Table des matières

Liste des abréviations

1D, 2D, 3D	Une, Deux, Trois Dimension(s)
AA	Acides Aminés
ADMETox	Absorption, Distribution, Métabolisme, Excrétion et Toxicité
AE	Auto-Encodeurs
ALK	*Anaplastic Lymphoma Kinase* (kinase du lymphome anaplasique)
AMM	Autorisation de Mise sur le Marché
ANSM	Agence National de la Sécurité du Médicament
ATP	Adénosine TriPhosphate
AurKA	*Aurora kinase A*
BCR-ABL1	*Breakpoint Cluster Region-Abelson*
BDD	Base de données
BFP	*Bit fingerprint* (empreinte moléculaire composée de bits)
BLAs	*Biologics License Applications* (biomédicaments)
CDK	*Cyclin Dependent Kinase* (Protéines kinase dépendante des cyclines)
ChEMBL	*Chemical database of the European Molecular Biology Institute*
CLK	*Cdc2-Like Kinase*
CNN	*Convolutional Neural Networks* (Réseaux de neurones à convolutions)
CNS MPO	*Central Nervous System Multiparameter Optimization*
CSS	*Cascading Style Sheets*
DL	*Deep Learning* (Apprentissage par réseaux de neurones profonds)
DNS	*Domain Name System* (système de noms de domaine)
DRL	*Deep Reinforcement Learning* (Apprentissage par renforcement profond)
DYRK	*Dual-specificity tyrosine regulated kinase 1A*
ECFP	*Extended Connectivity FingerPrint* (Empreinte Moléculaire à Connectivité Etendue)
F2D	Frags2Drugs
FBDD	*Fragment-Based Drug Design* (Conception d'inhibiteurs basée sur les fragments)
FDA	*Food and Drug Administration*
GAN	*Generative Adversarial Networks* (Réseaux Antagonistes Génératifs)
Gunicorn	*Green Unicorn*
Haspin	*Haploid germ cell–specific nuclear protein*
HBA	*Hydrogen Bond Acceptors* (Nombre d'accepteurs de liaison hydrogène)
HBD	*Hydrogen Bond Donnors* (Nombre de donneurs de liaison hydrogène)
HTML	*HyperText Markup Language*
HTS	*High Throughput Screening* (Criblage à haut débit)
HTTP	*Hypertext Transfer Protocol* (protocole de transfert hypertexte)
HTTPS	*Hypertext Transfer Protocol Secure* (protocole de transfert hypertexte sécurisé)
IA	Intelligence Artificielle
ID	Identifiant
IFP	*Interaction fingerprint*
InChI	*International Chemical Identifier*
IP	protocole internet
IUPAC	*International Union of Pure and Applied Chemistry*
JS	JavaScript
JSME	JavaScript *Molecular Editor*
KLIFS	*Kinase-Ligand Interaction Fingerprints and Structures*
LabEx	Laboratoire d'excellence
LE	*Ligand Efficiency*
LLE	*Ligand-lipophilicity efficiency*
LogP	Coefficient de partage octanol/eau

MDL	*Molecular Design Limited*
ML	*Machine Learning* (Apprentissage Automatique)
MOL file	*MOLecule file*
MTV	*Model-Template-View* (Modèle-Gabarit-Vue)
MVC	Modèle-vue-contrôleur
MW	*Molecular Weight* (Masse Molaire)
NHA	*Number of Heavy Atoms* (Nombre d'atomes lourds)
NMEs	*New Molecular Entities* (Nouvelles Entités Moléculaires)
NRB	*Number of Rotatable Bonds* (Nombre de liaisons rotables)
ORM	*Object-Relational Mapper*
PAINS	*Pan Assay INterference compoundS*
PDB	*Protein Data Bank*
PHP	Hypertext Preprocessor
Pim	Proviral integration site for Moloney murine leukemia virus
PKI	*Protein Kinase Inhibitor* (Inhibiteur de Protéine Kinase)
PKIDB	*Protein Kinase Inhibitor Database*
PLK1	*Polo-like kinase 1*
QED	*Quantitative Estimate of Druglikeness*
QSAR	*Quantitative Structure Activity Relationship* (Relation Quantitative Structure-Activité)
RCSB	*Research Collaboratory for Structural Bioinformatics*
RDKit	*Rational Discovery Kit*
RL	*Reinforcement Learning* (Apprentissage par renforcement)
RMN	Résonance Magnétique Nucléaire
RNN	*Recurent Neural Networks* (Réseaux de neurones récurrents)
RO3	*Rule of three* (Règle de 3)
RO5	*Rule of five* (Règle de 5)
Score SA	*Synthetic Accessibility Score* (Score de facilité de synthèse)
SDF	*Structure Data File*
SGBD	Système de gestion de base de données
SMILES	*Simplified Molecular Input Line Entry Specification*
SQL	*Structure Query Language*
URL	*Uniform Resource Locator* (adresse web)
vHTS	*virtual High throughput Screening* (Criblage virtuel à Haut Débit)
www	*World Wide Web* (toile mondiale)
ZINC	*ZINC Is Not Commercial*

Liste des figures

Liste des tableaux

Avant-propos

J'ai effectué ma thèse à l'Institut de Chimie Organique et Analytique (ICOA) d'Orléans. Elle a été cofinancée par la région Centre Val de Loire et le Laboratoire d'Excellence (LabEx) Synthèse Organique (SynOrg).

L'ICOA est un laboratoire de recherche créé en 1995. Placé sous les tutelles de l'Université d'Orléans et du CNRS, il s'agit de l'unité mixte de recherche (UMR) 7311. L'ICOA est impliqué dans deux LabEx : SynOrg et IRON. Ce laboratoire a été dirigé de 2016 à mars 2021 par le Pr Pascal Bonnet. C'est aujourd'hui le Pr Sylvain Routier qui en assure la direction par intérim. Ce laboratoire a pour but de modéliser, concevoir, synthétiser et analyser des molécules à visée thérapeutique et cosmétique. Cette recherche se décline en différents axes de recherche répartis en cinq équipes :

1. Bioinformatique Structurale et Chémoinformatique (SB&C)
2. Glycobio&chimie
3. Hétérocycles, Nucléosides et Agents d'Imagerie (HNAI)
4. Méthodologies, chimie hétérocyclique, chimie verte (MCHCV)
5. Stratégies analytiques, affinités et bioactifs (SAAB)

J'ai effectué mon stage de master 2, puis ma thèse de doctorat dans l'équipe SB&C. Cette équipe est dirigée, depuis neuf ans, par le Pr Pascal Bonnet. Elle est constituée de cinq membres permanents, le Pr Pascal Bonnet, le Dr Samia Aci-Sèche, le Dr Stéphane Bourg, l'ingénieur d'étude Pascal Krezel et le Pr Caroline West, cette dernière partageant son temps recherche entre l'équipe SB&C (50%) et l'équipe SAAB (50%). L'équipe SB&C se compose aussi de post-doctorants, de doctorants et de stagiaires (https://www.icoa.fr/fr/bonnet/equipe). La recherche effectuée dans cette équipe concerne la modélisation et l'obtention de molécules bioactives par la conception et l'application de méthodes informatiques. L'équipe SB&C a une expertise tournée vers l'étude de la famille des protéines kinases, bien que les travaux initiés actuellement tendent à cibler d'autres familles de protéines. Parmi les projets développés dans l'équipe, il y a la prédiction de la cinétique des inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI), la mise en place d'un outil de construction de molécules *in silico* à partir de fragments moléculaires ainsi que diverses applications de méthodes d'amarrage moléculaire avec l'analyse des poses obtenues pour aider les chimistes au développement rationnel de candidats médicaments prometteurs. L'activité biologique de ces candidats médicaments y est aussi évaluée par des méthodes d'apprentissage profond et de simulations de dynamiques moléculaires. Les différents services développés par l'équipe sont regroupés sur une plateforme web (http://sbc.icoa.fr/) donnant l'accès aux outils à tout chercheur.

J'ai réalisé, au cours de ma licence en biologie, que l'informatique pouvait être utilisée pour mieux comprendre le fonctionnement de l'organisme, des pathologies et accélérer la découverte de nouveaux médicaments. Cela a créé en moi une vocation pour la bioinformatique et m'a poussé à continuer mes études dans ce domaine, en passant par la chémoinformatique, jusqu'au doctorat. J'espère que par la lecture de cette thèse je vous transmettrai ma passion et créerai de nouvelles vocations. La puissance de l'informatique permet de gagner du temps dans le processus de découverte des nouveaux médicaments. Cela implique un gain d'argent pour les entreprises souhaitant commercialiser le futur médicament. Grâce à l'utilisation de l'informatique les échecs sont réduits pendant les tests biologiques et les essais cliniques effectués sur les candidats médicaments. L'apport de connaissances en chimie à celles que j'avais en biologie et en informatique m'a permis de mener à bien cette thèse. Des disciplines existent à l'interface de chacun de ces domaines pris deux à deux (Figure 1), je situerai ma thèse à

l'interface des 3, entre la chémoinformatique et la biologie. En effet, au cours de celle-ci, j'ai utilisé et développé des programmes informatiques faisant appel à des connaissances de chimie pour inhiber des cibles biologiques.

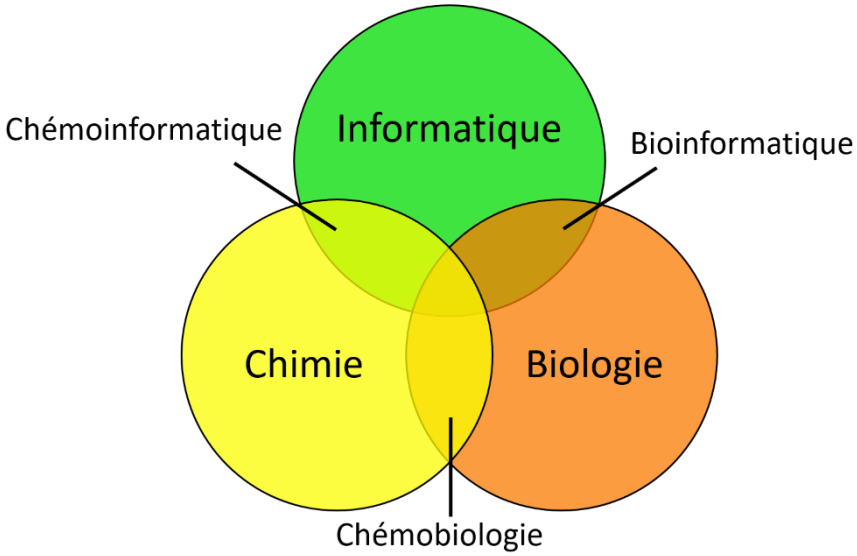


Figure 1 : Diagramme de Venn des différents domaines abordés au cours de cette thèse.

La méthode de découverte de nouveaux PKI principalement utilisée dans cette thèse repose sur des fragments de molécules. Semblables à des pièces de « Lego », ils permettent par un nouvel assemblage de découvrir des molécules aux propriétés intéressantes. Différentes applications de cet outil seront présentées au cours de cette thèse. Cet outil a été initié lors de la thèse du Dr José-Manuel Gally[1], puis développé pendant celle du Dr Colin Bournez[2] avec l'aide de l'ingénieur d'étude Pascal Krezel. Les travaux effectués par mes prédécesseurs ont abouti à un outil fonctionnel pouvant être appliqué sur n'importe quelle protéine kinase. Des molécules ont d'ailleurs été conçues, synthétisées et testées biologiquement ayant des activités nanomolaires au cours de la thèse du Dr Colin Bournez, validant cet outil[2]. À la suite de mon stage de master 2, j'ai collaboré avec lui pendant ma première année de thèse pour terminer le développement, apprendre à utiliser et analyser cet outil informatique.

Dans ce manuscrit une introduction présentera, tout d'abord, toutes les notions à connaître pour la compréhension de cette thèse. Vous y trouverez des explications de la découverte de nouveaux médicaments, leur conception basée sur les fragments et une présentation de la famille des protéines kinases. Vous y découvrirez aussi comment sont définis et catégorisés les PKI, puis, enfin, un article que j'ai co-écrit sur une comparaison des PKI déjà commercialisés ou en cours d'essais cliniques avec les PKI présents dans la *Chemical database of the European Molecular Biology Institute* (ChEMBL). Le second chapitre exposera d'abord, sous la forme d'un article, l'outil de conception de PKI principalement utilisé pendant cette thèse, puis les améliorations que j'y ai apporté. Le troisième chapitre abordera la mise en place de la plateforme web regroupant les différents outils développés par l'équipe SB&C. Ce chapitre contiendra aussi un article présentant l'outil que j'ai développé durant mon stage de master 2 et terminé pendant ma thèse. Le quatrième chapitre de cette thèse montrera l'application de méthodes *in silico* se basant sur les fragments pour la conception de nouveaux PKI. La découverte de molécules s'y déclinera en 4 projets, 3 ciblant des protéines kinases impliquées en cancérologie, en maladies neurodégénératives et 1 ayant pour but de trouver des sondes fluorescentes. Enfin une conclusion générale viendra terminer cette thèse avec une mise en perspective pour la continuité de ces travaux.

Chapitre 1 : Introduction

1.1 Chémoinformatique

1.1.1 Définition

La première définition de la chémoinformatique date de 1998, où F.K. Brown indique que la chémoinformatique est le mélange d'informations issues du processus de découverte de médicaments pour obtenir des connaissances permettant d'identifier et d'optimiser, plus rapidement, de meilleurs candidats médicaments[3].

Cette discipline a été officiellement déclarée en Europe en 2006, pendant un congrès à Obernai (France) « *Workshop chemoinformatics in Europe : Research and Teaching* ». La définition complète qui a été donnée pour la chémoinformatique est :

« Chemoinformatics is a scientific discipline that has evolved in the last 40 years at the interface between chemistry and computer science. It has been realized that in many areas of chemistry, the huge amount of data and information produced by chemical research can only be processed and analyzed by computer methods. Furthermore, many of the problems faced in chemistry are so complex that novel approaches utilising solutions that are based on informatics methods are needed. Thus, methods were developed for building databases on chemical compounds and reactions, for the prediction of physical, chemical and biological properties of compounds and materials, for drug design, for structure elucidation, for the prediction of chemical reactions and for the design of organic syntheses. Research and development in chemoinformatics is essential:

- *For increasing our understanding of chemical phenomena*
- *For industry to remain competitive in a global economy*

Chemoinformatics methods can be applied in any field of chemistry, from analytical chemistry to organic chemistry. It is of particular importance in drug design and development […] »[4].

Pour résumer cette définition en une phrase, la chémoinformatique est une discipline permettant l'analyse de données chimiques à partir d'outils informatiques. Les méthodes chémoinformatiques reposent sur :

- La construction de chimiothèques (bases de données faites de composés et de réactions chimiques)
- La conception de médicaments
- La détermination de structures chimiques
- La prédiction de :
    - Propriétés physiques, chimiques et biologiques de composés
    - Réactions chimiques
    - Synthèse de composés organiques

À la suite de ce congrès et de la première journée nationale en chémoinformatique, en 2007, naît la Société Française de Chémoinformatique. Elle regroupe plus de 100 chémoinformaticiens francophones, industriels (27 sociétés privées) et académiques (35 laboratoires académiques)[5]. Cette association est actuellement présidée par le Pr Matthieu Montes.

L'apport de l'informatique à la chimie permet de manipuler de grands nombres d'informations, ce qui aurait été impossible autrement. Les bases de données (BDD) contiennent souvent plusieurs

millions de molécules et les criblages virtuels à haut débit (*virtual High Throughput Screening,* vHTS) génèrent autant de résultats à analyser que de molécules testées. L'informatique permet d'estimer la relation quantitative structure-activité (*Quantitative Structure Activity Relationship*, QSAR) pour prédire l'activité de composés par des modèles statistiques[6,7]. Ces modèles QSAR sont difficilement applicables à de nouveaux composés[8]. L'intelligence artificielle (IA) est utilisée pour estimer l'activité de nouveaux composés sur de nouvelles cibles[9]. L'apprentissage par réseaux de neurones profonds est de plus en plus utilisé en chémoinformatique, apportant de meilleurs outils de prédiction de bioactivité, des outils de conception de molécules *de novo* et des outils de prédiction de synthèse de composés[10]. La chémoinformatique permet, comme je l'évoquais en avant-propos, un gain considérable de temps qui entraîne aussi un gain d'argent. Différents échecs peuvent survenir lors des différentes étapes de la conception d'un médicament. Il s'agit par exemple du manque d'efficacité du médicament ou de la présence de toxicité obligeant à arrêter et recommencer les essais cliniques menés. En réduisant ces échecs par des calculs ou prédictions, l'argent non dépensé est économisé[11].

La première problématique résolue par cette discipline et que nous allons voir dans la prochaine partie a été :

Comment faire pour qu'un ordinateur puisse interpréter une molécule, habituellement dessinée par un chimiste, afin de pouvoir effectuer les calculs précédemment listés ?

1.1.2 Représentations moléculaires

Une molécule est « une entité électriquement neutre faite de plus d'un atome (n>1) »[12]. Les atomes composant une molécule peuvent être identiques ou non et sont liés par liaisons covalentes.

De nombreuses représentations moléculaires existent, la plus simple étant la formule brute qui indique combien d'atomes de chaque type constituent la molécule. Cette représentation ne précise pas le positionnement de chaque atome et une même formule brute peut représenter différentes molécules. Ainsi, pour observer la manière dont sont liés les atomes d'une molécule, il est possible de la représenter en deux dimensions (2D), par exemple, sous sa formule développée ou topologique. D'autres représentations, comme celles de Cram, permettent d'obtenir une représentation de la molécule en trois dimensions (3D). Bien que ces représentations soient utiles pour qu'un humain puisse interpréter les molécules, elles ne le sont pas pour un ordinateur.

Les programmes informatiques peuvent interagir avec des molécules lorsque celles-ci sont stockées sous la forme de graphes moléculaires. Dans ce type de graphe, les nœuds correspondent aux atomes et les arêtes aux liaisons chimiques entre eux. Ces représentations informatiques des molécules peuvent contenir plus ou moins d'informations, allant de la 2D à la 3D.

Les différents formats utilisés pour transcrire ces représentations sont des chaînes de caractères ou des tables de connectivité. Les chaînes de caractères sont traitées en tant que tableaux à une dimension (1D) ou vecteurs. Les tables de connectivité sont des matrices ou tableaux 2D.

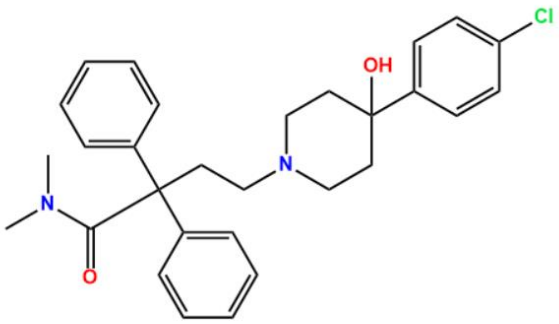
1.1.2.1 Représentations à partir de chaînes de caractères

Les chaînes de caractères encodent de manière légère la représentation 2D d'une molécule. Parmi les formats couramment utilisés, il y a la nomenclature *International Union of Pure and Applied Chemistry* (IUPAC)[13], *l'International Chemical Identifier* (InChI)[14] et le *Simplified Molecular Input Line Entry Specification* (SMILES)[15]. Il existe de nombreux autres formats pour encoder la représentation 2D d'une molécule, mais dans ce paragraphe, je me focaliserai sur ceux que j'ai utilisé au cours de ma thèse. Un exemple pour chacun de ces formats est montré en Tableau 1 pour le lopéramide, principe actif de l'antidiarrhéique Imodium®.

La nomenclature IUPAC énumère chaque groupe fonctionnel présent dans la molécule[13]. Elle permet par sa lecture de dessiner la molécule. La nomenclature IUPAC est faite pour être comprise par l'humain. Elle repose sur trop de règles et de manières différentes de décrire une même molécule pour être optimisée pour son utilisation par un programme informatique.

Le format SMILES est un format initialement propriétaire, développé par Daylight Chemical Information Systems[15], devenu libre de droits depuis 2007 avec le projet OpenSMILES. Il est compréhensible et interprétable par l'humain et la machine. Il repose sur un nombre minimal de règles simples pour encoder le graphe d'une molécule. Lors de son utilisation dans une BDD, où les redondances doivent être évitées, il peut être converti en SMILES canonique pour éviter d'avoir deux représentations SMILES encodant la même molécule. Cependant, la forme canonique dépend de l'algorithme utilisé pour l'obtenir, elle n'est donc pas unique pour une même molécule.

Tableau 1 : Différentes chaînes de caractères chémoinformatiques pour encoder la molécule lopéramide. La nomenclature IUPAC a été obtenue par LexiChem 2.6.6, l'InChI et l'InChIKey par InChI 1.0.5 et le SMILES canonique par OEChem 2.1.5

Nom de format	Valeur
Molécule (structure 2D)	
Nomenclature IUPAC	4-[4-(4-chlorophenyl)-4-hydroxypiperidin-1-yl]-N,N-dimethyl-2,2-diphenylbutanamide
SMILES canonique	CN(C)C(=O)C(CCN1CCC(CC1)(C2=CC=C(C=C2)Cl)O)(C3=CC=CC=C3)C4=CC=CC=C4
InChI	InChI=1S/C29H33ClN2O2/c1-31(2)27(33)29(24-9-5-3-6-10-24,25-11-7-4-8-12-25)19-22-32-20-17-28(34,18-21-32)23-13-15-26(30)16-14-23/h3-16,34H,17-22H2,1-2H3
InChIKey	RDOIQAHITMMDAJ-UHFFFAOYSA-N

L'InChI est un autre format libre de droits, développé par l'IUPAC en 2005. C'est une chaîne de caractères ASCII (pour *American Standard Code for Information Interchange*), semblable à un code-barres, permettant de retrouver la représentation graphe de la molécule qu'elle encode[14]. Ce format est composé de quatre parties principales séparées par une barre oblique. D'abord le numéro de version et un « S » pour préciser s'il s'agit d'un InChI standard, ensuite la formule brute de la molécule, puis les connexions entre les atomes lourds et enfin les connexions avec les atomes d'hydrogène. D'autres informations peuvent ensuite être ajoutées comme le nombre de protons, la stéréochimie et l'isotopie. Le problème principal de l'InChI est sa longueur, qui augmente drastiquement avec le nombre d'atomes de la molécule représentée. Pour pallier ce problème inhérent aux représentations par chaînes de caractères, des méthodes comme le « hachage » sont utilisées.

Le hachage est la transformation d'une chaîne de caractères de n'importe quelle taille en une clé plus petite, de longueur fixe. Le format InChIKey est obtenu par un hachage de l'InChI visant à obtenir une chaîne d'une longueur fixe de 27 caractères. L'InChIKey se décompose en trois blocs, séparés par des traits d'union. Les 14 premiers caractères encodent les parties principales de l'InChI. Les 9 caractères suivants encodent les autres informations de l'InChI et la dernière lettre indique le nombre de protons de la molécule[14]. À la différence de l'InChI qui est unique pour chaque molécule, un même InChIKey peut encoder plusieurs InChI et donc différentes molécules. Ce phénomène, appelé « collision », est dû aux algorithmes de hachage. Cependant, entre 2007 et 2012 seuls deux cas de collisions ont été observés, ce qui nous induit à penser que l'InChIKey peut être utilisé efficacement[16]. Contrairement aux trois autres représentations (IUPAC, SMILES et InChI), à cause des algorithmes de hachage utilisés, un humain ne peut pas reconstruire une molécule directement à partir d'une représentation InChIKey.

1.1.2.2 Représentations à partir de matrices

Les tables de connectivité permettent d'encoder la tri-dimensionnalité des molécules en précisant les coordonnées de chaque atome les constituant. Ce format popularisé par *Molecular Design Limited* (MDL) est plus lourd qu'une simple chaîne de caractères, mais apporte plus d'informations[17]. La table de connectivité de MDL contient 6 sous-parties, tout d'abord une ligne des comptes indiquant le nombre d'atomes, de liaisons et de listes d'atomes, ainsi que la chiralité et la version de la table de connectivité. Ensuite le bloc des atomes indiquant pour chaque atome ses coordonnées, son symbole, sa différence de masse (par rapport au même élément dans le tableau périodique), sa charge, sa stéréochimie et les atomes d'hydrogène qui lui sont associés. La troisième sous-partie est le bloc des liaisons indiquant quels atomes sont reliés et par quels types de liaisons, en plus de la stéréochimie et de la topologie des liaisons (cycliques ou non). Les trois premières sous-parties suffisent pour obtenir la structure 3D d'une molécule. Les deux sous-parties suivantes sont moins fréquentes, il s'agit du bloc des listes d'atomes et d'un bloc descriptif spécifique à certains logiciels. En dernière position, il y a le bloc des propriétés où des informations complémentaires peuvent être renseignées.

Le fichier texte MDL *molecule file* (MOL file) permet de sauvegarder une molécule à l'aide d'un en-tête et d'une table de connectivité. L'en-tête se compose de trois lignes permettant d'identifier le nom de la molécule, l'utilisateur, le logiciel, la date et d'indiquer un commentaire d'une ligne[17].

Le format *Structure Data File* (SDF) permet de sauvegarder plusieurs molécules (tables de connectivité) dans un même fichier texte. Les différentes molécules y sont séparées par 4 symboles « $$$$ » et des informations complémentaires peuvent être ajoutées à chaque molécule[17]. Ces informations peuvent par exemple être la valeur de descripteurs moléculaires, qui seront détaillés dans la sous-partie 1.1.3 de cette thèse. Il faut bien noter que même si les fichiers MOL file et SDF sont majoritairement utilisés pour représenter la structure 3D des molécules grâce aux coordonnées « x, y et z » de chaque atome, il suffit d'assigner la valeur 0 à toute la colonne des coordonnées « z » pour obtenir une représentation 2D des molécules. Le fichier SDF obtenu pour la molécule lopéramide, contenant la table de connexion correspondante, est montré en Figure 2.

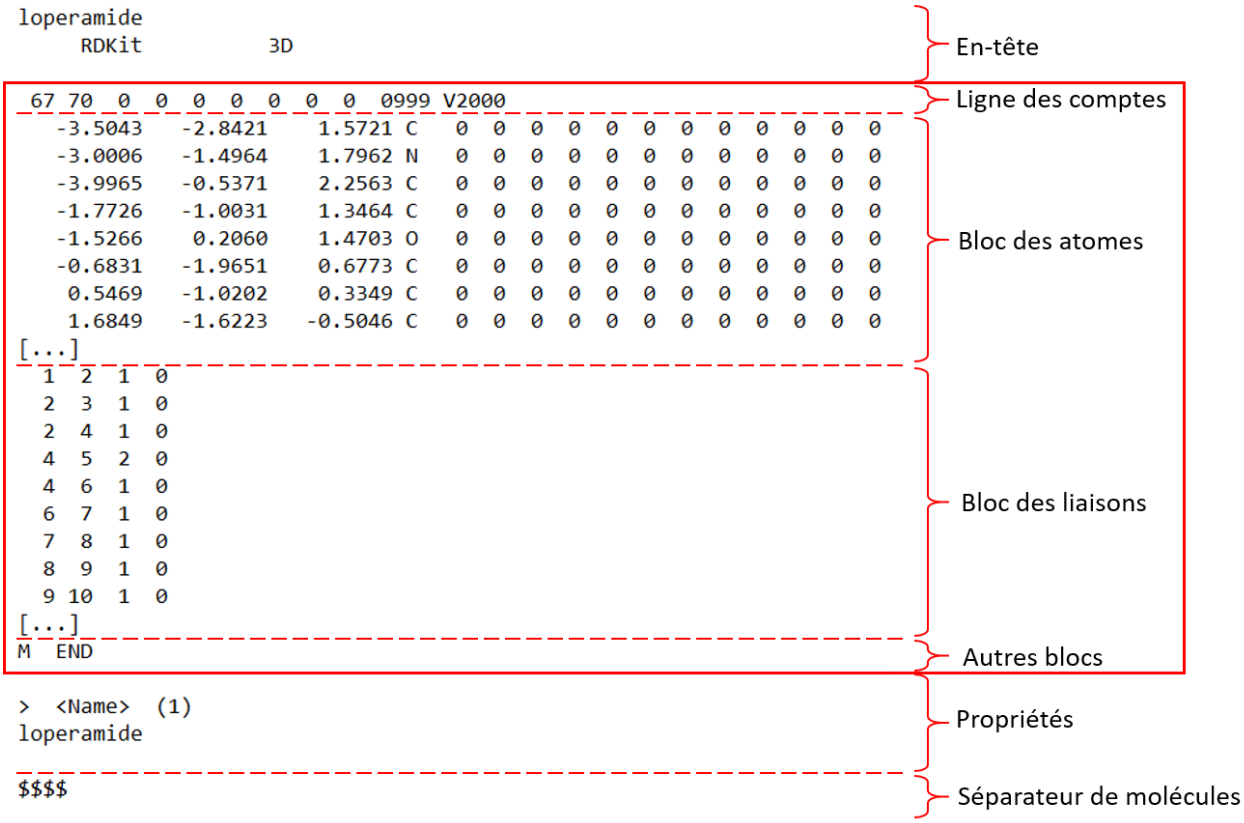


Figure 2 : Exemple de fichier SDF obtenu pour la molécule lopéramide. La table de connexion est encadrée en rouge. « Propriétés » désigne les propriétés du format SDF et non le bloc des propriétés de la table de connexion. Pour condenser cette représentation, des parties des blocs des atomes et des liaisons ont été tronquées et remplacée par « […] », ce symbole ne fait pas partie de la table de connexion.

1.1.3 Descripteurs moléculaires

Après avoir brièvement défini la chémoinformatique et avoir vu différentes représentations moléculaires, nous allons apprendre quels calculs peuvent être effectués à partir de ces représentations. Des algorithmes de chémoinformatique permettent de passer d'une représentation moléculaire à une ou plusieurs valeurs numériques[18]. Il s'agit là de descripteurs moléculaires « théoriques », mais il existe aussi des descripteurs expérimentaux obtenus par manipulations en laboratoire[19]. Le coefficient de partage octanol/eau (LogP) est un exemple de descripteur expérimental permettant d'estimer l'hydrophobicité d'une molécule.

Il existe plusieurs milliers de descripteurs moléculaires « théoriques », et de nouveaux descripteurs peuvent toujours être définis pour des applications particulières. Ils se décomposent en trois catégories selon les représentations moléculaires utilisées pour les obtenir[20]. Les descripteurs 1D sont obtenus à partir de la formule brute de la molécule. Les descripteurs 2D nécessitent de connaître la structure 2D de la molécule. Les descripteurs 3D prennent en compte la conformation tridimensionnelle de la molécule. Les descripteurs 1D et 2D, très rapides à calculer sont souvent utilisés comme filtres moléculaires, alors que les descripteurs 2D et 3D sont plus couramment utilisés dans la construction de modèles QSAR ou QSPR (pour relations quantitatives structures-propriétés). Le Tableau 2 montre quelques exemples de descripteurs classés par dimension avec la valeur obtenue sur le composé lopéramide.

Tableau 2 : Exemple de descripteurs moléculaires pouvant être calculés sur le composé lopéramide.

Catégorie	Nom
1D	Masse molaire
2D	Nombre d'atomes chiraux
3D	Premier moment principal d'inertie

Les empreintes moléculaires (ou *fingerprints*) sont des descripteurs particulièrement optimisés pour des calculs informatiques complexes, comme les prédictions de nouvelles propriétés par apprentissage automatique (*Machine Learning, ML*). En effet, ces descripteurs sont encodés en 1D dans des vecteurs de bits. L'information représentée par ces bits peut provenir initialement d'une représentation 2D ou 3D. Dans cette introduction, nous nous intéresserons aux empreintes moléculaires 2D, utilisées pour l'estimation de la similarité entre deux molécules. Ces empreintes moléculaires encodent dans chacun de leurs bits la présence ou l'absence de certaines sous-structures dans la molécule. Ces fragments peuvent être définis à l'avance ou relevés en parcourant la molécule. L'empreinte moléculaire MACCS (pour *Molecular ACCess System*) est un exemple utilisant des fragments prédéfinis. Il s'agit de 166 sous-structures permettant de distinguer efficacement les molécules[21]. L'empreinte moléculaire à connectivité étendue (*Extended Connectivity FingerPrint*, ECFP) repose sur le parcours de la molécule à l'aide d'un cercle de diamètre croissant permettant d'obtenir des sous-structures représentatives de la molécule[22]. L'obtention d'une ECFP est illustrée Figure 3. Pour l'exemple, l'empreinte a une longueur de 8 bits et les cercles vont jusqu'au diamètre de 4 atomes, en prenant les diamètres de 0, 2 et 4 atomes. Ce diamètre est couramment utilisé pour une recherche par similarité ou un regroupement (*clustering*) de molécules. Les méthodes de ML nécessitent parfois des diamètres plus élevés, jusqu'à 8 atomes[22]. La longueur du vecteur de bits est habituellement beaucoup plus grande que celle utilisée pour l'illustration, avec une taille de 1024 ou 2048 bits.

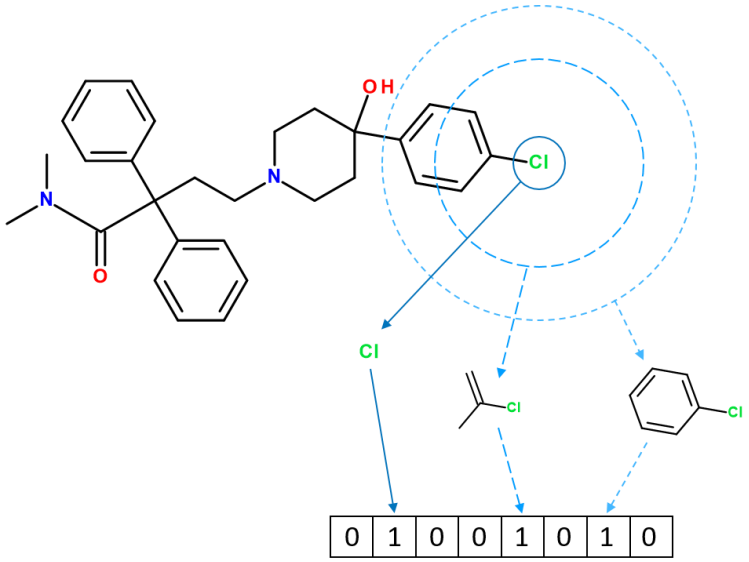


Figure 3 : Obtention d'une empreinte moléculaire de type ECFP. Le diamètre maximal autorisé ici est 4, les cercles successifs ont des diamètres de 0, 2 et 4. La présence des sous-structures obtenues est stockée dans un vecteur de bits, ici d'une longueur de 8.

Face aux différents formats permettant l'encodage des molécules et au nombre de descripteurs pouvant être calculés, nous pourrions penser qu'il est compliqué de passer d'un format à l'autre, de calculer divers descripteurs ou empreintes moléculaires. Certaines librairies écrites dans le langage de

programmation Python intègrent déjà le code nécessaire pour effectuer ces calculs. Par exemple, la librairie *Rational Discovery Kit* (RDKit), libre de droits et qui bénéficie d'une communauté très active, permet d'effectuer toutes ces actions sur des molécules virtuelles[23]. D'autres initiatives permettent un apprentissage facilité de la chémoinformatique, comme TeachOpenCADD[24]. Il s'agit de tutoriels orientés pour des néophytes proposant des exemples d'applications concrètes utilisant la chémoinformatique.

1.1.4 Bases de données en chémoinformatique

Une chimiothèque est une collection de différentes molécules, souvent formée de plusieurs milliers de molécules. Par exemple, la chimiothèque nationale, en France, contient 85 000 composés et extraits naturels[25]. La chimiothèque de l'ICOA contient 14 789 molécules.

Les molécules en chémoinformatique peuvent être regroupées dans des BDD. Le nombre de molécules dans ces chimiothèques virtuelles peut-être de plusieurs millions. L'avantage de l'utilisation de ces BDD de molécules virtuelles est de pouvoir accéder facilement aux informations qui y sont associées, sans devoir lire des articles et revues pour en extraire manuellement les informations de chaque molécule. À part les chimiothèques proposées par les fournisseurs commerciaux, les molécules d'une BDD proviennent de la littérature scientifique. Les descripteurs et les empreintes moléculaires permettent de filtrer, de trier et d'effectuer des recherches par similarité ou sous-structure sur les molécules contenues dans ces BDD.

Les molécules peuvent être retrouvées dans les BDD à partir d'identifiants spécifiques. Un exemple d'identifiant du lopéramide, pour chaque BDD, est montré dans le Tableau 3. L'utilisation des BDD privées est restreinte aux utilisateurs pouvant en payer l'accès. Je n'en ai pas utilisé au cours de ma thèse, mais la plus célèbre est *SciFinder,* développée par CAS (pour *Chemical Abstract Service*). Initiée en 1995, elle contient aujourd'hui plus de 161 millions de molécules extraites de la littérature et référencées à l'aide du CASRN (pour *Chemical Abstract Service Registry Number*)[26]. Les avantages de cette BDD privée sont que les informations qu'elle comporte sont relues par des experts, et basées sur un algorithme permettant d'éviter des doublons ou des erreurs. Parmi les désavantages, il y a l'accès privé ne favorisant pas l'avancée de la recherche et les collaborations entre équipes académiques.

Les BDD de molécules publiques donnent accès à tous aux informations extraites de la littérature. Il en existe de nombreuses, mais nous ne présenterons ici que celles qui ont été utilisées dans cette thèse. PubChem est une BDD libre d'accès contenant des informations sur 111 millions de molécules uniques[27]. La référence aux molécules dans cette BDD se fait par un nombre entier appelé PubChem CID (pour *Compound Identifier*).

La *Chemical database of the European Molecular Biology Institute* (ChEMBL) regroupe plus de deux millions de molécules dont l'activité a été mesurée expérimentalement et publiée[28]. Un identifiant (ID) nommé CHEMBL ID est attribué à chaque élément (molécules, activité, cible, référence, …). La ChEMBL repose sur des services web pour permettre l'accès aux informations qu'elle contient[29]. Un des inconvénients de cette BDD est que ses données ne sont pas parfaitement revues et standardisées. Malgré cela, elle s'améliore constamment avec l'aide des retours que font les utilisateurs[28].

La BDD *ZINC Is Not Commercial* (ZINC) est une autre BDD publique proposant environ 200 millions de composés proposés par plusieurs fournisseurs commerciaux, cette BDD peut aussi être utilisée pour des criblages virtuels[30]. Depuis sa dernière mise à jour, la BDD ZINC intègre des informations sur les cibles et les activités biologiques permettant une utilisation par une communauté plus large[31]. Comme la BDD ZINC, Ambinter (Greenpharma SAS) propose un libre accès à des composés commerciaux. Ambinter contient 36 millions de composés disponibles directement chez les fournisseurs.

Tableau 3 : Différents identifiants du lopéramide selon différentes bases de données chémoinformatiques

Base de données	Nom d'identifiant	Identifiant
SciFinder	CASRN	53179-11-6
PubChem	PubChem CID	3955
ChEMBL	ChEMBL ID	CHEMBL841
ZINC	ZINC ID	ZINC537928
Ambinter	Ambinter ID	Amb2231667

Une BDD de bioinformatique structurale largement utilisée est la *Research Collaboratory for Structural Bioinformatics* (RCSB) *Protein Data Bank* (PDB). Il s'agit d'une BDD publique contenant les structures tridimensionnelles de protéines et de complexes protéine-ligand et d'acides nucléiques[32]. Les molécules y sont référencées par un identifiant PDB de ligand à 3 caractères. L'identifiant PDB de la protéine ou du complexe protéine-ligand est appelé PDB ID, il est fait de 4 caractères. En mai 2021, la PDB contient plus de 175 977 structures tridimensionnelles (protéine seule ou co-cristallisée avec un ligand). Un même ligand peut y être présent plusieurs fois, cristallisé dans différentes protéines ou plusieurs fois dans la même protéine. La BDD PDBeChem permet d'accéder aux informations associées aux ligands de la PDB, à partir de plus de 33 390 ligands[33].

Enfin, des BDD peuvent-être constituées en intégrant les informations provenant de différentes sources, de différentes BDD, dans le but de regrouper des connaissances sur une famille de protéine particulière. C'est le cas de la BDD des inhibiteurs de protéines kinases actuellement sur le marché ou en cours d'essais cliniques, ayant une dénomination commune internationale : *Protein Kinase Inhibitor Database* (PKIDB)[34,35]. Cette BDD a été développée dans l'équipe de recherche SB&C où j'ai effectué ma thèse.

1.2 Les médicaments

1.2.1 Définition d'un médicament

Selon le Larousse médical, un médicament est une substance ou composition ayant une activité pharmacologique : « utilisée pour prévenir, diagnostiquer, soigner une maladie, un traumatisme ou pour restaurer, corriger, modifier les fonctions organiques »[36]. Cette définition résume très bien celle plus complète indiquée par le Code de la Santé Publique dans l'article L5111-1[37]. La conception de médicaments est communément appelée par l'anglicisme « *drug design* » et la recherche de médicaments au sens large est appelée « drug discovery ».

Un médicament est constitué de deux éléments, un principe actif et un (ou plusieurs) excipient(s). Le principe actif permet au médicament d'effectuer son action, il provient majoritairement de synthèses chimiques, mais peut aussi être issu de sources biologiques. Les excipients sont les autres composants du médicament en dehors du principe actif. Les excipients doivent, dans l'idéal avoir une parfaite innocuité, sinon ils seront des excipients à effets notoires. Parmi leurs nombreux rôles, les excipients contribuent à :

- La stabilité du principe actif
- L'absorption du principe actif par l'organisme
- La forme galénique du médicament
- Au goût du médicament

Les biomédicaments sont des médicaments dont le principe actif provient d'une source biologique[38]. Les biotechnologies, à la base de la production de ce type de médicaments, sont en constante amélioration depuis les années 1970. Ils peuvent être créés par des bactéries ou des lignées cellulaires modifiées génétiquement, dans le but de produire en quantité des protéines ou des vaccins. C'est le cas par exemple de la production de l'insuline données aux patients diabétiques[38]. La modification génétique des microorganismes ou des cellules s'effectue aujourd'hui à l'aide de la méthode CRISPR-CAS9 (pour *Clustered Regularly Interspaced Short Palindromic Repeats* et *CRISPR-associated protein 9*). La mise en place de la méthode CRISPR-CAS9 a valu à E. Charpentier et à J. Doudna le prix Nobel de chimie en 2020[39]. Les biomédicaments permettent aussi de cibler très précisément les mécanismes pathologiques, grâce notamment aux anticorps monoclonaux, dont l'apport dans le traitement des cancers par immunothérapie a valu cette fois à J. Allison et à T. Honjo le prix Nobel de médecine en 2018[40].

1.2.2 Généralités sur les médicaments

Il existe différentes organisations pouvant autoriser un médicament à être mis sur le marché, selon les pays ou des regroupements de pays. Il y a par exemple l'EMA (pour agence Européenne des médicaments), l'ANSM (pour Agence Nationale de la Sécurité du Médicament) en France et la NMPA (pour *National Medical Products Administration*) en Chine. Cependant, la plupart des informations trouvées dans la littérature concernent les médicaments approuvés par la *Food and Drug Administration* (FDA) aux Etats-Unis d'Amérique.

La FDA approuve en moyenne 32 médicaments par an depuis 1993, avec une tendance à la hausse depuis 2007[41], comme illustré Figure 4 où le nombre de nouveaux biomédicaments approuvés, sous le nom de *Biologics License Applications* (BLAs), est distingué du nombre de médicaments issus de synthèse chimique, Nouvelles Entités Moléculaires (NMEs). L'année 2014 marque un changement dans le nombre moyen de BLAs, avec un passage de 4 à 12 biomédicaments approuvés en moyenne par an. Cela se traduit aussi par une augmentation dans la répartition entre les BLAs et les NMEs. Avant 2014, les BLAs représentaient en moyenne 17% du nombre d'approbations annuelles, alors qu'après cette année, ils représentent 35% du nombre moyen de nouveaux médicaments par an.

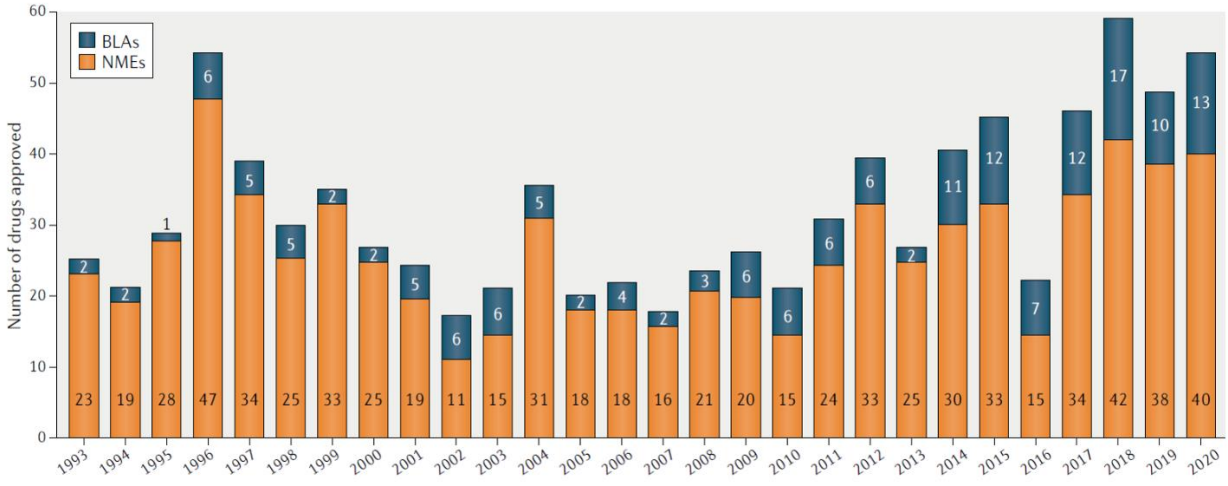


Figure 4 : Nombre de médicaments approuvés par an par la FDA. En orange, les NMEs et en bleu foncé, les BLAs. A. Mullard[41]

Bien que la tendance soit à la hausse chaque année depuis 2007, il y a relativement peu de nouveaux médicaments à être mis sur le marché tous les ans. En effet, il est compliqué pour une molécule de devenir un médicament. Nous allons maintenant détailler ce processus de conception de médicaments.

1.3 Conception de médicaments

La conception de médicaments est encadrée par une règlementation très précise. Elle se décompose en 3 étapes principales : d'abord la recherche, puis le développement et enfin la commercialisation. Le développement se découpe lui-même en 5 phases : le développement préclinique, suivi des 4 phases des essais cliniques, dont les 3 premières se font avant l'arrivée du médicament sur le marché (commercialisation). Pendant la commercialisation, une surveillance est effectuée sur les patients prenant le médicament, considérée comme la 4ème phase des essais cliniques et appelée « pharmacovigilance ».

Le dépôt du brevet garantissant la propriété intellectuelle sur la molécule appelée à devenir un médicament intervient pendant l'étape de recherche. Chaque étape ensuite, de la recherche à l'arrivée du médicament sur le marché, dure plusieurs années (Figure 5). La conception d'un médicament dure en moyenne douze ans au total[42]. Le brevet en assure la protection commerciale pendant 20 à 25 ans, c'est à dire qu'il couvre une période d'environ 10 ans après l'arrivée du médicament sur le marché[42].

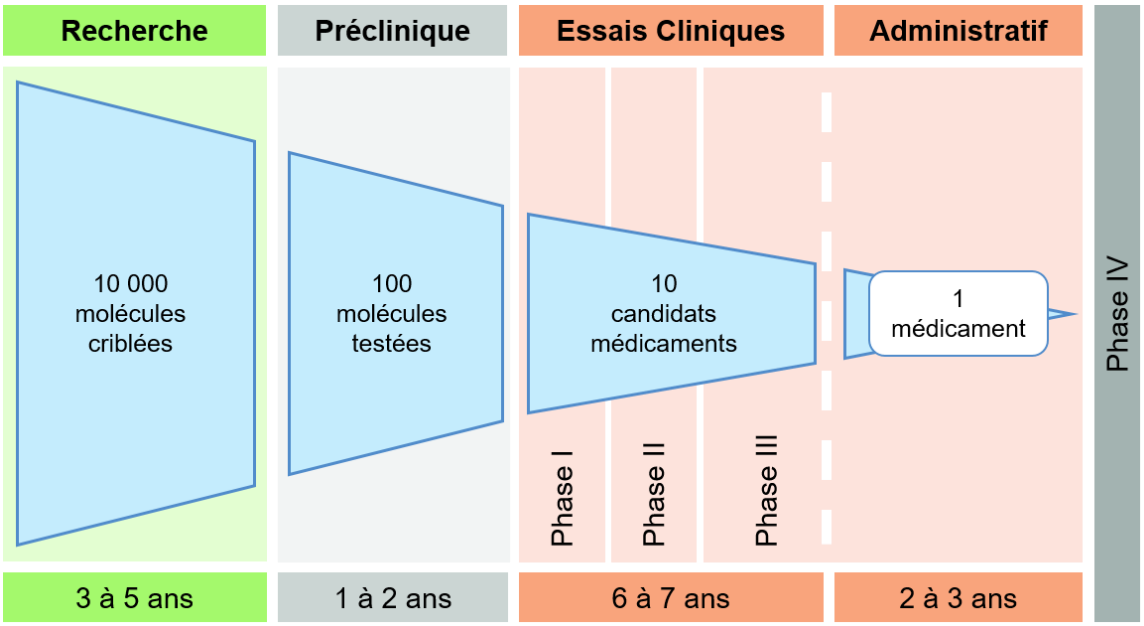


Figure 5 : Différentes étapes de la conception d'un médicament. Adapté d'après Matthews et al.[43] et Leem[42].

En plus d'être long, le déroulé de la conception de médicaments est coûteux[44] et se heurte à de nombreux échecs[45]. Entre 2006 et 2015, Thomas *et al.* ont observé que 62,2% des molécules sont passées de la phase I à la phase II, que 30,7% des molécules en essais cliniques sont allées de la phase II à la phase III et que 49,6% des molécules en phase III ont été approuvées[46].

Le coût moyen de la recherche et du développement d'un nouveau médicament fait l'objet de débats dans la littérature. DiMasi *et al.* estimaient à 1,1 milliards de dollars ce coût en 2003[11], et à 2,8 milliards de dollars en 2016[47]. Une autre étude basée sur les données publiques des médicaments approuvés par la *Food and Drug Administration* (FDA) de 2009 à 2018 estime la médiane de ce coût à 985 millions de dollars et la moyenne à 1,3 milliards[44]. En 2017, en France, la recherche et le développement dans les entreprises pharmaceutiques représentent 9,8% du chiffre d'affaires équivalant à 4,5 milliards d'euros[42].

La chémoinformatique intervient dès l'étape de recherche pour accélérer l'ensemble du processus de conception de médicaments et réduire les coûts. Son but est d'identifier, de la façon la

plus précise possible, les molécules qui auront le plus de chance de passer avec succès les étapes suivantes. Les étapes de la conception de médicaments permettent d'étudier l'efficacité de la molécule et ses effets secondaires. Les prochaines sous-parties vont détailler le contenu de ces étapes afin de comprendre comment passer d'une molécule à un médicament sur le marché.

1.3.1 Recherche

L'objectif de l'étape de recherche est de réduire un grand nombre de molécules à une centaine de molécules appelées « têtes de série », qui seront ensuite testées. La recherche de nouvelles molécules commence par la compréhension de la maladie à traiter. Le but est de comprendre des gènes aux protéines, des protéines aux cellules, des cellules aux tissus, puis au corps entier, comment la maladie effectue son action néfaste. Cela permet d'aboutir à l'identification d'une cible thérapeutique. Cette cible, souvent un gène ou une protéine, doit avoir une action pouvant être modifiée par un médicament.

Une fois la cible thérapeutique identifiée, il faut trouver des « têtes de série » qui seront les molécules ayant le plus de chances de devenir des candidats médicaments lors de l'étape de développement. Pour cela, des substances aux propriétés thérapeutiques intéressantes peuvent être extraites de la nature. L'extraction de produits naturels est d'ailleurs utilisée depuis le début du XIXème siècle, avec par exemple l'extraction de la morphine à partir de l'opium[48]. Mais aujourd'hui, il existe des méthodes plus sophistiquées pour trouver des têtes de série. Par exemple, des bactéries peuvent être génétiquement modifiées pour produire une hormone, comme l'insuline évoquée en partie 1.2.1. Depuis quelques années, grâce aux progrès en chémoinformatique et en intelligence artificielle (IA), des molécules peuvent être conçues *de novo* à partir d'ordinateurs, nous y reviendrons dans la partie 1.5.

Les touches, précurseurs des têtes de séries (*leads*) non optimisées, sont le plus souvent issues de criblages à haut débit (*High Throughput Screening*, HTS). Ces HTS sont faits par des robots testant des centaines de milliers de composés pour identifier les molécules montrant une interaction avec la cible. Les HTS peuvent être effectués par chémoinformatique (vHTS) pour tester virtuellement des millions de molécules. Les touches sont des molécules qui nécessitent des tests complémentaires et optimisations pour devenir des têtes de série. Ces tests permettront de prioriser certaines touches par rapport à d'autres, par exemple en effectuant d'autres criblages, en évaluant leurs propriétés physico-chimiques ou encore en cherchant la présence de propriété intellectuelle les concernant.

Quand les séries chimiques sont identifiées, elles sont optimisées pour les rendre plus efficaces et moins nocives possible. Leurs propriétés d'absorption, distribution, métabolisme, excrétion et toxicité (ADMETox) sont évaluées par chémoinformatique et expérimentalement. Ensuite, des petites modifications sont effectuées de manière à générer des centaines de molécules pour chaque série chimique. Tout au long du processus de recherche, une collaboration a lieu entre les biologistes testant les molécules sur leurs cibles et les chimistes effectuant des modifications pour améliorer le résultat de ces tests.

1.3.2 Développement

L'étape du développement commence par des études précliniques dont le but est de savoir si les têtes de séries retenues pourront devenir un candidat médicament, qui sera testé sur l'être humain. Ensuite, les essais cliniques se déroulent en 3 phases. A l'issue du développement, si le médicament retenu obtient une autorisation de mise sur le marché (AMM), il sera commercialisé.

1.3.2.1 Études précliniques

Les têtes de série ayant les propriétés les plus intéressantes vont subir des tests complémentaires *in vitro* et *in vivo* pour identifier les molécules ayant une innocuité clinique, les moins dangereuses lors

d'essais cliniques menés sur les humains. Ces tests sont plus précis que pendant la dernière étape de recherche (optimisation des têtes de série) et les modèles animaux utilisés sont plus proches de l'humain. Pendant ces études précliniques, les chercheurs évaluent aussi la capacité du potentiel médicament à être produit en grande quantité. Les quelques molécules restantes, souvent une ou deux, sont des candidats médicament prêts à être testés cliniquement.

1.3.2.2 Essais cliniques

Pour mener à bien les essais cliniques, les bonnes pratiques cliniques doivent être respectées sous la responsabilité des investigateurs de l'études (médecins experts de la maladie étudiée) et de l'agence de régulation. Ces études cliniques se déroulent en 3 phases :

- Lors de la phase I, la pharmacocinétique est évaluée sur une cohorte de volontaires sains (ou malades s'il s'agit de traitement anticancéreux par exemple) pour étudier la toxicité du candidat médicament. Cette étape permet aussi de déterminer les doses permettant l'innocuité du médicament.
- Au cours de la Phase II, la pharmacodynamie du candidat médicament est évaluée. Il s'agit de comprendre le mécanisme d'action du médicament sur un groupe de patients porteurs de la maladie. L'efficacité (activité biologique) du futur médicament est évaluée, tout en continuant d'observer l'innocuité de la molécule. Le but est de connaître la dose optimale à administrer pour soigner les patients.
- Pendant la phase III, aussi appelée étude pivot, l'efficacité et l'innocuité de la molécule sont testées à plus grande échelle. Les cohortes de patients sont plus nombreuses et souvent réparties dans plusieurs centres nationaux et internationaux. Cette phase III s'approche le plus possible des conditions du traitement final. Elle permet de confirmer la posologie du médicament, d'obtenir la balance bénéfices/risques pour les patients et de comparer les résultats obtenus avec le traitement de référence, ou un placebo si le traitement est nouveau.

Une phase 0 peut parfois se dérouler en amont de la phase I, dans le cas d'essais cliniques menés sur un nouveau composé. Elle permet d'estimer au plus tôt la pharmacocinétique et la pharmacodynamique du médicament, à l'aide de faibles doses administrées à un petit groupe de patients et pendant quelques jours[49].

Pour accélérer la recherche de médicaments, les 3 premières phases peuvent être combinées deux à deux en phases I/II ou phases II/III. Si un candidat médicament passe avec succès ces 3 premières phases, il sera évalué par les autorités de santé du pays correspondant (ANSM, FDA, NMPA, …) dans le but d'obtenir une AMM et être commercialisé.

1.3.3 Commercialisation

Lorsque le médicament est commercialisé, la dernière phase des essais cliniques, la phase IV, commence. Elle évalue la sécurité du médicament (pharmacovigilance) en assurant le suivi des accidents qui pourraient être liés à la prise de ce médicament.

Depuis une trentaine d'années, lors de l'étape de recherche, des criblages HTS sont effectués avec succès pour identifier les touches à optimiser afin d'obtenir des candidats médicaments[50]. Cependant, appliqués sur certaines cibles complexes, les HTS, même virtuels, donnent peu de touches[51]. Les HTS peuvent aussi, dans certains cas, amener des faux-positifs ou des artéfacts sous la forme d'agrégats[52]. Une autre méthode de conception de médicaments s'est donc développée en parallèle des HTS. Elle se base sur les fragments moléculaires, des molécules généralement d'une masse molaire (*Molecular Weight*, MW) inférieure à 300 Daltons (Da). Nous allons maintenant en exposer tous les détails.

1.4 Conception de médicaments basée sur les fragments

1.4.1 Définition

La conception de médicaments basée sur les fragments est plus connue sous l'acronyme FBDD, abréviation de l'anglais *Fragment Based Drug Design*. Cette méthode consiste à identifier des touches par criblages de fragments, tandis que dans les criblages à haut débit (HTS) des ligands avec une MW plus élevée sont criblés. Les fragments se lient avec des affinités plus faibles que les inhibiteurs classiques sur leurs cibles, mais avec une efficacité plus élevée. En effet, ils peuvent aller plus en profondeur, dans des endroits difficilement atteignables par des inhibiteurs classiques[53] (Figure 6). De plus, en utilisant des fragments au lieu d'inhibiteurs plus grands, il y a moins de chances que des sous-structures des molécules forment des interactions non favorables avec la cible[54].

Un autre avantage de l'utilisation de la méthode FBDD par rapport au HTS est la couverture plus grande de l'espace chimique. Les chimiothèques de fragments contiennent moins de molécules que celles utilisées pour le HTS, quelques milliers contre plusieurs millions. Grâce aux nombreuses possibilités pouvant émerger des combinaisons entre ces fragments, un grand espace chimique peut être exploré plus rapidement que par HTS[55].

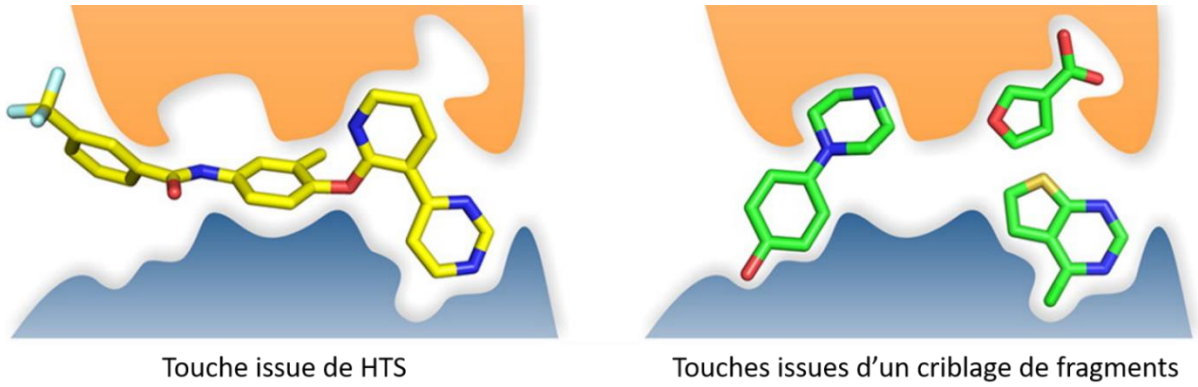


Figure 6 : Comparaison de touches obtenues par HTS ou criblage de fragments. D'après Scott et al.[53]

La méthode FBDD se décompose en quatre étapes, similaires, à l'exception de la taille des molécules, aux étapes de campagnes HTS :

1. Création d'une chimiothèque
2. Criblage sur la cible
3. Identification des meilleures touches
4. Optimisation des touches vers une tête de série

Grâce à la faible masse molaire des fragments, la dernière étape est relativement plus simple à effectuer, de même que l'optimisation des propriétés physico chimiques et du profil de toxicité des touches. Nous allons maintenant voir un historique des débuts de la méthode de FBDD, avant de détailler les étapes qui la composent.

1.4.2 Historique

En 1981, Jencks met en évidence que les énergies de liaisons entre les molécules et leurs cibles sont additives[56]. Cela marque le début des recherches effectuées à partir de fragments, bien qu'il ait été prouvé plus tard que l'ajout de nouveaux fragments peut aussi apporter un effet non additif[57]. L'énergie de liaison entre une molécule et sa cible est égale à la somme des énergies de liaison des sous-

structures de la même molécule avec la même cible[56]. Ces sous-structures doivent être composées d'au moins un groupe fonctionnel. Il s'agit de fragments moléculaires.

Une dizaine d'années plus tard, en 1992, Verlinde *et al.* tentent pour la première fois de trouver des inhibiteurs sélectifs de la triose-phosphate isomérase de *Trypanosoma brucei* par une approche basée sur les fragments[58].

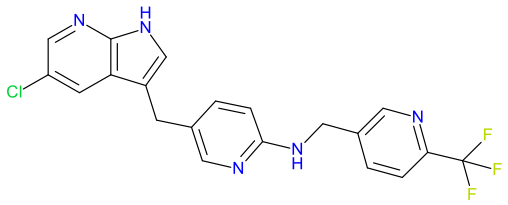
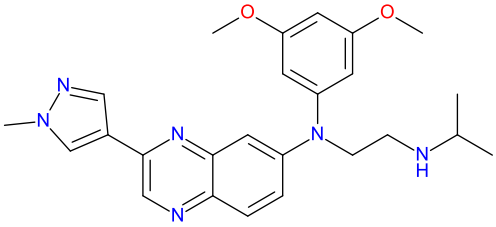
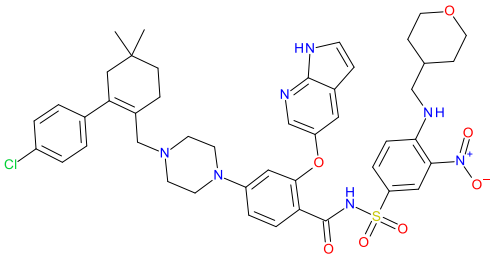
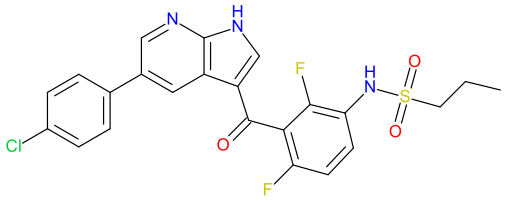
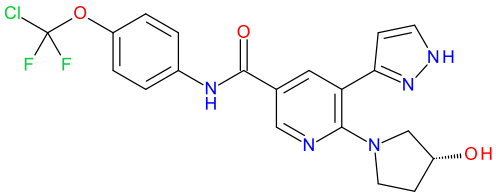
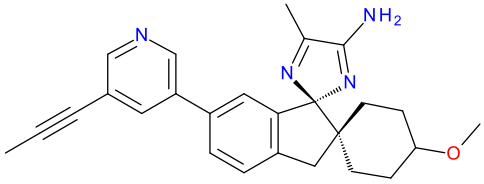
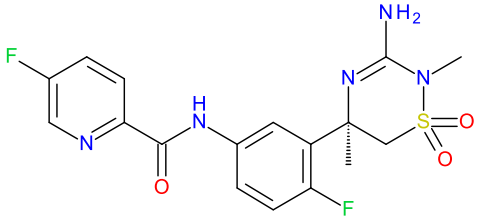
En 1996, une première approche expérimentale permet de trouver des inhibiteurs de la protéine FK506 à l'aide d'un criblage de fragments évalué par résonance magnétique nucléaire (RMN)[59]. C'est le début de nombreuses innovations sur les technologies et les principes associés à la méthode de FBDD[60]. Nous reviendrons sur celles-ci dans les prochaines sous-parties. Quinze ans après la première approche expérimentale, en 2011, le premier inhibiteur conçu par FBDD, découvert par Bollag *et al.,* est approuvé par la *Food and Drug Administration* (FDA)[61].

Les méthodes *in silico* pour la méthode de FBDD se sont développées parallèlement aux méthodes expérimentales. En 1985, GRID est le premier logiciel permettant de positionner virtuellement des groupes fonctionnels dans une cible afin de cartographier les meilleurs sites de liaison[62]. En 1991, MCSS (pour *Multiple Copy Simultaneous Search*) permet de cribler virtuellement des fragments et d'évaluer leur positionnement dans le site actif par un calcul d'énergie minimale[63]. L'année suivante, LUDI permet, en plus de positionner et d'évaluer l'emplacement des fragments, de les relier au sein d'une nouvelle molécule[64]. Comme pour son pendant expérimental, de nombreux logiciels de FBDD ont vu le jour jusqu'à aujourd'hui[65], nous les détaillerons dans la partie 1.4.6.

Les méthodes expérimentales et *in silico* sont complémentaires et permettent, utilisées ensemble, d'améliorer la recherche de nouveaux traitements. Aujourd'hui la méthode de FBDD est couramment utilisée dans les secteurs académique et privé[55]. Les dernières nouveautés concernant les fragments sont recensées chaque semaine par Erlanson et accessibles à partir du site http://practicalfragments.blogspot.com/. Il y collecte aussi l'ensemble des inhibiteurs issus de FBDD actuellement approuvés ou en cours d'essais cliniques. Au 14 avril 2021, 49 inhibiteurs y sont présents, dont 3 en phase III et 4 en phase IV (Tableau 4). Trois inhibiteurs sur les quatre en phase IV sont des inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI). Nous reviendrons en détails sur ce type d'inhibiteurs dans la partie 1.7. D. Erlanson rappelle bien que cette liste, faite manuellement, reste incomplète et peut être améliorée par les contributions dans les commentaires du blog.

L'approbation des molécules issues de FBDD peut être rapide, seulement 6 ans pour le vemurafenib et le pexidartinib. Mais dans d'autres cas, la durée correspond au temps classique de la conception de médicament. L'erdafintinib a été approuvé en 13 ans et le venetoclax l'a été en 20 ans. Après avoir vu l'historique de la méthode de FBDD, nous allons maintenant détailler les quatre étapes qui la constituent.

Tableau 4 : Molécules issues de FBDD en phase III ou phase IV des essais cliniques.

Molécule	Structure	Entreprise	Cible	Phase
Pexidartinib		Plexxikon	CSF1R, KIT	IV (08/2019)
Erdafitinib		Astex/ Janssen	FGFR1-4	IV (04/2019)
Venetoclax		AbbVie/ Genentech	BCL-2	IV (04/2016)
Vemurafenib		Plexxikon/ Roche	BRAF (V600E)	IV (08/2011)
Asciminib		Novartis	BCR-ABL1	III
Lanabecestat		Astex/ AstraZeneca/ Lilly	BACE1	III
Verubecestat		Merck	BACE1	III

1.4.3 Création d'une chimiothèque de fragments

La méthode de FBDD commence par la création d'une chimiothèque servant de point de départ pour le criblage par fragments. Une cible thérapeutique sera identifiée par des méthodes bioinformatiques (*in silico*) ou expérimentales (*in vitro, in vivo*).

Nous avons déjà évoqué que les fragments constituant cette chimiothèque sont souvent des composés d'une MW < 300 Da. De façon plus précise, un fragment peut être caractérisé par la règle de trois (*Rule of three*, RO3) indiquant les valeurs attendues pour certains descripteurs physico-chimiques[66] (Tableau 5).

Cette RO3 s'inspire de la règle de cinq (*Rule of five*, RO5) de Lipinski, qui permet d'évaluer la probabilité qu'une molécule administrée par voie orale ait une bonne biodisponibilité[67]. La RO5 repose sur les mêmes descripteurs : MW, Nombre de donneurs de liaison hydrogène (*Hydrogen Bond Donnors*, HBD), Nombre d'accepteurs de liaison hydrogène (*Hydrogen Bond Acceptors*, HBA) et le coefficient de partage octanol/eau (LogP), mais les valeurs y sont plus élevées (Tableau 5). D. Veber a ensuite amélioré cette RO5 en remettant en question la MW de 500 g/mol et en précisant les valeurs pour le nombre de liaisons rotables (*Number of Rotatable Bonds*, NRB) et la surface polaire[68] (Tableau 5). La RO3 permet d'obtenir des fragments qui, une fois optimisés, auront des valeurs de descripteurs qui tendront vers celles de la RO5. Il faut savoir que ces règles restent des recommandations et peuvent parfois ne pas être respectées, même pour des molécules ayant une bonne biodisponibilité[69]. C'est pourquoi il existe des chimiothèques de fragments ne respectant pas scrupuleusement la RO3[70]. Cela correspond aussi au fait qu'une molécule correspondant à la RO5 puisse ne pas respecter 2 de ces critères.

Tableau 5 : Valeurs des descripteurs physico-chimiques attendues dans la RO3, la RO5 et la règle de Veber

Descripteurs	RO3	RO5 et Veber
MW (g/mol)	≤ 300	≤ 500
HBD	≤ 3	≤ 5
HBA	≤ 3	≤ 10
LogP	≤ 3	≤ 5
NRB	≤ 3	≤ 10
Surface Polaire ($Å^2$)	≤ 30	≤ 140

Les fragments dans une chimiothèque ont une faible complexité moléculaire. Comme indiqué par Hann, plus une molécule est complexe, plus son affinité avec la cible sera faible[71]. Il faut donc trouver le juste milieu pour que le MW soit assez faible, tout en ayant assez de groupes fonctionnels pour que le fragment puisse se lier avec la cible[72].

Les chimiothèques expérimentales contiennent environ 1 000 à 10 000 fragments, alors que les chimiothèques virtuelles peuvent contenir jusqu'à un million de fragments. Elles doivent être assez diverses pour représenter au mieux l'espace chimique ciblé. Les chimiothèques doivent aussi être choisies en accord avec la cible, par exemple, prendre des fragments issus de PKI pour obtenir une librairie dédiée à la recherche de nouveaux PKI, avec le risque de ne pas avoir suffisamment d'originalité dans les molécules obtenues. En effet, une des deux manières d'obtenir une chimiothèque virtuelle pour la méthode de FBDD est de fragmenter des composés déjà existants. L'autre manière est d'utiliser

des chimiothèques déjà composées de fragments (commerciaux, produits naturels ou non-commerciaux).

A l'aide d'outils chémoinformatiques, les chimiothèques peuvent être combinées, standardisées et filtrées avant d'être criblées sur une cible. Les filtres sont la suppression de doublons, la suppression de composés toxiques, le choix selon certains descripteurs physico-chimiques (en appliquant la RO3 par exemple) et la sélection de composés divers de la chimiothèque. D'autres composés comportant certaines sous-structures peuvent être supprimés, notamment quand ces sous-structures amènent des liaisons non spécifiques, covalentes, des effets d'oxydo-réduction, de l'auto-fluorescence ou des dégradations. Les composés contenant ces sous-structures sont connus sous le nom de *Pan Assay INterference compoundS* (PAINS)[73]. Après avoir constitué une chimiothèque de fragments, le criblage peut être effectué sur la protéine cible.

1.4.4 Criblage de fragments

Le criblage des fragments conservés consiste à tester, pour chacun, s'il peut se lier avec la protéine cible, grâce à des méthodes *in vitro* ou *in silico*. La plupart du temps, lors des criblages, les fragments se lient de manière non spécifique sur des cibles diverses. De plus, les méthodes expérimentales nécessitent de fortes concentrations en fragments qui peuvent induire des effets d'agrégats. C'est pourquoi il faut coupler plusieurs méthodes, de manière orthogonale, pour s'assurer que les fragments retenus soient de réelles touches et non de faux positifs[74].

Tout d'abord, des méthodes de ciblages virtuels peuvent être appliquées en utilisant des chimiothèques de fragments virtuels. L'avantage est qu'elles ne nécessitent que la structure cristallographique de la protéine cible pour fonctionner. Ces méthodes *in silico* se basent sur l'amarrage moléculaire des fragments sur la protéine cible, communément appelé *docking*. Après avoir calculé toutes les positions pouvant être prises par le fragment dans la protéine, des fonctions de scores permettent de déterminer celles s'approchant le plus (théoriquement) de la réalité. Différentes revues reviennent sur les méthodes de criblages virtuels[75–77] et il existe plus de 60 logiciels d'amarrage moléculaire[78]. Bien que certains logiciels aient été développés spécifiquement pour l'amarrage de fragments, il n'y a pas de différence significative entre l'amarrage d'un fragment ou d'une molécule plus grande[79]. Il faut bien garder à l'esprit que ces méthodes restent théoriques et doivent être couplées à des méthodes expérimentales pour être validées[80].

Les méthodes utilisées pour cribler expérimentalement les fragments reposent sur la biophysique. Parmi celles-ci, il y a la résonance plasmonique de surface (*Surface Plasmon Resonance*)[81], la mesure de la variation de la température de dénaturation d'une protéine (*Thermal Shift Assay*)[82] ou la thermophorèse à micro-échelle (*Microscale thermophoresis*)[83]. Deux autres méthodes donnant des informations détaillées sur l'interaction que forment les fragments avec la protéine cible sont la cristallographie aux rayons X[84,85] et la RMN[86].

En appliquant plusieurs des méthodes citées ci-dessus, le criblage de fragments permet d'identifier des touches. Pour obtenir une tête de série, il faut trouver les meilleures touches parmi celles criblées.

1.4.5 Identification des meilleures touches

L'évaluation de l'activité biologique des fragments sur leur cible va permettre de trouver ceux à conserver pour la future optimisation vers une tête de série. Pour cela, l'efficacité du ligand (*Ligand* Efficiency, LE) peut être évaluée. Il s'agit du rapport entre l'énergie libre de liaison (énergie de Gibbs) et le nombre d'atomes lourds (Number of Heavy Atoms, NHA)[87] *cf* Equation (1). Ce calcul peut être exprimé

en utilisant d'autres termes, par exemple le pIC_{50} ou le pourcentage d'inhibition[88]. Le nombre d'atomes lourds peut aussi être remplacé par la MW ou la surface polaire totale[80].

$$LE = \frac{-RT\ln K_i}{NHA} \quad (1)$$

L'efficacité lipophile (aussi appelée *Ligand-lipophilicity efficiency*, LLE) est calculée en prenant en compte le LogP : *cf* Equation (2). Le LLE permet de privilégier les touches peu hydrophobes, car lors de l'optimisation, les groupements ajoutés sont souvent lipophiles.

$$LLE = pIC_{50} - \log P \quad (2)$$

Pour identifier les meilleures touches, il faut aussi prendre en compte la solubilité, la faisabilité de synthèse, la disponibilité d'analogues commerciaux et la présence de données structurelles sur le mode de liaison. La cristallographie aux rayons X, la cryo-microscopie électronique et la RMN donnent des informations sur le mode de liaison entre le fragment et la cible.

L'évaluation de l'activité biologique est cruciale dans la conception de médicaments. Mais dans le cas d'une conception basée sur les fragments, leur faible affinité pour la cible pose problème. En effet, de fortes concentrations de fragments sont nécessaires mais peuvent être problématiques[80]. Des méthodes *in silico* peuvent aussi être utilisées pour prédire l'activité biologique, en allant des plus anciennes telles que la relation structure-activité[89] aux plus récentes utilisant l'intelligence artificielle (IA)[90], en passant par l'amarrage moléculaire (ou *docking*).

Pour observer expérimentalement l'inhibition d'une cible par les fragments sélectionnés, une méthode comme la spectrométrie de masse peut être utilisée. Dans d'autres cas, des tests d'activité biologique basés sur la fluorescence sont utilisés. Après avoir identifié les meilleures touches, celles-ci vont être optimisées pour obtenir des inhibiteurs avec un MW plus élevé, susceptible de devenir par la suite un candidat médicament.

Après avoir identifié les meilleures touches, celles-ci vont être optimisées pour obtenir des têtes de séries.

1.4.6 Optimisation des touches vers une tête de série

L'optimisation consiste en l'ajout de nouveaux fragments ou de nouveaux substituants pour rendre la molécule plus affine de sa cible ou pour répondre à d'autres problématiques (solubilité, métabolisme, etc.). L'optimisation des touches se fait de trois manières différentes : agrandissement de fragments (*fragment growing*), liaison de fragments (*fragment linking*), ou fusion de fragments (*fragment merging*). Ces trois méthodes sont présentées dans la suite de cette sous-partie.

1.4.6.1 Agrandissement de fragments

L'agrandissement de fragments est la méthode la plus couramment utilisée en FBDD[91]. C'est, par exemple, grâce à celle-ci que le vemurafenib a été découvert[61]. Le but est d'ajouter de nouveaux atomes, groupements fonctionnels ou d'autres fragments, au fragment initial, appelé graine ou *seed*, de manière à augmenter le nombre d'interactions que le ligand forme avec la cible (Figure 7). Pour cela, il faut identifier des vecteurs d'agrandissement indiquant une direction vers laquelle grandir à partir d'une position sur la graine.

Figure 7 : Schéma représentant la technique d'agrandissement de fragments.

Adapté d'après Souza et al.[65]

Bien qu'il soit possible d'effectuer un agrandissement de fragment sans connaître les structures 3D du ligand et de sa cible, il est préférable d'en avoir connaissance pour réussir à trouver un nouvel inhibiteur. Sans information 3D, des analogues d'une graine peuvent être synthétisés et leurs activités biologiques évaluées expérimentalement. Si l'ajout d'un substituant à partir d'une position permet d'augmenter l'interaction, cette position constitue la position de départ pour l'agrandissement. Après différentes itérations de synthèse d'analogues et de tests d'activité biologique, des molécules plus actives peuvent être trouvées[92]. L'apport des données 3D permet de mieux comprendre dans quelle direction il faut agrandir les fragments et permet de savoir quel type de substituant privilégier (polaire, apolaire, etc.), par rapport aux interactions qu'il pourra former avec les acides aminés du site actif de la cible.

En utilisant la chémoinformatique, les tests d'activités biologiques sont remplacés par des scores basés, par exemple, sur l'amarrage des molécules ou par des calculs de minimisation d'énergie. Les premiers programmes, dans les années 1990, agrandissaient les molécules « atome par atome ». C'est notamment le cas des programmes LEGEND[93], GenStar[94], ou encore du dernier en date RASSE[95]. Le premier programme d'agrandissement basé sur les fragments est GroupBuild[96]. Pendant trois décennies, une trentaine de logiciels sont apparus pour concevoir des inhibiteurs à partir d'un agrandissement de fragments (Tableau 6).

Tableau 6 : Programmes permettant d'obtenir des inhibiteurs par agrandissement de fragments

Année	Programme	DOI	Auteurs
1991	LEGEND[93]	10.1016/S0040-4020(01)86503-0	Y. Nishibata and A. Itai
1991	GROW[97]	10.1002/prot.340110409	J. B. Moon and W. J. Howe
1993	SPROUT[98]	10.1007/BF00126441	V. Gillet, A. P. Johnson, P. Mata, S. Sike & P. Williams
1993	GenStar[94]	10.1007/BF00141573	S. H. Rotstein and M. A. Murcko
1993	GroupBuild[96]	10.1021/jm00064a003	S. H. Rotstein and M. A. Murcko
1993	CONCEPTS[99]	10.1002/jcc.540141008	D. A. Pearlman and M. A. Murcko
1993	[100]	10.1016/0263-7855(93)87001-L	A.W.R. Payne and R.C. Glen
1994	GrowMol[101]	10.1021/ja00092a006	R. S. Bohacek and C. McMartin
1995	DLD[102]	10.1002/prot.340230403	A. Miranker and M. Karplus
1996	RASSE[95]	10.1021/ci950277w	Z. Luo, R. Wang, and L. Lai
1996	CONCERTS[103]	10.1021/jm950792l	D. A. Pearlman and M. A. Murcko
1996	SMoG[104]	10.1021/ja960751u	R. S. DeWitte and E. I. Shakhnovich
1997	Skelgen[105]	10.1023/A:1008042711516	N.P. Todorov and P.M. Dean
2000	TOPAS[89]	10.1023/a:1008184403558	G. Schneider, M.-L. Lee, M. Stahl and P. Schneider
2000	LEA[106]	10.1023/A:1008108423895	D. Douguet, E. Thoreau and G. Grassy
2001	F-DycoBlock[107]	10.1023/A:1014817911249	J. Zhu, H. Fan, H. Liu and Y. Shi
2001	ADAPT[108]	10.1023/A:1014389729000	S. C.-H. Pegg, J. J. Haresco & I. D. Kuntz
2001	PEP[109]	10.2174/1386207013330652	N. Budin, S. Ahmed, N. Majeux, A. Caflisch
2003	SYNOPSIS[110]	10.1021/jm030809x	H. M. Vinkers, M. R. de Jonge, F. F. D. Daeyaert, J. Heeres, L. M. H. Koymans, J. H. van Lenthe, P. J. Lewi, H. Timmerman, K. Van Aken, P. A. J. Janssen
2004	CoG[111]	10.1021/ci034290p	N. Brown, B. McKay, F. Gilardoni and J. Gasteiger
2004	BOMB[112]	10.1126/science.1096361	W. L. Jorgensen
2005	LEA3D[113]	10.1021/jm0492296	D. Douguet, H. Munier-Lehmann, G. Labesse and S. Pochet
2006	Molecule Evoluator[114]	10.1021/ci050369d	E.-W. Lameijer, J. N. Kok, T. Bäck and A. P. IJzerman
2008	FOG[115]	10.1021/ci9000458	P. S. Kutchukian, D. Lou, and E. I. Shakhnovich
2009	MEGA[116]	10.1021/ci800308h	C. A. Nicolaou, J. Apostolakis, and C. S. Pattichis
2009	AutoGrow[117]	10.1111/j.1747-0285.2008.00761.x	J. D. Durrant, R. E. Amaro and J. A. McCammon
2011	EvoMD	10.1109/TCBB.2010.100	S.S.Y. Wong; W. Luo; K.C.C. Chan
2012	DOGS[118]	10.1371/journal.pcbi.1002380	M. Hartenfeller,H. Zettl,M. Walter,M. Rupp,F. Reisen,E. Proschak,S. Weggen,H. Stark and G. Schneider
2013	AutoGrow3.0[119]	10.1016/j.jmgm.2013.05.006	J. D. Durrant, S. Lindert and J. A. McCammon
2015	LEADOPT[120]	10.1016/j.ejmech.2015.02.019	G.-B. Li, S. Ji, L.-L. Yang, R.-J. Zhang, K. Chen, L. Zhong, S. Ma and S.-Y. Yang
2016	OpenGrowth[121]	10.1021/acs.jmedchem.5b00886	N. Chéron, N. Jasty, and E. I. Shakhnovich
2018	LeadOp+R[122]	10.3389/fphar.2018.00096	F.-Y. Lin, E. X. Esposito, and Y. J. Tseng
2020	AutoGrow4[123]	10.1186/s13321-020-00429-4	J. O. Spiegel and J. D. Durrant

Les molécules obtenues par ces logiciels sont parfois difficiles à synthétiser[124]. C'est pourquoi les programmes les plus récents prennent en compte la difficulté de synthèse et la rétro-synthèse pour créer de meilleures molécules.

1.4.6.2 Liaison de fragments

La liaison de fragments consiste à regrouper au sein d'une même molécule deux fragments, ou plus, déjà positionnés dans le site actif de l'enzyme ciblée (Figure 8). Ce regroupement s'effectue à l'aide d'un autre fragment nommé *linker*.

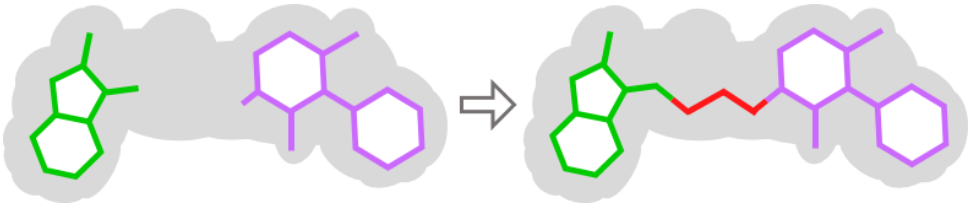


Figure 8 : Schéma représentant la technique de liaison de fragments. Adapté d'après Souza et al.[65]

Il n'est pas toujours facile de trouver expérimentalement le bon *linker*, permettant d'orienter correctement les fragments sans trop de flexibilité[80]. Cependant, le fait de regrouper des fragments au sein d'une même molécule peut amener une meilleure affinité que la somme de chaque affinité prises séparément[125].

En parallèle des programmes informatiques basés sur l'agrandissement de fragments mentionnés précédemment, une vingtaine de programmes basés sur la liaison de fragments a vu le jour (Tableau 7). Certaines méthodes de liaison permettent de remplacer la partie de la molécule sur laquelle repose la propriété intellectuelle (*scaffold*), pour obtenir un « nouveau » médicament conservant les substituants interagissant avec la protéine. Cette méthode s'appelle « *scaffold hopping* » et peut être considérée comme une sous partie des méthodes de liaison de fragments. ReCore[126] est un exemple de méthode de *scaffold hopping*.

Tableau 7 : Programmes permettant d'obtenir des inhibiteurs par liaison de fragments.

Année	Programme	DOI	Auteurs
1989	CAVEAT[127]	ISBN : 978-0-85186-796-0	E.A. Bartlett, G.T. Shea, S.J. Teller and S. Waterman
1992	LUDI[64]	10.1007/BF00124387	H.-J. Böhm
1992	Builder[128]	10.1016/0263-7855(92)80059-M	R. A. Lewis, D. C. Roe, C. Huang, T. E. Ferrin, R. Langridge and I. D. Kuntz
1992	CLIX[129]	10.1002/prot.340120105	M. C. Lawrence and P. C. Davis
1992	Linked fragment Approach[58]	10.1007/BF00129424	C.L.M.J. Verlinde, G. Rudenko and W.G.J. Hol
1993	SPLICE[130]	10.1007/BF00125322	C. M.W. Ho and G. R. Marshall
1993	NEWLEAD[131]	10.1021/jm00076a016	V. Tschinke and N. C. Cohen
1994	[132]	10.1007/BF00126746	A.R. Leach and S.R. Kilvington
1994	HOOK[133]	10.1002/prot.340190305	M. B. Eisen, D. C. Wiley, M. Karplus and R. E. Hubbard
1995	Builder v.2[134]	10.1007/BF00124457	D. C. Roe and I. D. Kuntz
1995	PRO-LIGAND[135]	10.1007/BF00117275	D. E. Clark, D. Frenkel, S. A. Levy, J. Li, C. W. Murray, B. Robson, B. Waszkowycz & D. R. Westhead
2006	Flux (1)[136]	10.1021/ci0503560	U. Fechner and G. Schneider
2007	ReCore[126]	10.1021/ci060094h	P. Maass, T. Schulz-Gasch, M. Stahl, and M. Rarey
2007	Flux (2)[137]	10.1021/ci6005307	U. Fechner and G. Schneider
2008	CONFIRM[138]	10.1007/s10822-008-9221-8	D. C. Thompson, R. A. Denny, R. Nilakantan, C. Humblet, D. Joseph-McCarthy & E. Feyfant
2008	GANDI[139]	10.1021/ci700424b	F. Dey and A. Caflisch
2010	PhDD[140]	10.1016/j.jmgm.2010.02.002	Q. Huang, L.-L. Li and S.-Y. Yang
2011	LeadOp[141]	10.1021/ci200136j	F.-Y. Lin, and Y. J. Tseng
2016	ACFIS[142]	10.1093/nar/gkw393	G.-F. Hao, W. Jiang, Y.-N. Ye, F.-X. Wu, X.-L. Zhu, F.-B. Guo and G.-F. Yang
2018	LeadOp+R[122]	10.3389/fphar.2018.00096	F.-Y. Lin, E. X. Esposito and Y. J. Tseng

D'autres programmes informatiques permettent aussi d'effectuer à la fois l'agrandissement et la liaison des fragments (Tableau 8).

*Tableau 8 : Programmes permettant d'obtenir des inhibiteurs à la fois par agrandissement ou liaison de fragments. *FlexNovo est un programme d'agrandissement via la liaison de fragments.*

Année	Programme	DOI	Auteurs
2000	LigBuilder[143]	10.1007/s0089400060498	R. Wang, Y. Gao and L. Lai
2006	FlexNovo*[144]	10.1002/cmdc.200500102	J. Degen and M. Rarey
2007	EAISFD[145]	10.1021/jm070750k	Q. Liu, B. Masek, K. Smith and J. Smith
2011	LigBuilder 2.0[146]	10.1021/ci100350u	Y. Yuan, J. Pei and L. Lai
2011	BIBuilder[147]	10.1002/minf.201000122	M. Teodoro and I. Muegge
2012	IADE[148]	10.1007/s10822-012-9609-3	P. Ertl and R. Lewis
2013	LiGen[149]	10.1021/ci400078g	A. R. Beccari, C. Cavazzoni, C. Beato and G. Costantino
2020	Ligbuilder 3.0[150]	10.3389/fchem.2020.00142	Y. Yuan, J. Pei and L. Lai

Parmi les programmes informatiques d'agrandissement ou de liaison de fragments précédemment listés, certains reposent sur des algorithmes génétiques ou des algorithmes d'évolution[89,151]. Inspiré de la théorie de l'évolution en biologie, le principe est de générer une population de nouvelles structures chimiques (enfants) à partir de mutations et recombinaisons appliquées sur la population de structures initiale (parents). A l'aide de fonctions de scores, certaines structures sont sélectionnées parmi les enfants pour être les parents de l'itération suivante. Nous pouvons par exemple citer LigBuilder[143], LEA[106], ADAPT[108], PEP[109], SYNOPSIS[110], LEA3D[113], GANDI[139], TOPAS[89], Flux[136,137], MEGA[116] et EvoMD[152] dans cette catégorie.

1.4.6.3 Fusion de fragments

La troisième et dernière manière de concevoir des inhibiteurs à partir de fragments est la fusion de fragments. Pour cela, il faut identifier deux fragments se liant à la protéine, très proches l'un de l'autre et ayant des sous-structures similaires. Ces deux fragments pourront être fusionnés à partir des sous-structures se recouvrant (Figure 9).

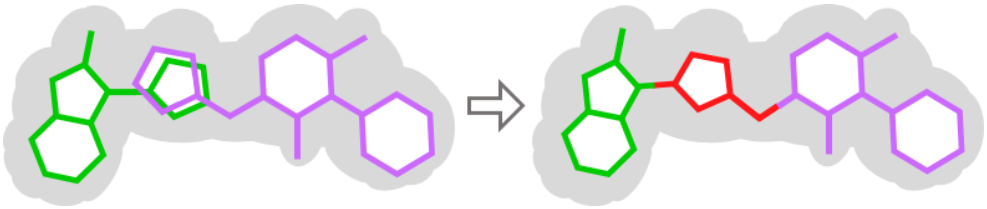


Figure 9 : Schéma représentant la technique de fusion de fragments. Adapté d'après Souza et al.[65]

La cristallographie aux rayons X et la RMN permettent de connaître les structures des fragments de manière fiable et précise, ce qui est nécessaire pour effectuer une fusion de fragments[80]. Cependant, en cas d'impossibilité d'appliquer ces techniques expérimentales, il est possible d'effectuer *in silico* une prédiction de la position des fragments par amarrage moléculaire.

Les programmes chémoinformatiques permettant d'appliquer cette méthode sont plus récents. Le premier, BREED, est apparu en 2004[153] et 4 autres programmes de fusion de fragments sont apparus depuis 2011 (Tableau 9).

*Tableau 9 : Programmes permettant d'obtenir des inhibiteurs par fusion de fragments. *GVM est un programme d'agrandissement via la fusion de fragments.*

Année	Programme	DOI	Auteurs
2004	BREED[153]	10.1021/jm030543u	A. C. Pierce, G. Rao and G. W. Bemis
2011	AutotT&T[154]	10.1021/ci200036m	Y. Li, Y. Zhao, Z. Liu and R. Wang
2012	LigMerge[155]	10.1111/j.1747-0285.2012.01414.x	S. Lindert, J. D. Durrant and J. A. McCammon
2016	AutoT&T v.2[156]	10.1021/acs.jcim.5b00691	Y. Li, Y. Zhao, Z. Liu, M. Su and R. Wang
2018	GVM*[157]	10.1021/acs.jmedchem.7b01558	F. Chevillard, H. Rimmer, C. Betti, E. Pardon,S.Ballet, N. van Hilten, J. Steyaert, W. E. Diederich and P. Kolb

La conception d'inhibiteurs *de novo* est un terme plus générique, comprenant certains outils parmi ceux précédemment cités en *in silico FBDD*. Le terme « *de novo* » signifie que la conception d'inhibiteur ne se fait à partir d'aucune donnée expérimentale[158]. Dans ce domaine, l'intelligence artificielle (IA) permet aussi, en plus des méthodes de conception d'inhibiteurs basées sur les fragments, de trouver de nouveaux inhibiteurs.

1.5 L'intelligence artificielle dans la conception d'inhibiteurs *de novo*

L'IA est un domaine de recherche dont le but est d'amener les machines à imiter les fonctions cognitives humaines telles que l'apprentissage ou la résolution de problèmes[159]. L'IA inclue différentes parties comme par exemple le traitement automatique du langage naturel, le raisonnement automatisé ou encore l'apprentissage automatique (*Machine Learning*, ML). L'organisation des différentes parties de l'IA est montrée Figure 10.

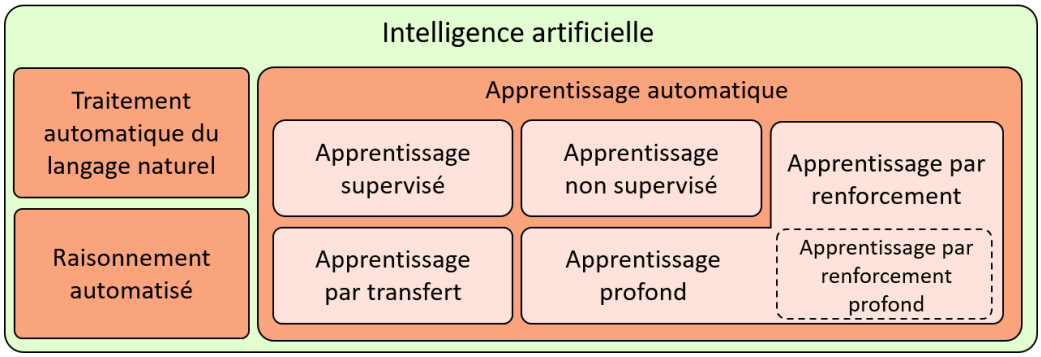


Figure 10 : Organisation de différentes parties de l'intelligence artificielle

Le ML se définit comme la capacité d'une machine à effectuer des prédictions de résultats en apprenant à identifier des modèles statistiques à partir de séries de données[160]. Les résultats peuvent, par exemple, être de nouvelles molécules dont la structure est prédite à partir d'une chimiothèque servant à l'apprentissage. Le ML a été utilisé à maintes reprises dans le domaine de conception d'inhibiteurs[161]. Le ML est un domaine se divisant lui-même en différentes sous-parties telles que les apprentissages supervisés, non-supervisés, par transfert, par renforcement (*Reinforcement Learning, RL*) et par réseaux de neurones profonds (*Deep Learning,* DL). Le RL repose sur un système de récompense obtenue si la décision prise par l'algorithme est celle qui était attendue[162]. Le DL a aussi été utilisé à de nombreuses reprises dans la recherche de médicaments[10]. Le DL est appliqué à partir de jeux de données massifs et repose sur différents réseaux de neurones artificiels faits de couches multiples[163].

L'apprentissage par renforcement profond (*Deep Reinforcement Learning,* DRL) est une combinaison de DL et de RL. Le DRL est la méthode la plus récente utilisée dans la conception d'inhibiteurs *de novo*. Elle fonctionne à l'aide d'un générateur et d'un agent. Le générateur est un réseau de neurones artificiel profond (*Deep Neural Network*, DNN), prenant en entrée des *Simplified Molecular Input Line Entry Specification* (SMILES) ou des graphes moléculaires et renvoyant les molécules générées en sortie[164]. L'agent permet d'améliorer les performances obtenues par les molécules générées en les modifiant de manière à obtenir de meilleures récompenses lors de l'apprentissage (RL). Les molécules générées sont issues d'un processus itératif d'apprentissage et d'évaluations[165] (Figure 11).

Générateur :
Réseau de neurones profonds

Agent :
Apprentissage par renforcement

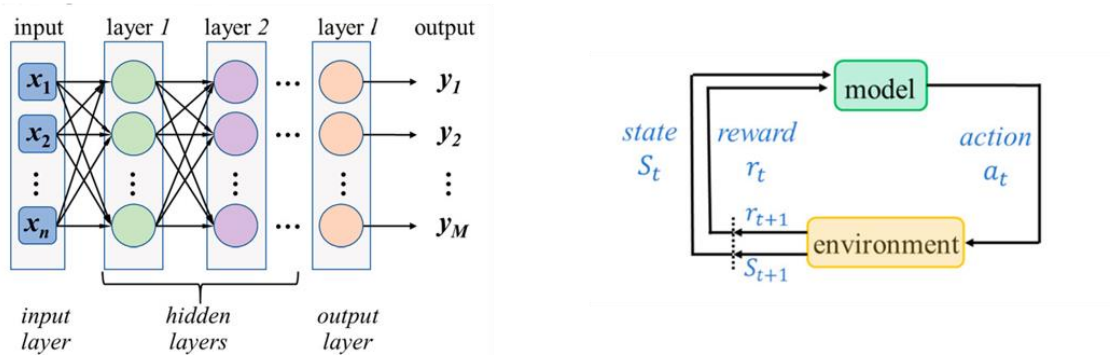


Figure 11 : Architecture de l'apprentissage par renforcement profond. D'après Mouchlis et al.[151] et Yang et al.[166]

Récemment, l'augmentation du volume de données disponibles et l'amélioration continue de la puissance des ordinateurs ont permis à certaines méthodes de DL, initiées plusieurs années auparavant, d'émerger.

Par exemple, les réseaux de neurones récurrents (*Recurrent Neural Networks*, RNN) sont des réseaux de neurones artificiels faits de connexions cycliques[167,168] (Figure 12). Les RNN utilisant du RL pour l'apprentissage, alors situés dans le DRL, ont été utilisés avec succès pour la conception d'inhibiteurs *de novo*[169–173]. Ces modèles apprennent à générer de nouvelles molécules à partir des représentations SMILES qu'ils considèrent comme une phrase dont l'enchaînement des caractères (atomes) repose sur une règle qu'ils déterminent. Les molécules générées sont faites de SMILES valides et conservant les propriétés des molécules utilisées lors de l'apprentissage[174,175]. Une utilisation des RNN en association avec des méthodes de RL a aussi été appliquée dans la conception d'inhibiteurs basée sur des fragments[176]. Les RNN peuvent aussi être utilisés en combinaison avec un apprentissage par transfert. L'apprentissage par transfert permet d'entraîner un modèle sur une première tâche, avant de l'entraîner à nouveau sur une nouvelle tâche[177]. L'utilisation de cette méthode d'apprentissage est pratique, par exemple, pour améliorer les performances des modèles utilisés pour trouver des inhibiteurs actifs contre une famille de protéines précisément[178,179].

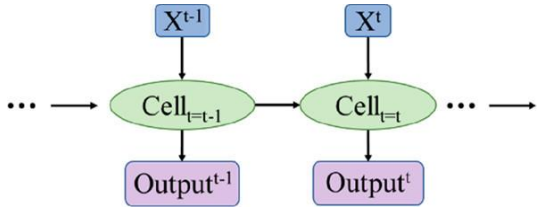


Figure 12 : Architecture d'un réseau de neurones récurrents. Yang et al.[166]

La deuxième méthode de DL utilisée pour la génération de nouvelles molécules repose sur les réseaux de neurones à convolutions (*Convolutional Neural Networks*, CNN). L'architecture de ces réseaux s'inspire du fonctionnement biologique du cortex visuel (Figure 13). L'architecture des CNN n'est pas cyclique, comme pour les RNN, mais est faite de couches de neurones, de convolutions et de regroupements[180]. Les CNN sont utilisés dans la découverte de nouveaux médicaments, par exemple pour prédire des interactions entre les médicaments et leurs cibles[181,182]. Les CNN ont aussi été appliqués pour la génération de molécules *de novo*, par exemple avec DeepScaffold qui se base sur des

molécules représentées sous forme de graphes 2D[183]. A partir du même type de représentation moléculaire, la méthode DeepGraphMolGen repose sur un CNN combiné à un RL[184].

