Conception d'un logiciel pour la recherche de nouvelles molécules bioactives

Colin Bournez

UNIVERSITÉ D'ORLÉANS

ÉCOLE DOCTORALE
SANTE, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Institut de Chimie Organique et Analytique

THÈSE présentée par :

Colin BOURNEZ

soutenue le : **04 décembre 2019**

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

Discipline / Spécialité : Chimie / Chémoinformatique

Conception d'un logiciel pour la recherche de nouvelles molécules bioactives

THÈSE dirigée par :

M. BONNET Pascal	Professeur, Université d'Orléans
M. BERNARD Philippe	Docteur, PDG, Greenpharma S.A.S

RAPPORTEURS :

Mme KELLENBERGER Esther	Professeur, Université de Strasbourg
M. MORELLI Xavier	Directeur de recherche, Université de Marseille

---

PRÉSIDENT DE JURY :

M. BESSON Thierry	Professeur, Université de Rouen

JURY :

Mme KELLENBERGER Esther	Professeur, Université de Strasbourg
M. MORELLI Xavier	Directeur de recherche, Université de Marseille
M. BESSON Thierry	Professeur, Université de Rouen
Mme ACI-SECHE Samia	Chargé de recherche, CNRS Orléans
M. BERNARD Philippe	Docteur, PDG, Greenpharma S.A.S
M. BONNET Pascal	Professeur, Université d'Orléans

« Fais le ou ne le fais pas. Il n'y a pas d'essai. »

Maître Yoda à Luke Skywalker, Star Wars, épisode V : L'Empire contre-attaque (1980)

Remerciements

De par ces mots j'achevai ce manuscrit, de par ces mots sa lecture débutera…

Mes premiers remerciements s'adressent aux membres de mon jury qui ont bien voulu lire mais surtout évaluer les 3 années de mon doctorat condensées dans ce manuscrit. Je remercie donc le Pr Esther Kellenberger, le Dr Xavier Morelli, le Pr Thierry Besson, le Dr Samia Aci-Sèche, le Dr Phillipe Bernard et enfin le Pr Pascal Bonnet. Merci à vous pour le temps et l'énergie consacrés pour relire et évaluer cet ouvrage mais aussi pour venir assister à ma soutenance à Orléans.

En outre, Mme Kellenberger et M. Morelli je vous remercie plus particulièrement d'avoir accepté d'être les rapporteurs de ce manuscrit. Il n'est pas forcément aisé de trouver le temps nécessaire à cela, surtout à cette période de l'année et j'apprécie grandement votre implication pour mon travail.

Enfin, Mme Kellenberger, au risque de me répéter, je tiens à vous remercier également pour m'avoir envoyé cette offre de thèse il y a un peu plus de 3 ans et avoir appuyé ma candidature sans quoi je n'en serais pas là.

Je tiens ensuite à remercier la Région Centre-Val de Loire et la société Greenpharma, dirigé par le Dr Philippe Bernard, pour avoir financé ma thèse et ainsi permis le bon déroulement de mes travaux. Merci aussi au Dr Quoc-Tuan Do, de Greenpharma pour le soutien ponctuel quand j'en avais besoin.

A présent, au tour des nombreux acteurs qui ont été indispensables à ma thèse :

Tout d'abord, Pascal, c'est à toi que j'adresse mes sincères remerciements. Il y a un peu plus de 3 ans tu as accepté de me faire confiance et de m'intégrer à ton équipe, merci. Merci pour la confiance et l'autonomie que tu m'as donné tout au long de ces travaux. Merci d'avoir été toujours là quand il le fallait (pas forcément physiquement mais au moins électroniquement) et d'avoir toujours été réactif à mes nombreuses interrogations et ce, malgré les nombreuses autres tâches administratives et managériales t'incombant du fait de ta position à la direction de l'ICOA.

Ensuite, Samia, un grand merci pour avoir co-encadré mes travaux et m'avoir aussi laissé cette liberté de travail. Excuse-moi pour les fameuses réunions bihebdomadaires du jeudi où je n'étais jamais tout à fait prêt et pouvait être confondu avec un touriste. J'ai toujours ressenti ta confiance en moi malgré cela et je t'en remercie. Merci pour ta grande disponibilité à mon égard et tes conseils avisés (PS : je ne doute pas que tu ferais une excellente directrice de thèse, je te souhaite bien évidemment bonne chance pour ton HDR).

Je pense ensuite évidemment à tous les autres membres de l'équipe SB&C, actuels ou passés.

Stéphane, le plus matinal de l'équipe et donc logiquement pas le plus enjoué… Plaisanterie à part, merci à toi pour ta franchise, ta rigueur et ta lutte sans faille contre les fautes d'orthographe. J'ai beaucoup apprécié les nombreuses discussions que l'on a pu avoir professionnelles ou non et je tiens à souligner ta maitrise de la langue et des calembours. Et encore merci pour ce moment merveilleux de la « currywurst » gravé dans ma mémoire lors d'un congrès à Berlin.

José, mon premier contact au sein du laboratoire. Tu m'as transmis le projet F2D avec entrain et m'a permis de bien débuter mon doctorat. Merci pour ta disponibilité et ta bonne humeur quotidienne.

Pascal K., même si je n'arrive pas toujours à suivre ce qu'il se passe dans ta tête, merci à toi. Ton aide et ton expérience des graphes a été plus que précieuse pour mes travaux et je te suis très reconnaissant pour ça. Promis, je continuerais à m'intéresser aux boules d'énergie et à la fusion froide.

Gautier, merci de faire continuer à vivre le projet et pour ton aide durant ma rédaction. Tu as parfaitement su reprendre le programme et je te souhaite beaucoup de succès avec.

Merci aux stagiaires passés dans l'équipe durant ma thèse, et particulièrement Thomas pour son apport en tant qu'informaticien et les améliorations qu'il a su amené à mes projets. Je suis sûr que les deux plats quotidiens du CNRS doivent te manquer…

Un petit mot pour les anciens dont je vais bientôt faire partie. Notamment, je remercie les anciens post-docs pour leurs conseils. Julien, merci pour ta patate évidemment. En revanche, je ne peux en dire autant de ta musique, il te reste quelques progrès à faire là-dessus… 2 ans que j'attendais ce moment, mais je peux enfin l'écrire : de collègue tu es passé à ami. J'ai hâte que l'on se retrouve autour d'une bonne bière. Merci aussi à Jade, pour sa bonne humeur et son rire communicatif.

Les anciens doctorants, je ne vous oublie pas non plus. Merci à toi Abdennour pour toutes les sensations fortes procurées à chaque montée dans ton véhicule, ton humeur et tes boutades quotidiennes. Fabrice, je te remercie pour tes nombreux conseils à propos de Python et pour tout ce que tu m'as apporté à moi et à l'équipe. Merci également à Sonia pour ta sincérité, ton professionnalisme et tes connaissances partagées sur la Tunisie et les pays du Maghreb en général. Et merci à toi Baptiste, pour ces quelques soirées mémorables, pardonne-moi pour ton citronnier…

De manière globale, merci aussi pour tous les moments autres que ceux passés au laboratoire, autour d'un verre, plusieurs verres, ou d'une assiette ou d'une table.

Au sein du laboratoire je remercie également Laurent Robin pour son aide sur divers aspects informatiques et ses nombreuses histoires sur la vie d'Olivet et merci aussi à nos équipes administratives, notamment Marie-Madeleine et Sophie.

Enfin, je remercie le Professeur Gérald Guillaumet pour sa collaboration à mon projet de thèse, notamment pour la synthèse des molécules, ainsi que Mohsine Driowya et Sophie Front-Deschamps.

J'étends ces remerciements de manière générale à tous les autres membres du laboratoire.

Et comme il n'y a pas que la science dans la vie, je tiens à remercier ces quelques personnes de mon entourage sur lesquels j'ai pu compter aussi durant ces années. Chloé, merci d'avoir été là durant mes études supérieures et d'avoir cru en mes projets. Je te souhaite bonne chance pour la fin de ton doctorat et une belle carrière scientifique. Alex, tout pareil, j'espère pouvoir compter sur toi pour faire valoir mes bon-droits si besoin se fait ressentir un jour. Merci à toute la bande de Besançon, ils se reconnaitront…

Un merci tout spécial pour toi Caroline, merci d'être là, de toujours croire en moi et de m'avoir toujours soutenu et poussé pour que je termine la rédaction de ce manuscrit. Les rôles vont désormais s'inverser et c'est moi qui vais te coller aux basques pour que tu rédiges ton mémoire à ton tour.

On dit souvent que les chiens ne font pas des chats… Pour ma part, si j'en suis là ce n'est pas un hasard total non plus et une part du mérite revient bien évidemment à mes parents. Papa, Maman, merci pour tout ce que vous avez fait pour moi, merci d'avoir toujours cru en moi et merci de votre soutien sans failles (*et non maman, ceci n'est pas mon rapport de DUT, je suis bien en doctorat désormais*). Evidemment ces remerciements ne seraient pas complet sans le reste de ma famille, mes frères et sœurs, Pernelle, Célio et Siloé sur qui je sais que je peux compter pour tout, sauf la chémoinformatique ! Pour finir, merci à la dernière arrivée, ma nièce Lia, qui a eu la bonne idée de pointer le bout de son nez à l'autre bout du monde pile pendant ma période de rédaction. Sache que tu m'as coûté un mois de rédaction, je te revaudrais ça en te lisant ce manuscrit pour t'endormir…

Et merci à mon ordinateur de m'avoir supporté pendant ces 3 années sans broncher malgré les nombreuses sollicitations de ma part.

Table des matières

Liste des abréviations

1D, 2D ou 3D : 1, 2 ou 3 dimensions

ABL1 : *Abelson tyrosine-protein kinase 1*

ACP : Analyses en composantes principales

ADME-Tox : Absorption, Distribution, Métabolisme, Excrétion et Toxicité

ADN : Acide désoxyribonucléique

ADP : Adénosine diphosphate

Afssaps : Agence française de sécurité sanitaire des produits de santé

AGC : Protéines kinases A, G et C

ALK : *Anaplastic lymphoma kinase*

AMM : Autorisation de mise sur le marché

ANSM : Agence nationale de sécurité du médicament et des produits de santé

API : Interface de programmation

aPK : Protéines kinases atypiques

ARN : Acide ribonucléique

ASP : *Astex Statistical Potential*

ATP : Adénosine triphosphate

AUC : *Area under the curve* / Aire sous la courbe

AxK : Ala – x – Lys

Bcl-2 : *B-cell lymphoma 2*

BDD : Base de données

BFS : *Breadth-First Search*

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

CAS : *Chemical Abstracts Service*

CCLE : *Cancer Cell Line Encyclopedia*

CCP : Certificat complémentaire de protection

CDK2 : *cyclin-dependent kinase 2*

ChEMBL : *Chemical database of the European Molecular Biology Laboratory*

CLogP : coefficient de partage octanol/eau calculé

CNRS : Centre national de la recherche scientifique

COX : Cyclooxygénase

Da : Dalton

DFG : Asp – Phe – Gly

DFS : *Depth-First Search*

DLL : Degrés de libertés des liaisons

DUD-E : *Database of Useful (Docking) Decoys - Enhanced*

DUT : Diplôme universitaire de technologie

DYRK : *Dual specificity tyrosine-phosphorylation-regulated kinase*

ECACC : *European Collection of Authenticated Cell Cultures*

ECFP : *Extended Connectivity Fingerprints*

EF : Facteur d'enrichissement

EGFR : *Epidermal Growth Factor Receptor*

EM : Microscopie électronique

EMA : *European Medicines Agency* / Agence européenne des médicaments

EPHB2 : *Ephrin type-B receptor 2*

ePK : Protéines kinases eucaryotes

F2D : Frags2Drugs

FBDD : *Fragment-based drug design*

mM, nM ou µM : milli, nano ou micromole

MSCS : *Multiple solvent crystal structures*

FDA : *Food and Drug Administration* / Administration américaine des denrées alimentaires et des médicaments

FGFR : *Fibroblast growth factor receptor*

FP : Nombre de faux positifs

FPR : Taux de faux positifs

Go : Gigaoctet

Gxe : Graphe d'exclusion

Gxi : Graphe d'inclusion

HBA : Nombre de donneurs de liaisons hydrogène

HBD : Nombre d'accepteurs de liaisons hydrogène

HDF5 : *Hierarchical Data Format 5*

HRD : His – Arg – Asp

HTS : Criblage à haut débit

IC_{50} : Concentration inhibitrice médiane

ICH : *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* / Conseil international d'harmonisation des exigences techniques pour l'enregistrement des médicaments à usage humain

ICOA : Institut de Chimie Organique et Analytique

InChi : *International Chemical Identifier* / Identifiant chimique international

InsR : Récepteur à l'insuline

IRON : *Innovative Radiopharmaceuticals in Oncology and Neurology*

IUPAC : *International Union of Pure and Applied Chemistry*

IUT : Institut universitaire de technologie

JAK : *Just Another Kinase*

K_d : Constante de dissociation

kg : kilogramme

KLIFS : *Kinase-Ligand Interaction Fingerprints and Structures database*

LE : Ligand efficiency

LMC : Leucémie Myéloïde Chronique

MACCS : *Molecular ACCess System*

MCSS : *Multiple Copy Simultaneous Search*

MEK : *Mitogen-activated Extracellular signal-Regulated Kinase*

MMFF94 : *Merck Molecular Force Field 94*

MM-GBSA : *Molecular mechanics energies generalized Born and surface area continuum solvation*

MOE : *Molecular operating environment*

GOLD : *Genetic Optimisation for Ligand Docking*

MS : Spectrométrie de masse

MW : Masse moléculaire

NAR : Nombre de cycles aromatiques

NCA : Nombre d'atomes chiraux

NHA : Nombre d'atomes lourds

NOSQL : *Not only SQL*

NRB : Nombre de liaisons à rotation libre

PAINS : *Pan Assay Interference Compounds*

PDB : *Protein Data Bank*

PFAM : *Protein Families Database*

PhD : *Philosophiæ doctor*

PIM1 : *Proto-oncogene serine/threonine-protein kinase Pim-1*

PKC : Protéine kinase C

PKI : *Protein kinase inhibitor* / / Inhibiteur de protéines kinases

PKIDB : *Protein kinase inhibitor Database* / Base de données d'inhibiteurs de protéines kinases

PMF : *Potential of mean force*

PMI : Principaux moments d'inertie

PRKACA : *protein kinase cAMP-activated catalytic subunit alpha*

QSAR : Relation quantitative structure-activité

R&D : Recherche et développement

RAM : *Random Access Memory*

RCPG : Récepteur couplé aux protéines G

RCSB : *Research Collaboratory for Structural Bioinformatics*

RECAP : *Retrosynthetic Combinatorial Analysis Procedure*

rIFN : Interféron recombinant

RMN : Résonance magnétique nucléaire

RMSD : *Root-Mean-Square Deviation*

ROC : *Receiver operating characteristic*

ROS : *Proto-oncogene tyrosine-protein kinase ROS*

SA_Score : *Synthetic Accessibility Score*

SB&C : Bioinformatique Structurale et Chémoinformatique

SDF : *Structure-data file*

SFCi : Société Française de Chémoinformatique

SIFt : *Structural Interaction Fingerprint*

SMARTS : *SMILES arbitrary target specification*

SMILES : *Simplified Molecular Input Line Entry Specification*

SPLIF : *Structural Protein–Ligand Interaction Fingerprints*

SPR : Résonance plasmonique de surface

SQL : *Structured Query Language* / Langage de requête structurée

SYK : *Spleen tyrosine kinase*

TK : Tyrosine kinase

TPR : Taux de vrais positifs

TPSA : Surface polaire topologique

TSA : *Thermal shift assay*

UMR : Unité mixte de recherche

UICPA : Union internationale de chimie pure et appliquée

VP : Nombre de vrais positif

Liste des figures

Liste des tables

Liste des équations

Avant-propos

Malgré le terme populaire de PhD (*Philosophiæ doctor*) pour désigner un titulaire de doctorat, ce manuscrit n'a guère à voir avec cette matière, il a été écrit afin de permettre à son auteur d'obtenir le grade de docteur en chimie. Aussi, vous trouverez tout au long de ce texte une étude visant à répondre aux différentes problématiques concernant la création de nouveaux médicaments, et plus particulièrement ce que l'on appelle l'approche par fragments.

Mon doctorat a été réalisé à l'Institut de Chimie Organique et Analytique (ICOA), un laboratoire de l'Université d'Orléans fondé en 1996. L'ICOA est une unité mixte de recherche (UMR 7311) sous les tutelles de l'Université d'Orléans et du CNRS. Il est dirigé par le professeur Pascal Bonnet depuis 2016. Il s'agit d'un laboratoire qui a pour mission la conception, la synthèse et l'analyse de nouvelles molécules bioactives avec des applications thérapeutiques et cosmétiques. L'ICOA est notamment partenaire des Laboratoires d'Excellence en réseau SynOrg et IRON. Il est divisé en cinq axes de recherches principaux :

- Chémoinformatique, modélisation.
- Glycomolécules, de la synthèse à l'enzymologie.
- Synthèse hétérocyclique et chimie thérapeutique.
- Nucléosides modifiés.
- Extraction, analyse de molécules bioactives.

Mes travaux ont été effectués au sein de l'équipe de Bioinformatique Structurale et Chémoinformatique (SB&C). Cette équipe est dirigée par le Pr Pascal Bonnet depuis 2012. Elle est à ce jour composée de trois membres permanents à temps plein et un membre permanent partiellement : le Pr Pascal Bonnet, le Dr Samia Aci-Sèche, le Dr Stéphane Bourg et le Dr Caroline West (50 %). Il accueille régulièrement des post-doctorants, des doctorants et des stagiaires de différents niveaux (http://www.icoa.fr/fr/bonnet/anciens). La spécialité de l'équipe SB&C est la conception de molécules actives par l'utilisation et le développement de méthodes numériques. La grande majorité de ces recherches se focalisent sur une même famille de protéines : la famille des protéines kinase. Dans ce but, plusieurs travaux de recherche sont en cours, de la prédiction de la cinétique des inhibiteurs de protéines kinases au développement d'une méthode de construction de molécules *in silico* à partir de fragments en passant par l'usage des méthodes d'amarrage moléculaires et l'analyse des poses obtenues. Les méthodes développées sont ensuite appliquées à d'autres familles de protéines. Depuis 2018, l'équipe SB&C a mis à disposition un serveur web regroupant les différents services développés en son sein et permettant aux chercheurs extérieurs d'y avoir accès (http://sbc.icoa.fr/).

