Développement de nouvelles approches protéo-chimiométriques appliquées à l'étude des interactions et de la sélectivité des inhibiteurs de kinases

Nicolas Bosc

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE SANTE, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

LABORATOIRE ICOA

THÈSE présentée par :

Nicolas Bosc

soutenue le : **20 novembre 2015**

pour obtenir le grade de : **Docteur de l'Université d'Orléans**

Discipline/ Spécialité : Chimie / Bioinformatique structurale et Chémoinformatique

Développement de nouvelles approches protéo-chimiométriques appliquées à l'étude des interactions et de la sélectivité des inhibiteurs de kinases

Directeur de thèse : **Prof. Pascal Bonnet**
Co-encadrant : **Dr. Christophe Meyer**

JURY

Prof. Bernard Offmann	UFIP, Université de Nantes	**Président**
Prof. Anne-Claude Camproux	MTi, Université Paris Diderot	**Rapporteur**
Dr. Philippe Roche	iSCB, Université Aix-Marseille	**Rapporteur**
Dr. Alexandre De Brevern	INSERM, Université Paris Diderot	**Examinateur**
Dr. Christophe Meyer	Janssen-Cilag	**Co-encadrant**
Prof. Pascal Bonnet	ICOA, Université d'Orléans	**Directeur**

« You always pass failure on your way to success. »

Mickey Rooney

« All models are wrong... but some are useful. »

George E. P. Box

Remerciements

Je tiens à exprimer mes sincères remerciements à la région Centre-Val de Loire et à l'entreprise Janssen, dont les financements m'ont permis de mener à bien ce travail de thèse. Merci également au Professeur Luc Morin-Allory d'avoir participé à l'élaboration du projet de recherche.

Il aurait été bien difficile de finir cette thèse sans le concours de bien des personnes, qu'il est maintenant temps de remercier.

Tout d'abord, je remercie les membres de mon jury d'avoir accepté de lire mon manuscrit et d'évaluer mon travail : le Professeur Anne-Claude Camproux, le Docteur Philippe Roche, le Docteur Alexandre De Brevern et le Professeur Bernard Offmann. Merci aussi d'avoir pris le temps de vous déplacer jusqu'à Orléans.

Un grand merci au Professeur Pascal Bonnet de m'avoir choisi pour l'accompagner dans sa première direction de thèse à l'Université d'Orléans. C'était un choix risqué pour toi, comme pour moi, mais je pense que le résultat en valait la peine. Merci de m'avoir transmis tes connaissances sur les protéines kinases. J'en suis arrivé à un point où je les déteste, sans pour autant cesser d'en être fasciné. Merci d'avoir été présent tout au long de ces trois années, d'avoir répondu à mes questions et de m'avoir soutenu. Merci.

Je remercie également le Docteur Christophe Meyer avec qui j'ai travaillé principalement à distance, mais qui a su se rendre disponible dans les moments clés. Merci pour tes conseils et merci d'avoir œuvré pour je puisse aller effectuer des stages chez Janssen.

Je remercie le Professeur Olivier Martin de m'avoir accepté au sein de l'Institut de Chimie Organique et Analytique.

Merci à tous les membres de l'équipe du groupe de Bioinformatique Structurale & Chémoinformatique pour l'aide et le soutien que vous avez pu m'apporter. Merci Stéphane et Samia pour votre aide à différents niveaux scientifiques, pour vos conseils avisés et pour les discussions en tout genre sur la vie. Merci pour votre franchise et votre humour.

Merci aux doctorants pour ces bons moments passés attablés autour d'un bon repas universitaire, ou autour d'un bon verre ! Merci José-Manuel de m'avoir accompagné depuis ton stage et pour m'avoir montré que ma seule présence pouvait parfois résoudre tes problèmes. Je te remercie d'avoir partagé avec nous tes goûts pour le cinéma (promis, je vais essayer de ne me souvenir que des bons films). Merci Baptiste pour ta banane quotidienne et pour avoir embrassé la voie de ~~modéli~~modélo, alors que la chimie analytique t'attendait les bras ouverts ! Merci Abdennour pour ton humour, tes répliques cultes et ton

soutien inconditionnel à Nicolas Cage, malgré ses erreurs. Merci Sonia d'avoir partagé avec nous certaines de tes histoires que j'ai beaucoup aimé écouter et merci pour ta franchise. Je regrette que tu sois arrivée si tard dans l'équipe. Enfin merci Fabrice de reprendre le flambeau de la PCM dans l'équipe et bon courage pour l'amener au-delà des limites (tout en restant de le domaine d'applicabilité, bien entendu). Merci à tous pour votre gentillesse et bonne chance pour la suite de votre thèse.

Merci aux post-docs qui sont passés par l'équipe au cours de ces trois dernières années, et notamment Emilie pour ses coups de gueules et Jade pour sa bonne humeur contagieuse. Un grand merci à Felipe pour son aide avec R et pour m'avoir accompagné un temps à l'aviron. Je me souviendrai que nous avons failli tomber à l'eau, mais heureusement, seulement failli. Merci également à Rohit.

Merci à tous les stagiaires qui nous ont rejoints : Stéphanie, Alexandre, Mandy, Alan et Emeline. Certaines de nos discussions resteront gravées dans ma mémoire.

Aussi, je n'oublie pas l'aide que m'ont apporté le Professeur Christel Vrain ainsi que le Docteur Lionel Martin, du Laboratoire d'Informatique Fondamentale d'Orléans, au sujet des méthodes d'apprentissages automatisées et notamment pour les machines à vecteurs de support. Merci à vous.

Mes sincères remerciements, également, à Janssen-Cilag et Janssen Pharmaceutica, et notamment aux chercheurs m'ayant accueilli à Beerse : Berthold Wroblowski, Edgar Jacoby, Gary Tresadern et Herman Van Vlijmen. Merci à Berthold pour ses idées et merci à Edgar pour m'avoir permis de tester mes modèles de manière expérimentale. Merci à Benny pour m'avoir permis de perdre le moins de temps possible à installer mes machines.

Merci à ceux qui ont œuvré dans l'ombre pour que je ne puisse me consacrer qu'à la recherche. Merci à Virginie Loisel et Diane Verkuringen pour avoir organisé mes déplacements à Beerse. Un énorme merci à Marie-Madeleine et Yann pour toute l'aide administrative que vous m'avez apportée et pour votre gentillesse. Merci à Alain-Michel pour avoir fait en sorte, entre autres, que nos bureaux restent le plus accueillant possible.

Je remercie, sans pouvoir toutes les nommer, les personnes qui m'ont donné goût à la science et à la recherche pharmaceutique. Parmi elles se trouvent des enseignants, des professeurs et ces personnes que j'ai côtoyées au cours de mes stages.

Merci à Chantal et Edgard pour m'avoir accueilli quand j'en avais besoin, pour ces bons moments passés en votre compagnie et pour m'avoir gâté, culinairement parlant.

Pour finir, il y a mes amis et ma famille. Il aurait été bien difficile d'en arriver là aujourd'hui, sans le soutien inconditionnel de mes parents, mais également sans le support sans faille de mon frère. Quelle joie que la famille se soit agrandit au cours de ces trois années !

Et pour vraiment finir, merci à toi Aline, pour m’avoir supporté, pour ton aide, pour avoir été là, pour tout.

Table des matières

Liste des abréviations

ABL1 : *Abelson tyrosine-protein kinase 1*

ACP : Analyse en composantes principales

AD : Maladie d'Alzheimer

ADN : Acide Désoxyribose Nucléique

ALH : Accepteur de Liaison Hydrogène

ATP: Adénosine-5'-Triphosphate

BMS : Bristol-Myers Squibb

BRAF : *Serine/threonine-protein kinase B-raf*

CADD : *Computer-aided Drug Design /* Drug Design assisté par ordinateur

CDK : *Cyclin-Dependent Kinase*

CML : Leucémie myéloïde chronique

CoMFA : *Comparative molecular field analysis*

CoMSIA : *Comparative molecular similarity indices analysis*

CTD : Composition, Transition, Distribution

CV : validation croisée

DDN : Demande de Drogue Nouvelle

DLH : Donneur de Liaison Hydrogène

EBI : Institut Européen de Bioinformatique

FDA : *Food and Drug Administration*

FN : Faux négatif

FP : Faux positif

GSK : GlaxoSmithKline

IC50 : concentration inhibitrice médiane

IGF1R : *Insulin-like Growth Factor 1 Receptor*

InChi : *International Chemical Identifier*

INSR : *Insulin receptor*

IUPAC : *International Union of Pure and Applied Chemistry* / Union internationale de chimie pure et appliquée

JAK : *Janus Kinase*

Kd : constante de dissociation

Ki : constante d'inhibition

LMO : *Leave-Multiple-out*

LOO : *Leave-One-Out*

mCRC : Cancer colorectal métastasique

mNSCLC : Cancer métastasique non à petites cellules du poumon

MAPK : *Mitogen-Activated Protein Kinase*

MTC : Cancer médullaire de la thyroïde

NME : *New Molecular Entity* / Nouvelle Entité Moléculaire

NR : *Nuclear receptor*

PCM : Protéo-chimiométrique

PDB : Protein Data Bank

PLS : *Partial Least Squares* / Moindres Carrés Partiels

pKd : -log(Kd)

pKi : -log(Ki)

PKI : Inhibiteurs de Protéines Kinases

PI3K : Phosphatidylinositol 4,5-bisphosphate 3-kinase

QSAM : *Quantitative Sequence-Activity Modelling /* Modélisation Quantitative Séquence à Activité

QSAR : *Quantitative Structure-Activity Relationship* / Relation Quantitative Structure à Activité

QSPR : *Quantitative Structure-Property Relationship* / Relation Quantitative Structure à Propriété

RBF : *Radial Basis Function*

RCPG : Récepteur couplé aux protéines G

RF : *Random Forest* / Forêt aléatoire

R^2 : coefficient de détermination

RMSE : Erreur quadratique moyenne

RCSB : *Research Collaboratory for Structural Bioinformatics*

σ : ecart-type

SH2 : *Src homology 2*

TFP : Taux de Faux Positifs

VN : Vrai Négatif

VP : Vrai Positif

Liste des figures

Liste des tables

I. Avant-propos

A. Présentation des lieux de travail

Cette thèse s'est déroulée au sein de l'Institut de Chimie Organique et Analytique (ICOA) de l'Université d'Orléans et est le résultat d'une collaboration entre l'Université d'Orléans et la société Janssen.

L'ICOA fut créé en 1995 après la fusion du Laboratoire de Chimie Bio-organique et Analytique et du Laboratoire de Biochimie Structurale, et est actuellement dirigé par le professeur Olivier Martin. Depuis peu, l'ICOA a intégré le Labex IRON et le Labex SynOrg. Ce dernier repose sur l'association de quatre laboratoires d'excellence en synthèse organique et bio-organique situés en Normandie et en région Centre-Val de Loire. Ces laboratoires ont pour objectif de devenir des acteurs majeurs dans le développement du premier bassin de recherche pharmaceutique en Europe.

La démarche scientifique de l'ICOA s'étend de la conception de nouvelles structures par modélisation moléculaire, à la synthèse de nouvelles molécules organiques (composés hétérocycliques, dérivés de sucres et analogues de nucléosides), en passant par l'extraction du milieu naturel (plantes) par les techniques séparatives les plus performantes et l'analyse par spectrométrie de masse, et à l'enzymologie, pour connaître les récepteurs de certaines molécules bioactives. C'est au sein de cet établissement qu'évolue le groupe de Bioinformatique Structurale et Chémoinformatique (SB&C), dirigé par le Professeur Pascal Bonnet. Le groupe s'est spécialisé dans l'étude des protéines kinases et s'intéresse plus précisément à la conception de nouveaux inhibiteurs, à la validation de cibles, à l'étude conformationnelle des protéines, à la réalisation de criblages virtuels, à la prédiction de paramètres pharmacocinétique et enfin à l'étude des interactions protéines kinases – ligands.

La société Janssen (Janssen-Cilag en France et Janssen Pharmaceutica en Belgique) appartient au groupe américain Johnson & Johnson, premier groupe de santé mondial. Janssen-Cilag découle de la fusion de Janssen et de Cilag en 2010, après que Cilag ait rejoint Johnson & Johnson en 1959. Janssen Pharmaceutica fut créé en 1953 par le docteur Paul Janssen. L'entreprise rejoindra Johnson & Johnson huit ans après. Les deux sociétés pharmaceutiques se sont spécialisées dans l'infectiologie, l'immunologie, les neurosciences, les maladies cardiovasculaires et métaboliques et en oncologie. C'est notamment dans ce dernier domaine que les protéines kinases jouent un rôle prépondérant.

B. Objectifs de la thèse

Cette thèse a pour objectif de présenter plusieurs approches computationnelles visant à étudier différents aspects des interactions entre les protéines kinases et leurs inhibiteurs.

Le chapitre II s'attache à passer en revue les bouleversements récents de l'industrie pharmaceutique ainsi que ses défis à venir. Nous y présenterons également la famille des protéines kinases, ainsi que les inhibiteurs de cette famille actuellement sur le marché. A la fin de ce chapitre, nous parlerons des approches statistiques focalisées sur les protéines kinases.

Dans le chapitre III, nous nous intéresserons à l'estimation de la sélectivité d'inhibiteurs de protéines kinases à partir de données expérimentales d'interaction. En introduisant de nouvelles métriques, nous chercherons à proposer des manières efficaces de sélectionner une ou plusieurs molécules d'un panel, selon leur profil d'activité. En comparant ces nouvelles méthodes à celles déjà existantes, nous démontrerons leur utilité dans la découverte de molécules bioactives et sélectives.

Dans le chapitre IV, nous nous focaliserons exclusivement sur les inhibiteurs dits de Type II. A l'aide d'une approche protéométrique innovante, nous chercherons à identifier les résidus importants des protéines kinases qui pourraient être impliqués dans la liaison de cette classe d'inhibiteur. Les positions de ces résidus seront identifiées à l'aide d'une sélection des variables effectuée sur l'alignement des séquences des protéines. Nous montrerons que le modèle obtenu permet, en outre, de prédire l'activité d'inhibiteurs de Type II sur des protéines kinases d'intérêt.

Enfin, dans le chapitre V, nous introduirons un nouveau descripteur tridimensionnel de protéines kinases. Celui-ci servira notamment à l'établissement de modèles protéo-chimiométriques (PCM), nommés ici kino-chimiométrique (KCM). Ces modèles reposeront sur des données publiques ou privées et s'appuieront sur une large portion du kinome humain. Nous verrons également au cours de ce chapitre les limites actuelles de la PCM et les défis qu'il lui reste à relever pour s'imposer dans la prédiction d'inhibiteurs affins et sélectifs de protéines kinases.

Au cours de cette thèse, il nous est apparu que ces méthodes, bien que différentes dans leur approche, proposent un intérêt certain dans la découverte de nouvelles molécules thérapeutiques appliquées aux protéines kinases.

II. Les outils statistiques comme prédicteurs d'inhibiteurs de protéines kinases

A. L'émergence de l'industrie pharmaceutique

Depuis des siècles, l'Homme n'a eu de cesse d'essayer de soulager les maux qui le tourmentent. Pouvant être causés par l'environnement extérieur ou bien par des défaillances internes, ils sont généralement à l'origine de grandes souffrances qui ont poussé les civilisations du monde entier à tenter de les guérir et ce, depuis bien longtemps. Le papyrus d'Ebers, daté du deuxième millénaire avant notre ère, serait l'un des premiers écrits décrivant plusieurs centaines de remèdes pour la plupart à base de plantes.[1] La médecine traditionnelle chinoise utilise depuis plus de deux millénaires des remèdes à base de plantes et/ou de toucher sur la peau afin de traiter les maladies.[2] Plus tard, et notamment avec la découverte des Amériques, des plantes vont être importées en Europe et être utilisées, dans certains cas, à des fins thérapeutiques (le quinquina, le coca…). Dès lors, il est notable que la médecine, au sens large, s'est longtemps contentée d'employer les substances végétales auxquelles elle avait accès, sans pour autant véritablement comprendre les mécanismes biologiques et chimiques impliqués. C'est seulement au début du XIX^{ème} siècle, avec les avancées de la chimie, que vont être extraites les premières molécules. En 1804, Friedrich Wilhelm Sertürner est le premier à publier des découvertes sur l'extraction de la morphine à partir de l'opium.[3] Les décennies suivantes ont vu de nouvelles molécules être isolées, ainsi que les premières molécules synthétisées par l'Homme. L'acide acétylsalicylique, synthétisé pour la première fois au milieu du XIXe siècle, sera commercialisé 50 ans plus tard sous le nom commercial aspirine.[4] Il connaitra un succès remarquable en tant que premier analgésique mis sur le marché par les laboratoires Bayer. Dès lors, de nombreuses entreprises basant leur économie sur le développement et la production de médicaments ont vu le jour.

Aujourd'hui, ce secteur économique pèse plusieurs centaines de milliards de dollars au niveau mondial (source : IMS Health 2014 « *Global Outlook for Medicines Through 2018* »). En comparant les années 2008 et 2013, les dépenses en santé sont passées de 800 à quasiment 1000 milliards, soit une augmentation de 25% en seulement 5 ans. De plus, les analystes tablent sur une augmentation encore plus grande entre 2013 et 2018 (Figure 1), du fait du prolongement de la durée de vie moyenne et d'un rebond au niveau de l'économie mondiale. Des chiffres qui, à eux seuls, permettent de comprendre l'intérêt des entreprises pharmaceutiques pour ce secteur d'activité.

---

[1] Bryan, C. P. (1974) Ancient Egyptian medicine: the Papyrus Ebers; *Ares Publishers*: Chicago.
[2] Xu, J., Yang, Y. (2009), Traditional Chinese medicine in the Chinese health care system, *Health Policy*, 90, 133-39.
[3] Meyer, K. (2004), Dem Morphin auf der Spur, *Pharmazeutischen Zeitung*.
[4] Gerhardt, C. (1853), Recherches sur les acides organiques anhydrides, *Annales de Chimie et de Physique*, 37, 285-342.

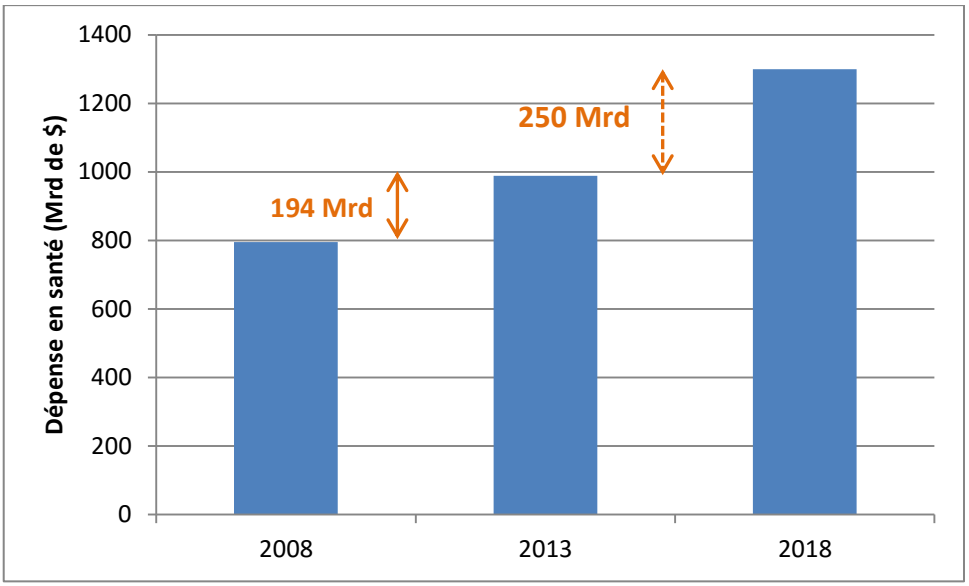


Figure 1 : Evolution des dépenses de santé au niveau mondial entre 2008 et 2018. Le montant annoncé pour 2018 est fondé sur des prédictions (source : IMS Health 2014 « *Global Outlook for Medicines Through 2018* »).

B. Les bouleversements de l'industrie pharmaceutique

Malgré cette bonne santé économique évidente, l'industrie pharmaceutique connait, depuis plusieurs années, des mutations majeures qui ont entrainé des modifications de son paysage. Le coût nécessaire à la recherche et au développement d'un nouveau médicament est souvent mis en avant par les entreprises pharmaceutiques pour justifier le prix des traitements. Les chiffres annoncés varient parfois du simple au double et il est difficile d'avoir une idée précise du coût réel de revient, même moyen, d'un médicament. La dernière étude en date annonce un chiffre ahurissant de 2,6 milliards de dollars.[5] Ce chiffre a rapidement été contesté par une grande partie des journaux scientifiques et économiques du fait de la méthodologie mise en place et de l'absence de données brutes.[6,7] En effet, l'étude n'est basée que sur un petit nombre d'entreprises pharmaceutiques et celles-ci n'ont pas rendu publique les données brutes des dépenses en R&D. Il est donc très difficile de vérifier ce qui est annoncé. Toutefois, un article plus ancien s'appuyant sur les données publiques disponibles avait estimé le coût de développement d'un nouveau médicament à 1 milliard de dollars.[8] A défaut de pouvoir estimer avec précision ce chiffre, il est irréfutable que le développement d'un nouveau médicament nécessite de mobiliser des ressources financières très importantes. En comptabilisant l'ensemble des dépenses en R&D des entreprises membres du PhRMA, ce montant passe de 15 milliards de dollars en 1995 à 50 milliards en 2013 (Figure 2). Les entreprises pharmaceutiques se placent ainsi parmi les entreprises investissant le plus en recherche.

[5] DiMasi, J. A. *et al.* (2015), The cost of drug development, *N. Engl. J. Med.*, 372, 1972.
[6] Avorn, J. (2015), The $2.6 Billion Pill — Methodologic and Policy Considerations, *N. Engl. J. Med.*, 372, 1877-79.
[7] The Economist. (29 novembre 2014), *The price of failure*.
[8] Adams, C. P., Brantner, V. V. (2010), Spending on new drug development, *Health Econ.*, 19, 130-41.

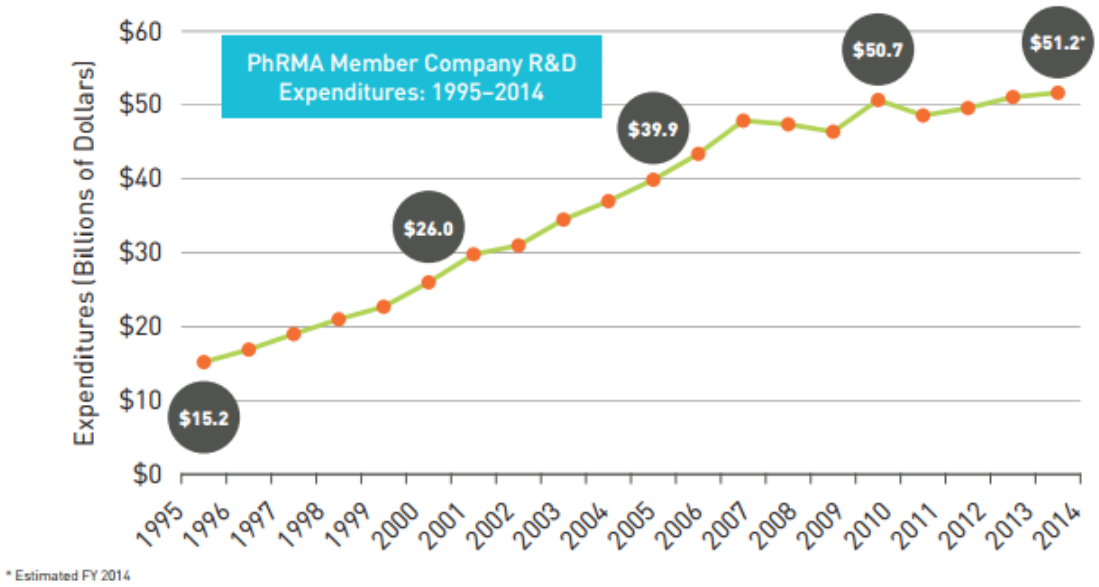


Figure 2 : Investissements en R&D des entreprises pharmaceutiques membres du PhRMA. Ceux-ci ont plus que triplé en moins de 20 ans (source : www.phrma.org/sites/default/files/pdf/2015_phrma_profile.pdf).

Les causes de cette augmentation sont multifactorielles. De manière générale, l'augmentation du coût de la vie est fortement corrélée à l'augmentation des salaires que les entreprises doivent intégrer dans leurs charges. Dans un cadre plus scientifique, les coûts liés à la justification de la non dangerosité des nouvelles entités moléculaires (NME) avant leur mise sur le marché ont fortement crû. Ce renforcement des règles de mise sur le marché est principalement dû aux rappels de plusieurs produits au début des années 2000.[9] Un des exemples marquants est le retrait mondial du Vioxx (Merck & Co's) en 2004 suite à de fortes présomptions de dangerosité, notamment au niveau cardiovasculaire.[10] Les études cliniques supplémentaires demandées par l'agence américaine du médicament (FDA) ont ainsi entrainé des retards importants de mise sur le marché, en plus des coûts inhérents à la recherche et au développement. En parallèle, la FDA a récemment décidé de modifier les règles sur lesquelles elle se base pour accepter une Demande de Drogue Nouvelle (DDN).[11] Dorénavant, l'institution s'attèle beaucoup à vérifier si la molécule en question présente un effet thérapeutique au moins équivalent à une molécule déjà présente sur le marché, avant de l'accepter.

De plus, les entreprises pharmaceutiques ont eu, et vont continuer, à faire face à l'expiration des brevets de nombreux *blockbusters* ou produits phares (terme désignant un médicament dont les revenus dépassent 1 milliard de dollars par an). Ces échéances entraînent systématiquement une chute vertigineuse

[9] Abou-Gharbia, M., Childers, W. E. (2014), Discovery of Innovative Therapeutics: Today's Realities and Tomorrow's Vision. 2. Pharma's Challenges and Their Commitment to Innovation, *J. Med. Chem.*, 57, 5525-53.
[10] FDA Public Health Advisory: Safety of Vioxx. (2014) http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106274.htm
[11] FDA's consideration of evidence from certain clinical trials. (2010) *United States Government Accountability Off ice*.

des revenus générés par les médicaments en question, comme le montre l'exemple du Lipitor® (atorvastatine) commercialisé par Pfizer, qui a vu son brevet expiré fin 2011. Cela a entraîné une chute de 50% des revenus générés par ce médicament en à peine 6 mois, du fait de l'arrivée sur le marché du générique de l'atorvastatine (Figure 3). Entre 2011 et 2014, 14 autres *blockbusters* ont ainsi vu leur brevet expiré ce qui, dans la plupart des cas, a conduit à une baisse très importante des revenus pour les entreprises pharmaceutiques concernées. Citons parmi ces médicaments, les exemples du Seroquel® d'AstraZeneca et du Plavix® de Sanofi.

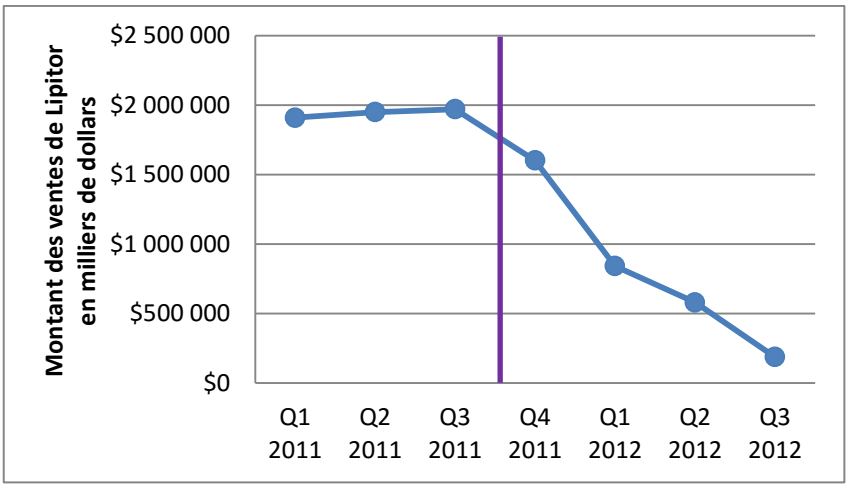


Figure 3 : Evolution des ventes de Lipitor (en milliers de dollars) aux Etats-Unis entre le premier trimestre 2011 et le troisième trimestre 2012. Le générique Atorvastatine a été introduit sur le marché américain à la fin du troisième trimestre 2011, symbolisée par un trait vertical violet (Inspiré d'après [8] | Source : http://www.drugs.com/stats/lipitor/).

Suite à ces bouleversements, il semble que les entreprises pharmaceutiques se soient adaptées à ces nouvelles contraintes. En témoigne le nombre de nouvelles entités moléculaires acceptées par la FDA ces dernières années qui est à un niveau similaire à celui de la fin des années 1990 (Figure 4). Toutefois, en regardant de plus près ces nouveaux médicaments mis sur le marché, nous remarquons qu'une part importante correspond à des médicaments orphelins, c'est-à-dire, des médicaments traitant des maladies rares (36% en moyenne entre 2011 et 2014). Une maladie rare est considérée comme n'affectant au plus qu'une personne sur 200000 aux Etats-Unis, contre une sur 2000 en Europe. Ces chiffres tendent à montrer que les entreprises pharmaceutiques s'adaptent et n'ignorent plus les maladies rares, comme cela leur est souvent reproché. D'ailleurs, cela représente plusieurs intérêts pour elles. D'une part, les ressources financières utilisées sont parfois moindres (essais cliniques sur de petites populations et d'une durée moindre, mesures législatives incitatives). D'autre part, elles peuvent vendre ces traitements à des prix élevés. Néanmoins, en raison du faible nombre de patients visé, ces produits n'atteignent pas les seuils de rentabilité des *blockbusters*.

De plus, certaines entreprises pharmaceutiques ont récemment effectué des découvertes majeures, notamment en immunothérapie, qui pourraient mener à de nouveaux *blockbusters*. Ainsi BMS et Merck & Co's, avec leurs anticorps monoclonaux respectifs, pembrolizumab et nivolumab, pourraient bien aider à sauver de nombreux patients, tout en dégageant chacun plus de 3 milliards de dollars en 2019.[12]

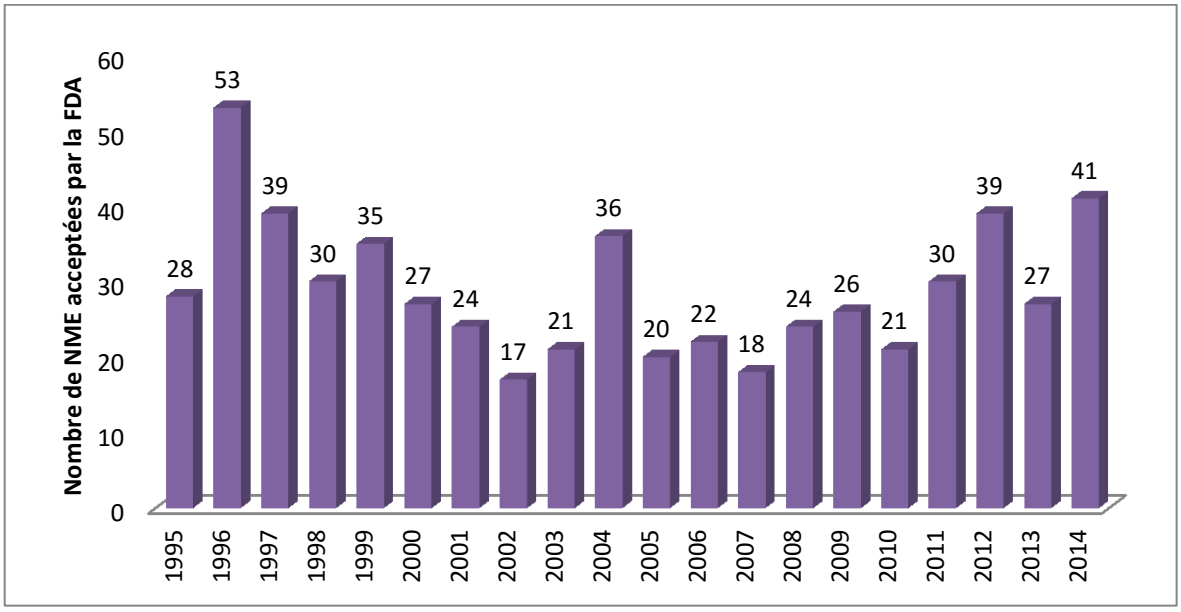


Figure 4 : Evolution du nombre de NME acceptées par la FDA entre 1995 et 2014. Alors que pendant la période 1995-2001, 34 NME étaient acceptées en moyenne chaque année, ce chiffre est tombé à 22 entre 2002 et 2007, avant de remonter à 30 pendant la période 2008-2014.

C. Vers une remise en question de la recherche pharmaceutique

Les acteurs travaillant à l'élaboration de médicaments, qu'ils soient Big Pharma, start-up ou académiques, ont tendances à tous suivre le même processus de recherche et développement. Ces dernières années ont vu se généraliser les approches centrées autour de la ou des cibles (généralement des protéines) impliquées dans des voies de signalisation liées à la maladie à traiter. L'objectif est souvent le même, trouver une molécule capable de se lier à la cible pour l'inhiber, ou dans certains cas, l'activer, afin de permettre un retour de l'activité physiologique à un seuil normal. Ainsi, l'identification de la cible est déterminante car elle est le point de départ de décisions qui seront prises par la suite.

La recherche de cibles pharmaceutiques a connu un bond énorme depuis les années 1980 grâce à l'arrivée de nouvelles technologies qui ont, entre autres, permis de séquencer les génomes et de moduler, voire d'éteindre complétement un gène afin d'étudier des maladies. Ces avancées ont été déterminantes dans la compréhension des mécanismes moléculaires et de la signalétique cellulaire. Parallèlement,

[12] Mullard, A. (2015), 2014 FDA drug approvals, *Nat. Rev. Drug Discov.*, 14, 77-81.

l'apparition des méthodes de criblage à haut débit a permis de tester un très grand nombre de molécules sur un large panel de cibles, afin d'identifier les molécules qui entreront dans le processus de développement. Cependant, cette approche tend à montrer ses limites et nombreuses sont les molécules qui n'atteignent finalement pas l'autorisation de mise sur le marché. Les raisons de ces échecs sont multiples,[13] avec dans la majeure partie des cas l'efficacité de la molécule qui est insuffisante en phase 2 ou phase 3 (Figure 5). [14,15] Dès lors, l'intérêt soudain pour les maladies orphelines prend en partie son sens, les médicaments développés pour traiter ces maladies venant généralement combler un vide pour des patients jusque-là démunis face à leur maladie. Le rapport bénéfice/risque de ces médicaments peut ainsi être moins élevé, comparé à celui d'un médicament n'apportant pas réellement un nouveau bénéfice pour le patient.

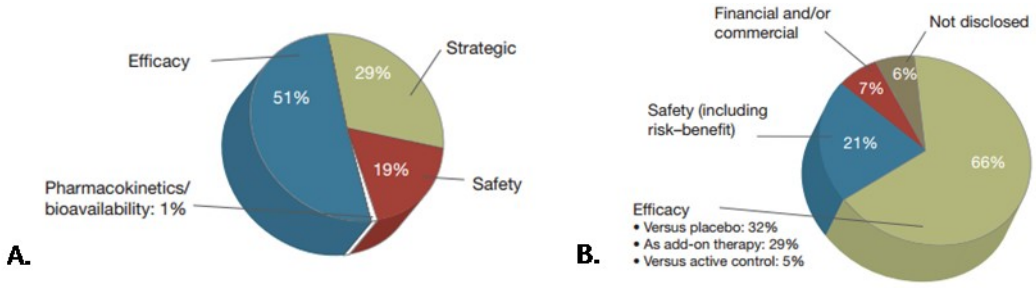


Figure 5 : A/ Causes d'échecs au cours des phases cliniques 2 entre 2007 et 2010. B/ Causes d'échecs au cours des phases cliniques 3 entre 2007 et 2010. (Données récoltées par la division Life Sciences de l'agence Thomson Reuters, présentées initialement par J. Arrowsmith)

Ces échecs, qui voient leur nombre augmenter notamment à cause des mesures prises par la FDA (partie II.B), amènent à remettre de plus en plus en question le paradigme «un gène, un médicament, une maladie», largement diffusé dans le domaine pharmaceutique. Des études ont montré l'implication des réseaux biologiques dans la réponse aux médicaments et dans la réponse aux modifications intracellulaires.[16,17] La tendance est aujourd'hui au retour de l'emploi de tests phénotypiques par les acteurs de l'industrie pharmaceutique.[18] Cette approche présente l'avantage de pouvoir tester des molécules sans pour autant connaitre leur mécanisme d'action, seulement en observant les effets du produit sur le phénotype de lignées cellulaires testées. Cependant, sans connaissance précise de la cible, il

[13] Allison, M. (2012), Reinventing clinical trials, *Nat. Biotechnol*, 30, 41-49.
[14] Arrowsmith, J. (2011), Trial watch: Phase II failures: 2008–2010, *Nat. Rev. Drug Discov.*, 10, 328-29.
[15] Arrowsmith, J. (2011), Trial watch: phase III and submission failures: 2007-2010, *Nat. Rev. Drug Discov.*, 10, 87.
[16] Mu, T.-W. *et al.* (2008), Chemical and Biological Approaches Synergize to Ameliorate Synergize to Ameliorate, *Cell*, 134, 769-81.
[17] Maslov, S., Ispolatov, I. (2007), Propagation of large concentration changes in reversible protein-binding networks, *PNAS*, 104, 13655-60.
[18] Kotz, J. (2012), Phenotypic screening, take two, *SciBX*, 5.

est très difficile d'optimiser les molécules sélectionnées.[19] Ce sont d'ailleurs ces raisons qui avaient poussé certaines entreprises à délaisser cette approche avant de revenir sur leurs pas.[18] Cette approche est aujourd'hui très utilisée pour connaitre l'effet sélectif d'une molécule en fonction des différentes voies de signalisation présentes dans différentes lignées cellulaires.

La stratégie par cible reste néanmoins largement utilisée car, malgré ses défauts tels qu'un faible taux de touches lors d'un criblage à haut débit, elle permet une approche rationnelle de la découverte de molécules actives. De plus, elle a été fortement aidée par les améliorations apportées aux méthodes se basant sur la structure des molécules comme les techniques de résolution de structures 3D des protéines (cristallographie par diffraction aux rayons X, Résonnance Magnétique Nucléaire, microscopie électronique 3D, cryométrie) ou les approches computationnelles (amarrage moléculaire ou *docking*, modélisation par homologie). C'est notamment par le biais de ces méthodes qu'ont été découverts les premiers inhibiteurs de protéines kinases.[19]

Sur approximativement 30000 protéines que compterait le protéome humain, entre 10% et 14% d'entre elles sont susceptibles d'être inhibées par un médicament.[20] En analysant la base de données Drugbank, une étude de 2011 a pour sa part démontré qu'environ 1% de ces protéines était la cible de médicaments déjà sur le marché (Figure 6 A).[21,22] En 2006, parmi les familles de protéines déjà ciblées par des molécules thérapeutiques approuvées, les récepteurs couplés aux protéines G (RCPG) représentaient un quart des cibles thérapeutiques, suivis par les récepteurs nucléaires (NR) comptant pour un huitième des cibles (Figure 6 B).[23] Notons qu'à cette époque, les protéines kinases ne figuraient pas parmi les protéines les plus ciblées par les médicaments. En comparant ces chiffres avec ceux de 2011, il est notable que la part des RCPG et des NR a baissé proportionnellement à d'autres familles de cibles thérapeutiques. Les canaux ioniques ligand-dépendants semblent avoir été plus ciblés en 5 ans, mais ce sont surtout les protéines kinases qui ont vu leur proportion augmenter et apparaissent désormais parmi les cibles thérapeutiques majeures. Il est important de souligner qu'avec l'arrivée sur le marché de 19 inhibiteurs de protéines kinases entre 2011 et le début de l'année 2015, nul doute que les protéines kinases ne tarderont pas à rattraper les RCPG. Une hypothèse d'autant plus fondée que les protéines kinases représenteraient 22% du protéome susceptibles d'être la cible d'un médicament.

---

[19] Swinney, D. C., Anthony, J. (2011), How were new medicines discovered?, *Nat. Rev. Drug Discov.*, 10, 507-19.
[20] Hopkins, A. *et al.* (2002), The druggable genome, *Nat. Rev. Drug Discov.*, 1, 727-30.
[21] Rask-Andersen, M. *et al.* (2011), Trends in the exploitation of novel drug targets, *Nat. Rev. Drug Discov.*, 10, 579-90.
[22] Law, V. *et al.* (2014), DrugBank 4.0: shedding new light on drug metabolism, *Nuc. Acids Res.*, 42, D1091-97.
[23] Overington, J. P. *et al.* (2006), How many drug targets are there?, *Nat. Rev. Drug Discov.*, 5, 993-96.

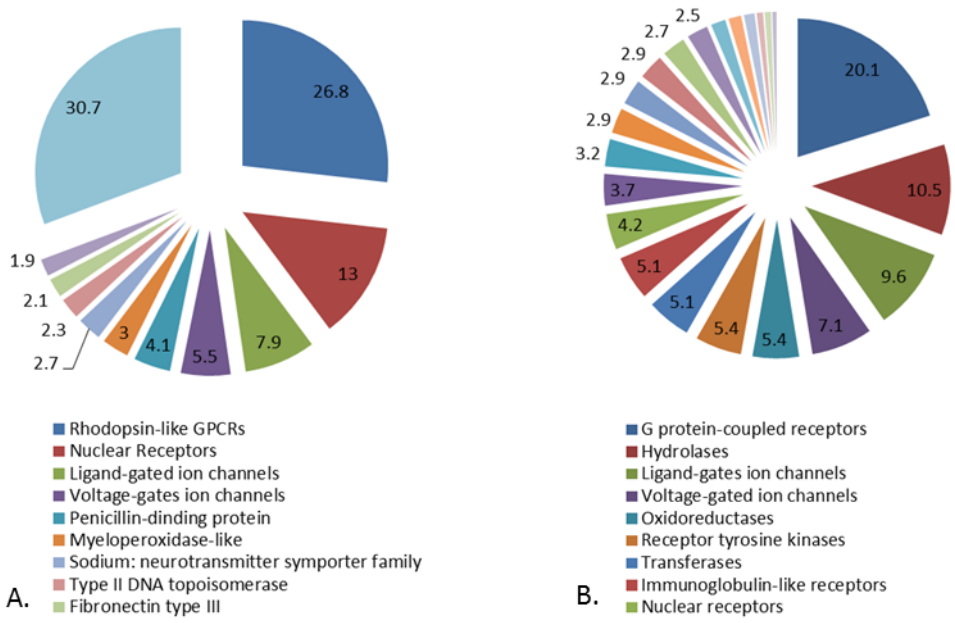


Figure 6 : A/ Distribution des familles de protéines en fonction des médicaments approuvés par la FDA avant 2006 qui les ciblent. Diagramme inspiré de Overington *et al.*[23] B/ Distribution des classes de protéines selon les médicaments approuvés par la FDA avant 2011. Diagramme inspiré de Rask-Andersen *et al.*[21]

D. La famille des protéines kinases humaines

1. Implication physiologique

Le mot kinase provient du grec ancien κινέω, qui signifie mouvoir, suivi du suffixe –ase pour indiquer la fonction d'enzyme de cette famille de protéine. Appartenant au groupe des transférases, les protéines kinases catalysent le transfert du phosphate en position gamma de l'adénosine triphosphate (ATP) vers un résidu hydroxylé (sérine, thréonine ou tyrosine) d'une cible substrat. Dans le cadre de cette thèse, il ne sera fait mention que des kinases phosphorylant d'autres protéines ou bien des lipides. La réaction inverse de la phosphorylation, la déphosphorylation, est catalysée par les protéines de la famille des phosphatases (Figure 7). La phosphorylation de la protéine cible va induire une modification structurale de cette dernière qui conduira à son activation ou à sa désactivation.

La première phosphorylation d'une protéine a été décrite en 1954. A cette époque, Gene Kennedy décrit une enzyme du foie catalysant la phosphorylation de la caséine en étudiant le transfert d'un phosphate radioactif (P^{32}) vers la protéine.[24] Peu après, les travaux successifs de Fisher et Krebs vont permettre de mettre en évidence le rôle d'une enzyme, qu'ils appelleront alors « phosphorylase kinase », capable de transférer le phosphate gamma de l'ATP vers un résidu sérine d'une protéine impliquée dans la

[24] Burnett, G., Kennedy, E. P. (1954), The enzymatic phosphorylation of proteins, *J. Biol. Chem.*, 211, 969-80.

glycogénolyse.[25] La protéine cAMP-dependent protein kinase (Prkaca), couramment appelée PKA, et son implication dans la phosphorylation de plusieurs protéines, notamment la phosphorylase kinase, fut découverte quelques années plus tard.[26] Cette découverte posa les premiers jalons de nos connaissances de la signalisation cellulaire, avec une succession de plusieurs phosphorylations permettant la transduction de signaux moléculaires. Par la suite, des travaux mettront en évidence une multitude de substrats pouvant être phosphorylés, par plusieurs protéines kinases, et ce, dans différentes compartiments de l'organisme.[27] Les découvertes de réactions de phosphorylation en cascade vont se succéder à partir des années 1990, laissant alors apparaître les implications multiples de cette famille de protéines.

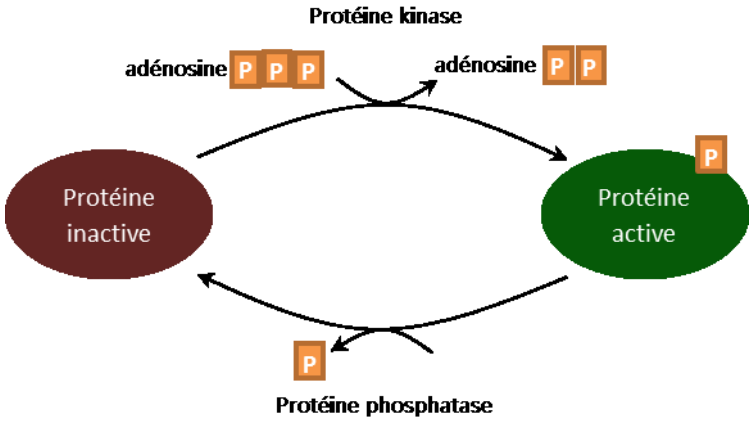


Figure 7 : Schéma de phosphorylation d'une protéine cible par une protéine kinase. Le phosphate de l'ATP en position gamma est transféré vers la protéine cible. La réaction inverse, catalysée par une phosphatase, libère le phosphate.

A ce jour, trois travaux de recherche sur les kinases et le processus de phosphorylation ont été récompensés par le Comité Nobel :

- 1992, Prix Nobel de Physiologie ou de Médecine attribué à Edmond H. Fischer et Edwin G. Krebs pour leurs découvertes concernant « la phosphorylation réversible des protéines comme un mécanisme biologique de régulation ».
- 2000, Prix Nobel de Physiologie ou de Médecine attribué à Arvid Carlsson, Paul Greengard et Eric R. Kandel pour leurs découvertes sur «la transduction de signaux dans le système nerveux ».
- 2001, Prix Nobel de Physiologie ou de Médecine attribué à Leland H. Hartwell, R. Timoty Hunt et Paul M. Nurse pour leurs découvertes concernant « les régulateurs clés du cycle cellulaire ».

---

[25] Fischer, E. H. *et al.* (1959), Structure of the site phosphorylated in the phosphorylase b to a reaction, *J. Biol. Chem.*, 234, 1698-704.
[26] Walsh, D. A. *et al.* (1968), An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle, *J. Biol. Chem.*, 243, 1763-3765.
[27] Cohen, P. (2002), The origins of protein phosphorylation, *Nat. Cell Biol.*, 4, E127-30.

Il serait trop long de citer toutes les voies de signalisation impliquant des protéines kinases, certaines sont d'ailleurs toujours inconnues, c'est pourquoi seuls trois exemples seront succinctement présentés ici.

a. Le rôle central des protéines de la famille CDK dans le cycle cellulaire

Plusieurs protéines kinases appartenant à la famille des Kinases Cycline-Dépendantes (CDK) participent à la régulation du cycle cellulaire des mammifères. Elles jouent des rôles primordiaux à différentes étapes clés. Ainsi, CDK4 et CDK6, en complexes avec des protéines cyclines de type D, concourent à l'inactivation de protéines répressives de la transcription et vont de la sorte sortir la cellule de sa quiescence (phase G_0) pour l'initier dans la phase G1. Le complexe CDK2-cycline E, contribue à enclencher les mécanismes de duplication du génome en activant des gènes de réplication de l'ADN. Les protéines CDK1 et CDK2 préparent la division cellulaire. CDK1-cycline A va participer à la condensation des chromosomes, puis CDK1-cycline B intervient dans les mécanismes de la mitose (Figure 8).[28]

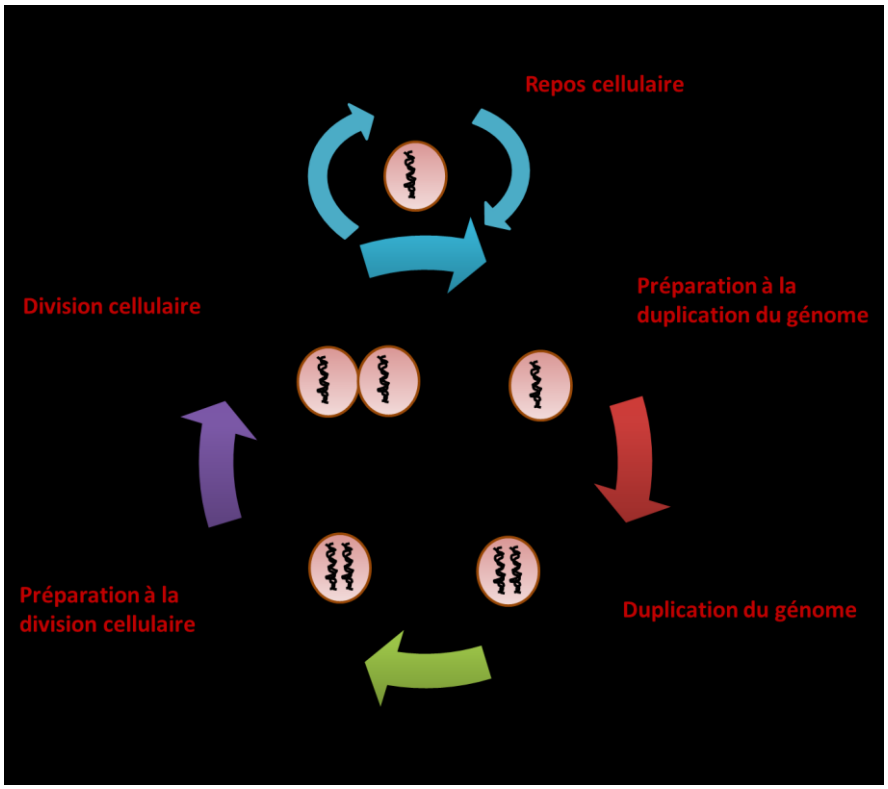


Figure 8 : Implication des protéines CDK dans le cycle cellulaire des mammifères. En association avec différents types de protéines cyclines, les CDKs jouent des rôles très importants à plusieurs étapes de la vie de la cellule.

b. L'implication des protéines kinases dans la réponse à l'insuline

L'insuline, hormone synthétisée par les cellules β des îlots de Langherans du pancréas, provoque une réponse coordonnée de certaines cellules (cellules hépatiques, cellules musculaires striées, neurones),

[28] Lapenna, S., Giordano, A. (2009), Cell cycle kinases as therapeutic targets for cancer, *Nat. Rev. Drug Discov.*, 8, 547-66.

en réponse à une augmentation de la glycémie. L'insuline va tout d'abord se lier à son récepteur extracellulaire (INSR) ou bien au récepteur à l'IGF-1 (IGF1R), récepteurs qui contiennent chacun un domaine intracellulaire tyrosine kinase (TK). Cela entraîne une autophosphorylation des récepteurs, suivie de cascades de phosphorylations participant à l'activation de voies de signalisation auxquelles contribuent des protéines de la famille des phosphatidylinositol 4.5-biphosphate 3 –kinases (PI3K) et des *mitogen-activated protein kinases* (MAPK). Ces deux voies agissent en synergie pour coordonner la translocation de vésicules contenant des transporteurs au glucose vers la membrane plasmique. Elles aident également à la synthèse de protéines, l'activation et l'inactivation d'enzymes et l'expression de gènes. Tous ces phénomènes jouent des rôles primordiaux dans la régulation de la glycémie (Figure 9).[29]

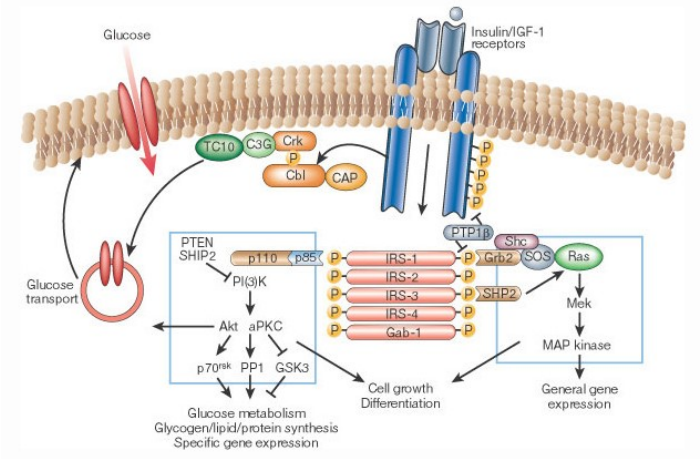


Figure 9 : Schéma de la transduction du signal de l'insuline et implication des protéines kinases dans la réponse à l'augmentation de la glycémie.[29]

c. La participation des protéines kinases à la survie neuronale

La maladie d'Alzheimer (AD) se manifeste par des déficiences cognitives. Au niveau physiologique, elle se caractérise par la formation de plaques séniles, une dégénérescence neurofibrillaire et une perte neuronale importante. C'est sur ce dernier point qu'intervient la voie de signalisation JAK2.[30] En effet, il semblerait que la liaison des peptides humanine ou coliveline sur un récepteur hétérodimérique situé sur la membrane plasmique des neurones, entraine la phosphorylation de JAK2. Cette kinase activerait alors plusieurs voies de signalisation dont les voies anti-apoptotique PI3K/Akt et MEK/ERK. Elle phosphorylerait aussi la protéine STAT3 qui est un facteur de transcription de gènes de la survie cellulaire (Figure 10).[30] Dans le cadre d'une AD, il semble qu'il y ait des dysfonctionnements dans ces voies qui entraineraient alors la mort cellulaire.

---

[29] Saltiel, A. R., Kahn, C. R. (2001), Insulin signalling and the regulation of glucose and lipid metabolism, *Nature*, 414, 799-806.

[30] Chiba, T. *et al.* (2009), Targeting the JAK2/STAT3 axis in Alzheimer's disease, *Expert Opin. Ther. Targets*, 13, 1155-67.

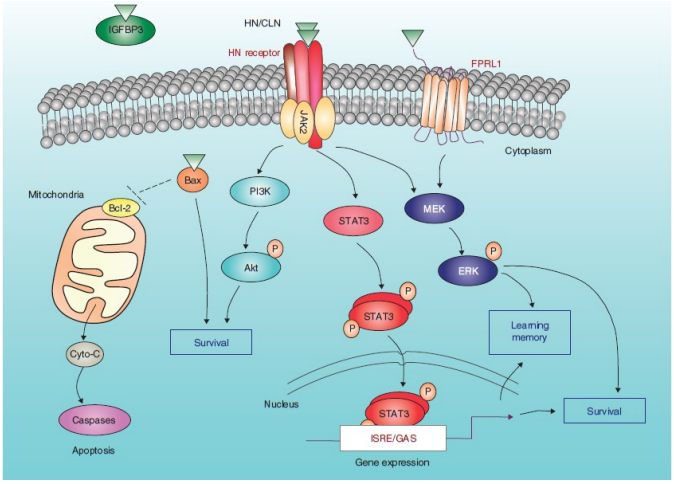


Figure 10 : Implication de JAK2 dans la survie cellulaire par l'intermédiaire des voies PI3K/Akt, MEK/ERK et STAT3.[30]

2. Classification des protéines kinases

De 1001 protéines kinases supposées chez l'Homme à la fin des années 1990,[31] les avancées dans le domaine du séquençage génomique, et notamment les travaux de Manning et de ses collaborateurs ramènent ce nombre à 518 en 2002.[32] Cependant quelques erreurs d'identification ont depuis été repérées,[33] sans pour autant que la communauté scientifique ne remette en question le nombre de 518 protéines kinases. Ainsi, la famille des kinases représente environ 1.7% du génome humain, soit approximativement deux fois moins que la famille des GPCRs.

En se basant sur les séquences des domaines catalytiques des protéines kinases, communément appelés domaines kinases, ainsi qu'en s'aidant d'autres régions connues pour leur rôle biologique, Manning et ses collaborateurs ont proposé une classification du kinome humain.[32] En tout, 478 protéines ont été classées dans la superfamille des protéines kinases eucaryotes (ePK), les 40 restantes étant qualifiées de protéines kinases atypiques (aPK), c'est-à-dire qu'une activité kinase a expérimentalement été détectée pour ces protéines, mais pour autant, leur similarité avec les ePK est très faible.[32] Ramené au nombre de domaines kinases (une protéine peut posséder plus d'un seul domaine kinase), il existe 491 domaines ePK et 40 domaines aPK (Table 1). Ces derniers ont été regroupés en 9 groupes (AGC, CAMK, CMGC, CK1, RGC, TK, TKL, STE et Autres) selon leur similarité. Précisons qu'une protéine kinase est une protéine qui possède au moins un domaine kinase, c'est pour cela qu'il a été dénombré plus de domaines kinases (531) qu'il n'y a de protéines kinases recensées (518).

---

[31] Hunter, T. (1987), A thousand and one protein kinases, *Cell*, 50, 823-29.
[32] Manning, G. *et al.* (2002), The protein kinase complement of the human genome, *Science*, 298, 1912-34.
[33] Braconi-Quintaje, S., Orchard, S. (2008), The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot, *Mol. Cell. Proteomics*, 7, 1409-19.

Superfamille	Groupe	Nombre de domaines kinases
ePK	AGC	63
	CAMK	82
	CMGC	61
	CK1	12
	STE	48
	TK	94
	TKL	43
	RGC	5
	Autres	83
aPK	A6	2
	ABC1	5
	Alpha	6
	BRD	4
	Other	9
	PDHK	5
	PIKK	6
	RIO	3

Table 1 : Répartition des 531 domaines kinases dans les différents groupes.[32]

Bien qu'une grande partie des travaux de recherche sur les protéines kinases semble ne s'intéresser qu'au domaine kinase, il est important de noter que ce domaine ne représente souvent qu'une petite partie de la longueur totale des protéines kinases.[34] En effet, d'autres domaines des protéines kinases participent à la modulation des dites protéines. Par exemple, le domaine *Src homology 2* (SH2), situé en N-terminal du domaine catalytique, peut jouer le rôle de recruteur d'autres protéines ou encore participer à la régulation de l'activité kinase.[35,36] Le domaine SH2 étant particulièrement conservé à travers la famille des protéines kinase du groupe TK, il est également possible d'effectuer une classification phylogénétique (Figure 11).

[34] Filippakopoulos, P. *et al.* (2009), SH2 domains: modulators of nonreceptor tyrosine kinase activity, *Curr. Opin. Struct. Biol.*, 19, 643-49.
[35] Mayer, B. J. *et al.* (1995), Evidence that SH2 domains promote processive phosphorylation by protein-tyrosine kinases, *Curr. Biol.*, 5, 296-305.
[36] Kuriyan, J., Eisenberg, D. (2007), The origin of protein interactions and allostery in colocalization, *Nature*, 450, 983-90.

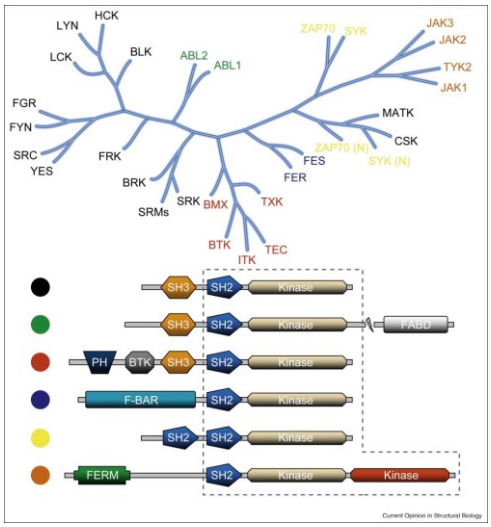


Figure 11 : Arbre phylogénétique basé sur le domaine SH2 de protéines TKs (non-récepteurs).[34]

De manière générale, les relations entre protéines kinases sont représentées par un arbre phylogénétique, rendu célèbre par les travaux de Manning *et al.* (Figure 12).[32]

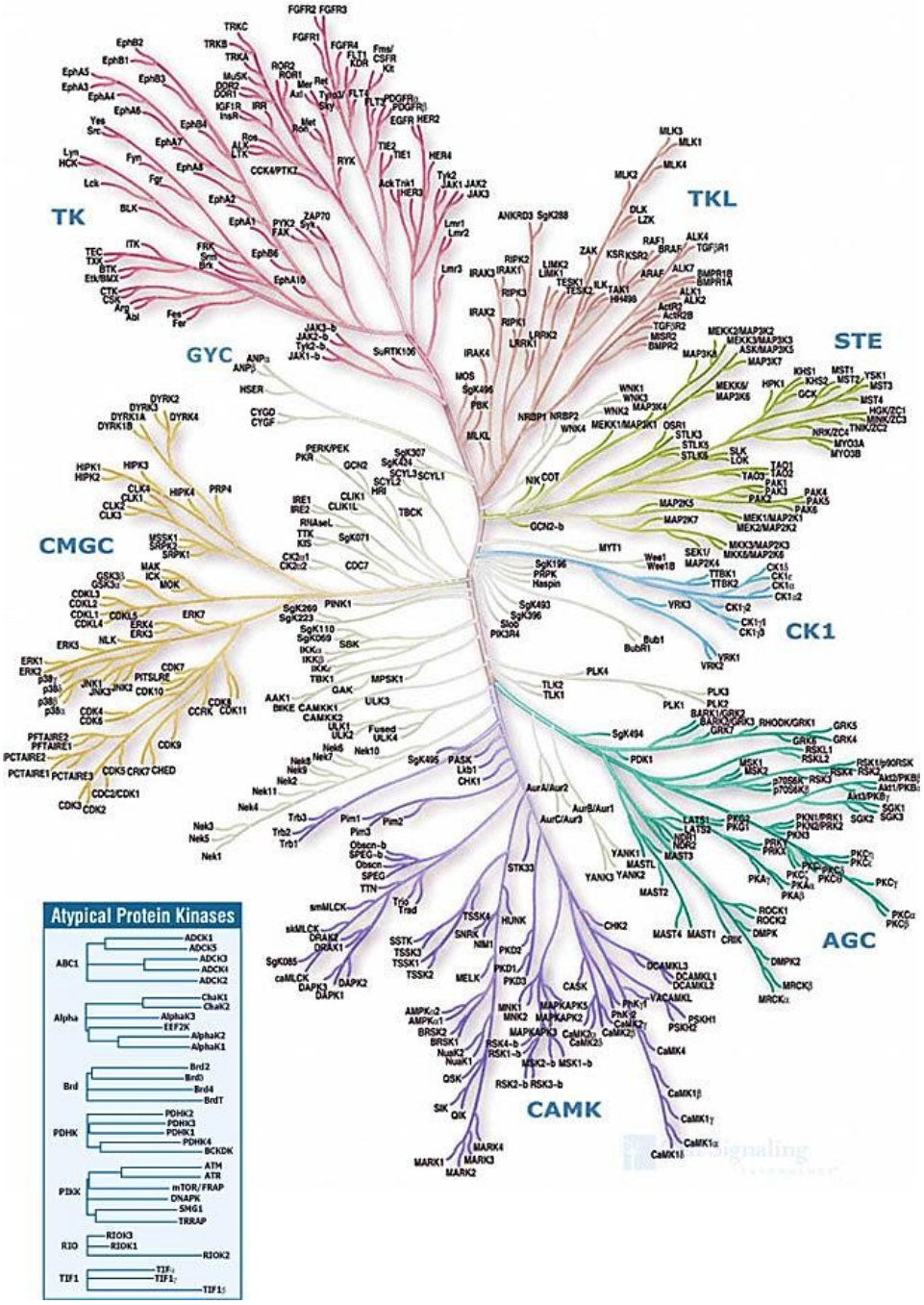


Figure 12: Arbre phylogénétique représentant les relations phylogénétiques entre les 491 domaines kinases. Les branches grisées représentent le groupe « Autres ». Les aPK sont placées dans un cadre séparé et sont elles-mêmes classées en 7 groupes.

3. Les structures des protéines kinases

Avec l'émergence de la cristallographie par diffraction aux rayons X, chaque année un nombre important de structures de protéines kinases humaines sont ajoutées dans la PDB (Protein Data Bank) de la

RCSB (Research Collaboratory for Structural Bioinformatics).[37] Depuis 2006, plus de 200 structures sont enregistrées annuellement pour atteindre, fin novembre 2014, un total de 2539 entrées dans la PDB correspondant à des protéines kinases humaines (Figure 13). Parmi ces structures, environ 350 correspondent à la protéine CDK2, 200 à la protéine MAPK14 et 100 à la protéine CHEK1. Malgré ces protéines très représentées dans la base de données, nous n'en avons dénombré que 22 ayant au moins 30 structures. Seules 68 protéines kinases n'ont qu'une seule structure disponible. Enfin, nous avons recensé 224 protéines kinases, soit moins de 50% de la totalité des protéines kinases. Ainsi, il reste encore une majorité de protéines kinases pour lesquelles nous ne connaissons pas encore la structure tridimensionnelle (Figure 14).

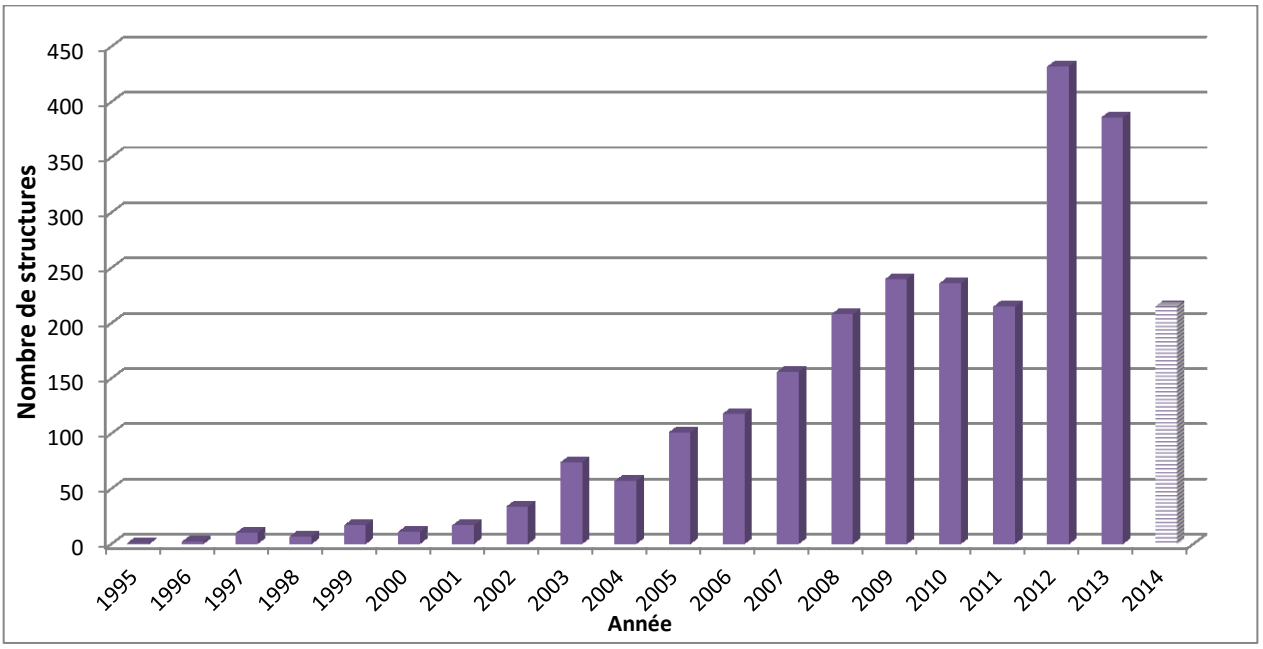


Figure 13 : Evolution du nombre de protéines kinases humaines ajoutées chaque année dans la PDB, entre février 1995 et novembre 2014.

[37] Berman, H. M. *et al.* (2000), The Protein Data Bank, *Nucl. Acids Res.*, 28, 235-42.

