these used to build loops, to abort predicted non-binding conformations for systems where the rate of valid conformations is near 100%, which is common for systems with only one tail or linker.

With further work, we aim at establishing MoMA-FReSa as a fast, reliable and robust method for the estimation of effective concentrations, facilitating a deeper understanding of linker behavior in the context of multi-site biomolecular interactions.

3.4 Structural analysis of linkers in multi-domain proteins

3.4.1 General presentation

The CORNFLEX project is a collaborative research initiative, involving the *Laboratoire d'Analyse et d'Architecture des Systèmes* of the *Centre National de la Recherche Scientifique* (LAAS-CNRS), the *Centre de Biologie Structurale* (CBS), the *Institut de Mathématiques de Toulouse* (IMT), *Toulouse Biotechnology Institute* (TBI) and the *Laboratoire de Recherches en Sciences Végétales* (LRSV) (<https://anr.fr/Project-ANR-22-CE45-0003>). Its main goal is the development of computational design methods for Intrinsically Disordered Proteins (IDPs) and Regions (IDRs) with specific properties. The project particularly focuses on the design of flexible linkers in multi-modular enzymes. In the context of this project, MoMA-FReSa was used to generate conformations for multiple linker sequences for chimeric multi-modular proteins involving a CBM domain (PDB ID: 1XDB) and a GH11 domain (PDB ID: 2C1F). One of these constructions was used as Case 1B in our benchmark in Chapter 2, illustrated in Figure 3.15.

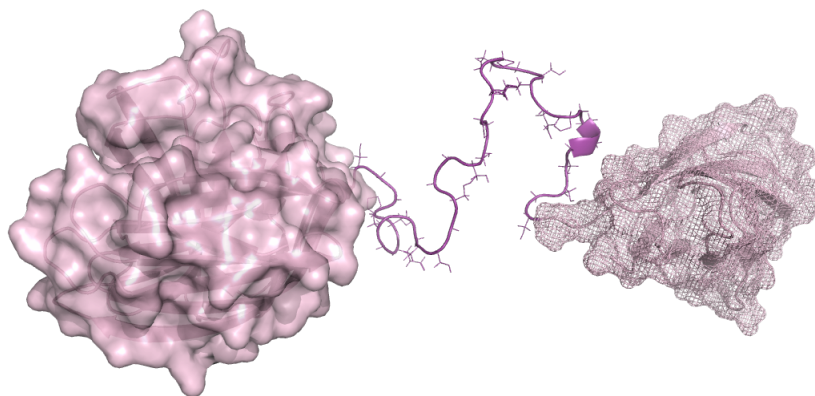


Figure 3.15: Illustration of the a chimeric linker of the CORNFLEX project, used as Case 1B in Chapter 2.

Here, we used two example constructions of the CORNFLEX project to provide a better sketch of some of the post-processing tools of MoMA-FReSa. One of the sequences used for the linker was the *Regular Sequence* of 37 residues of Case 1B:

SGGPSEGS GSSSGSGGSSGPSGPSKSKEPGSGGSSGS

The other one was the *Structured Sequence*, an *a priori* more structured sequence of similar

length (39 residues):

NSGGNGGQQQQQPPSNNNNNNNNNGGQQQQQPPSNNNGG

Note that both sequences were extracted from natural enzymes.

Post-processing tools and energy computation with MoMA-FReSa

As mentioned in Chapter 2, MoMA-FReSa offers valuable post-processing tools that provide crucial insights into the generated ensembles, particularly relevant for the CORNFLEX project. Here, we specifically focus on the two following tools: i) analysis of the relative pose of domains connected by linkers, and ii) energy computation.

At its core, MoMA-FReSa represents the pose of each residue using a 4×4 homogeneous transformation matrix. This matrix captures both the position and orientation of the residue in 3D space. This representation allows MoMA-FReSa to report the relative pose of the two domains at the two ends of a flexible linker. In this case, the moving and fixed domains are identified by the position of their closest residue to the linker (first or last residue of the domain depending on the direction of the linker). When MoMA-FReSa reaches a valid conformation, it can provide a report containing this relative pose information for each linker in a multi-domain protein architecture.

In addition to linker analysis, this section explores the energy computation provided by MoMA-FReSa. The function currently used is a customized version of a hydropathy scale (HPS) [57] [171]. MoMA-FReSa takes advantage of the segmentation in regions to optimize HPS calculations, facilitating faster overall computations. This optimization notably involves omitting calculations between residues within the same rigid region, as their structure (and energy) remains constant.

3.4.2 Post-processing analysis

Comparison of the spatial distribution between the different linkers

Our initial analysis focused on comparing the spatial distribution of two linker sequences: *Regular Sequence* and *Structured Sequence*. We generated ensembles of 10000 conformations for each sequence using MoMA-FReSa with default parameters. We notably used Single-Residue-based Sampling (SRS) strategy, defined in Chapter 2, on all the flexible regions.

To visualize the distribution of the relative position of the two ends of a linker, we employed a Python code to generate voxel maps with a resolution of 20\AA . In these maps, the origin represents the position of the fixed rigid domain. Each colored voxel indicates that the mobile rigid domain position (position of the closest residue of to the linker) of at least one conformation of the ensembles was found within that voxel. The results are presented in Figure 3.16. The color coding differentiates voxels corresponding to the *Regular Sequence* (red), *Structured Sequence* (blue), or both (black).

The initial comparison using SRS revealed slight differences between the two sequences of similar lengths. These results suggested that the *Structured Sequence* sampled a larger volume, maybe due to the two additional amino-acids or due to the sequence that allows more extended conformations.

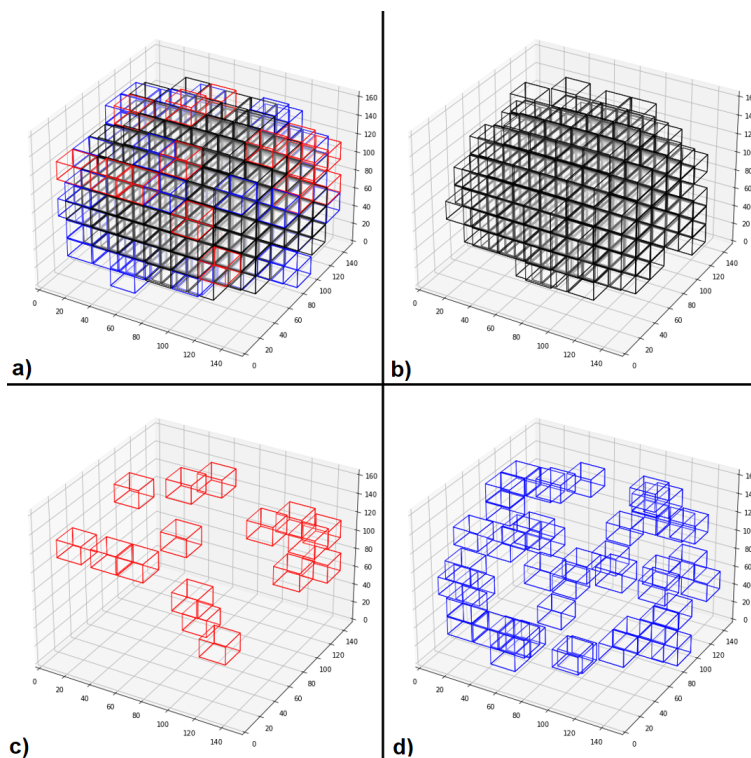


Figure 3.16: Voxel maps used to compare the spatial distribution of the linkers with *Regular Sequence* and *Structured Sequence*. Ensembles generated in SRS. a) All the voxels - b) Common voxels - c) *Regular Sequence* exclusive voxels - d) *Structured Sequence* exclusive voxels.

In effect, the LS2P profiles [66] of these two sequences (Figure 3.17 for the *Regular Sequence* and Figure 3.18 for the *Structured Sequence*) confirmed strong structural dissimilarities. Indeed, as expected the *Regular Sequence* was quite flexible and the *Structured Sequence* presented several regions predicted to be partially structured. In particular, this last sequence presented extended structural propensities in some regions. Note that, conversely, the *Regular Sequence* was richer in Glycines that tend to adopt more compact structures.

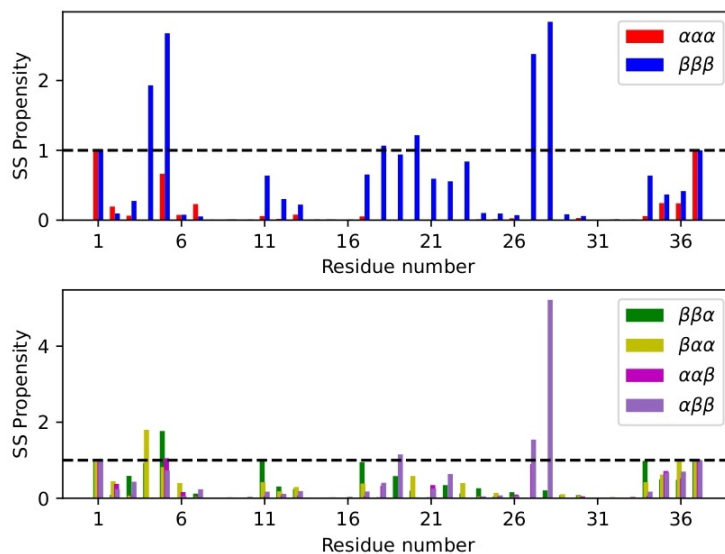


Figure 3.17: LS2P profile of the *Regular Sequence*.

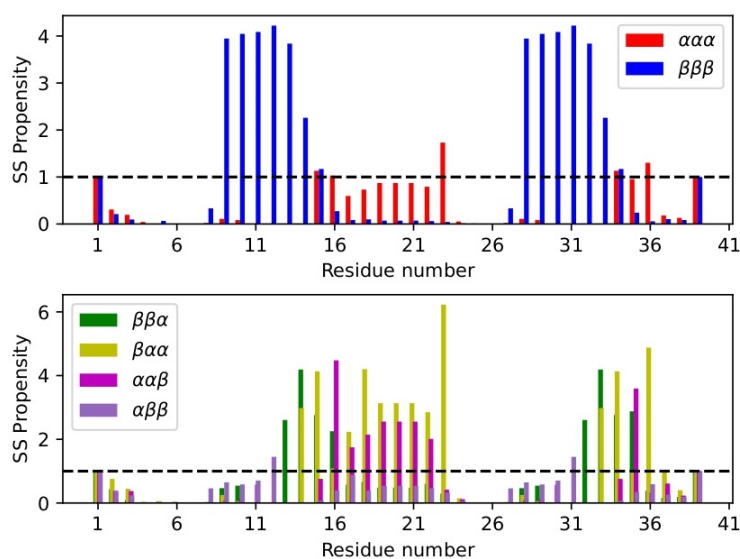


Figure 3.18: LS2P profile of the *Structured Sequence*.

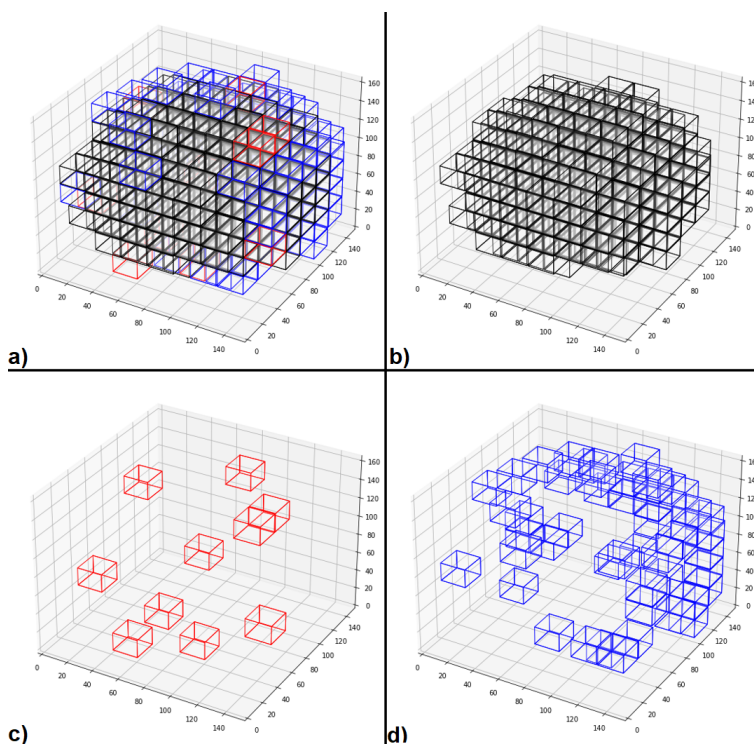


Figure 3.19: Voxel maps used to compare the spatial distribution of the linkers with *Regular Sequence* and *Structured Sequence*. Ensembles generated in TRS. a) All the voxels - b) Common voxels - c) *Regular Sequence* exclusive voxels - d) *Structured Sequence* exclusive voxels.

MoMA-FReSa offers a specialized Three-Residue-based Sampling (TRS) strategy, detailed in Chapter 2, that produces more locally-structured ensembles. Given the different secondary structure propensities between the *Regular Sequence* and the *Structured Sequence*, TRS seemed a more suitable approach for this comparison. We then generated new ensembles of 10000 conformations for both sequences using TRS and compared the spatial distributions through the voxel maps presented in Figure 3.19.

As expected, the results obtained with TRS exhibited a clearer distinction compared to SRS. The less structured *Regular Sequence* (red voxels) seemed to more often explore the central region close to the fixed domain, while the *Structured Sequence* (blue voxels) reached more distant areas due to its inherent structure. This comparison using TRS effectively revealed the distinct spatial distribution of these two linker sequences, exhibiting structural preferences.

Energy based computations

Building upon the precedent spatial distribution analysis, we used the energy reports generated by MoMA-FReSa to create energy maps for both the *Regular Sequence* and *Structured Sequence* TRS ensembles. In these maps, in Figure 3.20, each point represents the position of a mobile domain relative to the fixed one, and the color encodes the energy of the corresponding global conformation. For this analysis, we focused on two extreme cases: the most stable conformations with the lowest energy (below -2900 J) and the least stable ones with the highest energy (above -400 J).

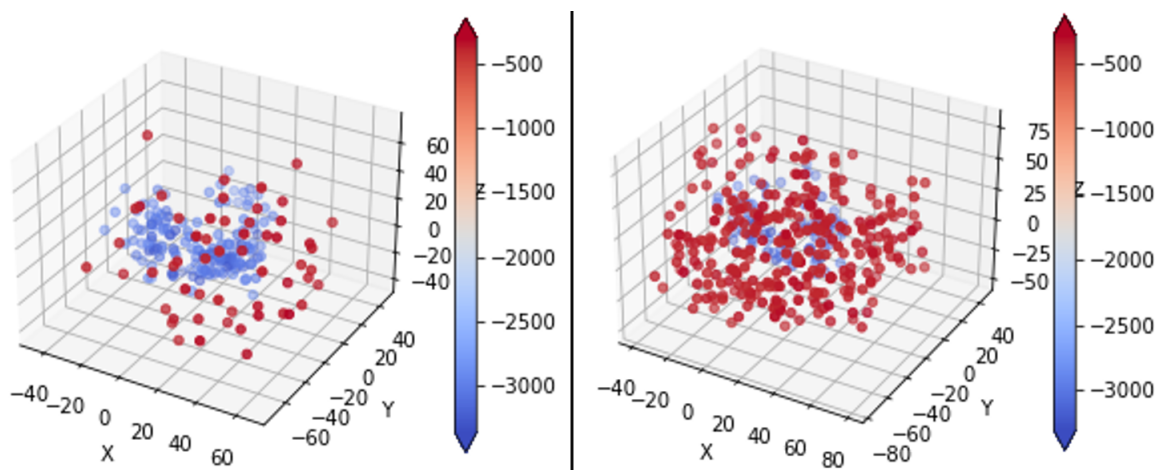


Figure 3.20: Energy maps of the *Regular Sequence* (left) and *Structured Sequence* (right) of the more extreme conformations (below -2900 J or above -400 J). The axes are in \AA and the scale in Joules.

As expected, using the very simple HPS potential, the most energetically favorable conformations corresponded to short distances between the two domains. Due to its flexibility, we saw that the *Regular Sequence* more frequently explored conformations for which the domains were closer to each other. As a result, its ensemble contained a significantly high number of low-energy conformations (156) and only 56 high-energy ones. In contrast, due to its structural preferences, the ensemble of the *Structured Sequence* tended to adopt conformations that sampled a larger volume. This tendency led to a larger proportion of high-energy conformations (281) within its ensemble and only 87 low-energy ones.

3.4.3 Discussion and perspectives

In the context of the collaboration within the CORNFLEX consortium, this work explored the capabilities of the post-processing tools of MoMA-FReSa for analysing linker conformations in multi-domain proteins. We employed two key functionalities: spatial distribution analysis and energy landscape calculations.

To elucidate the structural differences between a highly flexible linker sequence and a more structured one, we employed the Three Residue Sampling (TRS) method of MoMA-FReSa. The spatial distribution analysis demonstrated the effectiveness of TRS in revealing these structural variations for linkers of similar length.

The study of energy landscapes complemented the spatial distribution results, highlighting the connection between flexibility, spatial exploration, and energy distribution within the ensembles generated by TRS for the two linker sequences. Future work could explore the incorporation of more accurate energy functions, such as CALVADOS [199], within MoMA-FReSa to address more diverse research needs.

These preliminary analyses pave the way to more complex investigations. Detailed spatial distribution of relative poses and energy landscape information are relevant to understand the behavior of linkers within multi-modular proteins. In the context of CORNFLEX, the post-processing capabilities of MoMA-FReSa hold significant potential for linker design aiming to control the activity of enzymes.

3.5 Sampling of structural motifs

3.5.1 General presentation

Climate change presents a multifaceted challenge for plants. Rising temperatures, lead to decreased oxygen availability, triggering the production of ethanol within plant cells [59]. Recent studies, suggest that this endogenous ethanol might act as a signal molecule, potentially interacting with ethylene receptors [40]. This research suggests that even low concentrations of ethanol can influence plant growth through its potential interaction with ethylene receptors, adding another layer of complexity to plant signaling pathways.

Ethylene, a gaseous plant hormone, plays a vital role in regulating numerous developmental processes. Its perception relies on a family of transmembrane ethylene receptors (ETR) with a crucial, yet poorly understood, component: the transmembrane domain. The transmembrane domain is a region of the protein embedded within the cell membrane and it is believed to contain 3 or 4 α -helices responsible for interaction with other molecules [41]. However, the exact number and arrangement of these helices remain a subject of debate.

Understanding how ethanol interacts with the transmembrane domain is a challenge in plant signaling comprehension, especially considering its potential role as a signaling molecule highlighted by recent research. A group of the *École Nationale Supérieure Agronomique de Toulouse* (INP-ENSAT) students, decided to study this topic in the context of a bioinformatics challenge.

Case definition

The ENSAT team focused the study on the second ethylene receptor (ETR2) of *Vitis vinifera*. Three models of the *Vitis vinifera* ETR2 transmembrane domain structure were constructed using the software AlphaFold2 [100], I-TASSER [225], and Swiss-Model [85]. The comparison of these models allowed them to identify both conserved and variable domains within the structure.