Au cours de mon doctorat, j'ai aussi eu l'opportunité de collaborer avec l'entreprise Greenpharma, une société de biotechnologie spécialiste des substances naturelles visant à découvrir de nouveaux ingrédients pour l'industrie pharmaceutique, cosmétique ou encore agroalimentaire (https://www.greenpharma.com/). Elle a été fondée en 2000 par le Dr Philippe Bernard. Depuis 2011, Greenpharma propose aussi de la vente de molécules via la société Ambinter qu'elle a acquise, lui fournissant un catalogue d'environ 30 millions de références (http://www.ambinter.com/).

Qu'est-ce qu'un fragment? Quelles sont les spécificités des protéines kinases? Comment les outils numériques peuvent-ils nous aider dans la recherche de médicaments? C'est en partie à ces interrogations que je vais essayer de répondre dans ce manuscrit, tout en développant mes travaux personnels et les résultats obtenus. Ainsi, vous trouverez dans le premier chapitre un avant-goût de mon travail avec une introduction sur la fabrication des médicaments, les différents concepts utiles à la compréhension de ce manuscrit et une présentation de la famille des protéines sur lesquelles j'ai travaillé : les protéines kinases. Ce premier chapitre se terminera par un premier article portant sur une étude entre les inhibiteurs de protéines kinases approuvés ou en essai clinique et les inhibiteurs de protéines kinases que l'on trouve dans une base de données publique. Le deuxième chapitre relate la découverte de nouveaux produits naturels dans le domaine de la cosmétique par amarrage moléculaire. Il se termine également par une publication résumant mon travail et mes résultats. Le chapitre trois se consacre à l'état de l'art sur l'approche par fragments dans le domaine de la création de médicaments, particulièrement *in silico*. Il se conclue par une revue décrivant les différentes étapes et les différents outils informatiques disponibles pour un projet de recherche de médicaments par l'approche des fragments. Le quatrième chapitre décrit le fruit de mon travail principal, à savoir le développement d'un nouveau logiciel ayant pour but de concevoir de nouvelles molécules thérapeutiques prometteuses à partir de fragments. Enfin, le dernier chapitre se veut le bilan de mes trois années de thèse suivi d'une perspective quant au futur de ces travaux qui je l'espère ne sont que le début d'une belle histoire…

En accompagnement de ce manuscrit, je vous conseille le thé vert *Sencha*, d'origine japonaise et préférentiellement biologique, *en évitant toutefois de le renverser*, car c'est la boisson favorite de ma maman et c'est à elle que je dédie cet ouvrage.

Le Lapin Blanc mit ses lunettes. *« Par où commencerai-je, s'il plaît à Votre Majesté ? »* demanda-t-il.

« Commencez par le commencement, » dit d'un ton emprunt de gravité, le Roi…

Chapitre 1 : Introduction

1.1 Genèse d'un médicament

En France, un médicament est défini par l'article L5111-1 du Code de la santé publique : « On entend par médicament toute substance ou composition présentée comme possédant des propriétés curatives ou préventives à l'égard des maladies humaines ou animales, ainsi que toute substance ou composition pouvant être utilisée chez l'homme ou chez l'animal ou pouvant leur être administrée, en vue d'établir un diagnostic médical ou de restaurer, corriger ou modifier leurs fonctions physiologiques en exerçant une action pharmacologique, immunologique ou métabolique.

Sont notamment considérés comme des médicaments les produits diététiques qui renferment dans leur composition des substances chimiques ou biologiques ne constituant pas elles-mêmes des aliments, mais dont la présence confère à ces produits, soit des propriétés spéciales recherchées en thérapeutique diététique, soit des propriétés de repas d'épreuve.

Les produits utilisés pour la désinfection des locaux et pour la prothèse dentaire ne sont pas considérés comme des médicaments.

Lorsque, eu égard à l'ensemble de ses caractéristiques, un produit est susceptible de répondre à la fois à la définition du médicament prévue au premier alinéa et à celle d'autres catégories de produits régies par le droit communautaire ou national, il est, en cas de doute, considéré comme un médicament. »[1].

Le processus de recherche et développement d'un médicament, aussi appelé le « Drug Discovery », est constitué d'un ensemble d'étapes règlementées et dure en moyenne de 8 à 15 ans (Figure 1). Le but de ces différentes étapes est d'arriver à caractériser le plus précisément possible le profil pharmacologique du futur médicament et d'éliminer les molécules présentant des risques pour la santé. Il s'agit d'un procédé très onéreux : on estime souvent à environ un milliard d'euros la mise sur le marché d'un médicament[2,3]. En réalité, ce coût serait plutôt de l'ordre de deux milliards de nos jours, notamment en prenant en compte les molécules ayant échoué à devenir un médicament[4,5]. En effet, pour amener un nouveau médicament sur le marché, il y a peu d'élus pour beaucoup d'appelés. Classiquement, au début du projet de recherche, on peut tester plusieurs dizaines de milliers de molécules dont la grande majorité vont être écartées au fur et à mesure des différentes phases de l'avancée du processus. Ces phases se comptent généralement au nombre de quatre : recherche, développement clinique, mise sur le marché et enfin pharmacovigilance.

[1] « Code de la santé publique - Article L5111-1 », L5111-1 Code de la santé publique § (s. d.), consulté le 5 juin 2019.
[2] Joseph A DiMasi, Ronald W Hansen, et Henry G Grabowski, « The price of innovation: new estimates of drug development costs », *Journal of Health Economics* 22, nº 2 (1 mars 2003): 151-85, https://doi.org/10.1016/S0167-6296(02)00126-1.
[3] « L'innovation thérapeutique, un processus long et coûteux », consulté le 5 mars 2019, https://www.leem.org/linnovation-therapeutique-un-processus-long-et-couteux-0.
[4] Craig W. Lindsley, « New Statistics on the Cost of New Drug Development and the Trouble with CNS Drugs », *ACS Chemical Neuroscience* 5, nº 12 (17 décembre 2014): 1142-1142, https://doi.org/10.1021/cn500298z.
[5] Joseph A. DiMasi, Henry G. Grabowski, et Ronald W. Hansen, « Innovation in the pharmaceutical industry: New estimates of R&D costs », *Journal of Health Economics* 47 (1 mai 2016): 20-33, https://doi.org/10.1016/j.jhealeco.2016.01.012.

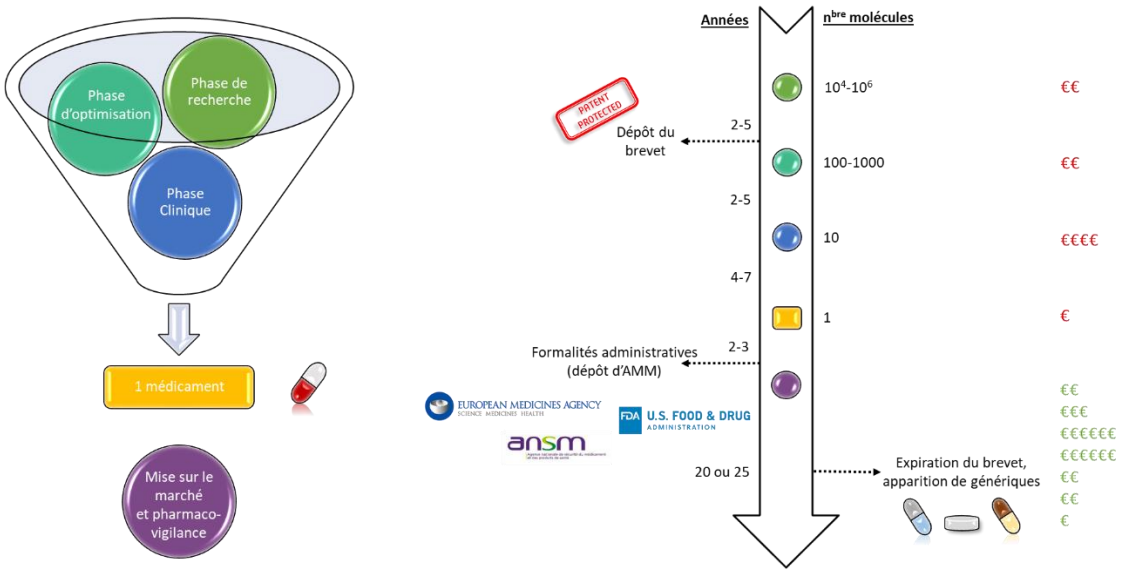


Figure 1 : Schéma récapitulatif des différentes étapes nécessaires de la recherche d'un médicament à l'apparition de génériques.

1.1.1 La recherche, le point de départ

À l'origine d'un médicament, il faut commencer par rechercher et trouver une cible potentiellement intéressante pour une maladie donnée. Pour cela, plusieurs paramètres peuvent guider le choix final des chercheurs en fonction des stratégies de la structure (publique ou privée) de recherche et/ou des besoins médicaux. Pour caractériser la cible, on peut s'appuyer sur les avancées en recherche fondamentale et les connaissances sur la maladie. Une fois cette cible choisie et caractérisée, le but est de trouver des composés qui vont influencer son mode d'action. On parle alors d'inhibiteur, ou antagoniste, si la molécule empêche la cible de réaliser son action initiale, ou au contraire d'activateur, ou agoniste, si la molécule permet d'activer ou de stimuler l'action de la cible. Par exemple, l'aspirine est un inhibiteur des cyclooxygénases (COX) 1 et 2, cette action amène à la réduction de la production de prostaglandines et de thromboxanes lui conférant ainsi ses propriétés antalgiques, antipyrétiques et anti-inflammatoires[6].

Il existe plusieurs méthodes pour trouver des composés actifs. Une des plus utilisées dans l'industrie pharmaceutique est le criblage à haut débit (HTS) d'une chimiothèque de plusieurs dizaines de milliers, voire de million(s) de molécules. Ce criblage est effectué par des robots qui vont tester l'affinité de chaque composé directement sur la cible, en regardant plus particulièrement sa capacité à se lier avec elle. Les composés sélectionnés, appelés touches (ou « hits ») vont alors être analysés plus en détail et confirmés par des tests d'activité plus poussés. Une fois les touches validées, celles-ci vont alors passer en phase d'optimisation.

1.1.2 Le développement d'un candidat

1.1.2.1 La préparation du composé principal ou lead

La phase d'optimisation commence par l'amélioration des touches issues de la phase de recherche afin d'obtenir une ou plusieurs tête(s) de série, ou « lead(s) ». Cette amélioration est

[6] J. R. Vane et R. M. Botting, « The Mechanism of Action of Aspirin », *Thrombosis Research* 110, nº 5 (15 juin 2003): 255-58, https://doi.org/10.1016/S0049-3848(03)00379-7.

nécessaire car il est rare que les molécules initialement identifiées soient d'emblées adaptées à l'organisme, que ce soit au niveau de leur absorption, de leur métabolisme ou de leur toxicité. C'est le travail des chimistes médicinaux de transformer ces touches en composés assimilables et non directement dégradés par l'organisme ou toxiques pour lui. En plus de l'optimisation de leurs propriétés ADME-Tox (Absorption, Distribution, Métabolisme, Excrétion et Toxicité), il faut aussi penser à la brevetabilité et à la future synthèse en quantité industrielle des composés sélectionnés. Ainsi, une molécule présentant des très bons résultats d'inhibition sur la cible mais que l'on ne peut breveter ou que l'on ne peut créer à l'échelle commerciale (kg ou tonnes) ne sera pas forcément intéressante pour une entreprise pharmaceutique car le retour sur investissement ne sera pas acceptable. Une fois tous ces paramètres pris en compte et les têtes de série améliorées pour obtenir un bon profil pharmacologique, on parle alors de molécules candidates, ou « lead avancé », le développement se poursuit en phase préclinique.

1.1.2.2 La phase préclinique

Il s'agit de la phase avant les essais sur l'homme. A ce stade, il reste une dizaine, voire une centaine de composés, qu'il faut donc caractériser afin de connaitre leurs effets potentiels *in vivo*. Il faut particulièrement veiller à ce que le candidat médicament ne présente aucun signe de carcinogénicité, ne soit pas un perturbateur endocrinien et n'entraine aucune modification génétique chez le patient. Pour cela, les investigations continuent *in vitro*, avec dans un premier temps, des tests sur des cellules puis sur des tissus cellulaires mimant le comportement physiologique de la cible dans son environnement. Enfin, les composés prometteurs vont être testés *in vivo* chez l'animal. Cette phase est essentielle avant de passer chez l'humain car elle va permettre de dresser un « passeport » complet de chaque composé restant et notamment les prérequis pharmacologiques nécessaires aux essais cliniques. A l'issue des tests précliniques, pour chaque molécule sélectionnée, on dispose de la toxicité aigüe ou chronique, de la dose maximale tolérée, des potentiels d'interaction avec d'autres protéines, des propriétés ADME-Tox, pharmacocinétiques, etc. Toutes ces recherches ne se font pas à l'appréciation personnelle de chacun mais sont codifiées internationalement par les directives de l'ICH (« International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use »)[7]. Sans cette base, impossible de poursuivre les expérimentations chez l'homme. Toutes les données collectées seront présentées dans le futur dossier d'autorisation de mise sur le marché (AMM). Cette phase dure en moyenne 2 à 5 ans et seulement une dizaine ou moins de molécules seront envoyées en phase clinique. Tous les résultats de la phase préclinique vont servir à élaborer la dose de départ et la marge posologique pour la suite en phase clinique. Elle suggère également les signes cliniques à rechercher afin de détecter tout effet indésirable lors des essais.

1.1.2.3 Le parcours clinique

C'est à partir de cette étape, après le feu vert des autorités compétentes, que les tests commencent à se dérouler chez l'être humain. Elle se décompose en trois phases principales appelées phase 1, phase 2 et phase 3.

1.1.2.3.1 Phase 1 : évaluation de la tolérance et posologie

La première phase des essais cliniques se déroule sur des individus sains, volontaires et bien évidemment avertis. Les tests se déroulent sur une cohorte de patients relativement faible (n = 10 – 100), dans des centres spécialisés et sous étroite surveillance. Elle dure moins d'un an durant lequel des doses croissantes du candidat médicament vont être administrées

[7] « Guidelines : ICH », consulté le 18 mars 2019, http://www.ich.org/products/guidelines.

différemment aux patients. Le but de cette phase est de connaitre notamment la dose maximale tolérée et les effets secondaires possibles afin d'ajuster la posologie ainsi que la méthode d'administration (orale ou parentérale, gélule ou comprimé, …). Environ 70 % des candidats médicaments à ce stade atteignent la phase suivante[8]. Dans certains cas, notamment en oncologie, les individus de cette phase peuvent être des individus atteints par la pathologie ciblée.

1.1.2.3.2 Phase 2 : efficacité de la molécule sur un petit groupe

Cette étape marque un tournant dans le développement puisque c'est celle qui permet l'essai du candidat médicament chez des individus présentant la pathologie pour laquelle il est développé. Elle permet donc de démontrer réellement la preuve de concept de l'étude démarrée des années auparavant et l'efficacité du candidat médicament. Cette fois-ci, la cohorte de patients malades peut atteindre plusieurs centaines de patients et dure plus d'un an. Le principal but ici est de trouver la dose optimale à administrer (meilleur rapport bénéfice/risque). Cette phase ne prend pas assez d'individus en compte pour pouvoir démontrer suffisamment l'efficacité du composé et le commercialiser directement. Cependant, elle permet aux chercheurs d'obtenir des premiers résultats et de mettre en place un protocole d'administration précis pour la suite. Environ 33 % des candidats médicaments vont atteindre la phase 3[8].

1.1.2.3.3 Phase 3 : confirmation de l'efficacité et bénéfice clinique

Aussi appelée phase « pivot », il s'agit d'un essai thérapeutique à grande échelle, avec une cohorte de plusieurs milliers de patients répartis dans différents centres géographiques. L'efficacité du traitement est comparée à un placebo (ou à un traitement de référence, notamment en oncologie). Cette phase s'étale sur plusieurs années permettant ainsi d'obtenir des informations sur les effets secondaires du candidat médicament à plus long terme. De plus, la taille de la cohorte permet aussi de connaitre les potentiels effets plus rares, qui ne vont toucher que quelques patients. Enfin, les risques d'interactions avec d'autres traitements peuvent aussi être détectés durant cette étape. Si la confirmation de l'efficacité de la molécule ainsi que son innocuité sont bien démontrées, toutes les informations regroupées depuis le début du projet de recherche sont intégrées dans le dossier d'AMM et transmises aux autorités sanitaires. Environ 30 % des candidats médicaments restants vont faire l'objet d'une demande d'AMM[8].

Toutes les informations sur les essais cliniques en cours partout dans le monde sont répertoriées sur le site suivant : https://clinicaltrials.gov/.

1.1.3 La commercialisation

Une fois toutes les étapes de recherche et développement (R&D) validées, le candidat médicament fait l'objet d'une demande d'AMM auprès des autorités compétentes qui vont autoriser ou non sa commercialisation[9]. En France, cette autorisation est délivrée par les autorités compétentes européennes (EMA, « European Medicines Agency ») ou nationales (Agence nationale de sécurité du médicament et des produits de santé, ANSM, qui a remplacé l'Afssaps en 2012). Aux Etats-Unis, il faut passer par la « Food and Drug Administration » (FDA). Sans AMM, impossible de commercialiser un médicament. Ces différentes

[8] Office of the Commissioner, « The Drug Development Process - Step 3: Clinical Research », WebContent, consulté le 20 mars 2019, https://www.fda.gov/forpatients/approvals/drugs/ucm405622.htm.
[9] « Demande initiale d'AMM - ANSM : Agence nationale de sécurité du médicament et des produits de santé », consulté le 21 mars 2019, https://www.ansm.sante.fr/Activites/Autorisations-de-Mise-sur-le-Marche-AMM/Demande-initiale-d-AMM/(offset)/1.

organisations sont indépendantes, ainsi, un traitement peut très bien être accepté en Europe et refusée aux Etats-Unis, comme par exemple le tivozanib, Fotivda de son nom commercial, utilisé pour le traitement du carcinome à cellules rénales[10,11]. Une fois l'AMM acceptée, le médicament rentre en phase 4, aussi appelée pharmacovigilance. Durant cette phase, même s'il est commercialisé, le médicament reste sous surveillance constante, pour vérifier en particulier qu'il ne présente pas d'effets indésirables graves à long terme. En cas de risque avéré pour la santé, il se peut tout à fait que le produit soit retiré du marché. On estime à environ 10 par an le nombre de médicaments ainsi désapprouvés localement ou globalement[12]. Il se peut aussi que de nouvelles indications thérapeutiques soient explorées par le fabricant pouvant déboucher sur une nouvelle demande d'AMM pour une autre application. On appelle cela le repositionnement de médicament. L'exemple le plus connu est celui du sildénafil, plus répandu sous son nom commercial Viagra, qui devait à l'origine traiter l'angine de poitrine. Durant les études cliniques l'efficacité fut finalement plus faible qu'escomptée. Cependant, un effet secondaire se fit remarquer. En effet, le sildéfanil provoquait chez les patients une érection (effet que l'on retrouvait uniquement chez les patients masculins de manière évidente). Pfizer, l'entreprise qui le développait eut alors la bonne (et très lucrative) idée d'utiliser cette molécule pour les troubles de l'érection avec le succès qu'on lui connait[13]. Il s'agit d'un bel exemple à la fois de sérendipité et de repositionnement de médicament. La pratique est assez courante, en repositionnant un médicament on bénéficie de la connaissance des études cliniques déjà existantes et on économise ainsi beaucoup de temps et d'argent[14].