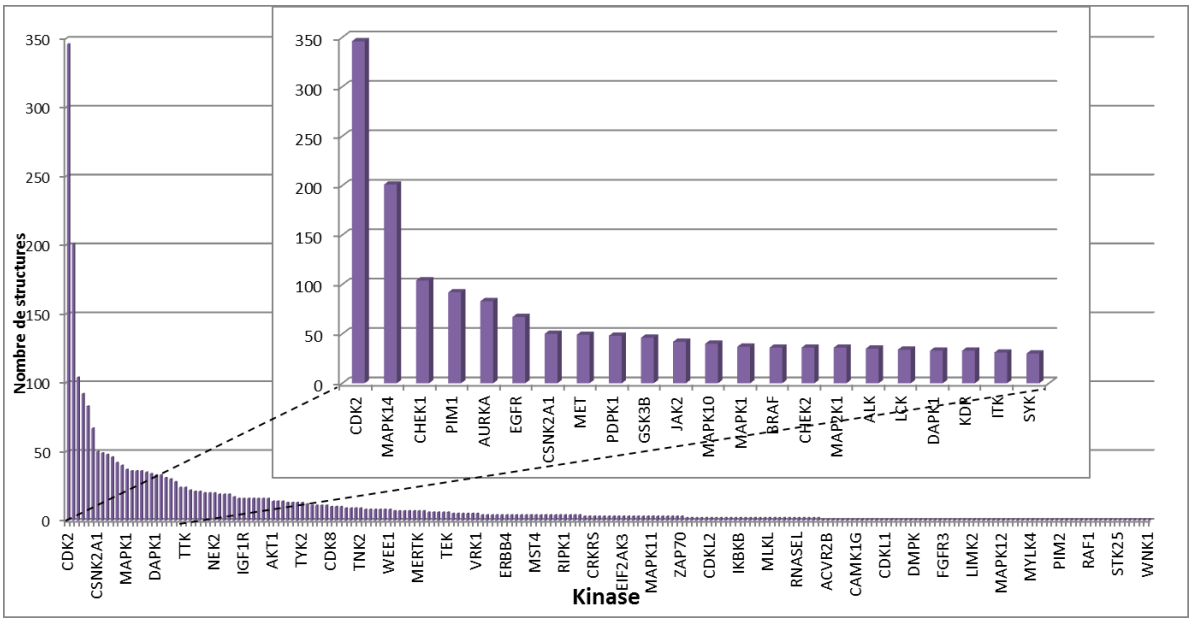


Figure 14 : Nombre de structures pour chacune des protéines kinases humaines présentes dans la PDB entre février 1995 et novembre 2014. Encadré : les 22 protéines kinases ayant plus de 30 structures.

A cela, ajoutons que les protéines kinases peuvent posséder plusieurs conformations dites actives ou inactives, comme nous allons le voir dans le prochain paragraphe. En effet, l'implication des protéines kinases dans la transmission de messages intracellulaires requière que ces dernières puissent voir leur activité régulée.

a. La conformation active

C'est dans cette conformation que les protéines kinases expriment leur activité phosphorylante. Les premières structures tridimensionnelles ont été résolues au début des années 1990. En 1991, la structure apo de Prkaca (groupe AGC) murine est déposée dans la PDB (code PDB : 2CPK). Puis, en 1993, la même protéine complexée à l'ATP et à un peptide substrat la rejoint (code PDB : 1ATP).[38,39] Les années suivantes, d'autres protéines kinases seront cristallisées avec en premier lieu, INSR (TK) puis CK-1 (CK1).[40,41] Dès lors, il est clairement apparu que les protéines kinases présentaient une grande similarité structurale intergroupes.

[38] Knighton, D. R. *et al.* (1991), Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase, *Science*, 253, 404-14.
[39] Zheng, J. *et al.* (1993), 2.2 A refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor, *Acta Crystallogr.*, 49, 362-65.
[40] Hubbard, S. R. *et al.* (1994), Crystal structure of the tyrosine kinase domain of the human insulin receptor, *Nature*, 372, 746-54.
[41] Xu, R. M. *et al.* (1995), Crystal structure of casein kinase-1, a phosphate-directed protein kinase., *EMBO J.*, 14, 1015-23.

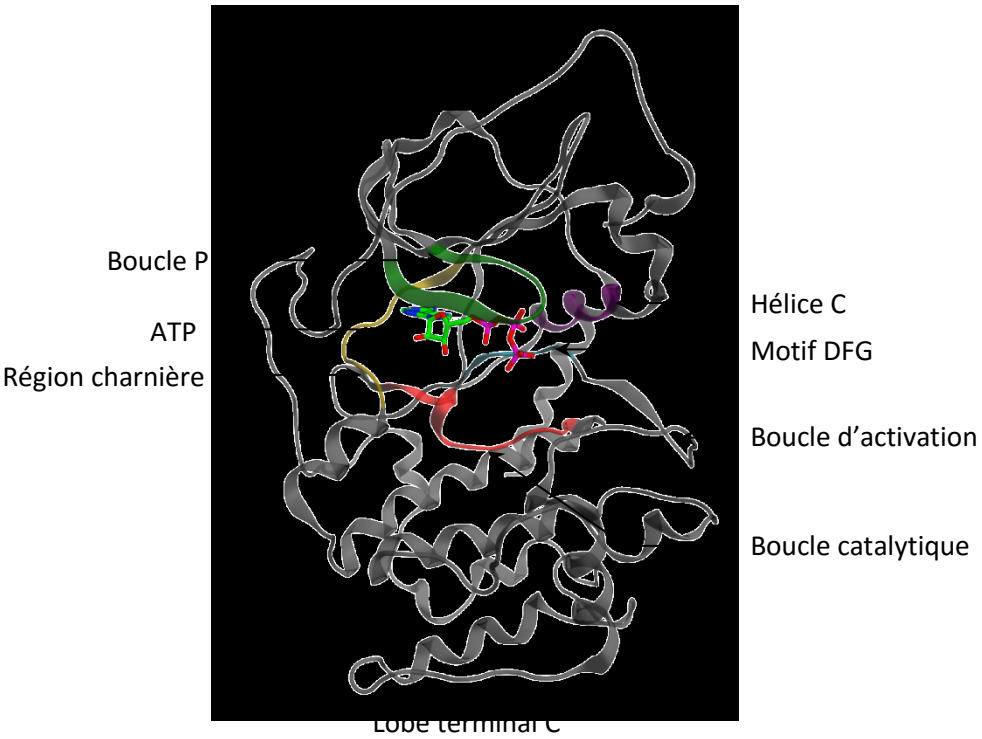


Figure 15 : Représentation en ruban de la protéine kinase Prkaca sous sa forme active (code PDB : 1ATP).

Le domaine kinase mesure en moyenne entre 220 et 260 acides aminés. De manière générale, le domaine catalytique des ePK est constitué de deux lobes (Figure 15). Le lobe terminal N est formé de cinq feuillets β antiparallèles (β1 – β5) et d'une hélice α (hélice C ou αC). La boucle P (aussi appelée boucle riche en glycines) est située entre les brins β1 et β2 et participe à la liaison et à l'orientation de l'ATP. Le lobe terminal C est le plus gros des deux lobes. Dans la conformation active, il est constitué de quatre brins β (β6-β9) et d'hélices α dont le nombre varie en fonction des structures (entre six et huit). Les brins β6 et β7 contiennent la boucle catalytique avec le motif HRD (His-Arg-Asp) à son extrémité N-terminale. Le motif DFG (Asp-Phe-Gly) se situe quant à lui entre les brins β8 et β9. Directement après se trouve la boucle d'activation (*A-loop*) responsable en partie de l'état d'activation des protéines kinases. Une boucle reliant le brin β5 et l'hélice D connecte les deux lobes entre eux. Cette région charnière (ou *hinge)* confère la flexibilité entre les deux lobes.[42,43,44]

Pour faciliter la lecture de cette partie, nous allons utiliser la numérotation des résidus présents dans Prkaca (code PDB : 1ATP). L'ATP se lie au domaine kinase au niveau de la cavité située entre les deux lobes C et N. Le groupement adénine est entouré de résidus hydrophobes (Val57, Val123, Leu173, Phe327)

[42] Cowan-Jacob, S. W. (2006), Structural biology of protein tyrosine kinases, *Cell. Mol. Life Sci.*, 63, 2608-25.
[43] Nolen, B. *et al.* (2004), Regulation of protein kinases; controlling activity through activation segment conformation, *Mol. Cell*, 15, 661-75.
[44] Taylor, S. S., Kornev, A. P. (2011), Protein kinases: evolution of dynamic regulatory proteins, *Trends Biochem. Sci.*, 36, 65-77.

et réalise des liaisons hydrogène avec des résidus de la région charnière (Met120, Glu121, Val123, Thr183).[45] Le ribose est stabilisé par des liaisons hydrogène avec des résidus du lobe C-terminal (Glu127, Glu170). La Lys72 (lysine catalytique) du motif AXK situé sur le brin β3 forme un pont salin avec la Glu91 de l'hélice C, ce qui contribue au bon positionnement des phosphates α et β de l'ATP.[45] La boucle P, très flexible, positionne le phosphate γ de l'ATP en vue de son transfert vers la protéine cible.[45] L'Asp166 du motif HRD localisé sur la boucle catalytique joue le rôle d'accepteur pour le transfert du phosphate γ vers la protéine cible. L'Asp184 du motif DFG lie un ion Mg^{2+} ou Mn^{2+} situé dans la cavité qui participe au bon positionnement des phosphates β et γ en vue du transfert de ce dernier.[45] La boucle d'activation permet de positionner le groupement hydroxyle du peptide substrat (Figure 16). Le peptide substrat, pour sa part, va se lier au niveau de l'hélice F du lobe C terminal, à partir du moment où la boucle d'activation est correctement positionnée.[46] Une fois le transfert du phosphate γ effectué, l'ADP résultant et le substrat nouvellement phosphorylé sont libérés.

Le résidu N-terminal de la région charnière est communément appelé le *gatekeeper* (Met120, Figure 16). Son rôle est primordial pour l'accès des ligands au site actif, à tel point qu'il a été montré qu'un *gatekeeper* trop grand empêche l'action de certains inhibiteurs.[47,48]

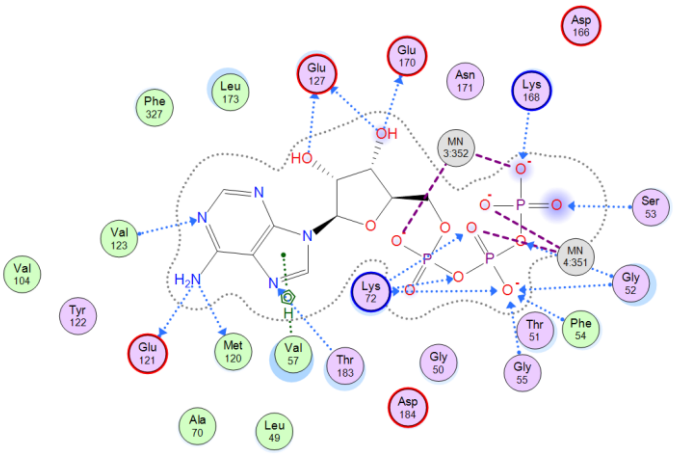


Figure 16 : Schéma représentant les interactions entre l'ATP et les résidus du domaine catalytique de Prcaka (code PDB : 1ATP). L'adénine, entourée de résidus hydrophobes, est maintenue en place par des liaisons hydrogène (tirets bleu clair). Les phosphates réalisent de nombreuses interactions non covalentes avec des résidus du site actif et les ions présents.

[45] Fabbro, D. *et al.* (2015), Ten things you should knowabout protein kinases: IUPHAR Review 14, *Br. J. Pharmacol.*, 172, 2675-700.
[46] Ubersax, J. A., Ferrell, J. E. (2007), Mechanisms of specificity in protein phosphorylation, *Nat. Rev. Mol. Cell Biol.*, 8, 530-41.
[47] Tanramluk, D. *et al.* (2009), On the Origins of Enzyme Inhibitor Selectivity and Promiscuity: A Case Study of Protein Kinase Binding to Staurosporine, *Chem. Biol. Drug Des.*, 74, 16-24.
[48] Blencke, S. *et al.* (2004), Characterization of a Conserved Structural Determinant Controlling Protein Kinase Sensitivity to Selective Inhibitors, *Chem. Biol.*, 11, 691-701.

En plus des motifs structuraux des domaines catalytiques cités précédemment, il existe des motifs originaux qui ne sont la conséquence, ni de la séquence, ni de la géométrie des protéines. Les épines hydrophobes (*hydrophobic spines)* ont été mises en évidence en 2008 à l'aide d'outils bioinformatiques.[49] Elles sont formées de résidus hydrophobes discontinus sur la séquence, localisés dans les deux lobes et participent à la plasticité des protéines kinases. La première épine hydrophobe, noté R pour régulation (*R spine*), est constituée de quatre acides aminés et sa formation est dépendante de la conformation de la boucle d'activation. En effet, l'épine R n'est présente que dans les structures en conformation active. L'épine hydrophobe C, pour catalytique (*C spine)*, se forme quant à elle uniquement en présence d'ATP dans le site actif. Elle semble agir tel un connecteur entre les deux résidus adjacents des lobes C et N terminaux (Figure 17).

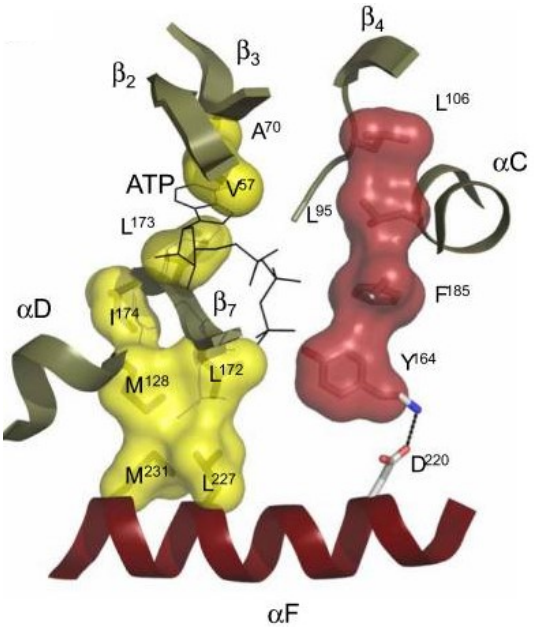


Figure 17 : Représentation en surface des épines R (en rouge) et C (en jaune) sur une structure de Prkaca. Le groupement adénine de l'ATP s'intercale entre les deux lobes de la protéine pour stabiliser l'épine C.[49]

Il est à noter que parmi les ePK se trouvent également des protéines avec des domaines kinases sans activité catalytique. Les domaines pseudokinases sont pourtant similaires aux domaines kinases d'un point de vue structural. Ils représentent environ 10% du kinome humain et possèdent les principaux motifs des domaines kinases. Cependant, ils possèdent des différences concernant certains résidus importants dans l'activité catalytique.[32] Supposés inactifs dans un premier temps, certains domaines pseudokinases semblent pourtant pouvoir participer à la création de complexes protéiques, tandis que d'autres présentent une activité catalytique selon des mécanismes différents de celui présenté dans le chapitre

[49] Kornev, A. P. *et al.* (2008), A helix scaffold for the assembly of active protein kinases, *Proc. Natl. Acad. Sci. U. S. A.*, 105, 14377-82.

1.4.1. C'est le cas par exemple de la pseudokinase WNK qui ne possède pas de lysine catalytique, mais qui en contrepartie utilise une autre lysine, présente sur le brin β2, qui remplit la même fonction (code PDB : 3FPQ).[50]

L'activation du domaine kinase peut être constitutive, mais elle passe en général par sa phosphorylation. Celle-ci peut avoir lieu à divers endroits de la protéine en fonction des cas. Ainsi Prkaca est phosphorylée sur la Thr197,[51] alors qu'INSR est activée par la phosphorylation de trois résidus tyrosine de la boucle d'activation.[52] Cela provoque un déplacement de la boucle et sa stabilisation, permettant ainsi au peptide substrat de s'approcher. L'activation des protéines kinases peut aussi passer par de plus grands déplacements de ses sous-domaines. Le domaine SH2 participe, entre autres, à l'activation des protéines CSK et ABL1. Dans les deux cas, il interagit avec le lobe terminal N, maintenant ainsi la protéine dans une conformation active (code PDB : 1K9A).[53,54]

Les sites de phosphorylation peuvent également être localisés sur d'autres domaines de la protéine. C'est le cas notamment de la protéine EPHB2 qui est phosphorylée sur son domaine juxtamembranaire. Cela a pour effet de déplacer ce domaine, qui autrement bloque la boucle d'activation.[55] D'autres modifications structurales interviennent dans le processus d'activation et celles-ci dépendent de l'état d'inactivation dans lequel était la protéine. Les structures présentes dans la PDB montrent qu'une variété de conformations inactives existe.[56] Nous allons passer en revue certaines de ces conformations dans la prochaine sous-partie.

b. Les conformations inactives et la régulation des protéines kinases

Nous l'avons vu plus haut, les protéines kinases jouent un rôle fondamental dans la signalisation intracellulaire. Dès lors, une régulation précise de ces dernières est primordiale pour éviter des dysfonctionnements du système. Ainsi, les protéines kinases passent régulièrement d'un état actif à un état inactif en fonction du mécanisme auquel elles participent. Il existe plusieurs voies intrinsèques pour stopper l'activité catalytique de ces protéines kinases. Toutes ne sont pas encore connues, de même qu'il est question que certains mécanismes d'inactivation ne s'appliquent qu'à une partie du kinome. Il existe

---

[50] Xu, B. *et al.* (200), WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II, *J. Biol. Chem.*, 275, 16795-801.
[51] Adams, J. A. (2001), Kinetic and Catalytic Mechanisms of Protein Kinases, *Chem. Rev.*, 101, 2271-90.
[52] Hubbard, S. R. (1997), Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog, *EMBO J.*, 16, 5572-81.
[53] Ogawa, A. *et al.* (2002), Structure of the Carboxyl-terminal Src Kinase, Csk, *J. Biol. Chem.*, 277, 14351-54.
[54] Nagar, B. *et al.* (2006), Organization of the SH3–SH2 unit in active and inactive forms of the c-Abl tyrosine kinase, *Mol. Cell*, 21, 787-98.
[55] Wybenga-Groot, L. E. *et al.* (2001), Structural basis for autoinhibition of the Ephb2 receptor tyrosine kinase by the unphosphorylated juxtamembrane region, *Cell*, 106, 745-57.
[56] Huse, M., Kuriyan, J. (2002), The conformational plasticity of protein kinases, *Cell*, 109, 275-82.

toutefois deux modifications conformationnelles qui mènent à une inactivation et qui ont été caractérisées chez plusieurs protéines kinases.

En 1995 une structure atypique du domaine kinase de INSR a été publiée.[40] Dans cette conformation, plus tard appelée *DFG-out*, la boucle d'activation, et notamment le motif DFG, est considérablement déplacée, comparée aux structures connues à l'époque. Plus précisément, le résidu Phe du motif DFG subit une rotation de l'ordre de 180° et se retrouve dans une poche située au niveau du site de liaison de l'adénosine. La totalité de la boucle d'activation est alors déplacée, empêchant de ce fait la fixation, aussi bien de l'ATP, que du peptide substrat (Figure 18).

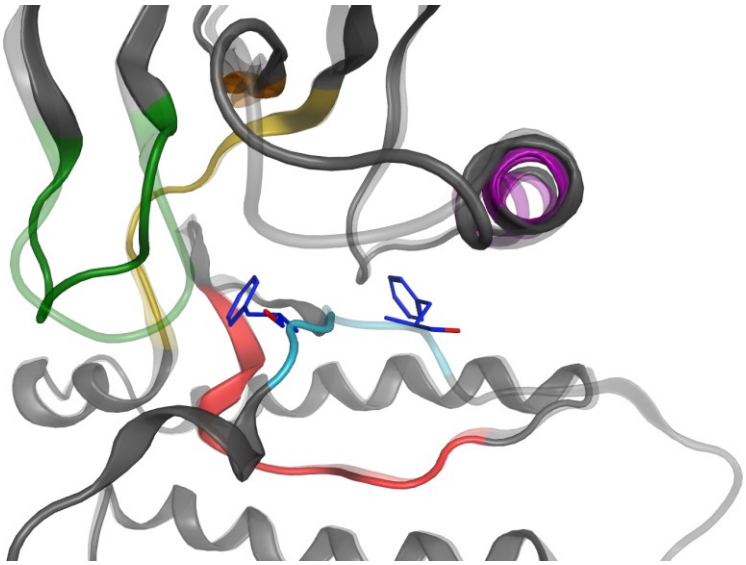


Figure 18 : Représentation en ruban des structures de ABL1 en conformation *DFG-out* (code PDB : 2HIW) et en conformation *DFG-in* (code PDB : 2GQG). La seconde est présentée en transparent pour souligner le déplacement de la rotation de la Phe (en bleu foncé) du motif DFG et de la boucle d'activation.

Depuis, d'autres structures de protéines kinases en conformation *DFG-out*, avec ou sans ligand, ont été résolues, parmi lesquelles *Abelson tyrosine-protein kinase 1* (ABL1) et s*erine/threonine-protein kinase B-raf* (BRAF) dont nous parlerons plus en détails dans la partie II.D.4 à propos des inhibiteurs de protéines kinases. En l'absence de protéines kinases cristallisées dans cette conformation, la question est maintenant de savoir si toutes les protéines du kinome sont capables d'adopter cette conformation. Malheureusement, à l'heure actuelle nous n'avons pas encore de réponse.

La conformation inactive *αC-helix out* a été décrite pour la première fois en 1996 avec une structure cristallisée de CDK2 (voir son rôle dans la partie a) (code PDB : 1HCL).[57] Cet état se manifeste par une cavité inter-lobale plus étroite due à un mouvement relatif des deux lobes. De plus, la partie N-

[57] Schulze-Gahmen, U. *et al.* (1996), High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design, *J. Med. Chem.*, 39, 4540-46.

terminale de l'hélice C est en retrait par rapport au site catalytique. Cela se traduit par la rupture du pont salin entre le Glu51 de l'hélice C et la Lys33 catalytique du brin β3. Enfin, la boucle d'activation se replie sur elle-même et son extrémité N-terminale va même jusqu'à former une petite hélice tournée vers l'hélice C (Figure 19). Cette conformation empêche le peptide substrat de se lier, a contrario de l'ATP, comme l'a démontré la structure *αC-helix out* de CDK2 en complexe avec ce ligand (code PDB : 1HCK).[57] L'activation de cette protéine passe alors par la complexation avec une des protéines de la famille des cyclines.

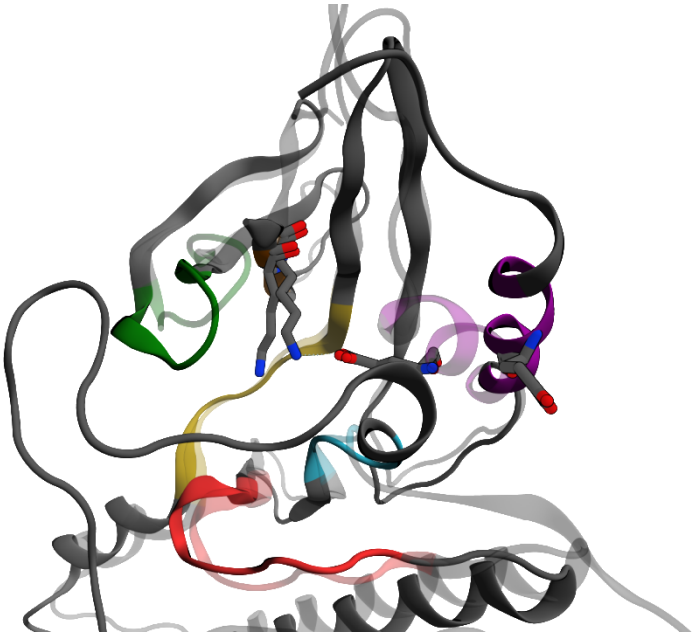


Figure 19 : Représentation en ruban des structures de CDK2 en conformation *αC-helix out* (code PDB : 1HCL) et en conformation *αC-helix in* (code PDB : 1JST). La seconde est présentée en transparent pour souligner le déplacement de l'hélice C et la modification de la boucle d'activation.

Depuis, d'autres protéines kinases ont été cristallisées dans cette conformation. C'est notamment le cas de Src (code PDB : 1FMK) où une tyrosine phosphorylée en C-terminal du lobe C entraîne un rapprochement des domaines SH2 et SH3 vers le domaine catalytique. Cela a pour conséquence finale de déplacer l'hélice C.[58]

En plus de la conformation active *DFG-in/αC-helix in*, les domaines catalytiques des protéines kinases peuvent aussi adopter les conformations *DFG-in/αC-helix out*, *DFG-out*/αC-helix in et *DFG-out/αC-helix out*. Cependant, d'après les conformations présentent dans la PDB, il semble que toutes ne se limitent pas à ces extrêmes. En effet, la plasticité des protéines kinases mène à des conformations bien particulières, caractérisées entre autres par une Phe du motif DFG dans des positions intermédiaires comparées aux deux conformations classiques *DFG-in* et *DFG-out* (Figure 20). Des travaux menés au sein du

[58] Xu, W. *et al.* (1997), Three-dimensional structure of the tyrosine kinase c-Src, *Nature*, 385, 595-602.

laboratoire ont proposé une nouvelle classification des domaines catalytiques des protéines kinases afin de nuancer le dualisme *–in /–out* du domaine DFG et de prendre en compte d'autres paramètres structuraux dont les épines.[59]

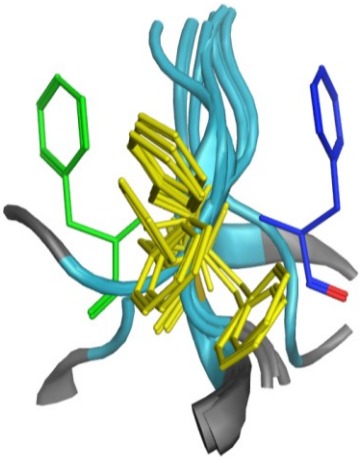


Figure 20 : Représentation en ruban des conformations intermédiaires du résidu Phe du motif DFG (en jaune) (code PDB : 1M7N, 1P4O, 2JIT, 2JIU, 2W5B, 2W5H, 3IKA et 4FZF). Les conformations extrêmes de la Phe en *DFG-in* et *DFG-out* sont représentées respectivement en vert et bleu.

4. Les inhibiteurs de protéines kinases (PKI)

Nous avons vu plus haut que les protéines kinases participent à de nombreux mécanismes biologiques. Dès lors, avec un rôle si important, la moindre modification de leur activité peut entraîner de graves troubles physiologiques. De nombreuses preuves ont été accumulées et tendent à montrer que les protéines kinases sont impliquées dans diverses maladies graves. Les cancers représentent à l'heure actuelle le domaine thérapeutique le plus dynamique quant à la recherche d'inhibiteurs. En effet, de nombreux indicateurs ont montré la responsabilité des protéines kinases dans plusieurs tumeurs cancéreuses (leucémie/ABL1, cancer du sein/EGFR, carcinome gastrique/MET, myélomes/IGFR, cancer colorectal/Src…).[60,61] Leur implication a également été montrée dans des maladies inflammatoires telle que l'arthrite rhumatoïde (JAK1/3), des troubles du système nerveux central telles que les maladies de Huntington ou d'Alzheimer (DYRK1A), ou encore des maladies cardiovasculaires et des complications liées au diabète.[62,63,64,65] La recherche d'inhibiteurs de protéines kinases est donc un domaine très actif et en

[59] Golib-Dzib, J. F. *et al.*, Towards an objective criterion for the determination of protein kinase conformations, *Submitted*.
[60] Huang, M. *et al.* (2014), Molecularly targeted cancer therapy: some lessons from the past decade, *Trends Pharmacol. Sci.*, 35, 41-50.
[61] Ma, W. W., Adjei, A. A. (2009), Novel Agents on the Horizon for Cancer Therapy, *CA Cancer J. Clin.*, 111-137, 59.
[62] Clark, J. D. *et al.* (2014), Discovery and Development of Janus Kinase (JAK) Inhibitors for Inflammatory Diseases, *J. Med. Chem.*, 57, 5023-38.

pleine expansion. Le nombre de molécules développées et conduites en essais cliniques afin de cibler cette famille de protéines, s'élevait à plus de 3000 à la fin de l'année 2014.[66]

Le développement d'inhibiteurs de protéines kinases a connu son essor à partir du début des années 2000, avec la mise sur le marché d'imatinib (Gleevec®), médicament traitant la leucémie myéloïde chronique. Le composé développé par Novartis avait été conçu spécifiquement pour se lier à la protéine de fusion Bcr-Abl, au niveau du site actif de ABL1.[67] L'arrivée sur le marché de cette molécule fut une véritable révolution car jusqu'à cette date, tous les inhibiteurs de protéines kinases souffraient d'une trop grande toxicité due à une mauvaise sélectivité. De plus, les entreprises pharmaceutiques pouvaient se montrer réticentes à l'idée de développer des inhibiteurs de protéine kinases, dont l'efficacité devait être importante à faible dose pour surpasser la compétition avec l'ATP, ligand intrinsèque.[68] Ainsi, après que l'imatinib ait montré qu'il était possible de surmonter ces difficultés, ce médicament sélectif s'est placé comme une référence sur laquelle se baseront d'autres entreprises pour développer de manière rationnelle leurs propres inhibiteurs.

Jusqu'en 2015, 28 petites molécules inhibitrices de protéines kinases ont été approuvées par la FDA (Figure 21). Parmi ces composés, idelalisib a été développé pour cibler une kinase lipidique. Nous ne nous attarderons pas plus sur cette molécule par la suite, préférant nous focaliser sur les protéines kinases qui ciblent d'autres protéines. Les inhibiteurs de protéines kinases peuvent être classés dans différentes catégories en fonction de leur mécanisme d'action. Les inhibiteurs de Type I sont des inhibiteurs compétitifs de l'ATP et ne se lient qu'à la forme active des protéines kinases. Les inhibiteurs de Type II, quant à eux, inhibent les protéines dans leur conformation inactive *DFG-out*, *i.e.* le résidu Phe du motif DFG de la boucle d'activation a effectué une rotation et ouvre l'accès à une poche hydrophobe allostérique, adjacente au site de liaison de l'ATP. Les inhibiteurs de Type I ½ ont un mode d'action intermédiaire entre les Type I et les Type II. Se liant à des structures en conformation active *DFG-in*, ils n'ont pas accès à la totalité de la poche hydrophobe allostérique et donc pénètrent moins profondément dans la protéine kinase que les Types II. Les inhibiteurs appartenant à ces trois classes sont compétitifs de l'ATP. Les

---

[63] Muth, F. *et al.* (2015), Tetra-substituted pyridinylimidazoles as dual inhibitors of p38α mitogen-activated protein kinase and c-Jun N-terminal kinase 3 for potential treatment of neurodegenerative diseases, *J. Med. Chem.*, 58, 443-56.

[64] Kikuchi, R. *et al.* (2014), An antiangiogenic isoform of VEGF-A contributes to impaired vascularization in peripheral artery disease, *Nat. Med.*, 20, 1464-71.

[65] Banks, A. S. *et al.* (2015), An ERK/Cdk5 axis controls the diabetogenic actions of PPARγ, *Nature*, 517, 391-95.

[66] Rask-Andersen, M. *et al.* (2014), Advances in kinase targeting: current clinical use and clinical trials, *Trands Pharmacol. Sci.*, 35, 606-20.

[67] Druker, B. J. *et al.* (2001), Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia, *N. Engl. J. Med.*, 344, 1031-37.

[68] Knight, Z. A., Shokat, K. M. (2005), Features of Selective Kinase Inhibitors, *Chem. Biol.*, 12, 621-37.

inhibiteurs de Type III n'occupent que la poche hydrophobe allostérique. Les inhibiteurs de Type IV se fixent sur un site allostérique non adjacent au site de liaison de l'ATP. Les inhibiteurs de ces deux classes sont non compétitifs de l'ATP. Ajoutons que depuis 2013, des molécules réalisant des liaisons covalentes ont été mises sur le marché. Nous allons voir dans les sous-parties suivantes que, malgré les possibilités offertes pour inhiber les protéines kinases, la part du kinome ciblée est encore relativement faible. A la fin de l'année 2014, il était estimé que seulement 58 protéines kinases étaient des cibles vérifiées d'inhibiteurs et 72 étaient impliquées dans des études cliniques.[66]

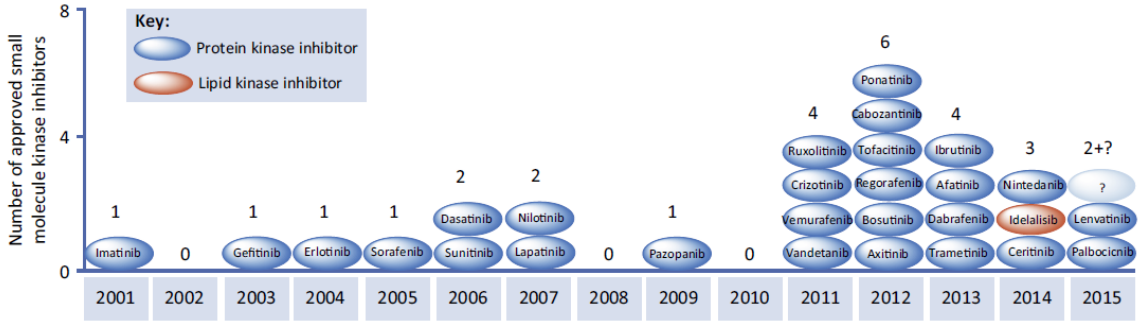


Figure 21 : Les 28 petites molécules inhibitrices de protéines kinases approuvées par la FDA jusqu'à juillet 2015. Idelalisib est, à ce jour, le seul inhibiteur de PI3K.[69]

a. Les inhibiteurs de Type I

Alors que le mode de liaison des premiers inhibiteurs de protéines kinases acceptés par la FDA n'était pas encore connu, les premiers travaux se sont orientés vers la recherche de molécules mimant la position de l'ATP afin d'en être compétitifs. Cette stratégie a sans doute été adoptée du fait que les tests enzymatiques étaient réalisés sur des protéines phosphorylées et donc dans la plupart des cas, actives. Appelés Type-I, ces inhibiteurs se lient dans la conformation *DFG-in* des protéines kinases et occupent la poche où se lie normalement l'ATP (Figure 22).[70] Généralement, les molécules de cette classe peuvent présenter jusqu'à trois liaisons hydrogène avec la région charnière de la protéine et miment les interactions hydrophobes du groupement adénine avec le brin β2 du lobe terminal C (Figure 23). L'utilisation de cette classe d'inhibiteur est particulièrement attrayante dans le cas de protéines adoptant majoritairement une conformation active. Cependant, la liaison de ces molécules s'effectuant avec une partie très conservée des protéines kinases, elle peut causer des problèmes de sélectivité.[70]

[69] Wu, P. *et al.* (2015), FDA-approved small-molecule kinase inhibitors, *Trends Pharmacol. Sci.*, 36, 422-39.
[70] Zuccotto, F. *et al.* (2010), Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation, *J. Med. Chem.*, 53, 2681-94.

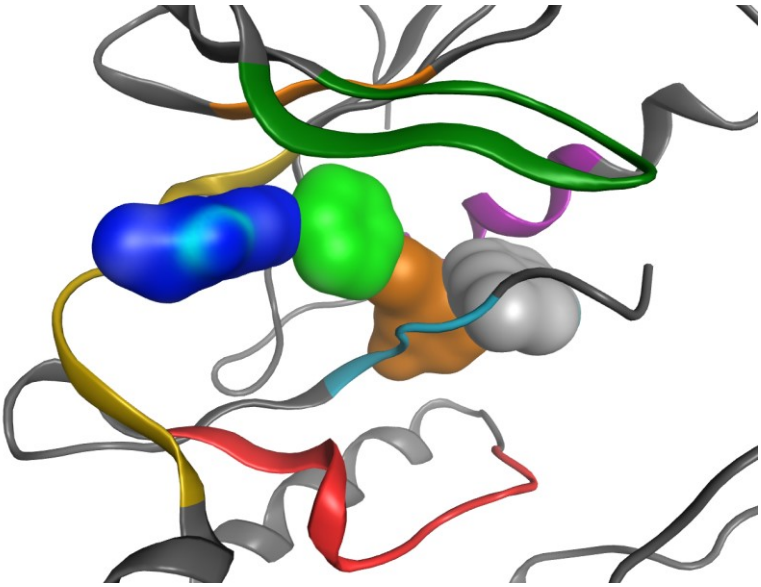


Figure 22 : Représentation de différentes poches situées dans le domaine kinase de ABL1 (code PDB : 1UWH). Le site de liaison de l'adénine est représenté en bleu, la poche hydrophobe additionnelle en vert, la poche hydrophobe adjacente en orange et la poche allostérique en gris.

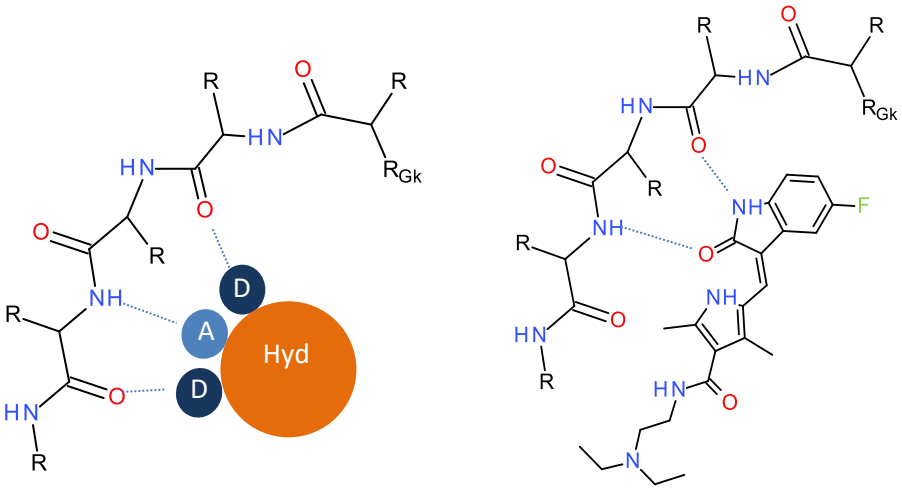
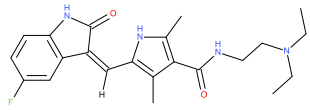
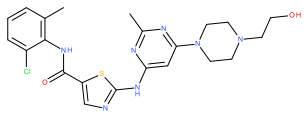
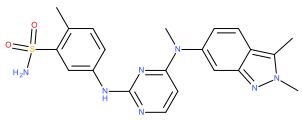
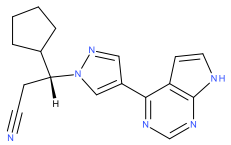
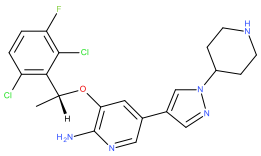
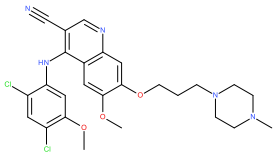


Figure 23 : Gauche : Représentation du pharmacophore des inhibiteurs de Type I. Hyd : région hydrophobe ; A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés bleus. Seule la région charnière de la protéine est représentée. Droite : Exemple des interactions réalisées entre l'inhibiteur de Type I sunitinib et la protéine KDR (PDB code : 4AGD).

A ce jour, dix inhibiteurs de Type I ont été acceptés par la FDA (Table 2), dont le premier fut le sunitinib approuvé en 2006 et utilisé dans des cas de cancers gastrointestinaux ou rénaux (Figure 23, Table 2). Parmi ces molécules, nintedanib apparait comme un cas atypique, du fait de sa forte similarité avec un inhibiteur de Type-II (voir partie II.D.4.c pour plus de détails sur cette classe d'inhibiteur). En effet, son groupement piperazine pourrait se lier dans la poche hydrophobe allostérique, à la manière de l'imatinib, tandis que le groupement amide agirait en tant que région connectrice. Toujours en émettant des hypothèses, le groupement oxindole ainsi que la fonction ester pourraient se lier dans la poche de l'adénine en effectuant des liaisons hydrogène avec la région charnière. Toutefois, comme le montre la

structure cristallographique de cette molécule liée dans KDR (code PDB : 3C7Q), la réalité est toute autre. Elle se lie en effet dans une orientation très différente. La piperazine et ce que nous décrivions comme région connectrice, sont exposées au solvant, alors que l'ester effectue une liaison hydrogène avec la Lys catalytique.[71] Ce que nous prenions pour un inhibiteur de Type II serait au final un inhibiteur de Type I classique car aucun groupement ne se lie dans la poche hydrophobe allostérique. Cependant, la protéine kinase se trouve dans une conformation *DFG-out* et non *DFG-in* comme tous les inhibiteurs de type I.

Structure chimique de l'inhibiteur	**Nom de la molécule** **Entreprise** **Année d'approbation par la FDA**	**Cibles principales**	**Indications**	**Codes PDB**
	Sunitinib Pfizer 2006	PDGFRα/β, VEGFR1/2/3, Kit, Flt3, CSF-1R, RET	Tumeur stromale gastro-intestinale Cancer avancé du rein	4QMZ, 4KS8, 4AGD, 3TI1, 2Y7J, 3MIY, 3G0E, 3G0F
	Dasatinib BMS 2006	BCR-Abl, Src, Lck, Yes, Fyn, Kit, EphA2, PDGFRβ	Leucémie myéloïde chronique (CML)	4QMS, 4XEY, 4XLI, 3QLG, 2Y6O, 3SXR, 3OHT, 3LFA, 3OCT, 3K54, 3G5D, 2ZVA, 2GQG
	Pazopanib GSK 2009	VEGFR1/2/3, PDGFRα/β, FGFR1/3, Kit, Lck, Fms, Itk	Cancer du rein	
	Ruxolitinib Incyte 2011	JAK1/2, Src	Myélofibrose	4U5J
	Crizotinib Pfizer 2011	ALK, c-Met (HGFR), ROS, MST1R	Cancer métastasique non à petites cellules du poumon (mNSCLC)	4C9W, 3ZBF, 4ANQ, 4ANS, 2YFX, 2XP2, 2WGJ
	Bosutinib Wyeth 2012	BCR-Abl, Src, Lyn, Hck	CML	4QMN, 5AJQ, 4OTW, 4MXY, 4MXO, 4MXX, 4MXZ, 3UE4, 3SOA

[71] Hilberg, F. *et al.* (2008), Structure of VEGFR2 kinase domain in complex with BIBF1120, *cancer Res.*, 68, 4774-82.

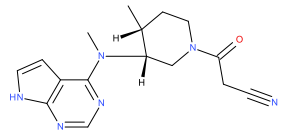
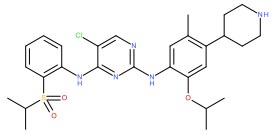
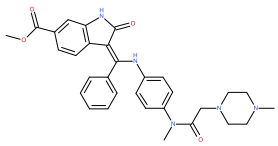
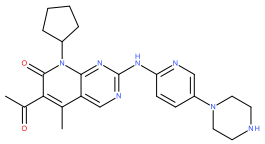
	Tofacitinib Pfizer 2012	JAK1/3	Arthrite rhumatoïde (RA)	4OTI, 3LXK, 3LXN, 3FUP, 3EYG
	Ceritinib Novartis 2014	ALK, IGF-1R, InsR, ROS1	mNSCLC	4MKC
	Nintedanib Boehringer Ingelheim 2014	FGFR1/2/3, PDGFRα/β, VEGFR1/2/3, Flt3	Fibrose pulmonaire idiopathique (IPF)	3C7Q
	Palbociclib Park Davis 2015	CDK4/6	Cancer du sein	2EUF

Table 2 : Les dix inhibiteurs de Type I approuvés par la FDA.

b. Les inhibiteurs de Type I ½

Les inhibiteurs de Type I ½ ciblent également la conformation active *DFG-in*, mais profitent de la petite taille du résidu *gatekeeper*, pour accéder à une poche hydrophobe additionnelle adjacente à celle formée par les résidus hydrophobes du brin β2 et située derrière le site de liaison de l'adénine (Figure 22). L'accès à cette poche, où l'ATP ne se lie pas, est dépendant du volume du *gatekeeper*, car trop gros, il en occupe l'espace. En accédant à cette poche, la molécule peut effectuer d'autres interactions, notamment avec les résidus du motif DFG. La liaison de ces inhibiteurs passe aussi par l'établissement de liaisons hydrogène avec la partie charnière des protéines kinases.[70] Les inhibiteurs de Type I ½ peuvent être divisés en deux sous-classes selon que la protéine ciblée est en conformation *αC-helix in* ou *out*. Le premier cas correspond à ce que nous venons de détailler (Figure 24). Nous avons dénombré quatre inhibiteurs de Type I ½ approuvés par la FDA (Table 3), parmi lesquels gefitinib quand il est lié à la protéine EGFR (Figure 24).

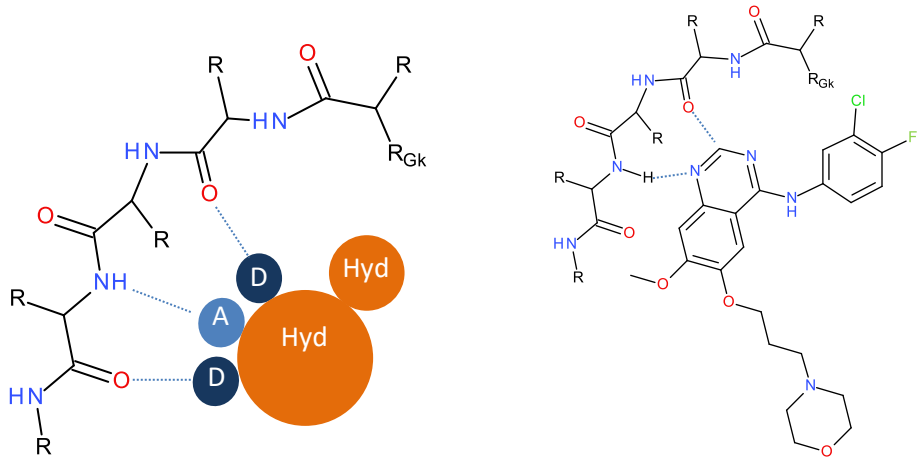


Figure 24 : Gauche : Représentation du pharmacophore des inhibiteurs de Type I ½ dans une protéine en conformation *αC-helix in*. Hyd : région hydrophobe ; A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés bleus. Seule la région charnière de la protéine est représentée. Droite : Exemple des interactions réalisées entre l'inhibiteur de Type I ½ gefitinib et la protéine EGFR (PDB code :2ITY).

D'autres inhibiteurs de Type I ½ profitent que l'hélice C soit en conformation *out* pour réaliser des interactions dans la poche hydrophobe qui se voit agrandie. L'accès à cette poche peut être un avantage pour le développement d'inhibiteurs sélectifs.[70] Vemurafenib, inhibiteur de BRAF, est un bon exemple de ce type d'inhibiteur. Dans BRAF muté (V600E), bien que la conformation soit de type *DFG-in*, la phénylalanine bloque l'accès à une grande partie de la poche allostérique, le groupement propylsulfonamide pénètre en profondeur et réalise une liaison hydrogène avec l'acide aspartique du motif DFG (Figure 25). A l'heure actuelle, seuls deux inhibiteurs approuvés par la FDA utilisent ce mode d'action, trois si l'on ajoute dabrafenib, non co-cristallisé avec une protéine kinase et dont la forte similarité avec un autre composé (code PDB : 4CQE), laisse penser qu'il se lierait à la manière de vemurafenib (Table 3).

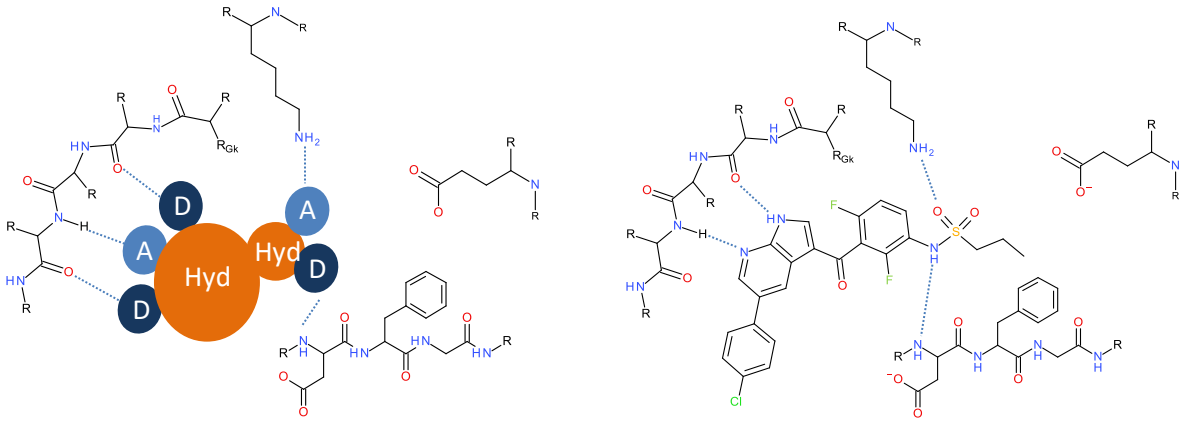


Figure 25 : Gauche : Représentation du pharmacophore des inhibiteurs de Type I ½ dans une protéine en conformation αC-helix out. Hyd : région hydrophobe ; A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés bleus. La région charnière de la protéine, ainsi que le motif DFG, la Glu de l'hélice C et la Lys catalytique sont représentées. Droite : Exemple des interactions réalisées entre l'inhibiteur de Type I ½ vemurafenib et la protéine BRAF (PDB code : 3OG7).

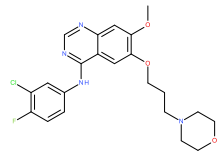
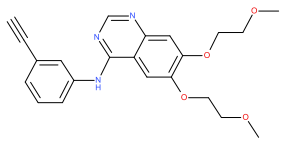
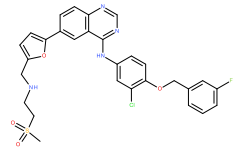
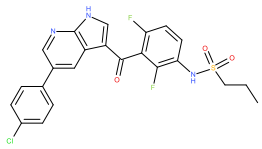
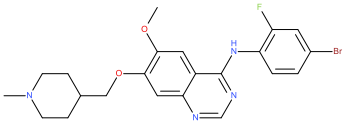
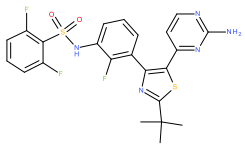
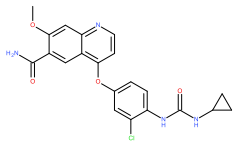
Structure chimique de l'inhibiteur	Nom de la molécule Entreprise Année d'approbation par la FDA	Cibles principales	Indications	Codes PDB	Hélice C
	Gefitinib Astrazeneca 2003-2005, 2015	EGFR	mNSCLC	4WKQ, 4I22, 3UG2, 2ITO, 2ITY, 2ITZ	in
	Erlotinib OSI 2004	EGFR	mNSCLC	1M17, 4HJO	in
	Lapatinib GSK 2007	EGFR, ErbB2	Cancer du sein métastasique	1XKK, 3BBT	out
	Vemurafenib Plexxikon/Genentech 2011	A/B/C-Raf, B-Raf (V600E), SRMS, ACK1, MAP4K5, FGR	Mélanome	3OG7	out
	Vandetanib IPR Pharms 2011	EGFRs, VEGFRs, RET, Brk, Tie2, EphRs, Src	Cancer médullaire de la thyroïde (MTC)	2IVU	in
	Dabrafenib GSK 2013	B-Raf	Mélanome		out
	Lenvatinib Easai 2015	VEGFRs, FGFRs, PDGFR, Kit, RET	Cancer différencié de la thyroïde	3WZD	in

Table 3 : Liste des sept inhibiteurs de Type I ½ approuvés par la FDA. Parmi ces molécules, trois lient des protéines kinases dans une conformation *αC-helix out*, les quatre autres en conformation *αC-helix in*.

c. Les inhibiteurs de Type II

La classe des inhibiteurs de Type II a été découverte en 2000, lorsqu'imatinib a été cristallisé dans la protéine Abl1.[72] La molécule, qui deviendra un an plus tard la première petite molécule inhibitrice de protéines kinases approuvée par la FDA, lie sa cible d'une manière jugée inattendue à l'époque, puisque qu'initialement développée en tant qu'inhibiteur de Type I.[73] En effet, en utilisant une protéine non phosphorylée, la boucle d'activation effectue une rotation et la position *out* de la phénylalanine du motif DFG empêche la fixation de l'ATP. Cependant, le déplacement de ce résidu a également pour effet de permettre l'accès à la poche hydrophobe allostérique adjacente au site de liaison de l'adénine (Figure 22). Le volume accessible est de ce fait bien plus grand que dans le cas d'une inhibition de Type I ½. Les inhibiteurs de Type II exploitent ainsi, à la fois l'accès à la région charnière pour effectuer des liaisons hydrogène, la poche de l'adénine, la poche hydrophobe adjacente à la poche précédente, et la poche hydrophobe allostérique (Figure 22). Entre les deux régions hydrophobes, le déplacement de la boucle d'activation permet également la formation de liaisons hydrogène entre une région connectrice de l'inhibiteur, le motif DFG et avec l'acide glutamique de l'hélice C (Figure 26). La présence d'un résidu *gatekeeper* de petite taille est encore aujourd'hui considérée comme un prérequis pour permettre l'accès à la poche hydrophobe adjacente, comme en témoigne le mode de liaison d'imatinib dans la protéine kinase SYK (code PDB : 1XBB). Cette dernière possède un acide aminé Met comme *gatekeeper* qui contraint l'imatinib à se replier pour adopter un mode d'interaction comparable aux inhibiteurs de Type I. Cependant, une structure de FLT3 co-cristallisée avec l'inhibiteur de Type II quizartinib montre que la présence d'un résidu phénylalanine comme *gatekeeper* n'est pas forcément rédhibitoire pour ce mode d'interaction (code PDB : 4XUF).

L'émergence des inhibiteurs de Type II a très vite été considérée comme une étape importante dans la découverte de molécules à la fois affines et sélectives.[74] Cependant, des analyses récentes de sélectivités réalisées sur 300 protéines kinases, ont montré que les inhibiteurs de Type I et de Type II pouvaient partager des profils de sélectivité similaires, allant de très sélectifs à peu sélectifs.[75] Les recherches actuelles se focalisent donc sur le développement d'inhibiteurs avec une plus grande diversité de chémotypes au niveau de la région connectrice pour accroître la sélectivité.[76] D'autres études tentent de

---

[72] Schindler, T. *et al.* (2000), Structural mechanism for STI-571 inhibition of abelson tyrosine kinase, *Science*, 289, 1938-42.
[73] Capdeville, R. *et al.* (2002), Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug, *Nat. Rev. Drug Discov.*, 1, 493-502.
[74] Mol, C. D. *et al.* (2004), Structural insights into the conformational selectivity of STI-571 and related kinase inhibitors, *Curr. Opin. Drug Discovery Dev.*, 7, 639-48.
[75] Davis, M. I. *et al.* (2011), Comprehensive analysis of kinase inhibitor selectivity, *Nat. Biotechnol.*, 29, 1046-51.
[76] Zhao, Z. *et al.* (2014), Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery?, *ACS Chem. Biol.*, 9, 1230-41.

comprendre quels sont les acides aminés des protéines kinases qui pourraient jouer un rôle dans ce type d'inhibition.

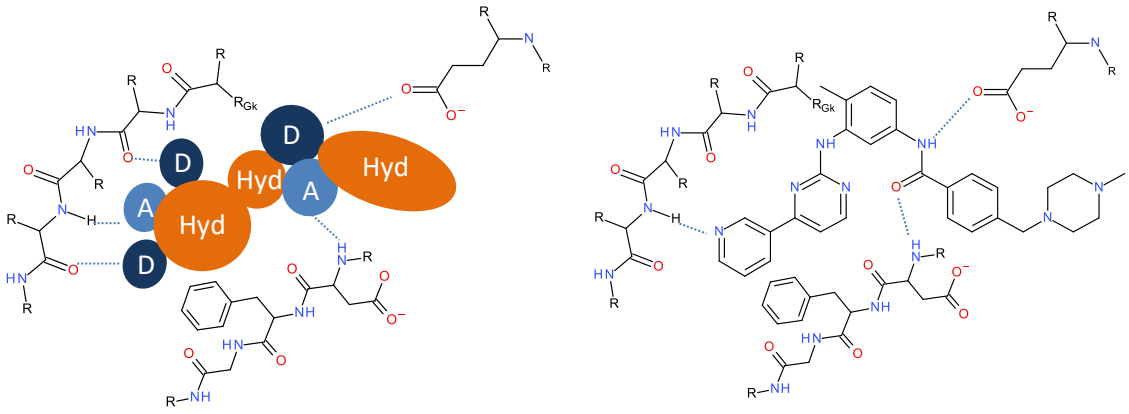
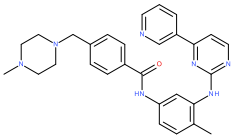
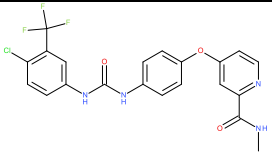
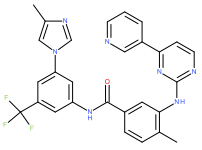


Figure 26 : Gauche : Représentation du pharmacophore des inhibiteurs de Type II dans une protéine en conformation DFG-out. Hyd : région hydrophobe ; A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés bleus. La région charnière de la protéine, ainsi que le motif DFG et la Glu de l'hélice C sont représentés. Droite : Exemple des interactions réalisées entre l'inhibiteur de Type II imatinib et la protéine Abl1 (PDB code : 2HYY).

A l'heure actuelle, sept inhibiteurs de Type II ont été approuvés par la FDA et notamment quatre en 2012, ce qui pouvait laisser supposer un tournant dans les stratégies d'inhibition de protéines kinases employées par les entreprises pharmaceutiques (Table 4). Cependant, depuis cette date, aucun nouvel inhibiteur de cette classe n'a été approuvé.

Structure chimique de l'inhibiteur	**Nom de la molécule** **Entreprise** **Année d'approbation par la FDA**	**Cibles principales**	**Indications**	**Codes PDB**
	Imatinib Novartis 2001	BCR-Abl, Kit, PDGFR	CML	4CSV, 4BKJ, 3OEZ, 3PYY, 3MS9, 3MSS, 3K5V, 3HEC, 3GVU, 3FW1, 2PL0, 2OIQ, 2HYY, 1XBB, 1T46, 1OPJ, 1IEP
	Sorafenib Bayer 2005	B/C-Raf, B-Raf (V600E), Kit, Flt3, RET, VEGFR1/2/3, PDGFRβ	Carcinome rénal	3WZE, 4ASD, 3RGF, 3HEG, 3GCS, 1UWH, 1UWJ,
	Nilotinib Novartis 2007	BCR-Abl, PDGFR, DDR1	CML	3CS9, 3GP0

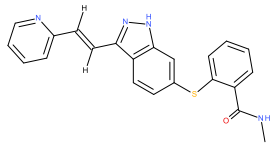
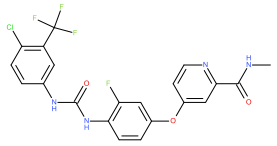
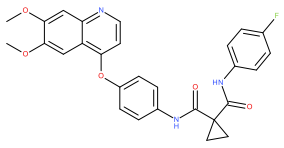
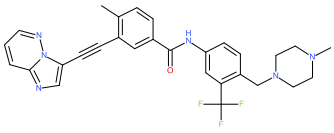
	Axitinib Pfizer 2012	VEGFR1/2/3, PDGFRβ	Carcinome rénal	4TWP, 4WA9, 4AG8, 4AGC
	Regorafenib Bayer 2012	VEGFR1/2/3, BCR-Abl, B-Raf, B-Raf (V600E), Kit, PDGFRα/β, RET, FGFR1/2, Tie2, Eph2A	Cancer colorectal métastasique (mCRC)	
	Cabozantinib Exelixis 2012	RET, Met, VEGFR1/2/3, Kit, TrkB, Flt3, Axl, Tie2	MTC	
	Ponatinib Ariad 2012	BCR-Abl, BCR-Abl T315I, VEGFR, PDGFR, FGFR, EphR, Src, Kit, RET, Tie2, Flt3	CML	4V01, 4V04, 4UXQ, 4QRC, 4U0I, 4TYJ, 4C8B, 3ZOS, 3OXZ, 3IK3

Table 4 : Les sept inhibiteurs de Type II approuvés par la FDA.

d. Les inhibiteurs de Type III

La classe des inhibiteurs de Type III regroupe les molécules se liant uniquement dans une poche adjacente au site de liaison de l'ATP (Figure 22). Ils diffèrent des inhibiteurs de Type II du fait qu'ils ne réalisent aucune liaison hydrogène avec la région charnière.[69] A cause de la grande conservation du domaine de liaison à l'ATP dans la famille des protéines kinases, l'idée de développer des inhibiteurs totalement allostériques a rapidement émergé.[77] De plus, du fait d'une sélectivité accrue, comparée aux inhibiteurs de Type I, I ½ et même II, de telles molécules employées en thérapie pourraient considérablement réduire les effets secondaires.[78] Cependant, la découverte de tels inhibiteurs semble compliquée. En effet, les techniques de criblage actuelles ont pour la plupart l'inconvénient de ne permettre que la détection de molécules compétitives de l'ATP, en utilisant des domaines kinases phosphorylés. Dès lors, la découverte de molécules ciblant des conformations autres que la conformation active semble plus difficile. De plus, la caractérisation des modes d'action, notamment pour identifier des inhibiteurs non compétitifs de l'ATP, requière des étapes et du temps supplémentaire.[79]

---

[77] Bogoyevitch, M. A., Fairlie, D. P. (2007), A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding, *Drug Discov Today*, 12, 622-33.

[78] Kirkland, L. O., McInnes, C. (2009), Non-ATP competitive protein kinase inhibitors as anti-tumor therapeutics, *Biochem. Pharmacol.*, 77, 1561-71.

[79] Gavrin, L. K., Saiah, E. (2013), Approaches to discover non-ATP site kinase inhibitors, *MedChemComm*, 4, 41-51.

Toutefois, de manière surprenante et malgré ces complications, plusieurs exemples d'inhibiteurs de Type III ont été recensés, bien que peu d'entre eux aient pour l'instant obtenu une autorisation de mise sur le marché. En effet, il semble que seule la molécule trametinib ciblant la protéine kinase MAP2K1, approuvée par la FDA en 2013, soit un inhibiteur de Type III (Table 5). Cependant, en l'absence de structure cristallographique il est difficile d'en être certain. Néanmoins, la forte similarité de trametinib avec son analogue Tak-733 cristallisé dans MAP2K1 (aussi appelé MEK1) dans une mode d'inhibition de Type III (code PDB : 3PP1), semble confirmer cette hypothèse. La liaison de Tak-733 dans la poche allostérique rendue accessible par la rotation du résidu Phe du motif DFG bloquerait la boucle d'activation dans cette conformation et ainsi empêcherait la liaison du peptide substrat, bien que l'ATP puisse accéder à son site de liaison. La lysine catalytique semble avoir un rôle central dans l'inhibition de Tak-733. Celle-ci est reliée au composé par deux liaisons hydrogène. De plus, elle forme une liaison hydrogène avec l'ATP, qui lui-même effectue une liaison hydrogène avec l'inhibiteur. Une autre liaison hydrogène avec une sérine située en C-terminal de la boucle d'activation finit de verrouiller la protéine dans une conformation inactive, maintenant rapprochée de la boucle d'activation du lobe N-terminal (Figure 27).[80]

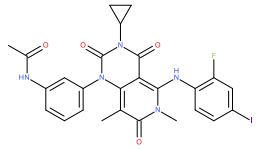
Structure chimique de l'inhibiteur	**Nom de la molécule** **Entreprise** **Année d'approbation par la FDA**	**Cibles principales**	**Indications**
	Trametinib GSK 2013	MAP2K1/2, BRAF	Mélanome

Table 5 : La seule petite molécule inhibitrice de protéines kinases supposée de type III approuvée par la FDA.

[80] Dong, Q. *et al.* (2011), Crystal Structure of the Human Mitogen-activated protein kinase kinase 1 (MEK 1) in complex with ligand and MgATP, *Bioorg. Med. Chem. Lett.*, 21, 1315-19.

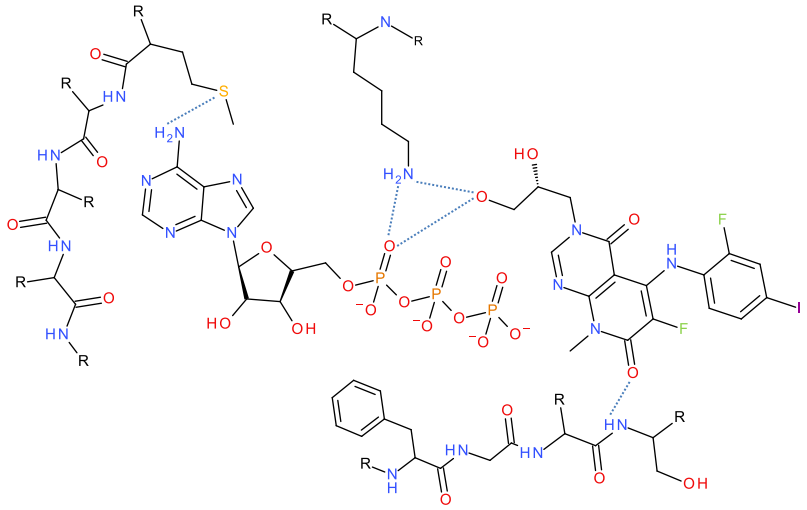


Figure 27 : Exemple des interactions réalisées entre l'inhibiteur de Type III Tak-733 (analogue de trametinib) et la protéine MAP2K1 (PDB code : 3PP1). Les liaisons hydrogène sont représentées en pointillés bleus. La région charnière de la protéine, ainsi que la portion C-terminale de la boucle d'activation, la Lys catalytique et une molécule d'ATP sont représentées.

e. Les inhibiteurs de Type IV

Les inhibiteurs de Type IV se lient aux protéines kinases dans une poche allostérique. A la différence des inhibiteurs de type III, cette poche est éloignée du site catalytique, et sa localisation dépend des domaines kinases. Elle peut se trouver aussi bien sur le lobe C que sur le lobe N terminal (Figure 28).[81] Dans tous les cas, la liaison dans un de ces sites va entraîner une modification de la conformation du site catalytique, menant à son inhibition.

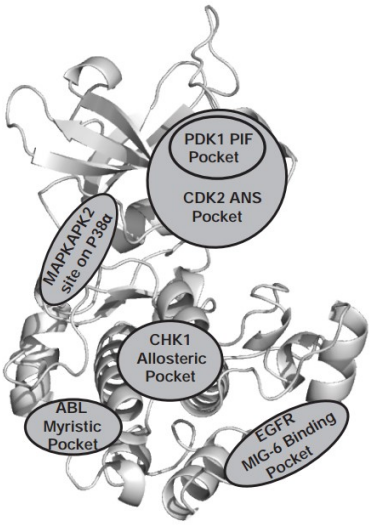


Figure 28 : Localisation des sites allostériques de différentes protéines kinases sur le même domaine kinase.[81]

[81] Lamba, V., Ghosh, I. (2012), New Directions in Targeting Protein Kinases: Focusing Upon True Allosteric and Bivalent Inhibitors, *Curr. Pharm. Des.*, 18, 2936-45.

Ainsi, dans le cas d'AKT1, une équipe a réussi à cristalliser cette kinase avec un inhibiteur allostérique lié dans une poche située entre les deux lobes, à environ 10 Å du site actif (code PDB : 3O96). Au cours de cette inhibition, l'intervention du domaine pleckstrin homology (PH), participant à la translocation de la protéine depuis le cytosol vers la membrane plasmique, est requise. En effet, le tryptophane du domaine PH effectue une interaction π-π avec l'imidazoquinoxaline de l'inhibiteur, bloquant ainsi la protéine dans une conformation inactive.[82]

D'autres exemples d'inhibitions allostériques sont présents dans la littérature (Figure 28). Il est notable que chaque poche implique des inhibiteurs appartenant à des chémotypes radicalement différents (Table 6).[81] Cependant, à notre connaissance, aucun inhibiteur de Type IV n'a encore été accepté par la FDA.

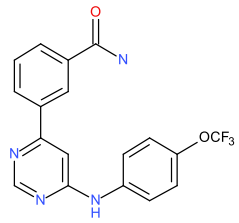
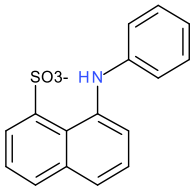
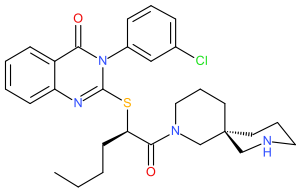
Structure chimique de l'inhibiteur	Nom de la molécule	Cible	Codes PDB	IC50 (nM)
	GNF-2	ABL	3K5V	267
	ANS	CDK2	3PXZ	37000
	38	CHEK1	3F9N	1300

Table 6 : Exemples d'inhibiteurs de Type IV.

f. Les inhibiteurs covalents

Toutes les molécules que nous avons présentées jusqu'à présent étaient non covalentes, c'est-à-dire qu'elles ne se lient à leurs cibles qu'avec des liaisons faibles (< 50 kJ.mol^{-1}). Des inhibiteurs covalents de protéines kinases ont été acceptés par la FDA pour la première fois en 2013 (Table 7). La liaison covalente

[82] Wu, W. I. *et al.* (2010), Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition, *PloS One*, 5, e12913.

s'effectue entre un accepteur de Michaël situé sur l'inhibiteur et un résidu nucléophile du site actif de la protéine, comme une cystéine. Afatinib, le premier d'entre eux, est un inhibiteur de EGFR tout comme gefitinib et erlotinib. A la différence de ces deux derniers, il est beaucoup moins sensible à la mutation T790M qui affecte le résidu *gatekeeper* de EGFR et qui entraîne une résistance au traitement chez certains patients.[83] Cette mutation n'empêche pas la fixation de ces inhibiteurs, mais associée à une autre mutation, elle augmente considérablement l'affinité de EGFR pour l'ATP ($K_{m[ATP]}$ = 8.4 µM pour EGFR L858R/T790M contre $K_{m[ATP]}$ = 5.2 µM pour EGFR sauvage).[83] Afatinib, qui se lie comme un inhibiteur de Type I tout en réalisant une réaction covalente avec une cystéine de la région charnière, bloque définitivement l'activité de la protéine, dans sa forme sauvage ou mutante (code PDB : 4G5P).