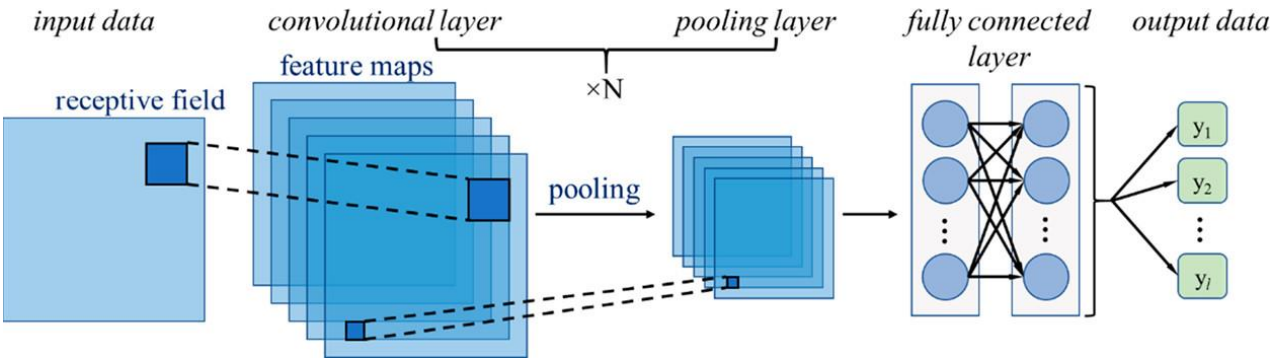


Figure 13: Architecture d'un réseau de neurones à convolutions. Yang et al.[166]

Le troisième type de méthode utilisée en DL sont les réseaux antagonistes génératifs (*Generative Adversarial Networks*, GAN). Il s'agit d'un modèle de réseaux dans lequel deux réseaux de neurones sont entraînés en même temps, l'un orienté vers la génération d'images et l'autre vers la discrimination[185] (Figure 14). Les GAN ont été développés dans le cadre de la création d'images à partir de textes, de l'augmentation de la résolution d'images ou de transformations d'une image vers une autre[186]. Pour la création de nouvelles molécules, différents types de GAN ont aussi été appliqués[187,188], incluant du RL pour certains[189].

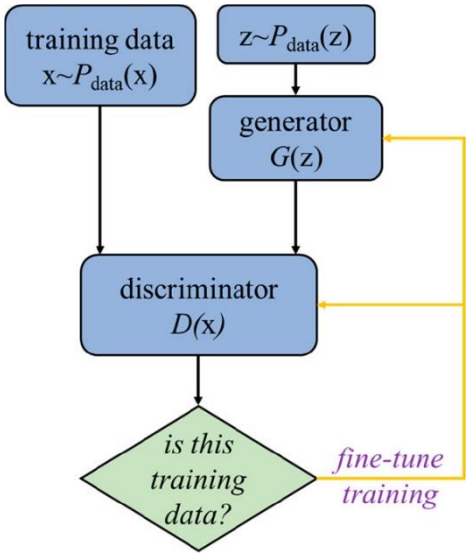


Figure 14 : Architecture d'un réseau antagoniste génératif. Yang et al.[166]

La quatrième et dernière méthode de DL permettant d'obtenir des molécules *de novo* sont les auto-encodeurs (AE). Ces réseaux de neurones apprennent, de manière non supervisée, à générer de nouvelles molécules en utilisant un encodage de molécules déjà connues. L'encodage permet de réduire la dimensionnalité de l'espace de représentation des molécules déjà connues en les représentant dans un espace latent d'un plus faible nombre de dimensions. Un décodeur est ensuite utilisé pour obtenir les nouvelles molécules à partir des représentations de molécules de l'espace latent[190,191]. Divers types d'AE ont été appliqués pour la génération de molécules *de novo* en association avec des CNN[192–194], des RNN[195,196] ou des GAN[197,198]. Les autres auto-encodeurs sont les auto-encodeurs

variationnels (*variational autoencoder*) et les auto-encodeurs contractifs (*adversarial autoencoder*). Les architectures de ces différents auto-encodeurs sont représentées en Figure 15.

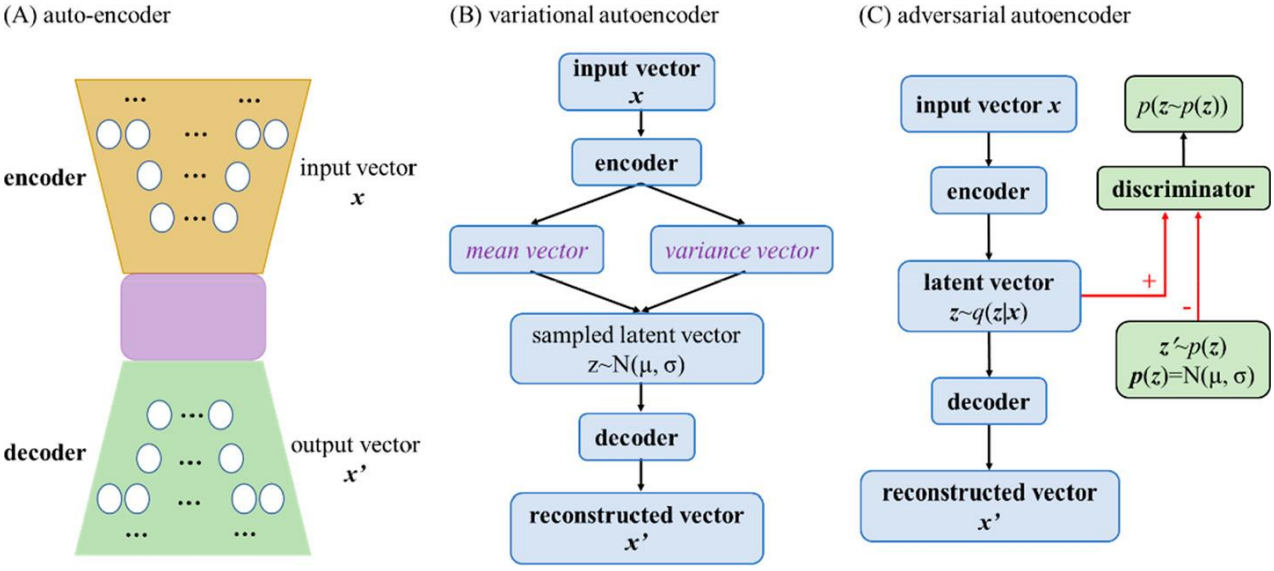


Figure 15 : Architectures des différents types d'auto-encodeurs. Yang et al.[166]

Dans cette partie, nous avons vu qu'un grand nombre de méthodes de DL viennent concurrencer les méthodes classiques de conception d'inhibiteurs basées sur les fragments. Les méthodes de DL ont même permis d'identifier de nouveaux inhibiteurs en seulement 21 jours[199]. Cependant, il existe des inconvénients à ces méthodes de DL. Elles prennent rarement en compte la représentation 3D des molécules, représentation qui est cruciale dans la méthode de conception d'inhibiteurs basée sur les fragments (Fragment-Based Drug Design, FBDD)[200]. En utilisant des méthodes de conception d'inhibiteurs sans DL, les inhibiteurs créés seront souvent plus précis et toutes les étapes lors de la génération seront comprises et maîtrisées. Cela s'oppose aux « boîtes noires » venant des couches cachées des réseaux de neurones[201]. De plus, les méthodes de DL nécessitent des jeux de données composés d'un très grand nombre d'informations. De telles quantités ne sont pas toujours accessibles dans la recherche d'inhibiteurs ciblant une famille de protéines précisément.

Après avoir vu les méthodes d'IA dans la conception d'inhibiteurs *de novo*, intéressons-nous maintenant à la famille les protéines kinases, puis à leurs inhibiteurs.

1.6 Les protéines kinases

Il est prédit que le génome humain doit coder pour 19 733 protéines, dont 17 874 (90,4%) avec une expression qui a été validée avec un haut niveau de qualité[202]. Cependant, la caractérisation, les implications et les rôles de chacune de ces protéines n'ont pas encore tous été trouvés. Selon les données venant de la base de données (BDD) UniProt[203], seulement 4 009 d'entre elles sont liées à des maladies. La BDD Drugbank[204] indique que 754 de ces protéines sont déjà la cible de médicaments approuvés par la *Food and Drug Administration* (FDA). Ce qui signifie que le nombre de cibles thérapeutiques pouvant être ciblées, selon la littérature et n'ayant pas encore d'inhibiteurs approuvés est de 1 326[205,206]. Une étude menée en 2017 par Santos et al.[207] révèle que sur les 667 protéines, alors identifiées comme cibles thérapeutiques, 127 (19%) sont des canaux ioniques, 80 (12%) sont des récepteurs couplés aux protéines G, 67 (10%) sont des protéines kinases, 20 (3%) sont des récepteurs nucléaires et les 373 (56%) restant correspondent à d'autres familles de protéines.

Les protéines kinases ne sont pas les premières de ce classement, mais ce sont les protéines les plus ciblées par les nouveaux médicaments approuvés[208]. En 2002 elles étaient considérées comme les

nouvelles cibles majeures du 21ème siècle[209]. En 2020, les protéines kinases sont toujours une cible d'intérêt, avec 8 inhibiteurs approuvés par la FDA. Nous reviendrons en détail sur les inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI) dans la partie 1.7.

1.6.1 Définition

Les protéines kinases sont une famille de protéines ayant une fonction d'enzyme. Une protéine kinase catalyse le transfert du groupement phosphate γ de l'adénosine triphosphate (ATP) sur le groupe hydroxyle de son substrat (Figure 16). L'ATP est une molécule constituée d'une base azotée adénine, d'un groupement ribose et de trois groupes phosphates nommés α, βet γ. Le groupement phosphate transféré est donc le troisième en partant du ribose (Figure 16).

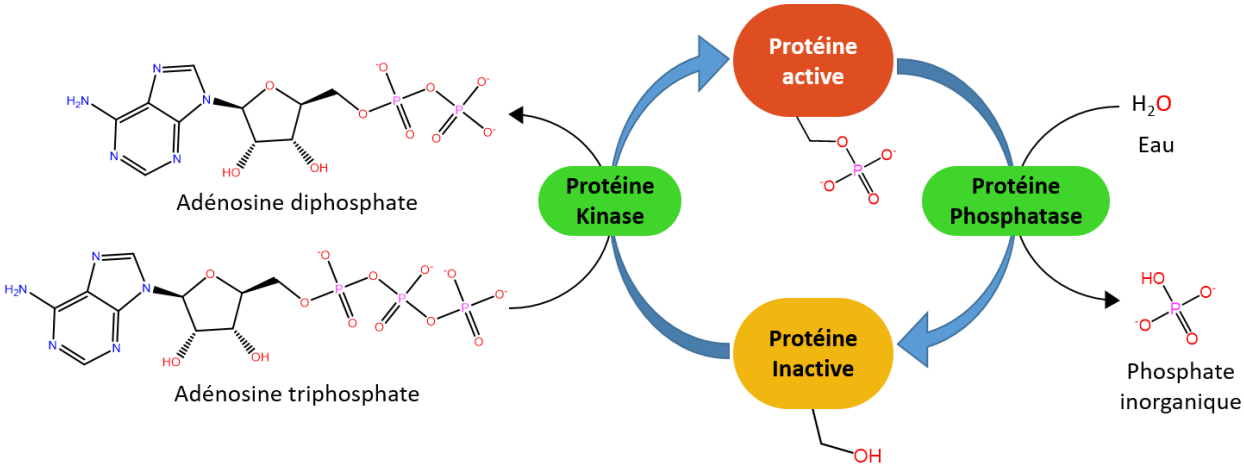


Figure 16 : Phosphorylation / déphosphorylation d'une protéine cible par une protéine kinase / phosphatase. Une fois phosphorylée, la protéine cible devient active. D'après Cohen et al.[210]

Le transfert du groupe phosphate de l'ATP sur une protéine cible s'appelle la phosphorylation. La phosphorylation est une modification post-traductionnelle que les protéines kinases effectuent le plus fréquemment pour activer leur protéine cible[210], mais la phosphorylation peut aussi rendre la protéine cible inactive. La réaction inverse, la déphosphorylation, effectuée par les protéines phosphatases, permet le plus souvent d'inactiver la protéine cible, libérant un groupement phosphate inorganique (Figure 16). La phosphorylation et la déphosphorylation sont deux mécanismes réversibles. Au niveau de la protéine cible, ces deux réactions s'effectuent sur les chaînes latérales des acides aminés contenant une fonction alcool : sérine, thréonine ou tyrosine.

Les phosphorylations faites par les protéines kinases ont des rôles clés dans la régulation de l'activité des protéines cibles impliquées dans de nombreuses fonctions cellulaires[209]. Les protéines kinases assurent le bon fonctionnement du métabolisme, de la transcription, du mouvement et des divisions cellulaires, de l'apoptose, du système immunitaire et du système nerveux[211]. La phosphorylation et les protéines kinases font l'objet de nombreuses découvertes depuis la fin du 19ème siècle. Voyons maintenant lesquelles, avant de nous intéresser aux implications pathologiques pouvant arriver en cas de dysfonctionnement de ces protéines.

1.6.2 Historique

Il y a plus de 130 ans, en 1883, une protéine phosphorylée est identifiée pour la première fois. Il s'agit de la caséine, découverte par O. Hammarsten[212]. La deuxième protéine phosphorylée, la vitelline, sera découverte en 1901 par P. A. Levene et C. Alsberg[213]. Près de 30 ans après, F. A. Lipmann et P. A. Levine trouvent que le groupement phosphate est ajouté sur une sérine de la vitelline[214]. Les autres acides aminés recevant un groupement phosphate seront ensuite identifiés en 1953 pour la

thréonine[215] et en 1979 pour la tyrosine[216]. La réaction de phosphorylation est démontrée par G. Burnett et E. P. Kennedy en 1954[217]. À la suite de ces découvertes préliminaires, en 1955, E. H. Fischer et E. G. Krebs[218,219] ainsi que E. W. Sutherland et W. D. Wosilait[220] mettent en évidence la première protéine kinase effectuant une phosphorylation à partir de l'ATP, permettant ainsi à une protéine phosphorylase de devenir plus active.

Depuis la découverte de cette famille de protéine, plusieurs prix Nobel de physiologie ou de médecine ont été attribués à des découvertes effectuées sur les protéines kinases. En 1992, E.H. Fischer et E.G. Krebs le reçoivent « pour leurs découvertes concernant la phosphorylation réversible d'une protéine en tant que mécanisme de régulation biologique »[221]. En 2000, A. Carlsson, P. Greengard et E. R. Kandel l'obtiennent « pour leurs découvertes sur la transmission du signal dans le système nerveux »[222], la protéine kinase A étant impliquée dans cette voie de signalisation. Enfin, en 2001, L. H. Hartwell, T. Hunt et P. M. Nurse gagnent le prix Nobel « pour leurs découvertes concernant la régulation du cycle cellulaire »[223], les CDK (pour *Cyclin-Dependent Kinase*, protéines kinases dépendantes des cyclines).

Les protéines kinases sont majoritairement étudiées pour leur rôle de phosphorylation, mais elles peuvent aussi avoir d'autres rôles en servant par exemple de protéines d'échafaudage, de compétiteur des interactions protéines-protéines ou en ayant des effets allostériques sur d'autres enzymes[224].

1.6.3 Implications pathologiques

Les nombreux rôles clés des protéines kinases précédemment évoqués font qu'en cas d'une activation anormale de ces dernières, des implications pathologiques peuvent apparaître. Ces pathologies sont les cancers, un dysfonctionnement du système immunitaire ou des maladies dégénératives[225].

Dans le cas des cancers, les protéines kinases ont des rôles dans de nombreux points clés de la vie des tumeurs[226]. Ces points clés sont la prolifération, la survie, la motilité, le métabolisme, l'angiogenèse, et l'échappement aux réponses immunitaires anti-tumorales[227]. En cas de dysfonctionnement du système immunitaire, les protéines kinases sont impliquées dans l'activation du système immunitaire, la synergie avec les inhibiteurs de points de contrôle des lymphocytes T, l'immunosuppression, les maladies auto-immunes et les maladies inflammatoires[225]. Concernant les maladies dégénératives, les protéines kinases ont des implications dans le stress du réticulum endoplasmique, l'angiogenèse et les maladies neurodégénératives[225].

Les protéines kinases ont de nombreuses implications pathologiques en cas de dysfonctionnement. Cibler ces protéines est alors un traitement de choix dans de nombreuses pathologies[228]. De plus, la plupart des protéines kinases sont encore peu étudiées, rendant la recherche de PKI encore intéressante[41,229–231].

1.6.4 Classification

Les protéines kinases forment une superfamille de protéines composée, chez l'Homme, de 518 membres, selon l'article de Manning *et al.* publié en 2002[228]. L'ensemble des protéines kinases est nommé kinome. Le nombre de protéines kinases chez l'Homme a été plus récemment revu à la hausse, dénombrant 555 membres en incluant les lipides kinases (protéines phosphorylant des lipides)[232]. En ne prenant en compte que les protéines kinases, le site http://kinase.com recense 529 membres dans le kinome humain. Parmi ces 529 protéines, 479 sont des protéines kinases eucaryotes et 50 sont des protéines kinases atypiques[228]. Ces 50 protéines kinases atypiques se distinguent des 479 autres par une similarité de séquence plus faible et les protéines kinases eucaryotes se classent en 9 groupes principaux composés par le rassemblement de sous-familles au sein des protéines kinases (Figure 17).

Le groupe TK (pour tyrosines kinases) contient des protéines effectuant leur phosphorylation exclusivement sur des tyrosines. Les *tyrosine kinase-like* (TKL), regroupent des familles de protéines dont la séquence est similaire aux TK mais effectuent leur phosphorylation sur des sérines ou des thréonines. Le 3ème groupe, nommé STE (pour *yeast sterile kinases*) se compose principalement de 3 familles de protéines kinases (STE7, STE11 et STE20) s'activant mutuellement pour mener à l'activation des protéines de la famille MAPK (pour *Mitogen-Activated Protein Kinases*). Le 4ème groupe, CK1 (pour *Casein Kinase 1*), est nommé ainsi pour la présence de la famille de protéines éponymes dans ce groupe. Le 5ème groupe contient des protéines kinases des familles protéines kinases A, protéines kinases G et protéines kinases C et se nomme donc AGC. Le 6ème groupe est composé d'un regroupement de CAMK (pour *Calcium/calmodulin-dependent protein kinases*). Le 7ème groupe contient des protéines kinases des familles CDK, MAPK, GSK (pour *Glycogen Synthase Kinases*) et *Cdc2-Like Kinases* (CLK) et se nomme CMGC. Le 8ème groupe contient des RGC (pour récepteurs à activité guanylate cyclase). Les familles de protéines kinases ne pouvant entrer dans ces 8 groupes sont placées dans un 9ème groupe appelé « autre ».

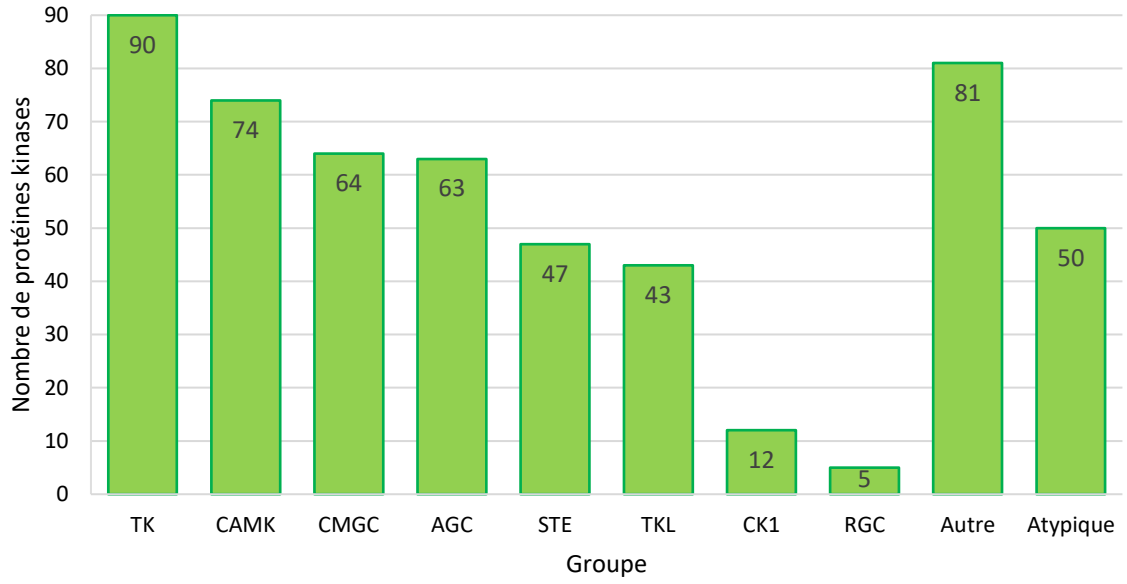


Figure 17 : Histogramme de la répartition des protéines kinases au sein des différents groupes de la classification

La classification des protéines kinases eucaryotes est montrée, à partir d'une illustration créée à partir d'un arbre phylogénétique, Figure 18. Les protéines kinases atypiques sont séparées de cette classification et sont présentées dans un encart spécifique dans la Figure 18. Malgré une faible similarité de séquence, certaines protéines kinases atypiques présentent une similarité dans leur structure tridimensionnelle avec les protéines kinases eucaryotes. En prenant en compte les lipides kinases, 26 protéines de cette catégorie sont des *protein kinase-like*[232,233] sur les 50 protéines kinases atypiques.

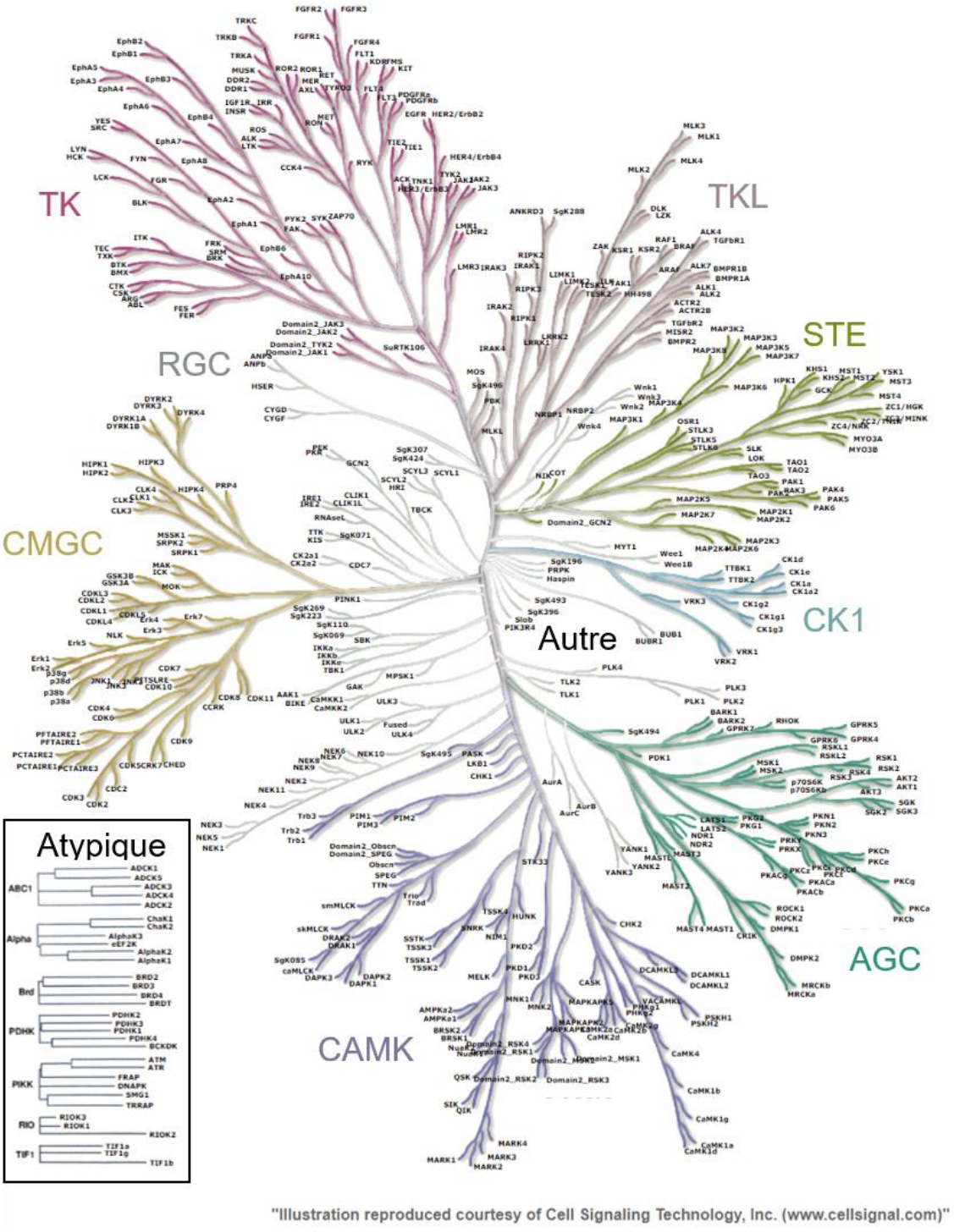


Figure 18 : Arbre phylogénétique du kinome humain. D'après Manning et al.[228] et Cell Signaling Technology[234]

Pour expliquer et visualiser la classification des protéines kinases, nous nous sommes concentrés sur les protéines humaines, mais les protéines kinases sont aussi retrouvées dans de nombreuses espèces, dont certaines présentant un intérêt scientifique ont leurs kinomes répertoriés sur la BDD KinBase[235]. Les espèces dont les kinomes sont recensés dans cette BDD sont les souris, les oursins, les drosophiles, les nématodes, les éponges *(Amphimedon queenslandica*), les coléoptères (*Meloe brevicollis*), les levures de boulanger, les champignons (*Coprinopsis cinerea*), les myxomycètes, les algues (*Tetrahymena sp.*), les parasites (*Giardia lamblia, Leishmania major, Trichomonas vaginalis*) et les lycophytes (*Selaginella moellendorffii*). En plus de présenter une grande similarité de séquence

entre-elles, les protéines kinases ont aussi une grande similarité de structure. Nous allons maintenant étudier la structure 3D de ces protéines kinases.

1.6.5 Structure 3D

Le nombre de structures 3D de protéines kinases, ajoutées à la *Research Collaboratory for Structural Bioinformatics* (RCSB) *Protein Data Bank* (PDB), augmente progressivement chaque année depuis 1993 (Figure 19). Cette année-là, une structure 3D comprenant la *protein kinase cAMP-Activated Catalytic Subunit Alpha* (PRKACA) co-cristallisée avec une molécule d'ATP, est incluse dans la RCSB PDB. Cette protéine kinase (code PDB : 1ATP[236]), est classée dans le groupe AGC chez la souris (*Mus musculus*) et a été obtenue par cristallographie aux rayons X. L'obtention de structures 3D de protéines seules ou de complexes protéine-ligand repose sur 3 méthodes principales : la cristallographie aux rayons X, la spectroscopie de RMN et la microscopie électronique.

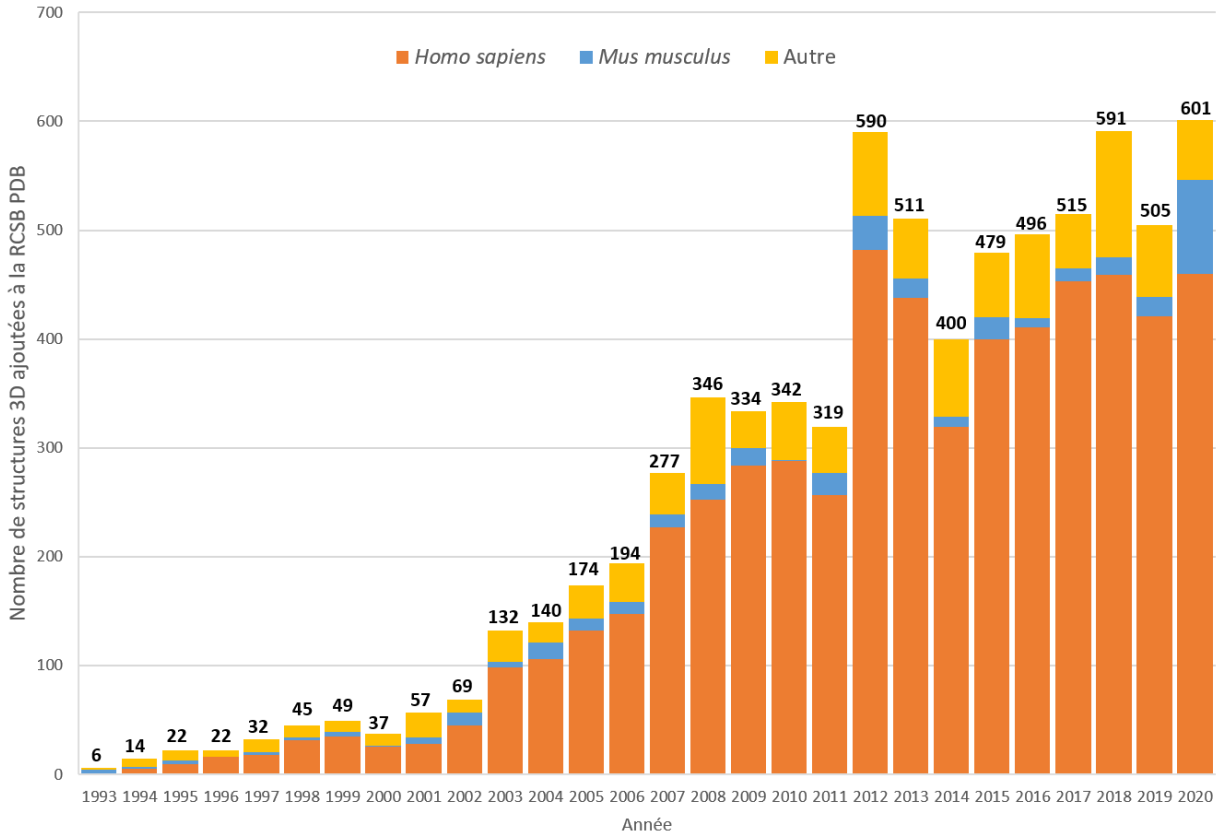


Figure 19 : Histogramme montrant le nombre de structures 3D ajoutées à la RCSB PDB par année de 1993 à 2020. Les structures cristallographiques comptabilisées ont été trouvées grâce à la base de données PFAM[237]. En prenant les identifiants PF07714 (protéine tyrosine et sérine-thréonine kinase) ou PF00069 (domaine protéine kinase).

Il existe aujourd'hui, toutes espèces confondues, plus de 7 300 structures 3D de protéines kinases dans la BDD RCSB PDB. Comme le montre la Figure 19, jusqu'au 31 décembre 2020, l'espèce *Homo sapiens* était majoritaire dans les structures de protéines kinases avec 80% (5 847) des structures ajoutées à la RCSB PDB. En deuxième position, *Mus musculus*, représentait 5% (362) des structures et toutes les autres espèces cumulées un total de 15% (1 090).

La BDD *Kinase-Ligand Interaction Fingerprints and Structures* (KLIFS) regroupe toutes les structures 3D de protéines kinases humaines (*Homo sapiens*) et de souris (*Mus musculus*)[238]. Malgré le

nombre élevé de structures 3D présentes dans la RCSB PDB pour ces 2 espèces, la BDD KLIFS indique qu'au 1er août 2020, seulement 307 protéines kinases sont représentées parmi toutes les structures 3D qu'elle référence (plus de 5 200)[239]. Il reste alors 222 protéines kinases dont la structure 3D n'est pas encore connue. Le nombre de structures 3D obtenues pour chaque protéine kinase est inégal, avec certaines protéines ayant été cristallisées plus de 150 fois (Tableau 10). Dans ce tableau, le nombre de PKI disposant d'une dénomination commune internationale (DCI) et approuvés pour les 6 protéines présentées va de 0 à 25.

Tableau 10 : Exemple de protéines kinases ayant plus de 150 structures 3D.

Protéine	Groupe	Nombre de structures 3D (humain ou souris)	Nombre de PKI en phase clinique (ayant une DCI)
CDK2	CMGC	421	3
MAPK14	CMGC	359	9
CSNK2A1	CMGC	197	1
EGFR	CMGC	187	25
AURKA	Autre	173	5
PIM1	CAMK	168	0

Cette répartition s'explique par la facilité d'accès à la structure 3D de chaque protéine kinase et aussi par l'intérêt biologique et pharmaceutique d'obtenir la structure de certaines protéines kinases. Par exemple, les cibles montrées dans le Tableau 10 ont toutes des implications en cancérologie.

Bien que la répartition du nombre de structures par protéine kinase soit inégale, les connaissances structurales sur cette famille de protéines sont vastes. Les protéines kinases bénéficient de beaucoup d'annotations et de validations facilitant leur étude.

Les protéines kinases sont toutes constituées d'un domaine kinase (ou domaine catalytique) composé de 250 à 300 acides aminés (AA). Le domaine kinase comporte 12 sous-domaines se repliant entre eux pour former la structure catalytique, typique des protéines kinases[240–243]. Ce repliement concerne à la fois les protéines kinases eucaryotes et atypiques[232].

Le domaine kinase possède deux lobes principaux (lobe N et lobe C) s'articulant autour d'une liaison charnière (*Hinge*). Le lobe N est majoritairement constitué de feuillets β, alors que le lobe C est lui surtout composé d'hélices α. La liaison charnière permet à l'ATP de se placer entre les lobes N et C, dans la poche située à proximité, en formant des interactions à partir de deux atomes donneur et accepteur de liaison hydrogènes situés sur sa base adénine[241] (Figure 20). Le premier résidu de la liaison charnière (*gatekeeper*), par l'encombrement stérique dû à sa taille et sa position, permet d'autoriser ou non l'accès aux 2 poches hydrophobes adjacentes à l'emplacement de fixation de l'adénine[244].

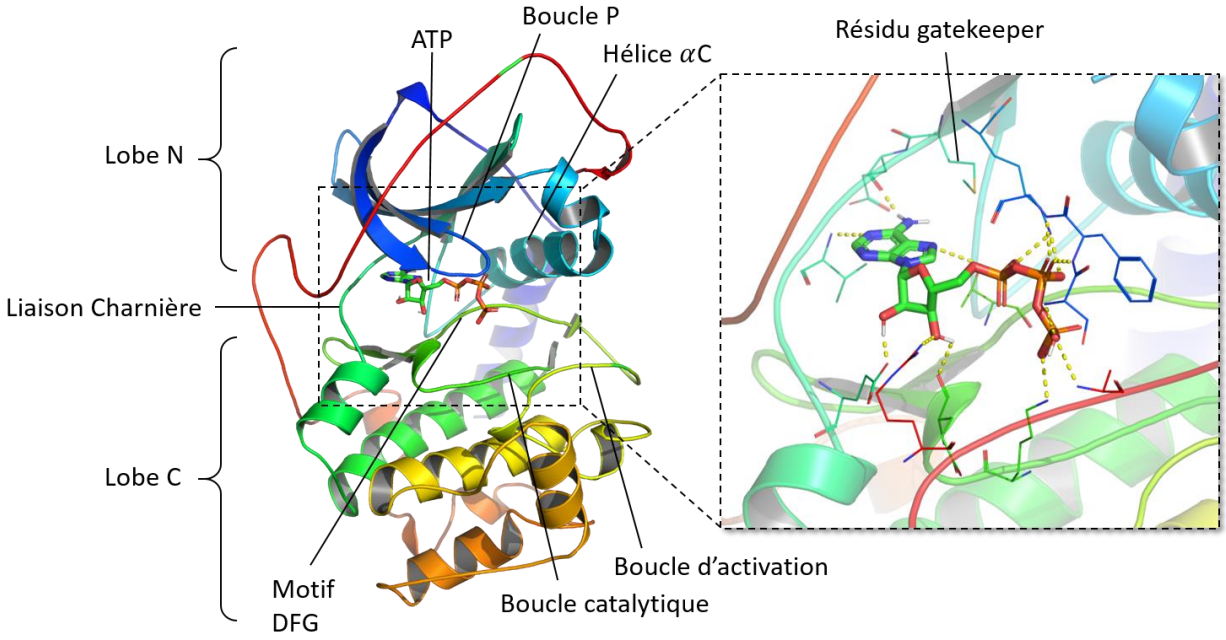


Figure 20 : Structure 3D typique d'un domaine kinase illustré par la protéine PRKACA (code PDB : 1ATP)

En plus de la liaison charnière, les autres motifs structuraux conservés entre les différents membres de la famille des protéines kinases sont le motif DFG, la boucle P, l'hélice αC, ainsi que les boucles catalytiques et d'activation. La boucle P a la particularité d'être riche en acides aminés glycine, donnant un encombrement stérique minimal et une grande flexibilité pour l'interaction avec le phosphate de l'ATP[245]. La boucle P est aussi appelée boucle G pour *glycine-rich loop*.

Le motif DFG est constitué des AA : acide aspartique, phénylalanine et glycine. Il est flexible et présente différentes conformations par rotation lui permettant d'activer ou d'inactiver la protéine kinase. Si l'AA phénylalanine du motif DFG est hors du site de liaison à l'ATP, orienté vers une poche adjacente, le motif DFG sera alors en conformation *in* (Figure 21 A). Cette conformation *in* laisse l'accès à l'ATP pour se placer dans le site actif de la protéine kinase. A l'inverse, en conformation *out*, la phénylalanine du motif DFG tourne pour se positionner dans le site de liaison à l'ATP, créant un encombrement stérique (Figure 21 A).

L'activation de la protéine kinase se fait aussi grâce à la conformation de l'hélice αC (Figure 21 B). Dans la conformation *in* de l'hélice αC, la glutamine présente sur celle-ci forme un pont salin avec la lysine catalytique. La glutamine peut aussi effectuer une rotation à 180° pour passer en conformation *out* (Figure 21 B).

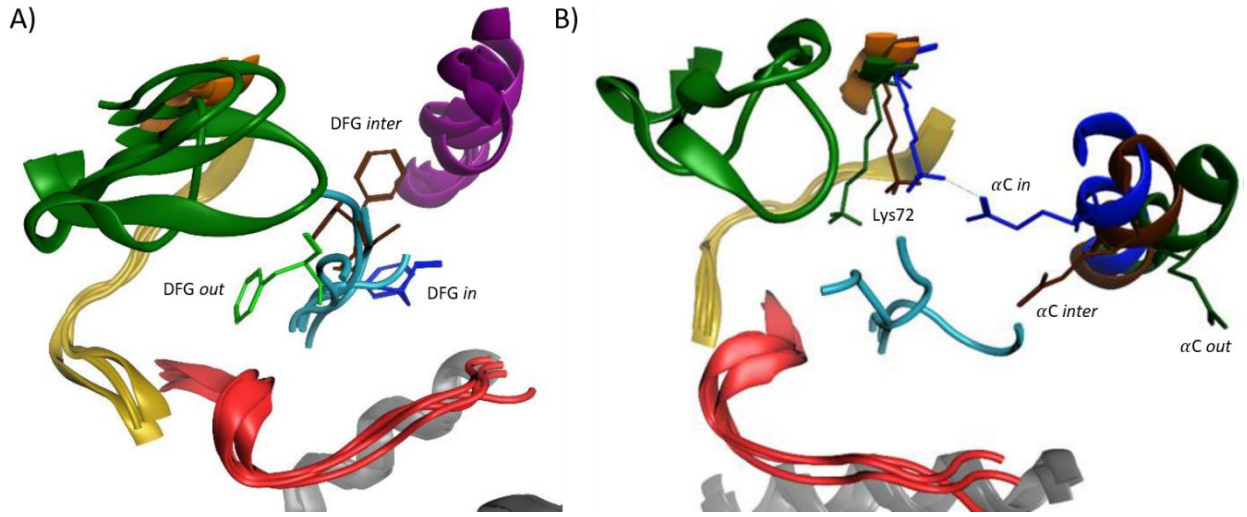


Figure 21 : Différentes conformations pouvant être prises par le motif DFG (A) et par l'hélice αC (B). La phénylalanine du motif DFG et la glutamine de l'hélice αC sont coloriées en bleu dans leur position « in », en vert dans leur position « out » et en marron dans leur position intermédiaire. D'après la thèse de Colin Bournez[2].

Pour que la protéine kinase soit active, le motif DFG et l'hélice αC doivent être tous les deux en conformation *in*. Si un des deux, ou les deux à la fois sont en conformation *out*, la protéine kinase sera inactive. Ces deux éléments structuraux peuvent aussi prendre une conformation à mi-chemin entre *in* et *out*, nommée alors conformation intermédiaire (*inter*). En observant les données de conformation indiquées par la BDD KLIFS (https://klifs.net/stats.php), il apparaît que 70,49% (3 927) des structures 3D de protéines kinases sont en conformation active et 25,15% en conformation inactive (1 401). Les 4,36 % (243) restant possèdent une conformation intermédiaire du motif DFG, et ce quelle que soit la conformation adoptée par l'hélice αC.

Tableau 11 : Nombre de structures 3D de protéines kinases (Homme et souris) pour chaque type de conformation du motif DFG et de l'hélice αC. Données obtenues à partir de la base de données KLIFS le 04/05/2021.

	DFG in	**DFG out**	**DFG inter**	**Total**
αC in	3927	428	174	4529
αC out	865	108	69	1042
Total	4792	536	243	5571

Le motif DFG joue aussi un rôle dans l'assemblage (et le désassemblage) des colonnes hydrophobes de la protéine kinase (*hydrophobic spines*), de façon directe dans la colonne dite régulatrice (*R-spine*) et de façon indirecte dans celle dite catalytique (*C-spine*)[246,247]. Ces deux colonnes hydrophobes sont constituées d'acides aminés hydrophobes, éloignés dans la séquence de la protéine, mais regroupés dans sa structure tridimensionnelle. Elles contribuent toutes deux au maintien du repliement de la protéine kinase en conformation active et participent ainsi à la phosphorylation[246–248].

Dans la structure 3D des protéines kinases, un autre motif est conservé. Il s'agit du motif HRD, constitué des acides aminés : histidine, arginine et acide aspartique. Situé sur la boucle catalytique, ce motif est important pour le transfert du phosphate de l'ATP vers le substrat[249].

Les protéines kinases contiennent des acides aminés pouvant être phosphorylés (sérine, thréonine et tyrosine) et il est très fréquent qu'une protéine kinase active une autre protéine kinase par phosphorylation. Lors d'activations anormales des protéines kinases, les pathologies déjà évoquées en partie 1.6.3 pourront apparaître. Une stratégie pour contrer ces maladies est de créer des PKI.

1.7 Les inhibiteurs de protéines kinases

1.7.1 Définition

Les inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI) sont de petites molécules interagissant avec les protéines kinases de manière à bloquer leur action. Principalement utilisés en cancérologie, ils appartiennent à la catégorie des thérapies ciblées. Parmi les thérapies ciblées, il y a d'autres types de traitements comme les anticorps monoclonaux. Les thérapies ciblées ont un mécanisme d'action différent des chimiothérapies classiques, même si les deux ont pour but d'empêcher la croissance et la progression des cellules malignes. Alors que les chimiothérapies classiques bloquent toutes les cellules en division rapide, amenant des effets cytotoxiques sur des cellules non cancéreuses, les PKI s'attaquent plus précisément à des protéines surexprimées ou mutées dans les voies de signalisation essentielles pour le développement des tumeurs[250]. Ils agissent de manière intracellulaire et permettent de réduire la toxicité des traitements, tout en ayant une meilleure efficacité[251]. Les PKI ont aussi l'avantage d'être administrés oralement. Cependant, les thérapies ciblées peuvent aussi amener des complications, en s'attaquant à d'autres cellules du corps[252]. En effet, la grande similarité structurale entre les différentes protéines kinases rend difficile la recherche de molécules sélectives. Cela s'explique aussi par les cibles thérapeutiques qui peuvent être parfois présentes dans des cellules saines.

Les PKI peuvent agir de différentes manières[253] :

- Par inhibition compétitive de l'adénosine triphosphate (ATP)
- En se plaçant à la place du substrat, par inhibition allostérique
- En se localisant dans d'autres poches, bloquant la protéine kinase dans une conformation non canonique.
- Par inhibition covalente

Les inhibiteurs de protéines kinases se classent ainsi en 7 types selon leur mode d'inhibition[254].

1.7.2 Catégories

Les PKI de type I se placent dans le site de fixation de l'ATP, près de la liaison charnière, là où l'adénine de l'ATP se positionne pour se lier par le biais d'interactions hydrogène. Les protéines kinases doivent être en conformation active (DFG *in*, hélice αC *in*) pour être inhibées par les PKI de type I[255].

Les PKI de type $I^{1/2}$ entrent aussi en compétition avec l'ATP, mais sont plus longs que les inhibiteurs de type I, s'étendant dans une poche hydrophobe adjacente située derrière l'emplacement de fixation de l'adénine. Pour lier des inhibiteurs de type $I^{1/2}$, les protéines kinases adoptent une conformation inactive (DFG *in*, hélice αC *out*)[256].

Les PKI de type II effectuent, comme les 2 premiers types, des interactions avec la liaison charnière, et accèdent, comme les PKI de type $I^{1/2}$, à la poche hydrophobe placée derrière le site de fixation de l'ATP. Les PKI de type II se lient en conformation inactive (DFG *out*, hélice αC *out*)[255].

Les PKI de type III n'entrent pas en compétition avec l'ATP, ils sont appelés « inhibiteurs allostériques ». Bien que certains de ces inhibiteurs allostériques aient été découverts en cherchant des inhibiteurs compétitifs de l'ATP[257], ils se placent dans le site allostérique adjacent au site de liaison à l'ATP[258].

Les PKI de type IV agissent également par inhibition allostérique en se plaçant dans un autre site que le site actif de la protéine kinase[258]. Ils sont appelés « vrais inhibiteurs allostériques », par

opposition aux inhibiteurs de type III. Les sites de liaison de ce type d'inhibiteurs sont divers et dépendent des protéines kinases ciblées.

Les PKI de type V sont des inhibiteurs bivalents pouvant se lier dans 2 régions distinctes du domaine kinase[259]. Ils peuvent par exemple inhiber la protéine kinase en se liant à la fois dans le site de liaison de l'ATP et dans la poche du substrat.

Les PKI de type VI sont des inhibiteurs formant une liaison covalente avec la protéine kinase ciblée[254]. Ces liaisons peuvent se former avec les acides aminés : cystéine, lysine ou tyrosine. À la différence des autres types d'inhibiteurs, qui forment des liaisons réversibles, ce type d'inhibiteur se lie par une liaison covalente.

1.7.3 PKI déjà connus

Certains PKI sont des dérivés de produits naturels comme la staurosporine dont l'activité sur la protéine kinase PKC a été prouvée en 1986[260]. Le fasudil est le premier PKI approuvé en 1995 par la *Pharmaceuticals and Medical Devices Agency* (agence du médicament du Japon). Le premier PKI autorisé à être mis sur le marché par la *Food and Drug Administration* (FDA) en 2001 est l'imatinib. Les structures 2D de ces trois molécules sont montrées en Figure 22.

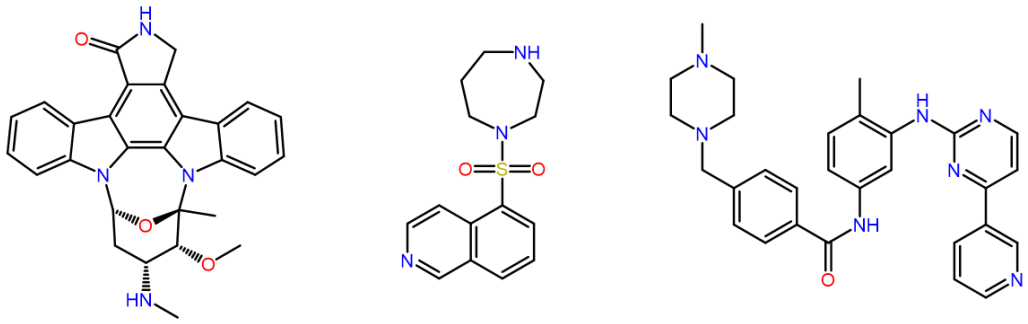


Figure 22 : Structures 2D des premiers inhibiteurs de protéines kinases.

L'imatinib a été développé par la société Novartis pour soigner la leucémie myéloïde chronique. À la suite de l'arrivée sur le marché de l'imatinib, 71 autres PKI ont été approuvés et 188 PKI ayant une dénomination commune internationale (DCI) sont en cours d'essais cliniques en mai 2021. L'ensemble des inhibiteurs actuellement sur le marché et en cours d'essais cliniques et disposant d'une DCI peuvent être consultés sur la BDD *Protein Kinase Inhibitor Database* (PKIDB, https://www.icoa.fr/pkidb/)[34,35].

Les PKI déjà sur le marché sont principalement dérivés de pyridine, pyrimidine, quinolone, quinazoline ou carboxamide[35,251]. Ils sont, le plus souvent, compétitifs de l'ATP (types I, $I^{1/2}$ et II) et moins efficaces lors de l'apparition de mutations amenant des résistances à ces traitements[261]. C'est par exemple le cas de la protéine kinase de fusion ciblée par l'imatinib (*Breakpoint Cluster Region-Abelson*, BCR-ABL1) pouvant muter sur son acide aminé *gatekeeper*. En cette position 315 une thréonine peut être modifiée en une isoleucine (T315I), empêchant l'imatinib d'accéder à la poche hydrophobe juxtaposée à l'emplacement de liaison de l'ATP[262]. Dans ce cas, de nouveaux inhibiteurs sont trouvés pour cibler la protéine kinase résistante au traitement initial. Par exemple, le ponatinib peut cibler spécifiquement la mutation T315I de la protéine BCR-ABL1.

L'article présenté à la fin de ce premier chapitre a été rédigé en décembre 2019. Il proposait une mise à jour des inhibiteurs approuvés par la FDA à cette date. En mai 2021, 13 nouveaux PKI ont été approuvés par la FDA : 9 inhibiteurs en 2020 et 4 PKI en 2021 (Figure 23). Parmi ces inhibiteurs, le type

d'inhibition du tirbanibulin n'a pas encore été déterminé, le selumetinib est un inhibiteur de type III, le ripretinib et le tivozanib sont respectivement des inhibiteurs de type II et les 9 autres sont de type I.

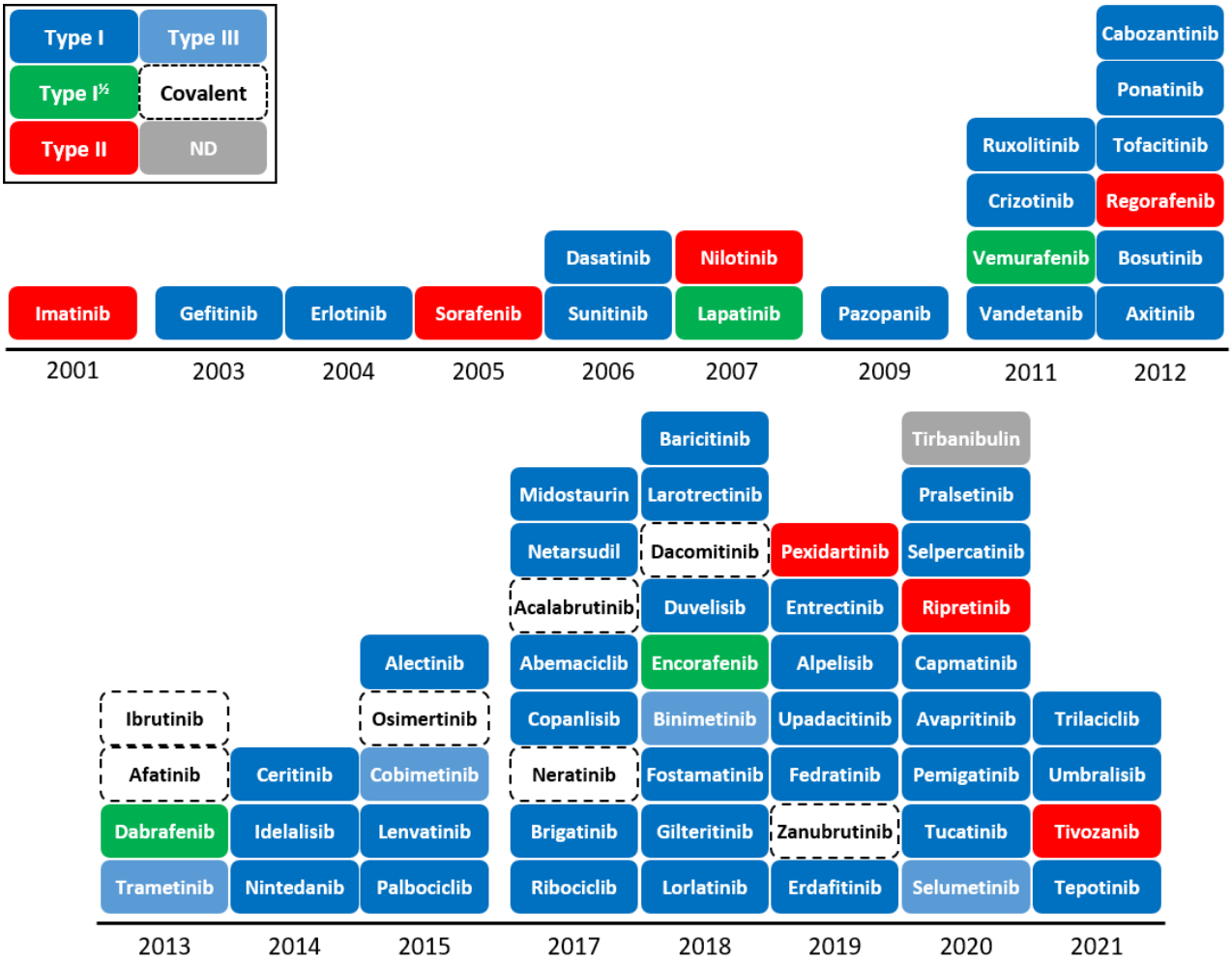


Figure 23 : Liste et types des inhibiteurs de protéines kinases approuvés par la FDA classés par année. D'après Attwood et al.[263]

Bien que l'espace chimique des PKI soit déjà encombré, l'ensemble des inhibiteurs de protéines kinases ne ciblent qu'une faible proportion du kinome. Sur les 212 protéines kinases potentiellement impliquées dans des pathologies recensées par Manning *et al.*[228], 46,7% (99) des protéines sont des cibles de PKI actuellement en cours d'essais cliniques ou sur le marché (Figure 24). Cela laisse encore l'opportunité de découvrir de nouvelles molécules inhibitrices intéressantes sur le plan pharmacologique. Par ailleurs, la découverte de PKI peut parfois générer un chiffre d'affaire annuel de plus d'un milliard de dollars aux laboratoires pharmaceutiques les commercialisant. C'est le cas de l'imatinib qui, dès 2003, a rapporté 1,128 milliards de dollars et 1,168 milliards de dollars l'année suivante[264].

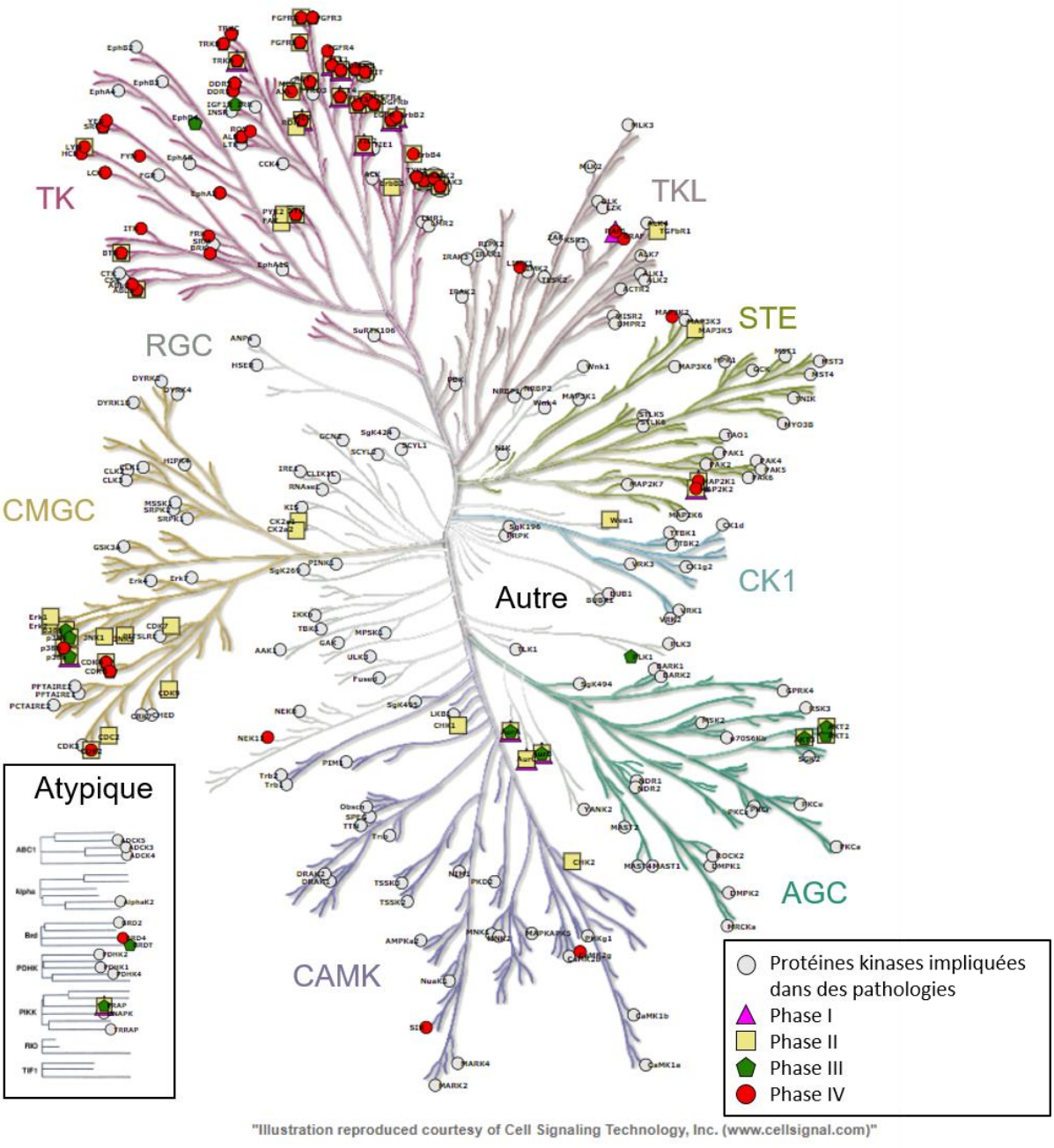


Figure 24 : Représentation des protéines kinases impliquées dans des pathologies et des cibles des PKI en phases cliniques et sur le marché. Figure générée avec l'outil KinMap (http://www.kinhub.org/kinmap/index.html)[265]. D'après Attwood et al.[263]

1.8 Étude des *scaffolds* des inhibiteurs de protéines kinases

Comme évoqué précédemment, les inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI) ayant une dénomination commune internationale actuellement approuvés ou en essais cliniques ont déjà été recensés par l'équipe SB&C de l'ICOA et regroupés au sein d'une base de données (BDD) fréquemment mise à jour[34]. J'ai pu contribuer à la rédaction d'un article intitulé « *Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB* » publié dans le journal *Molecules*[35]. Cet article présente une mise à jour des inhibiteurs de PKIDB, les *scaffolds* les plus fréquents des PKI, ainsi qu'une comparaison entre ceux de PKIDB et ceux trouvés dans la *Chemical database of the European Molecular Biology Institute* (ChEMBL). Il est intégré à cette thèse, après la conclusion de ce chapitre.

Un autre point clé de cet article est de fournir des recommandations à suivre pour les différents descripteurs des PKI. Ainsi, des recommandations, à la manière de celles de Lipinski ou de Veber, peuvent être appliquées lors de la recherche d'inhibiteurs spécifiques des protéines kinases. Les valeurs à respecter sont extraites de l'article et reportées dans le Tableau 12. Pour chaque descripteur, les valeurs ont été déterminées en conservant celles qui se situaient à moins de 2 écarts-types de la moyenne (95.4% de l'intervalle de confiance)[35].

Tableau 12 : Recommandations de valeurs de descripteurs à suivre pour filtrer des molécules de type PKI. D'après Bournez et al.[35]. Légende : Masse molaire (Molecular Weight, MW), surface topologique accessible au solvant (TPSA), le coefficient de partage octanol/eau calculé (ClogP), nombre de d'accepteurs et de donneurs de liaison hydrogène (HBA et HBD), nombre de liaisons rotables (NRB), nombre de cycles aromatiques (NAR), Nombre d'atomes chiraux (NCA).

	Minimum	Maximum
MW (g/mol)	314	613
TPSA (Å^2)	55	138
ClogP	0.7	6.3
HBA	3	10
HBD	0	4
NRB	1	11
NAR	1	5
NCA	0	2

Article

Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB

Colin Bournez [1], Fabrice Carles [1], Gautier Peyrat [1], Samia Aci-Sèche [1], Stéphane Bourg [1], Christophe Meyer [2] 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; colin.bournez@univ-orleans.fr (C.B.); fabrice.carles@univ-orleans.fr (F.C.); gautier.peyrat@univ-orleans.fr (G.P.); samia.aci-seche@univ-orleans.fr (S.A.-S.); stephane.bourg@univ-orleans.fr (S.B.)

2 Janssen-Cilag, Centre de Recherche Pharma, CS10615 - Chaussée du Vexin, 27106 Val-de-Reuil, France; cmeyer2@its.jnj.com

3 Correspondence: pascal.bonnet@univ-orleans.fr; Tel.: +33-238-417-254

Academic Editor: Christian Peifer
Received: 17 June 2020; Accepted: 10 July 2020; Published: 15 July 2020

Abstract: Since the first approval of a protein kinase inhibitor (PKI) by the Food and Drug Administration (FDA) in 2001, 55 new PKIs have reached the market, and many inhibitors are currently being evaluated in clinical trials. This is a clear indication that protein kinases still represent major drug targets for the pharmaceutical industry. In a previous work, we have introduced PKIDB, a publicly available database, gathering PKIs that have already been approved (Phase 4), as well as those currently in clinical trials (Phases 0 to 3). This database is updated frequently, and an analysis of the new data is presented here. In addition, we compared the set of PKIs present in PKIDB with the PKIs in early preclinical studies found in ChEMBL, the largest publicly available chemical database. For each dataset, the distribution of physicochemical descriptors related to drug-likeness is presented. From these results, updated guidelines to prioritize compounds for targeting protein kinases are proposed. The results of a principal component analysis (PCA) show that the PKIDB dataset is fully encompassed within all PKIs found in the public database. This observation is reinforced by a principal moments of inertia (PMI) analysis of all molecules. Interestingly, we notice that PKIs in clinical trials tend to explore new 3D chemical space. While a great majority of PKIs is located on the area of "flatland", we find few compounds exploring the 3D structural space. Finally, a scaffold diversity analysis of the two datasets, based on frequency counts was performed. The results give insight into the chemical space of PKIs, and can guide researchers to reach out new unexplored areas. PKIDB is freely accessible from the following website: http://www.icoa.fr/pkidb.

Keywords: protein kinase inhibitors; clinical trials; approved drugs; database; chemometrics analysis; kinome; molecular scaffolds; rings system

1. Introduction

The reversible phosphorylation of proteins plays a preeminent role in cell cycle regulation. This process, which consists in the transfer of a phosphoryl group PO_3^{2-} to the target substrate, is catalyzed by enzymes pertaining to the protein kinase family. Protein kinases constitute one of the largest protein families encoded by the human genome and counts 518 members (or 538 members when atypical kinases are included) [1–3]. Numerous studies have shown that deregulation or mutation of kinases is responsible for a variety of cancers [4], as well as for other diseases in the immune or neurological area [5,6]. The majority of protein kinases, however, have not yet been fully explored [7], and there is still a high potential for innovation in targeting the protein kinome for the treatment of cancer. The Food and Drug Administration (FDA) approved 55 small-molecule protein kinase inhibitors (PKIs) by end of 2019,

whereas the Chinese and European regulatory authorities have granted market access to five more compounds, namely anlotinib, apatinib, icotinib, fasudil, and tivozanib (Figure 1). It is worth mentioning that higher molecular weight inhibitors like macrocyclic lactones, such as sirolimus and temsirolimus, or kinase-targeted antibodies, such as cetuximab and trastuzumab, have been approved for the treatment of colorectal, head/neck, and breast cancers, respectively [8–10]. These large molecules were excluded from this study, which focuses on small-molecule PKIs targeting the kinase domain. The first PKI approved by the FDA was imatinib in 2001. Imatinib is a small-molecule type-II inhibitor containing a phenylamino-pyrimidine scaffold. It targets the inactive conformation of ABL1 kinase and is used against chronic myelogenous leukemia (CML) [11]. Since then, at least one new PKI reaches the market every year, with a significant acceleration since 2011. The exceptions to this rule are 2002, 2008, 2010, and 2016, with no compound approved in these years.

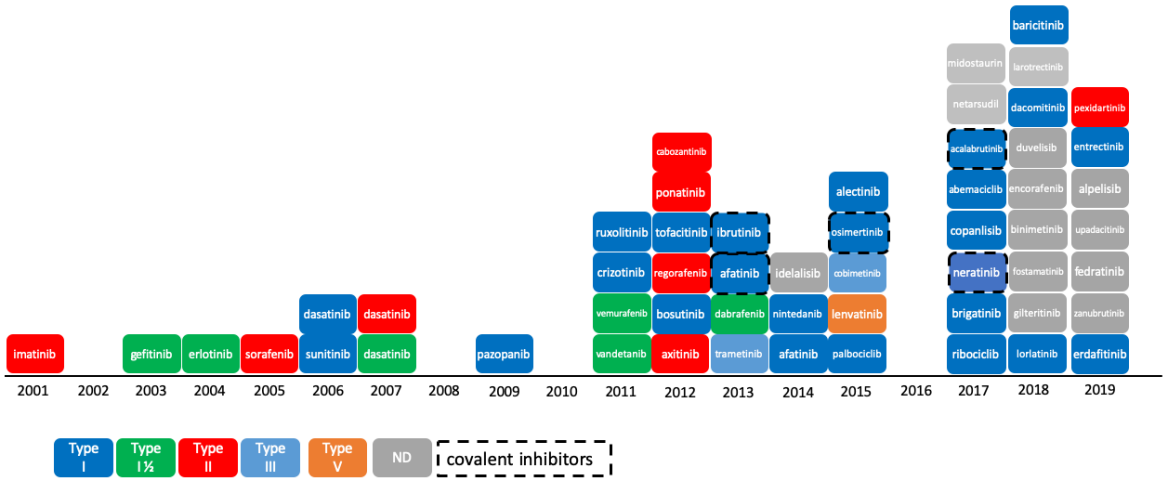


Figure 1. Progression of Food and Drug Administration (FDA)-approved protein kinase inhibitors (2001–2019) and their type of inhibition [12]. As of 11th December 2019, 55 kinase inhibitors were approved by the FDA. Not shown here: tivozanib, approved by European Medicines Agency (EMA) in 2017; anlotinib, apatinib, and icotinib, approved by the China Food and Drug Administration (CFDA) in 2018, 2014, and 2011, respectively; and fasudil, approved in China and in Japan in 1995. ND: not defined.

In addition to approved PKIs, many novel compounds are currently being evaluated in clinical trials throughout the pharmaceutical industry. Taken collectively, these compounds show new trends in terms of structures, physicochemical properties, and biological activities that foreshadows changes in the PKI landscape. To collect and organize this data as well as keep up to date with their evolution, we developed PKIDB [12], a curated, annotated and updated database of PKIs in clinical trials. In order to enter the PKIDB, compounds should be currently in one development phase (from Phase 0 to Phase 4), have a disclosed chemical structure, as well as an International Nonproprietary Name (INN) [13]. Each compound is provided with comprehensive descriptive data, as well as with links to external databases such as ChEMBL [14], PDB [15], PubChem [16], and others. The type of binding mode specified in PKIDB has been manually entered and comply with Roskoski's classification [12]. The database is freely accessible on a dedicated website (http://www.icoa.fr/pkidb). As of 11th of December 2019, it contains 218 inhibitors: 60 approved and 158 in various stages of clinical trials (from Phase 0 to Phase 3).

In this study, we compared PKIDB to a large dataset of 76,504 PKIs retrieved from ChEMBL (referred herein as "PKI_ChEMBL", see the Materials and Methods section). The objective is to be able to better select PKIs from public databases based on structural and physicochemical property information of PKIs already in clinical trials. Firstly, we performed a principal component analysis (PCA), and compared the projection of both datasets in a common factorial space. We also assessed the structural shape diversity of PKIs using a principal moments of inertia (PMI) analysis. Secondly, in addition to comparisons at the global molecular structure level, we performed a substructure analysis based on PKI scaffolds. In medicinal chemistry, scaffolds are mostly used to represent core structures of

bioactive compounds. They are relevant for the medicinal and/or computational chemists, and have proved to be useful in the identification of "privileged substructures" [17] in "scaffold hopping" [18] or in structure–activity relationships (SAR) analyses [19]. The concept of the scaffold was first defined by Bemis and Murcko, as frameworks that consist of rings and linkers, from which substituents are removed [20]. From these scaffolds, two levels of abstraction were derived: the heteroatom framework and the graph representation. The heteroatom framework only takes into account the atom type, without considering bond types or aromaticity, whereas the graph representation (also known as cyclic skeleton) turns every atom type to carbon and every bond type to a single bond, reducing the initial molecule to a simple graph [21]. Finally, unfused rings present in the molecules are separated by removing their connecting bonds.

The balance between the molecular diversity of scaffolds, and their frequency is an important parameter in a chemical database. A high frequency associated to a small number of scaffolds corresponds to a focused library composed of structurally similar molecules, bearing varying substituents. Contrarily, a low frequency associated to a large number of scaffolds reflects a high molecular diversity [12]. Thus, this criterion needs to be addressed when designing or selecting a chemical library depending on its forecasted usage. We assessed scaffold diversity for the PKIDB and PKI_ChEMBL datasets using the molecular Bemis and Murcko scaffolds and cyclic skeleton. The most represented scaffolds (frequency) and the comparison of their distribution within the two studied datasets are presented. Finally, an analysis of the rings of all molecules was performed. We first considered all the rings devoid of substituent (first attached atoms were replaced by hydrogen atoms). Then, we encoded the rings while retaining the position and atom type of their original substituents. This scaffold diversity analysis reflects the chemical space of PKIs and can be useful for the medicinal chemistry community to reach out new unexplored areas**.**

2. Results

2.1. Update on PKIDB

The description and analysis of PKIDB have been reported in a previous study by Carles et al. [22]. Referencing 218 molecules the 11th December 2019, PKIDB contains 38 more inhibitors (from phase 0 to phase 4) than the first release (abivertinib, adavosertib, alvocidib, asciminib, avapritinib, bemcentinib, berzosertib, bimiralisib, capivasertib, ceralasertib, derazantinib, dezapelisib, enzastaurin, fasudil, lazertinib, leniolisib, mavelertinib, midostaurin, nazartinib, neflamapimod, nemiralisib, netarsudil, ningetinib, parsaclisib, pralsetinib, ravoxertinib, ripasudil, ripretinib, rivoceranib, rogaratinib, ruboxistaurin, samotolisib, sotrastaurin, tomivosertib, umbralisib, vactosertib, verosudil, zanubrutinib).

Among these 38 compounds, nine were FDA-approved in 2017, eight in 2018, and seven in 2019. Fasudil, a ROCK inhibitor, approved in China and in Japan in 1995 was therefore the first kinase inhibitor that reached the market, but it is not FDA-approved. Those compounds were automatically added to PKIDB database thanks to their name stem. Indeed, since the first release of PKIDB, the INN made an update on the stems that assign the molecules with the "aurin" and "udil" suffixes to the kinase inhibitor class. Moreover, the stem 'cidib' was also updated and has been replaced by 'ciclib' (see cumulative USAM stem list from AMA [23]). However, we also kept the stem 'cidib' to retrieve information on alvocidib, not yet referenced as alvociclib.

In addition to those compounds, Table 1 gathers the eight and seven PKIs that reached phase 4 and were FDA-approved in 2018 and 2019, respectively. Among those 15 PKIs, all were previously in a phase lower than 4 in our database, except zanubrutinib, which was not in the first release. One should note that FDA recently approved avapritinib, a selective dual KIT and PDGFRα inhibitor, after the updated version of PKIDB, and is therefore not considered in this study.

This brings the total number of approved drugs on the market referenced in our database to 60. As described in PKIDB, most of the PKIs are targeting more than one protein kinases, and since the first version of PKIDB, new targets have emerged, such as the Wee1-like protein kinase inhibited by adavosertib [24].

Table 1. Protein kinase inhibitor (PKIs) approved in 2018 and 2019 with their respective targets extracted from DrugBank (Uniprot ID extracted from https://www.uniprot.org/.).

PKI	Unitprot ID	Gene Name
Alpelisib	P42336	PI3KCA
Binimetinib	Q02750	MAP2K1
Dacomitinib	P00533	EGFR
Duvelisib	O00329	PI3KCD
	P48736	PI3KCG
Encorafenib	P15056	BRAF
Entrectinib	P04629	NTRK1
	Q16620	NTRK2
	Q16288	NTRK3
	P08922	ROS1
	Q9UM73	ALK
Erdafitinib	P11362	FGFR1
Fedratinib	O60674	JAK2
	P36888	FLT3
	O60885	BRD4
Fostamatinib	P43405	SYK
Gilteritinib	P36888	FLT3
	P30530	AXL
	Q9UM73	ALK
Larotractinib	P04629	NTRK1
	Q16620	NTRK2
	Q16288	NTRK3
Lorlatinib	Q9UM73	ALK
	P08922	ROS1
Pexidartinib	P36888	FLT3
	P10721	KIT
	P07333	CSF1R
Upadacitinib	P23458	JAK1
Zanubrutinib	Q06187	BTK

2.2. Physicochemical Analysis of PKI Datasets

2.2.1. Distribution of Physicochemical Properties of PKIs

To describe a molecule, it is common to compute its physicochemical properties to obtain information on the size, the lipophilicity, the atomic composition, etc. Some descriptors, as described by Lipinski or Veber, are still widely used to evaluate the potential oral bioavailability of a compound [25,26]. Lipinski rule relies on four properties: molecular weight (MW) ≤ 500; number of hydrogen bond donors (HBD) ≤ 5; number of hydrogen bond acceptors (HBA) ≤ 10 and logP ≤ 5. Veber rule relies on the number of rotatable bonds (NRB) ≤ 10 and topological polar surface area (TPSA) ≤ 140 Å or the sum of HBD and HBA ≤ 12. In addition, since they are also important in drug design, the number of aromatic rings and the number of chiral atoms were also calculated [27,28]. During the search of a lead compound in a virtual screening campaign, such descriptors may serve as a filter to discard molecules, and therefore decrease the size of the chemical library, since a virtual library can be large. The distribution of these descriptors calculated from inhibitors extracted from PKIDB is shown in Figure 2.

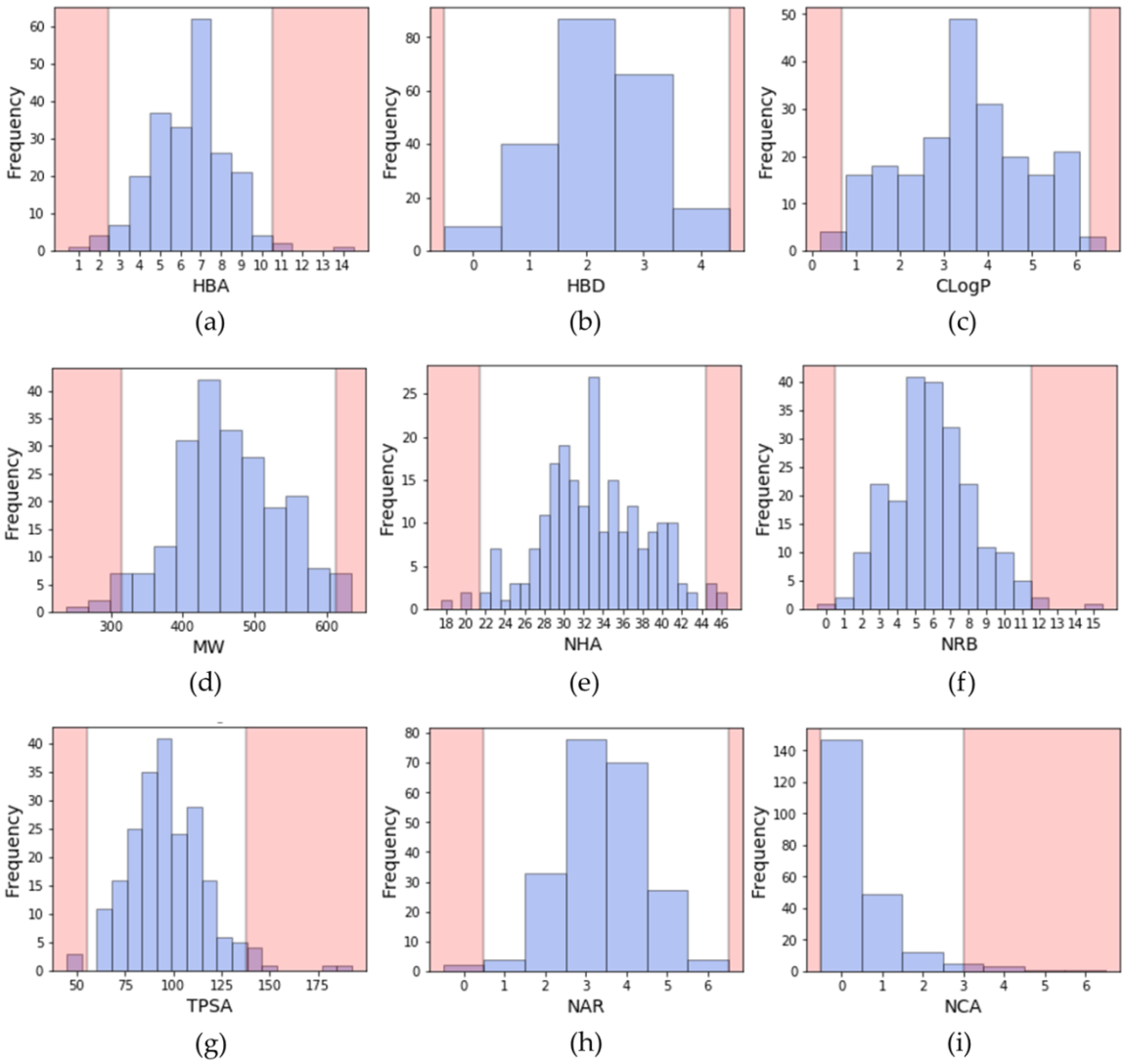


Figure 2. Distribution of physicochemical properties of PKIs: (**a**) number of hydrogen bond acceptors (HBA); (**b**) number of hydrogen bond donors (HBD); (**c**) ClogP (Rational Discovery Kit (RDKit)); (**d**) molecular weight (MW); (**e**) number of heavy atoms (NHA); (**f**) number of rotatable bonds (NRB); (**g**) topological polar surface area (TPSA); (**h**) number of aromatic rings (NAR); (**i**) number of chiral atoms (NCA). Pink areas represent values outside two standard deviation from the mean (95.4% confidence interval).

In a previous study [22], we analyzed the 'rule of five' descriptors detailed by Lipinski [25] for inhibitors in clinical trials and approved. Here, we updated the statistical analysis with new PKIs included in PKIDB and we compared them to the ChEMBL dataset (Table 2).

Table 2. Comparison of Lipinski's rules violation between PKIs approved, in clinical trials and in ChEMBL.

[1]	**0 Ro5 Violation**	**1 Ro5 Violation**	**2 Ro5 Violations**	**> 2 Ro5 Violations**
PKIs approved	33/60 (55.0%)	20/60 (33.0%)	7/60 (12.0%)	0/56 (0%)
PKIs in clinical trials	101/158 (64.0%)	41/158 (26.0%)	16/158 (10.0%)	0/158 (0%)
All PKIs	134/218 (61.5%)	61/218 (28.0%)	23/218 (10.5%)	0/218 (0%)

PKIs ChEMBL	51,858/76,504 (67.8%)	18,601/76,504 (24.3%)	5,876/76,504 (7.7%)	169/76,504 (0.2%)

[1] RDKit was used to calculate all descriptors including ClogP.

We found that 62% and 68% of PKIs in PKIDB and in ChEMBL, respectively, do not violate any Lipinki's rule. One single violation occurs in 28% and 24% of the PKIs for PKIDB and ChEMBL, respectively, and two violations occur for about 10% of the PKIs in the two datasets. Finally, few PKIs in ChEMBL dataset violates more than two rules (0.2%), and none for the PKIs in PKIDB. These results may vary, depending on how the LogP is calculated. Here, we used Wildman-Crippen approach [29]. Compared to the initial study, we removed the counter ion during the standardization of the molecules, such as the bromide ion in tarloxotinib. Despite the large different number of compounds in both datasets (76,504 molecules in ChEMBL and 218 in PKIDB) we revealed that the two datasets exhibit similar rule of five violation profiles.

The ratio of PKIs having descriptors out of the Lipinski's or Veber's rule are given in Table 3. Here, again, we found that there is no significant difference between the two kinase subsets over all the descriptors. Molecular weight (MW) and CLogP are the most discriminant descriptors. Interestingly, less than 5% of the PKIs have descriptors out of Veber's boundaries.

Table 3. Number of PKIs violating at least one Lipinski's or Veber's rule.

[1]	**MW > 500 Da**	**ClogP > 5**	**HBA > 10**	**HBD > 5**	**TPSA > 140 Å^2**	**NRB > 10**
PKIs approved	20/60 (33.3%)	12/60 (20.0%)	2/60 (3.3%)	0/60 (0%)	2/60 (3.3%)	2/60 (3.3%)
PKIs in clinical trials	46/158 (29.1%)	26/158 (16.5%)	1/158 (0.6%)	0/158 (0%)	4/158 (2.5%)	6/158 (3.8%)
All PKIs	66/218 (30.3%)	38/218 (17.4%)	3/218 (1.4%)	0/218 (0%)	6/218 (2.8%)	8/218 (3.7%)
PKIs ChEMBL	18,892/76,504 (24.7%)	10,897/76,504 (14.2%)	924/76,504 (1.2%)	208/76,504 (0.2%)	3695/76,504 (4.8%)	2,051/76,504 (2.7%)

[1] RDKit was used to calculate all descriptors including ClogP.

From these calculations, we propose a range of descriptors to guide the design of kinase inhibitors. The proposed ranges do not consider the property values beyond two standard deviations from the mean (95.4% confidence interval). Thus, the upper and lower limits of molecular descriptors better represent the current chemical space of kinase inhibitors, either approved or in clinical trials.

One can notice that despite new PKIs in PKIDB, these guidelines have not changed much compared to the ones presented in our first study. This shows that the define PKI chemical space seems well defined.

Considering all PKIs from PKIDB, the guidelines for prioritization are:

- A molecular weight (MW) between 314 and 613 Da (average of 463.4 Da);
- A ClogP (calculated with a Rational Discovery Kit (RDKit)) between 0.7 and 6.3 (average of 3.5);
- Between 0 and 4 hydrogen bond donors (HBD) (average of 2.2);
- Between 3 and 10 hydrogen bond acceptors (HBA) (average of 6.4);
- A topological polar surface area (TPSA) comprised between 55 and 138 Å^2 (average of 96.6 Å^2);
- Between 1 and 11 rotatable bonds (NRB) (average of 6.0);
- Number of aromatic rings (NAR) between 1 and 5 (average of 3.4);
- Number of chiral atoms (NCA) between 0 and 2 (average of 0.5).

2.2.2. Statistical Analysis of Protein Kinase Inhibitors

To compare the chemical space of the kinase inhibitors from PKIDB and from ChEMBL (PKI_ChEMBL), we performed a principal component analysis (PCA). Each molecule was described using 10 classical physicochemical descriptors (See Materials and Methods) well suited to characterize chemical structures. The goal here is to compare PKI_ChEMBL to PKIDB.

The PCA plot (Figure 3) illustrates the chemical space of PKIs in a 2D reference frame, represented by the two first principal components (PC1 and PC2).

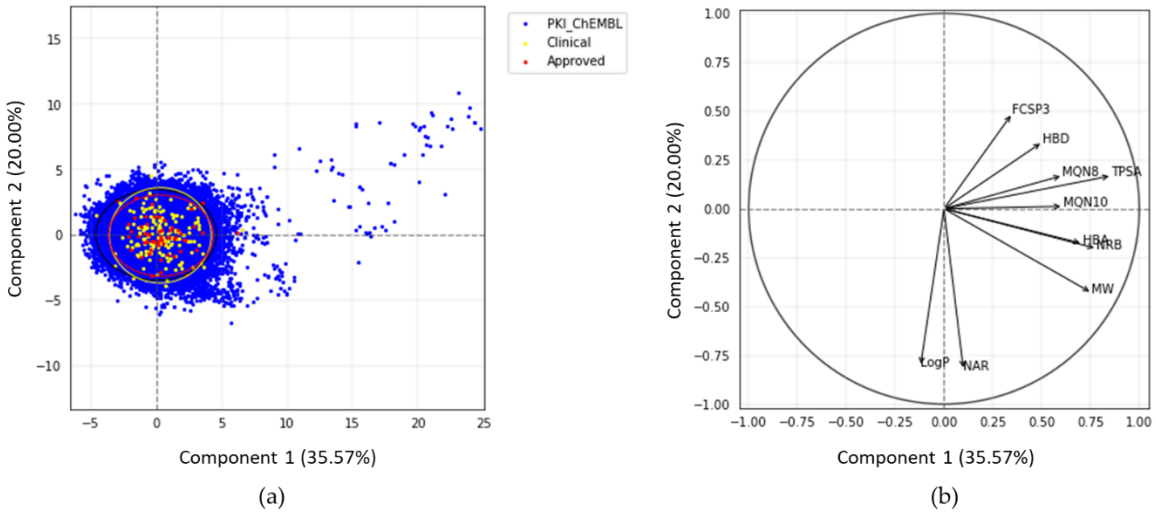


Figure 3. (**a**) Principal component analysis (PCA) of PKIs from ChEMBL and PKIDB, containing 76,504 and 209 compounds, respectively. Black, yellow, and red ellipses encompass 95% of the individuals from class "PKI_ChEMBL", "Clinical_PKI", and "Approved_PKI", respectively; (**b**) correlation circle.

The two first principal components explain 35.6% and 20.0% of the total variance respectively. PC3, not shown here, encompasses 13.2%. Thus, the 2D scatterplot of the factorial space illustrated here represents around 56% of the total variance (Figure 3).

Each dot on the graph (Figure 3a) represents a molecule. Few compounds from PKI_ChEMBL are projected in the upper right quadrant but none belongs to PKIDB. Most of the PKIDB compounds are centered in the PCA plot. Approved (red dots) and in clinical trials (yellow dots) PKIs are projected in the same chemical space. The graphical representation of normalized variables is shown in the correlation circle (Figure 3b). The angle between two vectors indicates the correlation between the two corresponding variables. A value close to 0° or 180° indicates positively or negatively correlated variables, respectively. A value near 90° indicates that the variables are not correlated. On the correlation circle (Figure 3b), one can see that the first factorial axis (PC1) is highly correlated with TPSA, NRB, and MW. These three variables contribute to PC1 at 20.6%, 17.1%, and 16.1%. The variables CLogP and NAR do not contribute to this axis, and are negatively correlated with the second factorial axis (PC2). Their contribution to PC2 are 32.6% and 34.0%, respectively. To a lesser extent, this axis is also positively correlated with FCSP3 and HBD (contributions of 11.8% and 5.8%, respectively). These two descriptors are correlated to PC3 (contributions of 24.7% and 29.7%, respectively). A weak angle between NAR and CLogP vectors is consistent with the fact that CLogP increases with the number of aromatic rings.

In view of these results, PCA confirms our preliminary observations that there are few outliers in PKI_ChEMBL dataset (dots on the upper right quadrant). It appears that these compounds correspond to either small-modified peptides or macrocyclic lactones with high TPSA values. These molecules, such as everolimus, were removed from PKIDB, since they do not inhibit protein kinases directly, however, the macrocycles in PKI_ChEMBL are active on protein kinases and, thus, were not removed from the dataset. Regarding the compounds in PKIDB, semaxanib, has the lowest MW (yellow dot, bottom-left). The two dots outside the circle and on the middle right of the quadrant corresponds to barasertib (Clinical_PKI in yellow) and fostamatinib (Approved_PKI in red). Both of these molecules contain phosphate group, increasing their TPSA, and therefore explaining their position on the PCA map.

2.2.3. Principal Moments of Inertia

Until now, we only analyzed the molecules using 2D descriptors; therefore, to evaluate the shape diversity, we represented the molecules on a principal moments of inertia (PMI) plot [30]. In a triangular PMI map, the three corners represent distinctive shapes: rod (represented by diacetylene), disk (benzene) and sphere (adamantane). Note that such a plot only describes molecular shapes, without any consideration of other properties. In order to escape from the flatland, compounds should get closer to the sphere [31].

The PMI plot (Figure 4) reveals a vast majority of kinase inhibitors are located along the rod-disc axis, indicating a preponderance of flat molecules, explained by the fact that all these molecules target a similar ATP active site. Indeed, most of the compounds in PKIDB are targeting the ATP site, thus, presenting a similar shape. Some of the MEK inhibitors are targeting an allosteric site adjacent to the ATP site. The three molecules from PKIDB closest to the extreme vertices are mubritinib near the rod, mavelertinib near the disc, and galunisertib (yellow dot in Figure 4) and idelalisib (red dot in Figure 4) near the sphere. They are all in clinical trials, in phase 1, 0, and 2, respectively. Unlike approved PKI, a few compounds in development tend to adopt a disc shape that explores a new molecular space in PKIs. We also observe some compounds from PKI_ChEMBL dataset getting closer to the sphere vertex, showing that some spherical molecules could also inhibit protein kinases. These ones could open the way to the exploration of a potential novel chemical space.

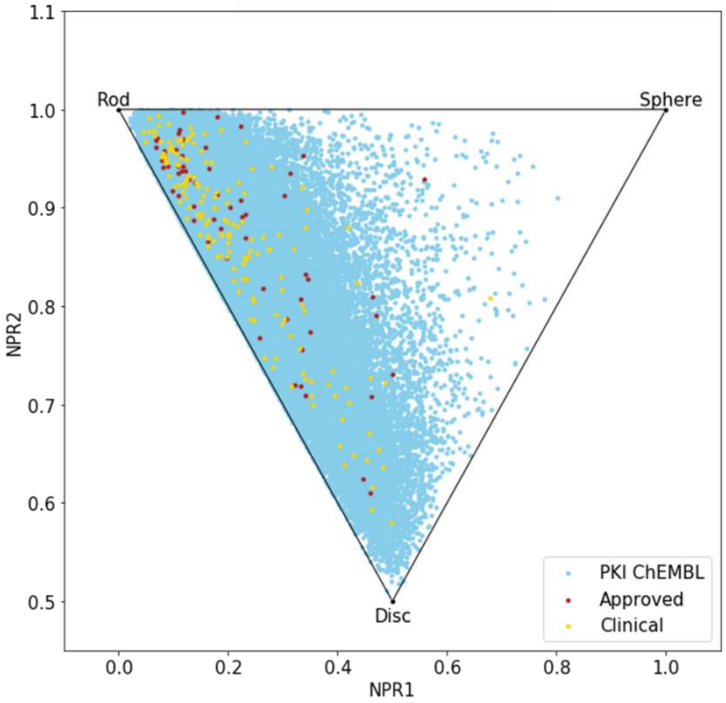


Figure 4. Principal moments of inertia (PMI) plot of PKIs in clinical trials (yellow), approved (red) and from ChEMBL database (light blue).

Here again, there is a great resemblance between the two datasets, PKIDB being well encompassed in PKI_ChEMBL regarding shape diversity.

2.3. Scaffold Diversity Assessment

2.3.1. Analysis of Molecular Scaffolds

To get a deeper insight on the molecular diversity of PKIs, we focused on scaffolds and ring systems of these compounds. The results of scaffold analysis are summarized in Table 4. First, we looked for the presence of macrocyclic molecules (rings > 12 atoms). In PKIDB, there are four macrocycles. Two of them

are approved drugs: icotinib, approved by CFDA and lorlatinib, and two are in phase 3: pacritinib and ruboxistaurin. This class of molecules might not be fully explored, since the percentage of macrocycles found in PKI_ChEMBL is very weak (<1%). As mentioned above, it is important to note that we excluded from PKIDB macrocycles containing the stem 'imus'. However, these compounds do not directly target a kinase binding site, but rather an upstream protein, causing a complex formation that inhibits the kinase [32].

The different types of molecular scaffolds are shown in Figure 5. For this study we used two types of scaffolds: Bemis and Murcko (BM) and a graph framework issued from BM. As a reminder, Bemis and Murcko scaffold corresponds to the core of a molecule after removing side chains [20]. The graph framework corresponds to BM scaffold, where each heteroatom is replaced by a carbon and each multiple bond was substituted by a single one. Therefore, such frameworks cover topologically equivalent BM scaffolds differentiated by heteroatom substitutions and bond types.

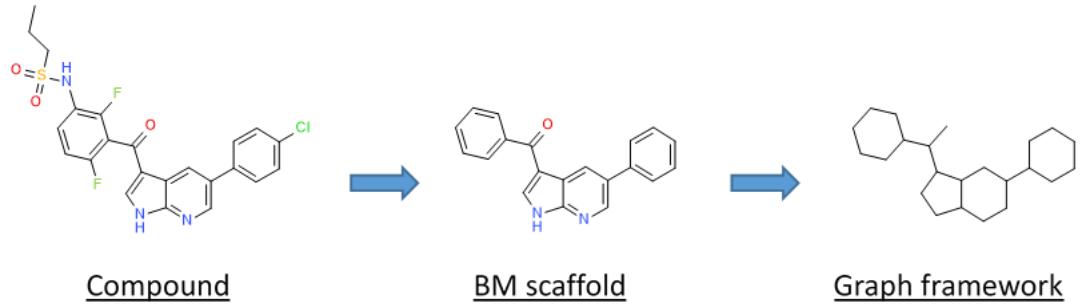


Figure 5. Representation of a molecular decomposition into scaffolds according to Bemis and Murcko (BM) and in graph framework.

In PKIDB dataset, among 218 molecules, 207 present a unique BM scaffold and 195 a unique graph framework (GF). Whereas, for the 76,504 PKIs present in ChEMBL, only 28,732 and 13,331 BM scaffolds and GF, respectively, are found (Table 4). In other words, in PKIDB almost each compound has a unique scaffold (218/207). The pairwise molecular similarity means between PKIDB and PKI_ChEMBL, calculated with MACCS keys, indicates that both datasets are diverse, with a mean of Tanimoto similarity of about 0.5 (Table 4). However, in the PKI_ChEMBL dataset, the scaffold diversity corresponding to the total number of molecules over the total number of BM scaffolds, is much lower with about a BM scaffold for about 2.7 molecules in average. Regarding the graph frameworks, their number tends to decrease compared to BM scaffolds i.e., one GF for 1.1 and 5.7 molecules in PKIDB and PKI_ChEMBL, respectively.

Table 4. Data obtained for the Bemis and Murcko scaffolds for the two datasets.

	No. Molecules	No. Macrocycles	No. BM Scaffolds	No. Graph Frameworks	Molecular Similarity Mean [a] (SD)
PKIDB	218	4 (1.8%)	207 (95.0%)	195 (89.5%)	0.51 (0.11)
PKI_ChEMBL	76,504	487 (0.64%)	28,732 (37.6%)	13,331 (17.4%)	0.49 (0.11)

[a] Calculated with MACCS keys (166 bits) and the Tanimoto coefficient.

The most represented BM scaffold in PKIDB, the indolinone derivative (Figure 6), is retrieved in three inhibitors and differs from the one in PKI_ChEMBL, which is found 644 times. This scaffold is prominent compared to others in PKI_ChEMBL: the second most retrieved scaffold, the quinazoline derivative, is only found 239 times. It shows the importance of that scaffold in PKIs, which is found only in erlotinib in PKIDB. The search for molecules containing PKIDB's highest occurrence of BM scaffold in PKI_ ChEMBL only returns 10 compounds, revealing a major difference between the two datasets.

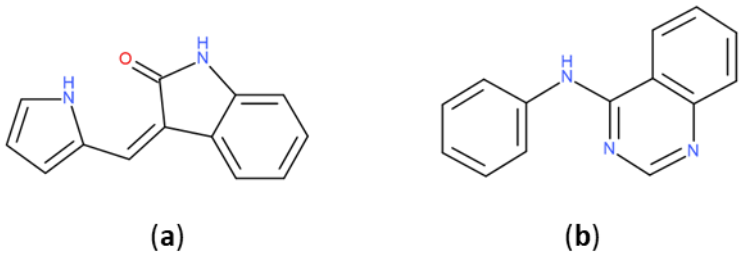


Figure 6. Most retrieved Bemis and Murcko scaffolds in PKIDB dataset (**a**): (3*Z*)-3-(1*H*-pyrrol-2-ylmethylene)indolin-2-one and in PKI_ChEMBL dataset (**b**): *N*-phenylquinazolin-4-amine.

Then, for each unique BM scaffold in PKIDB, we checked how many PKIs are obtained in PKI_ChEMBL. From the 207 unique BM scaffolds available in PKIDB, only 107 are present in PKI_ChEMBL, which represent 2423 molecules out of a total of 76,504 (3.2%). This result is surprising. Firstly, we might expect that many analogues would be systematically provided for each PKI and, thus, would be available in a public database. Secondly, because PKIDB covers similar chemical space to PKI_ChEMBL according to PCA and PMI comparisons. Finally, using all unique graph frameworks from PKIDB, we were able to match 7686 compounds (10.0%) in PKI_ChEMBL.

2.3.2. Ring Analysis

In PKIs, rings are making hydrogen bonds, van der Waals or π-stacking interactions with residues of the active site. As example, an heterocycle may form hydrogen bonds, as does adenine in ATP with protein kinase [33]. We applied a molecular decomposition method, using RDKit to fragment molecules and retain only rings (Figure 7). After collecting all rings for both datasets, we searched for the most represented ones by gathering them using their smiles representation. We focused on fused heteroaromatic rings, since such fragments are present as a main scaffold in most kinase inhibitors. Moreover, fused rings offer favorable interactions (van der Waals and hydrogen bonds) into the ATP binding site compared to non-fused rings [34].

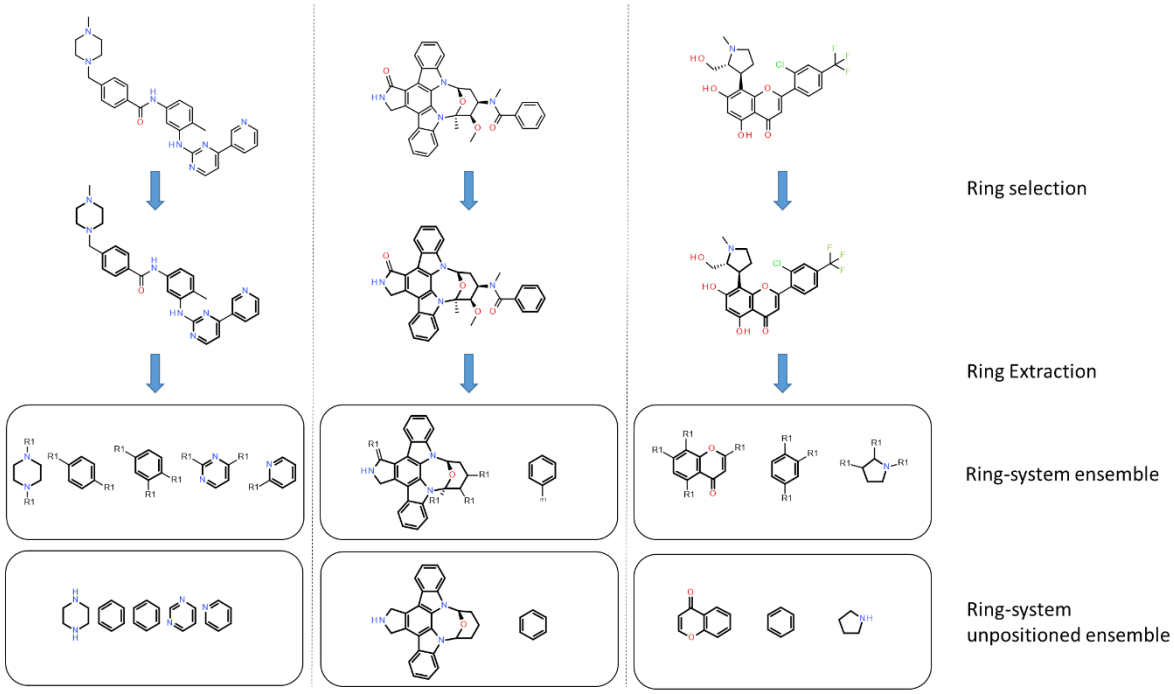


Figure 7. Application of the ring-system ensemble classification. Ring-system ensembles are obtained by removing substituents on acyclic bonds and by keeping attachment point (R1). The ring system unpositioned ensembles do not keep information on the attachment point. Rings are shown in bold.

In both datasets, we found bicycles in around 65% of the molecules, demonstrating their importance as a core during hit to lead or lead optimization steps. In PKIDB, we found 56 unique bicyclic scaffolds among the total of 172. More surprising, 31 out of these 56 bicycles are singletons, i.e., the bicyclic scaffold is found only once in the dataset. For the PKI_ChEMBL dataset, there are 918 unique bicycles for a total of 57,438. However, among those 918 unique bicycles, only 26 are singletons. Since the PKI_ChEMBL dataset contains more analogues of chemical series compared to PKIDB, this could explain the lowest ratio of unique fused rings.