All three models predicted the presence of four transmembrane helices. However, in some models, the fourth helix was identified as partially formed, raising questions about the accuracy of this prediction. Additionally, both I-TASSER and Swiss-Model rely on AlphaFold2 for model generation, potentially introducing a bias in the comparison.

These limitations encouraged the ENSAT team to explore MoMA-FReSa as an alternative method. Indeed, MoMA-FReSa can explore the conformational landscape of a region using different sampling strategies. The analysis of these ensembles can provide interesting complementary information with respect to other predictive methods.

MoMA-FReSa was applied for sampling the 33-residue long tail containing the putative fourth helix illustrated in Figure 3.21. Indeed, in this tail, the prediction confidence of AlphaFold2 falls below 80%. The tail was sampled in SRS, except the [A-2;D-18] region, encompassing the sub-region of interest for potential helix formation. This sub-region was sampled with SRS and TRS modes, but also with the structured-oriented sampling method (detailed in Chapter 2), which samples the residues of the potential helix only in the α -region of the Ramachandran plot (α -oriented sampling mode).

Beyond its initial goal to give clues about the presence of the fourth helix, this study was used complementarily to the one of the Section 3.4 to highlight differences between sampling strategies. Thus, through these investigations, we aimed at providing a better understanding of the effect of SRS and TRS but also structured-oriented sampling.

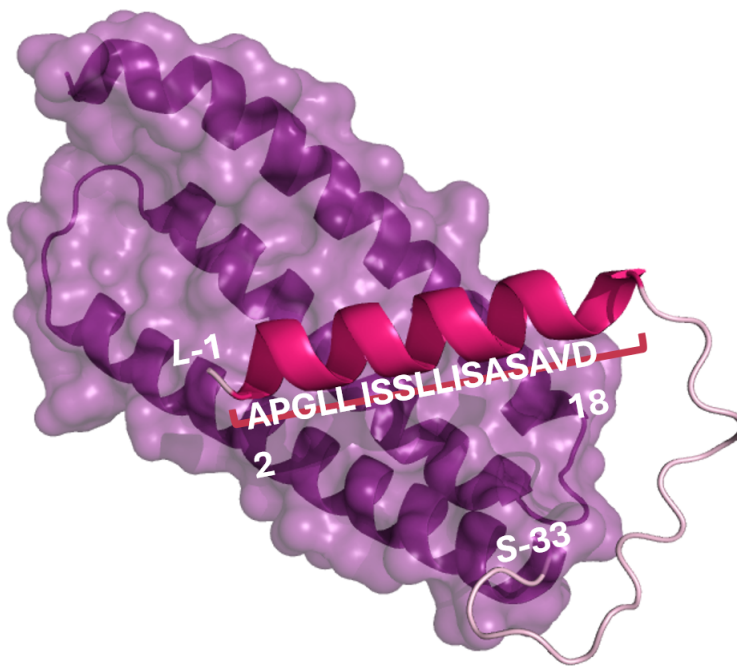


Figure 3.21: AlphaFold2 model of ETR2 of *Vitis vinifera*. The N-terminal region (residues 1-33) was predicted with lower confidence (pLDDT < 80%). It involves a helical sub-region of 17 residues.

3.5.2 Ensemble generation and analysis

Ensembles comparison between SRS and TRS sampling strategies

In a first time, we generated two ensembles of 10,000 conformations using SRS or TRS sampling strategies involving the putative helical sub-region (residues [A-2;D-18]) of the *V. vinifera* ETR2 protein. A few conformations from these two ensembles are presented in Figures 3.22 and 3.23.

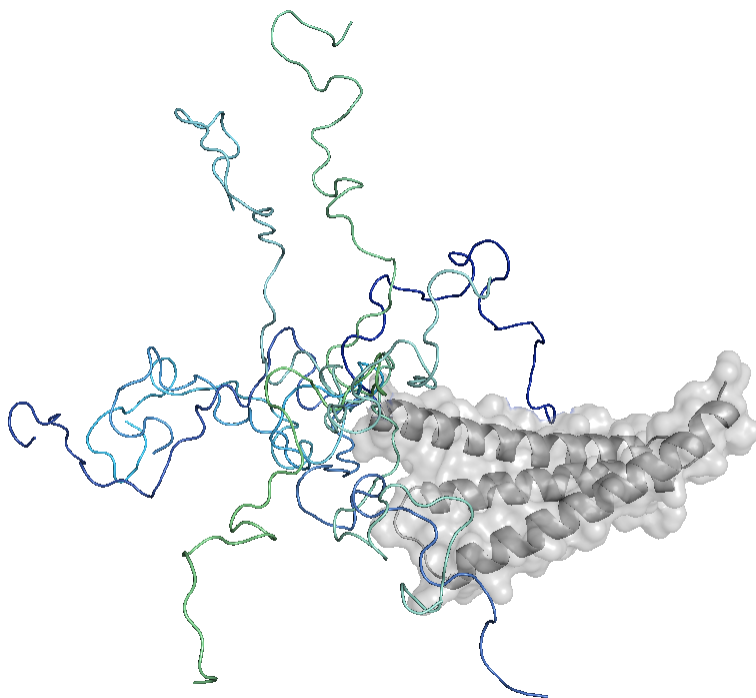


Figure 3.22: Example of conformations inside the ensemble generated with a SRS approach.

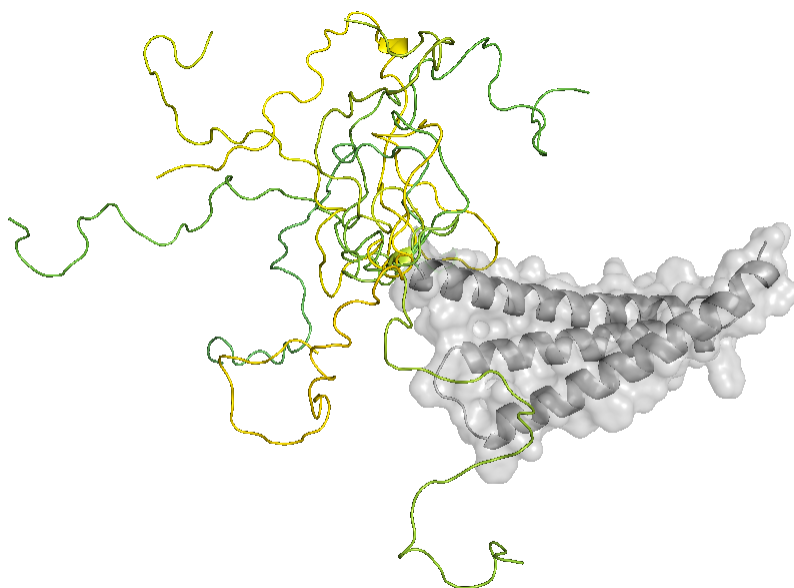


Figure 3.23: Example of conformations inside the ensemble generated with a TRS approach.

To study and compare the conformational ensembles generated by the TRS and SRS methods in MoMA-FReSa, we used a Wasserstein-based statistical tool called WASCO [83] introduced in Chapter 2. WASCO was employed to compare the two ensembles across the entire flexible region. The results are presented in Figure 3.24. The WASCO comparison revealed a local disparity specifically localized on the 17 N-terminal residues [A-2;D-18] sampled by both methods. This difference highlights the impact of the sampling strategy (SRS vs. TRS) on the resulting ensemble.

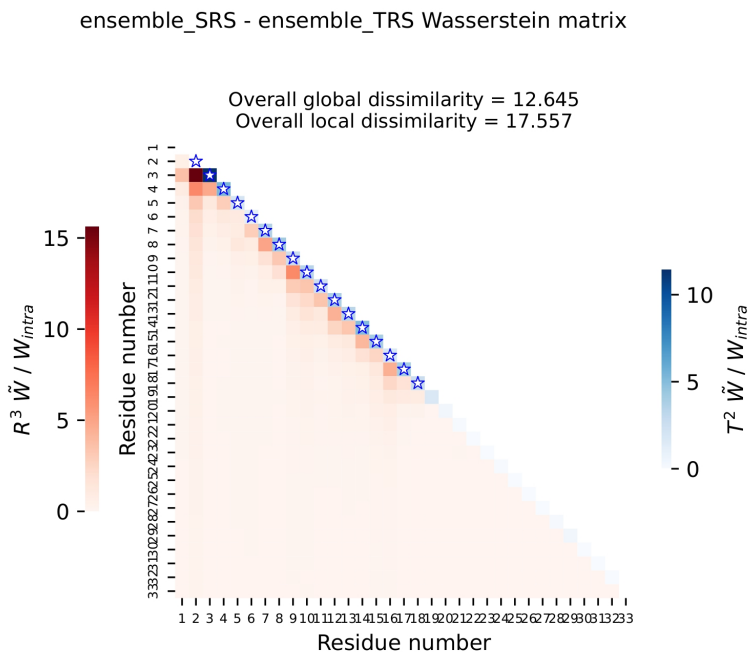


Figure 3.24: Comparison matrix generated by WASCO for ensembles of the flexible regions generated using the SRS and the TRS sampling strategies.

For a more focused comparison, a local analysis, corresponding to the diagonal of the matrix, was performed on the sub-region [A-2;D-18] in Figure 3.25. This analysis revealed a high local dissimilarity of 15.487 on this sub-region, particularly at the N-terminus, where the local maximum is reached for residue P-3 (second residue of the local figure).

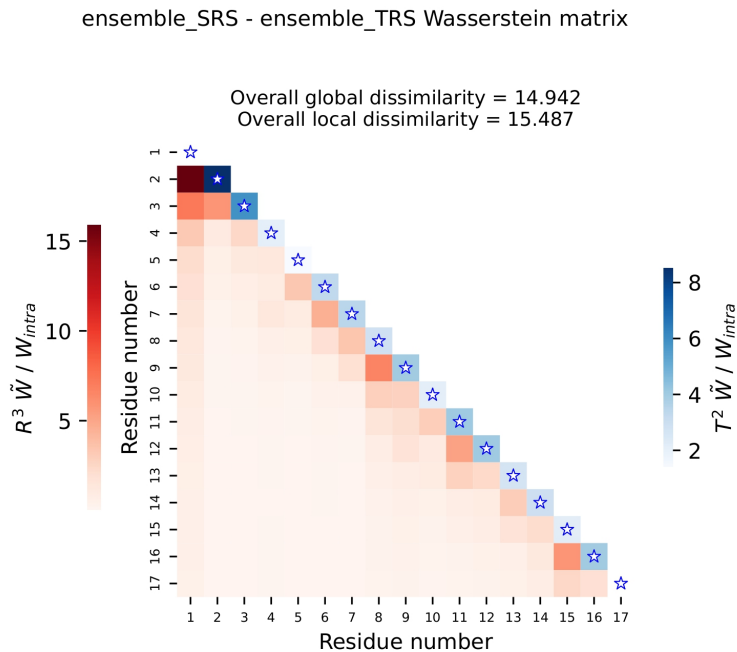


Figure 3.25: Comparison matrix generated by WASCO for ensembles of the flexible [A-2;D-18] sub-region computed using the SRS and the TRS sampling approaches (top part in Figure 3.24).

The global dissimilarity, corresponding to the off-diagonal elements of the matrix, between the ensembles was also high, indicating a substantial change in the overall structure of this sub-region. These differences were primarily observed between neighboring residues or residues very close to each other (“neighbors of neighbors”). Notably, residue A-2 (first in the local figure) had the lowest local dissimilarity but shows the most extensive global dissimilarities. For example, the difference between residues A-2 and P-3 reached 15 (ie. signifying a potentially change in the relative position of these two residues of 15 times larger than the uncertainty/noise in both ensembles).

It is interesting to note that the two residues with the highest local dissimilarity (P-3) and highest global dissimilarity (A-2) are located at the N-terminus of the sub-region. This observation, considering the C-to-N sampling direction of this tail, suggests a potentially stronger influence of the TRS method at the end of the “TRS sub-regions”. This point requires further investigation.

Ensembles comparison between TRS and α -oriented sampling

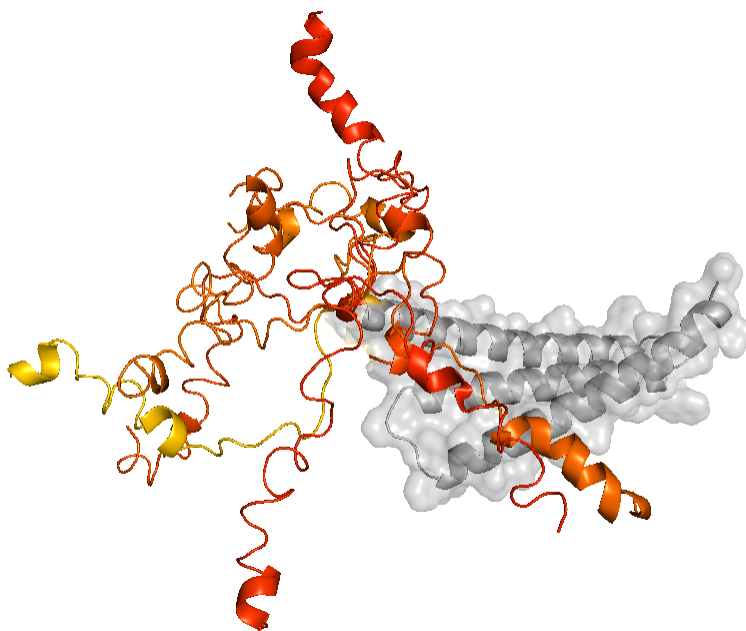


Figure 3.26: Example of conformations inside the ensemble generated with an α -oriented approach.

We generated another ensemble of 10,000 conformations this time by sampling the residues of the sub-region [A-2;D-18] only in the α region of the Ramachandran plot. Some conformations of this ensemble are presented in Figure 3.26. This ensemble was next compared to the TRS one using WASCO, focusing on the initial 17 residues. The matrix is presented in Figure 3.27.

This time, dissimilarities were more pronounced: more than 5 times higher for local dissimilarities and close to 4 times higher for the global ones. In addition of higher average dissimilarity values, we also observed more extended global dissimilarities impacting significantly further than the close neighbours. We can deduce that even if TRS is more adapted to generate partially formed secondary structures, for this case, only the use of TRS is not enough to build helical structures because the sequence of this region is unlikely to form them.

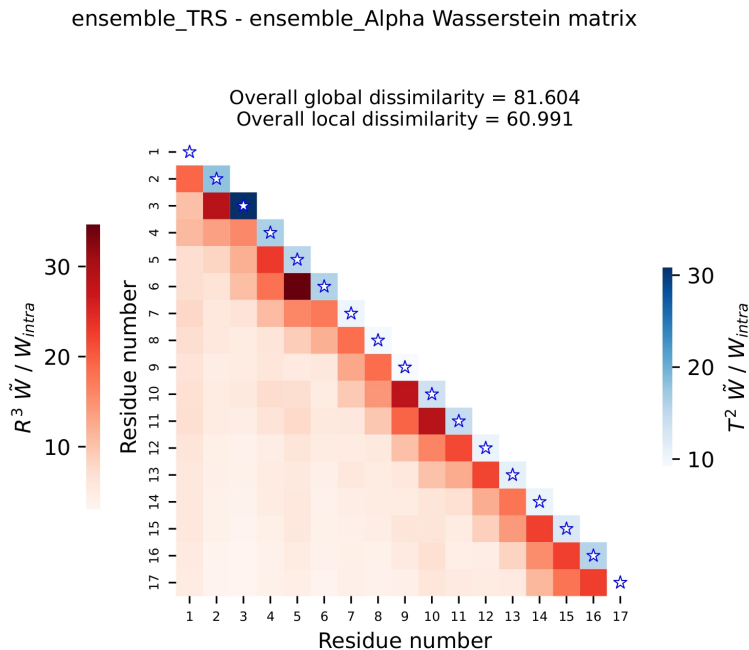


Figure 3.27: Comparison matrix generated by WASCO for the flexible $[AA_2 - AA_{18}]$ sub-region computed using the TRS and the α -oriented sampling approaches.

Deeper analysis of the ensembles for the ENSAT project

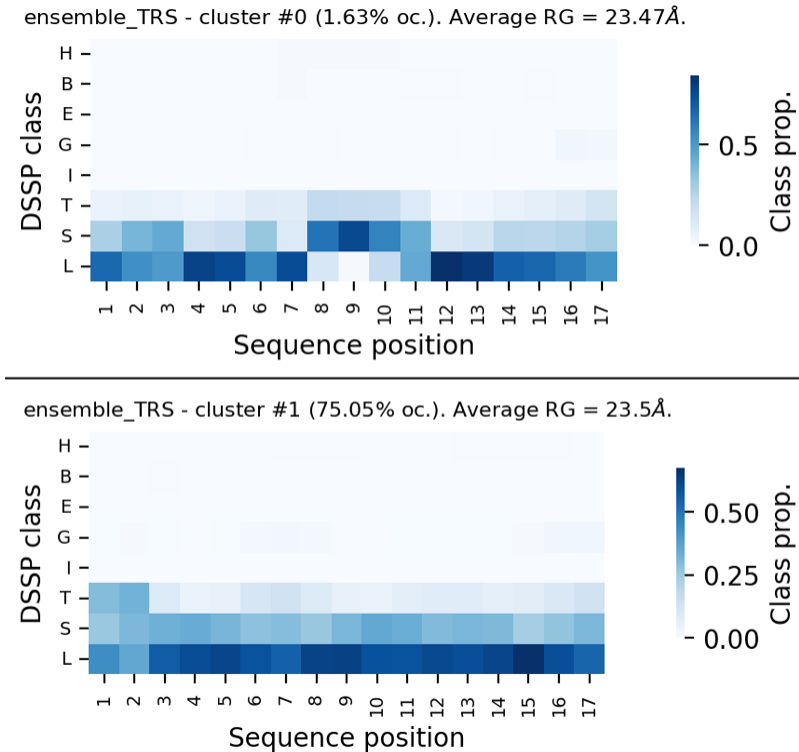


Figure 3.28: DSSP class distribution for the two clusters of the TRS ensemble. Helical propensities are on line H.

With the generated conformational ensembles, we tried to make a hypothesis on the presence of a fourth helix to help the ENSAT team. We used a very hypothetical approach on the basis of the TRS ensemble, which is appropriate to express structural preferences and not as biased as the α -oriented approach. This last ensemble was used as a reference ensemble.

The TRS ensemble did not display a strong presence of an α -helix within the sub-region [A-2;D-18]. This trend was comforted by the comparison between the TRS ensemble and the α -oriented one, where WASCO showed strong differences. To complement this analysis, we used another statistical tool WARIO [82], introduced in Chapter 2, to analyse and clusterize the TRS ensemble on the 17-residues sub region. Using default settings, WARIO identified two clusters in the conformational ensemble, one containing 75.05% of the conformations and the other one 1.63% (the remaining conformations were not clustered). Figure 3.28 shows the average secondary structure content for each cluster computed by DSSP [101]. These plots clearly show that the ensemble is highly disordered, and that the helical propensity is very low. This finding aligned with the hypothesis of a missing fourth helix. The AlphaFold2 model could be based on a form of the complex where the helix forms upon interacting with the other three helices, despite its low inherent propensity to form a helix. Based on our results, the ENSAT team incorporated our newly generated ensembles alongside the AlphaFold2 model for further analysis in their research project.