Finalement, après vingt ans d'exploitation, le brevet de la molécule expire. L'exploitant a alors la possibilité de prolonger de 5 ans la protection en demandant un certificat complémentaire de protection (CCP) mais une fois ce délai passé, le médicament n'est plus protégé. Il peut alors être copié et distribué par un autre fabricant. Dans ce cas, la copie est appelée générique.

1.2 Vue d'ensemble sur les médicaments et les cibles thérapeutiques

Lorsque que l'on analyse les médicaments approuvés ou leurs cibles, les données de la FDA font figure de référence dans les différentes études publiées, je vais donc parler en me basant sur ces données. Ces 25 dernières années, la majorité des médicaments approuvés par la FDA appartiennent à la catégorie des petites molécules (Figure 2).

[10] « Fotivda », Text, European Medicines Agency, 17 septembre 2018, https://www.ema.europa.eu/en/medicines/human/EPAR/fotivda.
[11] « Tivozanib for Kidney Cancer Rejected by FDA », Medscape, consulté le 21 mars 2019, http://www.medscape.com/viewarticle/805578.
[12] Vishal B. Siramshetty et al., « WITHDRAWN—a Resource for Withdrawn and Discontinued Drugs », *Nucleic Acids Research* 44, nº Database issue (4 janvier 2016): D1080, https://doi.org/10.1093/nar/gkv1192.
[13] « Viagra: How a Little Blue Pill Changed the World », Drugs.com, consulté le 17 avril 2019, https://www.drugs.com/slideshow/viagra-little-blue-pill-1043.
[14] Sudeep Pushpakom et al., « Drug Repurposing: Progress, Challenges and Recommendations », *Nature Reviews Drug Discovery* 18, nº 1 (janvier 2019): 41-58, https://doi.org/10.1038/nrd.2018.168.

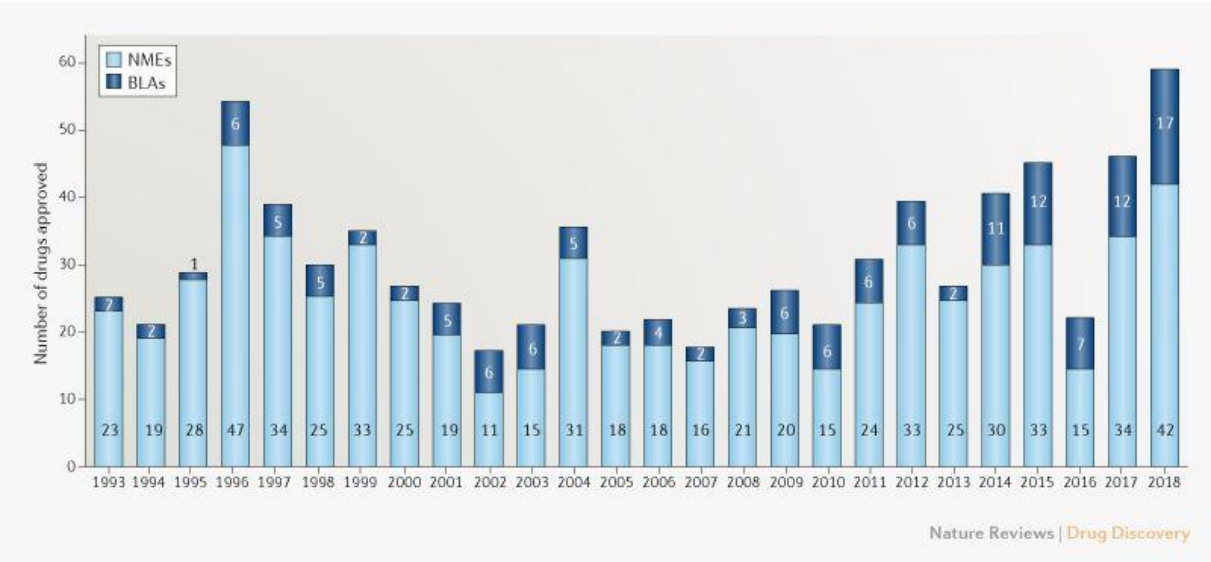


Figure 2 : Nombre de médicaments par catégorie approuvés par la FDA depuis 1993. *En bleu clair, les petites molécules, en bleu foncé, les médicaments biologiques incluant les anticorps monoclonaux. D'après A. Mullard* [15].

L'année 2018 fait office de record en termes de médicaments approuvés aussi bien pour les petites molécules que pour les produits biologiques. Cependant, comme on le voit sur la Figure 2, il est difficile d'estimer si cette tendance va se poursuivre ou si ce nombre va baisser au cours des prochaines années. Comme l'an passé, c'est dans le domaine de l'oncologie que l'on retrouve le plus de médicaments approuvés, suivi par celui des maladies infectieuses[16]. La nature reste encore une source d'inspiration majeure pour trouver des nouveaux composés puisque sur les 59 médicaments approuvés en 2018, 10 sont des dérivés de produits naturels. Enfin, 19 soit 32 % sont les premiers médicaments d'une nouvelle classe thérapeutique (« first-in-class »).

En revanche, on constate depuis peu que de nouveaux principes actifs se démocratisent notamment grâce aux investissements dans les biotechnologies. Ainsi, de plus en plus de produits biologiques sont désormais approuvés et les anticorps monoclonaux en particulier émergent comme alternative aux petites molécules. Cette tendance s'illustre aussi dans la recherche fondamentale avec le prix Nobel de Médecine 2018 attribué à deux chercheurs en immunologie, James Allison et Tasuku Honjo pour « leur découverte du traitement du cancer par inhibition de la régulation immunitaire négative », à l'aide d'anticorps[17]. Ces nouveaux principes actifs présentent de nombreux avantages par rapport aux molécules classiques. Tout d'abord ils présentent moins de limitations par rapport à la propriété intellectuelle et à la brevetabilité. De plus, ils sont aussi mieux protégés par rapport à l'apparition de génériques, que l'on appelle biosimilaire dans ce cas, de par leur nature. En effet, un médicament biologique se réalise à l'aide de cellules ou d'organismes vivants (levure, bactérie…). Une variabilité biologique peut ainsi grandement impacter les propriétés des produits et ils sont donc beaucoup plus difficiles à copier. Enfin, les produits biologiques vont présenter de meilleures performances en sélectivité car ils sont programmées pour reconnaitre et s'attaquer uniquement

[15] Asher Mullard, « 2018 FDA Drug Approvals », *Nature Reviews Drug Discovery* 18 (15 janvier 2019): 85, https://doi.org/10.1038/d41573-019-00014-x.
[16] Beatriz G. de la Torre et Fernando Albericio, « The Pharmaceutical Industry in 2018. An Analysis of FDA Drug Approvals from the Perspective of Molecules », *Molecules* 24, nº 4 (janvier 2019): 809, https://doi.org/10.3390/molecules24040809.
[17] « The Nobel Prize in Physiology or Medicine 2018 », NobelPrize.org, consulté le 5 juin 2019, https://www.nobelprize.org/prizes/medicine/2018/summary/.

à la cible désirée[18]. Toutefois, ces nouveaux produits ne sont pas exempts d'inconvénients et de nombreux problèmes restent encore un frein à cette innovation. D'abord de par leur taille (environ 150 kDa), les produits biologiques et en particulier les anticorps monoclonaux ne peuvent traverser la barrière hémato-encéphalique et sont ainsi inefficaces dans le cas de pathologies au niveau du cerveau[18]. De plus, la localisation de leur cible est restreinte à l'extérieure ou à la surface des cellules, ils ne peuvent la pénétrer[19]. Enfin, les coûts de développement et de production de ces produits restent très élevés rendant de ce fait les traitements très onéreux et pas forcément accessibles à tous. D'autres types de principes actifs biologiques comme les peptides, les nucléotides ou les microARNs sont également une alternative aux petites molécules avec un fort potentiel d'innovation[20]. Ainsi, en 2018 a été approuvé le premier ARN interférant, pour traiter et améliorer la neuropathie amyloïde familiale à transthyrétine[21].

D'un point de vue des cibles, établir le nombre de celles potentiellement modulables (« druggable ») par voie médicamenteuse demeure assez compliqué[22]. On estime à au moins 20 000 le nombres de protéines dans le corps humain[23]. Toutes ces protéines ne sont pas pour autant des cibles intéressantes d'un point de vue thérapeutique. Certaines ne sont pas impliquées dans des maladies, d'autres ne sont pas accessibles et enfin le rôle de la plupart n'est pas encore suffisamment caractérisé. Le projet « The Human Protein Atlas » fait part de 1 265 cibles potentielles dont 672 sont déjà la cible de médicaments sur le marché. Les cibles sont regroupées en famille selon leurs structures et leurs fonctions. Les principales familles de protéines humaines sont les récepteurs couplés aux protéines G (RCPGs) suivies des canaux ioniques puis des protéines kinases, que j'appellerai maintenant kinases pour simplifier et enfin des récepteurs nucléaires (Figure 3, gauche). Si l'on regarde la proportion par cibles de médicaments approuvés, les RCPGs restent premiers, suivi de près par les kinases, les canaux ioniques et les récepteurs nucléaires (Figure 3, droite). Cependant, les deux premières familles représentent à elles seules plus de la moitié de toutes les protéines cibles des médicaments approuvés. Si les RCPGs constituent depuis longtemps une famille de prédilection[24], le nombre de médicaments ciblant une protéine kinase n'a cessé d'augmenter depuis le début du siècle, si bien que cette famille est souvent considérée comme celle des nouvelles cibles du 21e siècle[25]. Le potentiel de cette famille reste encore sous-évalué puisque la plupart de ses membres n'ont encore jamais fait l'objet d'études cliniques[26]. De plus, il est désormais connu que les kinases sont des cibles de prédilections dans le domaine de l'oncologie. Je reviendrai plus en détails sur cette famille dans la suite de ce manuscrit.

[18] Kohzoh Imai et Akinori Takaoka, « Comparing Antibody and Small-Molecule Therapies for Cancer », *Nature Reviews Cancer* 6, nº 9 (septembre 2006): 714, https://doi.org/10.1038/nrc1913.
[19] Paul J. Carter, « Potent Antibody Therapeutics by Design », *Nature Reviews Immunology* 6, nº 5 (mai 2006): 343, https://doi.org/10.1038/nri1837.
[20] Keld Fosgerau et Torsten Hoffmann, « Peptide therapeutics: current status and future directions », *Drug Discovery Today* 20, nº 1 (1 janvier 2015): 122-28, https://doi.org/10.1016/j.drudis.2014.10.003.
[21] Arnt V Kristen et al., « Patisiran, an RNAi therapeutic for the treatment of hereditary transthyretin-mediated amyloidosis », *Neurodegenerative Disease Management* 9, nº 1 (27 novembre 2018): 5-23, https://doi.org/10.2217/nmt-2018-0033.
[22] John P. Overington, Bissan Al-Lazikani, et Andrew L. Hopkins, « How Many Drug Targets Are There? », *Nature Reviews Drug Discovery* 5, nº 12 (décembre 2006): 993, https://doi.org/10.1038/nrd2199.
[23] Elena A. Ponomarenko et al., « The Size of the Human Proteome: The Width and Depth », *International Journal of Analytical Chemistry* 2016 (2016), https://doi.org/10.1155/2016/7436849.
[24] Stephen J Hill, « G-protein-coupled receptors: past, present and future », *British Journal of Pharmacology* 147, nº Suppl 1 (janvier 2006): S27-37, https://doi.org/10.1038/sj.bjp.0706455.
[25] Philip Cohen, « Protein Kinases — the Major Drug Targets of the Twenty-First Century? », *Nature Reviews Drug Discovery* 1, nº 4 (avril 2002): 309-15, https://doi.org/10.1038/nrd773.
[26] Leah J. Wilson et al., « New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome », *Cancer Research* 78, nº 1 (1 janvier 2018): 15-29, https://doi.org/10.1158/0008-5472.CAN-17-2291.

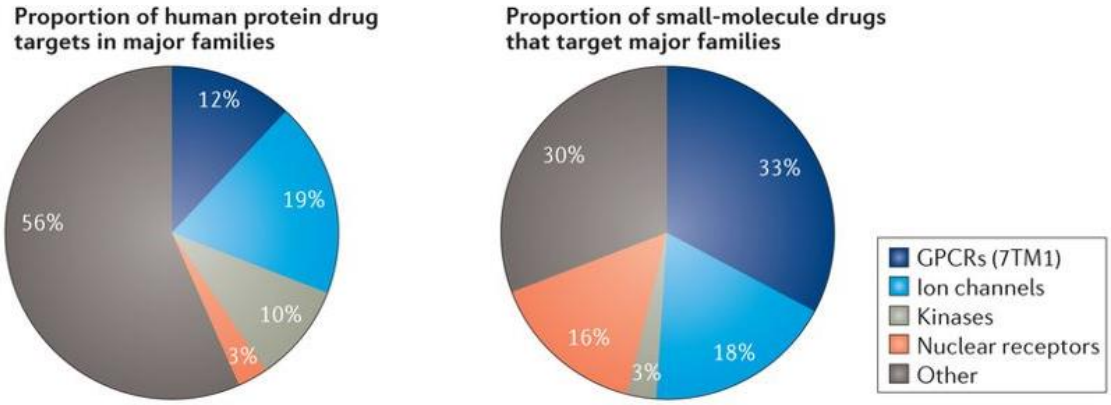


Figure 3 : Distribution des principales familles de cibles humaines (gauche) et distribution des principales cibles des médicaments approuvés (droite) . *D'après R. Santos et al.*[27].

1.3 Le futur de l'industrie pharmaceutique

Le secteur de l'industrie pharmaceutique est un poids lourd de l'économie mondiale. Le marché du médicament est un marché en constante croissance. Il a dépassé le seuil des 1 000 milliards de dollars de chiffre d'affaires en 2017[28] . La première entreprise mondiale dans le domaine de la santé est Johnson&Johnson qui se classe 9ème, en terme de capitalisation boursière, en 2019[29]. De fait, l'industrie pharmaceutique ne peut obtenir une rentabilité telle que les entreprises technologiques comme Google, Facebook ou Amazon. En effet, comme on l'a vu précédemment, à la différence d'un logiciel qui peut être développé très rapidement et sans énorme moyens financiers et humains, un médicament nécessite des dépenses en R&D bien plus importantes. Ainsi, les entreprises pharmaceutiques font partie des entreprises qui investissent le plus en recherche (9,8 % du chiffre d'affaires contre 4,8 % dans le secteur automobile par exemple)[30]. Il est estimé que ces coûts de R&D ont continuellement augmenté d'environ 10 % par an ces dernières années[15] (Figure 4). Cela s'explique par trois principaux facteurs :

- La complexification et l'allongement des études cliniques pour prouver l'efficacité du candidat médicament.

- Le taux d'échec élevé dans des nouveaux domaines d'études comme les maladies chroniques et dégénératives.

- La perte de brevets des « blockbusters » (on l'estime à 209 milliard pour la période 2009-2014)[31], ainsi que la difficulté à trouver de nouvelles molécules brevetables.

[27] Rita Santos et al., « A Comprehensive Map of Molecular Drug Targets », *Nature Reviews Drug Discovery* 16, nº 1 (janvier 2017): 19-34, https://doi.org/10.1038/nrd.2016.230.
[28] « Marché mondial », consulté le 17 avril 2019, https://www.leem.org/marche-mondial.
[29] PricewaterhouseCoopers, « Global Top 100 Companies 2019 », PwC, consulté le 17 octobre 2019, https://www.pwc.com/gx/en/services/audit-assurance/publications/global-top-100-companies-2019.html.
[30] « Recherche et développement », consulté le 21 mars 2019, https://www.leem.org/recherche-et-developpement.
[31] Ish Khanna, « Drug discovery in pharmaceutical industry: productivity challenges and trends », *Drug Discovery Today* 17, nº 19 (1 octobre 2012): 1088-1102, https://doi.org/10.1016/j.drudis.2012.05.007.

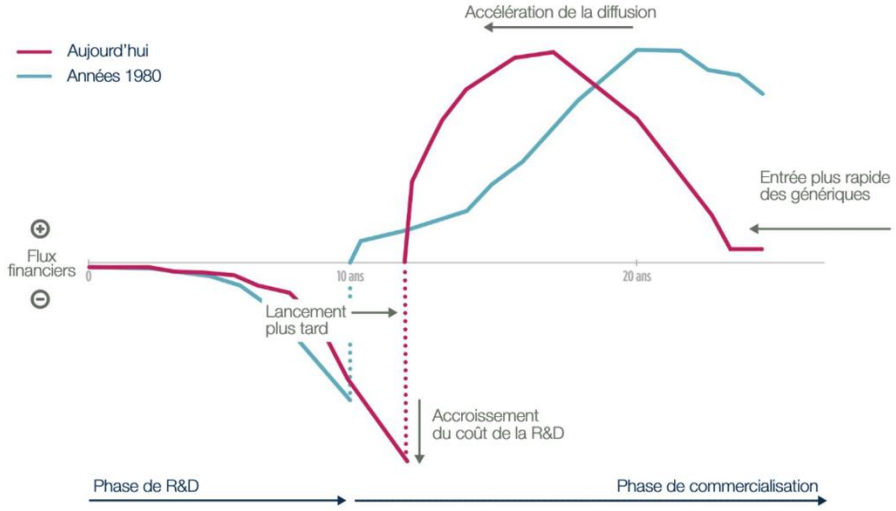


Figure 4 : Cycle de vie d'un médicament dans les années 1980 comparé à aujourd'hui. *Source Leem.*

Jusqu'à présent, c'est l'industrie pharmaceutique qui a le plus souvent défini elle-même la valeur de ses produits. Cependant, avec la hausse des coûts des soins médicaux, il lui est désormais de plus en plus difficile de continuer comme cela, les assurances santé se montrant de plus en plus réticentes à l'idée de rembourser des traitements très onéreux. On peut citer par exemple le cas récent du Sovaldi, développé par l'entreprise Gilead, un médicament contre l'hépatite. Le prix de ce traitement étant estimé trop cher, le gouvernement français est parvenu à un accord afin de l'encadrer[32]. Ainsi, les compagnies pharmaceutiques doivent s'adapter et proposer un nouveau modèle économique reposant non plus sur les blockbusters mais se dirigeant vers une médecine personnalisée tout en réduisant les coûts des soins médicaux. En effet, certains médicaments ne seront remboursés que s'il existe un réel bénéfice chez le patient, la thérapie ciblée devient donc de plus en plus visée. En outre, désormais le patient, tout comme le médecin, est beaucoup plus averti et méfiant face à cette « consommation de masse » des médicaments Enfin, comme dans tous les secteurs, l'avènement de l'intelligence artificielle et des données massives (« Big Data »), exige le recours à de nouvelles expertises et compétences, en passant notamment par la science des données et la chémoinformatique que je vais présenter plus en détail dans la prochaine partie.