The number and the frequency of the top 10 most retrieved bicycles are illustrated in Figure 8. In both datasets, the quinazoline scaffold is the most represented bicycle, it remains an important core, and its substituted analogues, such as the 4-anilinoquinazoline, have been extensively studied [35]. Examples of PKIs containing quinazoline scaffold are gefitinib, lapatinib, erlotinib, afatinib, and, more recently, canertinib [36]. Kinase inhibitors bearing this scaffold have mainly been designed to target EGFR. The second most represented bicyclic scaffold is the quinolone, another fused six-membered aromatic ring. It is worth noting that depending on the substituent types or the tautomeric form present in the molecules, a Rational Discovery Kit (RDKit) may break the aromaticity in the heterocyclic scaffolds. As an example, by removing the carbonyl functional group, considered as a substituent, in the indole scaffold, a non-aromatic indoline scaffold is kept, as shown in Figure 8. Most of the bicycles contain at least one heteroatom such as the nitrogen. This heteroatom allows H-bond interaction (acceptor or donor), with the hinge region of the kinase. Interestingly, the PKIDB and the PKI_ChEMBL datasets contain almost the same top ten bicyclic scaffolds. However, unlike BM scaffolds where more than half of them from PKIDB were not retrieved in PKI_ChEMBL, here, only three bicycles from PKIDB (not shown) are not retrieved in the PKI_ChEMBL dataset. We also performed an analysis of the bicyclic scaffolds by considering the attached atom position and atom type (Figure 9). Atoms involved in a double bond linked to the scaffold were not modified. However, all atoms were replaced by a dummy atom labelled differently according to the atom type (Figure 9). In this case, the 4,6,7-trisubstituted quinazoline is the most retrieved core in both datasets. Such a scaffold is found in twelve inhibitors in PKIDB, and an ether group (often a methoxy) is always attached on the 7 position. The second most retrieved bicycle is the 4,6,7-trisubstituted quinoline in PKIDB, and this is the third most represented scaffold in PKI_ChEMBL. Here again, the substituent in the 7 position is always an ether. Interestingly, the second most retrieved substituted bicycle in PKI_ChEMBL is not found in top tenth of PKIDB, as shown in Figure 8.

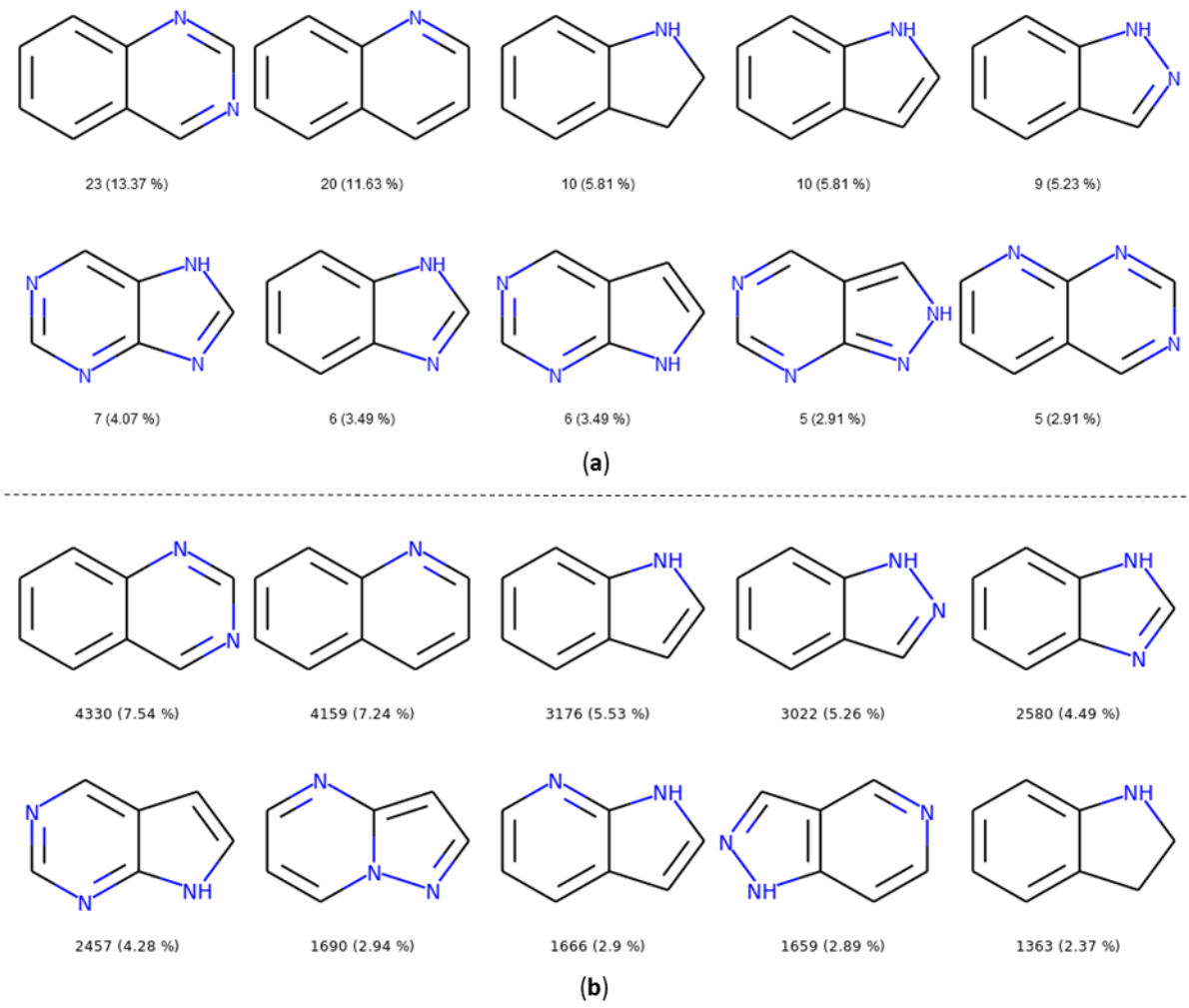


Figure 8. Top ten bicycles retrieved in the PKIDB dataset (**a**) and in PKI_ChEMBL (**b**) with their occurrence and their frequency in brackets. In PKIDB there are 172 bicycles (56 unique) and in PKI_ChEMBL, there are 57,439 bicycles (918 unique).

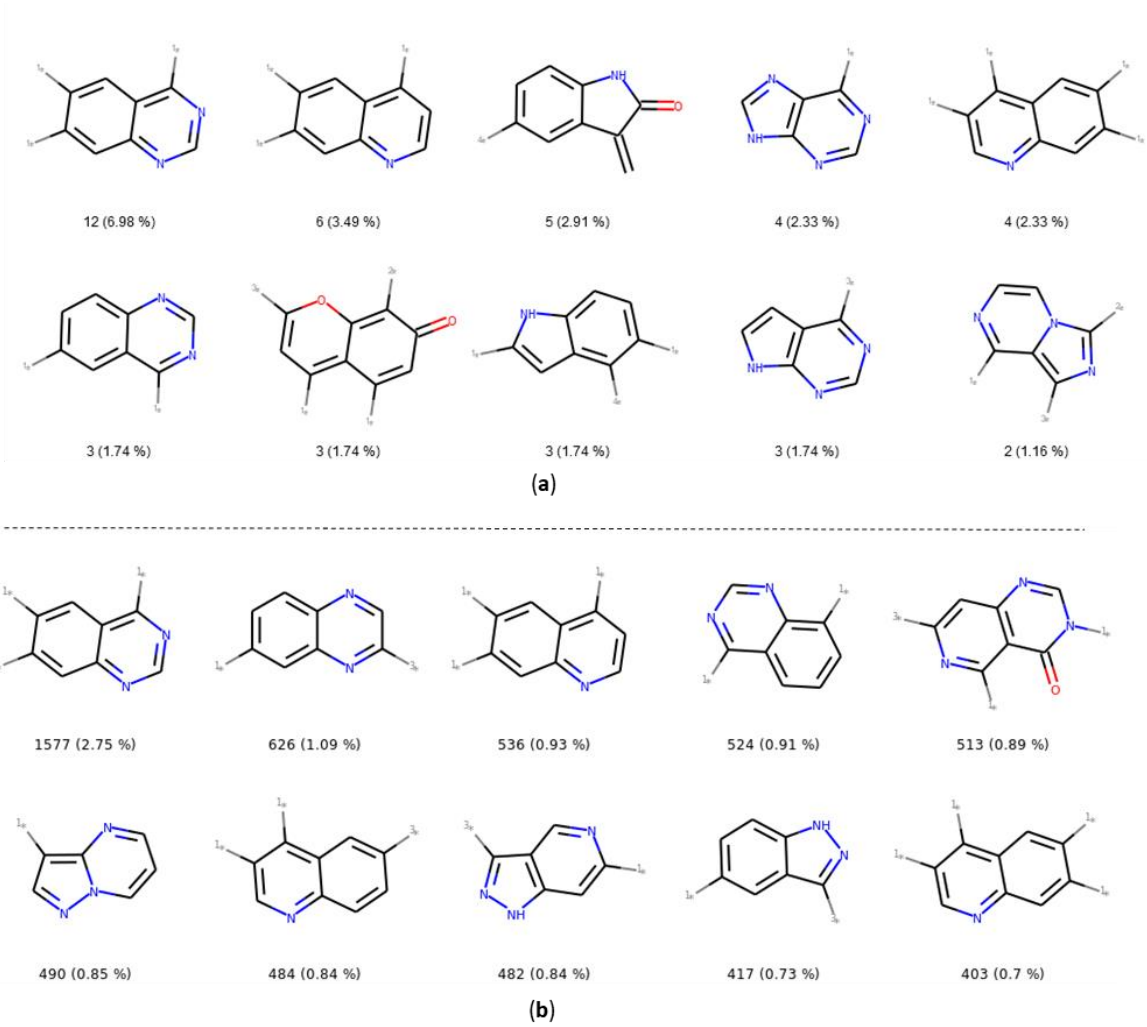


Figure 9. Top ten most retrieved bicycles with their substituents in the PKIDB dataset (**a**) and in PKI_ChEMBL (**b**) with their occurrence and their frequency in brackets. In PKIDB, there are 172 bicycles (129 unique) and in PKI_ChEMBL, there are 57,438 bicycles (4480 unique). 1*—connected to an atom not double bonded, not aromatic, not in a cycle and not halogen; 2*—connected to non aromatic ring; 3*—connected to aromatic atom; 4*—connected to an halogen.

In Figure 9, the great majority of bicycles are polysubstituted, confirming their use as core scaffolds to link substituents. By considering the substituents during the analysis, the frequency of the bicycles shows a different distribution in both datasets, and the top ten bicyclic scaffolds are different.

3. Discussion

PKIDB is a freely available database containing all kinase inhibitors on the market or in clinical trials gathered using their international nonproprietary name (INN). This database, regularly updated, contains information on the structure of the kinase inhibitors, their physicochemical properties, their protein kinase targets, as well as their therapeutic indications. It also contains links to various external databases. We analyzed this dataset and compared it to active PKIs found in the ChEMBL database. Classical physicochemical descriptors, such as Lipinski's or Veber's, showed that a significant number of protein kinase inhibitors, either approved or in clinical trials, do not follow the recommended drug-like thresholds, especially regarding molecular weight and calculated LogP. Moreover, all PKI present in PKIDB violate a maximum of only two Lipinski'rules. Therefore, for this typical class of compounds, we propose new boundaries to better characterize the chemical space of kinase inhibitors. Moreover, all PKIS in PKIDB have a maximum of two chiral centers and five aromatic rings.

The projection of the chemical space resulting from a principal component analysis shows that most of the inhibitors shared the same chemical space. However, the PKIs available in ChEMBL fill a larger chemical space in the PCA plot compared to PKIs in PKIDB. The distribution of the physicochemical descriptors for both datasets do not present major differences. This suggests that most active PKIs available in the ChEMBL have drug-like properties.

Concerning the molecular shape of the PKIs, the PMI plot reveals that PKIs from ChEMBL exhibit a larger shape diversity compared to the ones in PKIDB. However, the majority of PKIs remain clustered around the rod-disc axis because they target a common ATP binding site in the kinase domain, which is highly conserved in this protein family. Yet, PKIs under development tend to explore wider topology, particularly near the disc edge. More frequent macrocyclic structures could contribute to this specific molecular shape. Moreover, moving to new chemical space will help medicinal chemists to escape from a crowded intellectual property (IP) space. Regarding PKIs in ChEMBL, we also found some compounds escaping from this rod-disc axis and get closer to the spherical form. This information could be used to design new chemically-diverse kinase inhibitors.

Concerning the molecular scaffold analysis of the two datasets, it appears that PKIs in PKIDB exhibit a great molecular scaffold diversity compared to the ones in ChEMBL. More than 100 scaffolds from PKIDB are not present in the ChEMBL. Each molecule present in PKIDB and, more particularly, the corresponding scaffold, was patented, preventing the design of analogues. Thus, each molecule present in PKIDB is in fact a representative of a chemical series, but only one new molecular entity (NME) was selected to continue its development in clinical phases. Most pharmaceutical companies will not unveil all chemical analogues of the selected NMEs, limiting information on the chemical series. On the opposite, in a public database such as ChEMBL, there are often lots of available analogues for a specific scaffold. The ring analysis performed on the two datasets indicates a similar number of bicycle singletons despite the large size difference in the two datasets, 218 vs. 76,504 molecules for PKIDB and PKI_ChEMBL, respectively. By considering the position and the type of the substituents, a significant part of the scaffolds in PKIDB are absent in ChEMBL, because most of the structures of pharmaceutical companies are protected by patents.

The PKIDB database is regularly updated and is accessible from this website: http://www.icoa.fr/pkidb. We hope that this resource will assist researchers in their quest for novel kinase inhibitors.

4. Materials and Methods

For the creation and maintenance of PKIDB please refer to our previous study [22]. All experiments and calculations have been performed with Python 3.6. Molecular descriptors used for PCA (Table 5) and PMI were calculated with RDKit (version '2018-09-01', Palo Alto, USA). Scaffolds analysis and clustering were performed with RDKIT and with Butina algorithm [37] using Tanimoto similarity and Morgan Fingerprint, with a radius of two (equivalent of FCPF4). The PCA was calculated with an in-house library derived from Prince [38] and Scikit-learn [39] packages (Rocquencourt, France). For PMI analysis, 3D conformations were generated using ETKDG method [40] followed with an energy minimization using the MMFF94 forcefield [41]. To delimit the dots of the PMI triangle, three compounds (diacetylene, benzene and adamantane) were considered and added to the dataset. All the figures are made using matplotlib [42] and seaborn [43] packages. Molecules were drawn with Biovia Draw 2018 (Velizy, France).

The PKI_ChEMBL dataset results from ChEMBL (version 'ChEMBL_24', Cambridgeshire, UK). To be included in this dataset, a compound must have at least one recorded activity, either IC_{50}, Ki or Kd, on a protein kinase with a pchembl value > 6 (<1000 nM). We then removed duplicates, empty SMILES and molecules from PKIDB. It is composed of 76,504 molecules. Both datasets (PKIDB and PKI_ChEMBL) have been prepared and standardized with VSPrep [44], and for each compound we kept the best tautomer as defined in VSPrep.

Table 5. Descriptors used for PCA.

Name Variable	Descriptor
MW	Molecular weight
LogP	Wildman-Crippen LogP value
TPSA	Topological polar surface area
HBA	Number of Hydrogen Bond Acceptors
HBD	Number of Hydrogen Bond Donors
NRB	Number of Rotatable Bonds
NAR	Number of aromatic rings
FCSP3	Fraction of C atoms that are SP3 hybridized
MQN8	Molecular Quantum Numbers
MQN10	Molecular Quantum Numbers

Author Contributions: Data curation, C.B., F.C., G.P. and S.B.; Formal analysis, C.B., F.C. and G.P.; Funding acquisition, C.M. and P.B.; Investigation, C.B., Samia Aci-Sèche and P.B.; Methodology, C.B., F.C., S.B., C.M. and P.B.; Project administration, S.A.-S. and P.B.; Software, C.B., F.C. and G.P.; Supervision, S.A.-S., C.M. and P.B.; Validation, C.B., F.C. and G.P.; Visualization, C.B., F.C., G.P. and S.B.; Writing – original draft, C.B.; Writing – review & editing, F.C., G.P., Samia Aci-Sèche, S.B., C.M. and P.B..

Funding: This research was funded by Janssen and Région Centre Val de Loire grant number APR-IR isNatProd.

Acknowledgments: The authors wish to thank the Région Centre Val de Loire and Janssen for financial support. Authors also thank ChemAxon for providing academic license free of charge. G.P, S. A.-S., S.B. and P.B. are supported by LABEX SynOrg (ANR-11-LABX-0029). The authors also thank Laurent Robin for maintaining the website PKIDB.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Manning, G.; Whyte, D.B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298*, 1912–1934, doi:10.1126/science.1075762.
2. Bhullar, K.S.; Lagarón, N.O.; McGowan, E.M.; Parmar, I.; Jha, A.; Hubbard, B.P.; Rupasinghe, H.P.V. Kinase-targeted cancer therapies: progress, challenges and future directions. *Mol. Cancer* **2018**, *17*, 48, doi:10.1186/s12943-018-0804-2.
3. Fabbro, D.; Cowan-Jacob, S.W.; Moebitz, H. Ten things you should know about protein kinases: IUPHAR Review 14. *Br. J. Pharmacol.* **2015**, *172*, 2675–2700, doi:10.1111/bph.13096.
4. Giamas, G.; Stebbing, J.; Vorgias, C.E.; Knippschild, U. Protein kinases as targets for cancer treatment. *Pharmacogenomics* **2007**, *8*, 1005–1016, doi:10.2217/14622416.8.8.1005.
5. Mueller, B.K.; Mack, H.; Teusch, N. Rho kinase, a promising drug target for neurological disorders. *Nat. Rev. Drug Discov.* **2005**, *4*, 387–398, doi:10.1038/nrd1719.
6. Cohen, P. Immune diseases caused by mutations in kinases and components of the ubiquitin system. *Nat. Immunol.* **2014**, *15*, 521–529, doi:10.1038/ni.2892.
7. Dimova, D.; Bajorath, J. Assessing Scaffold Diversity of Kinase Inhibitors Using Alternative Scaffold Concepts and Estimating the Scaffold Hopping Potential for Different Kinases. *Molecules* **2017**, 22(5), 730; doi:10.3390/molecules22050730.
8. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors. *Pharmacol. Res.* **2019**, *144*, 19-50. doi:10.1016/j.phrs.2019.03.006.
9. Van Cutsem, E.; Köhne, C.-H.; Hitre, E.; Zaluski, J.; Chang Chien, C.-R.; Makhson, A.; D'Haens, G.; Pintér, T.; Lim, R.; Bodoky, G.; et al. Cetuximab and Chemotherapy as Initial Treatment for Metastatic Colorectal Cancer. *N. Engl. J. Med.* **2009**, *360*, 1408–1417, doi:10.1056/NEJMoa0805019.
10. Maximiano, S.; Magalhães, P.; Guerreiro, M.P.; Morgado, M. Trastuzumab in the Treatment of Breast Cancer. *BioDrugs* **2016**, *30*, 75–86, doi:10.1007/s40259-016-0162-9.
11. Cohen, M.H.; Williams, G.; Johnson, J.R.; Duan, J.; Gobburu, J.; Rahman, A.; Benson, K.; Leighton, J.; Kim, S.K.; Wood, R.; et al. Approval Summary for Imatinib Mesylate Capsules in the Treatment of Chronic Myelogenous Leukemia. *Clin. Cancer Res.* **2002**, *8*, 935–942.

12. Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* **2016**, *103*, 26–48, doi:10.1016/j.phrs.2015.10.021.
13. WHO INN stems. Available online: http://www.who.int/medicines/services/inn/stembook/en/ (accessed on Mar 20, 2019).
14. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954, doi:10.1093/nar/gkw1074.
15. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242, doi:10.1093/nar/28.1.235.
16. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109, doi:10.1093/nar/gky1033.
17. Schneider, P.; Schneider, G. Privileged Structures Revisited. *Angew. Chem. Int. Ed.* **2017**, *56*, 7971–7974, doi:10.1002/anie.201702816.
18. Methods and principles in medicinal chemistry. *Scaffold hopping in medicinal chemistry*; Brown, N., Ed.; Wiley-VCH-Verl: Weinheim, Germany 2014; ISBN 978-3-527-33364-6.
19. Dimova, D.; Stumpfe, D.; Bajorath, J. Computational design of new molecular scaffolds for medicinal chemistry, part II: generalization of analog series-based scaffolds. *Future Sci. OA* **2017**, *4*, doi:10.4155/fsoa-2017-0102.
20. Bemis, G.W.; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893, doi:10.1021/jm9602928.
21. Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Mol. Inform.* **2011**, *30*, 646–664, doi:10.1002/minf.201100078.
22. Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23*, 908, doi:10.3390/molecules23040908.
23. United States Adopted Names approved stems. Available online: https://www.ama-assn.org/about/united-states-adopted-names/united-states-adopted-names-approved-stems (accessed on Jun 26, 2019).
24. Hiroshi H.; Yoshikazu I.; Megumu O.; Tsuyoshi A.; Toshihide N.; Makiko K.; Toshifumi K.; Naoki K.; Junko O.; Kazunori Y.; et al. Small-molecule inhibition of Wee1 kinase by MK-1775 selectively sensitizes p53-deficient tumor cells to DNA-damaging agents. *Mol. Cancer Ther.* **2009**, 8, 11, 2992-3000. doi:10.1158/1535-7163.MCT-09-0463.
25. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26, doi:10.1016/S0169-409X(00)00129-0.
26. Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623, doi:10.1021/jm020017n.
27. Brooks, W.H.; Guida, W.C.; Daniel, K.G. The Significance of Chirality in Drug Design and Development. *Curr. Top. Med. Chem.* **2011**, *11*, 760–770.
28. Ward, S.E.; Beswick, P. What does the aromatic ring number mean for drug design? *Expert Opin. Drug Discov.* **2014**, *9*, 995–1003, doi:10.1517/17460441.2014.932346.
29. Wildman, S.A.; Crippen, G.M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873, doi:10.1021/ci990307l.
30. Sauer, W.H.B.; Schwarz, M.K. Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 987–1003, doi:10.1021/ci025599w.
31. Lovering, F.; Bikker, J.; Humblet, C. Escape from Flatland: Increasing Saturation as an Approach to Improving Clinical Success. *J. Med. Chem.* **2009**, *52*, 6752–6756, doi:10.1021/jm901241e.
32. Dowling, R.J.O.; Topisirovic, I.; Fonseca, B.D.; Sonenberg, N. Dissecting the role of mTOR: Lessons from mTOR inhibitors. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **2010**, *1804*, 433–439, doi:10.1016/j.bbapap.2009.12.001.
33. Zhang, J.; Yang, P.L.; Gray, N.S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28–39, doi:10.1038/nrc2559.
34. Zhao, H.; Caflisch, A. Current kinase inhibitors cover a tiny fraction of fragment space. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 2372–2376, doi:10.1016/j.bmcl.2015.04.005.
35. Conconi, M.T.; Marzaro, G.; Urbani, L.; Zanusso, I.; Di Liddo, R.; Castagliuolo, I.; Brun, P.; Tonus, F.; Ferrarese, A.; Guiotto, A.; et al. Quinazoline-based multi-tyrosine kinase inhibitors: Synthesis, modeling, antitumor and antiangiogenic properties. *Eur. J. Med. Chem.* **2013**, *67*, 373–383, doi:10.1016/j.ejmech.2013.06.057.

36. Smaill J.B.; Rewcastle G.W.; Loo J.A.; Greis K.D.; Chan O.H.; Reyner E.L.; Lipka E.; Showalter H.D.; Vincent P.W.; Elliott W.L.; Denny W.A. Tyrosine kinase inhibitors. 17. Irreversible inhibitors of the epidermal growth factor receptor: 4-(phenylamino)quinazoline- and 4-(phenylamino)pyrido [3,2-d]pyrimidine-6-acrylamides bearing additional solubilizing functions. *J Med Chem.* **2000**, 43,7, 1380-97. doi:10.1021/jm990482t.
37. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747–750, doi:10.1021/ci9803381.
38. Halford, M. :crown: Python factor analysis library (PCA, CA, MCA, MFA, FAMD): MaxHalford/prince; 2019; https://github.com/MaxHalford/prince.
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Riniker, S.; Landrum, G.A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574, doi:10.1021/acs.jcim.5b00654.
41. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519, doi:10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.
42. Thomas A Caswell; Michael Droettboom; John Hunter; Eric Firing; Antony Lee; David Stansby; Elliott Sales de Andrade; Jens Hedegaard Nielsen; Jody Klymak; Nelle Varoquaux; et al. *matplotlib/matplotlib v3.0.1*; Zenodo, 2018;
43. Michael Waskom; Olga Botvinnik; Drew O'Kane; Paul Hobson; Joel Ostblom; Saulius Lukauskas; David C Gemperline; Tom Augspurger; Yaroslav Halchenko; John B. Cole; et al. *mwaskom/seaborn: v0.9.0 (July 2018)*; Zenodo, 2018;
44. Gally José-Manuel; Bourg Stéphane; Do Quoc-Tuan; Aci-Sèche Samia; Bonnet Pascal VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Mol. Inform.* **2017**, *36*, 1700023, doi:10.1002/minf.201700023.

Sample Availability: Not available.

1.9 Conclusion

Au cours de ce premier chapitre introductif, nous avons défini, expliqué et illustré toutes les notions à connaître pour pouvoir comprendre cette thèse sur la conception d'inhibiteurs de protéines kinases à l'aide de méthodes *in silico* se basant sur les fragments.

Le prochain chapitre présentera le logiciel que j'ai principalement utilisé au cours de ma thèse : Frags2Drugs (F2D). Ce deuxième chapitre montrera aussi l'analyse et les améliorations que j'ai apportées à ce logiciel, ainsi que l'article scientifique le présentant.

Chapitre 2 : Frags2Drugs, un logiciel basé sur les fragments pour la conception d'inhibiteurs de protéines kinases

J'ai utilisé différents programmes et logiciels au cours de ma thèse pour concevoir des inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI). Cependant, comme je l'ai évoqué en avant-propos, mon outil principal a été le logiciel Frags2Drugs (F2D). Ce logiciel permet la création de PKI à partir de fragments venant de données expérimentales. Au cours de ce chapitre, nous allons étudier en détail F2D, en commençant par un article présentant sa conception, sa validation et quelques exemples d'utilisation. Ensuite, un autre article présentera un cas d'utilisation dirigé vers la découverte de macrocycles, à la suite de quoi je montrerai les améliorations que j'ai apportées à F2D.

2.1 Présentation de Frags2Drugs

2.1.1 Historique

Le développement de F2D a été initié pendant la thèse du Dr José-Manuel Gally[1], soutenue en 2017. Il a ensuite été poursuivi pendant celle du Dr Colin Bournez[2], soutenue en 2019, avec l'aide de l'ingénieur d'études Pascal Krezel. Je contribue au développement de F2D depuis le début de ma thèse. F2D a été créé dans un contexte où il existait déjà de nombreux programmes de conception d'inhibiteurs basée sur les fragments (Fragment-Based Drug Design, FBDD, Chapitre 1, partie 1.4.6). Malgré l'existence et la disponibilité de ces autres programmes de FBDD, l'équipe SB&C a fait le choix d'en développer un nouveau en interne.

Ce choix se justifie d'abord par des tests réalisés par le Dr Colin Bournez montrant l'incapacité de certains de ces logiciels à recréer un inhibiteur déjà connu à partir de ses propres fragments[2]. Les problèmes majeurs étaient :

1. La lenteur de la plupart des logiciels
2. L'impossibilité de prendre en compte des fragments issus des complexes PDB
3. La nécessité d'avoir une intervention humaine à chaque étape

De plus, certains de ces programmes ne sont plus mis à jour. Ensuite, le langage de programmation utilisé dans ces logiciels peut être un frein pour une bonne appropriation et une poursuite du développement par notre équipe qui souhaite n'utiliser que le langage Python pour une bonne homogénéité des programmes. L'absence d'une documentation accessible et fournie peut aussi empêcher leur utilisation, tout comme une incompatibilité avec certains systèmes d'exploitation. Ces logiciels peuvent être distribués sous licence commerciale, rendant leur utilisation payante et ne permettant pas de modifier certaines fonctionnalités. Enfin la confidentialité, qui doit être gardée sur les molécules recherchées, n'est pas permise par les interfaces web hébergées par des entreprises tierces.

Les travaux effectués par mes prédécesseurs ont fait de F2D un outil fonctionnel pour la découverte de PKI. Des PKI ont d'ailleurs été découverts, synthétisés et testés comme actifs et sélectifs au cours de la thèse de Colin Bournez, prouvant le succès de cet outil[2]. Aujourd'hui le dépôt d'une licence de logiciel est en cours et F2D fait l'objet d'un programme de valorisation porté par c-Valo

(https://www.youtube.com/watch?v=s0lvMsBNNVI). Au cours de ma thèse, j'ai pu apprendre à utiliser F2D et continuer son développement tout en l'analysant.

2.1.2 Conception de F2D

La conception de F2D est décrite dans l'article suivant, actuellement soumis pour publication dans le *Journal of Medicinal Chemistry*. Cet article explique le détail du fonctionnement de F2D pour obtenir des PKI à partir d'un fragment initial positionné dans le site actif d'une protéine kinase. Il explique aussi comment les résultats initiaux sont filtrés pour n'obtenir que les molécules ayant la plus grande probabilité de devenir des candidats médicaments. Trois exemples d'utilisation sont ensuite présentés. Ils ont permis de recréer des PKI déjà connus et d'obtenir de nouvelles molécules pour les protéines BCR-ABL1, BRAF et MELK. Dans l'exemple d'utilisation de F2D sur MELK, nous avons distingué l'obtention d'inhibiteurs de type I ou de type II. Enfin, l'article présente le site web (http://frags2drugs.icoa.fr) permettant une utilisation de F2D par tout chercheur, sans devoir utiliser des outils de programmation.

Le fonctionnement de F2D est résumé en Figure 25. Le logiciel se décompose en 3 parties principales : la création de la base de données (BDD) de fragments, la conception de molécules par agrandissement et liaison de fragments (cœur de F2D) et la sélection des meilleures molécules générées. Ce schéma d'organisation générale peut être modifié pour certains projets spécifiques. Par exemple, lors de la recherche de molécules devant passer la barrière hémato-encéphalique (BHE), l'estimation d'un score *Central Nervous System MultiParameter Optimisation* (CNS MPO) peut être calculé[266]. Les molécules ayant un score supérieur à 4 seront plus susceptibles de passer la BHE et d'agir sur des cibles localisées après le passage de cette barrière.

Lorsque la structure 3D d'une cible n'existe pas dans la BDD RCSB PDB, un modèle par homologie peut être conçu. Il s'agit d'une technique partant de la structure 3D d'une protéine ayant une très grande similarité de séquence et permettant de prédire la structure 3D d'une cible n'ayant jamais été cristallisée.

À la suite de la génération de molécules et après avoir appliqué les trois premiers filtres, nous avons ajouté une étape de recherche de l'existence de composés similaires dans des BDD de composés publiques. Quatre BDD sont utilisées pour vérifier si des molécules similaires aux résultats de F2D existent déjà. Il s'agit des BDD ChEMBL, PKIDB, ZINC et Ambinter. Les deux premières permettent de savoir s'il existe des composés similaires aux molécules trouvées ayant déjà été testés expérimentalement (ChEMBL) ou étant des PKI en cours d'essais cliniques ou sur le marché ayant une dénomination commune internationale (PKIDB). Les deux dernières permettent d'identifier des molécules similaires pouvant être achetées. Enfin les molécules provenant de F2D et évaluées comme étant « les meilleures » mais non disponibles commercialement pourront être synthétisées par des chimistes. Les autres sous-parties de l'organisation du programme sont expliquées dans l'article ci-dessous.

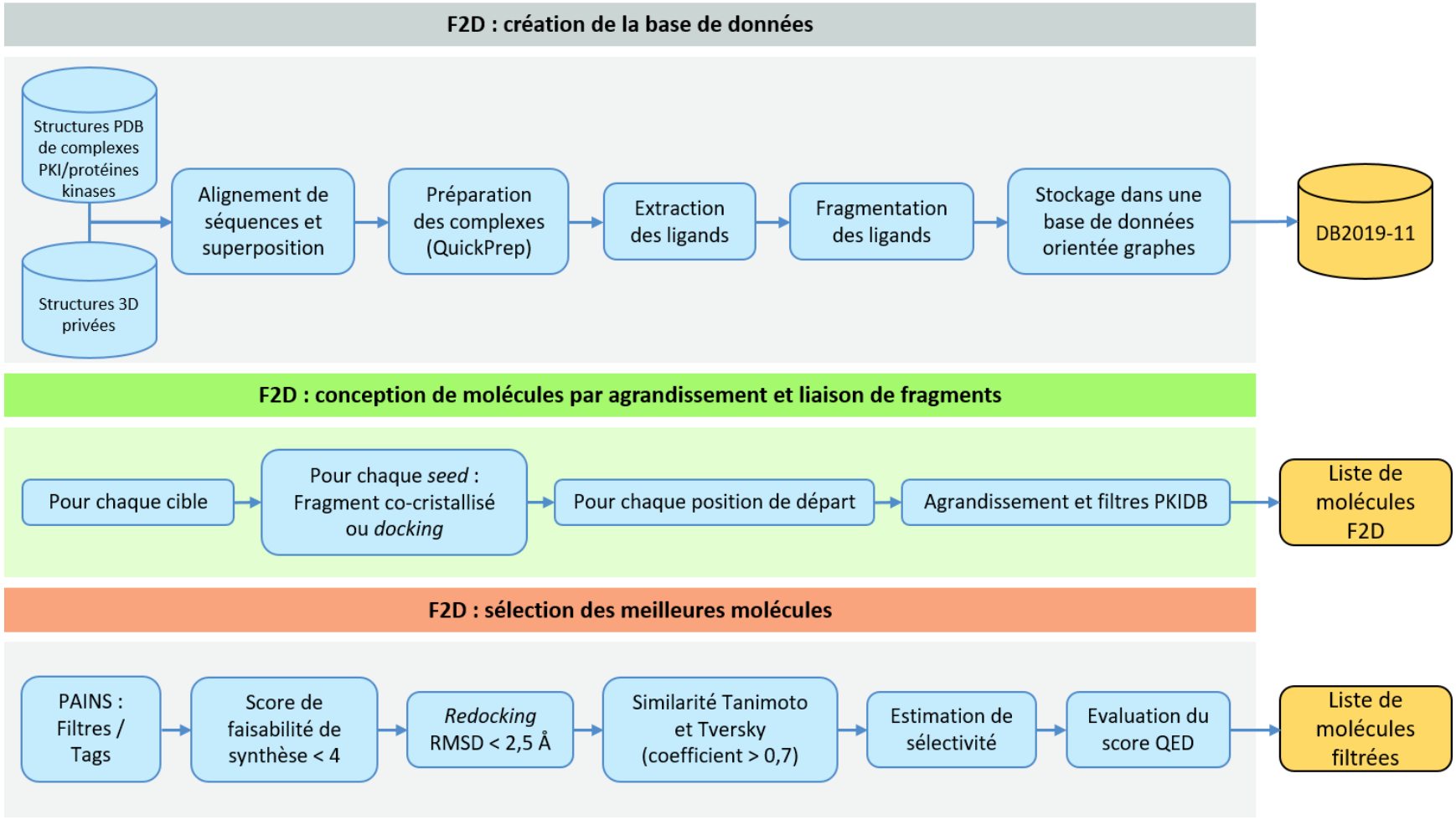


Figure 25 : Organisation des différentes parties et sous-parties du programme F2D. La préparation des complexes PKI/protéines kinases s'effectue par le module QuickPrep de la suite logiciel Moe (Chemical Computing Group, 2016_0802).

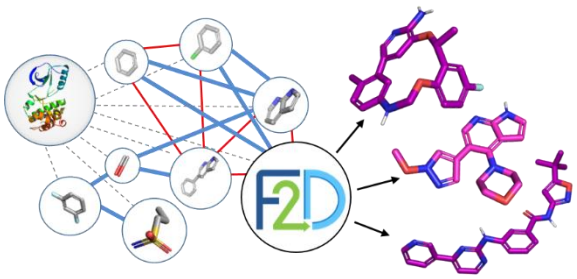
Frags2Drugs: a tool to discover novel protein kinase inhibitors from a 3D fragment network

Gautier Peyrat,‡ Colin Bournez,‡ Pascal Krezel, José-Manuel Gally, Stéphane Bourg, Samia Aci-Sèche and Pascal Bonnet*

Institut de Chimie Organique et Analytique (ICOA UMR 7311), Université d'Orléans - Pôle de chimie, rue de Chartres - BP 6759, 45067 Orléans Cedex 2, France

Keywords: Fragment growing, Fragment-based drug design, Network graph, Protein kinase inhibitor, Kinase research

ABSTRACT: Fragment-based approaches have been widely developed and employed in both academia and industry in the field of drug discovery. We present here an innovative *in silico* Fragment-Based Drug Design approach, aiming to design new kinase inhibitors directly into the ATP binding site. This tool, called Frags2Drugs (F2D), relies on an in-house three dimensional (3D) fragment library obtained from co-crystallized ligands. This library is stored in a graph-oriented database containing required information to link fragments together. F2D builds every possible molecule fitting in the given cavity in a minute scale. Molecules are then filtered to keep those presenting the best potential affinity. Several specific molecular filters are applied, including a specific in-house Protein Kinase Inhibitors (PKI) like filter. We validated our method by reconstructing existing co-crystallized ligands and known kinase inhibitors. In this article, we describe this new software program and provide examples of its usage to identify known and novel type I, type I½ and type II inhibitors on several protein kinases. F2D is freely accessible at http://frags2drugs.icoa.fr



INTRODUCTION

In recent decades, Fragment-Based Drug Design (FBDD) has emerged to become a major approach in the discovery of new chemical compounds.[1] The first success from FBDD dates from 2011, with the approval by the U.S. Food and Drug Administration (FDA) of vemurafenib (Zelboraf), a drug targeting the mutated protein kinase BRAF V600E.[2] Nowadays, several drugs approved or in advanced clinical trials have been discovered by using FBDD approaches.[3] Another recent example is erdafetinib (Balversa)[4] developed by Janssen and Astex and approved by FDA in 2019.

Fragments are an interesting alternative to traditional compounds used for high-throughput screening: they are smaller (< 300 Da) and can cover a bigger chemical space with less molecules[5]. Once identified during a screening campaign, fragment hits have to be optimized using medicinal chemistry and structural methods such as X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. Computational chemistry may also play a substantial role in this process. Multiple software applications have been recently developed in the field of *in silico* FBDD[6] that can be classified into three main categories: growing, linking and merging.

Growing methods aim to grow from an initial fragment, called the seed, by adding sequentially new fragments. Software programs such as AutoGrow[7–9], FOG[10] and Open-Growth[11] were developed using this strategy. Other tools use atoms rather than fragments to optimize the seed, as LEGEND[12] and Genstar[13]. However, atom-based methods remain less employed. The growth is usually performed in the binding site, considering three-dimensional coordinates of the original hit found experimentally or by docking methods. Thus, available space and characteristics of the binding site are important features for the optimization of the seed.

The second strategy, the linking, starts from multiple fragments rather than from a unique seed with the aim of connecting them together. This method is equivalent to the experimental Structure-Activity Relationships by NMR approach[14]. The target needs to present distinct pockets in its binding site allowing the starting fragments to bind at different key interaction sites. The fragment linking method is notably employed in LUDI[15],

LEA3D[16], GANDI[17] and Fragment-Shuffling[18]. More recently, DeLinker[19], a deep learning tool, was developed to generate linkers between fragments by using 3D information.

The third and perhaps less employed strategy is the merging approach. It relies on the addition of multiple fragments presenting overlapping groups. Common structural features of the fragments are combined to create a novel molecule.

One of the main challenges for computational FBDD is to design synthesizable compounds. Thereby, some software applications such as LigBuilder 2.0[20] or PINGUI[21] include the estimation of the synthetic accessibility with the help of chemical rules applied during the growth and/or with retro-synthesis analysis.

Most of the FBDD programs spend the major part of their calculation time in the systematic exploration of the binding site, as well as in energy minimization of the compound under construction to find a stable conformation. At the same time, the number of experimental protein structures with co-crystallized ligand is continuously expanding, providing increasingly available structural data. From this observation, it appears relevant to develop a method exploiting these data so that the time dedicated to cavity exploration and energy minimization would be considerably reduced. Frags2Drugs (F2D) program uses fragments from co-crystallized ligands, or from docking experiments, and searches possible linkage between them to construct new molecules. F2D is a method based on both growing and linking methods. The growing is performed by linking fragments already positioned close to each other. F2D requires several fragments, coming from diverse protein-ligand complexes, in a close position to create covalent bounds while keeping their initial position. During this work, we were particularly interested in a specific class of protein, named protein kinases, having a similar ATP binding site and sharing similar structural fold. Protein kinases are enzymes catalyzing the transfer of the gamma phosphate group of ATP to a protein substrate. They are critical for intracellular signal transduction and deregulation of a member of this protein family may lead to diverse pathologies, including cancer, metabolic, autoimmune or neurological troubles.[22] Because of their determining role in transduction pathways, kinases were considered as important drug target by pharmaceutical companies and are still strongly studied.[23] From a structural point of view, the kinases share a common catalytic domain composed of two distinct lobes: a N-terminal lobe (N-lobe) and a C-terminal lobe (C-lobe) connected by a hinge region. The cleft formed between these two lobes is the ATP binding site, which forms hydrogen bonds with the hinge region stabilizing its position in the catalytic site (Figure 1). Most of kinase inhibitors on the market or still in development targets the ATP binding site[24] but other druggable pockets are known, as the back pocket[25] and other allosteric sites.[26] For computational chemists, working in the kinase field is facilitated by the large amount of structural data: 5,004 structures of human kinase domains are deposited in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank[27] (PDB, https://www.rcsb.org/, December 2020).

PKI are divided in six types depending on their binding modes[28]. The three first types are non-covalent ATP competitive inhibitors. Type I inhibitors bind in DFG-in kinase conformation, type I½ in DFG-in conformation with an access to an adjacent hydrophobic pocket and type II inhibitors bind kinases in DFG-out conformation lying from ATP site to the adjacent allosteric pocket. Other types are allosteric inhibitors (type III and IV), bivalent inhibitors (type V) and covalent inhibitors (type VI).

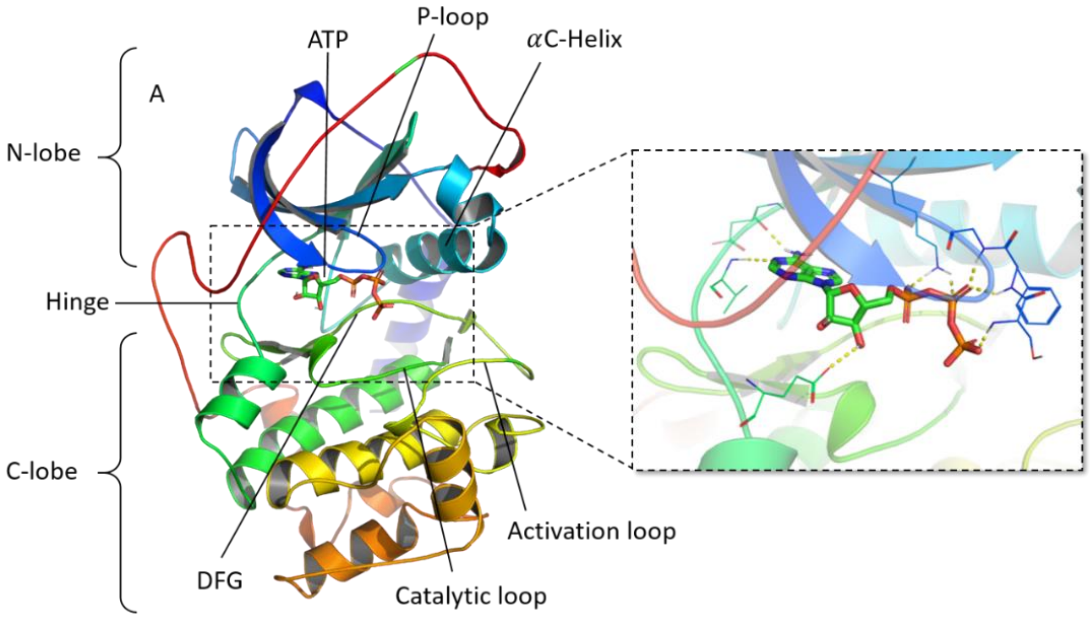


Figure 1. Representation of a protein kinase crystallographic structure (PDB ID: 1ATP, chain E) with a focus on the ATP binding site highlighting the interactions between ATP and the residues of the catalytic site.

RESULTS AND DISCUSSION

Protein kinase superimposition and fragment dataset

F2D is a FBDD program relying on 3D experimental data. Indeed, fragments obtained from 3D crystallographic protein kinase complexes fill the database used by F2D to generate ligands.

To create the fragment library of F2D, we first extracted protein kinase structures available from the RCSB PDB by keeping the kinase domains. We then performed a customized structural alignment on a reference which is the first crystallized kinase domain bound to ATP[29] (PDB ID: 1ATP, chain E, Figure 2). Once aligned, structures were cleaned to suppress ions, co-factors, crystallization agents and solvent molecules. Ligands were afterward extracted and fragmented while keeping their 3D coordinates, so their placement in the aligned active site is conserved and stored in the database.

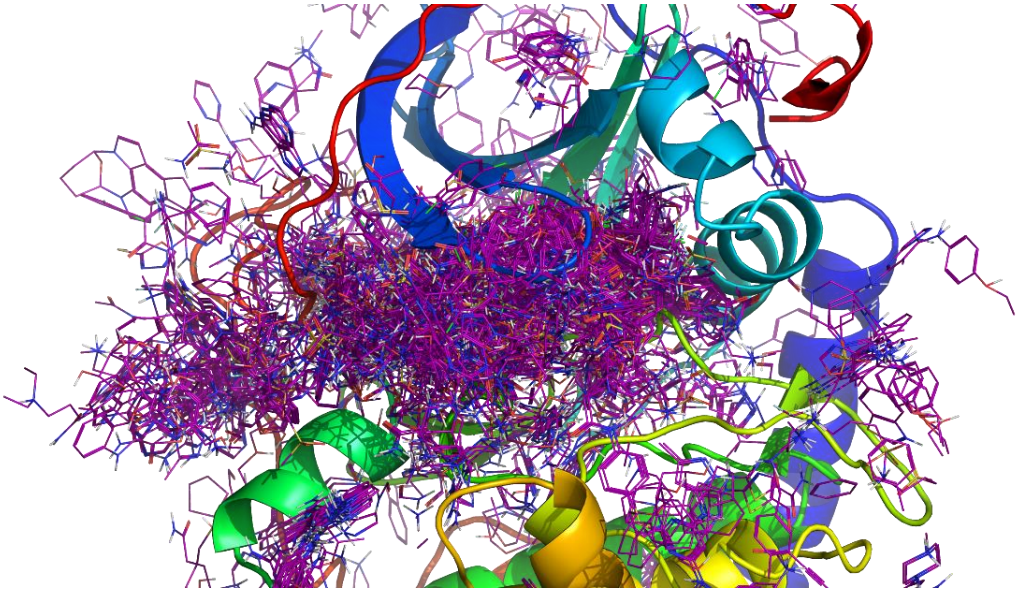


Figure 2: Graphical representation of all fragments after superimposition of all protein kinases on a common reference (PDB ID: 1ATP, chain E).

The Kinase–Ligand Interaction Fingerprints and Structure database (KLIFS) is composed of an alignment of 85 key residues from kinase binding sites[30]. A recent work applied a similar approach than the creation of the F2D database. Based on fragments extracted from KLIFS, they collected and fragmented co-crystallized ligands. This collection of fragments, called KinFragLib[31], was used to generate kinase inhibitors. Because it is based on KLIFS information, this work is limited to inhibitors from only human and mouse kinase-ligand complexes. In addition, fragments obtained in KinFragLib are restricted to type I and $I^{1/2}$ inhibitors. In F2D, we extended the database to all species available in the RSCB PDB and allow creating type II inhibitors in addition to types I and $I^{1/2}$.

The aim of F2D is to combine several fragments to create potential new protein kinase inhibitors (PKI). Thus, we decided to store each fragment in a graph database along with its 3D position and its ability to bind neighboring fragments. Indeed, a network is particularly suited to store information on connectable entities and thus, perfectly fulfils the aim of F2D. By exploring a graph of connected fragments, those fragments can be linked together to build a PKI.

Creation of the 3D fragment network

Fragments are represented by nodes in a 3D fragment network and are linked by edges (or relations) between fragment atoms. The graph database of F2D is composed of two types of nodes (fragment and protein) linked by three types of edges: inclusion, exclusion and compatibility. A sample of F2D graph database with only nodes and relations from the PDB ID 3OG7, and vemurafenib fragments is shown in Figure 3.

The first step to create the graph database consists in calculating the relations between all the fragments. These relations belong to two types: *inclusion* or *exclusion*. An *inclusion* relation means that the two considered fragments satisfy all the following conditions to bind together: it exists a couple of atoms, one from each fragment, having incomplete valence and far from a distance varying within a given interval. Moreover, their out of plane (OOP) angles, bond angles, and dihedral angles with their bond lengths also fit within predefined intervals. The definition of allowed intervals is based on force field values extracted from MMFF94[32–36] by taking into account the type and the hybridization of both involved atoms. This force field is usually used in drug design projects because it has already shown good performances on small molecules.

If the two fragments are too close or overlapped or if there is an unallowed angle value between them, an *exclusion* relation is established, and they cannot be linked. Furthermore, they cannot belong to a same built molecule, even if they have inclusion relations with a common intermediate fragment. If both fragments are too far apart, no relation is established between them. These rules are summarized in Table 1. Moreover, to avoid difficulties during synthesis of the built molecules,

other connection rules were added, such as prohibiting the creation of undesired substructures (see Methods).

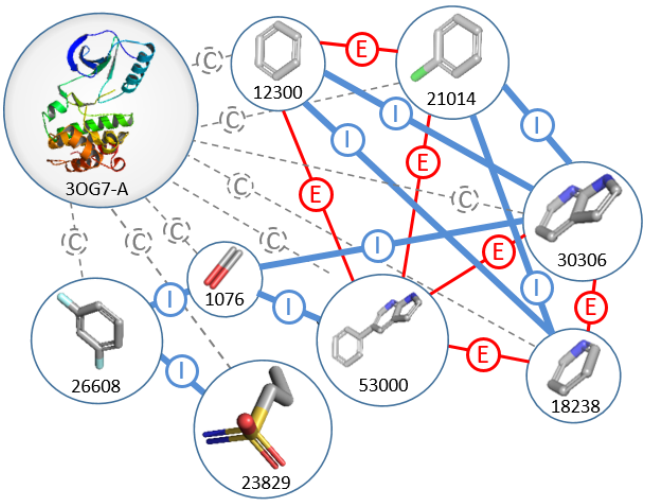
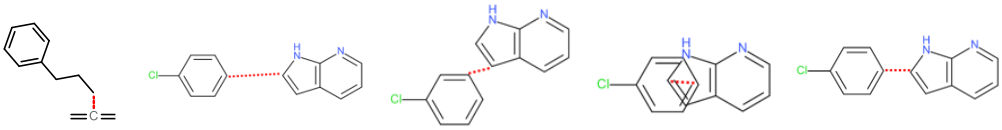


Figure 3. Representation of F2D graph database, focused on fragments obtained after fragmentation of vemurafenib. Nodes are 3D fragment structures and 3D target protein extracted from crystal structures. Edges are colored according to their type: exclusion (red, E), inclusion (blue, I) and compatibility (grey, C).

The second step of the creation of the graph database is the evaluation of allowed relations between the fragments and all the kinase residues. Fragments are kept if their atoms are far enough from all atoms in the residues. When a fragment is connected to a binding site with this third relation type, compatibility, this fragment is kept in the given cavity.

In the graph database, additional information is encoded in nodes and edges. The attribute Molecular Weight (MW) is assigned to fragment nodes and used to stop the addition of new fragments during future molecule generation. The inclusion edges, between fragment nodes that can be bound, contain complementary information. When an inclusion relation can be established between two fragments, both linkable atoms are saved, as well as the type of bond (simple, double or triple) that will be made. Linking fragments from molecules bound to different protein kinases may need a little distortion of angles and distance compared to reference values from the force field MMFF94[37]. We thus defined four parameters to set the thresholds of allowed distortions, three parameters for angles (bond, dihedral and OOP) and one for distance. By increasing these thresholds, F2D is able to generate more molecules which are in some cases slightly distorted.

Table 1. Conditions of the bond formation between two fragments.

Situation					
Valence	✗	✓	✓	✓	✓
Distance	✓	✗	✓	✓	✓
Angles (bond, dihedral and OOP)	✓	✓	✗	✓	✓
No steric hindrance	✓	✓	✓	✗	✓
Bond formation	✗	✗	✗	✗	✓

Molecules are drawn in 2D for visual purpose, they are in 3D in F2D database. On each situation, the dotted line represents the possible bond to form between both fragments.

Molecule generation and filtering process

F2D molecule generation is initialized with a fragment, referred as the seed, and a PDB structure of the targeted kinase. In drug design project, a seed can be drawn, in 2D, then added, in 3D, to F2D database after molecular docking. This approach is often used in bioisosteric replacement, in scaffold hopping or in novelty requirement for patent. Starting from the seed, F2D uses only fragments compatible with the targeted kinase.

Then, F2D searches all possible paths through bonding relations, and extracts the corresponding graph. To save time in further treatments, this graph is split in two subgraphs: the subgraph of inclusions (Gi) of nodes linked by inclusion relations and the subgraph of exclusions (Ge) with only exclusion relations between nodes. In Gi, F2D browses all possible paths starting from the seed. For each obtained path, all Ge subgraphs are removed from the graph. Once all possible paths

obtained, the construction of the molecules from fragments is performed. In order to build molecules of reasonable size, a threshold based of maximum molecular weight is defined. Beyond 650 Da, the growth is stopped, and no more fragments can be added.[38]

At this stage, several thousands of molecules could be constructed starting from one seed. However, due to the presence of similar positions of the same fragment extracted from different crystallized ligands, we remove duplicates based on the Root Mean Square Deviation (RMSD <= 1.0 Å) of the 3D conformation of the identical built molecules. We then applied molecular filters to keep only compounds belonging in the PKI chemical space. Some rules already exist for oral drug assessment, such as the Lipinski's Rule of Five (Ro5)[39]. According to this Ro5, a compound will probably not be efficiently absorbed by the organism if at least two of the following constraints are violated: MW ≤ 500, calculated logP (ClogP) ≤ 5, number of hydrogen bond acceptors (HBA) ≤ 10, number of hydrogen bond donors (HBD) ≤ 5. Other physicochemical properties, such as topological polar surface area (TPSA) and number of rotatable bonds (NRB), can also be used in drug design project (TPSA ≤ 140 Å² and NRB ≤ 10)[40]. In F2D, we applied the PKI-like filters from Protein Kinase Inhibitor Database (PKIDB).[38,41] In the same way as Lipinski's Ro5, we authorize the violation for 2 of the 8 rules (Table 2). After applying PKI-like filters, results from F2D are downloadable as an SDF file.

In F2D, the molecules are built into the binding site of the kinase target. No minimization is performed during the building process of the 3D molecules. Therefore, there is a need to check that the conformation of the molecules has reasonable strain energy. For this purpose, each molecule is docked into the bind site and the conformation of the docking pose is then compared to the molecules built by F2D. Conformations having a RMSD value less than 2.5 Å are kept for further analysis.

Table 2. PKI-like guidelines used to filter molecules built by Frags2Drugs

	Minimum	Maximum
MW (Da)	314	613
TPSA (Å²)	55	138
ClogP	0.7	6.3
HBA	3	10
HBD	0	4
NRB	1	11
NAR	1	5
NCA	0	2

NAR: Number of Aromatic Rings, NCA: Number of Chiral Atoms.

Optional post-processing tools have also been developed to reduce and optimize the number of compounds proposed by F2D. Molecules having Pan-Assay Interference Compounds (PAINS)[42] substructures can be identified and removed or tagged. To evaluate the synthetic feasibility of built molecules, a Synthetic Accessibility (SA) scores[43] can be calculated. Molecules with a SA score < 4 are selected. Similarly, the Central Nervous System Multiparameter Optimization (CNS MPO) score estimates the ability of molecules to pass through the Blood-Brain Barrier.[44] An upper threshold value of 4 is set to keep compounds. Finally, as the objective of F2D is to identify active and innovative compounds, we implemented a similarity search to compounds available in ChEMBL[45] database, in ZINC[46] or Ambinter[47] databases of purchasable molecules. We extended this search to PKI in clinical trials in the PKIDB database. Figure 4 summarizes the general F2D workflow. At the end of the process, a SDF file containing all the information on the origin of each fragment present in each built molecule is provided.

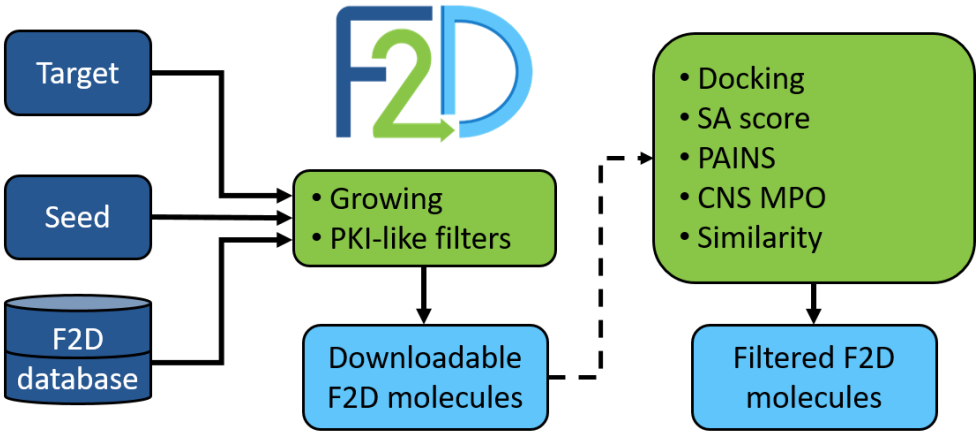


Figure 4. General F2D workflow. Three inputs are required to run F2D, a target and a seed (which can be either picked from F2D database or a new docked fragment) and F2D database. Inputs are colored in dark blue, F2D functions are colored in light green and outputs are colored in light blue. The dotted arrow indicates optional filtering steps.

Validation of F2D

In the first validation step of F2D, we aimed at reconstructing the 3,218 PKI-like co-crystallized ligands in their native protein kinase extracted from the RCSB PDB. Each ligand was fragmented to create a 3D fragment graph. Then, F2D started the growing process from the fragment bound to the hinge region. The tool succeeded in reconstructing 88.13% of the 3,218 co-crystallized kinase inhibitors. Figure 5 shows an example of the reconstruction of vemurafenib starting from the 7-azaindole seed.

382 ligands could not be reconstructed by F2D. However, these failures do not seem to be due to the F2D process itself but to the ligand fragmentation method or to errors in the crystal structures such as incorrect angles. Indeed, we observed in several cases the absence of fragment in the database required to reconstruct the ligand. Fragmentation methods employed in F2D sometimes lead to fragments with MW > 300 Da which are removed in the process.

After this first validation step, we used F2D in internal research programs to ensure is ability to find innovative and active compounds. Firstly, we searched for BCR-ABL inhibitors, secondly for BRAF V600E inhibitors and thirdly for MELK inhibitors (type I and type II inhibitors). These three applications of F2D are detailed below.

Discovery of BCR-ABL inhibitors

Breakpoint Cluster Region gene- Abelson proto-oncogene (BCR-ABL) is a member of the tyrosine kinase family and is involved in a mechanism which can lead to chronic myeloid leukemia (CML).[48] Before discovering its involvement in CML, classical anticancer treatments were given to patients such as busulfan and hydroxyurea[49]. Due to the cytotoxic effects of these compounds, recombinant interferon-alfa has also been employed[50]. However, thanks to the evolution of the knowledge on CML and with the availability of structural data[51], the first inhibitor targeting specifically BCR-ABL, imatinib[52], reached the market in 2001 becoming the reference treatment of CML. Because of the resistance to this drug observed in several patients, other inhibitors have been developed since then, such as dasatinib, nilotinib, ponatinib and bosutinib[53]. As BCR-ABL is today one of the most studied kinase, we chose this protein as target of our first F2D trial. The structure used to assess the F2D process is the wild-type ABL kinase domain co-crystallized with imatinib (PDB ID: 2HYY, chain A). The pyridine group of imatinib, which is bound to the hinge region, has been selected as the seed (Figure 6A). We aimed to grow the molecule from this seed toward the hydrophobic pocket of the cavity. Therefore, after a visual inspection of the seed position in the 3D complex, we specified as unique starting point the C atom on position 3 (Figure 6).

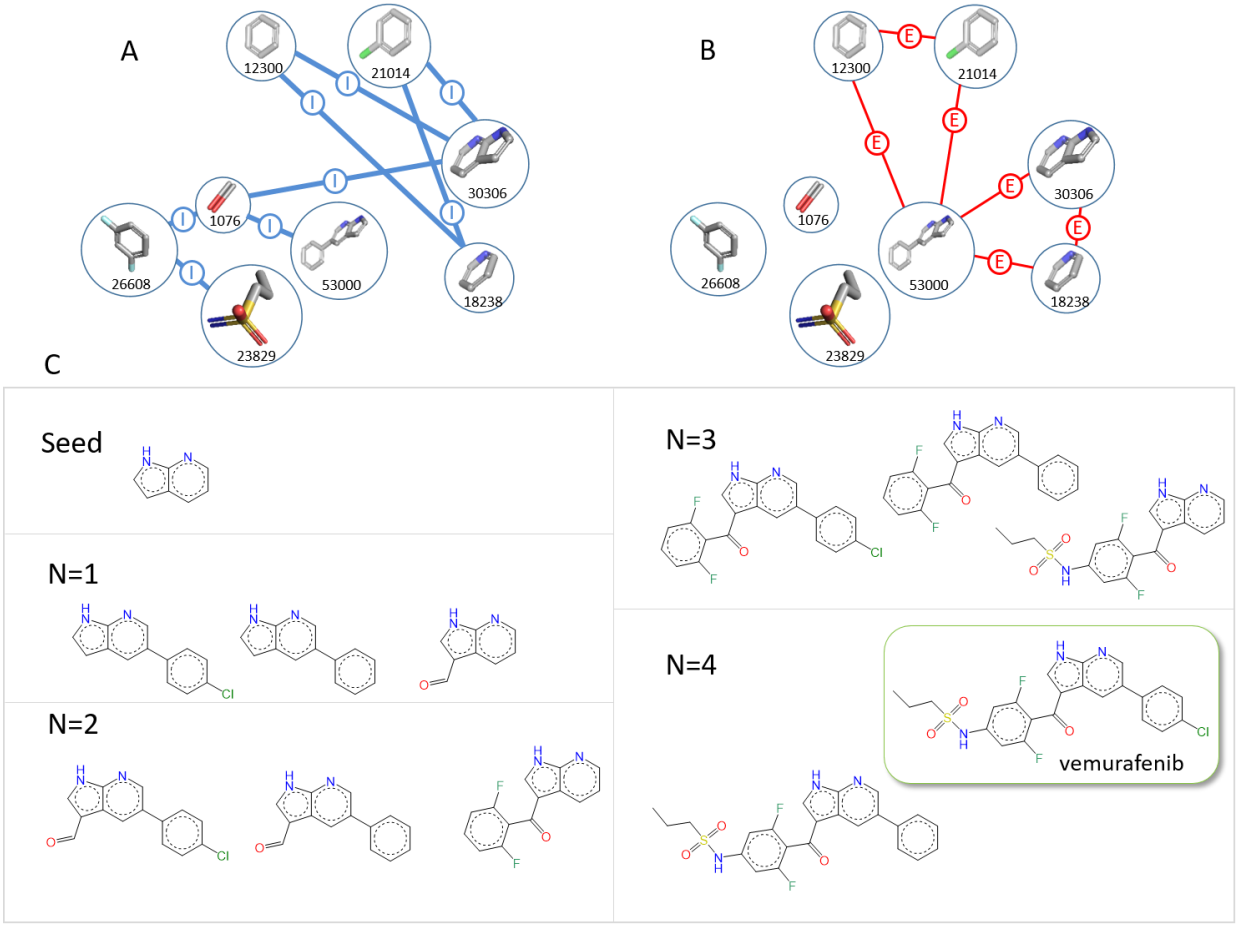


Figure 5. Example of validation of the reconstruction of a molecule, vemurafenib (PDB ID: 3OG7). (A) Detailed Gi. (B) Detailed Ge. (C) Representation of created molecules for each addition (N) of a new fragment, starting from the seed 7-azaindole. For a better visualization, molecules are shown in 2D.

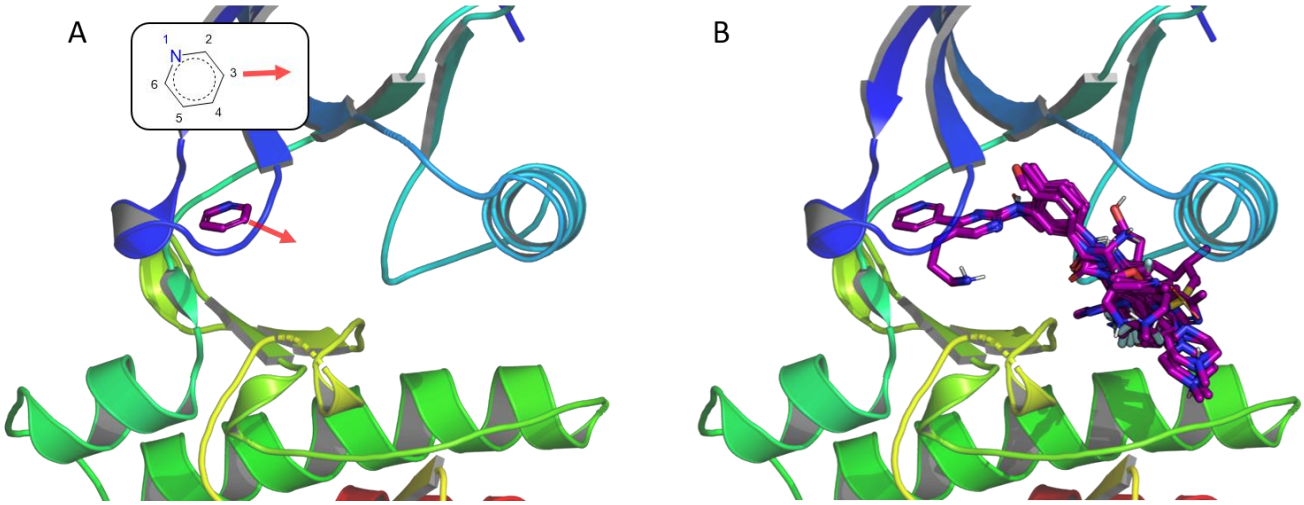


Figure 6. Example of F2D on BCR-ABL (PDB ID: 2HYY, chain A). (A) Initial position of the seed from co-crystallized complex, the red arrow indicates the starting atom and the direction for growing. (B) Representation of all the 50 built molecules obtained by F2D.

F2D generated 2,595 molecules and clustered them (RMSD = 1.0 Å) resulting in 1,128 molecules, 514 of which follow PKI-like filters as described in Table 2. Then, we applied the filtering process of F2D workflow. Among the 514 molecules, 9 molecules were removed because of unwanted PAINS substructure and 495 molecules have a SA score < 4. These 495 remaining molecules were docked to check if the combination of several fragments from different PDB complexes corresponds to a favorable conformation in the kinase active site. 138 molecules obtained a RMSD value less than 2.5 Å between the docking poses and the molecules built by F2D. In most of these compounds, we retrieved the core substructure of imatinib such as N-phenyl-4-(3-pyridyl)pyrimidin-2-amine.

We then searched for similar molecules in the ChEMBL, ZINC, Ambinter and PKIDB databases to check the novelty of the built compounds and found 50 new molecules among the 138 retained. The 3D superimposition inside the binding site is shown on Figure 6B. The 5 best molecules, according to the Quantitative Estimate of Druglikeness (QED) score[54], are shown in Figure 7A. Their QED scores range from 0.63 to 0.79. The expertise of medicinal chemists is then required to select and optimize these molecules.

The 88 other molecules present similarities with at least one molecule, including imatinib, nilotinib, radotinib, flumatinib, bafetinib and masitinib, by considering a similarity threshold of 0.7. Two of them correspond to a known PKI: imatinib and nilotinib (Figure 7B).

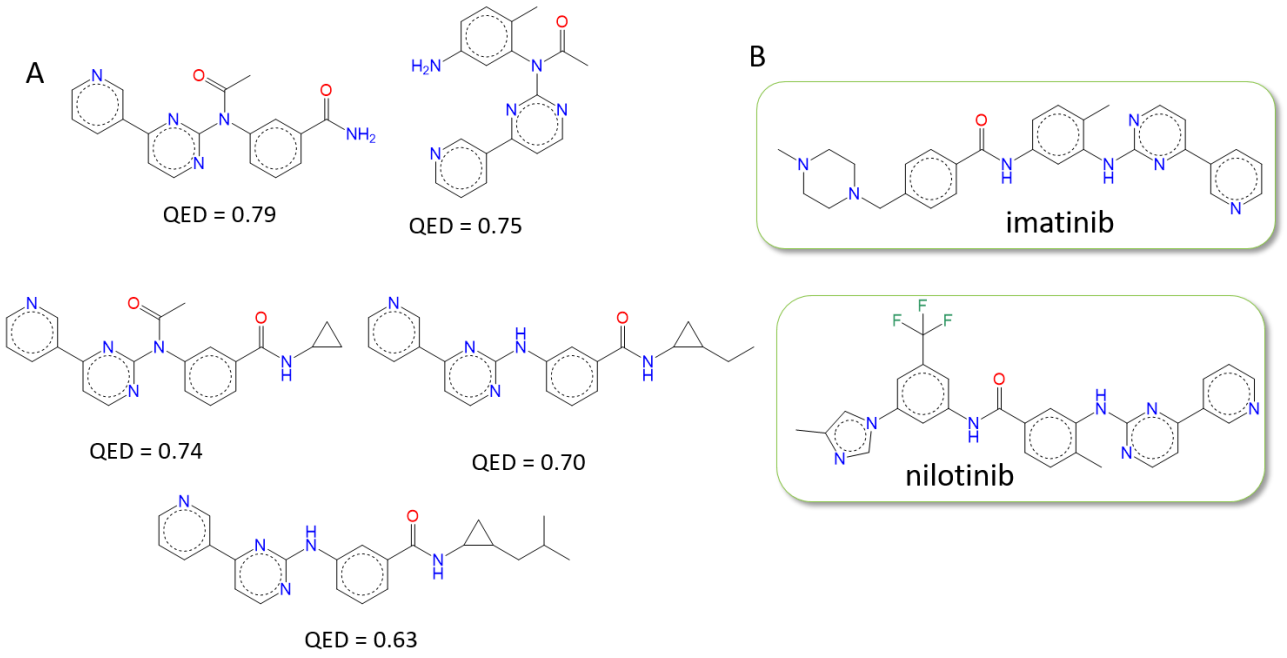


Figure 7. Structures of molecules obtained by generating molecules with F2D on BCR-ABL from a pyridine. (A) Top 5 of the 50 new molecules sorted by QED score. (B) Two known PKI found among the results: imatinib and nilotinib.

Discovery of BRAF V600E inhibitors B–Rapidly Accelerated Fibrosarcoma (BRAF) is a serine/threonine kinase and a member of the RAF family[55]. BRAF is mutated in several cancers[56], the common mutation V600E is mostly related to melanoma[57]. Vemurafenib (Zelboraf®) is targeting BRAF V600E[2], it was found by using structural information from sorafenib (Nexavar®), a dual BRAF and CRAF inhibitor[58] having a low inhibition on BRAF V600E. As vemurafenib is the first approved drug derived from FBDD, we chose BRAF as second example of target for F2D. The complex structure provided as input to F2D is PDB ID 3OG7, (chain A), which corresponds to the structure of BRAF V600E co-crystallized with vemurafenib. The 7-azaindole bound to the hinge was selected as a seed (Figure 8A).

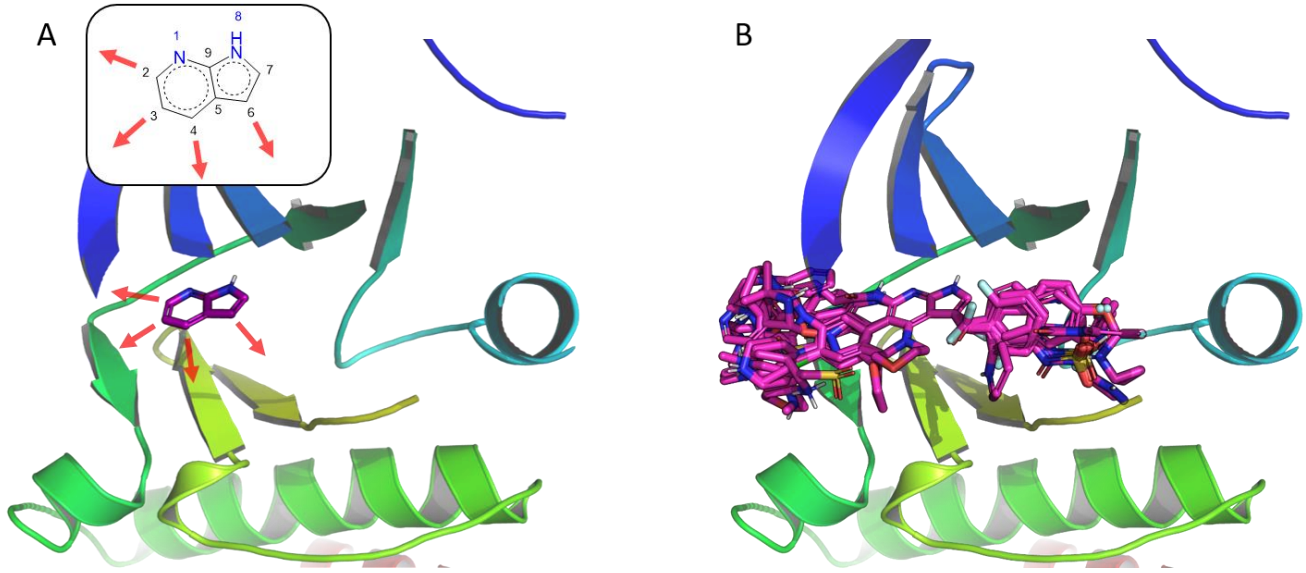


Figure 8. Example of F2D on BRAF V600E (PDB ID: 3OG7, chain A). (A) Initial position of the seed from co-crystallized structure, the red arrows indicate starting atoms and directions for growing. (B) All the 510 new molecules obtained from F2D.

We obtained 69,868 unique PKI-like molecules. We applied F2D filtering steps, removing PAINS substructures to reduce the number of results to 69,563 and keeping 12,378 molecules with SA score < 4. After docking, 525 molecules present a RMSD value < 2.5 Å compared to built molecules from F2D.

Among the 525 molecules, 510 have no match in ChEMBL, ZINC, Ambinter or PKIDB databases using a similarity threshold of 0.7 (Figure 8B). The only common substructure is the pyrrolo[2,3-*b*]pyridine seed. The higher number of novel molecules found on BRAF V600E target compared to BCR-ABL can be explained by the four growing atoms. 15 molecules among the 510 have a QED >= 0.75 and are shown on Figure 9.

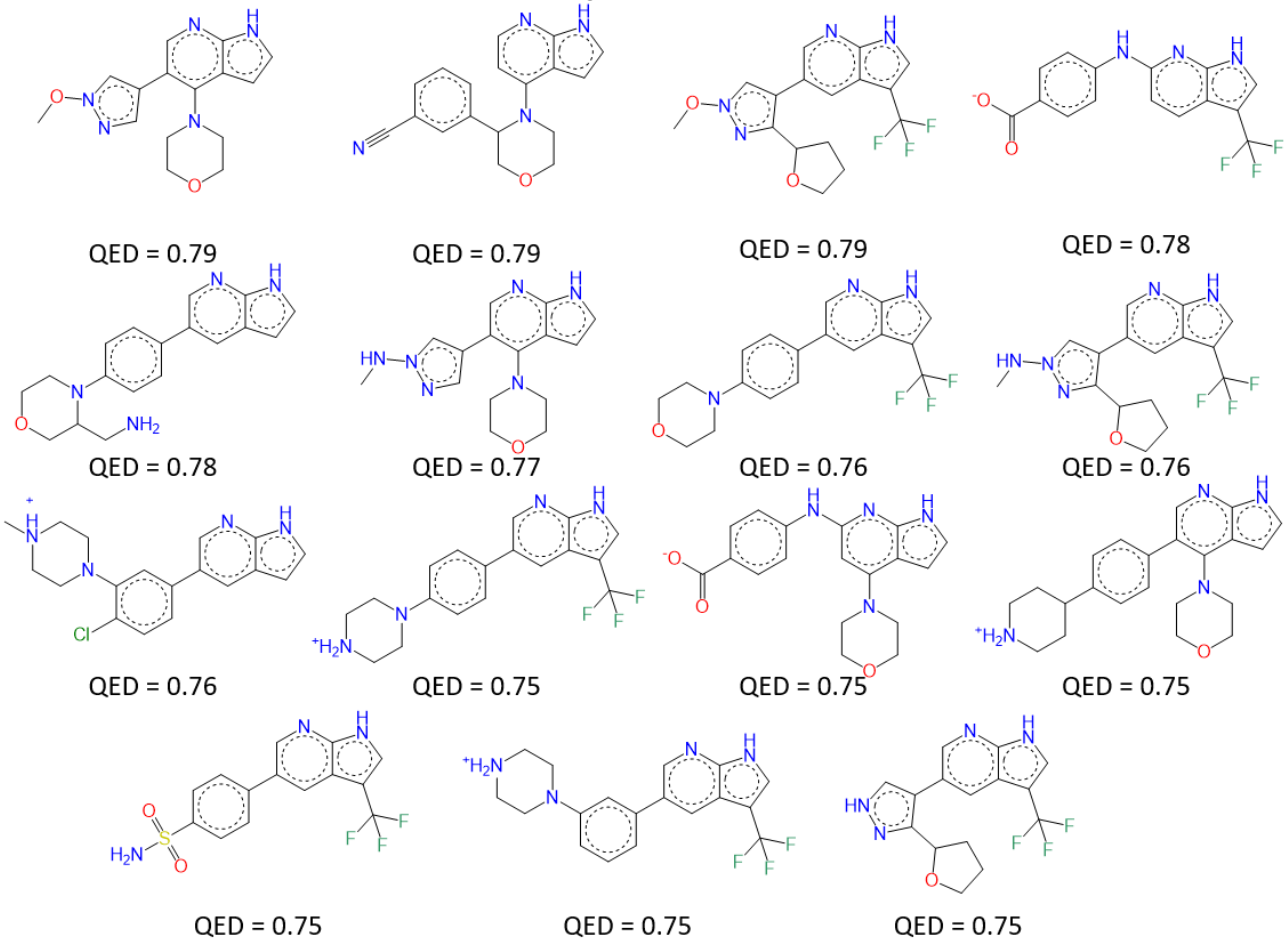


Figure 9. Structures of the 15 molecules with the highest QED scores generated by F2D on BRAF V600E using the pyrrolo[2,3-*b*]pyridine as a seed.

Discovery MELK inhibitors

Maternal Embryonic Leucine zipper Kinase (MELK) is a serine/threonine kinase, member of the CAMK family[59]. Its deregulation seems to be involved in many cancers[60], even if this point is a controversial subject in the literature[61]. Type I and Type II MELK Inhibitors have already been discovered[62], also from FBDD approaches.[63,64]

Type I inhibitors

We first searched new type I inhibitors of MELK. The target and the seed were chosen from structure of the PDB ID 4UMQ, which is a structure of MELK co-crystallized with 3-{5-[(3-hydroxy-5-methoxyphenyl) amino] -2- (phenylcarbamoyl) phenoxy} propan-1-aminium. This structure has both DFG motif and αC-helix in *in*-conformation. The seed positioned in the protein active site and the starting growing directions are shown in Figure 10A.

87,961 compounds were generated, reducing to 86,669 without PAINS substructures and only 3,934 molecules with a SA score < 4. After the docking step, this number decreased to 41 molecules. Their 3D structures are presented in Figure 10B. Visualization of the molecules confirms their type 1 binding mode.

All the 41 molecules, selected after applying F2D filtering steps, are new molecules. They have no similar compound in ChEMBL, Ambinter, ZINC or PKIDB. Molecules with the highest QED scores (QED >= 0.6) are drawn in Figure 11.

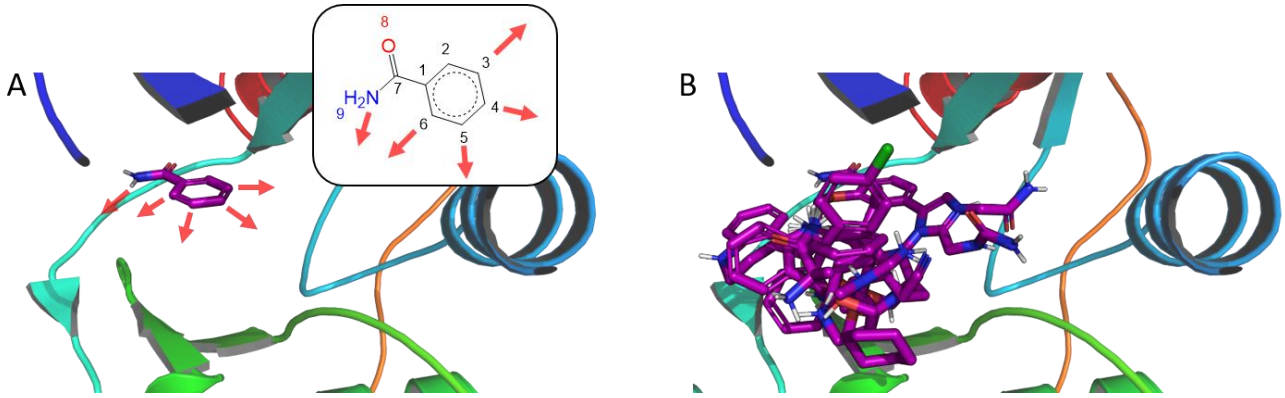


Figure 10. Example of F2D to discover type I inhibitors on MELK (PDB ID: 4UMQ, chain A). (A) Initial position of the benzamide seed, the red arrows indicate starting atoms and directions for growing. (B) All the 41 new molecules obtained from F2D.

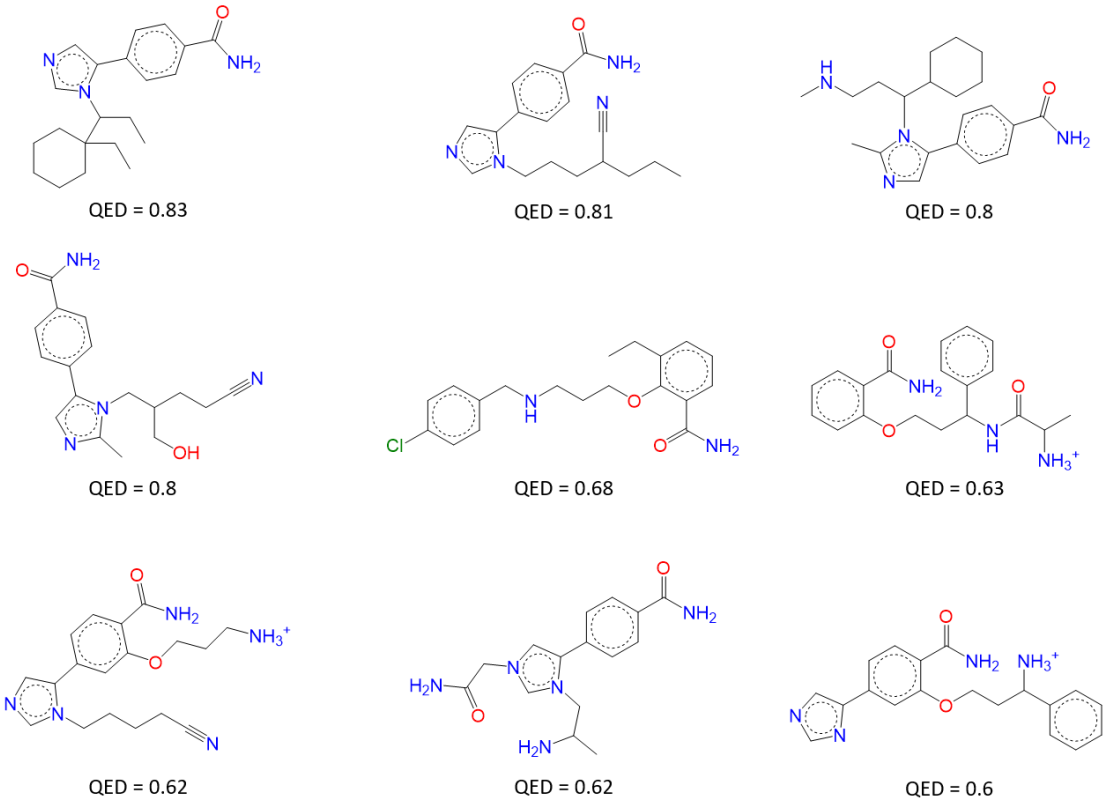


Figure 11. Structures of the 9 type I MELK inhibitors, having the highest QED scores, generated by F2D using a benzamide seed.

Type II inhibitors

Secondly, we focused on the discovery of new type 2 inhibitors of MELK. We selected the PDB ID 4UMT as a protein target, which is the structure of MELK co-crystallized with 1-(4-{[3-(isoquinolin-7-yl)prop-2-yn-1-yl]oxy}-2-methoxybenzyl)piperazinediium.[64] The conformation of the target has a DFG-*in* motif and an *out*-form of the αC helix. We started the fragment growing from the 3-(isoquinolin-7-yl)prop-2-yn-1-ol seed (Figure 12A).

We found 63 unique PKI-like molecules. We then applied F2D filtering, 3 molecules were removed because they contained PAINS substructures and 18 molecules were kept with SA scores < 4. 42 molecules were docked into MELK active site, 14 docking poses provided an RMSD value < 2.5 Å compared to molecules built in F2D.

Over the 14 molecules, 6 molecules are similar to at least one compound in ChEMBL, and none in the other databases. Five molecules were evaluated[64] *in vitro* (CHEMBL3355060, CHEMBL3355061, CHEMBL3355062, CHEMBL3355063, and CHEMBL3355066). The sixth molecule is CHEMBL215098.

The eight other molecules are novel, since there is no similar compound in ChEMBL, ZINC, Ambinter or PKIDB databases (Figure 12B). Their QED scores vary

from 0.45 to 0.73. The 2D structures of molecules with QED score ≥ 0.5 are shown in Figure 13.

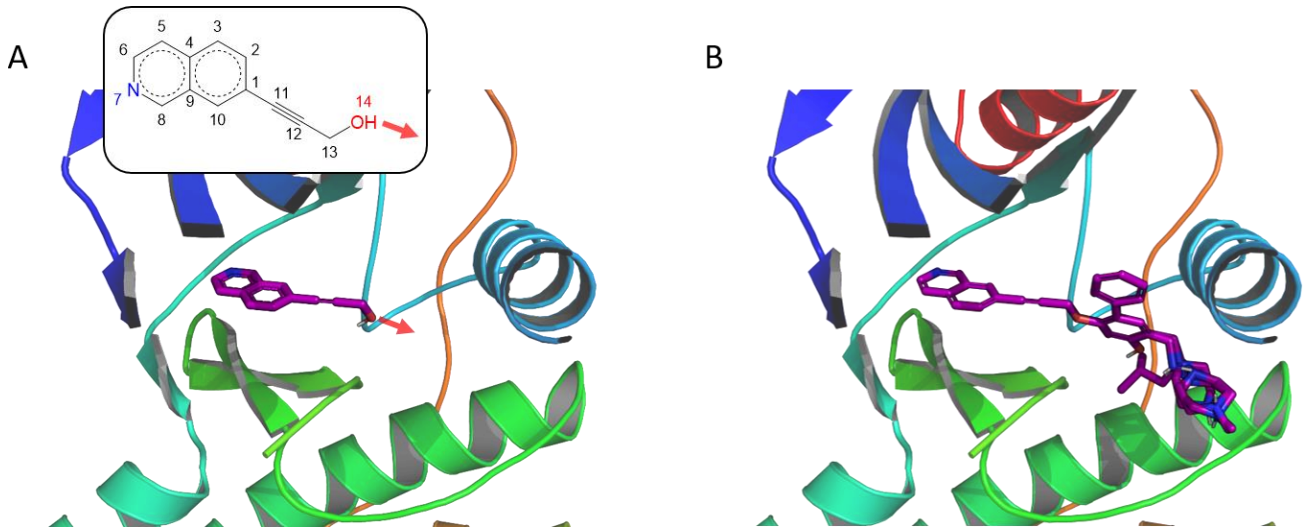


Figure 12. Example of F2D on MELK (PDB ID: 4UMT, chain A). (A) Initial position of the 3-(isoquinolin-7-yl)prop-2-yn-1-ol seed, the red arrow indicates the starting atom and the direction for growing. (B) Eight new molecules obtained by F2D in MELK.

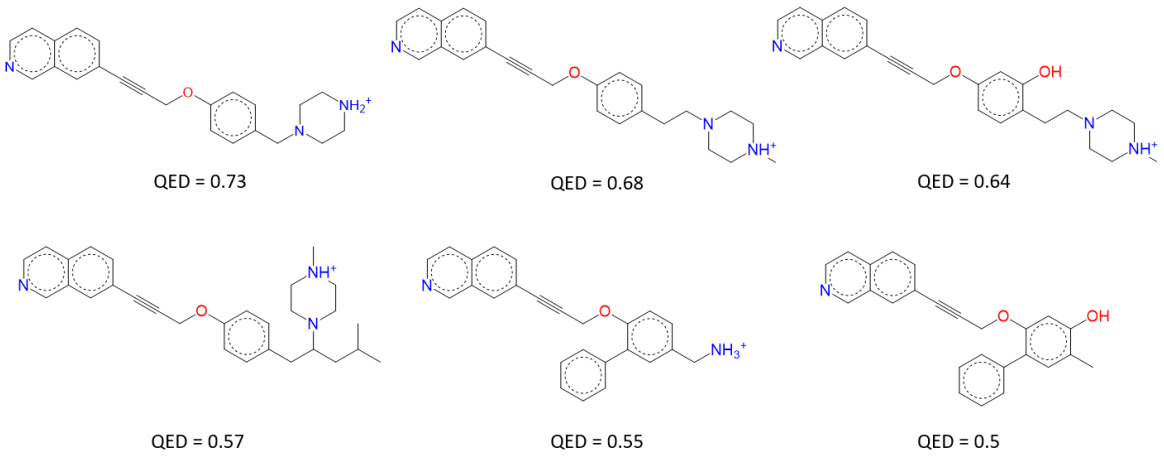


Figure 13. Structures of the 6 type II inhibitors, having the highest QED score, generated by F2D in MELK from the 3-(isoquinolin-7-yl)prop-2-yn-1-ol seed.

Frags2Drugs Website

F2D has been created to help researchers (medicinal and computational chemists, biologists) to find new protein kinase inhibitors. Thus, we provide an online interface to use Frags2Drugs. This interface is freely available at http://frags2drugs.icoa.fr. F2D needs several input data which is a target structure, a seed structure and, optionally, the definition of the atoms for growing. This leads to four initial steps required to launch F2D, presented in a user guide. Firstly, the user has to select the desired target using the protein name (e.g. BRAF). A table of the PDB ID corresponding to this protein name thus appears and the user can select the target structure. Finally, the user chooses a seed and the atoms from which the growing will start. The four selection steps are provided in a user-friendly interface.

Once validated, F2D runs and provides an URL to retrieve the results. Once available, the results are displayed as shown in Figure 14. On the top of this result page, the used parameters and the number of generated molecules are detailed in a table. The 2D representation of built molecules are displayed below in an 8 by 8 matrix. The user may select a molecule to get additional details. Results are downloadable as SDF file to allow researchers for further treatments. All results launched on F2D website are accessible through the home page during one month.

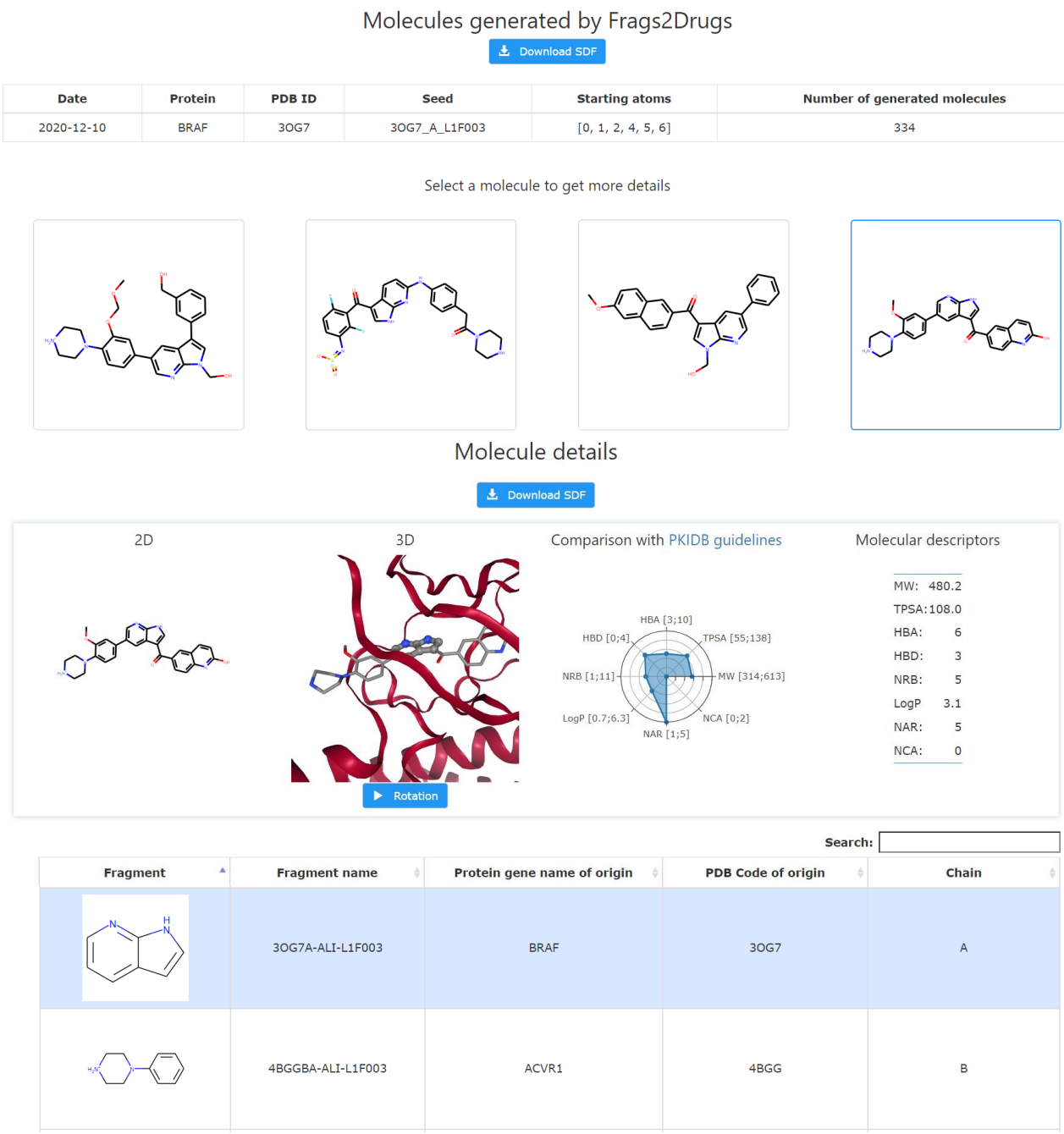


Figure 14. Example of results obtained from Frags2Drugs website. This figure shows 4 of the 334 resulting molecules. On the website, all the molecules are available. The table showing the details of each fragment from the selected molecule is presented.

CONCLUSIONS

In this article we have presented Frags2Drugs, a new *in silico* tool based on a 3D fragment network and a mixed growing/linking methodology to find novel protein kinase inhibitors. We have demonstrated its successfulness on several examples to generate type I and type II inhibitors of protein kinases. The main advantage of F2D is to work directly with 3D experimental structures without the need of energy minimization after each fragment growing step. Furthermore, no systematic cavity exploration is needed, the results can be obtained generally in few hours.

F2D is available at http://frags2drugs.icoa.fr/. In few time, the program provides a set of PKI-like molecules built from a chosen seed inside the desired target.

F2D is able to create type I, type I$^{1/2}$ and type II inhibitors of protein kinases.

LIMITATIONS AND PERSPECTIVES

F2D is currently focused on the design of protein kinase inhibitors. A great improvement would be to broaden

the application of F2D to any family of proteins. For this purpose, we need to deeply revise the protocol. Several methods may be envisaged as aligning cavities from different family of proteins[65] or moving fragment from an active site to another using subpockets[66] or shape-based descriptors.[67]

A key element of F2D database creation is the alignment of all protein kinases, so that the binding sites of solved structures are well superimposed. This imposes that the seed position is defined inside the binding site and leads the building of type I, type I$^{1/2}$ and type II inhibitors. Therefore, allosteric inhibitors outside of the active site cannot be discovered by F2D. Once again, F2D could be improved by adding more fragments in allosteric sites by using docking experiments.

The use of F2D is also restricted to proteins with available crystallographic structures. All the protein kinase space could be covered by using bioinformatics methods such as homology modelling to predict the 3D structure of kinases.

The fragmentation of molecules is another key point of the database creation. However, despite the use of published fragmentation algorithms to achieve this step, several known inhibitors could not be reconstructed because of fragmentation issues. New fragmentation methods need to be developed to override this limitation. This could be achieved by modifying current implemented methods or using other established methods such as DAIM[68] or Ftrees-FS.[69]

We have implemented a filtering procedure to reduce the number of molecules given by F2D, based on structural and physicochemical features. However, the estimation of the binding affinity of the molecules is not currently considered. To guide the choice of the most interesting compounds, it could be wise to implement a method predicting protein-ligand binding affinities by using deep learning-based tools, such as PAFnucy.[70]

METHODS

Protein kinases superimposition and fragment dataset

PDB ID of protein kinases were retrieved by using a request based on specific PFAM[71] identifiers: the PFAM id "PF07714" indicates protein tyrosine kinase and the PFAM id "PF00069" refers to protein kinase domain. These PFAM accession numbers gave us access to 3,809 structures, regardless of the origin of the organism (July 2016). These structures were then split by chain using Biopython tools[72] (1.72). Structures having alternative positions of the ligand in the crystal structure were duplicated to treat cases independently. The alignment of all protein kinase structures on the reference protein kinase with the PDB ID: 1ATP chain E was performed with Molecular Operating Environment (MOE, Chemical Computing Group, 2016_0802).[73]

Several algorithms already implemented in KNIME[74] were used to fragment the ligands : RECAP[75], BRICS[76] and Scaffold Tree.[77] By using three methods, we aimed at providing more fragments for molecular generation. Fragments were then standardized using VSPrep[78], a KNIME workflow recently developed. At this stage, a filter is then applied on fragments. Fragment having a MW > 300 Da, a phosphate, a carbohydrate, an ester, N-N, N-O and O-O substructures are removed. The next step consists of removing 3D duplicated fragments. The fragments are first grouped using their SMILES formula. Then, the raw RMSD matrix is calculated between each member in a group. For each pair of fragments having a RMSD value less than 0.25 Å, only the first one was kept.

Starting from 6,204 initial ligands, the database finally contains 72,480 fragments. Each of these fragments is labelled with a unique identifier and information about its origin (PDB ID, chain, alternative location, Uniprot gene name and species).

Creation of the 3D fragment network

We used Python (3.6.8), neo4j (3.5.11) and the Python library networkx (2.2) to create the fragment network.

Calculations of compatibility relationship are performed in two steps. Firstly for each protein, its binding site cavity and the surrounding residues are identified using FPocket[79], an open-source cavity detection tool based on Voronoi tessellation. Secondly a fragment is defined as "compatible" with a cavity if the distance between the nearest atoms of the residue and the fragment range between 1.4 Å and 3.5 Å. Thus, we avoid both clashes with the protein and building compounds far from the active site of the target.

Between each fragment to be covalently linked, the four calculated distortion parameters for bond length, bond angle, dihedral angle and out of plane angle range from 0 to 45, 0 meaning there is no distortion compared to reference value from the force field. For angles, increasing the parameter value by one means an increase of one degree, while for distance, it corresponds to an increase percentage of allowed deviation from the reference value.

Molecule generation and filtering process

Nodes from the graph database are associated to corresponding structures of fragments with the use of Pandas dataframe (1.0.5) and RDKit[80] (2020.03.3). RDKit is also used to link fragments together to generate molecules.

During the filtering process, substructure recognition relies on SMARTS patterns[81] and are applied by substructure match from RDKit. The PKI-like filter removes molecules containing undesired substructures after the combination of several fragments. These substructures are ester, aldehyde, acyclic N-N, N-O and O-O.

The molecular docking step is performed with rDock[82] (2013.1) software. Before proceeding to docking, molecules are standardized. To avoid using the 3D conformations of the built molecules, new molecular coordinates are generated using the RDKit method of conformation generation ETKDG[83] with the MMFF94 force field refinement.[37] RMSD between molecules from F2D and the docking poses is calculated with RDKit. We chose a threshold of 2.5 Å from minimal RMSD values; this value was obtained by applying the same procedure on all known kinase-ligand crystal structures from the PDB. SA score and PAINS substructure are evaluated

using RDKit modules. CNS MPO is calculated using ChemAxon's Calculator plugins, Marvin 15.2.9, 2015, ChemAxon (https://www.chemaxon.com), and RDKit.

Similarity searches performed on whole databases uses FPSim2.[84] The similarity search is based on both Tanimoto[85] and Tversky coefficient[86] with a similarity threshold of 0.7 and the parameters values α=1 and β=0.5 for Tversky coefficient. The versions of the four databases are ChEMBL 27 (access June 2020), ZINC 15 (access June 2020), Ambinter Feb. 2020 and PKIDB 2020-12-09 http://www.icoa.fr/pkidb/.