La même année, le composé ibrutinib est le second inhibiteur covalent approuvé par la FDA. Bien qu'il n'existe pas encore de structure de la molécule liée à sa cible BTK, la forte similarité d'ibrutinib avec le composé B43 permet de déduire que le groupement N-acryloylpiperidine se comporte comme un accepteur de Michaël dans une réaction avec la Cys481 voisine (code PDB : 3GEN). Ainsi, l'utilisation d'inhibiteurs covalents semble être une bonne parade pour augmenter le temps de résidence de l'inhibiteur dans sa cible. Cependant, il a été noté chez certains patients traités à l'ibrutinib, qu'une résistance pouvait apparaitre. Celle-ci est due à la mutation C481S qui empêche la réalisation de la liaison covalente.[84]

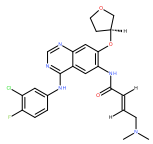
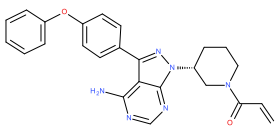
Structure chimique de l'inhibiteur	**Nom de la molécule** **Entreprise** **Année d'approbation par la FDA**	**Cibles principales**	**Indications**	**Codes PDB**
	Afatinib Boehringer Ingelheim 2013	EGFR, ErbB2, ErbB4	mNSCLC	4G5J, 4G5P
	Ibrutinib Pharmacyclics / JNJ 2013	Bruton's kinase	Lymphome à cellules du manteau	

Table 7 : Les deux petites molécules covalentes inhibitrices de protéines kinases approuvées par la FDA.

[83] Yun, C. H. *et al.* (2008), The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP, *Proc. Natl. Acad. Sci. USA*, 105, 2070-75.
[84] Woyach, J. A. *et al.* (2014), Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib, *N. Engl. J. Med.*, 370, 2286-94.

g. Bilan sur les inhibiteurs actuels de protéine kinases

Actuellement, la majorité des inhibiteurs de protéines kinases sur le marché se répartit entre les classes de Type I, Type I ½ et Type II, des inhibiteurs compétitifs de l'ATP dans les trois cas. Pour contrer des mutations qui pourraient entraîner une meilleure affinité de la protéine pour l'ATP et donc venir gêner la liaison de ces inhibiteurs, nous avons vu que, jusqu'à présent, une des solutions a été de développer des inhibiteurs covalents (afatinib, ibrutinib). Ces inhibiteurs ont l'avantage de rester liés à leurs cibles une fois fixés.

Les mutations peuvent également venir complétement bloquer la liaison de l'inhibiteur. C'est par exemple le cas de l'imatinib et du nilotinib, que se lient à la forme sauvage de Abl1, mais sont incapable d'inhiber la forme mutée du *gatekeeper* T315I. La taille de l'isoleucine est trop grande comparée à celle de la thréonine, ce qui bloque l'accès à la poche hydrophobe allostérique, en plus d'empêcher la formation d'une liaison hydrogène. Ponatinib a alors été développé pour venir inhiber la forme mutée de Abl1, en remplaçant le groupement aminopyrimidine qui provoquait un encombrement stérique avec le *gatekeeper* muté, par un groupement alcyne, plus petit.[85]

Un autre aspect qui est pris très tôt en considération lors du développement d'inhibiteurs de protéines kinases est la sélectivité. Les inhibiteurs de Type II étaient supposés être plus sélectifs que les inhibiteurs de Type I et I ½ car interagissant avec une poche allostérique moins conservée des protéines kinases. Toutefois, il s'est avéré que ces inhibiteurs peuvent également se montrer peu sélectifs, tout du moins au sein d'un même groupe de la famille des protéines kinases.[75] Dès lors, l'utilisation d'inhibiteurs de Type III et IV présente beaucoup d'avantages. Ils n'entrent pas en compétition avec l'ATP et se fixent dans des poches dont les acides aminés sont moins conservés dans le kinome. Cependant, la première génération d'inhibiteurs de Type IV souffre d'un manque d'affinité associé aux sites de liaisons qu'elle vise, sites qui sont généralement plutôt plats et de ce fait offrent peu de possibilités d'interactions.

Des inhibiteurs combinant plusieurs caractéristiques des différentes classes d'inhibiteurs susmentionnées et donc liant plusieurs sites pourraient alors voir le jour. Ces inhibiteurs bivalents, aussi appelés Type V, viseraient à la fois le site actif des protéines kinases et un site allostérique. L'intérêt de tels inhibiteurs serait que l'énergie libre de liaison d'une molécule AB serait moindre que la somme des énergies libres de liaison de A et de B.[86] Actuellement, les stratégies visant à développer des inhibiteurs de Type V se divisent en deux catégories. La première regroupe les inhibiteurs analogues de deux substrats des protéines kinases, l'ATP et le peptide substrat. La seconde catégorie rassemble les inhibiteurs se liant à la

[85] Zhou, T. *et al.* (2011), Structural mechanism of the Pan-BCR-ABL inhibitor ponatinib (AP24534): lessons for overcoming kinase inhibitor resistance, *Chem. Biol. Drug Des.*, 77, 1-11.
[86] Jencks, W. P. (1981), On the attribution and additivity of binding energies, *Proc. Natl. Acad. Sci. USA*, 78, 4046-50.

fois dans le site de liaison de l'ATP et dans une poche située ailleurs sur la protéine, à l'exception des sites des substrats.[81]

L'apparition de nouveaux inhibiteurs de protéines kinases peut entraîner *de facto* de nouvelles classes. Il peut parfois paraître ardu de retrouver à quelle classe appartient une molécule, d'autant que dans la plupart des cas, la résolution d'une structure du composé lié à sa cible est obligatoire afin de la confirmer. Le cas de nintedanib illustre parfaitement cette difficulté. Lenvatinib est un autre exemple d'inhibiteur dont le mode de liaison a pu paraître singulier. Les auteurs de l'unique structure existante à ce jour (code PDB : 3WZD) sont même allés jusqu'à annoncer une nouvelle classe d'inhibiteur de protéine kinase (Type V). Pourtant, en comparant son mode de liaison avec celui d'autres inhibiteurs, il semble correct de classer lenvatinib parmi les inhibiteurs de Type I ½. La classification des inhibiteurs de protéines kinases recèle encore quelques difficultés et l'arrivée probable de nouveaux types d'inhibiteurs risque de complexifier la tâche.

En conclusion de cette partie, nous constatons qu'en quatorze ans et vingt-sept molécules approuvées, la recherche thérapeutique ciblant les protéines kinases a fait un grand bond en avant. Dans cet intervalle de temps, la famille des protéines kinases s'est imposée comme un axe d'étude majeur dans le traitement de maladies souvent graves. Cependant, les inhibiteurs actuels ne ciblent qu'environ 10% du kinome humain, laissant de côté un grand nombre de protéines potentiellement impliquées dans diverses maladies.[66] La recherche de nouveaux inhibiteurs, qu'elle soit basée sur des études phénotypiques ou bien orientée par une cible, est aujourd'hui grandement aidée par des outils *in silico*. Que ce soit pour étudier la dynamique d'une protéine, établir un modèle de structure protéique, étudier les mécanismes de régulation cellulaire ou encore prédire des interactions protéines-ligands, ces outils sont d'une aide précieuse en plus de permettre de réduire les coûts humains et financiers. Nous allons nous attacher dans la prochaine partie à décrire les approches qui ont été développées ces dernières années, pour prédire les interactions de ligands potentiels avec des protéines kinases.

E. Les approches statistiques prédictives orientées vers la recherche d'inhibiteurs de protéines kinases

L'étude des interactions protéines kinases – ligands et la recherche de nouveaux inhibiteurs de protéines kinases (PKI) sont des domaines en plein essor depuis la commercialisation de l'imatinib.[67,87] De ce fait, les problèmes cités précédemment concernant les difficultés liées à la recherche de nouvelles molécules actives ressurgissent. Les approches prédictives *in-silico* permettent de répondre en partie à ces

[87] Cohen, P. (2002), Protein kinases--the major drug targets of the twenty-first century?, *Nat. Rev. Drug Discov.*, 1, 309-15.

problèmes, notamment au sujet des coûts de la recherche de nouveaux médicaments. Les domaines *in silico* ont beaucoup profité de l'augmentation de la puissance de calcul des ordinateurs et des avancées en informatique. Utilisée à de multiples étapes du processus de découverte de médicaments (*drug discovery*), l'informatique fait sans conteste partie de ces grandes évolutions qui ont permis la découverte de nouvelles molécules actives. Dans le cadre de cette thèse, nous nous sommes intéressés aux modèles statistiques prédictifs basés sur un apprentissage automatique (*machine learning*).[88] Ceux-ci sont particulièrement efficaces pour s'atteler aux problèmes liés aux propriétés physico-chimiques, à l'activité biologique, à la sélectivité et à la propriété intellectuelle des molécules. Tout au long de ce manuscrit nous allons faire appel à des notions de chémoinformatique et de bioinformatique, qu'il nous semble important de définir. Nous aborderons ensuite les modèles statistiques prédictifs élaborés pour trouver de nouveaux inhibiteurs de kinases.

1. Bioinformatique

La bioinformatique est un domaine multidisciplinaire à l'interface, comme son nom l'indique, entre la biologie et l'informatique (Figure 29). Créée à la suite de la révolution qu'a été le séquençage à haut débit, elle est utilisée pour stocker les informations biologiques (séquences nucléotidiques ou d'acides aminés), les annoter et permettre de les retrouver, dans le but d'aider à la compréhension des phénomènes biologiques. Toutes ces informations sont stockées dans des bases de données afin d'être accessibles et utilisables via des outils informatiques. La grande difficulté de ce domaine est la gestion d'une quantité très importante de données (multiplicité des espèces, cellules, gènes, protéines…), qu'il faut standardiser et nettoyer. A cela, ajoutons qu'un grand soin doit être apporté à la gestion des données, afin d'éviter toute redondance des informations dans les bases de données.

Une fois ces données traitées, elles doivent être rendues accessibles à travers des plateformes faciles à prendre en main par les équipes de recherche. L'analyse de ces données et l'exploitation qui en est faite doit ainsi permettre d'identifier les cibles d'intérêt thérapeutique, de regrouper les protéines au sein de familles et de comprendre les modifications cellulaires.[32] Nombreuses sont les bases de données bioinformatiques. Les plus utilisées dans le cadre de ces travaux ont été Uniprot, pour les séquences de protéines, et la PDB pour les structures tridimensionnelles des protéines.[89,37]

[88] Mitchell, J. B. O. (2014), Machine learning methods in chemoinformatics, *WIREs Comput. Mol. Sci.*, 4, 468-81.
[89] The Uniprot Consortium. (2014), Activities at the Universal Protein Resource (UniProt), *Nucleic Acids Research*, 42, D191-98.

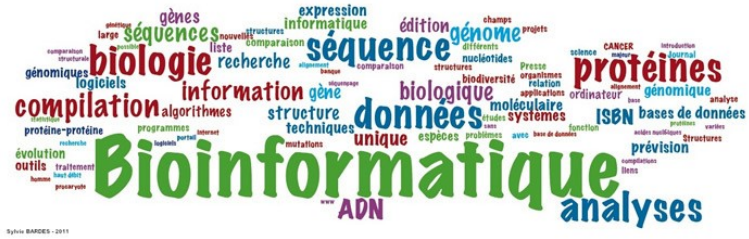


Figure 29: La bioinformatique, un domaine à l'interface de la biologie et de l'informatique (Sylvie Bardes, 2011).

2. Chémoinformatique

La chémoinformatique est à la chimie ce que la bioinformatique est à la biologie. Son but premier est le stockage, l'analyse, la recherche d'entités chimiques et le développement d'outils prédictifs d'inhibiteurs. Tout comme la bioinformatique, la chémoinformatique nécessite de gérer une grande quantité d'informations, qu'il faut standardiser et dont la redondance doit être exclue. La chémoinformatique est aussi vue comme l'outil d'étude et d'exploration de l'espace chimique. Espace chimique dans lequel les molécules sont représentées par des descripteurs moléculaires.[90]

En chimie thérapeutique, l'espace chimique décrit l'ensemble des petites molécules organiques, contenant au maximum 30 atomes lourds, qu'il est théoriquement possible de synthétiser (Figure 30).[91,92] Il contiendrait 10^{60} entités, contre 10^{23} étoiles dans notre univers observable.[93] Il est important de noter qu'à l'heure actuelle, plus de 99,9% des molécules de l'espace chimique n'a jamais été synthétisé.[94,] En septembre 2015, la base de données CAS référençait plus 100 millions de molécules enregistrées. De ce vaste espace, des librairies de molécules peuvent être générées selon les besoins afin, par exemple, de regrouper toutes les molécules d'un même chémotype, ou bien d'identifier tous les isomères d'une même formule brute. De la volonté de créer des bases de données de molécules, a vite découlé le besoin de les comparer. Pour ce faire, la notion de similarité a vu le jour. Les modèles prédictifs se basent sur elle pour stipuler que des molécules structuralement similaires ont des activités similaires sur une même protéine. Bien qu'aujourd'hui très contestée, c'est sur cette hypothèse que repose une grande partie des décisions prisent en chimie médicinale.[95]

---

[90] Baskin, I., Varnek, A. (2008) Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening; *RCS*: Cambrige.

[91] Reymond, J.-L. *et al.* (2011), Exploring the chemical space of known and unknown organic small molecules at www.gdb.unibe.ch, *Chimia*, 65, 863-67.

[92] Reymond, J.-L. *et al.* (2012), The enumeration of chemical space, *WIREs Comput. Mol. Sci.*, 2, 717-33.

[93] Kirkpatrick, P., Ellis, C. (2004), Chemical space, *Nature*, 432, 823.

[94] Reymond, J. L., Awale, M. (2012), Exploring chemical space for drug discovery using the chemical universe database, *ACS Chem. Neurosci.*, 3, 649-57.

[95] Martin, Y. C. *et al.* (2002), Do Structurally Similar Molecules Have Similar Biological Activity?, *J. Med. Chem.*, 2002, 4350-58.

Afin d'accéder aux propriétés des molécules, il est d'usage de passer par des bases de données. Ces dernières se cantonnent principalement à référencer des molécules qui ont été synthétisées et testées. Elles fournissent ainsi les informations d'activité nécessaires à la création de modèles statistiques. Elles peuvent aussi renseigner sur les fournisseurs à contacter pour se procurer lesdites molécules. C'est notamment le cas de la ChEMBL et de PubChem, qui ont toutes deux été largement exploitées dans le cadre de cette thèse.[96,97] Avec une quantité de données qui ne cesse de croître chaque année, nul doute que l'importance de ces bases de données publiques ne va faire que s'accentuer.[98] Les données qu'elles contiennent peuvent, notamment, être utilisées afin de travailler sur le développement d'outils prédictifs. Nous avons choisi de les répartir en trois catégories : chimiométrie, protéométrie et protéo-chimiométrie, qui diffèrent les unes des autres en fonction de leurs domaines d'application. Ces outils peuvent être utilisés en amont de la recherche thérapeutique et notamment afin de trouver de nouveaux inhibiteurs de protéines kinases.

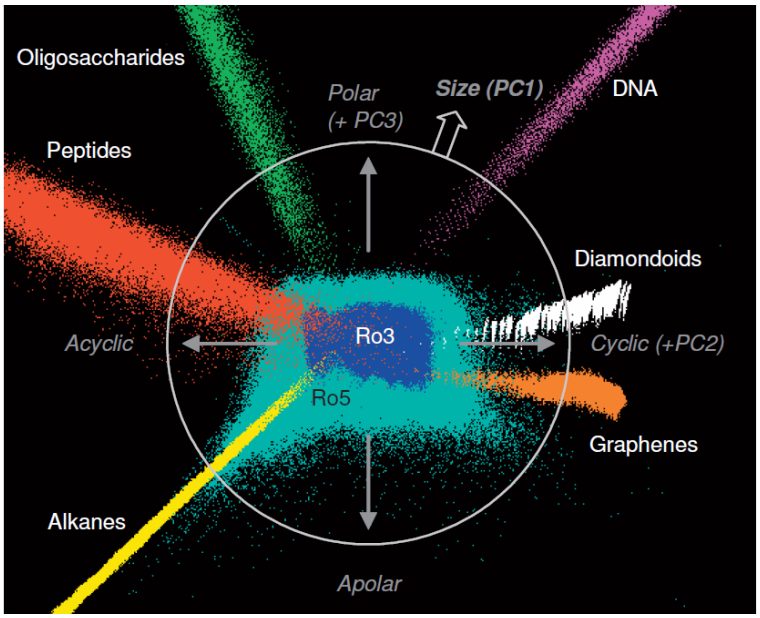


Figure 30: Représentation des molécules de la base de données PubChem dans un espace chimique à deux dimensions. Les dimensions ont été obtenues à l'aide d'une analyse en composantes principales. Ro3 désigne les fragments trouvés dans PubChem, et Ro5 les molécules suivant la règle des 5 de Lipinski. Les autres catégories représentées correspondent à des collections virtuelles des familles de molécules désignées.[91]

[96] Bento, A. P. *et al.* (2014), The ChEMBL bioactivity database: an update, *Nucleic Acids Res.*, 42, 1083-90.
[97] Bolton, E. *et al.* (2008) Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities; *Elsevier*; Vol. 4, pp 217-241.
[98] Gaulton, A. *et al.* (2012), ChEMBL: a large-scale bioactivity database for drug discovery, *Nucleic Acids Res*, 40, D1100-07.

3. La chimiométrie

a. Historique et définition

La chimiométrie (ou *chemometrics* en anglais) englobe toutes les applications mathématiques et statistiques appliquées à la mesure en chimie. Au début des années 1960, Hansch *et al.* parviennent à développer une équation capable de prédire le coefficient de partition octanol/eau (log P).[99,100] C'est le premier exemple de quantification d'une relation entre une structure moléculaire et une propriété biologique. Ces travaux seront plus tard considérés comme le premier jalon du domaine d'étude des relations quantitatives structure à propriété (QSPR) et du domaine qui nous intéresse ici, celui des relations quantitatives structure à activité (QSAR). Au fil des années, le domaine de la chimiométrie s'est développé, profitant des évolutions de l'informatique, pour s'imposer comme un outil à part entière du *drug design* assisté par ordinateur (CADD). Aujourd'hui, son représentant le plus utilisé en modélisation moléculaire, est le QSAR. Nous allons voir en détail que cette approche a été utilisée dans l'optique de rechercher de nouveaux inhibiteurs de protéine kinases.

Le QSAR peut être défini comme un modèle statistique étudiant les interactions d'un groupe de molécules (jeu d'apprentissage) sur une cible (généralement une protéine), les interactions étant représentées par des valeurs mesurées expérimentalement (Figure 31). L'ensemble de ces valeurs sert de référence pour établir un modèle utilisé pour prédire les interactions entre la cible et les molécules non testées (jeu de test). Comme tout modèle statistique, les prédictions des modèles QSAR reposent tout particulièrement sur la fonction d'apprentissage automatique sous-jacente et la fiabilité des données employées. Elles sont également liées aux descripteurs moléculaires utilisés (suite de valeurs numériques pouvant être employées par le programme informatique). Afin de pouvoir juger de la capacité du modèle à réaliser de bonnes prédictions, il est crucial de le valider statistiquement. Finalement, la notion de domaine d'applicabilité va définir la capacité du modèle à faire des prédictions correctes pour des molécules comprises dans un espace chimique délimité par des molécules similaires à celles du jeu d'apprentissage (Figure 32).

---

[99] Hansch, C., Fujita, T. (1964), p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, *J. Am. Chem. Soc*, 86, 1616-26.

[100] Hansch, C. (1969), A quantitative approach to biochemical structure-activity relationships, *Acc. Chem. Res.*, 2, 232-39.

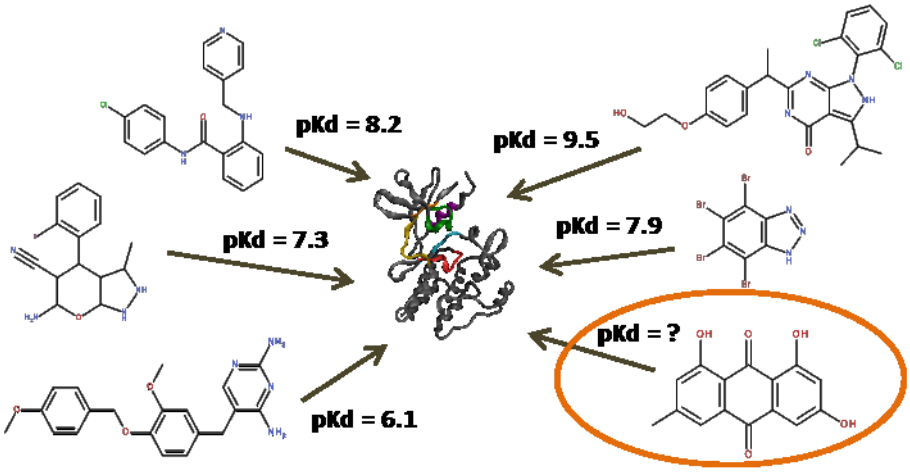


Figure 31 : Représentation des interactions entre différentes molécules et une protéine. Les interactions sont dans cet exemple caractérisées par des valeurs logarithmiques de constantes de dissociation (pKd). Les modèles QSAR s'attèlent à utiliser les valeurs connues d'affinité pour prédire celle d'une autre molécule entourée en orange.

b. Modèles QSAR appliqués à la recherche d'inhibiteurs de protéines kinases

De nombreuses études basées sur des approches QSAR ont été entreprises dans le but de découvrir de nouveaux inhibiteurs de protéines kinases. Jeux de données, descripteurs, fonctions mathématiques, validations et estimations du domaine d'applicabilité, sont les paramètres qui varient selon les cas. Nous allons voir que les méthodes de validation peuvent beaucoup différer dans leurs approches et dans leurs exigences. La notion de domaine d'applicabilité a quant à elle, fait son apparition tardivement dans les publications traitant de QSAR.[101] Nous allons voir par la suite quelques exemples de modèles QSAR qui ont été développés (Table 8).

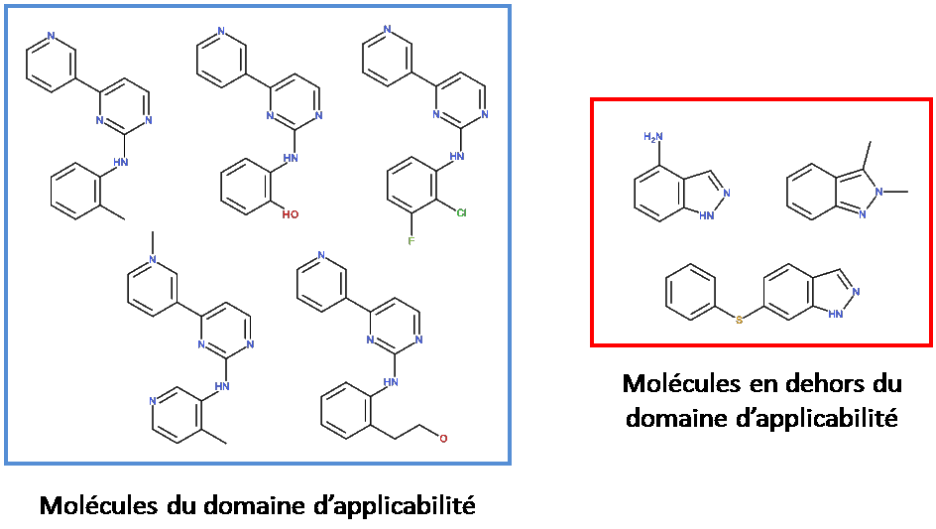


Figure 32 : Exemple de domaine d'applicabilité basé sur la similarité des molécules.

[101] Sahigara, F. *et al.* (2012), Comparison of Different Approaches to Define the Applicability Domain of QSAR Models, *Molecules*, 17, 4791-810.

i. Les jeux de données expérimentaux

Avant même de penser à la conception du modèle, il est nécessaire de collecter un jeu de données. Dans le cadre d'un modèle QSAR, celui-ci consiste en un nombre plus ou moins élevé de molécules, toutes testées sur une seule protéine. Malgré le nombre important de cibles d'études concernant la famille des protéines kinase, les publications concernées (Table 8) montrent que certaines semblent être privilégiées (SRC, GSK3B). Quoi qu'il en soit, le jeu de données utilisé doit être le plus fiable possible. Cependant, cette fiabilité est difficile à estimer à cause de protocoles expérimentaux parfois insuffisamment détaillés. En effet, rares sont les études présentant des valeurs d'inhibition avec un écart-type. Dans ces cas, il faudra prendre en compte qu'une incertitude inconnue existe déjà pour les données du jeu d'apprentissage. Une incertitude supplémentaire s'applique également si plusieurs sources de données sont mélangées ; une approche couramment employée pour augmenter la taille du jeu de données. Une étude s'est intéressée à la quantifier dans le cas de données d'IC50 et il semblerait que celle-ci soit raisonnable (σ_{pIC50}=0,68).[102] Dans le cas des modèles présentés (Table 8), hormis celui de Sprous *et al,*[103]*.* tous reposent sur des valeurs d'IC50, mais aucun ne communique sur l'incertitude des données expérimentales. L'ensemble de ces IC50 représente la variable que le modèle tentera de prédire.

ii. Les descripteurs moléculaires

Un descripteur moléculaire est le résultat final d'une procédure mathématique et logique qui transforme l'information chimique encodée dans la représentation symbolique d'une molécule en une valeur numérique utile ou en un résultat d'une expérience standardisée.[104] Le choix de la représentation des molécules dans le modèle et de ce fait, le choix des descripteurs, peut représenter un dilemme pour le modélisateur. Choisir entre une représentation à deux (2D) ou trois dimensions (3D), revient à sélectionner le niveau de détails à attribuer aux molécules. Tandis qu'une description 2D consiste généralement en une conversion du diagramme moléculaire en une suite de valeurs numériques, ainsi qu'au calcul de propriétés pharmacophysicochimiques, les descripteurs 3D tiennent compte de la position des atomes entre eux dans l'espace. Bien que non exhaustifs dans notre sélection de modèles, nous constatons que la plupart des modèles QSAR développés sur des protéines kinases utilisent des descripteurs moléculaires 3D. Parmi ces descripteurs, il semblerait qu'il y ait des préférences pour les descripteurs CoMFA/CoMSIA[105,106,107] et les

---

[102] Kalliokoski, T. *et al.* (2013), Comparability of mixed IC50 data – A statistical analysis, *PLoS One*, 8, e61007.
[103] Sprous, D. G. *et al.* (2006), Kinase inhibitor recognition by use of a multivariable QSAR model, *J. Mol. Graph. Model.*, 24, 278-95.
[104] Todeschini, R., Consonni, V. (2008) Handbook of molecular descriptors; *Wiley-VCH*.
[105] Zhu, L. L. *et al.* (2001), 3D QSAR analyses of novel tyrosine kinase inhibitors based on pharmacophore alignment., *J Chem Inf. Comput. Sci.*, 41, 1032-40.
[106] Lescot, E. *et al.* (2005), 3D-QSAR and docking studies of selective GSK-3beta inhibitors. Comparison with a thieno[2,3-b]pyrrolizinone derivative, a new potential lead for GSK-3beta ligands., *J. Chem Inf. Model.*, 45, 708-15.

descripteurs pharmacophoriques.[108,109] Les descripteurs 3D requièrent au préalable un alignement conformationnel de toutes molécules utilisées dans le modèle. Développés par Tripos® (société rachetée par Certara®), les descripteurs CoMFA (C*omparative Molecular Field Analysis*) prédisent, pour un conformère donné, les champs électrostatiques et stériques qui l'entourent. Pour cela, une sonde atomique (généralement un carbone sp^3 chargé +1 ou -1) est placée à chaque maille d'une grille centrée sur la molécule (généralement d'un pas de 1 ou 2 Å). En utilisant des principes de mécanique moléculaire, les champs sont mesurés à chacune des positions de la sonde. Les descripteurs CoMSIA (*Comparative Molecular SImilarity Indices Analysis)* tendent à améliorer les CoMFA car ils visent à pallier les changements trop brusques de potentiels énergétiques qui peuvent être induits par le pas entre chaque maille de la grille.[105]

Les descripteurs pharmacophoriques, comme ceux mentionnés dans la Table 8 reposent sur la notion de pharmacophore. Il est définit par l'Union Internationale de Chimie Pure et Appliquée (IUPAC) comme l'« ensemble de propriétés stériques et électroniques qui est nécessaire pour la réalisation d'interactions supramoléculaires optimales, avec une cible biologique spécifique, pour déclencher (ou bloquer) sa réponse biologique ».[110] En résumé, le pharmacophore d'une molécule regroupe généralement, les donneurs et accepteurs de liaisons hydrogène (respectivement DLH, ALH), les groupements hydrophobes, les noyaux aromatiques, ainsi que les groupements chargés. Dans le cadre d'un modèle QSAR, n'ayant pas forcément accès aux structures des complexes cristallisés de la cible et des molécules du jeu d'apprentissage, il est nécessaire d'émettre des hypothèses sur ce à quoi pourrait ressembler le pharmacophore.

Nous notons également que certains modèles ont été réalisés à l'aide de descripteurs 2D.[103,111,108,112,113] Bien que les raisons du choix du type de descripteurs ne soient que rarement expliquées, nous pouvons supposer que la nécessité d'aligner les molécules, dans le cas de descripteurs 3D, peut être ardue. En effet, il peut être difficile d'aligner correctement des dizaines, voire des centaines de molécules,

---

[107] Thaimattam, R. *et al.* (2005), 3D-QSAR studies on c-Src kinase inhibitors and docking analyses of a potent dual kinase inhibitor of c-Src and c-Abl kinases, *Bioorg. Med. Chem.*, 13, 4704-12.
[108] Lather, V. *et al.* (2008), QSAR models for prediction of Glycogen Synthase Kinase-3b inhibitory activity of indirubin derivatives, *QSAR Com. Sci.*, 27, 718-28.
[109] Kirubakaran, P. *et al.* (2012), Ligand-based Pharmacophore Modeling; Atom-based 3D-QSAR Analysis and Molecular Docking Studies of Phosphoinositide-Dependent Kinase-1 Inhibitors, *Indian J. Pharm. Sci.*, 74, 141-51.
[110] Wermuth, C. G. *et al.* (1998), Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998), *Pure Appl. Chem.*, 70, 1129-43.
[111] Shi, W. M. *et al.* (2007), QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression, *Eur. J. Med. Chem.*, 42, 81-86.
[112] Comelli, N. C. *et al.* (2014), Conformation-independent QSAR on c-Src tyrosine kinase inhibitors, *Chemom. Intell. Lab. Syst.*, 134, 47-52.
[113] Martin, E. *et al.* (2011), Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity, *J. Chem. Inf. Model.*, 51, 1942-56.

lorsqu'elles n'ont pas le même chémotype. A l'opposé, les descripteurs 2D ne nécessitent pas d'alignement et sont généralement plus rapides à calculer. Nous n'avons pas la prétention ici, de référencer tous les descripteurs 2D existants. Toutefois en analysant ceux utilisés dans les modèles QSAR listés, nous pouvons différencier les descripteurs décrivant des propriétés de molécules (DLH, ALH, poids moléculaire, volume, lipophilie etc…) des descripteurs d'empreintes moléculaires, plus connus sous leur dénomination anglaise *fingerprints*. Tandis que les premiers vont prendre la forme de valeurs numériques, les seconds sont habituellement des suites de valeurs binaires. Il existe autant de *fingerprints* que d'algorithmes pour les construire.[114]

Dans les exemples présentés, certains groupes ont fait le choix d'utiliser des descripteurs physico-chimiques qu'ils ont calculé à l'aide de logiciels propriétaires (SYBYL et Dragon).[Erreur ! Signet non défini.,112]. A l'inverse, dans leur publication, Martin *et al.* ont fait le choix d'employer un *fingerprint* circulaire appelé *Functional-Class FingerPrint* (FCFP), de diamètre 6 (FCFP6), développé par la société Accelrys® depuis rachetée par Dassault Systèmes®. Ce *fingerprint* encode l'environnement circulaire de chaque atome de la molécule jusqu'à une distance de trois liaisons depuis l'atome central (Figure 33). Retranscrivant la molécule sous la forme d'une combinaison de sous-structures, le *fingerprint* se voit attribuer une valeur de 0 ou de 1 à chacune de ses positions (appelées bits) en fonction de la présence (1) ou de l'absence (0) de la sous-structure dans la molécule.

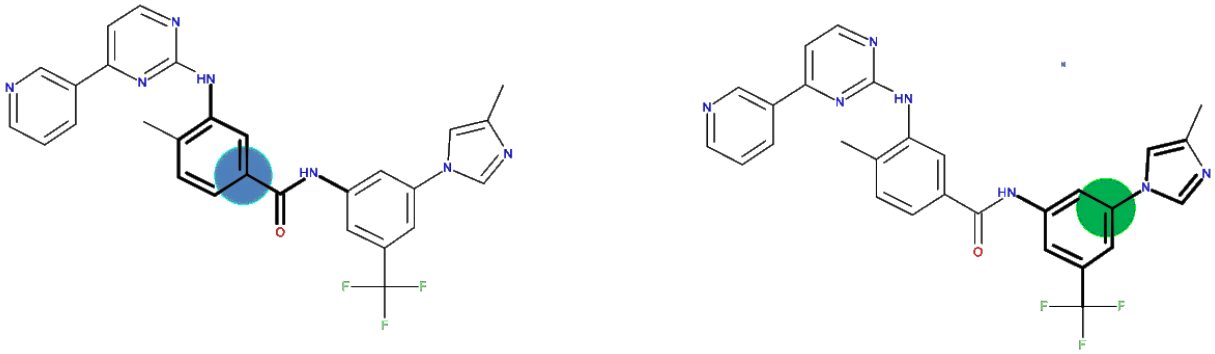


Figure 33 : Représentation des environnements circulaires de deux atomes (bleu et vert) pour des chemins de longueur deux et trois respectivement.

Ajoutons que dans la conception d'un modèle QSAR (cela s'appliquera aussi aux approches que nous verrons dans les prochaines parties), le choix des descripteurs va avoir pour effet de générer plus ou moins de variables descriptives et donc de complexifier plus ou moins le modèle.

[114] Riniker, S., Landrum, G. A. (2013), Open-source platform to benchmark fingerprints for ligand-based virtual screening, *J. Cheminfor.*, 5, 26.

Référence	Dimension	Méthode	Jeu de données	Descripteurs	Validation	Meilleurs résultats	Domaine d'applicabilité
Zhu *et al.*[105]	3D	PLS	31 molécules testées sur ERBB2	CoMFA	LOO	Q^2 : 0.656	Non estimé
				CoMSIA		Q^2 : 0.710	
Lescot *et al.*[106]	3D	PLS	74 molécules testées sur GSK3B	CoMSIA	LOO Validation externe	Q^2 : 0.688 R^2 : 0.539	Non estimé
Thaimattam *et al.*[107]	3D	PLS	87 molécules testées sur SRC	CoMFA	LOO Validation externe *Training set* : 84% *Test set* : 16%	Q^2 : 0.612 R^2 : 0.591	Non estimé
				CoMSIA		Q^2 : 0.688 R^2 : 0.539	
Sprous *et al.* Erreur ! Signet non défini.	2D	PLS	258 PKI 230 non PKI	Propriétés physiques, fragment	Validation externe *Training set* : 30% *Test set* : 70%	Précision : 86% TFP : 10%	Non estimé
Shi *et al.*[111]	2D	Algorithme de Colonies de fourmis	61 molécules testées sur EGFR	Divers	LOO Validation externe	Q^2 : 0.635 R^2 : 0.722	Non estimé
Lather *et al.*[108]	2D	PLS	44 molécules testées sur GSK3B	Divers	Validation externe *Training set* : 80% *Test set* : 20%	R^2 : 0.60	Non estimé
	3D			Descripteurs pharmacophoriques		R^2 : 0.97	
Martin *et al.*[113]	2D	Naïve Bayésienne	92 protéines Au moins 600 molécules par protéine	*Fingerprint* circulaire	Validations externe *Training set* : 75% *Test set* : 25%	R^2moyen : 0.59	Non estimé
Kirubakaran *et al.*[109]	3D	PLS	82 molécules testées sur PDPK1	Descripteurs pharmacophoriques	Validation externe *Training set* : 80% *Test set* : 20%	R^2 : 0.751	Estimé par approche *leverage*
Comelli *et al.*[112]	2D	Régression linéaire, optimisation de Monte Carlo	80 molécules testées sur SRC	Divers	LOO Validation externe Randomisation d'Y	Q^2 : 0.64 R^2 : 0.68	Estimé par approche *leverage*

Table 8 : Neuf exemples de modèles QSAR appliqués à la prédiction de nouveaux inhibiteurs de protéines kinases. PLS : Moindres Carrés Partiels ; LOO : *Leave-One-Out ; training set*: jeu d'apprentissage; *test set*: jeu de test.

iii. Régression et classification

Une fois les données choisies, nettoyées puis décrites, il faut choisir un algorithme d'apprentissage automatique qui permettra d'effectuer, au choix, une classification ou une régression.

Dans une classification, la variable d'activité est discrète, représentée la plupart du temps par deux classes actifs/inactifs, mais des classes supplémentaires peuvent être créée. L'algorithme va ensuite générer un modèle en fonction des variables du jeu d'apprentissage. Ce modèle va alors être appliqué au jeu de test pour classer les molécules dans une des deux classes en fonction de leurs descripteurs.

Les performances d'un modèle reposant sur une classification peuvent être estimées en calculant divers coefficients. Ils reposent tous sur l'estimation de la capacité du modèle à avoir correctement classé les molécules du jeu de test. Dans tous les cas, un tableau de contingence doit être établi. Il permet de répartir les prédictions dans quatre groupes selon que le modèle ait prédit des : Vrais Positifs (VP), Vrais Négatifs (VN), Faux Positifs (FP), Faux Négatifs (FN) ; comme illustrés dans la Table 9.

		Classes prédites	
		Actifs	Inactifs
Classes expérimentales	Actifs	VP	FN
	Inactifs	FP	VN

Table 9 : Exemple de tableau de contingence.

A partir de cette table, il est possible de calculer plusieurs paramètres de validation, notamment la sensibilité, la spécificité et la précision du modèle (Figure 34). Les valeurs de ces paramètres sont toutes comprises entre 0 et 1, 1 étant idéal. La sensibilité et la spécificité représentent les pourcentages, respectivement, d'actifs et d'inactifs correctement prédits, tandis que la précision est un indice globale de la capacité du modèle à correctement classé les individus.

$$Sensibilité = \frac{VP}{VP + FN}$$

$$Spécificité = \frac{VN}{VN + FP}$$

$$Précision = \frac{VP + VN}{VP + FN + VN + FP}$$

Figure 34 : Définitions de quelques paramètres de validation d'un modèle QSAR basé sur une classification.

Dans un modèle QSAR basé sur une régression, la ou les variables à expliquer sont quantitatives. En se basant sur les entrées du jeu d'apprentissage, le modèle va prédire cette variable pour celles du jeu de test. Il est alors possible d'estimer sa performance en comparant les valeurs prédites et les valeurs expérimentales. Deux paramètres sont alors couramment calculés. Il s'agit du coefficient de détermination (R^2) et de l'erreur quadratique moyenne (RMSE) (Figure 35). La valeur de R^2 ne peut excéder 1 qui est sa valeur optimale, tandis que plus le RMSE sera faible et plus le modèle sera performant. Anciennement, mais cela arrive encore parfois aujourd'hui, l'erreur du modèle était représentée par l'écart-type (σ).

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

Figure 35 : Définitions des paramètres de validation R^2 et RMSE de modèle QSAR basé sur une régression. y_i représente chacune des valeurs expérimentales du jeu de test, $\hat{y}_i$ les valeurs prédites, $\bar{y}_i$ la moyenne des valeurs expérimentales et *n* le nombre d'observations expérimentales dans le jeu de test.

Parmi les exemples de modèles QSAR que nous avons trouvés dans le contexte de la prédiction d'inhibiteurs de protéines kinases, un seul était basé sur une classification.[111] En utilisant la méthode des moindres carrés partiels (PLS), les auteurs ont établi un modèle visant à différencier les inhibiteurs des molécules non inhibitrices de protéines kinases. Les performances des différents modèles créés ont ensuite été estimées en calculant la sensibilité et le taux de faux positifs (1 - spécificité).

A l'inverse, tous les autres exemples trouvés effectuent dans leur protocole une régression linéaire. Tous, sauf trois, ont fait le choix d'utiliser la méthode PLS.[105,106,107,108,109] Shi *et al.* ont opté pour un algorithme de colonies de fourmis,[111] Martin *et al.* ont utilisé une approche Naïve Bayesienne,[113] et Comelli *et al.* ont, quant à eux, effectué une régression linéaire simple avec optimisation des paramètres selon la méthode de Monte Carlo.[112] La validation des modèles passe à chaque fois par le calcul du R^2, parfois de σ, et du Q^2 que nous allons décrire dans la partie suivante.

iv. La validation

Valider son modèle est une étape cruciale car elle permet de donner de la valeur et de la confiance aux prédictions qui seront faites. Nous avons vu dans la partie II.E.3.i que le jeu de données initial est découpé en un jeu d'apprentissage et en un jeu de test. Puis dans la partie II.E.3.iii, nous avons détaillé le calcul du R^2, basé sur la différence entre les valeurs prédites et les valeurs expérimentales. La création du

jeu d'apprentissage et du jeu de test étant généralement faite de manière aléatoire, celle-ci peut entraîner un biais au moment de la validation. Pour remédier à ce problème, il est recommandé d'effectuer une validation croisée. La validation croisée peut prendre plusieurs formes en fonction de la taille du jeu de données et du temps imparti. En QSAR, lorsque que le nombre de molécules est relativement faible (moins d'une centaine), il est courant d'effectuer une validation croisée de type *leave-one-out* (LOO). Cette validation consiste à retirer une molécule du jeu d'apprentissage et à générer le modèle avec le nouveau jeu d'apprentissage ainsi obtenu, puis à prédire l'activité de cette molécule. Cette étape est effectuée autant de fois qu'il le faut pour que chaque molécule ait été retirée et prédite. Le paramètre Q^2 (parfois noté R^2_{CV}) peut alors être calculé (Figure 36).

Figure 36 : Définition du paramètre de validation Q^2 suite à une validation croisée. y_i représente chacune des valeurs expérimentales du jeu de test, $\hat{y}_{i/i}$ la valeur prédite à chaque itération, $\bar{y}_i$ la moyenne des valeurs expérimentales et *n* le nombre d'observations expérimentales dans le jeu de test.

Dans les cas où il y a trop de molécules dans le jeu d'apprentissage, il peut être souhaitable de faire moins d'itérations en effectuant une validation croisée de type *leave-multiple-out* (LMO). Dans ce cas, le jeu d'apprentissage est coupé en plusieurs parties (généralement 5 ou 10) et à chaque itération, un modèle est créé puis validé avec la partie laissée de côté.[115] Notons que le terme validation croisée est plus fréquemment utilisé. Le modèle ainsi obtenu, une fois ces paramètres optimisés, peut être évalué sur le jeu de test. On parle alors de validation externe. La méthode dite de la randomisation d'Y (*Y-scrambling*), où Y représente la variable d'activité, peut également être utilisée, en supplément de la validation croisée. Elle consiste à mélanger les valeurs d'activité pour chacune des molécules. En générant un modèle sur ce jeu de données « imaginaires », les R^2 et Q^2 sont censés être très inférieurs à ceux du modèle initial. En effet, obtenir des paramètres équivalents, voir supérieurs serait le signe d'un surapprentissage du modèle, ainsi que d'une absence totale de relation structure-activité. Enfin, un modèle ne saurait être considéré comme valide avec certitude, si certaines de ses prédictions ne sont pas évaluées expérimentalement. Malheureusement, sans doute pour des questions de coûts, cette ultime validation n'est que rarement faite. Selon nous, une alternative économique pourrait être de prédire les interactions d'un jeu de données étranger à celui utilisé aussi bien pour générer le modèle que pour la validation externe. Il se pose alors la question de savoir si ces données sont comprises dans le domaine d'applicabilité du modèle.

[115] Tropsha, A. *et al.* (2003), The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models, *QSAR Comb. Sci.*, 22, 69-77.

Dans les articles pris en exemple, les auteurs ont pour la plupart effectués une LOO sur le jeu d'apprentissage suivit d'une validation externe sur les données du jeu de test. Seule l'étude la plus récente évalue son modèle avec 10000 randomisations d'Y.[112]

v. Le domaine d'applicabilité

L'étude du domaine d'applicabilité consiste à déterminer l'espace chimique délimité par les molécules du jeu d'apprentissage et à évaluer si les molécules qui seront évaluées par le modèle en font partie. Bien que le problème du domaine d'applicabilité ait été abordé depuis de nombreuses années, beaucoup d'articles présentant des modèles QSAR appliqués aux inhibiteurs de protéines kinases ont éludé la question, en témoignent les modèles présentés dans la Table 8.[115,101] Plusieurs méthodes pour calculer le domaine d'applicabilité ont déjà été présentées, sans qu'aucune ne soit véritablement adoptée à l'unanimité. Parmi ces méthodes, nous allons nous attarder sur deux d'entre elles.

L'approche dite par *bounding box* consiste à définir un hyper rectangle à *p* dimensions (*p* étant le nombre de descripteurs) en se basant sur les valeurs maximales et minimales de chaque descripteur utilisé dans le modèle. Cette technique présente l'inconvénient de ne pas prendre en compte la corrélation entre les différents descripteurs. Il sera alors préférable, au préalable, de réaliser une analyse en composantes principales (ACP) sur les descripteurs, afin d'en réduire la dimension. Le problème est alors de choisir le nombre de composantes à garder. De plus, l'approche par *bounding box*, qu'elle soit appliquée directement sur les descripteurs, ou bien après réduction par ACP, peut gommer des vides au sein des descripteurs. En effet, un descripteur peut ne pas prendre toutes les valeurs de l'intervalle sur lequel il est défini.

La méthode par similarité (ou par distance) repose sur la comparaison des descripteurs des molécules du jeu de test avec ceux des molécules du jeu d'apprentissage. Le point critique est le choix d'un seuil qui permet de discriminer les molécules dans le domaine d'applicabilité, de celles à l'extérieur. La similarité se calcule différemment en fonction des données à comparer. Dans le cas de molécules décrites par des *fingerprints*, la similarité entre les molécules est souvent mesurée à l'aide du coefficient de Tanimoto qui repose sur le nombre de 1 (la présence d'une caractéristique) partagés. Celui-ci est spécifiquement conçu pour comparer des données binaires et s'échelonne entre 0 et 1, 1 indiquant une similarité parfaite. Dans le cas où les descripteurs de molécules sont des valeurs réelles, de nombreuses métriques existent pour calculer la distance entre molécules (Euclidienne, Manhattan, Cosine).[116] Une autre approche consiste à utiliser la méthode des k plus proches voisins (ou *k nearest neighbors)* pour prédire la similarité de nouvelles molécules avec celles du jeu d'apprentissage. Elle revient à calculer les distances

[116] Weaver, S., Gleeson, M. P. (2008), The importance of the domain of applicability in QSAR modeling, *J. Mol. Graph. Model.*, 26, 1315-26.

entre les premières et leurs k plus proches voisines du domaine d'applicabilité. Enfin, citons également le calcul de distance dite *leverage* qui consiste à calculer la valeur h_i pour chaque molécule du jeu de test (Figure 37), où x_i représente les descripteurs de la molécule à tester et X la matrice globale des descripteurs.

$$h_i = x_i^T (X^T X)^{-1} x_i$$

Figure 37 : Définition du calcul de h_i

Parmi les modèles QSAR présentés, seuls deux présentent une évaluation du domaine d'applicabilité. Les articles de Kirubakaran *et al.*, ainsi que Comelli *et al.* font tous deux mention du calcul du *leverage* pour chacune des molécules du jeu de test.[109,112]

vi. Inconvénients des modèles QSAR

Les approches par modèle QSAR présentent un intérêt certain lorsqu'il s'agit d'étudier une seule protéine et un nombre restreint de molécules. Cependant, dès que nous sortons de ce cadre fixé, des limites importantes apparaissent.

La limite principale de l'approche QSAR est sans conteste qu'elle ne tient compte des interactions de molécules que sur une seule protéine à la fois. De ce fait, elle est incapable de prédire les interactions de ces mêmes molécules sur une autre protéine. Dès lors, l'étude de molécules sur plusieurs protéines cibles requiert d'avoir suffisamment de données expérimentales pour établir un modèle QSAR pour chacune d'entre elles. Malheureusement, c'est rarement le cas, en particulier pour des cibles nouvelles. Celles-ci souffrent généralement d'un manque d'expériences réalisées.

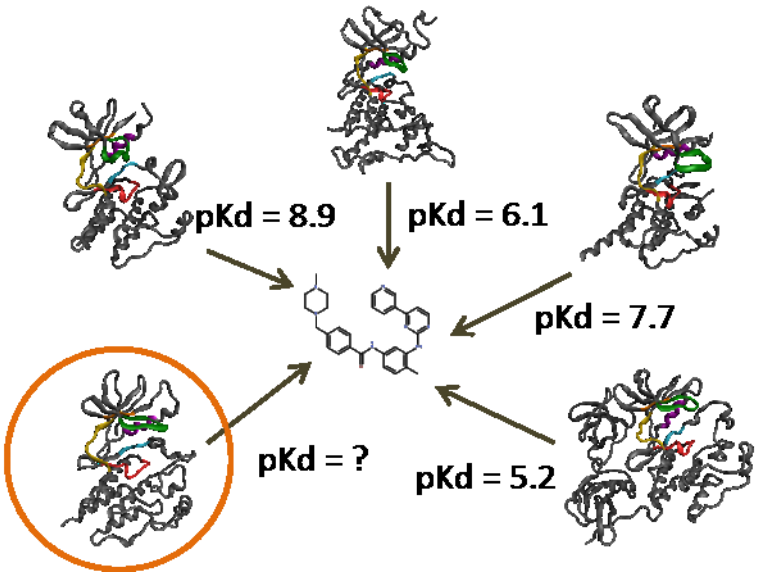
La seconde limite en QSAR est l'utilisation unique des descripteurs de molécules. Or l'interaction d'un ligand sur sa cible dépend, bien entendu, du ligand en lui-même, mais également du site actif de la cible. Un modèle ne va donc pas être capable de décrire efficacement tous les aspects des interactions en jeu. Cet aspect intervient également dans la limite du domaine d'applicabilité. Celui-ci est restreint car limité aux molécules proches de celles des jeux d'apprentissage et de test.

Enfin, en plus des problèmes de prédiction, se pose la question de la sélectivité. Celle-ci découlant de l'activité de molécules sur plusieurs cibles, il semble évident que l'approche QSAR n'est pas la meilleure méthode pour étudier cet aspect.

4. La protéométrie

Les approches protéométriques sont aux protéines, ce que les approches QSAR sont aux molécules. En effet, ce domaine s'attelle à l'étude et à la compréhension des interactions de plusieurs protéines sur une seule molécule (Figure 38). Très proche dans sa définition des approches de Modélisation Quantitative Séquence à Activité (QSAM), nous préférons cependant distinguer les deux démarches. Plusieurs études basées sur des modèles QSAM ont déjà été présentées et leurs objectifs peuvent différer. Par exemple, le terme QSAM a initialement été employé pour décrire des outils prédictifs appliqués aux séquences d'ADN.[117] Plus récemment, il a été utilisé dans le cadre de la prédiction de l'activité de peptides sur des cibles macromoléculaires.[118,119] Le terme QSAM tend donc à désigner des modèles QSAR dont l'objectif est de prédire les interactions entre une cible et de grosses molécules. De notre point de vue, la protéométrie doit permettre de comprendre quelles sont les propriétés des résidus responsables des interactions protéines-ligands, en regardant du côté des cibles protéiques et non plus du côté des ligands.

Aucun exemple répondant à notre définition n'a été jusqu'à alors présenté dans la littérature. Néanmoins, nous verrons dans la partie IV.B une approche protéométrique développée au sein de notre laboratoire, dans le but d'identifier les résidus des protéines kinases considérés comme importants pour la liaison d'inhibiteurs, en particulier des inhibiteurs de Type II.



[117] Jonsson, J. *et al.* (1993), Quantitative sequence-activity models (QSAM)--tools for sequence design, *Nucleic Acids Res.*, 21, 733-39.
[118] van Westen, G. J. P. *et al.* (2013), Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets, *J. Cheminform.*, 5, 41.
[119] Zhou, P. *et al.* (2010), Gaussian process: an alternative approach for QSAM modeling of peptides, *Amino Acids*, 38, 199-212.

Figure 38 : Schéma représentant les interactions entre plusieurs protéines et une molécule. Les interactions sont dans cet exemple caractérisées par des valeurs logarithmiques de constantes de dissociation (pKd). Dans notre exemple, la protéométrie s'attelle à utiliser les valeurs connues d'affinité pour prédire celle d'une autre protéine, entourée en orange.

Précisons tout de même qu'un modèle protéométrique partage avec un modèle QSAR les mêmes paramètres à prendre en compte. Une attention toute particulière doit, là aussi, être apportée au choix du jeu de données employé. Celui-ci doit bien entendu être le plus fiable possible afin de pouvoir accorder un maximum de confiance aux résultats. Un choix doit également être fait sur les descripteurs de protéines. Une analyse comparative de treize descripteurs d'acides aminés a été effectuée récemment par van Westen *et al*.[118] En effectuant une ACP, les auteurs montrent que plusieurs descripteurs appartiennent aux même clusters (MSWHIM avec T-scales et ST-sclaes ; VHSE avec FASGAI et ProtFP). Finissons en ajoutant que les mêmes méthodes d'apprentissage qu'en QSAR peuvent également être appliquées ici, tout comme la validation (R^2 et Q^2) et l'évaluation du domaine d'applicabilité.

5. La protéo-chimiométrie (PCM)

a. Historique et définition

Introduite en 2001 par Lapinsh *et al.*, la modélisation par approche PCM peut être considérée comme un mélange des deux approches présentées précédemment (Parties II.E.3 et II.E.4), où les activités biologiques à modéliser sont associées à des paires protéine-ligand.[120,121] En prenant à la fois en compte les descripteurs des protéines et ceux des molécules, la PCM doit permettre de détailler avec un maximum de précision les processus mis en place au cours de la liaison des ligands à leurs cibles. Beaucoup d'espoirs sont fondés sur la capacité de la PCM à permettre une meilleure compréhension des interactions protéines-ligands, pour mieux les appréhender. La PCM est particulièrement adaptée pour répondre à des questions de sélectivité, ainsi que pour rapidement estimer la polypharmacologie des molécules.[122]

Parce qu'une telle approche considère plusieurs protéines, différenciées par leurs descripteurs respectifs, elle est plus apte que l'approche QSAR, à réaliser de bonnes prédictions entre des molécules du jeu d'apprentissage et des protéines du jeu de test (Figure 39 A et B). En revanche, les mêmes problèmes qu'en QSAR existent lorsqu'il s'agit de prédire des interactions d'une paire, pour laquelle, soit la protéine, soit la molécule est absente du jeu d'apprentissage. Dès lors, une évaluation du domaine d'applicabilité est nécessaire afin de s'assurer du niveau de confiance des valeurs prédites (Figure 39 C).

---

[120] Lapinsh, M. *et al.* (2001), Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions, *Biochim. Biophys. Acta - Gen. Subj.*, 1525, 180-90.

[121] Lapinsh, M. *et al.* (2005), Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions, *Bioinformatics*, 21, 4289-96.

[122] Cortés-Ciriano, I. *et al.* (2015), Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects, *MedChemComm*, 6, 24-50.

Aujourd'hui, de nombreuses familles de protéines ont été étudiées par approche PCM. Nous avons notamment dénombré des RCPGs,[121,123,124] des protéines impliquées en épigénétique,[125] des protéines virales,[126] des cyclooxygénases,[127] des protéases,[128] et bien sûr les kinases. La prochaine partie s'intéressera uniquement à cette dernière famille de protéines.

b. Les approches PCM appliquées à la découverte d'inhibiteurs de protéines kinases

La famille des protéines kinases représente un sujet d'étude particulièrement intéressant pour l'approche PCM. En effet, des études ont montré que les protéines kinases qui partagent plus de 60% d'identité dans leur séquence, ont tendance à être inhibées par les mêmes molécules.[129] Cela prouve bien que, pour comprendre les phénomènes de sélectivité au sein des protéines kinases, l'espace chimique et l'espace biologique doivent être analysés simultanément.

Nombreux sont les modèles PCM à avoir été développés au cours de ces cinq dernières années.[122] Néanmoins, d'après nos observations et d'après une étude récente,[122] seulement quatre modèles avaient pour objectif l'étude des protéines kinases (Table 10). Ainsi, en dépit de l'intérêt important autour de cette famille de protéine, en particulier, pour la recherche d'inhibiteurs sélectifs, il semble que malgré ses promesses, la PCM ne se soit pas encore totalement exploitée et que des études supplémentaires doivent être effectuées. Quelles sont les raisons de cet accueil mitigé ? Premièrement, la PCM est une méthode relativement ressente avec environ 15 années d'existences. Ce chiffre est à comparer aux plus de 50 ans du QSAR. Deuxièmement, le manque de jeux de données adaptés à la PCM peut également expliquer le faible nombre d'études entreprises. A l'heure actuelle, à quelques exceptions près, seules les grands groupes pharmaceutiques ont les ressources techniques et financières suffisantes pour générer des jeux de données comprenant un grand nombre de composés testés sur une portion significative du kinome. Cependant, dans la plupart des cas, elles gardent ces données de manière confidentielles pour des raisons stratégiques. Nous allons voir dans la prochaine partie qu'il existe plusieurs

---

[123] Jacob, L. *et al.* (2008), Virtual screening of GPCRs: An in silico chemogenomics approach, *Bmc Bioinf.*, 9, 363.
[124] Weill, N., Rognan, D. (2009), Development and validation of a novel protein−ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands, *J. Chem. Inf. Model.*, 49, 1049-62.
[125] Wu, D. *et al.* (2012), Screening of selective histone deacetylase inhibitors by proteochemometric modeling, *BMC Bioinf.*, 13, 212.
[126] van Westen, G. J. P. *et al.* (2011), Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development., *PLoS One*, 6, e27518.
[127] Cortés-Ciriano, I. *et al.* (2015), Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling, *J. Cheminf.*, 7, 1.
[128] Ain, Q. U. *et al.* (2014), Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features, *Integr. Biol.*, 6, 1023-33.
[129] Vieth, M. *et al.* (2005), Kinomics: characterizing the therapeutically validated kinase space, *Drug Discov. Today*, 10, 839-46.

données rendues publiques pouvant permettre de créer des modèles PCM centrés sur les protéines kinases.

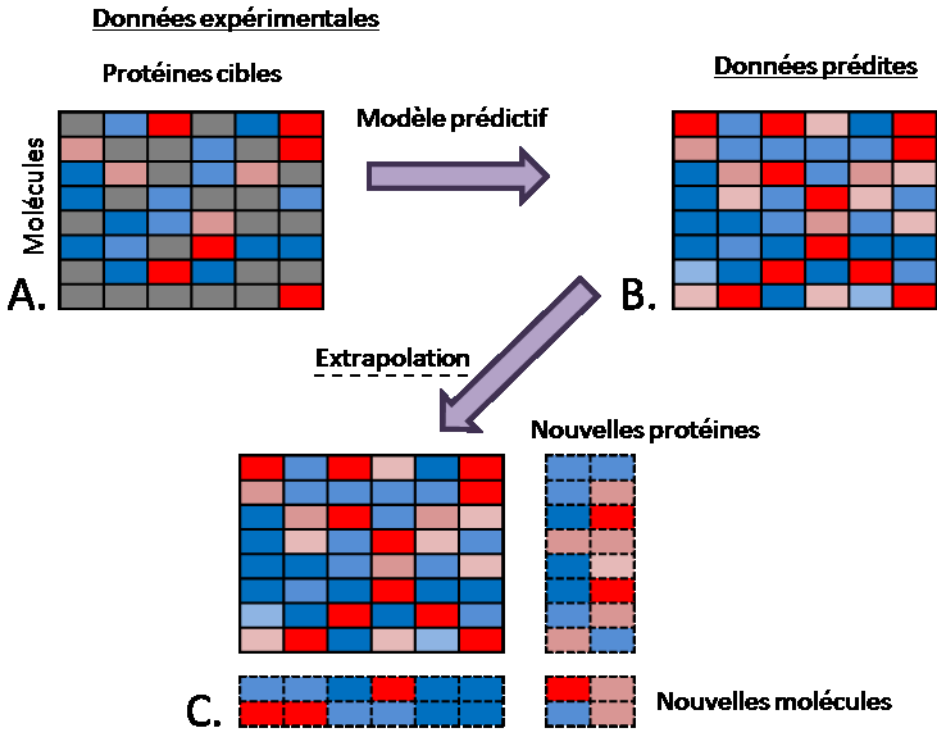


Figure 39 : Espace des interactions protéines-ligands représenté sous la forme de tables avec en ligne les molécules et en colonne les protéines. Chaque interaction protéine - molécule est représentée par une couleur (rouge : bonne | bleue : mauvaise). A/ Espace tel que défini dans le jeu d'apprentissage, avec présence éventuelle de valeurs manquantes. B/ Espace tel que défini après avoir appliqué une approche PCM sur les valeurs manquantes. C/ Espace après extrapolation par utilisation du modèle pour prédire des interactions entre des molécules et/ou des protéines absentes de l'espace initial.

i. Les jeux de données adaptés à la PCM sur les protéines kinases

Il existe plusieurs sources de données qu'il est possible d'utiliser pour développer des approches PCM autour des protéines kinases. Concernant les protéines kinases, d'un côté, comme nous l'avons vu dans la partie II.D.3, beaucoup d'entre elles ont été cristallisées, de l'autre, leur séquence primaire est maintenant connue.[32] Les données indispensables afin de générer des descripteurs de protéines sont donc disponibles. Au sujet des molécules, les chimiothèques contiennent toutes les informations nécessaires afin de récupérer leur structure. Cependant, les données d'inhibition et d'interaction peuvent encore paraître insuffisantes.

Les données d'inhibition (ou d'affinité) ont vu leur nombre croître considérablement au cours des dernières années. En effectuant une recherche dans la base de données ChEMBL (version 18), nous avons récupéré plus de 524000 paires protéines kinases – ligands, tous types de données confondus. En nous limitant seulement aux constantes d'inhibition (Ki), aux IC50 et aux constantes de dissociation (Kd), environ

261000 paires ont été retrouvées. Néanmoins, ce chiffre ne doit pas faire oublier qu'une grande inégalité de traitement existe en fonction des protéines kinases (Figure 40). Ainsi, un gouffre de connaissance sépare les protéines très étudiées, telles que MAPK1, EGFR ou ABL1, des protéines pour lesquelles peu d'activités biologiques sont accessibles. (BUB1, ARAF, MAST1…).

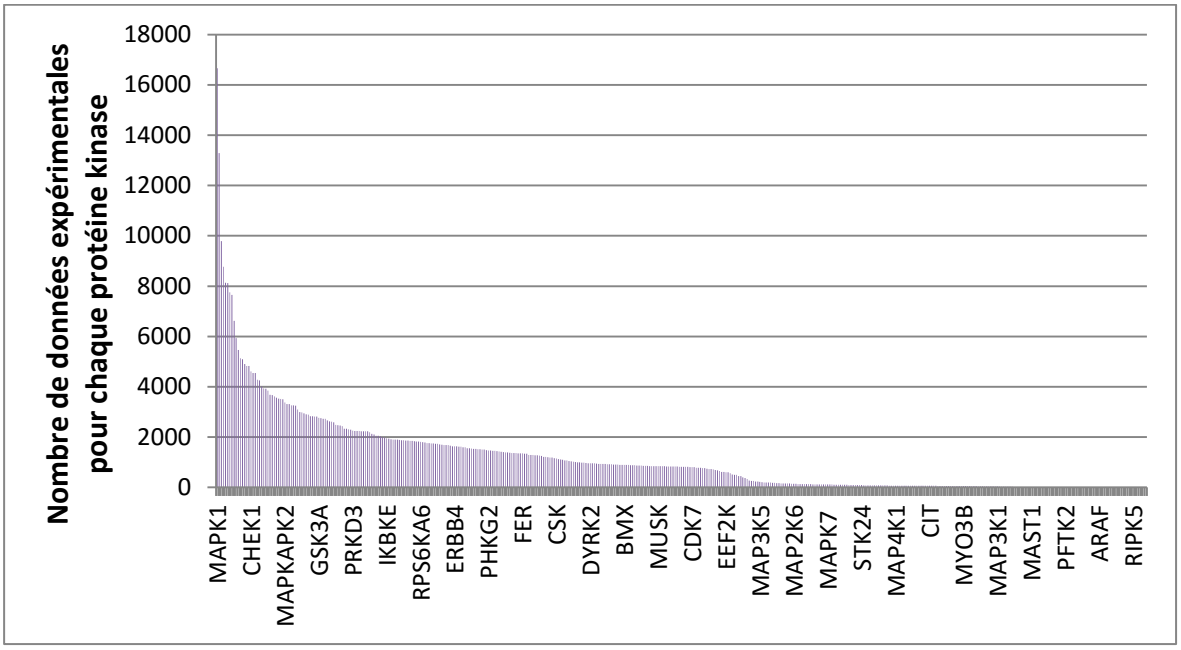


Figure 40 : Nombre de données expérimentales relatives à l'activité biologique (tous types confondus) retrouvées dans la base de données ChEMBL (version 18) pour chaque protéine kinase humaine référencée.

Référence	Méthode	Jeu de données	Descripteurs protéines	Descripteurs molécules	Validation	Meilleurs résultats	Domaine d'applicabilité
Fernandez *et al.*[130]	Machines à vecteurs de support (SVM)	62 protéines et 2233 molécules (matrice complète à 2,6%)	*Sequences-based structural fragments and amino acid sequence autocorrelation*	Enumération de fragments structuraux, vecteurs d'autocorrélation	CV, validation externe	Sensibilité : 0,87 Spécificité : 0,78 Précision : 0,81	Non estimé
Lapins *et al.*[131]	Arbre décisionnel, plus proches voisins, SVM, forêt aléatoire (RF), PLS	317 protéines et 38 molécules (matrice complète à 100%)	z-scales, CTD, auto et cross-covariances des z-scales	Physico-chimiques, géométriques, moléculaires	Double CV	Q^2 : 0,73	Non estimé
Cao *et al.*[132]	RF, Naïve Bayésienne	372 protéines et 22229 molécules (matrice complète à 0,7%)	Proportion de chaque acide aminé dans les séquences et descripteurs CTD	*Fingerprints* topologiques	Partitionnement, CV, validation externe	Sensibilité : 0,91 Spécificité : 0,94 Précision : 0,92	Non estimé
Subramanian *et al.*[133]	PLS	50 protéines et 80 molécules (matrice complète à 24%)	*Knowledge-based field*, watermap	Mold², open babel, volsurf	CV, LOO, validation externe, randomisation d'Y, validation prospective	R^2 : 0,66 Q^2 : 0,47 RMSE : 0,63	Non estimé

Table 10 : Quatre exemples d'approches PCM développées autour des protéines kinases. CV : *cross-validation* ; LOO : *Leave-One-Out* ; CTD : composition, transition et distribution

[130] Fernandez, M. *et al.* (2010), Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines, *J. Chem. Inf. Model.*, 50, 1179-88.
[131] Lapins, M., Wikberg, J. E. S. (2010), Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques, *BMC Bioinf.*, 11, 339.
[132] Cao, D.-H. *et al.* (2013), Large-scale prediction of human kinase–inhibitor interactions using protein sequences and molecular topological structures, *Anal. Chim. Acta*, 792, 10-18.
[133] Subramanian, V. *et al.* (2013), Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics, *J. Chem. Inf. Model.*, 53, 3021-30.

D'autre part, rares sont les études ayant criblé un grand nombre de molécules sur un large panel de protéines. En effet, peu d'équipes académiques possèdent suffisamment de moyens pour tester expérimentalement des milliers de molécules sur, ne serait-ce que la moitié du kinome humain. Le coût d'un point pour une molécule varie de 4 à 43 euros selon l'entreprise de criblage et le protocole expérimental, mais ce prix peut évoluer en fonction du nombre de molécules testées (σ ± 13€). Ces tests à grande échelle ont pourtant un intérêt crucial, notamment en PCM, car ils permettent de s'assurer que les mesures ont été réalisées dans les conditions les plus similaires possibles et limitent ainsi la variabilité au sein des données. Une étude s'est attelée à réaliser des modèles QSAR sur différents jeux de données. Les auteurs ont noté que les résultats obtenus diffèrent d'un jeu de données à l'autre, bien qu'ils partagent nombre de paires protéines kinases – ligands identiques. Cette observation provient de la variabilité entre les jeux de données.[134] Les auteurs concluent alors que les résultats obtenus à partir de modèles basés sur des combinaisons de jeux de données doivent être considérés avec tout le recul nécessaire.

Notons quand même, que plusieurs jeux de données de profilage ont été publiés dernièrement (Table 11). Ceux-ci diffèrent aussi bien par les protocoles expérimentaux utilisés que par les méthodes expérimentales, que par le nombre et la sélection des protéines ou encore que par le nombre et la sélection des molécules testées. Enfin, ajoutons qu'une grande quantité de données est conservée dans les bases de données de la plupart des grands groupes pharmaceutiques à des fins de recherche et développement. Bien que ces données soient confidentielles pour des raisons de propriété intellectuelles, il arrive parfois, que des entreprises décident de partager une partie de leur connaissance avec la communauté scientifique (GSK).[135]

Les jeux de données de profilage, homogènes si pris individuellement, sont une mine d'informations considérable pour l'étude des interactions entre les protéines kinases et les molécules testées. Ils sont donc idéals pour la conception de modèles statistiques et donc, bien entendu, conviennent tout particulièrement aux approches PCM.