A noter que si les entreprises pharmaceutiques continuent d'innover, l'apport de nouvelles technologies dans le domaine réduit le temps de développement des nouveaux médicaments, il faut donc également que les autorités régulatrices puissent s'adapter rapidement.

1.4 L'essor de la chémoinformatique

A la différence de son homologue en biologie, la bioinformatique, la chémoinformatique est un domaine plus méconnu du grand public. J'ai eu l'occasion de vérifier cette affirmation moi-même au cours de différentes discussions, avec des proches ou des inconnus, lors desquelles j'abordais cette thématique, qui en général les laissaient pantois. Pourtant, la chémoinformatique n'a rien à envier à la bioinformatique, il s'agit d'ailleurs de deux spécialités complémentaires plutôt que rivales. En effet, là où la bioinformatique va plutôt se focaliser sur l'étude des protéines et des données dites « omiques » (génomique, protéomique,

[32] « Coût des nouveaux traitements de lutte contre l'hépatite C - Sénat », consulté le 11 juin 2019, https://www.senat.fr/questions/base/2014/qSEQ140712580.html.

transcriptomique, etc.), la chémoinformatique va se servir des outils informatiques pour étudier les données issues des différents domaines de la chimie (de la synthèse à l'analyse).

1.4.1 Historique et domaine d'application

En Europe, la chémoinformatique a été officiellement introduite en 2006 lors du congrès d'Obernai : « Workshop Chemoinformatics in Europe: Research and Teaching ». Elle est définie comme l'interface entre la chimie et l'informatique ayant pour but de procurer des outils et des méthodes pour analyser et traiter les données issues des différents domaines de la chimie[33]. Cependant, il existe une première définition plus ancienne proposée par F.K. Brown en 1998 : *"Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization."*[34]. En France, elle est représentée par la Société Française de Chémoinformatique (SFCi), actuellement présidée par le Pr Matthieu Montes, qui regroupe « plus de 100 scientifiques français (60 % académiques, 40 % industriels) issus de 35 laboratoires académiques (dont 4 à l'étranger) et de 27 sociétés privées »[35]. En 2009 est apparu un journal spécifique à la chémoinformatique : Journal of Chemoinformatics améliorant la visibilité de cette discipline.

Les domaines d'applications de la chémoinformatique sont divers. Elle sert tout d'abord à la création et à la gestion de bases de données de molécules, pouvant en contenir plusieurs millions, ou de réactions chimiques. Elle est aussi utilisée dans la prédiction de propriétés physiques, chimiques ou biologiques à l'aide de méthodes statistiques telles que le QSAR (relation quantitative structure à activité) ou le criblage virtuel. On va désormais retrouver la chémoinformatique dans toutes les étapes de création d'un nouveau médicament, que ce soit pour l'analyse de données générées par un criblage HTS - assurément qui dit criblage virtuel de 200 000 molécules dit 200 000 résultats à exploiter, ingérable sans outils informatiques – ou lors de la sélection des leads pour poursuivre les investigations grâce aux différentes prédictions de propriétés guidant le choix final. Ainsi, grâce aux différentes prédictions et modélisations, la chémoinformatique permet à la fois une réduction des coûts mais aussi un gain de temps, un facteur non négligeable avant expiration du brevet comme expliqué précédemment. C'est pour cela qu'on trouve maintenant au moins un département ou une équipe de chémoinformaticiens dans toutes les grandes et moyennes entreprises de conception de médicaments[36]. A noter qu'un chémoinformaticien est souvent un chimiste de formation.

L'un des fondements de la chémoinformatique fut de trouver comment représenter informatiquement une molécule. En effet, un ordinateur, ou plus précisément un programme informatique, a besoin de données dans un format structuré pour pouvoir fonctionner. Ainsi, en étant un peu caricatural, on ne peut pas lui fournir une photo du cahier de laboratoire et espérer qu'il nous retourne les propriétés physico-chimiques des molécules dessinées. Cela viendra sûrement mais aujourd'hui les données doivent être transcrites dans un format spécifique, compréhensible par un programme informatique afin d'être traitées.

---

[33] « The Obernai Declaration », *Workshop Chemoinformatics in Europe: Research and Teaching*, (2006), 2.
[34] Frank K. Brown, « Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. », in *Annual Reports in Medicinal Chemistry*, éd. par James A. Bristol, vol. 33 (Academic Press, 1998), 375-84, https://doi.org/10.1016/S0065-7743(08)61100-8.
[35] « Société Française de Chémoinformatique », consulté le 3 avril 2019, http://www.chemoinformatique.fr/.
[36] Alexander Hillisch, Nikolaus Heinrich, et Hanno Wild, « Computational Chemistry in the Pharmaceutical Industry: From Childhood to Adolescence », *ChemMedChem* 10, nº 12 (2015): 1958-62, https://doi.org/10.1002/cmdc.201500346.

1.4.2 La représentation d'une molécule

Telle que définie par l'Union internationale de chimie pure et appliquée (UICPA), une molécule est « une entité électriquement neutre comprenant plus d'un atome »[37]. Cette définition vague ne permet pas facilement de se représenter une molécule. Plus particulièrement, on peut la définir comme un ensemble d'atomes, similaires ou non, unis entre eux par des liaisons chimiques. La composition en atomes d'une molécule est donnée par sa formule chimique. La formule chimique ne donne pas d'information sur la structure, on ne peut donc pas s'en servir pour décrire précisément une molécule (Figure 5). Lorsque l'on dessine une molécule, on va utiliser, entre autre, la formule développée ou la formule de Lewis pour obtenir une représentation en 2D. En chémoinformatique, la molécule est représentée sous forme de graphe pour permettre à l'ordinateur de la comprendre et la traiter. Le choix du format de ce graphe va dépendre de plusieurs facteurs tels que le besoin d'informations structurales ou la taille de stockage disponible. Historiquement, de nombreux formats propriétaires ont été développés. On peut les distinguer en trois catégories principales, selon la dimension de représentation des molécules : 1D, 2D ou 3D.

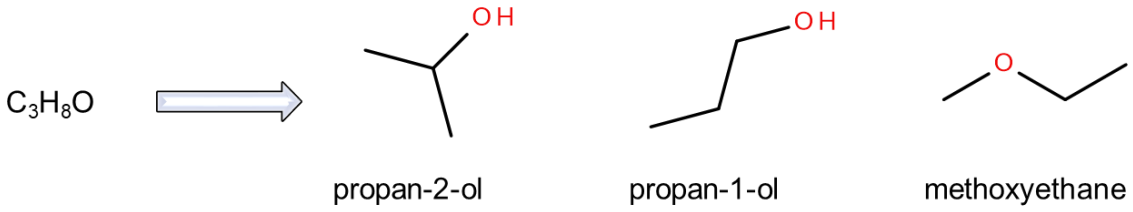


Figure 5 : Représentation de différentes molécules avec la même formule chimique.

- Les formats 1D sont une représentation textuelle des molécules. Ils retranscrivent uniquement l'information topologique (types d'atomes et types de liaisons entre eux). Il s'agit de formats compacts, en règle générale ils décrivent une molécule par ligne et sont donc adaptés pour le stockage car ils ne sont pas volumineux. On retrouve par exemple dans cette catégorie les formats suivants : SMILES, SMARTS, InChi, IUPAC…

- Les formats 2D stockent les molécules sous forme de table de connexion à partir de leur graphe initial. La différence majeure avec les formats 1D est qu'ils contiennent les coordonnées x, y de chaque atome. En règle générale, la conversion 1D - 2D et inversement se fait très facilement par tous les outils de chémoinformatique.

- Les formats 3D sont les formats les plus volumineux. Ils retranscrivent précisément le positionnement dans l'espace de chaque atome de la molécule, ainsi que les caractéristiques des liaisons (distance, angle, dièdre). Une molécule étant un objet flexible, elle peut avoir plusieurs conformations spatiales. Ces conformations peuvent être soit calculées à partir des molécules en 2D en utilisant un champ de force respectant les règles géométriques d'une molécule[38], soit être déterminées par l'environnement dans lequel se situe la molécule (par exemple le site actif d'une protéine). Dans le cas

[37] P. Muller, « Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994) », *Pure and Applied Chemistry* 66, nº 5 (2009): 1077–1184, https://doi.org/10.1351/pac199466051077.
[38] Sereina Riniker et Gregory A. Landrum, « Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation », *Journal of Chemical Information and Modeling* 55, nº 12 (28 décembre 2015): 2562-74, https://doi.org/10.1021/acs.jcim.5b00654.

d'études de poses cristallographiques ou de résultats d'amarrages moléculaires, les données 3D sont indispensables pour étudier les interactions entre le ligand et sa cible.

En plus des formats descriptifs, existent aussi les identifiants moléculaires. Il s'agit d'un identifiant unique relié à la molécule permettant de la retrouver facilement dans les bases de données. Ces identifiants peuvent être propriétaires (numéro CAS) ou publics comme les identifiants propres des bases de données publiques. Ainsi, l'aspirine aura comme identifiant « CHEMBL25 », ou « DB00945 » selon la base de données et comme numéro CAS : « 50-78-2 ».

Avant de voir d'autre méthodes de descriptions moléculaires, il est important de garder en tête que peu importe le format de représentation, il s'agit toujours de modèle virtuel de molécule, comme le rappelait M. Hann[39] (Figure 6).

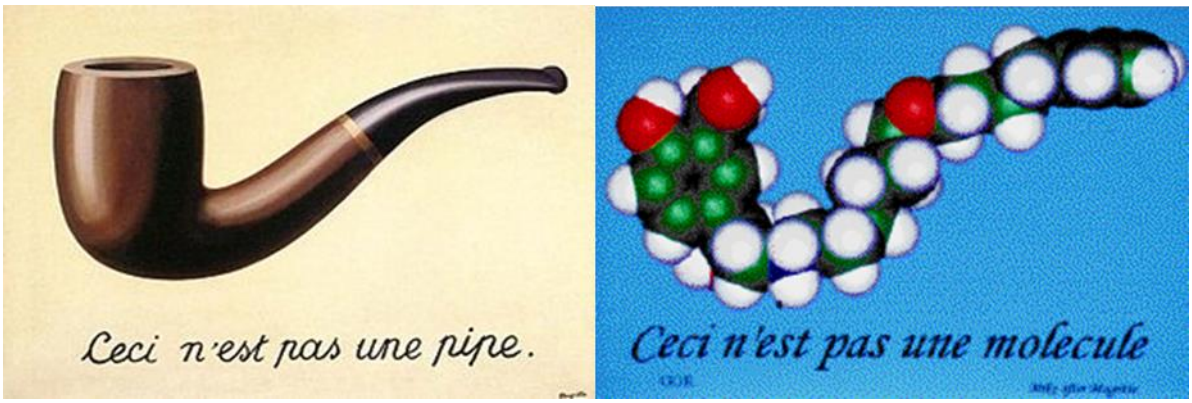


Figure 6 : Ceci n'est pas une molecule. *A gauche, « La Trahison des images » de René Magritte (1929) ; à droite, "Ceci n'est pas une molecule," qui pour Mike Hann, "serves to remind us that all of the graphics images presented here are not molecules, not even pictures of molecules, but pictures of icons which we believe represent some aspects of the molecule's properties".*

1.4.3 Les descripteurs moléculaires

Comme son nom l'indique, un descripteur moléculaire va servir à décrire une molécule. Une définition plus précise stipule qu'il s'agit du résultat final d'une procédure logique et mathématique qui transforme une information chimique encodée dans la représentation symbolique d'une molécule en une valeur numérique, ou du résultat d'une expérience standardisée[40]. Ainsi, les descripteurs moléculaires sont divisés en deux catégories : les descripteurs déterminés expérimentalement, notamment les propriétés physico-chimiques, ou les descripteurs théoriques, calculés par une formule mathématique ou un algorithme. Les descripteurs moléculaires font appel à plusieurs domaines scientifiques : chimie quantique, chimie organique, théorie des graphes, etc. Ils sont utilisés pour modéliser différentes propriétés moléculaires afin de reproduire les données expérimentales et prédire des données inconnues. Il en existe à ce jour pléthore[41] et il convient au chémoinformaticien de choisir les bons en rapport avec ses besoins. On peut les diviser en plusieurs classes selon la dimension utilisée :

[39] David S. Goodsell, Teresa A. Larsen, et T. J. O'Donnell, « 1994 Molecular Graphics Art Show and Video Show », *Journal of Molecular Graphics* 13, nº 4 (1 août 1995): 223-34, https://doi.org/10.1016/0263-7855(95)00036-6.
[40] « Molecular Descriptors for Chemoinformatics | Methods and Principles in Medicinal Chemistry », consulté le 19 avril 2019, https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766.
[41] « Molecular Descriptors Software », consulté le 19 avril 2019, http://www.moleculardescriptors.eu/softwares/softwares.htm.

- Les plus simples, les descripteurs 0D, vont décrire les propriétés macroscopiques de la molécule, tels que le poids moléculaire ou le nombre d'atomes de carbone. Ils sont très utilisés en chémoinformatique car ils sont très rapides et très simples à calculer. De plus, l'information sur la structure de la molécule n'est pas nécessaire pour les déterminer.

- Les descripteurs 1D sont des descripteurs fondés sur la recherche de sous fragments dans la molécule. Là encore, les descripteurs de cette catégorie sont rapides à calculer et les plus couramment utilisés sont les empreintes moléculaires, souvent appelées « fingerprints ». On les utilise notamment pour des études de similarité ou au contraire de diversité entre composés. Les empreintes moléculaires encodent la molécule sous forme de chaine binaire dont chaque bit représente la présence (1) ou l'absence (0) d'un ou plusieurs motifs structuraux. Ainsi, l'information est stockée de manière efficace et optimale pour communiquer avec l'ordinateur. En comparant les empreintes, on peut évaluer le niveau de similarité entre molécules. Les empreintes moléculaires peuvent être soit fondées sur des motifs prédéfinis et avoir une longueur déterminée par le nombre de motifs, soit construits à partir des motifs propres des molécules avec une longueur définie par le programmeur. L'exemple le plus connu de motifs prédéfinis est l'empreinte MACCS qui encode 166 motifs structuraux communément rencontrés en chimie thérapeutique[42]. L'avantage de celle-ci est la vitesse de calcul, même pour des milliers de molécules. L'inconvénient est que comme les motifs sont prédéterminés, ils peuvent ne pas être adaptés dans certaines situations. Dans le cas où les fragments sont calculés à la volée, les empreintes moléculaires de hashage procèdent en découpant de manière linéaire ou circulaire chaque molécule selon une longueur définie. L'avantage est que l'empreinte reflète bien les motifs propres de la molécule. L'inconvénient est que l'on ne peut pas connaitre à quoi chaque bit correspond et qu'il faut bien ajuster sa longueur pour bien décrire toute la molécule. On trouve dans cette catégorie les empreintes ECFPs[43]. La Figure 7 illustre les deux différents types d'empreintes moléculaires rencontrés.

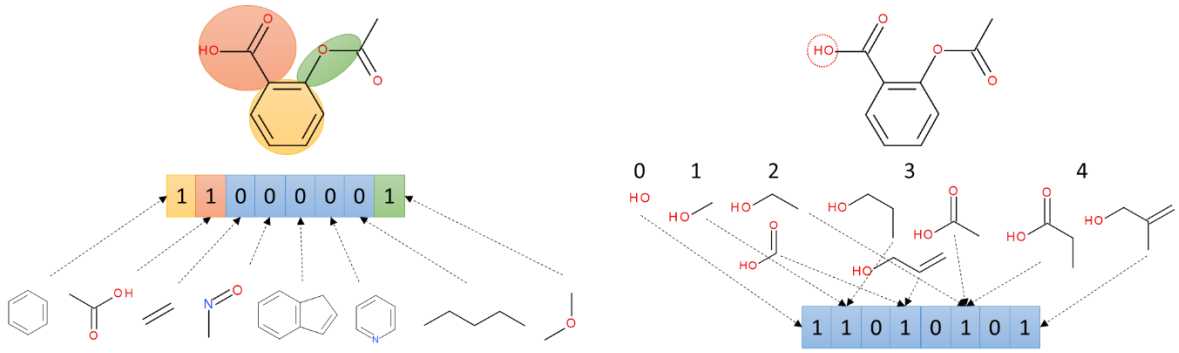


Figure 7 : Représentation de deux types d'une empreinte moléculaire de 8 bits. *A gauche une empreinte moléculaire par sous-structures (motifs prédéterminés). A droite, une empreinte topologique avec hashage circulaire de la molécule jusqu'à une distance de 4 (seuls les fragments partant de l'atome d'oxygène entourés en rouge sont montrés). On remarque que plusieurs fragments encodent le même bit (collision), cela signifie que la longueur de l'empreinte n'est pas adaptée.*

- Les descripteurs 2D et 3D fournissent le plus de détails en s'appuyant sur la topologie de la molécule et en la caractérisant spatialement. Ils sont basés sur une représentation

[42] Joseph L. Durant et al., « Reoptimization of MDL Keys for Use in Drug Discovery », *Journal of Chemical Information and Computer Sciences* 42, nº 6 (1 novembre 2002): 1273-80, https://doi.org/10.1021/ci010132r.
[43] David Rogers et Mathew Hahn, « Extended-Connectivity Fingerprints », *Journal of Chemical Information and Modeling* 50, nº 5 (24 mai 2010): 742-54, https://doi.org/10.1021/ci100050t.

de la molécule sous forme de graphe. On peut retrouver par exemple dans ces descripteurs les distances ou le nombre de connexions entre atomes et le volume ou la forme moléculaire.

Finalement, le choix des descripteurs repose avant tout sur du bon sens et de la rigueur scientifique et doit être adapté à la problématique. Ainsi, si l'on veut regrouper des molécules par leur forme, on va plutôt choisir des descripteurs 3D. A l'inverse, si le but est de séparer des composés selon leur composition, des descripteurs 0D et 1D suffiront. Les descripteurs moléculaires sont très utilisés en chémoinformatique, notamment pour créer des modèles prédictifs (QSAR), pour étudier la diversité d'un échantillon de molécules ou bien pour filtrer et caractériser des molécules.