Validation of F2D methodology

From the 6,204 initial ligands, we first selected 4,338 by keeping PKI-like molecules and removing staurosporine-like molecules (PDB ligands ID: STU, UCN and LY4). We removed these compounds because, due to the fragmentation method, they could not be rebuilt. Indeed, these compounds contain a glycosylated indolocarbazole which is not fragmented by the algorithms. Due to its MW > 300Da, this group is too heavy to be kept in the database. This led to 3,218 ligands to be reconstructed for F2D validation.

For each of the 3,218 ligands, we selected in the fragment network the nodes coming from its fragmentation. We ran F2D on this selection of fragments, using each fragment alternatively as a seed, to generate molecules until the ligand is found. After each generation of molecules, we check the presence of the original ligand among the created molecules. If the ligand is not found, we repeat the process using another seed.

F2D Website

F2D is freely accessible at http://frags2drugs.icoa.fr, this website is based on several Docker 20.10.5 containers communicating through docker-compose 1.28.6. The python web framework Django LTS 2.2 and Celery 4.4.6 are used to run the website and F2D calculations in a task queue. The database relies on PostgreSQL 12.2 and the RDKit cartridge 2020.03.

Discovery of inhibitors

The 3D position of the seeds used in the four examples has been obtained by removing all atoms, except those of the seed, from the co-crystallized molecules.

The four distortion parameters used while growing molecules on the first four examples presented in this article are presented in Table 3.

Table 3. Distortion parameters used in F2D during fragment growing.

	Torsion angle	Out of plane angle	Dihedral angle	Distance
BCR-ABL	15	15	15	10
BRAF V600E	10	10	10	10
MELK Type I	10	10	10	10
MELK Type II	14	14	14	12

AUTHOR INFORMATION

Corresponding Author

Pascal Bonnet – Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311 Université d'Orléans BP 6759, 45067 Orléans CEDEX 2, France; *Email :* pas*cal.bonnet@univ-orleans.fr; Phone*: +33 2 38 41 72 54

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript. ‡G.P., C.B.: These authors contributed equally.

Funding Sources

N Bosc, F Carles, J. Fogha and P Bonnet are grateful to the Région Centre Val de Loire for financial support. S. Aci-Sèche, S. Bourg and P. Bonnet are supported by LABEX SynOrg (ANR-11-LABX-0029). We gratefully acknowledged Financial support by the "Ligue contre le Cancer, Conseil Interrégional Grand Ouest (CSIRRGO)". The authors also thank the "Cancéropôle Grand Ouest" (axis: natural sea products in cancer treatment), IBiSA (French Infrastructures in living sciences: biology, health and agronomy) and Biogenouest (Western France life science and environment core facility network) for supporting the KISSf screening facility.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The authors thank Laurent Robin, Juliette Douare and Romain Launay for technical assistance in the web development.
This project was financially supported by Région Centre Val de Loire, LABEX SynOrg (ANR-11-LABX-0029).

ABBREVIATIONS

BRAF, B–rapidly accelerated fibrosarcoma; BCR-ABL, breakpoint cluster region gene Abelson proto-oncogene; ClogP, calculated logP; CNS MPO, central nervous system multiparameter optimization; CML, chronic myeloid leukemia; FBDD, fragment-based drug design; F2D, Frags2Drugs; Ge, graph of exclusions; Gi, graph of inclusions; KLIFS, kinase-ligand interaction fingerprints and structure database; Ro5, Lipinski's rule of five; MELK, maternal embryonic leucine zipper kinase; MW, molecular weight; NMR, nuclear magnetic resonance; NAR, number of aromatic rings; NCA, number of chiral atoms; HBA, number of hydrogen bond acceptors; HBD, number of hydrogen bond donors; NRB, number of rotatable bonds; OOP, out of plane; PAINS, pan-assay interference compounds; PDB, protein data bank; PKIDB, protein kinase inhibitor database; PKI, protein kinase inhibitors; QED, quantitative estimate of druglikeness; RCSB, research collaboratory for structural bioinformatics; RMSD, root mean square deviation; SA, synthetic accessibility; 3D, three dimensional; TPSA, topological polar surface area; FDA, U.S. food and drug administration

REFERENCES

(1) Scott, D. E.; Coyne, A. G.; Hudson, S. A.; Abell, C. Fragment-Based Approaches in Drug Discovery and Chemical Biology. *Biochemistry* **2012**, *51* (25), 4990–5003. https://doi.org/10.1021/bi3005126.

(2) Bollag, G.; Tsai, J.; Zhang, J.; Zhang, C.; Ibrahim, P.; Nolop, K.; Hirth, P. Vemurafenib: The First Drug Approved for *BRAF*-Mutant Cancer. *Nature Reviews Drug Discovery* **2012**, *11* (11), 873–886. https://doi.org/10.1038/nrd3847.

(3) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nature Reviews Drug Discovery* **2016**, *15* (9), 605–619. https://doi.org/10.1038/nrd.2016.109.

(4) Perera, T. P. S.; Jovcheva, E.; Mevellec, L.; Vialard, J.; Lange, D. D.; Verhulst, T.; Paulussen, C.; Ven, K. V. D.; King, P.; Freyne, E.; Rees, D. C.; Squires, M.; Saxty, G.; Page, M.; Murray, C. W.; Gilissen, R.; Ward, G.; Thompson, N. T.; Newell, D. R.; Cheng, N.; Xie, L.; Yang, J.; Platero, S. J.; Karkera, J. D.; Moy, C.; Angibaud, P.; Laquerre, S.; Lorenzi, M. V. Discovery and Pharmacological Characterization of JNJ-42756493 (Erdafitinib), a Functionally Selective Small-Molecule FGFR Family Inhibitor. *Mol Cancer Ther* **2017**, *16* (6), 1010–1020. https://doi.org/10.1158/1535-7163.MCT-16-0589.

(5) Hall, R. J.; Mortenson, P. N.; Murray, C. W. Efficient Exploration of Chemical Space by Fragment-Based Screening. *Progress in Biophysics and Molecular Biology* **2014**, *116* (2), 82–91. https://doi.org/10.1016/j.pbiomolbio.2014.09.007.

(6) Grove, L. E.; Vajda, S.; Kozakov, D. Computational Methods to Support Fragment-Based Drug Discovery. In *Fragment-based Drug Discovery Lessons and Outlook*; John Wiley & Sons, Ltd, 2016; pp 197–222. https://doi.org/10.1002/9783527683604.ch09.

(7) Durrant, J. D.; Amaro, R. E.; McCammon, J. A. AutoGrow: A Novel Algorithm for Protein Inhibitor Design. *Chemical Biology & Drug Design* **2009**, *73* (2), 168–178. https://doi.org/10.1111/j.1747-0285.2008.00761.x.

(8) Durrant, J. D.; Lindert, S.; McCammon, J. A. AutoGrow 3.0: An Improved Algorithm for Chemically Tractable, Semi-Automated Protein Inhibitor Design. *Journal of Molecular Graphics and Modelling* **2013**, *44*, 104–112. https://doi.org/10.1016/j.jmgm.2013.05.006.

(9) Spiegel, J. O.; Durrant, J. D. AutoGrow4: An Open-Source Genetic Algorithm for de Novo Drug Design and Lead Optimization. *Journal of Cheminformatics* **2020**, *12* (1), 25. https://doi.org/10.1186/s13321-020-00429-4.

(10) Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space. *J. Chem. Inf. Model.* **2009**, *49* (7), 1630–1642. https://doi.org/10.1021/ci9000458.

(11) Chéron, N.; Jasty, N.; Shakhnovich, E. I. OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *J. Med. Chem.* **2016**, *59* (9), 4171–4188. https://doi.org/10.1021/acs.jmedchem.5b00886.

(12) Nishibata, Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47* (43), 8985–8990. https://doi.org/10.1016/S0040-4020(01)86503-0.

(13) Rotstein, S. H.; Murcko, M. A. GenStar: A Method for de Novo Drug Design. *J. Comput. Aided Mol. Des.* **1993**, *7* (1), 23–43.

(14) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–1534. https://doi.org/10.1126/science.274.5292.1531.

(15) Böhm, H.-J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J Computer-Aided Mol Des* **1992**, *6* (1), 61–78. https://doi.org/10.1007/BF00124387.

(16) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48* (7), 2457–2468. https://doi.org/10.1021/jm0492296.

(17) Dey, F.; Caflisch, A. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48* (3), 679–690. https://doi.org/10.1021/ci700424b.

(18) Nisius, B.; Rester, U. Fragment Shuffling: An Automated Workflow for Three-Dimensional Fragment-Based Ligand Design. *J. Chem. Inf. Model.* **2009**, *49* (5), 1211–1222. https://doi.org/10.1021/ci8004572.

(19) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60* (4), 1983–1995. https://doi.org/10.1021/acs.jcim.9b01120.

(20) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.* **2011**, *51* (5), 1083–1091. https://doi.org/10.1021/ci100350u.

(21) Chevillard, F.; Rimmer, H.; Betti, C.; Pardon, E.; Ballet, S.; van Hilten, N.; Steyaert, J.; Diederich, W. E.; Kolb, P. Binding-Site Compatible Fragment Growing Applied to the Design of B2-Adrenergic Receptor Ligands. *J. Med. Chem.* **2018**, *61* (3), 1118–1129. https://doi.org/10.1021/acs.jmedchem.7b01558.

(22) Zhang, J.; Yang, P. L.; Gray, N. S. Targeting Cancer with Small Molecule Kinase Inhibitors. *Nature Reviews Cancer* **2009**, *9* (1), 28–39. https://doi.org/10.1038/nrc2559.

(23) Cohen, P. Protein Kinases — the Major Drug Targets of the Twenty-First Century? *Nature Reviews Drug Discovery* **2002**, *1* (4), 309–315. https://doi.org/10.1038/nrd773.

(24) Noble, M. E. M.; Endicott, J. A.; Johnson, L. N. Protein Kinase Inhibitors: Insights into Drug Design from Structure. *Science* **2004**, *303* (5665), 1800–1805. https://doi.org/10.1126/science.1095920.

(25) Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacological Research* **2016**, *103*, 26–48. https://doi.org/10.1016/j.phrs.2015.10.021.

(26) Fogha, J.; Diharce, J.; Obled, A.; Aci-Sèche, S.; Bonnet, P. Computational Analysis of Crystallization Additives for the Identification of New Allosteric Sites. *ACS Omega* **2020**, *5* (5), 2114–2122. https://doi.org/10.1021/acsomega.9b02697.

(27) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.

(28) Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacological Research* **2016**, *103*, 26–48. https://doi.org/10.1016/j.phrs.2015.10.021.

(29) Zheng, J.; Trafny, E. A.; Knighton, D. R.; Xuong, N.; Taylor, S. S.; Ten Eyck, L. F.; Sowadski, J. M. 2.2 Å Refined Crystal Structure of the Catalytic Subunit of CAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor. *Acta Cryst D, Acta Cryst Sect D, Acta Crystallogr D, Acta Crystallogr Sect D, Acta Crystallogr D Biol Crystallogr, Acta Crystallogr Sect D Biol Crystallogr* **1993**, *49* (3), 362–365. https://doi.org/10.1107/S0907444993000423.

(30) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Research* **2016**, *44* (D1), D365–D371. https://doi.org/10.1093/nar/gkv1082.

(31) Sydow, D.; Schmiel, P.; Mortier, J.; Volkamer, A. KinFragLib: Exploring the Kinase Inhibitor Space Using Subpocket-Focused Fragmentation and Recombination. *J. Chem. Inf. Model.* **2020**. https://doi.org/10.1021/acs.jcim.0c00839.

(32) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5–6), 490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

(33) Halgren, T. A. Merck Molecular Force Field. II. MMFF94 van Der Waals and Electrostatic Parameters for Intermolecular Interactions. *Journal of Computational Chemistry* **1996**, *17* (5–6), 520–552. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<520::AID-JCC2>3.0.CO;2-W.

(34) Halgren, T. A. Merck Molecular Force Field. III. Molecular Geometries and Vibrational Frequencies for MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5–6), 553–586. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<553::AID-JCC3>3.0.CO;2-T.

(35) Halgren, T. A. Merck Molecular Force Field. V. Extension of MMFF94 Using Experimental Data, Additional Computational Data, and Empirical Rules. *Journal of Computational Chemistry* **1996**, *17* (5–6), 616–641. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<616::AID-JCC5>3.0.CO;2-X.

(36) Halgren, T. A.; Nachbar, R. B. Merck Molecular Force Field. IV. Conformational Energies and Geometries for MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5–6), 587–615. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<587::AID-JCC4>3.0.CO;2-Q.

(37) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17* (5-6), 490–519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

(38) Bournez, C.; Carles, F.; Peyrat, G.; Aci-Sèche, S.; Bourg, S.; Meyer, C.; Bonnet, P. Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB. *Molecules* **2020**, *25* (14), 3226. https://doi.org/10.3390/molecules25143226.

(39) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings1PII of Original Article: S0169-409X(96)00423-1. The Article Was Originally Published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **2001**, *46* (1), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(40) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.

(41) Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23* (4), 908. https://doi.org/10.3390/molecules23040908.

(42) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.

(43) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *Journal of Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.

(44) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach to Enable Alignment of Druglike Properties. *ACS Chem Neurosci* **2010**, *1* (6), 435–449. https://doi.org/10.1021/cn100008c.

(45) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(46) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.

(47) Ambinter http://www.ambinter.com/ (accessed 2021 -04 -07).

(48) Deininger, M. W. N.; Goldman, J. M.; Melo, J. V. The Molecular Biology of Chronic Myeloid Leukemia. *Blood* **2000**, *96* (10), 3343–3356. https://doi.org/10.1182/blood.V96.10.3343.

(49) Bolin, R. W.; Robinson, W. A.; Sutherland, J.; Hamman, R. F. Busulfan versus Hydroxyurea in Long-Term Therapy of Chronic Myelogenous Leukemia. *Cancer* **1982**, *50* (9), 1683–1686. https://doi.org/10.1002/1097-0142(19821101)50:9<1683::AID-CNCR2820500904>3.0.CO;2-X.

(50) An, X.; Tiwari, A. K.; Sun, Y.; Ding, P.-R.; Ashby, C. R.; Chen, Z.-S. BCR-ABL Tyrosine Kinase Inhibitors in the Treatment of Philadelphia Chromosome Positive Chronic Myeloid Leukemia: A Review. *Leukemia Research* **2010**, *34* (10), 1255–1268. https://doi.org/10.1016/j.leukres.2010.04.016.

(51) Cowan-Jacob, S. W.; Fendrich, G.; Floersheimer, A.; Furet, P.; Liebetanz, J.; Rummel, G.; Rheinberger, P.; Centeleghe, M.; Fabbro, D.; Manley, P. W. Structural Biology Contributions to the Discovery of Drugs to Treat Chronic Myelogenous Leukaemia. *Acta Cryst D, Acta Cryst Sect D, Acta Crystallogr D, Acta Crystallogr Sect D, Acta Crystallogr D Biol Crystallogr, Acta Crystallogr Sect D Biol Crystallogr* **2007**, *63* (1), 80–93. https://doi.org/10.1107/S0907444906047287.

(52) Savage, D. G.; Antman, K. H. Imatinib Mesylate—a New Oral Targeted Therapy. *New England Journal of Medicine* **2002**, *346* (9), 683–693.

(53) Bitencourt, R.; Zalcberg, I.; Louro, I. D. Imatinib Resistance: A Review of Alternative Inhibitors in Chronic Myeloid Leukemia. *Rev Bras Hematol Hemoter* **2011**, *33* (6), 470–475. https://doi.org/10.5581/1516-8484.20110124.

(54) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nature Chemistry* **2012**, *4* (2), 90–98. https://doi.org/10.1038/nchem.1243.

(55) Ikawa, S.; Fukui, M.; Ueyama, Y.; Tamaoki, N.; Yamamoto, T.; Toyoshima, K. B-Raf, a New Member of the Raf Family, Is Activated by DNA Rearrangement. *Mol Cell Biol* **1988**, *8* (6), 2651–2654.

(56) Davies, H.; Bignell, G. R.; Cox, C.; Stephens, P.; Edkins, S.; Clegg, S.; Teague, J.; Woffendin, H.; Garnett, M. J.; Bottomley, W.; Davis, N.; Dicks, E.; Ewing, R.; Floyd, Y.; Gray, K.; Hall, S.; Hawes, R.; Hughes, J.; Kosmidou, V.; Menzies, A.; Mould, C.; Parker, A.; Stevens, C.; Watt, S.; Hooper, S.; Wilson, R.; Jayatilake, H.; Gusterson, B. A.; Cooper, C.; Shipley, J.; Hargrave, D.; Pritchard-Jones, K.; Maitland, N.; Chenevix-Trench, G.; Riggins, G. J.; Bigner, D. D.; Palmieri, G.; Cossu, A.; Flanagan, A.; Nicholson, A.; Ho, J. W. C.; Leung, S. Y.; Yuen, S. T.; Weber, B. L.; Seigler, H. F.; Darrow, T. L.; Paterson, H.; Marais, R.; Marshall, C. J.; Wooster, R.; Stratton, M. R.; Futreal, P. A. Mutations of the BRAF Gene in Human Cancer. *Nature* **2002**, *417* (6892), 949–954. https://doi.org/10.1038/nature00766.

(57) Pollock, P. M.; Meltzer, P. S. A Genome-Based Strategy Uncovers Frequent BRAF Mutations in Melanoma. *Cancer Cell* **2002**, *2* (1), 5–7. https://doi.org/10.1016/S1535-6108(02)00089-2.

(58) Wilhelm, S. M.; Carter, C.; Tang, L.; Wilkie, D.; McNabola, A.; Rong, H.; Chen, C.; Zhang, X.; Vincent, P.; McHugh, M.; Cao, Y.; Shujath, J.; Gawlak, S.; Eveleigh, D.;

Rowley, B.; Liu, L.; Adnane, L.; Lynch, M.; Auclair, D.; Taylor, I.; Gedrich, R.; Voznesensky, A.; Riedl, B.; Post, L. E.; Bollag, G.; Trail, P. A. BAY 43-9006 Exhibits Broad Spectrum Oral Antitumor Activity and Targets the RAF/MEK/ERK Pathway and Receptor Tyrosine Kinases Involved in Tumor Progression and Angiogenesis. *Cancer Res* **2004**, *64* (19), 7099–7109. https://doi.org/10.1158/0008-5472.CAN-04-1443.

(59) Gil, M.; Yang, Y.; Lee, Y.; Choi, I.; Ha, H. Cloning and Expression of a CDNA Encoding a Novel Protein Serine/Threonine Kinase Predominantly Expressed in Hematopoietic Cells. *Gene* **1997**, *195* (2), 295–301. https://doi.org/10.1016/S0378-1119(97)00181-9.

(60) Gray, D.; Jubb, A. M.; Hogue, D.; Dowd, P.; Kljavin, N.; Yi, S.; Bai, W.; Frantz, G.; Zhang, Z.; Koeppen, H.; Sauvage, F. J. de; Davis, D. P. Maternal Embryonic Leucine Zipper Kinase/Murine Protein Serine-Threonine Kinase 38 Is a Promising Therapeutic Target for Multiple Cancers. *Cancer Res* **2005**, *65* (21), 9751–9761. https://doi.org/10.1158/0008-5472.CAN-04-4531.

(61) Settleman, J.; Sawyers, C. L.; Hunter, T. Challenges in Validating Candidate Therapeutic Targets in Cancer. *eLife* *7*. https://doi.org/10.7554/eLife.32402.

(62) Chung, S.; Suzuki, H.; Miyamoto, T.; Takamatsu, N.; Tatsuguchi, A.; Ueda, K.; Kijima, K.; Nakamura, Y.; Matsuo, Y. Development of an Orally-Administrative MELK-Targeting Inhibitor That Suppresses the Growth of Various Types of Human Cancer. *Oncotarget* **2012**, *3* (12), 1629–1640.

(63) Johnson, C. N.; Berdini, V.; Beke, L.; Bonnet, P.; Brehmer, D.; Coyle, J. E.; Day, P. J.; Frederickson, M.; Freyne, E. J. E.; Gilissen, R. A. H. J.; Hamlett, C. C. F.; Howard, S.; Meerpoel, L.; McMenamin, R.; Patel, S.; Rees, D. C.; Sharff, A.; Sommen, F.; Wu, T.; Linders, J. T. M. Fragment-Based Discovery of Type I Inhibitors of Maternal Embryonic Leucine Zipper Kinase. *ACS Med. Chem. Lett.* **2015**, *6* (1), 25–30. https://doi.org/10.1021/ml5001245.

(64) Johnson, C. N.; Adelinet, C.; Berdini, V.; Beke, L.; Bonnet, P.; Brehmer, D.; Calo, F.; Coyle, J. E.; Day, P. J.; Frederickson, M.; Freyne, E. J. E.; Gilissen, R. A. H. J.; Hamlett, C. C. F.; Howard, S.; Meerpoel, L.; Mevellec, L.; McMenamin, R.; Pasquier, E.; Patel, S.; Rees, D. C.; Linders, J. T. M. Structure-Based Design of Type II Inhibitors Applied to Maternal Embryonic Leucine Zipper Kinase. *ACS Med. Chem. Lett.* **2015**, *6* (1), 31–36. https://doi.org/10.1021/ml5001273.

(65) Eguida, M.; Rognan, D. A Computer Vision Approach to Align and Compare Protein Cavities: Application to Fragment-Based Drug Design. *J. Med. Chem.* **2020**, *63* (13), 7127–7142. https://doi.org/10.1021/acs.jmedchem.0c00422.

(66) Kalliokoski, T.; Olsson, T. S. G.; Vulpetti, A. Subpocket Analysis Method for Fragment-Based Drug Discovery. *J. Chem. Inf. Model.* **2013**, *53* (1), 131–141. https://doi.org/10.1021/ci300523r.

(67) Penner, P.; Martiny, V.; Gohier, A.; Gastreich, M.; Ducrot, P.; Brown, D.; Rarey, M. Shape-Based Descriptors for Efficient Structure-Based Fragment Growing. *J. Chem. Inf. Model.* **2020**, *60* (12), 6269–6281. https://doi.org/10.1021/acs.jcim.0c00920.

(68) Kolb, P.; Caflisch, A. Automatic and Efficient Decomposition of Two-Dimensional Structures of Small Molecules for Fragment-Based High-Throughput Docking. *J. Med. Chem.* **2006**, *49* (25), 7384–7392. https://doi.org/10.1021/jm060838i.

(69) Rarey, M.; Stahl, M. Similarity Searching in Large Combinatorial Chemistry Spaces. *J Comput Aided Mol Des* **2001**, *15* (6), 497–520. https://doi.org/10.1023/A:1011144622059.

(70) Stepniewska-Dziubinska, M. M.; Zielenkiewicz, P.; Siedlecki, P. Development and Evaluation of a Deep Learning Model for Protein–Ligand Binding Affinity Prediction. *Bioinformatics* **2018**, *34* (21), 3666–3674. https://doi.org/10.1093/bioinformatics/bty374.

(71) Finn, R. D.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Mistry, J.; Mitchell, A. L.; Potter, S. C.; Punta, M.; Qureshi, M.; Sangrador-Vegas, A.; Salazar, G. A.; Tate, J.; Bateman, A. The Pfam Protein Families Database: Towards a More Sustainable Future. *Nucleic Acids Res* **2016**, *44* (D1), D279–D285. https://doi.org/10.1093/nar/gkv1344.

(72) Cock, P. J. A.; Antao, T.; Chang, J. T.; Chapman, B. A.; Cox, C. J.; Dalke, A.; Friedberg, I.; Hamelryck, T.; Kauff, F.; Wilczynski, B.; Hoon, D.; L, M. J. Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* **2009**, *25* (11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163.

(73) *Molecular Operating Environment (MOE), 2016_0802*; Chemical Computing Group ULC: 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2021.

(74) 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 *Data Analysis, Machine Learning and Applications*; Preisach, C., Burkhardt, H., Schmidt-Thieme, L., Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319–326.

(75) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511–522. https://doi.org/10.1021/ci970429i.

(76) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507. https://doi.org/10.1002/cmdc.200800178.

(77) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47* (1), 47–58. https://doi.org/10.1021/ci600338x.

(78) Gally José-Manuel; Bourg Stéphane; Do Quoc-Tuan; Aci-Sèche Samia; Bonnet Pascal. VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening. *Molecular Informatics* **2017**, *36* (10), 1700023. https://doi.org/10.1002/minf.201700023.

(79) Le Guilloux, V.; Schmidtke, P.; Tuffery, P. Fpocket: An Open Source Platform for Ligand Pocket Detection. *BMC Bioinformatics* **2009**, *10*, 168. https://doi.org/10.1186/1471-2105-10-168.

(80) RDKit: Open-source cheminformatics https://www.rdkit.org/ (accessed 2020 -11 -06).

(81) Daylight Theory: SMARTS - A Language for Describing Molecular Patterns https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35 (accessed 2020 -11 -06).

(82) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Computational Biology* **2014**, *10* (4), e1003571. https://doi.org/10.1371/journal.pcbi.1003571.

(83) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55* (12), 2562–2574. https://doi.org/10.1021/acs.jcim.5b00654.

(84) *Chembl/FPSim2*; The ChEMBL Group, 2021.

(85) Tanimoto, T. T. *An Elementary Mathematical Theory of Classification and Prediction*; New York, International Business Machines Corporation, 1958.

(86) Tversky, A. Features of Similarity. *Psychological Review* **1977**, *84* (4), 327–352. https://doi.org/10.1037/0033-295X.84.4.327.

2.1.3 Recherche de macrocycles

À la suite de la rédaction de l'article présentant F2D, un deuxième article est focalisé sur la recherche de macrocycles avec ce logiciel. Cet article m'a permis de pouvoir présenter ce travail dans une communication orale au congrès « Journées Ouvertes en Biologie, Informatique et Mathématiques » (JOBIM) en 2021. Dans cet acte de congrès, la protéine cible utilisée est la kinase du lymphome anaplasique (*Anaplastic Lymphoma Kinase*, ALK) car elle est déjà inhibée par un PKI de la catégorie des macrocycles : le lorlatinib. Le fragment de départ utilisé est une aminopyridine à partir de laquelle 153 nouveaux macrocycles ont été obtenus. À cause de problèmes de fragmentation, le lorlatinib n'a pas pu être reconstruit, mais une molécule très similaire a été obtenue. J'ai aussi pu présenter la découverte de macrocycles par l'utilisation de F2D sous la forme d'une affiche dans le congrès « 55èmes Rencontres Internationales de Chimie Thérapeutique » (RICT) en 2019 à Nantes.

Dans le cas de la recherche de macrocycles, l'organisation du programme F2D doit être modifiée pour obtenir ce type de molécules. Pendant l'agrandissement des molécules, seuls les parcours de graphe aboutissant à des cycles sont pris en compte. De plus, pour prendre en compte la complexité plus élevée de la synthèse des macrocycles, nous avons modifié le seuil du SA score. Au lieu de garder les molécules avec un score de facilité de synthèse < 4, nous conservons des molécules avec un score SA < 5. Cette valeur de SA score adaptée aux macrocycles, a été validée en calculant les scores SA pour les macrocycles présents dans la BDD *Protein Kinase Inhibitor Database* (PKIDB)[34,35] (Figure 26). Ces scores ont une valeur moyenne de 4,61 et un écart-type de 0,69 d'où le fait que nous ayons augmenté la valeur seuil de score SA à 5 pour les macrocycles.

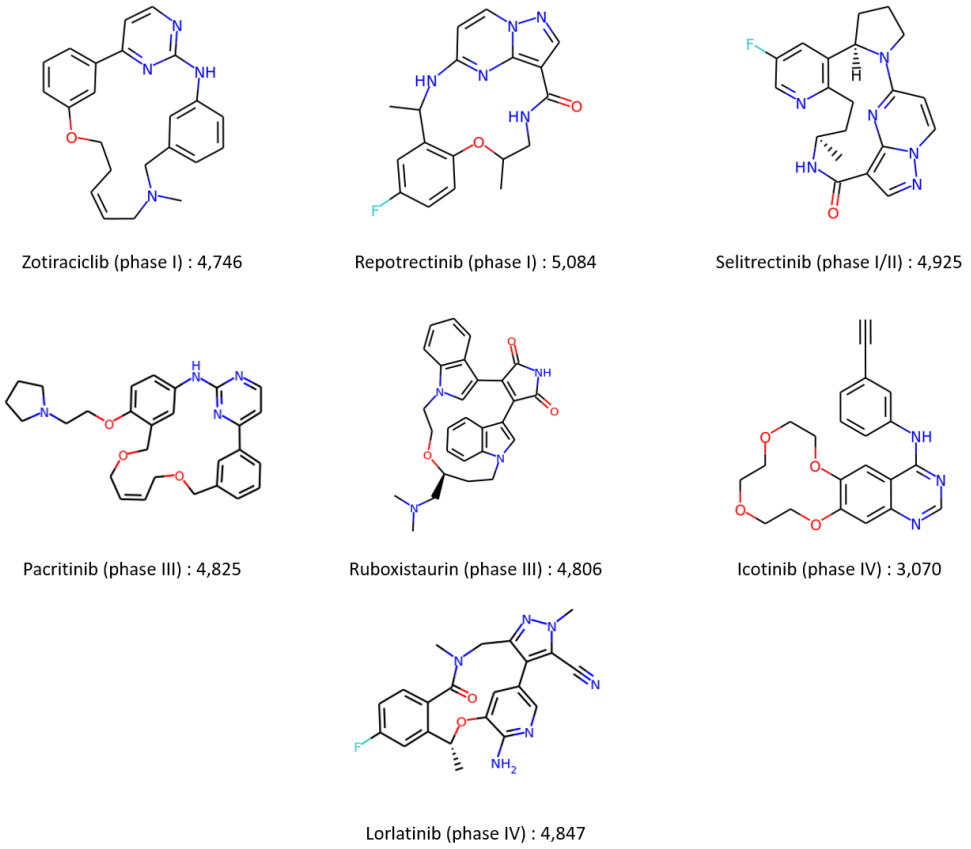


Figure 26 : Structures 2D et valeurs de score SA pour les macrocycles de PKIDB

Discovery of macrocyclic inhibitors of ALK using Frags2Drugs, a fragment-based drug design *in silico* tool

Gautier PEYRAT[1], Colin BOURNEZ[1], Pascal KREZEL[1], Stéphane BOURG[1], Samia ACI-SECHE[1], Pascal BONNET[1]

[1] Institut de Chimie Organique et Analytique (ICOA UMR 7311), Université d'Orléans - Pôle de chimie, rue de Chartres - BP 6759, 45067 Orléans Cedex 2, France

Corresponding Author: pascal.bonnet@univ-orleans.fr

Abstract *We developed an innovative in silico Fragment Based Drug Design (FBDD) approach, aiming to design new kinase inhibitors directly into their ATP binding site. This software program, called Frags2Drugs (F2D), relies on a fragment library obtained from X-ray crystal structures of protein-ligand complexes. This library is stored in a graph-oriented database containing required information to link fragments together. F2D builds every possible molecule fitting in the given cavity in a minute scale. Molecules are then filtered to keep those presenting the best potential affinity. Several specific molecular filters are applied, including an in-house Protein Kinase Inhibitors (PKI) like filter. We validated our method by reconstructing existing co-crystallized ligands. We describe here an application of F2D to find macrocyclic inhibitors of the protein kinase ALK.*

Keywords **Frags2drugs, Fragment-based drug design, Network graph, Macrocycle, Protein kinase**

Introduction

In the field of drug discovery, fragments are an interesting alternative to traditional compounds used for high-throughput screening: they are smaller (usually ~ 350 Da) and can cover a larger chemical space with less molecules[1]. Once identified during a screening campaign, fragment hits have to be optimized using medicinal chemistry expertise, biophysical and structural biology approaches such as X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy. Computational chemistry and structural bioinformatics may also play a substantial role in this process. Multiple software applications have been recently developed in the field of in silico FBDD[2].

F2D is a program that uses fragments from co-crystallized ligands (or from docking experiments) and combines between them to design new molecules. F2D requires many fragments, coming from diverse protein-ligand complexes, properly separated to create bounds while keeping their initial position.

We first validated F2D on protein kinases. From a structural point of view, protein kinases share a common catalytic domain composed of two distinct lobes: an N-terminal lobe (N-lobe) and a C-terminal lobe (C-lobe) connected by a hinge region. The cleft between these two lobes is the binding site of their cofactor, the ATP, which forms hydrogen bonds with the hinge region stabilizing its position in the catalytic site. Most of kinase inhibitors in development or already on the market target the ATP binding site[3] but other druggable pockets are known, as the allosteric site and the back pocket[4]. Because of the large amount of available protein kinase crystal structures – 5,004 structures of human kinases domains are deposited in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Databank[5] (PDB, https://www.rcsb.org/, December 2020) – computational chemists can more easily validate their new tool on this protein family.

Molecules are called “macrocycles” when they have a cycle composed of at least 12 atoms in a ring architecture[6]. Macrocycles represent an interesting class of compounds able to inhibit challenging targets[6,7]. Macrocycles often provide a better affinity and selectivity than other compounds on protein

kinases [6].

As a proof of concept for the discovery of macrocycles, we focused on Anaplastic Lymphoma Kinase (ALK) target and its inhibitor lorlatinib. ALK is a tyrosine kinase belonging to the insulin receptor subfamily implicated in the nervous system's development and function[8]. Chromosomal translocations involving ALK result in several fusion proteins identified in cancers [9–11] as well as the amplification and mutations of the full ALK gene[12–15].

Methods

F2D relies on 3D structural data of protein kinase complexes to generate new inhibitors. All ligands are extracted and fragmented from superimposed PDB structures. Several algorithms already implemented in KNIME[16] were used to fragment the ligands : RECAP[17], BRICS[18] and Scaffold Tree[19]. The use of the three methods provided more fragments, allowing a more exhaustive decomposition of ligands. We store each fragment in a graph database along with its position and its ability to bind neighboring fragments. Indeed, a network is particularly suited to store information about connectable entities and thus, perfectly fulfils the aim of F2D.

By exploring a graph of connected fragments, those fragments can be linked together to form a possible PKI. F2D needs three inputs to generate kinase inhibitors: a fragment, referred as the seed, the PDB structure of the targeted kinase and the F2D graph database (Fig 1). The graph database of F2D is composed of two types of nodes (fragment and protein) linked by three types of edges: inclusion, exclusion and compatibility. Nodes from the graph database are associated to corresponding structures of fragments with the use of Pandas dataframe (1.0.5) and RDKit library[20] (2020.03.3). The RDKit library is also used to link fragments together in order to generate molecules.

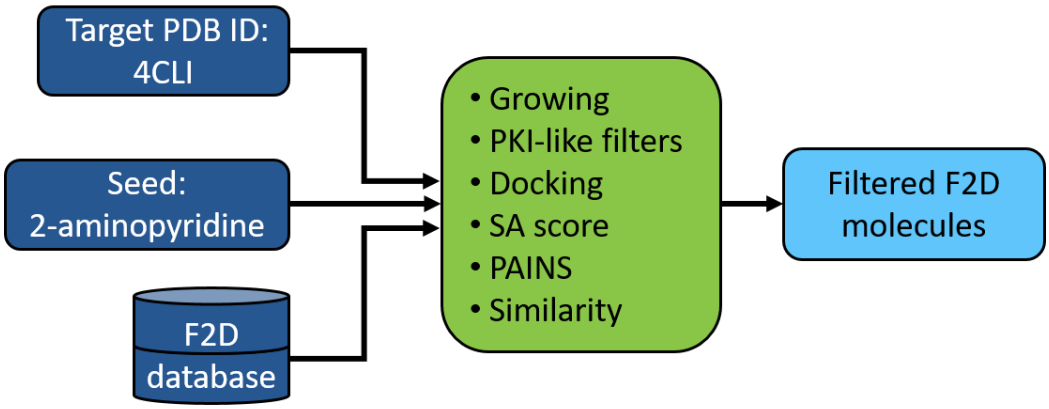


Fig 1. F2D workflow applied to discover macrocycles on ALK. Inputs are colored in dark blue, F2D functions are colored in green and outputs are colored in light blue.

The inclusion or exclusion relations characterized the potential link between fragments, while compatibility edges refer to the interaction between proteins and fragments. An inclusion relation means that the two considered fragments satisfy all the following conditions to bind together: it exists a couple of atoms, one from each fragment, having incomplete valences and separated by an acceptable distance. Moreover, their out of plane, dihedral and bond angles with their direct neighbors fulfill predefined allowed intervals. The definition of allowed intervals is based on force field values extracted from MMFF94[21] taking into account the type and the hybridization of both atoms involved. If the two fragments are too close, overlapped or if there is a non-realistic angle value between them, an exclusion relation is thus established so they cannot be bound. Furthermore, they cannot belong to a same built molecule, even if they have inclusion relations with a common intermediate fragment. We used Python (3.6.8), neo4j (3.5.11) and mainly the Python library networkx (2.2) to create the fragment network. Calculation of compatibility relationships is performed in two steps. Firstly, for each protein, its binding site cavity and the surrounding atoms are identified using FPocket[22], an open-source cavity detection tool based on Voronoi tessellation. Secondly, a fragment is defined as compatible with a cavity if the

distances between fragment and residue atoms fit in an acceptable range. Thus, we avoid both clashes, one between two fragments and one between fragment and residue from the active site of the target.

After the growing step, several filters are applied to remove undesirable molecules. The first one is a PKI-like filter recently developed in team [23,24]. During this filtering, substructure recognition relies on SMARTS patterns[25] and are applied by substructure match from RDKit. The PKI-like filter removes molecules containing undesired substructures after the combination of several fragments. These substructures are, as example, ester, aldehyde, acyclic N-N, N-O and O-O.

Then, we check that the molecules bind efficiently the target by molecular docking carried on with rDock[26] (2013.1). Before proceeding to docking, molecules are standardized, chirality is checked, and a conformational search is finally performed with MOE (Chemical Computing Group, 2019_0101) suite[27]. The root mean square deviation (RMSD) between the conformation obtained by F2D and the docking poses is calculated with the help of RDKit library. All macrocycles having a RMSD higher than 2.5 Å are removed. The threshold of 2.5 Å was chosen after validating the docking protocol on all kinase-ligand PDB structures.

To evaluate the synthetic feasibility of newly built molecules, a Synthetic Accessibility (SA) score[28] is calculated. Pan-Assay INterference compoundS (PAINS)[29] are tagged using RDKit.

As F2D needs to find active and innovative compounds, we implemented a similarity search to compounds present in ChEMBL27 database (accessed June 2020)[30], compounds available in purchasable compound database such as ZINC 15 (accessed June 2020)[31] or Ambinter (Accessed February 2020)[32]. We extended this search to protein kinase inhibitors already marketed or in clinical trials by searching PKIDB 2020-12-09[23]. Similarity searches performed on all molecular databases uses FPSim2[33]. The similarity search is based on both Tanimoto[34] and Tversky measures[35] with a Tanimoto coefficient of 0.7 and for Tversky coefficients of α=1 and β=0.5.

Results and discussion

Several inhibitors have already been described against ALK, such as crizotinib[36] and lorlatinib, a macrocycle targeting ALK wild type and L1196M mutation[37]. We selected the X-ray structure (PDB ID 4CLI) as the target for F2D. To start the growing step, we selected the 3 carbon atoms on positions 3, 4 and 5 of the 2-aminopyridine (Fig 2.A). The 3D position of the 2-aminopyridine seed is obtained by deleting all other atoms, except those of the seed, from the co-crystallized molecule.

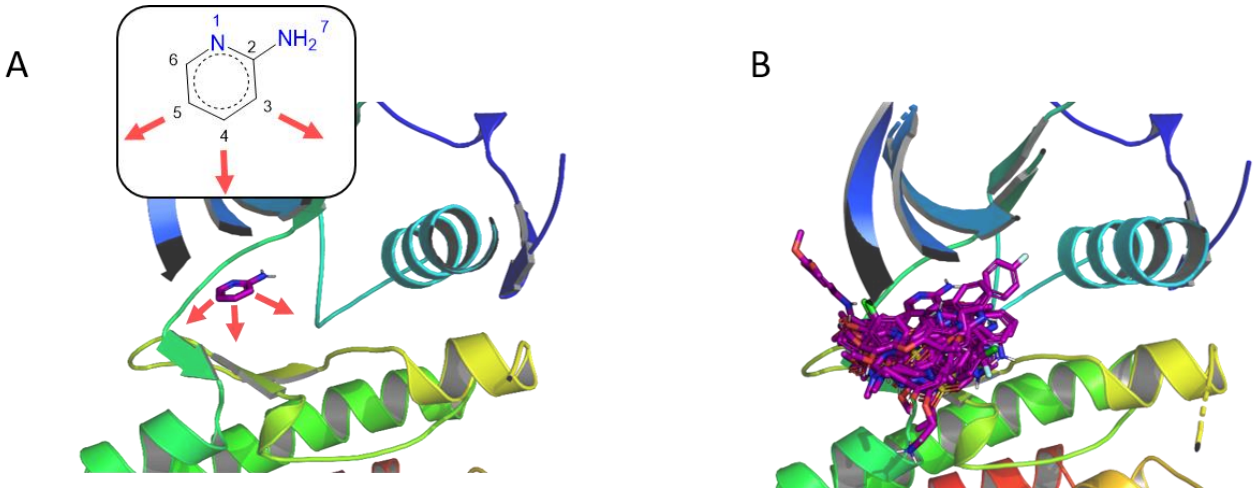


Fig 2. Example of F2D applied to the discovery of ALK macrocyclic inhibitors. (A) Initial position of the 2-aminopyridine seed in the active site (PDB ID: 4CLI, chain A), the red arrows indicate the vector positions for growing. (B) New ALK macrocycles generated by F2D.

Then, we selected molecules with at least 11 atoms in their largest ring corresponding to a macrocycle. We obtained 592 molecules having macrocycles with lengths from 11 to 15 atoms. 35 molecules were discarded as PAINS substructure. Because synthesis of macrocycles is more complex than linear molecules, we used an SA score threshold of 5, to retain the most synthetic feasible macrocycles. Thus, we obtained 174 molecules having an SA score < 5. After performing docking, 153 macrocycles showed a minimal RMSD value < 2.5 Å when comparing to the molecules generated from F2D. No similar compound was found in the ChEMBL, Ambinter, ZINC or PKIDB databases, leading to 153 new ALK macrocycles identified by F2D (Fig 2.B).

To get the best molecules among those selected, we estimated their Quantitative Estimation of Druglikeness (QED) scores[38]. This score ranging from 0 to 1 indicates which molecules are the most likely to become lead compounds, 1 being the best value. Nine macrocycles have a QED score >= 0.7 indicating that they could become interesting leads after few optimization steps (Fig 3). By visual inspection, we observed that one compound seems to be very close to lorlatinib, even if its Tanimoto coefficient is only of 0.41 (Fig 4).

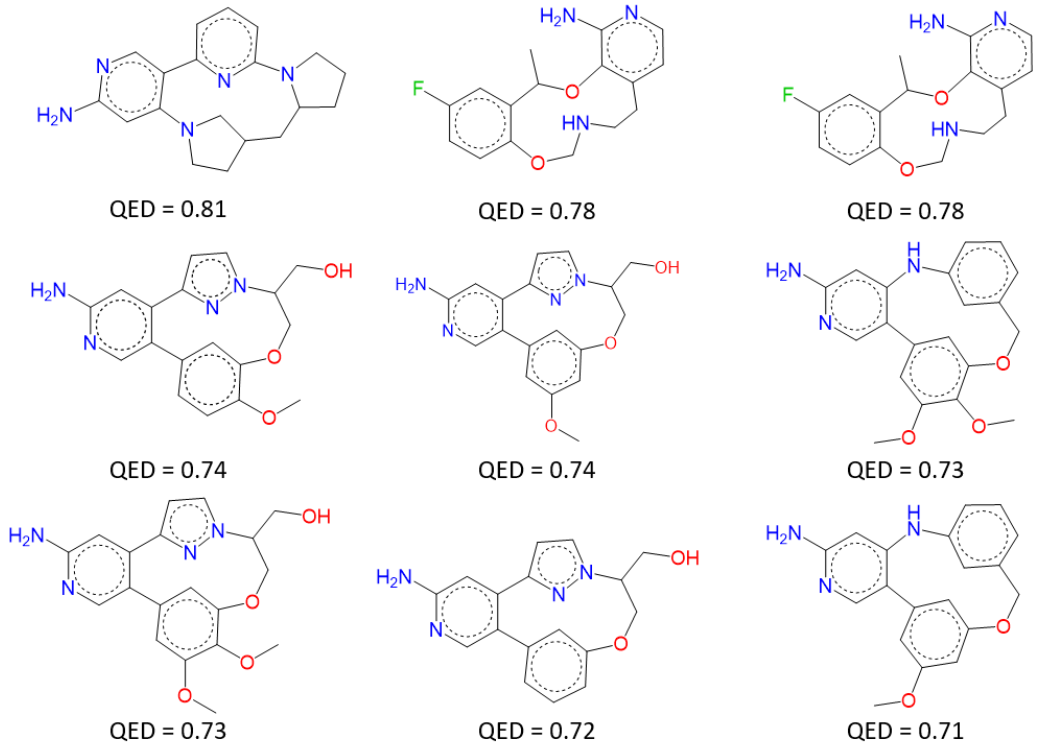


Fig 3. Structures of the 9 macromolecules having the best QED scores generated by F2D from a 2-aminopyridine seed bound to ALK (PDB ID 4CLI).

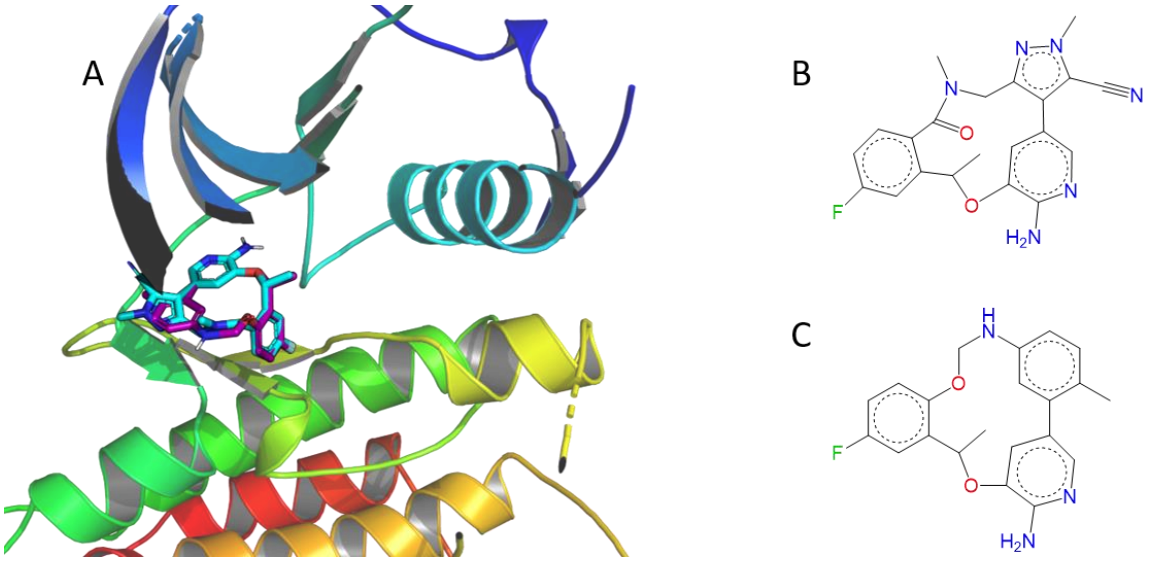


Fig 4. Comparison between lorlatinib and the most similar inhibitor found by Frags2Drugs. (A) 3D structures of both inhibitors bound in the active site of ALK (PDB ID 4CLI), lorlatinib in cyan and F2D inhibitor in purple. (B) 2D structure of lorlatinib. (C) 2D structure of the most similar F2D inhibitor compared to lorlatinib (Tanimoto coefficient of 0.41).

CONCLUSIONS AND PERSPECTIVES

In this article, we used Frags2Drugs, an *in silico* program based on a 3D fragment network to find efficient protein kinase inhibitors. We demonstrated its efficiency to discover macrocyclic inhibitors of the protein kinase ALK. The main advantage of F2D is to work directly with 3D experimental structures without the need of energy minimization during fragment growing. Furthermore, no systematic cavity exploration is needed, thus, results can be obtained in a scale of hours.

We have implemented a filtering procedure to reduce the number of molecules given by F2D, based on structural and physicochemical properties of compounds. However, the potential bioactivity of molecules is not currently considered. To guide the choice of the most interesting compounds, we aim at implementing a predicting method of protein-ligand binding affinities by using recent deep learning-based methods.

In this example, we discovered several interesting macrocycles that will now require synthesis and biological validation. We aim at applying F2D on new protein kinase targets.

Acknowledgements:

Authors gratefully acknowledge financial support from Région Centre Val de Loire and FEDER under project FEDER-FSE 2014-2020-EX003677, and from LabEx SYNORG (ANR-11-LABX-0029) and IRON (ANR-11-LABX-0018-01).

REFERENCES

1. Hall RJ, Mortenson PN, Murray CW. Efficient exploration of chemical space by fragment-based screening. Progress in Biophysics and Molecular Biology 2014;116(2):82–91.
2. Grove LE, Vajda S, Kozakov D. Computational Methods to Support Fragment-based Drug Discovery [Internet]. In: Fragment-based Drug Discovery Lessons and Outlook. John Wiley & Sons, Ltd; 2016 [cited 2019 Jan 28]. page 197–222.Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9783527683604.ch09
3. Noble MEM, Endicott JA, Johnson LN. Protein Kinase Inhibitors: Insights into Drug Design from Structure. Science 2004;303(5665):1800–5.

4. Roskoski R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. Pharmacological Research 2016;103:26–48.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235–42.
6. Driggers EM, Hale SP, Lee J, Terrett NK. The exploration of macrocycles for drug discovery — an underexploited structural class. Nature Reviews Drug Discovery 2008;7(7):608–24.
7. Marsault E, Peterson ML. Macrocycles Are Great Cycles: Applications, Opportunities, and Challenges of Synthetic Macrocycles in Drug Discovery. J Med Chem 2011;54(7):1961–2004.
8. Morris SW, Naeve C, Mathew P, James PL, Kirstein MN, Cui X, et al. ALK , the chromosome 2 gene locus altered by the t(2;5) in non-Hodgkin's lymphoma, encodes a novel neural receptor tyrosine kinase that is highly related to leukocyte tyrosine kinase (LTK). Oncogene 1997;14(18):2175–88.
9. Morris SW, Kirstein MN, Valentine MB, Dittmer KG, Shapiro DN, Saltman DL, et al. Fusion of a kinase gene, ALK, to a nucleolar protein gene, NPM, in non-Hodgkin's lymphoma. Science 1994;263(5151):1281–4.
10. Mano H. Non-solid oncogenes in solid tumors: EML4–ALK fusion genes in lung cancer. Cancer Science 2008;99(12):2349–55.
11. Griffin CA, Hawkins AL, Dvorak C, Henkle C, Ellingham T, Perlman EJ. Recurrent Involvement of 2p23 in Inflammatory Myofibroblastic Tumors. Cancer Res 1999;59(12):2776–80.
12. Tuma RS. ALK Gene Amplified in Most Inflammatory Breast Cancers. JNCI: Journal of the National Cancer Institute 2012;104(2):87–8.
13. Ren H, Tan Z-P, Zhu X, Crosby K, Haack H, Ren J-M, et al. Identification of Anaplastic Lymphoma Kinase as a Potential Therapeutic Target in Ovarian Cancer. Cancer Res 2012;72(13):3312–23.
14. Azarova AM, Gautam G, George RE. Emerging importance of ALK in neuroblastoma. Seminars in Cancer Biology 2011;21(4):267–75.
15. Bergethon K, Shaw AT, Ignatius Ou S-H, Katayama R, Lovly CM, McDonald NT, et al. ROS1 Rearrangements Define a Unique Molecular Class of Lung Cancers. J Clin Oncol 2012;30(8):863–70.
16. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, et al. KNIME: The Konstanz Information Miner. In: Preisach C, Burkhardt H, Schmidt-Thieme L, Decker R, editors. Data Analysis, Machine Learning and Applications. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. page 319–26.
17. Lewell XQ, Judd DB, Watson SP, Hann MM. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. J Chem Inf Comput Sci 1998;38(3):511–22.
18. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. ChemMedChem 2008;3(10):1503–7.
19. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. J Chem Inf Model 2007;47(1):47–58.
20. RDKit: Open-source cheminformatics [Internet]. [cited 2020 Nov 6];Available from: https://www.rdkit.org/
21. Halgren TA. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry 1996;17(5–6):490–519.

22. Le Guilloux V, Schmidtke P, Tuffery P. Fpocket: An open source platform for ligand pocket detection. BMC Bioinformatics 2009;10:168.
23. Carles F, Bourg S, Meyer C, Bonnet P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. Molecules 2018;23(4):908.
24. Bournez C, Carles F, Peyrat G, Aci-Sèche S, Bourg S, Meyer C, et al. Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB. Molecules 2020;25(14):3226.
25. Daylight Theory: SMARTS - A Language for Describing Molecular Patterns [Internet]. [cited 2020 Nov 6];Available from: https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35
26. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, et al. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. PLOS Computational Biology 2014;10(4):e1003571.
27. Molecular Operating Environment (MOE), 2019_0101 [Internet]. 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7: Chemical Computing Group ULC; 2021 [cited 2021 Apr 6]. Available from: https://www.chemcomp.com/Research-Citing_MOE.htm
28. Ertl P, Schuffenhauer A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. Journal of Cheminformatics 2009;1(1):8.
29. Baell JB, Holloway GA. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. J Med Chem 2010;53(7):2719–40.
30. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, et al. The ChEMBL database in 2017. Nucleic Acids Res 2017;45(Database issue):D945–54.
31. Sterling T, Irwin JJ. ZINC 15 – Ligand Discovery for Everyone. J Chem Inf Model 2015;55(11):2324–37.
32. Ambinter [Internet]. [cited 2021 Apr 7];Available from: http://www.ambinter.com/
33. chembl/FPSim2 [Internet]. The ChEMBL Group; 2021 [cited 2021 Mar 11]. Available from: https://github.com/chembl/FPSim2
34. Tanimoto TT. An elementary mathematical theory of classification and prediction. New York, International Business Machines Corporation; 1958.
35. Tversky A. Features of similarity. Psychological Review 1977;84(4):327–52.
36. Cui JJ, Tran-Dubé M, Shen H, Nambu M, Kung P-P, Pairish M, et al. Structure Based Drug Design of Crizotinib (PF-02341066), a Potent and Selective Dual Inhibitor of Mesenchymal–Epithelial Transition Factor (c-MET) Kinase and Anaplastic Lymphoma Kinase (ALK). J Med Chem 2011;54(18):6342–63.
37. Johnson TW, Richardson PF, Bailey S, Brooun A, Burke BJ, Collins MR, et al. Discovery of (10R)-7-Amino-12-fluoro-2,10,16-trimethyl-15-oxo-10,15,16,17-tetrahydro-2H-8,4-(metheno)pyrazolo[4,3-h][2,5,11]-benzoxadiazacyclotetradecine-3-carbonitrile (PF-06463922), a Macrocyclic Inhibitor of Anaplastic Lymphoma Kinase (ALK) and c-ros Oncogene 1 (ROS1) with Preclinical Brain Exposure and Broad-Spectrum Potency against ALK-Resistant Mutations. J Med Chem 2014;57(11):4720–44.
38. Bickerton GR, Paolini GV, Besnard J, Muresan S, Hopkins AL. Quantifying the chemical beauty of drugs. Nature Chemistry 2012;4(2):90–8.

2.2 Analyse de F2D et poursuite du développement

2.2.1 Analyse par reconstruction des ligands d'origine

J'ai mis en place une procédure d'évaluation de la base de données (BDD) orientée graphes permettant de vérifier qu'à partir de celle-ci et avec les paramètres standards, le programme Frags2Drugs (F2D) est capable de reconstruire les inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI) qui ont été fragmentés pour la créer. Cette procédure d'évaluation consiste d'abord à sélectionner, pour chaque ligand d'un jeu de données de validation, les fragments issus de sa fragmentation, puis à vérifier si F2D est capable de reconstruire le ligand d'origine à partir de ces fragments. En cas d'échec, il faut comprendre l'origine du problème qui peut venir du ligand à fragmenter, de la BDD orientée graphes ou de la fragmentation appliquée.

2.2.1.1 Problèmes issus du ligand à fragmenter

En appliquant la procédure d'évaluation de la capacité de F2D à reconstruire les ligands ayant été fragmentés, nous avons observé des problèmes venant de la structure 3D des ligands avant leur fragmentation. Les 33 identifiants PDB des complexes contenant ces ligands sont : 1Y91, 2QU6, 3BV2, 3BYS, 3BYU, 3FI2, 3IKA, 3JY9, 3K54, 3NAY, 3OCT, 3Q4T, 3RVG, 3TT0, 3UG2, 3VS2, 3VS3, 4AGU, 4BC6, 4BCH, 4BCM, 4D1S, 4JVG, 4O6L, 4QTC, 5D7A, 5FDX, 5I3O, 5IKW, 5IUG, 5JQ8, 5X2C, 6H0O. La Figure 27 montre quelques exemples de ces structures 3D. Dans la structure 3D de l'identifiant PDB 5FDX, le benzène n'est pas plan, 3VS2 présente une sous structure dont les valeurs des angles et distances entre atomes ne sont pas physiquement acceptables.

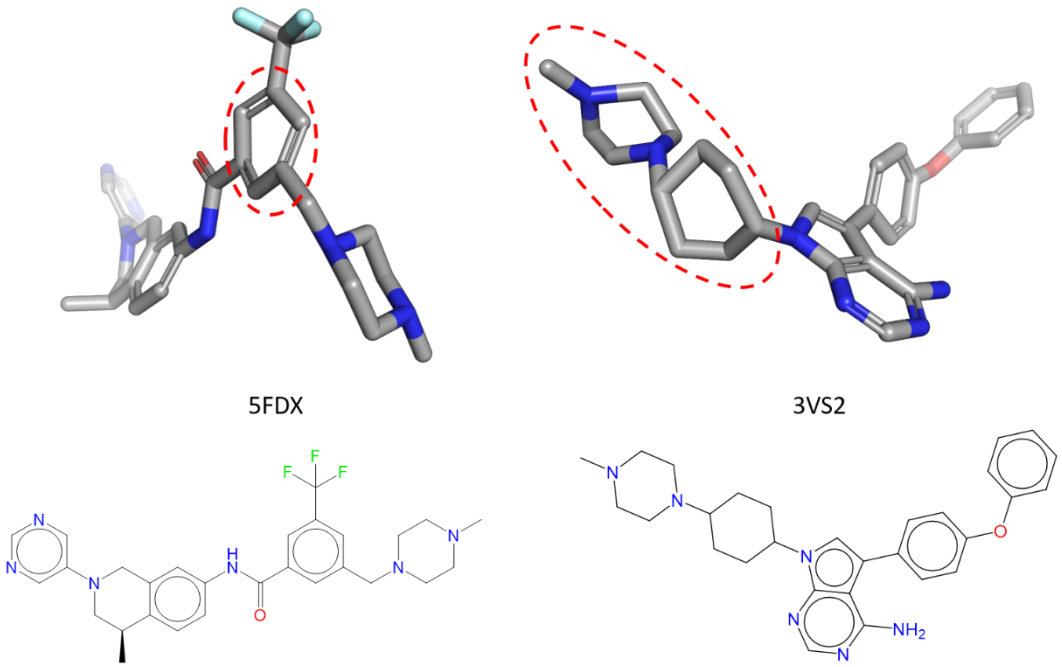


Figure 27 : Exemples de structures 2D et 3D de ligands présentant des problèmes de géométrie 3D.

La structure 2D du ligand enregistré dans la *Research Collaboratory for Structural Bioinformatics* (RCSB) *Protein Data Bank* (PDB) ne correspond pas toujours à celle initialement indiquée par les auteurs de la publication de référence. C'est le cas pour le complexe PDB 3NAY dont le ligand contient une fonction alcyne entourée en rouge sur la Figure 28.

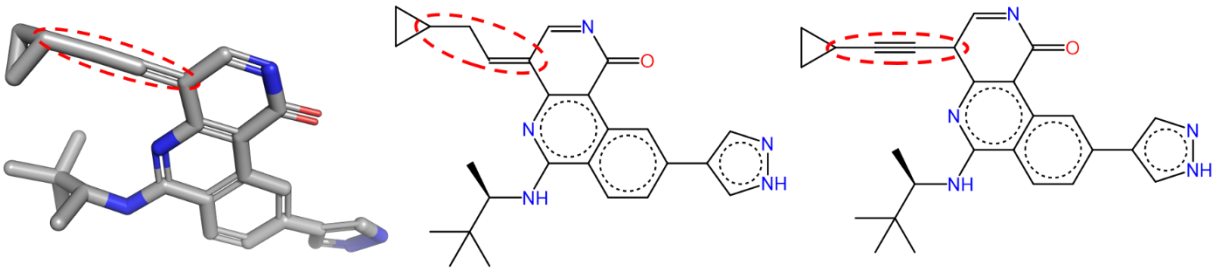


Figure 28 : Différentes structures 3D et 2D du ligand issu du complexe PDB 3NAY

Dans les représentations 3D et 2D provenant de la RCSB PDB, cette fonction est représentée de la façon suivante : deux liaisons simples, puis une liaison double, à la place de l'enchaînement : simple, triple, simple. La géométrie linéaire des quatre carbones montre que les structures enregistrées sur la RCSB PDB pour ce complexe ne sont pas correctes.

En raison du nombre de structures 3D de ligands utilisées dans F2D, il n'est pas possible de toutes les vérifier manuellement. Quand un ligand présente une erreur dans sa structure 3D, celui-ci est retiré du jeu de données de validation. Cependant, nous conservons quand même les « bons » fragments issus de ces ligands dans la BDD de F2D pour ne pas trop réduire le nombre total de fragments.

2.2.1.2 Problèmes issus de la BDD orientée graphes

Comme expliqué en partie 2.1.2, la BDD orientée graphes de F2D est constituée de nœuds (fragments ou protéines) et d'arêtes (relations d'inclusion ou d'exclusion). Les paramètres utilisés lors de la création de la BDD influent sur la possibilité d'associer des fragments voisins au sein d'une même molécule. En cas d'impossibilité de reconstruire un ligand, les paramètres peuvent être rendus plus permissifs pour autoriser la reconstruction. L'évaluation ne s'effectue qu'à partir des fragments provenant du ligand à valider. L'utilisation de F2D dans les projets de recherche de nouveaux PKI prend en compte beaucoup plus de fragments. Pour ne pas obtenir de molécules comportant des valeurs d'angles trop extrêmes, il est nécessaire de rester assez restrictif dans les valeurs d'acceptation des paramètres de création de la BDD.

Pour illustrer l'ajustement du paramètre de distance interatomique, prenons l'exemple du ligand SR-3562 dont la structure 3D est issue du complexe PDB 3KVX (Figure 29). La fragmentation de ce ligand donne des nœuds liés dans le graphe par des relations d'inclusion et une relation d'exclusion entre deux nœuds non liés dans la molécule d'origine. Cette relation d'exclusion entre deux nœuds empêche de regrouper ces fragments au sein d'une même molécule et donc de valider le ligand SR-3562. Le paramètre de distance interatomique est un seuil permettant d'éviter des encombrements stériques entre fragments. Au cours du développement de F2D, mes prédécesseurs avaient défini la valeur de ce paramètre à 2,3 Å. En diminuant la valeur minimale de ce seuil à 2,25 Å nous avons supprimé la relation d'exclusion entre les fragments 2 et 5 de la Figure 29 et permis la validation de ce ligand.

Il faut bien noter qu'en modifiant ce seuil, l'ensemble de la BDD est impacté. Ainsi, d'autres ligands dans le même cas ont aussi pu être validés, mais les possibilités d'associations de fragments provenant de molécules différentes ont aussi augmenté. C'est pourquoi nous n'avons pas diminué le seuil de plus de 0,05 Å.

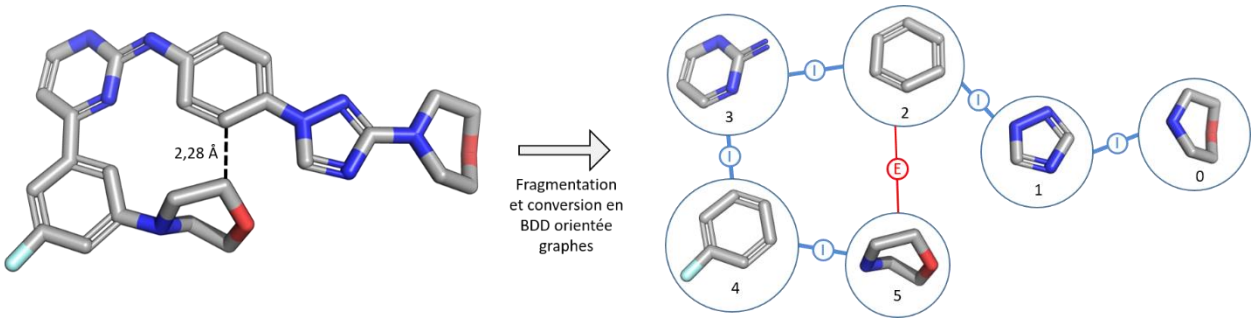


Figure 29 : Structure 3D du ligand SR-3562 (ID PDB : 3KVX) et graphe obtenu après sa fragmentation. I : relation d'inclusion, en bleu ; E : relation d'exclusion, en rouge. Le nœud de la protéine cible et les relations de compatibilité entre les fragments et la protéine ne sont pas montrés. Tous les fragments sont compatibles avec la protéine. D'après la thèse du Dr Colin Bournez[2].

Lors de la création de la BDD de F2D, nous définissons aussi les valeurs d'acceptation pour les angles de liaison, dièdres et hors plan. Il s'agit d'autres paramètres influant sur les relations entre les nœuds de la BDD. Ces paramètres permettent de prendre en compte les déformations des molécules lorsque celles-ci sont contraintes par les interactions qu'elles forment avec les atomes du site actif de leurs cibles.

2.2.1.3 Évaluation des différentes versions de la BDD de F2D

Au cours de mon analyse de F2D, j'ai pu évaluer trois versions de la BDD appelées DB2016-07, DB2019-07 et DB2020-02 selon l'année et le mois où elles ont été créées (Tableau 13).

Tableau 13 : Évaluation des différentes BDD de F2D

Nom de BDD	Nombre de fragments	Fragmentation	Nombre de ligands fragmentés	Nombre de ligands à valider	Nombre de ligands validés	Pourcentage de validation
DB2016-07	72 480	KNIME (RECAP[267], BRICS[268] et Scaffold Tree[269])	6 205	3 218	2 836	88,13%
DB2019-07	21 835	Inter-cycles	6 205	3 218	3 218	100,00%
DB2020-02	43 300	Inter-cycles et cycles fusionnés	9 113	6 874	6 799	98,91%

Les BDD correspondent aussi à l'application de différentes méthodes de fragmentation qui sont illustrées à partir de la molécule vemurafenib en Figure 30.

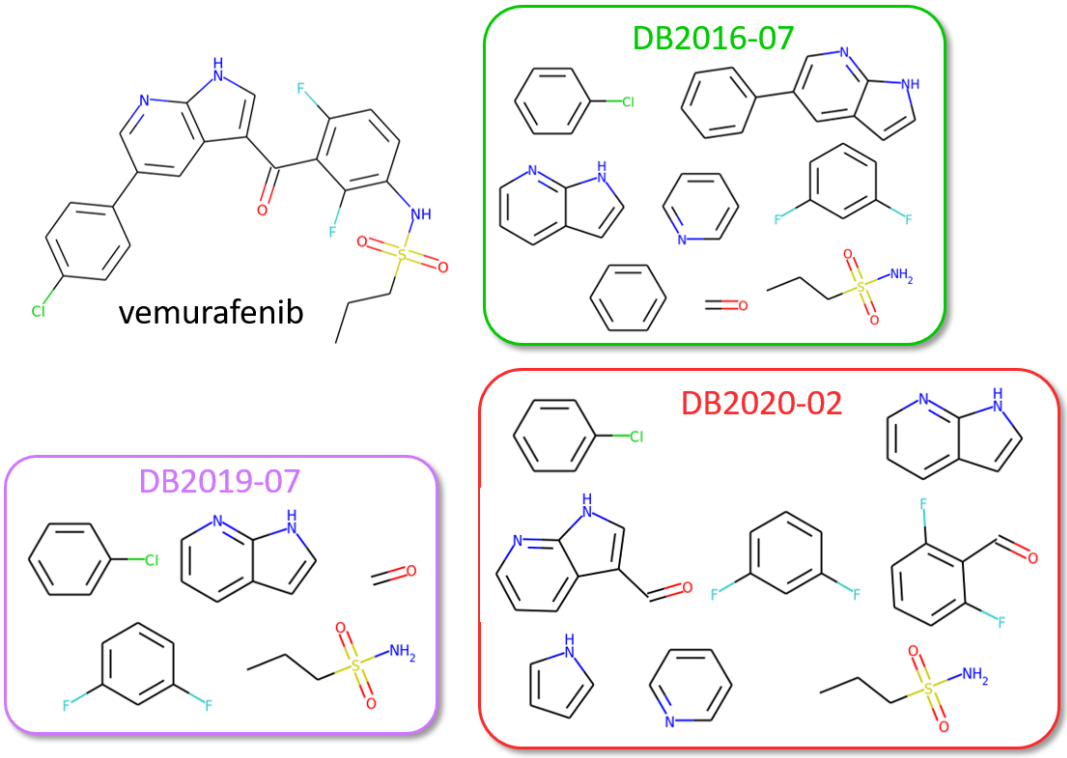


Figure 30 : Évolution des différentes méthodes de fragmentation utilisées par F2D pour générer les différentes BDD. La molécule vemurafenib est utilisée pour l'illustration des différentes fragmentations. Les molécules sont représentées en 2D, mais sont réellement en 3D dans les BDD de F2D.

La première BDD, DB2016-07, concerne les ligands collectés en juillet 2016, fragmentés par différents outils implémentés dans Konstanz Information Miner (KNIME). À partir de DB2016-07, F2D est capable de valider 88,13% du jeu de données de validation (Tableau 13). Il s'agit de la BDD présentée et utilisée dans les 2 articles des parties précédentes, ainsi que par le site web de F2D. La version du code source de F2D utilisée en association avec DB2016-07 dans ces articles et sur le site web date de février 2020.

La deuxième BDD, DB2019-07, utilise les mêmes ligands (collectés en juillet 2016) mais repose sur une fragmentation différente qui est détaillée dans la thèse de Colin Bournez[2]. Cette fragmentation ne repose plus sur KNIME, mais sur le langage de programmation Python et la librairie *Rational Discovery Kit* (RDKit). Elle consiste en une fragmentation nommée « inter-cycles » fragmentant les ligands autours des cycles, en évitant de laisser des atomes seuls et en ne fragmentant que des liaisons simples.

En utilisant DB2019-07, F2D est capable de reconstruire 100% des PKI du jeu de données de validation. Au cours du développement de cette BDD, à chaque échec de reconstruction d'un ligand, nous avons adapté le programme de création de la BDD pour qu'elle autorise la connexion de tous les fragments du ligand quelles que soient les valeurs des paramètres d'angles et de distances entre les atomes des fragments. Cela fonctionne en ne prenant que les fragments issus du ligand à valider. Cependant, pour obtenir un tel pourcentage de validation, la BDD est constituée de trop de relations d'inclusion et pas assez de relations d'exclusion. Lors du calcul des relations entre fragments provenant de ligands différents, pour la création de nouveaux inhibiteurs, cette version de la BDD donnait la possibilité de créer de trop nombreuses molécules. Certaines exceptions étaient devenues une norme d'acceptation. Cela amenait une impossibilité d'utiliser le programme pour obtenir de nouveaux

inhibiteurs de manière efficiente. Pour résoudre ce problème, nous avons implémenté une troisième version de la BDD, DB2020-02.

DB2020-02, correspond à une amélioration de la méthode de fragmentation « inter-cycles » associée à une mise à jour de la BDD de ligands à fragmenter. En ajoutant les nouvelles protéines kinases de la RCSB PDB, le nombre de ligands est passé de 6 205 à 9 113 en août 2019. La méthode de fragmentation a été améliorée pour pouvoir découper les molécules au niveau des liaisons doubles et triples et pouvoir également fragmenter les cycles fusionnés. Cette dernière méthode de fragmentation est nommée « inter-cycles et cycles fusionnés ». Il s'agit de cette BDD qui a été utilisée pour réaliser les différentes recherches d'inhibiteurs montrées dans le quatrième chapitre de cette thèse.

2.2.2 Ajout de branches Git au code source de F2D

Comme expliqué dans le premier article de ce chapitre, l'utilisation de F2D est possible par tout chercheur *via* le site http://frags2drugs.icoa.fr/.

Le programme F2D, bien que déjà utilisé pour la découverte de nouveaux PKI, est parallèlement en constante amélioration. Ces manipulations parallèles du programme F2D doivent être soigneusement structurées afin d'éviter des conflits dans le code source du programme. En effet, les dernières innovations apportées au code source peuvent gêner le bon déroulement de la recherche de PKI, en se révélant incompatibles avec d'autres parties essentielles du code.

Pour pallier ce problème, nous utilisons le logiciel Git en tant que contrôleur de versions du code de F2D. Git est un logiciel libre de droits (distribué sous la licence publique générale GNU) permettant à chaque développeur de conserver l'historique des changements apportés aux fichiers. La gestion des versions grâce à Git est décentralisée : chaque programmeur a sur son ordinateur une copie du code source (dépôt local) avec tout l'historique des modifications apportées à chaque fichier. Git donne la possibilité de :

- Retourner à une version précédente en cas d'erreur
- Collaborer à plusieurs sans risquer de supprimer les modifications apportées par les autres
- Suivre l'historique des modifications étape après étape

Git permet d'apporter de nouvelles fonctionnalités sans affecter le bon déroulement des projets dans lesquels F2D est utilisé.

Pour partager plus facilement les fichiers du code source entre les différents développeurs, le dépôt est hébergé sur un serveur distant. GitLab est le logiciel que nous utilisons pour conserver une copie délocalisée de l'historique du code de F2D (dépôt distant). Il existe deux versions de GitLab, la première est libre de droits (sous licence MIT) ; il s'agit de celle que nous utilisons. La seconde est propriétaire, payante et propose plus de fonctionnalités. Grâce à GitLab, les modifications apportées par chaque développeur sont centralisées. De manière semblable à une sauvegarde de fichiers dématérialisée « en nuage » (iCloud, Google Drive, Microsoft OneDrive …), le stockage du code dans GitLab évite de perdre des données en cas de problème sur son propre ordinateur. GitLab donne aussi une interface facilitant l'utilisation de Git pour la mise en commun du code.

2.2.3 Analyse de l'architecture du code source de F2D

Le code source de F2D est composé de 28 156 lignes écrites en langage Python et réparties dans 46 fichiers (Tableau 14). Ce décompte a été effectué après la réorganisation du code, en novembre 2020, au cours de laquelle j'ai unifié différentes versions du code enregistrées sur divers ordinateurs.

Tableau 14 : Détails sur chaque fichier de code source de F2D. La couleur des lignes indique l'emplacement du fichier dans l'architecture de F2D. En jaune : SBCTools/f2d, en rouge : frags2drugs/Database_Creation/Create_Database, en violet : frags2drugs/Database_Creation/Create_Database/on_cluster, en bleu : frags2drugs/Selection, en orange : frags2drugs/Growing/breadth_first_search, en vert : frags2drugs/Growing/depth_search, en rose : frags2drugs/Graph_analyse.

Fichier	Explication	Nombre de lignes
logging.py	Gestion de l'affichage de messages	43
numpy.py	Fonctions basées sur la librairie NumPy	174
basic.py	Fonctions utilitaires généralistes	1286
geo.py	Calculs géométriques	1834
networkx.py	Fonctions basées sur la librairie Networkx	276
rdkit.py	Fonctions basées sur la librairie RDKit	1960
rdkit_is.py	Tests effectués à partir de molécules	49
cypher.py	Requêtes Cypher	389
neo4j.py	Interaction avec Neo4j	678
pymol.py	Interaction avec PyMol	417
biopandas.py	Fonctions basées sur la librairie Biopandas	1303
show.py	Affichage de molécules ou graphes	467
angle.py	Calculs d'angles	69
peptide.py	Tests effectués à partir de peptides	208
rdock.py	Interactions avec rDock	1307
fragment.py	Fragmentation des molécules	1879
fragment_pep.py	Fragmentation des peptides	1006
rmsd.py	Calculs de RMSD	586
clustering.py	Regroupement de molécules	914
analyse.py	Fonctions utilitaires pour la sélection de molécules	578
tools.py	Fonctions "boîte à outils"	1173
base.py	Création de la base de données orientée graphes	1612
base_angles.py	Intégration du champ de force MMFF94	1180
cypher.py	Requêtes Cypher	476
dfi.py	Création du DataFrame des fragments	418
neo4j.py	Interactions avec Neo4j	220
niv_angles.py	Calcul de paramètres de distorsion des angles	255
target.py	Analyse et création des cibles (relations de compatibilité)	408
create_EI_csv.py	Passage de la création de la BDD sur cluster de calcul	136
find_EI.py	Passage de la création de la BDD sur cluster de calcul	86
filtering.py	Filtres PKI-like	226
community_status.py	Fonctions complémentaires de NetworkX	79
community.py	Fonctions complémentaires de NetworkX	479
grow.py	Agrandissement des fragments par parcours en largeur	2075
generate_lgmol_on_cluster.py	Passage de l'agrandissement des fragments sur cluster	30
exceptions.py	Création d'exceptions pour la gestion d'erreurs	110

node.py	Fonctions pour le parcours de la BDD orientée graphes	542
network.py	Fonctions pour le parcours de la BDD orientée graphes	434
growing_path.py	Recherche en profondeur	195
rules.py	Prise en compte des paramètres de l'utilisateur	495
fragment.py	Fonctions associées aux fragments	65
database.py	Interactions avec la BDD orientée graphes	361
generation.py	Création des molécules à partir de nœuds du graphe	560
tools.py	Fonctions utilitaires	313
frags2drugs.py	Appel des fonctions pour l'agrandissement des fragments	475
visualization.py	Affichage d'éléments du graphe	330

Les 46 fichiers de codes sont répartis en 2 modules principaux : « SBC Tools F2D » s'ils peuvent être utilisés dans d'autres projets ou « Frags2Drugs » s'ils sont spécifiques à F2D (Figure 31). Le module « SBC Tools » contient 21 fichiers de code source. Dans l'autre module « Frags2Drugs », les fichiers de codes sont répartis selon leur application. Ainsi, un premier sous-module regroupe 9 fichiers dédiés à la création de la BDD, un deuxième comprend 14 fichiers concernant l'agrandissement des molécules, et les deux autres, chacun composé d'un fichier, permettent respectivement l'analyse du graphe de la BDD et la sélection des meilleures molécules.

Les dépendances entre ces différents fichiers sont nombreuses, elle sont représentées par les flèches de la Figure 31. Par exemple, des fonctions du fichier « *logging* », permettant l'affichage de messages pour suivre le déroulement de l'exécution du programme, sont appelées dans 39 autres fichiers.

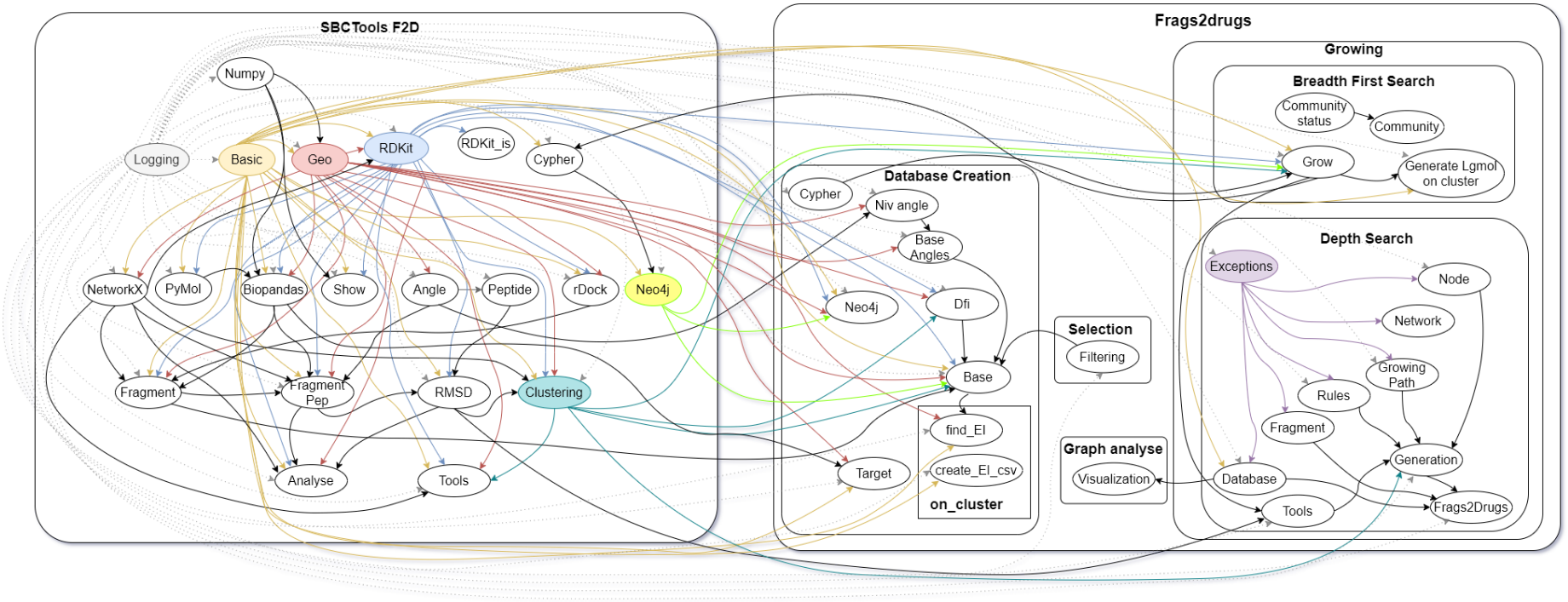


Figure 31 : Organisation des différents modules du code source de F2D

Mes contributions au code source de F2D ont permis d'optimiser son organisation pour gagner en clarté et en simplicité. En effet, par cette réorganisation j'ai supprimé des doublons dans le code ou des fonctions devenues obsolètes. Cette réorganisation a aussi permis de supprimer des bogues liés à des imports circulaires. Pour fonctionner correctement, une fonction repose sur « l'appel » d'autres fonctions venant de fichiers développés en amont. Une erreur sera engendrée si une nouvelle fonction est appelée dans un fichier dont elle dépendait déjà : il s'agit d'un import circulaire.

2.2.4 Mise en place d'une installation facilitée via Docker

Pour permettre une installation facile de F2D sur n'importe quel ordinateur, j'ai utilisé le logiciel Docker. Docker est un logiciel libre de « virtualisation légère » reposant sur quatre notions : les conteneurs, les images, les fichiers de configuration et les volumes (Figure 32). Similaire aux machines virtuelles, Docker permet d'installer et d'utiliser un logiciel dans un environnement virtuel spécifique, sans toucher aux fichiers systèmes de l'ordinateur. L'avantage d'utiliser Docker par rapport aux machines virtuelles est qu'il est plus léger, ne nécessitant pas la réinstallation d'un nouveau système d'exploitation pour chaque application[270].

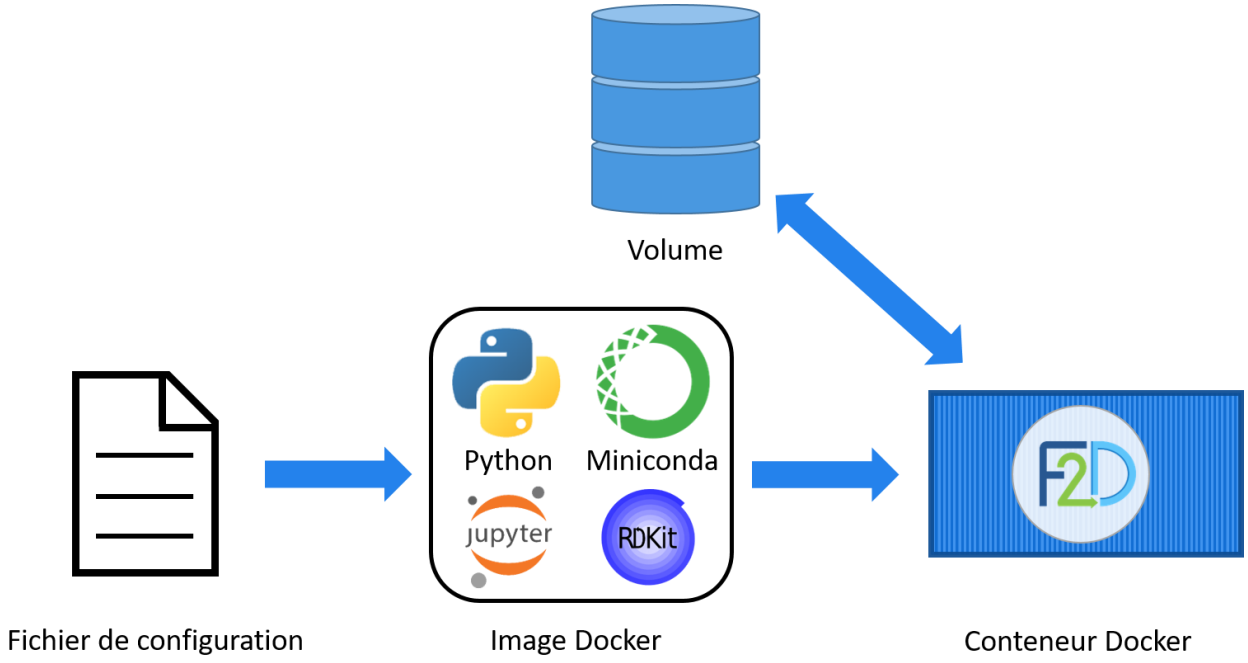


Figure 32 : Représentation des différents éléments nécessaires pour le fonctionnement de F2D via Docker. Les quatre langages de programmation et librairies listées dans l'image Docker sont un exemple de certaines technologies sur lesquelles reposent F2D.

Les conteneurs Docker sont des unités élémentaires de logiciels. Ils contiennent l'environnement virtuel nécessaire pour faire fonctionner un programme. Un conteneur peut être utilisé sur n'importe quel ordinateur, quel que soit son système d'exploitation, si Docker est installé dessus. Ainsi, un logiciel reposant sur Linux peut être exécuté sur Windows par exemple, sans devoir développer une version spécifique pour cela. Une image Docker est un système de fichiers permettant de créer un conteneur. À partir d'une image, le même conteneur pourra être créé plusieurs fois. Le fichier de configuration de Docker permet la création de l'image ainsi que la modification de celle-ci. Enfin, les volumes permettent d'associer des données aux conteneurs. Sans l'utilisation des volumes, les données sont perdues à chaque suppression du conteneur.

L'utilisation de F2D repose sur des cahiers électroniques Jupyter appelés « *notebooks* ». Ces *notebooks* permettent d'appeler les fonctions du code source de F2D pour les exécuter et en afficher

les résultats dans une même interface. Des commentaires sont ajoutés pour gagner en clarté sur le fonctionnement de chaque *notebook*. Différentes librairies Python sont utilisées dans F2D, en particulier pour la gestion des données (Pandas), le calcul scientifique (NumPy, SciPy) et la chémoinformatique (RDKit). Pour faire cohabiter toutes les différentes librairies utilisées dans F2D, le gestionnaire de paquets Anaconda est utilisé, par son installation la plus légère possible appelée « Miniconda ».

Le conteneur Docker que j'ai mis en place intègre donc ces différentes librairies *via* un environnement Miniconda. Docker-compose est un outil permettant le lancement et la communication entre différents conteneurs. Dans le cadre de F2D, il n'y a qu'un conteneur, mais docker-compose est utilisé pour ajouter un degré de simplicité supplémentaire au lancement de ce logiciel. Au lieu de devoir préciser à chaque lancement les volumes et les ports à utiliser, ceux-ci sont précisés dans un fichier de configuration spécifique à docker-compose. Le lancement de F2D via les *notebooks* Jupyter peut alors se faire en une seule instruction « docker-compose up ».

2.3 Conclusion et perspectives

L'ensemble du travail effectué, depuis la première mise en place de Frags2Drugs (F2D) en 2017 jusqu'à aujourd'hui, a fourni un logiciel fonctionnel et performant pour la recherche de nouveaux inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI). J'ai évalué F2D pour le rendre plus robuste et lui permettre de créer les meilleures molécules possibles. J'ai aussi apporté à ce programme une meilleure organisation et un meilleur suivi des modifications lors du développement. Enfin, j'ai contribué au développement de ce logiciel par l'apport de nouvelles fonctionnalités et la correction de bogues.

Des améliorations peuvent encore être apportées à ce logiciel. En premier lieu, il faudrait compléter la documentation associée au code source et y supprimer encore certains doublons. Cela rendrait le code de F2D encore plus performant et faciliterait sa valorisation.

Le développement de F2D ayant été initié en 2017 et les technologies informatiques évoluant très rapidement, certaines parties du code peuvent être revues pour gagner en performance et en rapidité d'exécution. De plus, une autre amélioration serait de ne plus stocker les relations d'exclusion dans la BDD orientée graphes, mais de les calculer « à la volée ». Les encombrements stériques pourraient être calculés au moment d'associer 2 fragments au lieu de rechercher la présence d'une relation d'exclusions préalablement calculée.

Il semble néanmoins, selon les retours que nous avons obtenus des collaborateurs (académiques et industriels), que la vitesse d'exécution du programme ne soit pas l'élément primordial de leurs attentes vis-à-vis d'un programme comme F2D.

Il serait aussi possible de s'affranchir de l'alignement effectué entre les différentes protéines kinases. Les fragments pourraient être positionnés dans une protéine cible à l'aide de méthodes basées sur l'identification de sous-cavités dans la cible[65]. L'utilisation de sous-cavités permettrait aussi de cibler d'autres familles que les protéines kinases. Une thèse poursuivant le développement de F2D en ce sens va d'ailleurs démarrer en octobre 2021 dans l'équipe SB&C de l'ICOA.

Aujourd'hui, des méthodes de conception de nouveaux inhibiteurs sont créées en se basant sur l'apprentissage automatique (*Machine Learning,* ML) ou profond (*Deep Learning,* DL, Chapitre 1 partie 1.5). La plupart de ces méthodes reposent sur l'utilisation de molécules 2D et n'intègrent pas réellement la 3D dans le processus de création de molécule. La seule qui s'approcherait d'une méthode de conception d'inhibiteurs basée sur les fragments (Fragment-Based Drug Design, FBDD) est DeLinker, mais il s'agit d'un logiciel de liaison de fragments et pas d'agrandissement[200]. Avec les récentes avancées

en intelligence artificielle (IA), il ne serait pas étonnant de voir apparaître un programme de FBDD basé sur l'IA faisant concurrence aux programmes « classiques ».

L'ajout de DL à F2D pourrait permettre d'améliorer la sélection des molécules, en estimant leurs voies de rétro-synthèse. Cette estimation a déjà donné des résultats probants dans la littérature[170,271–273] et sera ajoutée à F2D dans la thèse qui fera suite à la mienne. Le DL permettra aussi d'estimer l'activité biologique des molécules créées. F2D pourrait ainsi être utilisé en amont d'un autre outil en cours de développement à l'ICOA, qui fait l'objet de la thèse du Dr Pierre-Yves Libouban intitulée « Prédictions d'affinités de liaison de complexes protéines-ligands par combinaisons de simulations de dynamiques moléculaires et d'algorithmes d'apprentissage profond ».

F2D fait partie de la plateforme web que j'ai mise en place pour donner accès à l'utilisation des différents outils développés dans l'équipe SB&C de l'ICOA. Le prochain chapitre présentera la mise en place de cette plateforme et les différents outils qui y sont présents.

Chapitre 3 : Plateforme d'outils SB&C - Ouverture à l'utilisation des outils de l'équipe SB&C

L'utilisation des outils développés au sein de l'équipe SB&C de l'ICOA nécessite certaines connaissances préalables en informatique. Par exemple, pour utiliser Frags2Drugs (F2D), il faut être capable de télécharger les fichiers, de les décompresser, de les organiser comme indiqué dans la documentation, d'installer Docker, Docker compose, de lancer le programme, puis d'accéder au *notebook* Jupyter. À partir de là, il faut quelques connaissances en Python pour appeler les bonnes fonctions avec les bons paramètres afin d'obtenir le résultat escompté.

Bien que la complexité de l'installation soit réduite grâce à Docker et Anaconda, cette procédure pourrait décourager un chercheur ayant peu de connaissances en programmation informatique. C'est pourquoi, depuis mon arrivée en stage de master 2 à l'ICOA, j'ai développé une plateforme permettant une utilisation par les chercheurs des outils développés par l'équipe SB&C, sans avoir besoin de connaissance informatique (Figure 33).

Figure 33 : Page d'accueil de la plateforme SB&C tools. http://sbc.icoa.fr/

Pour mieux comprendre ce chapitre, voici quelques définitions. La toile mondiale (*World Wide Web,* web*)* est le système permettant d'accéder aux ressources du réseau informatique mondial : internet. L'accès aux pages web contenant ces ressources se fait par un site web, lui-même référencé par une adresse web. Un site web est hébergé sur un ordinateur répondant à des requêtes web (appelé serveur web). Ces requêtes passent par le protocole de transfert hypertexte (*Hypertext Transfer*

Protocol, HTTP) entre clients (navigateurs web) et serveurs web. L'équivalent sécurisé du protocole HTTP est HTTPS, qui y ajoute une couche de chiffrement pour garantir la confidentialité et l'intégrité des données échangées entre l'utilisateur et le serveur web. Il existe une deuxième définition d'un serveur web, à savoir un logiciel fournissant les ressources web. Enfin, une plateforme web est un service web permettant d'accéder à différents programmes (ou outils informatiques) via d'autres sites web.

Habituellement, pour accéder à un site web, l'utilisateur saisi un nom de domaine dans son navigateur web. Ce nom de domaine est traduit en une adresse de protocole internet (adresse IP) par un système de noms de domaine (Domain Name System, DNS). Cette adresse IP fait référence au serveur web hébergeant le site web et toutes ses ressources. La recherche de la bonne adresse IP par le serveur DNS se fait progressivement, en utilisant différents serveur DNS pour localiser le bon serveur web (Figure 34).

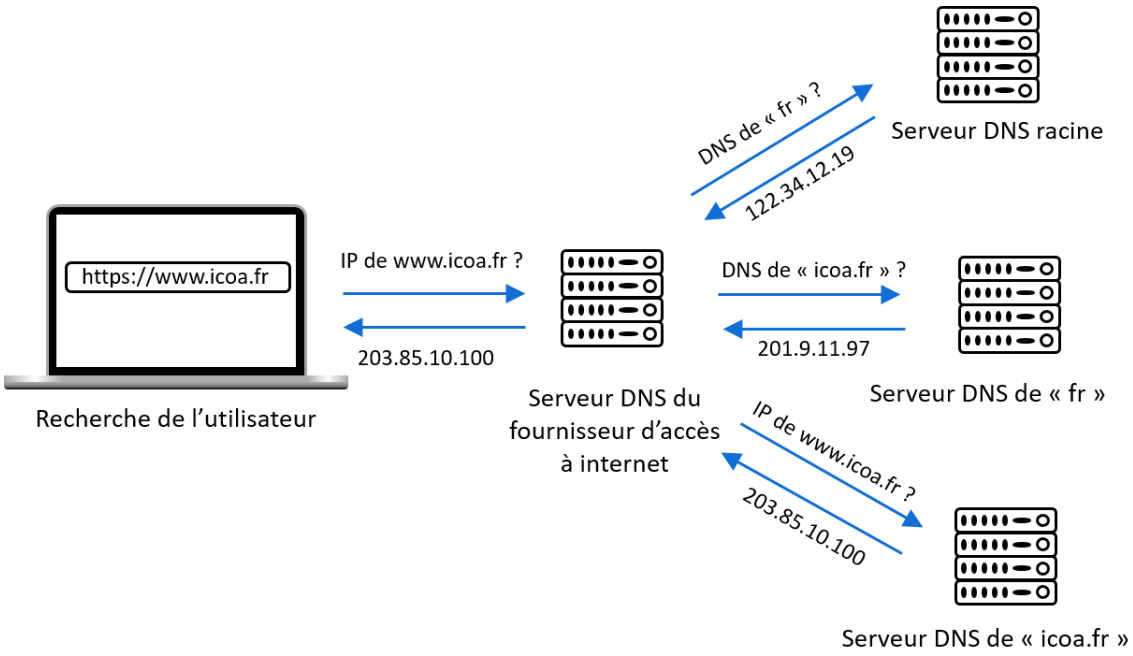


Figure 34 : Schéma du chemin parcouru au travers des serveurs DNS pour obtenir l'adresse IP d'un site web. D'après https://openclassrooms.com/fr/courses/1243111-gerer-son-nom-de-domaine

Lors de la création d'un site web, il y a trois étapes majeures : le développement, les tests et la mise en production (ou déploiement). Pendant la phase de développement, le site web fonctionne avec un serveur web local qui n'est pas exposé à internet. Le site doit ensuite passer des tests qui concernent, par exemple, la sécurité, la montée en charge, l'optimisation et la correction d'erreurs. Ensuite, la phase de production consiste à fournir un accès au site à partir d'internet. La principale difficulté de la mise en production est d'être capable de fournir cet accès pour une période infinie. Cet accès passe par le serveur web (logiciel) qui doit être plus sécurisé et plus performant que celui de la phase de développement.

La plateforme web que j'ai mise en place, nommée « *SB&C tools* », est accessible à l'adresse http://sbc.icoa.fr/ et est hébergée sur un serveur *web* de l'ICOA. Le développement de cette plateforme et des outils qui y sont regroupés m'a permis de collaborer avec plusieurs personnes de l'équipe SB&C. J'ai ainsi encadré trois stagiaires (étudiants de licence 3 informatique et master 2 bioinformatique) au cours de la mise en place de la plateforme *SB&C tools*. Parmi ces trois stagiaires, Juliette Douare a conçu les logos des outils de la plateforme web (Figure 33).

3.1 Présentation des outils

Sept outils sont présents au sein de la plateforme *SB&C tools* (Figure 33). En plus de la création de la plateforme, j'ai particulièrement travaillé sur quatre d'entre eux : Moldesc, Frags2Drugs, KinoMine et MetaPredict. Je reviendrai en détail sur ces quatre sites, mais pour les introduire, Moldesc permet le calcul de descripteurs moléculaires à partir de molécules dessinées *via* son interface. Frags2Drugs permet la recherche d'inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI) à partir d'une protéine kinase cible. KinoMine permet d'explorer les connaissances provenant de différentes bases de données (BDD) sur le kinome humain. MetaPredict permet le calcul de propriétés d'absorption, distribution, métabolisme, excrétion et toxicité (ADMETox) à partir de modèles statistiques.

Je n'ai pas participé au développement des trois autres outils. Parmi ces derniers, *Protein Kinase Inhibitor Database* (PKIDB)[34,35] est le premier outil de la plateforme à avoir été mis en ligne. Comme déjà évoqué dans le Chapitre 1 partie 1.7.3, cette BDD présente des informations sur les PKI actuellement sur le marché ou en cours d'essais cliniques disposant d'une DCI[34,35]. PKIDB est accessible depuis une adresse formée différemment de celles des autres outils (https://www.icoa.fr/pkidb/). Les six autres outils sont accessibles en écrivant le nom de l'outil en préfixe du nom de domaine « icoa.fr ».

ADMETox network est un site montrant l'organisation entre les différentes bases données utilisées en chémoinformatique et particulièrement sur l'ADMETox. Ce site est accessible, depuis la plateforme *SB&C Tools,* à l'adresse http://admetnetwork.icoa.fr.

Au cours de ce chapitre, je vais commencer par présenter l'environnement technique général permettant à la plateforme *web* de fonctionner. Je détaillerai ensuite les contributions apportées aux outils de la plateforme. Je terminerai en présentant les améliorations envisageables pour poursuivre le développement de la plateforme *SB&C tools*.

3.2 Environnement technique de la plateforme SB&C

3.2.1 Reverse proxy (serveur mandataire inverse)

Pour mettre en place la plateforme web, j'ai opté pour l'utilisation d'un serveur mandataire inverse, couramment appelé *reverse proxy* (Figure 35). Il s'agit d'un serveur par lequel un utilisateur doit passer pour accéder aux serveurs *web* internes. Les serveurs web internes donnent accès aux différents outils de la plateforme web. L'avantage d'utiliser cet intermédiaire est qu'il donne une seule ouverture vers internet permettant une sécurité plus grande que si chaque outil de la plateforme avait sa propre ouverture vers internet. Le *reverse proxy* fait correspondre les adresses locales aux noms de domaines saisis par l'utilisateur.