Parmi les quatre approches PCM que nous avons recensées, deux d'entre elles ont utilisé le jeu de données publié par Karaman *et al.* (Table 11).[131,133] L'étude publiée par Fernandez *et al.*[130] a utilisé des données issues de la base de données ProLINT.[136] Cette base contient des dizaines de milliers de données d'interactions concernant différentes familles de protéines et sont extraites de la littérature. L'étude de Cao *et al.*[132] a, quant à elle, choisi d'utiliser des données d'interaction provenant de Kinase SARfari, un sous ensemble de la ChEMBL ne contenant que des données sur les protéines kinases provenant de la littérature scientifique.

---

[134] Sheridan, R. *et al.* (2009), QSAR Models for Predicting the Similarity in Binding Profiles for Pairs of Protein Kinases and the Variation of Models between Experimental Data Sets, *J. Chem. Inf. Model.*, 49, 1974-85.
[135] Drewry, D. *et al.* (2014), Seeding Collaborations to Advance Kinase Science with the GSK Published Kinase Inhibitor Set (PKIS), *Curr. Top. Med. Chem.*, 14, 340-42.
[136] Ahmad, S. *et al.* (2003), Protein-Ligand Interactions: ProLINT Database and QSAR Analysis, *Genome Inf.*, 14, 537-38.

En choisissant de collecter un maximum de données, les auteurs ont également fait le choix de générer beaucoup de vide dans leur ensemble de données. En effet, avec 372 protéines et 22229 molécules, mais seulement 54012 données d'interactions, ils ont obtenu une matrice vide à plus de 99%. Le risque d'une telle approche est que le modèle ne puisse pas correctement intégrer les données de similarité liant les paires protéines – ligands, s'exposant ainsi à un phénomène de surapprentissage. De plus, dans ce cas comme dans le précédent, les auteurs créent leurs jeux d'apprentissages et de test à partir de données hétérogènes provenant de divers expérimentateurs.

Référence	Nombre de protéines	Nombre de molécules	Matrice complète	Données mesurées
Bain *et al*.[137]	80	65	Non	%inh
Fedorov *et al*.[138]	60	156	Non	*Thermal shift*
Karaman *et al*.[139]	317	38	Oui	Kd
Anastassiadis *et al*.[140]	300	178	Non	%inh
Davis *et al*.[75]	442	72	Oui	Kd
Metz *et al*. [141]	158	1458	Non	Ki
Drewry *et al*.[135]	200	364	Non	%inh

Table 11 : Aperçus des jeux de données publics d'interactions protéines kinases – ligands. %inh : pourcentage d'inhibition ; Kd : constante de dissociation.

ii. Les descripteurs

Tout comme en QSAR, le développement de modèles PCM nécessite que les données à modéliser soient décrites. Toutefois, en plus des descripteurs de molécules, l'utilisation de descripteurs de protéines est également obligatoire.

Descripteurs de ligands. Nous avons déjà vu dans la partie consacrée aux modèles QSAR (Partie II.E.3.b.ii) que différents types de descripteurs moléculaires existent. Ceux-ci peuvent, bien entendu, être

[137] Bain, J. *et al.* (2007), The selectivity of protein kinase inhibitors: a further update, *Biochem. J.*, 408, 297-315.
[138] Fedorov, O. *et al.* (2007), A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases, *PNAS*, 104, 20523-28.
[139] Karaman, M. W. *et al.* (2008), A quantitative analysis of kinase inhibitor selectivity, *Nat. Biotechnol.*, 26, 127-32.
[140] Anastassiadis, T. *et al.* (2011), Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity, *Nat. Biotechnol.*, 29, 1039-45.
[141] Metz, J. T. *et al.* (2011), Navigating the kinome, *Nat. Chem. Biol.*, 7, 200-02.

également utilisés en PCM. De manière spécifique aux exemples de modèles PCM, nous avons noté l'utilisation de descripteurs structuraux des ligands.[130] Le logiciel Dragon permet, entre autres, de compter un nombre défini de fragments présents dans les structures moléculaires.[142] Ce même logiciel a également été utilisé au cours d'une autre approche, mais cette fois pour calculer divers paramètres structuraux, physico-chimiques et géométriques, sur des ligands dont la conformation la plus stable avait été initialement générée.[131] Un exemple montre l'utilisation de *fingerprints* topologiques.[132] Enfin, le dernier exemple en date a testé différents descripteurs 1D, 2D et 3D des ligands.[133] L'exemple de Lapins *et al.*,[131] et celui de Subramaniane *et al.*[133] introduisent donc une composante 3D au sein des descripteurs moléculaires. Celle-ci est censée mieux décrire le mode d'interaction des ligands. Néanmoins, les résultats de Subramaniane *et al.* montrent des performances inférieures à celles obtenues avec un *fingerprint*.

Descripteurs de protéines. Nous introduisons ici les descripteurs de protéines. Nous l'avons vu précédemment (II.E.4), il existe plusieurs manières de décrire la séquence des protéines. Une étude s'est déjà intéressée à comparer certaines d'entre elles.[118] Nous allons voir ici les descripteurs déjà utilisés par le passé afin de décrire les protéines kinases.

Fernandez *et al.* ont utilisé des descripteurs structuraux de fragments basés sur la séquence en acides aminés. Afin d'éviter de générer un alignement des séquences de protéines kinases, les auteurs ont fait le choix d'utiliser des vecteurs d'autocorrélation. En effet, l'emploi de ces vecteurs est indépendant de l'index initial des acides aminés.

Lapins *et al.*[131] ont expérimenté plusieurs façons de décrire les séquences des protéines, dépendantes ou non de leur alignement. Ils ont en particulier fait usage des descripteurs z-scales introduits par Sandberg *et al*.[143] Ces descripteurs ont été obtenus en effectuant une ACP sur 26 descripteurs physico-chimiques mesurés sur 87 acides aminés dont les 20 acides aminés constitutifs des protéines. Les trois premières composantes principales permettent à elles seules d'expliquer 70% de la variance, tandis que l'utilisation des cinq premières permet d'atteindre 87% de la variance. Chacune des composantes décrit plus particulièrement certaines propriétés. Ainsi z1 explique l'hydrophobicité, z2 le volume et la polarisabilité, z3 la polarité. Le détail pour z4 et z5 est plus vague du fait de la condensation des variables générées par l'ACP, mais il semble que ces composantes décrivent plus particulièrement les propriétés électroniques des acides aminés. L'utilisation des z-scales requièrt, au préalable, que les séquences primaires des protéines aient été alignées. D'une manière similaire à l'article de Fernandez *et al.*,[130] les auteurs ont également testé l'emploi de vecteurs d'autocorrelation. Parmi les autres descripteurs testés, ils ont également analysé l'effet des descripteurs de

[142] Mauri, A. *et al.* (2006), DRAGON Software: An Easy Approach to Molecular Descriptor Calculations, *MATCH*, 56, 237-48.
[143] Sandberg, M. *et al.* (1998), New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids, *J. Med. Chem.*, 41, 2481-91.

composition, transition et distribution (CTD) des acides aminés.[144] Ces descripteurs sont, eux aussi, indépendants de l'alignement des séquences. La partie composition calcule les pourcentages de différentes classes prédéfinies et basées sur sept propriétés des acides aminés, chacune d'elle étant subdivisée en trois sous catégories : hydrophobe, neutre, polaire. La partie transition représente les fréquences avec lesquelles les classes changent durant le parcours de la séquence (par exemple : fréquence de passage d'un acide aminé hydrophobe à un acide aminé neutre). Enfin, la partie distribution s'intéresse à la distribution des classes et des attributs le long de la séquence. Au final, les auteurs ont pu étudier l'effet de chacun des descripteurs testés sur leur modèle.

Dans leur article, Cao *et al.* ont utilisé les descripteurs CTD en plus de calculer le pourcentage présence de chacun des vingt acides aminés dans chaque séquence des protéines kinases de leur modèle.

D'après nos informations, Subramanian *et al.* sont les premiers à avoir utilisé des descripteurs tridimensionnels de protéines dans une approche PCM appliquée aux protéines kinases. Les modèles basés sur des descripteurs de protéines 3D ont déjà montré des performances supérieures à des modèles 2D.[145] Dans le cadre de cette étude, les descripteurs utilisés rappellent les descripteurs moléculaires CoMFA et CoMSIA (ii), à la différence cette fois, que la grille est disposée autour du site de liaison de chacune des protéines kinases, au préalable alignées structuralement entre elles. Ils ont ainsi pu calculer les champs polaires et hydrophobiques à l'aide de sondes atomiques situées à chaque point de la grille. En plus de ces champs, les auteurs ont également utilisé l'outil WaterMap de Schrödinger qui détermine les endroits favorables aux molécules d'eau au sein des sites de liaison.[146] L'utilisation de ces différentes grandeurs permet d'observer directement l'impact de chacun des descripteurs, sur les structures, et de mieux comprendre la sélectivité des inhibiteurs. Afin de pouvoir utiliser tous ces descripteurs dans leurs modèles, les auteurs ont eu besoin de réduire le nombre de variables et ont choisi d'effectuer une ACP sur les descripteurs de protéines. Néanmoins, les performances des modèles générés lors de cette étude semblent en–deçà de celles obtenues par Lapins *et al.*[131] (Q^2 de respectivement 0,5 et 0,7).

Descripteurs cross-term. Enfin, nous allons aborder le sujet des descripteurs identifiés sous l'expression *cross-term*. Ces descripteurs doivent servir à intégrer une part de non linéarité au modèle, dans le cas où la méthode d'apprentissage est linéaire comme c'est notamment le cas dans une PLS. En effet, certaines composantes d'interaction entre la protéine et le ligand sont considérées comme non linéaires. Le

---

[144] Dubchak, I. *et al.* (1995), Prediction of protein folding class using global description of amino acid sequences, *Proc. Natl. Acad. Sci. USA*, 92, 8700-04.
[145] Meslamani, J., Rognan, D. (2011), Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel, *J. Chem. Inf. Model.*, 51, 1593-603.
[146] Robinson, D. D. *et al.* (2010), Understanding kinase selectivity through energetic analysis of binding site waters, *ChemMedChem*, 5, 618-27.

nombre de liaisons hydrogène qui peuvent être créées dans différentes protéines est clairement un paramètre non linéaire. Les descripteurs *cross-term* sont généralement obtenus en multipliant les descripteurs des protéines et les descripteurs des ligands, augmentant considérablement la dimension du jeu de données.[122]

iii. Les méthodes d'apprentissage utilisées en PCM

Nombreuses sont les méthodes d'apprentissage qui ont été utilisées dans des approches PCM. Nous l'avons vu précédemment (II.E.3.b.iii), ces méthodes permettent, en fonction du type de données à prédire, d'effectuer des régressions ou bien des classifications. Dans le cadre de cette thèse, nous allons nous intéresser aux trois méthodes les plus utilisées dans le cadre des approches PCM appliquées aux protéines kinases.

La méthode des moindres carrés partiels (PLS), est très proche de l'ACP dans son fonctionnement. En effet, l'objectif de la PLS est de trouver, pour un ensemble d'observations n, une relation multifactorielle entre les variables descriptives x_i, et la variable à expliquer, notée Y.[147] Pour y parvenir, la PLS projette les x_i et Y dans un espace formé par des variables latentes x_i' et trouve une relation linéaire entre les x_i' et la variable Y. Là où son fonctionnement se rapproche de celui de l'ACP, c'est que la PLS maximise la variance des variables x_i en utilisant plus ou moins de variables latentes et maximise la corrélation entre x_i et Y. Finalement, la variable Y se retrouve expliquée par une équation du type : $Y = k + a_1x_1' + a_2x_2' + \cdots + a_nx_n'$, où a_i représentent les coefficients des x_i' et *k* est une constante. Plus la valeur de a_i est grande et plus sa variable associée est considérée comme importante par le modèle pour expliquer Y. Cette équation peut ensuite être utilisée afin de prédire de nouvelles valeurs Y, pour de nouvelles observations.

La PLS est souvent utilisée en QSAR comme en PCM. Bien qu'il lui arrive de surclasser d'autres techniques, son fonctionnement linéaire oblige les modélisateurs à générer des descripteurs *cross-term* pour induire de la non linéarité. Le calcul de ces descripteurs ajoute généralement une quantité d'informations non négligeable et leur interprétation est compliquée. Parmi les modèles PCM appliqués aux protéines kinases, Subramian *et al.* ont entièrement basé leurs modèles sur des PLS, en combinant les descripteurs de protéines et les descripteurs de ligands à des *cross-terms.* Lapins *et al.* ont, quant à eux, opté pour une comparaison de plusieurs méthodes d'apprentissage, dont la PLS, mais également les méthodes RF et SVM que nous allons décrire dans les prochaines parties.

***La méthode des forêts aléatoires (random forest* (RF))**, est une méthode non linéaire, à la différence de la PLS. Pouvant au choix être appliquée à de la classification ou à de la régression, RF repose sur des arbres

[147] Wold, S. *et al.* (2001), PLS-regression: a basic tool of chemometrics, *Chemom. Intell. Lab. Syst.*, 58, 109-30.

décisionnels et sur le principe du *bagging*.[148] Ce terme définit une approche qui consiste à combiner plusieurs modèles d'apprentissage afin d'augmenter la précision globale et de moyenner le bruit dans le but de créer un modèle avec une faible variance. En suivant ce principe, l'algorithme sur lequel repose RF va générer plusieurs arbres décisionnels (Figure 41) tous décorrélés les uns des autres. C'est-à-dire que chaque arbre est basé sur un sous échantillonnage unique des variables du jeu de données d'apprentissage. Chaque échantillon est généré aléatoirement. Les observations à prédire sont ensuite évaluées par chaque arbre et le résultat de la prédiction est définit selon un vote, lors d'une classification, ou en faisant la moyenne des valeurs prédites par chaque arbre, dans le cas de la régression. Du fait que les résultats reposent sur des arbres décisionnels, ils sont facilement interprétables. De plus la paramétrisation d'un modèle RF est relativement aisée, l'utilisateur n'a besoin que de choisir le nombre d'arbres décisionnel qu'il veut générer et le nombre de variables, choisies aléatoirement, qu'il souhaite attribuer à chaque arbre. Ces raisons expliquent en grande partie pourquoi cette méthode est particulièrement plébiscitée en PCM, et notamment pour l'étude des protéines kinases.[131,132]

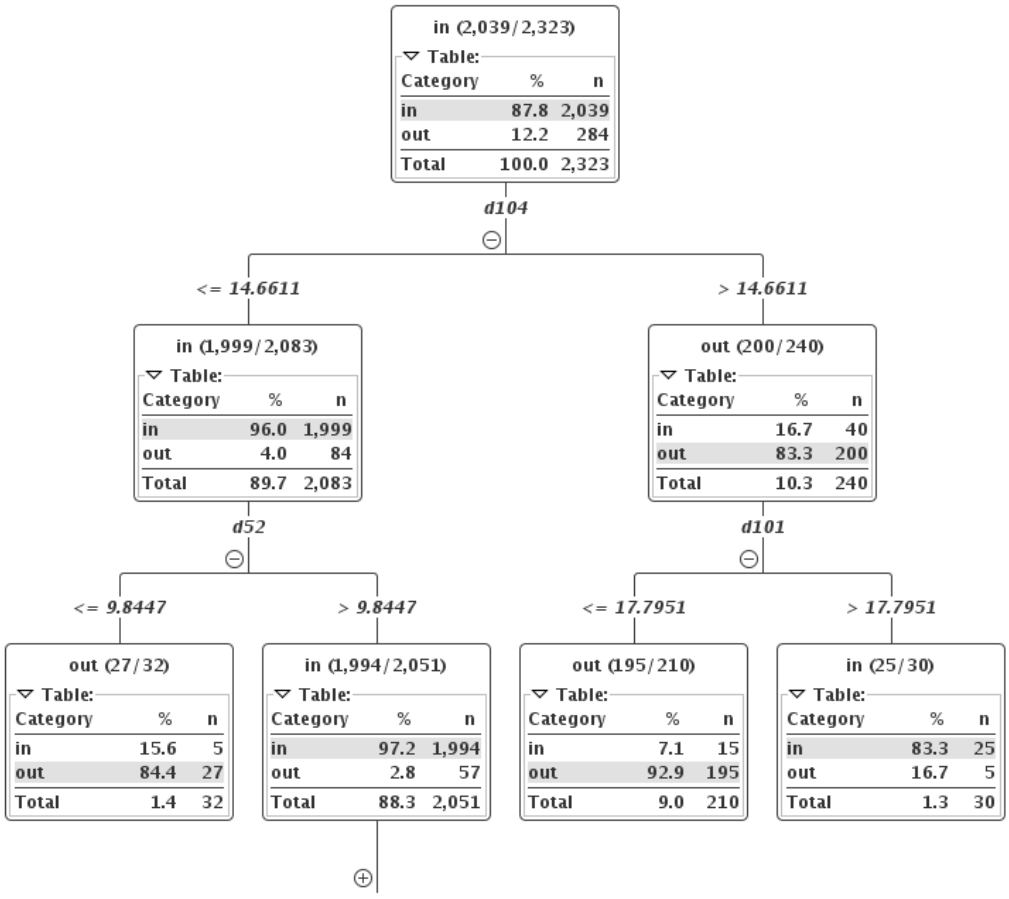


Figure 41 : Exemple d'une sous branche d'un arbre décisionnel. Il a été utilisé pour différencier les conformations *DFG-in/DFG-out*.

[148] Liaw, A., Wiener, M. (202), Calssification and Regression by randomForest, *R News*, 2/3, 18-22.

***Les machines à vecteurs de support* (*Support Vector Machines* (SVM))** forment un groupe de méthodes d'apprentissage supervisées pouvant être appliquées à de la classification comme à de la régression. Tout comme la méthode RF, SVM est une méthode non linéaire. Elle repose sur deux principes. Le premier revient à déterminer un hyperplan capable de séparer correctement toutes les observations. La détermination du meilleur hyperplan se fait en sélectionnant celui qui maximise au mieux les marges entre l'hyperplan et les classes (Figure 42).[149] Pour chaque classe, l'élément le plus proche de l'hyperplan est appelé vecteur de support. Le deuxième principe prend tout son sens lorsque la relation séparant les deux classes n'est pas linéaire. La méthode SVM est dans ce cas, capable de représenter les données dans un espace de plus grandes dimensions (théoriquement infini) dans lequel il sera possible de séparer les classes de façon linéaire. Cette transformation fait appel à une fonction noyau. La fonction la plus utilisé en PCM est la *Radial Basis Function* (RBF), mais d'autres existent telles que la fonction linéaire ou la fonction polynomiale (Figure 42).

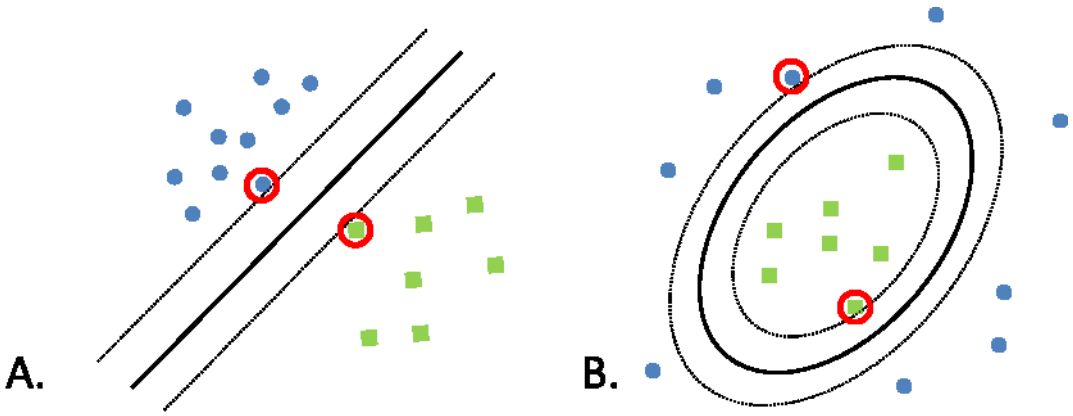


Figure 42 : Exemple de classifications SVM. Les classes sont représentées par des points de couleurs différentes. A/ Classification effectuée avec une fonction noyau linéaire. B/Classification effectuée avec une fonction noyau polynomiale non linéaire. Dans les deux cas, un hyperplan est représenté en un trait noir plein. Les vecteurs supports sont entourés en rouge. Les marges sont représentées en pointillés noirs.

En pratique, pour obtenir le meilleur modèle SVM, il faut d'une part choisir la fonction noyau la plus appropriée, et d'autre part sélectionner les meilleurs paramètres du modèle. Pour ce faire, la méthode la plus souvent utilisée revient à utiliser une grille exponentielle de recherche. Chaque maille de la grille est associée à un couple, voir un triplé de paramètres. Une multitude de modèles vont alors être entrainés en variant les paramètres pour chacun d'eux. A partir du modèle offrant les meilleures performances, il est alors possible de déduire les paramètres optimaux. Associée généralement à une validation croisée, cette approche est coûteuse en temps de calcul et représente l'un des plus gros inconvénients de la méthode. Toutefois, récemment des chercheurs ont publié une approche Bayésienne capable de donner plus rapidement les bons

[149] Ivanciuc, O. (2007), Applications to Support Vector Machines in Chemistry, *Reviews in Computational Chemistry*, 23, 291-400.

paramètres du modèle.[150] Enfin, il est important de noter que les méthodes SVM sont très difficiles à interpréter. En effet, elles ne renvoient pas de coefficients associés aux variables, comme c'est le cas par exemple en PLS.

L'étude de Lapins *et al.*,[131] ainsi que celle de Fernandez *et al.*[130], ont toutes deux fait usage de méthodes SVM. Dans le premier exemple, les auteurs ont comparés plusieurs méthodes d'apprentissage automatique et ont obtenu de meilleurs résultats avec une méthode SVM. Cette observation a également été faite dans le cas d'une étude PCM portant sur d'autres familles de protéines.[151]

iv. Validation et domaine d'applicabilité

Il n'existe quasiment aucune différence entre la validation d'un modèle QSAR et la validation d'un modèle PCM. Les approches détaillées dans la partie II.E.3.b.v s'appliquent également ici. Que ce soit pour une classification ou pour une régression, les mêmes coefficients qu'en QSAR peuvent être calculés pour estimer la performance d'un modèle PCM.

En plus de la validation habituelle, des articles traitant de PCM ont également introduits des études de la contribution des descripteurs sur les performances du modèle. L'idée est notamment d'utiliser la propriété des *fingerprints* circulaires pour étudier l'effet de la présence ou de l'absence d'une certaine sous-structure.[128,151] En effet, pour la totalité des *fingerprints* utilisés dans le modèle, passer la valeur d'un bit donné de 1 à 0 revient à virtuellement retirer la sous-structure correspondante des molécules. Il est alors possible d'entrainer un nouveau modèle sur ces données modifiées et d'observer l'effet sur les valeurs prédites en fonction de la présence ou l'absence de ladite sous-structure (Figure 43).[128]

Le domaine d'applicabilité dans le cas d'un modèle PCM nécessite de s'intéresser à la fois à l'espace chimique des ligands et à l'espace biologique des protéines. Concernant ce dernier, la forte similarité entre toutes les protéines kinases humaines fait qu'elles sont toutes théoriquement dans le domaine d'applicabilité biologique. En ce qui concerne le domaine d'applicabilité chimique, les mêmes approches que celles décrites dans la partie v peuvent s'appliquer. Toutefois, ajoutons que récemment une étude PCM a montré que l'utilisation des Processus Gaussiens, en tant que méthode statistique d'apprentissage, permet d'établir des prédictions tenant compte de la variance du modèle.[151]

---

[150] Czarnecki, W. M. *et al.* (2015), Robust optimization of SVM hyperparameters in the classification of bioactive compounds, *J. Cheminf.*, 7, 38.
[151] Cortes-Ciriano, I. *et al.* (2014), Proteochemometric modeling in a Bayesian framework, *J. Cheminf.*, 6, 35.

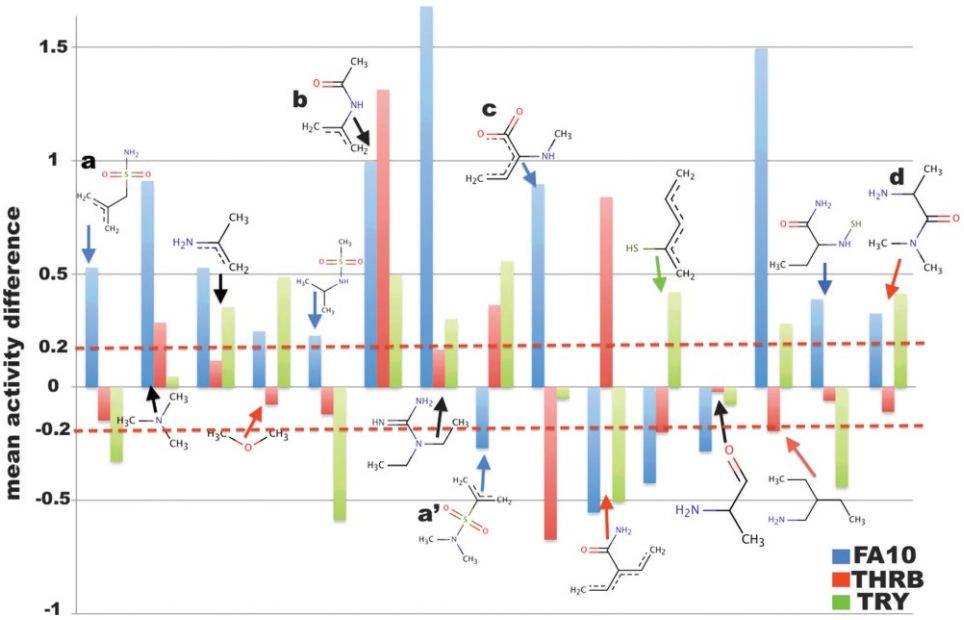


Figure 43 : Effet de la présence ou de l'absence de sous-structures moléculaires, représentées par les bits des *fingerprints*, sur les prédictions du modèle. Les auteurs de la figure présentent les résultats pour le facteur de coagulation X (FA10), la thrombine (THTB) et la trypsine (TRY). L'effet est quantifié via la différence moyenne entre les valeurs prédites avec ou sans la sous-structure en abscisse.

III. Analyse de la sélectivité d'inhibiteurs de protéines kinases basée sur des jeux de données d'activité biologiques

La sélectivité est un aspect délicat à évaluer dans le développement des inhibiteurs de protéines kinases. Il est rapidement apparu comme important de tester les molécules sur une portion du kinome la plus grande et la plus hétéroclite possible, et ce, le plus tôt possible dans le développement du médicament. Bien qu'il soit idéal d'évaluer la sélectivité d'un inhibiteur sur un maximum de protéines kinases humaines, les coûts engendrés (Partie II.E.5.b.i) sont tels qu'il est parfois préférable de sélectionner quelques protéines de chaque groupe. Dans le cas du développement d'un inhibiteur ciblé, il est indispensable de tester la molécule sur la cible d'intérêt (*on-target*), mais aussi sur les protéines avec lesquelles elle ne doit pas interagir (*off-target*). Dans leur article, Van Rompaey *et al.* présentent la caractérisation préclinique d'un inhibiteur de la protéine kinase JAK1 (filgotinib).[152] La difficulté principale a été, semble-t-il, d'inhiber JAK1 sans pour autant inhiber les autres protéines de la famille JAK (JAK2, JAK3 et TYK2). Des essais cliniques ont montré que l'inhibition non désirée de JAK2 peut entrainer une anémie, une neutropénie et une thrombopénie.[153] Les auteurs ont donc mené toutes les étapes du développement en comparant à chaque fois l'activité de l'inhibiteur sur les quatre protéines *in vitro*, mais aussi *in vivo*. Finalement leurs résultats montrent des activités semblables sur JAK1 et JAK2 et supérieurs à celles sur JAK3 et TYK2, mais les expériences menées sur des cellules T montrent une sélectivité trente fois supérieure de la voie dépendante de JAK1 sur la voie dépendante de JAK2. Cet exemple illustre bien les différences que peuvent présenter les tests biochimiques sur des cibles isolées et les tests phénotypiques (Partie II.C)

A. Les jeux de données d'inhibition

Nous avons vu plus haut (Partie II.E.5.b.i) qu'il existe aujourd'hui plusieurs jeux de données d'activité biologique basés sur une grande partie du kinome humain. Ils sont, pour la plupart, le résultat de la mise en place par des sociétés de tests enzymatiques permettant de connaitre rapidement les profils d'activité de molécules. Ces tests peuvent être réalisés aussi bien sur des protéines purifiées que sur des lignées cellulaires. Les sociétés Reaction Biology Corp, DiscoveRx, mais aussi Merck Millipore, communiquent toutes les trois sur plus de 300 protéines kinases disponibles pour du criblage expérimentale et mettent en avant leurs propres technologies. En plus des protéines en conformation active, ces prestataires proposent aussi d'effectuer des tests enzymatiques sur des protéines inactives, ainsi que sur des mutants. Les données rendues publiques par ces entreprises font partie des rares sources d'informations accessibles à toute la communauté scientifique.

---

[152] Van Rompaey, L. *et al.* (2013), Preclinical characterization of GLPG0634, a selective inhibitor of JAK1, for the treatment of inflammatory diseases, *J. Immunol.*, 191, 3568-77.
[153] O'Shea, J. J., Plenge, R. (2012), JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease, *Immunity*, 36, 542-50.

Rassemblés dans un jeu de données unique, ces données peuvent être très utiles pour l'étude des interactions entre les protéines kinases et leurs ligands, mais aussi pour l'établissement de modèles statistiques permettant de prédire de nouvelles interactions.[122] Toutefois, la multiplication des jeux de données entraîne *de facto* des problèmes de standardisation. En effet, il peut être utile de comparer des jeux de données entre eux, ne serait-ce que pour générer une base de données globale ou pour étudier la corrélation qui peut exister entre les valeurs mesurées. Ces travaux sont rendus compliqués par les différences qui sont rencontrées dans les noms des protéines ou dans ceux des molécules employées (Table 12 et Table 13), et qui empêchent d'utiliser des outils informatiques simples pour parcourir les données.

Pour éviter ces désagréments et faciliter la recherche de données d'activité biologique, l'Institut Européen de Bioinformatique (EBI) a créé la base de données ChEMBL regroupant des millions de données d'interactions. [98] Une interaction correspond ici à l'activité ou à l'affinité biologique d'une molécule sur une protéine. Cette base de données est accessible notamment au travers d'un site web (https://www.ebi.ac.uk/ChEMBL/). De plus, son service Kinase SARfari est intégralement centré autour des protéines kinases. Ces deux sources de données sont ouvertes à tous et peuvent s'avérer utiles dans la recherche de données d'interaction, aussi bien en effectuant des recherches depuis la cible, que depuis la molécule. Cependant, nous ne les avons pas trouvées pratiques lorsqu'il a été question d'effectuer des recherches de jeux de données d'inhibition d'une même origine. De même, leur interface manque d'intuitivité soit pour récupérer les informations d'inhibiteurs sur des cibles particulières, soit pour trouver les protéines inhibées par une molécule en particulier.

Millipore	DiscoverX	GSK	Reaction Biology	SARfari	Uniprot_id	Synonymes
Aurora-A	AURKA	AURA	Aurora A	hAURa_2362	AURKA_HUMAN	AIK, ARK1, AURORA2, BTAK, MGC34538, STK15, STK6

Table 12 : Exemples de termes utilisés pour désigner la protéine kinase Aurora-A selon les entreprises Merck Millipore,[154] DiscoverX,[75] GSK,[135] Reaction Biology,[140] ou dans les bases de données SARfari[155] et Uniprot.[89] Des synonymes trouvés dans d'autres jeux de données figurent également dans la table.

[154] Gao, Y. *et al.* (2013), A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery, *Biochem. J.*, 451, 313-28.
[155] Kinase SARfari. https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari (accessed 2014).

PubChem id	DiscoveRx	Reaction biology	Université d'Oxford	ChEMBL	Synonymes
11617559	GW-2580	cFMS Receptor Tyrosine Kinase Inhibitor	Diaminopyrimidine1	CHEMBL261849	CSF-1 Receptor Inhibitor, Kinome_3757, CHEMBL261849, HMS3229E11, HMS3303G12, HMS3305H21, GW632580X, IN1333

Table 13 : Exemples de termes utilisés pour désigner une molécule inhibitrice de la protéine kinase CSF1R selon les entreprises DiscoverX,[75] et Reaction Biology,[140] dans les bases de données PubChem[97] et ChEMBL,[98] et dans un article publié par l'Université d'Oxford.[138] Des synonymes trouvés dans d'autres jeux de données figurent également dans la table.

La base de données ChEMBL peut être téléchargée sous différents formats afin d'être directement intégrée dans des systèmes de gestion de bases de données internes. Cependant, d'une version à l'autre, le schéma de la base de données relationnelle peut être changé par ses auteurs et les requêtes SQL établies au préalable sont alors inopérantes, ce qui pose un problème pour la maintenance du système.

Au sujet de Kinase SARfari, nous avons noté un problème préoccupant. En regardant ses données brutes, il nous est apparu que certains champs devant contenir le nom des protéines étaient corrompus et ne permettaient pas d'identifier la cible. Nous avons contacté les auteurs de Kinase SARfari début 2013 pour leur signaler le problème et il nous a été répondu que le problème serait résolu au plus vite. Cependant, à l'heure actuelle, le problème est toujours présent (Table 14).

ACTIVITY_ID	NAME	COMPOUND_ID	ACTIVITY_TYPE	RELATION	STANDARD_VALUE	STANDARD_UNIT
2368415	hEGFR_1553	120564	Log IC50	=	4.06	
2368416	hEGFR_1553	84512	IC50	>	100000	nM
2368417	Starlite Functional	325156	IC50	=	310	nM
2368418	hEGFR_1553	168266	IC50	=	2820	nM
2368421	hLCK_2238	71541	IC50	<	3	nM
2368422	hEGFR_1553	29197	IC50	=	4	nM
2368423	Starlite ADMET	92	Log 10 cell kill	=	3.3	
2368424	hSRC_2204	67268	IC50	=	4000	nM
2368424	mCSK_2667	67268	IC50	=	4000	nM
2368428	hMEK1_307	327188	IC50	>	1000	nM
2368429	Starlite Functional	428690	IC50	=	500	nM
2368430	Starlite Functional	279472	MST	=	14.3	days

2368431	Starlite Functional	148889	IC50	>	5200	nM
2368432	hPDGFRa_27	7810	IC50	>	10000	nM
2368432	hPDGFRb_23	7810	IC50	>	10000	nM

Table 14 : Aperçu des données d'inhibition extraites depuis Kinase SARfari. Les cellules en rouge montrent les données anormales où devraient, en réalité, être indiqués les identifiants des noms des protéines kinases.

Pour pallier ce problème et disposer d'un outil opérationnel et pratique sur lequel nous aurions la main, nous avons choisi de créer une base de données interne au groupe de Bioinformatique Structurale et Chémoinformatique (SB&C) de l'ICOA.

B. La base de données d'inhibiteurs de protéines kinases

La base de données d'inhibiteurs de protéines kinases n'a pas pour vocation à concurrencer celle de ChEMBL. Elle ne contient, pour l'instant, que les données issues des jeux de données mentionnés dans la Table 11. Cela représente néanmoins environ 570 000 données d'interactions nettoyées et homogénéisées. Pour l'instant, cette base n'est constituée que de deux tables : une pour les activités et une pour les inhibiteurs de protéines kinases testées (2172 entrées uniques) (Table 15).

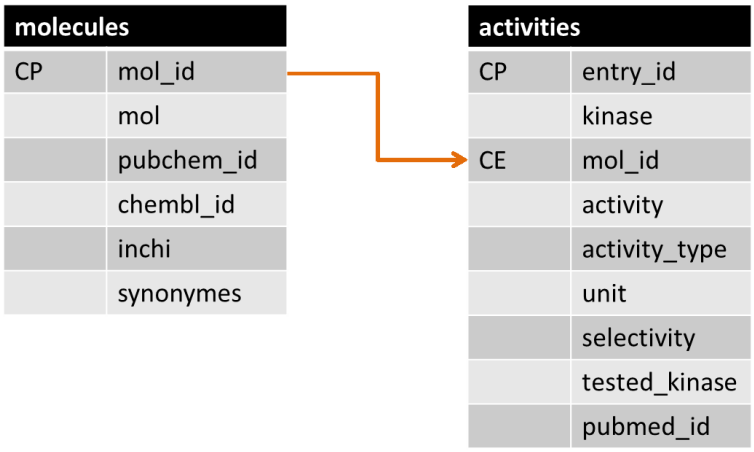


Table 15 : Schéma de la base de données d'inhibiteurs de protéines kinases stockée au sein du groupe de Bioinformatique Structurale et Chémoinformatique de l'ICOA. CP : clé primaire de la table ; CE : clé étrangère de la table.

Pour l'heure, elle ne contient que les données qui ont été nécessaires pour les projets mentionnés dans cette thèse. Ainsi, la table « *molecules* » contient tous les inhibiteurs de protéines kinases testés dans les jeux de données mentionnés dans la Table 11. Chaque entrée est rattachée à un identifiant unique, propre à notre base de données, contenu dans la colonne *mol_id*. En plus, s'ils existent, les identifiants des bases de données PubChem et ChEMBL ont été intégrés en vue de croiser les informations. L'InChi (*International Chemical Identifier*), introduit par IUPAC, est un identifiant standard qui permet un regroupement facile de différentes sources de données de molécules. Il est établi à partir de la structure des molécules. L'avantage de l'InChi est qu'il est théoriquement le même pour deux représentations différentes d'une même molécule. Pour

cela, il numérote de manière canonique les atomes et prend en charge un grand nombre de propriétés chimiques telles que la tautomérie ou la stéréoisomérie. L'intégration de l'InChi dans la base avait été anticipée dans l'hypothèse de la croiser avec des bases de données qui contiendraient également cet identifiant. Ce travail est actuellement en cours dans un autre projet de recherche de l'équipe SB&C. La colonne *synonymes* regroupe les différents identifiants des molécules qui ont pu être rencontrés au cours de la préparation de la base. Enfin, la colonne *mol* contient la structure au format mol.

La table « *activities* » contient toutes les données d'interaction que nous avons récoltées. Chaque paire protéine kinase – ligand testée est rattachée à un identifiant unique. La protéine kinase est identifiée par un nom unique, tandis que la molécule est reconnue par son *mol_id* provenant de la table « molecules ». Les colonnes *activity*, *activity_type* et *unit* contiennent respectivement, la valeur d'activité donnée par la source, le type d'activité (pIC50, pKd, pKi, pourcentage d'inhibition, *thermal shift*) et son unité. La colonne *selectivity* contient une valeur représentant la sélectivité de la molécule telle que calculée sur le panel de protéines testées (*tested_kinase*). Enfin, la colonne *pubmed_id* permet d'identifier l'origine bibliographique de l'activité mesurée. Bien entendu, une paire protéine kinase – ligand a pu être testée dans des sources différentes, mais chaque test possède bien un identifiant unique dans la table.

Afin d'exploiter cette base de données, des *workflows* Knime ont été développés.[156] L'un d'entre eux permet d'identifier rapidement les inhibiteurs pour des protéines kinases d'intérêt, mais aussi les molécules inactives. A partir de ces molécules, une option permet aussi d'identifier celles qui interagissent avec une autre protéine kinase donnée. Un *workflow* a également été développé pour récupérer la structure des molécules dans l'objectif de les cribler virtuellement. Le nombre de molécules choisies peut être réduit en ne gardant que des molécules structuralement distantes entre elles, et ce dans le but de gagner du temps lors du criblage.

Finalement, les jeux de données rassemblés dans cette base de données nous ont permis d'effectuer une étude comparative de la sélectivité des inhibiteurs de protéines kinases. Celle-ci fait l'objet d'une publication en cours de soumission. Nous avons introduit de nouvelles métriques de sélectivité, que nous avons comparée à d'autres, déjà existantes.

C. Article en cours de soumission : *The use of various selectivity scores in kinase research*

[156] Berthold, M. R. *et al.* (2007) KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization; *Springer*; pp 319-326.

The use of various selectivity scores in kinase research

Nicolas Bosc[a], Christophe Meyer[b] and Pascal Bonnet[*,a]

[a] *Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France*
[b] *Centre de Recherche Janssen-Cilag, Campus de Maigremont - CS 10615, 27106 Val de Reuil CEDEX*

[*]Author to whom all correspondence should be addressed:
Pascal Bonnet: Tel: +33 238 417 254, Fax: +33 238 417 254, E-mail: pascal.bonnet@univ-orleans.fr

ABSTRACT

Selectivity of inhibitor is an important issue when developing a new drug. Protein kinases are particularly concerned by this issue because of the high structural similarity they shared. In the worst case, a lake of selectivity may lead to toxicity. This property may be early assessed using data from protein kinase profiling panels. In this article, we proposed two new selectivity metrics. The window score and the ranking score can be applied to standard *in vitro* data (Ki, IC50 or percentage of inhibition), *in cellulo* (percentage of effect, EC50) or even kinetics data (Kd, Kon and Koff). They are both easy to compute and offer different viewpoints to consider molecule selectivity. We performed a comparative analysis of their results with those of already published selectivity metrics to analyze how they could influence the compound selection.

KEYWORDS : Selectivity, protein kinase, kinase inhibitor, Gini score, entropy score, window score, ranking score

INTRODUCTION

The assessment of compound selectivity is of major interest in drug discovery. This important parameter is carefully used in drug discovery projects to prevent potential toxicity of compounds. Selectivity assessment is often used in the definition of Target Product Profile (TPP) of a lead to highlight the potential toxicity of a compound in reference to the targeted protein(s). Poor selectivity profile is one of the reasons of numerous failures in clinical trials.[1] Research for the development of protein kinase inhibitors is particularly concerned by this problem. Protein kinase family represents more than 500 members of the human proteome[2]. Because they all share a well conserved ATP binding site, the design of selective inhibitors remains a challenge and has negatively impacted the progression of drug candidate to late-stage clinical development. The recent failures in early clinical trials of the p38 MAPK inhibitors,[3] as well as the difficulty of a CDK inhibitors (dinaciclib, PD-0332991) to reach phase III[4] clinical trial are noteworthy examples of the selectivity issue in the kinase field. Nonetheless, several studies have also shown that there are benefits to have selective but non-specific inhibitors as drugs.[5,6,7] The success of the multi-targeted kinase inhibitors imatinib and gefitinib shows that the safety of a drug does not rely on an selectivity, as long as its dosage remains in the therapeutic window.[8,9,10] Furthermore, the relative selectivity of kinase inhibitors could be an added value to a compound and bring advantageous polypharmacological treatments in oncology.[11] For instance, dual inhibitors of the PI3K/mTOR pathway have shown to be more efficient on breast cancer cells than specific inhibitor of PI3K or mTOR.[12,13] This demonstrates the importance of studying the selectivity profile of compounds as soon as possible in early discovery.

First attempts to evaluate protein kinase inhibitor effectiveness were first made on a few number of selected kinases. Thanks to the recent improvement of biochemical assays, several companies offer robust and reliable technologies for screening compound collection on large kinase panels.[14] Hence, several studies published broad kinase profiling in which dozen to hundreds of compounds were tested on more than half of the human kinome.[15,16,17,18] From these experimental data, various information are available such as the inhibition or binding affinity on the primary kinase target as well as the selectivity profile of the screened compounds. However, how the selectivity is defined or computed may highly influence our perception of molecules of interest. In the past decades, several selectivity metrics have been used and published. Since they differ in their

equation and in their ease of understanding and usability, the metrics give different results on the compound selectivity profile.

In this study, we propose two new metrics to assess compound selectivity and we compare the results with the well-known selectivity scores largely used in drug discovery. Comparative assessment on the selectivity scores has been performed on three different published datasets containing different assay types.. Two datasets were obtained from enzymatic assays performed on a large part of the kinome. For the third one, we used the results from a cellular assay to assess the performance of these selectivity metrics. The proposed metrics provide novel insights on the selectivity profile of tested molecules and can be used in drug discovery projects in making key decisions about the design and selection of compounds.

METHODS

Datasets

Among all the publicly available biological datasets containing protein kinases, we chosen the two datasets having the highest percentage of completeness. Such selection is particularly important to insure a reliable comparison between various selectivity metrics. The first dataset was provided by Davis *et al.*,[15] in which 72 inhibitors were tested on 442 wild-type and mutated protein kinases (completeness of 100%). A dissociation constant (Kd) was measured for each protein-ligand pair only when an activity was first detected above 10 μM compound concentration. Among the tested protein kinases, we removed three bacterial proteins leading to a total of 439 protein kinases. The second dataset was published by Anastassiadis *et al.*,[16] where the percentages of inhibition of 178 compounds, tested at 0.5 μM, were measured on 300 wild-type and mutated protein kinases (completeness of 99%).

The use of a unique metric to estimate the selectivity of a compound amongst different datasets, regardless of the experimental method used to generate those data would be of great value. Indeed it would ease data analysis by avoiding activity conversions, and it would provide a universal tool that would allow the direct comparison of different selectivity profiles. The experimental screening techniques used in both datasets are different (competition binding assay[19] and "HotSpot" assay[16]) and have been applied on a relatively important common protein kinase – ligand pairs. These two datasets were used for direct comparison of various selectivity scores. Therefore, we standardized protein kinase names using the UniProt[20] code and compound names using

InChI[21] identifier in both datasets for direct comparisons. We counted 4720 pairs present in both publications (249 protein kinases and 19 compounds). For convenience, we will refer the Davis *et al.*, and Anastassiadis *et al.* datasets, as Ambit and Reaction Biology datasets respectively.

While most of the selectivity screening assays are performed on recombinant enzymes in biochemical assays, cellular assays can provide complementary information such as signaling pathways, cellular potency and cell permeation.[22] Moreover, recent studies have shown that assessment of compound selectivity using cellular or enzymatic assays may provide different results.[23] In order to evaluate the broad applicability of the approach, we have also studied the different selectivity scores on cellular activities. The NCI-60 Developmental Therapeutics Program Human Tumor Cell Line Screen was initiated in the late 1980s. A total of 60 cell lines were assembled to represent nine tumor types: breast, CNS, colon, leukemia, lung, melanoma, ovarian, prostate and renal.[24] Thousands of compounds have been screened on this panel to estimate their inhibition rates. Nowadays, 25 protein kinase inhibitors approved by the Food and Drug Administration (FDA) or in clinical trial have already been tested in these cell lines. We collected these data using CellMiner™, a web application developed by the Genomics and Bioinformatics Group of the NCI.[25,26]

Standard selectivity score

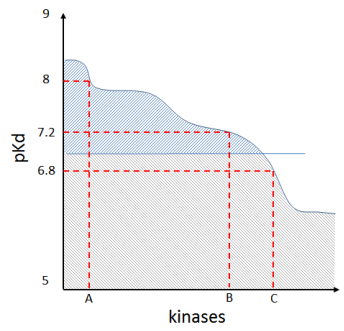
Given a compound, the standard selectivity score $S(x)$, where x represents the activity threshold, is calculated by dividing the number of inhibited protein kinases having an activity greater than x, by the total number of tested protein kinases, as shown in the equation (1).

$$S(x) = \frac{no.of\ values \geq x}{total\ no.of\ values} \quad (1)$$

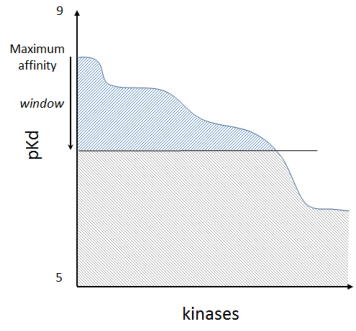
Depending on the experimental method used in the assay, x may also express a logarithmic affinity value. This metric is quantitative, easily computable and comparable between different profiling results. In this study, we have evaluated the variation of the standard selectivity score, $S(x)$, by using three different thresholds. For the Ambit dataset we calculated S(pKd5), S(pKd6) and S(pKd7), for 5, 6 and 7 logarithmic Kd thresholds respectively. For the Reaction Biology dataset, we selected similar range of activities S(50%), S(70%) and S(80%) for 50%, 70% and 80% of inhibition relative to control thresholds respectively. A low value

of $S(x)$ depicts a high compound selectivity whereas a high value of $S(x)$ illustrates a poor selectivity. The thresholds were chosen to depict respectively a low, medium and high biological interaction.

The main drawback of this metric is the applied threshold which needs to be defined and which has an important impact on the final result. For instance, we can represent the results of a profiling panel in which we take a compound tested on 100 protein kinases. We calculate the standard selectivity score S(pKd6) corresponding to a threshold of 6 for pKd. If we count three proteins with an affinity above the threshold, we may have in appearance a selective compound (3 inhibited proteins out of 100). Indeed, S(pKd6) score would be 0.03. However this selectivity score does not give any information on the strength of inhibition. Therefore, without this information, we would suppose that the three inhibited proteins will have similar effects on cells, leading to potential side effects. At the opposite, with the same conditions and three hits, we may have an affinity at a nanomolar level and the other just above the threshold. In this case, we would obtain the same standard selectivity score, however this case is completely different. The compound would bind preferentially one protein and so do the two others, but not at the same level, which would have different results at a cellular level. Moreover, as recalled by Cheng *et al.*,[27] the standard selectivity score does not capture any nuance. The use of the threshold will split the data into two categories and even if the values are close together around the threshold, they will belong to well distinct categories. Figure 1A illustrates this specific issue where, in the proposed example, the pKd threshold of 7 sets kinase B (pKd = 7.2) and kinase C (pKd = 6.8) in two different categories. While kinase A (pKd = 8) is in the same category as kinase B, though these two kinases have a larger affinity difference than kinase B and C. This is the classical issue in using a threshold for separating data and defining classes.



A.



B.

C.

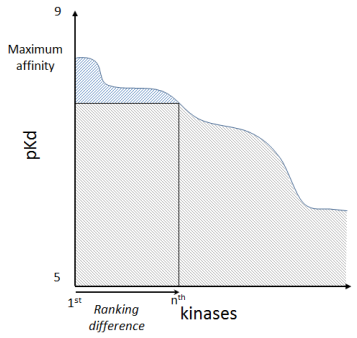


Figure 1: Schematic representation of three selectivity metrics. (A): standard selectivity score; (B) window score; (C) ranking score.

Gini coefficient

The Gini selectivity metric is particularly designed for percentage of inhibition. According to Graczyk the data have to be first sorted in ascending order to generate the curve of the cumulative fraction of inhibition. The Gini coefficient is then calculated using the area under the curve. For further details of the calculation, the reader may refer to the original article.[28] Unlike the standard selectivity score, this metric does not need a threshold to be used. However, the use of percentages of inhibition instead of a constant of inhibition (Ki) or a constant of dissociation (Kd), makes the Gini coefficient experimental condition dependent. Since the percentage of inhibition of a compound on a target is strongly dependent to its concentration, the Gini coefficient could vary at different compound concentrations.

Selectivity Entropy

The selectivity entropy (Ssel) is based on a thermodynamics approach to measure the compound selectivity.[29] The authors assume that the system theoretically contains all protein targets from the assay panel without any competitive molecules such as ATP. Then the inhibitor is added in a way that it does not saturate any target, but all inhibitor molecules bind a target. In such system, a selective inhibitor will bind approximatively one specific protein, so will have low entropy, while a non-selective molecule will bind many targets and so represents high entropy. Applying thermodynamic principals, they translate this theory using the Boltzmann law and obtain a selectivity value based on the entropy of the system.

Partition Index

The Partition Index was developed to discriminate selective compounds in small panels.[27] Using Kd values and a thermodynamic approach (such as the entropy score), this metric estimates the fraction of the

binding of a compound onto a reference protein kinase in comparison to the remaining kinases that are in the panel. This metric returns partition index values P ranging from 0 to 1. A compound with a P value close to 1 would indicate a promiscuous compound that binds almost exclusively to a unique protein. While this selectivity score is very useful within a unique drug discovery project, the use of a reference protein kinase to evaluate the partition index could be a limited factor if one wants to compare selectivity scores between compounds originating from various projects having different primary targets. As mentioned by the authors, the partition index is mainly suitable for a hit-to-lead process where few selected proteins are used, but it may not be applicable for a larger protein panel. To bypass this issue, a P_{MAX} index was introduced to represent the inhibitor partitioning to the most potently inhibited protein kinase.

Window Score

The following two novel metrics, respectively windows score and ranking score, require also a user-defined threshold. Nevertheless, they both meet three essential conditions. They are easy to calculate, they take into account the ranking of the experimental biological data and, for each compound, they consider the distance between the maximum affinity amongst all the biological data and another selected affinity defined by a threshold. Indeed, this information is important in the way many may look for optimizing the hit selection depending on these distances. Importantly, these two metrics can be applied on any data types such as affinity or activity data. In the current study, we applied the two selectivity scores on two datasets using pKd and percentage of inhibition data, but they can also be used on other data types such as thermal stability shift assays.[18]

For a given compound in a dataset, the window score requires the affinities to be ranked in descending order (from the highest pKd to the lowest). By choosing a window threshold, we count all the affinities that are included in the interval between the highest pKd and the highest pKd minus the defined window (Figure 1B). The windows score ranges from almost zero (technically it cannot be lower than 1/total no. of affinity) to one. The higher is the window, the worst is the selectivity. This number is then divided by the total number of tested proteins (2). Thus, the lower the window score is, the more selective the compound is.

$$Ws(x) = \frac{number\ \ of\ affinity\ \geq (\max(affinity) - x)}{total\ number\ of\ affinity} \quad (2)$$

where x represents the selected window. Like for the standard selectivity score mentioned above, the window score requires a threshold defined by the user. However, this threshold is not abstract because if defines the activity gap one may want to obtain between the most inhibited protein, often being the primary target, and the last inhibited protein in the given window. To evaluate the influence of the threshold on the window score, we chose three different windows: 2 log, 1 log and 0.5 log of Kd (respectively annotated WS(pKd2), WS(pKd1), WS(pKd0.5)) for the Ambit dataset, and 20, 10 and 5 percent of inhibition (respectively annotated WS(20%), WS(10%) and WS(5%)) for the Reaction Biology dataset.

Ranking Score

For a given compound in a dataset, the ranking score requires the affinities to be ordered from the best to the worst. By choosing a rank number, we subtract the affinity value at the position defined by the rank number from the best affinity (3) (Figure 1C). Therefore, the higher the ranking score, the more selective is the compound.

$$Rs(x) = max(affinity) - affinity^{x} \quad (3)$$

where x represents the affinity rank. As seen in Equation (3), a threshold also needs to be defined. The user only needs to define the rank corresponding to the activity difference between the activities of the x^{th} inhibited protein and the most inhibited protein. In this study, we evaluate the influence of the threshold on the ranking score by selecting three different thresholds: 20, 10 and 5 (respectively RS(20), RS(10) and RS(5)) for both datasets. Since this score uses a rank number it is independent of data types. Therefore, it can be applied for comparing compound selectivity profiles from a panel of proteins using different experimental techniques.

RESULTS AND DISCUSSION

We introduced two new metrics to assess the selectivity of compounds and we compared the results with already published selectivity scores. Comparative assessment on the selectivity scores has been performed on three different published datasets. For each dataset, the calculation of each aforementioned metrics was performed with a Knime workflow (available upon request) and each molecule was then ranked according to each metric.

Since the Gini coefficient, the entropy score and the partition index work exclusively with particular datatypes (percentages of inhibition for the Gini coefficient and Kd/Ki/IC50 for the two others) the calculation of all these metrics for each dataset required we converted the data using an approximation of the Hill formula (4). As example, Equation 4 was used to convert percentage of inhibition to an approximation IC50:

$$IC_{50} = [c](\frac{100}{\%inh} - 1) \quad (4)$$

where [c] is the screening compound concentration and %inh is the measured percentage of inhibition.

Results from Ambit dataset

Ambit dataset contains 72 protein kinase inhibitors tested on 439 human protein kinases. For each pair, a quantitative dissociation constant (Kd) was measured if the primary affinity was below compound concentration of 10 μM. For the non-confirmed compounds from first screen, we assigned a value of 9.99 μM, considering the theoretical maximum affinity. This approximation allows us to have a data matrix of protein – ligand pairs with a degree of completeness of 100%. Therefore, the selectivity scores calculated for each compound are performed on the same number of protein kinases. To ease the comparison, all metrics have been ordered according to the order of WS(pKd2) (Supplementary Table 1).

Depending of the applied threshold values, some selectivity scores cannot be computed. For instance, we noticed two compounds (CI-1040, BMS-345541) for which the standard selectivity score S(pKd7) could not be calculated because the cut-off was too low. As a consequence, S(pKd7) was equal to zero since no protein kinase was inhibited above this threshold. This illustrates the concept that a compound may seem selective when the applied cut-off is too high. In other words, the compound is not active on protein kinases while the selectivity could suggest the opposite. To avoid any confusion, we kept these compounds but we did not assign them a rank (Supplementary Table 1, empty cells). We compared the different metrics and observed that none of them offers the same exact selectivity order. The standard selectivity scores (S(pKd5), S(pKd6) and S(pKd7)) present similar orders as shown by the pairwise correlation $r > 0.85$ (Figure 2).

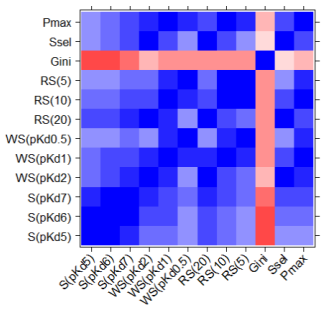


Figure 2: Ambit selectivity metric correlations. Values from -1 (red) to 1 (blue)

In the top four molecules selected from each selectivity score S(pKd5), S(pKd6) and S(pKd7), we retrieved the same four molecules namely GW-2580[30], CI-1040[31], MLN-120B[32] and SGX-523.[33] CI-1040 is not present in the top four molecules using S(pKd7) due to the absence of strong affinity on the tested protein kinases. Indeed, despite a very high selectivity score with S(pKd6), this compound presents a maximum affinity pKd of 6.9 on MAP2K1 amongst all protein kinases. The correlation coefficient between S(pKd5) and S(pKd7) is 0.85, slightly lower than the ones observed between S(pKd5) and S(pKd6), and S(pKd6) and S(pKd7) (respectively 0.97 and 0.93, Figure 2). It might be explained by the presence of several equally selective compounds caused by the high threshold that seems to avoid compound selectivity discrimination. The other metrics using score thresholds (window score and ranking score) are also very sensitive to the threshold, but as an advantage they rely on the maximum of affinity. In that way, unlike the standard selectivity score, they cannot return zero even with a stringent threshold. We retrieved similar ranking for WS(pKd2) and WS(pKd1) (Figure 2, r = 0.73), and for WS(pKd1) and WS(pKd0.5) (Figure 2, r = 0.75). However, a larger cut-off variation involves important differences in the rankings (Figure 2, r = 0.37 between WS(pKd2) and WS(pKd0.5)). While BMS-387032/SNS-032[34], nilotinib[35] and CHIR-265/RAF-265 are ranked respectively 14th, 24th and 28th by WS(pKd2), they are all ranked first according to WS(pKd0.5). Therefore, we can conclude these three compounds have both many affinity values in their windows as shown in equation 3 $\max(affinity) - x$, when x is high, but few values when x is low. This is an advantage of this metric that allows a rapid identification of the affinity window between the best hit and the following. We observed the same trend with the ranking score type. The correlations between the ranking scores are high when the thresholds are similar while larger differences in the threshold led to low correlations (Figure 2, r = 0.45 between RS(20) and RS(5), r = 0.64 between RS(20) and RS(10), r = 0.89 between Rs(10) and Rs(5)).

Regarding the metric Ssel, which does not use any threshold, it presents good correlation with several scores such as S(pKd7) (r = 0.63), WS(pKd2) (r = 0.90), RS(20) (r = 0.87) and P_{MAX} (r = 0.71). P_{MAX} is in good agreement with almost every other metrics, in particular with WS(pKd2) (r = 0.74), WS(pKd1) (r = 0.86), WS(pKd0.5) (r = 0.79), Rs(10) (r = 0.88), RS(5) (r = 0.84) and Ssel (r = 0.71). In general, all the metrics used in this study seem to select the same molecules as being selective, and the same molecules as being unselective, with some specific differences in the ranking (Supplementary information 1). Interestingly, only Gscore presents a very singular order characterized by several correlation coefficients below zero with every other metrics (Figure 2).

When we analyzed the most selective compounds, we retrieved GSK-461364A[36] at the first position with the use of WS(pKd2), WS(pKd1), WS(pKd0.5), Ssel and P_{MAX}. SGX-523[33] is at the first position with WS(pKd0.5), RS(20), RS(10) and RS(5). PLX-4720 is ranked first by the three window scores. It is important to note that due to their mathematical form, some selectivity scores will provide an identical score for several molecules. MLN-120B is ranked first by S(pKd7), WS(pKd1) and WS(pKd0.5). AZD-6244/ARRY-886 is ranked first by S(pKd7), S(pKd6) and WS(pKd0.5). GW-2580[30] is ranked first by S(pKd5), WS(pKd1) and WS(pKd0.5), though the global ranking of the compounds over each metric is quite different. VX-745[37] is ranked first by S(pKd6), WS(pKd1) and WS(pKd0.5).

Some compounds show a good consensus among several metrics. Indeed, we found that GSK-461364A, PLX-4720, SGX-523, GDC-0879, BI-2536, CP-690550, GW-2580 and VX-745 are in the top ten for at least eight selectivity metrics. The selectivity entropy (Ssel) and the partition index (P_{MAX}) contain the same top three molecules, and their top ten are in good agreement, but their order differ slightly for the reamining molecules (Figure 2, r = 0.71).

We noticed that the use of a cut-off for some metrics can induce a lack of sensitivity for some molecules since the same score is calculated for several molecules. This is observed mainly with low cut-off and with the standard selectivity scores and the window scores by looking at the ranks of molecules. For instance, two and three compounds are ranked 13^{th} and 14^{th} respectively with S(pKd5). However, S(pKd6) identifies three compounds ranked first, two compounds ranked second and four compounds ranked 14^{th}. Like the standard

selectivity score, the window score can also suffer from this trend and fifteen compounds are ranked first by WS(pKd0.5). Analyzing the affinities of these compounds, we noticed that their second best affinity is outside the 0.5 window and so, their respective window only contains one value, corresponding to the best affinity. However, though they have the same score, they do not bind the protein kinases with the same promiscuity. The maximum pKd for INCB18424[38] is 10.4 on JAK2 kinase domain, while for AZD-6244/ARRY-886[39] it is only 7 on MAP2K1. Therefore, WS(pKd0.5) is well designed to identify compounds that bind preferentially to a protein kinase, but it is not suited to distinguish highly to low active compounds.

Interestingly, the Gini coefficient is the metric providing the largest ranking differences compare to all other metrics (Figure 1) with r values always closed to 0 with the other metrics. As mentioned previously, this could be related to the conversion of Kd to percentage of inhibition, necessary to calculate the Gini coefficient.[29] When we plot the values of the S(pKd5) against the values of the Gini coefficients, we clearly see a parabolic relation that cannot be caught by a linear correlation (Figure 3).

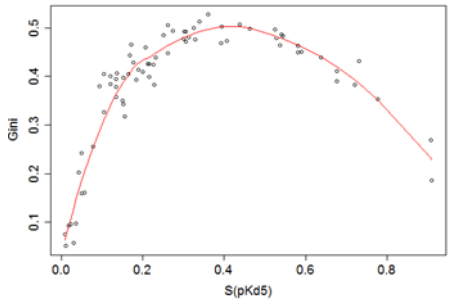


Figure 3: Correlation between the Gini coefficients and the standard selectivity score.

Surprisingly, we noticed that staurosporine, known to be a potent and unselective inhibitor of protein kinases,[19] is ranked third with WS(pKd0.5) and 12th with WS(pKd1) (Supplementary Table 1) in contrast to the other metrics that rather rank this molecule as highly promiscuous. This clearly shows that the use of low thresholds for selectivity scores can lead to erroneous conclusion. In fact, staurosporine is ranked 47th, 55th and 61th by WS(pKd7), WS(pKd6) and WS(pKd5) respectively. Finally, while vandetanib, a FDA-approved drug for EGFRs, VEGFRs protein kinases, is classified as non-selective, by all selectivity metrics GSK-461364A is identifies as a selective inhibitor

Reaction Biology dataset

The Reaction Biology dataset contains 178 protein kinase inhibitors tested on 300 protein kinases. Percentages of inhibition (instead of Kd) of each pair has been obtained at a 0.5 µM compound concentration. Since the authors did not evaluate experimentally the inhibition profile for all protein-ligand pairs, the percentage of completeness is 99%. For the comparison between all metrics, they all have been ordered according to WS(10%) (Supplementary Table 2). As mentioned previously, we noticed that some compounds do not have activity greater than the thresholds of the standard selectivity scores.. Therefore, we have calculated the twelve metrics on109 out 178 compounds.

The analysis of the results shows a global tendency for the employed metrics. The color coded applied on Table S2 clearly shows that same molecules populate respectively same ranking positions on the top and bottom parts of the table for almost all metrics except for the Gini coefficient, the selectivity entropy and the partition index. Figure 4 shows good pairwise correlations ($r > 0.7$) between the standard selectivity, the windows and the ranking scores. The Gini coefficient shows low correlations with the other metrics though they are still higher than with the previous dataset. Hence, it seems that the Gini coefficient always ranks compounds differently compared to other metrics whatever the biological data types. PD 174265 is found as the most selective compound by S(70%), S(80%), WS(10%) and WS(5%). Moreover, this compound is ranked in the top ten by four other metrics (S(50%), WS(20%), RS(10) and RS(5)), and in the top twenty by three other metrics (RS(20), Ssel and P_{MAX}). Surprisingly, this compound is ranked 131^{th} by the Gini coefficient. The compound mentioned in the article as EGFR inhibitor, presents a good consensus over the metrics. It is systematically in the top ten of the tested compounds but P_{MAX} (11^{th}) and Sel (12^{th}). p38 MAP Kinase Inhibitor III seems also rather selective as it is ordered first by S(70%), S(80%), WS(10%) and WS(5%), and it appears in the top twenty by S(50%), WS(20%), RS(20), RS(10). Similarly Tandutinib is ranked first by S(70%), S(80%), WS(10%) and WS(5%) and in the top twenty by WS(20%), RS(20), RS(10) and Gini score. We retrieved the same first nine molecules between P_{MAX} and Ssel with almost identical order, but these compounds are far to be the most selective according to other metrics. As mentioned earlier, we converted the percentages of inhibition into IC50 to calculate these two metrics. These transformed values may be responsible for such important differences. The Gini coefficient proposes considerable ranking differences as already reported.[29]

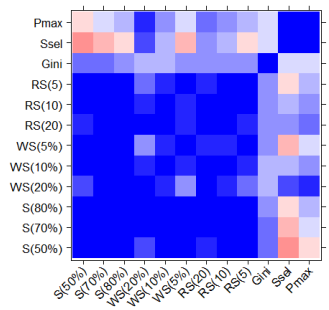


Figure 4: Reaction Biology selectivity metric correlations. Values from -1 (red) to 1 (blue)

Importantly, most of the metrics returns the same set of compounds as being promiscuous. Sunitinib[40], SU11652,[41] JAK3 Inhibitor VI, GW 6976[42], Indirubin Derivative E804, PKR Inhibitor,[43] CDK1/2 Inhibitor II,[44] SB 2180778[45], K-252a[46] and Staurosporine are always ranked amongst the least selective compounds by all metrics except by Ssel and P_{MAX}. This suggests that, in general, the metrics employing a threshold present important similarities in ranking compounds as suggested by their correlation coefficients. This conclusion differs slightly from what was observed with Ambit dataset. This may be due to the use of percentage of inhibition which does not fully discriminate inactive, weakly active and highly active compounds compare to pKd or pIC_{50}. Therefore, the effect of the threshold used with percentage of inhibition (5, 10, 20, 50, 70 and 80%) has lower impact in discriminating selective compounds than with pKd (0.5, 1, 2, 5, 6 and 7). However, for both dataset, the Gini coefficient offers singular rankings. Since the Gini coefficient is calculated using cumulative inhibition of the compounds on the tested proteins, it is particularly sensitive to compound concentration and to the number of screened protein.

Selectivity metrics analysis of shared protein-ligand pairs

4720 protein-ligand pairs shared by Ambit and Reaction Biology datasets were retrieved. This represents 19 molecules and 249 protein kinases. Because of the already mentioned threshold issue with the standard selectivity scores, we could not calculate the selectivity of few compounds using these metrics. The different rankings according to the metrics and the datasets are available in Supplementary Table 3.

Although the selection of these 19 compounds relied only on our will to study the impact of the activity type, the direct consequence of this low number of molecules is the increase of the ranking sensitivity. Indeed, small differences will have much impact when comparing two metrics. In a first part, we focused on the 19

molecules results obtains using the Ambit data. In a second part, we did the same analysis but using only the Reaction Biology data. In a third part, we first performed a pairwise comparison of each metric and analyzed the influence of the datasets. For the comparison we had to differentiate the metrics that rely directly (Gini score, entropy score, partition index) or indirectly (standard selectivity score, window score) on the type of measured activity. Only the ranking score does not rely on activity type.

The results obtained using Ambit data confirmed the previous observations made on the full molecule set. S(pKd5) S(pKd6) and S(pKd7) present all similar ranking (Supplementary Table 3A) with pairwise correlations of at least 0.83 (Figure 5). We also noticed similar rankings between WS(pKd2), RS(20), RS(10) and Ssel with good pairwise correlations ($r > 0.85$). Additionally, Ssel presents excellent correlations with WS(pKd1) ($r = 0.91$) and P_{MAX} ($r = 0.91$).