1.4.4 Les bases de données en chimie

Une collection de molécules, appelée chimiothèque, peut être réelle ou virtuelle. Dans ce dernier cas, les molécules sont stockées sous forme de données électroniques dans une base de données informatique. Il existe des chimiothèques générales ou des chimiothèques spécifiques, uniquement constituées par exemple d'inhibiteurs de kinases, d'inhibiteurs d'interaction protéine-protéine ou encore de produits naturels. Avec l'essor de la chimie combinatoire, le nombre de composés chimiques n'a cessé d'augmenter, chaque entreprise pharmaceutique qui pratique le criblage à haut débit possédant une collection de plusieurs milliers/millions de molécules. Très vite, cet essor a créé le besoin d'outils informatiques pour stocker et interroger toutes ces données accumulées. Le premier programme permettant d'effectuer des recherches dans une librairie de molécules virtuelles est paru en 1957[44]. Depuis, de nouveaux programmes ont émergés et ne cessent de s'améliorer suivant les progrès et les avancées en informatique.

En chimie il existe plusieurs types de bases de données spécifiques que ce soit pour la recherche bibliographique, la recherche de composés ou bien la recherche de structures cristallographiques. La majorité des informations des bases de données provient de la bibliographie scientifique qui renseigne sur l'activité de nombreux composés chimiques. Afin d'éviter un long travail fastidieux de recherche et d'extraction manuelle, les bases de données regroupent toutes ces informations et les rendent facilement accessibles au travers d'une interface graphique. Ainsi, de nombreuses bases de données sont disponibles pour permettre d'accéder par exemple aux données de bioactivité des molécules que l'on étudie. On citera les plus utilisées : ChEMBL[45], PubChem[46] (publiques) et SciFinder (Chemical Abstracts Service, privée). Un extrait d'une page d'un composé dans la ChEMBL est visible Figure 8. Une autre base de données importante à connaitre est la RCSB PDB[47], qui regroupe les structures cristallographiques existantes des protéines seules ou en complexe avec leur ligand. C'est notamment dans cette dernière base que l'on va aller retrouver des structures tridimensionnelles de protéines à l'aide d'un code PDB à 4 lettres.

[44] Louis C. Ray et Russell A. Kirsch, « Finding Chemical Records by Digital Computers », *Science* 126, nº 3278 (25 octobre 1957): 814-19, https://doi.org/10.1126/science.126.3278.814.
[45] Anna Gaulton et al., « The ChEMBL Database in 2017 », *Nucleic Acids Research* 45, nº D1 (4 janvier 2017): D945-54, https://doi.org/10.1093/nar/gkw1074.
[46] Sunghwan Kim et al., « PubChem 2019 Update: Improved Access to Chemical Data », *Nucleic Acids Research* 47, nº D1 (8 janvier 2019): D1102-9, https://doi.org/10.1093/nar/gky1033.
[47] Helen M. Berman et al., « The Protein Data Bank », *Nucleic Acids Research* 28, nº 1 (1 janvier 2000): 235-42, https://doi.org/10.1093/nar/28.1.235.

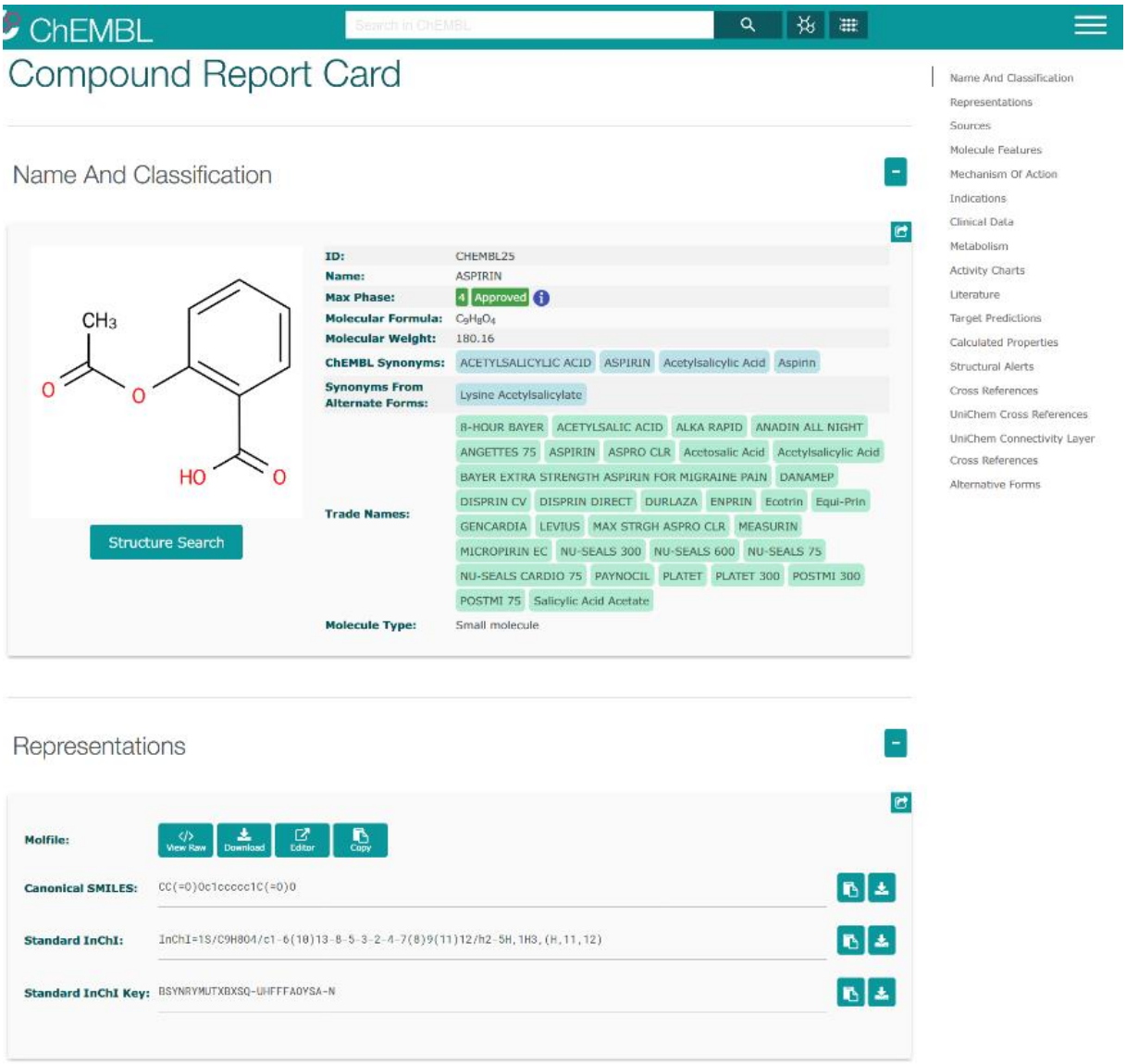


Figure 8 : Extrait de la fiche descriptive de l'aspirine dans la base de données ChEMBL. *Plusieurs formats de téléchargement sont disponibles, ainsi que de multiples informations à travers le menu déroulant en haut à droite.*

Mon introduction générale résumant les concepts à connaitre pour la suite se termine ici. Je vais à présent présenter plus en détail la famille des protéines kinases qui est la famille sur laquelle j'ai réalisé mes travaux de thèse.

1.5 L'émergence des protéines kinases

Des nombreuses cibles potentielles du corps humain, les kinases comptent parmi celles qui ont été le plus étudiées ces 30 dernières années[48]. La raison en est plutôt simple : les kinases jouent un rôle dans pratiquement tous les aspects de la vie cellulaire. Elles peuvent contrôler le métabolisme, la transcription, la transduction, la division et enfin la mort programmée de la cellule[49]. Une dérégulation ou une mutation de l'une de ces fonctions peut ainsi avoir des

[48] Fleur M. Ferguson et Nathanael S. Gray, « Kinase Inhibitors: The Road Ahead », *Nature Reviews Drug Discovery* 17, nº 5 (16 mars 2018): 353-77, https://doi.org/10.1038/nrd.2018.21.
[49] G. Manning et al., « The Protein Kinase Complement of the Human Genome », *Science* 298, nº 5600 (6 décembre 2002): 1912-34, https://doi.org/10.1126/science.1075762.

conséquences sur le système immunitaire ou nerveux[50], mais aussi être à l'origine de cancers[51]. La Figure 9 montre les différentes implications des kinases dans les différents évènements de la vie cellulaire.

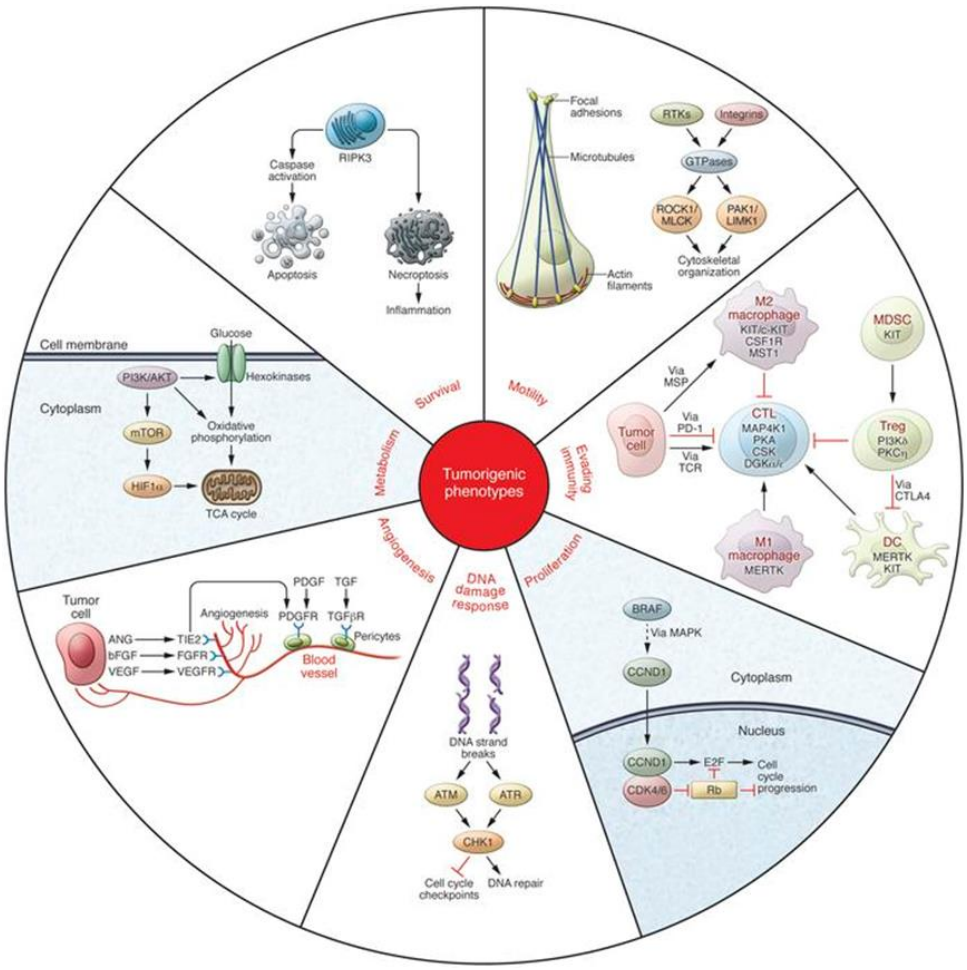


Figure 9 : Régulation des différents facteurs tumoraux par activation de protéines kinases. *D'après F.M. Ferguson & al.*[52].

1.5.1 Historique et rôle physiologique

Pour commencer, qu'appelle-t-on précisément une protéine kinase ? Il s'agit d'une enzyme catalysant le transfert d'un groupe phosphoryle d'un substrat à un autre. Les kinases font donc partie de la famille d'enzymes des transférases et sont responsables du transfert du phosphate γ de l'ATP sur le groupement hydroxyle (-OH) de leur substrat. Ce substrat peut être un lipide, un sucre ou une autre protéine. Dans notre étude, nous nous intéresserons uniquement aux protéines. Chez les eucaryotes, le transfert du groupement phosphate s'effectue spécifiquement sur les résidus de la chaine latérale possédant une fonction alcool : la sérine (Ser), la thréonine (Thr) ou la tyrosine (Tyr). On appelle cette réaction chimique la phosphorylation, il s'agit d'une modification post-traductionnelle. C'est à travers cette

[50] Peter B. Crino, « The MTOR Signalling Cascade: Paving New Roads to Cure Neurological Disease », *Nature Reviews Neurology* 12, nº 7 (juillet 2016): 379-92, https://doi.org/10.1038/nrneurol.2016.81.
[51] Stefan Gross et al., « Targeting Cancer with Kinase Inhibitors », *The Journal of Clinical Investigation* 125, nº 5 (1 mai 2015): 1780-89, https://doi.org/10.1172/JCI76094.
[52] Fleur M. Ferguson et Nathanael S. Gray, « Kinase Inhibitors: The Road Ahead », *Nature Reviews Drug Discovery* 17, no 5 (16 mars 2018): 353-77, https://doi.org/10.1038/nrd.2018.21.

phosphorylation que les kinases vont réguler l'activité des protéines ciblées, on parle alors d'activation. La réaction de déphosphorylation afin de désactiver la protéine est effectuée par une phosphatase qui va catalyser le retrait du groupement phosphate de la protéine (Figure 10).

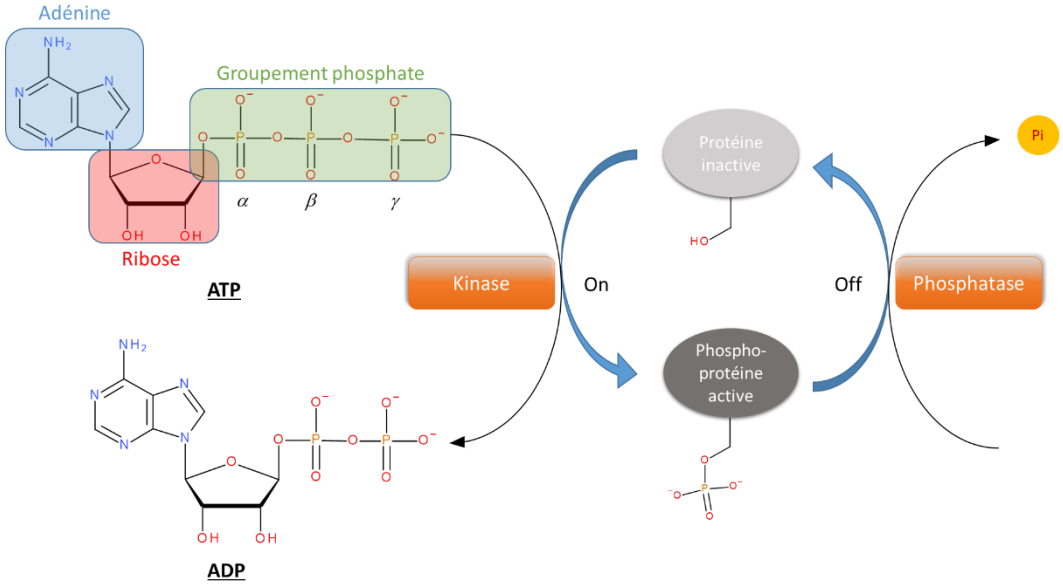


Figure 10 : Représentation schématique de la phosphorylation/déphosphorylation d'une protéine cible par une protéine kinase ou une phosphatase. *Le phosphate en position γ de l'ATP est transféré sur la protéine cible entrainant son activation. La désactivation se fait par une phosphatase et entraine la libération du phosphate sous forme inorganique (Pi).*

Les premières traces de phosphate retrouvées dans une protéine, la caséine, présente dans le lait, remontent à 1883[53]. Puis en 1906, on en trouve aussi dans la vitelline, une protéine contenue dans le jaune d'œuf[54]. Ensuite, à partir de 1932, F.A. Lipmann et P.A. Levene mettent en évidence le premier acide aminé phosphaté : la phosphosérine, toujours dans la vitelline[55]. A noter qu'à cette époque la thréonine n'était pas encore connue, il faudra attendre 1953 pour identifier la phosphothréonine dans la caséine[56]. Cependant, ce n'est que suite aux travaux de G. Burnett et E.P. Kennedy que la réaction de phosphorylation d'une protéine sera décrite en 1954[57]. Finalement, ce seront E.H. Fischer et E.G. Krebs qui vont nommer l'enzyme responsable de cette réaction : la « phosphorylase kinase »[58]. S'en suivent alors de nombreuses découvertes permettant de mettre en évidence d'une part, un très grand nombre de représentants de la famille des kinases, d'autre part leur implication dans la transduction du signal cellulaire à travers des réactions en cascades (Figure 11)[59]. Comme preuve de l'importance des kinases dans les fonctions du vivant, trois prix Nobel de physiologie ou de médecine ont récompensé des recherches sur ces protéines :

[53] Olof Hammarsten, « Zur Frage, ob das Caseïn ein einheitlicher Stoff sei. », *Zeitschrift für physiologische Chemie* 7, nº 3 (1883): 227–273, https://doi.org/10.1515/bchm1.1883.7.3.227.
[54] P. A. Levene et C. L. Alsberg, « The Cleavage Products of Vitellin », *Journal of Biological Chemistry* 2, nº 1 (8 janvier 1906): 127-33.
[55] Fritz A. Lipmann et P. A. Levene, « Serinephosphoric Acid Obtained on Hydrolysis of Vitellinic Acid », *Journal of Biological Chemistry* 98, nº 1 (10 janvier 1932): 109-14.
[56] C. H. De Verdier, « Isolation of Phosphothreonine from Bovine Casein », *Nature* 170, nº 4332 (8 novembre 1952): 804-5.
[57] George Burnett et Eugene P. Kennedy, « The Enzymatic Phosphorylation of Proteins », *Journal of Biological Chemistry* 211, nº 2 (12 janvier 1954): 969-80.
[58] Edwin G. Krebs, Donald J. Graves, et Edmond H. Fischer, « Factors Affecting the Activity of Muscle Phosphorylase b Kinase », *Journal of Biological Chemistry* 234, nº 11 (11 janvier 1959): 2867-73.
[59] Robert Roskoski, « A historical overview of protein kinases and their targeted small molecule inhibitors », *Pharmacological Research* 100 (1 octobre 2015): 1-23, https://doi.org/10.1016/j.phrs.2015.07.010.

- En 1992 : Edmond H. Fischer et Edwin G. Krebs* pour « leurs découvertes concernant les phosphorylations réversibles de protéines comme un mécanisme de régulation biologique »[60].