3.5.3 Discussion and perspectives

Our collaboration with the ENSAT students aimed at validating the presence of a putative fourth trans-membrane helix in the *V. vinefera* ETR2 protein. The TRS method of MoMA-FReSa was employed to analyse the suspected helical sub-region, addressing limitations encountered with AlphaFold2 predictions. We used WASCO and WARIO, two statistical tools developed in our group, to study this ensemble and compare it to the one with the hypothetical helical region sampled in α -region. This exploratory approach tended to validate the absence of the helix and the newly generated ensembles, along with the AlphaFold2 model, were incorporated by the students into their research project for further analysis. This study revealed a potential use of MoMA-FReSa to overcome potential biases in initial predictions of AlphaFold2 and provide alternative ensemble representations, which is an interesting perspective for future developments of our software. This field warrants further work to position MoMA-FReSa as a complementary tool of AlphaFold2.

Beyond this analysis, in collaboration with the ENSAT students, we conducted a more general study on the sampling method. Notably, we used this context to test the structure-oriented sampling method, which is another option of MoMA-FReSa beyond SRS and TRS to offer a more personalized sampling adapted to specific needs. The three highly different ensembles resulting from this study highlight the influence of the sampling method on the resulting protein structures and exemplify the benefit of different sampling methods for investigating alternative structural possibilities, particularly in regions with strong secondary-structure tendencies.

During this study, the fact that the N-terminal residues exhibited the highest dissimilarities suggests a potential influence of the C-to-N sampling direction on the ensembles. The impact of the sampling direction of the resulting ensembles was addressed in the original WASCO article to conclude to a non relevant impact for the discussed sequence [83]. In other cases, for example for Huntingtin [64], the direction of sampling revealed to have an impact. Our results warrants further investigation to gain a deeper understanding of sampling behavior especially with the use of neighbour-dependent approaches such as TRS, or structured-oriented

sampling. Notably, this result could stress the interest of stochastic approaches that sample in different directions when it is possible, as it can be done with MoMA-FReSa.

It is important to acknowledge a limitation within the current analysis. The utilized database consists of tripeptide fragments from proteins in aqueous environments. However, transmembrane helices, the focus of this study, reside in a lipid environment. This database limitation, which can be easily addressed due to the independent nature of the database in the program, could strongly impact the accuracy of the results. Therefore, while the present study may not provide definitive answers to the initial question regarding transmembrane helices, it successfully served its primary purpose: to illustrate the various sampling strategies of MoMA-FReSa and new potential uses of the method.

3.6 Conclusion and perspectives

This chapter explored the potential of MoMA-FReSa to address diverse questions related to disordered protein regions. We successfully demonstrated its capabilities in the area of protein design. In effect, the preliminary work within the iGEM competition exhibited promising perspectives of MoMA-FReSa for the conception of complex multi-modular systems. The CORNFLEX project further explored this potential and demonstrated the usefulness of MoMA-FReSa to study structural effects of disordered linkers in multi-modular enzymes.

Collaborations with ENSAT students and the CORNFLEX project highlighted the strengths of the different sampling strategies (TRS, SRS, and structured-oriented sampling) in addressing various structural analysis challenges. In Chapter 4, we will further explore the application of the structured-oriented sampling strategy within a for reinforcement learning method.

MoMA-FReSa has multiple post-processing tools that can be improved in future work. The most relevant one is probably energy calculation, which can play a crucial role in sorting conformations and guiding sampling within learning strategies. The preliminary work on estimation of effective concentration is also very promising direction to pursue, since current approaches based on simple polymer models are inaccurate [210].

Overall, while this chapter presented exploratory/preliminary work, the promising results from our collaborations suggest options to pursue for future research.

A Reinforcement Learning method for MoMA-FReSa

Contents

4.1	Introduction	89
4.2	Context and implementation choices	90
4.2.1	A learning approach suited to our sampling method	90
4.2.2	Learning methods employed	91
4.3	Implementation in MoMA-FReSa	93
4.3.1	Use of the learning method during the sampling procedure	93
4.3.2	Buffering and backtracking integration to the learning procedure	95
4.3.3	Global architecture	96
4.4	Ongoing work and perspectives	98
4.4.1	Ongoing work	98
4.4.2	Perspectives and areas of exploration	99
4.5	Conclusion	100

4.1 Introduction

In recent years, our team has focused on improving and generalizing algorithms for modeling Intrinsically Disordered Proteins (IDPs) and Regions (IDRs). An inherent challenge in this field is loop sampling, as we observed with MoMA-FReSa in Chapter 2. A previous study by A. Barozet [9] explored a simple model-based Reinforcement Learning (RL) technique to improve the performance of a sampling algorithm to generate large conformational ensembles of loops. Barozet’s method demonstrated the potential of RL in this domain, but its practical application was limited to simple cases due to an exponential memory requirement with the complexity of the problem. We aim to develop an alternative, model-free RL approach for MoMA-FReSa.

This project was conducted in collaboration with engineers from the *Programme National de Recherche en IA* (PNRIA), an engineer network supporting the implementation of artificial intelligence (AI) algorithms funded by the *Centre National de la Recherche Scientifique* (CNRS), the *Institut national de recherche en sciences et technologies du numérique* (Inria), and *Commissariat à l’énergie atomique et aux énergies alternatives* (CEA).

During this short project, we successfully established the foundations for an original learning strategy for MoMA-FReSa. This chapter details this preliminary work and explores its future developments. Section 4.2 presents the context, the starting point of the project and the chosen learning methods. Section 4.3 describes our method and its concrete implementation in

more detail. While this initial work has not lead to specific results, Section 4.4 discusses ongoing efforts and potential improvements for the coming months. Finally, Section 4.5 concludes this chapter.

4.2 Context and implementation choices

Our initial goal was to introduce a learning approach within MoMA-FReSa specifically for the improvement of loop sampling. However, we strategically shifted our focus towards a more general approach that enhanced the overall sampling capabilities of MoMA-FReSa when flexible regions are constrained due to the presence of rigid regions. This method needs to be versatile and applicable to all system types handled by MoMA-FReSa. In this section, we present the foundation upon which we built our method and the specific choices we made regarding its implementation.

4.2.1 A learning approach suited to our sampling method

Using the decomposition in regions

MoMA-FReSa, as detailed in Chapter 2, uses a hierarchical architecture with regional decomposition. To capitalize on this specific feature, we designed our learning approach around these regions. Each flexible region acts as an agent equipped with its own neural network, responsible for improving its own sampling process.

While the properties of different flexible region types (loops, linkers, tails, or IDPs) may subtly influence the learning method (e.g., buffer filling strategies), the core principles of the approach remain consistent across all regions.

Acting at the residue level

Our sampling approach relies on the three backbone dihedral angles (ϕ , ψ , ω) of the flexible residues. While all three angles are technically used in our sampling method, ϕ and ψ hold the majority of the conformational information. Therefore, our learning method is focused on these two angles. Furthermore, we hypothesize that the identity of the neighboring residues plays a crucial role in optimizing the sampling performance [67]. Thus, when machine learning is activated, the sampling process always operates in a Three Residue Sampling (TRS) mode.

A structural region approach

The challenge here lies in identifying suitable pairs of ϕ and ψ angles for each residue within its specific environment. The Ramachandran plot, visualized in Figure 4.1, depicts the distribution of these dihedral angles. This distribution is not uniform, and can be categorized into three main regions (α , β , γ), each one corresponding to distinct protein secondary structures.

Our learning approach focuses on refining the probability distribution of these structural regions for each flexible residue in its actual context, considering its amino acid type, its neighboring residues (amino acid type but also structural region if the residue is already sampled), and its spatial position. We use the structure-oriented sampling method of MoMA-FReSa to enforce the region in the Ramachandran plot during the sampling.

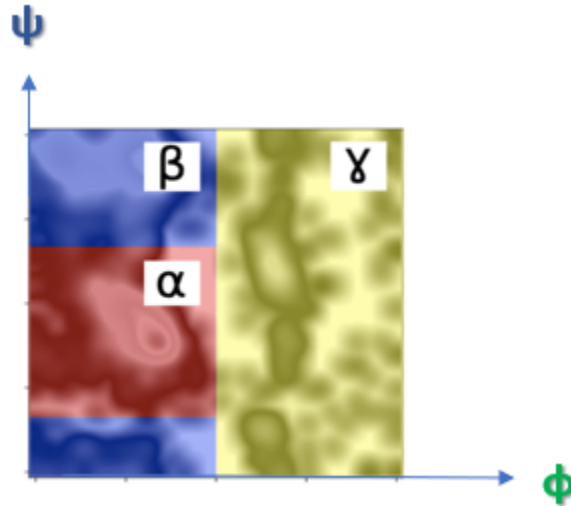


Figure 4.1: Ramachandran plot where the main secondary structures regions are plotted as a function of ϕ and ψ .

4.2.2 Learning methods employed

Reinforcement Learning (RL)

We chose to base our learning approach on RL [163]. RL is a machine learning technique that operates through trial and error. When faced with a decision, an agent selects an action based on its previous experiences and learning. The result of this action is then evaluated and a reward is issued. A value function uses this reward, either positive or negative, to update the “intelligence” of the agent and guide its future choices. The objective of the agent is to maximize the accumulated reward as iterations progress. The principle of RL is depicted in Figure 4.2. An advantage of RL in this context is its ability to learn autonomously, without the need for explicit training data. This stands in stark contrast to traditional approaches that rely on large datasets for training before being applied. This characteristic makes RL particularly well-suited for its integration in sampling methods.

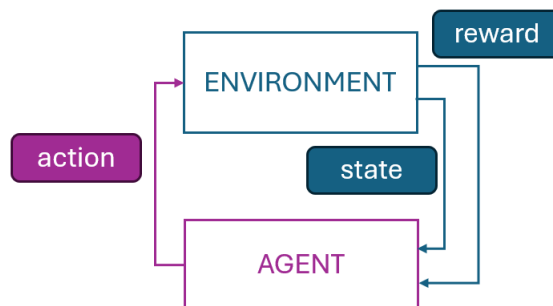


Figure 4.2: Schema of the Reinforcement Learning (RL) approach.

Due to the difficulty in our problem to figure out the entire environment (in construction during the sampling process), we opted for a model-free approach [163, 183]. Model-free RL methods learn directly by interacting with the environment, without explicitly modelling it. These methods rely on trial and error to learn the best actions to take in different situations.

They are more adapted for handling complex and unknown environments but require a longer exploration.

Actor-Critic strategy

The trade-offs between different RL algorithms should be carefully evaluated. In traditional Monte Carlo Reinforcement Learning (MC-RL), the reward signal is calculated only at the end of an episode. Imagine a game of tic-tac-toe as an episode, where the terminal state is reached upon a win, loss, or draw. In this scenario, the average reward considers all actions taken throughout the episode. If most actions were successful with only a few failures, the overall reward might still be positive [217]. Translating this concept to our MoMA-FReSa application, an episode could be considered the final conformation generated by the sampling method, whether it is a valid conformation or a failure due to excessive backtracks. However, this approach discards valuable information about the numerous intermediate steps before backtracking, in case of overall success or failure. Due to this limitation, MC-RL approaches did not seem well-suited to our strategy. Consequently, we opted to pursue alternative methods.

To develop an approach more adapted to the philosophy of MoMA-FReSa, and provide more fine grained feedback, we proposed incorporating an Actor-Critic strategy [112, 163], a Temporal-Difference (TD) Learning method [194]. TD Learning strategies evolves every step, as opposed to episodic MC-RL methods. The Actor-Critic strategy utilizes two neural networks working in tandem:

- **Actor:** Responsible for selecting the next action based on the current state.
- **Critic:** Approximates the value function and uses it to guide the decision-making process of the Actor.

Initially, the Actor might randomly select actions due to a lack of knowledge. The Critic observes these actions and provides feedback based on their outcome. Through continuous learning from this feedback, the Actor refines its policy to improve its performance in the sampling process. Simultaneously, the Critic also updates its own evaluation methods to provide more accurate feedback in the future. The core principle of Actor-Critic is illustrated in Figure 4.3.

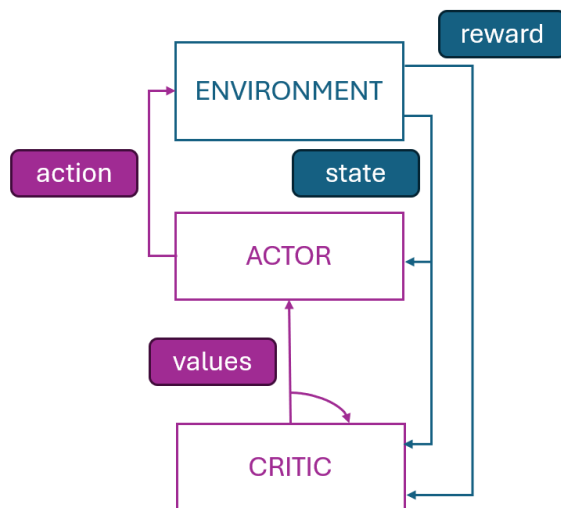


Figure 4.3: Schema of the Actor-Critic model.

Parallel implementation

The multithreaded implementation of MoMA-FReSa presents an opportunity for optimization during the learning process. To capitalize on this, we propose a system with a global network and independent workers (one per thread). Each worker is equipped with its own local copy of the network.

For Actor-Critic agents, two primary major implementation strategies exist [22]:

- **Advantage Actor-Critic (A2C):** All actors complete their work and update the global network in a synchronized fashion.
- **Asynchronous Advantage Actor-Critic (A3C):** Multiple actors operate in parallel, independently collecting experiences and updating their local copies of the network. These local updates are periodically applied to the global network.

Asynchronous methods, such as A3C, can introduce inconsistencies if workers learn using outdated versions of the network. To address this challenge, we opted for the A2C approach, prioritizing stability. The A2C coordinator employs a synchronized update mechanism and waits for all parallel actors to complete their work before updating the global parameters. This involves calculating and averaging the gradients from each worker to update the global network. Consequently, each iteration begins with all actors operating on the same network policy. This synchronized approach not only enhances stability but also has the potential to accelerate convergence as all actors use the most recent weights of the neural network.

Curiosity Driven

A last aspect in our implementation is the incorporation of a curiosity-driven module. Curiosity-driven learning presents an exploration method where the learning agent develops its internal reward function, essentially acting as a self-learner [36, 153, 74]. This eliminates the need for meticulously crafted external rewards, which can be challenging in complex scenarios like MoMA-FReSa. Additionally, curiosity-driven learning can mitigate overfitting by encouraging the agent to explore a broader range of valid conformations, rather than solely focusing on a single "optimal" solution. This aligns perfectly with the objective of MoMA-FReSa, which is not to identify optimal conformations, but rather to explore the vast ensemble of possible conformations a molecule might adopt. To achieve this goal, we propose incorporating Random Network Distillation, a technique which encourages exploration during learning, potentially leading the agent to discover a more diverse set of conformations [37].

4.3 Implementation in MoMA-FReSa

This section details the implementation of the RL approach within MoMA-FReSa. This implementation is unique as it adapts to the specificities of our method, making it inherently exploratory. While some elements will likely undergo changes based on upcoming testing and refinement of our implementation, the overall framework is relatively well defined.

4.3.1 Use of the learning method during the sampling procedure

Agent-Region correspondence

Each flexible region within the protein is represented by a single agent, corresponding to a single global neural network. This network can be mirrored by multiple workers operating in

a synchronous fashion using the A2C approach (described in Subsection 4.2.2).

Action space

For each step involving a flexible region, the agent selects a residue, AA_k , conforming to its sampling procedure. A corresponding structural region, sr_k , is defined among the three possible structural region α , β and γ . Once the structural region is chosen by the network, a tripeptide is randomly selected from the database that matches the chosen structural region for the central residue in TRS mode.

State representation

To decide the structure of the residue AA_k (i.e. the action), the agent considers its state, which includes the following information:

- The amino-acid type of AA_{k-1} encoded in binary encoding.
- The amino-acid type of AA_k encoded in binary encoding.
- The amino-acid type of AA_{k+1} encoded in binary encoding.
- The structural region of the precedent residue (sr_{k-1} or sr_{k+1} depending on the building direction).
- Position in the space (x, y, z) of the precedent residue (AA_{k-1} or AA_{k+1} depending on the building direction).

This comprehensive state representation of size 7 aims at capturing the local environment surrounding the residue of interest in term of both structure and position.

Reward function

Following an action (selecting a structural region sr_k), n conformations inside the designated structural region from the database are tested for the chosen residue (AA_k). The external reward (simply called reward in this chapter) is determined based on the following criteria:

- **Existence Test:** If a valid TRS entry does not exist for this tripeptide with a given structural region for the precedent sampled residue (sr_{k-1} or sr_{k+1} depending on the building direction) and sr_k for the central residue, the process leads to a fatal error. A backtracking procedure is then triggered without any other tests.
- **Validation Tests:** All standard validation tests used in MoMA-FReSa (described in Chapter 2) are performed, except for the TRS test, which is replaced by the next evaluation.
- **TRS Plausibility:** A new test is conducted to evaluate TRS plausibility. Depending on the building direction, the ratio of conformations that satisfy the TRS criteria is calculated:
 - In N-to-C: by the number of entries in the database of the tripeptide (AA_{k-2}, AA_{k-1}, AA_k) with the associated structural regions (sr_{k-2}, sr_{k-1}, sr_k) divided by the number of entries in the database of the same tripeptide with this time its associated structural regions ($sr_{k-2}, sr_{k-1}, \text{any}$).

- In C-to-N: by the number of entries in the database of the tripeptide (AA_k, AA_{k+1}, AA_{k+2}) with the associated structural regions (sr_k, sr_{k+1}, sr_{k+2}) divided by the number of entries in the database of the same tripeptide with this time its associated structural regions (any, sr_{k+1}, sr_{k+2}).

Conformations with higher representation in the database translate to higher TRS plausibility scores.

In absence of fatal error, the final reward is a combination of the rate of the number of conformations passing the validation tests (n_s) over the n tries and the TRS plausibility score for the given sr_k :

$$\text{Reward} = \lambda_{POS} \times \left(1 + \text{TRSPlausibility} + \frac{n_s}{n}\right).$$

As we want to favor TRS Plausibility and achieving a high rate of successful conformations, the reward function is a positive constant λ_{POS} multiplied by a term superior to 1 increasing with these two factors.

A fatal error leads to a negative reward value λ_{NEG} . Both λ values (λ_{POS} and λ_{NEG}) are constants defined at the initialisation.

4.3.2 Buffering and backtracking integration to the learning procedure

Buffering during sampling

The sampling process of MoMA-FReSa incorporates a buffering system to address the temporal nature of evaluations. Choosing a structural region sr_k for a residue AA_k might initially appear promising (high initial reward), but can systematically lead to collisions later in the chain, triggering backtracking. To account for this, MoMA-FReSa employs two buffer categories, representing the memory of the agent:

- **Draft Buffer:** This temporary buffer stores information for the most recent choices with an initial reward, as long as the reward for an action is not finalized. The initial reward here refers to the reward calculated before considering the downstream effects on the chain in Subsection 4.3.1. The size of the buffer is the number of residues in the region, each flexible residue has an assigned location in the buffer.
- **Real Buffer:** This buffer contains only validated (final) information. The network is trained solely on data from the real buffer. This buffer acts more as a classic buffer, and its size is a fixed parameter.