Gefitinib is ranked first in six cases (S(pKd7), WS(pKd2), WS(pKd1), WS(pKd0.5), Ssel and P_{MAX}) and always in the top five for all other metrics. The only exception is with the Gini coefficient that ranks this molecule 10th. Lapatinib[47] and GW-2580 are also ranked systematically in top five by the different metrics, except by the Gini coefficient that ranks the two compounds 19th and 18th respectively. Erlotinib appears rather selective in this panel since it is in the top five of 9 out 12 metrics.[48] A consensus appears for staurosporine which is identified unselective by all calculated metrics. Finally, SKI-606 has an interesting profile since it is considered more or less selective depending on the employed metrics.[49] Indeed, with a pKd affinity greater than 8 on 45 out of 249 tested protein kinases, the three standard selectivity scores rank this molecule among the least selective compounds. On the opposite, Ws0.5 ranks SKI-606 first because there is only one activity present in the corresponding window. More precisely, it shows a very high affinity on ABL1 (pKd = 10.2) and the second highest affinity is greater than the 0.5 log threshold (pKd = 9.3 on MPK4K5). This compound is also relatively well ranked by WS(pKd1) (third), RS(5) (6th), Gini (4th), Ssel (6th) and P_{MAX} (4th). Hence, SKI-606 clearly shows that the use of a selectivity scores is subjective and depends on the applied metrics. The use of 2 different metrics types such as Ranking or Windows scores and Gini, entropy score or partition index, provides greater confidence in the selectivity of a compound.

Concerning the 19 compounds from Reaction Biology data, we observed a similar tendency as with all the compounds of this dataset. Except with the Gini coefficient, the entropy score and the partition index, all pairwise metrics correlate very well (Figure 5). We retrieved gefitinib, erlotinib, MLN-518,[50] SB-203580,[51] imatinib[52] and GW-2580, almost each time in top five by the standard selectivity scores and the windows scores. AB-1010 is ranked first by Gini, Ssel and P_{MAX} but other metrics rank this compound with medium selectivity score.[53] Standard selectivity scores, windows scores and ranking scores are in agreement to consider staurosporine, dasitinib[54] and SKI-606 as poorly selective. Notably, Gini, Ssel and P_{MAX} consider the PI3K inhibitor, PI-103,[55] as a promiscuous compound. PI-103 is a special case since it has been designed to target an atypical family of the kinome not present in the published dataset. Therefore, it inhibits weakly most of the typical protein kinases of the panel as shown by the absence of ranking score for the standard selectivity scores (it does not inhibit any kinases at more than 50%). However, WS(10%) and specially WS(5%) rank this compound as particularly selective since an important activity gap is observed between the first and second most inhibited kinases (CSNK2A2 at 44% and NUAK2 at 25%). Interestingly, SKI-606, which was well ranked by WS(pKd1) and WS(pKd0.5) on Ambit dataset, is ranked last by the two equivalent metrics WS(10%) and WS(5%) on the Reaction Biology dataset .

Finally, to determine if one of these metrics could return approximately a similar rank regardless the dataset origin and the measured activity type, we performed pairwise comparison of the compound ranks given by each metric (Supplement Table 3A and 3B). S(pKd5) and S(50%) highly correlate (r = 0.92), like S(pKd6) and S(70%) (r = 0.90) and S(pKd7) and S(80%) (r = 0.91) (Figure 5). Additionally, S(pKd7), calculated on the Ambit dataset, correlates with the standard selectivity scores, the windows scores and the ranking scores calculated on the Reaction Biology dataset. S(pKd6) and S(pKd7) are more correlated with the window scores and the ranking scores calculated on the Reaction biology dataset than those calculated on Ambit dataset probably because the compounds have been tested at 0.5 μM, and a 50% inhibition of a compound roughly corresponds to a pIC50 (or pKd) of 6.3. The window scores calculated on the Ambit dataset differ slightly from those calculated on the Reaction Biology dataset. When the threshold diminishes for the windows score, the correlation between the two datasets decreases also, for instance the correlation between WS(pKd2) and

WS(50%) is equal to 0.75, it decreases to 0.46 between WS(pKd1) and WS(20%), and drops to 0 between WS(pKd0.5) and WS(10%) (Figure 5).

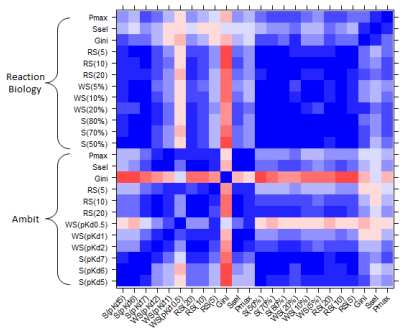


Figure 5: Ambit – Reaction Biology overlap metrics correlations. Values from -1 (red) to 1 (blue)

Here again, this might be due to the correlation between pKd and percentage of inhibition ranges and a 2-log, 1-log and 0.5-log gaps correspond to 50%, 20% and 10% activity gaps respectively. About the ranking scores, we noted interesting correlations between the two datasets using RS(20) (r = 0.74) and RS(10) (r = 0.71). These medium correlations clearly show significant differences in the selectivity results of the ranked compounds when two types of experimental data are used. Several outliers have also been identified.[56] A thorough analysis on each compound reveals that the major differences seem to come from the highest activities, as reflected by the low correlation of the RS(5) (r = 0.34) between the two datasets. Finally, as seen previously, there is no correlation between the Gini coefficients calculated (r = -0.15), the Ssel (r = 0.13) nor the PMAX (r = 0.25) probably due to activity data conversion necessary for these metrics (see methods). In conclusion, none of these three metrics returns the same compound ranking when measured activities are provided by different experimental techniques. Nevertheless, the standard selectivity score and the window scores returned similar ranking for the 19 studied molecules. It would be interesting to carry out this comparison on larger datasets to confirm this trend.[57]

Selectivity analysis of kinase inhibitors from cellular screening data

We evaluated the twelve metrics on cellular activities obtained from the US National Cancer Institute (NCI) panel.[24] We retrieved 25 protein kinase inhibitors tested on up to 60 human cancer cell lines. To avoid any bias, we removed the inhibitor Lestaurtinib since it was tested on only 40 cell lines, and the melanoma cell

line MDA_N used on only two inhibitors. For each metric, we ranked the 24 protein kinase inhibitors (Supplementary Table 4).

First, each metric offers a unique order. Imatinib was ranked first or second by all metric except by the Gini coefficient (6^{th}). Nilotinib also presents a good selectivity on these cell lines.[35] It is present in the top three of all metrics with the exception of S($pIC_{50}5$) and Gini. Regarding the results of bosutinib[49] and sunitib,[41] they are more contrasted. These two compounds are scored second and first by S($pIC_{50}7$), WS($pIC_{50}1$) and WS($pIC_{50}0.5$), fourth and third by RS(5), third and fourth by P_{MAX}, but 14^{th} and 13^{th} by S($pIC_{50}5$), 14^{th} and 17^{th} by Gini, and 10^{th} and 12^{th} by Ssel respectively. Regarding the compounds with the lowest selectivity scores, sorafenib[58] and trametinib seem to be the least selective kinase inhibitor amongst the 24 kinase inhibitors.[59,60] It is important to note that the MEK inhibitor trametinib is highly selective on the *in-vitro* biochemical kinase panel.[59] Interestingly, sorafenib is a drug approved by the (FDA) for the treatment of renal cell and hepatocellular carcinoma and also for locally recurrent of metastatic, progressive differentiated thyroid carcinoma. Nevertheless, in these assays, sorafenib inhibits the 59 tumor cell lines at less than 10 µM, but its maximum inhibition is relatively low (1.25 µM on the breast cancer cell line MDA_MB_231), the most sensitive cell line.

The correlation matrix allows the detection of novel patterns regarding the selectivity metrics differences (Figure 6). Firstly, the standard selectivity score S($pIC_{50}5$) poorly correlates with most of the metrics, even with the S($pIC_{50}6$) (r = 0.27) and S(p $IC_{50}7$) (r = 0). On the opposite, S($pIC_{50}7$) highly correlates with most of the other metrics, but this is mainly because these correlations were calculated taking into account only compounds for which S($pIC_{50}7$) could be estimated (17 out of 24 compounds). This is due to the lower number of high activities on cells for these compounds compared to those observed on enzymatic assays. WS($pIC_{50}2$) does not correlate with any of the other window scores (r = 0.54 with WS($pIC_{50}1$) and r = 0.23 with WS($pIC_{50}0.5$)) due to unique profiles for some compounds. Hence, bosutinib and sunitinib are ranked 11^{th} and 14^{th} respectively with WS($pIC_{50}2$) but second and first with WS($pIC_{50}1$) and WS($pIC_{50}0.5$). An analysis of the biological profile of sunitinib shows that the highest activity is on KM12 cell line with a pIC50 of 7.4, followed HOP_92 with a pIC50 of 6.4. The presence of only one activity in the 1 and 0.5 log windows for sunitinib explains its excellent rankings for WS($pIC_{50}1$) and WS($pIC_{50}0.5$). However, we counted 47 out of 59 activities

in the 2 log window, resulting a poor selectivity score for WS(pIC_{50}2). There are excellent correlations of 0.91 and 0.93 between WS(pIC_{50}2) and RS(20) and WS(pIC_{50}2) and Ssel respectively. WS(pIC_{50}1) and WS(pIC_{50}0.5) are both well correlated with RS(10), RS(5) and in particular with P_{MAX} (r = 0.92 and 0.87). RS(20) shows similar ranking especially with WS(pIC_{50}2) and Ssel, and in a less extend with RS(10) and P_{MAX}. RS(10) and RS(5) present good agreement with most of the metrics except with S(pIC_{50}5) and the Gini coefficient. Actually, we observed here a new illustration of the singularity of the Gini coefficient as it is in general poorly correlated with most of the other metrics. Ssel correlates well with WS(pIC_{50}2), WS(pIC_{50}1), RS(20), RS(10) and P_{MAX}. Finally, P_{MAX} is indubitably the metric correlating the most with the other metrics with r > 0.5 in all cases except with S(pIC_{50}5), S(pIC_{50}6) and Gini.

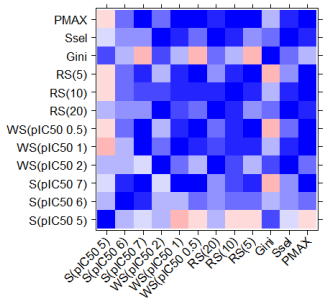


Figure 6: NCI-60 selectivity metric correlations. Values goes from -1 (red) to 1 (blue)

Therefore, the analysis of the twelve metrics on this dataset containing tumor cell lines confirmed our previous observations on enzymatic dataset. Several metrics seem to be unique in the way of identifying selective compounds. The application of the metrics is highly dependent of the need of the user such as the selectivity profile defined in the target product profile (TPP).

CONCLUSION

Is there a better metric to measure the selectivity of your compound? Definitely, the answer is negative. Although we prefer easily computable and understandable metrics, we think the chosen metric must not be selected randomly. The use of several metrics may help in finding a ranking consensus, but in this case, we suggest using non correlated metrics, like the WS or RS with P_{MAX} or Ssel. We do not recommend applying the Gini coefficient on a dataset with many protein kinases. As we saw, its ranking is particularly singular and not easy to understand.

The standard selectivity score, along with the window score and the ranking score have the advantage of being very easy to compute. Therefore the results they provide are simple to understand and interpret. Nevertheless, the threshold of the standard selectivity score may be at the origin of abnormality when its value is set too low, as emphasized in the text. To avoid this phenomenon, we recommend looking at the activity/affinity range before choosing the threshold. The window score also required thresholds but their repercussion is not as much important as with the standard selectivity score, because this metric cannot drop to zero. To differentiate compounds with these metrics, we recommend using at least two different thresholds. A thin window will identify compounds that bind preferentially one protein kinase, while a large cut-off will confirm this same compound do not bind plenty of other kinases with affinities still important. The same conclusion can be drawn for the ranking score. The cut-off retained had to represent the activity gap you are ready to accept between the main hit and the following hits.

To compare two datasets obtained with techniques returning different activity type, the ranking score is indubitably the metric that suit the best. As it is based only on the rank determined directly by the data, it does not require any data conversion that could lead to bias.

ACKOWLEDGMENTS

N Bosc and P Bonnet are grateful to the Région Centre Val de Loire and Janssen-Cilag for financial supports.

REFERENCES

1. Kola, I. & Landis, J. Opinion: Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discov.* **3,** 711–716 (2004).
2. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298,** 1912–1934 (2002).
3. Ghoreschi, K., Laurence, A. & O'Shea, J. J. Selectivity and therapeutic inhibition of kinases: to be or not to be? *Nat. Immunol.* **10,** 356–360 (2009).
4. Guha, M. Cyclin-dependent kinase inhibitors move into Phase III. *Nat. Rev. Drug Discov.* **11,** 892–894 (2012).
5. Hopkins, A., Mason, J. & Overington, J. Can we rationally design promiscuous drugs? *Curr. Opin. Struct. Biol.* **16,** 127–136 (2006).
6. Anighoro, A., Bajorath, J. & Rastelli, G. Polypharmacology: Challenges and Opportunities in Drug Discovery: Miniperspective. *J. Med. Chem.* **57,** 7874–7887 (2014).
7. Giordano, S. & Petrelli, A. From Single- to Multi-Target Drugs in Cancer Therapy: When Aspecificity Becomes an Advantage. *Curr. Med. Chem.* **15,** 422–432 (2008).
8. Knight, Z. A., Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nat. Rev. Cancer* **10,** 130–137 (2010).
9. Bamborough, P., Drewry, D., Harper, G., Smith, G. K. & Schneider, K. Assessment of Chemical Coverage of Kinome Space and Its Implications for Kinase Drug Discovery. *J. Med. Chem.* **51,** 7898–7914 (2008).
10. Brehmer, D. *et al.* Cellular targets of gefitinib. *Cancer Res.* **65,** 379–382 (2005).
11. Dar, A. C., Das, T. K., Shokat, K. M. & Cagan, R. L. Chemical genetic discovery of targets and anti-targets for cancer polypharmacology. *Nature* **486,** 80–84 (2012).
12. Saurat, T. *et al.* Design, Synthesis, and Biological Activity of Pyridopyrimidine Scaffolds as Novel PI3K/mTOR Dual Inhibitors. *J. Med. Chem.* **57,** 613–631 (2014).
13. Tang, K. D. & Ling, M.-T. Targeting Drug-Resistant Prostate Cancer with Dual PI3K/mTOR Inhibition. *Curr. Med. Chem.* **21,** 3048–3056 (2014).

14. Ma, H., Deacon, S. & Horiuchi, K. The challenge of selecting protein kinase assays for lead discovery optimization. *Expert Opin. Drug Discov.* **3,** 607–621 (2008).

15. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29,** 1046–1051 (2011).

16. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29,** 1039–1045 (2011).

17. Metz, J. T. *et al.* Navigating the kinome. *Nat. Chem. Biol.* **7,** 200–202 (2011).

18. Fedorov, O. *et al.* A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *Proc. Natl. Acad. Sci.* **104,** 20523–20528 (2007).

19. Fabian, M. A. *et al.* A small molecule–kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **23,** 329–336 (2005).

20. The UniProt Consortium. Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.* **42,** D191–D198 (2014).

21. The IUPAC International Chemical Identifier (InChI). (2015). at <www.iupac.org/home/publications/e-resources/inchi.html>

22. Hancock, M. K., Lebakken, C. S., Wang, J. & Bi, K. Multi-pathway cellular analysis of compound selectivity. *Mol. Biosyst.* **6,** 1834 (2010).

23. Bain, J. *et al.* The selectivity of protein kinase inhibitors: a further update. *Biochem. J.* **408,** 297–315 (2007).

24. Shoemaker, R. H. The NCI60 human tumour cell line anticancer drug screen. *Nat. Rev. Cancer* **6,** 813–823 (2006).

25. Shankavaram, U. T. *et al.* CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics* **10,** 277 (2009).

26. Reinhold, W. C. *et al.* CellMiner: A Web-Based Suite of Genomic and Pharmacologic Tools to Explore Transcript and Drug Patterns in the NCI-60 Cell Line Set. *Cancer Res.* **72,** 3499–3511 (2012).

27. Cheng, A. C., Eksterowicz, J., Geuns-Meyer, S. & Sun, Y. Analysis of Kinase Inhibitor Selectivity using a Thermodynamics-Based Partition Index. *J. Med. Chem.* **53,** 4502–4510 (2010).

28. Graczyk, P. P. Gini Coefficient: A New Way To Express Selectivity of Kinase Inhibitors against a Family of Kinases. *J. Med. Chem.* **50,** 5773–5779 (2007).

29. Uitdehaag, J. C. & Zaman, G. J. A theoretical entropy score as a single value to express inhibitor selectivity. *BMC Bioinformatics* **12,** 94 (2011).

30. Conway, J. G. *et al.* Inhibition of colony-stimulating-factor-1 signaling in vivo with the orally bioavailable cFMS kinase inhibitor GW2580. *Proc. Natl. Acad. Sci.* **102,** 16078–16083 (2005).

31. Barrett, S. D. *et al.* The discovery of the benzhydroxamate MEK inhibitors CI-1040 and PD 0325901. *Bioorg. Med. Chem. Lett.* **18,** 6501–6504 (2008).

32. Hideshima, T. *et al.* MLN120B, a Novel I B Kinase Inhibitor, Blocks Multiple Myeloma Cell Growth In vitro and In vivo. *Clin. Cancer Res.* **12,** 5887–5894 (2006).

33. Buchanan, S. G. *et al.* SGX523 is an exquisitely selective, ATP-competitive inhibitor of the MET receptor tyrosine kinase with antitumor activity in vivo. *Mol. Cancer Ther.* **8,** 3181–3190 (2009).

34. Heath, E. I., Bible, K., Martell, R. E., Adelman, D. C. & LoRusso, P. M. A phase 1 study of SNS-032 (formerly BMS-387032), a potent inhibitor of cyclin-dependent kinases 2, 7 and 9 administered as a single oral dose and weekly infusion in patients with metastatic refractory solid tumors. *Invest. New Drugs* **26,** 59–65 (2008).

35. Weisberg, E. *et al.* Characterization of AMN107, a selective inhibitor of native and mutant Bcr-Abl. *Cancer Cell* **7,** 129–141 (2005).

36. Degenhardt, Y. *et al.* Sensitivity of Cancer Cells to Plk1 Inhibitor GSK461364A Is Associated with Loss of p53 Function and Chromosome Instability. *Mol. Cancer Ther.* **9,** 2079–2089 (2010).

37. Duffy, J. P. *et al.* The Discovery of VX-745: A Novel and Selective p38α Kinase Inhibitor. *ACS Med. Chem. Lett.* **2,** 758–763 (2011).

38. Quintas-Cardama, A. *et al.* Preclinical characterization of the selective JAK1/2 inhibitor INCB018424: therapeutic implications for the treatment of myeloproliferative neoplasms. *Blood* **115,** 3109–3117 (2010).

39. Kolb, E. A. *et al.* Initial testing (stage 1) of AZD6244 (ARRY-142886) by the pediatric preclinical testing program. *Pediatr. Blood Cancer* **55,** 668–677 (2010).

40. Mendel, D. B. *et al.* In Vivo Antitumor Activity of SU11248, a Novel Tyrosine Kinase Inhibitor Targeting Vascular Endothelial Growth Factor and Platelet-derived Growth Factor Receptors: Determination of a Pharmacokinetic/Pharmacodynamic Relationship. *Clin. Cancer Res.* **9,** 327–337 (2003).

41. Ellegaard, A.-M. *et al.* Sunitinib and SU11652 Inhibit Acid Sphingomyelinase, Destabilize Lysosomes, and Inhibit Multidrug Resistance. *Mol. Cancer Ther.* **12,** 2018–2030 (2013).

42. Martiny-Baron, G. *et al.* Selective inhibition of protein kinase C isozymes by the indolocarbazole Gö 6976. *J. Biol. Chem.* **268,** 9194–9197 (1993).

43. Jammi, N. V., Whitby, L. R. & Beal, P. A. Small molecule inhibitors of the RNA-dependent protein kinase. *Biochem. Biophys. Res. Commun.* **308,** 50–57 (2003).

44. Pratt, D. J. *et al.* Dissecting the Determinants of Cyclin-Dependent Kinase 2 and Cyclin-Dependent Kinase 4 Inhibitor Selectivity †. *J. Med. Chem.* **49,** 5470–5477 (2006).

45. Jackson, J. R. *et al.* An Indolocarbazole Inhibitor of Human Checkpoint Kinase (Chk1) Abrogates Cell Cycle Arrest Caused by DNA Damage. *Cancer Res.* **60,** 566–572 (2000).

46. Carrasco, M. A. *et al.* Regulation of glycinergic and GABAergic synaptogenesis by brain-derived neurotrophic factor in developing spinal neurons. *Neuroscience* **145,** 484–494 (2007).

47. Burris, H. A. Dual Kinase Inhibition in the Treatment of Breast Cancer: Initial Experience with the EGFR/ErbB-2 Inhibitor Lapatinib. *The Oncologist* **9,** 10–15 (2004).

48. Ng, S. S. W., Tsao, M.-S., Nicklee, T. & Hedley, D. W. Effects of the epidermal growth factor receptor inhibitor OSI-774, Tarceva, on downstream signaling pathways and apoptosis in human pancreatic adenocarcinoma. *Mol. Cancer Ther.* **1,** 777–783 (2002).

49. Golas, J. M. SKI-606, a Src/Abl Inhibitor with In vivo Activity in Colon Tumor Xenograft Models. *Cancer Res.* **65,** 5358–5364 (2005).

50. Corbin, A. S. Sensitivity of oncogenic KIT mutants to the kinase inhibitors MLN518 and PD180970. *Blood* **104,** 3754–3757 (2004).

51. Barancík, M. *et al.* SB203580, a specific inhibitor of p38-MAPK pathway, is a new reversal agent of P-glycoprotein-mediated multidrug resistance. *Eur. J. Pharm. Sci. Off. J. Eur. Fed. Pharm. Sci.* **14,** 29–36 (2001).

52. Capdeville, R., Buchdunger, E., Zimmermann, J. & Matter, A. Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. *Nat. Rev. Drug Discov.* **1,** 493–502 (2002).

53. Hahn, K. A. *et al.* Masitinib is Safe and Effective for the Treatment of Canine Mast Cell Tumors. *J. Vet. Intern. Med.* **22,** 1301–1309 (2008).

54. O'Hare, T. In vitro Activity of Bcr-Abl Inhibitors AMN107 and BMS-354825 against Clinically Relevant Imatinib-Resistant Abl Kinase Domain Mutants. *Cancer Res.* **65,** 4500–4505 (2005).

55. López-Fauqued, M. *et al.* The dual PI3K/mTOR inhibitor PI-103 promotes immunosupression, *in vivo* tumor growth and increases survival of sorafenib-treated melanoma cells. *Int. J. Cancer* **26,** 1549–1561 (2010).

56. Zhang, C., Habets, G. & Bollag, G. Interrogating the kinome. *Nat. Biotechnol.* **29,** 981–983 (2011).

57. Jacoby, E. *et al.* Extending kinome coverage by analysis of kinase inhibitor broad profiling data. *Drug Discov. Today* **20,** 652-658 (2015).

58. Keating, G. M. & Santoro, A. Sorafenib: A Review of its Use in Advanced Hepatocellular Carcinoma. *Drugs* **69,** 223–240 (2009).

59. Gilmartin, A. G. *et al.* GSK1120212 (JTP-74057) Is an Inhibitor of MEK Activity and Activation with Favorable Pharmacokinetic Properties for Sustained In Vivo Pathway Inhibition. *Clin. Cancer Res.* **17,** 989–1000 (2011).

60. Sakai, T. Antitumor activities of JTP-74057 (GSK1120212), a novel MEK1/2 inhibitor, on colorectal cancer cell lines in vitro and in vivo. *Int. J. Oncol.* **39,** 23-31 (2011).

SUPPLEMENTARY INFORMATION

Table S1: Ranking of the different metrics computed for the Ambit dataset. The first two columns give the Puchem identifier and the common name of the molecules, the twelve other columns give the ranks for each metric. For each ranking column, we colored the cells from red to blue, from the most to the worst selective, respectively. Empty cells represent compounds for which we could not calculate a selectivity value for the given metric.

pubchem_id	Compound_Name	Rank_S(pKd5)	Rank_S(pKd6)	Rank_S(pKd7)	Rank_WS(pKd2)	Rank_WS(pKd1)	Rank_WS(pKd0.5)	Rank_RS(20)	Rank_RS(10)	Rank_RS(5)	Rank_Gini	Rank_Ssel	Rank_P_{MAX}
16050824	GSK-461364A	18	8	3	1	1	1	3	3	4	59	1	1
24180719	PLX-4720	30	11	5	1	1	1	10	7	7	51	2	2
3038525	VX-745	6	1	2	2	1	1	9	8	6	67	9	4
11617559	GW-2580	1	2	2	3	1	1	6	4	3	70	8	5
24779724	SGX-523	4	2	3	3	2	1	1	1	1	68	4	6
10461815	PHA-665752	45	33	14	3	3	3	17	11	8	19	10	11
11717001	GDC-0879	8	3	3	3	2	2	2	2	2	66	5	12
25126798	INCB18424	38	31	14	4	1	1	7	6	5	15	3	3
11364421	BI-2536	28	14	5	4	2	1	11	9	9	44	6	7
9926791	CP-690550	10	6	4	5	3	1	4	5	10	61	7	8
9929127	MLN-120B	3	3	1	6	1	1	16	10	11	69	56	13
156422	BIRB-796	27	22	10	7	3	1	18	12	16	34	11	9
5287969	Flavopiridol	43	30	14	8	3	1	24	14	19	21	13	10
151194	PTK-787	7	5	5	8	5	2	12	17	22	63	21	16
25182616	GSK-1838705A	22	21	15	9	3	3	21	19	17	33	12	15
208908	Lapatinib	8	7	10	9	11	6	5	46	48	62	18	31
123631	Gefitinib	29	20	11	9	10	9	13	60	60	36	19	56
9869779	Ki-20227	26	25	17	10	6	6	20	21	43	25	14	26
17755052	GDC-0941	14	10	11	10	11	6	14	53	62	54	20	41
10127622	AZD-6244/ARRY-886	5	1	1	10	2	1	34	15	12	71	71	46
156414	CI-1033	40	27	23	10	12	12	19	65	57	18	28	54
11314340	A-674563	39	35	22	11	5	3	25	20	24	6	15	22
176870	Erlotinib	35	29	20	11	11	8	23	56	65	17	24	48
10184653	BIBW-2992	17	15	13	11	11	10	8	66	68	45	23	61
153999	LY-333531	33	26	6	12	4	2	33	16	14	28	25	18
3038522	MLN-518	14	14	11	12	9	6	22	26	50	53	26	33
6918852	R547	15	18	18	13	8	6	27	29	64	40	17	49
9884685	PI-103	12	9	14	13	13	11	15	67	71	58	31	64
3025986	BMS-387032/SNS-032	24	19	14	14	3	1	26	22	15	37	22	14
11667893	AMG-706	19	18	16	14	10	3	28	37	26	42	32	24
10113978	Pazopanib	37	32	15	14	9	5	31	27	47	20	33	39
11712649	MLN-8054	16	12	4	15	4	2	32	18	20	56	36	20
5291	Imatinib	13	14	18	15	5	3	29	25	23	49	16	21
9933475	AZD-2171	33	34	29	15	10	8	30	51	67	4	27	55
11626560	Crizotinib	44	39	28	16	6	2	35	24	25	5	37	19
5329102	Sunitinib	56	50	40	17	7	3	39	32	31	38	40	25
6450551	AG-013736	37	31	23	17	10	6	40	47	44	11	29	40
5328940	SKI-606	52	43	38	17	12	9	44	61	63	14	35	59
25033099	AT-7519	11	13	20	18	9	4	41	33	39	52	30	28
10427712	TG-100-115	28	23	12	19	4	4	38	23	18	35	34	23
176155	SB-203580	17	8	7	19	9	6	37	34	41	57	52	36
24889392	AC220	13	16	21	20	16	6	55	55	45	43	45	47
10138259	SU-14813	55	46	35	20	11	8	42	48	52	30	42	50
11485656	ABT-869	32	28	26	21	11	5	52	52	40	13	44	32
16007391	AZD-1152HQPA	31	20	13	21	8	6	45	31	46	31	41	37
9915743	HKI-272	34	31	27	21	13	10	49	63	61	9	39	60
16725726	GSK-690693	20	24	17	22	13	6	43	45	66	29	38	53
11234052	BMS-540215	14	15	10	23	9	4	48	35	38	46	46	29
176167	LY-317615	23	14	9	23	10	4	47	40	37	47	51	35
644241	Nilotinib	21	28	24	24	5	1	51	30	21	22	43	17
10074640	AB-1010	12	17	19	24	14	5	54	59	44	41	47	43
216239	Sorafenib	36	31	25	25	9	3	56	36	36	10	49	27
9886808	CHIR-258/TKI-258	47	40	32	25	10	4	46	38	32	7	48	30
3062316	Dasatinib	42	37	36	26	15	11	62	64	54	1	50	63
11676971	JNJ-28312141	51	41	33	27	15	5	63	50	42	12	54	51
11656518	CHIR-265/RAF-265	25	21	8	28	10	1	57	49	34	39	65	34
16722836	TG-101348	58	45	34	29	6	3	53	28	28	32	53	42
451705	Staurosporine	61	55	47	30	12	3	60	39	27	64	55	44
42642645	EXEL-2880/GSK-1363089	50	48	44	31	11	3	61	42	29	23	58	38
11213558	R406	56	52	43	32	14	5	64	43	55	48	59	58
447077	PD-173955	46	42	39	32	23	13	70	70	70	3	61	72
9809715	BIBF-1120 (derivative)	53	49	42	33	11	3	59	41	30	27	60	45
24905147	PP-242	53	44	30	34	14	7	58	57	56	24	57	62
11409972	AST-487	54	47	41	34	17	8	66	62	51	26	63	65
24202429	PKC-412	49	42	31	35	17	2	65	54	35	16	67	52
3081361	Vandetanib	41	36	30	36	19	14	72	69	59	2	64	70
5494449	VX-680/MK-0457	48	38	30	36	20	15	71	71	69	8	62	71
16038120	TAE-684	59	53	45	37	18	12	68	68	58	55	66	67
11427553	KW-2449	57	51	37	38	21	6	69	58	53	50	68	68
126565	CEP-701	60	54	46	39	22	6	67	60	49	60	69	69
6918454	CI-1040	2	1		40	2	2	36	13	13	72	72	57
9813758	BMS-345541	9	4		40	10	3	50	44	33	65	70	66

Table S2: Ranking of the different metrics computed for the Reaction Biology dataset. The first two columns give the Puchem identifier and the common name of the molecules, the twelve other columns give the ranks for each metric. For each ranking column, we colored the cells from red to blue, from the most to the worst selective, respectively. Empty cells symbolize compounds for which we could not calculate a selectivity value for the given metric.

pubchem_id	Compound_Name	Rank_S(50%)	Rank_S(70%)	Rank_S(80%)	Rank_WS(20%)	Rank_WS(10%)	Rank_WS(5%)	Rank_RS(20)	Rank_RS(10)	Rank_RS(5)	Rank_Gini	Rank_SseI	Rank_P_{MAX}
3038522	Tandutinib	27	1	1	15	1	1	14	9	39	5	43	37
5174	SC-68376	1			97	1	1	76	54	25	49	134	113
123631	Gefitinib	31	10	1	17	1	1	25	24	31	61	33	20
6419739	p38 MAP Kinase Inhibitor III	18	1	1	10	1	1	19	15	21	102	78	38
4709	PD 174265	6	1	1	2	1	1	12	6	5	131	18	13
766949	JNK Inhibitor IX	2			13	2	2	49	28	13	18	104	71
5162	SB 202474				118	2	2	88	69	40	23	122	117
11624601	JNK Inhibitor VIII	7	5	2	4	2	2	4	3	2	39	25	16
4707	PD 158780	19	5	2	11	2	2	18	18	34	135	28	15
2051	AG 1478	38	18	2	29	2	2	38	53	46	138	11	10
9549299	EGFR Inhibitor	4	2	4	1	3	6	1	1	1	7	12	11
16760284	Akt Inhibitor X	4			33	3	6	56	35	18	11	102	81
5278396	ATM Kinase Inhibitor				114	3	6	93	70	45	27	121	123
5353593	DMBI	13	2	4	14	3	6	22	17	15	43	79	49
448014	GSK-3b Inhibitor VIII	9	2	4	6	3	6	17	10	7	52	67	33
3385203	GTP-14564	30	2	4	23	3	6	37	27	33	97	103	70
2044	AG 1024	5			36	4	7	55	45	30	47	112	99
2427	BPIQ-I	36	3	5	32	4	7	30	42	29	77	51	24
9969021	Chk2 Inhibitor II	14	3	5	7	4	7	9	8	6	110	9	7
11665831	JNK Inhibitor	15	1		26	5	1	41	33	27	103	117	91
6711154	PKCbII/EGFR Inhibitor	6	4	1	3	5	1	23	14	11	113	96	56
10907042	PDGF Receptor Tyrosine Kinase Inhibitor III	15	4	1	17	5	1	35	26	14	118	111	78
9549300	Tpl2 Kinase Inhibitor	10	4		28	5	8	51	40	28	119	127	102
2048	AG 1295				120	5	8	94	73	52	133	171	139
5291	Imatinib	32	11	7	18	6	9	8	22	23	12	32	26
5228	SKF-86002	16	5	7	8	6	2	10	5	10	16	36	23
176870	Erlotinib	28	5	2	20	6	2	28	23	19	54	46	25
16079009	TGF-b RI Inhibitor III	7			43	6	9	58	36	16	78	126	106
312145	Wortmannin	11	11	7	8	6	2	2	2	9	145	16	14
5329468	VEGF Receptor 2 Kinase Inhibitor IV	24	6	3	30	7	3	47	43	49	160	132	85
9549295	KRN633	29	7	9	22	8	5	27	25	35	107	90	53
2857	Compound 56	42	16	10	35	9	6	42	56	41	72	31	19
176155	SB 203580	22	4	1	24	10	8	26	32	38	30	74	65
16760635	Rho Kinase Inhibitor IV	27	14	6	19	10	8	16	31	32	31	19	29
208908	Lapatinib	11	11	13	4	11	14	13	7	3	9	22	39
4665	p38 MAP Kinase Inhibitor	23	15	13	18	11	14	21	29	37	115	64	46
10341154	VX-702	12	12	8	5	12	3	5	4	4	40	23	21
6419834	VEGF Receptor 2 Kinase Inhibitor I	24	12	3	21	12	3	43	30	24	156	125	87
9817165	AGL 2043	13	8		27	13	10	46	39	17	24	93	88
6539732	GSK-3b Inhibitor II	4			90	13	6	71	46	22	46	119	108
11493598	FR180204	9			57	13	6	68	61	50	56	113	115
2422	Bohemine	9			89	13	10	74	64	53	96	147	129
11617559	cFMS Receptor Tyrosine Kinase Inhibitor	14	9	11	9	14	11	20	13	8	20	52	57
10020713	GSK-3b Inhibitor XI	26	13	11	16	14	11	29	21	36	109	92	62
5353940	SB 202190	34	14	6	31	15	8	24	52	55	13	55	48
151194	Vatalanib	15			46	15	13	60	58	51	37	101	119
644354	Rho Kinase Inhibitor III	6			66	15	8	73	57	54	50	114	122
160355	Roscovitine	15	4		15	15	8	34	20	20	57	89	83
4712	PD 169316	43	17	12	40	15	13	40	59	62	85	68	74
5287855	PI 3-Kg Inhibitor II	15	4		12	15	8	36	19	12	111	107	90
448042	Y-27632	15	1		37	15	8	44	49	43	121	124	112
9797919	VEGF Receptor Tyrosine Kinase Inhibitor II	22	10	1	24	15	13	32	38	48	136	108	92
5289247	PI 3-Kg Inhibitor	32	15	13	18	16	9	11	16	42	95	21	32
9884685	PI-103				124	16	2	105	86	59	155	176	156
3973	LY 294002	3			110	17	4	81	62	57	73	146	127
9843206	EGFR/ErbB-2 Inhibitor	33	16	14	27	18	10	31	41	47	3	44	59
9549303	Aurora Kinase Inhibitor III	13			101	18	16	83	80	61	81	141	135
5288600	IC261	4			111	18	10	85	65	56	83	144	130
449241	H-89	42	22	17	35	18	6	39	48	44	88	1	1
3547	Fasudil	25	2		39	18	6	52	44	60	105	123	110
6419753	D4476	17			41	19	17	57	37	26	70	109	107
6918386	Cdk2 Inhibitor III	17			64	19	11	70	55	58	91	131	125
5164	SB220025	22	17	19	15	20	13	3	12	98	10	10	22
9818231	Tofacitinib	18	17	19	12	20	13	7	11	78	34	13	27
3820	Kenpaullone	39	20	16	34	20	13	45	68	81	53	59	68
2063	AG 9				121	20	1	108	82	65	94	161	143
9860529	DNA-PK Inhibitor II				117	21	10	89	76	64	33	118	128
16760346	TBCA	33	19	17	50	21	22	62	50	99	137	105	105
11772958	Flt-3 Inhibitor III	49	27	11	51	22	7	66	63	63	89	62	35
10074640	Masitinib	37	23	23	31	23	13	15	51	74	1	15	18
3467590	STO-609	45	23	12	46	23	13	59	60	67	45	58	60
9549296	Syk Inhibitor II	45	17	12	59	23	13	69	78	66	149	110	116
5154691	NM-PP1	54	33	20	54	24	14	72	74	68	80	29	28
5329155	VEGF Receptor 2 Kinase Inhibitor II	40	25	24	38	25	21	54	67	86	142	65	63
1048845	Flt-3 Inhibitor	21			88	26	25	78	66	139	75	116	131
11422035	JNK Inhibitor V	43	31	19	40	27	19	53	81	91	74	37	40
5329098	VEGF Receptor 2 Kinase Inhibitor III	59	32	23	63	27	13	80	79	79	130	73	58
9797370	PDGF Receptor Tyrosine Kinase Inhibitor IV	41	21	7	49	28	23	67	77	110	86	99	114
11644425	MNK1 Inhibitor	8			109	29	15	90	87	76	112	157	146
11566580	EGFR/ErbB-2/ErbB-4 Inhibitor	35	26	26	25	30	34	6	34	161	2	5	6
11983295	IRAK-1/4 Inhibitor	33	19	4	44	30	22	50	47	102	58	91	95
9549301	Cdk/Crk Inhibitor	47	35	28	45	31	31	33	83	116	42	3	3
2794188	CGP74514A	46	28	22	52	32	18	63	71	101	104	81	69
5312122	KN-93				143	33	1	121	92	71	68	156	133

447966	TGF-b RI Kinase Inhibitor	20			98	34	24	84	85	127	67	120	136
1893668	TG003	35	24	14	47	35	25	61	72	123	63	87	121
2856	Compound 52	50	30	15	51	36	17	77	84	73	129	97	82
1474853	Src Kinase Inhibitor I	51	36	29	48	37	32	65	89	118	6	2	2
4987	Purvalanol A	53	32	25	53	37	19	75	88	89	65	45	51
2399	BisindolylmaleimideIV	48	29	17	62	38	22	79	90	88	87	98	111
3795	JAK3 Inhibitor II	48	34	26	55	38	22	64	75	107	106	47	42
11601743	Flt-3 Inhibitor II	67	42	26	76	38	30	97	91	113	117	88	80
4124851	GSK-3b Inhibitor I				128	39	11	115	93	70	66	150	138
672296	Syk Inhibitor III				122	40	8	119	96	75	84	158	150
5289419	SU9516	69	40	30	70	40	32	86	95	112	101	4	4
5328804	AG 112				127	40	13	107	97	96	120	165	152
16760619	PKR Inhibitor, Negative Control	55	36	29	61	40	32	87	105	138	132	75	86
5113385	GNF-Pf-3800	65	44	30	67	40	41	99	145	154	165	56	75
3004085	NSC 625987				123	41	16	116	94	72	126	167	163
5330797	Cdk4 Inhibitor	74	39	27	83	41	34	91	109	128	154	95	97
216239	Sorafenib	45	32	32	42	42	36	48	118	148	4	20	41
3641059	RP106	44	8		89	43	16	96	113	87	127	143	144
9903786	IKK-2 Inhibitor IV	56	37	27	60	43	30	82	107	155	150	83	103
5472558	Cdk1 Inhibitor				130	44	12	122	102	85	123	168	160
16760609	PDGF RTK Inhibitor	59	44	32	61	45	32	103	120	131	76	39	45
4879	PP3				133	46	3	118	98	80	162	177	165
9549297	Met Kinase Inhibitor	71	48	34	75	47	37	112	122	143	108	53	43
16760502	Herbimycin A				129	48	22	124	99	100	98	162	155
16760391	DNA-PK Inhibitor V				143	49	17	128	100	82	41	139	137
5326739	Indirubin-3?-monoxime	75	47	33	79	49	26	110	108	132	159	70	76
5494449	Tozasertib	62	40	35	68	50	38	98	125	163	69	60	98
3971	LY 303511				124	51	2	125	101	69	62	152	147
5326843	RP107	52	24	4	87	52	25	106	124	115	152	145	142
6419741	Isogranulatimide	81	49	31	88	52	22	104	106	105	167	84	47
73292	Fascaplysin				119	53	17	117	104	77	141	170	168
8515	JNK Inhibitor II	60	35	18	78	53	31	111	115	114	147	115	132
10113978	Pazopanib	61	38	29	69	54	38	101	114	145	51	85	124
6538818	GSK-3 Inhibitor X	64	41	21	82	55	26	92	103	120	151	106	109
6419755	PKCb Inhibitor	66	43	37	72	56	41	102	147	141	64	50	73
644241	Nilotinib	58	46	38	58	57	39	109	143	150	8	30	55
2049	AG 1296				125	57	20	129	110	93	166	178	175
3499	G? 6983	63	45	39	65	58	40	100	138	157	139	54	67
5084	Ro-32-0432	57	43	37	56	59	49	95	163	174	14	35	84
6419766	GSK-3 Inhibitor XIII	76	47	36	80	60	35	114	131	121	157	76	72
6604931	Aminopurvalanol A	72	49	41	73	61	42	113	154	153	116	8	8
10202471	NU6140	77	49	39	81	61	46	120	162	164	161	69	93
3081361	Vandetanib	70	51	40	74	62	45	127	148	151	21	57	96
5005498	Alsterpaullone	68	47	42	71	63	43	123	146	130	92	40	36
481747	Cdk4 Inhibitor III				126	64	17	126	117	95	148	172	171
16760286	Alsterpaullone-2	80	56	44	86	65	46	139	151	142	122	48	44
6603792	Lck Inhibitor	55	50	46	56	66	53	158	171	168	17	27	66
509554	NF-kB Activation Inhibitor				142	66	8	131	116	84	38	140	141
5312137	SU6656	78	53	43	85	66	44	134	149	146	90	66	77
2396	BisindolylmaleimideI	73	52	45	77	67	50	135	164	160	134	34	34
11382492	MK2a Inhibitor				134	68	14	132	112	83	48	149	145
657356	Diacylglycerol Kinase Inhibitor II				141	69	22	130	111	92	26	130	134
9549289	TWS119	83	54	48	91	70	47	145	156	137	114	7	9
5330790	Aurora Kinase/Cdk Inhibitor	82	55	50	92	71	46	137	152	158	140	17	17
438981	Cdk1/5 Inhibitor				139	71	25	133	121	111	153	174	172
11524144	Compound C	85	58	47	94	71	40	148	141	133	158	14	12
6419747	Syk Inhibitor	86	57	49	96	72	51	153	157	165	163	63	89
5353431	BAY 11-7082				135	73	10	138	130	94	60	151	154
5287844	GSK-3 Inhibitor IX	84	60	51	93	74	48	156	159	152	146	6	5
9977819	Dovitinib	86	61	52	95	75	56	161	169	169	125	71	101
10196499	Akti-1/2				139	76	10	136	119	90	15	128	140
5494425	JAK Inhibitor I	87	59	53	99	76	54	164	167	167	164	42	52
2703	Chelerythrine Chloride				138	77	27	142	123	119	124	169	170
4713	PD 98059				132	78	19	140	126	109	143	173	173
9859309	DNA-PK Inhibitor III				135	79	34	141	129	124	35	137	148
9951490	MEK Inhibitor I				132	80	41	150	144	122	32	135	162
753704	Sphingosine Kinase Inhibitor				132	81	13	143	128	103	36	142	149
5329102	Sunitinib	89	63	54	102	82	57	165	166	175	169	100	126
6419764	Indirubin Derivative E804	90	65	55	104	83	59	168	168	166	173	82	100
16760627	Staurosporine, N-benzoyl-	91	66	61	103	84	20	159	140	108	168	38	30
5328779	AG 490				134	85	33	144	132	140	82	160	167
3062316	Dasatinib	79	62	56	84	86	64	177	175	173	25	26	61
5311510	Bcr-abl Inhibitor				135	87	40	146	142	125	99	163	169
6605258	ATM/ATR Kinase Inhibitor				131	88	34	147	136	126	22	129	153
5328940	Bosutinib	88	64	57	100	89	61	170	173	177	128	72	118
16760631	Rapamycin				134	90	14	149	127	97	44	148	151
9549284	MEK1/2 Inhibitor				137	91	28	151	139	117	29	138	158
389898	MEK Inhibitor II				136	92	29	155	135	136	28	133	159
6490494	PKR Inhibitor	95	69	60	108	93	62	169	165	162	174	24	31
3501	G? 6976	93	70	58	106	94	65	173	176	178	172	49	64
16760529	KN-62				143	95	14	154	133	104	55	153	157
5329103	SU11652	94	67	59	107	96	63	172	172	171	170	94	104
16760524	JAK3 Inhibitor VI	92	68	62	105	97	66	174	177	172	171	80	79
5719375	Akt Inhibitor IV				139	98	30	152	134	134	71	154	166
6444692	Mubritinib				143	99	19	157	137	106	59	155	161
5330548	PDGF Receptor Tyrosine Kinase Inhibitor II				143	100	52	160	153	144	144	175	178
9549305	IGF-1R Inhibitor II				140	101	55	162	161	156	79	159	174
3387354	SB 218078	96	71	63	113	102	68	178	178	176	176	61	54
5330812	Cdk1/2 Inhibitor III	97	72	64	112	103	69	176	174	170	175	41	120
490561	K-252a	98	73	65	115	104	67	171	160	135	177	86	94
5339183	ERK Inhibitor III				139	105	60	167	155	147	100	164	177
290486	Triciribine				139	106	54	163	150	129	19	136	164
176406	JAK3 Inhibitor IV				142	107	58	166	158	149	93	166	176
451705	Staurosporine	99	74	66	116	108	70	175	170	159	178	77	50

Table S3: Ranking of the different metrics computed for the shared compounds between Ambit dataset (A) and Reaction Biology dataset (B). The first two columns give the Puchem identifier and the common name of the molecules, the twelve other columns give the ranks for each metric. For each ranking column, we colored the cells from red to blue, from the most to the worst selective, respectively. Empty cells symbolize compounds for which we could not calculate a selectivity value for the given metric.

A.

pubchem_id	Compound_Name	Rank_S(pKd5)	Rank_S(pKd6)	Rank_S(pKd7)	Rank_WS(pKd2)	Rank_WS(pKd1)	Rank_WS(pKd0.5)	Rank_RS(20)	Rank_RS(10)	Rank_RS(5)	Rank_Gini	Rank_Ssel	Rank_P_{MAX}
123631	Gefitinib	10	7	1	1	1	1	3	5	3	10	1	1
176870	Erlotinib	11	9	4	2	1	1	5	4	4	8	2	2
3038522	MLN-518	7	5	4	5	6	7	4	6	5	14	5	9
176155	SB-203580	8	6	7	9	9	9	11	11	17	12	13	13
5291	Imatinib	6	5	6	8	8	10	8	8	18	13	10	15
11617559	GW-2580	1	3	2	4	2	3	1	1	1	18	3	3
208908	Lapatinib	2	2	3	3	4	5	2	2	2	19	4	4
10074640	AB-1010	5	8	7	11	8	4	10	13	14	11	11	12
10113978	Pazopanib	12	10	8	8	7	4	7	9	10	6	8	8
216239	Sorafenib	13	11	9	12	8	6	13	12	15	5	12	10
5494449	VX-680/MK-0457	16	14	11	15	9	10	14	15	19	3	14	18
3081361	Vandetanib	14	13	11	16	7	4	16	14	9	2	16	11
644241	Nilotinib	9	12	10	14	11	8	15	18	16	7	17	17
5329102	Sunitinib	18	17	14	7	5	2	9	7	7	9	7	6
3062316	Dasatinib	15	15	12	13	12	4	19	19	13	1	15	14
5328940	SKI-606	17	16	13	10	3	1	12	10	6	4	6	5
451705	Staurosporine	19	18	15	17	10	6	17	17	12	17	18	16
151194	PTK-787	3	4	5	6	7	4	6	3	8	15	9	7
9884685	PI-103	4	1		18	10	6	18	16	11	16	19	19

B.

pubchem_id	Compound_Name	Rank_S(50%)	Rank_S(70%)	Rank_S(80%)	Rank_WS(20%)	Rank_WS(10%)	Rank_WS(5%)	Rank_RS(20)	Rank_RS(10)	Rank_RS(5)	Rank_Gini	Rank_Ssel	Rank_Pmax
123631	Gefitinib	7	4	1	5	1	1	6	6	6	14	7	2
176870	Erlotinib	6	2	1	6	2	1	8	5	3	12	9	3
3038522	MLN-518	5	1	1	3	1	1	4	2	8	3	8	5
176155	SB-203580	4	2	1	6	3	2	7	7	4	8	13	11
5291	Imatinib	6	4	2	4	2	2	1	4	5	7	5	4
11617559	GW-2580	3	3	3	2	5	3	5	3	2	6	10	9
208908	Lapatinib	2	5	4	1	4	5	2	1	1	4	3	6
10074640	AB-1010	8	6	5	7	6	4	3	8	10	1	1	1
10113978	Pazopanib	11	8	6	11	8	7	11	12	11	13	16	17
216239	Sorafenib	9	7	7	9	7	6	9	11	12	2	2	7
5494449	VX-680/MK-0457	12	9	8	12	8	8	12	13	16	15	12	14
3081361	Vandetanib	13	11	9	13	10	10	15	15	14	10	11	13
644241	Nilotinib	10	10	10	10	9	9	14	14	13	5	6	10
5329102	Sunitinib	15	13	11	16	11	11	16	16	18	18	18	18
3062316	Dasatinib	14	12	12	14	12	13	19	19	17	11	4	12
5328940	SKI-606	16	14	13	15	13	12	17	18	19	16	15	16
451705	Staurosporine	17	15	14	17	14	14	18	17	15	19	14	8
151194	PTK-787	1			8	3	4	10	9	7	9	17	15
9884685	PI-103				18	3	1	13	10	9	17	19	19

Table S4: Ranking of the different metrics computed for the compounds tested in the NCI-60 assay. The first column gives the common name of the molecules, the twelve other columns give the ranks for each metric. For each ranking column, we colored the cells from red to blue, from the most to the worst selective, respectively. Empty cells symbolize compounds for which we could not calculate a selectivity value for the given metric.

Molecule_Name	Rank_S(pKd5)	Rank_S(pKd6)	Rank_S(pKd7)	Rank_WS(pKd2)	Rank_WS(pKd1)	Rank_WS(pKd0.5)	Rank_RS(20)	Rank_RS(10)	Rank_RS(5)	Rank_Gini	Rank_Ssel	Rank_P_{MAX}
Imatinib	2	1	1	1	1	1	2	1	1	6	1	1
Nilotinib	11	2	1	2	1	1	3	2	2	15	2	2
Bosutinib	14	11	2	11	2	2	12	9	4	14	10	3
Sunitinib	13	5	1	14	1	1	13	6	3	17	13	4
Cabozantinib	15	6	3	8	4	3	8	5	7	13	7	5
Ibrutinib	8	7	4	5	6	5	4	4	10	4	3	6
Gefitinib	13	7	3	12	4	3	11	8	8	10	11	7
Ponatinib	15	16	4	17	3	1	15	12	6	24	17	8
Crizotinib	15	8	3	16	3	3	14	13	5	20	14	9
Vemurafenib	15	8	6	3	8	7	5	3	13	8	6	10
Lapatinib	10	7	4	9	6	5	9	7	9	7	8	11
Erlotinib	6	7	1	9	9	7	10	14	15	3	12	12
Afatinib	15	13	5	7	5	6	7	10	17	18	4	13
Axitinib	7	4		15	7	6	16	15	14	5	15	14
Selumetinib	5	12	8	6	11	11	6	19	20	2	9	15
Dabrafenib	3	10	7	4	10	9	1	11	23	1	5	16
Orantinib	1			18	18	4	18	16	11	21	22	17
Dasatinib	12	14	9	10	12	11	17	20	18	11	16	18
Regorafenib	15	3		18	16	6	20	17	12	22	21	19
Vandetanib	15	9		18	14	10	19	21	19	19	20	20
Fostamatinib	4			18	17	12	21	22	21	9	19	21
Pazopanib	9			18	15	8	22	18	16	16	23	22
Trametinib	13	15	10	13	13	13	24	24	23	12	18	23
Sorafenib	15			18	19	14	23	23	22	23	24	24

IV. Identification des résidus favorisant la liaison d'inhibiteurs de protéines kinases de Type II

A. La conception d'inhibiteurs de Type II : un challenge difficile à atteindre

Nous l'avons vu dans la partie consacrée (Partie II.D.4.c), les inhibiteurs de protéines kinases de Type II inhibent leurs cibles dans la conformation inactive DFG-out. Dans celle-ci, la phénylalanine du motif DFG, située sur la boucle d'activation, effectue une rotation d'environ 180° et permet ainsi l'accès à une poche hydrophobe dans laquelle les inhibiteurs de Type II peuvent se lier.[157] Lancé en 2001, l'imatinib a ouvert la voie au développement de tels inhibiteurs.[73] Après quatorze ans, six autres inhibiteurs de Type II ont été approuvés par la FDA et force est de constater que, ce qui aurait pu devenir la classe prédominante d'inhibiteurs, ne semble pas convenir pour toutes les protéines kinases. Bien que longtemps considérés comme une alternative plus sélective aux inhibiteurs de Type I, du fait d'une moindre conservation des résidus situés dans la poche hydrophobe, de récentes études mettent en doute cette idée préconçue.[75,76] Si les inhibiteurs de Type II ne représentent pas forcément la source de molécules sélectives tant attendues, ils n'en demeurent pas moins particulièrement intéressants à la vue de la diversité chimique qu'ils apportent, un point à prendre à compte, notamment au niveau de la propriété intellectuelle.

De manière intéressante, après avoir analysé notre base de données (Partie III.B), nous avons noté que certaines protéines kinases ne semblent jamais être inhibées (Ki, Kd ou IC50 < 1 µM, ou pourcentage d'inhibition > 50%) par des inhibiteurs de Type II (Table 16). Parmi ces protéines, on retrouve, entre autres, les protéines AKT, DYRK et SYK. Ces résultats peuvent s'expliquer de deux manières. La première hypothèse est que la diversité chimique des inhibiteurs de Type II n'est, à l'heure actuelle, pas assez importante. En reprenant l'analogie de la serrure et de la clef, souvent utilisée pour décrire le phénomène de liaison entre une protéine et un ligand, il est possible que la communauté scientifique n'ait pas encore trouvé les clés qui siéent à ces serrures. Nul doute dans ce cas que, dans un futur proche, nous devrions voir apparaître des inhibiteurs de Type II structuralement différents et ciblant de nouvelles protéines kinases. La seconde hypothèse est que certaines protéines kinases ne peuvent pas adopter la conformation *DFG-out*, essentielle pour les inhibiteurs de Type II puissent s'y lier correctement. Officieusement, deux courants de pensées s'opposent à ce sujet. Certains pensant que toutes les protéines du kinome évoluent dans un équilibre conformationel *DFG-in*/*DFG-out* essentiel à la régulation de leur activité. L'absence de structures *DFG-out*, pour certaines protéines kinases, pourrait donc, en théorie, disparaitre avec le temps. D'autres, en revanche, considèrent que l'absence de cette

[157] Treiber, D., Shah, N. (2013), Ins and Outs of Kinase DFG Motifs, *Chem. Biol.*, 20, 745-46.

conformation pour une large tranche du kinome, est le signe de leur incapacité à passer d'une conformation à l'autre et qu'il existe des résidus des protéines kinases qui seraient responsables de ce mécanisme structural.

AAK1	CAMK4	CLK4	ERBB2	JAK2	MAP3K9	MKNK1	PAK5	PKMYT1	RIOK1	SRPK3	TNK1
ACVR1	CAMKK1	CSNK1A1	ERBB3	JAK3	MAP4K1	MKNK2	PAK6	PKN1	RIOK2	STK10	TNK2
ACVR1B	CAMKK2	CSNK1A1L	ERN1	KIT	MAP4K2	MLCK	PAK7	PKN2	RIOK3	STK11	TNNI3K
ACVR2A	CASK	CSNK1D	FER	LATS1	MAP4K3	MOS	PASK	PLK1	RIPK1	STK16	TRPM6
ACVR2B	CDC2	CSNK1E	FES	LATS2	MAP4K4	MST1R	PBK	PLK2	RIPK4	STK17A	TSSK1B
ACVRL1	CDC2L1	CSNK1G1	FGFR3	LIMK1	MAP4K5	MST4	PCTK1	PLK3	RIPK5	STK17B	TSSK2
ADCK3	CDC2L2	CSNK1G2	FGFR4	LIMK2	MAPK1	MYLK	PCTK2	PNCK	ROCK1	STK24	TSSK3
ADCK4	CDC2L5	CSNK1G3	FRAP1	LRRK2	MAPK10	MYLK2	PCTK3	PRKAA1	ROCK2	STK25	TTK
ADRBK1	CDC2L6	CSNK2A1	GAK	LTK	MAPK12	MYLK4	PDK1	PRKAA2	ROS1	STK3	TXK
ADRBK2	CDC42BPA	CSNK2A2	GRK1	LYN B	MAPK13	MYO3A	PDK2	PRKACA	RPS6KA1	STK32A	TYK2
AKT1	CDC42BPB	DAPK1	GRK4	MAK	MAPK3	MYO3B	PDK4	PRKACB	RPS6KA2	STK32B	TYRO3
AKT2	CDC42BPG	DAPK2	GRK5	MAP2K1	MAPK4	NEK1	PDPK1	PRKACG	RPS6KA3	STK32C	ULK1
AKT3	CDC7	DAPK3	GRK6	MAP2K2	MAPK6	NEK11	PFTK1	PRKCA	RPS6KA4	STK33	ULK2
ALK	CDK2	DCLK1	GRK7	MAP2K4	MAPK7	NEK2	PFTK2	PRKCB	RPS6KA5	STK35	ULK3
ANKK1	CDK3	DCLK2	GSG2	MAP2K5	MAPK8	NEK3	PHKG1	PRKCD	RPS6KA6	STK36	VRK1
AXL	CDK4	DCLK3	GSK3A	MAP2K7	MAPK9	NEK4	PHKG2	PRKCE	RPS6KB1	STK38	VRK2
BMP2K	CDK5	DMPK	GSK3B	MAP3K1	MAPKAPK2	NEK5	PI4KB	PRKCG	RPS6KB2	STK38L	VRK3
BMPR1A	CDK6	DYRK1A	HIPK1	MAP3K10	MAPKAPK3	NEK6	PIK3C2B	PRKCH	SBK1	STK39	WEE1
BMPR1B	CDK7	DYRK1B	HUNK	MAP3K11	MAPKAPK5	NEK7	PIK3C2G	PRKCI	SBK2	STK4	WEE2
BMPR2	CDK9	DYRK2	ICK	MAP3K12	MARK1	NEK9	PIK3CA	PRKCQ	SGK1	SYK	WNK1
BRSK1	CDKL1	DYRK3	IGF1R	MAP3K13	MARK2	NIM1	PIK3CB	PRKCZ	SGK2	TAOK1	WNK2
BRSK2	CDKL2	DYRK4	IKBKB	MAP3K14	MARK3	NLK	PIK3CD	PRKD1	SGK3	TAOK3	WNK3
BTK	CDKL3	EEF2K	IKBKE	MAP3K15	MARK4	NUAK1	PIK3CG	PRKD2	SIK3	TBK1	ZAP70
CAMK1	CDKL5	EGFR	INSR	MAP3K2	MAST1	NUAK2	PIM1	PRKD3	SLK	TEC	
CAMK1D	CHEK1	EIF2AK1	INSRR	MAP3K3	MATK	OSR1	PIM2	PRKG1	SNF1LK	TESK1	
CAMK1G	CHEK2	EIF2AK2	IRAK1	MAP3K4	MELK	OXSR1	PIM3	PRKG1b	SNF1LK2	TGFbR1	
CAMK2A	CHUK	EIF2AK3	IRAK3	MAP3K5	MERTK	PAK1	PIP4K2B	PRKG2	SNRK	TGFbR2	
CAMK2B	CIT	EIF2AK4	IRAK4	MAP3K6	MET	PAK2	PIP5K1A	PRKX	SRMS	TLK1	
CAMK2D	CLK2	EPHA1	ITK	MAP3K7	MINK	PAK3	PIP5K1C	PRPF4B	SRPK1	TLK2	
CAMK2G	CLK3	EPHB6	JAK1	MAP3K8	MINK1	PAK4	PIP5K2C	PTK2	SRPK2	TNIK	

Table 16 : Liste des protéines kinases non inhibées par les sept inhibiteurs de Type II approuvés par la FDA. Les données proviennent de notre base de données interne, dans laquelle ont été assemblés plusieurs jeux de données publics (Table 15).

Pour vérifier la première hypothèse il nous faudra peut-être attendre encore quelques années et voir si des inhibiteurs de Type II émergent et parviennent à bloquer l'activité de ces protéines kinases, pour l'instant insensibles à cette classe. Il sera alors intéressant d'étudier si leurs structures sont radicalement différentes de celles déjà existantes. De même, les progrès de la cristallographie aux rayons X amèneront peut-être à résoudre la structure de nouvelles protéines en conformation *DFG-out*. Nous nous sommes donc penchés sur la deuxième hypothèse. En utilisant une approche protéométrique, nous avons tenté d'identifier les résidus qui pourraient être responsables de l'absence de conformation *DFG-out*. Il est possible de répondre à cette question en prenant en compte les résidus qui sont impliqués dans la liaison des inhibiteurs de Type II et qui discriminent les protéines kinases actives des protéines kinases inactives. Ces résultats ont été valorisés par une publication récemment acceptée par le journal *ACS Chemical Biology*. Notons qu'à notre connaissance, il

s'agit de la première étude cherchant à déterminer l'effet de résidus de protéines kinases dans la liaison d'inhibiteurs, à partir d'un modèle d'apprentissage automatique.

B. Publication : *A Proteometric Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues*

A Proteometric Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues

Nicolas Bosc[1], Berthold Wroblowski[2], Samia Aci-Sèche[1], Christophe Meyer[3], and Pascal Bonnet[1*]

[1]*Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France*
[2]*Janssen Research & Development, a division of Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium*
[3]*Centre de Recherche Janssen-Cilag, Campus de Maigremont - CS 10615, 27106 Val de Reuil CEDEX*

[*]Author to whom all correspondence should be addressed:
Pascal Bonnet: Tel: +33 238 417 254, Fax: +33 238 417 254, E-mail: pascal.bonnet@univ-orleans.fr

ABSTRACT

Since the success of imatinib, the first Type-II kinase inhibitor approved by the FDA in 2001, sustained efforts have been made by the pharmaceutical industry to discover novel compounds stabilizing the inactive conformation of protein kinases. On the seven Type-II inhibitors having reached the market, four were released in 2012 suggesting an acceleration of the research of such class of compounds. Still, they represent less than a third of the protein kinase inhibitors available to patients today. The identification of key residues involved in the binding of this type of ligands in the kinase active site might ease the design of potent and selective Type-II inhibitors. In order to identify those discriminant residues, we have developed a proteometrics approach combining residue descriptors of protein kinase sequences and biological activities of various Type-II kinase inhibitors. We applied Partial Least Squares (PLS) regression to identify 29 key residues that influence the binding of four Type-II inhibitors to most proteins of the kinome. The gatekeeper residue was found to be the most relevant confirming an essential role in ligand binding as well as in protein kinase conformational changes. Using the newly developed proteometrics model, we predicted the propensity of each protein kinase to be inhibited by Type-II ligands. The model was further validated using an external dataset of protein/ligand activity pairs. Other residues present in the kinase domain, and more specifically in the binding site, have been highlighted by this approach, but their role in biological mechanisms is still unknown.

KEYWORDS: proteometrics / protein kinases / sequence / protein conformation / Type-II inhibitors / partial least squares regression / discriminant residue

INTRODUCTION

Protein kinases are involved in the mechanism of protein phosphorylation and contribute to many biological processes. (*1*,*2*) Variation in protein kinase expression may occur when their regulation pathway is modified, or when their nucleic acid sequence is mutated. Such modification of protein expression may lead to serious diseases such as cancer, diabetes, neuronal dysfunction or inflammatory diseases. (*3*,*4*) Hence, significant resources have been allocated since several years to find potent and selective protein kinase inhibitors with improved therapeutic efficacy.

Since the first protein kinase inhibitor approved by the FDA in 2001, 28 molecules targeting this protein family have been marketed. (*5*) So far, we count about twenty primary targets inhibited by these compounds out of a total of 518 protein kinases identified in the human genome. (*2*) The high degree of sequence similarity between all kinase domains makes it challenging to design potent and selective inhibitors. Protein kinase inhibitors have been classified depending on their mode of interaction. While few kinase inhibitors were developed to covalently bind protein kinases, (*6*) such as afatinib (*7*) or ibrutinib, (*8*) all other inhibitor classes bind non-covalently to the ATP binding site (Type-I and Type-II) or in an allosteric pocket (Type-III) such as trametinib. (*9*) Type-I inhibitors bind into the binding site of the active conformation of protein kinases by mimicking the ATP whereas Type-II inhibitors extend from the ATP site to an allosteric back pocket. Thus, Type-I and Type-II protein kinase inhibitors bind into protein kinases in two different manners due to important differences in their chemical structure. While Type-I inhibitors imitate the adenine moiety of ATP in protein kinase active site, (*10*) Type-II ligands in addition lay in a hydrophobic pocket adjacent to the ATP site. This pocket is not always reachable and its accessibility depends on the activation state of the protein kinase. The active conformation depends on the orientation of the DFG (Asp-Phe-Gly) motif, though other regions such as αC-helix or P-loop also seem to be involved in the process. In the inactive conformational state, Phe rotates away the nearby αC-helix (DFG-out) and projects into the ATP pocket revealing a hydrophobic cavity. (*11*) Combining the results of five screening studies, (*12*,*13*,*14*,*15*,*16*) around 1400 Type-I kinase inhibitors were found to be active (Ki, Kd and IC50 < 1μM, percentage of inhibition > 50%) on 496 protein kinases present in the panel. Interestingly, in the dataset published by Davis *et al.*, (*13*) 10 tested Type-II kinase inhibitors are never found to be active on almost 200 protein kinases such as AKT, DYRK or SYK, suggesting specific

structural motifs present in these proteins preventing the binding of Type-II ligands. Therefore, two hypotheses are proposed: the absence of an inactive state that prevents the opening of the hydrophobic pocket due to the presence of specific amino acids, or the existence of specific residues in the active site that prevents binding of the inhibitors. Unlike Type-I inhibitors that still represent the majority of current protein kinase inhibitors on the market with 17 drugs, only 7 Type-II inhibitors have been approved by the FDA so far (imatinib (2001), sorafenib (2005), nilotinib (2007), axitinib (2012), cabozantinib (2012), regorafenib (2012) and ponatinib (2012)). Because of their ability to bind a neighbor pocket of the ATP site, Type-II kinase inhibitors have been considered as being more selective than the Type-I class. However, a recent study demonstrated that Type-I and Type-II inhibitors may have similar selectivity profiles. (*13*) In addition, Zhao *et al.* found that Type-II inhibitors can target more than 200 protein kinases. However, some protein kinases appear to bind preferentially this type of inhibitors. (*17*) In this paper, we focused on the identification of specific discriminant residues that are present in protein kinases inhibited by Type-II inhibitors.

To improve our understanding of specificity upon binding of Type-II inhibitors and the propensity of protein kinases to selectively trap this inhibitor class, we studied amino acid occurrences that are the most significant for the binding of Type-II inhibitors using a novel proteometric approach. Proteometric methods can be compared to chemometrics which consists in using mathematical and statistical tools to rationalize chemical information variations. (*18*) Quantitative Structure-Activity Relationship (QSAR) is one application of chemoinformatics, a subfield of chemometrics (*19*) that focuses on the study of the interaction of a set of molecules with a protein target. QSAR uses every explanatory variable, such as molecular descriptors, to build a statistical model aiming at predicting a dependent variable such as the biological activity of the molecules. On the opposite, proteometrics takes into account information of one compound tested on several targets. If the proteins have similar three-dimensional structure and aligned sequences, the proteins can be described by residue descriptors (*20*) such as z-scales (*21*), VHSE (*22*), FASGAI (*23*) or MS-WHIM (*24*). Therefore, protein sequence descriptors can be associated to the affinity of a molecule in the development of statistical models.

Here, we analyzed the biological activities of four Type-II inhibitors on 263 protein kinases using a Partial Least Squares (PLS) regression, a suitable statistical method able to identify significant variables of linear systems. We identified the residues located in the kinase domain which influence most the receptor

inhibitory activity. We then mapped these residues on three-dimensional structures of protein kinases and attempted to infer their role upon Type-II inhibitor binding. Finally, we used the derived statistical model to predict affinities of these four Type-II inhibitors against non-tested protein kinases absent from the initial dataset. We found that our proteometrics model provides excellent predicted affinities compared to published experimental data.

METHODS

Data preparation

In this study, we generated a statistical model using a dataset of protein kinase-inhibitor interaction pairs published by Anastassiadis *et al.* (*12*) It contains 178 molecules tested against 300 protein kinases. They evaluated the percent remaining kinase activity for each kinase at 0.5μM compound concentration.