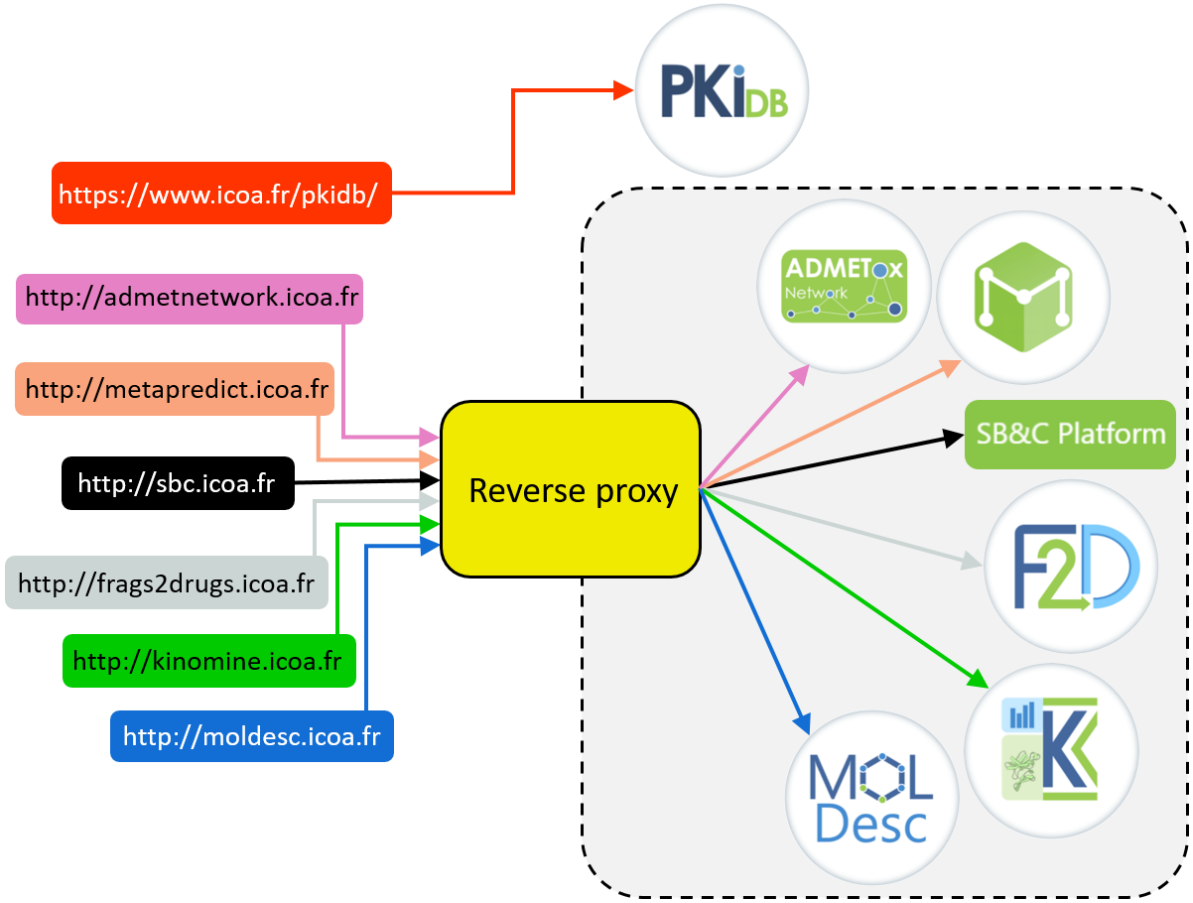


Figure 35 : Organisation par reverse proxy des outils de la plateforme SB&C.

Les serveurs web (logiciels) les plus populaires sont Nginx et Apache avec une représentation respectivement de 34% et 33,4% des sites web dont le serveur est connu au 26 mai 2021. Ces deux serveurs web proposent une utilisation en *reverse proxy*. J'ai choisi d'utiliser Nginx en tant que *reverse proxy*, car il requiert moins de mémoire et de ressources qu'Apache et bénéficie d'une documentation abondante pour le mettre en place avec Docker. Nous allons maintenant voir quelle est l'utilisation de Docker au sein de la plateforme SB&C.

3.2.2 Docker et Docker Compose

Le principe du fonctionnement de Docker a déjà été expliqué dans le chapitre 2, partie 2.2.4, pour son utilisation dans Frags2Drugs. Il permet de faciliter l'installation de ce logiciel dans un nouvel ordinateur. Au sein de la plateforme web, Docker est utilisé pour unifier les différents environnements des outils composés de différents serveurs web et langages de programmation (Tableau 15). Les langages de programmation sont : *HyperText Markup Language* (HTML), *Cascading Style Sheets* (CSS), *JavaScript* (JS), *Hypertext Preprocessor* (PHP), *Structure Query Language* (SQL), Python, Java et PostgreSQL. Les serveurs web utilisés sont *Green Unicorn* (Gunicorn) et Apache.

Tableau 15 : Serveurs web et langages utilisés dans les outils de la plateforme SB&C.

Outil	Serveur web	Langages web	Autres langages
SB&C tools	Apache	HTML, CSS, JS	
Admetox network	Apache	HTML, CSS, JS	
MetaPredict	Apache	HTML, CSS, JS	PHP, Java, Python
KinoMine	Gunicorn	HTML, CSS, JS	Python, PostgreSQL
Frags2Drugs	Gunicorn	HTML, CSS, JS	Python, PostgreSQL
Moldesc	Gunicorn	HTML, CSS, JS	Python
Protsong	Apache	HTML, CSS, JS	PHP

Docker permet un passage très simple du développement à la mise en production grâce à une architecture en microservices. Les microservices sont des composant logiciels indépendants échangeant un faible nombre d'informations et de manière asynchrone. Dans Docker, ces microservices sont nommés « conteneurs », ils sont individuels facilement remplaçables et conçus par rapport à leur utilisation.

Un réseau Docker nommé « nginx-proxy » est défini pour permettre la communication entre les différents containers via Docker Compose. Le *reverse proxy* est un conteneur créé à partir d'une image pouvant être récupérée depuis le dépôt Docker *hub* (https://hub.docker.com/r/jwilder/nginx-proxy). Le *reverse proxy* attend l'apparition d'autres conteneurs sur ce réseau interne pour permettre un accès vers eux. Chaque site web de la plateforme est composé d'au moins un conteneur, mais il peut aussi y en avoir plusieurs comme nous le verrons dans les parties 3.5 et 3.6.

Docker Compose est adapté à un projet de petite taille, comme la mise en place de la plateforme SB&C, mais si une grande quantité d'utilisateurs venait à utiliser simultanément les outils, il faudrait s'adapter à cette montée en charge. Pour cela, les conteneurs pourraient être dupliqués pour répartir le nombre d'utilisateurs par conteneurs. À partir d'un grand nombre de conteneurs, des outils d'orchestration de conteneurs tels que Docker Swarm ou Kubernetes peuvent être utilisés.

Nous avons vu que l'architecture de fonctionnement de la plateforme web est en *reverse proxy* et que la technologie principalement employée pour la faire fonctionner est Docker. Nous allons maintenant découvrir en détail les quatre sites web que j'ai développé et/ou rendu accessibles, en commençant par MetaPredict.

3.3 MetaPredict

3.3.1 Description

MetaPredict est un logiciel développé au cours de la thèse de Baptiste Canault intitulée « Développement d'une plateforme de prédiction *in silico* des propriétés ADME-Tox », soutenue en 2018. MetaPredict permet de créer automatiquement des modèles statistiques de relations quantitatives structures-activités (QSAR) pour la prédiction de propriétés d'absorption, distribution, métabolisme, excrétion et toxicité (ADMETox). MetaPredict permet d'optimiser chacune des étapes de création de ces modèles et d'améliorer leur qualité et leur robustesse[274]. Pendant cette thèse, une interface web a été développée, permettant de faciliter l'utilisation des modèles QSAR, l'interprétation des résultats et de s'adapter aux spécificités de chaque projet de recherche.

En plus de l'interface web, sur laquelle nous nous concentrerons dans cette partie, un *notebook* Jupyter permet de créer de nouveaux modèles statistiques afin de calculer de nouvelles propriétés ADMETox.

3.3.2 Mise en production

Le site web ne pouvant être lancé que localement, j'ai créé un conteneur Docker pour permettre son ajout sur la plateforme web. Pour cela j'ai utilisé une image Docker ayant déjà PHP et le serveur web Apache installés. Puis j'ai rédigé un fichier de configuration indiquant l'installation de Java, d'Anaconda, de l'environnement Python et des programmes ChemAxon, VolSurf[275], ISIDA, et CDK[276]. Le fichier de configuration permet la création de l'image « metapredict_web » qui rend l'outil accessible. Pour cela, il faut construire le conteneur Docker à partir de cette image, en indiquant au *reverse proxy* le port et l'adresse auquel il est accessible (http://metapredict.icoa.fr). Parmi les programmes cités précédemment, ISIDA est distribué gratuitement, sans licence, CDK est libre de droits (sous licence GNU), alors que VolSurf et ChemAxon sont des logiciels commerciaux. Le site web de MetaPredict étant libre d'accès, nous avons ajouté dans le pied de page une phrase, en accord avec les entreprises distribuant ces logiciels, précisant que l'utilisation du site ne doit pas être effectuée pour une application commerciale.

Pour simplifier le lancement de cette interface, l'adresse du site et le port sont indiqués dans un autre fichier de configuration, celui de Docker Compose. Ainsi, lorsque le *reverse proxy* fonctionne, MetaPredict peut être lancé par la simple commande « docker-compose up ». Ce second fichier de configuration permet aussi de définir correctement les volumes contenant les données des modèles QSAR présentés sur le site.

Une image Docker est censée occuper un espace assez limité en mémoire, de l'ordre de plusieurs méga octets (Mo) à quelques giga octets (Go) au maximum. Cela apporte différents avantages comme un gain de temps lors du développement. L'ajout des données (modèles statistiques de MetaPredict) directement dans l'image occupait plus de 40 Go. Alors, pour réduire l'espace utilisé par l'image, j'ai placé les données en dehors, à l'aide de volumes externes. Cela a permis de réduire son poids à 6,55 Go. Avoir les données hors de l'image donne la possibilité d'utiliser les mêmes fichiers au sein de différents conteneurs. La définition externe des volumes a ensuite été utilisée dans tous les autres outils de la plateforme, par exemple pour les fichiers statiques (images et mise en forme du site). En séparant l'image Docker et les modèles statistiques, ces modèles peuvent être recréés ou modifiés sans avoir besoin de recréer l'image.

J'ai mis MetaPredict en production tel qu'il avait été développé au cours de sa thèse par Baptiste Canault. Cependant, d'autres améliorations, plus chronophages, étaient aussi nécessaires pour faire fonctionner parfaitement cet outil. C'est une des tâches qu'a pu effectuer My Nina Hong que j'ai encadré pendant son stage de licence 3 informatique.

3.3.3 Améliorations apportées

Les fichiers permettant au site web de MetaPredict de fonctionner sont générés à partir d'un script Python. Ce script était exécuté manuellement avant de pouvoir rendre le site web accessible par une copie de ces fichiers sur le serveur web. La plupart des fichiers servant à la mise en page étaient toujours recréés à l'identique. Cette fonctionnalité a été modifiée pour créer uniquement les fichiers changeant selon l'utilisation du site. De plus, ce script est exécuté automatiquement à chaque lancement du site web pour s'adapter aux modèles statistiques à inclure.

My Nina Hong a aussi modifié l'interface de lancement des prédictions, qui était composée d'une unique page regroupant toutes les différentes parties. Les explications et documentations ont été déplacées dans d'autres pages (à propos et guide de l'utilisateur). L'interface de MetaPredict est ainsi plus uniforme avec les autres outils de la plateforme. L'entête et le pied de page du site ont aussi été modifiés pour être unifiés avec le reste de la plateforme.

Le choix des modèles statistiques à utiliser pour calculer chaque propriété ADMETox a été réduit à un modèle par propriété pour utiliser seulement celui présentant les meilleures performances (Figure 36). De plus, ce choix ne peut s'effectuer que lorsqu'au moins une molécule a été saisie par l'utilisateur.

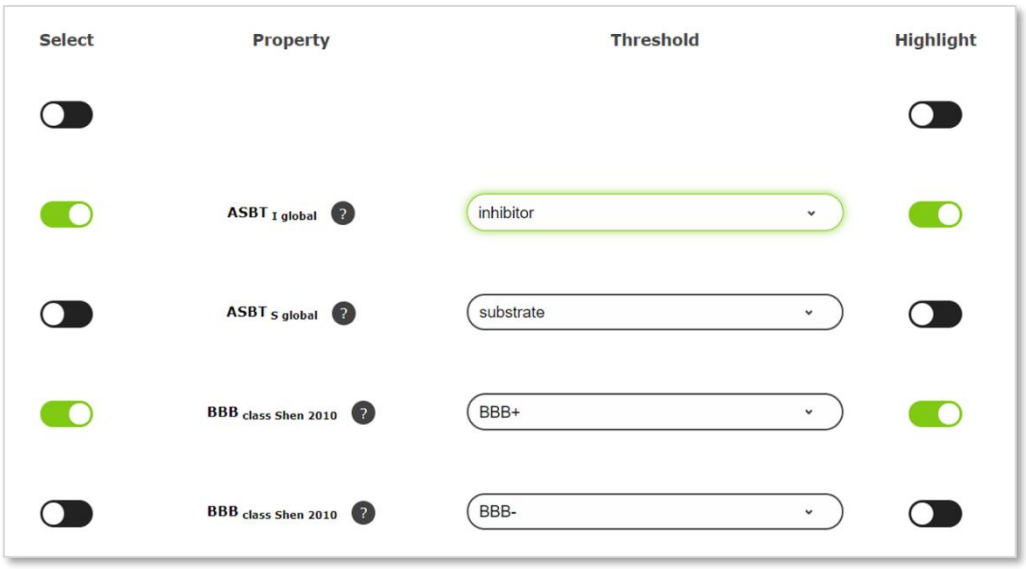


Figure 36 : Exemple de propriétés pouvant être sélectionnées dans MetaPredict

Les différentes erreurs pouvant être obtenues au cours de l'utilisation de ce site sont clairement affichées à l'utilisateur à l'aide de bannières (Figure 37). Cela apporte une interaction avec l'utilisateur qui peut ainsi mieux comprendre ce qui est attendu par le site pour qu'il fonctionne correctement.

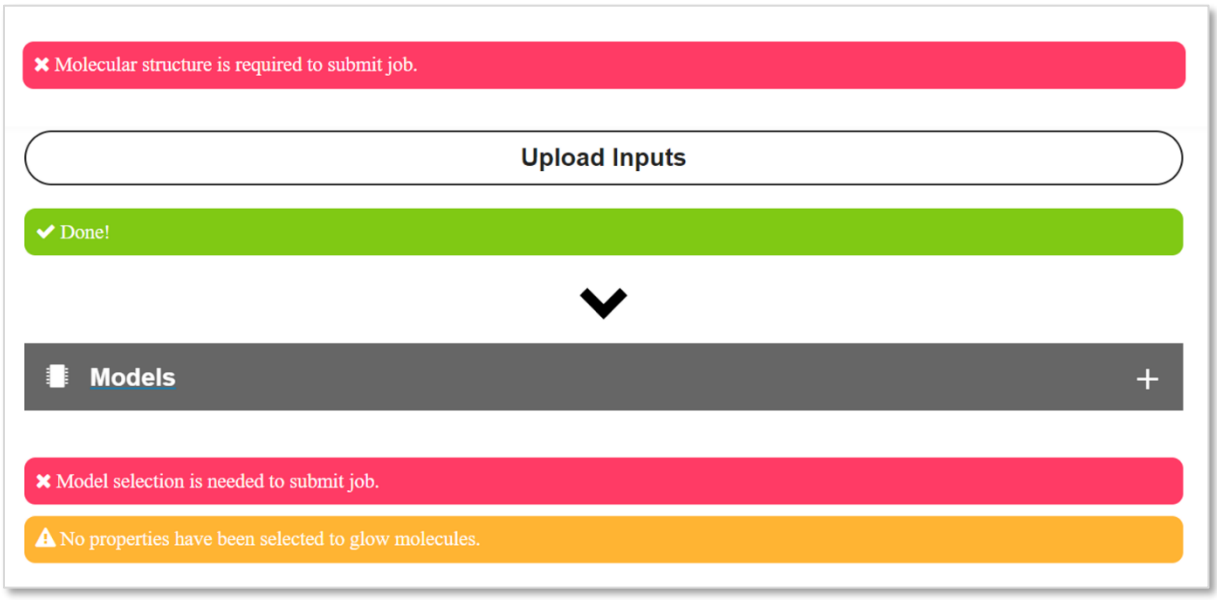


Figure 37 : Exemples de différents avertissements et différentes erreurs affichées dans MetaPredict. Les messages de succès sont affichés en vert, les avertissements sont affichés en orange et les erreurs en rouge.

Les résultats obtenus sur ce site sont présentés dans un tableau sur une autre page (Figure 38). Celui-ci a été transformé à l'aide de la librairie DataTables (https://DataTables.net/) permettant de ne plus créer de fichiers temporaires lors de l'export vers le format CSV ou HTML du tableau.

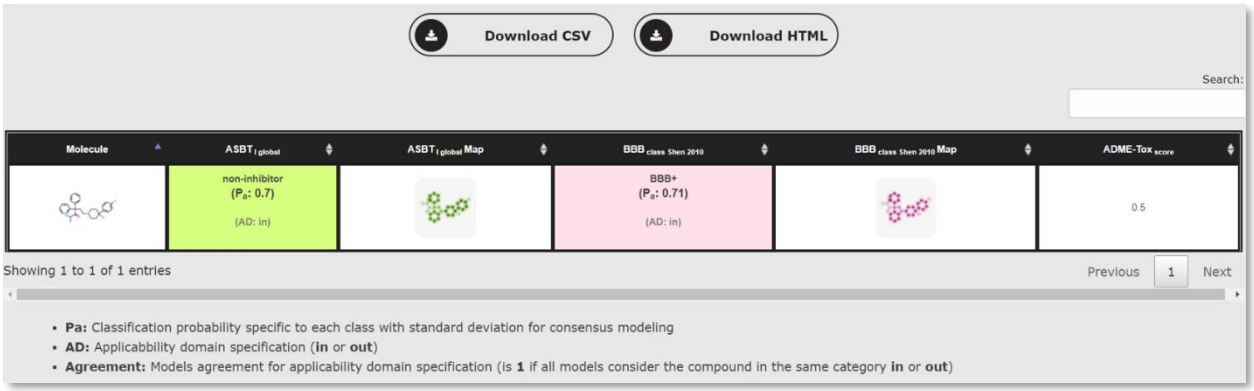


Figure 38 : Tableau de résultats de MetaPredict

Comme évoqué précédemment, de nouveaux modèles statistiques peuvent être ajoutés en plus de ceux déjà présents sur le site web de MetaPredict. Pour créer ces nouveaux modèles, un *notebook* Jupyter a été mis en place. Cependant, son installation nécessite un environnement virtuel avec des versions de librairies précises et parfois anciennes, ce qui présentait des problèmes de compatibilité lors de son installation et de son utilisation sur un nouvel ordinateur. Pour supprimer ces problèmes, My Nina Hong a créé une image Docker contenant toutes les librairies nécessaires et figeant leurs versions. Cette image permet la maintenance au cours du temps et le développement ou le portage sur différentes machines de MetaPredict.

3.3.4 Améliorations possibles

Metapredict peut être amélioré en supprimant automatiquement les fichiers créés temporairement. Pour cela, il faudrait mettre en place des tâches récurrentes, par exemple avec la librairie Celery qui a été utilisée pour Frags2Drugs (F2D) web et qui sera présentée dans la partie 3.6.2.

Le site web de MetaPredict peut aussi être amélioré en lui ajoutant une base de données (BDD) et en utilisant un « cadre de développement web ». Pendant la suite de ma thèse j'utiliserai l'anglicisme « *framework* web » à la place du terme précédent. Actuellement, une page web statique est créée à partir des modèles présents dans un dossier du site. MetaPredict pourrait être rendu dynamique, en gérant la liste des modèles et les informations associées à partir d'une BDD.

Après avoir créé une BDD pour MetaPredict, nous pourrions l'intégrer au site web via une architecture modèle-vue-contrôleur (MVC). L'architecture MVC repose sur trois modules ayant chacun une action permettant le bon fonctionnement d'un site web dynamique. Le modèle contient les données et leur organisation dans la BDD, la vue concerne l'affichage des données via une interface graphique et le contrôleur permet la manipulation et l'obtention des données par l'utilisateur (Figure 39).

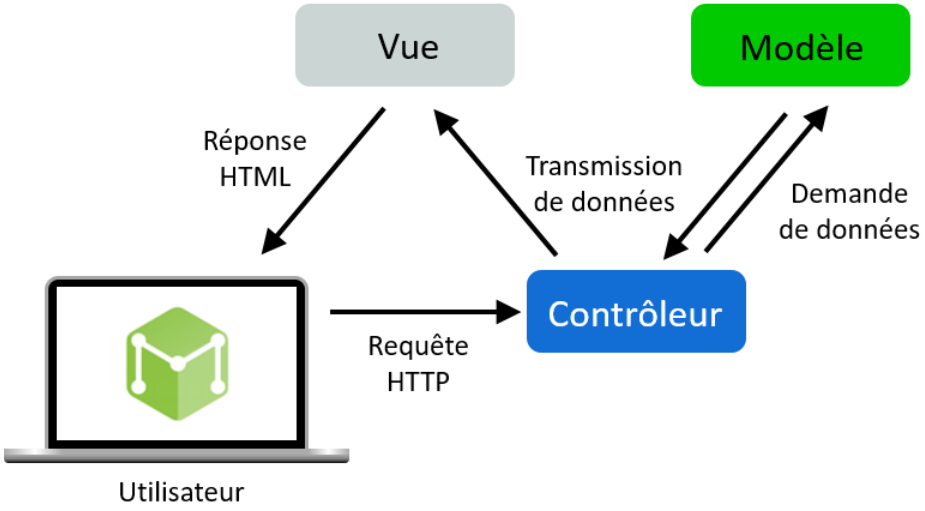


Figure 39 : Architecture MVC d'une application web. D'après https://symfony.com/legacy/doc/jobeet/1_4/en/04?orm=Propel

Les *frameworks* web, tels que Symfony (PHP) ou Django (Python), permettent de standardiser le développement et la mise en production d'applications web. Un *framework* web se base sur l'architecture MVC et sur la réutilisation du code avec une gestion des pages HTML par des gabarits. Nous pourrions utiliser un *framework* web pour ajouter l'architecture MVC et une BDD à MetaPredict.

Après avoir décrit l'outil MetaPredict, voyons maintenant MolDesc, le premier des 3 sites web que j'ai conçus, développés et mis en production à l'aide du *framework web* Django.

3.4 Moldesc

3.4.1 Description

MolDesc est un outil de la plateforme web SB&C permettant de calculer les valeurs de descripteurs physico-chimiques à partir d'une ou plusieurs molécules. Comme dans MetaPredict ou KinoMine, les molécules peuvent être dessinées à partir du module JavaScript Molecular Editor (JSME)[277] de manière à récupérer leurs codes *Simplified Molecular Input Line Entry Specification* (SMILES). Les molécules peuvent aussi être téléchargées vers le serveur à partir d'un fichier *Structure Data File* (SDF) avant d'effectuer les calculs. La Figure 40 représente le SMILES récupéré à partir du dessin de la structure du lopéramide dans le module JSME.

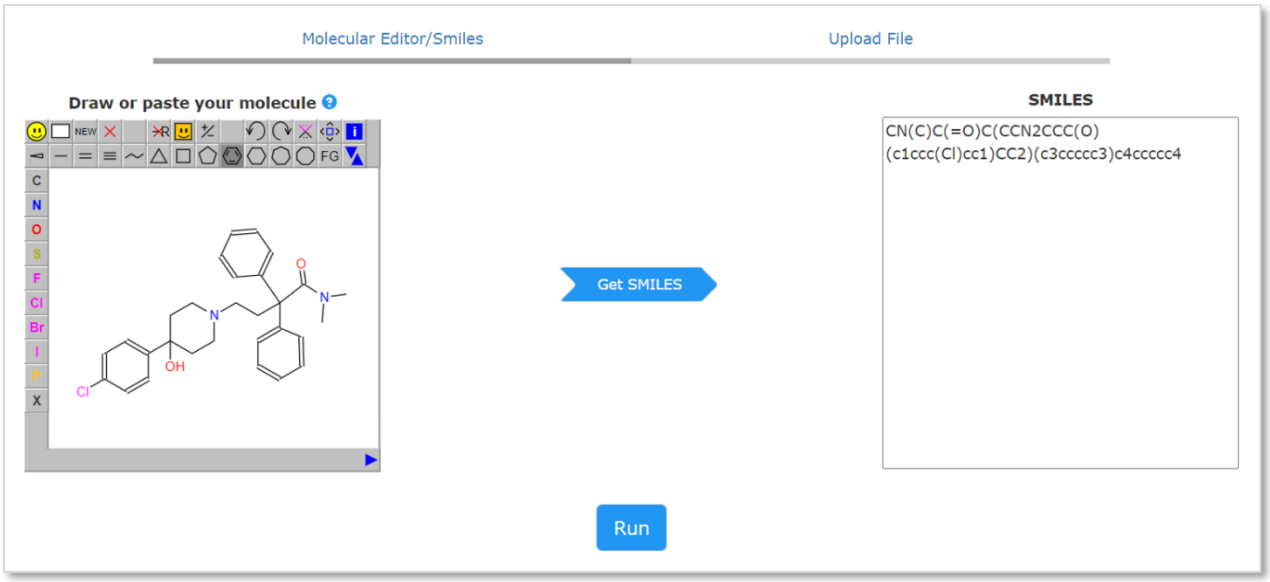


Figure 40 : Page de chargement des molécules sur MolDesc. Cette page illustre la représentation SMILES du lopéramide à partir du dessin de sa structure et le module JSME.

Les descripteurs calculés dans MolDesc sont présentés dans le Tableau 16. Ils sont calculés à l'aide des librairies *Rational Discovery Kit* (RDKit) ou ChemAxon. Les descripteurs calculés par la librairie RDKit sont : *Simplified Molecular Input Line Entry Specification* (SMILES), la masse molaire (Molecular Weight, MW), le coefficient de partage octanol/eau (LogP), la surface topologique accessible au solvant (TPSA), le nombre de d'accepteurs et de donneurs de liaison hydrogène (HBA et HBD), le nombre de liaisons rotables (NRB), le nombre d'atomes lourds (NHA), le nombre de règles de cinq (Ro5) violées, le nombre de cycles aromatiques (NAR), le nombre d'atomes chiraux (NCA), la fraction en carbones SP3 (FCSP3). Les autres descripteurs calculés par la librairie ChemAxon sont LogD, pKa et le score *Central Nervous System Multiparameter Optimization* (CNS MPO)[266]. Les sous-structures *pan assay interference compounds* (PAINS)[73], déjà présentées dans le chapitre précédent, sont identifiées, en plus des autres descripteurs, à l'aide d'une recherche par sous-structure à l'aide de RDKit.

À partir des valeurs obtenues pour chaque descripteur, différents scores sont calculés et des filtres sont mis en évidence pour interpréter plus facilement les résultats. Les seuils des valeurs attendues pour chaque filtre sont présentées dans le Tableau 16. Le filtre PKI-like (Tableau 12) provenant de PKIDB[35] et utilisé dans Frags2Drugs (F2D) est aussi présent dans MolDesc.

Tableau 16 : Seuils fixés pour les valeurs des descripteurs utilisés dans les différents filtres de MolDesc.

	MW (Da)	LogP	NRB	TPSA (Å^2)	HBD	HBA
Fragment-like[278]	<= 250	<= 3.5	<= 5			
Lead-like[279]	<= 350 et >=250	<= 3.5	<= 7			
Drug-like[280]	<= 500 et >= 150	<= 5	<= 7	< 150	<= 5	<= 10
GSK 4/400[281]	< 400	< 4				
Pfizer 3/75[282]		> 3		< 75		

Enfin, une recherche des composés identiques dans différentes bases de données (BDD) publiques est effectuée. La recherche s'effectue à partir de l'identifiant InchiKey de la molécule et de la BDD UniChem[283] permettant de retrouver toutes les occurrences des molécules dans les BDD

ChEMBL, PDBe, ZINC, SureChEMBL et PubChem. Lorsque la molécule est présente dans une de ces BDD, un lien est créé pour accéder à la page de la molécule correspondante.

Les résultats obtenus sont présentés sous la forme d'un tableau affiché grâce à la librairie DataTables. Ce tableau présente l'avantage de pouvoir être exporté aux formats HTML, CSV, Excel et SDF pour utiliser les molécules dans d'autres logiciels ou pour partager les molécules et leurs descripteurs.

3.4.2 Django

De manière similaire aux outils KinoMine et Frags2Drugs, qui seront présentés par la suite, MolDesc a été développé à l'aide du *framework web* Django, en langage de programmation Python. Le choix de ce langage de programmation se justifie par la volonté d'uniformiser tous les programmes développés dans l'équipe SB&C sous le langage Python. Django est le *framework web* Python libre de droits le plus complet offrant sécurité, rapidité et bénéficiant d'une documentation détaillée.

Django se base sur une adaptation de l'architecture modèle-vue-contrôleur (MVC) en une architecture modèle-gabarit-vue (model-template-view, MTV, Figure 41).

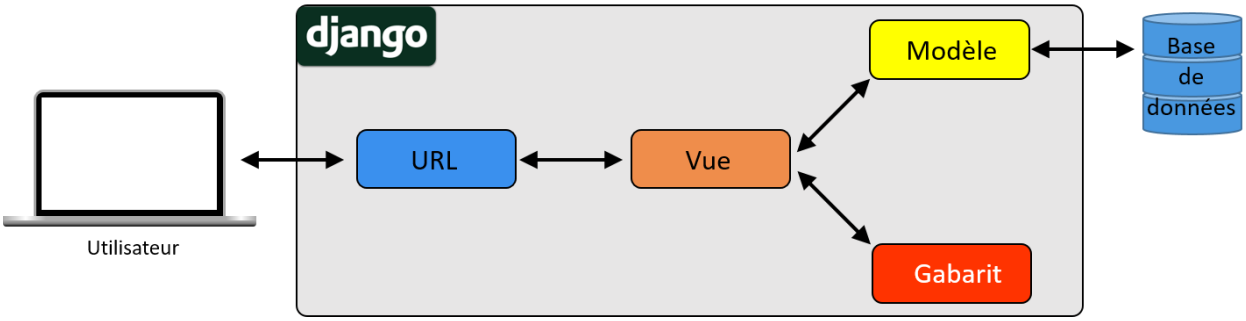


Figure 41 : Schéma de l'architecture classique MTV de Django.

Dans cette architecture le modèle a toujours pour rôle d'interagir avec la BDD. Le gabarit correspond à la vue du modèle MVC. Il s'agit d'une page HTML permettant d'éviter des redondances de code et pouvant contenir des objets Python. La vue de Django correspond au contrôleur du modèle MVC, elle a un rôle fondamental, elle reçoit la requête HTTP et effectue toutes les actions nécessaires (récupération de données, appel d'un gabarit) avant de renvoyer la réponse au navigateur web. La communication entre les requêtes de l'utilisateur et les bonnes vues dans Django se fait par l'intermédiaire d'adresses web (*Uniform Resource Locator*, URL).

MolDesc, au contraire des sites web KinoMine et Frags2Drugs, présente la particularité de ne pas reposer sur une BDD dédiée. Ce site ne stocke aucune molécule, il effectue des calculs et affiche les résultats sans les conserver. MolDesc nécessite alors une architecture sans modèle (Figure 42).

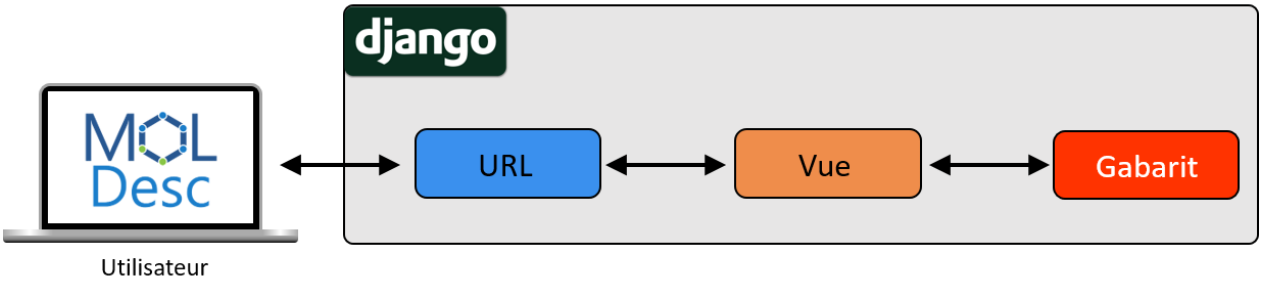


Figure 42 : Schéma de l'architecture de Django adaptée pour MolDesc.

Lors du développement du site web, Django utilise son propre serveur web permettant d'afficher les pages HTML et d'effectuer des requêtes dans un réseau local. Pour déployer et le rendre accessible depuis l'extérieur, il faut utiliser un serveur web plus robuste. N'importe quel serveur web peut être choisi, en passant par une *Web Server Gateway Interface* pour la communication entre les serveurs et les applications web en langage Python. Pour mettre en production MolDesc, j'ai choisi le serveur web *Green Unicorn* (Gunicorn) car il s'intègre bien avec Django, il est bien documenté et simple à mettre en place.

3.4.3 Améliorations apportées

Pendant le développement de MolDesc, j'ai encadré deux stagiaires de licence 3 informatique qui ont amélioré cet outil. Il s'agit de Juliette Douare (25/03/2019 – 02/08/2019) et My Nina Hong (15/06/2020 – 31/07/2020). Tout d'abord elles ont ajouté des fonctionnalités à ce site, puis elles ont participé à la mise en production, en se basant sur le déploiement de KinoMine que j'avais effectué auparavant et qui sera présenté dans la partie 3.5.

Juliette Douare a commencé par mettre à jour la version de Django de 1.11 à 2.2. Avoir la version la plus récente du *framework* permet de garantir une meilleure sécurité et un meilleur fonctionnement. Ensuite, elle a développé diverses fonctions pour vérifier que les codes SMILES saisis par les utilisateurs soient corrects. La possibilité de charger des molécules contenues dans un fichier (aux formats SD, SDF, TXT, CSV ou MOL) a aussi été ajoutée.

Au niveau de l'affichage des résultats, Juliette Douare a ajouté une colonne pour afficher la structure 2D de chaque molécule, avec la possibilité d'agrandir l'image. La colonne « filtres » a été ajoutée par Juliette Douare et complétée par My Nina Hong. Dans cette colonne, l'affichage en vert ou rouge des étiquettes indique si la molécule est retenue ou non par les filtres. Les PAINS sont aussi affichés de la même manière, selon trois catégories « A », « B » ou « C ». Ces catégories correspondent au nombre de touches que chaque sous structure PAINS fournissait dans la publication originale[73]. L'affichage de la présence de la molécule dans diverses BDD externes a été implémenté par My Nina Hong. My Nina Hong a aussi ajouté des modules de la suite logicielle ChemAxon à l'image Docker de MolDesc. Ces modules permettent le calcul des colonnes LogD et pKa pour pouvoir obtenir le score *Central Nervous System Multiparameter Optimization* (CNS MPO)[266]. La fraction en carbones hybridés SP3 (FCSP3) a aussi été ajoutée au tableau de résultats, comme les filtres GSK 4/400 et Pfizer 3/75 (Figure 43).

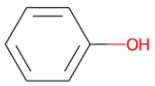
Molecule	SMILES	Exact Molecular Weight	LogP	LogD	Topological Polar Surface Area	...	Number of rule of five violations	Number of aromatic cycles	Number of chiral atoms	pKa	FCSP3	CNS MPO score	Filters	PAINS tag	External Databases
	Oc1ccccc1	94.0	1.4	1.7	20.2	...	0	1	0	-5.5	0	4.84	Fragment-like Lead-like Drug-like PKI-like GSK 4/400 Pfizer 3/75	A B C	ChEMBL : CHEMBL14060 PDBe : IPH ZINC : ZINC000005133329 SureChEMBL : SCHEMBL48 PubChem : 20488062 PubChem : 996

Figure 43 : Tableau de résultats obtenu pour un phénol sur MolDesc. Les colonnes "Number of hydrogen bond acceptors", "Number of hydrogen bond donors", "Number of rotatable bonds", "Number of heavy atoms" ont été masquées sur la capture d'écran.

En plus de la page de chargement des molécules et d'affichage des résultats, de nombreuses informations ont été ajoutées dans une page « à propos » et un « guide de l'utilisateur ». Celles-ci intègrent les références bibliographiques et des explications sur les différents descripteurs et filtres appliqués dans MolDesc.

My Nina Hong a aussi fait en sorte de pouvoir exporter le tableau de résultats aux formats HTML et Excel en y intégrant la représentation 2D des molécules et au format CSV et SDF avec un affichage correct des filtres.

3.4.4 Améliorations possibles

Au fur et à mesure des améliorations apportées à MolDesc, j'ai relevé des points qui restent améliorables. Le premier est qu'il faudrait afficher clairement à l'utilisateur les molécules qu'il a saisi de manière erronée. Pour cela, des bannières comme dans MetaPredict (Figure 37) pourraient lui indiquer les erreurs au lieu d'ignorer les molécules comme c'est le cas actuellement.

Une seconde amélioration concerne l'augmentation du nombre de molécules pouvant être prises en charge par MolDesc. Aujourd'hui, la taille maximale des fichiers de molécules pouvant être envoyés est de 1,5 Mo. Cette taille a été choisie pour éviter d'attendre trop longtemps que les calculs se terminent avant d'afficher le tableau de résultats. Ce temps de calcul est problématique car certains navigateurs (comme Firefox) arrêtent d'attendre la réponse de Django et n'affichent jamais les résultats en cas de temps trop long. La solution que nous envisageons serait d'utiliser Celery, un logiciel de gestion de file d'attente de tâches et de travaux asynchrones. Celery pourrait renvoyer une réponse au navigateur, lui indiquant que le calcul est encore en cours et celui-ci se rafraîchirait lorsque les résultats arriveraient. Nous découvrirons Celery plus en détail dans la partie 3.6.2 car il a été implémenté dans le site web de F2D.

Une autre amélioration pourrait être de pouvoir sélectionner les propriétés à calculer, après la saisie de molécules et avant d'afficher les résultats. Cela pourrait réduire le nombre de colonnes affichées et éviterait à certains utilisateurs d'effectuer des calculs pour des propriétés qui ne les intéressent pas.

Nous venons de voir les détails du site MolDesc, passons maintenant au site KinoMine, qui est chronologiquement le premier site de la plateforme SB&C que j'ai développé à l'aide du *framework* web Django.

3.5 KinoMine

3.5.1 Description

KinoMine est un outil permettant d'explorer les connaissances acquises sur les protéines kinases humaines et leurs inhibiteurs. Sur ce site, les données présentées proviennent de trois bases de données (BDD) principales : UniProt, *Chemical database of the European Molecular Biology Institute* (ChEMBL) et *Research Collaboratory for Structural Bioinformatics* (RCSB) *Protein Data Bank* (PDB). Les données de KinoMine ont été nettoyées par des méthodes implémentées au cours du développement d'algorithmes de ML, pour la prédiction d'activités biologiques d'inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI), dans la thèse de Fabrice Carles intitulée « Développement d'une approche protéo-chimiométrique tridimensionnelle pour l'identification d'inhibiteurs de protéines kinases ».

J'ai initié le développement de ce site au cours de mon stage de master 2 sous l'encadrement du doctorant Fabrice Carles et je l'ai mis en production dès le début de ma thèse. De nombreuses

améliorations ont été apportées à KinoMine pendant ma thèse et j'ai aussi pu encadrer Romain Launay pour cela pendant son stage de master 2 (09/03/2020 – 02/09/2020).

KinoMine permet d'abord d'accéder aux activités biologiques mesurées entre une sélection de PKI et une sélection de cibles. Ce site permet aussi d'observer les interactions formées entre les ligands et leurs cibles. Enfin, KinoMine donne pour chaque couple kinase/PKI les détails des annotations faites sur la séquence de la protéine, sur leurs complexes 3D et l'ensemble des bioactivités recensées pour ce couple.

En comparant ce site à d'autres déjà existants, KinoMine apporte une interface plus interactive et un niveau de détail supérieur concernant les cibles. En effet, les protéines comportant certaines mutations peuvent être distinguées les unes des autres. Les différentes pages et fonctionnalités de KinoMine sont montrées dans l'article le présentant en partie 3.5.3.

3.5.2 Base de données

3.5.2.1 Implémentation

À l'instar du site web de Frags2Drugs, qui sera présenté dans le Chapitre 3 partie 3.6, KinoMine repose sur une BDD relationnelle. L'architecture de Django pour ces deux sites est alors du type modèle-gabarit-vue (model-template-view, MTV, Figure 41). Par défaut Django utilise le moteur de BDD SQLite. La documentation de Django indique que ce moteur de BDD, le plus simple, est déjà inclus dans Python, mais que d'autres BDD comme PostgreSQL sont plus adaptées pour la mise en production[284]. Dans KinoMine et le site web de Frags2Drugs (F2D), j'ai remplacé SQLite par PostgreSQL. Django présente l'avantage d'utiliser un *Object-Relational Mapper* (ORM) transformant les lignes de code Python en requêtes SQL. Grâce à cet ORM, les requêtes n'ont pas à être réécrites en changeant de système de gestion de BDD (SGBD).

PostgreSQL est un SGBD relationnel, objet et libre de droits. Il contient tous les types de données SQL classiques (booléen, caractère, entier, etc), ainsi que d'autres types spécifiques de PostgreSQL dont certains sont déjà implémentés dans Django, comme *ArrayField* que j'ai utilisé pour stocker des listes de données. En plus de présenter une grande stabilité, PostgreSQL permet de définir de nouveaux types de variables et de nouvelles fonctions. J'ai choisi d'utiliser ce SGBD plutôt qu'un autre car il existe le *cartridge* RDKit permettant d'interagir avec les molécules dans PostgreSQL (https://rdkit.org/docs/Cartridge.html).

Le *cartridge* RDKit ajoute les types de RDKit dans PostgreSQL, ainsi les molécules y apparaissent en tant qu'objets RDKit (ROMol). Dans KinoMine, le *cartridge* RDKit permet de générer les empreintes moléculaires des molécules. Grâce à ces types, les recherches par sous-structures et similarité peuvent être effectuées directement dans la BDD. Un programme aussi utilisé par KinoMine permet d'intégrer ces types et fonctions dans les modèles de Django : django-rdkit (https://github.com/rdkit/django-rdkit).

Les empreintes moléculaires calculées et utilisées dans KinoMine sont présentées dans le Tableau 17. Parmi celles-ci, les trois premières empreintes du tableau sont déjà implémentées dans RDKit, en tant qu'empreintes moléculaires composées de bits (*bit fingerprint,* BFP). Elles sont aussi ajoutées à la BDD PostgreSQL grâce au *cartridge* RDKit. Le quatrième et dernier type d'empreinte moléculaire n'est pas présent par défaut dans RDKit, mais nous avons généré nos propres BFP pour l'ajouter.

Tableau 17 : Empreintes moléculaires utilisées dans KinoMine

Type d'empreinte	Abréviation	Taille (bits)	Définition
Morgan fingerprint	Mfp2	1024	Extended Connectivity FingerPrint (ECFP) basé sur sept propriétés atomiques[22]
Morgan fingerprint	Ffp2	1024	ECFP basé sur des pharmacophores[285]
Torsion bitvector fingerprint	Torsionbv	32	Empreinte moléculaire de l'angle de torsion, calculée à partir de 4 atomes lourds consécutifs[286]
Interaction fingerprint	IFP	680	Empreinte moléculaire calculée à partir de 8 interactions 3D entre protéine et ligand pour chaque résidu du site actif des protéines kinases

3.5.2.2 Structure de la base de données

La structure de la BDD de KinoMine a été établie après différentes étapes de conception et est présentée en Figure 44. Cette organisation de la BDD permet de garder la séparation entre les trois sources de données : structures 3D (RCSB PDB), séquence de protéine (Uniprot) et bioactivités (ChEMBL). Une autre table est séparée des autres et contient des informations générales sur les versions des BDD au moment de la dernière récupération des données. A chaque nouvelle mise à jour des données de KinoMine, cette table est aussi mise à jour avec la version de chaque BDD.

La longueur de chaque variable a été choisie pour éviter des conflits ou autres problèmes pendant l'import des données. Les types spécifiques de RDKit (ROMol, BFP) ont aussi été ajoutés à la structure de la BDD. Pour gagner en simplicité, seule l'empreinte moléculaire mfp2 est utilisée dans la recherche par similarité entre molécules (parmi mfp2, ffp2 et torsionbv) et des index ont été ajoutés à la BDD. En cas de nécessité, les autres empreintes pourront être facilement ajoutées, puisqu'elles sont déjà calculées. L'IFP est utilisé pour les recherches par similarité d'interaction entre ligand et protéine.

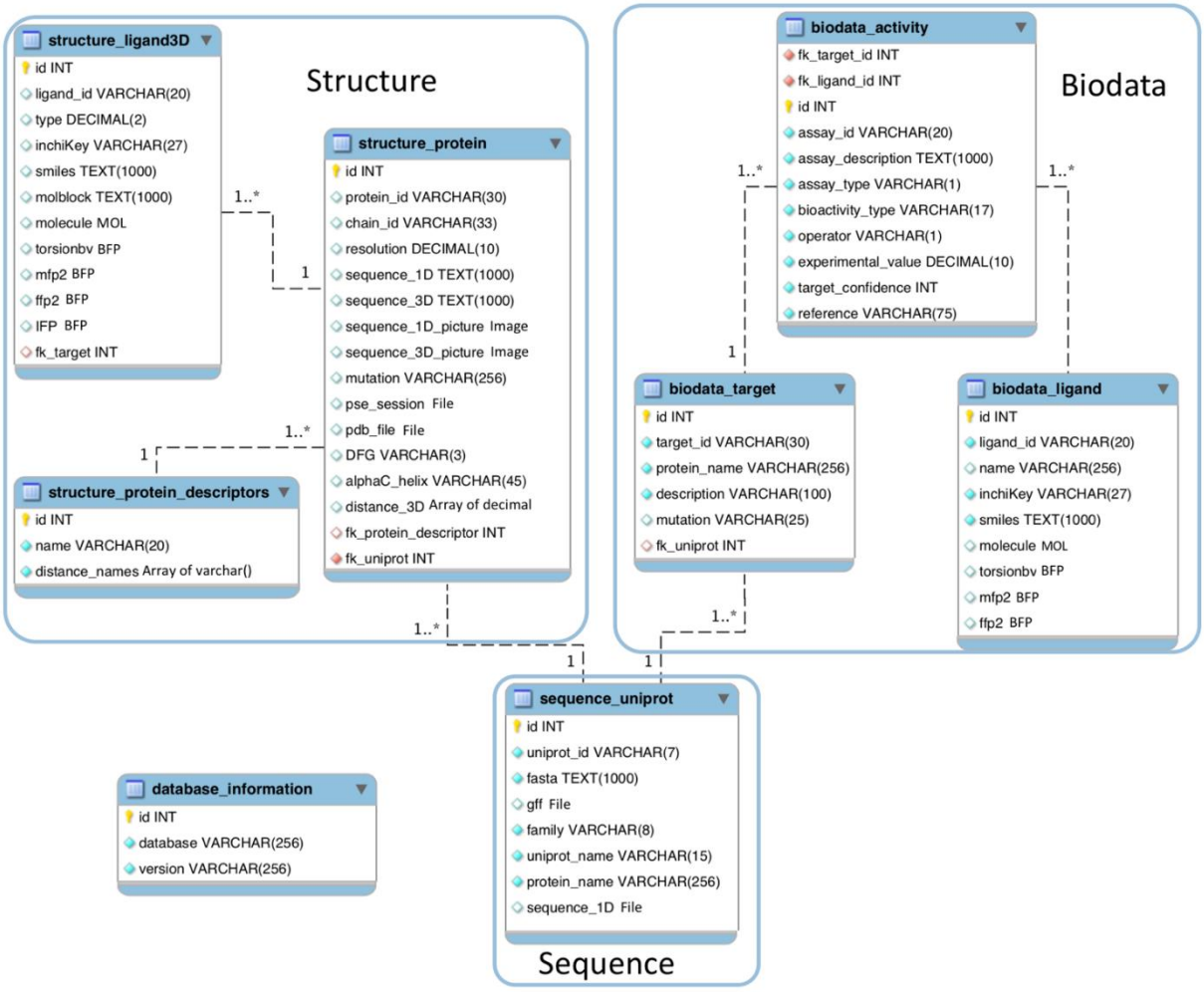


Figure 44 : Schéma de la structure de la BDD de KinoMine

3.5.3 Article de présentation de KinoMine

Le travail que j'ai effectué en stage de master 2 et continué pendant la suite de ma thèse, m'a permis de soumettre un article « *Application notes* » dans le journal *Bioinformatics* dans l'*Issue Databases and ontologies*. Après avoir fait tester le site par différents chercheurs, KinoMine correspond aux attentes des bioinformaticiens, chémoinformaticiens et chimistes médicinaux susceptibles de l'utiliser.

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Application Notes

Databases and ontologies

KinoMine: A new platform for data mining and data visualization in kinase research

Gautier Peyrat[1], Fabrice Carles[1], Christophe Meyer[2] and Pascal Bonnet[1,*]

[1]UMR CNRS-Université d'Orléans 7311, Institut de Chimie Organique et Analytique, Orléans, 45067, France, [2]Janssen-Cilag, Centre de Recherche Pharma, Val-de-Reuil, 27106, France.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein kinases are one of the major drug targets. Recently enhanced by artificial intelligence, the discovery of new protein kinase inhibitors requires data with strong veracity and curation.

Results: KinoMine, a web server and database to search and extract relevant chemical and biological kinase knowledge, such as selectivity/polypharmacology profile or 3D binding mode characteristics, both on wild type or mutated kinase. Through interactive interfaces, KinoMine provides the most curated kinase data in a single interface to help researchers in decision making.

Availability: KinoMine is freely accessible at http://kinomine.icoa.fr.

Contact: pascal.bonnet@univ-orleans.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Introduction

Protein kinase family contains 518 proteins in human (Manning et al., 2002), those proteins are involved in many cancer (Fleuren et al., 2016). Thus, protein kinases are particularly studied in the field of drug discovery. Nowadays in life science, Artificial Intelligence (AI) and big data, improve the way of discovering potent and selective inhibitors. Biological activities of molecules are predicted thanks to AI algorithms which rely on descriptors (Bosc et al., 2016; Gomes et al., 2017; Qiu et al., 2016; Stepniewska-Dziubinska et al., 2018; Subramanian et al., 2017). Those descriptors encode the large amount of information collected on the interaction between molecules and their targets.

In academic community, open data improves researcher decisions, but data is often underexploited. Actually, only a limited community of experts can exploit the entire value of this data using programming skills or proprietary software. For example, target assignment in ChEMBL database (Gaulton et al., 2012; Bento et al., 2014; Gaulton et al., 2017), version 27, does not differentiate wild type and mutated protein in the API and web interface (See Supplementary Information section "Comparison to already existing databases").

In a previous work we proposed a curated, annotated and updated database of Protein Kinase Inhibitors http://www.icoa.fr/pkidb (Carles *et al.*, 2018; Bournez *et al.*, 2020). We present here KinoMine (http://kinomine.icoa.fr/), an interactive web platform, freely accessible, which extends the chemical space available in PKIDB to biological and structural spaces. KinoMine provides data visualization in user-friendly 3D interfaces accessible to all researchers (See supplementary information section "KinoMine implementation").

KinoMine webserver

KinoMine is a search engine using either biological targets, chemical IDs, names, similarity or substructure searches. Through two different kind of heatmaps, KinoMine allows either the exploration of compound selectivity and polypharmacology (Anighoro et al., 2014; Bain et al., 2007; Bosc et al., 2017; Rastelli and Pinzi, 2015) or the exploration of 3D binding mode characteristics (See supplementary information section "Search engine").

On the selectivity heatmap, each row is a ligand and each column is a target, according to the selections from the search engine. This selectivity heatmap shows for each ligand-target couple either the number of bioactivities registered or the mean of the values corresponding to one bioactivity type. To access to more details (3D interactions and biological knowledge) on one ligand-target couple, researchers can click on the heatmap.

The non-bounded interaction heatmap is based on interaction fingerprints (IFP) (Deng et al., 2004). The ATP binding site of protein kinases is composed of 85 common residues, generally conserved for all protein kinases, as defined in KLIFS (Kooistra

et al., 2016). The panel of 85 residues forming the ATP binding site is generally in contact with the ligand in the majority of 3D kinase-ligand complexes. This panel is considered as a mask covering conserved interactions between protein kinases and ligands, and is used to compute interaction fingerprints. Each IFP has the same 680-bit length composed of 85 concatenated residue fingerprints of 8-bit long each Clicking on a row of the non-bounded interaction heatmap selects the associated protein-ligand complex and brings to the visualization page to observe detailed 3D interactions, chemical and biological knowledge. (See supplementary section "Heatmaps").

KinoMine Database

Data provided in KinoMine is freely available for download, to continue the research by using other tools. The database used by KinoMine is composed of information related on human protein kinases. The database consist firstly in the data integration of:

- Protein-ligand 3D structural data from the Protein Data Bank (PDB)
- Biological knowledge from Uniprot and literature
- Bioactivity and selectivity data from ChEMBL

Then, the data is filtered, curated and annotated (see Supplementary Information section "Preparation of kinase data"). KinoMine database is composed of 519 protein sequences and annotations from Uniprot. There are 738 proteins from ChEMBL which include both wild-type and mutated proteins. 133,075 ligands have been kept during data curation, as well as 711,982 bioactivities. Keeping mutated proteins allows searching for activity measure having specific mutations, which is not possible on the original ChEMBL and RCSB PDB databases. 6,249 ligands have a 3D structure from the RCSB PDB in KinoMine and 7,166 protein-ligand complexes can be visualized.

Visualization of chemical, biological and 3D structural details of crizotinib, an inhibitor of L1196M mutated ALK

We illustrate the request "get all data associated to the protein ALK L1196M and its inhibitor crizotinib" in **Fig. 1**. There are four panels to visualize: the sequence of the protein, the 3D protein-ligand complex, the ligand and associated bioactivity measurements. Thus, it allows exploring in the same page chemical, biological knowledge and activities as well as 3D structural spaces.

The protein feature viewer (https://github.com/calipho-sib/feature-viewer), on the top of the page, shows the annotated protein kinase sequence using both Uniprot and literature kinase knowledge. Annotations can be shown or hidden and the sequence can be zoomed. 3D protein-ligand structures are displayed in a table and the first row is loaded the 3D NGL viewer (Rose and Hildebrand, 2015; Rose *et al.*, 2018). 3D biological annotations like kinase motifs, mutated residues or known natural variants can be explored for each complex by downloading freely the PyMOL Session (PSE) file. The PDB file of the 3D structure superimposed on a reference (PRKCA) can also be downloaded. The two other panels contain information from ChEMBL. Firstly, on the left, the ligand panel shows information on the 2D ligand (i.e. structure and ID). Secondly, on the bottom, all the details about bioactivity data for the given ligand and the given protein and considering mutation are shown. In this panel, KinoMine allows selecting which column to display, exporting to CSV or xlsx, searching words in table and sorting ascending or descending any column.

Several cross-referencing links can redirect to the original databases. For example, redirection can be done through Uniprot identifier, ChEMBL ID, and bibliographic references.

KinoMine provides more accurate details in a unique webserver than those already existing. Kinase inhibitors research is improved through KinoMine search engine, heatmaps and interactive data visualization.

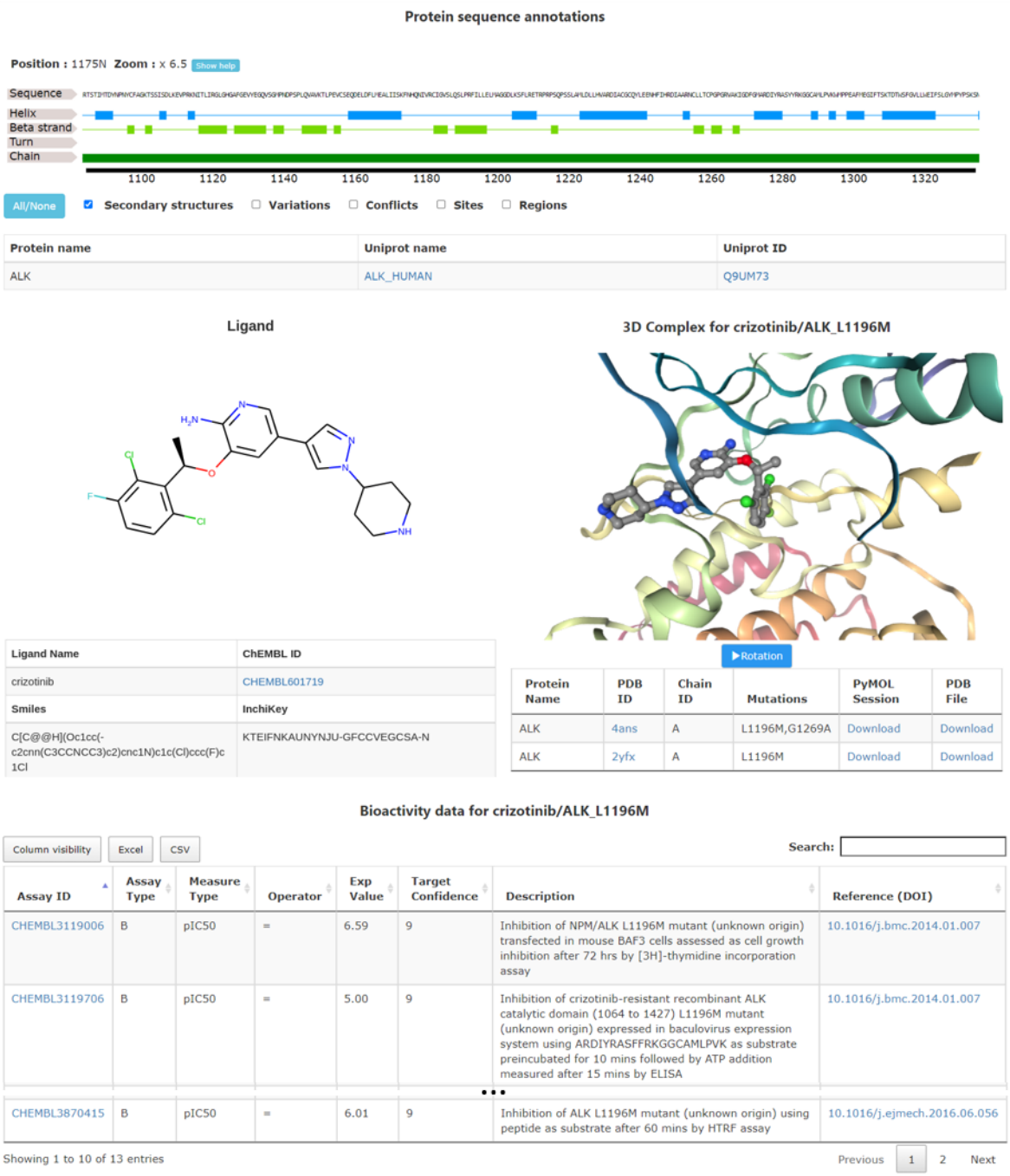


Fig. 1. Details from KinoMine between crizotinib and its target ALK with mutation L1196M. Data come from the curation of three major databases (ChEMBL, PDB and Uniprot). B means binding and F means functional for assay types. URLs allow going back to original data source. Sequence of protein and protein-ligand complexes are interactively visualized. PDB or PyMOL sessions files can also be downloaded.

Acknowledgements

The authors would like to thank Laurent Robin for his help in matter of security and maintenance during the production of the webserver. Juliette Douare and Romain Launay for their help during the development of the webserver.

Funding

This work has been supported by the …..

Conflict of Interest: none declared.

References

Anighoro,A. *et al.* (2014) Polypharmacology: Challenges and Opportunities in Drug Discovery. *J. Med. Chem.*, **57**, 7874–7887.

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

Bento,A.P. *et al.* (2014) The ChEMBL bioactivity database: an update. *Nucleic Acids Res*, **42**, D1083–D1090.

Bosc,N. *et al.* (2016) Prediction of Protein Kinase–Ligand Interactions through 2.5D Kinochemometrics. *J. Chem. Inf. Model.*

Bosc,N. *et al.* (2017) The use of novel selectivity metrics in kinase research. *BMC Bioinformatics*, **18**.

Bournez,C. *et al.* (2020) Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB. *Molecules*, **25**, 3226.

Carles,F. *et al.* (2018) PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules*, **23**, 908.

Deng,Z. *et al.* (2004) Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *Journal of Medicinal Chemistry*, **47**, 337–344.

Fleuren,E.D.G. *et al.* (2016) The kinome 'at large' in cancer. *Nature Reviews Cancer*, **16**, 83–98.

Gaulton,A. *et al.* (2012) ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, **40**, D1100-1107.

Gaulton,A. *et al.* (2017) The ChEMBL database in 2017. *Nucleic Acids Res*, **45**, D945–D954.

Gomes,J. *et al.* (2017) Atomic Convolutional Networks for Predicting Protein-Ligand Binding Affinity. *arXiv preprint arXiv:1703.10603*.

Kooistra,A.J. *et al.* (2016) KLIFS: a structural kinase-ligand interaction database. *Nucleic Acids Research*, **44**, D365–D371.

Manning,G. *et al.* (2002) The Protein Kinase Complement of the Human Genome. *Science*, **298**, 1912–1934.

Qiu,T. *et al.* (2016) The recent progress in proteochemometric modelling: focusing on target descriptors, cross-term descriptors and application scope. *Brief Bioinform*, bbw004.

Rastelli,G. and Pinzi,L. (2015) Computational polypharmacology comes of age. *Frontiers in Pharmacology*, **6**.

Rose,A.S. *et al.* (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.

Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Research*, **43**, W576–W579.

Stepniewska-Dziubinska,M.M. *et al.* (2018) Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*.

Subramanian,V. *et al.* (2017) 3D proteochemometrics: using three-dimensional information of proteins and ligands to address aspects of the selectivity of serine proteases. *MedChemComm*, **8**, 1037–1045.

Supplementary Information

KinoMine: A new platform for data mining and data visualization in kinase research

Gautier Peyrat[1], Fabrice Carles[1], Christophe Meyer[2] and Pascal Bonnet[1]

[1]UMR CNRS-Université d'Orléans 7311, Institut de Chimie Organique et Analytique, Orléans, 45067, France,

[2]Janssen-Cilag, Centre de Recherche Pharma, Val-de-Reuil, 27106, France.

Comparison to already existing databases

Several public or proprietary software applications and websites have already been developed. But, none of them provides freely data integration, curation and visualization in a unique interface. OpenPhacts (Williams *et al.*, 2012) valuably contributed to improve data interconnections and visualization between databases but it requires KNIME software platform or Pipeline Pilot, which are not straightforward or not free. Other cross-referencing such as SIFT (Velankar *et al.*, 2013; Dana *et al.*, 2019) include notably Uniprot and PDB data, but they only provide links between databases without data check. Another contribution is ChemProt database (Kringelum *et al.*, 2016) which provides a powerful bioactivity and chemical biology diseases mapping, but without structural and binding knowledge. At the opposite, KLIFS database (Kooistra *et al.*, 2016) takes up the challenge to pre-process the ligand and protein structural space of kinase and provides a very powerful database for kinase structural exploration of binding mode. It also adds bioactivity data from ChEMBL but it keeps the veracity limitations and do not provides an interactive and visual exploration of selectivity. 3D-e-Chem (McGuire *et al.*, 2017) is a similar project but requires to install complex tools through a virtual machine which is not straightforward for a non-specialized researchers. The latest ChEMBL database provides data visualization and exploration of bioactivity data thanks to interactive interfaces (Mendez *et al.*, 2019), but it does not provide 3D information and still contains inconsistencies mentioned above. The web portal KInhibition (Bello and Gujral, 2018) solved these inconsistencies and provides similar interfaces but for a limited number of kinase assay and without 2D molecular structures depiction. On the contrary, ProteinPlus server (Fährrolfes *et al.*, 2017) provides a variety of powerful structure-based molecular modelling tools but, as in KLIFS database, without data visualization of associated bioactivities data. The Protein Data Bank in Europe PDBe has started to provide visualization of biological knowledge in 3D structure of protein by linking Pfam domain annotations in 3D webGL visualizer. But unfortunately, Pfam knowledge is very limited compared to Uniprot. Some proprietary software and websites like MOE (Chemical Computing Group), 3DM (Bio-prodict) or 3decision (Discngine) already provide a part of those missing functionalities but require subscription to expensive licenses and thus cannot be democratized to the full scientific community.

KinoMine implementation

KinoMine has been developed with Python 3.7.7 thanks to Django framework (2.2 LTS). Django is an open source Python project, which allows a fast development and a secured deployment for websites. The data used in KinoMine is stored in a PostgreSQL database (12.2). This database includes six tables, mainly: three for protein kinase (containing information on 3D structures, sequence or mutations), two for ligands (3D or 2D structures) and one for associated protein-ligand bioactivities. Target, biological and chemical space are separated into different tables regarding their original sources as they contain a different knowledge. Tables are interconnected using standard target and ligand IDs.

Docker (1.13.1) images have been used to develop this website using heterogeneous technologies. Docker allows to develop services (website, database) in a computing environment similar to requirements of production servers. Two containers are required to run KinoMine, one for the website (front-end) and the other for the database (back-end). They communicate through Docker compose (1.21.0) orchestration tool.

On the back-end container, database is extended with RDKit PostgreSQL cartridge (2020.03.3) to directly store molecules with cheminformatics functionalities. It allows substructures or similarity search of a given query molecule against the whole KinoMine database. Similarity searches for molecules, as well as, binding mode patterns can be performed according to Tanimoto coefficient on the basis of binary molecular fingerprints (Bajusz *et al.*, 2015) and interaction fingerprints (Rácz *et al.*, 2018). This coefficient can rationalize and transform the molecular similarity into a value allowing to search for the most similar molecules to a given query in our database. Similarity is quantified with the equation (1), where a and b are the number of features present in compounds (or "on bits" from IFP) A and B respectively, and c is the number of features shared by A and B. The Tc lies between 0 and 1, where 1 is the biggest similarity between the two compounds.

$$\mathrm{Tc}(\mathrm{A},\mathrm{B}) = \frac{\mathrm{c}}{\mathrm{a}+\mathrm{b}-\mathrm{c}} \qquad (1)$$

On the front-end container we used RDKit (2020.03.4) for 2D ligand depiction and on the fly format conversion. Pandas data analysis library (1.0.5) is also used in this container to process data. The django-rdkit project (https://github.com/rdkit/django-rdkit) is used to facilitate communication between PostgreSQL database and Django models concerning molecular objects and substructure and similarity searches.

On the front-end, JSME molecular editor (Bienfait and Ertl, 2013) is used to draw query molecules. For bioactivity and selectivity search, when a query sends several values for the same ligand on the same target, the mean of the different values is calculated. Results are displayed on interactive heatmaps thanks to Plotly JavaScript data visualization tool (v1.37.1 https://plot.ly). On the final visualization page, the kinase-ligand complex detailed data is shown thanks to DataTables JavaScript library (https://datatables.net/), neXtProt feature viewer (v1.0.6, https://github.com/calipho-sib/feature-viewer) and NGL (Rose and Hildebrand, 2015; Rose *et al.*, 2018)

Search engine

KinoMine allows exploring both polypharmacology through the "bioactivities" tab or 3D kinase-ligand interactions through the "3D structures and binding mode" tab. In both cases an interactive heatmap shows the details of bioactivity data points or 3D knowledge between a given list of ligands and a given list of kinase targets. KinoMine provides a user-friendly search engine to execute a query or a combination of several queries from biological, chemical or structural identifiers. The detail of each choice field is shown in **Fig. S1**, clicking on one of these menus interactively displays the different input fields. This avoids the filling in of too many input fields which could enter in collision.

"Bioactivities" tab refers to bioactivity measures from ChEMBL 27. In this menu, the fields send three different requests. The first for ligand selection, the second for kinase selection and the last for measure type selection. To search ligands by similarity or substructure, a SMILES is required. A molecule can be obtained by drawing its structure or pasting a SMILES through JSME molecular sketcher. After structure edition, the molecular sketcher converts the drawing to SMILES in the correct input fields. A query containing several ligands can be selected by uploading a SDF file or pasting a list of several ligands identifiers using a case insensitive text areas (identifiers have to be separated by line feeds). "Bioactivity Type" selection proposes the

different types of bioactivity measures. Measures with the same units can be visualized together or one by one.

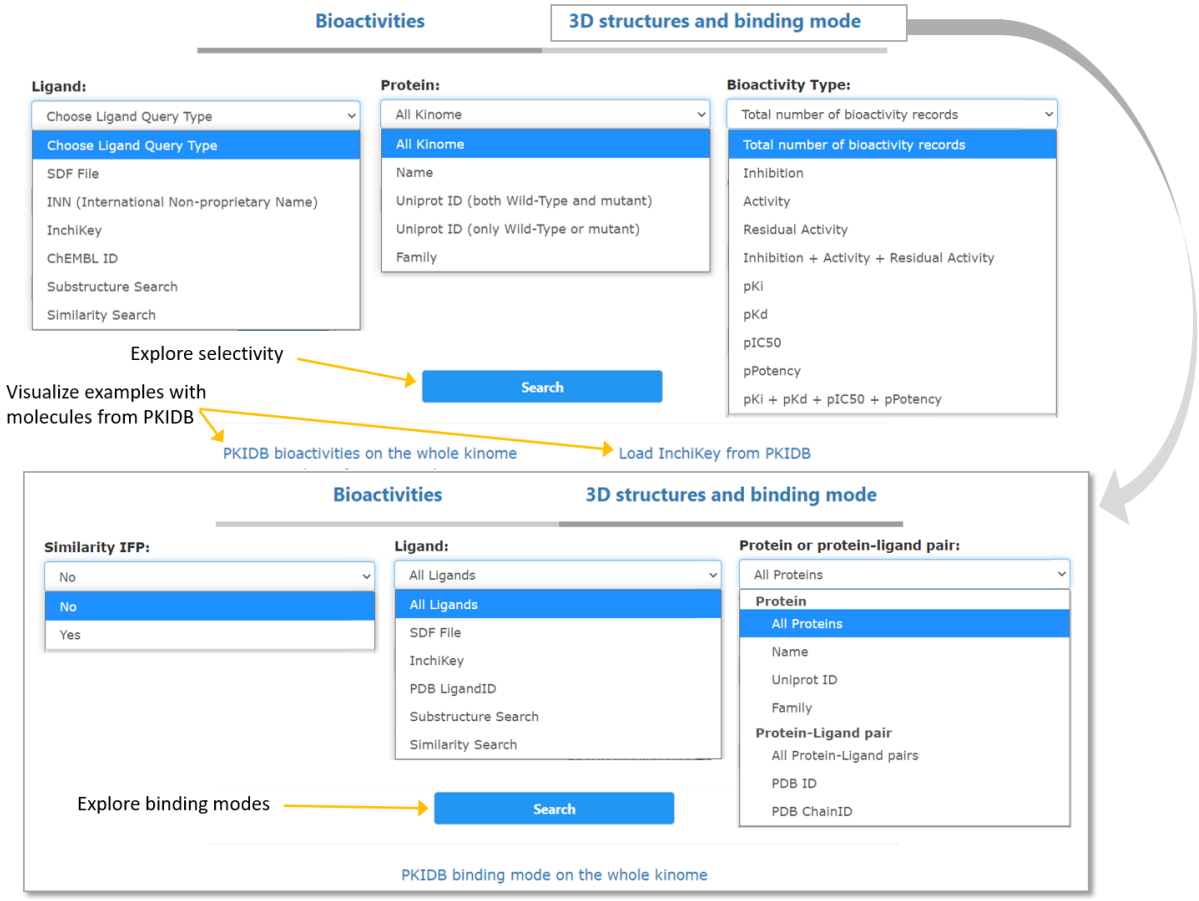


Fig. S1. KinoMine search engine. Inputs fields are interactively shown or hidden, depending on the selected input types.

"3D structures and binding mode" tab refers to the 3D structural data coming from the PDB. In this tab, three fields allow selecting ligand-protein complexes. These complexes can be found thanks to similarity of their binding modes (based on their IFP). They also can be reached using either ligands or kinases implicated in the pair. The fields combine to keep only the common ligand-protein pairs between each selected list. In the two next sections, examples of queries have been edited with a selection of ligands in the list of PKIs already approved or under development from PKIDB.

Heatmaps

For interactive selectivity exploration, a heatmap presents a visual and interactive representation of data. In contrast with the classical kinome tree allowing to visualize the selectivity of only one compound at a time, a heatmap represents all bioactivities values, for a given activity type, a given lists of ligands and a given list of targets in a single visual interface. Thus, in combination with interactive data visualization tools, this representation allows exploring selectivity and poly-pharmacological profile with the maximum amount of public knowledge. An example of request composed of ligands selected from PKIDB and activity measures (pKi, pKd, pIC_{50} and pPotency) on the Tyrosine Kinase protein family is shown **(Fig. S2 A)**.

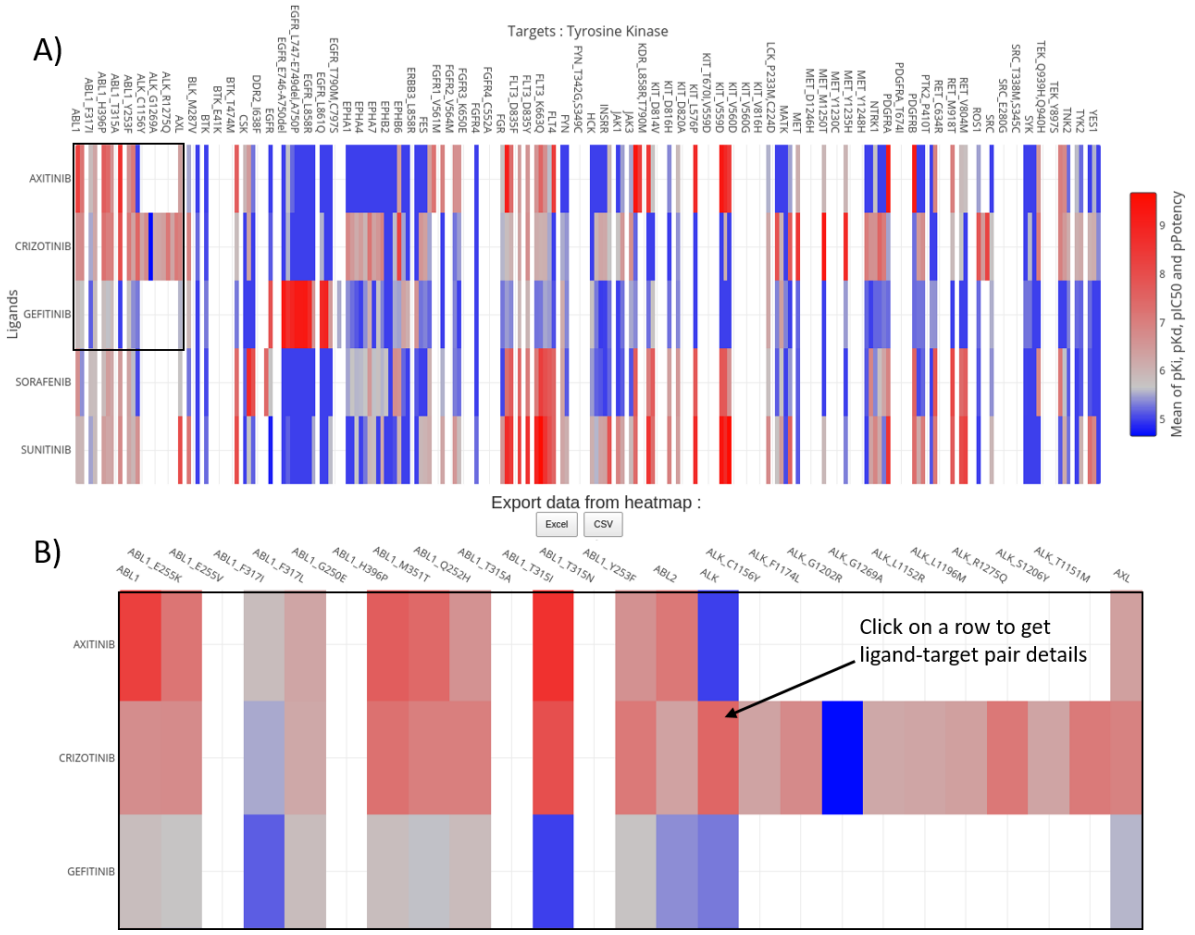


Fig. S2. A) Heatmap of interactive selectivity exploration. The query is to find all the activity measures of pKi, pKd, pIC50 and pPotency measures for a selection of ligands from PKIDB on Tyrosine Kinase family. **B) Zoom on a smaller portion.** When several bioactivity data are available for ligand-target pairs, the mean is calculated. Inactive molecules are colored in blue and active molecules are colored in red, while missing data are not colored.

When several experimental values have been measured between one ligand and a given target, average of the data points is used, and thus, could provide a more reliable data point. Final values are encoded through a color gradient in the heatmap. It helps to have a quick visual distinction between active ligands in red and inactive in blue. If no measure value is available, the white color of the background is displayed. The heatmap interactively displays the measured values. When there are too many targets to be displayed as vertical legends, only few of them are displayed. The heatmap can be zoomed on and the legends dynamically adapt showing more or less target names(**Fig. S2 B**). Hovering the heatmap shows the corresponding ligand and targets names and the measured value between them. Every heatmap values in KinoMine can be exported to Excel or CSV file for further external analysis and backup. Visualizing separately the measurements made on wild type targets and their associated mutated forms, is the major advantage of KinoMine. Mutations are indicated just after the target name separated by an underscore. Thus, it is easy to observe that the same ligand can be active on wild type and inactive on some mutation. For example, Gefitinib seems to be particularly affected by ABL1 T317I and T315I mutations (**Fig. S2 B**). Regardless of the input type of ligands, if its International Non-proprietary Name (INN) name is present in KinoMine database, it is displayed on the left of the heatmap. Otherwise, the corresponding ChEMBL ID is shown.

Interactive 3D structures and binding mode exploration uses a different heatmap (**Fig. S3**). Non-bonded interactions between the chosen set of ligands and a conserved set of binding site residues are displayed. Each protein-ligand structure has a unique ID composed of: Protein PDB ID, chain, ligand PDB ID, and finally unique number for similar ligand PDB ID in the same chains.

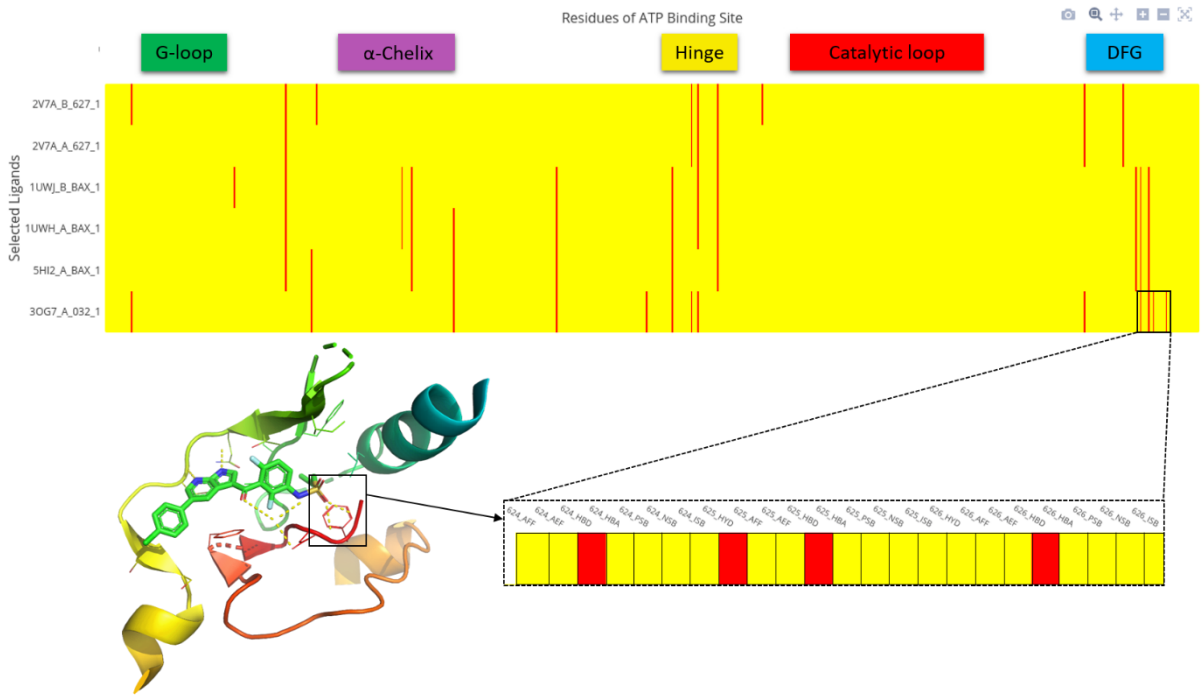


Fig. S3. Heatmap of non-bonded interactions between ligands and binding site of their kinase targets. A zoom is done on BRAF protein and its ligand vemurafenib. This zoom corresponds to the DFG motif. The binding mode is encoded with 8 interactions for each binding site residue: HYD hydrophobic contacts, AFF aromatic face to face, AEF aromatic edge to face, HBD Hydrogen bond (protein as donor), HBA Hydrogen bond (protein as acceptor), PSB Salt bridge (protein positively charged), NSB Salt bridge (protein negatively charged), ISB Salt bridge (ionic) bond with metal ion. In red the interaction is present, in yellow it is absent.

IFP can be visualized in a heatmap (**Fig S3**), two colors encode the binary information where each bit represents the presence (ON bit) or absence (OFF bit) of a non-bonded interaction. The heatmap is also interactive. Only a part of the full residue names list is shown, however the heatmap can be zoomed to observe the details of interactions encoded. Data can also be downloaded as CSV or xlsx file. Hovering the heatmap shows if the interaction is present "1" or absent "0" between the ligand and the corresponding residue.

Preparation of kinase data

The first KinoMine database release relies on ChEMBL version 27, RCSB PDB and Uniprot data collected September 23, 2020. Curation and annotations process improve the quality of injected data (**Fig. S4**).

Uniprot sequences and associated biological annotations have been collected from Uniprot database in fasta file format and General Feature Format (GFF) respectively. The starting point to define the human kinome was the original paper of manning (Manning *et al.*, 2002) updated by the pkinfam UniProt - Swiss-Prot Protein Knowledgebase at https://www.uniprot.org/docs/pkinfam.txt. Additional annotations about motif specific to kinase family (DFG, hinge, G-Loop ect …) were extracted from literature (Fabbro *et al.*, 2015) and converted to GFF format using a multiple sequence alignment of 490 kinase domain from möbitz paper (Möbitz, 2015). Then each data (e.g. sequence and annotation) were merged together using the biopython (Cock *et al.*, 2009) and bcbio-gff (Chapman, 2018) python package. Finally, all human kinase fasta sequences are concatenated and used to create a local BLAST database of the human kinome thanks to the last reimplemented and optimised BLAST+ C++ application (Camacho *et al.*, 2009).

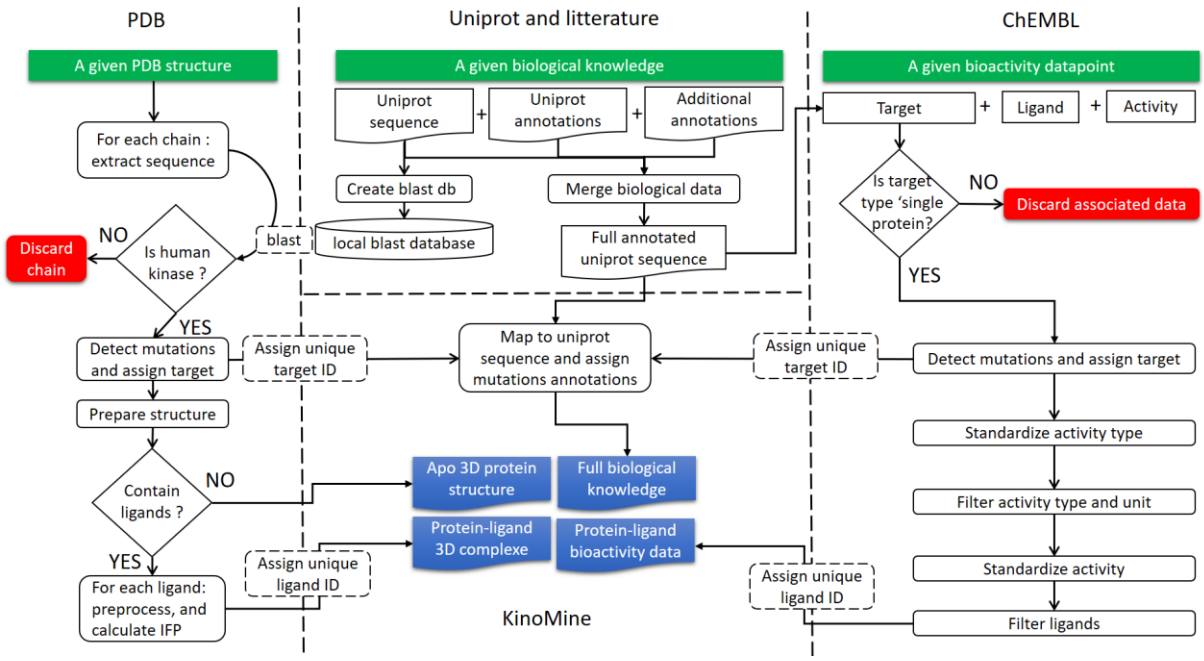


Fig. S4. Flowchart for preparation of kinase data. The process includes data ingestion of: (i) protein-ligand 3D structural data from the Protein Data Bank (PDB), (ii) biological knowledge from Uniprot and literature, and (iii) bioactivity and selectivity data from ChEMBL (green boxes). The curation and annotation process are described in white boxes. Discarded data in red and final data in blue.

For PDB curation and annotation process, similarly to PDBbind author's observations (Liu *et al.*, 2015), we developed a classification scheme completely based on the structural information given in the PDB file. In order to control all the processing steps "from scratch" we used original PDB data together with open source software. Thus, already pre-processed and cross-referencing data from, for example, KLIFS or SIFT (Kooistra *et al.*, 2016; Velankar *et al.*, 2013) were voluntary ignored to allow future processing of both public or private data without limitation. For 3D structural preparations all steps are achieved using PyMOL version 2.2.3 except for ligand preparation where OpenBabel version 2.4 through Pybel python wrapper (O'Boyle *et al.*, 2008) and RDKit version 2017.09.3 are used (Landrum). For each chain of a given PDB, fasta extraction is done using PyMOL API extended by a mapping between 635 non-canonical amino-acids three letters code and their respective standard parent one letter code, extracted from the Chemical Component Dictionary (Westbrook *et al.*, 2015). Then, a

local BLAST is used to identify protein kinase chains contained in a given PDB structure file. This process is inspired by the SIFT protocol (Velankar *et al.*, 2013) to assign the corresponding Uniprot ID of the best BLAST hit. Briefly, a BLAST sequence alignment result provides an E-value and a threshold of 0.04 allowing accurate distinction between kinase and other proteins. The SIFT protocol is also used to detect and annotate mutations. According to Velankar et al., a threshold over 85% of sequence identity is considered to ensure that any engineered mutations, tags or isoforms, result in the correct identification. Finally, after identification, biological knowledge is transferred in PyMOL 3D structure visualizer. We mapped mutations specific kinase family motifs (DFG, hinge, G-Loop etc …) (Fabbro *et al.*, 2015) as well as natural variants from Uniprot. Then, each kinase chain is superposed on the same coordinate space using one of the first PDB entry of kinase as reference: PRKCA (PDB code 1ATP). Eighty-five conserved residues of the binding site (Kooistra *et al.*, 2016) are extracted and renumbered with canonical universal index to get a uniform residue numbering for every protein kinase binding site. For each ligand, in each PDB chain, a unique ligand ID is assigned. During ligand preparation, RDKit was used to assign bond orders to the PDB ligand from the SDF file available in PDB.

Interactions involved within the binding site are computed as interaction fingerprint (IFP) (Deng *et al.*, 2004) using ODDT python package (Wójcikowski *et al.*, 2015). Eight interaction types were collected:

(1) Hydrophobic contacts
(2) Aromatic face to face
(3) Aromatic edge to face
(4) Hydrogen bond (protein as donor)
(5) Hydrogen bond (protein as acceptor)
(6) Salt bridge (protein positively charged)
(7) Salt bridge (protein negatively charged)
(8) Salt bridge (ionic) bond with metal ion

Each residue fingerprint is 8-bit long and the total IFP is finally composed of 85 residue fingerprints concatenated. Thus, each IFP has the same length (680 bits). Finally, this 3D knowledge is saved in KinoMine database, including PyMOL session files (pse) with biological annotations.

Curation and pre-processing of ChEMBL data related to kinase are required for a more comprehensive analysis of the selectivity of protein kinase inhibitors. Since the ChEMBL Kinase SARfari subset is no longer supported (Mendez *et al.*, 2018), we started data ingestion of kinase data directly from an SQLite ChEMBL data dump. This allows to quickly extract all kinase bioactivities using the list of kinase Uniprot ID collected in the previous step. Collected targets are filtered with ‘single protein’ target type. Kinase mutations are extracted from assay table description and a unique target id is created by concatenation of Uniprot id and a standardized mutation string. Thus, wild type ABL1 target id is now ‘P00519’ while ABL1 target id with an isoleucine substitution in position 315 is ‘P00519_T315I’.

After this step, we had standardized activity type in the STANDARD_TYPE of the Activities table since it could contain many non-standard activity type in original ChEMBL database (e.g. ‘Residual Activity’ vs ‘Residual activity’). Then we selected the main activity type 'IC50', 'Ki', 'Kd', 'Potency', 'Inhibition', 'Activity' and 'Residual Activity' and for each activity type, we

extracted and applied a specific data transformation of activity value from the ‘STANDARD_VALUE’ of the ‘ACTIVITIES’ table. Those steps included (i) selection of acceptable normal range [-51; 150] % or [0; 1.10^9] nM for 'Residual Activity', 'Inhibition', 'Activity' and 'IC50', 'Ki', 'Kd', 'Potency' respectively. (ii) Rescaling of activity values in standard range [0; 100] % for 'Residual Activity', 'Inhibition' and 'Activity' or transformation of activity value to the logarithmic scale for 'IC50', 'Ki', 'Kd' and 'Potency' (e.g. pIC50, pKd, etc.). Moreover, ‘Residual Activity' values were fitted to the same distribution than the one of 'Inhibition' and ‘Activity’ by transforming each ‘Residual Activity’ value by its opposite (e.g. 100 – x). All these steps allow a comprehensive selectivity exploration of kinase inhibitors.

The final processing of ChEMBL data included the filtering of several ligands without structures or having molecular weight outside the range of [150; 800] Dalton, thus not considered as small molecule. In the standardization process, some activity records were also removed due to data uncertainty, e.g. data outside normal range -51, +150 or 0, 1.10^9 nM. All pre-processed data (protein-ligand 3D complex, biological knowledge and bioactivity data) are finally aggregated in the KinoMine database. Final bioactivity data consist of 711,982 data points.

References

Bajusz,D. *et al.* (2015) Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, **7**, 20.

Bello,T. and Gujral,T.S. (2018) KInhibition: A Kinase Inhibitor Selection Portal. *iScience*, **8**, 49–53.

Bienfait,B. and Ertl,P. (2013) JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics*, **5**, 24.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Chapman,B. (2018) Incubator for useful bioinformatics code, primarily in Python and R: chapmanb/bcbb.

Cock,P.J.A. *et al.* (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.

Dana,J.M. *et al.* (2019) SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. *Nucleic Acids Res*, **47**, D482–D489.

Deng,Z. *et al.* (2004) Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions. *Journal of Medicinal Chemistry*, **47**, 337–344.

Fabbro,D. *et al.* (2015) Ten things you should know about protein kinases: IUPHAR Review 14: Ten things you should know about protein kinases. *British Journal of Pharmacology*, **172**, 2675–2700.

Fährrolfes,R. *et al.* (2017) ProteinsPlus: a web portal for structure analysis of macromolecules. *Nucleic Acids Res*, **45**, W337–W343.

Kooistra,A.J. *et al.* (2016) KLIFS: a structural kinase-ligand interaction database. *Nucleic Acids Research*, **44**, D365–D371.

Kringelum,J. *et al.* (2016) ChemProt-3.0: a global chemical biology diseases mapping. *Database (Oxford)*, **2016**.

Landrum,G. RDKit: cheminformatics and machine learning software. http://www.rdkit.org.

Liu,Z. *et al.* (2015) PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, **31**, 405–412.

Manning,G. *et al.* (2002) The Protein Kinase Complement of the Human Genome. *Science*, **298**, 1912–1934.
McGuire,R. *et al.* (2017) 3D-e-Chem-VM: Structural cheminformatics research infrastructure in a freely available Virtual Machine. *J. Chem. Inf. Model.*
Mendez,D. *et al.* (2019) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*, **47**, D930–D940.
Mendez,D. *et al.* (2018) ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res*.
Möbitz,H. (2015) The ABC of protein kinase conformations. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, **1854**, 1555–1566.
O'Boyle,N.M. *et al.* (2008) Pybel: a Python wrapper for the OpenBabel cheminformatics toolkit. *Chemistry Central Journal*, **2**, 5.
Rácz,A. *et al.* (2018) Life beyond the Tanimoto coefficient: similarity measures for interaction fingerprints. *J Cheminform*, **10**, 48.
Rose,A.S. *et al.* (2018) NGL viewer: web-based molecular graphics for large complexes. *Bioinformatics*, **34**, 3755–3758.
Rose,A.S. and Hildebrand,P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Research*, **43**, W576–W579.
Velankar,S. *et al.* (2013) SIFTS: Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res*, **41**, D483–D489.
Westbrook,J.D. *et al.* (2015) The chemical component dictionary: complete descriptions of constituent molecules in experimentally determined 3D macromolecules in the Protein Data Bank. *Bioinformatics*, **31**, 1274–1278.
Williams,A.J. *et al.* (2012) Open PHACTS: semantic interoperability for drug discovery. *Drug Discov. Today*, **17**, 1188–1198.
Wójcikowski,M. *et al.* (2015) Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. *J Cheminform*, **7**.

3.5.4 Améliorations apportées

KinoMine propose maintenant une interactivité avec l'utilisateur apportée par les différents outils implémentés par Romain Launay, que j'ai pu encadrer pendant son stage de master 2.

Dans la page de visualisation des résultats, une image « statique » était affichée pour présenter les annotations des séquences des protéines. Romain Launay a remplacé cet affichage par un module interactif nommé *nextProt Feature Viewer* (https://github.com/calipho-sib/feature-viewer). Cela permet d'afficher plus ou moins d'annotations et de zoomer sur une partie précise de la séquence.

Dans cette même page, Romain Launay a aussi intégré le module NGL viewer (https://github.com/nglviewer/ngl)[287,288] permettant d'afficher, de zoomer et de déplacer la structure 3D du complexe protéine-ligand directement depuis le site web de KinoMine.

Romain Launay a aussi apporté diverses améliorations graphiques, comme par exemple l'affichage des aides pour la saisie des identifiants de ligands ou protéines sous formes de bulles d'informations (Figure 45).

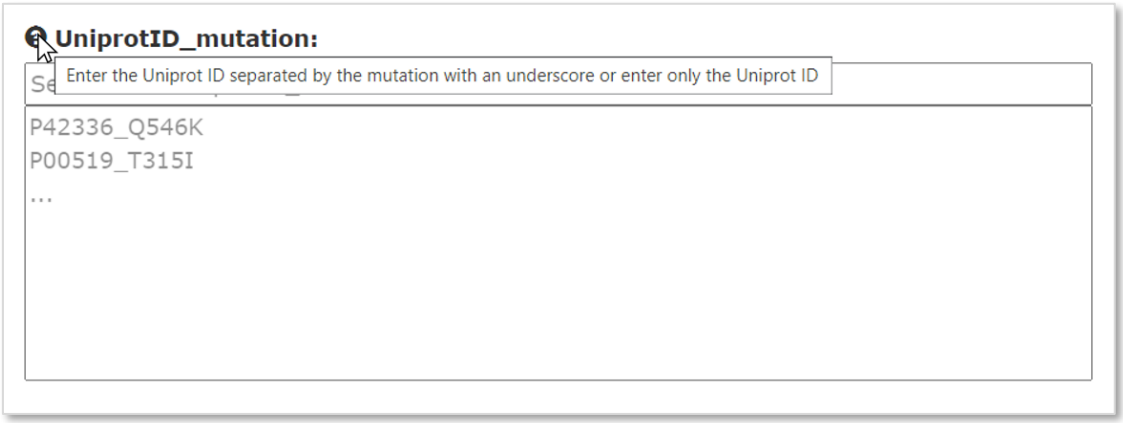


Figure 45 : Exemple d'une bulle d'information affichée dans KinoMine.

Lorsque j'ai proposé à de potentiels utilisateurs de KinoMine de tester l'interface, il m'a été suggéré d'ajouter une fonctionnalité pour pré-afficher les identifiants correspondant à la saisie de l'utilisateur. Cela permet de s'assurer qu'il n'y a pas de faute de frappe et que l'identifiant est bien présent dans la BDD (Figure 46). Pour ajouter cette fonctionnalité, nommée « auto complétion », j'ai utilisé le module *autocomplete* de la librairie JavaScript jQuery. J'ai aussi utilisé le langage de balisage extensible asynchrone AJAX (pour *asynchronous JavaScript and* eXtensible Markup Language, XML) ainsi que des vues Django. Les recherches d'identifiants dans la BDD sont alors effectuées à chaque nouvelle saisie de l'utilisateur sans recharger toute la page web.

Figure 46 : Exemple de pré-affichage d'identifiants lors de la saisie dans KinoMine.

3.5.5 Améliorations possibles

Le défaut principal de KinoMine reste la mise à jour des données. Bien que celle-ci soit semi-automatisée à l'aide de *notebooks* Jupyter, cette mise à jour ne se fait pas encore automatiquement. Une première amélioration de KinoMine serait de créer des routines permettant d'intégrer les nouvelles données automatiquement lorsque celles-ci sont déposées dans les BDD UniProt, ChEMBL ou RCSB PDB.

Une autre amélioration serait d'ajouter le profil d'activité des molécules sous la forme d'un histogramme en barres sur les différentes cibles sélectionnées. Cela donnerait un niveau de détail supplémentaire par rapport aux cartes de chaleur, qui sont plus adaptées pour visualiser l'activité de plusieurs molécules sur plusieurs cibles.

De plus, certaines améliorations apportées aux autres outils de la plateforme pourraient aussi être intégrées à KinoMine, comme l'affichage d'erreurs sous la forme de bannières (Figure 37). Les mêmes vérifications que celles faites sur MolDesc pourraient aussi être ajoutées à KinoMine concernant l'import de fichiers SDF.

3.5.6 Conclusion

KinoMine a été créé dans le cadre de la conception de nouveaux médicaments. Cet outil propose une interface agréable à utiliser pour parcourir les données et les connaissances acquises sur le kinome humain. Cela permet de savoir comment concevoir un nouvel inhibiteur en observant les interactions que forment les inhibiteurs déjà connus avec leurs cibles. KinoMine permet aussi de savoir si des cibles thérapeutiques sont déjà la cible d'inhibiteurs ou non. Nous allons maintenant détailler la conception et les améliorations qui ont été apportées au site web de F2D.

3.6 Frags2Drugs

Nous avons déjà présenté le programme Frags2Drugs (F2D) dans le chapitre 2 de cette thèse et dans l'article présenté en partie 2.1.2. Nous nous concentrerons ici sur l'implémentation de son interface web que j'appellerai « F2D web ». Comme pour KinoMine, F2D web a été développé à l'aide de Django selon l'organisation modèle-gabarit-vue (model-template-view, MTV). Au cours du développement et du déploiement de F2D web, j'ai encadré Juliette Douare (25/03/2019 – 02/08/2019) et Romain Launay (09/03/2020 – 02/09/2020) qui ont rendu ce site simple d'utilisation, fonctionnel et accessible.

La page d'accueil de F2D web contient l'historique des calculs déjà effectué sur ce site. Cet historique permet d'éviter d'effectuer plusieurs fois un même calcul. Si la requête a déjà été effectuée par un autre utilisateur, alors la page de résultats sera directement chargée. Le tableau affichant l'historique est présenté grâce au module jQuery « DataTables » (Figure 47).

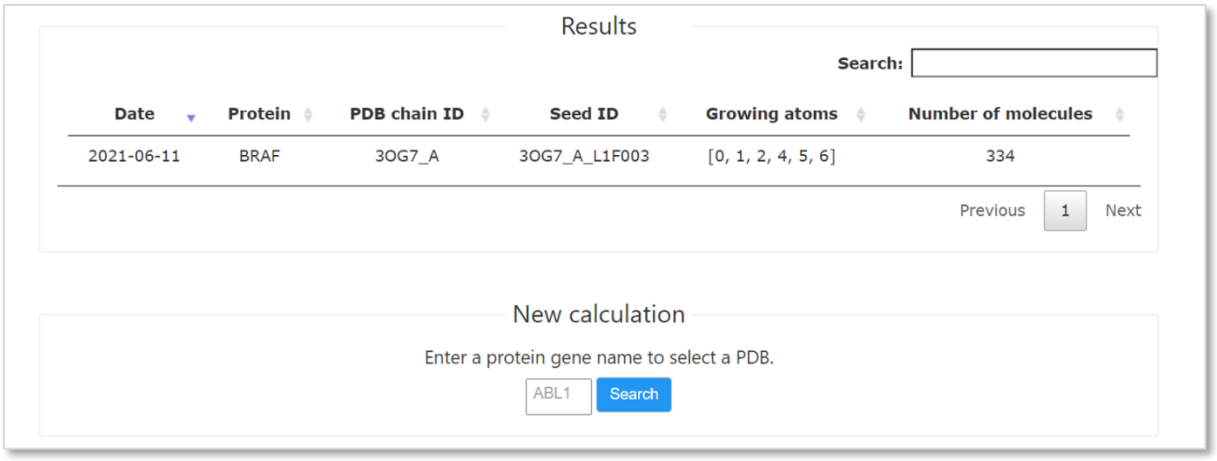


Figure 47 : Page d'accueil de F2D web.