If the entire protein conformation is successfully validated, all the entries from the draft buffer of each agent are transferred to their respective real buffer for training. This process ensures that good rewards lead ultimately to a valid final conformation.

Integration of backtracking

Backtracking occurs when a chosen action leads to a building problem later in the chain. In this case, all entries in the draft buffer corresponding to the backtracked residues are penalized: the reward is greatly lowered. We consider the step of backtracking b_{step} : the first residue to backtrack in the region has $b_{step} = 1$, the next one $b_{step} = 2$, etc... The responsibility of the

residue decreases during the backtrack. So the penalty is:

$$\text{Penalty} = \lambda_{PEN} \times \left(1 + \frac{1}{\log(1 + b_{step})} \right),$$

a negative value produced from a negative constant λ_{PEN} (defined at the initialisation) and a value larger than 1 decreasing with b_{step} increasing. This negative penalty is added to the reward to obtain a new lower reward: $\text{Reward} = \text{Reward} + \text{Penalty}$.

These penalized entries are transferred to the real buffer and will be overwritten in the draft buffer with new information during subsequent sampling attempts for the backtracked residues. This buffering system with backtracking penalty ensures that the network learns from validated actions and penalizes choices that ultimately lead to dead ends, guiding the learning process towards more efficient generation of conformations.

4.3.3 Global architecture

The overall architecture of each agent involves two distinct networks (one for the critic and one for the actor) that share a common feature extraction layer based on an encoder, which takes as entry a state of size 7 previously defined. Note that in our implementation, the two networks can be called independently. The actor network returns a distribution of probabilities for each structural region using a softmax function from the output layer. Figure 4.4 illustrates this architecture. Note that the internal structure of the network (number and type of layers) can vary and we are currently testing different compositions.

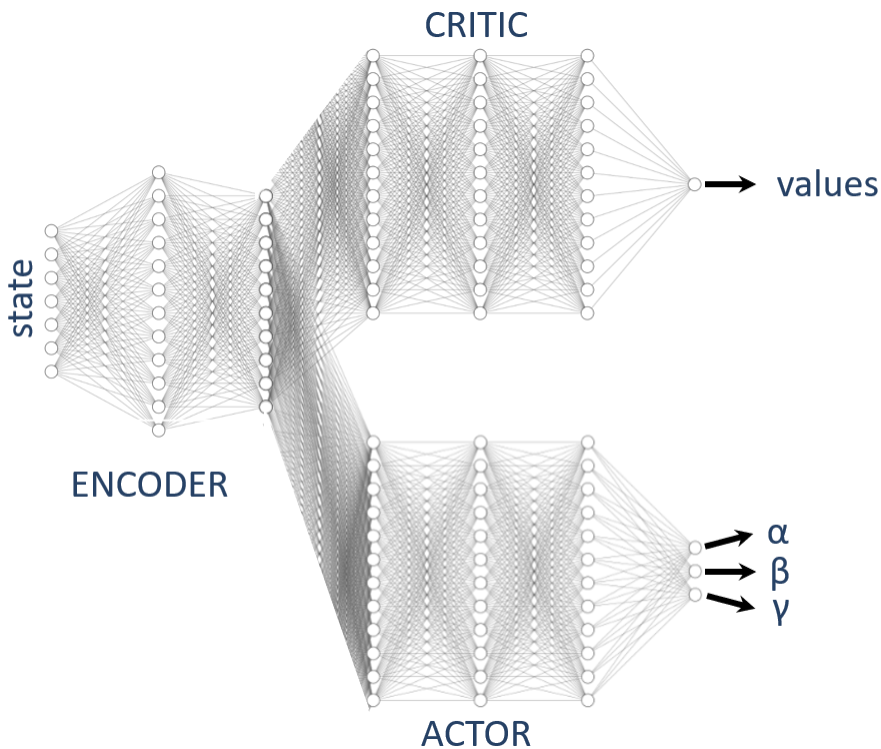


Figure 4.4: Representation of the network architecture. The state is given to an encoder, which provides entries to the actor and critic networks. The probabilities for each step are returned by the actor using a softmax function from the output layer and values of critic are issued from the critic network.

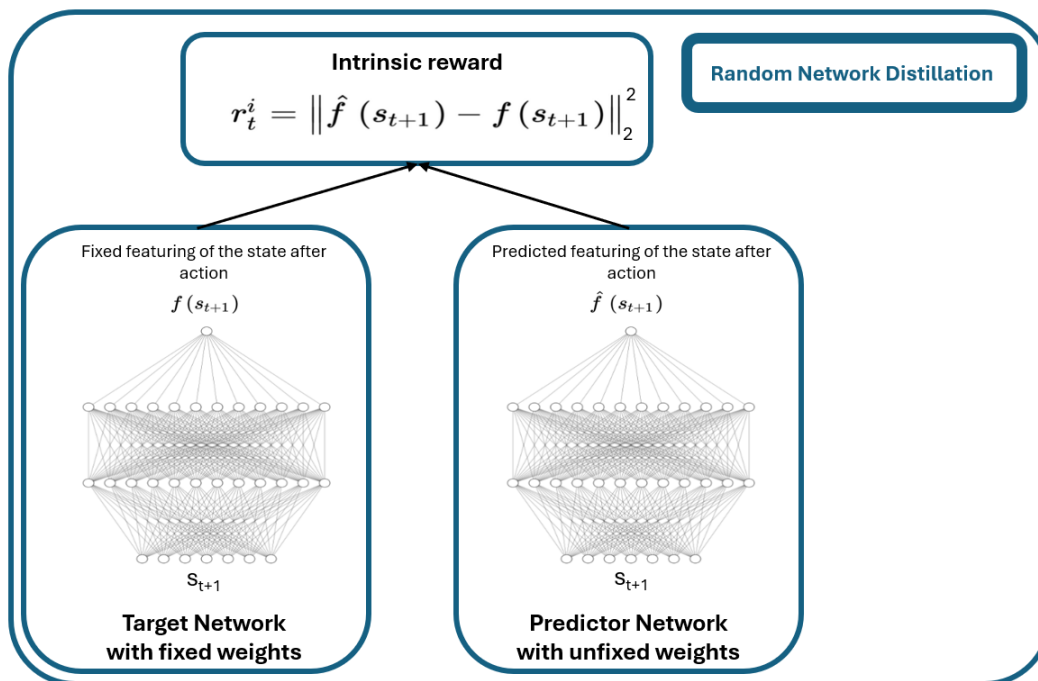


Figure 4.5: Illustration of the computation of the intrinsic reward by a Random Network Distillation. s_{t+1} represents the next state.

The RL process progresses following these steps:

1. **Predict from the State:** At each sampling step, the current state is provided to the architecture of Figure 4.4, encoded and given to the two networks actor and critic. The actor network provides the three probabilities for the structural regions as output. The critic network returns values from the evaluation of this action. These values are used in subsequent operations to update the weights of the network (see below).
2. **Take an Action:** An action is taken by a stochastic process following the three probabilities.
3. **Apply the Action:** After n tries with this policy, the action is evaluated with a first reward as explained in Subsection 4.3.1. We also obtain the next state.
4. **Computation of the Intrinsic Reward:** On the other hand, the internal reward (also called intrinsic reward) of the curiosity driven approach is calculated from the next state by a Random Network Distillation, as illustrated in Figure 4.5.
5. **Storage in the Draft Buffer:** The overall data (current and next states, action taken, outputs of the actor-critic network, intrinsic and initial rewards) of this process is stored in the draft buffer, at a reserved emplacement for this given residue, looking forward to a reevaluation or a validation of the initial reward.
6. **Transfer to the Real Buffer:** In case of backtracking the reward is penalised. The entries of the draft buffer associated to a backtracked residue are given to the real buffer. In case of success all the entries of the draft buffer are transferred to the real buffer without penalisation of the reward.

7. **Estimate Advantage:** Information inside the real buffer serve for a global training involving all the workers of the agent when the buffer is full. We estimate two advantages (one intrinsic and one extrinsic) thanks to the values of the critic and the respective rewards.
8. **Update the Networks:** These advantages are used to compute loss functions which guide the gradient descent of the backpropagation process to update the weights of our architecture. During this process, the feedback of the critic is used to improve the network of the actor and its own network. Random Network Distillation is also updated thanks to a loss function computed from the mean of the intrinsic rewards.

4.4 Ongoing work and perspectives

Our initial efforts resulted in a fully functional version of the RL-enhanced MoMA-FReSa. However, due to time constraints, we were unable to fine tune the implementation and to develop a testing procedure for this method. Consequently, we do not have results to present within this manuscript.

However, we are actively working on improvements to the foundational aspects of this new method to complete the base established in the previous section. This section focuses on the ongoing development and future potential of the learning approach of MoMA-FReSa.

4.4.1 Ongoing work

Parametrization and discussion about our choices

It is important to acknowledge that machine learning models often require a time-consuming parameterization phase to achieve optimal performance. This parameterization process is currently ongoing for the MoMA-FReSa learning method. This parametrization should provide an optimization of the remaining hyperparameters (λ values in the reward, size of the real buffer, composition of the networks) and evaluate the contribution and performances of the learning add-on at its best configuration. The results of this process might lead to reevaluations of certain choices, such as employing the A3C implementation instead of A2C. While A2C is our initial choice, we remain open to exploring the A3C method in the future, particularly if time becomes a critical factor.

A reevaluation could also lead to discuss alternative implementation choices. The current configuration with one agent per region might be restrictive for systems with multiple flexible regions that exhibit interdependencies. Multi-loop systems could serve as an ideal test to evaluate the need to redefine this aspect. Future exploration could involve models with inter-agent communication or a more complex model with one agent per system. We could also discuss about the implementation in TRS, and try a mixed learning method using both SRS and TRS for sampling to avoid non desired local structures.

Sequencing learning steps: Balancing success and failure

A critical consideration in our current method is the balance between learning from successes and failures. Notably, during the initial stages, even short loops can experience a high number of backtracks before generating a valid conformation. For instance, Case 2A in Chapter 2 (a system with a short loop of 9 residues) exhibits an average of 96 backtracks per sampled conformation.

Focusing solely on learning from failures during this initial phase might be counterproductive. Therefore, we propose a sequenced learning approach that incorporates a pre-training phase using only successful conformations. This pre-training phase can be defined by a set of conformations (e.g., the first 100) or by a desired learning rate progression.

Introducing randomness for improved exploration

In addition to the sequenced learning approach, we propose incorporating randomness into the current learning process to enhance exploration. After the pre-training phase, a normal learning phase (as described in Section 4.3) starts with a modification during the selection of the structural region of a residue. A randomization rate, p_r , is defined, where the method has a p_r probability of choosing sr_k randomly and a $(1 - p_r)$ probability of using the learning network to define sr_k . This randomization rate gradually decreases as the sampling progresses, once again based on the number of sampled conformations or the learning rate evolution.

4.4.2 Perspectives and areas of exploration

Improved structural representation

The initial area of focus for improvement lies in refining the current depiction of the structural regions associated with residues. Existing literature suggests more elaborated classifications based on ϕ and ψ angles. Implementing such classifications with a higher number of classes could significantly enhance the learning and sampling process. However, careful consideration is necessary for this new approach, as it would necessitate a complete redefinition of our existing database.

Expanding the state description

Another promising avenue for exploration involves refining the current state representation used by the learning method. We currently use a TRS approach, considering the two neighboring residues. Expanding this scope to include a more precise state with additional neighbors can be explored. A preliminary approach within the TRS framework could involve incorporating an additional preceding residue (either AA_{k-2} or AA_{k+2} depending on the sampling direction). This would result in a state encompassing a complete sampled tripeptide (e.g., $AA_{k-2} - AA_{k-1} - AA_k$ or $AA_k - AA_{k+1} - AA_{k+2}$). Furthermore, the current state representation does not account for the sampling direction, which might be detrimental to properly represent the sampling context. We could explore effective ways to integrate this information into the state description.

Transfer learning exploration

Finally, investigating the potential of transfer learning for our current method is crucial. Given the diverse nature of systems handled by MoMA-FReSa and the highly specific states used in our current approach, the direct implementation of transfer learning might be challenging. However, its potential to transfer learning from quickly sampled short systems to accelerate knowledge acquisition for similar, but more complex, systems is highly attractive. A comprehensive study will be required to assess its feasibility and potential benefits.

Experimental Data or Energy Computations

While the RL approach of MoMA-FReSa avoids dependence on experimental data, its availability could be advantageous. Integrating these data has the potential to accelerate the learning process. Furthermore, we can exploit intrinsic capabilities of MoMA-FReSa to enhance learning efficiency. By using the energy calculations, we can potentially rank the actions taken by the agent. This ranking could then be used to provide more informative feedback to the learning algorithm, ultimately improving its decision-making capabilities.

4.5 Conclusion

In this project, we have addressed the implementation of a challenging and innovative AI-based method for MoMA-FReSa. We specifically considered the unique characteristics of MoMA-FReSa to design a tailored model. To ensure unbiased and efficient exploration of the entire ensemble of possible conformations, we maintained a degree of randomness in the sampling process and incorporated specific modules within our learning approach to preserve curiosity. This add-on could offer promising perspectives to MoMA-FReSa, especially in loop sampling and for the generation of conformations of complex systems involving multiple chains.

Currently, we are working to finalize the implementation with the addition of promising features and to perform an evaluation of the methods on a benchmark set. Furthermore, many perspectives are already considered to improve our model. The results of the upcoming evaluation should help pointing the strengths, weaknesses and needs of our current model, guiding future development efforts.

Conclusions and perspectives

Summary of the thesis: The potential of MoMA-FReSa

This thesis introduced MoMA-FReSa (Molecular Motion Algorithms - Flexible Regions Sampler), a novel and versatile tool specifically designed to address the challenges of conformational sampling for Intrinsically Disordered Proteins (IDPs) and Intrinsically Disordered Regions (IDRs). Chapter 1 highlighted the limitations of existing methods for IDP/IDR modeling. MoMA-FReSa emerges as a solution, offering an all-atom conformational sampling method applicable to a broad range of flexible biomolecular systems.

Chapter 2 developed the core functionalities of MoMA-FReSa that contribute to its effectiveness in conformational sampling for IDPs and IDRs. A key strength lies in the hierarchical decomposition of protein structures into regions. This decomposition, coupled with region scheduling, means that MoMA-FReSa can be categorized as a state-space search algorithm. This approach enables MoMA-FReSa to efficiently navigate the vast conformational space of IDPs and IDRs, a crucial aspect for tackling their inherent complexity. Furthermore, the introduction of stochastic elements injects non-determinism into the process. This non-determinism fosters unbiased exploration of the conformational space.

Chapter 2 also presented the performances of MoMA-FReSa on a benchmark set, highlighting its versatility across various protein systems. In addition, Chapter 3 demonstrated the interest of MoMA-FReSa in various applications. These applications included the conception of complex multi-modular systems, structural analyses of linkers in multi-domain proteins, sampling of structural motifs, and the estimation of effective concentrations in biomolecular interactions. The successful application of MoMA-FReSa in these diverse areas, combined with the good performance on the benchmark set, underlines the significant potential of MoMA-FReSa and its associated post-processing tools for advancing in the understanding of the functional roles of intrinsic disorder in proteins.

While MoMA-FReSa demonstrated promising results, Chapter 2 acknowledged the ongoing challenge of loop sampling in protein structures. The introduction of an improved stochastic process significantly improved loop building within MoMA-FReSa. However, dealing with long and/or multiple loops remains an issue. To further enhance the ability of MoMA-FReSa to handle these intricate regions, Chapter 4 presented a novel Reinforcement Learning (RL) method specifically designed for improved sampling of disordered regions under strong constraints, as is the case for loops. This RL approach is still under development, but it holds potential for tackling even the most challenging scenarios.

Current limitations and work in progress: A tool still under development

While MoMA-FReSa has demonstrated significant potential, there are several areas where further development can enhance its capabilities. The machine learning approach introduced in Chapter 4, holds significant promise, but further development will be necessary. While this RL approach is very well-suited for MoMA-FReSa, this close fit presents both advantages and disadvantages. On the positive side, this close coupling allows MoMA-FReSa to use the strengths of the RL method effectively. However, the uniqueness of the architecture necessitates careful parameterization without clear precedents to guide us. This optimization process is ongoing, and the results will orient future decisions to optimize our machine learning approach.

The tripeptide database defined in Chapter 2 is a dynamic element with room for improvement. The current database, due to its structure and lack of optimization, requires a long loading time when the program is initiated. Optimizing the database structure could significantly reduce loading times but also enhance functionality for future applications. For instance, this could enable more refined selection within the machine learning process and expanding sorting options beyond the current limitations of the Ramachandran plot as suggested in Chapter 4. Additionally, creating specialized databases tailored for specific problems could extend the applicability of MoMA-FReSa to a wider range of scenarios, for example transmembrane regions, as in the systems presented in Chapter 3, in the context of collaborations with the *École Nationale Supérieure Agronomique de Toulouse* (INP-ENSAT) and the iGEM (International Genetically Engineered Machine) project.

MoMA-FReSa currently uses a simplified hydropathy scale (HPS) [57] [171] to evaluate energy. This could be significantly improved by incorporating more complex energy functions, such as CALVADOS [199]. The structural representation used by MoMA-FReSa allows for optimized energy computations avoiding redundant calculations within rigid regions of the protein (intra-region computations) or between rigid regions spatially fixed together (inter-regions computations). By incorporating these more sophisticated energy models, MoMA-FReSa could generate more physically-meaningful conformational ensembles, with more accurate associated weights to each sampled state.

Perspectives and wider horizons for MoMA-FReSa

MoMA-FReSa can be used to model intricate biological systems with numerous disordered regions, pushing the boundaries of current methods. Indeed, MoMA-FReSa will have the capability to model extremely complex biological system, such as the transcriptional machinery or some signaling events. Moreover, MoMA-FReSa can act as a valuable complement to other methods like coarse-grained or atomistic molecular dynamics simulations. By providing structurally diverse initial states for simulations, MoMA-FReSa can significantly enhance the capability of these methods to cover the huge conformational space of disordered systems.

While MoMA-FReSa possesses significant potential, its current implementation lacks user-friendliness. To address this and unlock its full potential for the scientific community, development efforts are focused on creating an intuitive user interface built in Python to serve a dual purpose. A free web application will provide a user-friendly browser interface, allowing researchers to easily utilize MoMA-FReSa without requiring technical expertise. Additionally, a freely available Apptainer container will cater to users who prefer local installations. This option will include clear guidance for installation and use, allowing researchers to access the whole set of parameters of MoMA-FReSa. The user interface will be designed to facilitate specifications on the studied systems: adjust flexible regions (size and nature), specify spatial constraints, *etc...* This level of user control aligns with the details described in the construction section of Chapter 2, allowing researchers to tailor the functionalities of MoMA-FReSa to their specific research questions.

In parallel to the development of the user interface, we are actively exploring ways to combine the AlphaFold Protein Structure Database [211] with MoMA-FReSa. With over 200 millions protein structure predictions covering a broad range of the biological diversity, this vast and publicly available resource has significantly accelerated protein structure research. While AlphaFold2 excels at single conformation predictions for well-folded proteins, it has limitations in representing IDPs and IDRs, which require ensembles of conformations for accurate representation. MoMA-FReSa can address this gap by generating ensembles of conformations

for AlphaFold2 predictions flagged as having low confidence (often corresponding to flexible regions [177]). The user interface will allow researchers to adjust sampling parameters specific to their needs and the desired minimal level of confidence for structured regions. This integration will significantly improve the AlphaFold Protein Structure Database by providing a more comprehensive picture of protein structures involving highly-flexible regions.