- En 2000 : Arvid Carlsson, Paul Greengard et Eric R. Kandel pour « leurs découvertes concernant la transduction du signal dans le système nerveux »[60].

- En 2001 : Leland H. Hartwell, Tim Hunt et Sir Paul M. Nurse pour « leurs découvertes de régulateurs clés du cycle cellulaire »[60].

* Attention à ne pas confondre Edwin G. Krebs avec Hans A. Krebs qui a également reçu le Prix Nobel en 1953 pour la découverte du cycle de l'acide citrique, de là à voir un lien de corrélation entre un nom et être récompensé du prix Nobel, je vous laisse seul juge.

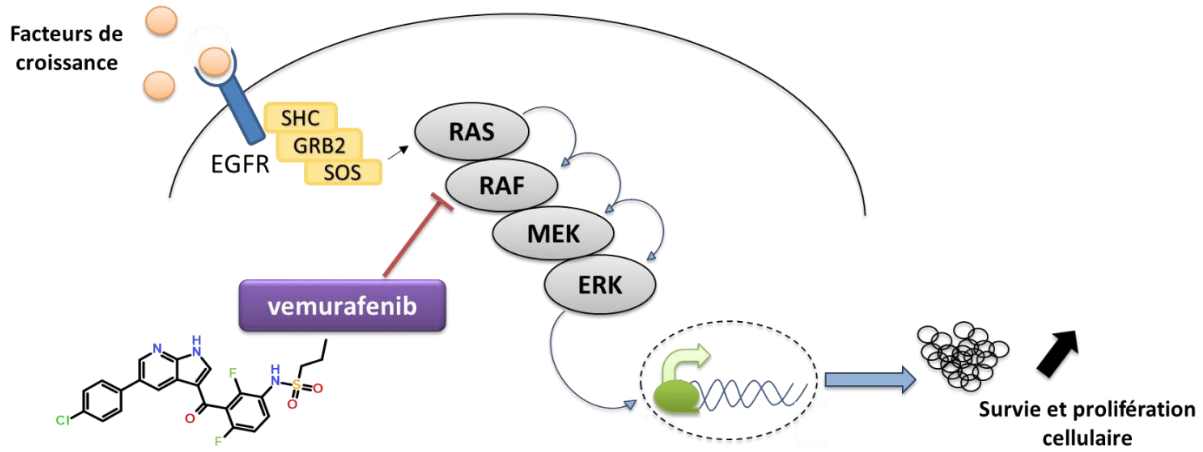


Figure 11 : Exemple d'une cascade de protéines kinases pour réguler la prolifération cellulaire.

1.5.2 La classification des kinases

Les kinases représentent une famille de plusieurs centaines de membres regroupés sous le terme de kinome. Les premières estimations font état de 1 001 kinases dans les années 1980[61]. Plus tard, les travaux de G. Manning en 2002, qui font désormais office de référence, redéfinissent ce chiffre à 518 : 478 protéines kinases eucaryotes (ePK) et 40 protéines kinases atypiques (aPK)[62]. Une kinase atypique présente une similarité de séquence plus faible avec les ePK. Depuis, quelques erreurs ont été trouvées et les publications récentes font état de 538 protéines kinases[63,64]. Enfin, une même kinase peut posséder plusieurs domaines kinases. Ainsi on dénombre au total 491 domaines kinases ePK et 40 domaines aPK. Les domaines kinases sont classés selon leur degré de parenté de séquence en 9 groupes principaux, on les représente sous la forme d'un arbre phylogénétique (Figure 12). Par exemple, dans le groupe des tyrosines kinases (TK), on regroupe la famille des Janus kinases (JAK) composé de 4 sous-familles (JAK1, JAK2, JAK3 et TYK2). Les protéines kinases sont caractérisées par la présence d'au

[60] « All Nobel Prizes in Physiology or Medicine », NobelPrize.org, consulté le 12 avril 2019, https://www.nobelprize.org/prizes/lists/all-nobel-laureates-in-physiology-or-medicine/.
[61] T. Hunter, « A Thousand and One Protein Kinases », *Cell* 50, nº 6 (11 septembre 1987): 823-29.
[62] Manning et al., « The Protein Kinase Complement of the Human Genome ».
[63] Silvia Braconi Quintaje et Sandra Orchard, « The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot: One Small Step in Manual Annotation, One Giant Leap for Full Comprehension of Genomes », *Molecular & Cellular Proteomics* 7, nº 8 (1 août 2008): 1409-19, https://doi.org/10.1074/mcp.R700001-MCP200.
[64] Doriano Fabbro, Sandra W Cowan-Jacob, et Henrik Moebitz, « Ten things you should know about protein kinases: IUPHAR Review 14 », *British Journal of Pharmacology* 172, nº 11 (juin 2015): 2675-2700, https://doi.org/10.1111/bph.13096.

moins un domaine catalytique (domaine kinase) mais peuvent aussi posséder d'autres domaines pouvant interagir avec d'autres protéines ou participer à la régulation de la kinase[65,66].

[65] B. J. Mayer, H. Hirai, et R. Sakai, « Evidence That SH2 Domains Promote Processive Phosphorylation by Protein-Tyrosine Kinases », *Current Biology: CB* 5, nº 3 (1 mars 1995): 296-305.
[66] John Kuriyan et David Eisenberg, « The Origin of Protein Interactions and Allostery in Colocalization », *Nature* 450 (12 décembre 2007): 983-90, https://doi.org/10.1038/nature06524.

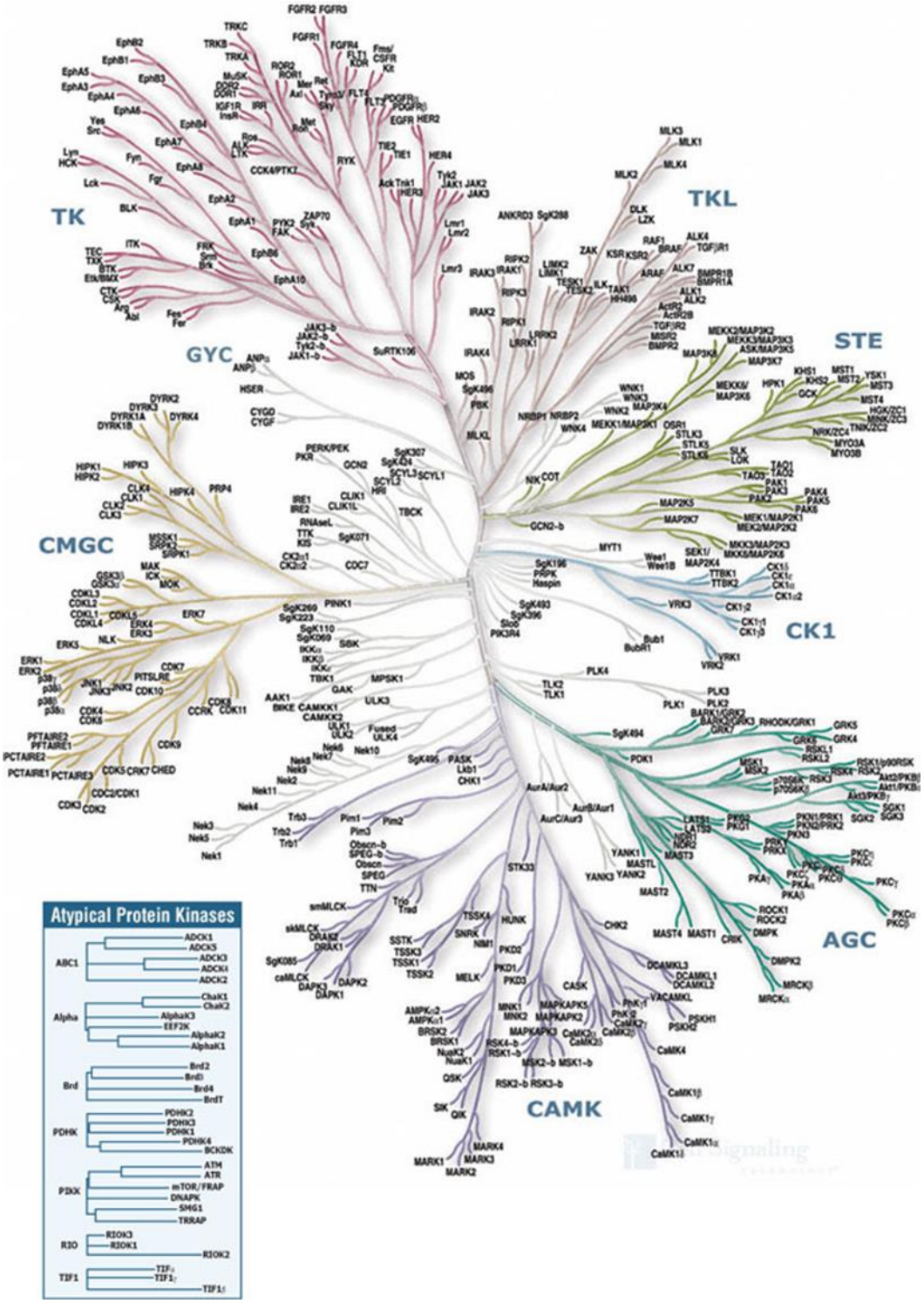


Figure 12 : Arbre phylogénétique du kinome humain. *Il représente les liens de parenté entre les 491 domaines kinases eucaryotes (ePK) et les 40 domaines kinases atypiques (aPK) répertoriés à ce jour. Les protéines kinases atypiques sont présentées à part et se divisent elles-mêmes en 8 groupes, la partie « Autres » des kinases atypiques n'étant pas indiquée ici. Source Cell Signaling Technology*[67].

[67] « Protein Kinases: Human Protein Kinases Overview | CST », consulté le 5 juin 2019, https://www.cellsignal.com/contents/science-protein-kinases/protein-kinases-human-protein-kinases-overview/kinases-human-protein.

1.5.3 La structure des kinases

La première structure tridimensionnelle de kinase en complexe avec l'ATP est parue en 1993 (code PBD : 1ATP)[68]. Il s'agit d'une protéine kinase appartenant à la famille PRKACA (groupe AGC) chez la souris, obtenue par cristallographie aux rayons X. Depuis, suite aux avancées des techniques pour obtenir une structure tridimensionnelle (cristallographie aux rayons X, résonance magnétique nucléaire et microscopie électronique principalement), couplées à l'engouement général pour les kinases, de nombreuses autres structures sont apparues (Figure 13). Ainsi, le nombre de publications par année, toutes espèces confondues, n'a cessé de croître jusqu'aux années 2010 puis est devenu stable avec environ 350 nouveaux articles par an. Cependant, ces chiffres ne reflètent pas tout le kinome humain. En effet, malgré plus de 4 000 structures tridimensionnelles de protéines kinases humaines, seuls 295 domaines kinases sont représentés sur les 518 (ou 538). Ainsi, certaines kinases sont présentes de nombreuses fois comme par exemple CDK2 que l'on peut retrouver dans plus de 370 structures. La raison de cette disparité s'explique notamment par le rôle de certaines kinases démontré en oncologie. Pour CDK2, on sait que celle-ci est impliquée dans la régulation du cycle cellulaire et elle présente donc un fort intérêt pharmaceutique.

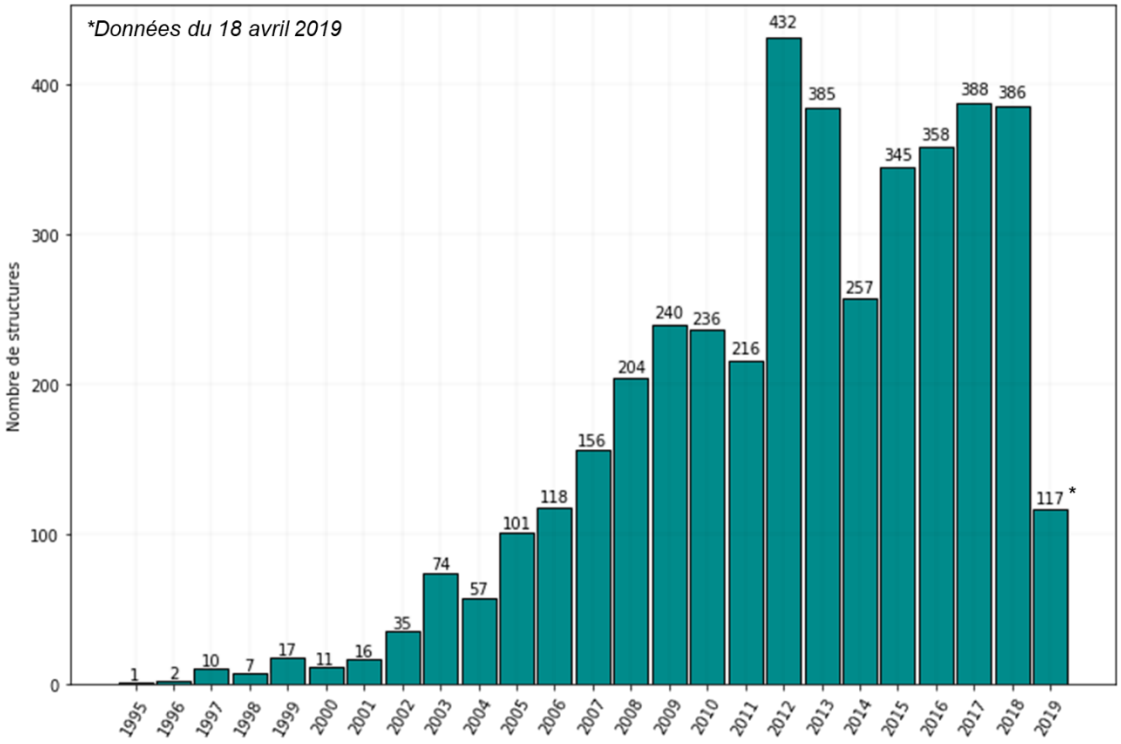


Figure 13 : Evolution du nombre de structures de domaines kinases dans la RCSB de 1995 à avril 2019. *Le nombre présenté correspond aux structures comportant les domaines PFAM PF07714 et/ou PF00069.*

Il est apparu au fil des structures disponibles que non seulement les différents sites catalytiques des protéines kinases présentaient tous une similarité de séquence, mais surtout un repliement tridimensionnel commun avec des éléments structuraux clés pour la fixation du ligand. Pour la description du domaine kinase et du site de fixation à l'ATP, je parlerai

[68] J. Zheng et al., « 2.2 Å Refined Crystal Structure of the Catalytic Subunit of CAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor », *Acta Crystallographica Section D: Biological Crystallography* 49, nº 3 (1 mai 1993): 362-65, https://doi.org/10.1107/S0907444993000423.

uniquement du domaine kinase eucaryote et je prendrai comme référence la structure ayant pour code PDB 1ATP, co-cristallisée en conformation active avec l'ATP (Figure 14).

Le domaine kinase est constitué de 220 à 260 acides aminés répartis en deux lobes séparés par la région charnière (« hinge », en jaune). Le plus petit lobe, appelé N-terminal ou N-lobe comporte cinq brins β antiparallèles (β1 – β5) et une hélice α (hélice C ou αC, en violet). Le lobe C-terminal, ou C-lobe, est environ deux fois plus gros et comporte quatre brins β (β6 – β9) et six à huit hélices α selon les kinases. Au niveau de la région charnière, les lobes sont connectés via le brin β5 du lobe N-terminal et l'hélice αD du lobe C-terminal, créant une boucle conférant une flexibilité à la structure. Le premier résidu de la région charnière, en partant du brin β5, est nommé gatekeeper, et contrôle l'accès à deux poches hydrophobes adjacentes au site de fixation de l'adénine par la taille de sa chaîne latérale. Les autres éléments structuraux importants sont : la boucle P (aussi appelée boucle riche en glycines, en vert) située entre les brins β1 et β2, la boucle catalytique avec le motif HRD (His-Arg-Asp, en rouge) située sur les brins β6 et β7, le motif DFG (Asp-Phe-Gly, en bleu) entre les brins β8 et β9 et le motif conservé AxK (Ala – x - Lys) sur le brin β3. Enfin, la boucle d'activation (A-loop) est responsable en partie de l'état d'activation des protéines kinases via le motif DFG. Chaque kinase pouvant être activée ou désactivée, elles existent donc sous deux conformations majoritaires : forme active ou inactive.

Dans le site catalytique l'ATP se fixe au niveau de la région charnière située entre les deux lobes C et N. Sachant que le rôle de phosphorylation est le même chez toutes les protéines kinases, leur structure est encore plus conservée que leur séquence[69]. La base adénine du ligand est entourée de résidus hydrophobes (Val57, Val123, Leu173, Phe327) et se lie par liaison hydrogène avec les résidus de la région charnière : Met120, Glu121, Val123, Thr183. La partie ribose va quant à elle se lier, aussi par des liaisons hydrogène, avec des résidus du lobe C-terminal (Glu127, Glu170). Enfin les groupements phosphates α et β sont stabilisés grâce à un pont salin entre l'acide aminé Lys72 du motif AxK situé sur le brin β3, aussi appelé lysine catalytique, et le résidu Glu91 faisant partie de l'hélice αC. En plus de ce pont salin, la position d'un ion métallique, Mg^{2+} ou $Mn^{2+,}$ lié par l'Asp184 du motif DFG (en bleu) va aussi participer à cette stabilité. Le dernier groupement phosphate (γ) est ainsi positionné pour être transféré vers la cible. C'est l'Asp166 du motif HRD, localisé sur la boucle catalytique, qui va servir d'accepteur pour le transfert vers la protéine cible. Cette protéine cible, ou substrat, est liée à la kinase au niveau de l'hélice F du lobe C (non visible sur l'image), à partir du moment où la boucle d'activation est correctement positionnée. Une fois que la phosphorylation est effectuée, les produits de la catalyse, l'ADP et le substrat phosphorylé, sont libérés. Dans certains cas, les molécules d'eau présentes dans le site actif peuvent avoir un rôle important dans la réaction de phosphorylation[70].

[69] Kristoffer Illergård, David H. Ardell, et Arne Elofsson, « Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores », *Proteins* 77, nº 3 (15 novembre 2009): 499-508, https://doi.org/10.1002/prot.22458.

[70] Soreen Cyphers et al., « A Water-Mediated Allosteric Network Governs Activation of Aurora Kinase A », *Nature Chemical Biology* 13, nº 4 (avril 2017): 402-8, https://doi.org/10.1038/nchembio.2296.