La page d'accueil permet aussi d'effectuer une nouvelle recherche d'inhibiteurs de protéines kinases (*protein kinase inhibitors*, PKI), en saisissant d'abord le nom du gène codant pour la protéine cible. La saisie d'un premier caractère entraînera la proposition de tous les noms contenant ce caractère par auto complétion, comme dans les champs de KinoMine (Figure 46). Après avoir sélectionné la protéine cible, il faut choisir la structure 3D à partir des identifiants RCSB PDB proposés (Figure 48). En sélectionnant une des lignes de ce tableau, les fragments disponibles pour exécuter F2D sont présentés. Il faut alors cliquer sur un de ces fragments pour accéder à la page de lancement de F2D.

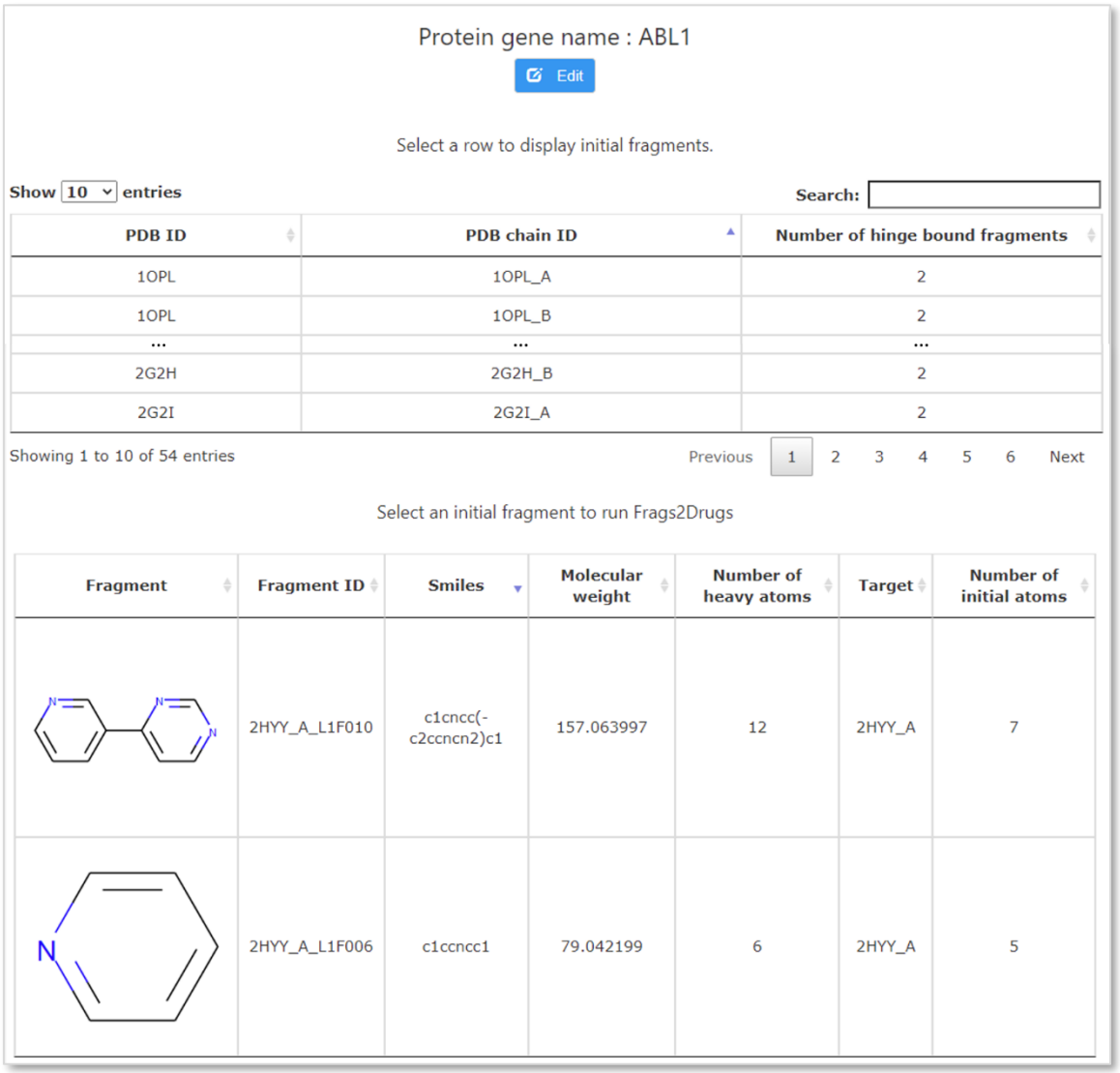


Figure 48 : Sélection de l'identifiant RCSB PDB et du fragment de départ de F2D web.

La page de lancement de F2D rappelle d'abord des informations (identifiants, SMILES, MW, NHA) sur le complexe protéine-fragment choisi (Figure 49). Les différents atomes de départ pour l'agrandissement sont montrés en 2D, à côté de la structure 3D de l'emplacement du fragment dans la protéine. L'utilisateur peut alors choisir précisément à partir de quels atomes lancer F2D.

La page de résultats s'affiche alors en deux temps. D'abord, pendant que F2D effectue son calcul, un message indique que le calcul est en train de s'effectuer. Une fois le calcul terminé, toutes les molécules obtenues s'afficheront et pourront être téléchargées, comme montré dans l'article présentant F2D (*cf* chapitre 2).

L'adresse URL de la page de résultats contient tous les éléments nécessaires au lancement de F2D : cible, fragment et atomes de départ. Ainsi, en la conservant, l'utilisateur peut vérifier plus tard si son calcul est terminé et partager ses résultats une fois ceux-ci obtenus.

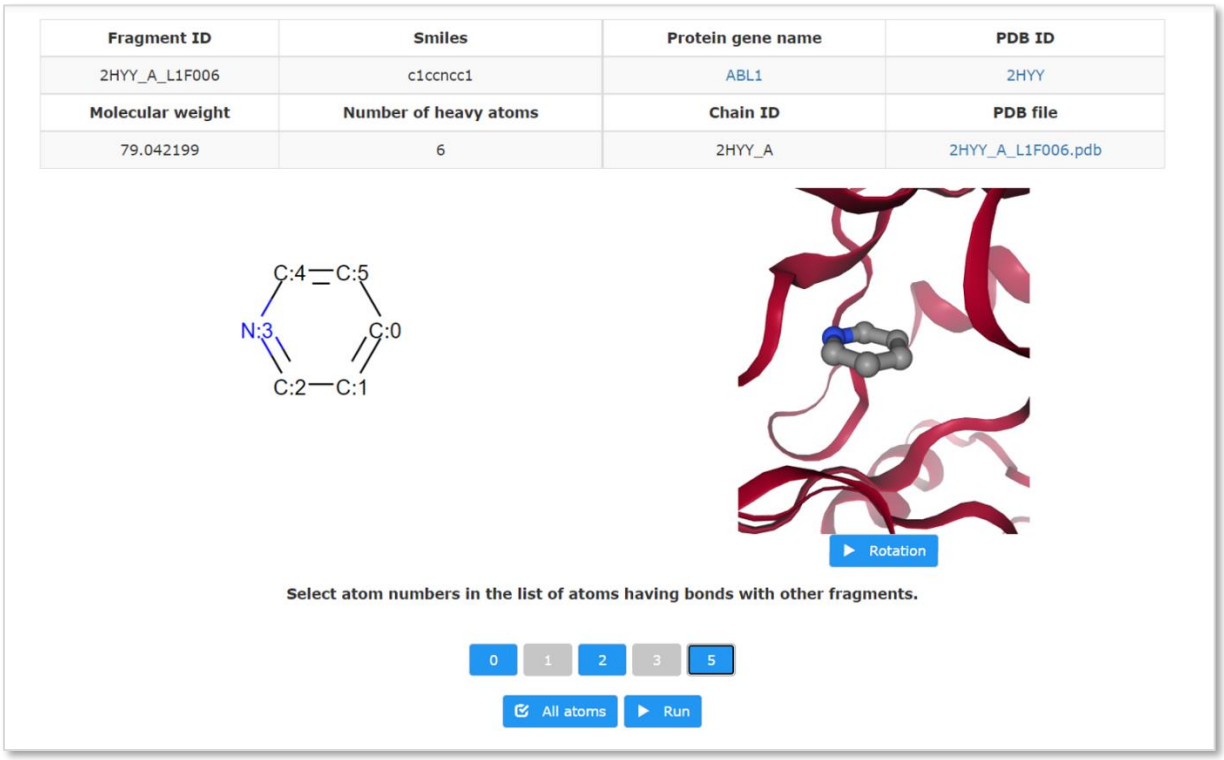


Figure 49 : Sélection des atomes de départ de F2D web.

De manière uniforme avec les autres outils de la plateforme SB&C, F2D web bénéficie des mêmes en-têtes et pieds de page, ainsi que des pages « guide de l'utilisateur » et « à propos ». F2D web contient plus de technologies que les autres outils de la plateforme pour son fonctionnement. En effet, il repose sur différentes bases de données (BDD) relationnelles et orientées graphes et sur Celery pour le lancement des calculs d'agrandissement de fragments.

3.6.1 Bases de données

Le fonctionnement de F2D repose sur une BDD orientée graphes, or Django ne permet pas d'intégrer directement ce type de BDD *via* ses modèles. C'est pourquoi l'architecture de F2D web contient deux BDD : une orientée graphe et l'autre relationnelle. La BDD relationnelle contient deux tables présentées en Figure 50, la première concerne les protéines et la seconde les fragments. La table des protéines contient 3 identifiants pour accéder aux protéines cibles, le quatrième champ nommé « id » est présent dans toute les tables et est nécessaire au bon fonctionnement de Django. La table des fragments ne concerne que les graines de départ pour le programme F2D. Tous les autres fragments pouvant être inclus dans une molécule lors de l'agrandissement sont contenues dans la BDD orientée graphes. Dans la BDD relationnelle, une protéine peut être reliée à plusieurs fragments. Les fragments sont inclus dans la BDD à l'aide de différents identifiants les rendant uniques (seed_id). D'autres identifiants qui leurs sont associés permettent d'effectuer la correspondance avec les fragments de la BDD orientée graphes (fragment_id). Les champs des fragments dans la BDD relationnelle contiennent aussi des représentations de la molécule (2D et 3D) et quelques descripteurs.

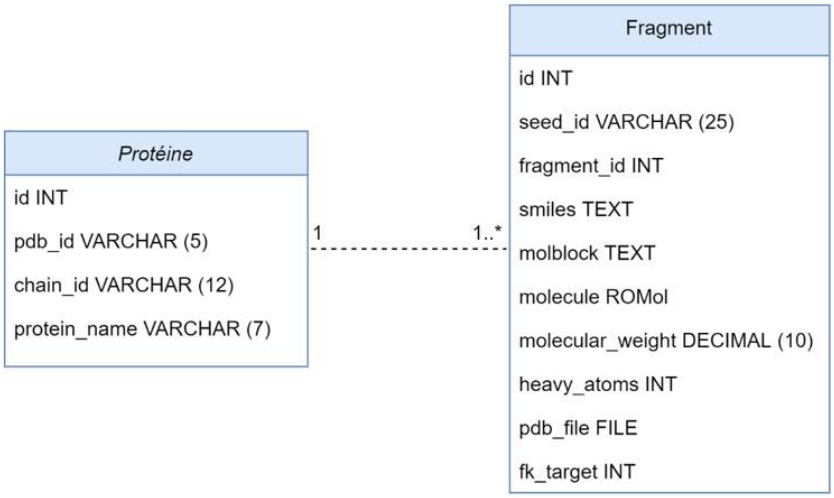


Figure 50 : Schéma de la structure de la BDD relationnelle de Frags2Drugs web.

La seconde BDD est orientée graphes, elle a déjà été présentée dans le chapitre 2, en partie 2.1.2. Elle contient tous les éléments pour le bon fonctionnement de F2D. L'architecture de F2D web utilise les modèles de Django pour communiquer avec la BDD relationnelle, et une communication directe avec la BDD orientée graphes depuis les vues Django (Figure 51).

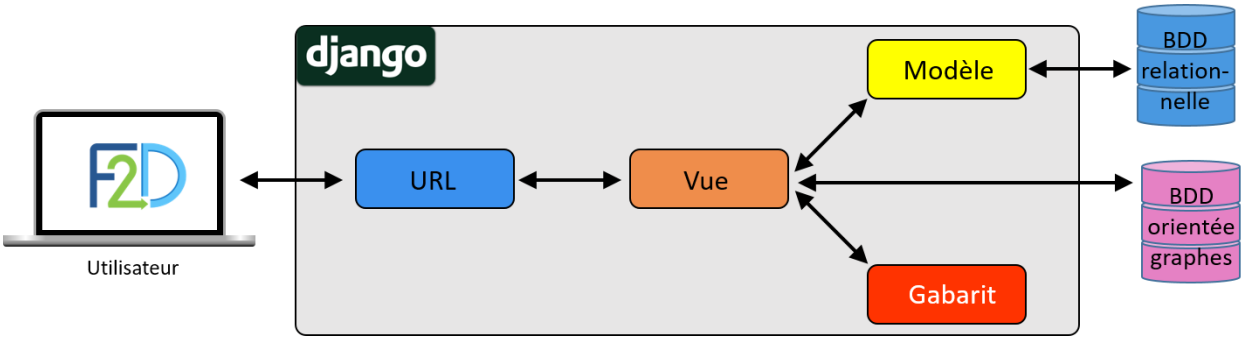


Figure 51 : Schéma de l'architecture MTV de F2D avec deux BDD.

Les calculs effectués sur F2D web pouvant durer plusieurs heures pour certains exemples, il a été nécessaire d'intégrer Celery à l'environnement pour permettre au site de fonctionner sans se retrouver bloqué dans l'attente des résultats d'un calcul.

3.6.2 Celery

Celery (https://docs.celeryproject.org/en/stable/) est une librairie Python permettant d'exécuter des tâches asynchrones ou des tâches récurrentes. Lors de l'exécution asynchrone, le calcul pourra tourner sans bloquer le site web. Les tâches sont placées dans une file d'attente pour que leur exécution soit lancée au moment opportun.

Les tâches asynchrones de Celery fonctionnent à l'aide de messages distribués entre trois éléments : le client, l'agent (*broker*) et le travailleur (*worker*). Le client (site web Django) envoie un

message dans la file d'attente, l'agent transmet alors ce message au travailleur pour l'exécution du calcul (Figure 52). Il peut y avoir plusieurs agents et travailleurs pour effectuer plusieurs calculs à la fois. Redis (https://redis.io/) est l'agent que nous avons choisi d'utiliser pour sa documentation plus abondante que les autres outils similaires. Redis et Celery sont des programmes libres de droits, sous licence BSD (pour *Berkeley Software Distribution*), permettant une utilisation du code sans restriction.

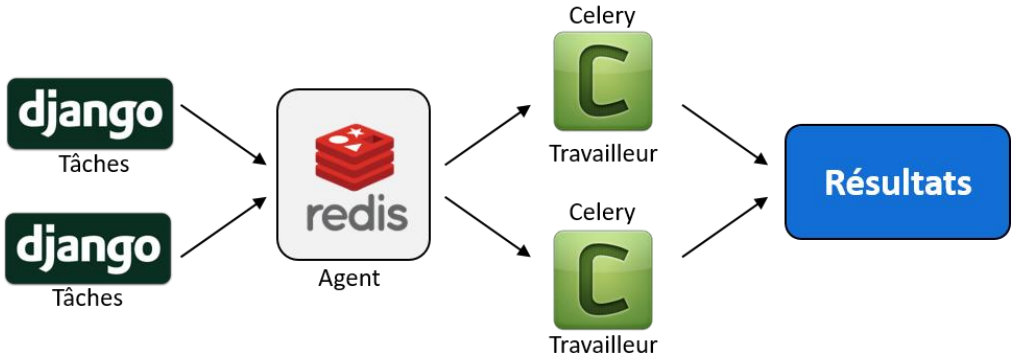


Figure 52 : Exécution des tâches asynchrones par Celery. Les flèches noires représentent les messages distribués lors de l'exécution.

Celery permet aussi d'exécuter des tâches récurrentes, grâce à Celery-Beat. Ces tâches ne sont pas de longs calculs, mais des tâches de maintenance, comme par exemple la suppression mensuelle des fichiers de résultats stockés par F2D sur le serveur web. Le fonctionnement de ces tâches passe aussi par l'agent Redis qui redistribuera les messages aux travailleurs Celery pour supprimer les fichiers une fois par mois.

La mise en place de ce système par Romain Launay a entraîné un changement dans l'architecture Docker de F2D web. En effet, l'ajout de Celery amène trois nouveaux conteneurs : Celery, Celery-Beat et Redis. Ainsi, F2D web fonctionne à partir de cinq conteneurs Docker, les 3 précédemment cités plus ceux de Django et de la BDD relationnelle.

Grâce à l'intervention de Juliette Douare et Romain Launay, F2D web est un site fonctionnel mis en production, qui peut cependant encore être amélioré.

3.6.3 Améliorations possibles

Comme pour KinoMine, une première amélioration de F2D web serait d'automatiser la mise à jour des deux BDD, relationnelle et orientée graphe. Cette mise à jour pourrait être effectuée par une tâche récurrente lancée à l'aide de Celery.

Une seconde amélioration serait un meilleur suivi de l'avancée de l'exécution des calculs. Actuellement la page de résultats affiche une page d'attente jusqu'à l'obtention du résultat. Cependant une barre de progression serait plus appropriée pour que l'utilisateur puisse suivre l'avancée des calculs. De plus, l'envoi d'un courriel à l'obtention des résultats pourrait éviter à l'utilisateur de devoir retourner sur la page des résultats pour vérifier si le calcul s'est terminé.

Une autre amélioration pour F2D web serait de proposer à l'utilisateur de charger sa propre structure de protéine cible ou son propre fragment de départ. Cela nécessiterait d'effectuer des vérifications sur la protéine et le ligand pour les standardiser. Ensuite, la protéine pourrait être superposée dans le référentiel commun, puis le fragment de départ positionné dans le site actif par amarrage moléculaire. Il s'agirait ici d'une mise à jour conséquente avec l'intégration d'outils de standardisation, de superposition et d'amarrage moléculaire.

Après avoir présenté tous les outils de la plateforme SB&C sur lesquels j'ai travaillé, je vais clore ce chapitre par une conclusion et une mise en perspective pour l'ensemble de la plateforme.

3.7 Conclusion et perspectives

La plateforme d'outils SB&C qui a été présentée dans ce chapitre permet aux chercheurs d'utiliser facilement les outils développés par notre équipe de recherche sans avoir besoin de maîtriser un langage de programmation au préalable ou d'effectuer une quelconque installation. Au 14 juin 2021, 6 outils sont accessibles sur les 7 constituant la plateforme.

La mise en place de cette plateforme m'a permis d'acquérir des connaissances en développement web et aussi de mise en production. J'ai ainsi pu travailler avec les langages ou librairies : Python, HTML, CSS, JavaScript, jQuery, Django, PostgreSQL, Docker, Docker-compose ou Celery. J'ai aussi utilisé des modules JavaScript tels que DataTables, Plotly, *nextProt Feature Viewer* ou *NGL.*

J'ai aussi eu l'opportunité d'encadrer trois stagiaires à qui j'ai pu apprendre à utiliser ces différentes technologies. Ils m'ont aussi appris à utiliser certaines de ces technologies, comme par exemple Celery, et apporté des connaissances supplémentaires en programmation web.

L'ensemble de la plateforme pourrait être amélioré avec le passage sur un site sécurisé en basculant du protocole HTTP vers HTTPS. Pour cela il faudrait ajouter des certificats électroniques à chaque adresse de la plateforme. Ce certificat est une sorte de carte d'identité du site web et est nommé TLS (pour *Transport Layer Security*).

La plateforme web pourrait aussi être améliorée par l'ajout d'un système permettant de mieux suivre l'utilisation des sites. Pour cela, des messages pourraient être paramétrés et récupérés via une journalisation et une analyse de l'historique des évènements.

Enfin, pour gagner en clarté sur l'utilisation des sites, des guides vidéos pourraient être ajoutés à chaque guide de l'utilisateur. Ils ont le mérite d'être plus clairs et plus efficients pour apprendre à chaque utilisateur le fonctionnement des outils.

Après avoir décrit la plateforme web SB&C, nous allons maintenant voir, dans le chapitre suivant, comment obtenir des PKI à l'aide de différents programmes *in silico,* dont principalement Frags2Drugs (F2D).

Chapitre 4 : Découverte de nouvelles molécules bioactives à l'aide d'outils *in silico*

4.1 Introduction

Ce chapitre concerne l'application d'outils *in silico* à la découverte de nouvelles molécules bioactives effectuée en collaboration avec quatre équipes de recherche. Je présenterai brièvement ces équipes au cours du chapitre, mais en introduction, je parlerai de leurs caractéristiques communes.

Ces équipes de recherche sont toutes composées de chimistes médicinaux et ont des problématiques similaires. Nous souhaitons proposer aux chercheurs un guide, une aide à la priorisation dans la synthèse de molécules inhibitrices. Dans ces quatre projets, des fragments initiaux dont les chercheurs maîtrisent la synthèse, appelés échafaudages, seront utilisés comme point de départ pour l'agrandissement. Dans la suite de ma thèse nous nommerons cet échafaudage ou fragment initial : « graine ». Selon les projets, la structure de la graine peut être confidentielle ou non.

Mon but est de concevoir *in silico* des molécules à partir des graines proposées par les chimistes pour inhiber leurs cibles. Je vais présenter dans ce chapitre l'application du programme d'agrandissement de fragments Frags2Drugs (F2D, présenté dans le Chapitre 2) associé à d'autres outils *in silico* pour l'obtention d'inhibiteurs de protéines kinases (PKI) à synthétiser dans les différents projets.

La validation de la reconstruction des PKI fragmentés pour créer la base de données (BDD) de F2D (Chapitre 2, partie 2.1.2) se base uniquement sur les fragments issus de chaque ligand à reconstruire. Les paramètres (distance, angles de liaison, hors plan et dièdres) utilisés pour la reconstruction ont reçu des valeurs élevées et cette forte tolérance permet la reconstruction de 98,91% des molécules du jeu de données de validation (6 799 validées sur 6 874). Cependant, lorsque tous les nœuds de la BDD orientée graphes de F2D sont pris en compte, ces paramètres génèrent trop de possibilités de molécules pour pouvoir être exploitées.

Nous avons alors cherché quelles valeurs permettraient au moins 90% de reconstruction du jeu de validation, tout en étant assez faibles pour générer les molécules les plus correctes possible. Ainsi, au sein de tous les projets d'application de F2D, nous utiliserons les valeurs seuils présentées dans le Tableau 18 pour les quatre paramètres d'angle et de distance.

Tableau 18 : Valeurs seuils utilisées pour les paramètres lors de la validation et de l'agrandissement de fragments par F2D

Paramètre	Valeur seuil (validation)	Valeur seuil (application)
Distance	±45%	±9%
Angle de liaison	±45°	±17°
Angle hors plan	±45°	±15°
Angle dièdre	±90°	±45°

Au cours de ce chapitre, je présenterai d'abord les trois projets à visée thérapeutique issus du laboratoire d'excellence (LabEx) SynOrg, puis la quatrième collaboration qui a été effectuée pour la recherche de sondes fluorescentes.

4.2 ICOA – Orléans (LabEx SynOrg)

Pour commencer, j'ai appliqué Frags2Drugs (F2D) en collaboration avec le groupe « chimie hétérocyclique pour l'innovation en thérapeutique et imagerie TEP » dirigé par le Pr Sylvain Routier. Ce groupe de recherche fait partie de l'équipe « Hétérocycles, Nucléosides et Agents d'Imagerie » (HNAI) de l'institut de chimie organique et analytique (ICOA) d'Orléans.

4.2.1 Contexte biologique

Au cours de ce projet, j'ai utilisé F2D pour deux applications thérapeutiques distinctes. La première, en cancérologie, concerne la protéine kinase *haploid germ cell–specific nuclear protein* (Haspin). La seconde, sur la maladie d'Alzheimer, concerne les protéines *dual-specificity tyrosine regulated kinase 1A* (DYRK1A) et *Cdc2-like kinase 1* (CLK1).

4.2.1.1 Haspin

Haspin est une protéine kinase atypique de la famille des sérines/thréonines kinases. L'ARN messager codant pour cette protéine a été découvert en 1994 dans des cellules germinales de souris[289]. La protéine Haspin est une protéine à activation constitutive, capable de s'activer seule. Elle est localisée dans le thymus, la moelle osseuse, le foie du fœtus et aussi de manière plus faible, dans l'intestin, la rate, les poumons et d'autres tissus fœtaux[290]. Haspin joue des rôles importants dans la régulation de la mitose, en particulier dans la phosphorylation de l'histone H3 et pour l'alignement des chromosomes pendant la métaphase[291]. De plus, l'ARN messager d'Haspin est exprimé dans les cellules en prolifération et il n'est pas exprimé dans les cellules qui ne se divisent pas[292].

En cancérologie, Haspin est une cible thérapeutique surexprimée dans le lymphome de Burkitt[293], la leucémie lymphoïde chronique[294] et le cancer du pancréas[295]. L'inhibition d'Haspin pourrait amener moins d'effets secondaires que l'inhibition d'autres protéines kinases « classiques » car il s'agit d'une protéine kinase atypique[296,297].

La structure 3D d'Haspin que j'ai utilisée était confidentielle au moment où j'ai exécuté F2D, mais elle a été publiée depuis dans la *Protein DataBank* (PDB) avec l'ID 7AVQ (Figure 53). En observant les positions des acides aminés Tyr 688 et Glu 535, nous voyons que la protéine a été cristallisée en conformation DFG-in, hélice αC-in. Le ligand co-cristallisé est de type I et forme des interactions hydrogène avec les résidus Glu 606 et Gly 608 de la région charnière, ainsi qu'une interaction hydrogène avec la lysine catalytique (Lys 511). La structure de ce ligand nous servira de base pour dessiner les graines de départ pour l'exécution de F2D, comme détaillé en partie 4.2.2.

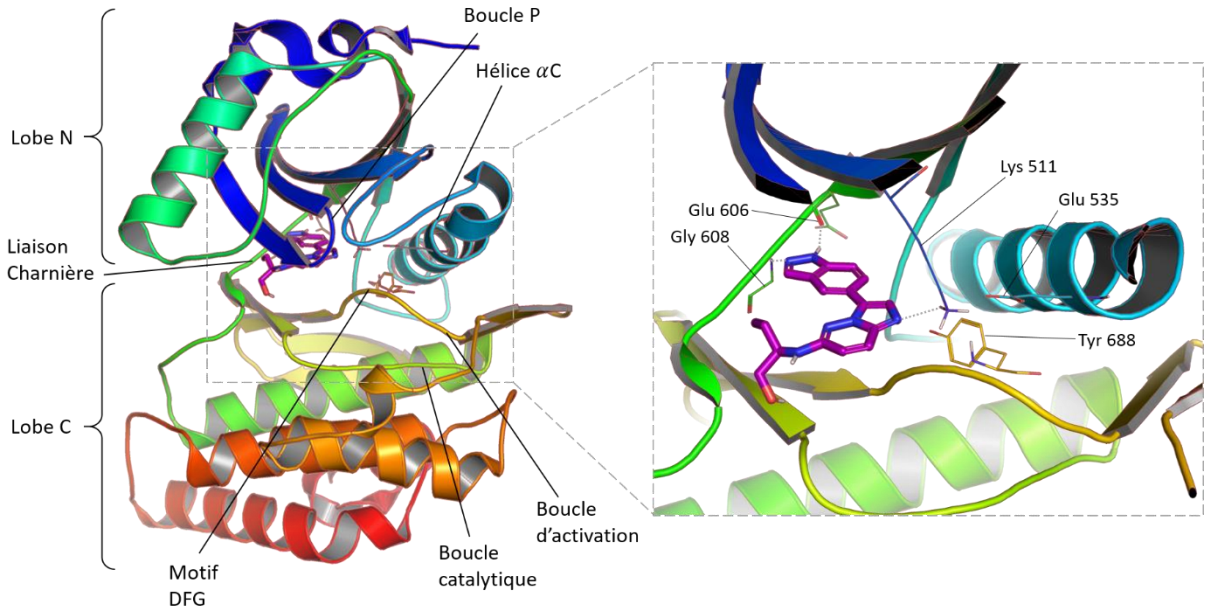


Figure 53 : Structure 3D de la protéine Haspin en complexe avec un inhibiteur imidazopyridazine. (PDB ID : 7AVQ). D'après J. Elie et al.[298]

La structure 3D de la protéine Haspin présente plusieurs particularités. D'abord, le lobe N-terminal est positionné sous une autre partie de la protéine faite d'une extension N-terminale et de deux insertions (Figure 54). Ensuite, il y a une différence avec la structure habituelle des protéines kinase au niveau de la boucle d'activation[299]. Il y a aussi deux brins β supplémentaires et la suppression de l'hélice αG

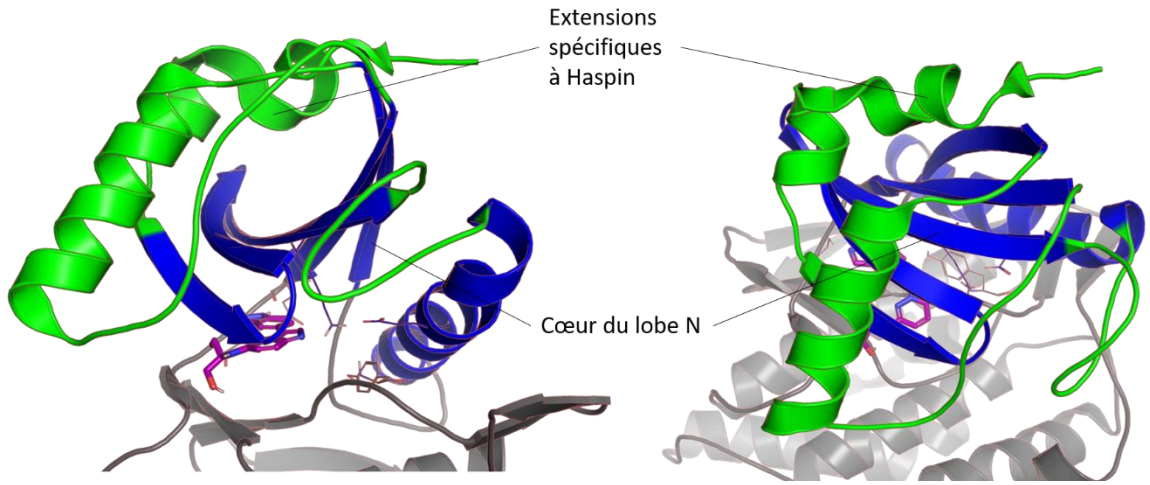


Figure 54 : Lobe et extensions N-terminales de la protéine Haspin. En bleu : Le cœur du lobe N, en vert : les extensions couvrant le lobe N. (PDB ID : 7AVQ)

La Figure 55 présente les structures 2D de quelques exemples d'inhibiteurs d'Haspin qui ont déjà été décrits. Parmi ces inhibiteurs, CHR-6494 a servi de structure de départ pour la conception d'inhibiteurs sélectifs d'Haspin dérivés d'imidazo[1,2-*b*]pyridazine[298]. SGI-1776 a terminé deux essais cliniques de phase I, mais a été retiré suite à l'apparition d'effets secondaires indésirables. Une étude de phase I est en cours de recrutement pour l'Harmine. La molécule LDN-211898 est un dérivé de l'Harmine.

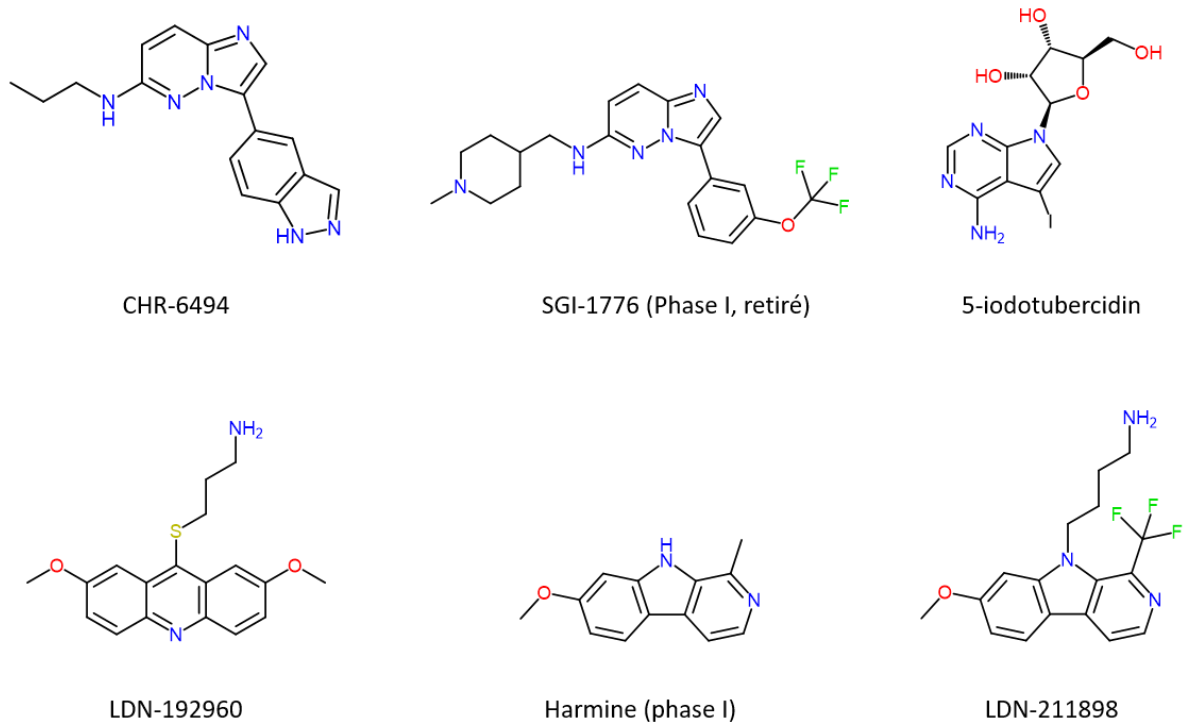


Figure 55 : Exemples de structures d'inhibiteurs d'Haspin déjà connus. D'après J. Elie et al.[298]

Après avoir étudié le contexte biologique de la cible Haspin, nous allons nous concentrer sur celui des protéines CLK1 et DYRK1A.

4.2.1.2 CLK1 et DYRK1A

Les protéines CLK1 et DYRK1A font partie de la famille des protéines kinase CMGC. La sous-famille DYRK est composée de 5 membres et la sous famille CLK est composée de 4 membres. Il s'agit de protéines kinase effectuant leurs phosphorylations sur des sérines et des thréonines. DYRK1A et CLK1 sont les deux protéines qui ont été les plus étudiées de leurs sous-familles respectives. Ces protéines jouent différents rôles physiologiques importants dans l'épissage alternatif de l'ARN messager et au niveau cellulaire dans la réparation de l'ADN, la survie, le contrôle du cycle et la différenciation. L'épissage alternatif de l'ARN messager est un processus biologique très important permettant d'augmenter le nombre de protéines pouvant être synthétisées par le génome humain[300].

Par leurs nombreuses implications physiologiques, les deux protéines CLK1 et DYRK1A sont impliquées dans diverses pathologies. CLK1 est par exemple impliquée dans la myopathie de Duchenne, des cancers et les infections par différents virus[301]. DYRK1A est impliquée dans 20 pathologies différentes dont le syndrome de Down (trisomie 21), des maladies neurodégénératives comme celles d'Alzheimer ou de Parkinson, le diabète et différents cancers[301].

Les structures cristallographiques de complexes protéines-ligands sont enregistrées dans la *Protein Databank* (PDB) et un identifiant « PDB ID » permet d'y faire référence. Dans la suite de mon manuscrit, pour simplifier la lecture, j'utiliserai « l'ID PDB 3OG7 » pour signifier « les structures cristallographiques du complexe protéine-ligand ayant pour PDB ID 3OG7 ».

Il existe de nombreuses structures cristallographiques de CLK1 et DYRK1A dans la PDB. Dans le cadre de ce projet, j'ai utilisé l'ID PDB 7OPG contenant la protéine CLK1 co-cristallisée avec un inhibiteur ayant une sous-structure imidazothiadiazole. Cette sous-structure, interagissant avec la lysine catalytique de CLK1, m'a permis de dessiner la deuxième partie des graines de départ pour l'utilisation

de F2D dans ce projet. La structure 3D de la protéine CLK1 est en conformation active DFG-in et hélice αC-in (Figure 56).

Pour l'agrandissement dans DYRK1A, l'ID PDB 2WO6 a été choisi. Il contient la protéine DYRK1A, un peptide substrat et un inhibiteur (Figure 56). Cette structure de la protéine est aussi cristallisée en conformation active (DFG-in, hélice αC-in). Le ligand forme plusieurs liaisons hydrogène avec Leu 241 (donneur et accepteur) et Glu 239 (accepteur) dans la région charnière.

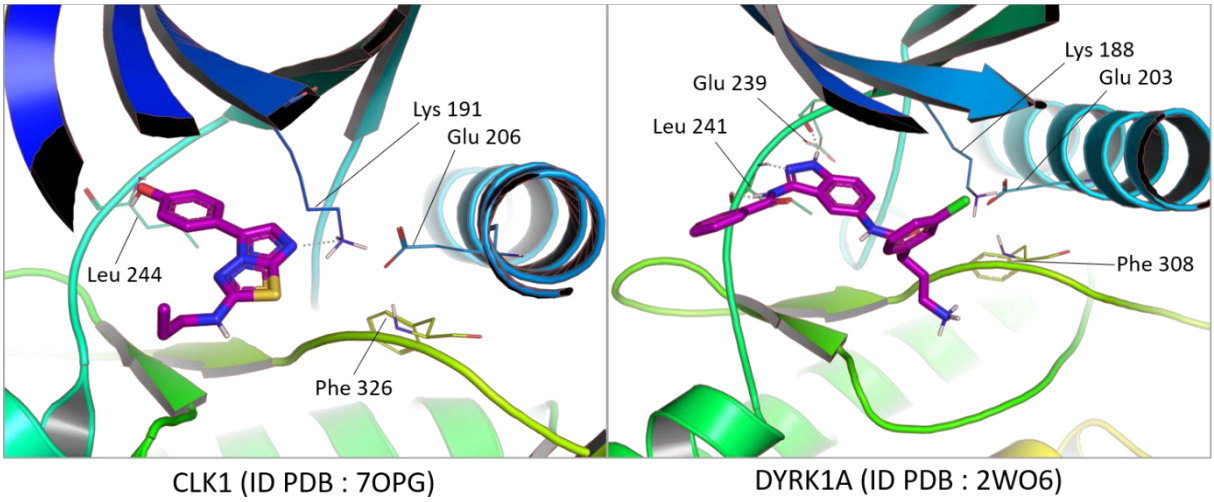


Figure 56 : Structures 3D des sites actifs des protéines CLK1 et DYRK1A utilisées pour l'agrandissement par F2D.

L'inhibition double des protéines DYRK1A et CLK1 par un même inhibiteur est une stratégie utilisée pour traiter la maladie d'Alzheimer, le syndrome de Down, différents cancers et pour augmenter le nombre de cellules β du pancréas dans le diabète de type 1 et de type 2[302]. Différents inhibiteurs ont déjà été découverts avec une activité simultanée sur ces 2 protéines (Figure 57)[301]. Comme pour la protéine Haspin, parmi les inhibiteurs de DYRK1A et CLK1, l'harmine et l'un de ses dérivés (AnnH75) sont retrouvés. Le lorecivivint est un inhibiteur en essai clinique en phase III. La molécule EGCG (épigallocatéchine-O-gallate), également un inhibiteur double de CLK1 et DYRK1A, est un produit naturel et le flavanol le plus abondant du thé. Cet inhibiteur a été soumis à de nombreux essais cliniques allant jusqu'à la phase IV.

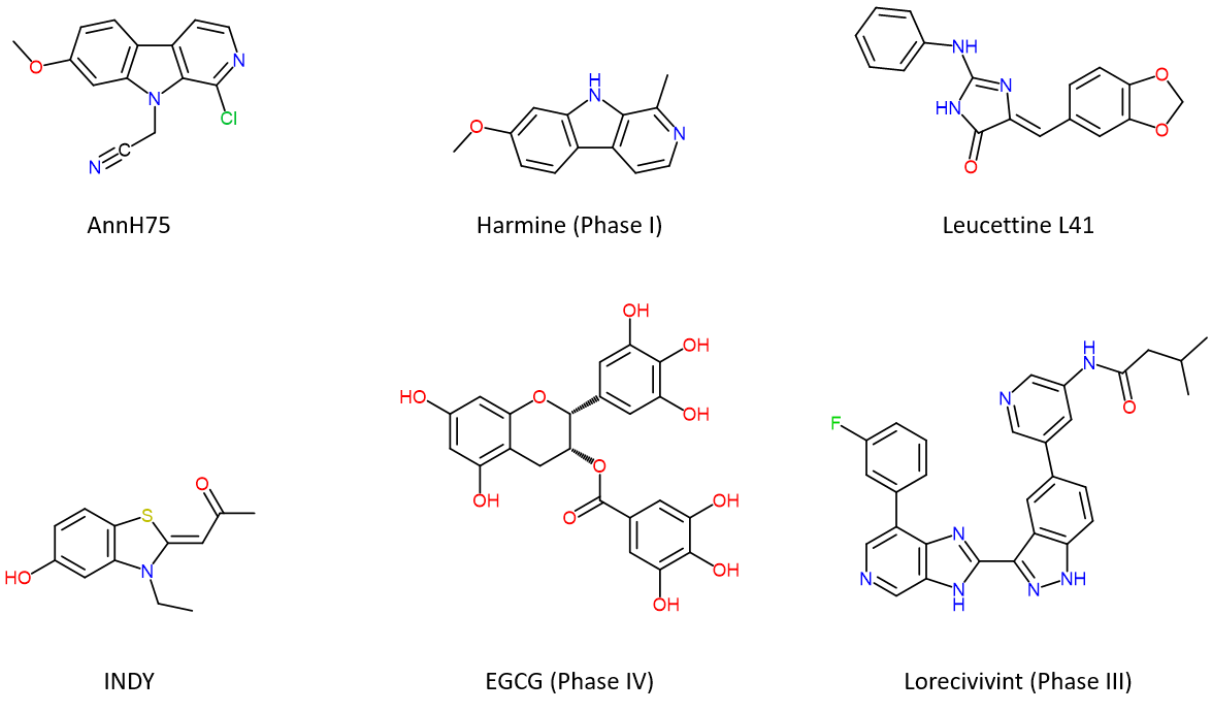


Figure 57 : Exemples de structures d'inhibiteurs doubles de CLK1 et DYRK1A déjà connus. D'après M. Lindberg et al.[301]

La première partie du projet effectué avec le groupe du Pr Sylvain Routier concerne l'inhibition d'Haspin. La deuxième partie de ce projet consiste à trouver des inhibiteurs dual CLK1/DYRK1A. Nous allons maintenant présenter les graines qui ont été utilisées pour l'agrandissement à partir de F2D dans ces cibles.

4.2.2 Graines utilisées

La recherche d'inhibiteurs de protéines kinases (PKI) par F2D dans les trois cibles (Haspin, DYRK1A et CLK1) se fait à partir de 14 graines de départ. Pour les obtenir, j'ai d'abord extrait 3 graines (fragments initiaux) provenant de 2 molécules co-cristallisées dans CLK1 et Haspin, puis j'ai modifié leurs structures afin d'obtenir les 14 graines de départ proposées par mes collègues chimistes. Les 2 molécules co-cristallisées viennent de 2 structures cristallographiques qui étaient confidentielles au moment où j'ai effectué ce travail. Aujourd'hui, ces 2 structures cristallographiques sont publiées dans la PDB avec les ID 7OPG pour CLK1 et 7AVQ pour Haspin. La Figure 58 montre les structures 2D des deux molécules provenant des structures cristallographiques d'Haspin et de CLK1, ainsi que les 3 graines initiales. La Figure 59 représente les structures 2D des 14 graines qui ont été obtenues à partir des modifications effectuées sur les 3 graines initiales.

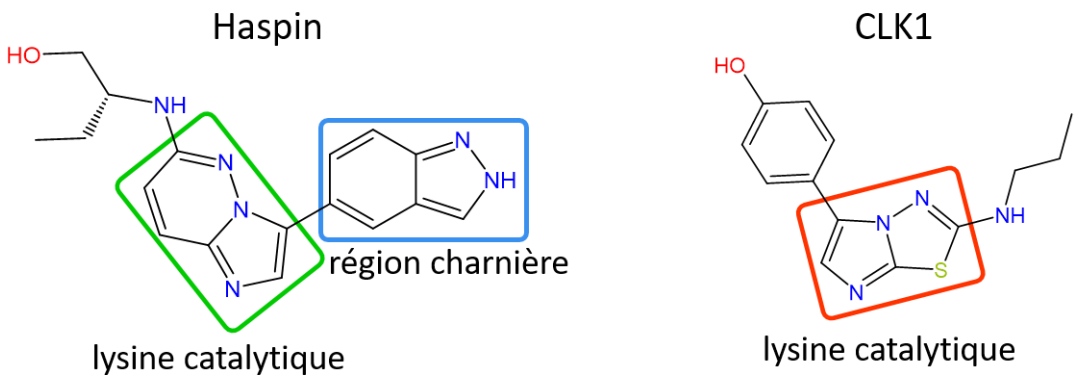


Figure 58 : Structures 2D des deux molécules co-cristallisées dans Haspin et CLK1. Les 3 graines initiales ayant servi de base pour obtenir les 14 graines utilisées dans ce projet sont encadrées. En vert : fragment associé à la lysine catalytique d'Haspin, en bleu : fragment associé à la région charnière d'Haspin et en rouge : fragment associé à la lysine catalytique de CLK1.

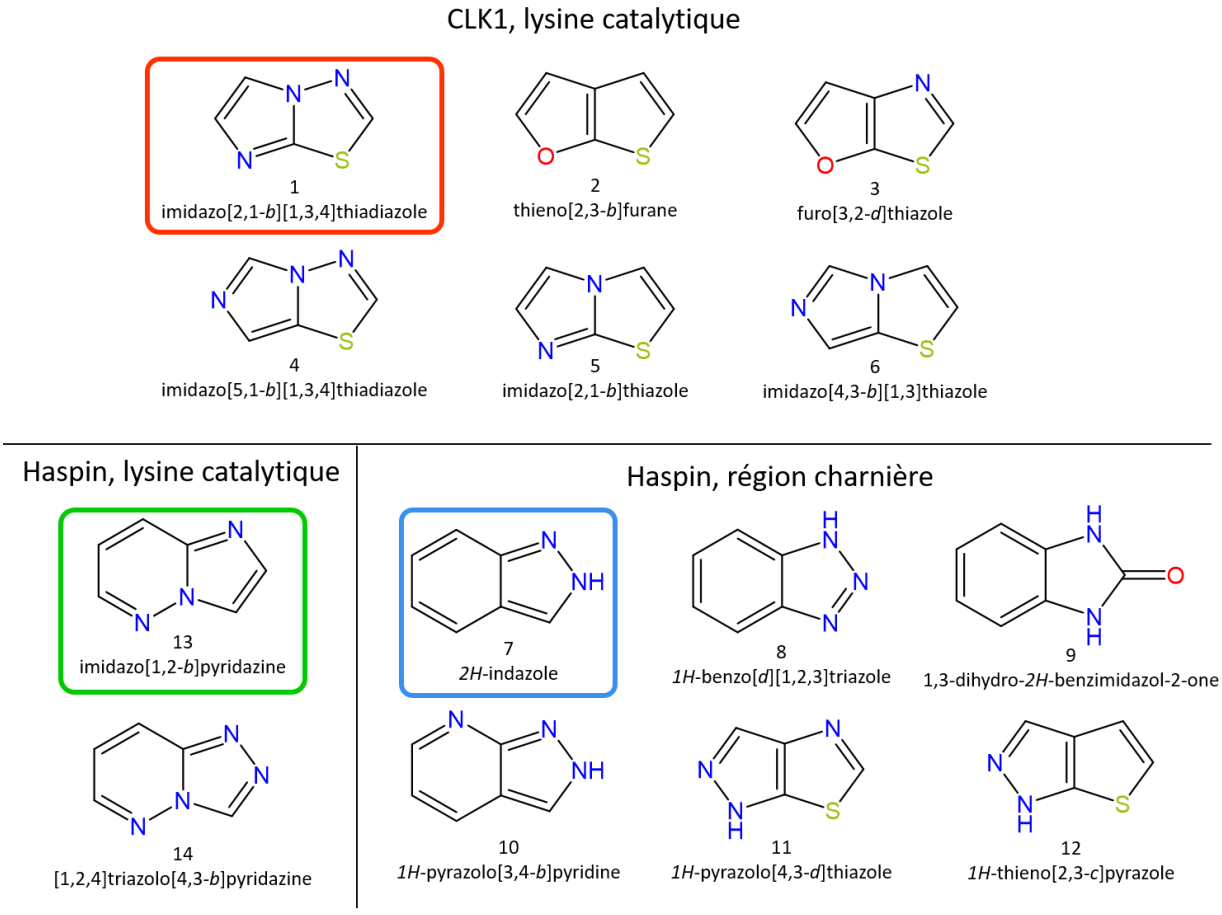


Figure 59 : Quatorze graines de départ utilisées dans le projet d'application de F2D sur Haspin et DYRK1A/CLK1.Six graines ont été obtenues par la modification de la structure moléculaire entourée en rouge qui était liée à la lysine catalytique de CLK1, six autres graines proviennent de celle entourée en bleu qui était associée à la région charnière d'Haspin et les deux dernières viennent de celle entourée en vert et sont liées à la lysine catalytique d'Haspin.

F2D nécessite de connaître l'emplacement 3D des graines de départ pour pouvoir les agrandir. Comme nous disposons des 3 graines initiales qui viennent de structures cristallographiques, il n'y a pas besoin d'effectuer de simulation de *docking* pour positionner les graines. Pour obtenir les structures 3D des 14 graines, j'ai effectué les modifications des atomes sur les structures 3D *via* le logiciel MOE (version 2019.0101). Pour les 14 graines, il y a 4 groupes de positions 3D représentatives au sein desquelles les graines ne diffèrent que par la nature des atomes autour des cycles :

- Groupe 1 : graines n°1 à n°6 (initialement associées à la lysine catalytique de CLK1)
- Groupe 2 : graines n°7 à n°10 (initialement associées à la région charnière d'Haspin)
- Groupe 3 : graines n°11 et n°12 (initialement associées à la région charnière d'Haspin)
- Groupe 4 : graines n°13 et n°14 (initialement associées à la lysine catalytique d'Haspin)

Les positions 3D des graines sont représentées par une graine de chaque groupe (Figure 60) pour Haspin et (Figure 61) pour DYRK1A et CLK1.

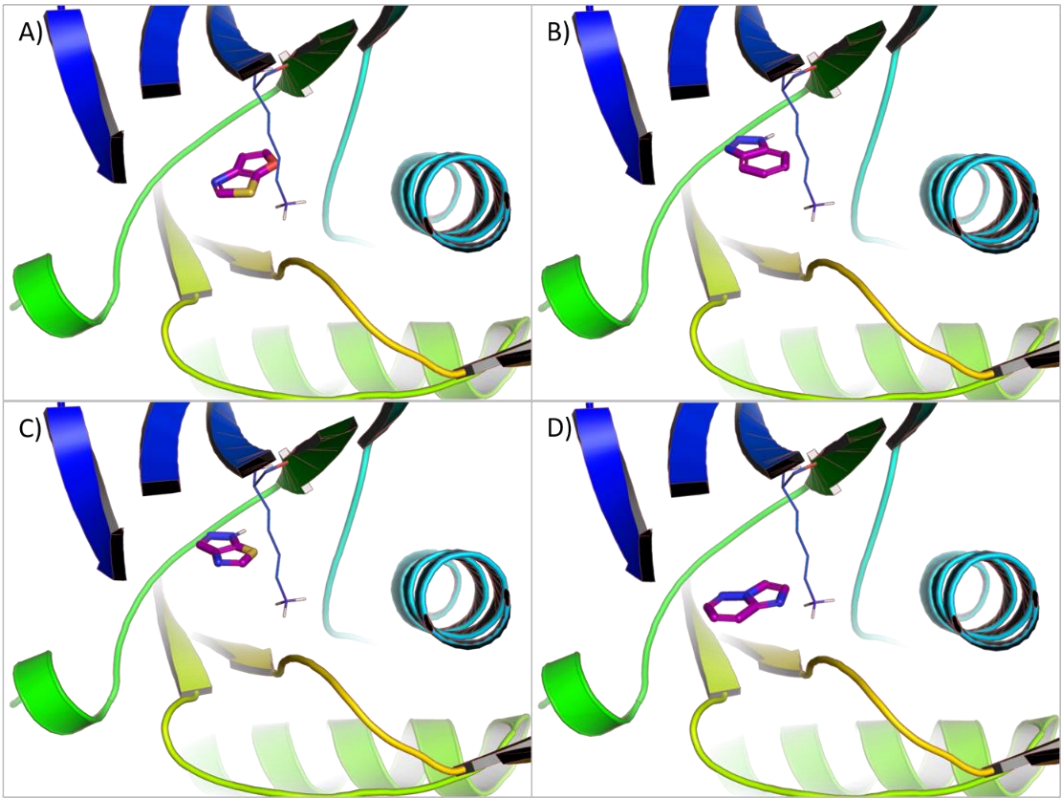


Figure 60 : Positions 3D des graines dans la protéine kinase Haspin. A) Positions 3D des graines du groupe 1. B) Positions 3D des graines du groupe 2. C) Positions 3D des graines du groupe 3. D) Positions 3D des graines du groupe 4.

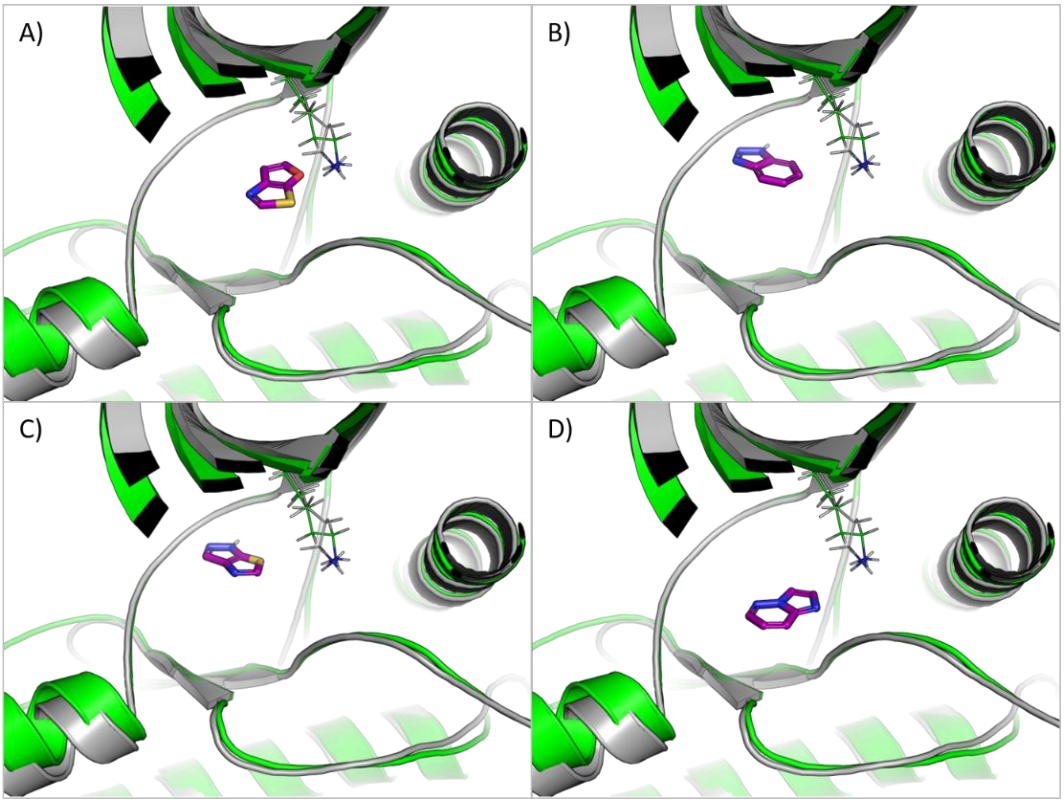


Figure 61 : Positions 3D des graines dans les protéines CLK1 (en vert) et DYRK1A (en gris, ID PDB : 2WO6). A) Positions 3D des graines du groupe. 1 B) Positions 3D des graines du groupe 2. C) Positions 3D des graines du groupe 3. D) Positions 3D des graines du groupe 4.

Grâce à la superposition préalable des protéines kinases, les fragments de la base de données (BDD) de F2D sont tous positionnés dans un référentiel commun (Chapitre 2 partie 2.1.2). Pour ajouter les graines et les nouvelles cibles (structures confidentielles d'Haspin et de CLK1) dans la BDD il faut procéder en deux étapes : d'abord calculer les relations d'inclusion et d'exclusion entre les nouvelles graines et tous les fragments de la BDD, puis calculer les relations de compatibilité entre tous les fragments de la BDD (comprenant maintenant les graines) et les nouvelles cibles. L'agrandissement pourra alors commencer car les critères nécessaires au fonctionnement de F2D seront réunis.

4.2.3 Agrandissement des fragments

Les valeurs seuils des paramètres utilisés et présentées dans le Tableau 18, autorisant de trop nombreuses liaisons entre fragments lors de l'agrandissement, rendent impossible d'effectuer l'agrandissement jusqu'à la masse molaire maximale. En effet, le nombre de molécules pouvant potentiellement être obtenues est trop grand et nécessite trop de mémoire vive (mémoire RAM) pour que le programme puisse fonctionner correctement. Dans ce cas, le programme F2D s'arrête sans fournir de résultats. Pour éviter cela, j'ai effectué l'agrandissement des fragments en procédant par cycles permettant de réduire le nombre de solutions potentiellement générées par F2D. Chaque cycle consiste en une boucle faite des 4 étapes suivantes :

1. Ajout des graines à la BDD de F2D
2. Agrandissement en se limitant à 1 nouveau fragment
3. Pré-sélection des molécules par informatique
4. Sélection par des chimistes médicinaux

Lors de la pré-sélection informatique (étape n°3), certaines méthodes de filtres (Chapitre 2 partie 2.1.2, Figure 25) ne peuvent pas être appliquées. Par exemple, lors du *redocking*, l'ajout d'un nouveau

fragment pourrait modifier la pose d'une molécule et faire qu'une molécule rejetée, composée de 2 fragments, soit conservée avec l'ajout d'un troisième fragment formant des interactions plus favorables. Ainsi, la pré-sélection s'effectue en calculant les descripteurs moléculaires et en appliquant les filtres *Protein Kinase Inhibitor Database* (PKIDB)[34,35]. Les molécules sont ensuite discutées avec nos partenaires pour sélectionner les plus intéressantes (étape n°4). Cette sélection apporte un certain biais car elle est effectuée sur des critères spécifiques à chaque chimiste, mais elle est primordiale pour s'assurer que les molécules générées correspondent bien à leurs attentes.

Après chaque cycle, les molécules sont sauvegardées et deviennent les graines pour le prochain agrandissement. Ces cycles sont reproduits jusqu'à l'obtention d'au moins une molécule avec une masse molaire de 650 g/mol. Cette limite de masse molaire correspond au maximum des PKI recensés sur PKIDB, elle permet d'effectuer l'agrandissement de fragments de manière exhaustive. Lorsque cette condition d'arrêt des cycles est atteinte, toutes les molécules obtenues pendant tous les cycles sont filtrées et évaluées pour sélectionner les meilleures.

J'ai appliqué cette procédure d'agrandissement dans trois des quatre projets de recherche d'inhibiteurs de protéines kinases présentés dans cette thèse, en commençant par les cibles Haspin, CLK1 et DYRK1A.

4.2.3.1 Agrandissements et sélections dans Haspin

Après deux cycles d'agrandissement, 1 409 molécules composées de trois fragments ont été générées (Tableau 19).

Tableau 19 : Nombre de molécules obtenues et sélectionnées sur la cible Haspin.

Nombre de fragments	Nombre de molécules pré-sélectionnées ou initiales	Nombre de molécules sélectionnées
1	14	14
2	275	97
3	1 409	1 295

Ce nombre de molécules est trop élevé pour qu'elles soient passées en revue une à une par des chimistes médicinaux. J'ai donc poursuivi l'agrandissement sans l'étape n°4 en la remplaçant par le calcul d'un score de facilité de synthèse nommé Score SA (pour Synthetic Accessibility Score)[303]. Ce score varie de 0 à 10, une molécule facile à synthétiser aura un Score SA plus faible qu'une molécule plus complexe à synthétiser. La valeur seuil que nous utilisons dans le programme F2D est Score SA < 4. Nous utilisons cette valeur car 92% des molécules présentes dans PKIDB ont un score SA < 4 (261 molécules / 283 en juillet 2021, Figure 62).

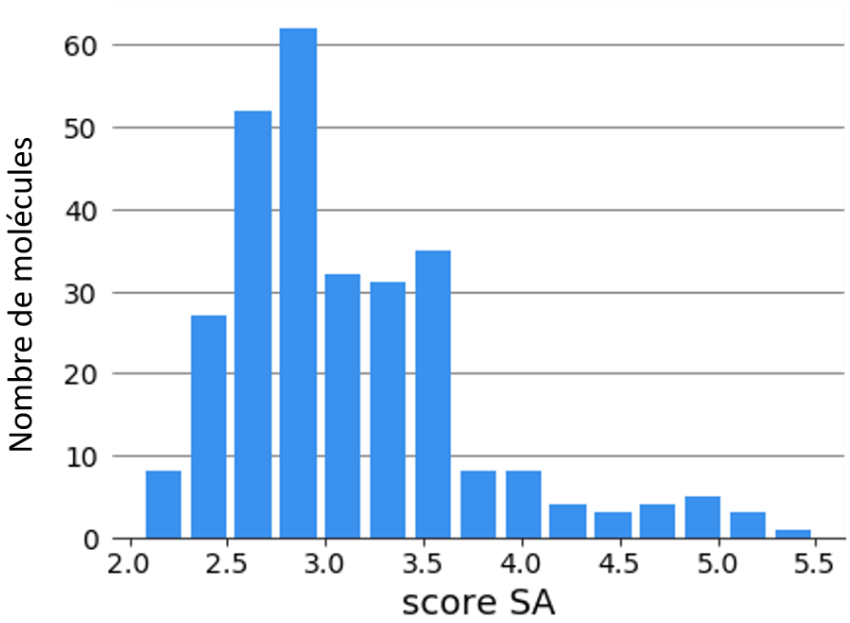


Figure 62 : Distribution du score SA des molécules de PKIDB.

À partir des 1 409 molécules composées de 3 fragments, j'ai conservé 1 295 molécules avec un Score SA < 4 (Tableau 19). Cette sélection est moins drastique que celle effectuée par les chimistes, mais elle a l'avantage de pouvoir être appliquée rapidement sur un grand nombre de molécules.

Pour être certains de n'oublier aucune solution, nous souhaitons atteindre la condition d'arrêt de l'agrandissement. Cela permet de s'assurer de ne pas manquer l'ajout d'un nouveau fragment qui améliorerait les résultats obtenus. Dans le but d'obtenir rapidement cette condition, j'ai ajouté 4 fragments au lieu d'un seul au cours d'un cycle. Cependant plusieurs problèmes ont été rencontrés. Plus de 5 500 000 molécules composées de 7 fragments ont été obtenues, parmi lesquelles aucune molécule n'a atteint une masse molaire de 650 g/mol. Ces molécules sont réparties en 1195 fichiers représentant au total 18,7 Go et ont été obtenues en 4 mois. Ce temps n'est pas uniquement celui de l'exécution de F2D, le programme s'arrêtant parfois de fonctionner car la mémoire vive (64 Go) de mon ordinateur était saturée. J'ai ainsi parfois dû relancer F2D pour qu'il puisse continuer l'agrandissement. Cela montre qu'un des objectifs initiaux qui était la rapidité d'exécution du programme, sans intervention humaine n'a pas été atteint par F2D pour cette application.

Lors de l'analyse des molécules obtenues, la quantité de mémoire vive disponible est encore problématique, puisqu'elle ne permet pas de charger tous les résultats en une fois. Une solution pourrait être de ne charger qu'une partie des molécules, puis de faire une sélection avant de passer à une partie suivante ; mais au final, j'ai choisi de recommencer l'agrandissement à partir de 3 fragments en n'ajoutant qu'un fragment à chaque cycle. Au moment de la rédaction de cette thèse, à cause des problèmes constatés, les calculs doivent être relancés.

4.2.3.2 Agrandissements et sélections dans CLK1 et DYRK1A

Le programme F2D ne permet pas de prendre en compte plusieurs cibles pour un même agrandissement. Pour découvrir des inhibiteurs doublement spécifiques de CLK1 et de DYRK1A, j'ai d'abord effectué l'agrandissement dans ces deux cibles séparément, puis conservé uniquement les molécules communes aux deux cibles. La structure 3D de CLK1 provient d'un fichier confidentiel qui m'a été fourni, alors que j'ai utilisé l'ID PDB 2WO6[304] pour DYRK1A.

En procédant de la même manière que pour l'agrandissement dans Haspin, j'ai obtenu 1 312 molécules (Tableau 3). Ce nombre de molécules est trop élevé pour que les chimistes médicinaux puissent effectuer leur sélection. En appliquant la sélection basée sur le Score SA, nous avons supprimé 110 molécules qui semblent trop complexes à synthétiser.

Tableau 20 : Nombre de molécules doublement spécifiques de CLK1 et DYRK1A obtenues et sélectionnées.

Nombre de fragments	Nombre de molécules pré-sélectionnées ou initiales	Nombre de molécules sélectionnées
1	14	14
2	294	96
3	1 312	1 202

Les 1 202 molécules ainsi sélectionnées sont toujours trop nombreuses pour être passées en revue une à une. J'ai effectué les agrandissements dans CLK1 et DYRK1A en parallèle de ceux dans Haspin jusqu'à l'obtention de molécules ayant 3 fragments pour chaque cible. Ensuite, je me suis focalisé sur Haspin pour terminer les agrandissements avant de terminer ceux effectués dans CLK1 et DYRK1A. Comme j'ai rencontré des problèmes pour le premier agrandissement dans Haspin, je n'ai pas pu terminer les suivants effectués dans CLK1 et DYRK1A.

4.2.4 Bilan du projet LabEx ICOA

Les résultats obtenus dans ce premier projet ne sont pas ceux escomptés, nous n'avons pas pu effectuer une recherche exhaustive d'inhibiteurs à proposer aux chimistes pour les 3 cibles. Le programme F2D a généré au total 1 406 molécules pour Haspin et 1 312 molécules pour CLK1/DYRK1A. Je souhaite poursuivre l'agrandissement, en procédant par cycles d'ajout d'un fragment et de sélections informatiques pour obtenir la condition d'arrêt du programme F2D, sans avoir les problèmes de mémoire RAM. Les sélections des meilleures molécules basées sur les simulations de docking, des recherches par similarité de molécules déjà connues, des estimations de sélectivité et des calculs de score *Quantitative Estimate of Druglikeness* (QED) pourront alors être effectuées car nous aurons obtenu toutes les possibilités de structures à partir de F2D.

Ce projet a néanmoins permis d'adapter la méthode utilisée pour la découverte de PKI afin qu'elle puisse mieux fonctionner pour les autres projets. Parallèlement au projet du LabEx SynOrg effectué avec l'ICOA, j'ai travaillé sur une deuxième application de F2D pour un autre laboratoire que je vais maintenant présenter.

4.3 COBRA – Rouen (LabEx SynOrg)

Le deuxième projet d'utilisation de Frags2Drugs (F2D) a été réalisé avec le Pr Thierry Besson de l'équipe « hétérocycles » du laboratoire « chimie organique bioorganique réactivité et analyse » (COBRA) situé à Rouen. J'ai utilisé F2D pour trouver des molécules ciblant les protéines *Dual-specificity tyrosine regulated kinase* (DYRK) 1A et DYRK2 dans le cadre d'une demande de financement pour un projet de thèse. L'obtention de résultats dans ce deuxième projet avait une contrainte de temps, les molécules devant être obtenues le plus rapidement possible pour le dépôt de la demande.

4.3.1 Contexte biologique

DYRK1A et DYRK2 sont deux membres de la sous-famille de protéines DYRK. Chez l'humain, cette sous-famille est composée de 5 membres répartis en 2 classes[305] :

- DYRK1A, DYRK1B (classe I)
- DYRK2, DYRK3, DYRK4 (classe II)

Comme nous l'avons vu précédemment (partie 4.2.1) DYRK1A est une protéine impliquée dans de nombreuses pathologies comme des maladies neurodégénératives, des cancers et le diabète[301]. Il s'agit de la première des deux cibles que nous utiliserons dans l'application de F2D avec le laboratoire COBRA.

DYRK2 a de nombreux rôles physiologiques et est la protéine la plus étudiée de la classe II pour ses nombreuses implications en cancérologie et dans le développement des tissus. DYRK2 phosphoryle de nombreux substrats, lui donnant, comme pour DYRK1A, deux rôles contradictoires : à la fois anti et pro-tumoraux[306]. Dans le cancer du foie, le cancer colorectal et la leucémie myéloïde chronique DYRK2 a un rôle anti-tumoral. Dans les cancers du cerveau, des poumons, des ovaires et du sein, DYRK2 joue les deux rôles divergents. Dans le myélome multiple et le cancer du sein triple-négatif, DYRK2 joue un rôle uniquement pro-oncogène[307].

À cause de leurs implications dans de nombreuses pathologies, DYRK1A et DYRK2 sont des cibles thérapeutiques d'intérêt. Cependant, il subsiste toujours des controverses concernant les rôles divergents de ces deux protéines[308].

Pour utiliser F2D afin de découvrir des inhibiteurs de DYRK1A et DYRK2, j'ai choisi 2 complexes différents de ceux déjà utilisés dans le premier projet mené avec l'institut de chimie organique et analytique (ICOA) car ils ont été publiés par le Pr Thierry Besson. Pour DYRK1A, l'ID PDB 6QU2 contient un complexe entre cette protéine et l'inhibiteur FC162. La protéine DYRK2 dans l'ID PDB 5LXD est en complexe avec la molécule EHT 1610[309] (Figure 63). Les deux structures sont en conformation active DFG-in, hélice αC-in. Dans les deux sites actifs, les inhibiteurs sont de type I et interagissent avec la leucine de la région charnière et la lysine catalytique.

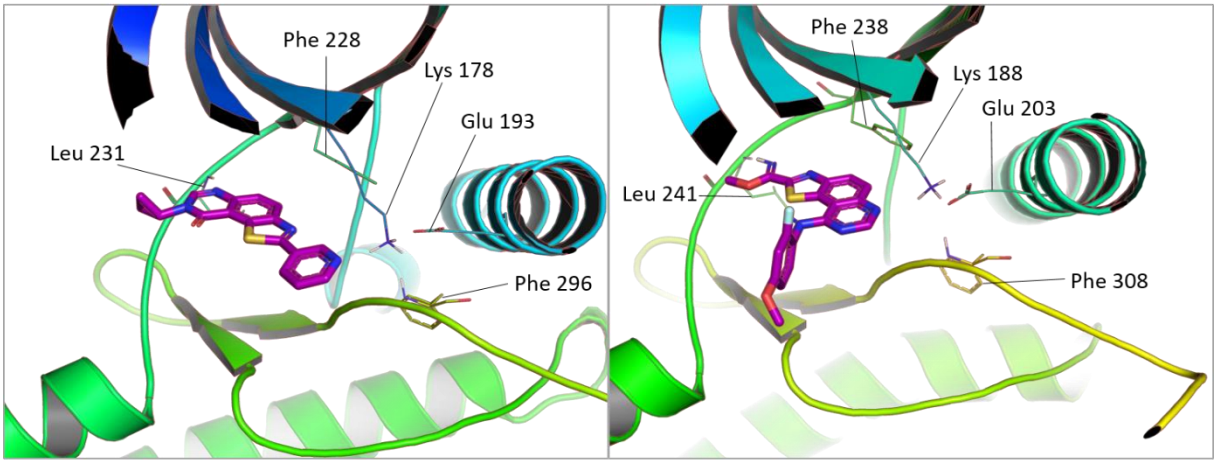


Figure 63 : Structures 3D des sites actifs des protéines DYRK1A et DYRK2 utilisées pour l'agrandissement par F2D.

De nombreux inhibiteurs doubles ont déjà été décrits pour les protéines DYRK1A et DYRK2[307]. C'est le cas par exemple de AnnH75, de l'harmine, de la Leucettine L41 ou encore de l'INDY qui sont aussi des inhibiteurs de CLK1 (Figure 57). D'autres molécules, n'ayant pas encore atteint les essais cliniques, ont été présentées dans la littérature scientifique. C'est par exemple le cas de EHT1610, RD0392, GSK626616, EHT 5372, FLuoro-DANDY analog 5g et SC97202 (Figure 64).

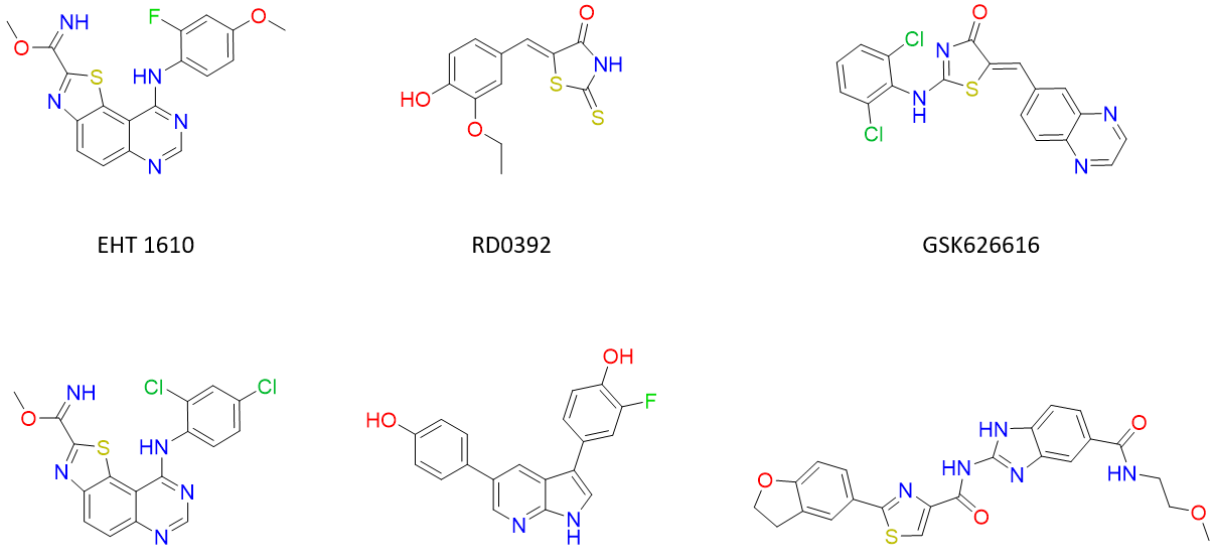


Figure 64 : Exemples de structures d'inhibiteurs de DYRK1A et DYRK2 déjà connus. D'après V. Tandon et al.[307]

4.3.2 Graines utilisées

En tant que points de départ pour l'utilisation de F2D sur les cibles DYRK1A et DYRK2, j'ai reçu 16 graines dessinées par le Pr Thierry Besson (Figure 65).

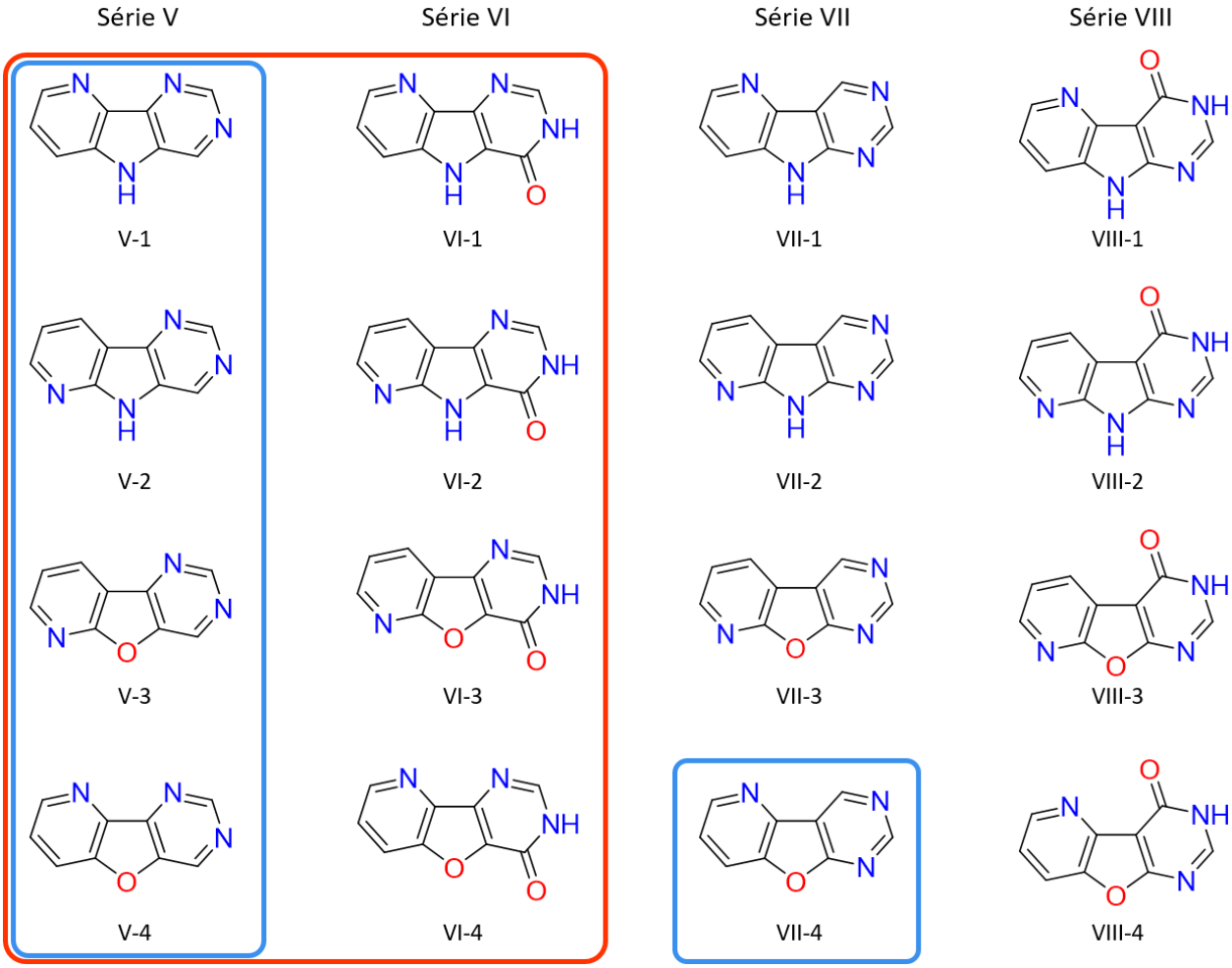


Figure 65 : Graines initialement proposées pour l'utilisation de F2D sur DYRK1A et DYRK2. Les graines entourées en bleu et rouge ont été sélectionnées après simulation de docking pour l'agrandissement dans DYRK1A et DYRK2 respectivement. Les noms de séries correspondent aux dénominations données par les chimistes de synthèse.

Il n'existe pas de structures 3D de ces graines co-cristallisées avec DYRK1A ou DYRK2. Pour indiquer leurs positions 3D au logiciel F2D, j'ai effectué des simulations de *docking* de ces 16 graines. Afin d'effectuer les simulations de *docking* les plus exhaustives possible, j'ai généré 500 poses pour chaque molécule à l'aide du logiciel rDock (2013.1)[310] et j'ai supprimé les doublons 3D, par un regroupement avec un RMSD < 0,5 Å. F2D peut générer de nombreuses molécules à partir des graines initiales. Pour réduire ces possibilités, j'ai souhaité réduire le nombre de poses à utiliser comme graines. J'ai conservé les 5 poses ayant le score « SCORE.INTER » le plus faible pour chaque graine, amenant le nombre total de graines à explorer à 80. Pour s'assurer que les graines soient correctement positionnées en 3D par rapport aux sites actifs des protéines DYRK1A et DYRK2, elles ont été sélectionnées si elles formaient une liaison hydrogène avec la région charnière et/ou la lysine catalytique. Ainsi, sept poses ont été conservées pour 5 graines dans DYRK1A (Figure 66) et 11 poses l'ont été pour 8 graines dans DYRK2 (Figure 67). Ces 19 poses obtenues par simulations de *docking* constituent les graines qui ont été ajoutées à la BDD de F2D pour pouvoir effectuer l'agrandissement de fragments.

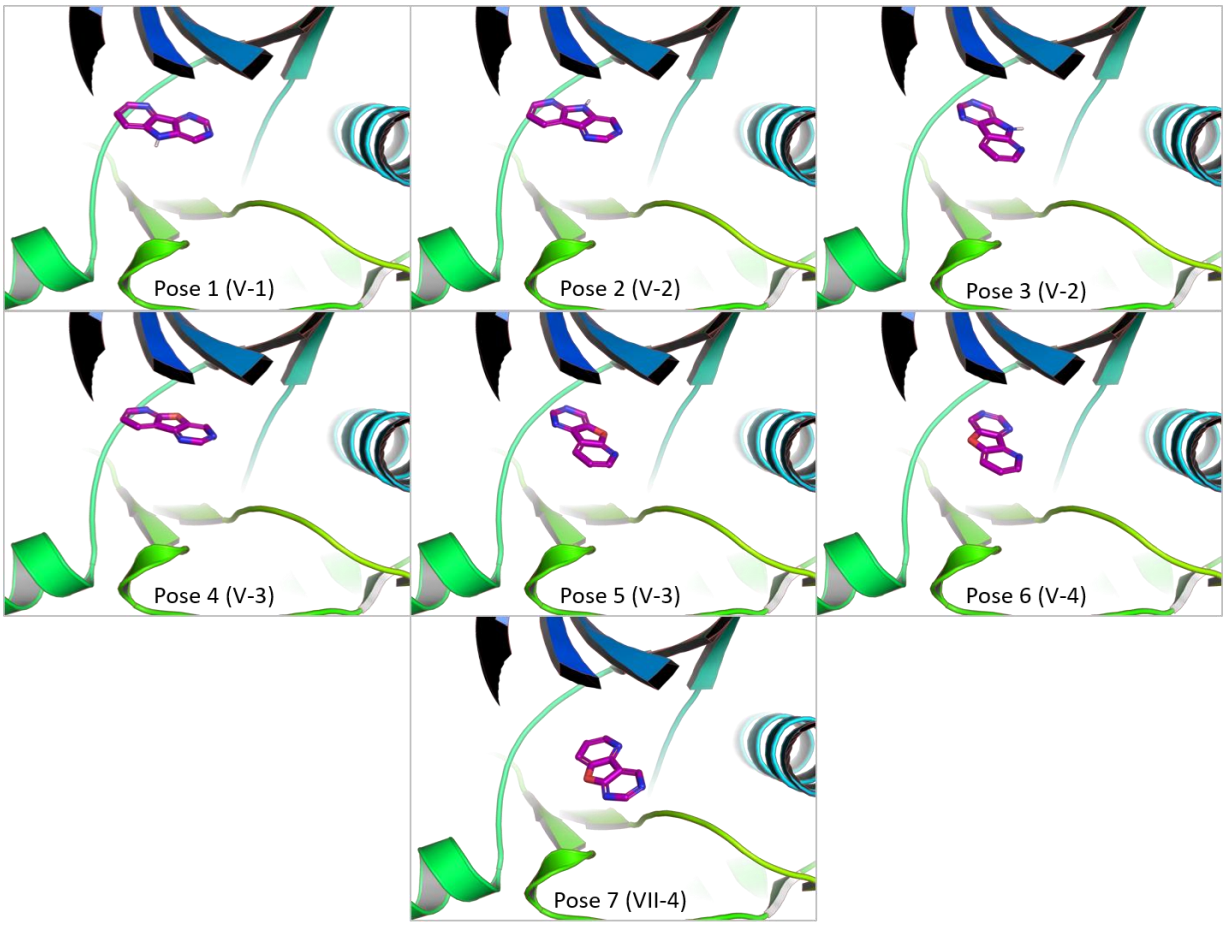


Figure 66 : Sept poses obtenues par simulation de docking et conservées pour initier l'agrandissement dans DYRK1A.

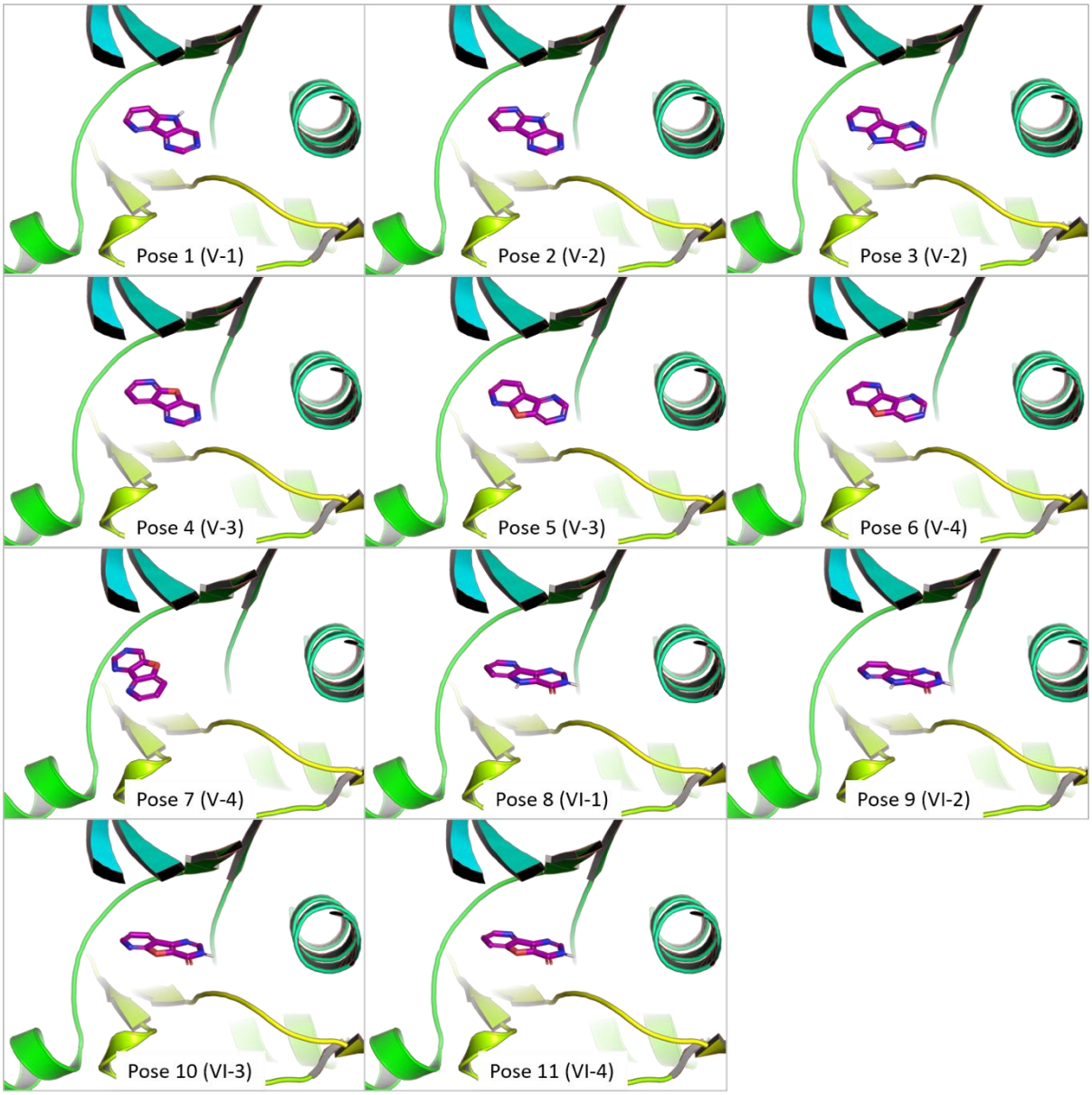


Figure 67 : Onze poses obtenues par simulation de docking et conservées pour initier l'agrandissement dans DYRK2.

4.3.3 Agrandissement des fragments

J'ai effectué l'agrandissement dans DYRK1A et dans DYRK2 en utilisant les valeurs seuils montrées dans le Tableau 18 comme paramètres de F2D. Cependant, en raison des difficultés rencontrées pour l'agrandissement dans Haspin (partie 4.2.3.1), j'ignorais si l'agrandissement pourrait s'effectuer jusqu'à la condition d'arrêt sans devoir effectuer des sélections intermédiaires. J'ai donc sauvegardé des fichiers de résultats à chaque nouvel ajout de fragment. J'ai arrêté l'agrandissement après l'ajout de 2 fragments aux graines initiales, car l'ajout d'un troisième fragment allongeait le temps de calcul et que le temps accordé pour déposer la demande de financement de thèse liée à ce projet était limité.

4.3.3.1 Agrandissements et sélections dans DYRK1A

1 817 molécules ont été obtenues après l'ajout de deux fragments aux 7 graines de départ dans DYRK1A. Parmi ces molécules, 1 388 ont un score de facilité de synthèse (score SA) < 4. La masse molaire maximale obtenue parmi ces molécules est de 495,61 g/mol. Parmi ces molécules constituées au total de trois fragments, six forment des interactions avec la région charnière et ont été sélectionnées *via*

l'utilisation d'un pharmacophore accepteur. La Figure 68 présente ces molécules en 2D et 3D. Les trois premières molécules viennent de l'agrandissement effectué à partir de la pose n°1 de la Figure 66. La quatrième molécule est issue de la pose n°2, la cinquième de la pose n°4 et la dernière de la pose n°7.

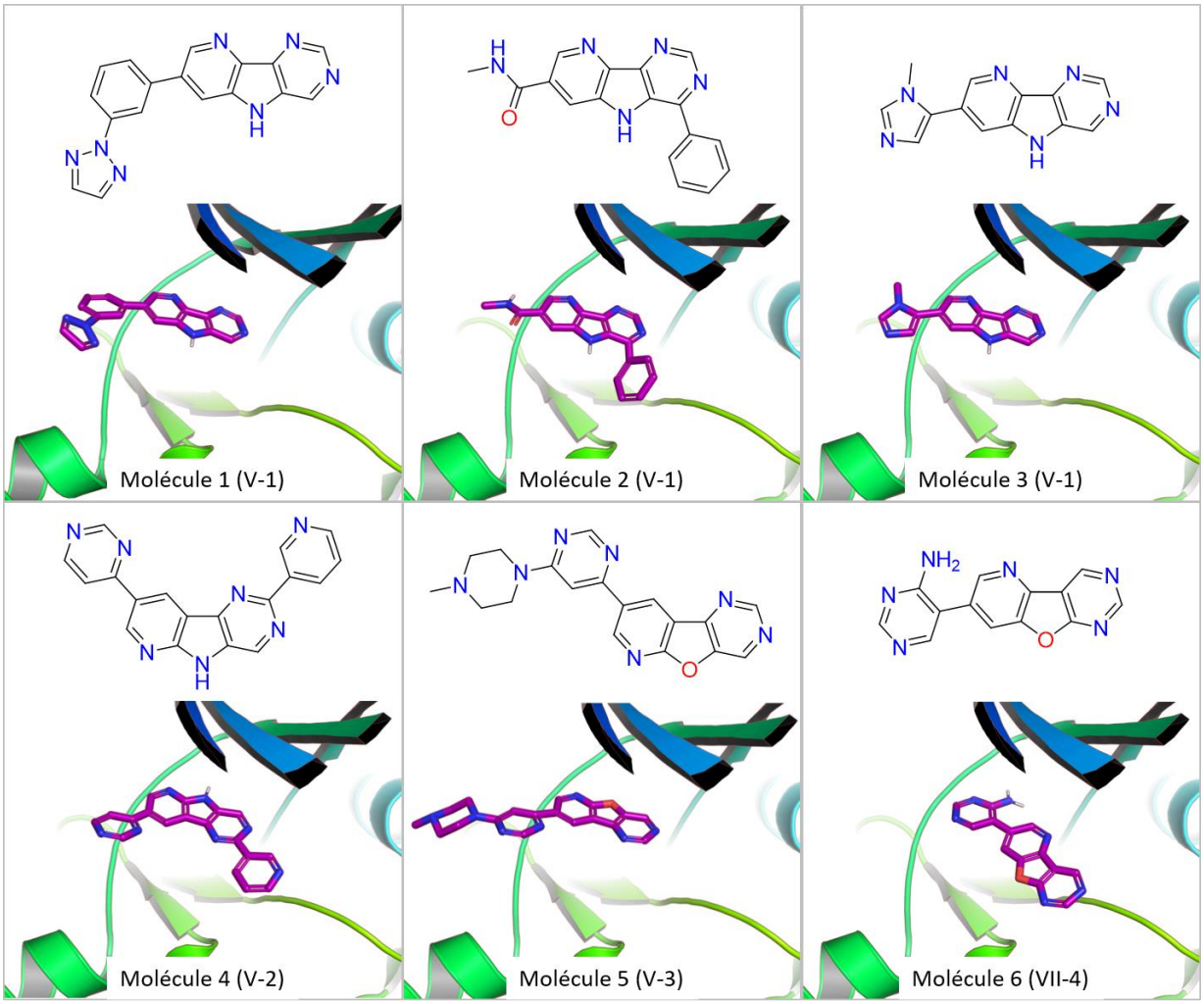


Figure 68 : Six molécules obtenues et sélectionnées lors de l'application de F2D sur DYRK1A.

4.3.3.2 Agrandissements et sélections dans DYRK2

Dans DYRK2, l'agrandissement a été fait à partir des 11 graines de départ. L'ajout de 2 fragments a généré 2 005 molécules. Nous obtenons 1 513 molécules en ne conservant que les molécules avec un score SA < 4. Comme pour la recherche de molécules dans DYRK1A, deux molécules ont été sélectionnées car elles forment des interactions similaires à celles du ligand EHT1610 avec la région charnière (Figure 69). Ces molécules correspondent toutes les deux à la pose n°1 de la simulation de *docking* effectuée dans DYRK2. Dans les deux molécules, l'interaction avec la région charnière est formée par l'oxygène de l'amide, alors que la graine interagit avec la lysine catalytique de la cible.

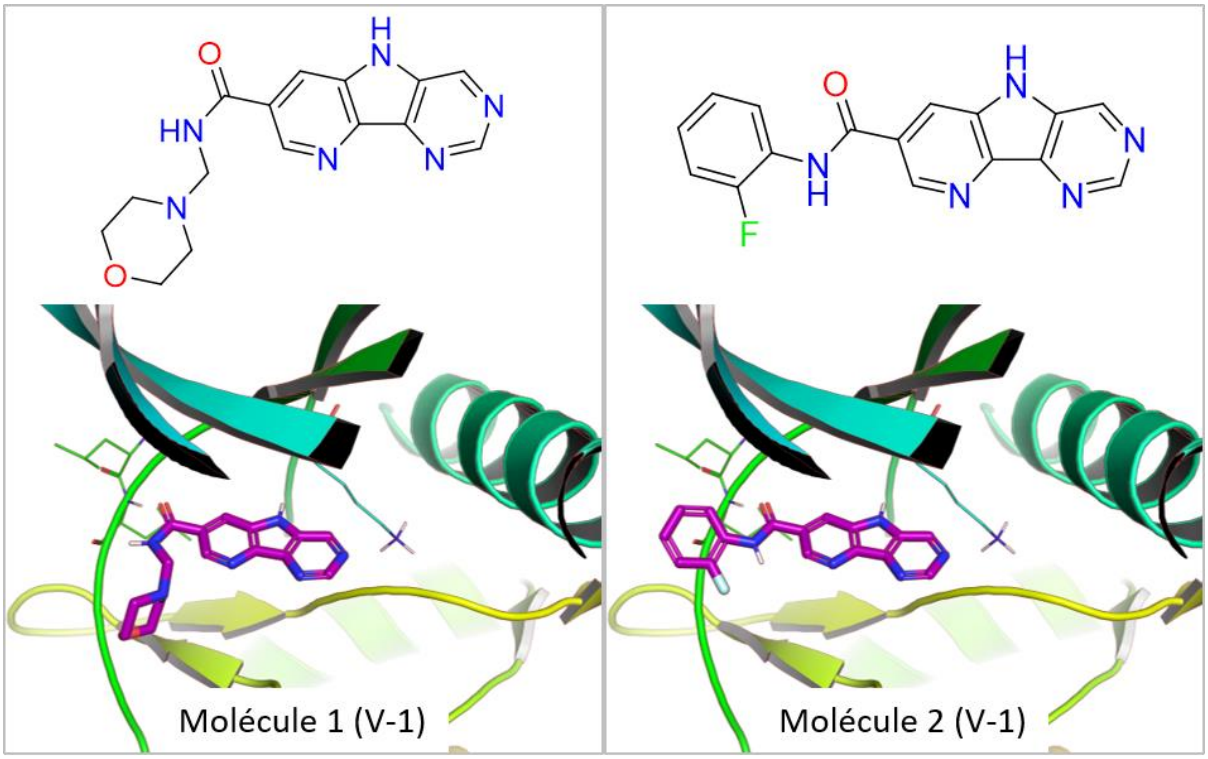


Figure 69 : Deux molécules obtenues et sélectionnées lors de l'application de F2D sur DYRK2.

4.3.4 Bilan du projet LabEx COBRA

Dans ce projet, comme dans celui effectué avec l'ICOA, présenté en partie 4.2, la recherche de molécules par agrandissement n'a pas été exhaustive, la condition d'arrêt n'ayant pas été atteinte. Cependant, sur les 1 388 et 1 513 molécules trouvées pour inhiber respectivement DYRK1A et DYRK2, 8 ont été intégrées dans la demande de financement pour le projet de thèse et cette demande a été acceptée.

La méthode utilisée dans ce projet a confirmé la difficulté d'agrandir les molécules sans passer par les cycles d'agrandissement décrits en partie 4.2.3, à cause de la saturation de la mémoire RAM. La méthode basée sur les cycles d'agrandissement a été appliquée avec succès lors de la collaboration avec le Groupe Innovation et Ciblage Cellulaire (GICC) de Tours, que je vais maintenant vous exposer.

4.4 GICC – Tours (LabEx SynOrg)

Pour la dernière utilisation de Frags2Drugs (F2D) au sein d'un projet issu du laboratoire d'excellence (LabEx) SynOrg, j'ai travaillé avec le Dr Caroline Denevault de l'équipe Innovation Moléculaire et Thérapeutique (IMT) du GICC situé à Tours.

4.4.1 Contexte biologique

Dans ce projet, nous ciblons la protéine *proviral integration site for Moloney murine leukemia virus* (Pim-1). Cette protéine fait partie de la sous famille des trois protéines kinases Pim qui sont toutes des protéines kinases sérines/thréonines proto-oncogènes[311]. Ces protéines ont une activité constitutive, c'est-à-dire qu'elles sont actives sans la présence d'un ligand.

Pim-1 est impliquée dans de nombreuses fonctions biologiques telles que la régulation du cycle cellulaire, l'apoptose, la survie, la prolifération et la différenciation cellulaire[312,313]. Il s'agit d'une cible thérapeutique en cancérologie car elle agit en tant que facteur de survie oncogénique, elle inhibe l'apoptose et régule positivement la progression du cycle cellulaire[314]. Cette protéine est surexprimée dans de nombreux cancers, tels que des hémopathies malignes[315,316] et des cancers solides[317–321]. De plus, chez des souris triple *knockout*, la suppression de la protéine Pim-1 n'altère pas la viabilité ou la fertilité des souris[322]. Cela indique qu'il y aurait peu ou pas d'effet négatif dû à l'inhibition de cette protéine. Enfin, Pim-1 agit aussi sur la résistance multiple aux médicaments des cellules cancéreuses rendant les chimiothérapies moins efficaces en activant des voies de survie ou l'expression de transporteurs d'efflux à médicaments multiples[323].

Le site actif de Pim-1 présente la particularité d'avoir un résidu proline en position 123 dans sa région charnière, là où dans les autres protéines kinases l'ATP forme une de ses deux liaisons hydrogène. La proline 123 ne forme pas de liaison hydrogène et l'ATP présente dans Pim-1 une seule liaison avec la région charnière[324]. C'est l'acide glutamique 121 qui forme la liaison hydrogène permettant l'interaction avec l'ATP (Figure 70). Une autre liaison hydrogène est aussi formée avec la lysine catalytique en position 67. La protéine Pim-1 est co-cristallisée avec l'ATP dans l'ID PDB 3A99 que nous utiliserons pour faire les agrandissements de ce projet.

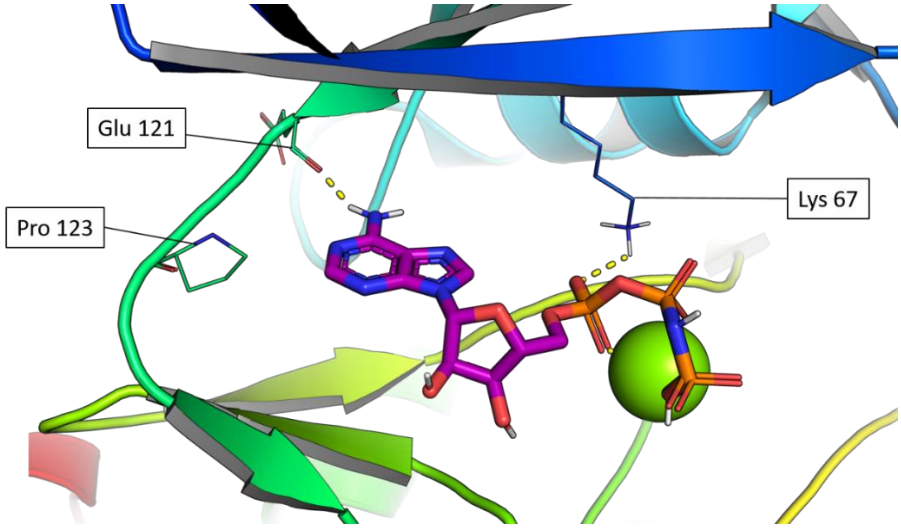


Figure 70 : Interactions formées entre le site actif de Pim-1 et l'ATP (ID PDB : 3A99).