Building on the applications presented in Chapter 3, the MoMA-FReSa potential extends to several exciting areas. One particularly promising area is the computation of effective concentration, which could become a powerful tool for understanding protein-protein interactions involving IDRs. MoMA-FReSa also showed capabilities in protein design that warrant further investigation. This line of development holds significant promise for advancements in designing flexible proteins with specific functionalities.

Résumé en Français

Nous fournissons ici un résumé en langue française des travaux présentés dans ce manuscrit de thèse.

A.1 Chapitre 1: Introduction

L'importance biologique des protéines dites désordonnées (IDPs) et des régions désordonnées (IDRs) souligne l'intérêt de disposer de modèles structuraux détaillés de cette classe de protéines et de leurs complexes. Ces modèles garantissent une perspective moléculaire des processus cellulaires clés et d'éventuelles interventions rationnelles à visée pharmacologique [3]. La coexistence d'un nombre astronomique de conformations et la nature moyenne des données expérimentales qui peuvent être enregistrées pour les IDPs/IDRs rendent l'utilisation de méthodes computationnelles inévitable. Les immenses défis dans ce domaine sont illustrés par l'étude des gouttelettes de type liquide, qui ont suscité l'intérêt d'une large communauté issue de divers domaines scientifiques [24, 2]. Ces condensats de protéines hautement concentrés sont intrinsèquement désordonnés et présentent des interactions intermoléculaires faibles et multivalentes qui sont modulées par des paramètres externes tels que le pH, la température ou les états de phosphorylation [156]. Ils présentent donc de multiples défis pour la modélisation informatique.

L'intérêt croissant de la communauté de la bioinformatique structurale pour relever les défis posés par les IDPs/IDRs est encourageant. L'amélioration des champs de force, tant pour les simulations tout-atome que pour les simulations de type gros grains, afin de les adapter aux états désordonnés, le développement de stratégies d'échantillonnage améliorées, ainsi que la généralisation des logiciels parallélisés et l'utilisation des GPU sont les indices les plus marquants de ces développements. L'augmentation du nombre d'études expérimentales axées sur les IDPs/IDRs est également cruciale car elle permet d'identifier de nouveaux mécanismes biologiques. En outre, les bases de données et les dépôts de données expérimentales et omiques améliorent notre connaissance structurale et fonctionnelle de ces protéines et offrent de nouvelles possibilités de développer et de valider les méthodes théoriques [88, 120]. Ces nouvelles données sont riches en informations et peuvent être utilisées, par exemple, pour améliorer les champs de force actuels, ou peuvent être exploitées pour concevoir des méthodes d'échantillonnage conformationnel plus précises [67]. L'utilisation de méthodes d'exploration de données et d'apprentissage automatique pour analyser et exploiter les informations pertinentes de ces bases de données est une voie très prometteuse pour l'amélioration des approches de modélisation moléculaire prédictive et pour le développement de nouveaux outils permettant d'aborder les questions difficiles posées par les protéines désordonnées et leurs complexes.

Dans ce contexte, cette thèse présente MoMA-FReSa (Molecular Motion Algorithms - Flexible Regions Sampler), une méthode innovante pour explorer les états conformationnels d'entités biomoléculaires désordonnées. Comme le souligne le chapitre 1, les méthodes existantes s'adressent souvent à des types de systèmes spécifiques et/ou sont très coûteuses en

termes de calcul. MoMA-FReSa, cependant, aspire à être un outil plus général et plus accessible sur le plan informatique, capable d’explorer le vaste espace conformationnel d’une large gamme de systèmes et de combler la lacune critique dans le domaine de la modélisation IDP/IDR. En outre, nous visons à avoir la capacité de modéliser les systèmes biomoléculaires les plus complexes, qui ne peuvent pas être étudiés par des méthodes de simulation moléculaire plus conventionnelles. Globalement, notre objectif principal est de fournir aux chercheurs un outil polyvalent qui dépasse les limites des méthodes actuelles.

En fournissant une feuille de route détaillée, les sous-sections suivantes décrivent la structure de ce manuscrit et soulignent les principaux résultats et contributions de cette thèse.

A.2 Chapitre 2 : Une description exhaustive du MoMA-FReSa

MoMA-FReSa est une nouvelle méthode d’échantillonnage stochastique à l’échelle atomique, conçue pour l’exploration de l’espace conformationnel des IDPs et IDRs. Elle s’appuie sur des méthodes établies telles que Flexible-Meccano [149] [16] en utilisant les angles dièdres des acides aminés pour générer des modèles d’ensemble. Pour développer un aspect plus polyvalent, MoMA-FReSa élargit certains aspects de ces méthodes traditionnelles. Par exemple, contrairement à Flexible-Meccano [16], les trois angles dièdres du squelette sont pris en compte (pas seulement ϕ et ψ) et les deux directions d’échantillonnage sont prises en compte en fonction des systèmes (pas seulement N-to-C).

Notre approche d’échantillonnage s’aligne sur des méthodes récentes également basées sur la construction du squelette par la sélection d’angles dièdres comme IDPConformerGenerator [198]. Bien que IDPConformerGenerator ait également introduit le module Local Disordered Region Sampling (LDRS) pour générer des ensembles d’IDRs [127], ses limites dans la diversité des systèmes qui peuvent être traités et son efficacité de calcul soulignent la nécessité d’une méthode plus générale et optimisée.

MoMA-FReSa s’aligne sur la philosophie de la suite MoMA (<https://moma.laas.fr/>) développée au Laboratoire d’Analyse et d’Architecture des Systèmes du Centre National de la Recherche Scientifique (LAAS-CNRS). Cette approche suit en particulier les travaux d’Alejandro Estaña [67] sur les approches des fragments à trois résidus, et d’Amélie Barozet [9] sur la modélisation des boucles. La sélection des angles dièdres repose sur une base de données de fragments à trois résidus [68] extraits de structures de protéines déterminées expérimentalement [73].

Le premier objectif de MoMA-FReSa est d’établir un algorithme unifié capable de traiter divers systèmes biomoléculaires avec des parties désordonnées. Pour ce faire, la méthode emploie une étape de prétraitement qui décompose hiérarchiquement le système en sous-régions de résidus d’acides aminés. Cette étape initiale permet non seulement à MoMA-FReSa d’être polyvalent, mais aussi d’optimiser le processus d’échantillonnage ultérieur. Au cours de cette phase d’échantillonnage, MoMA-FReSa agit comme un algorithme de recherche dans l’espace d’état visant à trouver une conformation réalisable par une approche de résolution de problèmes non déterministe. En effet, pour gérer de manière optimale l’exploration du vaste espace des conformations possibles, MoMA-FReSa utilise un processus stochastique hiérarchique.

Le chapitre 2 présente la méthode MoMA-FReSa et fournit une évaluation de ses capacités. Le chapitre démontre la polyvalence de MoMA-FReSa en l’appliquant à une série de systèmes de référence. Les résultats de ce chapitre fournissent des preuves convaincantes du potentiel

de MoMA-FReSa.

A.3 Chapitre 3 : Applications dans un cadre collaboratif

Après avoir établi la méthodologie et les capacités de MoMA-FReSa dans le chapitre précédent, nous appliquons le programme à de multiples cas. Ce chapitre montre la polyvalence de MoMA-FReSa au-delà de l'échantillonnage traditionnel grâce à divers projets dans un cadre collaboratif :

- **Conception des systèmes multimodulaires dans le cadre de la compétition iGEM:** la méthode MoMA-FReSa a été utilisée pour la conception d'une protéine multimodulaire dans le cadre du concours de biologie synthétique iGEM (International Genetically Engineered Machine). Une équipe de l'Institut national des sciences appliquées de Toulouse - Université Paul Sabatier (INSA-UPS) a fait appel à MoMA-FReSa pour développer CALIPSO, un système d'administration ciblée de médicaments destiné au traitement du cancer (<https://2023.igem.wiki/toulouse-insa-ups/home>).
- **Estimation des concentrations effectives dans les interactions moléculaires:** Une collaboration a été initiée avec le groupe du Dr. Lucia Chemes à l'Université de San Martin (Argentine) pour estimer les concentrations effectives dans les interactions intramoléculaires et intermoléculaires avec MoMA-FReSa. Cette collaboration initiale s'est concentrée sur un système particulièrement difficile [164].
- **L'analyse structurelle dans le cadre du projet CORNFLEX:** Les capacités de MoMA-FReSa pour l'analyse structurelle ont été explorées dans le contexte du projet CORNFLEX, une initiative de recherche collaborative impliquant plusieurs institutions françaises (<https://anr.fr/Project-ANR-22-CE45-0003>). Cette collaboration a également permis d'illustrer certaines des capacités d'analyse post-traitement de MoMA-FReSa.
- **Échantillonnage des motifs structurels par l'étude du récepteur 2 de l'éthylène chez *Vitis vinifera*:** Dans le cadre d'une collaboration avec des étudiants de l'École Nationale Supérieure Agronomique de Toulouse (INP-ENSAT), nous avons appliqué MoMA-FReSa pour étudier une région flexible du second récepteur de l'éthylène (ETR2) de *Vitis vinifera* [41]. Cette étude nous a permis de relever le défi de la modélisation des motifs structuraux. Ce travail a également servi à démontrer les diverses méthodes d'échantillonnage conformationnel offertes par MoMA-FReSa.

A.4 Chapitre 4 : Une approche d'apprentissage profond pour améliorer MoMA-FReSA

Enfin, le chapitre 4 explore une version améliorée de MoMA-FReSa, intégrant une méthode d'apprentissage par renforcement (RL) sans modèle [163, 183]. Les techniques d'intelligence artificielle (IA) sont devenues des outils de plus en plus puissants dans la modélisation et la conception des protéines. De nombreuses méthodes d'échantillonnage intègrent désormais des éléments d'apprentissage automatique ou sont entièrement basées sur des approches d'apprentissage profond [228, 234, 97, 233]. Suivant cette tendance prometteuse, nous avons travaillé sur une approche d'apprentissage automatique pour améliorer les capacités de

MoMA-FReSa. Parmi toutes les approches d'apprentissage automatique, la méthode RL était bien adaptée car elle ne nécessite pas de données expérimentales pour l'apprentissage. Au lieu de cela, l'apprentissage est effectué simultanément à l'échantillonnage conformationnel.

L'objectif principal était d'améliorer l'échantillonnage des systèmes contenant des boucles. En effet, la modélisation des boucles est un défi. Même si des méthodes de modélisation pour les boucles multiples sont apparues, la plupart des méthodes actuelles d'échantillonnage des boucles se concentrent sur la modélisation des protéines à boucle unique. MoMA-FReSa a la capacité d'échantillonner des systèmes impliquant une ou plusieurs boucles ainsi que d'autres régions flexibles, comme décrit dans le chapitre 2. Cependant, les systèmes impliquant des boucles sont ceux qui prennent le plus de temps pour MoMA-FReSa. En raison des défis importants qu'elles représentent pour les méthodes traditionnelles [151], les boucles sont particulièrement bien adaptées aux approches d'apprentissage automatique. Dans la continuité des méthodes précédentes [9], nous avons mis en œuvre une méthode RL personnalisée spécialement adaptée à notre approche. Cependant, l'objectif de cette implémentation va au-delà des boucles ; nous visons à améliorer les performances de MoMA-FReSa pour tous les types de systèmes, en particulier ceux qui sont très complexes.

Notre approche utilise une stratégie d'apprentissage par renforcement axée sur la curiosité [74]. Notre implémentation se concentre sur les angles dièdres et, plus spécifiquement, sur les propensions de la structure secondaire. Pour ce faire, MoMA-FReSa bénéficie d'une approche d'échantillonnage spéciale, qui prend en compte le type structural du résidu d'acide aminé (α , β , ou γ) au cours du processus d'échantillonnage. Bien que les premiers résultats présentés dans le chapitre 4 ne soient pas encore concluants, nous restons optimistes quant au potentiel du modèle mis en œuvre. Nous pensons que la poursuite du développement est prometteuse pour transformer cet ajout en un outil précieux. Ce travail est en cours, et plusieurs possibilités d'amélioration sont actuellement explorées.

A.5 Disponibilité du logiciel

MoMA-FReSa a été développé dans le but de devenir un outil largement accessible à la communauté scientifique. Bien qu'une version entièrement fonctionnelle ait déjà été utilisée dans le cadre d'une recherche collaborative au chapitre 3, il manque encore une interface pour faciliter l'utilisation par des personnes extérieures à l'équipe.

Le développement est en cours et la création d'une interface utilisateur intuitive construite en Python est l'objectif principal actuel. Cette interface utilisateur servira de base à une interface web, qui aboutira finalement à une application web librement accessible aux chercheurs. En outre, il est prévu d'offrir un conteneur Apptainer disponible gratuitement pour les utilisateurs qui préfèrent une installation locale. Nous discuterons plus en détail de ce travail en cours dans les conclusions de la thèse.

Bibliography

- [1] C. Abrams and G. Bussi. Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration. *Entropy*, 16(1):163–199, 2014. (Cited in page 4.)
- [2] S. Alberti and D. Dormann. Liquid–liquid phase separation in disease. *Annual review of genetics*, 53:171–194, 2019. (Cited in pages 14 and 105.)
- [3] S. Ambadipudi and M. Zweckstetter. Targeting intrinsically disordered proteins in rational drug discovery. *Expert Opin Drug Discov*, 11(1):65–77, 2016. PMID: 26549326. (Cited in pages 14 and 105.)
- [4] G. Apic, W. Huber, and S. A. Teichmann. Multi-domain protein families and domain pairs: comparison with known structures and a random model of domain recombination. *J Struct Funct Genomics*, 4(2-3):67–78, 2003. (Cited in page 5.)
- [5] M. Arbesú, G. Iruela, H. Fuentes, J. M. C. Teixeira, and M. Pons. Intramolecular fuzzy interactions involving intrinsically disordered domains. *Front Mol Biosci*, 5:39, 2018. (Cited in page 9.)
- [6] M. Arbesú, M. Maffei, T. N. Cordeiro, J. M. C. Teixeira, Y. Pérez, P. Bernadó, S. Roche, and M. Pons. The unique domain forms a fuzzy intramolecular complex in Src family kinases. *Structure*, 25:630–640, 2017. (Cited in page 9.)
- [7] A. Bah and J. D. Forman-Kay. Modulation of intrinsically disordered protein function by post-translational modifications. *J Biol Chem*, 291:6696–6705, 2016. (Cited in page 3.)
- [8] A. Barducci, M. Bonomi, and M. Parrinello. Metadynamics. *WIREs Comput Mol Sci*, 1(5):826–843, 2011. (Cited in page 4.)
- [9] A. Barozet, K. Molloy, M. Vaisset, T. Simeon, and J. Cortés. A Reinforcement-Learning-Based Approach to Enhance Exhaustive Protein Loop Sampling. *Bioinformatics*, 36(4):1099–1106, Feb. 2020. (Cited in pages 11, 13, 15, 16, 48, 49, 89, 106, and 108.)
- [10] P. Barthe, C. Roumestand, M. J. Canova, L. Kremer, C. Hurard, V. Molle, and M. Cohen-Gonsaud. Dynamic and structural characterization of a bacterial FHA protein reveals a new autoinhibition mechanism. *Structure*, 17(4):568–578, 2009. (Cited in page 9.)
- [11] R. Bayliss, T. Littlewood, and M. Stewart. Structural basis for the interaction between fxfp nucleoporin repeats and importin- β in nuclear trafficking. *Cell*, 102(1):99–108, 2000. (Cited in page 10.)
- [12] Z. Benayad, S. von Bülow, L. S. Stelzl, and G. Hummer. Simulation of fus protein condensates with an adapted coarse-grained model. *J Chem Theory Comput*, 17(1):525–537, 2021. PMID: 33307683. (Cited in page 12.)
- [13] T. Berau and M. Deserno. Generic coarse-grained model for protein folding and aggregation. *J Chem Phys*, 130(23):235106, 2009. (Cited in page 4.)

- [14] L. Bergeron-Sandoval, N. Safaee, and S. Michnick. Mechanisms and consequences of macromolecular phase separation. *Cell*, 165:1067–1079, 2016. (Cited in pages 3 and 12.)
- [15] P. Bernadó and M. Blackledge. Proteins in dynamic equilibrium. *Nature*, 468:1046–1048, 2010. (Cited in page 3.)
- [16] P. Bernadó, L. Blanchard, P. Timmins, D. Marion, R. W. H. Ruigrok, and M. Blackledge. A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. *Proceedings of the National Academy of Sciences*, 102(47):17002–17007, 2005. (Cited in pages 5, 11, 14, and 106.)
- [17] P. Bernadó, K. Modig, P. Grela, D. I. Svergun, M. Tchorzewski, M. Pons, and M. Akke. Structure and dynamics of ribosomal protein l12: An ensemble model based on saxs and nmr relaxation. *Biophysical journal*, 98(10):2374–2382, May 2010. (Cited in pages 7 and 49.)
- [18] P. Bernadó, E. Mylonas, M. V. Petoukhov, M. Blackledge, and D. I. Svergun. Structural characterization of flexible proteins using small-angle x-ray scattering. *J Am Chem Soc*, 129(17):5656–5664, 2007. PMID: 17411046. (Cited in page 5.)
- [19] I. Bertini, A. Giachetti, C. Luchinat, G. Parigi, M. V. Petoukhov, R. Pierattelli, E. Ravera, and D. I. Svergun. Conformational space of flexible biological macromolecules from average data. *J Am Chem Soc*, 132(38):13553–13558, 2010. (Cited in page 5.)
- [20] R. B. Best. Computational and theoretical advances in studies of intrinsically disordered proteins. *Curr Opin Struct Biol*, 42:147 – 154, 2017. (Cited in pages 3, 4, and 12.)
- [21] R. B. Best, W. Zheng, and J. Mittal. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J Chem Theory Comput*, 10(11):5113–5124, 2014. (Cited in page 4.)
- [22] T. Beysolow II. *Applied Reinforcement Learning with Python: With OpenAI Gym, Tensorflow, and Keras*. Apress, 2019. (Cited in page 93.)
- [23] S. Bhattacharya and X. Lin. Recent advances in computational protocols addressing intrinsically disordered proteins. *Biomolecules*, 9(4), 2019. (Cited in pages 3 and 12.)
- [24] S. Boeynaems, S. Alberti, N. L. Fawzi, T. Mittag, M. Polymenidou, F. Rousseau, J. Schymkowitz, J. Shorter, B. Wolozin, L. Van Den Bosch, P. Tompa, and M. Fuxreiter. Protein phase separation: A new phase in cell biology. *Trends Cell Biol*, 28(6):420–435, 2018. (Cited in pages 12, 14, and 105.)
- [25] M. Bonomi, C. Camilloni, A. Cavalli, and M. Vendruscolo. Metainference: A bayesian inference method for heterogeneous systems. *Sci Adv*, 2(1), 2016. (Cited in page 5.)
- [26] M. Bonomi, G. T. Heller, C. Camilloni, and M. Vendruscolo. Principles of protein structural ensemble determination. *Curr Opin Struct Biol*, 42:106–116, 2017. (Cited in page 5.)
- [27] W. Boomsma, J. Ferkinghoff-Borg, and K. Lindorff-Larsen. Combining experiments and simulations using the maximum entropy principle. *PLoS Comput Biol*, 10(2):1–9, 02 2014. (Cited in page 5.)