When some protein kinases were found in duplicates due to the presence of regulatory protein such as cyclin in CDK5/p25 and CDK5/p35, we retained only one complex. Indeed, the low standard deviation of biological activities between these complexes (5.7 percent) indicates a small influence of the cyclin domain. Hence, we kept the complex with the highest inhibition. Same rational was used for other kinases containing different cyclin proteins such as CDC2, CDK2, CDK3, CDK4, CDK6, CDK7 and CDK8 proteins. The maximum standard deviation observed amongst these complexes is only height percent of inhibition. After keeping one occurrence for each protein kinase name, we still had 277 remaining protein kinases.

As observed by Zhang *et al.* (*25*), inhibitors screened on FLT3, TRK, KHS and NUAK1 have possible large errors in their biological data and we therefore removed them from the dataset to avoid potential outliers. Additionally, we removed protein kinases having percentage of inhibition between 30 to 50% to avoid activity threshold bias. All data preparation steps have been performed using Knime pipelining software. **26** After this cleaning step, the final dataset containing a total 263 protein kinases corresponds to the training set used further for statistical modeling.

In order to validate our statistical model, we compared the predicted score with biological activities from an external set. (*13*) It contains 72 inhibitors tested against 442 protein kinases and an affinity (Kd) was measured for each pair. We standardized protein kinase names using UniProt (*27*) code and compound names with InChI (*28*) identifier in all databases for proper comparisons. By comparing protein names, we identified

144 protein kinases which were not included in the training set but for which Kd values are available. Therefore, these 144 proteins and their associated affinity data were used as an external test set to evaluate the robustness of the statistical model.

Protein alignment

Protein sequences included in the training and the external test sets were retrieved and aligned for the statistical modeling step. Sequences of the 518 human kinase domains were retrieved in FASTA format from the KinBase database, (*2*,*29*) out of which 444 typical kinase domain sequences were identified (Supplementary Table 1). The typical kinase domain sequences were read using MOE (Molecular Operating Environment, Chemical Computing Group Inc.) (*30*) sequence editor and annotated with the kinase annotation tool in order to identify the conserved motifs. Sequences were aligned with the protein alignment protocol in MOE and a progressive strategy was employed to build the alignment using kinase constraints. Blosum100 matrix was used to score the alignment, for 100 iterations using the default values for Gap Start and Gap Extend. The aligned sequences were saved in FASTA format and the accuracy of the overall alignment was tested by visually comparing conserved motifs.

In addition, to analyze the effect of the robustness of the statistical model on small modifications in the kinase sequence, we used structural information of protein kinase mutant sequences and we predicted their biological activities. These mutated sequences were obtained by manually replacing the mutated residues into the wild type sequences involved in well-known disease-related mutations such as T315I in ABL1, T790M in EGFR or V600E in BRAF.

Protein descriptors

Z-scale descriptors established by Sandberg *et al.* (*21*) were used to characterize each amino acid (Supplementary Table 2). These descriptors were obtained by applying principal component analysis on 26 physicochemical descriptors measured or calculated on 87 amino acids including the 20 natural amino acids. The first three principal components (z1 to z3) explained around 70% of the total variance of the original descriptors, while five components (z1 to z5) reach 87%. Thereby, the components may be interpreted as follow:

- z1: hydrophobicity

- z2: size and polarizability

- z3: polarity

- z4 and z5: electronic properties

We replaced each amino acid in the multiple sequence alignment by the five numerical z-scales and the gaps with five null values, resulting in a total of 4150 variables. On the training set, we removed variables with low variance (*31*) *i.e.* if the numerical values for each variable are identical at least 256 times. As a result, our training set containing 263 protein kinases was represented by 1995 remaining variables.

Type-II inhibitor selection

We identified the Type-II inhibitors present in each set. Because of their different mode of binding in protein kinases, Type-II inhibitors are structurally different from Type-I. (*32*) Several *in silico* approaches have been developed to identify Type-II inhibitors and to discriminate them from Type-I in a compound library. These strategies usually use primarily shape and pharmacophore-based virtual screening approaches. (*33*,*34*) To identify unknown Type-II kinase inhibitors from the training set, we calculated the chemical similarities between the compounds from this training set and the compounds co-crystallized in DFG-out conformation of protein kinase crystal structures. The X-ray structures were extracted from the Kinase Database available in MOE. (*30*) We calculated the Tanimoto similarities (Ts) between the compounds from the training set and the ligands from the crystallographic database using the ECFP6 fingerprint. We retrieved three molecules (imatinib, nilotinib and sorafenib) and one molecule with Ts of 0.70 (PDB entry 1T46 (*35*)) for masitinib. By visual inspection of their mode of binding in the crystal structure we tag them as Type-II inhibitors in the training set. By analyzing the biological activities measured by Anastassiadis *et al*., using an activity threshold of 50 percent of inhibition, we found that nilotinib is the least selective Type-II inhibitor with the highest number of inhibited protein kinases (25 inhibited proteins) (Figure 1). Looking at the available crystallographic structures containing Type-II inhibitors, we assumed that all Type-II inhibitors used in this study have similar mode of binding, *i.e.* they reach the allosteric pocket available upon the rotation of the DFG motif. Finally, for each kinase, we kept the highest biological activity amongst the four Type-II inhibitors and this value was used as the explained response in the statistical model (*Y* variable, see below).

Using the same procedure, we identified ten Type-II inhibitors in the external test set (AB-1010, ABT-869, AC220, AST-487, BIRB-796, EXEL-2880, imatinib, Ki-20227, nilotinib and sorafenib).

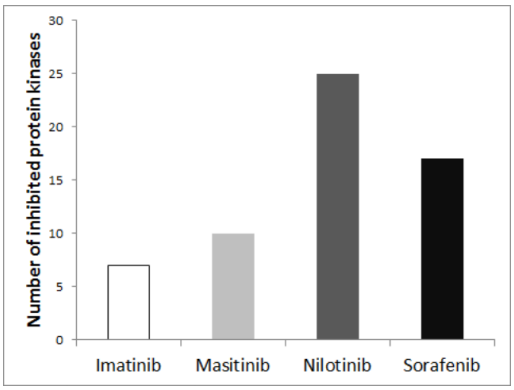


Figure 1: Bar Charts representing the number of protein kinases inhibited by each Type-II inhibitor identified in the dataset provided by Anastassiadis et al., (12) at a concentration of 0.5 µM. The inhibition threshold is set at 50%.

Statistical Approach

Partial Least Squares (PLS) regression is a statistical method that may be considered as a regression extension of Principal Component Analysis (PCA). (*36*,*37*) It finds a mathematical relation between two matrices X and Y of quantitative variables, by a linear multivariate model containing n observations. X is the matrix of predictive variables and Y the matrix of the dependent variable. The PLS projects the two matrices in a new multidimensional space constituted of latent variables and finds a linear regression between them. It tries to maximize the variance of the predictive variables x_i of X and maximizes the correlation between X and Y variables. Finally, the variable Y is explained by the x_i variables, giving such equation: $Y = k + a_1x_1 + a_2x_2 + \cdots + a_nx_n$, where a_i are the coefficients of x_i and k is a constant. The higher the absolute value of the coefficient a_i is, the more important it is for explaining Y. The equation can then be used to predict new Y values using new observations. We built the PLS model using the R statistical software, (*38*) and scripts from the 'chemometrics' package. (*31*,*39*,*40*) The scripts performed a repeated double cross-validation (rdCV (*31*,*41*)) to estimate the optimum number of components. This method estimates the optimal number of components using a first cross-validation then performs a second cross-validation to evaluate the performance of the model. As a reminder, a cross-validation starts by dividing the data into n segments. One is left out as an internal validation set and the remaining n-1 segments are used to generate the model. The performance of the

model is then estimated on the internal validation set. The process is repeated n times so that each segment is used as the internal validation set once. Using the training set containing 263 protein kinases we built a PLS model and evaluated its robustness with rdCV method. We selected the variables with the highest absolute regression coefficients based on the assumption that these variables contribute most to the Type-II activity. Because we had no indication on the number of absolute regression coefficients required to obtain the best model, we created several subsets by selecting sequentially 20 to 300 variables having the highest absolute coefficients. A first model was therefore established using the 1995 variables of the training set and then, by using the coefficient assigned to each variable, different numbers of variables have been selected. For each subset, we performed a rdCV and we evaluated the performance of each model with an error criteria. Here, we calculated the Standard Error of Prediction (SEP).

To estimate the robustness of our approach to predict percentages of inhibition, we performed an internal validation procedure. By selecting randomly 75% of the proteins from the training set, we created a PLS model and then predicted the percentages of inhibition of the 25% remaining proteins. The accuracy of the model was evaluated using the coefficient of determination (R2). To avoid any statistical bias, we performed 100 iterations by randomly selecting the 25% proteins. Finally, we used the statistical model to predict the tendency of 200 protein kinases, not included in the training set, to be inhibited by Type-II inhibitors. For 144 of these proteins we compared the predicted result with the external test set. Therefore, the predicted data will provide information on the most discriminant residues regarding the binding of the Type-II kinase inhibitors as well as the probability of protein kinases to be inhibited by those inhibitors.

Structure preparation

To visualize the position of the residues in protein kinase structures, we used 4 X-ray structures from the Protein Data Bank (PDB) using MOE. (*30*) Two structures are in a DFG-out conformation and two adopt a DFG-in conformation. One of the DFG-out conformations represents sorafenib bound to BRAF (PDB entry 1UWH (*42*)). Since the activation loop of this structure was missing, we built a homology model using the BRAF human sequence (UniProt entry P15056 (*27*)) and the PDB entry 1UWH as the template structure. The second structure (PDB entry 2HYY (*43*)) contains ABL1 co-crystallized with imatinib. Then, we selected two protein structures in DFG-in conformation, DYRK1A and SYK (PDB entries 2VX3 (*44*) and 1XBB (*45*)

respectively). While to our knowledge, DYRK1A has never been inhibited by Type-II inhibitors, SYK has been co-crystallized with imatinib. However, the crystal structure of the complex shows that imatinib binds weakly with a Type-I binding mode, adopting a characteristic U-shape instead of the extended conformation typically observed for Type-II inhibitors. (*32*) Therefore, it shows that Type-II inhibitor does not always trigger a conformational modification of its target. By mapping the residues onto the kinase structures, we can interpret their likely role during binding of Type-II inhibitors.

RESULTS AND DISCUSSION

Identification of important residues

The initial dataset contains remaining activities of 178 molecules tested at 0.5 μM against 300 protein kinases. After cleaning the dataset (see chapter Methods), 263 proteins have been retained and described using the z-scale descriptors. (*21*) For each protein kinase, we kept the maximum percentage of inhibition amongst the four Type-II inhibitors in the dataset (Table 1).

To determine which descriptors are the most important to discriminate the ability of protein kinases to bind any of the four Type-II inhibitors, we performed a PLS regression with the rdCV approach. We used the optimum parameters ascertained during the process where number of components and rdCV repetitions equal 2 and 30 respectively. A first model was built in which the low variance variables have been removed. The best variable selection was estimated by selecting successively the 20 to 300 variables with the highest absolute regression coefficients, and finally the total number of 1995 variables. Then each of these sets was validated using the rdCV method. We observed that 200 variables (Figure 2, Supplementary Table 3) gave better performance than 1995 variables (SEP= 12 versus SEP=20). Removing more variables involved a decrease of the model performance. With only 20 variables the performance of the statistical model brought down to the same level as using all 1995 variables.

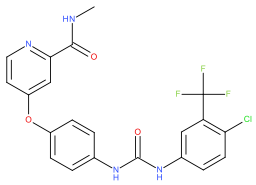
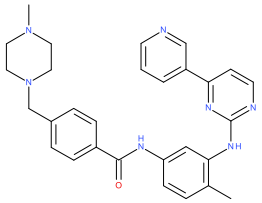
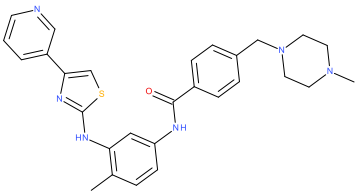
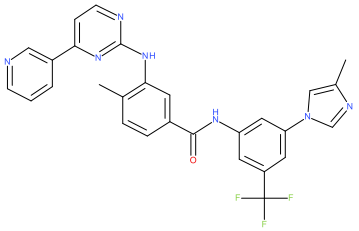
Name	Structure
sorafenib	
imatinib	
masitinib	
nilotinib	

Table 1: Type-II inhibitors identified among the 178 molecules tested by Anastassiadis et al. (12) and used in the study.

By looking at the sequence alignment, these 200 variables cover a total of 133 residue positions representing 13.6% of the global alignment. In the next part of the discussion, we will refer to the residue position according to the ABL1 sequence from our alignment (Supplementary Figure 4). To visualize only the most relevant information on protein kinase crystal structures, we applied two filters on residue selection. First, we kept only the residues containing at least two z-scale variables from the best 200 selected variables. We assume that the more a residue is represented by its z-scale descriptors the more it is important for the activity of the Type-II inhibitors. Secondly, we removed the z-scale variables when more than 20% of the position on the aligned sequences is represented by gaps. As a result we obtained 29 residue positions, characterized by only 70 regression coefficients (Table 2).

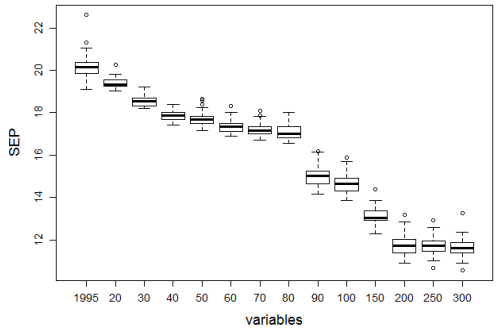


Figure 2: Effect of the variable removal on the Standard Error of Prediction (SEP) in the PLS model.

It is important to note that we also used the same approach to build four independent statistical models, each using activities of only one Type-II inhibitor. We applied the same protocol to select the most representative variables and obtained similar results (Supplementary Table 5) suggesting that the combination of the activities of the four Type-II inhibitors is a valid approach. In addition, to estimate the influence of the type of measured data, we used an external test set composed by experimental Kd provided by Davis et al., (*13*) (see Methods). Using the same method as previously described, with the same four Type-II inhibitors; we identified the 30 most important residues. Results are also similar when percentages of inhibition were used (Supplementary Table 6).

Residue positions	occurrences	corresponding variables	Rank sum_abs_coeff	Rank mean_diff	Rank max_diff
res409	**2**	**res409z2, res409z4**	**29**	**1**	**1**
res325	**4**	**res325z1, res325z2, res325z3, res325z4**	**1**	**2**	**2**
res414	**2**	**res414z1, res414z2**	**16**	**3**	**3**
res187	**2**	**res187z1, res187z5**	**14**	**4**	**4**
res426	**3**	**res426z3, res426z4, res426z5**	**2**	**10**	**5**
res296	**2**	**res296z2, res296z5**	**21**	**8**	**6**
res492	2	res492z2, res492z5	26	14	7
res226	3	res226z2, res226z3, res226z4	4	11	8
res778	3	res778z3, res778z4, res778z5	6	13	9
res192	**2**	**res192z4, res192z5**	**28**	**7**	**10**
res295	2	res295z1, res295z3	19	16	11
res785	3	res785z1, res785z3, res785z5	5	15	12
res118	2	res118z1, res118z3	15	9	13
res675	2	res675z3, res675z4	13	5	14
res329	**2**	**res329z4, res329z5**	**25**	**17**	**15**
res802	2	res802z1, res802z4	17	12	16
res577	2	res577z1, res577z3	18	18	17
res119	2	res119z1, res119z3	23	20	18
res622	2	res622z1, res622z3	12	19	19
res402	3	res402z3, res402z4, res402z5	8	6	20
res493	3	res493z1, res493z2, res493z3	11	26	21
res114	2	res114z3, res114z5	24	25	22
res239	2	res239z3, res239z4	22	21	23
res799	2	res799z3, res799z5	9	22	24
res365	2	res365z4, res365z5	20	24	25
res331	4	res331z1, res331z2, res331z3, res331z5	3	27	26
res450	2	res450z1, res450z3	27	23	27
res665	3	res665z2, res665z3, res665z4	7	28	28
res606	3	res606z2, res606z4, res606z5	10	29	29

Table 2: Residue selection from the training set based on the best 200 variables selected by PLS regression. Only residue positions containing at least 2 z-scale descriptors and less than 20% of gaps were kept. We calculated three different rankings. Ranking sum_abs_coeff was obtained by summing the absolute coefficients of each residue position. Rank mean_diff was calculated according to the mean percentage difference for each residue at each position between protein kinases that bind the four Type-II inhibitors and those that do not bind the inhibitors. Rank max_diff was determined using the maximum difference for each residue. In bold are the most important identified positions.

We then analyzed the residues at these positions for all the protein kinases of the training set and we mapped them on the four selected crystal structures. Among the positions identified as most important, we noticed that two residues are represented by four out of five regression coefficients or z-scales (res325 and res331) (Table 2). These residues are located on the hinge region which connects the two lobes of protein kinases. Interestingly, the residue at position 325 corresponds to the gatekeeper. This important residue which often involves key interactions with the inhibitors has been recently described as a major discriminating residue for Type-I and II kinase inhibitors. (*46*,*47*) More precisely, the size of the gatekeeper residue has been identified as the main discriminant factor. (*48*) Moreover it has been already pointed out the gatekeeper residue could influence the ability of protein kinases to adopt the inactive conformation. (*17*,*49*) The order of the four coefficients representing the gatekeeper is also particularly interesting with res325z1, res325z4, res325z2 and res325z3 ranked respectively 1st, 2nd, 5th and 25th amongst a total of 1995 descriptors. Hence, with the corresponding interpretation of each z-scale, we can argue that the hydrophobicity, the electronic properties, the polarizability and the volume are the most important physicochemical descriptors for the gatekeeper residue followed by the polarity. In any case, with three coefficients of the gatekeeper in the top 5, our approach confirms the significance of the gatekeeper residue over the other residues to explain the activity of these four compounds. Nevertheless, it would be pointless to assert one residue is entirely responsible for inhibitor binding. It is likely that the gatekeeper residue can prevent the binding of some Type-II inhibitors but other residues present in the ATP site or in the kinase domain are also essential for the binding of such inhibitors. For instance, the residue position 331 is represented by four descriptors ranked 15th (res331z5), 23rd (res331z1), 57th (res331z3) and 61st (res331z2). Therefore, the importance of this residue, which is part of the hinge segment, is likely due to its electronic properties. The residue at this position has previously been described and contributes to the flexibility between the two C and N-lobes.(*50*,*51*)

In addition, we investigated if some identified positions have been already described in the literature. Taking ABL1 as an example, we found that the positions 187 (Leu273), 295 (Cys305), 325 (Thr315), 409

(Met351) and 414 (Glu355) have been described in diseased-related mutations. L273M, (*52*) T315I, (*53*) E355G and E355A are reported as being associated with tumor resistance. (*54*,*55*), C305R was identified as a mutation that may be associated to Type-II inhibitor resistance in patients with chronic myeloid lymphoma (*56*) and M351T induces a diminution of the transformation potency of the ABL1 kinase domain. (*57*) Therefore our PLS model has identified several positions that prevent the binding of a Type-II inhibitor in ABL1.

Using our approach, we also intend to identify residues involved in biological mechanism such as the regulation of kinase activity. In this aim, we have selected important residues located in the ATP site and analyzed their relevance in the PLS model. Therefore we selected residues Asp, Phe and Gly, as well as His, Arg and Asp from the DFG and HRD regions respectively. The highly conserved DFG motif has been often highlighted in the literature because of the involvement of these three residues in kinase plasticity, especially the orientation of Asp and Phe is determinant for the activation of the protein kinases. (*58*) Compared to the active conformation of ABL1, its inactive conformation involves an approximately 180° rotation in the backbone torsion angle of Phe of the DFG motif. This movement leads the phenyl ring of Phe in the binding site of the ATP ligand. (*59*) It has been shown that Type-II inhibitors can bind protein kinases when they adopt a DFG-out conformation. (*60*) We also investigated the ranking of the three residues of the highly conserved HRD motif which is engaged in the kinase catalytic mechanism. (*61*) In particular, in protein kinases that are regulated by phosphorylation, the Arg interacts with the phosphate group attached to serine, threonine or tyrosine of the activation loop. (*62*,*63*) According to our sequence alignment, the three amino acids of the HRD motif are located at positions 429, 442 and 443, and the three residues of the DFG motif are occupying positions 500, 501 and 502 respectively. Because these residues are well conserved amongst the kinase family and we used molecular graph independent descriptors, we did not expect to find these descriptors in the top ranking coefficients (Table 3). Indeed, these residues are not present in the top 200 first variables according to their regression coefficient. Interestingly, both Asp and Gly of the DFG motif have been automatically removed during the data preparation step, when we have excluded the variables with low variance. Indeed, Asp is entirely conserved in the sequence alignment (Table 3).

Residue position	% Identity
429	89
442	84
443	100
500	99
501	87
502	96

Table 3: Percentage of identity of 429, 442, 443 and 500, 501, 502 residue positions corresponding to the HRD and the DFG motifs respectively. The multi-alignment has been performed with MOE using 444 typical kinase domain sequences. Blosum100 matrix was used to score the alignment.

Prediction of protein kinase ability to bind Type-II inhibitors

Using the training set reduced to the 200 variables stressed by the rdCV, we performed a 100-fold internal validation. For each fold, by selecting randomly 75% of the proteins from the training set, we created a PLS model and then predicted the percentages of inhibition for the 25% remaining proteins. We estimated the robustness of each prediction by calculating the coefficient of determination R2. We obtained a R2 of 0.90 and a Q2 of 0.80 calculated over all the iterations which demonstrate that the statistical model is sufficiently robust to accurately predict the percentages of inhibition of the four compounds on the proteins.

We then used this PLS model to predict the ability of proteins kinases that were excluded from the training set, to bind the four Type-II inhibitors. We have also included some well-known mutant sequences to study if our approach is accurate enough to take into account very small modifications in the protein sequences. Among these 200 proteins, 144 of them were present in the external test set. Therefore we used these proteins to validate the ability of our model to correctly predict the activity of the four Type-II inhibitors on these proteins. For each protein which was predicted to bind Type-II inhibitors, we checked the presence of an inactive conformation in the RCSB (*64*) or we analyzed the biological activities collected from another external dataset. (*13*) Several protein kinases having a high predicted score corresponding to high likelihood to bind Type-II inhibitor, were found in a DFG-out conformation complexed with a Type-II inhibitor in the RCSB database such as DDR1 (PDB entry 4BKJ (*65*)), ABL1 T315I (PDB entry 3QRJ (*66*)), or KIT (PDB entry 4HVS (*67*)). On the opposite, some protein kinases with a low predicted score have not yet been co-crystallized with any Type-II inhibitors to our knowledge. Indeed, we did not identify any DFG-out conformation in the crystal

structures for TTBK1, STK32A, MAPK6, MAPK7, PRKCB, GSG2, CDKL2, CDKL3, TTN, STK17B, ACVR2B, PLK4, AURKB and EIF2AK3 (see Supplementary Table 7 for the entire list).

Not all protein kinases present in our set have a crystal structure publicly available. Therefore, to enlarge the scope of validation, we also compared the predicted score with biological activities from another published dataset. (*13*) In the external test set, 72 inhibitors were tested against 442 protein kinases and an affinity (Kd) was measured for each pair. Ten Type-II inhibitors were found in this external test set; among them the four inhibitors of the training set. For each protein kinase inhibited by one or several Type-II inhibitors, we kept only the maximum affinity or inhibition. Figure 3 represents these biological values plotted against the predicted values obtained from our statistical model which was built using the training set. Since we are comparing different types of data which are not well correlated (percentage of inhibition versus Kd, Figure 3A) we do not intend to look for any correlation, but we aim at getting some trends between non-inhibited versus inhibited protein kinases by Type-II inhibitors. We clearly see on Figures 3B, that while the statistical model is able to identify the most inhibited protein kinases, there is no false positive, *i.e.* non-inhibited protein kinases (pKd < 6) predicted to be inhibited (% inhibition > 50) by the statistical model. The model seems to be robust enough to predict protein kinases inhibited by Type-II kinase inhibitors with few false negatives. Our model also shows a good predictive power for inactive compounds since all protein kinases with a pKd of 5 have measured and predicted activities less than 50% inhibition, (Figure 3B). However, for some protein kinases there is no correlation between the values returned by the two experimental techniques used for the training set (% inhibition) and for the external test (Kd's). There are a few protein kinases with a high pKd for Type-II kinase inhibitors and low percentage of inhibition such as CDK8, TIE1 and CDC2L6. We have also predicted few kinases with similar uncorrelated behavior. Some "false negatives" may be explained by the structural differences between the Type-II inhibitors used in the training set and those used in the Davis *et al.* external test dataset such as foretinib, or quizartinib. (*13*) These differences may also come from the two different screening techniques used to identify the compounds. (*25*) We noticed an outlier that corresponds to CDK8 (Figure 3B pKd=9 and % control <0). We believe that the peculiarity of CDK8, which contains a DMG instead of a DFG motif, could be due to the limitation of the application domain. Indeed activation of CDK8 seems to be different from other CDKs, with a glutamic acid in Cyclin C that could mimic the phosphoresidue that usually serves as

an anchor to adjust the orientation of three important Arg. (*68*) Importantly, our model did not overestimate the biological activities of pKd since the model does not provide high score values for low pKd.

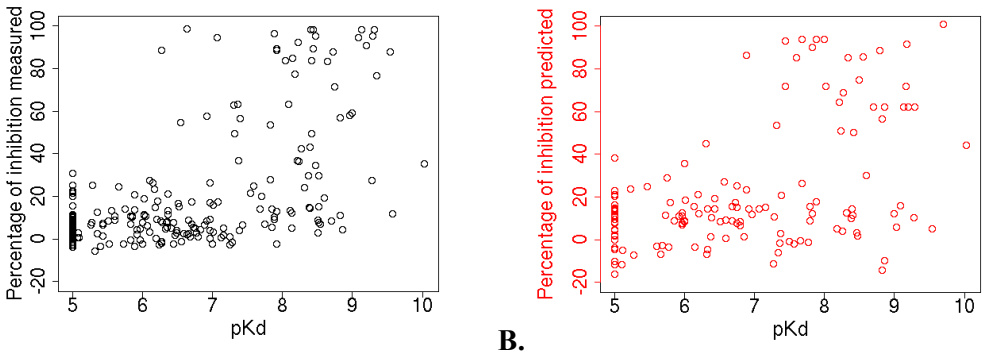


Figure 3 : (A) Scatter plot of the highest activity among the four Type-II inhibitors for each kinase measured by Davis et al. (13) against the activity measured by Anastassiadis et al. (12) (B) Scatter plot of the highest activity among the four Type-II inhibitors for each kinase measured by Davis et al. (13) against the activity predicted from the model with 200 variables.

Analysis of most discriminant residues

To analyze the importance of the residue positions in protein kinases able to bind Type-II inhibitors, we split the protein sequences of our training dataset in two subsets. In the first one, we included the proteins that bind Type-II inhibitors and in the second one the proteins that are never inhibited by Type-II inhibitors. To categorize the activity of a compound, we used a threshold value equal or greater to 50% for inhibited kinases and equal or lower to 30% for non-inhibited kinases.

According to the coefficients obtained from the PLS model, we investigated the residue distribution at each important position using the sequence alignment of 263 protein kinases. 29 positions were selected (Table 2, in bold). We ranked the positions using three different criteria. The first ranking (sum_abs_coeff) was obtained by summing the absolute coefficients of each residue position. The second ranking (mean_diff) was calculated according to the mean percentage difference for each residue at each position between protein kinases that bind the four Type-II inhibitors and those that do not bind the inhibitors. Finally, the last ranking (rank max_diff) was determined using the maximum difference for each residue. Taking into account the three ranking approaches (Table 2), we focused on the most important residues, *i.e.* positions 187, 192, 296, 325, 409, 414 and 426 (Figure 4, residues represented in blue ball and stick). We also investigated position 329 due to its critical location on the hinge region and its close proximity to the inhibitor. However, despite the fact that the position 450 is located on the activity loop and may interact with an inhibitor, we did not select it for further

investigation because of its bad ranks. Although the PLS method returns four descriptors for res331 we did not noticed important differences between the two groups of proteins. Indeed, a Gly residue is particularly well conserved in the eukaryotes protein kinases (data not shown). (*51*)

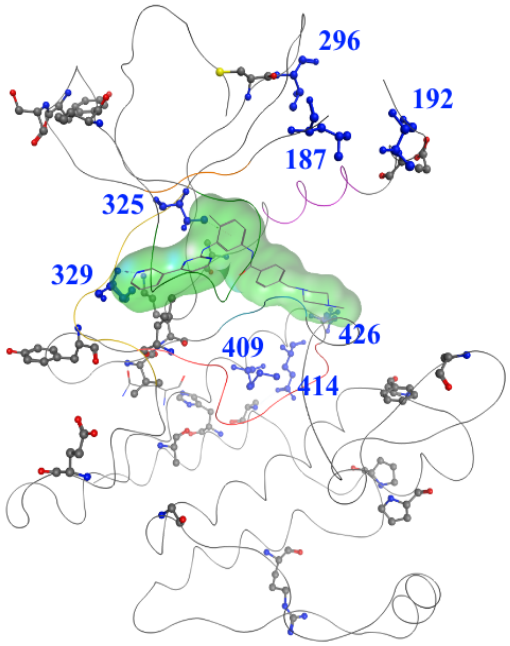


Figure 4 : Tridimensional representation of ABL1 with 29 selected residues identified by PLS regression and represented in atom-colored ball and stick. Residues at position 187, 192, 296, 325, 329, 409 414 and 426 are in dark blue ball and stick. Imatinib is represented in stick and its molecular surface is displayed in green.

Protein kinases inhibited by the four Type-II inhibitors bear almost exclusively a Leu at position 187 (Figure 5A). Met, Val and Ile taken individually do not exceed 10%. We found that in ABL1, a mutation may occur at this position (L273M) and induce a resistance to imatinib. (*52*) This may explain the low percentage of Met. However, we also found that Met is equally represented in protein kinases not inhibited by the Type-II inhibitors. This could mean that such amino acid at this position may also hinder the access of other class of inhibitors. In addition we found more than 30% of Leu and Ile and around 20% of Val.

At position 192, Thr (37%) and Met (20%) are the most abundant residues in the proteins binding Type-II inhibitors (Figure 5B). Only 8 amino acids and gaps represent the remaining 43%. For the other proteins, there is no clear evidence that a particular residue is required. Leu and Ile both represent slightly more than 10% but all other amino acids can also be present at low percentages.

For the protein kinases inhibited by Type-II inhibitors, we observed more than 50% of Thr, about 15% of Ser and less than 10% of Arg at position 296 (Figure 5C). On the opposite, we detected for the non-inhibited kinases, less than 30% of Glu and around 10% of Thr, Leu and Gln. All the other amino acids, but Cys, are also present but with very low occurrence.

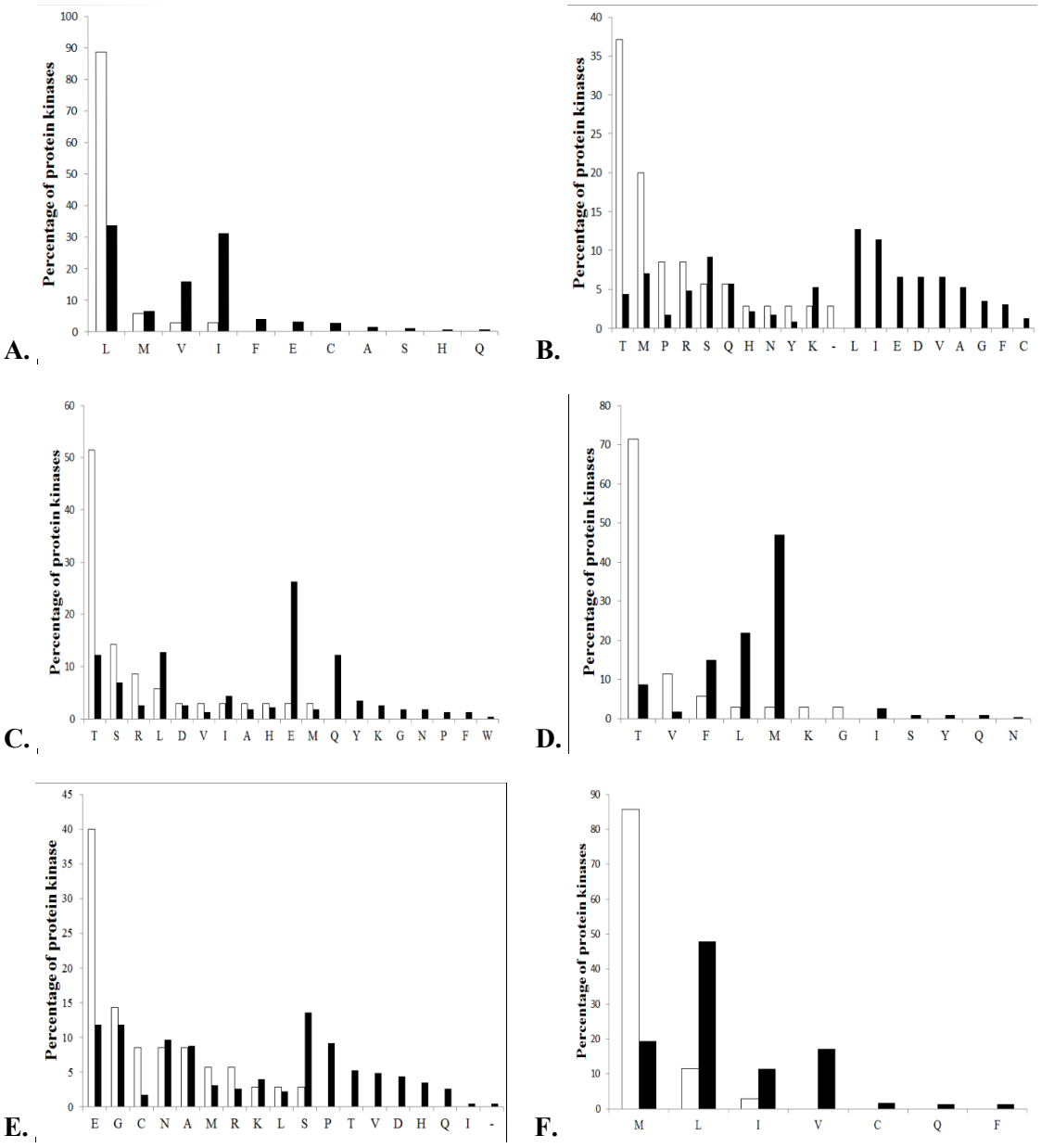
At position 325 (Figure 5D), the protein kinases from the first dataset count more than 70% of Thr and around 10% of Val. For the protein kinases included in the second set, we found about 50% of Met, more than 20% of Leu and around 15% of Phe. Thr and Val are relatively smaller than Met, Leu or Phe and indeed, res325z2 was ranked 5th among all the variables, indicating that the volume of the residue seems significant at this position. Res325z1 was 1st of this rank which means the hydrophobicity of the residues located at this position is highly important for the binding of Type-II inhibitors. As mentioned previously, this amino acid at position 325 is the important gatekeeper residue which is often forming hydrogen bond interaction with Type-II kinase inhibitors. For example, Thr315 of ABL1 forms a hydrogen bond with secondary amine NH of imatinib and nilotinib.

The highest occurrence difference is found at position 329 with 40% of Glu for the inhibited protein kinases against 12% for the non-inhibited (Figure 5E). Then, we found similar proportions for Gly (15-12%), Asn (~10%) and Ala (~10%). Cys was found almost exclusively in the inhibited protein kinases but at low percentage (9%). 15% of Ser, which is the most represented residue, was found for the non-inhibited protein kinases.

At position 409 (Figure 5F), we retrieved almost 85% of Met, 11% of Leu and around 4% for Ile in the first set. However for the protein kinases not inhibited by Type II inhibitors, we observed 50% of Leu at this position, 20% for Met and Val and 10% for Ile. Met, Leu and Ile are included in both categories but their proportions differ. Due to the high majority of Met at this position, it seems that this residue is essential for protein kinases trapping Type-II inhibitors.

At position 414 (Figure 5G), there are around 30% of Ala, 25% of Glu, 20% of His and Ser, and very small proportions of Lys and Met. The second set contains 80% of His, 10% of Glu. The other remaining amino acids are in minority.

Finally, we investigated the residues at position 426 (Figure 5H). Protein kinases inhibited by the four Type-II inhibitors have preferentially an Asn (58%), but also Lys (14%), Gly (14%) and to a lesser extend Ser, Ala and Asp. Regarding the non-inhibited protein kinases, we found at this position almost every amino acid, but Gly is the most represented with 35%, followed by Lys (15%), Gln and Arg (both 12%).



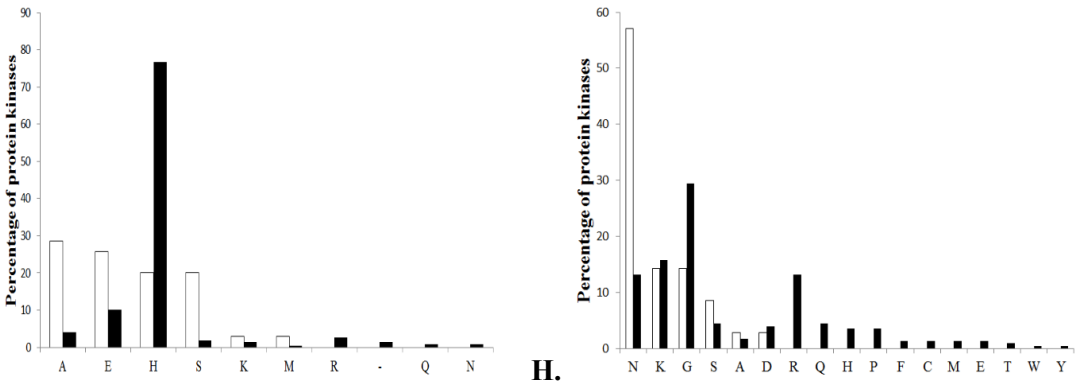


Figure 5 : Percentage of residues depending on whether the protein kinases are inhibited (white bars) or not (black bars) in presence of a Type-II inhibitor. Focus on the largest variations in residue distributions: (A) position 187, (B) position 192, (C) position 296, (D) position 325, (E) position 329, (F) position 409, (G) positon 414 and (H) position 426.

It is important to note that the binding of Type-II inhibitors to protein kinases is not due to the involvement of only one single amino acid but rather to a network of residues such as the recently discovered C- and R-spines. (*69*)

Mapping of the selected residues on X-ray structures

To facilitate the visualization of all the significant descriptors, we have highlighted their corresponding residues on protein kinase crystal structures. To analyze the interaction of the above described residues with some kinase inhibitors, we chose crystal structures containing two different Type-II inhibitors included in the training set such as imatinib and sorafenib. Additionally, we have selected two crystal structures of protein kinases which have never been described as targets inhibited by this specific class of inhibitors: SYK and DYRK1A. The visualization was performed in MOE.

Among the selected residue positions, the majority of them are located on both lobes, with no direct interaction with the ligands in the active site (Figure 4). Position 187 is located two residues before the catalytic Lys, while position 192 is in N-terminal of the Alpha-C helix. The position 296 is located on the N-terminal lobe between the α-C helix and the hinge region, and might be involved in the binding of the protein kinase SH3 domain. Hence it is not straightforward to interpret their role in the binding of Type-II inhibitors but the combined role of these two residues may suggest a role in DFG-out conformational change.

On the hinge region, near the ATP pocket, where Type-II inhibitors usually form a hydrogen bond, we have identified the positions 325 and 329. Among the eight residue positions we studied, only these two can

directly interact with a Type-II inhibitor. As mentioned previously, residue 325 is the gatekeeper residue. This residue has been described as being involved in compound binding and especially it can discriminate the binding of Type-I versus Type-II inhibitors. (*46*,*47*) Its size was shown to be a major discriminant parameter. (*48*) A relative small gatekeeper residue will open a hydrophobic pocket adjacent to the ATP active site, also known as the back pocket, first identified in CDKs. (*48*,*70*) Conversely, a bulky gatekeeper will mostly close the access to the hydrophobic pocket, and recent workaround has been successfully applied such as the incorporation of the alkyne moiety in ponatinib. The visualization of co-crystalized structures in complex with Type-II inhibitors confirms this statement. As seen on Figures 6A and 6B, the binding of imatinib or sorafenib in their respective protein kinase, ABL1 and BRAF, is possible because of the small size of the gatekeeper residue, since in both cases, the threonine is small enough to give access to the hydrophobic pocket. Figure 6C is an illustration of DYRK1A, a protein kinase that has never been identified so far to bind a Type-II inhibitor, nor crystallized in DFG-out conformation. (*71*) With a phenylalanine as a gatekeeper residue, the hydrophobic pocket is inaccessible by the four Type-II inhibitors used in this study. Figure 6D represents an atypical conformation of imatinib binding as Type-I to SYK protein kinase in an active conformation. This protein carries a methionine residue as a gatekeeper. This residue is larger than the threonine and prevents access to the hydrophobic pocket. We stress again, that the importance of the gatekeeper volume has been highlighted using 2D residue descriptors in a PLS model. Despite the fact that this observation has been proposed in several studies by visual inspection of crystal structures, we believe this is the first time that the importance of the size of the gatekeeper residue in protein kinase conformational change has been identified by proteometrics statistical approach.

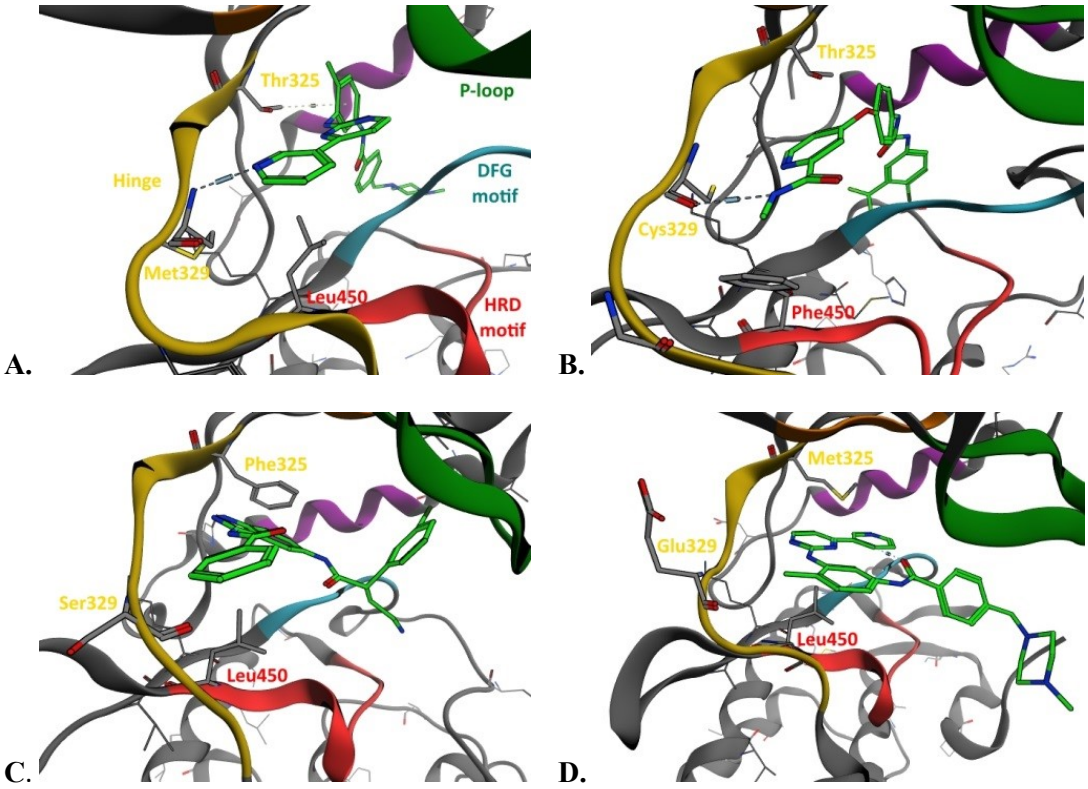


Figure 6 : Analysis of X-ray crystal structures. DFG-out conformations (A) Imatinib bound to ABL1 (PDB ID: 2HYY), (B) Sorafenib bound to BRAF (PDB ID: 1UWH). DFG-out conformations (C) indazole inhibitor bound to DYRK1A (PDB ID: 2VX3), (D) Imatinib bound to SYK (PDB ID: 1XBB). Only protein residues identified by PLS are displayed (stick representation with grey carbons). On each picture, the ligand is represented with green atoms. The light grey dash line indicates a hydrogen bond.

As mentioned above, ponatinib overcomes the problem of bulkiness of the gatekeeper residue to successfully inhibit ABL1 T315I. Using our proteometrics approach we repeated our protocol using inhibition data of ponatinib only. As this molecule was absent from the dataset of Anastassiadis *et al.,* (*12*) we used the dataset published by O'Hare *et al*. (*53*) After preparing the data, we applied the PLS method on 89 protein kinases, where each IC_{50} was converted into pIC_{50} before modeling. The model returns 200 regression coefficients that we analyzed as previously. The results show the gatekeeper residue at position 325 remains an important residue to explain the inhibition of ponatinib on these 89 protein kinases. However, while res325 was previously first or second with the three different rankings, here res325 is ranked 6^{th}, 4^{th} and 2^{nd} in our three rankings (Supplementary Figure 8). This result may be explained by the low affinity of ponatinib on several protein kinases containing a Met gatekeeper (AKT2, PTK2, SYK). We also noticed that ponatinib is less selective than the four other Type-II inhibitors as previously reported. (*72*)

The residues at positions 409, 414 and 426 are located close together in their 2D sequence as well as in their 3D structure. While positions 409 and 414 are located on the α-E helix of the C-terminal lobe, position 426

is on the activation loop. However, there is no clear evidence of their implication in the Type-II inhibitor binding process.

Additionally, we also investigated the ranking of the residues involved in the hydrophobic spines. Two important spines have been identified by Kornev *et al.* (*69*), using a bioinformatics tool able to identify conserved motifs that are related neither to protein sequence nor to protein geometry. They are formed by discontinuous residues which are located in both lobes of the protein kinases. Interestingly, it is suggested that the assembling and disassembling of these two spines may depend on the presence of ATP, the phosphorylation state of the activation loop or the α-C helix movement. Therefore they participate in protein kinase plasticity which is strongly linked with its activity. The R spine is composed of only four residues and its structure depends on the activation loop conformation. In the active conformation, the R spine connects the two lobes together and it is not present in an inactive conformation. Regarding the C spine, it is formed when ATP is bound to the protein kinase and ATP acts as a connector between the two adjacent residues in the N- and C-terminal lobes. In our sequence alignment, the C spine is represented by the residues at positions 110, 183, 359, 449, 450, 451, 641 and 645, and the R spine by the residues at positions 235, 291, 429 and 501 (Supplementary Figure 9). However, using our statistical approach we did not retrieve any residue of the spines with the variable selection approach. Such result is not surprising since the residues of the R spine are well conserved amongst the protein kinases. Moreover the formation of this spine is related to the orientation “in” and “out” of the conserved Phe of DFG motif and therefore is related to the active state of the kinase. As our descriptors only represent the 2D physicochemical properties of the residues, they do not take into account any spatial coordinates. Regarding the C spine, it is formed with the adenine moiety of ATP when bound to the protein kinase. Therefore to maintain this spine it is necessary that the kinase inhibitor contains a bioisostere of adenine forming hydrogen bond interactions with the hinge region, as it is often observed for Type-I kinase inhibitors and to a lesser extend for Type-II inhibitors.

In summary, our approach confirms the importance of the gatekeeper in the binding of the four Type-II inhibitors. Some others residues of the active site have been identified as discriminant for the binding of Type-II inhibitors, but their function is not yet clearly understood.

Gatekeeper analysis

We then studied if our observations based on protein kinase sequences could be translated onto the protein kinase structures. The co-crystal structures have been collected from the PDB structures using MOE Kinase Database updated in January 2014. We split this collection according to the same classification that we applied previously for the sequences: inactive class when the proteins are bound to a Type-II inhibitor or active class otherwise. For each co-crystal structure, we identified the gatekeeper residue using kinase sequence alignment.

Therefore for each class we enumerated the number of proteins with the same gatekeeper (Figure 7). We obtained similar results as compared to what we found previously at position 325 with the statistical approach. In proteins inhibited by a Type-II inhibitor, Thr is the amino acid with the highest occurrence (85%), followed by Val (9%). Regarding other protein kinases inhibited by Type-II kinase inhibitors and bearing a gatekeeper residue Phe, Leu, Lys or Gly, no crystal structure was available. This lack of structural information prevented us to draw any conclusion on the mode of binding of those inhibitors. Additionally, we noted a minority of crystal structures with Ile as gatekeeper. One example is the human ABL1 T315I gatekeeper mutant (*73*) in complex with Type-II inhibitors such as rebastinib (*66*) or ponatinib. (*53*) For the protein kinases that are not inhibited by Type-II inhibitors, the majority of residues at the gatekeeper position is Met (41%), followed by Leu (22%), Phe (17%), Thr (9%) and Ile (3%) (Figure 7). This proportion is similar to the one found by our statistical approach.

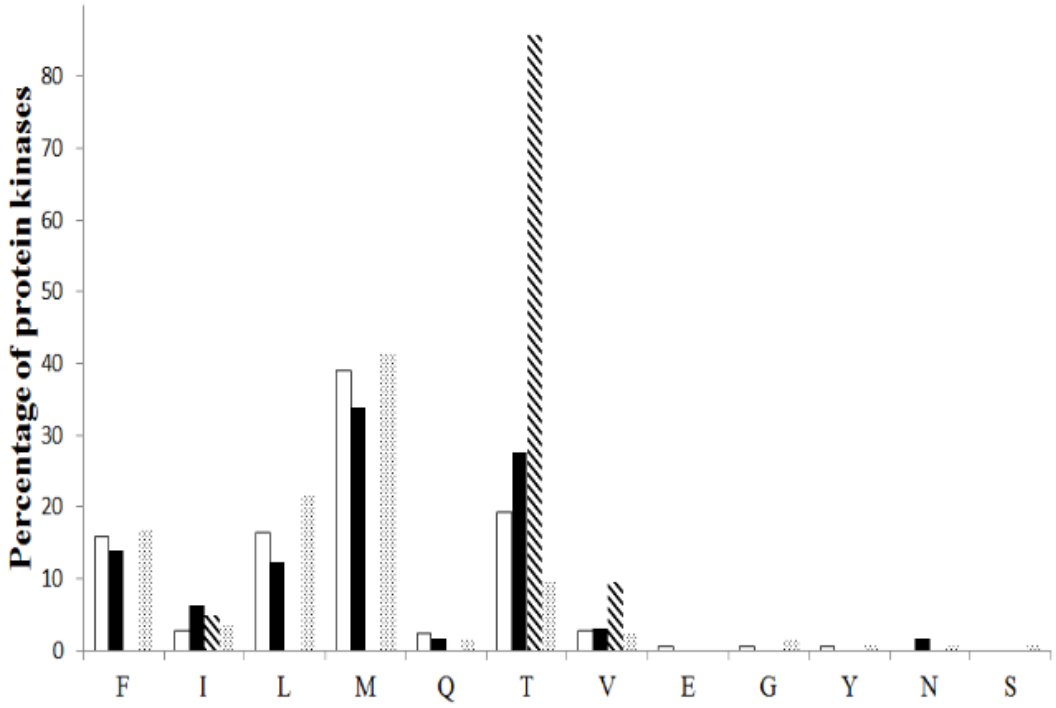


Figure 7 : Analysis of the gatekeeper residue in protein kinase crystal structures, depending on two classifications. The protein kinase structures are in DFG-in (white bars) or in DFG-out (black bars) conformation, or the protein kinases are inhibited (crosshatched bars) or not (bars with black dots).

Type-II inhibitors bind a protein kinase if its DFG motif is in an "out" conformation. Therefore, we decided to apply the same approach using a DFG based classification instead of structural Type-II classification. Using the same protein kinase database, we split protein kinase in two classes using the DFG parameter "in" and "out" calculated by MOE. Since the activation loop containing the DFG motif is highly flexible, this DFG parameter is missing for a large part of the available structures (422 out of 2657). Surprisingly, the results bring new information when compared to our first obtained results (Figure 5C). This classification showed that the amino acid nature of the gatekeeper residue is not correlated to the DFG conformation (Figure 7). For DFG-in and DFG-out, Met and Thr are the most represented residues respectively while a distinction could be seen by our approach. This might be due to some questionable DFG-out classification proposed by MOE for some protein kinases with a Met residue as gatekeeper (Supplementary Figure 10) or be biased by the kinases that have been crystallized so far.

We compared our observation with those made by Kufareva *et al*. (*11*) when they analyzed the protein kinase crystal structures in the PDB database and in particular the nature of the gatekeeper. The authors of the

DOLPHIN protocol indexed 10 different amino acids located at the gatekeeper position (Table 4). The most represented were Phe, Met, Leu and Thr. Altogether, 95 protein kinases had at least one DFG-in structure, and 23 have at least a DFG-out structure available in the PDB. With our protein kinase database updated on 2014 (Table 4), we counted 186 protein kinases that contain at least one DFG-in conformation, and 66 with at least one DFG-out conformation.. Comparing our results with those of Kufereva *et al.*, we noticed that almost every category has increased. Nevertheless, the only structure with a Ser gatekeeper in the DOLPHIN paper could not be classified by MOE. In addition, the authors did not identify any DFG-out structure with a Thr gatekeeper, but 6 years later four structures have been identified. The only structure with an Asn gatekeeper was classified as DFG-in in Kufareva *et al.* study, but it is tagged as DFG-out conformation in MOE. Finally, new protein kinases with Gly and Asp as gatekeeper residue have been crystallized since 2008. The different DFG-in/-out classification between Kufareva's study and MOE may be explained by the two dissimilar approaches used to classify the kinase conformations.

GK	Available DFG-in structures		Available DFG-out structures		DFG-in and DFG-out available structures	
	2008	**2014**	**2008**	**2014**	**2008**	**2014**
Y	1	1	NA	NA	NA	NA
F	12	25	3	6	1	4
E	NA	1	NA	NA	NA	NA
Q	2	3	NA	NA	NA	1
M	27	54	3	4	4	18
L	19	23	1	1	1	7
I	4	2	1	1	NA	3
N	1	NA	NA	1	NA	NA
V	2	4	1	1	NA	1
T	13	24	NA	4	7	14
S	1	NA	NA	NA	NA	NA
G	NA	1	NA	NA	NA	NA
	82	**138**	**9**	**18**	**13**	**48**

Table 4: Analysis of mammalian protein kinase crystal structures included in the Protein Data Bank and analyzed in June 2008 (11), and in January 2014. NA is for Non Available.

CONCLUSION

Identification of important amino acids involved in potency or selectivity of kinase inhibitors is of major interest in kinase drug discovery. The work described herein proposes a new method to identify discriminant residues in protein kinases responsible for the binding of four selected Type-II inhibitors. This original proteomics method relies on the statistical method PLS regression combined with residue descriptors

and allowed the identification of 29 discriminant residues involved in Type-II kinase inhibitor recognition. We applied this model on 200 protein kinases absent from the training set and proposed a predictive score to evaluate their ability to be inhibited by Type-II kinase inhibitors or to adopt a DFG-out conformation. The results were compared with an external test set containing experimental biological activities and showed excellent estimations suggesting that some protein kinases seem to be more amenable to be inhibited by Type-II kinase inhibitors. We also investigated the 29 important residue positions highlighted by the model with different approaches. We analyzed the nature of the amino acids localited at these positions according to protein kinase propensity to bind Type-II inhibitors. We confirmed the importance of the gatekeeper residue in agreement with previously published studies. The gatekeeper residue needs to be rather small in order to allow these Type-II inhibitors to access the hydrophobic pocket. Regarding the other positions, differences have been underlined but the nature of their influence on kinase conformation is still difficult to interpret. Further research could focus on these specific residues to understand their significance in the ability of protein kinases to trap Type-II inhibitors. As expected, our model did not retrieve highly conserved residues such as Asp-Phe-Gly of the DFG motif, neither the residues that shape the Spines. The results of this study could support structural biology efforts in the field of protein kinases, provide a better understanding of the protein kinase conformation changes and finally could help in designing novel Type-II kinase inhibitors by focusing on specific identified residues. Finally, these results demonstrate that proteometrics can also be very useful to predict activity of compounds on untested targets. Such characteristics may be very useful for future large scale profiling on the kinome by allowing a selection of kinases of interest.

ACKNOWLEDGMENTS

N Bosc, S Aci-Sèche and P Bonnet are grateful to the Région Centre Val de Loire and Janssen-Cilag for financial supports. The authors would like to acknowledge Jean-Marc Neefs for his assistance in retrieving public data and Rohit Arora for providing protein sequence alignment.

REFERENCES

1. Dhanasekaran, N., Premkumar Reddy, E. (1998) Signaling by dual specificity kinases. *Oncogene 17*, 1447-1455.

2. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S. (2002) The protein kinase complement of the human genome. *Science 298*, 1912-1934.

3. Cohen, P. (2002) Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov. 1*, 309-315.

4. Liu, M., Bender, S. A., Cuny, G. D., Sherman, W., Glicksman, M., Ray, S. S. (2013) Type II Kinase Inhibitors Show an Unexpected Inhibition Mode against Parkinson's Disease-Linked LRRK2 Mutant G2019S. *Biochemistry 52*, 1725-1736.

5. Roskoski, R. USFDA approved protein kinase inhibitors. http://www.brimr.org/PKI/PKIs.htm (accessed February 2015).

6. Gavrin, L. K., Saiah, E. (2013) Approaches to discover non-ATP site kinase inhibitors. *MedChemComm 4*, 41-51.

7. Solca, F., Dahl, G., Zoephel, A., Bader, G., Sanderson, M., Klein, C., Kraemer, O., Himmelsbach, F., Haaksma, E., Adolf, G. R. (2012) Target Binding Properties and Cellular Activity of Afatinib (BIBW 2992), an Irreversible ErbB Family Blocker. *J. Pharmacol. Exp. Ther. 343*, 342-350.

8. Pan, Z., Scheerens, H., Li, S.-J., Schultz, B., Sprengeler, P., Burrill, C. L., Mendonca, R., Sweeney, M., Scott, K. C. K., Grothaus, P., Jeffery, D., Spoerke, J., Honigberg, L., Young, P., Dalrymple, S., Palmer, J. (2007-01-15) Discovery of Selective Irreversible Inhibitors for Bruton's Tyrosine Kinase. *ChemMedChem 2*, 58-61.

9. Simard, J. R., Klüter, S., Grütter, C., Getlik, M., Rabiller, M., Rode, H. B., Rauh, D. (2009) A new screening assay for allosteric inhibitors of cSrc. *Nat. Chem. Biol. 5*, 394-396.

10. Zhang, J., Yang, P. L., Gray, N. S. (2009) Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer 9*, 28-39.

11. Kufareva, I., Abagyan, R. (2008) Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. *J. Med. Chem. 51*, 7921-7932.

12. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H., Peterson, J. R. (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol. 29*, 1039-1045.

13. Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., Zarrinkar, P. P. (2011) Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol. 29*, 1046-1051.

14. Metz, J. T., Johnson, E. F., Soni, N. B., Merta, P. J., Kifle, L., Hajduk, P. J. (2011) Navigating the kinome. *Nat.*

Chem. Biol. 7, 200-202.

15. Fedorov, O., Marsden, B., Rellos, P., Muller, S., Bullock, A. N., Schwaller, J., Sundstrom, M., Knapp, S. (2007) A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. *PNAS 104*, 20523-20528.

16. Bain, J., Plater, L., Elliott, M., Shpiro, N., Hastie, C. J., McLauchlan, H., Klevernic, I., Arthur, J. S. C., Alessi, D. R., Cohen, P. (2007) The selectivity of protein kinase inhibitors: a further update. *Biochem. J. 408*, 297-315.

17. Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., Gray, N. S. (2014) Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol. 9*, 1230-1241.

18. Caballero, J., Fernández, L., Garriga, M., Abreu, J. I., Collina, S., Fernández, M. (2007) Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. *J. Mol. Graphics Modell. 26*, 166-178.

19. Lavine, B. K., Ed. *Chemometrics and Chemoinformatics;* American Chemical Society, 2005; Vol. 894.

20. van Westen, G. J., Swier, R. F., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W., Bender, A. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. *J. Cheminf. 5*, 41.

21. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S. (1998) New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem. 41*, 2481-2491.

22. Mei, H., Liao, Z. H., Zhou, Y., Li, S. Z. (2005) A new set of amino acid descriptors and its application in peptide QSARs. *Biopolymers 80*, 775-786.

23. Liang, G., Li, Z. (2007) Factor Analysis Scale of Generalized Amino Acid Information as the Source of a New Set of Descriptors for Elucidating the Structure and Activity Relationships of Cationic Antimicrobial Peptides. *QSAR Comb. Sci. 26*, 754-763.

24. Zaliani, A., Gancia, E. (1999) MS-WHIM Scores for Amino Acids: A New 3D-Description for Peptide QSAR and QSPR Studies. *J. Chem. Inf. Model. 39*, 525-533.

25. Zhang, C., Habets, G., Bollag, G. (2011) Interrogating the kinome. *Nat. Biotechnol. 29*, 981-983.

26. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization;* Springer, 2007; pp 319-326.

27. The Uniprot Consortium. (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research 42*, D191-D198.

28. The IUPAC International Chemical Identifier (InChI). www.iupac.org/home/publications/e-resources/inchi.html (accessed February 2015).

29. Chen, M. J. KinBase. http://www.kinase.com/web/current/kinbase/ (accessed 2014).

30. Chemical Computing Group Inc. *Molecular Operating Environment (MOE), 2013.08;* 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, QC, Canada, 2014.

31. Varmuza, K., Filzmoser, P., Dehmer, M. (2013) Multivariate linear QSPR/QSAR models: Rigorous Evaluation of Variable Selection for PLS. *Comput. Struct. Biotechnol. J. 5*, 1-10.

32. Mucs, D., Bryce, R. A., Bonnet, P. (2011) Application of shape-based and pharmacophore-based in silico screens for identification of Type II protein kinase inhibitors. *J. Comput.-Aided Mol. Des. 25*, 569-581.

33. Liu, Y., Gray, N. S. (2006) Rational design of inhibitors that bind to inactive kinase conformations. *Nat. Chem. Biol. 2*, 358-364.

34. Bonnet, P., Mucs, D., Bryce, R. A. (2012) Targeting the inactive conformation of protein kinases: computational screening based on ligand conformation. *MedChemComm 3*, 434-440.

35. Mol, C. D., Dougan, D. R., Schneider, T. R., Skene, R. J., Kraus, M. L., Scheibe, D. N., Snell, G. P., Zou, H., Sang, B. C., Wilson, K. P. (2004) Structural basis for the autoinhibition and STI-571 inhibition of c-Kit tyrosine kinase. *J. Biol. Chem. 279*, 31655-31663.

36. Wold, S., Sjöström, M., Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. *Chemom. Intell. Lab. Syst. 58*, 109-130.

37. Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S. *Multi- and megavariate Data Analysis: principles and appliactions;* Umea, Sweden, 2001.

38. R Core Team. *R: A Language and Environment for Statistical Computing;* R Foundation for Statistical Computing: Vienna, Austria, 2014.

39. de Jong, S. (1993) SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst. 18*, 251-263.

40. Mevik, B.-H., Wehrens, R. (2007) The pls Package: Principal Component and Partial Least Squares Regression in R. *J. Stat. Soft. 18*, 1-24.

41. Ishibuchi, H., Nojima, Y. (2013) Repeated double cross-validation for choosing a single solution in evolutionary multi-objective fuzzy classifier design. *Knowledge-Based Systems 54*, 22-31.

42. Wan, P. T. C., Garnett, M. J., Roe, S. M., Lee, S., Niculescu-Duvaz, D., Good, V. M., Project, C. G., Jones, C. M., Marshall, C. J., Springer, C. J., Barford, D., Marais, R. (2004) Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF. *Cell 116*, 855-867.

43. Cowan-Jacob, S. W., Fendrich, G., Floersheimer, A., Furet, P., Liebetanz, J., Rummel, G., Rheinberger, P., Centeleghe, M., Fabbro, D., Manley, P. W. (2007) Structural biology contributions to the discovery of drugs to treat chronic myelogenous leukaemia. *Acta Cryst. 63*, 80-93.

44. Soundararajan, M., Roos, A., Savitsky, P., Filippakopoulos, P., Kettenbach, A., Olsen, J., Gerber, S., Eswaran, J., Knapp, S., Elkins, J. (2013) Structures of Down Syndrome Kinases, DYRKs, Reveal Mechanisms of Kinase Activation and Substrate Recognition. *Structure 21*, 986-996.

45. Atwell, S., Adams, J. M., Badger, J., Buchanan, M. D., Feil, I. K., Froning, K. J., Gao, X., Hendle, J., Keegan, K., Leon, B. C., Muller-Dieckmann, H. J., Nienaber, V. L., Noland, B. W., Post, K., Rajashankar, K. R., Ramos, A., Russell, M., Burley, S. K., Buchanan, S. G. (2004) A Novel Mode of Gleevec Binding Is Revealed by the Structure of Spleen Tyrosine Kinase. *J. Biol. Chem. 279*, 55827-55832.

46. Tanramluk, D., Schreyer, A., Pitt, W. R., Blundell, T. L. (2009) On the Origins of Enzyme Inhibitor Selectivity and Promiscuity: A Case Study of Protein Kinase Binding to Staurosporine. *Chem. Biol. Drug Des. 74*, 16-24.

47. Pargellis, C., Tong, L., Churchill, L., Cirillo, P. F., Gilmore, T., Graham, A. G., Grob, P. M., Hickey, E. R., Moss, N., Pav, S., Regan, J. (2002) Inhibition of p38 MAP kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol. 9*, 268-272.

48. Blencke, S., Zech, B., Engkvist, O., Greff, Z., Õrfi, L., Horvàth, Z., Kéri, G., Ullrich, A., Daub, H. (2004) Characterization of a Conserved Structural Determinant Controlling Protein Kinase Sensitivity to Selective Inhibitors. *Chem. Biol. 11*, 691-701.

49. Hari, S., Merritt, E., Maly, D. (2013) Sequence Determinants of a Specific Inactive Protein Kinase Conformation. *Chem. Biol. 20*, 806-815.

50. Martin, E., Mukherjee, P. (2012) Kinase-Kernel Models: Accurate In silico Screening of 4 Million Compounds Across the Entire Human Kinome. *J. Chem. Inf. Model. 52*, 156-170.

51. Kannan, N., Haste, N., Taylor, S. S., Neuwald, A. F. (2007) The hallmark of AGC kinase functional divergence is its C-terminal tail, a cis-acting regulatory module. *Proc. Natl. Acad. Sci. U. S. A. 104*, 1272-1277.

52. Lee, T.-S., Ma, W., Zhang, X., Giles, F., Cortes, J., Kantarjian, H., Albitar, M. (2008) BCR-ABL alternative splicing as a common mechanism for imatinib resistance: evidence from molecular dynamics simulations. *Mol. Cancer Ther. 7*, 3824-3841.

53. O'Hare, T., Shakespeare, W. C., Zhu, X., Eide, C. A., Rivera, V. M., Wang, F., Adrian, L. T., Zhou, T., Huang, W. S., Xu, Q., Metcalf, C. A., Tyner, J. W., Loriaux, M. M., Corbin, A. S., Wardwell, S., Ning, Y., Keats, J. A., Wang, Y., Sundaramoorthi, R., Thomas, M., Zhou, D., Snodgrass, J., Commodore, L., Sawyer, T. K., Dalgarno, D. C., Deininger, M. W., Druker, B. J., Clackson, T. (2009) AP24534, a pan-BCR-ABL inhibitor for chronic myeloid leukemia, potently inhibits the T315I mutant and overcomes mutation-based resistance. Cancer Cell. *Cancer Cell 16*, 401-412.

54. Bennour, A., Beaufils, N., Sennana, H., Meddeb, B., Saad, A., Gabert, J. (2010) E355G mutation appearing in a patient with e19a2 chronic myeloid leukaemia resistant to imatinib. *J. Clin. Pathol. 63*, 737-740.

55. Elias, M. H., Baba, A. A., Azlan, H., Rosline, H., Sim, G. A., Padmini, M., Fadilah, S. A., Ankathil, R. (2014) BCR-ABL kinase domain mutations, including 2 novel mutations in imatinib resistant Malaysian chronic myeloid leukemia patients-Frequency and clinical outcome. *Leuk. Res. 38*, 454-459.

56. Mascarenhas, C. C., Cunha, A. F., Miranda, E. C., Zulli, R., Silveira, R. A., Costa, F. F., Pagnano, K. B., De Souza, C. A. (2009) ew mutations detected by denaturing high performance liquid chromatography during screening of exon 6 bcr-abl mutations in patients with chronic myeloid leukemia treated with tyrosine kinase inhibitors. *Leuk. Lymphoma 50*, 1148-1154.

57. Griswold, I. J., MacPartlin, M., Bumm, T., Goss, V. L., O'Hare, T., Lee, K. A., Corbin, A. S., Stoffregen, E. P., Smith, C., Johnson, K., Moseson, E. M., Wood, L. J., Polakiewicz, R. D., Druker, B. J., Deininger, M. W. (2006) Kinase domain mutants of Bcr-Abl exhibit altered transformation potency, kinase activity, and substrate utilization, irrespective of sensitivity to imatinib. *Mol. Cell Biol. 26*, 6082-6093.

58. Atzori, A., Bruce, N. J., Burusco, K. K., Wroblowski, B., Bonnet, P., Bryce, R. A. (2014) Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics. *J. Chem. Inf. Model. 54*, 2764-2775.

59. Levinson, N. M., Kuchment, O., Shen, K., Young, M. A., Koldobskiy, M., Karplus, M., Cole, P. A., Kuriyan, J. (2006) A Src-Like Inactive Conformation in the Abl Tyrosine Kinase Domain. *PLoS Biol. 4*, e144.