Des essais cliniques ont été menés ou sont en cours sur certains inhibiteurs des protéines Pim. C'est le cas par exemple de SGI-1776[325], TP-3654[326], LGHPDB-447[327] et AZD1208[328] en phase I et uzansertib[329] et SEL24-B489[330] en phase I/II (Figure 71). La molécule LGH-447 a été co-cristallisée dans

l'ID PDB 5DWR[331]. Cependant, aucune de ces molécules ne contient la graine quinoxaline-2-acide carboxylique utilisée pour l'application de F2D dans Pim-1.

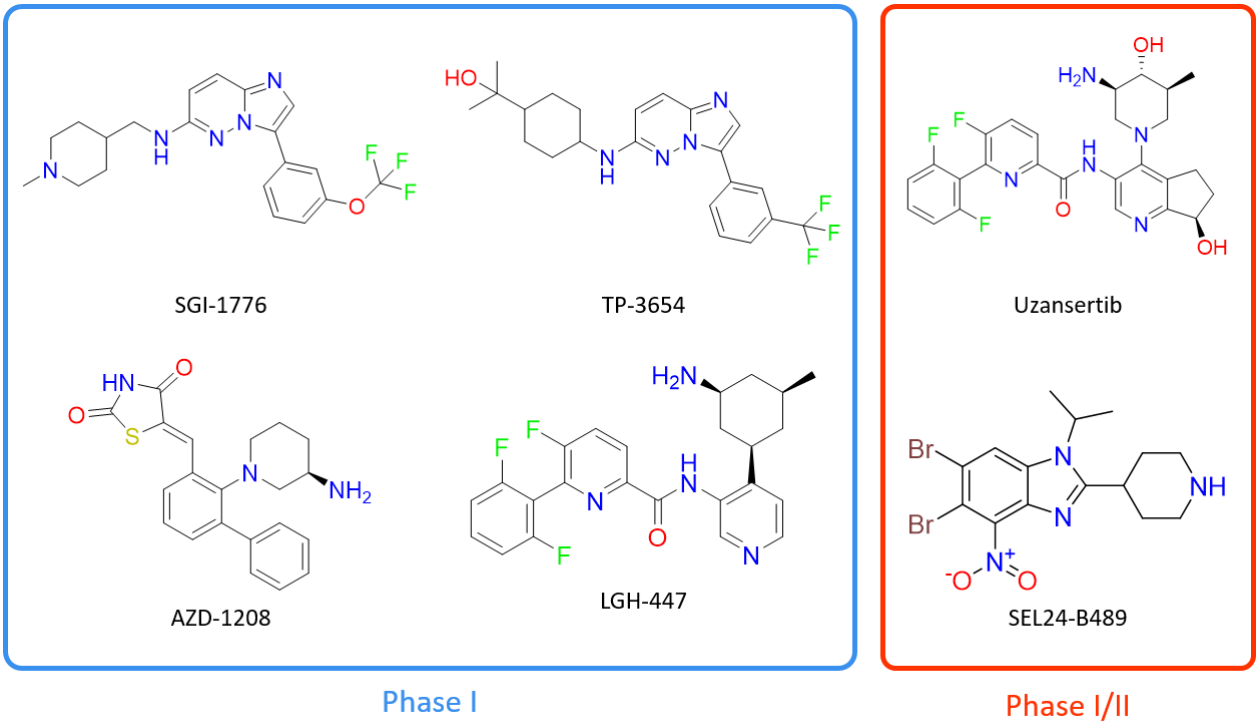


Figure 71 : Structures 2D des inhibiteurs des protéines Pim en phase I ou phase I/II des essais cliniques. D'après Oyallon et al.[332]

L'utilisation de F2D dans ce projet a été effectuée en deux temps, avec d'abord la recherche de petites molécules inhibitrices, puis une recherche de macrocycles.

4.4.2 Recherche de petites molécules inhibitrices

4.4.2.1 Graine utilisée

Le GICC a précédemment travaillé sur l'obtention d'inhibiteurs de Pim-1[332,333]. Ces molécules ont en commun le squelette acide quinoxaline-2-carboxylique qui me servira de graine pour obtenir de nouveaux inhibiteurs grâce à F2D (Figure 72).

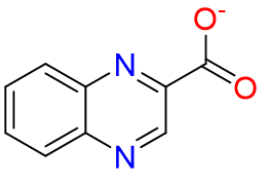


Figure 72 : Structure 2D de la acide quinoxaline-2-carboxylique.

Pour obtenir son emplacement 3D, j'ai effectué des simulations de *docking* en utilisant la protéine Pim-1 de l'ID PDB 3A99 comme cible. Pour effectuer la sélection des meilleures poses, j'ai utilisé un autre ID PDB (5NDT) contenant la protéine Pim-1 et une molécule ayant une sous-structure proche de la quinoxaline. J'ai observé que la molécule co-cristallisée dans l'ID PDB 5NDT forme une liaison hydrogène avec la lysine catalytique (Figure 73). J'ai sélectionné les 5 poses aux scores « SCORE.INTER » les plus faibles, en respectant la proximité avec la lysine catalytique dans l'ID PDB 3A99 (Figure 73).

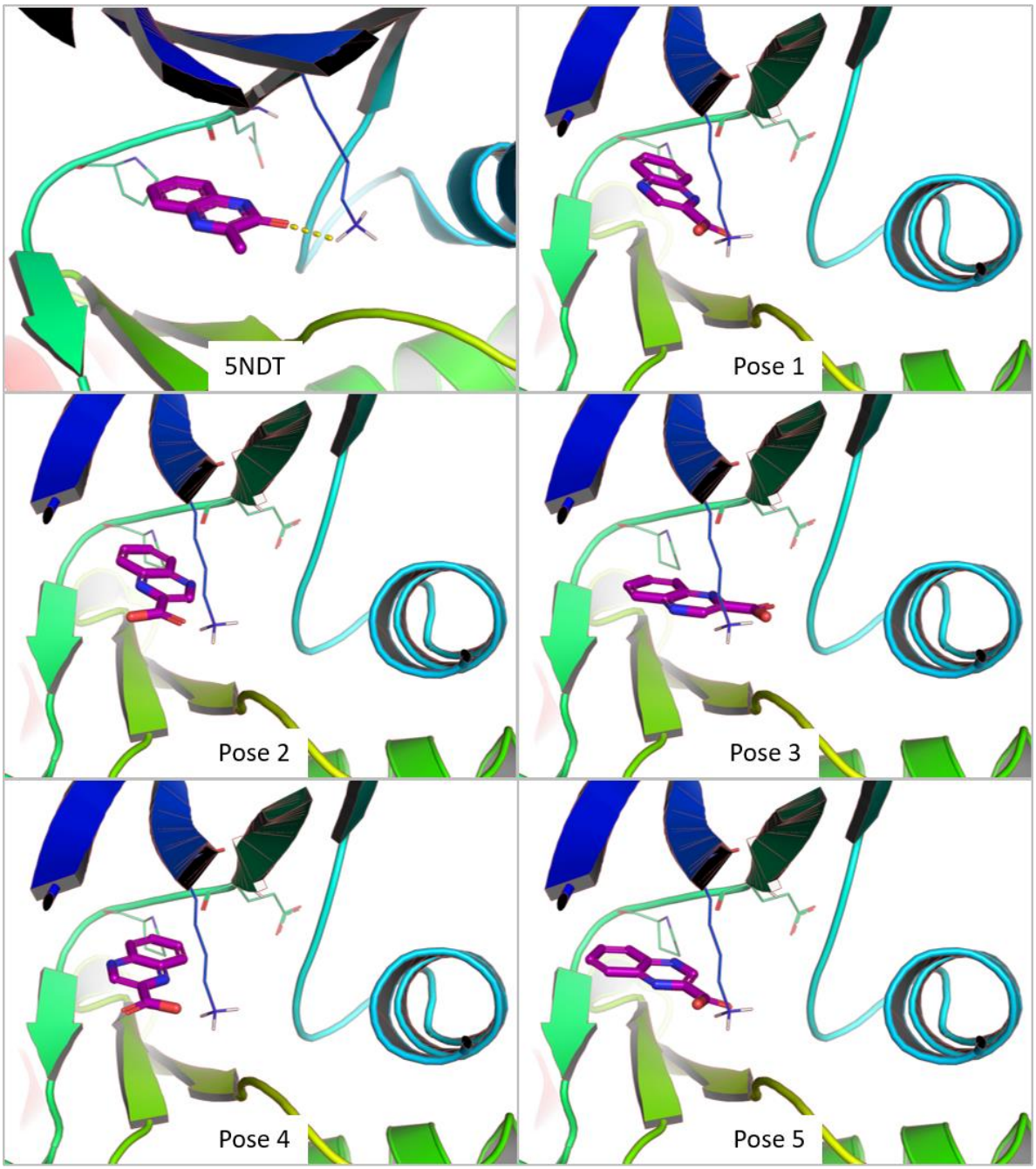


Figure 73 : Site actif du complexe PDB 5NDT et les cinq positions 3D obtenues et sélectionnées après simulations de docking dans Pim-1 (ID PDB : 3A99).

4.4.2.2 Agrandissement des fragments

J'ai ajouté les 5 poses en tant que graines à la BDD de F2D pour qu'elles puissent être agrandies en faisant des cycles d'agrandissement comme décrit en partie 4.2.3. J'ai obtenu les nombres de molécules indiqués dans le Tableau 21 jusqu'à l'ajout de 4 nouveaux fragments. Chaque cycle contient une sélection par les chimistes médicinaux impliqués dans le projet. J'ai pu poursuivre l'agrandissement à partir des molécules composées de 5 fragments jusqu'à la condition d'arrêt. J'ai sélectionné 26 molécules ayant un score de facilité de synthèse (score SA) < 4 parmi les 33 molécules constituées de cinq fragments. Puis, j'ai exécuté F2D pour faire une recherche exhaustive jusqu'à l'obtention d'au moins une molécule avec une masse molaire maximale de 650 g/mol. Nous allons présenter les

descripteurs moléculaires des molécules obtenues, puis observer comment l'application de filtres basés sur le score SA et le *redocking* améliorent ces descripteurs.

Tableau 21 : Nombre de petites molécules inhibitrices de Pim-1 obtenues par F2D et sélectionnées.

Nombre de fragments	Nombre de molécules obtenues ou initiales	Nombre de molécules sélectionnées
1	5	5
2	15	4
3	30	14
4	108	5
5	33	26

Dans ce troisième projet, F2D a obtenu toutes les solutions possibles. Nous pouvons alors analyser les molécules afin de sélectionner les meilleures. J'ai obtenu 12 284 structures 3D pour 8 225 molécules uniques composées de 6 à 12 fragments. La Figure 74 montre la distribution des descripteurs moléculaires calculés. La distribution du nombre d'atomes chiraux n'est pas montrée car seules trois molécules ont trois atomes chiraux.

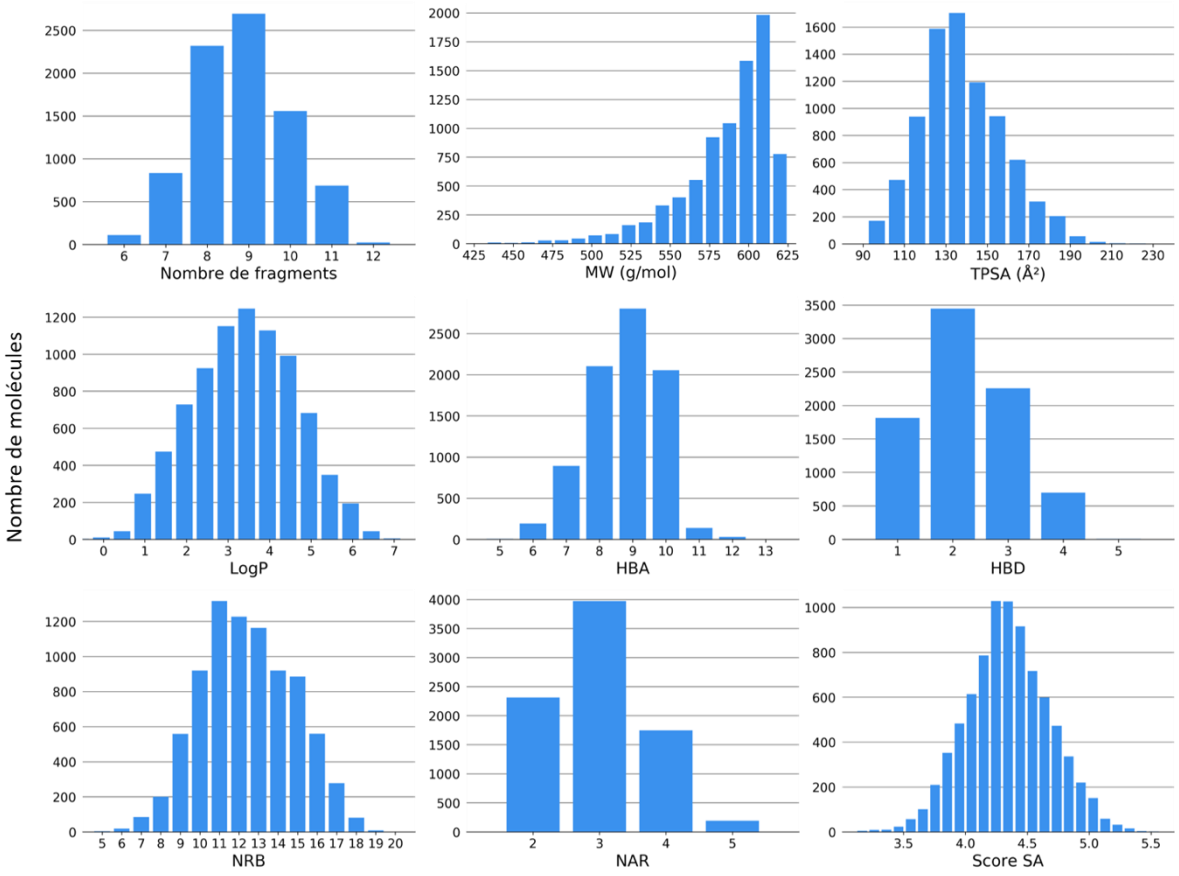


Figure 74 : Distributions des descripteurs moléculaires calculés sur les 8 225 molécules sélectionnées après l'exécution de F2D sur Pim-1.

Nous pouvons comparer les valeurs moyennes obtenues pour les descripteurs moléculaires des molécules générées par F2D aux valeurs des 283 PKI provenant de *Protein Kinase Inhibitor Database* (PKIDB)[34,35] (Tableau 22).

Tableau 22 : Comparaison entre les valeurs moyennes des descripteurs moléculaires de différent jeux de molécules. Le premier jeu de molécules est composé de PKI sur le marché ou en cours d'essais cliniques (PKIDB), le suivant des 8 225 molécules générées par F2D et le troisième des 1 196 molécules générées par F2D ayant un score de facilité de synthèse < 4.

Descripteur moléculaire	PKIDB	F2D	F2D (score SA < 4)
Nombre de molécules	283	8 225	1 196
MW (g/mol)	458,25 (± 76,68)	586,38 (± 29,27)	564,28 (± 40,01)
TPSA (Å²)	96,35 (± 20,31)	138,11 (± 19,66)	130,05 (± 20,79)
LogP	3,36 (± 1,36)	3,35 (± 1,24)	3,45 (± 1,32)
HBA	8,73 (± 1,86)	8,75 (± 1,10)	8,21 (± 1,20)
HBD	2,02 (± 0,88)	2,22 (± 0,89)	1,79 (± 0,74)
NRB	5,75 (± 2,41)	12,44 (± 2,38)	11,30 (± 2,34)
NAR	3,45 (± 1,00)	2,98 (± 0,76)	3,26 (± 0,67)
NCA	0,50 (± 0,90)	0 (± 0,0)	0 (± 0,0)
Score SA	3,10 (± 0,62)	4,34 (± ,033)	3,83 (± 0,14)

Nous observons que MW, TPSA et NRB sont en moyenne plus élevés pour les molécules de F2D que pour les molécules de PKIDB. Cela s'explique par le fait que nous ayons choisi comme point d'arrêt une masse molaire maximale et par le parcours de graphe sur lequel repose F2D. Le logiciel F2D génère de nombreuses solutions à chaque nouvel ajout de fragment et augmente ainsi le nombre de molécules obtenues avec des MW, TPSA et NRB élevés. La valeur moyenne de NRB est aussi plus élevée, en moyenne supérieure de 7 unités par rapport au NRB moyen des molécules PKIDB. Le NRB pourrait être utilisé comme une limite lors de l'agrandissement fait par F2D. F2D fournit des molécules qui ne sont pas optimisées pour être de bons PKI. Pour effectuer l'optimisation des molécules, l'expertise de chimistes médicinaux sera nécessaire et l'utilisation de la chémoinformatique pourra orienter les décisions. Selon la comparaison entre les valeurs moyennes des descripteurs de PKIDB et celles des molécules obtenues, l'optimisation devrait viser à réduire ce NRB. D'autres pistes pour l'optimisation des molécules F2D seraient de diminuer le score SA et d'augmenter le NAR. Les autres descripteurs : LogP, HBA, HBD et NCA sont assez similaires en moyenne entre les deux jeux de molécules.

En retirant les molécules contenant des sous-structures PAINS, nous enlevons 182 molécules et en obtenons 8 043. En conservant les molécules ayant un score SA < 4, nous gardons 1 196 molécules. Les valeurs moyennes des descripteurs calculés pour ces 1 196 molécules restent proches des valeurs moyennes des descripteurs des molécules PKIDB (Tableau 22). Bien que les valeurs pour MW, TPSA et NRB restent trop élevées, celles de NAR et de score SA tendent vers celles des PKI sur le marché ou en cours d'essais cliniques. Un filtre plus drastique peut être appliqué sur ces molécules comme la vérification des interactions formées entre les molécules et le site actif de Pim-1.

J'ai ensuite calculé les poses de ces molécules par simulations de *docking*. J'ai supprimé les molécules ayant un RMSD ≥ 3,5 Å entre la position 3D de la molécule issue de F2D et la pose la plus proche. Cette valeur de seuil de RMSD a été déterminée lors du *redocking* de tous les PKI provenant de la *Protein Databank* (PDB), selon lequel les molécules ayant un RMSD <2,5 Å sont à privilégier et celles avec un RMSD > 3,5 Å sont à supprimer. J'ai donc conservé 102 molécules uniques en deux groupes : « groupe 1 » celles ayant un RMSD < 2,5 Å (34 molécules) et « groupe 2 » celles ayant un RMSD compris entre 2,5 Å et 3,5 Å (68 molécules, Figure 75).

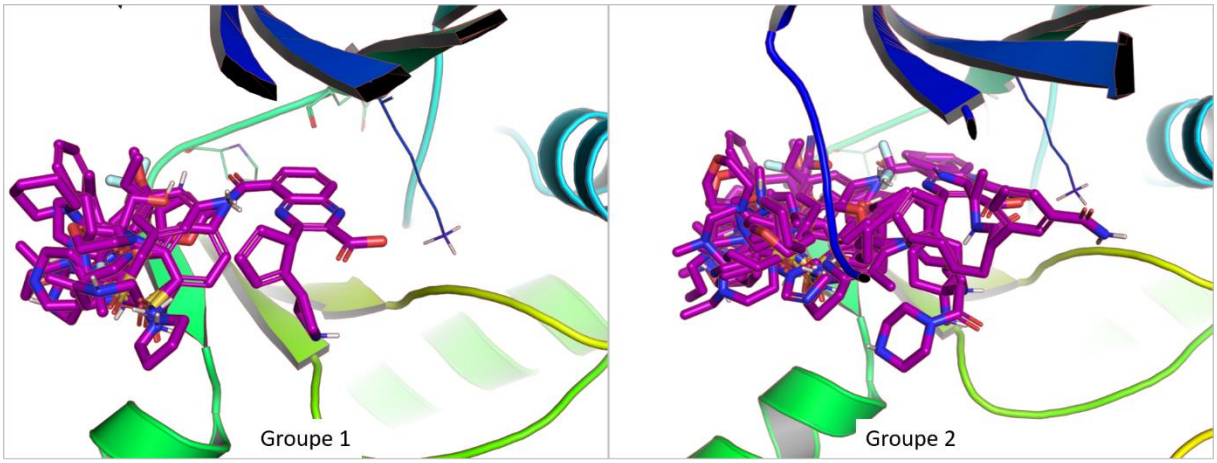


Figure 75 : Superposition 3D des molécules obtenues et sélectionnées au cours de l'application de F2D sur Pim-1.

Les valeurs moyennes obtenues pour les descripteurs moléculaires calculés sur les deux groupes de molécules se rapprochent des valeurs moyennes venant de PKIDB (Tableau 23). Cette tendance se confirme particulièrement pour le NRB qui est diminué par rapport à l'ensemble des molécules obtenues avec un score SA < 4.

Tableau 23 : Valeurs moyennes des descripteurs moléculaires des molécules F2D sélectionnées après simulations de docking dans la protéine Pim-1.

Descripteur moléculaire	Groupe 1	Groupe 2
Nombre de molécules	34	68
MW (g/mol)	511,90 (± 58,19)	531,52 (± 52,61)
TPSA (Å^2)	143,85 (± 30,90)	143,08 (± 29,01)
LogP	1,77 (± 0,95)	1,93 (± 1,08)
HBA	8,09 (± 1,31)	8,35 (± 1,42)
HBD	2,65 (± 0,69)	2,44 (± 0,78)
NRB	7,68 (± 1,85)	8,48 (± 1,79)
NAR	3,06 (± 0,24)	2,92 (± 0,35)
NCA	0 (± 0,0)	0 (± 0,0)
Score SA	3,81 (± 0,13)	3,83 (± 0,11)

En gardant les molécules ayant des RMSD < 3,5 Å, nous conservons plus de molécules qu'avec un RMSD < 2,5 Å. La distinction entre les deux groupes de molécules, basée sur le RMSD, permet de privilégier celles ayant un RMSD < 2,5 Å qui sont considérées comme meilleures. Toutes les molécules du groupe 1 comprennent un groupement fonctionnel azétidine. En voyant ce groupement, les chimistes médicinaux impliqués dans le projet n'ont pas envisagé de synthétiser ces molécules à cause de sa synthèse trop compliquée. La Figure 76 montre les structures 2D des 6 molécules du groupe 1 ayant les meilleurs scores *Quantitative estimate of druglikeness* (QED).

En observant les structures des molécules du groupe 2, nous en avons trouvé certaines qui ne contenaient pas de groupement azétidine. Le groupe 2 contient 59 molécules uniques, parmi lesquelles 13 structures ne contiennent pas d'azétidine. La Figure 77 montre les structures 2D des 7 molécules du groupe 2 ne contenant pas d'azétidine et ayant un score QED > 0,5. Aujourd'hui, ces molécules sont en cours d'observation par les chimistes médicinaux qui vont pouvoir en sélectionner certaines pour les synthétiser.

À la suite de la recherche de petites molécules pour inhiber Pim-1, nous avons décidé de nous focaliser sur la recherche de macrocycles pour cette cible. Bien qu'ils soient plus complexes à synthétiser, les macrocycles peuvent avoir une meilleure sélectivité que les autres composés sur les protéines kinase grâce à leur rigidité[334].

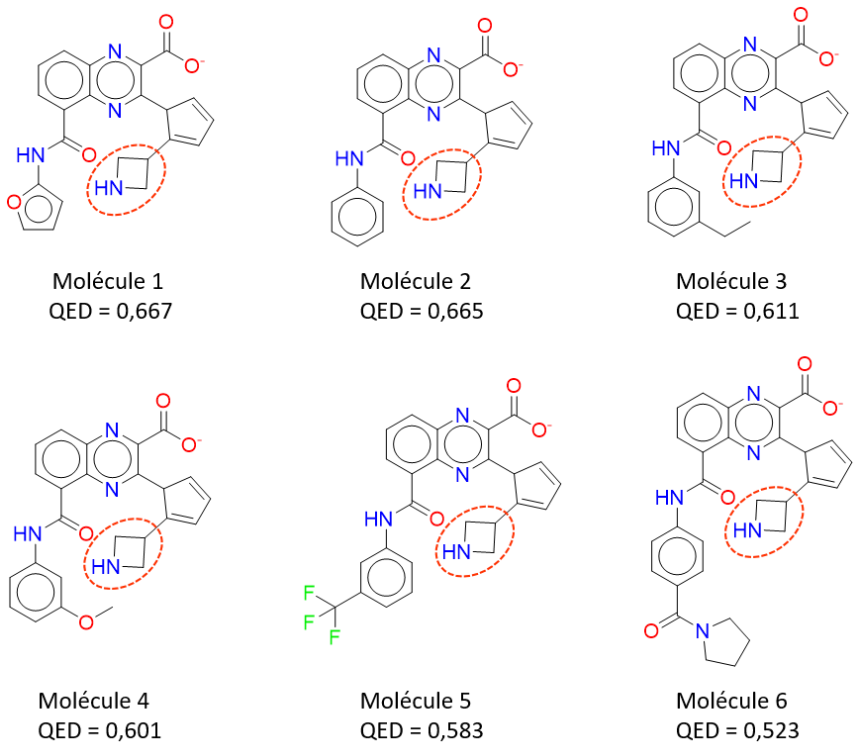


Figure 76 : Structures 2D des 6 molécules du groupe 1 ayant un score QED > 0,5. Le groupement azétidine qui pose problème pour la synthèse des molécules est entouré en rouge.

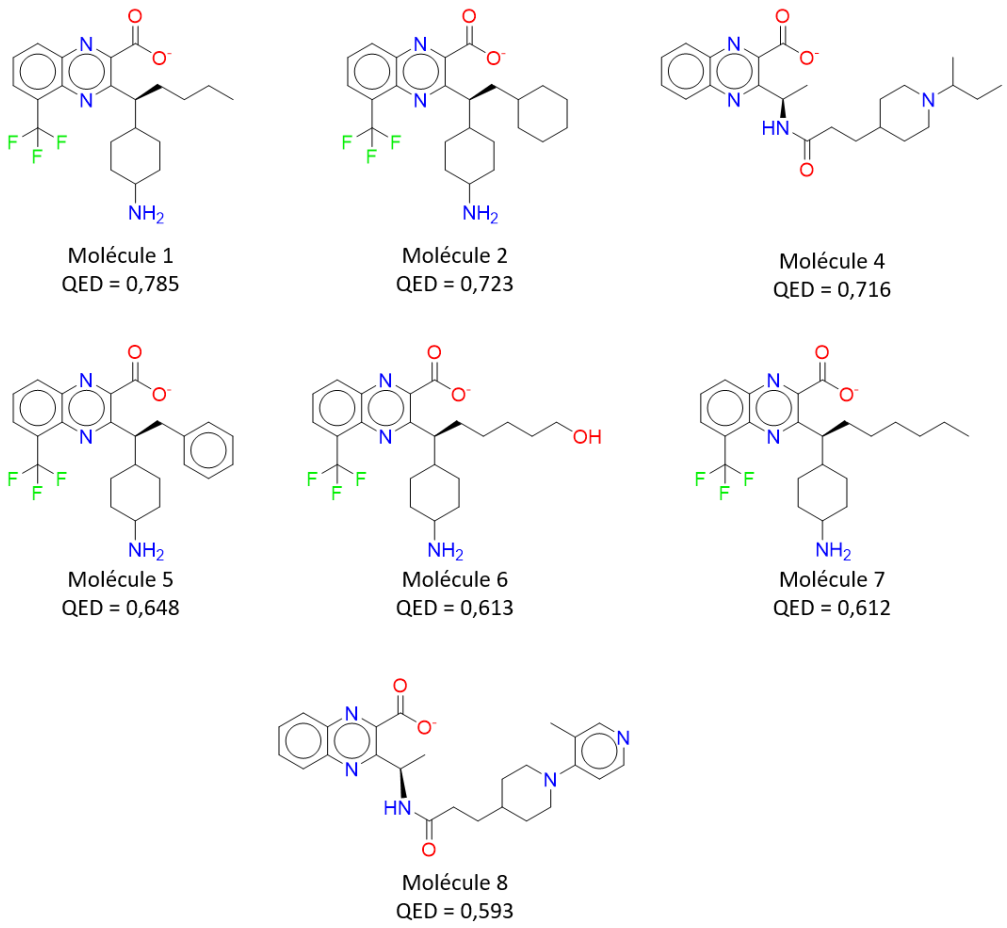


Figure 77 : Structures 2D des 7 molécules du groupe 2 ne contenant pas de groupement azétidine et ayant un score QED > 0,5.

4.4.3 Recherche de macrocycles

4.4.3.1 Graines utilisées

J'ai utilisé F2D pour proposer des structures de macrocycles pouvant inhiber les 3 protéines Pim, appelés « Pan-Pim ». La synthèse d'inhibiteurs macrocycliques « Pan-Pim » est un des objectifs du projet de thèse de Camille Blouet, doctorante au GICC depuis 2020. Nous avons dessiné 4 graines de départ à partir de l'acide quinoxaline-2-carboxylique qui ne peuvent pas être montrées pour des raisons de confidentialité.

La protéine cible utilisée est la même que pour la recherche de petites molécules, Pim-1 à partir de l'ID PDB 3A99. Pour trouver des inhibiteurs « Pan-Pim » alors que nous utilisons Pim-1 comme cible pour l'agrandissement, nous pourrons sélectionner les macrocycles formés de fragments ayant des relations de compatibilité avec Pim-2 et Pim-3. Cependant, la protéine Pim-3 n'a pas de structure 3D enregistrée dans la PDB au moment de la rédaction de cette thèse. Un modèle par homologie ou le programme Alpha Fold[335] pourraient pallier ce problème en donnant une structure 3D pour cette protéine.

Pour pouvoir lancer F2D à partir de ces quatre graines, j'ai effectué une simulation de *docking* exhaustive. J'ai conservé les 5 poses aux scores « SCORE.INTER » les plus faibles et je les ai observées de manière à en sélectionner une pour chaque graine. J'ai sélectionné 4 poses dans lesquelles l'acide carboxylique forme une interaction hydrogène avec la lysine catalytique. Dans ces 4 poses, le benzène de la quinoxaline s'oriente lui vers la région charnière. Après avoir sélectionné les poses de ces 4 graines, j'ai effectué un agrandissement spécifique en ne me focalisant que sur l'obtention de macrocycles.

4.4.3.2 *Agrandissement de fragments*

J'ai ajouté les 4 poses retenues à la BDD de fragments de F2D et calculé les relations d'inclusion et d'exclusion qu'elles forment avec les autres fragments déjà présents. J'ai aussi vérifié l'absence d'encombrement stérique, en calculant les relations de compatibilité avec le site actif de Pim-1.

Pour obtenir uniquement des macrocycles lors de l'agrandissement, j'ai modifié l'utilisation classique du logiciel F2D. Habituellement, F2D aurait cherché dans la BDD orientée graphes tous les chemins possibles à partir des graines pour générer les molécules et il aurait fallu sélectionner *a posteriori* les macrocycles parmi les résultats obtenus. Comme évoqué dans le chapitre 2 partie 2.1.3, pour la recherche de macrocyles par F2D, j'ai forcé l'obtention de macrocycles grâce à la librairie Python Networkx (version 2.2) et à la fonction *cycle_basis()*. Cette fonction a été créée à partir de l'algorithme CACM 491[336] et permet de trouver les cycles formés par des ensembles de nœuds reliés par des arêtes permettant de revenir au nœud d'origine. Ces cycles formés par des nœuds du graphe pourront correspondre à des macrocycles, lors de leur conversion en molécules, s'ils contiennent au moins 11 atomes.

Dans cette utilisation de F2D, nous ne procédons pas par différentes étapes d'ajout d'un fragment puis de sélection, nous obtenons directement les macrocycles. F2D a trouvé 205 macrocycles composés d'un cycle contenant de 11 à 15 atomes, macrocycles qui ont ensuite été filtrés selon les valeurs de leurs descripteurs moléculaires. Parmi les 205 macrocycles, 176 sont composés de 3 fragments et 29 sont constitués de 4 fragments. Aucun des macrocycles obtenus ne contient d'atome chiral. La Figure 78 montre les distributions des autres descripteurs moléculaires qui ont été calculés.

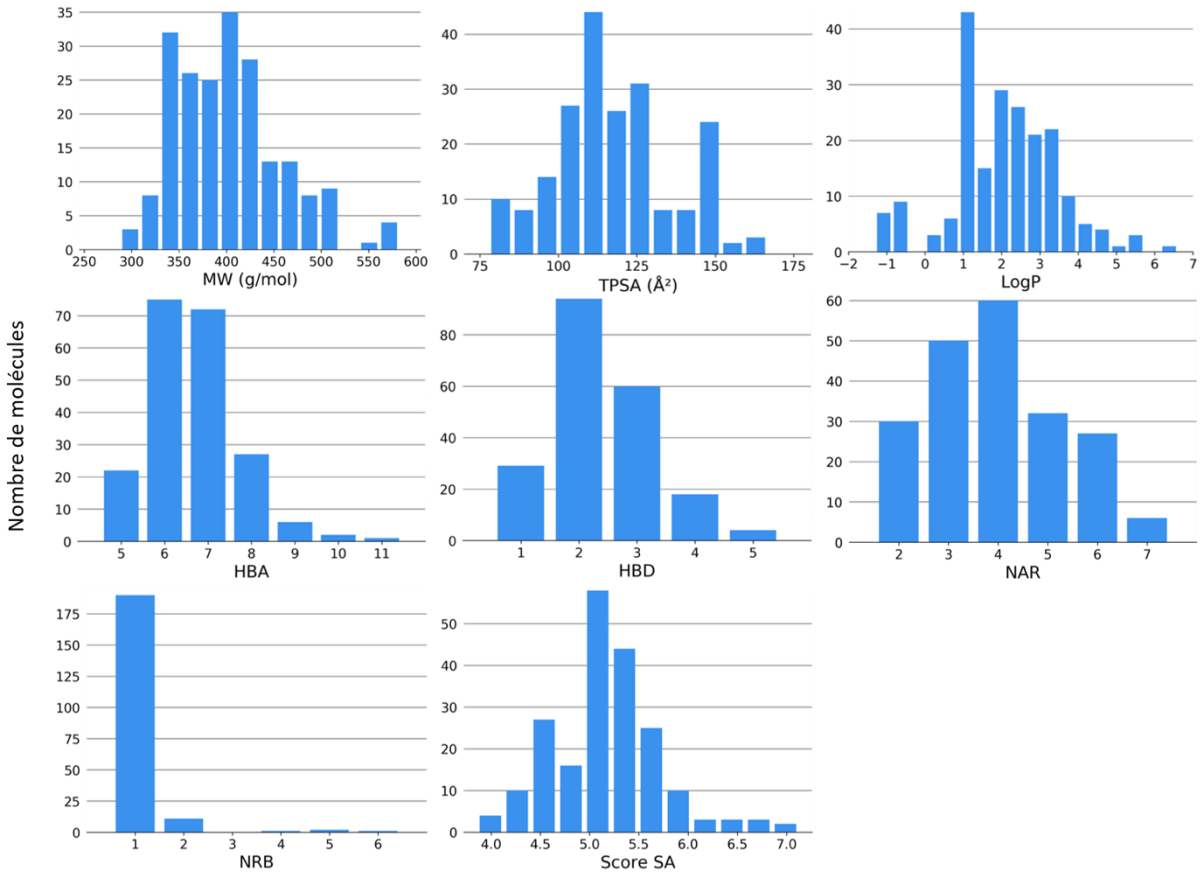


Figure 78 : Distributions des descripteurs moléculaires calculés sur les 205 macrocycles obtenus dans Pim-1.

Aucun des macrocycles obtenus ne contient de sous-structure PAINS. Il y a une seule liaison rotable pour 190 macrocycles sur les 205. Ce nombre correspond à la moyenne des NRB des 7 macrocycles provenant de PKIDB (1,14). Cette valeur de NRB est plus faible que la valeur moyenne de tous les inhibiteurs de PKIDB (5,75) et confirmerait que les macrocycles sont moins flexibles que les PKI classiques. Pour vérifier cette hypothèse, il faudrait faire une recherche conformationnelle sur ces macrocycles. Les valeurs moyennes des autres descripteurs moléculaires des macrocycles sont proches des valeurs calculées sur les macrocycles de PKIDB (Tableau 24), à part la moyenne de TPSA qui est supérieure à celle des macrocycles de PKIDB.

Le score SA est plus élevé en moyenne de 0,57 dans les 205 macrocycles générés par F2D en comparaison avec les macrocycles de PKIDB. Pour rappel, lors de la sélection des macrocycles basées sur le score SA, nous utilisons la valeur seuil de 5 au lieu de 4 habituellement (Chapitre 2 partie 2.1.3, Figure 26). Ainsi, à partir des 205 macrocycles, nous en sélectionnons 61 avec un score SA < 5.

Tableau 24 : Valeurs moyennes des descripteurs moléculaires calculés sur les 205 macrocycles obtenus dans Pim-1

Descripteur moléculaire	Molécules linéaires dans PKIDB	Macrocycles dans PKIDB	Macrocycles obtenus dans Pim-1
Nombre de molécules	276	7	205
MW (g/mol)	459,57 (± 76,89)	406,47 (± 46,31)	402,55 (± 56,36)
TPSA (Å²)	96,88 (± 20,12)	75,47 (± 18,05)	117,38 (± 18,69)
LogP	3,86 (± 1,36)	3,5 (± 0,98)	2,04 (± 1,38)
HBA	6,74 (± 1,88)	6,29 (± 0,76)	6,65 (± 1,05)
HBD	2,04 (± 0,88)	1,14 (± 1,57)	2,38 (± 0,90)
NAR	5,86 (± 1,01)	3,14 (± 0,38)	3,97 (± 0,60)
Score SA	3,06 (± 0,57)	4,61 (± 0,69)	5,18 (± 0,54)

J'ai ensuite comparé les poses prédites par simulations de *docking* pour ces 61 macrocycles à leurs positions 3D venant de l'assemblage des fragments de F2D. Dans le cas des simulations de *docking* de macrocycles, il est nécessaire d'effectuer une recherche conformationnelle avant de procéder au *docking*. En effet, le logiciel rDock ne permet pas de prendre en compte les différentes conformations possibles pour chaque macrocycle lorsqu'il génère ses coordonnées 3D. Pour cela, j'ai utilisé le logiciels MOE (version 2019.0101). MOE a permis de générer plusieurs conformations pour les macrocycles (avec la méthode *LowModeMD*) et rDock a été utilisé pour obtenir 50 poses par conformations de macrocycles obtenues. J'ai sélectionné les poses avec les RMSD les plus faible vis-à-vis des positions 3D des molécules provenant de F2D et obtenu 24 macrocycles uniques avec un RMSD < 3.5 Å. Ces molécules ont été présentées aux chimistes médicinaux qui en ont retenu 3 et envisagent aujourd'hui leur synthèse.

4.4.4 Bilan du projet LabEx GICC

Le projet de recherche d'inhibiteurs de Pim effectué en collaboration avec le GICC de Tours est, au moment de la rédaction de cette thèse, le plus abouti des 3 projets LabEx dans lesquels F2D a été utilisé. Grâce à l'utilisation de F2D et à l'interaction avec les chimistes médicinaux impliqués dans le projet, j'ai pu découvrir différents inhibiteurs potentiels de la protéine Pim-1. Ces inhibiteurs se classent en deux catégories : petites molécules ou macrocycles. L'équipe IMT du GICC passe actuellement en

revue les 13 petites molécules parmi les 156 qui ont été trouvées. De plus, les chimistes médicinaux envisagent la synthèse de 3 macrocycles parmi les 24 qui ont été découverts dans ce projet.

Le but initial de ce projet était d'obtenir des inhibiteurs des trois protéines Pim, mais F2D n'a été appliqué que sur la cible Pim-1. Il faut noter qu'en exécutant F2D sur les 2 autres cibles, d'autres possibilités de résultats pourraient être obtenues. Par la suite, je pourrai effectuer une estimation de la sélectivité des molécules obtenues dans chaque protéine Pim vis-à-vis des 3 protéines, cela donnera des inhibiteurs « Pan-Pim ».

En plus d'avoir utilisé F2D au sein de 3 projets de recherche de PKI du LabEx SynOrg, j'ai aussi participé à un projet concernant des sondes fluorescentes en collaboration avec une autre équipe de recherche de l'institut de chimie organique et analytique (ICOA) d'Orléans.

4.5 Recherche de sondes fluorescentes

La quatrième et dernière utilisation de Frags2Drugs (F2D) qui sera présentée au cours de ma thèse a été effectuée en collaboration avec le groupe « Chimie hétéroaromatique », dirigé par le Pr Franck Suzenet, de l'équipe « Hétérocycles, Nucléosides et Agents d'Imagerie » (HNAI) de l'ICOA.

4.5.1 Contexte biologique

Au cours de ce projet, nous souhaitons découvrir des sondes fluorescentes pour les protéines *Aurora kinase A* (AurKA) et *Polo-like kinase 1* (PLK1). Ces sondes fluorescentes sont de petites molécules qui permettront de pouvoir localiser ces deux protéines grâce à leur fluorescence lors de la réalisation d'expériences d'imagerie optique. Lorsque les sondes n'émettent pas de fluorescence elles-mêmes, mais doivent être couplées à un fluorochrome nous parlons de « sondes chimiques ».

4.5.1.1 AurKA

L'activation de la protéine AurKA contrôle différents évènements de la mitose tels que la maturation du centrosome, l'entrée en mitose, la formation du fuseau mitotique ou la cytodiérèse[337]. Bien qu'AurKA soit principalement étudiée pour ses rôles en lien avec la mitose, cette protéine est aussi impliquée dans l'organisation des microtubules, la migration et la polarité cellulaire[337]. AurKA est aussi impliquée dans la résorption du cil primaire[338] et le contrôle de la concentration de calcium intracellulaire[339].

En cas de surexpression d'AurKA différents types de cancers peuvent survenir, tels que le cancer colorectal, du sein, de l'ovaire, du col de l'utérus, de la prostate ou le neuroblastome[337]. AurKA est une cible thérapeutique pour laquelle de nombreux inhibiteurs ont été développés. Parmi ces inhibiteurs, certains sont des sondes chimiques et d'autres ont atteint la troisième phase des essais cliniques (Figure 79).

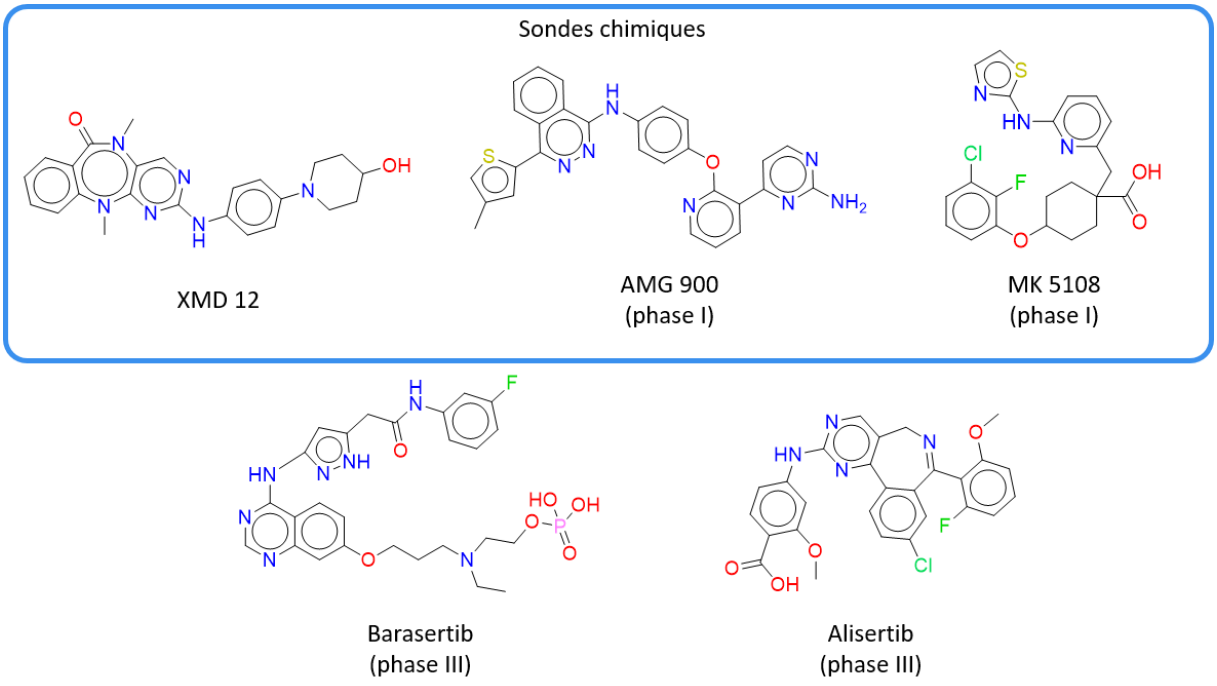


Figure 79 : Sondes chimiques et deux inhibiteurs d'AurKA ayant atteint la troisième phase des essais cliniques

Au 12 décembre 2019, 148 structures 3D existaient pour la protéine AurKA dans la *Protein Databank* (PDB). Nous avons décidé de nous concentrer sur l'ID PDB 6C2T qui contient AurKA en complexe avec l'inhibiteur LY3295668 (Figure 80). Ce choix a été réalisé en conservant les structures 3D contenant un ligand co-cristallisé, non allostérique et en triant les complexes PDB par résolution pour ne sélectionner que le complexe avec la plus faible résolution.

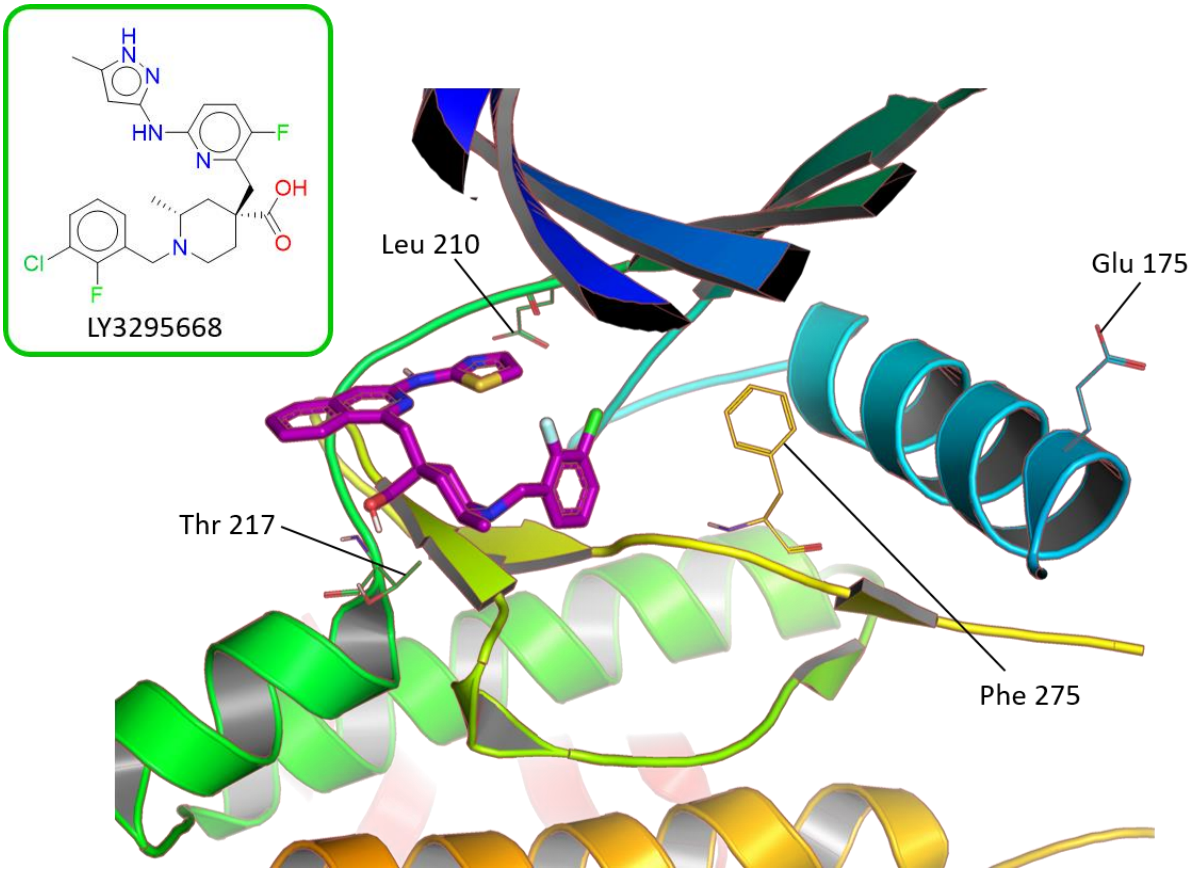


Figure 80 : Structure 3D du site actif de la protéine AurKA contenant la molécule LY3295668 (ID PDB : 6C2T).

Dans cette structure, nous observons grâce aux orientations de la Phe 275 et du Glu 175 que la protéine AurKA a une conformation DFG inter, hélice αC in. De plus, il apparaît que le résidu *gatekeeper* (Leu 210), de par sa chaîne latérale à faible encombrement stérique et son orientation, laisse un accès pour l'inhibiteur à la poche ATP. La Thr 217 forme une interaction importante avec l'acide carboxylique du ligand LY3295668 permettant à cette molécule d'être sélective d'AurKA vis-à-vis d'AurKB[340].

4.5.1.2 PLK1

Comme AurKA, la protéine PLK1 est aussi impliquée dans le cycle cellulaire et particulièrement dans la mitose. PLK1 agit sur la maturation du centrosome, la régulation du complexe de promotion de l'anaphase, sur la séparation des chromosomes, la formation du fuseau mitotique, la cytodiérèse et la régulation du contrôle des anomalies de réplication de l'ADN[341].

Les implications pathologiques de PLK1 sont similaires à celles d'AurKA avec une surexpression dans de nombreux cancers dont ceux du sein, des poumons, colorectal, des ovaires, du pancréas, de la prostate et ceux de la tête et du cou[342]. Pour mieux comprendre le rôle de PLK1 dans l'apparition de ces cancers, des agents d'imagerie peuvent être utilisés pour mettre cette protéine en évidence et suivre son évolution. En effet, PLK1 est un marqueur permettant de détecter l'apparition de cancers[343]. Certaines molécules ont déjà été développées et testées pour inhiber PLK1 telles que BI 2536, volasertib

(BI 6727) et rigosertib (Figure 81). En modifiant le squelette de BI 2536 avec un trans-cyclooctene, Budin et al. ont créé une sonde fluorescente pour PLK1 appelée BI 2536-TCO[344].

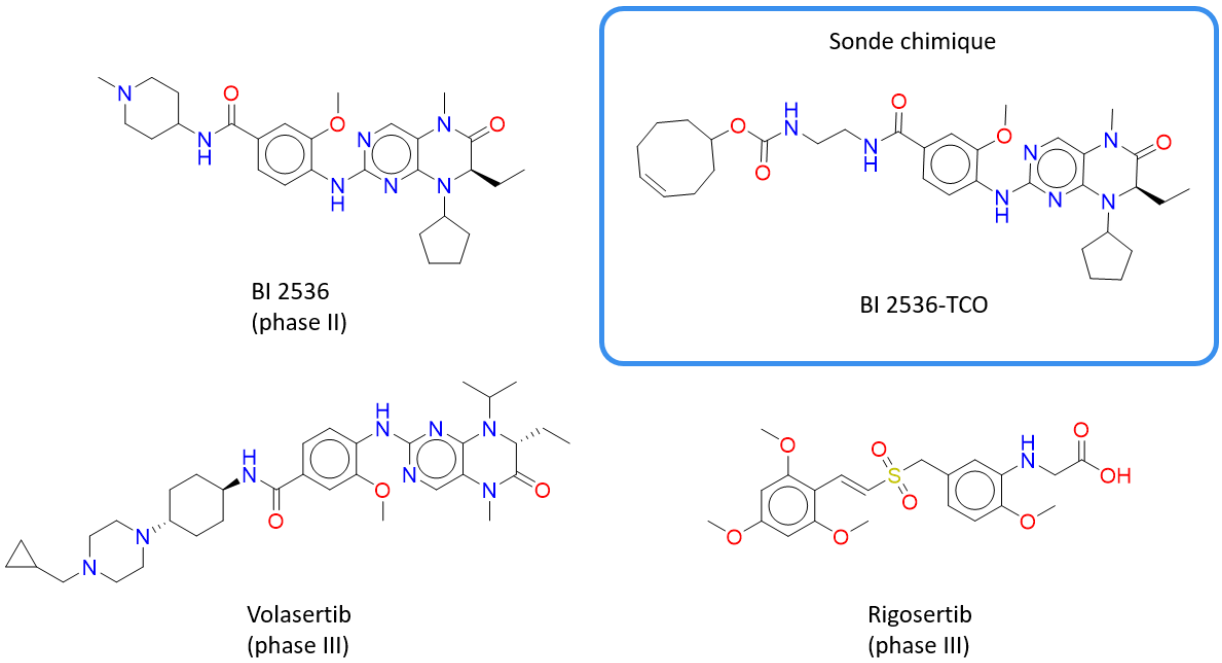


Figure 81 : Sondes chimiques et inhibiteurs de PLK1. Les 3 molécules sont en phase III des essais cliniques.

Au 11 décembre 2019, 14 structures 3D étaient recensées dans la PDB pour PLK1. Nous avons choisi l'ID PDB 2RKU contenant PLK1 avec le ligand BI2536. La structure 2D de ce ligand est présentée dans la Figure 81 et la Figure 82 montre la structure 3D de l'ID PDB 2RKU.

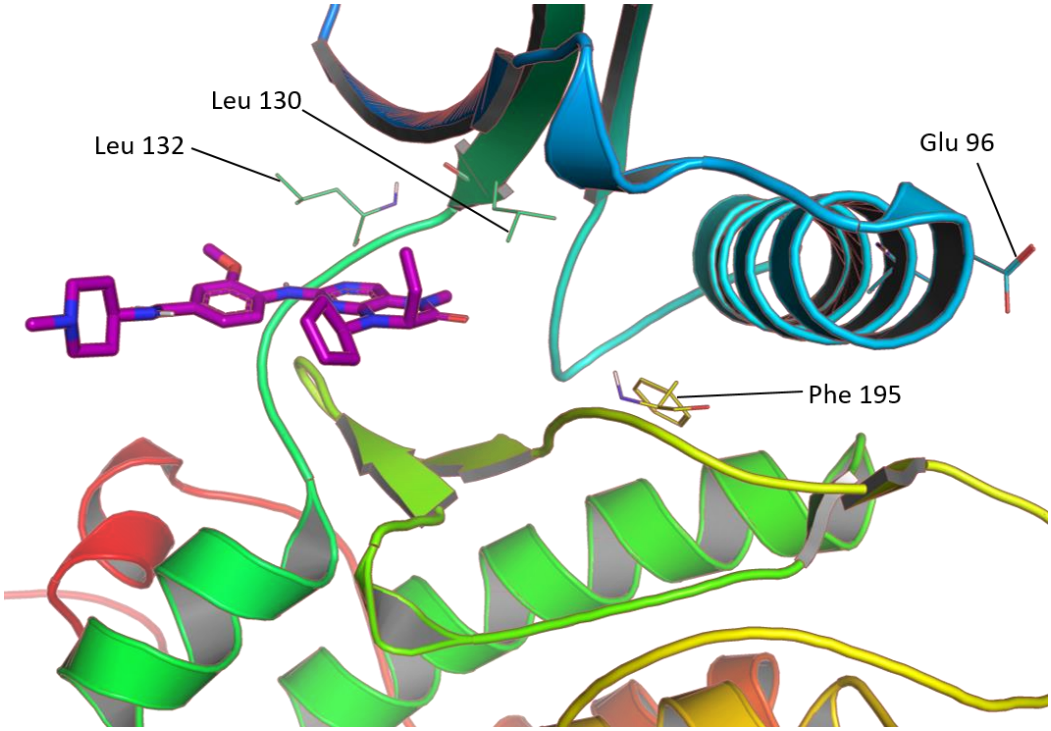


Figure 82 : Structure 3D du site actif de la protéine PLK1 en complexe avec la molécule BI 2536 (ID PDB : 2RKU).

La conformation du site actif de PLK1 est DFG in, hélice αC in. Nous pouvons observer cette conformation grâce à l'orientation des acides aminés Phe 195 et Glu 96 dans l'ID PDB 2RKU. Contrairement à ce que nous pouvions observer dans AurKA, le résidu *gatekeeper* (Leu 130) bloque l'entrée de la poche catalytique par l'orientation de sa chaîne latérale. La Leu 132 permet une interaction avec le groupement methoxy du ligand BI 2536, rendant cet inhibiteur spécifique des PLKs par rapport aux autres protéines kinases[345].

4.5.2 Graines utilisées

Le groupe « chimie hétéroaromatique » de l'ICOA a précédemment publié plusieurs molécules fluorescentes tricycliques dont deux molécules composées de diazines fusionnées à un triazapentalène[346]. Ce sont les deux graines que j'utiliserai pour la découverte de sondes fluorescentes dans AurKA et PLK1. Les deux cycles qui sont fusionnés au triazapentalène sont une pyrimidine (graine 1) et une pyrazine (graine 2, Figure 83).

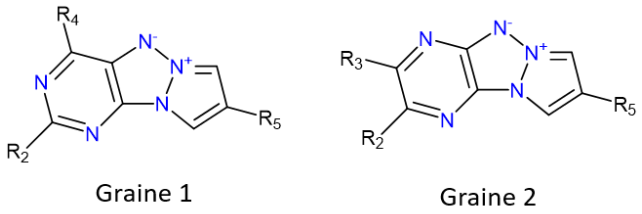


Figure 83 : Structures 2D des deux graines utilisées pour la découverte de sondes fluorescentes dans AurKA et PLK1.

Pour pouvoir être exécuté, F2D requiert de connaître les positions 3D des graines à agrandir. Je les ai obtenues par simulation de *docking* pour les deux graines dans les protéines PLK1 et AurKA. Pour chaque protéine et chaque graine, comme dans les projets précédents, j'ai obtenu les poses de ces graines par simulations de *docking* exhaustives. J'ai trié ces poses grâce au score « SCORE.INTER » du logiciel pour privilégier celles avec le score le plus faible. J'ai ensuite sélectionné visuellement les poses en prenant en compte leur proximité avec la région charnière des protéines cibles et l'orientation des 3 cycles. J'ai ainsi retenu 8 poses pour les 2 graines dans PLK1 et 27 dans AurKA. Pour plus de clarté, je n'illustrerai ces poses que par les 4 qui ont permis *in fine* d'obtenir des molécules à l'issue de l'agrandissement par F2D (Figure 84).

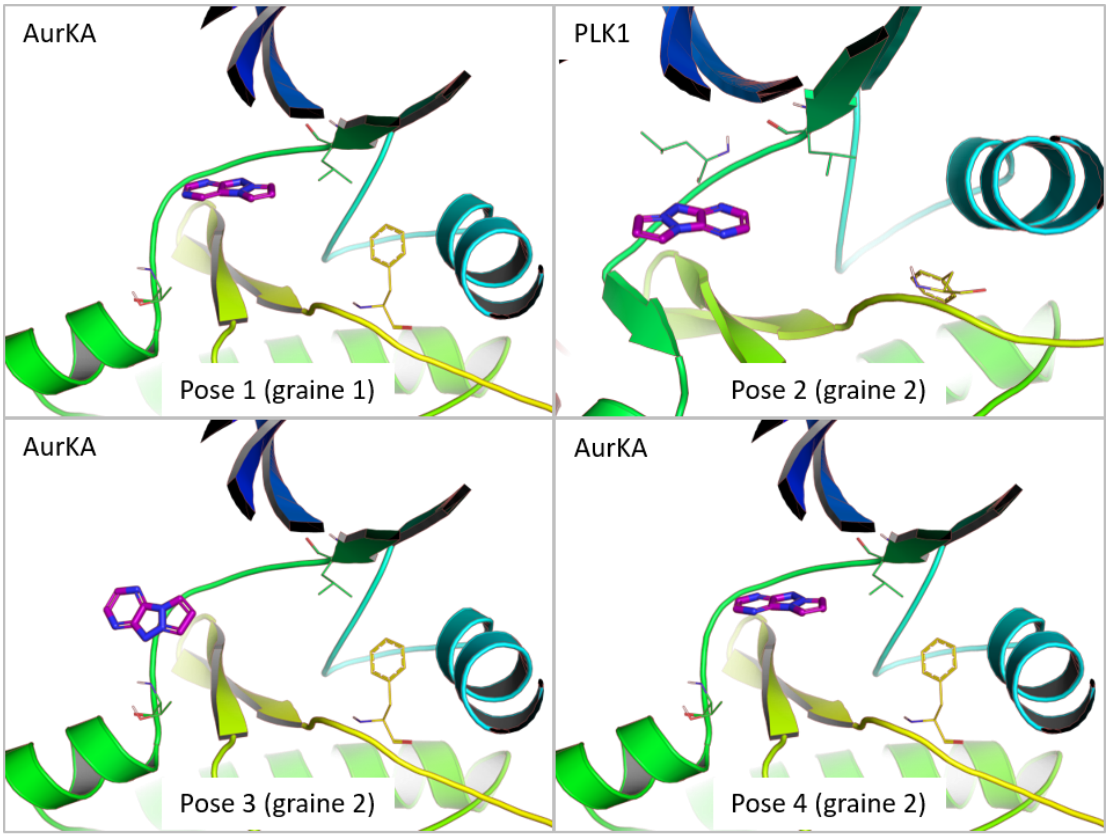


Figure 84 : Quatre poses ayant permis l'agrandissement de fragments par F2D parmi les 35 poses initiales pour les deux graines. Trois poses ont été obtenues dans la protéine AurKA (ID PDB : 6C2T) et la quatrième a été obtenue dans la protéine PLK1 (ID PDB : 2RKU).

4.5.3 Agrandissement des fragments

Pour agrandir les 2 graines dans les 2 protéines AurKA et PLK1, j'ai suivi la méthode par cycles (Partie 4.2.3), à la différence que la masse molaire maximale n'a pas été de 650 g/mol mais de 450 g/mol. Ce choix a été fait en concertation avec les chimistes impliqués dans le projet car nous souhaitions accélérer le temps d'obtention des sondes fluorescentes finales.

4.5.3.1 Agrandissements et sélections dans AurKA

Pour l'agrandissement dans AurKA, 27 poses ont été initialement sélectionnées puis agrandies au cours de cycles itératifs d'ajouts d'un fragment puis de sélections. Les nombres de molécules obtenues et sélectionnées à chaque étape sont indiquées dans le Tableau 25.

Tableau 25 : Nombre de sondes de la protéine AurKA obtenues par F2D et sélectionnées par les chimistes, après l'ajout de chaque fragment.

Nombre de fragments	Nombre de molécules obtenues ou initiales	Nombre de molécules sélectionnées
1	27	27
2	47	43
3	797	54
4	683	14

Contrairement aux projets précédemment présentés, l'ajout de nouveaux fragments sans sélections informatiques intermédiaires est possible avec la condition d'arrêt utilisée ici (masse molaire de 450 g/mol). À partir des 14 molécules composées de 4 fragments, j'ai obtenu 11 775 structures 3D de molécules qui sont de potentielles sondes fluorescentes pouvant cibler la protéine AurKA et qui contiennent de 5 à 9 fragments. Parmi ces structures 3D, il existe 5 243 molécules ayant une structure 2D unique. La Figure 85 montre les distributions des descripteurs de ces molécules.

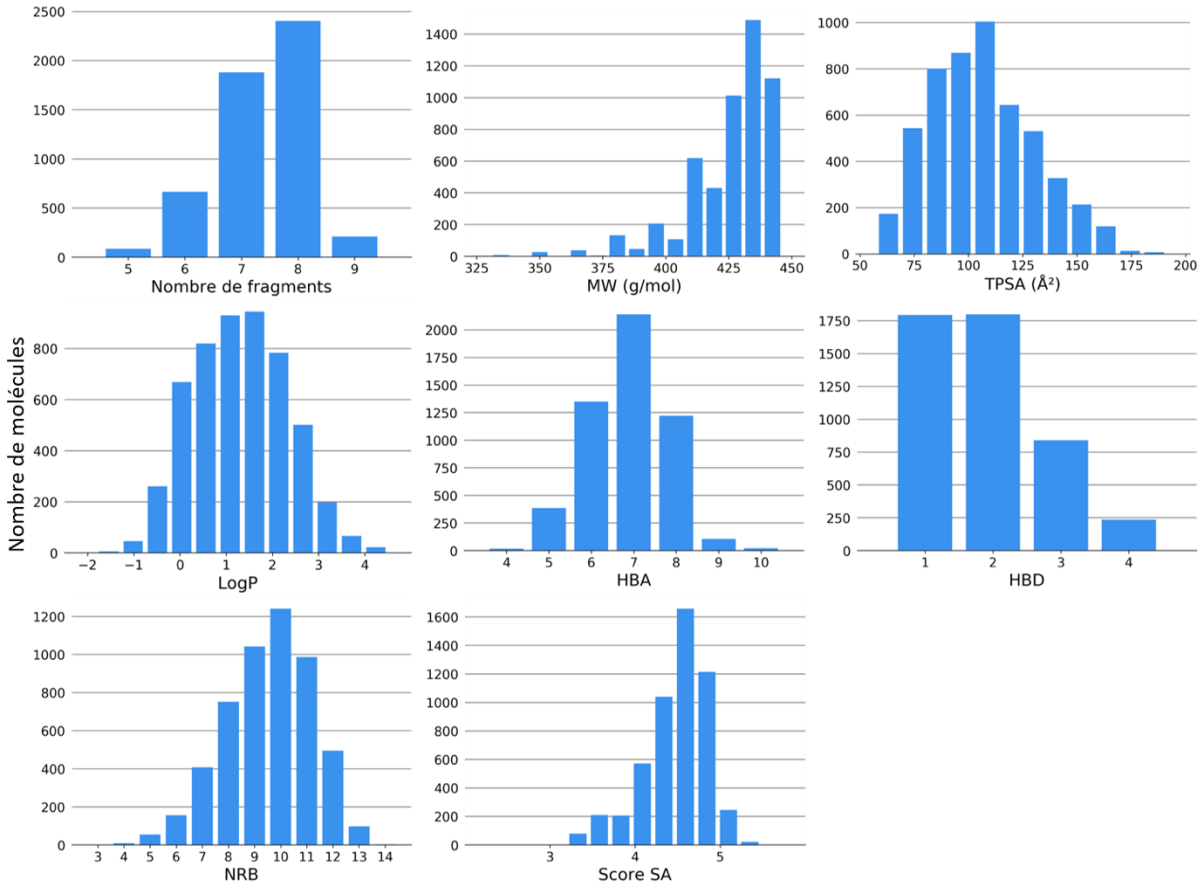


Figure 85 : Distributions des descripteurs moléculaires calculés sur les 5 243 molécules obtenues par l'exécution de F2D sur AurKA.

Les distributions des NAR et NCA ne sont pas montrées car il y a peu de valeurs différentes obtenues pour ces 2 descripteurs. En effet, 4 molécules sur les 5 243 ont une valeur de NCA > 0 avec un atome chiral pour deux molécules et deux atomes chiraux pour les 2 dernières. Pour les cycles aromatiques, 503 molécules ont une valeur de NAR = 5 quand les 4 740 autres ont une valeur de NAR = 4.

Pour mieux comprendre si les molécules obtenues peuvent être de bonnes sondes fluorescentes, nous pouvons comparer les valeurs moyennes de chaque descripteur aux valeurs moyennes des sondes chimiques provenant de la base de données (BDD) *chemical probes (https://new.chemicalprobes.org/)*[347]. Bien que cette BDD ne contienne pas que des sondes fluorescentes, il paraît plus pertinent d'effectuer la comparaison avec des sondes chimiques, qu'avec des inhibiteurs en cours d'essais cliniques. Cette comparaison, présentée dans le Tableau 26, révèle que les valeurs pour MW, TPSA, HBA et HBD sont assez similaires entre les 2 jeux de molécules. Par contre, les valeurs de LogP sont plus faibles pour les molécules de F2D et indiquent que ces molécules sont plus hydrophiles que les 121 sondes chimiques. La valeur de NRB est en moyenne supérieure de 3 unités

pour les molécules de F2D par rapport aux sondes chimiques et celle NAR est supérieure de 1 unité. L'optimisation des molécules obtenues devra alors viser à réduire ces valeurs, ainsi que celles du score SA.

Tableau 26 : Comparaison entre les valeurs moyennes des descripteurs moléculaires de différent jeux de molécules. Le premier jeu de molécules est composé de 121 sondes chimiques venant de chemicalprobes.org, le deuxième des 5 243 molécules uniques générées par F2D et le troisième des 269 molécules sélectionnées parmi celles de F2D.

Descripteur moléculaire	Chemical Probes	F2D	Sélection F2D
Nombre de molécules	121	5 243	269
MW (g/mol)	484,51 (± 121,08)	425,45 (± 17,01)	409,36 (± 30,20)
TPSA (Å²)	96,22 (± 32,12)	107,04 (± 23,47)	111,07 (± 21,65)
LogP	4,29 (± 1,52)	1,32 (± 1,02)	1,36 (± 1,10)
HBA	6,55 (± 2,16)	6,87 (± 0,96)	6,52 (± 1,13)
HBD	2,07 (± 1,23)	1,68 (± 1,01)	1,62 (± 0,73)
NRB	6,42 (± 3,12)	9,53 (± 1,67)	7,45 (± 1,97)
NAR	3,18 (± 1,17)	4,10 (± 0,29)	4,33 (± 0,47)
NCA	1,14 (± 2,19)	0 (± 0,04)	0 (± 0,0)
Score SA	3,22 (± 0,81)	4,48 (± 0,37)	3,64 (± 0,20)

Pour sélectionner les meilleures molécules parmi les 5 243 qui ont été générées, j'ai supprimé 19 molécules ayant des sous-structures pan assay interference compounds (PAINS) et 4 707 molécules ayant un score SA > 4. J'ai ainsi obtenu 536 molécules que j'ai replacé dans le site actif par simulations de *docking*. J'ai sélectionné les molécules ayant une pose avec un RMSD < 2,5 Å par rapport à la position de l'enchaînement des fragments de F2D. J'ai ainsi obtenu 269 molécules uniques pour lesquelles les valeurs moyennes des descripteurs moléculaires sont présentées dans le Tableau 26. Les valeurs moyennes obtenues changent peu par rapport à celles des 5 243 molécules initiales. Nous pouvons noter que les différences qu'il y avait entre les 121 molécules de chemicalprobes.org et les 5 243 molécules de F2D pour NRB et le score SA sont réduites pour les molécules sélectionnées. La valeur moyenne de LogP reste plus faible que celle des 121 sondes chimiques.

La Figure 86 montre les structures 2D et 3D des 6 meilleures molécules selon le score *Quantitative Estimate of Druglikeness* (QED). Les 6 molécules ont un premier enchaînement de fragments identiques après la graine avec une liaison amide reliant un benzène, dont les différentes substitutions créent les 6 molécules.

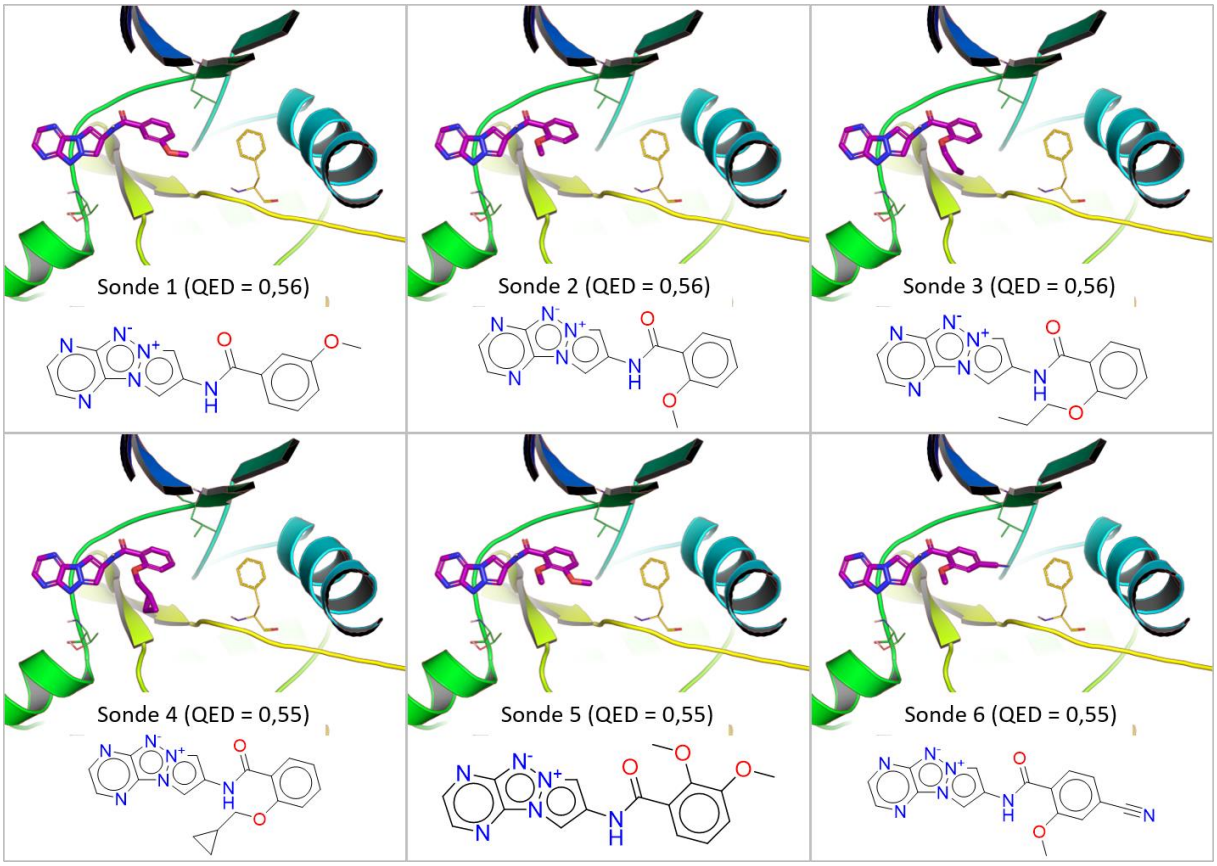


Figure 86 : Structures 2D et 3D des 6 meilleures sondes fluorescentes selon le score QED obtenues dans la protéine AurKA.

Parallèlement à la recherche de sondes fluorescentes pour AurKA, j'ai aussi cherché des sondes fluorescentes pour la protéine PLK1.

4.5.3.2 Agrandissements et sélections dans PLK1

F2D a été utilisé à partir de 8 poses de la même façon que dans AurKA. Les nombres de molécules uniques obtenues et sélectionnées au cours de 4 étapes d'ajouts de fragments sont indiquées dans le Tableau 25.

Tableau 27 : Nombre de sondes de la protéine PLK1 obtenues par F2D et sélectionnées par les chimistes, après l'ajout de chaque fragment.

Nombre de fragments	Nombre de molécules obtenues ou initiales	Nombre de molécules sélectionnées
1	8	8
2	57	41
3	459	55
4	103	16

En partant des 16 molécules sélectionnées par les chimistes, j'ai ajouté plusieurs fragments en une itération jusqu'à l'obtention de la masse molaire maximale de 450 g/mol. F2D a généré 506 structures 3D de molécules (224 structures 2D uniques) ciblant la protéine PLK1 et contenant de 5 à 8 fragments. La Figure 87 montre les distributions des descripteurs moléculaires de ces 224 molécules.

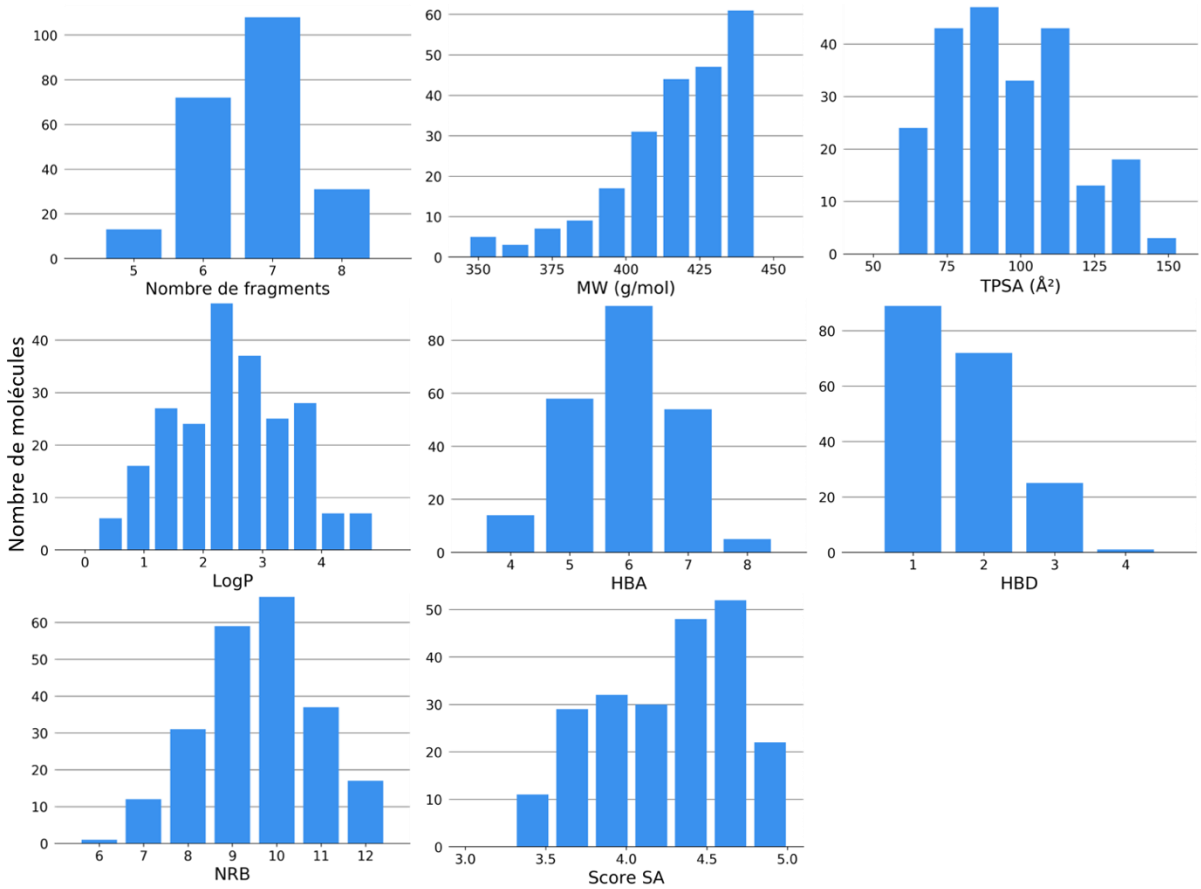


Figure 87 : Distributions des descripteurs moléculaires calculés sur les 224 molécules obtenues par l'exécution de F2D sur PLK1.

La distribution du nombre de cycles aromatiques n'est pas montrée car il n'y a que 2 valeurs : 17 molécules ont 5 cycles aromatiques et les 207 molécules restant en ont 4. Pour le nombre d'atomes chiraux, une seule molécule en a 2, alors que les 222 autres n'en ont pas. En effectuant la comparaison des valeurs moyennes des descripteurs moléculaires entre les 224 molécules et les sondes chimiques de la BDD *chemical probes* (Tableau 28), nous voyons que les valeurs de MW, LogP, HBA, HBD et NCA sont légèrement plus faibles pour les molécules obtenues par F2D. Les valeurs de TPSA sont quasiment identiques entre les 2 groupes de molécules. Les valeurs de NRB, NAR et Score SA sont plus élevées pour les molécules F2D que pour les 121 sondes.

Tableau 28 : Comparaison entre les valeurs moyennes des descripteurs moléculaires des différents jeux de molécules. Le premier jeu de molécules est composé des 121 sondes chimiques venant de chemicalprobes.org, le suivant des 224 molécules uniques générées par F2D dans PLK1 et le troisième des 46 molécules sélectionnées parmi celles de F2D.

Descripteur moléculaire	Chemical Probes	F2D	Sélection F2D
Nombre de molécules	121	224	46
MW (g/mol)	484,51 (± 121,08)	417,62 (± 21,23)	402,17 (± 30,09)
TPSA (Å²)	96,22 (± 32,12)	95,61 (± 21,91)	80,53 (± 23,61)
LogP	4,29 (± 1,52)	2,48 (± 1,00)	3,05 (± 1,11)
HBA	6,55 (± 2,16)	5,90 (± 0,91)	5,06 (± 0,83)
HBD	2,07 (± 1,23)	1,39 (± 0,91)	0,52 (± ,059)
NRB	6,42 (± 3,12)	9,60 (± 1,30)	8,43 (± 1,22)
NAR	3,18 (± 1,17)	4,08 (± 0,26)	4,19 (± 0,40)
NCA	1,14 (± 2,19)	0 (± 0,13)	0 (± 0,0)
Score SA	3,22 (± 0,81)	4,26 (± 0,42)	3,72 (± 0,17)

Une molécule parmi les 224 contient une sous-structure PAINS et a été supprimée. En retirant aussi les molécules ayant un score SA > 4, j'ai conservé 66 molécules. Pour savoir si ces molécules forment des interactions favorables dans le site actif de PLK1, j'ai comparé leurs poses prédites par simulations de *docking* avec la position qu'elles ont dans le logiciel F2D.

J'ai conservé 46 molécules ayant un RMSD < 2,5 Å entre pose et position venant de F2D. Les valeurs moyennes des descripteurs de ces molécules sont plus faibles pour MW, TPSA, HBA et HBD par rapport aux sondes connues. Bien qu'elles restent différentes, les moyennes de LogP, NRB et score SA tendent vers les valeurs moyennes des 121 molécules de référence.

La Figure 88 présente les structures 2D et 3D des 6 meilleures molécules classées par score QED. Nous observons que les scores QED associés à ces molécules sont plus faibles que pour les molécules trouvées dans AurKA. Ce qui n'est pas un problème car ce n'est pas l'aspect « *drug-like* » qui nous intéresse, mais l'aspect sondes fluorescentes. Les 6 molécules ont une sous structure commune composée de la graine 2 (Figure 83) et d'un pentoxymethylbenzene.

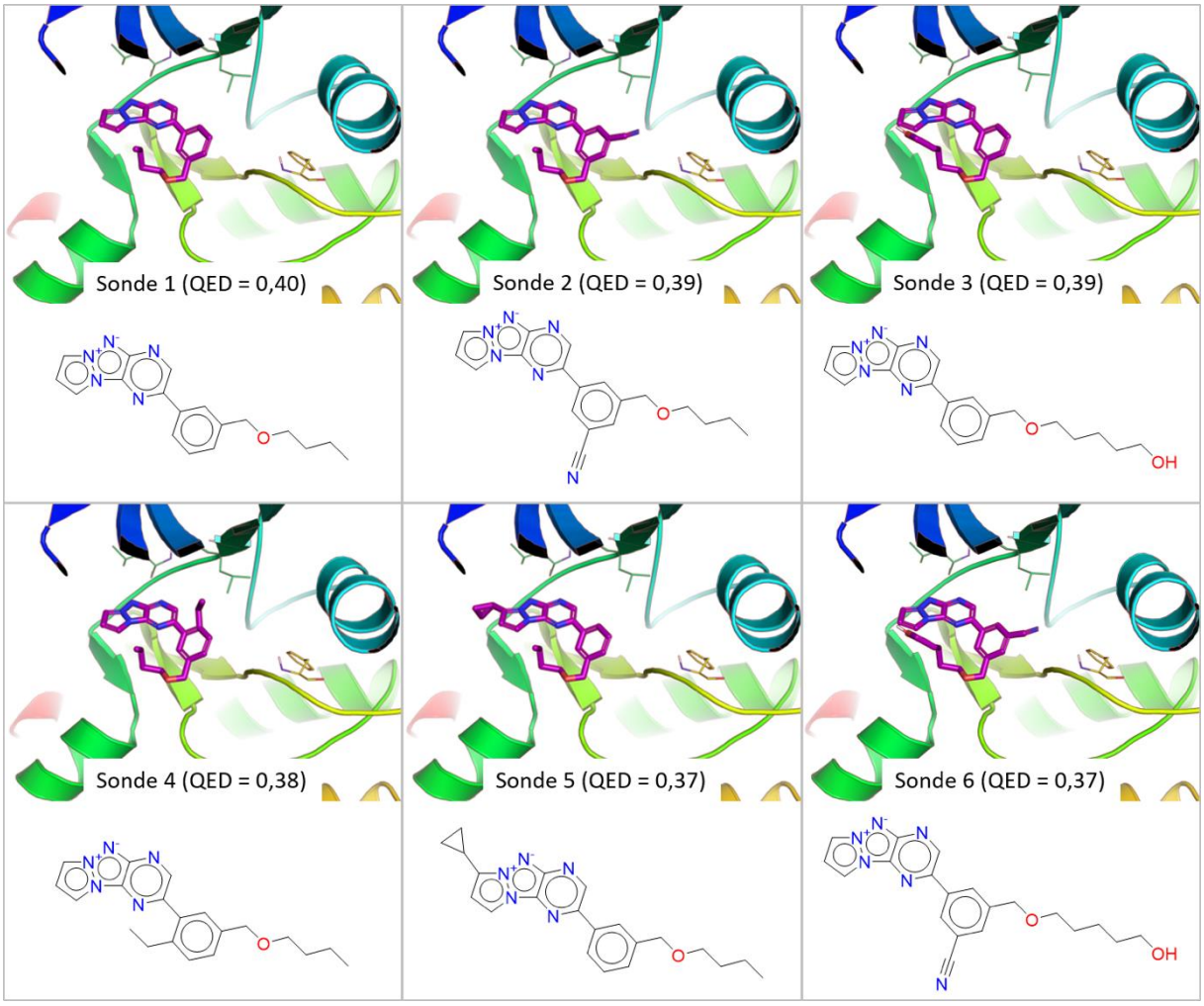


Figure 88 : Structures 2D et 3D des 6 meilleures sondes fluorescentes selon le score QED obtenues dans la protéine PLK1.

4.5.4 Bilan du projet de recherche de sondes fluorescentes

Dans ce projet, deux utilisations de F2D ont été effectuées avec succès pour l'obtention de sondes fluorescentes ciblant AurKA et PLK1. Au final, 407 molécules ont été proposées aux chimistes impliqués dans ce projet pour AurKA et 166 molécules l'ont été pour PLK1. Les possibilités de synthèse de ces potentielles sondes fluorescentes sont en cours d'évaluation.

Les sondes obtenues ne semblent pas optimales en se basant sur le score QED, particulièrement pour la protéine PLK1 dont les sondes ont un score QED maximal de 0,40. Ce n'est pas un souci car ces molécules n'ont pas pour but de devenir des inhibiteurs des protéines AurKa et PLK1, mais simplement d'être des sondes fluorescentes. Elles pourront être administrées sur des lignées cellulaires pour comprendre les voies de signalisation de ces protéines, mais ne pas être données à des humains.

4.6 Conclusion et perspectives

Au cours de ce chapitre, Frags2Drugs (F2D) a été appliqué sur 7 cibles au total (Haspin, CLK1, DYRK1A, DYRK2, PIM1, AurKA et PLK1). Ces recherches d'inhibiteurs ont été effectuées en collaborations avec 4 équipes de recherches.

Le premier projet sur les cibles Haspin et CLK1/DYRK1A n'a pas abouti. Il doit être poursuivi pour pouvoir appliquer tous les critères de sélection de F2D et ainsi obtenir des inhibiteurs de ces cibles. Ce

projet a permis de mettre en place une procédure qui a fait que les 3 autres projets ont pu être menés à terme.

L'utilisation de F2D en collaboration avec le COBRA de Rouen a permis de proposer des molécules pour l'obtention d'un financement de thèse. J'ai trouvé des petites molécules et des macrocycles pour PIM1 qui pourront être synthétisés par le GICC et testés biologiquement. F2D pourrait être de nouveau lancé sur Pim-2 et Pim-3 pour générer des inhibiteurs « Pan-Pim ». Pour les cibles AurKA et PLK1, j'ai obtenu de potentielles sondes fluorescentes qui seront évaluées par les chimistes médicinaux.

Les différentes utilisations de F2D ont mis en évidence un point d'amélioration pour une meilleure utilisation « en routine » du logiciel. Il s'agit de la gestion de la mémoire RAM qui est primordiale et est le premier facteur limitant dans F2D. Avec une bonne gestion de cette mémoire vive, F2D pourrait être systématiquement lancé de manière exhaustive jusqu'à l'obtention d'au moins une molécule avec une masse molaire maximale.

F2D pourrait aussi être amélioré par la prise en compte de l'environnement autour de chaque fragment de la base de données (BDD). Par exemple, nous pourrions prendre un triplet d'acides aminés autour de chaque fragment et intégrer cette information dans la BDD. Cela permettrait d'associer les fragments en prenant plus en compte le site actif de la protéine cible et ainsi de générer moins de molécules, mais plus pertinentes.

Chapitre 5 : Conclusion et perspectives générales

En mai 2021, vingt années après la mise sur le marché de l'imatinib, 98 inhibiteurs de protéines kinases (PKI) ont été approuvés dans le monde[263]. Parmi ces 98 PKI, il y a 71 petites molécules approuvées par la FDA. Au début de la création de Frags2Drugs (F2D), en 2016, 34 petites molécules étaient approuvées. Le nombre de petites molécules en phase IV qui a doublé en 5 ans démontre que les protéines kinases sont des cibles thérapeutiques d'intérêt majeur. Ces protéines sont impliquées principalement dans les cancers ainsi que dans le diabète et les maladies neurodégénératives où leur inhibition est un des traitements possibles.

Malgré le fait qu'il y ait de nombreux inhibiteurs ciblant les protéines kinases, une partie du kinome reste inexplorée. Le projet *NIH Illuminating the Druggable Genome (IDG)* utilise de nombreuses données venant de sources multiples pour mieux connaître les cibles thérapeutiques. Sur 635 protéines kinases recensées, IDG en a classé 3% (21 kinases) comme complètement inconnues et 30% (188 kinases) comme n'ayant aucun inhibiteur[263]. Il y a donc encore beaucoup à découvrir sur ces protéines kinases.

Au cours de ma thèse, j'ai contribué à la découverte de nouveaux PKI grâce au logiciel F2D. J'ai participé à son amélioration en l'analysant et en l'utilisant dans différents projets. La méthode F2D est présentée dans le premier article que j'ai rédigé. F2D est aussi capable de générer des macrocycles, comme le montre le deuxième article de ma thèse, focalisé sur cette classe de molécules. Bien qu'il subsiste des points d'amélioration pour ce logiciel, il résulte de cinq années de recherche faites de succès et d'échecs et il est aujourd'hui complètement fonctionnel. En utilisant F2D couplé à d'autres outils de chémoinformatique, j'ai pu proposer de nouvelles molécules pour orienter les décisions des chimistes impliqués dans quatre projets différents (Tableau 29).

Tableau 29 : Bilan des quatre projets d'applications de F2D.

	Projet LabEx SynOrg Orléans		Projet LabEx SynOrg Rouen		Projet LabEx SynOrg Tours		Projet sondes fluorescentes	
Cibles	Haspin	DYRK1A/ CLK1	DYRK1A	DYRK2	Pim-1 (petites molécules)	Pim-1 (macro-cycles)	AurKA	PLK1
Nombre de graines	14	14	7	11	5	4	27	8
Nombre de molécules	1 406	1 312	1 388	1 513	156	24	407	166

Les deux premiers projets issus du laboratoire d'excellence (LabEx) SynOrg Orléans et LabEx SynOrg Rouen n'ont pas été menés jusqu'à leur terme et doivent être poursuivis pour pouvoir fournir aux chimistes médicinaux des propositions d'inhibiteurs à synthétiser. Dans ces deux projets, le nombre de molécules obtenues est plus élevé que dans les deux autres. Les projets de recherche d'inhibiteurs pour Pim-1 et de sondes fluorescentes ont été concluants et ont permis d'obtenir des petites molécules et macrocycles qui ont été présentés à nos collaborateurs.

Bien que F2D soit fonctionnel et permette d'obtenir des inhibiteurs linéaires ou macrocycliques pour n'importe quelle protéine kinase, ce logiciel nécessite encore quelques améliorations. Un premier problème de ce logiciel est de présenter une saturation de mémoire RAM lorsqu'il est utilisé sans aucune intervention pour limiter le nombre de solutions qu'il peut produire. Pour pallier ce problème, il faudrait trouver une manière de réduire le nombre de différentes possibilités d'agrandissement en orientant la recherche dans la base de données orientée graphes directement vers les meilleurs futurs inhibiteurs. Pour cela, une intelligence artificielle pourrait être ajoutée au programme afin de prédire pendant l'agrandissement l'impact positif ou négatif de l'ajout d'un nouveau fragment.

F2D repose sur une superposition de tous les complexes protéines kinases-ligands, faisant que de nombreux fragments sont placés dans le site actif des protéines kinases permettant de créer des inhibiteurs de type I, $I^{1/2}$ et II. Pour pouvoir générer des inhibiteurs allostériques, il faudrait obtenir de nouvelles positions pour les fragments par simulations de *docking* dans les poches allostériques. L'utilisation de F2D est aussi limitée aux protéines avec au moins une structure cristallographique disponible. En effectuant des modèles par homologie pour prédire les structures des protéines dont seules les séquences sont connues, F2D pourrait être appliqué sur toutes les protéines kinases.

L'une des principales limitations de F2D est d'ailleurs son champ d'application actuel qui est limité aux protéines kinases. Une amélioration conséquente serait de permettre son utilisation sur l'ensemble des cibles protéiques ce qui fait l'objet d'une thèse qui débutera au laboratoire. Enfin, après l'agrandissement, nous pourrions ajouter à F2D une prédiction des affinités de liaisons des molécules générées pour mieux filtrer les résultats obtenus. Ce type de prédiction, basée sur l'apprentissage profond fait l'objet d'une autre thèse au laboratoire.

J'ai aussi mis en place un serveur web proposant l'utilisation de F2D et de nombreux autres outils chémoinformatiques développés par l'équipe SB&C (http://sbc.icoa.fr). La conception et le déploiement de cette plateforme web m'ont permis de recruter, former et encadrer trois stagiaires qui y ont apporté leurs contributions. Ce travail m'a apporté de nombreuses connaissances sur des logiciels et technologies souvent demandées dans des entreprises informatiques telles que Docker, Python ou Django par exemple. Cette plateforme web est aujourd'hui fonctionnelle et la publication de l'article sur l'outil KinoMine est en cours.

De nombreuses améliorations et mises à jour pourront encore être apportées aux trois sites que j'ai développés et mis en production. Par exemple, un suivi plus précis de l'avancée des calculs dans le site F2D web pourrait être ajouté. Certains des autres sites de la plateforme, comme *Protein Kinase Inhibitor Database* (PKIDB)[34,35], pourront être modifiés pour reposer sur les mêmes technologies que les trois sites que j'ai développés. Cela permettra d'unifier le fonctionnement des outils de la plateforme et de simplifier son développement.

Au fur et à mesure de mon cursus dans les études supérieures, j'ai découvert la chémoinformatique qui est un domaine qui aujourd'hui me passionne. Pourtant, mon parcours initial dans les études supérieures n'était pas directement orienté vers cette discipline. Dès ma première année de licence de biologie, quelques cours de bioinformatique nous ont été enseignés et m'ont donné un grand intérêt pour ce domaine. Ce n'est que pendant mon master de bioinformatique que j'ai découvert la chémoinformatique et que j'ai décidé de poursuivre en doctorat dans cette discipline. Après ces trois années, je suis convaincu de vouloir continuer à travailler à l'interface entre bioinformatique et chémoinformatique pour la découverte de nouveaux médicaments. Cet intérêt est d'autant plus grand depuis que, pendant la rédaction de cette thèse, la société *DeepMind* a publié dans *Nature* son outil *AlphaFold* permettant de prédire par apprentissage profond les structures de protéines

à partir de leur séquence[335]. Cette avancée majeure dans notre domaine décuple les possibilités de découverte de nouveaux médicaments.

Au cours de ma thèse, sur le plan personnel, j'ai beaucoup appris sur la gestion de projets scientifiques et à prioriser les objectifs quand plusieurs d'entre eux se chevauchent, en utilisant la méthode AGILE kanban. J'ai aussi amélioré mes connaissances dans la gestion de code informatique, par une utilisation avancée de Git et GitLab.

Au cours de ces trois années j'ai pu présenter mes travaux dans des congrès nationaux et internationaux. Cela m'a permis d'entendre les différentes remarques et critiques qui ont pu être émises et d'y répondre ou de les prendre en compte pour améliorer F2D. De plus, en participant à l'édition 2021 du concours « Ma thèse en 180 secondes », au niveau local et régional, j'ai beaucoup appris sur la gestion du stress et sur comment être synthétique et à l'aise à l'oral.

J'ai aussi donné des cours d'introduction à la chémoinformatique à des étudiants de master pendant 60 heures réparties sur trois années consécutives. Quand ma formation en chémoinformatique sera complète, je souhaiterai continuer l'enseignement de cette discipline.

Communications scientifiques

Communications orales

Conférences invitées dans un congrès international

- **Bonnet, P.** ; Bournez, C. ; Peyrat, G. ; Krezel, P. ; Aci-Sèche, S.
  An In silico Fragment Based Design Tool for the Discovery of Novel Kinase Inhibitors
  21st Romanian International Conference on Chemistry and Chemical Engineering (RICCCE 2019)
  Septembre 2019 - Constanta (Roumanie).

Communications orales dans un congrès national

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Gally, J.-M. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
  Application of Frags2Drugs for the fragment-based drug design of macrocyclic kinase inhibitors
  XXII congrès du GGMM (Group of Graphism and Molecular Modeling) & 10èmes journées de la SFCi (French Society of Chemoinformatics)
  Octobre 2021 - Lille.

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Gally, J.-M. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
  Discovery of macrocyclic inhibitors of ALK using Frags2Drugs, a fragment-based drug design in silico tool
  Journées Ouvertes en Biologie, Informatique et Mathématiques - JOBIM
  Juillet 2021 - Paris.

- **Bournez, C.** ; Peyrat, G. ; Gally, J.-M. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Fragment linking combined to graph-based approach in the discovery of novel kinase inhibitors
  9èmes Journées de la Société Française de Chémoinformatique - SFCI-2019
  Novembre 2019 - Paris.

Communications flash dans un congrès international

- **Carles, F.** ; Bourg, S. ; Peyrat, G. ; Meyer, C. ; Bonnet, P.
  Future perspective in data mining, data visualization and deep learning for kinase research
  22nd European Symposium on Quantitative Structure-Activity Relationships (EuroQSAR 2018)
  Septembre 2018 - Thessalonique (Grèce).

Communications flash dans un congrès national

- **Peyrat, G**. ; Carles, F. ; Bourg, S. ; Meyer, C. ; Bonnet, P.
  Perspective in data mining, data visualization and deep learning for kinase research
  9èmes Journées de la Société Française de Chémoinformatique - SFCI-2019
  Novembre 2019 - Paris.

- **Peyrat, G.** ; Bournez, C. ; Gally, J.-M. ; Krezel, P. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Frags2Drugs: discovery of kinase inhibitors from a fragment network
  9èmes Journées de la Société Française de Chémoinformatique - SFCI-2019
  Novembre 2019 - Paris.

Communications par affiches

Communications par affiches dans un congrès international

- **Peyrat, G.** ; Carles, F. ; Bourg, S. ; Meyer, C. ; Bonnet, P.
  Future perspective in data mining, data visualization and deep learning for kinase search
  55èmes Rencontres Internationales de Chimie Thérapeutique (RICT 2019)
  Juillet 2019 - Nantes.

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Design of macrocyclic kinase inhibitors from fragment
  55èmes Rencontres Internationales de Chimie Thérapeutique (RICT 2019)
  Juillet 2019 - Nantes.

- **Peyrat, G.** ; Bournez, C. ; Gally, J.-M. ; Krezel, P. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  FRAGS2DRUGS : Using fragment network to find new kinase inhibitors
  55èmes Rencontres Internationales de Chimie Thérapeutique (RICT 2019)
  Juillet 2019 - Nantes.

- **Bournez, C.** ; Peyrat, G. ; Krezel, P. ; Gally, J.-M. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Discovering new kinase inhibitors from fragment network: Frags2Drugs
  7th RSC-BMCS Fragment-based Drug Discovery meeting (Fragments 2019)
  Mars 2019 - Cambridge (Royaume Uni).

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Gally, J.-M. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Frags2Drugs: from fragment database to new potent and selective protein kinase inhibitors
  26th Young Research Fellow Meeting
  Février 2019 - Paris.

- **Carles, F.** ; Bourg, S. ; Peyrat, G. ; Meyer, C. ; Bonnet, P.
  Future perspective in data mining, data visualization and deep learning for kinase research
  22nd European Symposium on Quantitative Structure-Activity Relationships (EuroQSAR 2018)
  Septembre 2018 - Thessalonique (Grèce).

Communications par affiches dans un congrès national

- **Peyrat, G**. ; Carles, F. ; Bourg, S. ; Meyer, C. ; Bonnet, P.
  Perspective in data mining, data visualization and deep learning for kinase research
  9èmes Journées de la Société Française de Chémoinformatique - SFCI-2019
  Novembre 2019 - Paris.