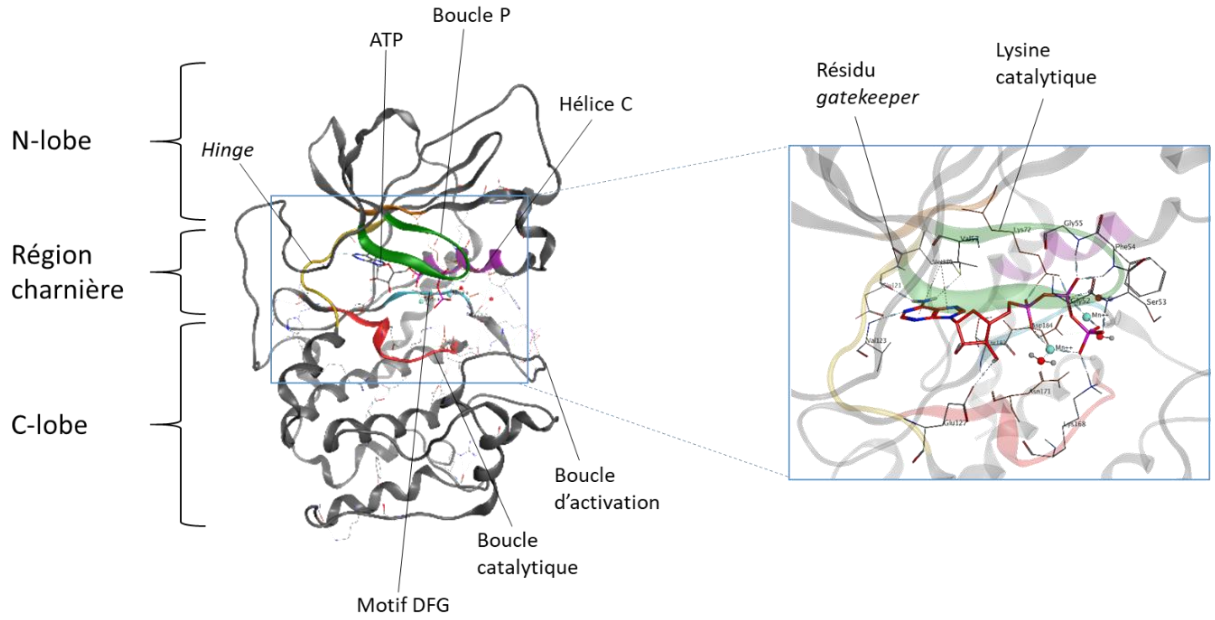


Figure 14 : Représentation en ruban du domaine kinase de PRKACA en conformation active. *A gauche, la structure complète avec les motifs conservés coloriés : boucle P (vert), région charnière (jaune), motif DFG (bleu), boucle catalytique (rouge), hélice C (violet). A droite, zoom sur le site catalytique présentant les interactions entre l'ATP et les résidus de la kinase. L'adénine entourée de résidus hydrophobes est maintenue en place par des liaisons hydrogène. Les groupements phosphates réalisent des interactions non covalentes avec les ions Mn^{2+} et les résidus du site actif. Code PDB : 1ATP.*

En plus des motifs structuraux décrits ci-dessus, d'autres éléments non visibles ont été découverts en 2008 par analyse de motifs spatiaux avec des outils bioinformatiques. Il s'agit de structures éphémères assemblées dynamiquement lorsque le domaine kinase adopte une conformation active, qui vont stabiliser la structure quaternaire des protéines kinases. Ces éléments sont appelés épines hydrophobes (« hydrophobic spines ») et catalytiques (« catalytic spines ») car ils sont composés de résidus hydrophobes discontinus dans la séquence protéique et alignés spatialement[71].

De par leurs différents rôles cellulaires fondamentaux, les kinases sont finement régulées. En règle générale, l'activation du domaine kinase est provoquée par sa phosphorylation. Dans mon exemple, PRKACA est phosphorylée sur la Thr197, mais cela reste un paramètre variable en fonction des différentes kinases. Dans le cas du récepteur à l'insuline InsR, l'activation passe par la phosphorylation de trois résidus Tyr présents sur la boucle d'activation[72]. Les sites de phosphorylation peuvent également être sur d'autres domaines comme c'est le cas pour EPHB2, phosphorylée sur son domaine juxta membranaire, ce qui permet de libérer la boucle d'activation[73]. Les domaines autres que kinases peuvent aussi participer à la régulation notamment selon leur position et leurs interactions avec le domaine kinase[74]. Ainsi, les kinases alternent régulièrement entre un état actif et un état inactif en fonction des besoins de la cellule. Il y a principalement deux changements majeurs dans la

[71] Alexandr P. Kornev, Susan S. Taylor, et Lynn F. Ten Eyck, « A Helix Scaffold for the Assembly of Active Protein Kinases », *Proceedings of the National Academy of Sciences of the United States of America* 105, nº 38 (23 septembre 2008): 14377-82, https://doi.org/10.1073/pnas.0807988105.
[72] Stevan R. Hubbard, « The Insulin Receptor: Both a Prototypical and Atypical Receptor Tyrosine Kinase », *Cold Spring Harbor Perspectives in Biology* 5, nº 3 (mars 2013), https://doi.org/10.1101/cshperspect.a008946.
[73] L. E. Wybenga-Groot et al., « Structural Basis for Autoinhibition of the Ephb2 Receptor Tyrosine Kinase by the Unphosphorylated Juxtamembrane Region », *Cell* 106, nº 6 (21 septembre 2001): 745-57.
[74] Bhushan Nagar et al., « Organization of the SH3-SH2 Unit in Active and Inactive Forms of the c-Abl Tyrosine Kinase », *Molecular Cell* 21, nº 6 (17 mars 2006): 787-98, https://doi.org/10.1016/j.molcel.2006.01.035.

conformation entre la forme active et inactive, au niveau du motif DFG et au niveau de l'hélice αC :

- Le motif DFG peut être orienté soit vers l'intérieur d'une poche adjacente au site de liaison, on parle dans ce cas de conformation « in », soit, suite à une rotation de 180° de son résidu Phe vers l'intérieur du site de liaison de l'ATP, on parle dans ce cas de conformation « out ». Ce dernier, ainsi que le substrat protéique, ne peuvent alors plus se fixer à la kinase. La Figure 15 montre les différentes possibilités de conformations du motif DFG.

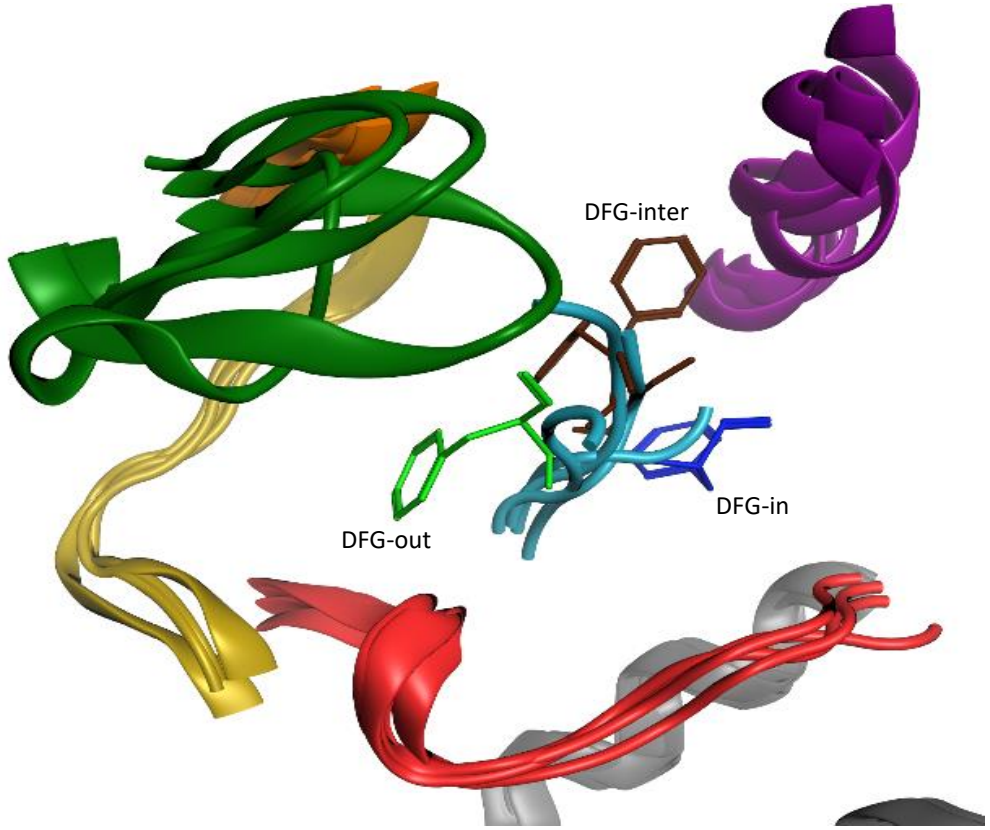


Figure 15 : Représentation des différentes conformations du résidu Phe du motif DFG chez les protéines kinases. *Le résidu Phe du motif DFG est coloré en fonction de son positionnement. En bleu : DFG-in (code PDB: 3OG7), en vert : DFG-out (code PDB : 2HYY) et en marron des exemples de positions intermédiaires notées DFG-inter (code PDB : 1MUO, 2J4Z).*

- Tout comme le motif DFG, l'hélice αC peut présenter deux conformations majeures « in » et « out » selon la position des résidus la composant et notamment selon que son résidu Glu forme ou non une interaction avec la lysine catalytique. La Figure 16 illustre les différentes possibilités de conformations de l'hélice αC.

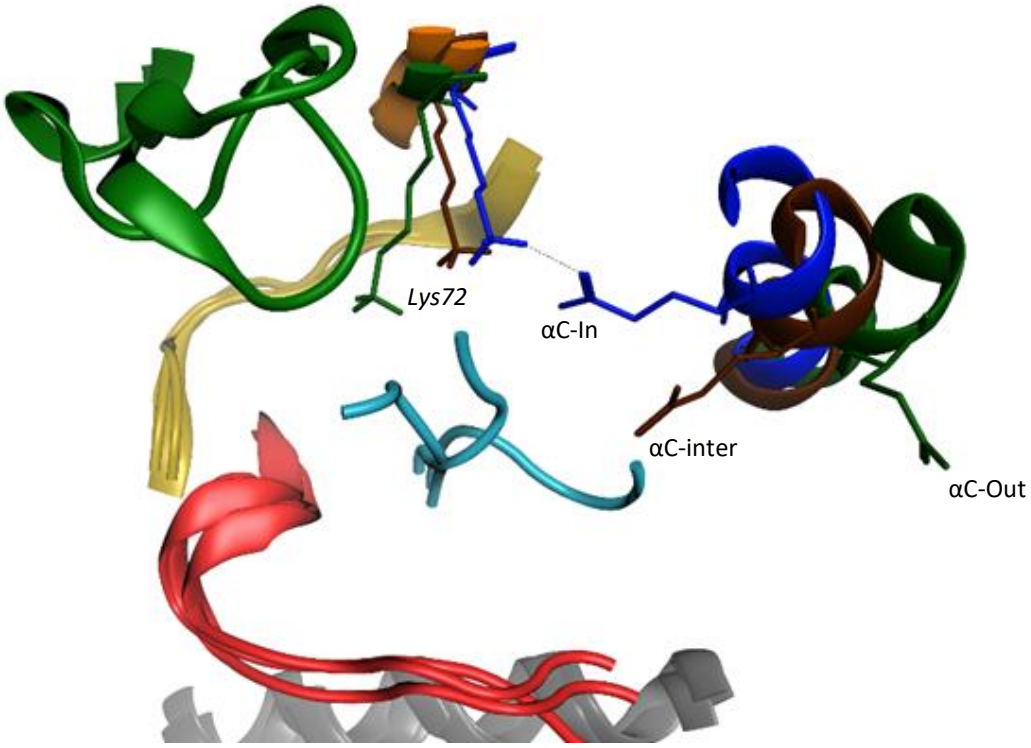


Figure 16 : Représentation des différentes conformations de l'hélice αC chez les protéines kinases. *En bleu : αC-in, le Glu 91 forme un pont salin avec la Lys catalytique 72 (code PDB : 1IEP). En vert : αC-out, le Glu 91 est orientée vers l'extérieur avec un déplacement de l'hélice αC (code PDB : 1HCL). En marron, une position intermédiaire αC-inter, observable notamment quand le Phe du motif DFG en position intermédiaire vient rompre le pont salin entre la Lys et le Glu (code PDB : 2J4Z).*

Dans la conformation active des kinases, le DFG et l'hélice αC sont tous deux en conformations « in », laissant ainsi la place à l'ATP et au substrat pour se fixer à la kinase. On pourrait en déduire que la conformation inactive est donc DFG-out, αC-out. Cela est vrai, cependant il existe aussi d'autres combinaisons possibles pour la conformation inactive : DFG-in, αC-out ou DFG-out, αC-in. Cette diversité de conformations a été observée au fil de l'accumulation des structures tridimensionnelles. Il s'est également avéré qu'il existe toute une variété de positions intermédiaires pour ces motifs structuraux, notamment dû à l'élasticité de la structure bilobée du domaine kinase[75,76]. En analysant les structures tridimensionnelles de protéines kinases présentes dans la PDB, on observe que celles-ci présentent en grande majorité une conformation DFG-in (Tableau 1). De plus, si l'on regarde plus en détail, on observe que dans le cas où la structure se présente en DFG-in, l'hélice αC est elle aussi le plus souvent en conformation in. Chez les structures en conformation DFG-out, l'hélice αC est généralement aussi en conformation out. Et enfin, quand la conformation du DFG est indéterminée (inter), alors l'hélice αC est elle aussi le plus souvent indéterminée.

[75] Morgan Huse et John Kuriyan, « The Conformational Plasticity of Protein Kinases », *Cell* 109, nº 3 (3 mai 2002): 275-82, https://doi.org/10.1016/S0092-8674(02)00741-9.
[76] Alessio Atzori et al., « Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics », *Journal of Chemical Information and Modeling* 54, nº 10 (27 octobre 2014): 2764-75, https://doi.org/10.1021/ci5003334.

Tableau 1 : Distribution des différentes conformations de structures tridimensionnelles de protéines kinases chez l'homme et la souris. *D'après les données de classification fournies par la base de données KLIFS*[77].

	DFG-in	DFG-out	DFG-inter	Total
αC-in	2921 (61 %)	9 (0 %)	50 (1 %)	2980 (62 %)
αC-out	748 (16 %)	352 (7 %)	97 (2 %)	1197 (25 %)
αC-inter	395 (8 %)	31 (1 %)	180 (4 %)	606 (13 %)
Total	4064 (85 %)	392 (8 %)	327 (7 %)	4783 (100 %)

1.6 Les inhibiteurs de kinase

1.6.1 Historique et premiers succès

Les entreprises pharmaceutiques ont commencé à s'intéresser aux protéines kinases à partir de la fin des années 1970 avec la découverte du premier oncogène qui se trouvait être une protéine kinase[78]. En 1981, les recherches sur les esters de phorbols connus comme étant des promoteurs tumoraux montrent que ceux-ci vont activer la protéine kinase C (PKC)[79], renforçant le rôle des kinases dans certains cancers. Cependant, l'inhibition des kinases présentait alors quelques difficultés. D'abord la forte concentration en ATP dans la cellule (de 1 à 5 mM) ne facilite pas l'inhibition compétitive[80], mais surtout, comme on l'a vu précédemment la grande similarité de séquence et des résidus composant leurs sites actifs rend très difficile la sélectivité d'un potentiel médicament et augmente donc les risques d'effets secondaires et de toxicité. Les premières molécules avec une activité inhibitrice de kinases sont des isoquinolinesulfonamides[81] (Figure 17, gauche). Peu après il a été montré que la staurosporine (Figure 17, milieu) était capable d'inhiber la PKC à des concentrations en nM[82]. Dès lors, remarquant que finalement on pouvait développer des inhibiteurs compétitifs, les entreprises pharmaceutiques se sont lancées dans la recherche active de telles molécules, s'inspirant beaucoup des premiers inhibiteurs. S'ensuit une liste d'évènements (cf Figure) jusqu'à l'avènement principal de la recherche d'inhibiteurs de kinase, en 2001, avec l'obtention d'une AMM, par la FDA, pour l'imanitib, développé par Novartis contre la leucémie myéloïde chronique (LMC). Le succès de l'imatinib (Figure 17, droite), classé comme blockbuster, a contribué au rayonnement international de Novartis.

[77] Albert J. Kooistra et al., « KLIFS: A Structural Kinase-Ligand Interaction Database », *Nucleic Acids Research* 44, nº D1 (4 janvier 2016): D365-71, https://doi.org/10.1093/nar/gkv1082.
[78] M. S. Collett et R. L. Erikson, « Protein Kinase Activity Associated with the Avian Sarcoma Virus Src Gene Product », *Proceedings of the National Academy of Sciences of the United States of America* 75, nº 4 (avril 1978): 2021-24, https://doi.org/10.1073/pnas.75.4.2021.
[79] M. Castagna et al., « Direct Activation of Calcium-Activated, Phospholipid-Dependent Protein Kinase by Tumor-Promoting Phorbol Esters », *The Journal of Biological Chemistry* 257, nº 13 (10 juillet 1982): 7847-51.
[80] Zachary A. Knight et Kevan M. Shokat, « Features of Selective Kinase Inhibitors », *Chemistry & Biology* 12, nº 6 (1 juin 2005): 621-37, https://doi.org/10.1016/j.chembiol.2005.04.011.
[81] Hiroyoshi Hidaka et al., « Isoquinolinesulfonamides, novel and potent inhibitors of cyclic nucleotide-dependent protein kinase and protein kinase C », *Biochemistry* 23, nº 21 (9 octobre 1984): 5036-41, https://doi.org/10.1021/bi00316a032.
[82] Tatsuya Tamaoki et al., « Staurosporine, a potent inhibitor of phospholipidCa++dependent protein kinase », *Biochemical and Biophysical Research Communications* 135, nº 2 (13 mars 1986): 397-402, https://doi.org/10.1016/0006-291X(86)90008-2.

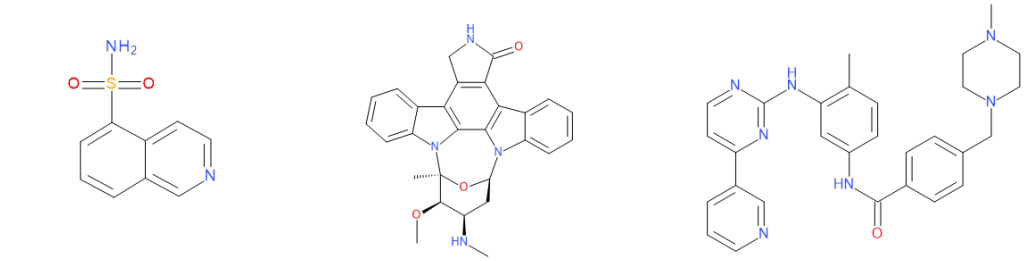


Figure 17 : Représentation des premiers inhibiteurs de kinases.