60. Treiber, D., Shah, N. (2013) Ins and Outs of Kinase DFG Motifs. *Chem. Biol. 20*, 745-746.

61. Adams, J. A. (2001) Kinetic and Catalytic Mechanisms of Protein Kinases. *Chem. Rev. 101*, 2271-2290.

62. Nolen, B., Taylor, S., Ghosh, G. (2004) Regulation of Protein KinasesControlling Activity through Activation Segment Conformation. *Molecular Cell 15*, 661-675.

63. Johnson, L. N., Lewis, R. J. (2001) Structural Basis for Control by Phosphorylation. *Chem. Rev. 101*, 2209-2242.

64. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000) The Protein Data Bank. *Nucl. Acids Res. 28*, 235-242.

65. Canning, P., Tan, L., Chu, K., Lee, S. W., Gray, N. S., Bullock, A. N. (2014) Structural Mechanisms Determining Inhibition of the Collagen Receptor DDR1 by Selective and Multi-Targeted Type II Kinase Inhibitors. *J. Mol. Biol. 426*, 2457-2470.

66. Chan, W. W., Wise, S. C., Kaufman, M. D., Ahn, Y. M., Ensinger, C. L., Haack, T., Hood, M. M., Jones, J., Lord, J. W., Lu, W. P., Miller, D., Patt, W. C., Smith, B. D., Petillo, P. A., Rutkoski, T. J., Telikepalli, H., Vogeti, L., Yao, T., Chun, L., Clark, R., Evangelista, P., Gavrilescu, L. C., Lazarides, K., Zaleskas, V. M., Stewart, L. J., Van Etten, R. A., Flynn, D. L. (2011) Conformational Control Inhibition of the BCR-ABL1 Tyrosine Kinase, Including the Gatekeeper T315I Mutant, by the Switch-Control Inhibitor DCC-2036. *Cancer Cell 19*, 556-568.

67. Zhang, C., Ibrahim, P. N., Zhang, J., Burton, E. A., Habets, G., Zhang, Y., Powell, B., West, B. L., Matusow, B., Tsang, G., Shellooe, R., Carias, H., Nguyen, H., Marimuthu, A., Zhang, K. Y. J., Oh, A., Bremer, R., Hurt, C. R., Artis, D. R., Wu, G., Nespi, M., Spevak, W., Lin, P., Nolop, K., Hirth, P., Tesch, G. H., Bollag, G. (2013) Design and pharmacology of a highly specific dual FMS and KIT kinase inhibitor. *PNAS 110*, 5689-5694.

68. Schneider, E. V., Böttcher, J., Blaesse, M., Neumann, L., Huber, R., Maskos, K. (2011) The Structure of CDK8/CycC Implicates Specificity in the CDK/Cyclin Family and Reveals Interaction with a Deep Pocket Binder. *J. Mol. Biol. 412*, 251-266.

69. Kornev, A. P., Taylor, S. S., Ten Eyck, L. F. (2008) A helix scaffold for the assembly of active protein kinases. *Proc. Natl. Acad. Sci. U. S. A. 105*, 14377-14382.

70. Azam, M., Seeliger, M. A., Gray, N. S., Kuriyan, J., Daley, G. Q. (2008) Activation of tyrosine kinases by mutation of the gatekeeper threonine. *Nat. Struct. Mol. Biol. 15*, 1109-1118.

71. Smith, B., Medda, F., Gokhale, V., Dunckley, T., Hulme, C. (2012) Recent Advances in the Design, Synthesis, and Biological Evaluation of Selective DYRK1A Inhibitors: A New Avenue for a Disease Modifying Treatment of Alzheimer's? *ACS Chem. Neurosci. 3*, 857-872.

72. Najjar, M., Suebsuwong, C., Ray, S. S., Thapa, R. J., Maki, J. L., Nogusa, S., Shah, S., Saleh, D., Gough, P. J., Bertin, J., Yuan, J., Balachandran, S., Cuny, G. D., Degterev, A. (2015) Structure Guided Design of Potent and Selective Ponatinib-Based Hybrid Inhibitors for RIPK1. *Cell Rep. 10*, 1850-1860.

73. Soverini, S., Hochhaus, A., Nicolini, F. E., Gruber, F., Lange, T., Saglio, G., Pane, F., Muller, M. C., Ernst, T., Rosti, G., Porkka, K., Baccarani, M., Cross, N. C. P., Martinelli, G. (2011) BCR-ABL kinase domain mutation analysis in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: recommendations from an expert panel on behalf of European LeukemiaNet. *Blood 118*, 1208-1215.

SUPPLEMENTARY INFORMATION

Table S1: List of the 444 kinase domain names used in the protein alignment.

AAK1	AATK	ABL1	ABL2	ACVR1	ACVR1B	ACVR1C	ACVR2A	ACVR2B
ACVRL1	ADRBK1	ADRBK2	AKT1	AKT2	AKT3	ALK	AMHR2	ANKK1
ARAF	AURKA	AURKB	AURKC	AXL	BLK	BMP2K	BMPR1A	BMPR1B
BMPR2	BMX	BRAF	BRSK1	BRSK2	BTK	BUB1	BUB1B	CAMK1
CAMK1D	CAMK1G	CAMK2A	CAMK2B	CAMK2D	CAMK2G	CAMK4	CAMKK1	CAMKK2
CCRK	CDC2	CDC2L2	CDC2L5	CDC2L6	CDC42BPA	CDC42BPB	CDC42BPG	CDC7
CDK10	CDK2	CDK3	CDK4	CDK5	CDK6	CDK7	CDK8	CDK9
CDKL1	CDKL2	CDKL3	CDKL4	CDKL5	CHEK1	CHEK2	CHUK	CIT
CLK1	CLK2	CLK3	CLK4	CRKRS	CSF1R	CSK	CSNK1A1	CSNK1A1L
CSNK1D	CSNK1E	CSNK1G1	CSNK1G2	CSNK1G3	CSNK2A1	CSNK2A2	DAPK1	DAPK2
DAPK3	DCLK1	DCLK2	DCLK3	DDR1	DDR2	DMPK	DYRK1A	DYRK1B
DYRK2	DYRK3	DYRK4	EGFR	EIF2AK1	EIF2AK2	EIF2AK3	EPHA1	EPHA2
EPHA3	EPHA4	EPHA5	EPHA6	EPHA7	EPHA8	EPHB1	EPHB2	EPHB3
EPHB4	ERBB2	ERBB4	ERN1	ERN2	FER	FES	FGFR1	FGFR2
FGFR3	FGFR4	FGR	FLT1	FLT3	FLT4	FRK	FYN	GAK
GCN2	GRK1	GRK4	GRK5	GRK6	GRK7	GSG2	GSK3A	GSK3B
HCK	HIPK1	HIPK2	HIPK3	HIPK4	HUNK	ICK	IGF1R	IKBKB
IKBKE	INSR	INSRR	IRAK1	IRAK3	IRAK4	ITK	JAK1	JAK2
JAK3	KALRN	KDR	KIAA1804	KIT	LATS1	LATS2	LCK	LIMK1
LIMK2	LMTK2	LMTK3	LRRK1	LRRK2	LTK	LYN	MAK	MAP2K1
MAP2K2	MAP2K3	MAP2K4	MAP2K5	MAP2K6	MAP2K7	MAP3K1	MAP3K10	MAP3K11
MAP3K12	MAP3K13	MAP3K14	MAP3K15	MAP3K2	MAP3K3	MAP3K4	MAP3K5	MAP3K6
MAP3K7	MAP3K8	MAP3K9	MAP4K1	MAP4K2	MAP4K3	MAP4K4	MAP4K5	MAPK1
MAPK10	MAPK11	MAPK12	MAPK13	MAPK14	MAPK15	MAPK3	MAPK4	MAPK6
MAPK7	MAPK8	MAPK9	MAPKAPK2	MAPKAPK3	MAPKAPK5	MARK1	MARK2	MARK3
MARK4	MAST1	MAST2	MAST3	MAST4	MASTL	MATK	MELK	MERTK
MET	MINK1	MKNK1	MKNK2	MLCK	MOS	RPS6KA5~d1	RPS6KA4~d1	MST1R
MST4	MUSK	MYLK	MYLK2	MYLK4	MYO3A	MYO3B	NEK1	NEK10
NEK11	NEK2	NEK3	NEK4	NEK5	NEK6	NEK7	NEK8	NEK9
NIM1	NLK	NRK1	NTRK1	NTRK2	NTRK3	NUAK1	NUAK2	OBSCN
OBSCN~b	OXSR1	PAK1	PAK2	PAK3	PAK4	PAK6	PAK7	PASK
PBK	PCTK1	PCTK2	PCTK3	PDGFRA	PDGFRB	PDIK1L	PDPK1	PFTK1
PFTK2	PHKG1	PHKG2	PIK3R4	PIM1	PIM2	PIM3	PINK1	PKDCC
PKMYT1	PKN1	PKN2	PKN3	PLK1	PLK2	PLK3	PLK4	PNCK
PRKAA1	PRKAA2	PRKACA	PRKACB	PRKACG	PRKCA	PRKCB	PRKCD	PRKCE
PRKCG	PRKCH	PRKCI	PRKCQ	PRKCZ	PRKD1	PRKD2	PRKD3	PRKG1
PRKG2	PRKX	PRKY	PRPF4B	PSKH1	PTK2	PTK2B	PTK6	RAF1
RAGE	RET	RIPK1	RIPK2	RIPK3	RIPK4	RIPK5	RNASEL	ROCK1
ROCK2	ROR1	ROR2	ROS1	RPS6KA1~d2	RPS6KA2~d2	RPS6KA3~d2	RPS6KA4~d2	RPS6KA5~d2
RPS6KA6~d2	RPS6KB1	RPS6KB2	RPS6KA1~d1	RPS6KA2~d1	RPS6KA3~d1	RPS6KA6~d1	RYK	SBK1
SBK2	SGK1	SGK110	SGK2	SGK3	SGK494	SIK3	SLK	SNF1LK
SNF1LK2	SNRK	SPEG	SPEG~b	SRC	SRMS	SRPK1	SRPK2	SRPK3
STK10	STK11	STK16	STK17A	STK17B	STK24	STK25	STK3	STK32A
STK32B	STK32C	STK33	STK35	STK36	STK38	STK38L	STK39	STK4
SYK	TAOK1	TAOK2	TAOK3	TBK1	TEC	TEK	TESK1	TESK2
TGFbR1	TGFbR2	TIE1	TLK1	TLK2	TNIK	TNK1	TNK2	TNNI3K
TP53RK	TRIO	TSSK1B	TSSK2	TSSK3	TSSK4	TSSK6	TTBK1	TTBK2
TTK	TTN	TXK	TYK2	TYRO3	UHMK1	ULK1	ULK2	ULK3
VRK1	VRK2	WEE1	WEE2	WNK1	WNK2	WNK3	WNK4	YES1
YSK4	ZAK	ZAP70						

Table S2: The first 5 z-scale components for the twenty natural amino-acids. *(21)*

Residue	z1	z2	z3	z4	z5
Ala	0.24	-2.32	0.60	-0.14	1.30
Arg	3.52	2.50	-3.50	1.99	-0.17
Asn	3.05	1.62	1.04	-1.15	1.61
Asp	3.98	0.93	1.93	-2.46	0.75
Cys	0.84	-1.67	3.71	0.18	-2.65
Gln	1.75	0.50	-1.44	-1.34	0.66
Glu	3.11	0.26	-0.11	-3.04	-0.25
Gly	2.05	-4.06	0.36	-0.82	-0.38
His	2.47	1.95	0.26	3.90	0.09
Ile	-3.89	-1.73	-1.71	-0.84	0.26
Leu	-4.28	-1.30	-1.49	-0.72	0.84
Lys	2.29	0.89	-2.49	1.49	0.31
Met	-2.85	-0.22	0.47	1.94	-0.98
Phe	-4.22	1.94	1.06	0.54	-0.62
Pro	-1.66	0.27	1.84	0.70	2.00
Ser	2.39	-1.07	1.15	-1.39	0.67
Thr	0.75	-2.18	-1.12	-1.46	-0.40
Trp	-4.36	3.94	0.59	3.44	-1.59
Tyr	-2.54	2.44	0.43	0.04	-1.47
Val	-2.59	-2.64	-1.54	-0.85	-0.02

Table S3: List of the 200 variables selected using the absolute regression coefficients of the PLS model. The variables are in decreasing order according to the absolute regression coefficients. In red are the variables kept after filtering on the total occurrence of the residue positions and on the gap proportions (gap percentage < 20%).

variables	absolute_coefficients
res325z1	0.90
res325z4	0.81
res426z5	0.70
res662z2	0.68
res226z2	0.67
res101z3	0.65
res804z2	0.64
res391z4	0.63
res388z5	0.62
res105z4	0.61
res103z2	0.61
res108z1	0.60
res325z2	0.60
res799z5	0.60
res241z1	0.59
res785z1	0.59
res230z5	0.59
res177z3	0.58
res722z3	0.58
res734z2	0.57
res426z3	0.56
res666z5	0.56
res722z5	0.55
res187z5	0.54
res691z2	0.53
res729z5	0.53
res389z1	0.52
res799z3	0.52
res622z3	0.51
res371z3	0.50
res363z1	0.49
res118z3	0.49
res389z5	0.49
res665z3	0.49
res378z2	0.49
res625z1	0.49
res754z4	0.49
res366z1	0.48
res234z1	0.47
res675z4	0.47
res240z2	0.47
res331z5	0.47
res330z5	0.46
res404z2	0.46
res567z2	0.46
res372z2	0.46
res746z2	0.45
res675z3	0.45
res725z5	0.45
res384z3	0.45
res567z5	0.45
res414z1	0.44
res802z4	0.44
res426z4	0.44
res371z5	0.44
res402z5	0.44
res366z2	0.44
res295z1	0.44
res397z4	0.43
res622z1	0.43
res331z1	0.43
res372z3	0.43
res226z3	0.43
es375z2	0.42
res367z4	0.42
res325z3	0.42
res365z4	0.42
res724z1	0.42
res414z2	0.42
res521z1	0.42
res739z5	0.42
res808z2	0.42
res249z4	0.41
res570z1	0.41
res389z2	0.41
res778z4	0.41
res567z1	0.41
res384z4	0.41
res782z3	0.40
res577z1	0.40
res402z4	0.40
res778z3	0.40
res296z5	0.40
res785z3	0.40
res367z3	0.40
res493z2	0.39
res798z2	0.39
res778z5	0.39
res663z4	0.39
res239z3	0.39
res615z2	0.39
res371z1	0.39
res740z4	0.39
res491z5	0.39
res121z2	0.39
res809z5	0.38
res194z1	0.38
res98z4	0.38
res577z3	0.38
res236z1	0.38
res366z5	0.38
res389z4	0.38
res802z1	0.38
res738z1	0.38
res181z5	0.38
res527z5	0.38
res118z1	0.38
res324z3	0.38
res316z4	0.38
res248z5	0.38
res196z1	0.38
res801z2	0.38
res606z2	0.37
res367z5	0.37
res809z1	0.37
res368z1	0.36
res254z1	0.36
res606z4	0.36
res605z2	0.36
res100z3	0.36
res196z2	0.36
res119z1	0.36
res665z4	0.36
res227z2	0.35
res725z2	0.35

res424z3	0.35
res329z5	0.35
res385z1	0.35
res425z4	0.35
res114z5	0.35
res748z2	0.35
res392z1	0.35
res120z5	0.35
res295z3	0.35
res407z4	0.34
res806z1	0.34
res179z4	0.34
res785z5	0.34
res242z3	0.34
res751z4	0.34
res427z5	0.34
res187z1	0.34
res401z4	0.34
res584z3	0.34
res362z3	0.34
res296z2	0.34
res189z5	0.34
res119z3	0.34
res396z1	0.34
res615z1	0.34
res665z2	0.34
res492z2	0.34
res492z5	0.34
res808z5	0.34
res192z5	0.33
res806z4	0.33
res400z3	0.33
res393z5	0.33
res114z3	0.33
res493z3	0.33
res331z3	0.33
res569z5	0.33
res374z5	0.33
res370z5	0.33

res808z4	0.33
res450z1	0.33
res388z3	0.33
res444z3	0.33
res329z4	0.32
res579z5	0.32
res450z3	0.32
res331z2	0.32
res416z2	0.32
res606z5	0.32
res662z5	0.32
res239z4	0.32
res787z1	0.32
res386z4	0.32
res365z5	0.32
res409z4	0.32
res528z5	0.32
res493z1	0.32
res184z2	0.31
res409z2	0.31
res620z3	0.31
res795z5	0.31
res617z4	0.31
res610z3	0.31
res196z3	0.31
res410z5	0.31
res116z3	0.31
res735z5	0.31
res192z4	0.31
res358z2	0.31
res567z3	0.31
res416z5	0.31
res402z3	0.31
res226z4	0.31
res371z2	0.31
res743z5	0.31

Figure S4: Aligned ABL1 sequence with the corresponding numbering used in the text. The residue positions selected by the model are in red.

Table S5: Comparison of the 29 residue positions found with the global statistical model including four Type-II kinase inhibitors with the residue positions found with each individual model built with one Type-II inhibitor.

Residue position	Occurences imatinib	rank imatinib	Occurences masitinib	Rank masitinib	Occurences nilotinib	rank nilotinib	Occurences sorafenib	rank sorafenib
res426	5	1	4	2	4	2	2	17
res325	2	23	3	7	4	1	3	2
res192	2	4	3	4	1	75	2	14
res606	3	3	2	13	2	21	1	67
res785					3	3	3	3
res414	1	50			3	13	3	1
res402	1	57	1	81	3	7	3	5
res226			2	32	3	9	2	11
res675	2	13	2	18	2	22	1	41
res802	4	2	2	12				
res239			3	14			3	7
res187					3	14	2	18
res331	3	7	1	66	2	24		
res118	1	52			3	11	1	57
res799			2	29	2	35		
res665							3	4
res492			1	60	2	28		
res409					1	62	2	23
res365			1	72	1	73	1	69
res329					2	29		
res114	2	20	1	51				
res450							2	20
res778					1	52	1	73
res295	1	46	1	49				
res577	1	42	1	59				
res622							1	32
res296					1	53		

Table S6: Residue selection from the dataset provided by Davis *et al.(13)* based on the best 200 variables selected by PLS regression. Only residue positions containing at least 2 z-scale descriptors and with less than 20% of gaps were kept. Asterisks symbolized the residue positions also found by applying the PLS model on the training set.

Residue positions	Occurrence	Corresponding variables	Sum of absolute coefficients
res226*	4	res226z2, res226z1, res226z4, res226z3	0.05
res325*	4	res325z4, res325z1, res325z2, res325z3	0.05
res606*	3	res606z4, res606z2, res606z5	0.04
res187*	3	res187z5, res187z1, res187z3	0.04
res665*	3	res665z3, res665z4, res665z2	0.04
res499	3	res499z4, res499z1, res499z2	0.04
res785*	3	res785z1, res785z3, res785z4	0.03
res397	2	res397z4, res397z2	0.03
res804	2	res804z1, res804z2	0.03
res799*	3	res799z5, res799z3, res799z2	0.03
res414*	2	res414z1, res414z2	0.02
res675*	3	res675z4, res675z1, res675z3	0.02
res108	2	res108z1, res108z2	0.02
res625	2	res625z4, res625z1	0.02
res787	2	res787z4, res787z1	0.02
res742	2	res742z5, res742z3	0.02
res360	2	res360z4, res360z3	0.02
res327	2	res327z2, res327z5	0.02
res425	2	res425z1, res425z2	0.02
res402*	2	res402z5, res402z4	0.02
res330	2	res330z4, res330z5	0.02
res321	2	res321z3, res321z5	0.02
res524	2	res524z2, res524z5	0.02
res181	2	res181z2, res181z1	0.02
res111	2	res111z3, res111z5	0.02
res118*	2	res118z3, res118z1	0.02
res331*	2	res331z2, res331z5	0.02
res622*	2	res622z4, res622z1	0.02
res404	2	res404z3, res404z1	0.02
res194	2	res194z5, res194z1	0.02

Table S7: Percentages of inhibition of Type-II inhibitors for each protein kinase. Non-bold rows represent the 263 protein kinases, used as training set for the PLS model and for which the activity was measured by Anastassiadis *et al*. Bold rows represent protein kinases for which the activity was predicted by our statistical model. The rows in which the predicted activity is given in parentheses represent protein kinases that had been removed from the training set, and for which their activities were measured by Anastassiadis *et al.(12)*(see Method)

Kinase	Percentage of inhibition
AAK1	**38.3**
AATK	**-12.0**
ABL1	96.5
ABL1/E255K/	**93.0**
ABL1/F317I/	**94.0**
ABL1/F317L/	**94.0**
ABL1/H396P/	**94.0**
ABL1/M351T/	**89.9**
ABL1/Q252H/	**85.3**
ABL1/T315I/	**85.2**
ABL1/Y253F/	**94.0**
ABL2	95.1
ACVR1	1.3
ACVR1B	0.2
ACVR1C	**17.5**
ACVR2A	**-0.3**
ACVR2B	**8.9**
ACVRL1	-3.1
ADRBK1	4.2
ADRBK2	4.9
AKT1	6.9
AKT2	-2.5
AKT3	25.2
ALK	4.9
AMHR2	**-16.1**
ANKK1	**12.3**
ARAF	93.0
AURKA	21.1
AURKB	36.87 **(2.9)**
AURKC	12.8
AXL	35.23 **(44.1)**
BLK	42.5 **(68.7)**
BMP2K	**20.5**
BMPR1A	**1.7**
BMPR1B	**-3.7**
BMPR2	**-4.9**
BMX	6.7
BRAF	88.5
BRAF/V600E/	**86.3**
BRSK1	4.2
BRSK2	0.1
BTK	0.4
BUB1	**58.0**
BUB1B	**11.9**
CAMK1	-2.4
CAMK1D	12.3
CAMK1G	1.6
CAMK2A	4.7
CAMK2B	5.0
CAMK2D	10.0
CAMK2G	1.4
CAMK4	1.4
CAMKK1	8.9
CAMKK2	2.5
CCRK	**-3.9**
CDC2	2.4
CDC2L2	**-6.8**
CDC2L5	**6.0**
CDC2L6	**-14.4**
CDC42BPA	14.8
CDC42BPB	2.9
CDC42BPG	**7.6**
CDC7	**46.1**
CDK10	**-12.9**
CDK2	5.3
CDK3/cyclinE/	**1.3**
CDK4	5.9
CDK5	8.3
CDK6	9.5
CDK7	-0.6
CDK8	**-9.7**
CDK9	5.3
CDKL1	**17.9**
CDKL2	**10.2**
CDKL3	**11.6**
CDKL4	**36.8**
CDKL5	**5.2**
CHEK1	9.1
CHEK2	0.9
CHUK	3.8
CIT	**10.6**
CLK1	26.2
CLK2	14.0
CLK3	6.4
CLK4	54.5
CRKRS	**4.5**
CSF1R	94.4
CSK	57.6

CSNK1A1	2.2
CSNK1A1L	**14.5**
CSNK1D	10.7
CSNK1E	7.6
CSNK1G1	7.6
CSNK1G2	11.1
CSNK1G3	3.2
CSNK2A1	0.8
CSNK2A2	13.3
DAPK1	2.5
DAPK2	1.5
DAPK3	8.5
DCLK1	**-4.0**
DCLK2	-1.7
DCLK3	**-4.6**
DDR1	**101.1**
DDR2	98.2
DMPK	6.8
DYRK1A	-1.7
DYRK1B	7.2
DYRK2	1.1
DYRK3	8.6
DYRK4	-0.4
EGFR	13.1
EGFR/E746_A750del/	**8.4**
EGFR/G719C/	**14.3**
EGFR/G719S/	**14.3**
EGFR/L747-T751del_Sins/	**8.4**
EGFR/L747_E749del_A750P/	**8.4**
EGFR/L747_S752del_P753S/	**8.8**
EGFR/L858R/	**17.3**
EGFR/L861Q/	**14.3**
EGFR/S752-I759del/	**12.5**
EGFR/T790M/	**14.3**
EGFR/T790M_L858R/	**17.3**
EIF2AK1	**18.5**
EIF2AK2	**6.8**
EIF2AK3	**-23.7**
EPHA1	36.8 **(64.5)**
EPHA2	89.2
EPHA3	59.2
EPHA4	88.0
EPHA5	77.4
EPHA6	58.0
EPHA7	19.3
EPHA8	71.6
EPHB1	89.4
EPHB2	98.5
EPHB3	94.4
EPHB4	88.6
ERBB2	-2.7
ERBB4	17.9
ERN1	**23.7**
ERN2	**12.4**

FER	6.4
FES	4.1
FGFR1	9.1
FGFR2	27.3
FGFR3	10.5
FGFR4	10.2
FGR	56.4
FLT1	49.5 **(50.3)**
FLT3	90.8 **(62.0)**
FLT3/D835H/	**62.0**
FLT3/D835Y/	**62.0**
FLT3/K663Q/	**62.0**
FLT3/N841I/	**56.6**
FLT3/R834Q/	**62.0**
FLT4	57.0
FRK	83.8
FYN	62.7
GAK	**35.7**
GRK1	**22.9**
GRK4	10.8
GRK5	5.0
GRK6	-2.6
GRK7	**-4.8**
GSG2	33.5 **(13.4)**
GSK3A	0.6
GSK3B	6.3
HCK	53.4
HIPK1	3.0
HIPK2	12.2
HIPK3	14.3
HIPK4	**95.1**
HUNK	**-7.4**
ICK	**-3.1**
IGF1R	11.6
IKBKB	17.2
IKBKE	10.7
INSR	1.0
INSRR	-1.5
IRAK1	6.7
IRAK3	**19.1**
IRAK4	19.8
ITK	1.8
KALRN	**-7.8**
KDR	63.1
KIAA1804	**24.7**
KIT	**71.8**
KIT/A829P/	**74.8**
KIT/D816H/	**71.8**
KIT/D816V/	**71.8**
LATS1	**-4.7**
LATS2	**-3.4**
LCK	92.4
LIMK1	23.2
LIMK2	**45.1**

LMTK2	**-1.5**
LMTK3	**-5.5**
LRRK1	**7.6**
LRRK2	18.9
LTK	0.1
LYN	83.8
MAK	**8.7**
MAP2K1	-2.5
MAP2K2	10.3
MAP2K3	**-6.8**
MAP2K4	**8.4**
MAP2K5	**30.0**
MAP2K6	3.7
MAP2K7	**-1.9**
MAP3K1	**28.8**
MAP3K10	-1.9
MAP3K11	1.7
MAP3K12	**-0.7**
MAP3K13	**8.9**
MAP3K14	9.0
MAP3K15	**7.7**
MAP3K2	-2.3
MAP3K3	-5.9
MAP3K4	**11.4**
MAP3K5	5.0
MAP3K6	**-11.7**
MAP3K7	10.8
MAP3K8	8.5
MAP3K9	-2.3
MAP4K1	**-1.7**
MAP4K2	-2.5
MAP4K3	**11.9**
MAP4K4	27.8
MAP4K5	43.0 **(14.4)**
MAPK1	10.1
MAPK10	9.2
MAPK11	84.7
MAPK12	7.0
MAPK13	9.3
MAPK14	76.7
MAPK15	**-6.1**
MAPK3	-1.5
MAPK4	**13.6**
MAPK6	**21.1**
MAPK7	**9.7**
MAPK8	7.9
MAPK9	5.2
MAPKAPK2	5.8
MAPKAPK3	2.2
MAPKAPK5	9.2
MARK1	25.2
MARK2	5.3
MARK3	6.2
MARK4	1.9

MAST1	**-16.1**
MAST2	**-18.6**
MAST3	**-13.1**
MAST4	**-9.4**
MASTL	**20.7**
MATK	20.8
MELK	15.2
MERTK	11.8
MET	4.2
MET/M1250T/	**12.3**
MET/Y1235D/	**15.9**
MINK1	13.3
MKNK1	24.9
MKNK2	29.7
MLCK	**10.2**
MOS	**26.0**
MST1R	15.3
MST4	3.0
MUSK	8.9
MYLK	7.4
MYLK2	3.9
MYLK4	**4.2**
MYO3A	**20.6**
MYO3B	17.4
NEK1	25.0
NEK10	**-1.9**
NEK11	4.9
NEK2	3.3
NEK3	4.1
NEK4	7.6
NEK5	**7.6**
NEK6	8.7
NEK7	4.5
NEK8	**1.2**
NEK9	4.6
NIM1	**14.1**
NLK	0.9
NRK1	**2.8**
NTRK1	23.9 **(12.5)**
NTRK2	14.6 **(9.8)**
NTRK3	34.5 **(1.6)**
NUAK1	1.2
NUAK2	1.6
OBSCN/b/	**31.6**
OXSR1	4.4
PAK1	-4.0
PAK2	5.8
PAK3	-3.4
PAK4	11.1
PAK6	-1.5
PAK7	8.0
PASK	1.1
PBK	2.5
PCTK1	**1.5**

PCTK2	**12.0**
PCTK3	**-11.4**
PDGFRA	98.4
PDGFRB	88.0
PDIK1L	**11.5**
PDPK1	15.4
PFTK1	**-1.4**
PFTK2	**-0.9**
PHKG1	9.0
PHKG2	30.9 **(7.5)**
PIK3R4	**-17.6**
PIM1	13.6
PIM2	1.9
PIM3	26.3
PINK1	**16.8**
PKDCC	**29.5**
PKMYT1	**16.4**
PKN1	1.1
PKN2	5.8
PKN3	**8.7**
PLK1	3.0
PLK2	2.1
PLK3	0.7
PLK4	**3.2**
PNCK	5.3
PRKAA1	**15.4**
PRKAA2	**11.1**
PRKACA	7.4
PRKACB	**9.1**
PRKACG	-0.1
PRKCA	16.5
PRKCB	**16.2**
PRKCD	2.1
PRKCE	3.6
PRKCG	18.3
PRKCH	-2.5
PRKCI	10.1
PRKCQ	24.6
PRKCZ	13.0
PRKD1	2.1
PRKD2	11.3
PRKD3	10.6
PRKG1	22.4
PRKG2	9.1
PRKX	2.3
PRKY	**3.6**
PRPF4B	**4.4**
PSKH1	**0.5**
PTK2	5.1
PTK2B	5.5
PTK6	29.6
RAF1	98.6
RAGE	**21.0**
RET	98.2

RET/M918T/	**91.6**
RET/V804L/	**88.6**
RET/V804M/	**85.6**
RIPK1	**15.6**
RIPK2	13.8
RIPK3	**45.2**
RIPK4	**15.0**
RIPK5	7.6
RNASEL	**38.3**
ROCK1	-0.5
ROCK2	4.9
ROR1	**16.1**
ROR2	**48.0**
ROS1	14.1
RPS6KA1/d2/	**2.0**
RPS6KA2/d2/	**21.1**
RPS6KA3/d2/	**10.9**
RPS6KA4/d2/	**4.7**
RPS6KA5/d2/	**-11.9**
RPS6KA6/d2/	**24.9**
RPS6KB1	8.6
RPS6KB2	0.0
RYK	**7.8**
SBK1	**-10.2**
SBK2	**13.3**
SGK1	4.8
SGK110	**36.2**
SGK2	-2.5
SGK3	4.4
SGK494	**15.6**
SIK3	**11.5**
SLK	7.6
SNF1LK	**6.5**
SNF1LK2	0.3
SNRK	**-28.0**
SPEG/b/	**2.2**
SRC	20.0
SRMS	7.1
SRPK1	1.9
SRPK2	-1.5
SRPK3	11.6
STK10	27.6
STK11	-2.2
STK16	6.6
STK17A	7.6
STK17B	**10.8**
STK24	-3.5
STK25	4.5
STK3	6.1
STK32A	**21.2**
STK32B	**27.0**
STK32C	**10.9**
STK33	16.1
STK35	**15.4**

STK36	**25.1**
STK38	4.1
STK38L	**-2.9**
STK39	6.8
STK4	-0.4
SYK	-1.1
TAOK1	21.5
TAOK2	63.1
TAOK3	36.5 **(50.8)**
TAOK3	36.5
TBK1	3.3
TEC	16.8
TEK	2.8
TESK1	**0.6**
TESK2	**-10.3**
TGFbR1	5.9
TGFbR2	2.3
TIE1	**5.2**
TLK1	**8.0**
TLK2	21.6
TNIK	**15.0**
TNK1	**26.5**
TNK2	17.4
TNNI3K	23.4
TP53RK	**-0.3**
TRIO	**-3.1**
TSSK1B	22.4
TSSK2	26.5
TSSK3	**8.9**
TSSK4	**17.0**
TSSK6	**-1.2**
TTBK1	**31.1**
TTBK2	**29.3**
TTK	2.1
TTN	**11.6**
TXK	-1.3
TYRO3	8.3
UHMK1	**34.7**
ULK1	6.4
ULK2	-2.5
ULK3	-2.8
VRK1	3.3
VRK2	**-0.2**
WEE1	3.0
WEE2	**14.8**
WNK1	**5.3**
WNK2	0.8
WNK3	0.8
WNK4	**11.6**
YES1	49.3 **(53.5)**
YES1	49.3
YSK4	**3.8**
ZAK	83.3
ZAP70	4.9

Figure S8: Residue selection from the PLS model established with the ponatinib datset. Only residue positions containing at least 2 z-scale descriptors and less than 20% of gaps were kept. We calculated three different rankings. Ranking sum_abs_coeff was obtained by summing the absolute coefficients of each residue position. Rank mean_diff was calculated according to the mean percentage difference for each residue at each position between protein kinases that bind the four Type-II inhibitors and those that do not bind the inhibitors. Rank max_diff was determined using the maximum difference for each residue. In bold are the positions we studied.

Residue position	Rank sum_abs_coeff	Rank Mean_diff	Rank Max_diff
res296	23	17	1
res426	27	10	2
res449	4	3	3
res325	2	6	4
res404	5	2	5
res184	106	4	6
res108	6	14	8
res450	1	1	9
res254	57	18	10
res746	89	20	11
res192	31	23	12
res492	3	7	13
res778	35	22	14
res491	55	16	16
res414	50	8	17
res250	11	15	19
res730	7	35	20
res252	54	12	21
res238	8	11	22
res572	14	29	23
res666	76	32	24
res327	12	19	25
res117	32	26	26
res100	26	34	28
res96	96	28	29
res226	10	39	30
res425	13	31	31
res233	37	36	32
res191	46	21	34
res97	30	45	35
res402	17	13	36
res779	24	25	37
res799	80	37	38
res390	15	24	39
res605	41	9	41
res112	93	33	44
res739	66	38	45
res320	29	49	46
res801	18	40	47
res741	42	48	48
res782	45	50	49
res621	51	47	50

Figure S9: Representation of the C and R spines on the kinase domain of ABL1. The hydrophobic part of the R spine is shown as a red molecular surface, while the C spine is colored in yellow. The 29 selected residues identified by the PLS regression are represented in dark blue stick. The spines residues are atom-colored.

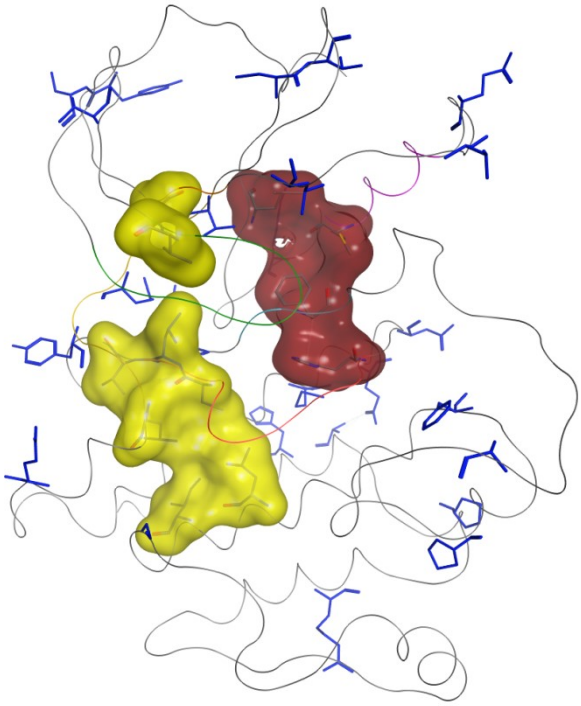
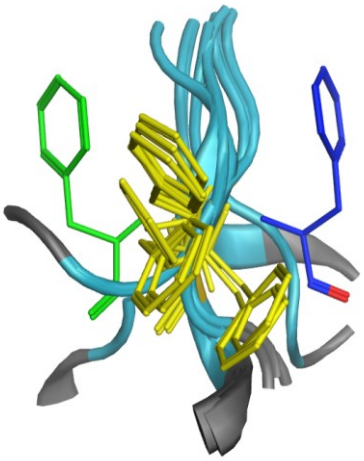


Figure S10: Representation of intermediate conformations of Phe of DFG motif (yellow) for protein kinases classified as DFG-out by MOE, with a Met as gatekeeper (PDB entries 1M7N, 1P4O, 2JIT, 2JIU, 2W5B, 2W5H, 3IKA and 4FZF). Phe conformations of DFG-out (green) and DFG-in (blue) of ABL1 crystal structures from PDB entries 2HYY and 2GQG respectively.



V. Une analyse kino-chimiométrique du kinome humain

Bien que les protéines kinases relèvent d'une importance considérable pour beaucoup d'acteurs du domaine pharmaceutique, leur utilisation en protéo-chimiométrie, ou comme nous préférons dire, en kino-chimiométrie (KCM), est encore rare comparée aux modèles QSAR. Comme déjà mentionné dans la partie (II.E.5.b.I), cet état de fait pourrait venir de la quantité de données qui, bien que déjà conséquente, pourrait ne pas être suffisante pour établir des modèles KCM suffisamment performants. Dans le cadre fixé par cette thèse, nous nous sommes intéressés à l'élaboration de modèles KCM tridimensionnels à partir de données qui n'ont pas encore été exploitées à cette fin. Pour ce faire nous avons développé un nouveau descripteur de protéines kinases basé sur leur structure 3D. Nous allons détailler ce descripteur, ainsi que son utilisation dans le cadre d'un modèle KCM basé sur un jeu de données publique, dans la prochaine partie (Partie V.A). Nous verrons ensuite dans cette partie que nous avons également tenté d'élaborer des modèles KCM à partir de données privées fournies par l'entreprise Janssen.

A. Publication en cours de soumission : *Prediction of protein kinase – ligand interactions through 2.5D Kinochemometrics*

Prediction of protein kinase – ligand interactions through 2.5D Kinochemometrics

Nicolas Bosc[1], Berthold Wroblowski[2], Christophe Meyer[3], and Pascal Bonnet[1*]

[1] *Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067 Orléans Cedex 2, France*
[2] *Janssen Research & Development, a division of Janssen Pharmaceutica N.V., Turnhoutseweg 30, 2340 Beerse, Belgium*
[3] *Centre de Recherche Janssen-Cilag, Campus de Maigremont - CS 10615, 27106 Val de Reuil CEDEX*

[*]Author to whom all correspondence should be addressed:
Pascal Bonnet: Tel: +33 238 417 254, Fax: +33 238 417 254, E-mail: pascal.bonnet@univ-orleans.fr

ABSTRACT

So far 518 protein kinases have been identified in the Human genome. They share a common mechanism of protein phosphorylation and are involved in many critical biological processes of eukaryotic cells. Deregulation of the kinase phosphorylation function induces severe diseases like cancer, diabetes or inflammatory diseases. Many actors in the pharmaceutical domain have made significant efforts to design new potent and selective protein kinase inhibitors as new potential drugs. Because the ATP binding site is highly conserved in the protein kinase family, the design of selective inhibitors remains a challenge and has negatively impacted the progression of drug candidate to late-stage clinical development. The work presented here comes within the context of a 2.5D kinochemometric (KCM) approach, derived from proteochemometrics (PCM), in which the protein kinases are depicted by a new 3D descriptor and the ligands by a 2D fingerprint. We demonstrate in two examples that the protein descriptor successfully classifies protein kinases based on their group membership and their DFG conformation. We also compared the performance of our models with those obtained from a full 2D KCM model. In both cases, the internal validations of the models demonstrated good capabilities to distinguish "active" from the "inactive" protein kinase – ligand pairs. However, the external validations performed on an independent dataset showed that our both approaches models tend to overestimate the number of "inactive" pairs.

INTRODUCTION

A well-recognized study has identified all 518 human protein kinases represent around 2% of the total human proteome. (*1*) They are involved in the mechanism of protein phosphorylation and regulate many biological processes such as cell division, insulin response or neuronal survival. Impaired kinase phosphorylation function is recognized to play a critical role in diseases like cancer, diabetes or neuronal depletion. (*2*,*3*,*4*) As a result, considerable efforts have been made to find or design protein kinase inhibitors as new potential drugs. (*5*) To date, since the approval of imatinib in 2002, 28 small molecule drugs have been approved by the US Food and Drug Administration (FDA). The research of new protein kinase inhibitors faces issues among which we can mention the selectivity that may lead to cellular toxicity.*(6)* This issue mainly comes from the kinase domain ATP binding site which is structurally highly conserved in its active conformation among the proteins of the kinome. Attempts to improve protein kinase inhibitor selectivity moved towards the development of ligands targeting secondary pockets of the protein kinases. While, Type I inhibitors competes directly ATP in a protein kinase active conformation, Type II inhibitors take advantage of the inactive DFG-out conformation that may adopt certain protein kinases. (*7*,*8*) In this conformation, a hydrophobic pocket, adjacent to the ATP binding site, may become accessible. Type II inhibitors bind in this pocket and block the protein kinase in its inactive conformation. (*9*) A popular belief proposes that Type II inhibitors may be more selective than Type I. However recent studies have shown that both Type I and Type II inhibitors may have similar selectivity profiles.*(10,11)* Hence, finding potent and selective protein kinase inhibitors remains an important challenge in drug discovery. (*11*)

Among the *in silico* approaches developed to help in finding new drugs, proteochemometrics (PCM) has recently emerged has a useful tool to predict the interactions between related proteins and several compounds. (*12*) Unlike quantitative structure-activity relationships (QSAR) that describe the interactions of a set of compounds on a unique target, only by considering the compound space, PCM takes into account the interactions from both compound and target information. Using the protein-ligand interaction space, PCM seems to be well designed to explore compound activities on several targets and to study the compound selectivity. Such approaches have already been applied on various protein families (*13*) and in particular on protein kinases. (*14*,*15*,*16*,*17*) Most of these studies based their approach on only a 2D representation of

protein kinase using protein sequences.(*13*,*14*,*15*) Although they have shown good performances both in regression and classification, such depiction, by definition, is not sufficient to characterize the structural properties of protein kinases that may be responsible for the ligand binding. Moreover, a 2D descriptor cannot be used to discriminate the different conformations of protein kinases, a critical point in the research of Type I and Type II inhibitors.

Here we intended to improve the kinase-based PCM, called kinochemometrics (KCM), and introduced a novel tridimensional (3D) protein kinase descriptor. It has been developed by using the physicochemical descriptors of several protein kinase residues and the residue pairwise distances. After statistical evaluation regarding the performance of the descriptor to classify protein kinases in their respective group or to distinguish DFG-in from DFG-out conformation, we applied the 3D descriptor, in combination with 2D molecular fingerprints for the ligands, to develop a novel KCM model. We compared the performance of our new approach to a more standard 2D approach.

METHODS

Datasets

Biological activities

The dataset used to generate the statistical models has been published in 2011 by Davis *et al.(10)* It comprised 72 protein kinase inhibitors tested on 439 human protein kinases from each group of the human kinome. The inhibitor set contains eleven approved kinase inhibitor drugs. For each protein-ligand pair, the authors performed full dose response to evaluate the affinity (Kd) if an initial activity was detected at 10 μM compound concentration. We converted each biological data into negative logarithmic value of the Kd (pKd). To obtain a full matrix, for each activity evaluated in the primary screen that was below the 10 μM compound concentration threshold, we assigned a pKd value of 5. To avoid any bias on protein information, we removed protein mutants and unphosphorylated protein states and kept only the phosphorylated wild type forms which represents the active proteins. Additionally, CDK4 were found in duplicate due to the presence of the regulatory cyclin proteins D1 and D3. The low mean standard deviation of biological affinities between these two complexes CDK4/CyclinD1 and CDK4/CyclinD3 (0.06 pKd unit) indicates a small influence of the cyclin domain. Hence, for each protein-ligand pair, we kept the one with the highest pKd

affinity. With 31752 distinct data points, this dataset offers the largest full matrix interaction data concerning protein kinases to our knowledge.

DUD-E

The Database of Useful Decoys – Enhanced (DUD-E) has been designed to help in benchmarking molecular docking programs by giving access to challenging decoys. (*18*) The database is divided into protein families among which, the protein kinases. Each of the 26 protein kinases available is associated with a set of active ligands (average of 224 ligands per target), retrieved from the ChEMBL database (version 09). A compound was considered as active on a target if the interaction value (IC50, EC50, Ki or Kd) was lower than 1 µM. For each active, 50 decoys with similar physico-chemical properties but with dissimilar 2D topology were collected from the ZINC database. (*19*) This dataset is particularly useful when testing docking software, because the scoring of decoys, with similar physico-chemical properties to the active compounds, is challenging. We used protein kinase dataset from DUD-E as an external validation set. In such validation, the data were never seen by the statistical models, *i.e.* they have not been included in the training set.

Description of protein kinases

3D descriptor

The first PCM models developed on the protein kinase family only considered the 2D representation of proteins using their amino acid sequence. (*14*,*15*,*16*) We can argue that a 2D depiction of protein kinase may induce major limitations. First, most of the protein kinases adopt several structural conformations which often change their biological function. As mentioned earlier, the binding of Type II inhibitors mainly depends on a protein conformational change that gives access to a specific allosteric pocket. The pocket accessibility depends on the activation loop of the protein kinase and in particular of the orientation of the DFG (Asp-Phe-Gly) motif, though other regions such as αC-helix or P-loop also seem to be strongly involved in the process. In the inactive conformational state, Phe rotates away the nearby αC-helix (DFG-out) and projects into the ATP pocket revealing a hydrophobic cavity. (*20*) A sequence-based descriptor of protein kinases cannot take into account such structural change. Secondly, since neither the 2D chemical

space of kinase inhibitors, (*21*) nor the 2D protein sequences cannot discriminate Type-I to Type-II inhibitors, a simple 2D representation does not provide enough ligand binding mode information.

For these reasons, we created a new 3D protein kinase-based descriptor to take into account kinase active states. We built this 3D descriptor in MOE (Molecular Operating Environment, Chemical Computing Group Inc. (*22*)) because this software already provides a prepared database of protein kinase X-ray structures from the Protein Data Bank (PDB). Moreover, the structures are all aligned on the same protein reference, which facilitate the development of the 3D descriptor. The descriptor is based on a selection of 16 important residues located in the ATP binding site that were identified in a previous study, using a genetic algorithm approach (Figure 1). (*23*) Many of these residues were mentioned in the literature as involved either in the binding of the ligands, (*24*) or in the inter protein interactions. (*23*) From this list of residue positions, we took into account two main dimensions of protein kinases: the residue pairwise distances reflecting 3D structural information and the 2D residue physicochemical properties. We calculated the pairwise distances for residues located on the different sequence motifs (for instance: hinge, catalytic loop, P-loop…), where the residue positions correspond to the position of the beta carbon (or center of mass for the glycine) for each residue. In addition, we described each amino acid using the z-scales descriptors. (*25*) These descriptors were obtained by applying principal component analysis on 26 physicochemical descriptors measured or calculated on 87 amino acids including the 20 natural amino acids. The first three principal components (z1 to z3) explained around 70% of the total variance of the original descriptors, while five components (z1 to z5) reach 87%. Thereby, the components may be interpreted as follow: hydrophobicity (z1), size and polarizability (z2), polarity (Z3), electronic properties (z4 and z5). Hence, for each protein structure, the 16 amino acids were identified according to their position, described by five numerical z-scales and their pairwise distances were measured. As a result, we obtained 190 numerical descriptors depicted by 2D sequence-based (80 z-scales) and 3D distance-based (110 distances) residues.

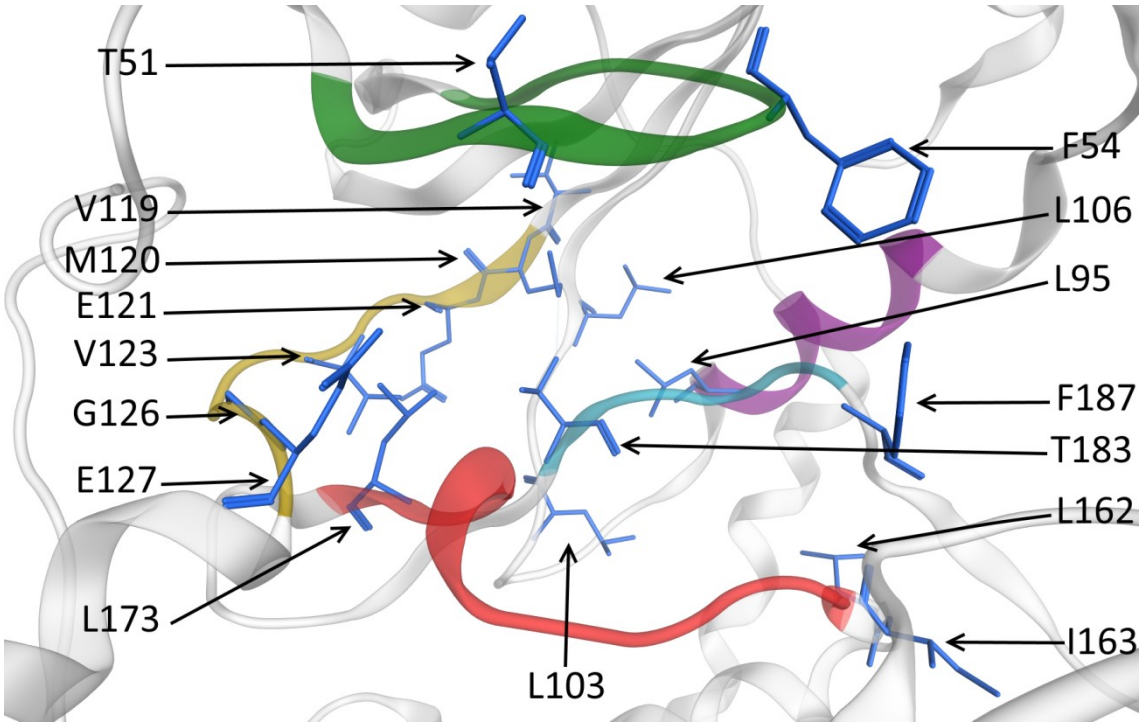


Figure 1: Mapping of the 16 protein kinase residue positions, identified by Martin et al., (23) on PKA (PDB ID: 1ATP) and used for residue descriptors.

Using the MOE kinome database updated in November 2014, we calculated the descriptor for 2549 unique protein structures associated to the human species. MOE provides a DFG-in/DFG-out classification of the kinase structures. In details, 2047 are DFG-in, 285 DFG-out and 201 could not be classified due to an "in-between" conformation or to missing residues in the DFG motif. In addition, we could not calculate the descriptor for 16 PDB codes due to important residues that were not resolved in the X-Ray structures. When a PDB code was associated to several chains, we kept the one with the best resolution. If several chains from the same PDB code belong to different DFG conformation, we kept the chain with the best resolution. We pulled out the structures with no DFG conformation identified.

Before using this new descriptor in our PCM models, we performed a Leave-One-Out (LOO) cross-validation on a set of protein descriptors where each protein kinase was represented once, *i.e.* for each protein kinase we kept the X-Ray structure with the best resolution. We obtained a total of 215 unique protein kinases. In LOO cross-validation, each protein was kept out of the protein set and a classification model was trained on the remaining proteins. The groups of the retained protein kinases were then predicted to estimate the accuracy of the model. The kinase groups refer to the classification established by Manning *et al.* (*1*) According to their sequence, a phylogenetic tree was established and eight kinase groups were

identified (AGC, CAMK, CK1, CMGC, Other, STE, TK and TKL). At the end of the LOO cross-validation process, we calculated the sensitivity, *i.e.* the propensity of the model to find the right kinase group for each retained protein. We also calculated the overall accuracy of each model, *i.e.* a parameter taking into account the right prediction (the sensitivity) and the wrong prediction rate for a given group. In addition, using the same approach, we estimated the sensitivity and the accuracy of the model to find the correct DFG conformation for each retained protein. The DFG-in/out classification for each protein was extracted from the 2014 MOE kinase database update in November 2014, it contains 2047 and 285 DFG-in and -out conformations respectively.

We performed the LOO cross-validation by comparing three different machine learning algorithms: Naïve Bayes (NB), Support Vector Machines (SVM) and Random Forest (RF). All statistical validations were performed using Knime pipelining software. *(26)*

2D descriptor

To compare the effect of the 3D descriptor on modelling performance, we also built statistical models using an alignment-based 2D protein kinase descriptor. For the multi alignment, the sequences of the 518 human kinase domains were retrieved in FASTA format from the KinBase database, (*1*,*27*) out of which 444 typical kinase domain sequences were identified. The typical kinase domain sequences were read using MOE sequence editor (Molecular Operating Environment, Chemical Computing Group Inc.) (*22*) and annotated with the kinase annotation tool in order to identify the conserved motifs. Sequences were aligned with the protein alignment protocol in MOE and a progressive strategy was employed to build the alignment using kinase constraints. Blosum100 matrix was used to score the alignment, for 100 iterations using the default values for Gap Start and Gap Extend. The accuracy of the overall alignment was tested by visually comparing conserved motifs.

Then we selected the residue positions according to their vicinity to three ligands in the corresponding three crystallographic structures representing the protein kinases SRC and ABL (in the DFG-in and DFG-out conformations), bound to ATP, dasatinib and imatinib, respectively (PDB ID 2G2F, 2GQG and 2HYY). The use of the natural substrate of protein kinases, and of two different types of inhibitors should cover the ensemble of residues involved in the majority of protein kinase-ligand interactions. We kept

only the residues within 8 Å around the ligands. We conserved a total of 110 residues after aligning all protein kinase structures. Finally, we replaced each amino acid in the subset of the multiple alignment, by the five numerical z-scales and the gaps with five null values. (*25*)

Description of kinase inhibitors

We described each chemical compound using RDKit Morgan fingerprint implemented in Knime, *(26)* with a radius of three. (*28*) The fingerprint takes into account the neighborhood of each atom within a molecule, by exploring the atoms in a vicinity of three bounds and defining a corresponding substructure. Each compound substructure is handled as a compound feature and mapped into a bit-vector array constituted of 0 and 1. Fingerprint descriptors have already been used in different QSAR and PCM approaches and have shown good results. (*29*,*30*)

Data processing

Protein descriptors calculated in MOE were imported in Knime and each protein kinase from the Ambit dataset, of which a crystal structure was available, was assigned by its corresponding 3D descriptor. When several crystal structures were available for one protein, we selected the structure having the best resolution. Hence, we first consider only the DFG-in conformation and kept the developed descriptor associated to the structure with the best resolution. As a result, we conserved 165 out the 439 protein kinases tested in the Ambit panel. This discrepancy demonstrates that roughly half of the human kinome has been today crystallized and a continuous effort from structural biology is still needed. Unfortunately this lack of crystallographic structures implies also a limited number of co-crystallized structures therefore the coupled information to generate a full matrix from both the structural information of ligand-bound to a protein kinase and the available biological data is very limited. This forced us to use 2D descriptors for the ligands and 2D and 3D descriptors for the proteins in order to maximize the amount of information from biological data. The alternative possibility regarding the use of 3D descriptors for the ligands is the use of docking to predict the binding mode of the ligands in each kinase or the use of a lowest energy conformer from a ligand conformational search. However, all these approaches will add supplementary inaccuracies to the statistical models. While interaction descriptor have been already used in some studies to detect similar binding site, this approach was not conceivable here. (*31*) Hence, we finally selected a 3D protein kinase descriptor from

crystal structures, a 2D descriptor from protein sequences and a 2D representation of the ligand to generate a 2.5D statistical model. Each of the 72 compounds of the dataset was described using the Morgan fingerprints.

For each protein kinase – ligand pairs, we assigned the corresponding protein and ligand descriptor. Hence, we described a total of 11'880 pairs, each of them being related to a unique pKd value. The biological data have been classified into 10 bins of 0.5 log unit each from a pKd < 5 to pKd > 9. The observations of the pKd distribution showed that a large majority corresponds to the bin "< 5" (Figure 2). This observation encouraged us to perform a classification instead of a regression.

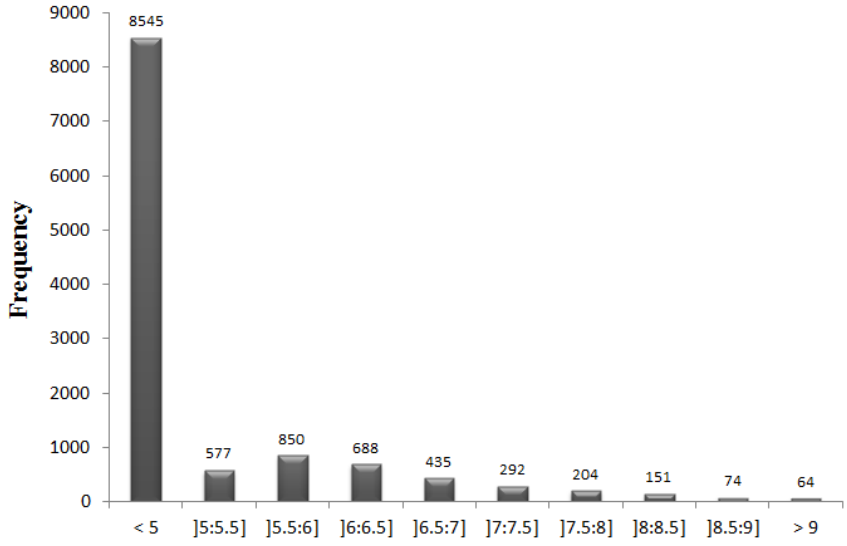


Figure 2: Distribution of the binned affinity values (pKd) from the Ambit dataset.*(10)*

The trained model was performed using R statistical package. (*32*) Before training, we scaled each variable such as the values of each descriptor are Gaussian distributed with a mean of 0.0 and a standard deviation of 1.0. Using statistical functions provided by the *camb* R package, (*33*) we removed the variables for which the variance was close to zero (var < 0.01) and we also removed variables when they were too correlated with others ($r^2 > 0.95$). Indeed, covariance within descriptors increases the dimensionality of the model that may lead to misinterpretation of the built-in model. The classification model was trained using the SVM implementation from the *caret* R package. (*34*) SVM is a non-linear modeling technique that has already been applied in several PCM studies, notably to model protein kinase – ligand interactions. (*14*,*15*)

This machine learning is particularly efficient to deal with high-dimensional sets, though the lack of interpretability is often stressed. (*13*) Activity classes have been created such that every protein kinase – ligand pair associated to a pKd greater than or equal to six was considered "active". On the opposite, the pairs that did not match this condition were considered as "inactive".

In parallel, we described the proteins mentioned in the Ambit dataset using the 2D protein descriptor and were able to describe 364 of them. In combination with the 2D ligand descriptor, the total of interaction pairs reached 24'912. The same protocol was applied to build full 2D KCM models.

Model validation

The partitioning of the 11'880 (2.5 KCM) or 24'912 (2D KCM) protein kinase – ligand pairs into a training (70% of the pairs) and a test sets (30% of the pairs) was performed in R. Nested cross-validation and grid search were performed for SVM parameter optimization (cost and gamma), by dividing the training set into five folds. A model was trained on four folds and used to predict the activity class ("active" or "inactive") of the fifth fold. We reiterated this process five times to test each fold once. A cross-validation was performed for each parameter combination as defined in the grid. The combination of cost and gamma parameters returning the highest sensitivity and specificity values along the cross-validation was considered as optimal. Therefore, a new SVM model was trained using the five folds and the best parameters, and the robustness was evaluated using the test set. The predictive power of the model on the test set was estimated using the sensitivity and the specificity that represent the rate of correctly predicted active and inactive pairs respectively.

In addition, we performed an external validation to estimate the predictive power (performance) of our model on new data not included in the training set. For this exercise, we used the protein kinase subset from the DUD-E database. We conserved 21 out of 26 protein kinases due to the absence of DFG-in structures for five of them (Supplementary Table 1). We investigated if our model was able to correctly classify the actives and the decoys.

RESULTS AND DISCUSSION

Protein kinase descriptor validation

We introduced a 3D protein kinase descriptor based on 16 residues located in the protein active site. We measured the residue pairwise distances and described each amino acid letter using five z-scale descriptors.

In order to evaluate the ability of our new descriptor to depict the residue properties of the kinase domains, we used a machine learning-based approach and performed LOO cross-validation on 215 protein kinases (each protein kinase was unified). We started by testing the ability of a supervised model to correctly assign kinase group to novel kinases not used in the training set. Therefore, we used all but one kinase as training set and build models using three different machine learning algorithms: NB, RF and SVM. The performance of each model was estimated by predicting the kinase group of the remaining protein kinase (Figure 3).

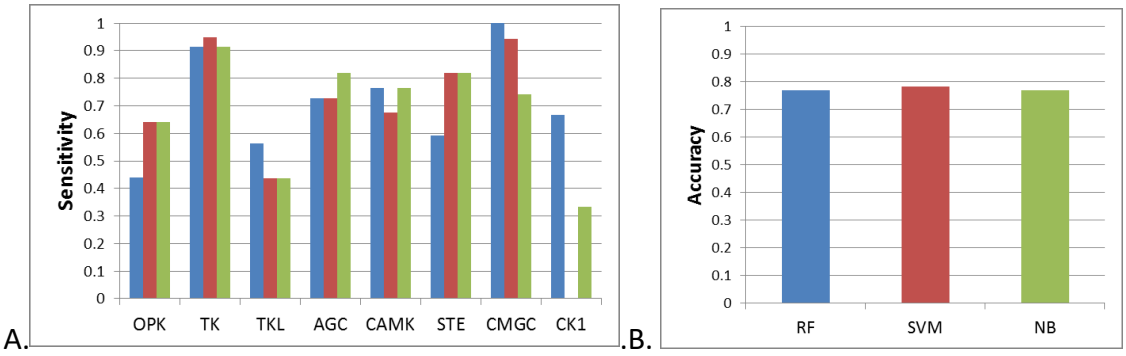


Figure 3: A/ Sensitivity obtained for LOO cross-validation to predict the protein kinase groups. B/ Overall accuracy obtained for LOO cross-validation. Blue, red and green bars correspond to RF, SVM and NB respectively.

Regarding the sensitivity of the models (Figure 3 A), we noted similar performances for the three algorithms for the TK (90%), AGC (72 to 82%) and CAMK (70%) groups. RF outperformed SVM and NB for the prediction of the TKL and CK1 groups, but failed retrieving more OPK proteins than RF and NB. RF and NB presented similar results for the prediction of OPK, TK, TKL, AGC, CAMK and STE proteins. However, SVM outperformed NB for the CMGC proteins, but at the opposite, NB outperformed SVM for CK1 group. The global lake of performance of the algorithms to predict TKL and CK1 is likely due to the few number of proteins belonging to both groups among the 215 protein kinases, with only sixteen and three

proteins respectively. However, while no statistical algorithm is performant on all kinase groups, clearly NB, RF and SVM can predict kinase group with remarkably good sensitivity by using 2D and 3D components for the proteins. We also measured the overall accuracy of the three algorithms for all combined kinase groups (Figure 3 B). The three algorithms showed good performances with an overall accuracy around 80%. These results indicate that the protein kinase descriptor contains enough information to model the biological space of the protein kinase – ligand interactions.

Additionally, we also performed a LOO cross-validation to determine if the applied protein descriptor could be used to discriminate DFG-in/DFG-out conformations. This is an important information for the identification of Type-I and Type-II kinase inhibitors binding to the DFG-in and DFG-out conformations respectively. However, if a crystal structure of a protein contains both DFG states (such as ABL1 or MAP2K1) we kept the two conformations and obtained a total of 244 proteins. We tested the same three machine learning algorithms (NB, RF, and SVM) and calculated the sensitivity and the accuracy.

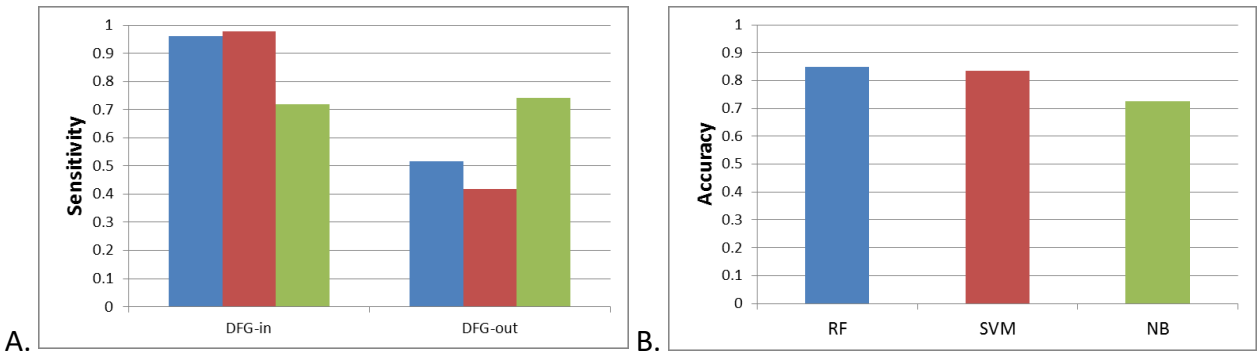


Figure 4: A/ Sensitivity obtained for LOO cross-validation to predict the protein kinase conformation. B/ Overall accuracy obtained for LOO cross-validation. Blue, red and green bars correspond for RF, SVM and NB respectively.

The calculation of the sensitivity showed very good performance for RF and SVM to classify the DFG-in conformations in the DFG-in class (both > 90%) (Figure 4 A). However, for the DFG-out conformation we observed a drop in performance for both algorithms (RF: 50%; SVM: 42%). These results suggest that around 50% of the DFG-out conformation was predicted as DFG-in. On the 244 proteins used in the LOO cross-validation, only 25% corresponds to a DFG-out conformation. This discrepancy may lead the algorithms to predict a large majority of the tested protein descriptors into the DFG-in class. It is also possible that the 3D descriptors used to describe protein conformations do not correctly take into account this

conformation and further descriptors might be needed. Interestingly, for the NB algorithm, it seems it is not influenced by this difference. Indeed, even if the sensitivity for the DFG-in class is lower compared to RF and SVM (74%), we obtained approximatively the same value for the DFG-out conformation (73 %). Finally, the scores of accuracy for RF and SVM, are partially influenced by the high sensitivity (Figure 4 B). NB obtained a lower accuracy, but still acceptable.

Hence, the protein kinase descriptor presented here has shown positive results. The good performance to predict the kinase group, combined with rather good ability to differentiate DFG-in from DFG-out conformation, comforted us to use the 2D and 3D protein descriptors combined with the 2D ligand descriptor into a 2.5D kinochemometrics approach.

Predictive results

We built 2.5D KCM models with the Ambit dataset, using the new 3D protein kinase descriptor in combination with a 2D molecular fingerprint for the ligands. In parallel, we used a more familiar aligned-based 2D descriptor using only z-scale descriptors, resulting in a 2D KCM modeling, to compare the influence of the different protein kinase descriptor on the SVM models.

2.5D KCM

We split the prepared dataset into a training set (70% of the pairs) and a test set (the remaining 30%). For the need of the classification, we stated that an “active” class is associated to pKd equals or greater than 6 to match the definition of an active in the DUD-E database. We built the first SVM model (Model 1) on the training set and by performing a nested five folds cross-validation and grid search, we defined the optimized cost and gamma parameters. The trained model was then used to predict the remaining 30% pairs. The results obtained in this internal validation are presented in Table 1. The sensitivity and the specificity are very encouraging with respectively 60% and 95%, but they underline a trend to better predict the “inactives” with respect to the actives pairs. This observation is also confirmed with the external validation for which we predicted the biological data of compounds from the DUD-E protein kinase set (Table 1). The statistical model was able to correctly predict only 19% of the “active” class, but successively retrieved 95% of the decoys corresponding to the “inactive” class. These results demonstrated the tendency of our model to better predict inactive compounds.

		Model 1[a]	Model 2[b]	Model3[c]
Internal validation	Sensitivity	0.60	0.74	0.73
	Specificity	0.95	0.95	0.94
External validation	Sensitivity	0.19	0.25	0.61
	Specificity	0.95	0.92	0.50

Table 1: Statistical values for different 2.5D KCM models built with SVM: (a) empty data have been replaced with a pKd value of 5; (b) pairs with pKd comprised between 5 and 6 were removed; (c) missing values were removed. Internal and external validations were both assessed by calculating the sensitivity and the specificity.

With an active-inactive threshold set to 6, we aimed to modulate the results of Model 1 and to avoid a threshold bias (and experimental errors), by removing interaction pairs comprised in an interval from 5 to 6 pKd. We built a second SVM model using the same protocol (Model 2) and the same training set/test set partition. We observed the sensitivity of interval validation increased to reach 74%, with a specificity staying at 95%. These good results demonstrated the positive effect of removing pairs around the defined activity threshold. However, the results obtained in the external validation still show the lack of performance to identify the "active" class of protein kinase – ligand pairs from the DUD-E, with a sensitivity of 25%, while the specificity stays high (92%). Although the values returned in the internal and the external validation of the Model 2 outperformed those obtained for the Model 1, the trend of our model to predict many "active" class as "inactive", especially for the DUD-E database, stays challenging. However, the high specificity in the internal and external validation demonstrates that our models correctly predict the large majority of inactive pairs, and therefore few inactive compounds were predicted active. This gives a good indication on how the prediction of actives by our model is reliable, a parameter highly important in *in silico* prediction.