- **Peyrat, G.** ; Bournez, C. ; Gally, J.-M. ; Krezel, P. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Frags2Drugs: discovery of kinase inhibitors from a fragment network
  9èmes Journées de la Société Française de Chémoinformatique - SFCI-2019
  Novembre 2019 - Paris.

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Gally, J.-M. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Frags2Drugs: from fragment database to new potent and selective protein kinase inhibitors

Journée Scientifique de la Fédération Physique et Chimie du Vivant- FR2708
Avril 2019 - Orléans.

- **Peyrat, G.** ; Bournez, C. ; Krezel, P. ; Gally, J.-M. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  Frags2Drugs: a novel in silico Fragment Based Drug Design tool
  30ème colloque Biotechnocentre
  Octobre 2018 - Seillac.

Bibliographie

(1) Gally, J.-M. Développement d'outils de Chémoinformatique Pour l'identification d'inhibiteurs de Protéines Kinases à Partir de Fragments. These de doctorat, Orléans, 2017.

(2) Bournez, C. Conception d'un Logiciel Pour La Recherche de Nouvelles Molécules Bioactives. These de doctorat, Orléans, 2019.

(3) Brown, F. K. Chapter 35 - Chemoinformatics: What Is It and How Does It Impact Drug Discovery. In *Annual Reports in Medicinal Chemistry*; Bristol, J. A., Ed.; Academic Press, 1998; Vol. 33, pp 375–384. https://doi.org/10.1016/S0065-7743(08)61100-8.

(4) The Obernai Declaration. 2.

(5) Société Française de Chémoinformatique http://www.chemoinformatique.fr/ (accessed 2021 -03 -22).

(6) Hansch, Corwin.; Fujita, Toshio. P-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.* **1964**, *86* (8), 1616–1626. https://doi.org/10.1021/ja01062a035.

(7) Hansch, C. Quantitative Approach to Biochemical Structure-Activity Relationships. *Acc. Chem. Res.* **1969**, *2* (8), 232–239. https://doi.org/10.1021/ar50020a002.

(8) Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. S. QSAR and Proteo-Chemometric Analysis of the Interaction of a Series of Organic Compounds with Melanocortin Receptor Subtypes. *J. Med. Chem.* **2003**, *46* (13), 2572–2579. https://doi.org/10.1021/jm020945m.

(9) van Westen, G. J. P.; Wegner, J. K.; IJzerman, A. P.; van Vlijmen, H. W. T.; Bender, A. Proteochemometric Modeling as a Tool to Design Selective Compounds and for Extrapolating to Novel Targets. *Med Chem Commun* **2011**, *2* (1), 16–30. https://doi.org/10.1039/C0MD00165A.

(10) Chen, H.; Engkvist, O.; Wang, Y.; Olivecrona, M.; Blaschke, T. The Rise of Deep Learning in Drug Discovery. *Drug Discov. Today* **2018**, *23* (6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039.

(11) DiMasi, J. A.; Hansen, R. W.; Grabowski, H. G. The Price of Innovation: New Estimates of Drug Development Costs. *J. Health Econ.* **2003**, *22* (2), 151–185. https://doi.org/10.1016/S0167-6296(02)00126-1.

(12) Muller, P. Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994). *Pure Appl. Chem.* **1994**, *66* (5), 1077–1184. https://doi.org/10.1351/pac199466051077.

(13) Favre, H. A.; Powell, W. H. *Nomenclature of Organic Chemistry*; 2013. https://doi.org/10.1039/9781849733069.

(14) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminformatics* **2013**, *5* (1), 7. https://doi.org/10.1186/1758-2946-5-7.

(15) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(16) Pletnev, I.; Erin, A.; McNaught, A.; Blinov, K.; Tchekhovskoi, D.; Heller, S. InChIKey Collision Resistance: An Experimental Testing. *J. Cheminformatics* **2012**, *4*, 39. https://doi.org/10.1186/1758-2946-4-39.

(17) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K. I.; Grier, D. L.; Leland, B. A.; Laufer, J. Description of Several Chemical Structure File Formats Used by Computer Programs Developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, *32* (3), 244–255. https://doi.org/10.1021/ci00007a012.

(18) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*, 1st ed.; Methods and Principles in Medicinal Chemistry; Wiley, 2000. https://doi.org/10.1002/9783527613106.

(19) Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*; Leszczynski, J., Kaczmarek-Kedziera, A., Puzyn, T., G. Papadopoulos, M., Reis, H., K.

Shukla, M., Eds.; Springer International Publishing: Cham, 2017; pp 2065–2093. https://doi.org/10.1007/978-3-319-27282-5_51.
(20) Xue, L.; Bajorath, J. Molecular Descriptors in Chemoinformatics, Computational Combinatorial Chemistry, and Virtual Screening. *Comb. Chem. High Throughput Screen.* **2000**, *3* (5), 363–372. https://doi.org/10.2174/1386207003331454.
(21) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280. https://doi.org/10.1021/ci010132r.
(22) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(23) RDKit: Open-source cheminformatics https://www.rdkit.org/ (accessed 2020 -11 -06).
(24) Sydow, D.; Morger, A.; Driller, M.; Volkamer, A. TeachOpenCADD: A Teaching Platform for Computer-Aided Drug Design Using Open Source Packages and Data. *J. Cheminformatics* **2019**, *11* (1), 29. https://doi.org/10.1186/s13321-019-0351-x.
(25) ChemBioFrance - Infrastructure de recherche https://chembiofrance.cn.cnrs.fr/fr/composante/chimiotheque (accessed 2021 -03 -25).
(26) Gabrielson, S. W. SciFinder. *J. Med. Libr. Assoc. JMLA* **2018**, *106* (4), 588–590. https://doi.org/10.5195/jmla.2018.515.
(27) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49* (D1), D1388–D1395. https://doi.org/10.1093/nar/gkaa971.
(28) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magariños, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (Database issue), D945–D954. https://doi.org/10.1093/nar/gkw1074.
(29) Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities. *Nucleic Acids Res.* **2015**, *43* (Web Server issue), W612–W620. https://doi.org/10.1093/nar/gkv352.
(30) Irwin, J. J.; Shoichet, B. K. ZINC − A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* **2005**, *45* (1), 177–182. https://doi.org/10.1021/ci049714+.
(31) Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55* (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.
(32) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242. https://doi.org/10.1093/nar/28.1.235.
(33) Golovin, A.; Oldfield, T. J.; Tate, J. G.; Velankar, S.; Barton, G. J.; Boutselakis, H.; Dimitropoulos, D.; Fillon, J.; Hussain, A.; Ionides, J. M. C.; John, M.; Keller, P. A.; Krissinel, E.; McNeil, P.; Naim, A.; Newman, R.; Pajon, A.; Pineda, J.; Rachedi, A.; Copeland, J.; Sitnov, A.; Sobhany, S.; Suarez-Uruena, A.; Swaminathan, G. J.; Tagari, M.; Tromm, S.; Vranken, W.; Henrick, K. E-MSD: An Integrated Data Resource for Bioinformatics. *Nucleic Acids Res.* **2004**, *32* (Database issue), D211-216. https://doi.org/10.1093/nar/gkh078.
(34) Carles, F.; Bourg, S.; Meyer, C.; Bonnet, P. PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials. *Molecules* **2018**, *23* (4), 908. https://doi.org/10.3390/molecules23040908.
(35) Bournez, C.; Carles, F.; Peyrat, G.; Aci-Sèche, S.; Bourg, S.; Meyer, C.; Bonnet, P. Comparative Assessment of Protein Kinase Inhibitors in Public Databases and in PKIDB. *Molecules* **2020**, *25* (14), 3226. https://doi.org/10.3390/molecules25143226.
(36) Larousse, É. médicament - LAROUSSE https://www.larousse.fr/encyclopedie/medical/m%C3%A9dicament/14453 (accessed 2021 -03 -26).

(37) Article L5111-1 - Code de la santé publique - Légifrance https://www.legifrance.gouv.fr/codes/article_lc/LEGIARTI000006689867/ (accessed 2021 -03 -26).
(38) Les biomédicaments, une nouvelle génération de traitements https://www.leem.org/les-biomedicaments-une-nouvelle-generation-de-traitements (accessed 2021 -04 -12).
(39) The Nobel Prize in Chemistry 2020 https://www.nobelprize.org/prizes/chemistry/2020/summary/ (accessed 2021 -04 -12).
(40) The Nobel Prize in Physiology or Medicine 2018 https://www.nobelprize.org/prizes/medicine/2018/summary/ (accessed 2021 -04 -12).
(41) Mullard, A. 2020 FDA Drug Approvals. *Nat. Rev. Drug Discov.* **2021**, *20* (2), 85–90. https://doi.org/10.1038/d41573-021-00002-0.
(42) Recherche et développement https://www.leem.org/recherche-et-developpement (accessed 2021 -03 -26).
(43) Matthews, H.; Hanison, J.; Nirmalan, N. "Omics"-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives. *Proteomes* **2016**, *4* (3), 28. https://doi.org/10.3390/proteomes4030028.
(44) Wouters, O. J.; McKee, M.; Luyten, J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA* **2020**, *323* (9), 844. https://doi.org/10.1001/jama.2020.1166.
(45) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2019**, *20* (2), 273–286. https://doi.org/10.1093/biostatistics/kxx069.
(46) Thomas, D. W.; Burns, J.; Audette, J.; Carroll, A.; Dow-Hygelund, C.; Hay, M. Clinical Development Success Rates 2006-2015.
(47) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.
(48) Sertuerner. Ueber Das Morphium, Eine Neue Salzfähige Grundlage, Und Die Mekonsäure, Als Hauptbestandtheile Des Opiums. *Ann. Phys.* **1817**, *55* (1), 56–89. https://doi.org/10.1002/andp.18170550104.
(49) Lancet, T. Phase 0 Trials: A Platform for Drug Development? *The Lancet* **2009**, *374* (9685), 176. https://doi.org/10.1016/S0140-6736(09)61309-X.
(50) Macarron, R.; Banks, M. N.; Bojanic, D.; Burns, D. J.; Cirovic, D. A.; Garyantes, T.; Green, D. V. S.; Hertzberg, R. P.; Janzen, W. P.; Paslay, J. W.; Schopfer, U.; Sittampalam, G. S. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10* (3), 188–195. https://doi.org/10.1038/nrd3368.
(51) Barker, A.; Kettle, J. G.; Nowak, T.; Pease, J. E. Expanding Medicinal Chemistry Space. *Drug Discov. Today* **2013**, *18* (5), 298–304. https://doi.org/10.1016/j.drudis.2012.10.008.
(52) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A Common Mechanism Underlying Promiscuous Inhibitors from Virtual and High-Throughput Screening. *J. Med. Chem.* **2002**, *45* (8), 1712–1722. https://doi.org/10.1021/jm010533y.
(53) Scott, D. E.; Coyne, A. G.; Hudson, S. A.; Abell, C. Fragment-Based Approaches in Drug Discovery and Chemical Biology. *Biochemistry* **2012**, *51* (25), 4990–5003. https://doi.org/10.1021/bi3005126.
(54) Leach, A. R.; Hann, M. M. Molecular Complexity and Fragment-Based Drug Discovery: Ten Years On. *Curr. Opin. Chem. Biol.* **2011**, *15* (4), 489–496. https://doi.org/10.1016/j.cbpa.2011.05.008.
(55) Erlanson, D. A.; Fesik, S. W.; Hubbard, R. E.; Jahnke, W.; Jhoti, H. Twenty Years on: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discov.* **2016**, *15* (9), 605–619. https://doi.org/10.1038/nrd.2016.109.
(56) Jencks, W. P. On the Attribution and Additivity of Binding Energies. *Proc. Natl. Acad. Sci.* **1981**, *78* (7), 4046–4050. https://doi.org/10.1073/pnas.78.7.4046.

(57) Patel, Y.; Gillet, V. J.; Howe, T.; Pastor, J.; Oyarzabal, J.; Willett, P. Assessment of Additive/Nonadditive Effects in Structure−Activity Relationships: Implications for Iterative Drug Design. *J. Med. Chem.* **2008**, *51* (23), 7552–7562. https://doi.org/10.1021/jm801070q.
(58) Verlinde, C. L. M. J.; Rudenko, G.; Hol, W. G. J. In Search of New Lead Compounds for Trypanosomiasis Drug Design: A Protein Structure-Based Linked-Fragment Approach. *J. Comput. Aided Mol. Des.* **1992**, *6* (2), 131–147. https://doi.org/10.1007/BF00129424.
(59) Shuker, S. B.; Hajduk, P. J.; Meadows, R. P.; Fesik, S. W. Discovering High-Affinity Ligands for Proteins: SAR by NMR. *Science* **1996**, *274* (5292), 1531–1534. https://doi.org/10.1126/science.274.5292.1531.
(60) Chessari, G.; Woodhead, A. J. From Fragment to Clinical Candidate—a Historical Perspective. *Drug Discov. Today* **2009**, *14* (13), 668–675. https://doi.org/10.1016/j.drudis.2009.04.007.
(61) Bollag, G.; Tsai, J.; Zhang, J.; Zhang, C.; Ibrahim, P.; Nolop, K.; Hirth, P. Vemurafenib: The First Drug Approved for BRAF-Mutant Cancer. *Nat. Rev. Drug Discov.* **2012**, *11* (11), 873–886. https://doi.org/10.1038/nrd3847.
(62) Goodford, P. J. A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules. *J. Med. Chem.* **1985**, *28* (7), 849–857. https://doi.org/10.1021/jm00145a002.
(63) Miranker, A.; Karplus, M. Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method. *Proteins Struct. Funct. Bioinforma.* **1991**, *11* (1), 29–34. https://doi.org/10.1002/prot.340110104.
(64) Böhm, H.-J. The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors. *J. Comput. Aided Mol. Des.* **1992**, *6* (1), 61–78. https://doi.org/10.1007/BF00124387.
(65) de Souza Neto, L. R.; Moreira-Filho, J. T.; Neves, B. J.; Maidana, R. L. B. R.; Guimarães, A. C. R.; Furnham, N.; Andrade, C. H.; Silva, F. P. In Silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Front. Chem.* **2020**, *8*. https://doi.org/10.3389/fchem.2020.00093.
(66) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A 'Rule of Three' for Fragment-Based Lead Discovery? *Drug Discov. Today* **2003**, *8* (19), 876–877. https://doi.org/10.1016/S1359-6446(03)02831-9.
(67) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **1997**, *23* (1), 3–25. https://doi.org/10.1016/S0169-409X(96)00423-1.
(68) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623. https://doi.org/10.1021/jm020017n.
(69) Doak, B. C.; Over, B.; Giordanetto, F.; Kihlberg, J. Oral Druggable Space beyond the Rule of 5: Insights from Drugs and Clinical Candidates. *Chem. Biol.* **2014**, *21* (9), 1115–1142. https://doi.org/10.1016/j.chembiol.2014.08.013.
(70) Keserű, G. M.; Erlanson, D. A.; Ferenczy, G. G.; Hann, M. M.; Murray, C. W.; Pickett, S. D. Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia. *J. Med. Chem.* **2016**, *59* (18), 8189–8206. https://doi.org/10.1021/acs.jmedchem.6b00197.
(71) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856–864. https://doi.org/10.1021/ci000403i.
(72) Jacquemard, C.; Kellenberger, E. A Bright Future for Fragment-Based Drug Discovery: What Does It Hold? *Expert Opin. Drug Discov.* **2019**, *14* (5), 413–416. https://doi.org/10.1080/17460441.2019.1583643.
(73) Baell, J. B.; Holloway, G. A. New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays. *J. Med. Chem.* **2010**, *53* (7), 2719–2740. https://doi.org/10.1021/jm901137j.
(74) Erlanson, D. A.; Jahnke, W.; Mannhold, R.; Kubinyi, H.; Folkers, G. *Fragment–based Drug Discovery: Lessons and Outlook*, 1er édition.; Wiley VCH: Weinheim, 2015.

(75) Singh, M.; Tam, B.; Akabayov, B. NMR-Fragment Based Virtual Screening: A Brief Overview. *Mol. J. Synth. Chem. Nat. Prod. Chem.* **2018**, *23* (2). https://doi.org/10.3390/molecules23020233.
(76) Gimeno, A.; Ojeda-Montes, M. J.; Tomás-Hernández, S.; Cereto-Massagué, A.; Beltrán-Debón, R.; Mulero, M.; Pujadas, G.; Garcia-Vallvé, S. The Light and Dark Sides of Virtual Screening: What Is There to Know? *Int. J. Mol. Sci.* **2019**, *20* (6), 1375. https://doi.org/10.3390/ijms20061375.
(77) da Silva Rocha, S. F. L.; Olanda, C. G.; Fokoue, H. H.; Sant'Anna, C. M. R. Virtual Screening Techniques in Drug Discovery: Review and Recent Applications. *Curr. Top. Med. Chem.* **2019**, *19* (19), 1751–1767. https://doi.org/10.2174/1568026619666190816101948.
(78) Pagadala, N. S.; Syed, K.; Tuszynski, J. Software for Molecular Docking: A Review. *Biophys. Rev.* **2017**, *9* (2), 91–102. https://doi.org/10.1007/s12551-016-0247-1.
(79) Verdonk, M. L.; Giangreco, I.; Hall, R. J.; Korb, O.; Mortenson, P. N.; Murray, C. W. Docking Performance of Fragments and Druglike Compounds. *J. Med. Chem.* **2011**, *54* (15), 5422–5431. https://doi.org/10.1021/jm200558u.
(80) Kirsch, P.; Hartman, A. M.; Hirsch, A. K. H.; Empting, M. Concepts and Core Principles of Fragment-Based Drug Design. *Molecules* **2019**, *24* (23), 4309. https://doi.org/10.3390/molecules24234309.
(81) Navratilova, I.; Hopkins, A. L. Fragment Screening by Surface Plasmon Resonance. *ACS Med. Chem. Lett.* **2010**, *1* (1), 44–48. https://doi.org/10.1021/ml900002k.
(82) Senisterra, G.; Chau, I.; Vedadi, M. Thermal Denaturation Assays in Chemical Biology. *ASSAY Drug Dev. Technol.* **2011**, *10* (2), 128–136. https://doi.org/10.1089/adt.2011.0390.
(83) Jerabek-Willemsen, M.; André, T.; Wanner, R.; Roth, H. M.; Duhr, S.; Baaske, P.; Breitsprecher, D. MicroScale Thermophoresis: Interaction Analysis and Beyond. *J. Mol. Struct.* **2014**, *1077*, 101–113. https://doi.org/10.1016/j.molstruc.2014.03.009.
(84) Chilingaryan, Z.; Yin, Z.; Oakley, A. J. Fragment-Based Screening by Protein Crystallography: Successes and Pitfalls. *Int. J. Mol. Sci.* **2012**, *13* (10), 12857–12879. https://doi.org/10.3390/ijms131012857.
(85) Davies, T. G.; Tickle, I. J. Fragment Screening Using X-Ray Crystallography. In *Fragment-Based Drug Discovery and X-Ray Crystallography*; Davies, T. G., Hyvönen, M., Eds.; Topics in Current Chemistry; Springer: Berlin, Heidelberg, 2012; pp 33–59. https://doi.org/10.1007/128_2011_179.
(86) Harner, M. J.; Frank, A. O.; Fesik, S. W. Fragment-Based Drug Discovery Using NMR Spectroscopy. *J. Biomol. NMR* **2013**, *56* (2), 65–75. https://doi.org/10.1007/s10858-013-9740-z.
(87) Hopkins, A. L.; Groom, C. R.; Alex, A. Ligand Efficiency: A Useful Metric for Lead Selection. *Drug Discov. Today* **2004**, *9* (10), 430–431. https://doi.org/10.1016/S1359-6446(04)03069-7.
(88) Shultz, M. D. Setting Expectations in Molecular Optimizations: Strengths and Limitations of Commonly Used Composite Parameters. *Bioorg. Med. Chem. Lett.* **2013**, *23* (21), 5980–5991. https://doi.org/10.1016/j.bmcl.2013.08.029.
(89) Schneider, G.; Clément-Chomienne, O.; Hilfiger, L.; Schneider, P.; Kirsch, S.; Böhm, H.-J.; Neidhart, W. Virtual Screening for Bioactive Molecules by Evolutionary De Novo Design. *Angew. Chem. Int. Ed.* **2000**, *39* (22), 4130–4133. https://doi.org/10.1002/1521-3773(20001117)39:22<4130::AID-ANIE4130>3.0.CO;2-E.
(90) Gupta, R.; Srivastava, D.; Sahu, M.; Tiwari, S.; Ambasta, R. K.; Kumar, P. Artificial Intelligence to Deep Learning: Machine Intelligence Approach for Drug Discovery. *Mol. Divers.* **2021**. https://doi.org/10.1007/s11030-021-10217-3.
(91) Lamoree, B.; Hubbard, R. E. Current Perspectives in Fragment-Based Lead Discovery (FBLD). *Essays Biochem.* **2017**, *61* (5), 453–464. https://doi.org/10.1042/EBC20170028.
(92) Kirsch, P.; Jakob, V.; Oberhausen, K.; Stein, S. C.; Cucarro, I.; Schulz, T. F.; Empting, M. Fragment-Based Discovery of a Qualified Hit Targeting the Latency-Associated Nuclear Antigen of the Oncogenic Kaposi's Sarcoma-Associated Herpesvirus/Human Herpesvirus 8. *J. Med. Chem.* **2019**, *62* (8), 3924–3939. https://doi.org/10.1021/acs.jmedchem.8b01827.
(93) Nishibata, Y.; Itai, A. Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation. *Tetrahedron* **1991**, *47* (43), 8985–8990. https://doi.org/10.1016/S0040-4020(01)86503-0.

(94) Rotstein, S. H.; Murcko, M. A. GenStar: A Method for de Novo Drug Design. *J. Comput. Aided Mol. Des.* **1993**, *7* (1), 23–43. https://doi.org/10.1007/BF00141573.
(95) Luo, Z.; Wang, R.; Lai, L. RASSE: A New Method for Structure-Based Drug Design. *J. Chem. Inf. Comput. Sci.* **1996**, *36* (6), 1187–1194. https://doi.org/10.1021/ci950277w.
(96) Rotstein, S. H.; Murcko, M. A. GroupBuild: A Fragment-Based Method for de Novo Drug Design. *J. Med. Chem.* **1993**, *36* (12), 1700–1710. https://doi.org/10.1021/jm00064a003.
(97) Moon, J. B.; Howe, W. J. Computer Design of Bioactive Molecules: A Method for Receptor-Based de Novo Ligand Design. *Proteins Struct. Funct. Genet.* **1991**, *11* (4), 314–328. https://doi.org/10.1002/prot.340110409.
(98) Gillet, V.; Johnson, A. P.; Mata, P.; Sike, S.; Williams, P. SPROUT: A Program for Structure Generation. *J. Comput. Aided Mol. Des.* **1993**, *7* (2), 127–153. https://doi.org/10.1007/BF00126441.
(99) Pearlman, D. A.; Murcko, M. A. CONCEPTS: New Dynamic Algorithm for de Novo Drug Suggestion. *J. Comput. Chem.* **1993**, *14* (10), 1184–1193. https://doi.org/10.1002/jcc.540141008.
(100) Payne, A. W. R.; Glen, R. C. Molecular Recognition Using a Binary Genetic Search Algorithm. *J. Mol. Graph.* **1993**, *11* (2), 74–91. https://doi.org/10.1016/0263-7855(93)87001-L.
(101) Bohacek, R. S.; McMartin, C. Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth. *J. Am. Chem. Soc.* **1994**, *116* (13), 5560–5571. https://doi.org/10.1021/ja00092a006.
(102) Miranker, A.; Karplus, M. An Automated Method for Dynamic Ligand Design. *Proteins Struct. Funct. Bioinforma.* **1995**, *23* (4), 472–490. https://doi.org/10.1002/prot.340230403.
(103) Pearlman, D. A.; Murcko, M. A. CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design. *J. Med. Chem.* **1996**, *39* (8), 1651–1663. https://doi.org/10.1021/jm950792l.
(104) DeWitte, R. S.; Shakhnovich, E. I. SMoG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 1. Methodology and Supporting Evidence. *J. Am. Chem. Soc.* **1996**, *118* (47), 11733–11744. https://doi.org/10.1021/ja960751u.
(105) Todorov, N. P.; Dean, P. M. Evaluation of a Method for Controlling Molecular Scaffold Diversity in de Novo Ligand Design. 18.
(106) Douguet, D.; Thoreau, E.; Grassy, G. A Genetic Algorithm for the Automated Generation of Small Organic Molecules: Drug Design Using an Evolutionary Algorithm. *J. Comput. Aided Mol. Des.* **2000**, *14* (5), 449–466. https://doi.org/10.1023/A:1008108423895.
(107) Zhu, J.; Fan, H.; Liu, H.; Shi, Y. Structure-Based Ligand Design for Flexible Proteins: Application of New F-DycoBlock. *J. Comput. Aided Mol. Des.* **2001**, *15* (11), 979–996. https://doi.org/10.1023/A:1014817911249.
(108) Pegg, S. C.-H.; Haresco, J. J.; Kuntz, I. D. A Genetic Algorithm for Structure-Based de Novo Design. *J. Comput. Aided Mol. Des.* **2001**, *15* (10), 911–933. https://doi.org/10.1023/A:1014389729000.
(109) Nicolas, B.; Shaheen, A.; Nicolas, M.; Amedeo, C. An Evolutionary Approach for Structure-Based Design of Natural and Non-Natural Peptidic Ligands. *Comb. Chem. High Throughput Screen.* **2001**, *4* (8), 661–673.
(110) Vinkers, H. M.; de Jonge, M. R.; Daeyaert, F. F. D.; Heeres, J.; Koymans, L. M. H.; van Lenthe, J. H.; Lewi, P. J.; Timmerman, H.; Van Aken, K.; Janssen, P. A. J. SYNOPSIS: SYNthesize and OPtimize System in Silico. *J. Med. Chem.* **2003**, *46* (13), 2765–2773. https://doi.org/10.1021/jm030809x.
(111) Brown, N.; McKay, B.; Gilardoni, F.; Gasteiger, J. A Graph-Based Genetic Algorithm and Its Application to the Multiobjective Evolution of Median Molecules. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 1079–1087. https://doi.org/10.1021/ci034290p.
(112) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303* (5665), 1813–1818. https://doi.org/10.1126/science.1096361.
(113) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48* (7), 2457–2468. https://doi.org/10.1021/jm0492296.

(114) Lameijer, E.-W.; Kok, J. N.; Bäck, T.; IJzerman, A. P. The Molecule Evoluator. An Interactive Evolutionary Algorithm for the Design of Drug-Like Molecules. *J. Chem. Inf. Model.* **2006**, *46* (2), 545–552. https://doi.org/10.1021/ci050369d.

(115) Kutchukian, P. S.; Lou, D.; Shakhnovich, E. I. FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space. *J. Chem. Inf. Model.* **2009**, *49* (7), 1630–1642. https://doi.org/10.1021/ci9000458.

(116) Nicolaou, C. A.; Apostolakis, J.; Pattichis, C. S. De Novo Drug Design Using Multiobjective Evolutionary Graphs. *J. Chem. Inf. Model.* **2009**, *49* (2), 295–307. https://doi.org/10.1021/ci800308h.

(117) Durrant, J. D.; Amaro, R. E.; McCammon, J. A. AutoGrow: A Novel Algorithm for Protein Inhibitor Design. *Chem. Biol. Drug Des.* **2009**, *73* (2), 168–178. https://doi.org/10.1111/j.1747-0285.2008.00761.x.

(118) Hartenfeller, M.; Zettl, H.; Walter, M.; Rupp, M.; Reisen, F.; Proschak, E.; Weggen, S.; Stark, H.; Schneider, G. DOGS: Reaction-Driven de Novo Design of Bioactive Compounds. *PLoS Comput. Biol.* **2012**, *8* (2), e1002380. https://doi.org/10.1371/journal.pcbi.1002380.

(119) Durrant, J. D.; Lindert, S.; McCammon, J. A. AutoGrow 3.0: An Improved Algorithm for Chemically Tractable, Semi-Automated Protein Inhibitor Design. *J. Mol. Graph. Model.* **2013**, *44*, 104–112. https://doi.org/10.1016/j.jmgm.2013.05.006.

(120) Li, G.-B.; Ji, S.; Yang, L.-L.; Zhang, R.-J.; Chen, K.; Zhong, L.; Ma, S.; Yang, S.-Y. LEADOPT: An Automatic Tool for Structure-Based Lead Optimization, and Its Application in Structural Optimizations of VEGFR2 and SYK Inhibitors. *Eur. J. Med. Chem.* **2015**, *93*, 523–538. https://doi.org/10.1016/j.ejmech.2015.02.019.

(121) Chéron, N.; Jasty, N.; Shakhnovich, E. I. OpenGrowth: An Automated and Rational Algorithm for Finding New Protein Ligands. *J. Med. Chem.* **2016**, *59* (9), 4171–4188. https://doi.org/10.1021/acs.jmedchem.5b00886.

(122) Lin, F.-Y.; Esposito, E. X.; Tseng, Y. J. LeadOp+R: Structure-Based Lead Optimization With Synthetic Accessibility. *Front. Pharmacol.* **2018**, *9*. https://doi.org/10.3389/fphar.2018.00096.

(123) Spiegel, J. O.; Durrant, J. D. AutoGrow4: An Open-Source Genetic Algorithm for de Novo Drug Design and Lead Optimization. *J. Cheminformatics* **2020**, *12* (1), 25. https://doi.org/10.1186/s13321-020-00429-4.

(124) Schneider, G.; Fechner, U. Computer-Based de Novo Design of Drug-like Molecules. *Nat. Rev. Drug Discov.* **2005**, *4* (8), 649–663. https://doi.org/10.1038/nrd1799.

(125) Murray, C. W.; Verdonk, M. L. The Consequences of Translational and Rotational Entropy Lost by Small Molecules on Binding to Proteins. *J. Comput. Aided Mol. Des.* **2002**, *16* (10), 741–753. https://doi.org/10.1023/A:1022446720849.

(126) Maass, P.; Schulz-Gasch, T.; Stahl, M.; Rarey, M. Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations. *J. Chem. Inf. Model.* **2007**, *47* (2), 390–399. https://doi.org/10.1021/ci060094h.

(127) Bartlett, P.; Shea, G.; Telfer, S.; Waterman, S. In Roberts, SM (Ed.) Molecular Recognition: Chemical and Biological Problems. *R. Soc. Lond.* **1989**, 182–196.

(128) Lewis, R. A.; Roe, D. C.; Huang, C.; Ferrin, T. E.; Langridge, R.; Kuntz, I. D. Automated Site-Directed Drug Design Using Molecular Lattices. *J. Mol. Graph.* **1992**, *10* (2), 66–78. https://doi.org/10.1016/0263-7855(92)80059-M.

(129) Lawrence, M. C.; Davis, P. C. CLIX: A Search Algorithm for Finding Novel Ligands Capable of Binding Proteins of Known Three-Dimensional Structure. *Proteins Struct. Funct. Bioinforma.* **1992**, *12* (1), 31–41. https://doi.org/10.1002/prot.340120105.

(130) Ho, C. M. W.; Marshall, G. R. SPLICE: A Program to Assemble Partial Query Solutions from Three-Dimensional Database Searches into Novel Ligands. *J. Comput. Aided Mol. Des.* **1993**, *7* (6), 623–647. https://doi.org/10.1007/BF00125322.

(131) Tschinke, V.; Cohen, N. C. The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses. *J. Med. Chem.* **1993**, *36* (24), 3863–3870. https://doi.org/10.1021/jm00076a016.

(132) Leach, A. R.; Kilvington, S. R. Automated Molecular Design: A New Fragment-Joining Algorithm. *J. Comput. Aided Mol. Des.* **1994**, *8* (3), 283–298. https://doi.org/10.1007/BF00126746.
(133) Eisen, M. B.; Wiley, D. C.; Karplus, M.; Hubbard, R. E. HOOK: A Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site. *Proteins* **1994**, *19* (3), 199–221. https://doi.org/10.1002/prot.340190305.
(134) Roe, D. C.; Kuntz, I. D. BUILDER v.2: Improving the Chemistry of a de Novo Design Strategy. *J. Comput. Aided Mol. Des.* **1995**, *9* (3), 269–282. https://doi.org/10.1007/BF00124457.
(135) Clark, D. E.; Frenkel, D.; Levy, S. A.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B.; Westhead, D. R. PRO_LIGAND: An Approach to de Novo Molecular Design. 1. Application to the Design of Organic Molecules. *J. Comput. Aided Mol. Des.* **1995**, *9* (1), 13–32. https://doi.org/10.1007/BF00117275.
(136) Fechner, U.; Schneider, G. Flux (1): A Virtual Synthesis Scheme for Fragment-Based de Novo Design. *J. Chem. Inf. Model.* **2006**, *46* (2), 699–707. https://doi.org/10.1021/ci0503560.
(137) Fechner, U.; Schneider, G. Flux (2): Comparison of Molecular Mutation and Crossover Operators for Ligand-Based de Novo Design. *J. Chem. Inf. Model.* **2007**, *47* (2), 656–667. https://doi.org/10.1021/ci6005307.
(138) Thompson, D. C.; Aldrin Denny, R.; Nilakantan, R.; Humblet, C.; Joseph-McCarthy, D.; Feyfant, E. CONFIRM: Connecting Fragments Found in Receptor Molecules. *J. Comput. Aided Mol. Des.* **2008**, *22* (10), 761. https://doi.org/10.1007/s10822-008-9221-8.
(139) Dey, F.; Caflisch, A. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48* (3), 679–690. https://doi.org/10.1021/ci700424b.
(140) Huang, Q.; Li, L.-L.; Yang, S.-Y. PhDD: A New Pharmacophore-Based de Novo Design Method of Drug-like Molecules Combined with Assessment of Synthetic Accessibility. *J. Mol. Graph. Model.* **2010**, *28* (8), 775–787. https://doi.org/10.1016/j.jmgm.2010.02.002.
(141) Lin, F.-Y.; Tseng, Y. J. Structure-Based Fragment Hopping for Lead Optimization Using Predocked Fragment Database. *J. Chem. Inf. Model.* **2011**, *51* (7), 1703–1715. https://doi.org/10.1021/ci200136j.
(142) Hao, G.-F.; Jiang, W.; Ye, Y.-N.; Wu, F.-X.; Zhu, X.-L.; Guo, F.-B.; Yang, G.-F. ACFIS: A Web Server for Fragment-Based Drug Discovery. *Nucleic Acids Res.* **2016**, *44* (W1), W550–W556. https://doi.org/10.1093/nar/gkw393.
(143) Wang, R.; Gao, Y.; Lai, L. LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design. *Mol. Model. Annu.* **2000**, *6* (7), 498–516. https://doi.org/10.1007/s0089400060498.
(144) Degen, J.; Rarey, M. FlexNovo: Structure-Based Searching in Large Fragment Spaces. *ChemMedChem* **2006**, *1* (8), 854–868. https://doi.org/10.1002/cmdc.200500102.
(145) Liu, Q.; Masek, B.; Smith, K.; Smith, J. Tagged Fragment Method for Evolutionary Structure-Based De Novo Lead Generation and Optimization. *J. Med. Chem.* **2007**, *50* (22), 5392–5402. https://doi.org/10.1021/jm070750k.
(146) Yuan, Y.; Pei, J.; Lai, L. LigBuilder 2: A Practical de Novo Drug Design Approach. *J. Chem. Inf. Model.* **2011**, *51* (5), 1083–1091. https://doi.org/10.1021/ci100350u.
(147) Teodoro, M.; Muegge, I. BIBuilder: Exhaustive Searching for De Novo Ligands. *Mol. Inform.* **2011**, *30* (1), 63–75. https://doi.org/10.1002/minf.201000122.
(148) Ertl, P.; Lewis, R. IADE: A System for Intelligent Automatic Design of Bioisosteric Analogs. *J. Comput. Aided Mol. Des.* **2012**, *26* (11), 1207–1215. https://doi.org/10.1007/s10822-012-9609-3.
(149) Beccari, A. R.; Cavazzoni, C.; Beato, C.; Costantino, G. LiGen: A High Performance Workflow for Chemistry Driven de Novo Design. *J. Chem. Inf. Model.* **2013**, *53* (6), 1518–1527. https://doi.org/10.1021/ci400078g.
(150) Yuan, Y.; Pei, J.; Lai, L. LigBuilder V3: A Multi-Target de Novo Drug Design Approach. *Front. Chem.* **2020**, *8*. https://doi.org/10.3389/fchem.2020.00142.
(151) Mouchlis, V. D.; Afantitis, A.; Serra, A.; Fratello, M.; Papadiamantis, A. G.; Aidinis, V.; Lynch, I.; Greco, D.; Melagraki, G. Advances in De Novo Drug Design: From Conventional to Machine Learning Methods. *Int. J. Mol. Sci.* **2021**, *22* (4). https://doi.org/10.3390/ijms22041676.

(152) Wong, S. S. Y.; Luo, W.; Chan, K. C. C. EvoMD: An Algorithm for Evolutionary Molecular Design. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2011**, *8* (4), 987–1003. https://doi.org/10.1109/TCBB.2010.100.
(153) Pierce, A. C.; Rao, G.; Bemis, G. W. BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease. *J. Med. Chem.* **2004**, *47* (11), 2768–2775. https://doi.org/10.1021/jm030543u.
(154) Li, Y.; Zhao, Y.; Liu, Z.; Wang, R. Automatic Tailoring and Transplanting: A Practical Method That Makes Virtual Screening More Useful. *J. Chem. Inf. Model.* **2011**, *51* (6), 1474–1491. https://doi.org/10.1021/ci200036m.
(155) Lindert, S.; Durrant, J. D.; McCammon, J. A. LigMerge: A Fast Algorithm to Generate Models of Novel Potential Ligands from Sets of Known Binders. *Chem. Biol. Drug Des.* **2012**, *80* (3), 358–365. https://doi.org/10.1111/j.1747-0285.2012.01414.x.
(156) Li, Y.; Zhao, Z.; Liu, Z.; Su, M.; Wang, R. AutoT&T v.2: An Efficient and Versatile Tool for Lead Structure Generation and Optimization. *J. Chem. Inf. Model.* **2016**, *56* (2), 435–453. https://doi.org/10.1021/acs.jcim.5b00691.
(157) Chevillard, F.; Rimmer, H.; Betti, C.; Pardon, E.; Ballet, S.; van Hilten, N.; Steyaert, J.; Diederich, W. E.; Kolb, P. Binding-Site Compatible Fragment Growing Applied to the Design of B2-Adrenergic Receptor Ligands. *J. Med. Chem.* **2018**, *61* (3), 1118–1129. https://doi.org/10.1021/acs.jmedchem.7b01558.
(158) Devi, R. V.; Sathya, S. S.; Coumar, M. S. Evolutionary Algorithms for de Novo Drug Design – A Survey. *Appl. Soft Comput.* **2015**, *27*, 543–552. https://doi.org/10.1016/j.asoc.2014.09.042.
(159) Russell, S.; Norvig, P. *Artificial Intelligence: A Modern Approach*, 4e édition.; Pearson, 2021.
(160) Chan, H. C. S.; Shan, H.; Dahoun, T.; Vogel, H.; Yuan, S. Advancing Drug Discovery via Artificial Intelligence. *Trends Pharmacol. Sci.* **2019**, *40* (8), 592–604. https://doi.org/10.1016/j.tips.2019.06.004.
(161) Klambauer, G.; Hochreiter, S.; Rarey, M. Machine Learning in Drug Discovery. *J. Chem. Inf. Model.* **2019**, *59* (3), 945–946. https://doi.org/10.1021/acs.jcim.9b00136.
(162) Han, M.; May, R.; Zhang, X.; Wang, X.; Pan, S.; Yan, D.; Jin, Y.; Xu, L. A Review of Reinforcement Learning Methodologies for Controlling Occupant Comfort in Buildings. *Sustain. Cities Soc.* **2019**, *51*, 101748. https://doi.org/10.1016/j.scs.2019.101748.
(163) LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. *Nature* **2015**, *521* (7553), 436–444. https://doi.org/10.1038/nature14539.
(164) David, L.; Thakkar, A.; Mercado, R.; Engkvist, O. Molecular Representations in AI-Driven Drug Discovery: A Review and Practical Guide. *J. Cheminformatics* **2020**, *12* (1), 56. https://doi.org/10.1186/s13321-020-00460-5.
(165) Liu, X.; Ye, K.; van Vlijmen, H. W. T.; IJzerman, A. P.; van Westen, G. J. P. An Exploration Strategy Improves the Diversity of de Novo Ligands Using Deep Reinforcement Learning: A Case for the Adenosine A2A Receptor. *J. Cheminformatics* **2019**, *11* (1), 35. https://doi.org/10.1186/s13321-019-0355-6.
(166) Yang, X.; Wang, Y.; Byrne, R.; Schneider, G.; Yang, S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. *Chem. Rev.* **2019**, *119* (18), 10520–10594. https://doi.org/10.1021/acs.chemrev.8b00728.
(167) Pineda, F. J. Generalization of Back-Propagation to Recurrent Neural Networks. *Phys. Rev. Lett.* **1987**, *59* (19), 2229–2232. https://doi.org/10.1103/PhysRevLett.59.2229.
(168) Lipton, Z. C.; Berkowitz, J.; Elkan, C. A Critical Review of Recurrent Neural Networks for Sequence Learning. *ArXiv150600019 Cs* **2015**.
(169) Lavecchia, A.; Di Giovanni, C. Virtual Screening Strategies in Drug Discovery: A Critical Review. *Curr. Med. Chem.* **2013**, *20* (23), 2839–2860. https://doi.org/10.2174/09298673113209990001.
(170) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23* (25), 5966–5971. https://doi.org/10.1002/chem.201605499.

(171) Merk, D.; Friedrich, L.; Grisoni, F.; Schneider, G. De Novo Design of Bioactive Small Molecules by Artificial Intelligence. *Mol. Inform.* **2018**, *37* (1–2), 1700153. https://doi.org/10.1002/minf.201700153.
(172) Popova, M.; Isayev, O.; Tropsha, A. Deep Reinforcement Learning for de Novo Drug Design. *Sci. Adv.* **2018**, *4* (7), eaap7885. https://doi.org/10.1126/sciadv.aap7885.
(173) Maragakis, P.; Nisonoff, H.; Cole, B.; Shaw, D. E. A Deep-Learning View of Chemical Space Designed to Facilitate Drug Discovery. *J. Chem. Inf. Model.* **2020**, *60* (10), 4487–4496. https://doi.org/10.1021/acs.jcim.0c00321.
(174) Kotsias, P.-C.; Arús-Pous, J.; Chen, H.; Engkvist, O.; Tyrchan, C.; Bjerrum, E. J. Direct Steering of de Novo Molecular Generation with Descriptor Conditional Recurrent Neural Networks. *Nat. Mach. Intell.* **2020**, *2* (5), 254–265. https://doi.org/10.1038/s42256-020-0174-5.
(175) Grisoni, F.; Moret, M.; Lingwood, R.; Schneider, G. Bidirectional Molecule Generation with Recurrent Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (3), 1175–1183. https://doi.org/10.1021/acs.jcim.9b00943.
(176) Ståhl, N.; Falkman, G.; Karlsson, A.; Mathiason, G.; Boström, J. Deep Reinforcement Learning for Multiparameter Optimization in de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59* (7), 3166–3176. https://doi.org/10.1021/acs.jcim.9b00325.
(177) Pan, S. J.; Yang, Q. A Survey on Transfer Learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22* (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
(178) Gupta, A.; Müller, A. T.; Huisman, B. J. H.; Fuchs, J. A.; Schneider, P.; Schneider, G. Generative Recurrent Networks for De Novo Drug Design. *Mol. Inform.* **2018**, *37* (1–2), 1700111. https://doi.org/10.1002/minf.201700111.
(179) Yasonik, J. Multiobjective de Novo Drug Design with Recurrent Neural Networks and Nondominated Sorting. *J. Cheminformatics* **2020**, *12* (1), 14. https://doi.org/10.1186/s13321-020-00419-6.
(180) Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* **1998**, *86* (11), 2278–2324. https://doi.org/10.1109/5.726791.
(181) Rifaioglu, A. S.; Nalbat, E.; Atalay, V.; Martin, M. J.; Cetin-Atalay, R.; Doğan, T. DEEPScreen: High Performance Drug–Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations. *Chem. Sci.* **2020**, *11* (9), 2531–2557. https://doi.org/10.1039/C9SC03414E.
(182) Sun, M.; Zhao, S.; Gilvary, C.; Elemento, O.; Zhou, J.; Wang, F. Graph Convolutional Networks for Computational Drug Development and Discovery. *Brief. Bioinform.* **2020**, *21* (3), 919–935. https://doi.org/10.1093/bib/bbz042.
(183) Li, Y.; Hu, J.; Wang, Y.; Zhou, J.; Zhang, L.; Liu, Z. DeepScaffold: A Comprehensive Tool for Scaffold-Based De Novo Drug Discovery Using Deep Learning. *J. Chem. Inf. Model.* **2020**, *60* (1), 77–91. https://doi.org/10.1021/acs.jcim.9b00727.
(184) Khemchandani, Y.; O'Hagan, S.; Samanta, S.; Swainston, N.; Roberts, T. J.; Bollegala, D.; Kell, D. B. DeepGraphMolGen, a Multi-Objective, Computational Strategy for Generating Molecules with Desirable Properties: A Graph Convolution and Reinforcement Learning Approach. *J. Cheminformatics* **2020**, *12* (1), 53. https://doi.org/10.1186/s13321-020-00454-3.
(185) Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. 9.
(186) Yi, X.; Walia, E.; Babyn, P. Generative Adversarial Network in Medical Imaging: A Review. *Med. Image Anal.* **2019**, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552.
(187) Putin, E.; Asadulaev, A.; Vanhaelen, Q.; Ivanenkov, Y.; Aladinskaya, A. V.; Aliper, A.; Zhavoronkov, A. Adversarial Threshold Neural Computer for Molecular de Novo Design. *Mol. Pharm.* **2018**, *15* (10), 4386–4397. https://doi.org/10.1021/acs.molpharmaceut.7b01137.
(188) Prykhodko, O.; Johansson, S. V.; Kotsias, P.-C.; Arús-Pous, J.; Bjerrum, E. J.; Engkvist, O.; Chen, H. A de Novo Molecular Generation Method Using Latent Vector Based Generative Adversarial Network. *J. Cheminformatics* **2019**, *11* (1), 74. https://doi.org/10.1186/s13321-019-0397-9.

(189) Putin, E.; Asadulaev, A.; Ivanenkov, Y.; Aladinskiy, V.; Sanchez-Lengeling, B.; Aspuru-Guzik, A.; Zhavoronkov, A. Reinforced Adversarial Neural Computer for de Novo Molecular Design. *J. Chem. Inf. Model.* **2018**, *58* (6), 1194–1204. https://doi.org/10.1021/acs.jcim.7b00690.

(190) Hinton, G. E.; Salakhutdinov, R. R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313* (5786), 504–507. https://doi.org/10.1126/science.1127647.

(191) Dong, G.; Liao, G.; Liu, H.; Kuang, G. A Review of the Autoencoder and Its Variants: A Comparative Perspective from Target Recognition in Synthetic-Aperture Radar Images. *IEEE Geosci. Remote Sens. Mag.* **2018**, *6* (3), 44–68. https://doi.org/10.1109/MGRS.2018.2853555.

(192) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4* (2), 268–276. https://doi.org/10.1021/acscentsci.7b00572.

(193) Lim, J.; Ryu, S.; Kim, J. W.; Kim, W. Y. Molecular Generative Model Based on Conditional Variational Autoencoder for de Novo Molecular Design. *J. Cheminformatics* **2018**, *10* (1), 31. https://doi.org/10.1186/s13321-018-0286-7.

(194) Skalic, M.; Jiménez, J.; Sabbadin, D.; De Fabritiis, G. Shape-Based Generative Modeling for de Novo Drug Design. *J. Chem. Inf. Model.* **2019**, *59* (3), 1205–1214. https://doi.org/10.1021/acs.jcim.8b00706.

(195) Bjerrum, E. J.; Sattarov, B. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules* **2018**, *8* (4), 131. https://doi.org/10.3390/biom8040131.

(196) Gao, K.; Nguyen, D. D.; Tu, M.; Wei, G.-W. Generative Network Complex for the Automated Generation of Drug-like Molecules. *J. Chem. Inf. Model.* **2020**, *60* (12), 5682–5698. https://doi.org/10.1021/acs.jcim.0c00599.

(197) Kadurin, A.; Aliper, A.; Kazennov, A.; Mamoshina, P.; Vanhaelen, Q.; Khrabrov, K.; Zhavoronkov, A. The Cornucopia of Meaningful Leads: Applying Deep Adversarial Autoencoders for New Molecule Development in Oncology. *Oncotarget* **2016**, *8* (7), 10883–10890. https://doi.org/10.18632/oncotarget.14073.

(198) Kadurin, A.; Nikolenko, S.; Khrabrov, K.; Aliper, A.; Zhavoronkov, A. DruGAN: An Advanced Generative Adversarial Autoencoder Model for de Novo Generation of New Molecules with Desired Molecular Properties in Silico. *Mol. Pharm.* **2017**, *14* (9), 3098–3104. https://doi.org/10.1021/acs.molpharmaceut.7b00346.

(199) Zhavoronkov, A.; Ivanenkov, Y. A.; Aliper, A.; Veselov, M. S.; Aladinskiy, V. A.; Aladinskaya, A. V.; Terentiev, V. A.; Polykovskiy, D. A.; Kuznetsov, M. D.; Asadulaev, A.; Volkov, Y.; Zholus, A.; Shayakhmetov, R. R.; Zhebrak, A.; Minaeva, L. I.; Zagribelnyy, B. A.; Lee, L. H.; Soll, R.; Madge, D.; Xing, L.; Guo, T.; Aspuru-Guzik, A. Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors. *Nat. Biotechnol.* **2019**, *37* (9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x.

(200) Imrie, F.; Bradley, A. R.; van der Schaar, M.; Deane, C. M. Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* **2020**, *60* (4), 1983–1995. https://doi.org/10.1021/acs.jcim.9b01120.

(201) Castelvecchi, D. Can We Open the Black Box of AI? *Nat. News* **2016**, *538* (7623), 20. https://doi.org/10.1038/538020a.

(202) Omenn, G. S.; Lane, L.; Overall, C. M.; Cristea, I. M.; Corrales, F. J.; Lindskog, C.; Paik, Y.-K.; Van Eyk, J. E.; Liu, S.; Pennington, S. R.; Snyder, M. P.; Baker, M. S.; Bandeira, N.; Aebersold, R.; Moritz, R. L.; Deutsch, E. W. Research on the Human Proteome Reaches a Major Milestone: >90% of Predicted Human Proteins Now Credibly Detected, According to the HUPO Human Proteome Project. *J. Proteome Res.* **2020**, *19* (12), 4735–4746. https://doi.org/10.1021/acs.jproteome.0c00485.

(203) UniProt Consortium, T. UniProt: The Universal Protein Knowledgebase. *Nucleic Acids Res.* **2018**, *46* (5), 2699. https://doi.org/10.1093/nar/gky092.

(204) Wishart, D. S. DrugBank: A Comprehensive Resource for in Silico Drug Discovery and Exploration. *Nucleic Acids Res.* **2006**, *34* (90001), D668–D672. https://doi.org/10.1093/nar/gkj067.
(205) Uhlén, M.; Fagerberg, L.; Hallström, B. M.; Lindskog, C.; Oksvold, P.; Mardinoglu, A.; Sivertsson, Å.; Kampf, C.; Sjöstedt, E.; Asplund, A.; Olsson, I.; Edlund, K.; Lundberg, E.; Navani, S.; Szigyarto, C. A.-K.; Odeberg, J.; Djureinovic, D.; Takanen, J. O.; Hober, S.; Alm, T.; Edqvist, P.-H.; Berling, H.; Tegel, H.; Mulder, J.; Rockberg, J.; Nilsson, P.; Schwenk, J. M.; Hamsten, M.; Feilitzen, K. von; Forsberg, M.; Persson, L.; Johansson, F.; Zwahlen, M.; Heijne, G. von; Nielsen, J.; Pontén, F. Tissue-Based Map of the Human Proteome. *Science* **2015**, *347* (6220). https://doi.org/10.1126/science.1260419.
(206) The druggable proteome https://www.proteinatlas.org/humanproteome/tissue/druggable (accessed 2021 -04 -30).
(207) Santos, R.; Ursu, O.; Gaulton, A.; Bento, A. P.; Donadi, R. S.; Bologa, C. G.; Karlsson, A.; Al-Lazikani, B.; Hersey, A.; Oprea, T. I.; Overington, J. P. A Comprehensive Map of Molecular Drug Targets. *Nat. Rev. Drug Discov.* **2017**, *16* (1), 19–34. https://doi.org/10.1038/nrd.2016.230.
(208) Overington, J. P.; Al-Lazikani, B.; Hopkins, A. L. How Many Drug Targets Are There? *Nat. Rev. Drug Discov.* **2006**, *5* (12), 993–996. https://doi.org/10.1038/nrd2199.
(209) Cohen, P. Protein Kinases — the Major Drug Targets of the Twenty-First Century? *Nat. Rev. Drug Discov.* **2002**, *1* (4), 309–315. https://doi.org/10.1038/nrd773.
(210) Cohen, P. The Origins of Protein Phosphorylation. *Nat. Cell Biol.* **2002**, *4* (5), E127–E130. https://doi.org/10.1038/ncb0502-e127.
(211) Roskoski, R. A Historical Overview of Protein Kinases and Their Targeted Small Molecule Inhibitors. *Pharmacol. Res.* **2015**, *100*, 1–23. https://doi.org/10.1016/j.phrs.2015.07.010.
(212) Hammarsten, O. Zur Frage, Ob Das Caseïn Ein Einheitlicher Stoff Sei. **1883**, *7* (3), 227–273. https://doi.org/10.1515/bchm1.1883.7.3.227.
(213) Levene, P. A.; Alsberg, C. Zur Chemie Der Paranucleinsäure. **1901**, *31* (5–6), 543–555. https://doi.org/10.1515/bchm2.1901.31.5-6.543.
(214) Lipmann, F. A.; Levene, P. A. SERINEPHOSPHORIC ACID OBTAINED ON HYDROLYSIS OF VITELLINIC ACID. *J. Biol. Chem.* **1932**, *98* (1), 109–114. https://doi.org/10.1016/S0021-9258(18)76142-5.
(215) De Verdier, C.-H. Isolation of Phosphothreonine from Bovine Casein. *Nature* **1952**, *170* (4332), 804–805. https://doi.org/10.1038/170804b0.
(216) Eckhart, W.; Hutchinson, M. A.; Hunter, T. An Activity Phosphorylating Tyrosine in Polyoma T Antigen Immunoprecipitates. *Cell* **1979**, *18* (4), 925–933. https://doi.org/10.1016/0092-8674(79)90205-8.
(217) Burnett, G.; Kennedy, E. P. THE ENZYMATIC PHOSPHORYLATION OF PROTEINS. *J. Biol. Chem.* **1954**, *211* (2), 969–980. https://doi.org/10.1016/S0021-9258(18)71184-8.
(218) Krebs, E. G.; Fischer, E. H. PHOSPHORYLASE ACTIVITY OF SKELETAL MUSCLE EXTRACTS. *J. Biol. Chem.* **1955**, *216* (1), 113–120. https://doi.org/10.1016/S0021-9258(19)52288-8.
(219) Fischer, E. H.; Krebs, E. G. CONVERSION OF PHOSPHORYLASE b TO PHOSPHORYLASE a IN MUSCLE EXTRACTS. *J. Biol. Chem.* **1955**, *216* (1), 121–132. https://doi.org/10.1016/S0021-9258(19)52289-X.
(220) Sutherland, E. W.; Wosilait, W. D. Inactivation and Activation of Liver Phosphorylase. *Nature* **1955**, *175* (4447), 169–170. https://doi.org/10.1038/175169a0.
(221) The Nobel Prize in Physiology or Medicine 1992 https://www.nobelprize.org/prizes/medicine/1992/summary/ (accessed 2021 -05 -03).
(222) The Nobel Prize in Physiology or Medicine 2000 https://www.nobelprize.org/prizes/medicine/2000/summary/ (accessed 2021 -05 -03).
(223) The Nobel Prize in Physiology or Medicine 2001 https://www.nobelprize.org/prizes/medicine/2001/summary/ (accessed 2021 -05 -03).
(224) Rauch, J.; Volinsky, N.; Romano, D.; Kolch, W. The Secret Life of Kinases: Functions beyond Catalysis. *Cell Commun. Signal.* **2011**, *9* (1), 23. https://doi.org/10.1186/1478-811X-9-23.

(225) Ferguson, F. M.; Gray, N. S. Kinase Inhibitors: The Road Ahead. *Nat. Rev. Drug Discov.* **2018**, *17* (5), 353–377. https://doi.org/10.1038/nrd.2018.21.
(226) Hanahan, D.; Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **2011**, *144* (5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013.
(227) Gross, S.; Rahal, R.; Stransky, N.; Lengauer, C.; Hoeflich, K. P. Targeting Cancer with Kinase Inhibitors. *J. Clin. Invest.* **2015**, *125* (5), 1780–1789. https://doi.org/10.1172/JCI76094.
(228) Manning, G. The Protein Kinase Complement of the Human Genome. *Science* **2002**, *298* (5600), 1912–1934. https://doi.org/10.1126/science.1075762.
(229) Understudied Proteins https://commonfund.nih.gov/idg/understudiedproteins (accessed 2021 -05 -01).
(230) Fedorov, O.; Müller, S.; Knapp, S. The (Un)Targeted Cancer Kinome. *Nat. Chem. Biol.* **2010**, *6* (3), 166–169. https://doi.org/10.1038/nchembio.297.
(231) New Frontiers in Kinases: Special Issue. *ACS Med. Chem. Lett.* **2014**, *5* (4), 270–270. https://doi.org/10.1021/ml500071m.
(232) Kanev, G. K.; de Graaf, C.; de Esch, I. J. P.; Leurs, R.; Würdinger, T.; Westerman, B. A.; Kooistra, A. J. The Landscape of Atypical and Eukaryotic Protein Kinases. *Trends Pharmacol. Sci.* **2019**, *40* (11), 818–832. https://doi.org/10.1016/j.tips.2019.09.002.
(233) Scheeff, E. D.; Bourne, P. E. Structural Evolution of the Protein Kinase–Like Superfamily. *PLOS Comput. Biol.* **2005**, *1* (5), e49. https://doi.org/10.1371/journal.pcbi.0010049.
(234) Human Protein Kinases Overview https://www.cellsignal.at/learn-and-support/protein-kinases/human-protein-kinases-overview (accessed 2021 -05 -01).
(235) KinBase: Kinase Database at Manning's Group http://kinase.com/web/current/kinbase/genes/SpeciesID/9606/ (accessed 2021 -05 -03).
(236) Zheng, J.; Trafny, E. A.; Knighton, D. R.; Xuong, N.; Taylor, S. S.; Eyck, L. F. T.; Sowadski, J. M. 2.2 Å Refined Crystal Structure of the Catalytic Subunit of CAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor. *Acta Crystallogr. Sect. D* **1993**, *49* (3), 362–365. https://doi.org/10.1107/S0907444993000423.
(237) Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G. A.; Sonnhammer, E. L. L.; Tosatto, S. C. E.; Paladin, L.; Raj, S.; Richardson, L. J.; Finn, R. D.; Bateman, A. Pfam: The Protein Families Database in 2021. *Nucleic Acids Res.* **2021**, *49* (D1), D412–D419. https://doi.org/10.1093/nar/gkaa913.
(238) Kooistra, A. J.; Kanev, G. K.; van Linden, O. P. J.; Leurs, R.; de Esch, I. J. P.; de Graaf, C. KLIFS: A Structural Kinase-Ligand Interaction Database. *Nucleic Acids Res.* **2016**, *44* (D1), D365–D371. https://doi.org/10.1093/nar/gkv1082.
(239) Kanev, G. K.; de Graaf, C.; Westerman, B. A.; de Esch, I. J. P.; Kooistra, A. J. KLIFS: An Overhaul after the First 5 Years of Supporting Kinase Research. *Nucleic Acids Res.* **2020**. https://doi.org/10.1093/nar/gkaa895.
(240) Hanks, S. K.; Quinn, A. M.; Hunter, T. The Protein Kinase Family: Conserved Features and Deduced Phylogeny of the Catalytic Domains. *Science* **1988**, *241* (4861), 42–52. https://doi.org/10.1126/science.3291115.
(241) Knighton, D. R.; Zheng, J. H.; Eyck, L. T.; Ashford, V. A.; Xuong, N. H.; Taylor, S. S.; Sowadski, J. M. Crystal Structure of the Catalytic Subunit of Cyclic Adenosine Monophosphate-Dependent Protein Kinase. *Science* **1991**, *253* (5018), 407–414. https://doi.org/10.1126/science.1862342.
(242) Taylor, S. S.; Knighton, D. R.; Zheng, J.; Ten Eyck, L. F.; Sowadski, J. M. Structural Framework for the Protein Kinase Family. *Annu. Rev. Cell Biol.* **1992**, *8* (1), 429–462. https://doi.org/10.1146/annurev.cb.08.110192.002241.
(243) Hanks, S. K.; Hunter, T. The Eukaryotic Protein Kinase Superfamily: Kinase (Catalytic) Domain Structure and Classification1. *FASEB J.* **1995**, *9* (8), 576–596. https://doi.org/10.1096/fasebj.9.8.7768349.
(244) Romano, V.; de Beer, T. A. P.; Schwede, T. A Computational Protocol to Evaluate the Effects of Protein Mutants in the Kinase Gatekeeper Position on the Binding of ATP Substrate Analogues. *BMC Res. Notes* **2017**, *10* (1), 104. https://doi.org/10.1186/s13104-017-2428-9.

(245) Hemmer, W.; McGlone, M.; Tsigelny, I.; Taylor, S. S. Role of the Glycine Triad in the ATP-Binding Site of CAMP-Dependent Protein Kinase*. *J. Biol. Chem.* **1997**, *272* (27), 16946–16954. https://doi.org/10.1074/jbc.272.27.16946.
(246) Kornev, A. P.; Haste, N. M.; Taylor, S. S.; Eyck, L. F. T. Surface Comparison of Active and Inactive Protein Kinases Identifies a Conserved Activation Mechanism. *Proc. Natl. Acad. Sci.* **2006**, *103* (47), 17783–17788. https://doi.org/10.1073/pnas.0607656103.
(247) Kornev, A. P.; Taylor, S. S.; Eyck, L. F. T. A Helix Scaffold for the Assembly of Active Protein Kinases. *Proc. Natl. Acad. Sci.* **2008**, *105* (38), 14377–14382. https://doi.org/10.1073/pnas.0807988105.
(248) Taylor, S. S.; Kornev, A. P. Protein Kinases: Evolution of Dynamic Regulatory Proteins. *Trends Biochem. Sci.* **2011**, *36* (2), 65–77. https://doi.org/10.1016/j.tibs.2010.09.006.
(249) Adams, J. A. Activation Loop Phosphorylation and Catalysis in Protein Kinases: Is There Functional Evidence for the Autoinhibitor Model? *Biochemistry* **2003**, *42* (3), 601–607. https://doi.org/10.1021/bi020617o.
(250) Krause, D. S.; Van Etten, R. A. Tyrosine Kinases as Targets for Cancer Therapy. *N. Engl. J. Med.* **2005**, *353* (2), 172–187. https://doi.org/10.1056/NEJMra044389.
(251) Gharwan, H.; Groninger, H. Kinase Inhibitors and Monoclonal Antibodies in Oncology: Clinical Implications. *Nat. Rev. Clin. Oncol.* **2016**, *13* (4), 209–227. https://doi.org/10.1038/nrclinonc.2015.213.
(252) Baldo, B. A.; Pham, N. H. Adverse Reactions to Targeted and Non-Targeted Chemotherapeutic Drugs with Emphasis on Hypersensitivity Responses and the Invasive Metastatic Switch. *Cancer Metastasis Rev.* **2013**, *32* (3), 723–761. https://doi.org/10.1007/s10555-013-9447-3.
(253) Hartmann, J. T.; Haap, M.; Kopp, H.-G.; Lipp, H.-P. Tyrosine Kinase Inhibitors - A Review on Pharmacology, Metabolism and Side Effects. *Curr. Drug Metab.* **2009**, *10* (5), 470–481. https://doi.org/10.2174/138920009788897975.
(254) Roskoski, R. Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes. *Pharmacol. Res.* **2016**, *103*, 26–48. https://doi.org/10.1016/j.phrs.2015.10.021.
(255) Dar, A. C.; Shokat, K. M. The Evolution of Protein Kinase Inhibitors from Antagonists to Agonists of Cellular Signaling. *Annu. Rev. Biochem.* **2011**, *80* (1), 769–795. https://doi.org/10.1146/annurev-biochem-090308-173656.
(256) Zuccotto, F.; Ardini, E.; Casale, E.; Angiolini, M. Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation. *J. Med. Chem.* **2010**, *53* (7), 2681–2694. https://doi.org/10.1021/jm901443h.
(257) Caunt, C. J.; Sale, M. J.; Smith, P. D.; Cook, S. J. MEK1 and MEK2 Inhibitors and Cancer Therapy: The Long and Winding Road. *Nat. Rev. Cancer* **2015**, *15* (10), 577–592. https://doi.org/10.1038/nrc4000.
(258) Krim Gavrin, L.; Saiah, E. Approaches to Discover Non-ATP Site Kinase Inhibitors. *MedChemComm* **2013**, *4* (1), 41–51. https://doi.org/10.1039/C2MD20180A.
(259) Vandana, L.; Indraneel, G. New Directions in Targeting Protein Kinases: Focusing Upon True Allosteric and Bivalent Inhibitors. *Curr. Pharm. Des.* **2012**, *18* (20), 2936–2945.
(260) Tamaoki, T.; Nomoto, H.; Takahashi, I.; Kato, Y.; Morimoto, M.; Tomita, F. Staurosporine, a Potent Inhibitor of PhospholipidCa++dependent Protein Kinase. *Biochem. Biophys. Res. Commun.* **1986**, *135* (2), 397–402. https://doi.org/10.1016/0006-291X(86)90008-2.
(261) Fabbro, D. 25 Years of Small Molecular Weight Kinase Inhibitors: Potentials and Limitations. *Mol. Pharmacol.* **2015**, *87* (5), 766–775. https://doi.org/10.1124/mol.114.095489.
(262) Gorre, M. E.; Mohammed, M.; Ellwood, K.; Hsu, N.; Paquette, R.; Rao, P. N.; Sawyers, C. L. Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification. *Science* **2001**, *293* (5531), 876–880. https://doi.org/10.1126/science.1062538.
(263) Attwood, M. M.; Fabbro, D.; Sokolov, A. V.; Knapp, S.; Schiöth, H. B. Trends in Kinase Drug Discovery: Targets, Indications and Inhibitor Design. *Nat. Rev. Drug Discov.* **2021**. https://doi.org/10.1038/s41573-021-00252-y.

(264) Melnikova, I.; Golden, J. Targeting Protein Kinases. *Nat. Rev. Drug Discov.* **2004**, *3* (12), 993–994. https://doi.org/10.1038/nrd1600.
(265) Eid, S.; Turk, S.; Volkamer, A.; Rippmann, F.; Fulle, S. KinMap: A Web-Based Tool for Interactive Navigation through Human Kinome Data. *BMC Bioinformatics* **2017**, *18* (1), 16. https://doi.org/10.1186/s12859-016-1433-7.
(266) Wager, T. T.; Hou, X.; Verhoest, P. R.; Villalobos, A. Moving beyond Rules: The Development of a Central Nervous System Multiparameter Optimization (CNS MPO) Approach to Enable Alignment of Druglike Properties. *ACS Chem. Neurosci.* **2010**, *1* (6), 435–449. https://doi.org/10.1021/cn100008c.
(267) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAPRetrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry. *J. Chem. Inf. Comput. Sci.* **1998**, *38* (3), 511–522. https://doi.org/10.1021/ci970429i.
(268) Degen, J.; Wegscheid-Gerlach, C.; Zaliani, A.; Rarey, M. On the Art of Compiling and Using "Drug-Like" Chemical Fragment Spaces. *ChemMedChem* **2008**, *3* (10), 1503–1507. https://doi.org/10.1002/cmdc.200800178.
(269) Schuffenhauer, A.; Ertl, P.; Roggo, S.; Wetzel, S.; Koch, M. A.; Waldmann, H. The Scaffold Tree − Visualization of the Scaffold Universe by Hierarchical Scaffold Classification. *J. Chem. Inf. Model.* **2007**, *47* (1), 47–58. https://doi.org/10.1021/ci600338x.
(270) What is a Container? | App Containerization | Docker https://www.docker.com/resources/what-container (accessed 2021 -05 -20).
(271) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610. https://doi.org/10.1038/nature25978.
(272) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5* (6), 970–981. https://doi.org/10.1021/acscentsci.9b00055.
(273) Baylon, J. L.; Cilfone, N. A.; Gulcher, J. R.; Chittenden, T. W. Enhancing Retrosynthetic Reaction Prediction with Deep Learning Using Multiscale Reaction Classification. *J. Chem. Inf. Model.* **2019**, *59* (2), 673–688. https://doi.org/10.1021/acs.jcim.8b00801.
(274) Canault, B. Développement d'une Plateforme de Prédiction in Silico Des Propriétés ADME-Tox. These de doctorat, Orléans, 2018.
(275) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A New Tool for the Pharmacokinetic Optimization of Lead Compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29–S39. https://doi.org/10.1016/S0928-0987(00)00162-7.
(276) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (2), 493–500. https://doi.org/10.1021/ci025584y.
(277) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Cheminformatics* **2013**, *5* (1), 24. https://doi.org/10.1186/1758-2946-5-24.
(278) Carr, R. A. E.; Congreve, M.; Murray, C. W.; Rees, D. C. Fragment-Based Lead Discovery: Leads by Design. *Drug Discov. Today* **2005**, *10* (14), 987–992. https://doi.org/10.1016/S1359-6446(05)03511-7.
(279) Teague, S. J.; Davis, A. M.; Leeson, P. D.; Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angew. Chem. Int. Ed.* **1999**, *38* (24), 3743–3748. https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U.
(280) Lipinski, C. A. Drug-like Properties and the Causes of Poor Solubility and Poor Permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44* (1), 235–249. https://doi.org/10.1016/S1056-8719(00)00107-6.
(281) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51* (4), 817–834. https://doi.org/10.1021/jm701122q.
(282) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager,

T.; Whiteley, L.; Zhang, Y. Physiochemical Drug Properties Associated with in Vivo Toxicological Outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18* (17), 4872–4875. https://doi.org/10.1016/j.bmcl.2008.07.071.

(283) Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. *J. Cheminformatics* **2013**, *5* (1), 3. https://doi.org/10.1186/1758-2946-5-3.

(284) Écriture de votre première application Django, 2ème partie | Documentation de Django | Django https://docs.djangoproject.com/fr/2.2/intro/tutorial02/ (accessed 2021 -06 -09).

(285) Gobbi, A.; Poppinger, D. Genetic Optimization of Combinatorial Libraries. *Biotechnol. Bioeng.* **1998**, *61* (1), 47–54. https://doi.org/10.1002/(SICI)1097-0290(199824)61:1<47::AID-BIT9>3.0.CO;2-Z.

(286) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27* (2), 82–85. https://doi.org/10.1021/ci00054a008.

(287) Rose, A. S.; Hildebrand, P. W. NGL Viewer: A Web Application for Molecular Visualization. *Nucleic Acids Res.* **2015**, *43* (W1), W576–W579. https://doi.org/10.1093/nar/gkv402.

(288) Rose, A. S.; Bradley, A. R.; Valasatava, Y.; Duarte, J. M.; Prlić, A.; Rose, P. W. NGL Viewer: Web-Based Molecular Graphics for Large Complexes. *Bioinformatics* **2018**, *34* (21), 3755–3758. https://doi.org/10.1093/bioinformatics/bty419.

(289) Tanaka, H.; Yoshimura, Y.; Nishina, Y.; Nozaki, M.; Nojima, H.; Nishimune, Y. Isolation and Characterization of CDNA Clones Specifically Expressed in Testicular Germ Cells. *FEBS Lett.* **1994**, *355* (1), 4–10. https://doi.org/10.1016/0014-5793(94)01155-9.

(290) Amoussou, N. G.; Bigot, A.; Roussakis, C.; Robert, J.-M. H. Haspin: A Promising Target for the Design of Inhibitors as Potent Anticancer Drugs. *Drug Discov. Today* **2018**, *23* (2), 409–415. https://doi.org/10.1016/j.drudis.2017.10.005.

(291) Dai, J.; Sultan, S.; Taylor, S. S.; Higgins, J. M. G. The Kinase Haspin Is Required for Mitotic Histone H3 Thr 3 Phosphorylation and Normal Metaphase Chromosome Alignment. *Genes Dev.* **2005**, *19* (4), 472–488. https://doi.org/10.1101/gad.1267105.

(292) Higgins, J. M. The Haspin Gene: Location in an Intron of the Integrin AlphaE Gene, Associated Transcription of an Integrin AlphaE-Derived RNA and Expression in Diploid as Well as Haploid Cells. *Gene* **2001**, *267* (1), 55–69. https://doi.org/10.1016/s0378-1119(01)00387-0.

(293) Dave, S. S.; Fu, K.; Wright, G. W.; Lam, L. T.; Kluin, P.; Boerma, E.-J.; Greiner, T. C.; Weisenburger, D. D.; Rosenwald, A.; Ott, G.; Müller-Hermelink, H.-K.; Gascoyne, R. D.; Delabie, J.; Rimsza, L. M.; Braziel, R. M.; Grogan, T. M.; Campo, E.; Jaffe, E. S.; Dave, B. J.; Sanger, W.; Bast, M.; Vose, J. M.; Armitage, J. O.; Connors, J. M.; Smeland, E. B.; Kvaloy, S.; Holte, H.; Fisher, R. I.; Miller, T. P.; Montserrat, E.; Wilson, W. H.; Bahl, M.; Zhao, H.; Yang, L.; Powell, J.; Simon, R.; Chan, W. C.; Staudt, L. M. Molecular Diagnosis of Burkitt's Lymphoma. *N. Engl. J. Med.* **2006**, *354* (23), 2431–2442. https://doi.org/10.1056/NEJMoa055759.

(294) Rosenwald, A.; Alizadeh, A. A.; Widhopf, G.; Simon, R.; Davis, R. E.; Yu, X.; Yang, L.; Pickeral, O. K.; Rassenti, L. Z.; Powell, J.; Botstein, D.; Byrd, J. C.; Grever, M. R.; Cheson, B. D.; Chiorazzi, N.; Wilson, W. H.; Kipps, T. J.; Brown, P. O.; Staudt, L. M. Relation of Gene Expression Phenotype to Immunoglobulin Mutation Genotype in B Cell Chronic Lymphocytic Leukemia. *J. Exp. Med.* **2001**, *194* (11), 1639–1648. https://doi.org/10.1084/jem.194.11.1639.

(295) Han, X.; Kuang, T.; Ren, Y.; Lu, Z.; Liao, Q.; Chen, W. Haspin Knockdown Can Inhibit Progression and Development of Pancreatic Cancer in Vitro and Vivo. *Exp. Cell Res.* **2019**, *385* (1), 111605. https://doi.org/10.1016/j.yexcr.2019.111605.

(296) Tanaka, H.; Yoshimura, Y.; Nozaki, M.; Yomogida, K.; Tsuchida, J.; Tosaka, Y.; Habu, T.; Nakanishi, T.; Okada, M.; Nojima, H.; Nishimune, Y. Identification and Characterization of a Haploid Germ Cell-Specific Nuclear ProteinKinase (Haspin) in Spermatid Nuclei and Its Effects on Somatic Cells*. *J. Biol. Chem.* **1999**, *274* (24), 17049–17057. https://doi.org/10.1074/jbc.274.24.17049.

(297) Higgins, J. M. G. Haspin-like Proteins: A New Family of Evolutionarily Conserved Putative Eukaryotic Protein Kinases. *Protein Sci.* **2001**, *10* (8), 1677–1684. https://doi.org/10.1110/ps.49901.

(298) Elie, J.; Feizbakhsh, O.; Desban, N.; Josselin, B.; Baratte, B.; Bescond, A.; Duez, J.; Fant, X.; Bach, S.; Marie, D.; Place, M.; Ben Salah, S.; Chartier, A.; Berteina-Raboin, S.; Chaikuad, A.; Knapp, S.; Carles, F.; Bonnet, P.; Buron, F.; Routier, S.; Ruchaud, S. Design of New Disubstituted Imidazo[1,2-b]Pyridazine Derivatives as Selective Haspin Inhibitors. Synthesis, Binding Mode and Anticancer Biological Evaluation. *J. Enzyme Inhib. Med. Chem.* **2020**, *35* (1), 1840–1853. https://doi.org/10.1080/14756366.2020.1825408.

(299) Villa, F.; Capasso, P.; Tortorici, M.; Forneris, F.; Marco, A. de; Mattevi, A.; Musacchio, A. Crystal Structure of the Catalytic Domain of Haspin, an Atypical Kinase Implicated in Chromatin Organization. *Proc. Natl. Acad. Sci.* **2009**, *106* (48), 20204–20209. https://doi.org/10.1073/pnas.0908485106.

(300) Black, D. L. Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annu. Rev. Biochem.* **2003**, *72* (1), 291–336. https://doi.org/10.1146/annurev.biochem.72.121801.161720.

(301) Lindberg, M. F.; Meijer, L. Dual-Specificity, Tyrosine Phosphorylation-Regulated Kinases (DYRKs) and Cdc2-Like Kinases (CLKs) in Human Disease, an Overview. *Int. J. Mol. Sci.* **2021**, *22* (11), 6047. https://doi.org/10.3390/ijms22116047.

(302) Kumar, K.; Suebsuwong, C.; Wang, P.; Garcia-Ocana, A.; Stewart, A. F.; DeVita, R. J. DYRK1A Inhibitors as Potential Therapeutics for β-Cell Regeneration for Diabetes. *J. Med. Chem.* **2021**, *64* (6), 2901–2922. https://doi.org/10.1021/acs.jmedchem.0c02050.

(303) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminformatics* **2009**, *1* (1), 8. https://doi.org/10.1186/1758-2946-1-8.

(304) Soundararajan, M.; Roos, A. K.; Savitsky, P.; Filippakopoulos, P.; Kettenbach, A. N.; Olsen, J. V.; Gerber, S. A.; Eswaran, J.; Knapp, S.; Elkins, J. M. Structures of Down Syndrome Kinases, DYRKs, Reveal Mechanisms of Kinase Activation and Substrate Recognition. *Structure* **2013**, *21* (6), 986–996. https://doi.org/10.1016/j.str.2013.03.012.

(305) Becker, W.; Joost, H.-G. Structural and Functional Characteristics of Dyrk, a Novel Subfamily of Protein Kinases with Dual Specificity. In *Progress in Nucleic Acid Research and Molecular Biology*; Moldave, K., Ed.; Academic Press, 1998; Vol. 62, pp 1–17. https://doi.org/10.1016/S0079-6603(08)60503-6.

(306) Yoshida, S.; Yoshida, K. Multiple Functions of DYRK2 in Cancer and Tissue Development. *FEBS Lett.* **2019**, *593* (21), 2953–2965. https://doi.org/10.1002/1873-3468.13601.

(307) Tandon, V.; de la Vega, L.; Banerjee, S. Emerging Roles of DYRK2 in Cancer. *J. Biol. Chem.* **2021**, *296*, 100233. https://doi.org/10.1074/jbc.REV120.015217.

(308) Boni, J.; Rubio-Perez, C.; López-Bigas, N.; Fillat, C.; de la Luna, S. The DYRK Family of Kinases in Cancer: Molecular Functions and Therapeutic Opportunities. *Cancers* **2020**, *12* (8), 2106. https://doi.org/10.3390/cancers12082106.

(309) Chaikuad, A.; Diharce, J.; Schröder, M.; Foucourt, A.; Leblond, B.; Casagrande, A.-S.; Désiré, L.; Bonnet, P.; Knapp, S.; Besson, T. An Unusual Binding Model of the Methyl 9-Anilinothiazolo[5,4-f] Quinazoline-2-Carbimidates (EHT 1610 and EHT 5372) Confers High Selectivity for Dual-Specificity Tyrosine Phosphorylation-Regulated Kinases. *J. Med. Chem.* **2016**, *59* (22), 10315–10321. https://doi.org/10.1021/acs.jmedchem.6b01083.

(310) Ruiz-Carmona, S.; Alvarez-Garcia, D.; Foloppe, N.; Garmendia-Doval, A. B.; Juhos, S.; Schmidtke, P.; Barril, X.; Hubbard, R. E.; Morley, S. D. RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLOS Comput. Biol.* **2014**, *10* (4), e1003571. https://doi.org/10.1371/journal.pcbi.1003571.

(311) Cuypers, H. T.; Selten, G.; Quint, W.; Zijlstra, M.; Maandag, E. R.; Boelens, W.; van Wezenbeek, P.; Melief, C.; Berns, A. Murine Leukemia Virus-Induced T-Cell Lymphomagenesis: Integration of Proviruses in a Distinct Chromosomal Region. *Cell* **1984**, *37* (1), 141–150. https://doi.org/10.1016/0092-8674(84)90309-X.

(312) Mochizuki, T.; Kitanaka, C.; Noguchi, K.; Muramatsu, T.; Asai, A.; Kuchino, Y. Physical and Functional Interactions between Pim-1 Kinase and Cdc25A Phosphatase: IMPLICATIONS FOR THE Pim-1-MEDIATED ACTIVATION OF THE c-Myc SIGNALING PATHWAY*. *J. Biol. Chem.* **1999**, *274* (26), 18659–18666. https://doi.org/10.1074/jbc.274.26.18659.

(313) Bachmann, M.; Kosan, C.; Xing, P. X.; Montenarh, M.; Hoffmann, I.; Möröy, T. The Oncogenic Serine/Threonine Kinase Pim-1 Directly Phosphorylates and Activates the G2/M Specific Phosphatase Cdc25C. *Int. J. Biochem. Cell Biol.* **2006**, *38* (3), 430–443. https://doi.org/10.1016/j.biocel.2005.10.010.

(314) Tursynbay, Y.; Zhang, J.; Li, Z.; Tokay, T.; Zhumadilov, Z.; Wu, D.; Xie, Y. Pim-1 Kinase as Cancer Drug Target: An Update (Review). *Biomed. Rep.* **2016**, *4* (2), 140–146. https://doi.org/10.3892/br.2015.561.

(315) Amson, R.; Sigaux, F.; Przedborski, S.; Flandrin, G.; Givol, D.; Telerman, A. The Human Protooncogene Product P33pim Is Expressed during Fetal Hematopoiesis and in Diverse Leukemias. *Proc. Natl. Acad. Sci.* **1989**, *86* (22), 8857–8861. https://doi.org/10.1073/pnas.86.22.8857.

(316) Hsi, E. D.; Jung, S.-H.; Lai, R.; Johnson, J. L.; Cook, J. R.; Jones, D.; Devos, S.; Cheson, B. D.; Damon, L. E.; Said, J. Ki67 and PIM1 Expression Predict Outcome in Mantle Cell Lymphoma Treated with High Dose Therapy, Stem Cell Transplantation and Rituximab: A Cancer and Leukemia Group B 59909 Correlative Science Study. *Leuk. Lymphoma* **2008**, *49* (11), 2081–2090. https://doi.org/10.1080/10428190802419640.

(317) Guo, S.; Mao, X.; Chen, J.; Huang, B.; Jin, C.; Xu, Z.; Qiu, S. Overexpression of Pim-1 in Bladder Cancer. *J. Exp. Clin. Cancer Res.* **2010**, *29* (1), 161. https://doi.org/10.1186/1756-9966-29-161.

(318) Dhanasekaran, S. M.; Barrette, T. R.; Ghosh, D.; Shah, R.; Varambally, S.; Kurachi, K.; Pienta, K. J.; Rubin, M. A.; Chinnaiyan, A. M. Delineation of Prognostic Biomarkers in Prostate Cancer. *Nature* **2001**, *412* (6849), 822–826. https://doi.org/10.1038/35090585.

(319) Brasó-Maristany, F.; Filosto, S.; Catchpole, S.; Marlow, R.; Quist, J.; Francesch-Domenech, E.; Plumb, D. A.; Zakka, L.; Gazinska, P.; Liccardi, G.; Meier, P.; Gris-Oliver, A.; Cheang, M. C. U.; Perdrix-Rosell, A.; Shafat, M.; Noël, E.; Patel, N.; McEachern, K.; Scaltriti, M.; Castel, P.; Noor, F.; Buus, R.; Mathew, S.; Watkins, J.; Serra, V.; Marra, P.; Grigoriadis, A.; Tutt, A. N. PIM1 Kinase Regulates Cell Death, Tumor Growth and Chemotherapy Response in Triple-Negative Breast Cancer. *Nat. Med.* **2016**, *22* (11), 1303–1313. https://doi.org/10.1038/nm.4198.

(320) Yan, B.; Yau, E. X.; Samanta, S.; Ong, C. W.; Yong, K. J.; Ng, L. K.; Bhattacharya, B.; Lim, K. H.; Soong, R.; Yeoh, K. G.; Deng, N.; Tan, P.; Lam, Y.; Salto-Tellez, M.; Singapore Gastric Cancer Consortium. Clinical and Therapeutic Relevance of PIM1 Kinase in Gastric Cancer. *Gastric Cancer* **2012**, *15* (2), 188–197. https://doi.org/10.1007/s10120-011-0097-2.

(321) Weirauch, U.; Beckmann, N.; Thomas, M.; Grünweller, A.; Huber, K.; Bracher, F.; Hartmann, R. K.; Aigner, A. Functional Role and Therapeutic Potential of the Pim-1 Kinase in Colon Carcinoma. *Neoplasia N. Y. N* **2013**, *15* (7), 783–794.

(322) Mikkers, H.; Nawijn, M.; Allen, J.; Brouwers, C.; Verhoeven, E.; Jonkers, J.; Berns, A. Mice Deficient for All PIM Kinases Display Reduced Body Size and Impaired Responses to Hematopoietic Growth Factors. *Mol. Cell. Biol.* **2004**, *24* (13), 6104–6115. https://doi.org/10.1128/MCB.24.13.6104-6115.2004.

(323) Isaac, M.; Siu, A.; Jongstra, J. The Oncogenic PIM Kinase Family Regulates Drug Resistance through Multiple Mechanisms. *Drug Resist. Updat.* **2011**, *14* (4), 203–211. https://doi.org/10.1016/j.drup.2011.04.002.

(324) Kumar, A.; Mandiyan, V.; Suzuki, Y.; Zhang, C.; Rice, J.; Tsai, J.; Artis, D. R.; Ibrahim, P.; Bremer, R. Crystal Structures of Proto-Oncogene Kinase Pim1: A Target of Aberrant Somatic Hypermutations in Diffuse Large Cell Lymphoma. *J. Mol. Biol.* **2005**, *348* (1), 183–193. https://doi.org/10.1016/j.jmb.2005.02.039.

(325) Batra, V.; Maris, J. M.; Kang, M. H.; Reynolds, C. P.; Houghton, P. J.; Alexander, D.; Kolb, E. A.; Gorlick, R.; Keir, S. T.; Carol, H.; Lock, R.; Billups, C. A.; Smith, M. A. Initial Testing (Stage 1) of SGI-

1776, a PIM1 Kinase Inhibitor, by the Pediatric Preclinical Testing Program. *Pediatr. Blood Cancer* **2012**, *59* (4), 749–752. https://doi.org/10.1002/pbc.23364.

(326) Lebedinsky, C.; Anthony, S. P.; Mohi, G.; Yang, H.; Mei, J.; Braendle, E. A Phase 1 Study of TP-3654, an Orally-Delivered PIM Kinase Inhibitor, in Patients with Intermediate-2 or High-Risk Primary or Secondary Myelofibrosis. *Blood* **2020**, *136*, 3–4. https://doi.org/10.1182/blood-2020-134039.

(327) Iida, S.; Sunami, K.; Minami, H.; Hatake, K.; Sekiguchi, R.; Natsume, K.; Ishikawa, N.; Rinne, M.; Taniwaki, M. A Phase I, Dose-Escalation Study of Oral PIM447 in Japanese Patients with Relapsed and/or Refractory Multiple Myeloma. *Int. J. Hematol.* **2021**, *113* (6), 797–806. https://doi.org/10.1007/s12185-021-03096-9.

(328) Cortes, J.; Tamura, K.; DeAngelo, D. J.; de Bono, J.; Lorente, D.; Minden, M.; Uy, G. L.; Kantarjian, H.; Chen, L. S.; Gandhi, V.; Godin, R.; Keating, K.; McEachern, K.; Vishwanathan, K.; Pease, J. E.; Dean, E. Phase I Studies of AZD1208, a Proviral Integration Moloney Virus Kinase Inhibitor in Solid and Haematological Cancers. *Br. J. Cancer* **2018**, *118* (11), 1425–1433. https://doi.org/10.1038/s41416-018-0082-1.

(329) Koblish, H.; Li, Y.; Shin, N.; Hall, L.; Wang, Q.; Wang, K.; Covington, M.; Marando, C.; Bowman, K.; Boer, J.; Burke, K.; Wynn, R.; Margulis, A.; Reuther, G. W.; Lambert, Q. T.; Roman, V. D.; Zhang, K.; Feng, H.; Xue, C.-B.; Diamond, S.; Hollis, G.; Yeleswaram, S.; Yao, W.; Huber, R.; Vaddi, K.; Scherle, P. Preclinical Characterization of INCB053914, a Novel Pan-PIM Kinase Inhibitor, Alone and in Combination with Anticancer Agents, in Models of Hematologic Malignancies. *PLOS ONE* **2018**, *13* (6), e0199108. https://doi.org/10.1371/journal.pone.0199108.

(330) Solomon, S. R.; Nazha, A.; Strickland, S. A.; Walter, R. B.; Valimberti, I.; Tagliavini, A.; Mazzei, P.; Fiesoli, C.; Scartoni, S.; Bellarosa, D.; Binaschi, M.; Chrom, P.; Baldini, S.; Brzózka, K.; Capriati, A.; Pellacani, A.; Ravandi, F. First in Human Study of SEL24/MEN1703, First in Class, Orally Available Dual PIM/FLT3 Kinase Inhibitor, in Patients with Acute Myeloid Leukemia. *Blood* **2019**, *134* (Supplement_1), 3920–3920. https://doi.org/10.1182/blood-2019-125878.

(331) Burger, M. T.; Nishiguchi, G.; Han, W.; Lan, J.; Simmons, R.; Atallah, G.; Ding, Y.; Tamez, V.; Zhang, Y.; Mathur, M.; Muller, K.; Bellamacina, C.; Lindvall, M. K.; Zang, R.; Huh, K.; Feucht, P.; Zavorotinskaya, T.; Dai, Y.; Basham, S.; Chan, J.; Ginn, E.; Aycinena, A.; Holash, J.; Castillo, J.; Langowski, J. L.; Wang, Y.; Chen, M. Y.; Lambert, A.; Fritsch, C.; Kauffmann, A.; Pfister, E.; Vanasse, K. G.; Garcia, P. D. Identification of N-(4-((1R,3S,5S)-3-Amino-5-Methylcyclohexyl)Pyridin-3-Yl)-6-(2,6-Difluorophenyl)-5-Fluoropicolinamide (PIM447), a Potent and Selective Proviral Insertion Site of Moloney Murine Leukemia (PIM) 1, 2, and 3 Kinase Inhibitor in Clinical Trials for Hematological Malignancies. *J. Med. Chem.* **2015**, *58* (21), 8373–8386. https://doi.org/10.1021/acs.jmedchem.5b01275.

(332) Oyallon, B.; Brachet-Botineau, M.; Logé, C.; Robert, T.; Bach, S.; Ibrahim, S.; Raoul, W.; Croix, C.; Berthelot, P.; Guillon, J.; Pinaud, N.; Gouilleux, F.; Viaud-Massuard, M.-C.; Denevault-Sabourin, C. New Quinoxaline Derivatives as Dual Pim-1/2 Kinase Inhibitors: Design, Synthesis and Biological Evaluation. *Molecules* **2021**, *26* (4), 867. https://doi.org/10.3390/molecules26040867.

(333) Oyallon, B.; Brachet-Botineau, M.; Logé, C.; Bonnet, P.; Souab, M.; Robert, T.; Ruchaud, S.; Bach, S.; Berthelot, P.; Gouilleux, F.; Viaud-Massuard, M.-C.; Denevault-Sabourin, C. Structure-Based Design of Novel Quinoxaline-2-Carboxylic Acids and Analogues as Pim-1 Inhibitors. *Eur. J. Med. Chem.* **2018**, *154*, 101–109. https://doi.org/10.1016/j.ejmech.2018.04.056.

(334) Driggers, E. M.; Hale, S. P.; Lee, J.; Terrett, N. K. The Exploration of Macrocycles for Drug Discovery — an Underexploited Structural Class. *Nat. Rev. Drug Discov.* **2008**, *7* (7), 608–624. https://doi.org/10.1038/nrd2590.

(335) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction

with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(336) Paton, K. An Algorithm for Finding a Fundamental Set of Cycles of a Graph. *Commun. ACM* **1969**, *12* (9), 514–518. https://doi.org/10.1145/363219.363232.

(337) Nikonova, A. S.; Astsaturov, I.; Serebriiskii, I. G.; Dunbrack, R. L.; Golemis, E. A. Aurora A Kinase (AURKA) in Normal and Pathological Cell Division. *Cell. Mol. Life Sci.* **2013**, *70* (4), 661–687. https://doi.org/10.1007/s00018-012-1073-7.

(338) Pugacheva, E. N.; Jablonski, S. A.; Hartman, T. R.; Henske, E. P.; Golemis, E. A. HEF1-Dependent Aurora A Activation Induces Disassembly of the Primary Cilium. *Cell* **2007**, *129* (7), 1351–1363. https://doi.org/10.1016/j.cell.2007.04.035.

(339) Plotnikova, O. V.; Golemis, E. A. Aurora A Kinase Activity Influences Calcium Signaling in Kidney Cells. *J. Cell Biol.* **2011**, *193* (6), 1021–1032. https://doi.org/10.1083/jcb.201012061.

(340) Gong, X.; Du, J.; Parsons, S. H.; Merzoug, F. F.; Webster, Y.; Iversen, P. W.; Chio, L.-C.; Horn, R. D. V.; Lin, X.; Blosser, W.; Han, B.; Jin, S.; Yao, S.; Bian, H.; Ficklin, C.; Fan, L.; Kapoor, A.; Antonysamy, S.; Nulty, A. M. M.; Froning, K.; Manglicmot, D.; Pustilnik, A.; Weichert, K.; Wasserman, S. R.; Dowless, M.; Marugán, C.; Baquero, C.; Lallena, M. J.; Eastman, S. W.; Hui, Y.-H.; Dieter, M. Z.; Doman, T.; Chu, S.; Qian, H.-R.; Ye, X. S.; Barda, D. A.; Plowman, G. D.; Reinhard, C.; Campbell, R. M.; Henry, J. R.; Buchanan, S. G. Aurora A Kinase Inhibition Is Synthetic Lethal with Loss of the RB1 Tumor Suppressor Gene. *Cancer Discov.* **2019**, *9* (2), 248–263. https://doi.org/10.1158/2159-8290.CD-18-0469.

(341) Nigg, E. A. Polo-like Kinases: Positive Regulators of Cell Division from Start to Finish. *Curr. Opin. Cell Biol.* **1998**, *10* (6), 776–783. https://doi.org/10.1016/s0955-0674(98)80121-x.

(342) Eckerdt, F.; Yuan, J.; Strebhardt, K. Polo-like Kinases and Oncogenesis. *Oncogene* **2005**, *24* (2), 267–276. https://doi.org/10.1038/sj.onc.1208273.

(343) Yuan, J.; Hörlin, A.; Hock, B.; Stutte, H. J.; Rübsamen-Waigmann, H.; Strebhardt, K. Polo-like Kinase, a Novel Marker for Cellular Proliferation. *Am. J. Pathol.* **1997**, *150* (4), 1165–1172.

(344) Budin, G.; Yang, K. S.; Reiner, T.; Weissleder, R. Bioorthogonal Probes for Polo-Like Kinase 1 Imaging and Quantification. *Angew. Chem. Int. Ed Engl.* **2011**, *50* (40), 9378–9381. https://doi.org/10.1002/anie.201103273.

(345) Kothe, M.; Kohls, D.; Low, S.; Coli, R.; Rennie, G. R.; Feru, F.; Kuhn, C.; Ding, Y.-H. Research Article: Selectivity-Determining Residues in Plk1. *Chem. Biol. Drug Des.* **2007**, *70* (6), 540–546. https://doi.org/10.1111/j.1747-0285.2007.00594.x.

(346) Sirbu, D.; Diharce, J.; Martinić, I.; Chopin, N.; Eliseeva, S. V.; Guillaumet, G.; Petoud, S.; Bonnet, P.; Suzenet, F. An Original Class of Small Sized Molecules as Versatile Fluorescent Probes for Cellular Imaging. *Chem. Commun.* **2019**, *55* (54), 7776–7779. https://doi.org/10.1039/C9CC03765A.

(347) Arrowsmith, C. H.; Audia, J. E.; Austin, C.; Baell, J.; Bennett, J.; Blagg, J.; Bountra, C.; Brennan, P. E.; Brown, P. J.; Bunnage, M. E.; Buser-Doepner, C.; Campbell, R. M.; Carter, A. J.; Cohen, P.; Copeland, R. A.; Cravatt, B.; Dahlin, J. L.; Dhanak, D.; Edwards, A. M.; Frederiksen, M.; Frye, S. V.; Gray, N.; Grimshaw, C. E.; Hepworth, D.; Howe, T.; Huber, K. V. M.; Jin, J.; Knapp, S.; Kotz, J. D.; Kruger, R. G.; Lowe, D.; Mader, M. M.; Marsden, B.; Mueller-Fahrnow, A.; Müller, S.; O'Hagan, R. C.; Overington, J. P.; Owen, D. R.; Rosenberg, S. H.; Ross, R.; Roth, B.; Schapira, M.; Schreiber, S. L.; Shoichet, B.; Sundström, M.; Superti-Furga, G.; Taunton, J.; Toledo-Sherman, L.; Walpole, C.; Walters, M. A.; Willson, T. M.; Workman, P.; Young, R. N.; Zuercher, W. J. The Promise and Peril of Chemical Probes. *Nat. Chem. Biol.* **2015**, *11* (8), 536–541. https://doi.org/10.1038/nchembio.1867.

Gautier PEYRAT

Conception d'inhibiteurs de protéines kinases à partir de méthodes *in silico* basées sur les fragments

Résumé :

Les protéines kinases sont une famille de protéines ayant des rôles clés dans la régulation de fonctions cellulaires. Ce sont des cibles thérapeutiques majeures étudiées principalement en oncologie et plus récemment dans d'autres pathologies. Cette famille de protéines est composée de plus de 500 membres qui présentent une grande similarité tridimensionnelle entre eux, ce qui complique leur inhibition spécifique. Les macrocycles sont une classe de molécules qui par leur rigidité peuvent apporter cette sélectivité. Frags2Drugs est un outil *in silico* qui a été récemment développé afin de concevoir des inhibiteurs de protéines kinases à partir de fragments moléculaires, composés organiques de faible poids moléculaire. Au cours de cette thèse, cet outil a été analysé et amélioré pour l'appliquer dans quatre projets de recherche d'inhibiteurs de protéines kinases. Ainsi, en collaboration avec des équipes de chimistes et biologistes, de nouveaux inhibiteurs et macrocycles ont été conçus sur des cibles impliquées en oncologie. Pour cela, Frags2Drugs a été combiné à d'autres méthodes pour la sélection des meilleurs inhibiteurs. Plusieurs molécules et macrocycles obtenus sont en cours de synthèse et seront ensuite évalués biologiquement. Un site internet (http://frags2drugs.icoa.fr) a été développé pour rendre accessible l'outil Frags2Drugs à la communauté scientifique. Les travaux de thèse ont aussi consisté à mettre en place deux autres outils, MolDesc (http://moldesc.icoa.fr) pour le calcul de descripteurs moléculaires et KinoMine (http://kinomine.icoa.fr) pour l'exploration du kinome humain.

Mots clés : Chémoinformatique, Chimie-informatique, Conception de médicament, Fragment, Macrocycle, Protéine kinase

Design of protein kinase inhibitors from *in silico* fragment-based methods

Summary:

Protein kinases are a family of proteins having key roles in cellular functions. They represent a major therapeutic target class, studied mainly in oncology and more recently in other pathologies. There are more than 500 members in the protein kinase family sharing a high similarity in their three-dimensional structures. Due to this high degree of similarity, targeting protein kinases by selective and potent kinase inhibitors is still very challenging. Due to their rigidity, macrocycles are a class of cyclic molecules, which could provide this selectivity. Frags2Drugs is a recently developed fragment-based *in silico* tool able to design new inhibitors from fragments, which are low molecular weight compounds. In this work, Frags2Drugs was improved and applied on four drug design projects targeting protein kinases. Thus, in collaboration with biologists and chemists, new inhibitors and macrocycles were discovered on several targets involved in oncology. For this purpose, Frags2Drugs was combined to other *in silico* tools for the selection of the best inhibitors. Several molecules and macrocycles obtained by this approach are currently being synthesized. A website (http://frags2drugs.icoa.fr) was developed to share Frags2Drugs with the scientific community. During this PhD, two other tools were also deployed: MolDesc (http://moldesc.icoa.fr) for the calculation of molecular descriptors and KinoMine (http://kinomine.icoa.fr) for the exploration of the human kinome.

Keywords: Chemoinformatics, Cheminformatics, Drug Design, Fragment, Macrocycle, Protein Kinase

Institut de Chimie Organique et Analytique
UMR 7311 CNRS Université d'Orléans
Pôle de chimie rue de Chartres
45067 Orléans