- [28] M. Borg, T. Mittag, T. Pawson, M. Tyers, J. D. Forman-Kay, and H. S. Chan. Polyelectrostatic interactions of disordered ligands suggest a physical basis for ultrasensitivity. *Proc Natl Acad Sci USA*, 104(23):9650–9655, 2007. (Cited in page 10.)
- [29] A. Borgia, M. Borgia, K. Bugge, V. Kissling, P. Heidarsson, C. Fernandes, A. Sottini, A. Soranno, K. Buholzer, D. Nettels, B. B. Kragelund, R. Best, and B. Schuler. Extreme disorder in an ultrahigh-affinity protein complex. *Nature*, 555:61–66, 03 2018. (Cited in pages 3 and 12.)
- [30] A. Borgia, W. Zheng, K. Buholzer, M. B. Borgia, A. Schüler, H. Hofmann, A. Soranno, D. Nettels, K. Gast, A. Grishaev, R. B. Best, and B. Schuler. Consistent view of polypeptide chain expansion in chemical denaturants from multiple experimental methods. *J Am Chem Soc*, 138(36):11714–11726, 2016. (Cited in page 5.)
- [31] S. Bottaro, T. Bengtson, and K. Lindorff-Larsen. Integrating molecular simulation and experimental data: a bayesian/maximum entropy reweighting approach. *Structural bioinformatics: methods and protocols*, pages 219–240, 2020. (Cited in page 5.)
- [32] S. Bottaro, T. Bengtson, and K. Lindorff-Larsen. Integrating molecular simulation and experimental data: a bayesian/maximum entropy reweighting approach. *Structural bioinformatics: methods and protocols*, pages 219–240, 2020. (Cited in page 5.)
- [33] C. P. Brangwynne, P. Tompa, and R. V. Pappu. Polymer physics of intracellular phase transitions. *Nat Phys*, 11(11):899–904, 2015. (Cited in page 12.)
- [34] H. Bret, J. Gao, D. J. Zea, J. Andreani, and R. Guerois. From interaction networks to interfaces, scanning intrinsically disordered regions using alphafold2. *Nature Communications*, 15(1):597, 2024. (Cited in pages 8 and 13.)
- [35] U. Brinkmann and R. E. Kontermann. The making of bispecific antibodies. *mAbs*, 9(2):182–212, 2017. (Cited in page 6.)
- [36] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*, 2018. (Cited in page 93.)
- [37] Y. Burda, H. Edwards, A. Storkey, and O. Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018. (Cited in page 93.)
- [38] G. Bussi, F. L. Gervasio, A. Laio, and M. Parrinello. Free-energy landscape for β hairpin folding from combined parallel tempering and metadynamics. *J Am Chem Soc*, 128(41):13435–13441, 2006. (Cited in page 4.)
- [39] C. Charlier, G. Bouvignies, P. Pelupessy, A. Walrant, R. Marquant, M. Kozlov, P. De Ioannes, N. Bolik-Coulon, S. Sagan, P. Cortes, A. K. Aggarwal, L. Carlier, and F. Ferrage. Structure and dynamics of an intrinsically disordered protein region that partially folds upon binding by chemical-exchange NMR. *J Am Chem Soc*, 139(35):12219–12227, 2017. (Cited in page 8.)
- [40] Y. Chen, R. Althiab Almasaud, E. Carrie, G. Desbrosses, B. M. Binder, and C. Chervin. Ethanol, at physiological concentrations, affects ethylene sensing in tomato germinating seeds and seedlings. *Plant Science*, 291:110368, 2020. (Cited in page 81.)

- [41] Y. Chen, J. Grimplet, K. David, S. D. Castellarin, J. Terol, D. C. Wong, Z. Luo, R. Schaffer, J.-M. Celton, M. Talon, G. A. Gambetta, and C. Chervin. Ethylene receptors and related proteins in climacteric and non-climacteric fruits. *Plant Science*, 276:63–72, 2018. (Cited in pages 16, 81, and 107.)
- [42] J.-M. Choi, F. Dar, and R. V. Pappu. LASSI: A lattice model for simulating phase transitions of multivalent proteins. *PLoS Comput Biol*, 15(10):e1007028, 2019. (Cited in page 12.)
- [43] A. E. Conicella, G. L. Dignon, G. H. Zerze, H. B. Schmidt, A. M. D’Ordine, Y. C. Kim, R. Rohatgi, Y. M. Ayala, J. Mittal, and N. L. Fawzi. TDP-43 α -helical structure tunes liquid-liquid phase separation and function. *Proc Natl Acad Sci USA*, 117(11):5883–5894, 2020. (Cited in page 12.)
- [44] T. N. Cordeiro, P.-C. Chen, A. De Biasio, N. Sibille, F. J. Blanco, J. S. Hub, R. Crehuet, and P. Bernadó. Disentangling polydispersity in the PCNA-p15PAF complex, a disordered, transient and multivalent macromolecular assembly. *Nucleic Acids Res*, 45(3):1501–1515, 11 2016. (Cited in page 3.)
- [45] T. N. Cordeiro, F. Herranz-Trillo, A. Urbanek, A. Estaña, J. Cortés, N. Sibille, and P. Bernadó. Small-angle scattering studies of intrinsically disordered proteins and their complexes. *Curr Opin Struct Biol*, 42:15 – 23, 2017. (Cited in page 3.)
- [46] T. N. Cordeiro, N. Sibille, P. Germain, P. Barthe, A. Boulahtouf, F. Allemand, R. Bailly, V. Vivat, C. Ebel, A. Barducci, W. Bourguet, A. le Maire, and P. Bernadó. Interplay of protein disorder in retinoic acid receptor heterodimer and its corepressor regulates gene expression. *Structure*, 27(8):1270–1285, 2019. (Cited in page 11.)
- [47] A. C.R.Martin, K. Toda, H. J. Stirk, and J. M. Thornton. Long loops in proteins. *Protein Engineering, Design and Selection*, 8(11):1093–1101, 11 1995. (Cited in page 49.)
- [48] V. Csizmok, A. V. Follis, R. W. Kriwacki, and J. D. Forman-Kay. Dynamic protein interaction networks and new structural paradigms in signaling. *Chem Rev*, 116(11):6424–6462, 2016. (Cited in page 1.)
- [49] V. Csizmok and J. D. Forman-Kay. Complex regulatory mechanisms mediated by the interplay of multiple post-translational modifications. *Curr Opin Struct Biol*, 48:58 – 67, 2018. (Cited in page 3.)
- [50] S. Das, Y.-H. Lin, R. M. Vernon, J. D. Forman-Kay, and H. S. Chan. Comparative roles of charge, π , and hydrophobic interactions in sequence-dependent phase separation of intrinsically disordered proteins. *Proc Natl Acad Sci USA*, 117(46):28795–28805, 2020. (Cited in page 12.)
- [51] N. E. Davey. The functional importance of structure in unstructured protein regions. *Curr Opin Struct Biol*, 56:155–163, 2019. (Cited in page 7.)
- [52] N. E. Davey, G. Travé, and T. J. Gibson. How viruses hijack cell regulation. *Trends Biochem Sci*, 36(3):159–169, 2011. (Cited in page 8.)
- [53] N. E. Davey, K. Van Roey, R. J. Weatheritt, G. Toedt, B. Uyar, B. Altenberg, A. Budd, F. Diella, H. Dinkel, and T. J. Gibson. Attributes of short linear motifs. *Mol BioSyst*, 8:268–281, 2012. (Cited in pages 3 and 7.)

- [54] V. R. de Angulo, J. Cortés, and J. M. Porta. Rigid-ell: Avoiding constant-distance computations in cell linked-lists algorithms. *Journal of Computational Chemistry*, 33(3):294–300, 2012. (Cited in page 47.)
- [55] D. H. de Jong, G. Singh, W. F. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman, and S. J. Marrink. Improved parameters for the martini coarse-grained protein force field. *J Chem Theory Comput*, 9(1):687–697, 2013. (Cited in page 4.)
- [56] I. M. S. de Vera, J. Zheng, S. Novick, J. Shang, T. S. Hughes, R. Brust, P. Munoz-Tello, W. J. Gardner, D. P. Marciano, X. Kong, P. R. Griffin, and D. J. Kojetin. Synergistic regulation of coregulator/nuclear receptor interaction by ligand and DNA. *Structure*, 25(10):1506–1518.e4, 2017. (Cited in page 11.)
- [57] G. L. Dignon, W. Zheng, Y. C. Kim, R. B. Best, and J. Mittal. Sequence determinants of protein phase behavior from a coarse-grained model. *PLoS Computational Biology*, 14, 2017. (Cited in pages 12, 46, 77, and 102.)
- [58] G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal. Temperature-controlled liquid-liquid phase separation of disordered proteins. *ACS Cent Sci*, 5(5):821–830, 2019. (Cited in page 12.)
- [59] A. Diot, G. Groth, S. Blanchet, and C. Chervin. Responses of animals and plants to physiological doses of ethanol: a way of sensing climate change? working paper or preprint, Mar. 2023. (Cited in page 81.)
- [60] Z. Dosztányi, B. Mészáros, and I. Simon. ANCHOR: web server for predicting protein binding regions in disordered proteins. *Bioinformatics*, 25(20):2745–2746, 08 2009. (Cited in page 4.)
- [61] E. F. Dudás, G. Pálffy, D. K. Menyhárd, F. Sebák, P. Ecsédi, L. Nyitray, and A. Bodor. Tumor-suppressor p53TAD(1-60) forms a fuzzy complex with metastasis-associated S100A4: Structural insights and dynamics by an NMR/MD approach. *ChemBioChem*, 21(21):3087–3095, 2020. (Cited in page 11.)
- [62] H. J. Dyson and P. E. Wright. Intrinsically unstructured proteins and their functions. *Nat Rev Mol Cell Biol*, 6:197–208, 2005. (Cited in page 1.)
- [63] H. J. Dyson and P. E. Wright. Nmr illuminates intrinsic disorder. *Curr Opin Struct Biol*, 70:44–52, 2021. (Cited in page 3.)
- [64] C. A. Elena-Real, A. Sagar, A. Urbanek, M. Popovic, A. Morató, A. Estaña, A. Fournet, C. Doucet, X. L. Lund, Z.-D. Shi, et al. The structure of pathogenic huntingtin exon 1 defines the bases of its aggregation propensity. *Nature structural & molecular biology*, 30(3):309–320, 2023. (Cited in page 87.)
- [65] J. R. Espinosa, J. A. Joseph, I. Sanchez-Burgos, A. Garaizar, D. Frenkel, and R. Collepardo-Guevara. Liquid network connectivity regulates the stability and composition of biomolecular condensates with many components. *Proc Natl Acad Sci USA*, 117(24):13238–13247, 2020. (Cited in page 12.)

- [66] A. N. Estaña, A. Barozet, A. Mouhand, M. Vaisset, C. Zanon, P. Fauret, N. Sibille, P. N. Bernadó, and J. Cortés. Predicting secondary structure propensities in IDPs using simple statistics from three-residue fragments. *Journal of Molecular Biology*, 342(19):5447–5459, Sept. 2020. (Cited in pages 47 and 78.)
- [67] A. N. Estaña, N. Sibille, E. Delaforge, M. Vaisset, J. Cortés, and P. Bernadó. Realistic Ensemble Models of Intrinsically Disordered Proteins Using a Structure-Encoding Coil Database. *Structure*, 27(5):381–391.e2, 2019. (Cited in pages 5, 14, 15, 42, 90, 105, and 106.)
- [68] A. Estaña, M. Ghallab, P. Bernadó, and J. Cortés. Investigating the formation of structural elements in proteins using local sequence-dependent information and a heuristic search algorithm. *Molecules*, 24(6), 2019. (Cited in pages 15 and 106.)
- [69] A. S. Ettayapuram Ramaprasad, S. Uddin, J. Casas-Finet, and D. J. Jacobs. Decomposing dynamical couplings in mutated scFv antibody fragments into stabilizing and destabilizing effects. *J Am Chem Soc*, 139(48):17508–17517, 2017. (Cited in page 6.)
- [70] R. Evans, M. O’Neill, A. Pritzel, N. Antropova, A. Senior, T. Green, A. Žídek, R. Bates, S. Blackwell, J. Yim, et al. Protein complex prediction with alphafold-multimer. *bioRxiv*, pages 2021–10, 2021. (Cited in page 8.)
- [71] H. J. Feldman and C. W. Hogue. Probabilistic sampling of protein conformations: New hope for brute force? *Proteins*, 46(1):8–23, 2002. (Cited in page 5.)
- [72] T. Flock, R. J. Weatheritt, N. S. Latysheva, and M. M. Babu. Controlling entropy to tune the functions of intrinsically disordered regions. *Curr Opin Struct Biol*, 26:62–72, 2014. New constructs and expression of proteins / Sequences and topology. (Cited in page 7.)
- [73] N. Fox, S. Brenner, and J.-M. Chandonia. Scope: Structural classification of proteins - extended, integrating scop and astral data and classification of new structures. *Nucleic acids research*, 42, 12 2013. (Cited in pages 15, 47, and 106.)
- [74] M. Frank, J. Leitner, M. Stollenga, A. Förster, and J. Schmidhuber. Curiosity driven reinforcement learning for motion planning on humanoids. *Frontiers in Neurorobotics*, 7, 2014. (Cited in pages 16, 93, and 108.)
- [75] H. Fukunishi, O. Watanabe, and S. Takada. On the hamiltonian replica exchange method for efficient sampling of biomolecular systems: Application to protein structure prediction. *J Chem Phys*, 116(20):9058–9067, 2002. (Cited in page 4.)
- [76] M. Fuxreiter. Classifying the binding modes of disordered proteins. *Int J Mol Sci*, 21(22), 2020. (Cited in pages 3 and 10.)
- [77] C. Geng, S. Narasimhan, J. P. Rodrigues, and A. M. Bonvin. Information-driven, ensemble flexible peptide docking using haddock. *Modeling Peptide-Protein Interactions: Methods and Protocols*, pages 109–138, 2017. (Cited in pages 4 and 8.)
- [78] R. A. George and J. Heringa. An analysis of protein domain linkers: their classification and role in protein folding. *Protein Eng Des Sel*, 15(11):871–879, 11 2002. (Cited in page 5.)

- [79] M. Ghallab, D. Nau, and P. Traverso. *Automated Planning: Theory and Practice*. The Morgan Kaufmann Series in Artificial Intelligence. Morgan Kaufmann, Amsterdam, 2004. (Cited in page 35.)
- [80] G.-N. W. Gomes, M. Krzeminski, A. Namini, E. W. Martin, T. Mittag, T. Head-Gordon, J. D. Forman-Kay, and C. C. Gradinaru. Conformational ensembles of an intrinsically disordered protein consistent with nmr, saxs, and single-molecule fret. *J Am Chem Soc*, 142(37):15697–15710, 2020. (Cited in page 5.)
- [81] N. S. Gonzalez-Foutel, W. M. Borchers, J. Glavina, S. Barrera-Vilarmau, A. Sagar, A. Estaña, A. Barozet, G. Fernandez-Ballester, C. Blanes-Mira, I. E. Sánchez, G. de Prat-Gay, J. Cortés, P. Bernadó, R. V. Pappu, A. S. Holehouse, G. W. Daughdrill, and L. B. Chemes. Conformational buffering underlies functional selection in intrinsically disordered protein regions. *bioRxiv*, 2021. (Cited in pages 49 and 70.)
- [82] J. González-Delgado, P. Bernadó, P. Neuvial, and J. Cortés. Wario: Weighted families of contact maps to characterize conformational ensembles of (highly-)flexible proteins. 2024. (Cited in pages 46 and 87.)
- [83] J. González-Delgado, A. Sagar, C. Zanon, K. Lindorff-Larsen, P. Bernadó, P. Neuvial, and J. Cortés. Wasco: A wasserstein-based statistical tool to compare conformational ensembles of intrinsically disordered proteins. *Journal of Molecular Biology*, 435(14):168053, 2023. *Computation Resources for Molecular Biology*. (Cited in pages 46, 83, and 87.)
- [84] J. Gsponer and M. Vendruscolo. Theoretical approaches to protein aggregation. *Protein Pept Lett*, 13(3):287–293, 2006. (Cited in page 12.)
- [85] N. Guex and M. C. Peitsch. Swiss-model and the swiss-pdb viewer: an environment for comparative protein modeling. *electrophoresis*, 18(15):2714–2723, 1997. (Cited in page 81.)
- [86] S. Guseva, V. Schnapka, W. Adamski, D. Maurin, R. W. Ruigrok, N. Salvi, and M. Blackledge. Liquid–liquid phase separation modifies the dynamic properties of intrinsically disordered proteins. *Journal of the American Chemical Society*, 145(19):10548–10563, 2023. (Cited in page 13.)
- [87] T. S. Harmon, A. S. Holehouse, M. K. Rosen, and R. V. Pappu. Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife*, 6:e30294, nov 2017. (Cited in page 12.)
- [88] A. Hatos, B. Hajdu-Soltész, A. M. Monzon, N. Palopoli, L. Álvarez, B. Aykac-Fas, C. Bassot, G. I. Benítez, M. Bevilacqua, A. Chasapi, L. Chemes, N. E. Davey, R. Davidović, A. Dunker, A. Elofsson, J. Gobeill, N. S. Foutel, G. Sudha, M. Guharoy, T. Horvath, V. Iglesias, A. V. Kajava, O. P. Kovacs, J. Lamb, M. Lambrugh, T. Lazar, J. Y. Leclercq, E. Leonardi, S. Macedo-Ribeiro, M. Macossay-Castillo, E. Maiani, J. A. Manso, C. Marino-Buslje, E. Martínez-Pérez, B. Mészáros, I. Mičetić, G. Minervini, N. Murvai, M. Necci, C. A. Ouzounis, M. Pajkos, L. Paladin, R. Pancsa, E. Papaleo, G. Parisi, E. Pasche, P. J. Barbosa Pereira, V. J. Promponas, J. Pujols, F. Quaglia, P. Ruch, M. Salvatore, E. Schad, B. Szabo, T. Szaniszló, S. Tamana, A. Tantos, N. Veljkovic, S. Ventura, W. Vranken, Z. Dosztányi, P. Tompa, S. C. E. Tosatto, and

- D. Piovesan. DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res*, 48(D1):D269–D276, 11 2019. (Cited in pages 14 and 105.)
- [89] M. K. Hazra and Y. Levy. Biophysics of phase separation of disordered proteins is governed by balance between short- and long-range interactions. *J Phys Chem B*, 125(9):2202–2211, 2021. PMID: 33629837. (Cited in page 12.)
- [90] E. D. Holmstrom, A. Holla, W. Zheng, D. Nettels, R. B. Best, and B. Schuler. Chapter ten - accurate transfer efficiencies, distance distributions, and ensembles of unfolded and intrinsically disordered proteins from single-molecule fret. In E. Rhoades, editor, *Intrinsically Disordered Proteins*, volume 611 of *Methods in Enzymology*, pages 287 – 325. Academic Press, 2018. (Cited in page 3.)
- [91] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B. L. de Groot, H. Grubmüller, and A. D. MacKerell. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods*, 14:71–73, 2017. (Cited in page 4.)
- [92] Q. Huang, M. Li, L. Lai, and Z. Liu. Allostery of multidomain proteins with disordered linkers. *Curr Opin Struct Biol*, 62:175 – 182, 2020. (Cited in page 6.)
- [93] G. Hummer and J. Köfinger. Bayesian ensemble refinement by replica simulations and reweighting. *J Chem Phys*, 143(24):243150, 2015. (Cited in page 5.)
- [94] I. M. Ilie and A. Caflisch. Simulation studies of amyloidogenic polypeptides and their aggregates. *Chem Rev*, 119(12):6956–6993, 2019. PMID: 30973229. (Cited in page 12.)
- [95] R. E. Ithuralde, A. E. Roitberg, and A. G. Turjanski. Structured and unstructured binding of an intrinsically disordered protein as revealed by atomistic simulations. *J Am Chem Soc*, 138(28):8742–8751, 2016. (Cited in page 9.)
- [96] M. P. Jacobson, D. L. Pincus, C. S. Rapp, T. J. Day, B. Honig, D. E. Shaw, and R. A. Friesner. A hierarchical approach to all-atom protein loop prediction. *Proteins: Structure, Function, and Bioinformatics*, 55(2):351–367, 2004. (Cited in page 49.)
- [97] G. Janson and M. Feig. Transferable deep generative modeling of intrinsically disordered protein conformations. *bioRxiv*, pages 2024–02, 2024. (Cited in pages 13, 16, and 107.)
- [98] J. A. Joseph, A. Reinhardt, A. Aguirre, P. Y. Chew, K. O. Russell, J. R. Espinosa, A. Garaizar, and R. Collepardo-Guevara. Physics-driven coarse-grained model for biomolecular phase separation with near-quantitative accuracy. *Nature Computational Science*, 1(11):732–743, 2021. (Cited in page 4.)
- [99] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. (Cited in pages 8 and 13.)
- [100] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, K. Tunyasuvunakool, O. Ronneberger, R. Bates, A. Žídek, A. Bridgland, et al. Alphafold 2. *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction*, 2020. (Cited in pages 13 and 81.)
- [101] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983. (Cited in pages 47 and 87.)