Figure 2 shows that the pairs associated to a pKd of 5 represent 72% of the full matrix of interaction pairs. As mentioned in the methods section, these values were included in order to entirely complete our data matrix. Such method was already tested for developing regression models focused on protein kinases and the authors presented good results (Q^2 of 0.73). (*14*) Even if we applied a different dataset in our study, our training set presents the same active/inactive ratio. Such difference in performance may be due to the novel 3D protein descriptor. In their study, the authors have chosen only a 2D representation of protein kinases, with the calculation of z-scales descriptors based on protein sequences. In our case, we also incorporated such descriptors in combination with the pairwise distances to add 3D information, but the selected 16

residues may probably not be enough. However, it is important to note that the statistical models described by Lapins *et al.* were unfortunately not validated on external dataset such as the DUD-E database.

To characterize the effect of the large proportion of inactive pairs, we built a third model (Model 3). We used the same descriptors, but reduced the data by keeping only the experimentally determined Kd, *i.e.* without replacing undetected interactions by a constant value of 5. In a way, we equilibrated the active/inactive with an almost 50/50 distribution. The internal validation returned similar sensitivity and specificity compared to the previous model (Table 1). However, the results from the external validation set are significantly worst with the specificity dropping to only 50%, while the sensitivity reaching 61%. This clearly suggests that the predictive power of the models is strongly correlated to the active/inactive distribution of the training set. The value obtained for the sensitivity is particularly interesting and indicates that the model would be able to find more than 60% of the active pairs. However, the low specificity value leads to a high number of inactive pairs wrongly predicted as actives. Finally, between the three models, the second one would be the more helpful to find new protein kinase inhibitor, despite its statistical limitations.

2D KCM

To investigate if the 3D descriptor of the protein kinase could be responsible of the lack of sensitivity of our 2.5D KCM models, we built a new model using only 2D representations of both proteins and ligands. The same protocol as for the 2.5 KCM modelling was applied. The first model (Model 4) was trained on the Ambit dataset in which every missing value was replaced by a pKd value of 5. The inner validation returned results slightly better compared to 2.5D KCM, with a sensitivity of 67% (against 60%), but the specificity is identical (95%) (Table 2). The external validation showed that the model suffers from the same trend, *i.e.* it detects almost perfectly the "inactive" class but on the opposite, retrieves only a quarter of the "active" class. The model 5 confirmed our observations, and despite very good metrics in the internal validation (sensitivity: 77%; specificity: 96%), the performances stays low for the external validation (sensitivity: 29%; specificity: 89%) (Table 2). Finally, the internal validation of the Model 6, for which we did not replace the undetected interactions (empty values), returned a significant lower specificity (63%) compared to the Models 4 (95%) and 5 (96%), but still a good sensitivity (76%). In the external validation

we observed the same trend that in the Model 3. The model is not able to correctly discriminate the "active" from the "inactive" class.

		Model 4[a]	Model 5[b]	Model 6[c]
Internal validation	Sensitivity	0.67	0.77	0.76
	Specificity	0.95	0.96	0.63
External validation	Sensitivity	0.24	0.29	0.55
	Specificity	0.95	0.89	0.58

Table 2: Statistical values for different 2D KCM models built with SVM: (a) empty data have been replaced with a pKd value of 5; (b) pairs with pKd comprised between 5 and 6 were removed; (c) missing values were removed. Internal and external validations were both assessed by calculating the sensitivity and the specificity.

CONCLUSION

We introduced a new protein kinase descriptor based on the physicochemical properties of 16 residues of the active site and their pairwise distances. Despite our descriptor validation suggested a depiction level enough to correctly classify the protein kinase in their group or to discriminate DFG-in from DFG-out conformations, when we used it in our KCM approach we realized it could not be enough to detect all the inhibitors that might be present in a future dataset. Moreover, the results were strongly correlated to active/inactive pair repartition of the training set. Distributing evenly the two classes helped in increasing the sensitivity of our model, nevertheless it also involved a fall of the specificity that could lead to many false positive predictions. These performances are very close to those of the full 2D KCM models. This encourages us in the idea that the 3D protein kinase descriptor is not entirely at the origin of these disappointing results. This also suggests the difficulties of the models may come from the DUD-E database which could be particularly difficult to model. It is likely that an interaction descriptor as we first imagined, could help in detecting the specific features of this dataset.

ACKOWLEDGMENTS

N Bosc and P Bonnet are grateful to the Région Centre Val de Loire and Janssen-Cilag for financial supports. The authors would like to acknowledge Lionel Martin for his help on the SVM models.

REFERENCES

1. Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002,** *298* (5600), 1912-1934.

2. Lapenna, S.; Giordano, A. Cell cycle kinases as therapeutic targets for cancer. *Nat. Rev. Drug Discov.* **2009,** *8,* 547-566.

3. Saltiel, A. R.; Kahn, C. R. Insulin signalling and the regulation of glucose and lipid metabolism. *Nature* **2001,** *414,* 799-806.

4. Chiba, T.; Yamada, M.; Aiso, S. Targeting the JAK2/STAT3 axis in Alzheimer's disease. *Expert Opin. Ther. Targets* **2009,** *13,* 1155-1167.

5. Cohen, P. Protein kinases--the major drug targets of the twenty-first century? *Nat. Rev. Drug Discov.* **2002,** *1* (4), 309-315.

6. Anastassiadis, T.; Deacon, S. W.; Devarajan, K.; Ma, H.; Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011,** *29,* 1039-1045.

7. Treiber, D.; Shah, N. Ins and Outs of Kinase DFG Motifs. *Chem. Biol.* **2013,** *20* (6), 745-746.

8. Bosc, N.; Wroblowski, B.; Aci-Sèche, S.; Meyer, C.; Bonnet, P. A Proteometric Analysis of Human Kinome: Insight into Discriminant Conformation-dependent Residues. *ACS Chem Biol* **2015,** just accepted.

9. Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation. *J. Med. Chem.* **2010,** *53,* 2681-2694.

10. Davis, M. I.; Hunt, J. P.; Herrgard, S.; Ciceri, P.; Wodicka, L. M.; Pallares, G.; Hocker, M.; Treiber, D. K.; Zarrinkar, P. P. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **2011,** *29,* 1046-1051.

11. Zhao, Z.; Wu, H.; Wang, L.; Liu, Y.; Knapp, S.; Liu, Q.; Gray, N. S. Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? *ACS Chem. Biol.*

2014, *9,* 1230-1241.

12. Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. S. Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions. *Bioinformatics* **2005,** *21,* 4289-4296.

13. Cortés-Ciriano, I.; Ain, Q. U.; Subramanian, V.; Lenselink, B. E.; Méndez-Lucio, O.; IJzerman, A. P.; Wohlfahrt, G.; Prusis, P.; Malliavin, T. E.; van Westen, G. J. P.; Bender, A. Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. *MedChemComm* **2015,** *6,* 24-50.

14. Lapins, M.; Wikberg, J. E. S. Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. *BMC Bioinf.* **2010,** *11,* 339.

15. Fernandez, M.; Ahmad, S.; Sarai, A. Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines. *J. Chem. Inf. Model.* **2010,** *50,* 1179-1188.

16. Cao, D.-H.; Zhou, G.-H.; Liu, S.; Zhang, L.-X.; Xu, Q.-S.; He, M.; Liang, Y.-Z. Large-scale prediction of human kinase–inhibitor interactions using protein sequences and molecular topological structures. *Anal. Chim. Acta* **2013,** *792,* 10-18.

17. Subramanian, V.; Prusis, P.; Pietilä, L.-O.; Xhaard, H.; Wohlfahrt, G. Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. *J. Chem. Inf. Model.* **2013,** *53,* 3021-3030.

18. Mysinger, M. M.; Carchia, M.; Irwin, J. J.; Shoichet, B. K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012,** *55,* 6582-6594.

19. Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005,** *45,* 177-182.

20. Kufareva, I.; Abagyan, R. Type-II Kinase Inhibitor Docking, Screening, and Profiling Using Modified Structures of Active Kinase States. *J. Med. Chem.* **2008,** *51* (24), 7921-7932.

21. Bonnet, P.; Mucs, D.; Bryce, R. A. Targeting the inactive conformation of protein kinases: computational screening based on ligand conformation. *MedChemComm* **2012,** *3* (4), 434-440.

22. Chemical Computing Group Inc. *Molecular Operating Environment (MOE), 2013.08;* 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, QC, Canada, 2014.

23. Martin, E.; Mukherjee, P. Kinase-Kernel Models: Accurate In silico Screening of 4 Million Compounds Across the Entire Human Kinome. *J. Chem. Inf. Model.* **2012,** *52,* 156-170.

24. Arora, R.; Di Michele, M.; Stes, E.; Vandermarliere, E.; Martens, L.; Gevaert, K.; Van Heerde, E.; Linders, J. T.; Brehmer, D.; Jacoby, E.; Bonnet, P. Structural investigation of B-Raf paradox breaker and inducer inhibitors. *J. Med. Chem.* **2015,** *58,* 1818-1831.

25. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. *J. Med. Chem.* **1998,** *41* (14), 2481-2491.

26. Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization;* Springer, 2007; pp 319-326.

27. Chen, M. J. KinBase. http://www.kinase.com/web/current/kinbase/ (accessed 2014).

28. RDKit, Open-Source Cheminformatics. http://www.rdkit.org.

29. Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. *J. Chem. Inf. Model.* **2011,** *51,* 1942-1956.

30. Cortes-Ciriano, I.; van Westen, G. J. P.; Lenselink, E. B.; Murell, D. S.; Bender, A.; Malliavin, T. Proteochemometric modeling in a Bayesian framework. *J. Cheminf.* **2014,** *6,* 35.

31. Chupakhin, V.; Marcou, G.; Gaspar, H.; A., V. Simple Ligand–Receptor Interaction Descriptor (SILIRID) for alignment-free binding site comparison. *Comput. Struct. Biotechnol. J.* **2014,** *10,* 33-37.

32. R Core Team. *R: A Language and Environment for Statistical Computing;* R Foundation for Statistical Computing: Vienna, Austria, 2014.

33. Murrell, D. S.; Cortes-Ciriano, I.; van Westen, G. J. P.; Stott, I. P.; Bender, A.; Malliavin, T. E.; Glen, R. C. Chemically Aware Model Builder (camb): an R package for property and bioactivity modelling of small molecules. *J. Cheminf.* **2015,** *7,* 45.

34. Kuhn, M.; Wing, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kendel, B. caret: Classification and Regression Training. **2015**.

SUPPLEMENTARY INFORMATION

Table S1: Details of the external validation set obtained from the DUD-E database. (*18*)

	Actives	Decoys
ABL1	295	10885
AKT1	423	16576
AKT2	190	6952
BRAF	251	10098
CDK2	798	28328
CSF1R	286	12433
EGFR	832	35442
FGFR1	242	494
IGF1R	226	9407
KDR	620	25279
KIT	252	10609
LCK	683	27856
MAP2K1	242	8241
MAPK10	186	6714
MAPK14	915	36432
MAPKAPK2	206	6244
MET	244	11433
PLK1	155	6879
PRKCB	248	8844
ROCK1	203	6377
SRC	831	34959

Table S1

B. La limitation des modèles KCM développés

Les exemples présentés dans la partie précédente montrent toute la difficulté que nous avons eu à développer des modèles KCM prédictifs. Le jeu de données publié par Ambit peut être à l'origine du problème rencontré lors de la validation externe.[75] En effet, la présence de seulement 72 molécules dans le jeu de données peut paraître insuffisante au regard de la base de données de la DUD-E que nous avons cherché à prédire.[158] Nous avons donc voulu élaborer des modèles KCM basés sur un jeu de données plus large et plus divers.

1. Le jeu de données de Janssen

Dans le cadre de la collaboration avec l'entreprise Janssen, nous avons eu accès à une partie non négligeable des données privées de l'entreprise. Celles-ci correspondent à des données biologiques d'affinité. Nous avons remarqué que parmi elles, de nombreuses valeurs n'avaient pas été déterminées avec précision et étaient associées à des symboles d'approximation (>, <, ~). Cela n'a rien de surprenant dans un tel jeu de données puisqu'il est évident qu'il n'est pas nécessaire de chercher à déterminer avec précision le Kd d'une molécule sur une cible si aucune activité n'a été détectée lors du criblage primaire. Au final, après nettoyage des données, ce sont tout de même plus 1,5 million de Kd déterminé avec précision qui a pu être extrait et utilisé dans plusieurs modèles. Cela représente les interactions biologiques de 4735 molécules mesurées sur 456 protéines, de manière non exhaustive, *i.e.* la matrice correspondante n'est pas entièrement pleine. Après n'avoir retenu que les protéines kinases pour lesquelles nous avions un descripteur 3D, puis retiré toutes les affinités comprises entre des pKd de 5 et 6 (les résultats obtenus sur le jeu de données Ambit montraient de meilleurs résultats ainsi (partie V.A), nous avons retenu 166 protéines kinases et 1105 ligands, pour un total de 166921 paires (indice de remplissage de 91%). Plusieurs modèles ont été générés à partir de ces données.

Dans un second temps, afin d'améliorer les performances de nos modèles, nous avons procédé à une étape de nettoyage sur les données, afin de retirer les données n'apportant que peu d'information, notamment sur la sélectivité des molécules. C'est pour cela qu'ont été retirées toutes les molécules dont aucun des Kd n'était associé à une valeur inférieure ou égale à 1 µM. De même, les protéines kinases présentant plus de 99% de Kd supérieurs ou égales à 10 µM n'ont pas été gardées. Enfin, afin de garder une certaine homogénéité au niveau du nombre de molécules associées à chaque protéine du panel, nous avons fait le choix de retirer toutes les molécules qui avaient été testées sur moins de 300 protéines kinases. Après description des protéines restantes, à l'aide du descripteur 3D de protéines, c'est finalement

[158] Mysinger, M. M. *et al.* (2012), Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, *J. Med. Chem.*, 55, 6582-94.

86588 paires (indice de remplissage de 94%) qui ont été utilisées afin d'établir des modèles KCM, après transformation des Kd en pKd. Au sein de ces données, constituées de 126 protéines kinases pour 734 molécules, une proportion de 97% correspond à des paires appartenant à la classe des « inactives » (pKd < 1 µM).

2. Résultats des modèles KCM développés à partir des données de Janssen

Pour les détails concernant la génération des modèles KCM, le lecteur peut se référer à la publication ci-dessus (V.A). Le jeu de données a été subdivisé en un jeu d'apprentissage et un jeu de test, en suivant un ratio 70%/30%. Dans le cadre de la classification que nous avons effectuée, la classe « active » définit les paires protéines kinases – ligands associées à une valeur de pKd supérieure à 6. En utilisant le jeu d'apprentissage, les paramètres optimaux (gamma et cost) de la méthode SVM ont été calculés à l'aide d'une validation croisée, combinée à une recherche selon une grille de valeurs. Une fois ces paramètres clairement identifiés, un nouveau modèle SVM est généré toujours à l'aide du jeu d'apprentissage. Une validation interne est alors effectuée en prédisant les affinités du jeu de test (Table 17).

		Modèle 7[a]	Modèle 8[b]	Modèle 9[c]
Validation interne	Sensibilité	0.50	0.50	0.53
	Spécificité	0.99	0.99	0.99
Validation externe	Sensibilité	0.09	0.10	0.16
	Spécificité	0.98	0.99	0.97

Table 17 : Résultats obtenus pour différents modèles KCM 2D générés avec la méthode SVM. Dans les trois modèles, les paires avec un pKd compris entre 5 et 6 ont été retirées. (a) La validation externe est faite en prédisant les données de la DUD-E ; (b-c) La validation externe est faite en prédisant les données d'Ambit ; (c) Les molécules ne présentant aucun pKd supérieur à 6 ont été retirées, les protéines pour lesquelles 99% des pKd associés est inférieur ou égale à 5 ont été supprimées, ainsi que les molécules testées sur moins de 300 protéines kinases.

Pour les modèles générés, les résultats sont moins bons à ceux obtenus à partir du jeu de données Ambit (Modèles 1, 2 et 3). La sensibilité peine à dépasser les 50%, malgré nos efforts pour nettoyer le jeu de données (Modèle 9). En revanche, la spécificité est à un niveau proche de la perfection avec, pour les trois modèles, plus de 99% des « inactive » retrouvés. La méthode SVM semble donc ne pas être en mesure de correctement détecter les particularités des paires « active », ce qui se traduit par 50% d'entre elles classées dans la mauvaise catégorie. Bien entendu, le jeu de données est plus conséquent que celui d'Ambit, avec quinze fois plus de molécules testées dans les modèles 7 et 8, et dix fois plus dans le modèle 9. Cela entraîne une proportion de paires inactives très grande (97%).

Malgré ces résultats peu encourageants, nous avons quand même effectué des validations externes. Les modèles 7 et 8 diffèrent seulement par le jeu de données externe employé, avec respectivement la DUD-E et le jeu d'Ambit. L'idée d'utiliser le jeu de données d'Ambit vient du fait que, tout comme celui de Janssen, ce sont des Kd qui ont été mesurées. La DUD-E, quant à elle, se base sur divers types de données pour identifier les molécules actives.[158] Les résultats de validation devraient donc être moins bons que ceux obtenus avec les données Ambit. En pratique, dans les deux cas, les valeurs de sensibilité et de spécificité montrent que les deux modèles respectifs n'arrivent pas à détecter les paires « active » (sensibilité de 9% et de 10% pour les modèles 7 et 8 respectivement). Ils ont donc tendance à prédire la grande majorité des paires dans la classe « inactive ». En somme, nous notons le même phénomène que pour les modèles basés sur le jeu de données Ambit (Modèles 1, 2 et 3). Enfin, la validation externe du modèle 9 n'a été faite qu'avec le jeu de données Ambit. En dépit d'une réduction du nombre de molécules et de protéines présentes dans le jeu d'apprentissage, ce modèle n'est guère plus à même d'identifier les paires de la classe « active ».

En résumé, il semble clair que modéliser le jeu de données de Janssen est particulièrement délicat. A la différence des modèles basés sur le jeu de données d'Ambit qui présentaient de bons résultats lors de la validation interne, ces nouveaux modèles n'ont même pas cet avantage.

C. Les limites actuelles de la KCM

Les résultats obtenus en modélisant les jeux de données d'Ambit et de Janssen tendent à montrer que le nombre de molécules et de protéines présents dans le jeu d'apprentissage n'est qu'un facteur parmi tant d'autres à prendre en compte lors de la génération d'un modèle. Il est évident qu'il est nécessaire d'utiliser des descripteurs adéquats aussi bien pour les protéines kinases, que pour les molécules. Nos résultats montrent que le choix fait d'utiliser des descripteurs indépendants l'un de l'autre, choix imposé par le manque de structures cristallographiques, n'a pas permis d'atteindre les objectifs fixés, à savoir, la prédiction de nouveau inhibiteurs de protéines kinases, ni d'utiliser les résultats pour en déduire la sélectivité des ligands.

Néanmoins, le modèle 2 a été utilisé au sein de Janssen dans le but de trouver de nouveaux inhibiteurs pour des cibles données. Ce modèle étant capable de trouver quasiment tous les inactifs, tout en classant correctement une part non négligeable des actifs, il a été considéré qu'il pouvait être utile. Les résultats que nous avons prédits ont été comparés à d'autres approches *in silico* développées en interne et les résultats expérimentaux sont en cours d'acquisition.

Tous les résultats n'ont pas été présentés, mais au cours du développement de nos modèles nous avons cherché à savoir si l'utilisation d'un autre descripteur de molécules ou d'une autre méthode

d'apprentissage automatique ne pouvait pas être la solution à nos problèmes. Malheureusement ni l'utilisation d'un *fingerprint* de type MACCS, ni l'emploi de descripteurs physicochimiques n'ont amélioré les résultats. De même, encouragé par les bons résultats de la méthode *Random Forest* pour classer les structures de protéines kinases à partir du nouveau descripteur 3D, nous avons utilisé cette méthode à la place de SVM. Encore une fois, les résultats ne nous ont pas paru concluants.

Finalement, une solution pourrait être de revoir la manière de décrire les interactions entre les protéines kinases et les ligands. Plutôt que d'avoir, d'un côté, le descripteur de protéine et de l'autre, le descripteur de molécule, un descripteur d'interaction pourrait avoir plusieurs avantages. D'une part, il tiendrait compte de la conformation du ligand la plus plausible pour permettre l'interaction. D'autre part, la poche de la protéine serait précisément décrite puisque nous saurions précisément où se trouve le ligand (Type I, Type II, Type III, Type IV ?). De ce fait, la conformation de la protéine serait également connue. Néanmoins, le chemin pour pouvoir appliquer ce genre de descriptions sur tout un jeu de données biologiques de protéines kinases est encore long. Les avancées de la cristallographie aux rayons X finiront forcément par nous donner accès à toutes les structures des protéines kinases. Cependant, aurons-nous, un jour, accès à toutes les structures des complexes qu'un jeu de données peut contenir ? Rien que pour le jeu de données d'Ambit, en ne tenant compte que des interactions détectées, cela représenterait déjà 9382 structures et des ressources, aussi bien financières que techniques, incroyables à l'heure actuelle. La chimie computationnelle, et notamment les méthodes de *docking*, pourraient donc venir combler ce gouffre qui existe entre la réalité et le souhaité. En mettant de côté les résultats de *docking* qui sont parfois discutables, la problématique ici est de savoir à l'avance sur quelle conformation il est préférable de cribler la molécule. Une étude interne au groupe SB&C a montré au cours de ma thèse que les protéines kinases existent majoritairement dans trois conformations.[59] Cribler chaque molécule sur trois conformations prendrait un temps non négligeable, mais les données seraient acquises une fois pour toute. Cependant, l'obtention de ces trois conformations pour chaque protéine n'est pas chose aisé, et comme nous le soulevions dans la partie IV.A, nous ne sommes même pas sûrs que toutes les protéines kinases adoptent les conformations *DFG-in* et *DFG-out*, alors de là à obtenir expérimentalement trois conformations différentes, la route est encore longue. Au sein de l'équipe, nous avons quand même cherché à obtenir des modèles par homologie de ces trois conformations pour chaque protéine du kinome et ce travail fera bien l'objet d'un papier. Tous les éléments nécessaires à l'élaboration d'un nouveau type de descripteur pourraient donc bien être réunis aujourd'hui.

VI. Conclusions et perspectives

Avec la fin des brevets de plusieurs de leurs blockbusters certains acteurs de l'industrie pharmaceutique doivent revoir leur stratégie de découverte de médicaments. Comme nous avons pu le découvrir dans la partie II.C, beaucoup d'entre eux se tournent vers les maladies orphelines qui présentent l'intérêt de bénéficier d'un rapport bénéfice/risque plus facile à atteindre. La recherche sur les protéines kinases, dont certaines sont d'ailleurs impliquées dans des maladies orphelines, a connu un regain d'intérêt ces dernières années, bien aidée par le succès de l'inhibiteur imatinib et des ceux qui ont rapidement suivi. Une étude datant de juin dernier a recensée que 363 entreprises pharmaceutiques avaient dans leur pipeline de recherche et développement, à cette date, 609 inhibiteurs de protéines kinases impliqués dans 2525 projets autour du cancer.[159] De nouveaux inhibiteurs pourraient donc bientôt venir rejoindre les 28 médicaments déjà approuvés par la FDA (partie II.D.4), et faire des protéines kinases la famille de protéine la plus ciblée du marché.

La découverte d'inhibiteurs de protéines kinases passe encore souvent par des criblages enzymatiques ou phénotypiques. L'analyse de chaque point mesuré permet de connaitre l'activité de la molécule sur la cible en question. En revanche, c'est l'analyse de tous les points qui permet de déterminer la sélectivité de la molécule. Cette sélectivité est un paramètre primordial à prendre en compte lors du développement d'un inhibiteur de protéines kinases. L'inhibition de nombreuses protéines entraine généralement de la toxicité cellulaire et donc l'abandon de la molécule concernée. Comment, dans ce cas, distinguer les inhibiteurs sélectifs ? Afin de répondre à cette question, nous avons introduit de nouvelles métriques de sélectivité. Le *window score* et le *ranking score* sont deux métriques faciles à calculer et faciles à interpréter. Elles peuvent être appliquées sur tout type de données, allant des tests enzymatiques aux tests cellulaires.

Les modèles statistiques multidimensionnelles présentent un intérêt évident dans la recherche de nouveaux inhibiteurs de protéines kinases. Il s'agit bien entendu de diminuer les coûts humain et financier qu'ils engendrent, comparés à ceux engendrés par des études expérimentales. Utiliser en amont de la recherche et du développement, ces modèles permettent d'effectuer un premier trie des molécules et de ne finalement tester que les plus prometteuses. Les approches QSAR ont grandement participé à l'adoption de ces méthodes *in silico* au sein des entreprises et des laboratoires académiques. Cependant, à l'heure où la polypharmacologie est un thème qui prend de plus en plus d'ampleur, il devient évident qu'un modèle basé sur une seule cible n'est pas suffisant. Les approches PCM tendent à répondre à cet inconvénient en

[159] Protein Kinase Inhibitors in Oncology Drug Pipeline Update 2015. http://www.prnewswire.com/news-releases/protein-kinase-inhibitors-in-oncology-drug-pipeline-update-2015-300103468.html (accessed juillet 2015).

incluant dans ses modèles l'information des protéines et des molécules dans le but d'étudier les interactions des unes sur les autres. Les protéines kinases se prêtent bien à l'utilisation de telles approches car les 518 protéines qui forment le kinome humain présentent de grandes similarités structurales. Malgré les sources de données conséquentes générées autour de cette famille de protéine, la taille du kinome humain fait qu'encore plus d'informations sont nécessaires afin de saisir les particularités de chaque protéine et de chaque inhibiteur.

Comprendre pourquoi certaines protéines kinases n'ont encore jamais été décrites comme étant inhibées par des inhibiteurs de Type II fait partie des informations manquantes. Pour tenter d'apporter un début de réponse, nous avons introduit une approche protéométrique innovante. En étudiant les interactions de quatre inhibiteurs de Type II connus testés sur un large panel de protéines kinases, nous avons identifié les résidus qui pourraient être responsables de la liaison de ces molécules. Le résidu *gatekeeper* a été mis en évidence et sa contribution au modèle protéométrique est particulièrement importante. En effet, sa taille participe à sélectionner les inhibiteurs pouvant accéder à la poche adjacente hydrophobe des protéines kinases nécessaire au mode de liaison des inhibiteurs de Type II. En plus de ce résidu, nous en avons identifié 28 autres dont certains sont reconnus comme participant au phénomène de résistance de certaines protéines à des inhibiteurs de Type II, notamment lorsque ces résidus mutent. Le modèle protéométrique établi peut également permettre de prédire l'activité des inhibiteurs de Type II sur des protéines absentes du modèle. Cette approche est particulièrement intéressante car elle permet, en partant d'un descripteur de dimension 1D des protéines, de mettre en évidence des implications 3D des résidus de la séquence. Nous espérons qu'elle sera adoptée à d'autres familles de protéines.

Enfin, nous avons cherché à développer nos propres modèles protéochimiométriques (PCM) appliqués aux protéines kinases et nommés *kinochemome*trics (KCM) en nous basant, d'une part, sur un jeu de données publiques, d'autre part, sur un jeu de données privés fourni par l'entreprise Janssen. Afin de prendre en compte la conformation des protéines kinases dans notre étude, nous avons imaginé et implémenté un nouveau descripteur 3D. Celui-ci se base sur les propriétés physicochimiques de certains résidus du site actif des protéines kinases, et sur les distances les séparant. La validation de notre descripteur a montré qu'il pouvait permettre de distinguer les protéines kinases aussi bien en fonction de leur groupe, que de leur conformation DFG. En revanche son utilisation dans des modèles KCM ne nous a pas paru concluante, malgré des performances obtenues similaires à des modèles basés uniquement sur un descripteur 2D des protéines. En pratique le modèle basé sur le jeu de données d'Ambit est performant pour prédire des données issues de cette même source. Par contre, il tend à prédire la majorité des données externes comme étant inactives, passant ainsi à côté d'une partie non négligeable des molécules actives, qui représentent pourtant des inhibiteurs de protéines kinases. Nous pensons que l'utilisation d'un

descripteur prenant en compte directement l'interaction entre les protéines et les ligands pourraient solutionner ce problème.

VII. BIBLIOGRAPHIE

1. Bryan, C. P. Ancient Egyptian medicine: the Papyrus Ebers; Ares Publishers: Chicago, 1974.

2. Xu, J., Yang, Y. (2009) Traditional Chinese medicine in the Chinese health care system. Health Policy 90, 133-139.

3. Meyer, K. (2004) Dem Morphin auf der Spur. Pharmazeutischen Zeitung.

4. Gerhardt, C. (1853) Recherches sur les acides organiques anhydrides. Annales de Chimie et de Physique 37, 285-342.

5. DiMasi, J. A., Grabowski, H. G., Hansen, R. W. (2015) The cost of drug development. N. Engl. J. Med. 372, 1972.

6. Avorn, J. (2015) The $2.6 Billion Pill — Methodologic and Policy Considerations. N. Engl. J. Med. 372, 1877-1879.

7. The Economist. (29 novembre 2014) The price of failure.

8. Adams, C. P., Brantner, V. V. (2010) Spending on new drug development. Health Econ. 19, 130-141.

9. Abou-Gharbia, M., Childers, W. E. (2014) Discovery of Innovative Therapeutics: Today's Realities and Tomorrow's Vision. 2. Pharma's Challenges and Their Commitment to Innovation. J. Med. Chem. 57, 5525-5553.

10. FDA Public Health Advisory: Safety of Vioxx. (2014). HYPERLINK "http://www.fda.gov/Drugs/DrugSafety/PostmarketDrugSafetyInformationforPatientsandProviders/ucm106274.htm" (accessed 2015 July).

11. FDA’s consideration of evidence from certain clinical trials; United States Government Accountability Office, 2010.

12. Mullard, A. (2015) 2014 FDA drug approvals. Nat. Rev. Drug Discov. 14, 77-81.

13. Allison, M. (2012) Reinventing clinical trials. Nat. Biotechnol 30, 41-49.

14. Arrowsmith, J. (2011) Trial watch: Phase II failures: 2008–2010. Nat. Rev. Drug Discov. 10, 328-329.

15. Arrowsmith, J. (2011) Trial watch: phase III and submission failures: 2007-2010. Nat. Rev. Drug Discov. 10, 87.

16. Mu, T.-W., Ong, D. S. T., Wang, Y.-J., Balch, W. E., Yates III, J. R., Segatori, L., Kelly, J. W. (2008) Chemical and Biological Approaches Synergize to Ameliorate Synergize to Ameliorate. Cell 134, 769-781.

17. Maslov, S., Ispolatov, I. (2007) Propagation of large concentration changes in reversible protein-binding networks. PNAS 104, 13655-13660.

18. Kotz, J. (2012) Phenotypic screening, take two. SciBX 5.

19. Swinney, D. C., Anthony, J. (2011) How were new medicines discovered? Nat. Rev. Drug Discov. 10, 507-519.

20. Hopkins, A., L., Groom, C. R. (2002) The druggable genome. Nat. Rev. Drug Discov. 1, 727-730.

21. Rask-Andersen, M., Almén, M. S., B. Schiöth, H. (2011) Trends in the exploitation of novel drug targets. Nat. Rev. Drug Discov. 10, 579-590.

22. Law, V., Knox, C., Djoumbou, Y., Jewison, T., Guo, A. C., Liu, Y., Maciejewski, A., Arndt, D., Wilson, M., Neveu, V., Tang, A., Gabriel, G., Ly, C., Adamjee, S., Dame, Z. T., Han, B., Zhou, Y. &. W. D. S. (2014) DrugBank 4.0: shedding new light on drug metabolism. Nuc. Acids Res. 42, D1091-D1097.

23. Overington, J. P., Al-Lazikani, B., Hopkins, A. L. (2006) How many drug targets are there? Nat. Rev. Drug Discov. 5, 993-996.

24. Burnette, G., Kennedy, E. P. (1954) The enzymatic phosphorylation of proteins. J. Biol. Chem. 211, 969-980.

25. Fischer, E. H., Graves, D. J., Crittenden, E. R. S., Krebs, E. G. (1959) Structure of the site phosphorylated in the phosphorylase b to a reaction. J. Biol. Chem. 234, 1698-1704.

26. Walsh, D. A., Perkins, J. P., Krebs, E. G. (1968) An adenosine 3',5'-monophosphate-dependant protein kinase from rabbit skeletal muscle. J. Biol. Chem. 243, 1763-3765.

27. Cohen, P. (2002) The origins of protein phosphorylation. Nat. Cell Biol. 4, E127-E130.

28. Lapenna, S., Giordano, A. (2009) Cell cycle kinases as therapeutic targets for cancer. Nat. Rev. Drug Discov. 8, 547-566.

29. Saltiel, A. R., Kahn, C. R. (2001) Insulin signalling and the regulation of glucose and lipid metabolism. Nature 414, 799-806.

30. Chiba, T., Yamada, M., Aiso, S. (2009) Targeting the JAK2/STAT3 axis in Alzheimer's disease. Expert Opin. Ther. Targets 13, 1155-1167.

31. Hunter, T. (1987) A thousand and one protein kinases. Cell 50, 823-829.

32. Manning, G., Whyte, D. B., Martinez, R., Hunter, T., Sudarsanam, S. (2002) The protein kinase complement of the human genome. Science 298, 1912-1934.

33. Braconi-Quintaje, S., Orchard, S. (2008) The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot. Mol. Cell. Proteomics 7, 1409-1419.

34. Filippakopoulos, P., Müller, S., Knapp, S. (2009) SH2 domains: modulators of nonreceptor tyrosine kinase activity. Curr. Opin. Struct. Biol. 19, 643-649.

35. Mayer, B. J., Hirai, H., Sakai, R. (1995) Evidence that SH2 domains promote processive phosphorylation by protein-tyrosine kinases. Curr. Biol. 5, 296-305.

36. Kuriyan, J., Eisenberg, D. (2007) The origin of protein interactions and allostery in colocalization. Nature 450, 983-990.

37. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., Bourne, P. E. (2000) The Protein Data Bank. Nucl. Acids Res. 28, 235-242.

38. Knighton, D. R., Zheng, J. H., Eyck, L. F. T., Ashford, V. A., Xuong, N. H., Taylor, S. S., Sowadski, J. M. (1991) Crystal structure of the catalytic subunit of cyclic adenosine monophosphate-dependent protein kinase. Science 253, 404-414.

39. Zheng, J., Trafny, E. A., Knighton, D. R., Xuong, N. H., Taylor, S. S., Ten Eyck, L. F., Sowadski, J. M. (1993) 2.2 A refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with MnATP and a peptide inhibitor. Acta Crystallogr. 49, 362-365.

40. Hubbard, S. R., Wei, L., Ellis, L., Hendrickson, W. A. (1994) Crystal structure of the tyrosine kinase domain of the human insulin receptor. Nature 372, 746-754.

41. Xu, R. M., Carmel, G., Sweet, R. M., Kuret, J., Cheng, X. (1995) Crystal structure of casein kinase-1, a phosphate-directed protein kinase. EMBO J. 14, 1015-1023.

42. Cowan-Jacob, S. W. (2006) Structural biology of protein tyrosine kinases. Cell. Mol. Life Sci. 63, 2608-2625.

43. Nolen, B., Taylor, S., Ghosh, G. (2004) Regulation of protein kinases; controlling activity through activation segment conformation. Mol. Cell 15, 661-675.

44. Taylor, S. S., Kornev, A. P. (2011) Protein kinases: evolution of dynamic regulatory proteins. Trends Biochem. Sci. 36, 65-77.

45. Fabbro, D., Cowan-Jacob, S. W., Moebitz, H. (2015) Ten things you should knowabout protein kinases: IUPHAR Review 14. Br. J. Pharmacol. 172, 2675-2700.

46. Ubersax, J. A., Ferrell, J. E. (2007) Mechanisms of specificity in protein phosphorylation. Nat. Rev. Mol. Cell Biol. 8, 530-541.

47. Tanramluk, D., Schreyer, A., Pitt, W. R., Blundell, T. L. (2009) On the Origins of Enzyme Inhibitor Selectivity and Promiscuity: A Case Study of Protein Kinase Binding to Staurosporine. Chem. Biol. Drug Des. 74, 16-24.

48. Blencke, S., Zech, B., Engkvist, O., Greff, Z., Õrfi, L., Horvàth, Z., Kéri, G., Ullrich, A., Daub, H. (2004) Characterization of a Conserved Structural Determinant Controlling Protein Kinase Sensitivity to Selective Inhibitors. Chem. Biol. 11, 691-701.

49. Kornev, A. P., Taylor, S. S., Ten Eyck, L. F. (2008) A helix scaffold for the assembly of active protein kinases. Proc. Natl. Acad. Sci. U. S. A. 105, 14377-14382.

50. Xu, B., English, J. M., Wilsbacher, J. L., S., S., Goldsmith, E. J., M.H., C. (200) WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II. J. Biol. Chem. 275, 16795-16801.

51. Adams, J. A. (2001) Kinetic and Catalytic Mechanisms of Protein Kinases. Chem. Rev. 101, 2271-2290.

52. Hubbard, S. R. (1997) Crystal structure of the activated insulin receptor tyrosine kinase in complex with peptide substrate and ATP analog. EMBO J. 16, 5572-5581.

53. Ogawa, A., Takayama, Y., Sakai, H., Chong, K. T., Takeuchi, S., Nakagawa, A., Nada, S., Okada, M., Tsukihara, T. (2002) Structure of the Carboxyl-terminal Src Kinase, Csk. J. Biol. Chem. 277, 14351-14354.

54. Nagar, B., Hantschel, O., Seeliger, M., Davies, J. M. . W. W. I. . S.-F. G., Kuriyan, J. (2006) Organization of the SH3–SH2 unit in active and inactive forms of the c-Abl tyrosine kinase. Mol. Cell 21, 787-798.

55. Wybenga-Groot, L. E., Baskin, B., Ong, S. H., Tong, J., Pawson, T., Sicheri, F. (2001) Structural basis for autoinhibition of the Ephb2 receptor tyrosine kinase by the unphosphorylated juxtamembrane region. Cell 106, 745-757.

56. Huse, M., Kuriyan, J. (2002) The conformational plasticity of protein kinases. Cell 109, 275-282.

57. Schulze-Gahmen, U., De Bondt, H. L., Kim, S. H. (1996) High-resolution crystal structures of human cyclin-dependent kinase 2 with and without ATP: bound waters and natural ligand as guides for inhibitor design. J. Med. Chem. 39, 4540-4546.

58. Xu, W., Harrison, S. C., Eck, M. J. (1997) Three-dimensional structure of the tyrosine kinase c-Src. Nature 385, 595-602.

59. Golib-Dzib, J. F., Arora, R., Bonnet, P. Towards an objective criterion for the determination of protein kinase conformations. Submitted.

60. Huang, M., Shen, A., Ding, J., Geng, M. (2014) Molecularly targeted cancer therapy: some lessons from the past decade. Trends Pharmacol. Sci. 35, 41-50.

61. Ma, W. W., Adjei, A. A. (2009) Novel Agents on the Horizon for Cancer Therapy. CA Cancer J. Clin. 111-137, 59.

62. Clark, J. D., Flanagan, M. E., Telliez, J.-B. (2014) Discovery and Development of Janus Kinase (JAK) Inhibitors for Inflammatory Diseases. J. Med. Chem. 57, 5023-5038.

63. Muth, F., Günther, M., Bauer, S. M., Döring, E., Fischer, S., Maier, J., P., D., Köppler, J., Trappe, J., Rothbauer, U., Koch, P., Laufer, S. A. (2015) Tetra-substituted pyridinylimidazoles as dual inhibitors of p38α mitogen-activated protein kinase and c-Jun N-terminal kinase 3 for potential treatment of neurodegenerative diseases. J. Med. Chem. 58, 443-456.

64. Kikuchi, R., Nakamura, K., MacLauchlan, S., Ngo, D. T., Shimizu, I., Fuster, J. J., Katanasaka, Y., Yoshida, S., Qiu, Y., Yamaguchi, T. P., Matsushita, T., Murohara, T., Gokce, N., Bates, D. O., Hamburg, N. M., Walsh, K. (2014) An antiangiogenic isoform of VEGF-A contributes to impaired vascularization in peripheral artery disease. Nat. Med. 20, 1464-1471.

65. Banks, A. S., McAllister, F. E., P., C. J., J., Z. P., Jurczak, M. J., Laznik-Bogoslavski, D., Shulman, G. I., Gygi, S. P., Spiegelman, B. M. (2015) An ERK/Cdk5 axis controls the diabetogenic actions of PPARγ. Nature 517, 391-395.

66. Rask-Andersen, M., Zhang, J., Fabbro, D., B., S. H. (2014) Advances in kinase targeting: current clinical use and clinical trials. Trands Pharmacol. Sci. 35, 606-620.

67. Druker, B. J., Talpaz, M., Resta, D. J., Peng, B., Buchdunger, E., Ford, J. M., Lydon, N. B., Kantarjian, H., Capdeville, R., Ohno-Jones, S., Sawyers, C. L. (2001) Efficacy and safety of a specific inhibitor of the BCR-ABL tyrosine kinase in chronic myeloid leukemia. N. Engl. J. Med. 344, 1031-1037.

68. Knight, Z. A., Shokat, K. M. (2005) Features of Selective Kinase Inhibitors. Chem. Biol. 12, 621-637.

69. Wu, P., Nielsen, T. E., Clausen, M. H. (2015) FDA-approved small-molecule kinase inhibitors. Trends Pharmacol. Sci. 36, 422-439.

70. Zuccotto, F., Ardini, E., Casale, E., Angiolini, M. (2010) Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation. J. Med. Chem. 53, 2681-2694.

71. Hilberg, F., Roth, G. J., Krssak, M., Kautschitsch, S., Sommergruber, W., Tontsch-Grunt, U., Garin-Chesa, P., Bader, G., Zoephel, A., Quant, J., Heckel, A., Rettig, W. J. (2008) Structure of VEGFR2 kinase domain in complex with BIBF1120. cancer Res. 68, 4774-4782.

72. Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B., Kuriyan, J. (2000) Structural mechanism for STI-571 inhibition of abelson tyrosine kinase. Science 289, 1938-1942.

73. Capdeville, R., Buchdunger, E., Zimmermann, J., Matter, A. (2002) Glivec (STI571, imatinib), a rationally developed, targeted anticancer drug. Nat. Rev. Drug Discov. 1, 493-502.

74. Mol, C. D., Fabbro, D., Hosfield, D. J. (2004) Structural insights into the conformational selectivity of STI-571 and related kinase inhibitors. Curr. Opin. Drug Discovery Dev. 7, 639-648.

75. Davis, M. I., Hunt, J. P., Herrgard, S., Ciceri, P., Wodicka, L. M., Pallares, G., Hocker, M., Treiber, D. K., Zarrinkar, P. P. (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1046-1051.

76. Zhao, Z., Wu, H., Wang, L., Liu, Y., Knapp, S., Liu, Q., Gray, N. S. (2014) Exploration of Type II Binding Mode: A Privileged Approach for Kinase Inhibitor Focused Drug Discovery? ACS Chem. Biol. 9, 1230-1241.

77. Bogoyevitch, M. A., Fairlie, D. P. (2007) A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. Drug Discov Today 12, 622-633.

78. Kirkland, L. O., McInnes, C. (2009) Non-ATP competitive protein kinase inhibitors as anti-tumor therapeutics. Biochem. Pharmacol. 77, 1561-1571.

79. Gavrin, L. K., Saiah, E. (2013) Approaches to discover non-ATP site kinase inhibitors. MedChemComm 4, 41-51.

80. Dong, Q., Dougan, D. R., Gong, X., Halkowycz, P., Jin, B., Kanouni, T., O'Connell, S. M., Scorah, N., Shi, L., Wallace, M. B., Zhou, F. (2011) Crystal Structure of the Human Mitogen-activated protein kinase kinase 1 (MEK 1) in complex with ligand and MgATP. Bioorg. Med. Chem. Lett. 21, 1315-1319.

81. Lamba, V., Ghosh, I. (2012) New Directions in Targeting Protein Kinases: Focusing Upon True Allosteric and Bivalent Inhibitors. Curr. Pharm. Des. 18, 2936-2945.

82. Wu, W. I., Voegtli, W. C., Sturgis, H. L., Dizon, F. P., Vigers, G. P., Brandhuber, B. J. (2010) Crystal structure of human AKT1 with an allosteric inhibitor reveals a new mode of kinase inhibition. PloS One 5, e12913.

83. Yun, C. H., Mengwasser, K. E., Toms, A. V., Woo, M. S., Greulich, H., K., W. K., Meyerson, M., Eck, M. J. (2008) The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. Proc. Natl. Acad. Sci. USA 105, 2070-2075.

84. Woyach, J. A., Furman, R. R., Liu, T.-M., Ozer, H. G., Zapatka, M., Ruppert, A. S., Xue, L., Li, D. H.-H., Steggerda, S. M., Versele, M., Zhang, J., Yilmaz, A. S., Jaglowski, S. M., Blum, K. A., Lozanski, A., Lozanski, G., James, D. F., Barrientos, J. C., Lichter, P., Stilgenbauer, S., Buggy, J. J., Chang, B. Y., Johnson, A. J., Byrd, J. C. (2014) Resistance Mechanisms for the Bruton's Tyrosine Kinase Inhibitor Ibrutinib. N. Engl. J. Med. 370, 2286-2294.

85. Zhou, T., Commodore, L., Huang, W. S., Wang, Y., Thomas, M., Keats, J., Xu, Q., Rivera, V. M., Shakespeare, W. C., Clackson, T., Dalgarno, D. C. . Z. X. (2011) Structural mechanism of the Pan-BCR-ABL inhibitor ponatinib (AP24534): lessons for overcoming kinase inhibitor resistance. Chem. Biol. Drug Des. 77, 1-11.

86. Jencks, W. P. (1981) On the attribution and additivity of binding energies. Proc. Natl. Acad. Sci. USA 78, 4046-4050.

87. Cohen, P. (2002) Protein kinases--the major drug targets of the twenty-first century? Nat. Rev. Drug Discov. 1, 309-315.

88. Mitchell, J. B. O. (2014) Machine learning methods in chemoinformatics. WIREs Comput. Mol. Sci. 4, 468-481.

89. The Uniprot Consortium. (2014) Activities at the Universal Protein Resource (UniProt). Nucleic Acids Research 42, D191-D198.

90. Baskin, I., Varnek, A. Fragment descriptors in SAR/QSAR/QSPR studies, molecular similarity analysis and in virtual screening; RCS: Cambrige, 2008.

91. Reymond, J.-L., Blum, L. C., van Deursen, R. (2011) Exploring the chemical space of known and unknown organic small molecules at www.gdb.unibe.ch. Chimia 65, 863-867.

92. Reymond, J.-L., Ruddigkeit, L., Blum, L., van Deursen, R. (2012) The enumeration of chemical space. WIREs Comput. Mol. Sci. 2, 717-733.

93. Kirkpatrick, P., Ellis, C. (2004) Chemical space. Nature 432, 823.

94. Reymond, J. L., Awale, M. (2012) Exploring chemical space for drug discovery using the chemical universe database. ACS Chem. Neurosci. 3, 649-657.

95. Martin, Y. C., Kofron, J. L., Traphagen, L. M. (2002) Do Structurally Similar Molecules Have Similar Biological Activity? J. Med. Chem. 2002, 4350-4358.

96. Bento, A. P., Gaulton, A., Hersey, A., Bellis, L. J., Chambers, J., Davies, M., Kruger, F. A., Light, Y., Mak, L., McGlinchey, S., Nowotka, M., Papadatos, G., Santos, R., Overington, J. P. (2014) The ChEMBL bioactivity database: an update. Nucleic Acids Res. 42, 1083-1090.

97. Bolton, E., Wang, Y., Thiessen, P. A., Bryant, S. H. Chapter 12 - PubChem: Integrated Platform of Small Molecules and Biological Activities; Elsevier, 2008; Vol. 4, pp 217-241.

98. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., Overington, J. P. (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40, D1100-D1107.

99. Hansch, C., Fujita, T. (1964) p-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. J. Am. Chem. Soc 86, 1616-1626.

100. Hansch, C. (1969) A quantitative approach to biochemical structure-activity relationships. Acc. Chem. Res. 2, 232-239.

101. Sahigara, F., Mansouri, K., Ballabio, D., Mauri, A., V., C., Todeschini, R. (2012) Comparison of Different Approaches to Define the Applicability Domain of QSAR Models. Molecules 17, 4791-4810.

102. Kalliokoski, T., Kramer, C., Vulpetti, A., Gedeck, P. (2013) Comparability of mixed IC50 data – A statistical analysis. PLoS One 8, e61007.

103. Sprous, D. G., Zhang, J., Zhang, L., Wang, Z., Tepper, M. A. (2006) Kinase inhibitor recognition by use of a multivariable QSAR model. J. Mol. Graph. Model. 24, 278-295.

104. Todeschini, R., Consonni, V. Handbook of molecular descriptors; Wiley-VCH, 2008.

105. Zhu, L. L., Hou, T. J., Chen, L. R., Xu, X. J. (2001) 3D QSAR analyses of novel tyrosine kinase inhibitors based on pharmacophore alignment. J Chem Inf. Comput. Sci. 41, 1032-1040.

106. Lescot, E., Bureau, R., Sopkova-de Oliveira Santos, J., Rochais, C., Lisowski, V., Lancelot, J. C., Rault, S. (2005) 3D-QSAR and docking studies of selective GSK-3beta inhibitors. Comparison with a thieno[2,3-b]pyrrolizinone derivative, a new potential lead for GSK-3beta ligands. J. Chem Inf. Model. 45, 708-715.

107. Thaimattam, R., Daga, P. R., Banerjee, R., Iqbal, J. (2005) 3D-QSAR studies on c-Src kinase inhibitors and docking analyses of a potent dual kinase inhibitor of c-Src and c-Abl kinases. Bioorg. Med. Chem. 13, 4704-4712.

108. Lather, V., Kristam, R., Saini, J. S., Kristam, R., Karthikeyan, N. A., Balaji, V. N. (2008) QSAR models for prediction of Glycogen Synthase Kinase-3b inhibitory activity of indirubin derivatives. QSAR Com. Sci. 27, 718-728.

109. Kirubakaran, P., Muthusamy, K., Singh, K. H., D., Nagamani, S. (2012) Ligand-based Pharmacophore Modeling; Atom-based 3D-QSAR Analysis and Molecular Docking Studies of Phosphoinositide-Dependent Kinase-1 Inhibitors. Indian J. Pharm. Sci. 74, 141-151.

110. Wermuth, C. G., Ganellin, C. R., Lindberg, P., Mitscher, L. A. (1998) Glossary of terms used in medicinal chemistry (IUPAC Recommendations 1998). Pure Appl. Chem. 70, 1129-1143.

111. Shi, W. M., Shen, Q., Kong, W., Ye, B. X. (2007) QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression. Eur. J. Med. Chem. 42, 81-86.

112. Comelli, N. C., Ortiz, E. V., Kolacz, M., Toropovad, A. P., Toropov, A. A., Duchowiczc, P. R. E. A. C. (2014) Conformation-independent QSAR on c-Src tyrosine kinase inhibitors. Chemom. Intell. Lab. Syst. 134, 47-52.

113. Martin, E., Mukherjee, P., Sullivan, D., Jansen, J. (2011) Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. J. Chem. Inf. Model. 51, 1942-1956.

114. Riniker, S., Landrum, G. A. (2013) Open-source platform to benchmark fingerprints for ligand-based virtual screening. J. Cheminfor. 5, 26.

115. Tropsha, A., Gramatica, P., Gombar, V. K. (2003) The Importance of Being Earnest: Validation is the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR Comb. Sci. 22, 69-77.

116. Weaver, S., Gleeson, M. P. (2008) The importance of the domain of applicability in QSAR modeling. J. Mol. Graph. Model. 26, 1315-1326.

117. Jonsson, J., Norberg, T., Carlsson, L., Gustafsson, C., Wold, S. (1993) Quantitative sequence-activity models (QSAM)--tools for sequence design. Nucleic Acids Res. 21, 733-739.

118. van Westen, G. J. P., Swier, R. F., Wegner, J. K., IJzerman, A. P., van Vlijmen, H. W., Bender, A. (2013) Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets. J. Cheminform. 5, 41.

119. Zhou, P., Chen, X., Wu, Y., Shang, Z. (2010) Gaussian process: an alternative approach for QSAM modeling of peptides. Amino Acids 38, 199-212.

120. Lapinsh, M., Prusis, P., Gutcaits, A., Lundstedt, T., Wikberg, J. E. S. (2001) Development of proteo-chemometrics: a novel technology for the analysis of drug-receptor interactions. Biochim. Biophys. Acta - Gen. Subj. 1525, 180-190.

121. Lapinsh, M., Prusis, P., Uhlén, S., Wikberg, J. E. S. (2005) Improved approach for proteochemometrics modeling: application to organic compound—amine G protein-coupled receptor interactions. Bioinformatics 21, 4289-4296.

122. Cortés-Ciriano, I., Ain, Q. U., Subramanian, V., Lenselink, B. E., Méndez-Lucio, O., IJzerman, A. P., Wohlfahrt, G., Prusis, P., Malliavin, T. E., van Westen, G. J. P., Bender, A. (2015) Polypharmacology modelling using proteochemometrics (PCM): recent methodological developments, applications to target families, and future prospects. MedChemComm 6, 24-50.

123. Jacob, L., Hoffmann, B., Stoven, V., Vert, J.-P. (2008) Virtual screening of GPCRs: An in silico chemogenomics approach. Bmc Bioinf. 9, 363.

124. Weill, N., Rognan, D. (2009) Development and validation of a novel protein−ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. J. Chem. Inf. Model. 49, 1049-1062.

125. Wu, D., Huang, Q., Zhang, Y., Zhang, Q., Liu, Q., Gao, J., Cao, Z., Zhu, R. (2012) Screening of selective histone deacetylase inhibitors by proteochemometric modeling. BMC Bioinf. 13, 212.

126. van Westen, G. J. P., Wegner, J. K., Geluykens, P., Kwanten, L., Vereycken, I., Peeters, A., Ijzerman, A. P., van Vlijmen, H. W., Bender, A. (2011) Which compound to select in lead optimization? Prospectively validated proteochemometric models guide preclinical development. PLoS One 6, e27518.

127. Cortés-Ciriano, I., Murrell, D. S., van Westen, G. J. P., Bender, A., Malliavin, T. E. (2015) Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. J. Cheminf. 7, 1.

128. Ain, Q. U., Méndez-Lucio, O., Cortés-Ciriano, I., Malliavin, T., van Westen, G. J. P., Bender, A. (2014) Modelling ligand selectivity of serine proteases using integrative proteochemometric approaches improves model performance and allows the multi-target dependent interpretation of features. Integr. Biol. 6, 1023-1033.

129. Vieth, M., Sutherland, J. J., Robertson, D. H., Campbell, R. M. (2005) Kinomics: characterizing the therapeutically validated kinase space. Drug Discov. Today 10, 839-846.

130. Fernandez, M., Ahmad, S., Sarai, A. (2010) Proteochemometric recognition of stable kinase inhibition complexes using topological autocorrelation and support vector machines. J. Chem. Inf. Model. 50, 1179-1188.

131. Lapins, M., Wikberg, J. E. S. (2010) Kinome-wide interaction modelling using alignment-based and alignment-independent approaches for kinase description and linear and non-linear data analysis techniques. BMC Bioinf. 11, 339.

132. Cao, D.-H., Zhou, G.-H., Liu, S., Zhang, L.-X., Xu, Q.-S., He, M., Liang, Y.-Z. (2013) Large-scale prediction of human kinase–inhibitor interactions using protein sequences and molecular topological structures. Anal. Chim. Acta 792, 10-18.

133. Subramanian, V., Prusis, P., Pietilä, L.-O., Xhaard, H., Wohlfahrt, G. (2013) Visually Interpretable Models of Kinase Selectivity Related Features Derived from Field-Based Proteochemometrics. J. Chem. Inf. Model. 53, 3021-3030.

134. Sheridan, R., Nam, K., Maiorov, V. N., McMasters, D. R., Cornell, W. D. (2009) QSAR Models for Predicting the Similarity in Binding Profiles for Pairs of Protein Kinases and the Variation of Models between Experimental Data Sets. J. Chem. Inf. Model. 49, 1974-1985.

135. Drewry, D., Willson, T., Zuercher, W. (2014) Seeding Collaborations to Advance Kinase Science with the GSK Published Kinase Inhibitor Set (PKIS). Curr. Top. Med. Chem. 14, 340-342.

136. Bain, J., Plater, L., Elliott, M., Shpiro, N., Hastie, C. J., McLauchlan, H., Klevernic, I., Arthur, J. S. C., Alessi, D. R., Cohen, P. (2007) The selectivity of protein kinase inhibitors: a further update. Biochem. J. 408, 297-315.

137. Fedorov, O., Marsden, B., Rellos, P., Muller, S., Bullock, A. N., Schwaller, J., Sundstrom, M., Knapp, S. (2007) A systematic interaction map of validated kinase inhibitors with Ser/Thr kinases. PNAS 104, 20523-20528.

138. Karaman, M. W., Herrgard, S., Treiber, D. K., Gallant, P., Atteridge, C. E., Campbell, B. T., Chan, K. W., Ciceri, P., Davis, M. I., Edeen, P. T., Faraoni, R., Floyd, M., Hunt, J. P., Lockhart, D. J., Milanov, Z. V., Morrison, M. J., Pallares, G., Patel, H. K., Pritchard, S., Wodicka, L. M., Zarrinkar, P. P. (2008) A quantitative analysis of kinase inhibitor selectivity. Nat. Biotechnol. 26, 127-132.

139. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H., Peterson, J. R. (2011) Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. Nat. Biotechnol. 29, 1039-1045.

140. Metz, J. T., Johnson, E. F., Soni, N. B., Merta, P. J., Kifle, L., Hajduk, P. J. (2011) Navigating the kinome. Nat. Chem. Biol. 7, 200-202.

141. Ahmad, S., Kitajima, K., Selvaraj, S., Kubodera, H., Sunada, S., An, J. (2003) Protein-Ligand Interactions: ProLINT Database and QSAR Analysis. Genome Inf. 14, 537-538.

142. Mauri, A., Consonni, V., Pavan, M., Todeschini, R. (2006) DRAGON Software: An Easy Approach to Molecular Descriptor Calculations. MATCH 56, 237-248.

143. Sandberg, M., Eriksson, L., Jonsson, J., Sjöström, M., Wold, S. (1998) New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids. J. Med. Chem. 41, 2481-2491.

144. Dubchak, I., Muchnik, I., Holbrook, S. R., Kim, S. H. (1995) Prediction of protein folding class using global description of amino acid sequences. Proc. Natl. Acad. Sci. USA 92, 8700-8704.

145. Meslamani, J., Rognan, D. (2011) Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel. J. Chem. Inf. Model. 51, 1593-1603.

146. Robinson, D. D., Sherman, W., Farid, R. (2010) Understanding kinase selectivity through energetic analysis of binding site waters. ChemMedChem 5, 618-627.

147. Wold, S., Sjöström, M., Eriksson, L. (2001) PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58, 109-130.

148. Liaw, A., Wiener, M. (202) Calssification and Regression by randomForest. R News 2/3, 18-22.

149. Ivanciuc, O. (2007) Applications to Support Vector Machines in Chemistry. Reviews in Computational Chemistry 23, 291-400.

150. Czarnecki, W. M., Podlewska, S., Bojarski, A. J. (2015) Robust optimization of SVM hyperparameters in the classification of bioactive compounds. J. Cheminf. 7, 38.

151. Cortes-Ciriano, I., van Westen, G. J. P., Lenselink, E. B., Murell, D. S., Bender, A., Malliavin, T. (2014) Proteochemometric modeling in a Bayesian framework. J. Cheminf. 6, 35.

152. Van Rompaey, L., Galien, R., van der Aar, E. M., Clement-Lacroix, P., Nelles, L., Smets, B., L., L., Christophe, T., Conrath, K., Vandeghinste, N., Vayssiere, B., De Vos, S., Fletcher, S., Brys, R., van 't Klooster, G., Feyen, J. H., Menet, C. (2013) Preclinical characterization of GLPG0634, a selective inhibitor of JAK1, for the treatment of inflammatory diseases. J. Immunol. 191, 3568-3577.

153. O'Shea, J. J., Plenge, R. (2012) JAK and STAT Signaling Molecules in Immunoregulation and Immune-Mediated Disease. Immunity 36, 542-550.

154. Gao, Y., Davies, S. P., Augustin, M., Woodward, A., Parel, U. A. . K. R., Harvey, K. J. (2013) A broad activity screen in support of a chemogenomic map for kinase signalling research and drug discovery. Biochem. J. 451, 313-328.

155. Kinase SARfari. HYPERLINK "https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari" https://www.ebi.ac.uk/chembl/sarfari/kinasesarfari (accessed 2014).

156. Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., Wiswedel, B. KNIME: The Konstanz Information Miner. Studies in Classification, Data Analysis, and Knowledge Organization; Springer, 2007; pp 319-326.

157. Treiber, D., Shah, N. (2013) Ins and Outs of Kinase DFG Motifs. Chem. Biol. 20, 745-746.

158. Mysinger, M. M., Carchia, M., Irwin, J. J., Shoichet, B. K. (2012) Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. J. Med. Chem. 55, 6582-6594.

159. Protein Kinase Inhibitors in Oncology Drug Pipeline Update 2015. HYPERLINK "http://www.prnewswire.com/news-releases/protein-kinase-inhibitors-in-oncology-drug-pipeline-update-2015-300103468.html" http://www.prnewswire.com/news-releases/protein-kinase-inhibitors-in-oncology-drug-pipeline-update-2015-300103468.html (accessed juillet 2015).

Communications scientifiques

Publications

- Bosc, N., Wroblowski, B., Aci-Sèche, S., Meyer, C. and Bonnet, P. A Proteometric Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues. *ACS Chemical Biology*.
- Bosc, N., Meyer, C. and Bonnet, P. The use of various selectivity scores in kinase research.
- Bosc, N., Meyer, C. and Bonnet, P. Prediction of protein kinase – ligand interactions using a 2.5D kinochemometrics approach.

Communications orales

- Bosc, N., Wroblowski, B., Aci-Sèche, S., Meyer, C. and Bonnet, P. A Proteometric analysis of human kinome: insight into discriminant conformation-dependent residues. *19ème congrès du Groupe de Graphisme et de Modélisation Moléculaire*. Mai 2015.
- Bosc, N., Meyer, C. and Bonnet, P. Proteometrics, chemometrics and proteochemometrics: helpful tools for kinase drug discovery. *22nd Young Research Fellows Meeting*. Février 2015.
- Bosc, N., Meyer, C. and Bonnet, P. Broad profiling prediction of protein kinase inhibitors via kinochemometrics approaches. *Colloque ITMO Bases Moléculaires et structurales du vivant*. Juin 2014.

Communications par affiches

- Pihan, E., Bosc, N., Bourg, S., Canault, B., Arora, R., Golib, F., Gally, J.-M., Aci Sèche, S. and Bonnet, P. In silico kinase platform (ISKP): a tool for the identification of selective kinase inhibitors. *7èmes journées de la Société Française de Chémoinformatique (SFCi)*. Octobre 2015.
- Bonnet, P., Bosc, N., Wroblowski, B., Aci Sèche, S. and Meyer, C. Proteometrics, chemometrics and proteochemometrics: helpful tools for kinase drug discovery. *Frontiers in Medicinal Chemistry*. Septembre 2015.
- Bosc, N. and Bonnet, P. The use of various selectivity scores in kinase research. *51th International Conference on Medicinal Chemistry (RICT 2015)*. Juillet 2015.
- Bourg, S., Pihan, E., Canault, B., Arora, R., Golib, F., Gally, J.-M., Bosc, N., Aci Sèche, S. and Bonnet, P. In silico kinase platform (ISKP): a tool for the identification of selective kinase inhibitors. *51th International Conference on Medicinal Chemistry (RICT 2015)*. Juillet 2015.
- Bosc, N., Meyer, C. and Bonnet, P. Proteometrics, chemometrics and proteochemometrics: helpful tools for kinase drug discovery. *Journée Jeunes Chercheurs de la Section Régionale Centre-Ouest Société Chimique de France*. Février 2015.

- Gally, J.-M., Bosc, N., Pihan, E., Arora, R., Canault, B., Bourg, S., Aci-Sèche, S. and Bonnet, P. Combining the strengths of freely accessible toolkits: a universal workflow for the preparation of molecules for virtual screening. *Journée Jeunes Chercheurs de la Section Régionale Centre-Ouest Société Chimique de France*. Février 2015.
- Bosc, N., Meyer, C. and Bonnet, P. Proteometrics, chemometrics and proteochemometrics: helpful tools for kinase drug discovery. *22nd Young Research Fellows Meeting*. Février 2015.
- Bosc, N., Meyer, C. and Bonnet, P. Broad profiling prediction of protein kinase inhibitors via kinochemometrics approaches. *27ème Colloque Biotechnocentre*. Octobre 2014.
- Gally, J.-M., Bosc, N., Pihan, E., Arora, R., Bourg, S., Aci-Sèche, S., Canault, B. and Bonnet, P. A free accessible workflow for preprocessing virtual screening experiments. *27ème Colloque Biotechnocentre*. Octobre 2014.
- Pihan, E., Canault, B., Arora, R., Golib-Dzib, J.-F., Bosc, N., Bourg, S. and Bonnet, P. Large scale kinase virtual screening plateform. *50th International Conference on Medicinal Chemistry (RICT 2014)*. Juillet 2014.
- Gally, J.-M., Bosc, N., Pihan, E., Arora, R., Bourg, S., Canault, B. and Bonnet, P. Development of a universal workflow for the preparation of molecular databases for virtual screening. *Journée scientifique de la Fédération de Recherche FR2708.* Juin 2014.
- Gally, J.-M., Bosc, N., Pihan, E., Arora, R., Bourg, S., Canault, B. and Bonnet, P. Development of a universal workflow for the preparation of molecular databases for virtual screening. *3rd Chemoinformatics Strasbourg Summer School*. Juin 2014.
- Bosc, N., Meyer, C. and Bonnet, P. Broad profiling prediction of protein kinase inhibitors via kinochemometrics approaches.*3rd Chemoinformatics Strasbourg Summer School.* Juin 2014.
- Bosc, N., Meyer, C. and Bonnet, P. Broad profiling prediction of protein kinase inhibitors via kinochemometrics approaches. *Colloque ITMO Bases Moléculaires et structurales du vivant.* Juin 2014.
- Bosc, N., Meyer, C. and Bonnet, P. Kinochemometrics : Un outil d'aide à l'identification d'inhibiteurs sélectifs de protéines kinases*. 6èmes journées de la Société Française de Chémoinformatique (SFCi)*. Octobre 2013.
- Bosc, N., Meyer, C. and Bonnet, P. Kinochemometrics : Identification rapide d'inhibiteurs affins et sélectifs de protéines kinases*. 26ème Colloque Biotechnocentre*. Octobre 2013.
- Bosc, N., Meyer, C. and Bonnet, P. Kinochemometrics: Une approche protéo-chimiométrique tridimensionnelle appliquée aux protéines kinases*. 20ème Journée Jeunes Chercheurs de la Société de Chimie Thérapeutique*. Février 2013.

Nicolas BOSC

Développement de nouvelles approches protéo-chimiométriques appliquées à l'étude des interactions et de la sélectivité des inhibiteurs de kinases

Résumé :

Le kinome humain comprend 518 protéines. Elles participent au processus de phosphorylation des protéines qui joue un rôle important dans les voies de signalisation cellulaire. Leur dérégulation est connue comme étant une cause de nombreuses maladies graves tels que les cancers. Du fait de leur grande similarité structurale des protéines kinases, il est difficile de développer des inhibiteurs qui soient à la fois efficaces et sélectifs. L'absence de sélectivité conduit le plus souvent à des effets secondaires particulièrement néfastes pour l'organisme. Au cours de cette thèse, nous avons d'abord développé de nouvelles métriques dont le but est de déterminer la sélectivité d'inhibiteurs à partir de données d'inhibition. Elles présentent l'avantage, comparées à d'autres métriques, d'être applicables sur n'importe quel type de données. Dans un deuxième temps, nous avons développé une approche protéométrique dans le but de comprendre pourquoi certaines protéines kinases ne sont jamais inhibées par des inhibiteurs de Type II. Le modèle statistique mis en place nous a permis d'identifier plusieurs résidus discriminants dont certains déjà décrits expérimentalement dans la littérature. Dans un troisième temps, nous avons développé un nouveau descripteur 3D de protéines kinases avec lequel nous avons mis en place et validé des modèles protéo-chimiométriques visant à étudier et découvrir de nouveaux inhibiteurs.

Mots clés : protéo-chimiométrie, protéine kinase, inhibiteur, sélectivité, protéométrie

Development of new proteo-chemometric approaches applied to the study of the interaction and the selectivity of kinase inhibitors

Summary:

The human kinome contains 518 proteins. They share a common mechanism of protein phosphorylation known to play an important role in cellular signaling pathways. Impaired kinase function is recognized to be involved in severe diseases like cancer. Due to high structural similarity between protein kinases, development of potent and selective kinase inhibitors is a challenging task. The selectivity of kinase inhibitors may lead to side effects potentially harmful. In this thesis, we first developed new selectivity metrics to determine inhibitor selectivity directly from biological inhibition data. Compared to existing metrics, the new selectivity scores can be applied on diverse inhibition data types. Second, we developed a proteometric approach in order to understand why some protein kinases are never inhibited by Type II inhibitors. The statistical model built for this purpose allowed us to identify several discriminant residues of which few of them correspond to experimentally described residues of interest. Third, using a new 3D protein kinase descriptor, we developed and validated novel proteo-chemometrics approaches to study and discover new kinase inhibitors.

Keywords: proteo-chemometrics, protein kinase, inhibitor, selectivity, proteometrics

Institut de Chimie Organique et Analytique
UMR CNRS-Université d'Orléans 7311
Université d'Orléans
Rue de Chartres
45067 Orléans