- [102] K. Kasahara, M. Shiina, J. Higo, K. Ogata, and H. Nakamura. Phosphorylation of an intrinsically disordered region of Ets1 shifts a multi-modal interaction ensemble to an auto-inhibitory state. *Nucleic Acids Res*, 46(5):2243–2251, 2018. (Cited in page 9.)
- [103] K. Kasahara, H. Terazawa, T. Takahashi, and J. Higo. Studies on molecular dynamics of intrinsically disordered proteins and their fuzzy complexes: A mini-review. *Comput Struct Biotechnol J*, 17:712 – 720, 2019. (Cited in pages 3 and 12.)
- [104] Y. C. Kim and G. Hummer. Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding. *J Mol Biol*, 375(5):1416–1433, 2008. (Cited in page 12.)
- [105] M. Kjaergaard. Estimation of effective concentrations enforced by complex linker architectures from conformational ensembles. *Biochemistry*, 61(3):171–182, 2022. PMID: 35061369. (Cited in pages 69 and 71.)
- [106] M. Kjaergaard, J. Glavina, and L. B. Chemes. Chapter six - predicting the effect of disordered linkers on effective concentrations and avidity with the “ceff calculator” app. In M. Merkx, editor, *Linkers in Biomacromolecules*, volume 647 of *Methods in Enzymology*, pages 145–171. Academic Press, 2021. (Cited in page 69.)
- [107] F. Klein, E. E. Barrera, and S. Pantano. Assessing SIRAH’s Capability to Simulate Intrinsically Disordered Proteins and Peptides. *J Chem Theory Comput*, 17(2):599–604, feb 2021. (Cited in page 4.)
- [108] J. S. Klein, S. Jiang, R. P. Galimidi, J. R. Keeffe, and P. J. Bjorkman. Design and characterization of structured protein linkers with differing flexibilities. *Protein Eng Des Sel*, 27(10):325–330, 10 2014. (Cited in page 6.)
- [109] P. Klein, T. Pawson, and M. Tyers. Mathematical modeling suggests cooperative interactions between a disordered polyvalent ligand and a single receptor site. *Curr Biol*, 13:1669–1678, 2003. (Cited in page 10.)
- [110] M. Knott and R. B. Best. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J Chem Phys*, 140(17):175102, 2014. (Cited in page 4.)
- [111] T. P. J. Knowles, M. Vendruscolo, and C. M. Dobson. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol*, 15(6):384–396, 2014. (Cited in page 12.)
- [112] V. Konda and J. Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing systems*, 12, 1999. (Cited in page 92.)
- [113] J. M. Krieger, G. Fusco, M. Lewitzky, P. C. Simister, J. Marchant, C. Camilloni, S. M. Feller, and A. De Simone. Conformational recognition of an intrinsically disordered protein. *Biophys J*, 106(8):1771–1779, 2014. (Cited in page 9.)
- [114] I. Krystkowiak and N. E. Davey. SLiMSearch: a framework for proteome-wide discovery and annotation of functional modules in intrinsically disordered regions. *Nucleic Acids Res*, 45(W1):W464–W469, 04 2017. (Cited in page 4.)

- [115] M. Krzeminski, J. A. Marsh, C. Neale, W.-Y. Choy, and J. D. Forman-Kay. Characterization of disordered proteins with ENSEMBLE. *Bioinformatics*, 29(3):398–399, 12 2012. (Cited in page 5.)
- [116] M. Kumar, M. Gouw, S. Michael, H. Sámano-Sánchez, R. Pancsa, J. Glavina, A. Diakogianni, J. A. Valverde, D. Bukirova, J. Čalyševa, N. Palopoli, N. E. Davey, L. B. Chemes, and T. J. Gibson. ELM—the eukaryotic linear motif resource in 2020. *Nucleic Acids Res*, 48(D1):D296–D306, 2019. (Cited in page 3.)
- [117] K. Kundert and T. Kortemme. Computational design of structured loops for new protein functions. *Biol Chem*, 400(3):275–288, 2019. (Cited in page 11.)
- [118] J. Köfinger, L. S. Stelzl, K. Reuter, C. Allande, K. Reichel, and G. Hummer. Efficient ensemble refinement by reweighting. *J Chem Theory Comput*, 15(5):3390–3401, 2019. (Cited in page 5.)
- [119] A. Laio and M. Parrinello. Escaping free-energy minima. *Proc Natl Acad Sci USA*, 99(20):12562–12566, 2002. (Cited in page 4.)
- [120] T. Lazar, E. Martínez-Pérez, F. Quaglia, A. Hatos, L. Chemes, J. A. Iserte, N. A. Méndez, N. A. Garrone, T. Saldaño, J. Marchetti, A. Rueda, P. Bernadó, M. Blackledge, T. N. Cordeiro, E. Fagerberg, J. D. Forman-Kay, M. Fornasari, T. J. Gibson, G.-N. W. Gomes, C. Gradinaru, T. Head-Gordon, M. R. Jensen, E. Lemke, S. Longhi, C. Marinobuslje, G. Minervini, T. Mittag, A. Monzon, R. V. Pappu, G. Parisi, S. Ricard-Blum, K. M. Ruff, E. Salladini, M. Skepö, D. Svergun, S. Vallet, M. Varadi, P. Tompa, S. C. E. Tosatto, and D. Piovesan. PED in 2021: A major update of the protein ensemble database for intrinsically disordered proteins. *Nucleic Acids Res*, 49(D1):D404–D411, 2020. (Cited in pages 3, 14, and 105.)
- [121] M. Levitt. A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59–107, 1976. (Cited in pages 45 and 48.)
- [122] M. Li, H. Cao, L. Lai, and Z. Liu. Disordered linkers in multidomain allosteric proteins: Entropic effect to favor the open state or enhanced local concentration to favor the closed state? *Protein Sci*, 27(9):1600–1610, 2018. (Cited in page 7.)
- [123] Y.-H. Lin, J. D. Forman-Kay, and H. S. Chan. Theories for sequence-dependent phase behaviors of biomolecular condensates. *Biochemistry*, 57(17):2499–2508, 2018. (Cited in page 12.)
- [124] K. Lindorff-Larsen and B. B. Kragelund. On the potential of machine learning to examine the relationship between sequence, structure, dynamics and function of intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167196, 2021. From Protein Sequence to Structure at Warp Speed: How AlphaFold Impacts Biology. (Cited in page 13.)
- [125] J. Liu and R. Nussinov. Molecular dynamics reveal the essential role of linker motions in the function of cullin–RING E3 ligases. *J Mol Biol*, 396(5):1508 – 1523, 2010. (Cited in page 6.)

- [126] Y. Liu, X. Wang, and B. Liu. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Brief Bioinform*, 20(1):330–346, 2017. (Cited in page 1.)
- [127] Z. H. Liu, J. M. C. Teixeira, O. Zhang, T. E. Tsangaris, J. Li, C. C. Gradinaru, T. Head-Gordon, and J. D. Forman-Kay. Local Disordered Region Sampling (LDRS) for ensemble modeling of proteins with experimentally undetermined or low confidence prediction segments. *Bioinformatics*, 39(12):btad739, 12 2023. (Cited in pages 13, 15, and 106.)
- [128] J. W. Locasale. Allovalency revisited: An analysis of multisite phosphorylation and substrate rebinding. *J Chem Phys*, 128(11):115106, 2008. (Cited in page 10.)
- [129] S. C. Lovell, J. M. Word, J. S. Richardson, and D. C. Richardson. The penultimate rotamer library. *Proteins: Structure, Function, and Bioinformatics*, 40(3):389–408, 2000. (Cited in page 46.)
- [130] B. Ma, C.-J. Tsai, T. Haliloğlu, and R. Nussinov. Dynamic allostery: Linkers are not merely flexible. *Structure*, 19(7):907 – 917, 2011. (Cited in page 6.)
- [131] M. Maffei, M. Arbesú, A.-L. Le Roux, I. Amata, S. Roche, and M. Pons. The SH3 domain acts as a scaffold for the n-terminal intrinsically disordered regions of c-Src. *Structure*, 23:893–902, 2015. (Cited in page 9.)
- [132] E. W. Martin, A. S. Holehouse, I. Peran, M. Farag, J. J. Incicco, A. Bremer, C. R. Grace, A. Soranno, R. V. Pappu, and T. Mittag. Valence and patterning of aromatic residues determine the phase behavior of prion-like domains. *Science*, 367(6478):694–699, 2020. (Cited in page 12.)
- [133] P. Mier, L. Paladin, S. Tamana, S. Petrosian, B. Hajdu-Soltész, A. Urbanek, A. Gruca, D. Plewczynski, M. Grynberg, P. Bernadó, Z. Gáspári, C. A. Ouzounis, V. J. Promponas, A. V. Kajava, J. M. Hancock, S. C. E. Tosatto, Z. Dosztanyi, and M. A. Andrade-Navarro. Disentangling the complexity of low complexity proteins. *Brief Bioinform*, 21(2):458–472, 2019. (Cited in page 1.)
- [134] S. Milles, D. Mercadante, I. Aramburu, M. Jensen, N. Banterle, C. Koehler, S. Tyagi, J. Clarke, S. Shammas, M. Blackledge, F. Gräter, and E. Lemke. Plasticity of an ultra-fast interaction between nucleoporins and nuclear transport receptors. *Cell*, 163(3):734–745, 2015. (Cited in page 10.)
- [135] S. Milles, N. Salvi, M. Blackledge, and M. R. Jensen. Characterization of intrinsically disordered proteins and their dynamic complexes: From in vitro to cell-like environments. *Prog Nucl Magn Reson Spectrosc*, 109:79 – 100, 2018. (Cited in page 3.)
- [136] M. Miskei, A. Horvath, M. Vendruscolo, and M. Fuxreiter. Sequence-based prediction of fuzzy protein interactions. *J Mol Biol*, 432(7):2289 – 2303, 2020. (Cited in page 3.)
- [137] T. Mittag, J. Marsh, A. Grishaev, S. Orlicky, H. Lin, F. Sicheri, M. Tyers, and J. D. Forman-Kay. Structure/function implications in a dynamic complex of the intrinsically disordered Sic1 with the Cdc4 subunit of an SCF ubiquitin ligase. *Structure*, 18:494–506, 2010. (Cited in page 10.)

- [138] T. Mittag, S. Orlicky, W.-Y. Choy, X. Tang, H. Lin, F. Sicheri, L. E. Kay, M. Tyers, and J. D. Forman-Kay. Dynamic equilibrium engagement of a polyvalent ligand with a single-site receptor. *Proc Natl Acad Sci USA*, 105(46):17772–17777, 2008. (Cited in page 10.)
- [139] A. Mittal, A. S. Holehouse, M. C. Cohan, and R. V. Pappu. Sequence-to-conformation relationships of disordered regions tethered to folded domains of proteins. *J Mol Biol*, 430(16):2403 – 2421, 2018. (Cited in page 6.)
- [140] A. Mittal, N. Lyle, T. S. Harmon, and R. V. Pappu. Hamiltonian switch metropolis monte carlo simulations for improved conformational sampling of intrinsically disordered regions tethered to ordered domains of proteins. *J Chem Theory Comput*, 10(8):3550–3562, 2014. (Cited in page 6.)
- [141] Z. Monahan, V. H. Ryan, A. M. Janke, K. A. Burke, S. N. Rhoads, G. H. Zerze, R. O’Meally, G. L. Dignon, A. E. Conicella, W. Zheng, R. B. Best, R. N. Cole, J. Mittal, F. Shewmaker, and N. L. Fawzi. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *EMBO J*, 36(20):2951–2967, 2017. (Cited in page 12.)
- [142] M. Nagulapalli, G. Parigi, J. Yuan, J. Gsponer, G. Deraos, V. V. Bamm, G. Harauz, J. Matsoukas, M. R. R. de Planque, I. P. Gerotheranassis, M. M. Babu, C. Luchinat, and A. G. Tzakos. Recognition pliability is coupled to structural heterogeneity: A calmodulin intrinsically disordered binding region complex. *Structure*, 20(3):522–533, 2012. (Cited in page 5.)
- [143] S. Naudi-Fabra, M. Tengo, M. R. Jensen, M. Blackledge, and S. Milles. Quantitative description of intrinsically disordered proteins using single-molecule fret, nmr, and saxs. *Journal of the American Chemical Society*, 143(48):20109–20121, 2021. PMID: 34817999. (Cited in page 3.)
- [144] V. Nguemaha and H.-X. Zhou. Liquid-liquid phase separation of patchy particles illuminates diverse effects of regulatory components on protein droplet formation. *Sci Rep*, 8(1):6728, 2018. (Cited in page 12.)
- [145] J. T. Nielsen and F. A. A. Mulder. Quality and bias of protein disorder predictors. *Sci Rep*, 9:5137, 2019. (Cited in page 1.)
- [146] G. Nodet, L. Salmon, V. Ozenne, S. Meier, M. R. Jensen, and M. Blackledge. Quantitative description of backbone conformational sampling of unfolded proteins at amino acid resolution from nmr residual dipolar couplings. *J Am Chem Soc*, 131(49):17908–17918, 2009. (Cited in page 5.)
- [147] C. J. Oldfield and A. K. Dunker. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*, 83(1):553–584, 2014. (Cited in page 1.)
- [148] J. G. Olsen, K. Teilum, and B. B. Kragelund. Behaviour of intrinsically disordered proteins in protein-protein complexes with an emphasis on fuzziness. *Cell Mol Life Sci*, 74:3175–3183, 2017. (Cited in pages 3 and 10.)
- [149] V. Ozenne, F. Bauer, L. Salmon, J.-r. Huang, M. R. Jensen, S. Segard, P. Bernadó, C. Charavay, and M. Blackledge. Flexible-meccano: a tool for the generation of explicit

- ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. *Bioinformatics*, 28(11):1463–1470, 05 2012. (Cited in pages 5, 11, 14, 34, 49, and 106.)
- [150] M. Paloni, R. Bailly, L. Ciandrini, and A. Barducci. Unraveling molecular interactions in liquid–liquid phase separation of disordered proteins by atomistic simulations. *J Phys Chem B*, 124(41):9009–9016, 2020. (Cited in page 13.)
- [151] F. Pan, Y. Zhang, C.-C. Lo, A. Mandal, X. Liu, and J. Zhang. Protein loop modeling and refinement using deep learning models. *bioRxiv*, 2021. (Cited in pages 13, 16, and 108.)
- [152] E. Papaleo, G. Saladino, M. Lambrughi, K. Lindorff-Larsen, F. L. Gervasio, and R. Nussinov. The role of protein loops and linkers in conformational dynamics and allostery. *Chem Rev*, 116(11):6391–6423, 2016. (Cited in pages 6 and 11.)
- [153] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pages 2778–2787. PMLR, 2017. (Cited in page 93.)
- [154] F. Paul, F. Noé, and T. R. Weikl. Identifying conformational-selection and induced-fit aspects in the binding-induced folding of PMI from Markov state modeling of atomistic simulations. *J Phys Chem B*, 122(21):5649–5656, 2018. (Cited in page 9.)
- [155] C. M. Payne, M. G. Resch, L. Chen, M. F. Crowley, M. E. Himmel, L. E. Taylor, M. Sandgren, J. Ståhlberg, I. Stals, Z. Tan, and G. T. Beckham. Glycosylated linkers in multimodular lignocellulose-degrading enzymes dynamically bind to cellulose. *Proc Natl Acad Sci USA*, 110(36):14646–14651, 2013. (Cited in page 7.)
- [156] I. Peran and T. Mittag. Molecular structure in biomolecular condensates. *Current opinion in structural biology*, 60:17–26, 2020. (Cited in pages 14 and 105.)
- [157] L. X. Peterson, A. Roy, C. Christoffer, G. Terashi, and D. Kihara. Modeling disordered protein interactions from biophysical principles. *PLoS Comput Biol*, 13(4):1–28, 04 2017. (Cited in pages 4 and 8.)
- [158] S. Piana, A. G. Donchev, P. Robustelli, and D. E. Shaw. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *J Phys Chem B*, 119(16):5113–5123, 2015. (Cited in page 4.)
- [159] S. Piana and A. Laio. A bias-exchange approach to protein folding. *J Phys Chem B*, 111(17):4553–4559, 2007. (Cited in page 4.)
- [160] S. Piana, P. Robustelli, D. Tan, S. Chen, and D. E. Shaw. Development of a force field for the simulation of single-chain proteins and protein-protein complexes. *J Chem Theory Comput*, 16(4):2494–2507, 2020. (Cited in page 4.)
- [161] L. M. Pietrek, L. S. Stelzl, and G. Hummer. Hierarchical ensembles of intrinsically disordered proteins at atomic resolution in molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 16(1):725–737, 2020. PMID: 31809054. (Cited in page 13.)

- [162] L. M. Pietrek, L. S. Stelzl, and G. Hummer. Structural ensembles of disordered proteins from hierarchical chain growth and simulation. *Current Opinion in Structural Biology*, 78:102501, 2023. (Cited in page 13.)
- [163] G.-C. Porusniuc. A comparative study on reinforcement learning methods for learning robot control behaviour in complex environments. Master’s thesis, Itä-Suomen yliopisto, 2023. (Cited in pages 16, 91, 92, and 107.)
- [164] S. Putta, L. Alvarez, S. Lüdtke, P. Sehr, G. A. Müller, S. M. Fernandez, S. Tripathi, J. Lewis, T. J. Gibson, L. B. Chemes, and S. M. Rubin. Structural basis for tunable affinity and specificity of lxcxe-dependent protein interactions with the retinoblastoma protein family. *Structure*, 30(9):1340–1353.e3, 2022. (Cited in pages 15, 70, and 107.)
- [165] R. Rangan, M. Bonomi, G. T. Heller, A. Cesari, G. Bussi, and M. Vendruscolo. Determination of structural ensembles of proteins: Restraining vs reweighting. *J Chem Theory Comput*, 14(12):6632–6641, 2018. (Cited in page 5.)
- [166] B. Raveh, N. London, L. Zimmerman, and O. Schueler-Furman. Rosetta FlexPepDock ab-initio: Simultaneous folding, docking and refinement of peptides onto their receptors. *PLoS One*, 6(4):1–10, 04 2011. (Cited in pages 4 and 8.)
- [167] E. Ravera, L. Sgheri, G. Parigi, and C. Luchinat. A critical assessment of methods to recover information from averaged data. *Phys Chem Chem Phys*, 18:5686–5701, 2016. (Cited in page 3.)
- [168] V. Receveur-Brechot and D. Durand. How random are intrinsically disordered proteins? a small angle scattering perspective. *Curr Protein Pept Sci*, 13(1):55–75, 2012. (Cited in page 3.)
- [169] V. P. Reddy Chichili, V. Kumar, and J. Sivaraman. Linkers in the structural biology of protein-protein interactions. *Protein Sci*, 22(2):153–167, 2013. (Cited in page 5.)
- [170] R. M. Regy, G. L. Dignon, W. Zheng, Y. C. Kim, and J. Mittal. Sequence dependent phase separation of protein-polynucleotide mixtures elucidated using molecular simulations. *Nucleic Acids Res*, 48(22):12593–12603, 12 2020. (Cited in page 12.)
- [171] R. M. Regy, J. Thompson, Y. C. Kim, and J. Mittal. Improved coarse-grained model for studying sequence dependent phase separation of disordered proteins. *Protein Science*, 30(7):1371–1379, 2021. (Cited in pages 46, 77, and 102.)
- [172] P. Robustelli, S. Piana, and D. E. Shaw. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci USA*, 115(21):E4758–E4766, 2018. (Cited in page 4.)
- [173] P. Robustelli, S. Piana, D. E. Shaw, and D. E. Shaw. Mechanism of coupled folding-upon-binding of an intrinsically disordered protein. *J Am Chem Soc*, 142(25):11092–11101, 2020. (Cited in page 9.)
- [174] N. Rochel, F. Ciesielski, J. Godet, E. Moman, M. Roessle, C. Peluso-Iltis, M. Moulin, M. Haertlein, P. Callow, Y. Mély, D. I. Svergun, and M. Dino. Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. *Nat Struct Mol Biol*, 18:564–570, 2011. (Cited in page 11.)

- [175] P. Romero, Z. Obradovic, X. Li, E. C. Garner, C. J. Brown, and A. K. Dunker. Sequence complexity of disordered protein. *Proteins*, 42(1):38–48, 2001. (Cited in page 1.)
- [176] B. Roux and J. Weare. On the statistical equivalence of restrained-ensemble simulations with the maximum entropy method. *J Chem Phys*, 138(8):084107, 2013. (Cited in page 5.)
- [177] K. M. Ruff and R. V. Pappu. AlphaFold and implications for intrinsically disordered proteins. *Journal of molecular biology*, 433(20):167208, 2021. (Cited in page 103.)
- [178] K. M. Ruff, R. V. Pappu, and A. S. Holehouse. Conformational preferences and phase behavior of intrinsically disordered low complexity sequences: insights from multiscale simulations. *Curr Opin Struct Biol*, 56:1 – 10, 2019. (Cited in pages 3 and 12.)
- [179] D. M. Ruiz, V. R. Turowski, and M. T. Murakami. Effects of the linker region on the structure and function of modular gh5 cellulases. *Sci Rep*, 6:28504, June 2016. (Cited in page 7.)
- [180] N. Salvi, A. Abyzov, and M. Blackledge. Multi-timescale dynamics in intrinsically disordered proteins from NMR relaxation and molecular simulation. *J Phys Chem Lett*, 7(13):2483–2489, 2016. (Cited in page 5.)
- [181] D. W. Sammond, C. M. Payne, R. Brunecky, M. E. Himmel, M. F. Crowley, and G. T. Beckham. Cellulase linkers are optimized based on domain type and function: Insights from sequence analysis, biophysical measurements, and molecular simulation. *PloS One*, 7(11):1–14, 11 2012. (Cited in page 7.)
- [182] L. Senicourt, A. le Maire, F. Allemand, J. E. Carvalho, L. Guee, P. Germain, M. Schubert, P. Bernadó, W. Bourguet, and N. Sibille. Structural insights into the interaction of the intrinsically disordered co-activator TIF2 with retinoic acid receptor heterodimer (RXR/RAR). *J Mol Biol*, page 166899, 2021. (Cited in page 11.)
- [183] Y. Seo, L. Chen, J. Shin, H. Lee, P. Abbeel, and K. Lee. State entropy maximization with random encoders for efficient exploration, 2021. (Cited in pages 16, 91, and 107.)
- [184] J.-E. Shea, R. B. Best, and J. Mittal. Physics-based computational and theoretical approaches to intrinsically disordered proteins. *Curr Opin Struct Biol*, 67:219–225, 2021. (Cited in pages 3 and 12.)
- [185] A. Shehu and L. E. Kaviraki. Modeling structures and motions of loops in protein molecules. *Entropy*, 14(12):252–290, 2012. (Cited in page 11.)
- [186] Y. Shin and C. P. Brangwynne. Liquid phase condensation in cell physiology and disease. *Science*, 357(6357), 2017. (Cited in pages 3 and 12.)
- [187] U. R. Shrestha, J. C. Smith, and L. Petridis. Full structural ensembles of intrinsically disordered proteins from unbiased molecular dynamics simulations. *Commun Biol*, 4:243, 2021. (Cited in page 4.)
- [188] J. Skolnick, M. Gao, H. Zhou, and S. Singh. AlphaFold 2: why it works and its implications for understanding the relationships of protein sequence, structure, and function. *Journal of chemical information and modeling*, 61(10):4827–4831, 2021. (Cited in page 13.)

- [189] D. Song, R. Luo, and H.-F. Chen. The IDP-specific force field ff14IDPSFF improves the conformer sampling of intrinsically disordered proteins. *J Chem Inf Model*, 57(5):1166–1178, 2017. PMID: 28448138. (Cited in page 4.)
- [190] A. Sottini, A. Borgia, M. Borgia, K. Bugge, D. Nettels, A. Chowdhury, P. Heidarsson, F. Zosel, R. Best, B. B. Kragelund, and B. Schuler. Polyelectrolyte interactions enable rapid association and dissociation in high-affinity disordered protein complexes. *Nat Commun*, 11:5736, 11 2020. (Cited in page 12.)
- [191] L. S. Stelzl, L. M. Pietrek, A. Holla, J. Oroz, M. Sikora, J. Köfinger, B. Schuler, M. Zweckstetter, and G. Hummer. Global structure of the intrinsically disordered protein tau emerges from its local structure. *JACS Au*, 2(3):673–686, 2022. (Cited in page 13.)
- [192] J. E. Stone, J. C. Phillips, P. L. Freddolino, D. J. Hardy, L. G. Trabuco, and K. Schulten. Accelerating molecular modeling applications with graphics processors. *J Comput Chem*, 28(16):2618–2640, 2007. (Cited in page 4.)
- [193] Y. Sugita and Y. Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chem Phys Lett*, 314(1):141–151, 1999. (Cited in page 4.)
- [194] R. S. Sutton and B. Tanner. Temporal-difference networks. *Advances in neural information processing systems*, 17, 2004. (Cited in page 92.)
- [195] Szabo, Horvath, Schad, Murvai, Tantos, Kalmar, Chemes, Han, and Tompa. Intrinsically disordered linkers impart processivity on enzymes by spatial confinement of binding domains. *Int J Mol Sci*, 20(9):2119, 2019. (Cited in page 7.)
- [196] K. Tang, S. W. Wong, J. S. Liu, J. Zhang, and J. Liang. Conformational sampling and structure prediction of multiple interacting loops in soluble and β -barrel membrane proteins using multi-loop distance-guided chain-growth Monte Carlo method. *Bioinformatics*, 31(16):2646–2652, 04 2015. (Cited in pages 13 and 16.)
- [197] X. Tang, S. Orlicky, T. Mittag, V. Csizmok, T. Pawson, J. D. Forman-Kay, F. Sicheri, and M. Tyers. Composite low affinity interactions dictate recognition of the cyclin-dependent kinase inhibitor Sic1 by the SCFCdc4 ubiquitin ligase. *Proc Natl Acad Sci USA*, 109(9):3287–3292, 2012. (Cited in page 10.)
- [198] J. M. C. Teixeira, Z. H. Liu, A. Namini, J. Li, R. M. Vernon, M. Krzeminski, A. A. Shamandy, O. Zhang, M. Haghighatlari, L. Yu, T. Head-Gordon, and J. D. Forman-Kay. Idpconformergenerator: A flexible software suite for sampling the conformational space of disordered protein states. *The Journal of Physical Chemistry A*, 126(35):5985–6003, 2022. PMID: 36030416. (Cited in pages 13, 14, 49, and 106.)
- [199] G. Tesei and K. Lindorff-Larsen. Improved predictions of phase behaviour of intrinsically disordered proteins by tuning the interaction range. *bioRxiv*, 2022. (Cited in pages 4, 81, and 102.)
- [200] G. Tesei, T. K. Schulze, R. Crehuet, and K. Lindorff-Larsen. Accurate model of liquid-liquid phase behavior of intrinsically disordered proteins from optimization of single-chain properties. *Proceedings of the National Academy of Sciences*, 118(44):e2111696118, 2021. (Cited in page 4.)

- [201] G. Tesei, A. I. Trolle, N. Jonsson, J. Betz, F. E. Knudsen, F. Pesce, K. E. Johansson, and K. Lindorff-Larsen. Conformational ensembles of the human intrinsically disordered proteome. *Nature*, 626(8000):897–904, 2024. (Cited in page 4.)
- [202] P. Tompa. The interplay between structure and function in intrinsically unstructured proteins. *FEBS Lett*, 579(15):3346–3354, 2005. (Cited in page 2.)
- [203] P. Tompa, N. Davey, T. Gibson, and M. Babu. A million peptide motifs for the molecular biologist. *Mol Cell*, 55(2):161–169, 2014. (Cited in page 3.)
- [204] P. Tompa, E. Schad, A. Tantos, and L. Kalmar. Intrinsically disordered proteins: emerging interaction specialists. *Curr Opin Struct Biol*, 35:49 – 59, 2015. (Cited in page 2.)
- [205] G. Tria, H. D. T. Mertens, M. Kachala, and D. I. Svergun. Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ*, 2(2):207–217, Mar 2015. (Cited in page 5.)
- [206] V. N. Uversky. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*, 11:739–756, 2002. (Cited in page 1.)
- [207] V. N. Uversky, C. J. Oldfield, and A. K. Dunker. Intrinsically disordered proteins in human diseases: Introducing the d2 concept. *Annu Rev Biophys*, 37(1):215–246, 2008. (Cited in page 3.)
- [208] R. van der Lee, M. Buljan, B. Lang, R. J. Weatheritt, G. W. Daughdrill, A. K. Dunker, M. Fuxreiter, J. Gough, J. Gsponer, D. T. Jones, P. M. Kim, R. W. Kriwacki, C. J. Oldfield, R. V. Pappu, P. Tompa, V. N. Uversky, P. E. Wright, and M. M. Babu. Classification of intrinsically disordered regions and proteins. *Chem Rev*, 114(13):6589–6631, 2014. (Cited in pages 1 and 7.)
- [209] K. Van Roey, B. Uyar, R. J. Weatheritt, H. Dinkel, M. Seiler, A. Budd, T. J. Gibson, and N. E. Davey. Short linear motifs: Ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev*, 114(13):6733–6778, 2014. (Cited in page 3.)
- [210] M. van Rosmalen, M. Krom, and M. Merks. Tuning the flexibility of glycine-serine linkers to allow rational design of multidomain proteins. *Biochemistry*, 56(50):6565–6574, 2017. PMID: 29168376. (Cited in pages 65 and 88.)
- [211] M. Varadi, S. Anyango, M. Deshpande, S. Nair, C. Natassia, G. Yordanova, D. Yuan, O. Stroe, G. Wood, A. Laydon, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research*, 50(D1):D439–D444, 2022. (Cited in page 102.)
- [212] S. I. Virtanen, A. M. Kiiirikki, K. M. Mikula, H. Iwai, and O. H. S. Ollila. Heterogeneous dynamics in partially disordered proteins. *Phys Chem Chem Phys*, 22:21185–21196, 2020. (Cited in pages 4 and 6.)
- [213] A. Vitalis and R. V. Pappu. ABSINTH: A new continuum solvation model for simulations of polypeptides in aqueous solutions. *J Comput Chem*, 30(5):673–699, 2009. (Cited in page 4.)

- [214] E. von Castelmur, M. Marino, D. I. Svergun, L. Kreplak, Z. Ucurum-Fotiadis, P. V. Konarev, A. G. Urzhumtsev, D. Labeit, S. Labeit, and O. Mayans. A regular pattern of ig super-motifs defines segmental flexibility as the elastic mechanism of the titin chain. *Proceedings of the National Academy of Sciences*, 105:1186 – 1191, 2008. (Cited in page 49.)
- [215] I. von Ossowski, J. T. Eaton, M. Czjzek, S. J. Perkins, T. P. Frandsen, M. SchÄElein, P. Panine, B. Henrissat, and V. Receveur-Bréchet. Protein disorder: Conformational distribution of the flexible linker in a chimeric double cellulase. *Biophys J*, 88(4):2823 – 2832, 2005. (Cited in page 7.)
- [216] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic. Flavors of protein disorder. *Proteins*, 52(4):573–584, 2003. (Cited in page 1.)
- [217] H. Wang, M. Emmerich, and A. Plaat. Monte carlo q-learning for general game playing. *arXiv preprint arXiv:1802.05944*, 2018. (Cited in page 92.)
- [218] J. Wang, J.-M. Choi, A. S. Holehouse, H. O. Lee, X. Zhang, M. Jahnel, S. Maharana, R. Lemaitre, A. Pozniakovsky, D. Drechsel, I. Poser, R. V. Pappu, S. Alberti, and A. A. Hyman. A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, 174(3):688–699.e16, 2018. (Cited in page 12.)
- [219] W. Wang and D. Wang. Extreme fuzziness: Direct interactions between two IDPs. *Biomolecules*, 9(3), 2019. (Cited in page 11.)
- [220] P. E. Wright and H. J. Dyson. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*, 16:18–29, 2015. (Cited in page 1.)
- [221] H. Wu. Higher-order assemblies in a new paradigm of signal transduction. *Cell*, 153:287–292, 2013. (Cited in page 8.)
- [222] H. Wu, P. G. Wolynes, and G. A. Papoian. AWSEM-IDP: A coarse-grained force field for intrinsically disordered proteins. *J Phys Chem B*, 122(49):11115–11125, 2018. (Cited in page 4.)
- [223] S. Wu, D. Wang, J. Liu, Y. Feng, J. Weng, Y. Li, X. Gao, J. Liu, and W. Wang. The dynamic multisite interactions between two intrinsically disordered proteins. *Angew Chem Int Ed*, 56(26):7515–7519, 2017. (Cited in page 12.)
- [224] H. Xie, S. Vucetic, L. M. Iakoucheva, C. J. Oldfield, A. K. Dunker, V. N. Uversky, and Z. Obradovic. Functional anthology of intrinsic disorder. 1. biological processes and functions of proteins with long disordered regions. *J Proteome Res*, 6(5):1882–1898, 2007. (Cited in page 1.)
- [225] J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, and Y. Zhang. The i-tasser suite: protein structure and function prediction. *Nature methods*, 12(1):7–8, 2015. (Cited in page 81.)
- [226] Y. I. Yang, Q. Shao, J. Zhang, L. Yang, and Y. Q. Gao. Enhanced sampling in molecular dynamics. *J Chem Phys*, 151(7):070902, 2019. (Cited in page 4.)
- [227] B. W. Zhang, D. Jasnow, and D. M. Zuckerman. Efficient and verified simulation of a path ensemble for conformational change in a united-residue model of calmodulin. *Proc Natl Acad Sci USA*, 104(46):18043–18048, 2007. (Cited in page 4.)

- [228] O. Zhang, M. Haghightalari, J. Li, Z. H. Liu, A. Namini, J. M. C. Teixeira, J. D. Forman-Kay, and T. Head-Gordon. Learning to evolve structural ensembles of unfolded and disordered proteins using experimental solution data. *The Journal of Chemical Physics*, 158(17), may 2023. (Cited in pages 13, 16, and 107.)
- [229] W. Zheng, G. L. Dignon, N. Jovic, X. Xu, R. M. Regy, N. L. Fawzi, Y. C. Kim, R. B. Best, and J. Mittal. Molecular details of protein condensates probed by microsecond long atomistic simulations. *J Phys Chem B*, 124(51):11671–11679, 2020. PMID: 33302617. (Cited in page 13.)
- [230] G. Zhou, G. A. Pantelopulos, S. Mukherjee, and V. A. Voelz. Bridging microscopic and macroscopic mechanisms of p53-MDM2 binding with kinetic network models. *Biophys J*, 113(4):785–793, 2017. (Cited in page 9.)
- [231] H.-X. Zhou. Quantitative account of the enhanced affinity of two linked scfvs specific for different epitopes on the same antigen. *J Mol Biol*, 329(1):1–8, 2003. (Cited in page 6.)
- [232] H.-X. Zhou. Polymer models of protein stability, folding, and interactions. *Biochemistry*, 43(8):2141–2154, 2004. PMID: 14979710. (Cited in page 69.)
- [233] J. Zhu, Z. Li, H. Tong, Z. Lu, N. Zhang, T. Wei, and H.-F. Chen. Phanto-IDP: compact model for precise intrinsically disordered protein backbone generation and enhanced sampling. *Briefings in Bioinformatics*, 25(1):bbad429, 11 2023. (Cited in pages 13, 16, and 107.)
- [234] J. Zhu, Z. Li, B. Zhang, Z. Zheng, B. Zhong, J. Bai, T. Wang, T. Wei, J. Yang, and H.-F. Chen. Precise generation of conformational ensembles for intrinsically disordered proteins using fine-tuned diffusion models. *bioRxiv*, 2024. (Cited in pages 13, 16, and 107.)
- [235] D. Zuckerman. *Statistical Physics of Biomolecules: An Introduction*. CRC Press, 2010. (Cited in page 59.)