Depuis cette date, chaque année au moins un inhibiteur de kinase est approuvé (excepté en 2016), on les retrouve tous dans la figure 1 de l'article concluant ce chapitre.

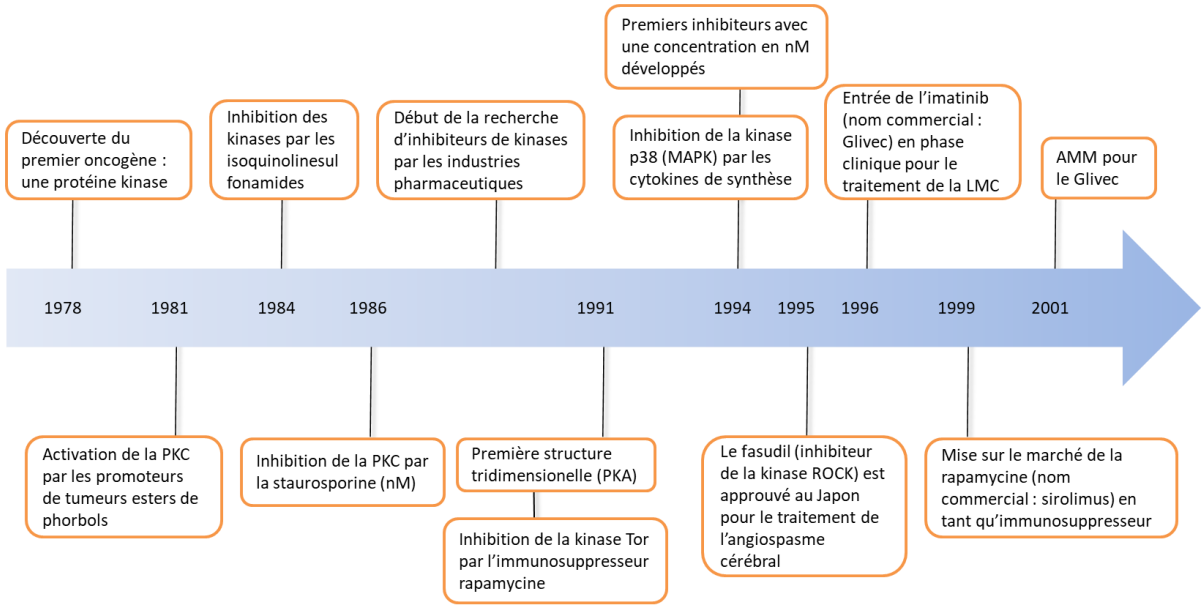


Figure 18 : Chronologie récapitulative des différents évènements dans le développement des inhibiteurs de kinase jusqu'à la mise sur le marché de l'imatinib en 2001. *Adapté de P. Cohen*[83].

1.6.2 Les différentes catégories d'inhibiteurs de kinase

Les structures cristallographiques de protéines kinases en complexe avec un inhibiteur ont permis de classer ceux-ci en différentes catégories selon leur mode d'interaction[84]. On trouve tout d'abord trois types d'inhibiteurs compétitifs (de l'ATP) non-covalents (types I, I½ et II). Puis deux types allostériques (types III et IV). Certains inhibiteurs peuvent se lier à la fois au site actif et à un autre domaine de la kinase (type V). Et enfin, des inhibiteurs covalents (type VI). La spécificité de chaque type est précisée ci-dessous :

- **Type I :** ils se lient au niveau du site de fixation de l'ATP, et notamment de l'adénine en formant des liaisons hydrogène avec la région charnière. Ils se fixent en conformation active DFG-in.

[83] Philip Cohen, « Protein Kinases — the Major Drug Targets of the Twenty-First Century? », *Nature Reviews Drug Discovery* 1, nº 4 (avril 2002): 309-15, https://doi.org/10.1038/nrd773.
[84] Robert Roskoski, « Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes », *Pharmacological Research* 103 (1 janvier 2016): 26-48, https://doi.org/10.1016/j.phrs.2015.10.021.

- **Type I½ :** ils se fixent également en conformation active DFG-in. De plus, ils ont accès à une poche hydrophobe additionnelle adjacente à celle formée par les résidus hydrophobes du brin β2 et située derrière le site de liaison de l'adénine.

- **Type II :** ils se fixent en conformation inactive DFG-out. Le ligand est ainsi stabilisé par des liaisons hydrogène avec la région charnière, la poche de l'ATP et la poche allostérique juxtaposée[85].

- **Type III :** ils se lient au site allostérique adjacent au site de l'ATP en conformation DFG-in ou out avec ou non présence de l'ATP.

- **Type IV :** ils se fixent dans une poche allostérique éloignée du site catalytique, ce qui leur vaut l'appellation de « true allosteric inhibitors ». La localisation de cette poche dépend des protéines kinases et elle peut aussi bien se situer sur le lobe N-terminal que sur le lobe C-terminal du domaine kinase.

- **Type V :** ils ont été ajoutés récemment et sont un mélange des caractéristiques des inhibiteurs de type I à IV, ils peuvent donc se lier à la fois au site de liaison de l'ATP et à un site allostérique, ce qui leur vaut l'appellation d'inhibiteurs bivalents.

- **Type VI :** ils peuvent être de n'importe quel type de ceux précisés ci-dessus, à la différence qu'ils portent un groupement capable de former une liaison covalente, avec un acide aminé nucléophile au sein du site actif (Cys, Lys ou Tyr), créant une liaison irréversible.

La classification des inhibiteurs de protéines kinases nécessite l'obtention d'une structure tridimensionnelle ainsi, certains inhibiteurs approuvés ne sont pas encore classés et leur type d'interactions dans le site actif n'est pas encore totalement caractérisé. De plus, cette classification peut aussi révéler des surprises, comme pour l'imatinib qui peut aussi se lier comme un inhibiteur de type I[86]. On retrouve cette configuration dans une structure de la kinase SYK, co-cristallisée avec l'imatinib (code PDB : 1XBB) en conformation DFG-in.

1.6.3 Bilan et perspectives

Au 1ier juin 2019, il y a 54 inhibiteurs de protéines kinases approuvés dans le monde et au moins 150 en phases d'essais cliniques (http://www.icoa.fr/pkidb/)[87]. La majorité sont de types I, I1/2 et II, soit des compétiteurs du substrat initial : l'ATP. Malheureusement, ce type d'inhibiteurs est plus enclin à favoriser l'apparition de résistance, dues aux mutations des kinases ciblées, chez les patients traités. En effet, ces mutations empêchent l'inhibiteur d'accéder au site actif[88]. Par exemple, chez la protéine kinase ABL1, la mutation du résidu 315,

[85] Daniel Mucs, Richard A. Bryce, et Pascal Bonnet, « Application of Shape-Based and Pharmacophore-Based in Silico Screens for Identification of Type II Protein Kinase Inhibitors », *Journal of Computer-Aided Molecular Design* 25, nº 6 (1 juin 2011): 569-81, https://doi.org/10.1007/s10822-011-9442-0.
[86] Shane Atwell et al., « A Novel Mode of Gleevec Binding Is Revealed by the Structure of Spleen Tyrosine Kinase », *The Journal of Biological Chemistry* 279, nº 53 (31 décembre 2004): 55827-32, https://doi.org/10.1074/jbc.M409792200.
[87] Fabrice Carles et al., « PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials », *Molecules* 23, nº 4 (15 avril 2018): 908, https://doi.org/10.3390/molecules23040908.
[88] Doriano Fabbro, « 25 Years of Small Molecular Weight Kinase Inhibitors: Potentials and Limitations », *Molecular Pharmacology* 87, nº 5 (1 mai 2015): 766-75, https://doi.org/10.1124/mol.114.095489.

une thréonine, en une isoleucine (mutation T315I) bloque l'accès de l'imatinib à la poche hydrophobe allostérique[89]. En effet, il s'agit d'une mutation sur un résidu clé, le gatekeeper, rendant l'action de l'imatinib caduque. L'autre difficulté avec ce type d'inhibiteurs est d'obtenir une bonne sélectivité. En ciblant le site actif de l'ATP très conservé parmi toutes les kinases, il est très compliqué d'obtenir des composés sélectifs. Enfin, de par l'étude massive sur cette famille de protéines, l'espace chimique disponible est devenu très réduit et obtenir des nouveaux chémotypes libres d'exploitation s'avère désormais très ardu.

Tout n'est pas pour autant noir. Les problèmes de résistance peuvent être contournés en développant de nouvelles molécules ciblant spécifiquement les protéines mutées. Ainsi, le ponatinib va pouvoir inhiber la forme mutée de la protéine ABL1 (T315I). On peut aussi citer l'abivertinib, un composé en phase clinique 3, qui inhibe quant à lui la forme mutée d'EGFR (T790M)[90]. Dans ce cas, la mutation permet même une meilleure sélectivité car les inhibiteurs n'auront pas d'effet sur la forme non mutée de la protéine. Le développement d'inhibiteurs allostériques permet aussi d'obtenir une meilleure sélectivité, les poches secondaires étant beaucoup moins conservées que le site de liaison de l'ATP. De plus, l'inhibition par allostérie permet d'ouvrir la voie au ciblage des protéines kinases atypiques et à un espace chimique beaucoup plus ouvert et libre d'exploitation, comme ce fut le cas pour les inhibiteurs de MEK[91]. A l'heure actuelle les inhibiteurs de kinases approuvés ou en phase clinique ne couvrent qu'environ 10-15 % de tout le kinome, il y a encore beaucoup de recherches et de possibilités dans ce domaine. La plupart des inhibiteurs de kinase mis sur le marché deviennent rapidement des blockbusters et rapportent des milliards de dollars par an aux entreprises pharmaceutiques.

1.7 Etude sur les squelettes moléculaires des inhibiteurs de kinase

Un premier travail de caractérisation et de recensement de tous les inhibiteurs de kinase a déjà été réalisé par l'équipe[92]. L'article concluant ce chapitre est la prolongation de ce travail et porte cette fois sur l'étude des squelettes moléculaires de ces inhibiteurs, notamment sur la différence entre ceux approuvés ou en phase clinique par rapport à ceux que l'on peut trouver dans les bases de données publiques.

[89] Mercedes E. Gorre et al., « Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification », *Science* 293, nº 5531 (3 août 2001): 876-80, https://doi.org/10.1126/science.1062538.
[90] Qing Zhou et al., « Safety and efficacy of abivertinib (AC0010), a third-generation EGFR tyrosine kinase inhibitor, in Chinese patients with EGFR-T790M positive non-small cell lung cancer (NCSLC). », *Journal of Clinical Oncology* 37, nº 15_suppl (20 mai 2019): 9091-9091, https://doi.org/10.1200/JCO.2019.37.15_suppl.9091.
[91] Zheng Zhao, Lei Xie, et Philip E. Bourne, « Insights into the binding mode of MEK type-III inhibitors. A step towards discovering and designing allosteric kinase inhibitors across the human kinome », *PLoS ONE* 12, nº 6 (19 juin 2017), https://doi.org/10.1371/journal.pone.0179936.
[92] Fabrice Carles et al., « PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials », *Molecules* 23, nº 4 (15 avril 2018): 908, https://doi.org/10.3390/molecules23040908.

Article

Comparative assessment of protein kinase inhibitors in public databases and in clinical trials

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
2 Janssen-Cilag, Centre de Recherche Pharma, CS10615 - Chaussée du Vexin, 27106 Val-de-Reuil, France

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

Academic Editor:
Received: date; Accepted: date; Published: date

Abstract: Since 2001 when the first protein kinase inhibitor (PKI) was FDA-approved, 54 PKIs reached the market. Protein kinases are still major drug targets in the pharmaceutical industries with many PKIs currently in clinical trials. In a previous study, we presented PKIDB, a freely available protein kinase inhibitor database, gathering PKIs already approved and in clinical trials (from phase 0 to 4). This database is frequently updated and new improvements are presented here. In this study, we focus on the comparison between PKIs in clinical trials from PKIDB and PKIs in early preclinical studies from ChEMBL, the largest publically available database. For each dataset, the distribution of the classical physicochemical descriptors is presented. From these results, updated guidelines to prioritize compounds for targeting protein kinases are proposed. Results of a statistical method show that the PKIDB dataset is perfectly encompassed within all PKIs found in the public database. This observation is reinforced by a Principal Moments of Inertia (PMI) analysis of all molecules. Moreover, 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 the cell cycle, particularly in regulation mechanisms. This process, which consists of the transfer of a phosphoryl group PO_3^{2-} to the target substrate, is catalyzed by the family of protein kinases. Protein kinases are a large family with 518 or 538 members, including atypical kinases, encoded by human genome [1–3]. Numerous studies showed that their deregulation and their mutations are responsible of various cancers [4] but also of other diseases in immune or neurological area [5,6]. However, a majority of protein kinases have not been fully explored [7] and there is still a high potential of innovation for targeting the protein kinome to treat cancer. Nowadays, FDA has approved 49 small molecule protein kinase inhibitors (PKIs), to which we can add anlotinib, apatinib, icotinib and fasudil approved by the Chinese regulatory authorities and tivozanib approved in Europe (Figure). It is also important to mention the macrocyclic lactones such as sirolimus or temsirolimus and kinase-targeted antibodies such as cetuximab or trastuzumab approved against colorectal, head/neck and breast cancers respectively [8–10]. These large molecules were not taken into account in this study, we focused only on small molecule protein kinase inhibitors (PKIs) targeting the kinase domain. The first PKI approved by the Food and Drug

Administration (FDA) was imatinib, in 2001, containing a pyridyl-anilinopyrimidine scaffold. Imatinib targets the inactive conformation of ABL1 kinase and is used against chronic myelogenous leukemia (CML) [11]. Since then, one or more PKIs reach the market almost every year, with a significant increase from 2011 except in 2016 when no PKI was approved. In order to stay up to date on the status of PKIs landscape, we developed PKIDB, which gathers data on PKIs currently in development or already approved (phase 0 to 4) and having an International Nonproprietary Name (INN) [12]. We collect useful information on each compound and provide links to different external databases such as ChEMBL [13], PDB [14], PubChem [15], etc. [16]. For each molecule, the type of binding mode specified in PKIDB has been manually classified according to Roskoski's review [17]. The database is freely accessible at a dedicated website (http://www.icoa.fr/pkidb). As of 24thof April 2019, it contains 212 inhibitors, 54 approved and 158 in different stages of clinical trials (from phase 0 to phase 3).

In this study, we compared PKIDB to a large dataset of PKIs retrieved from ChEMBL [13]. Firstly, we performed a Principal Component Analysis (PCA) using standard physicochemical descriptors and compared the projected space of the two datasets onto the loading plot. We also analyzed the structural shape diversity of PKIs using the Principle Moments of Inertia (PMI). Secondly, besides these comparisons based on the molecular structure, we performed a substructure analysis on all PKI scaffolds. Molecular scaffolds represent the main core of a chemical series; they are relevant information for the medicinal and/or computational chemists. Indeed, they are used to explore structure–activity relationships (SAR) information [18] which consists in substituting chemical groups on the scaffolds to improve a hit on a protein target. The concept of scaffold was first defined by Bemis and Murcko as frameworks which consist of rings and linkers, thus substituents are removed [19]. From these scaffolds, different levels of abstraction were applied: 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 single bond, reducing the initial molecule to a simple graph [20]. Finally, the rings are obtained by removing bonds between rings.

The balance between the molecular diversity of scaffolds and their frequency is an important feature in a chemical database. A high frequency associated to a small number of scaffolds corresponds to a focused library with structurally similar molecules bearing different substituents. On the opposite, a low frequency associated to a great number of scaffolds reflects a high molecular diversity [16]. Thus, this criterion needs to be addressed when designing or choosing a chemical library depending on its use. We assessed scaffold diversity for the two datasets, using the molecular Bemis and Murcko scaffolds and cyclic skeleton. The most represented ones and the distribution difference between the two studied datasets are presented. Finally, an analysis of the rings of all molecules was performed. We first considered all the rings without their substituents where each first attached atom was replaced by an hydrogen atom. Then, we encoded the rings by considering the position and atom type of their 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.

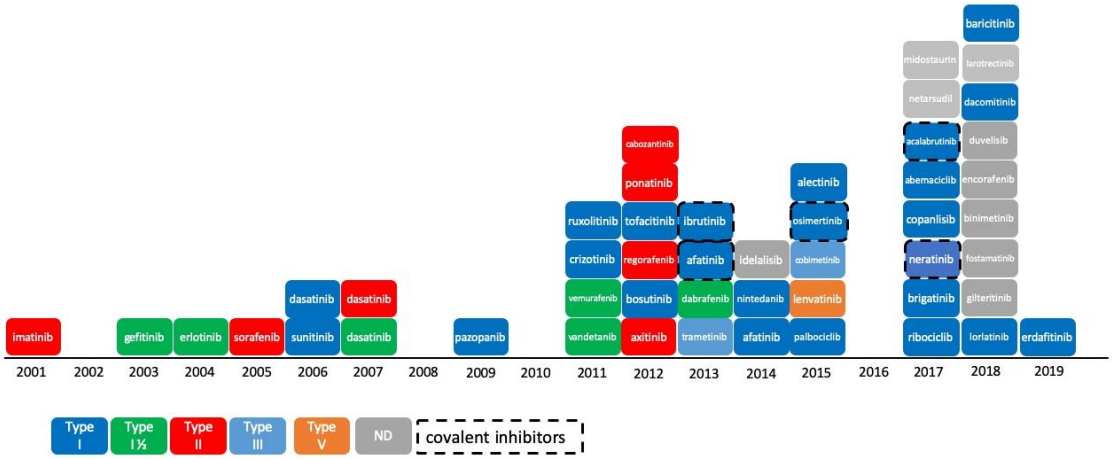


Figure 1. Progression of FDA-approved protein kinase inhibitors (2001-2019) and their type of inhibition. As of 24th April 2019, there are 49 FDA-approved kinase inhibitors. Not shown here: tivozanib approved by EMA (European Medicines Agency) in 2017, anlotinib, apatinib and icotinib approved by CFDA (China Food and Drug Administration) in 2018, 2014 and 2011 respectively and fasudil approved in China and in Japan in 1995.

2. Results

2.1. Update on PKIDB

The content and the description of PKIDB were provided in a previous study by Carles *et al.* [16]. Referencing 212 molecules at the end of April 2019, PKIDB contains 32 more inhibitors (from phase 0 to phase 4) than the first release (abivertinib, adavosertib, alvocidib, asciminib, avapritinib, avitinib, bemcentinib, berzosertib, bimiralisib, capivasertib, dezapelisib, enzastaurin, fasudil, leniolisib, mavelertinib, midostaurin, nazartinib, neflamapimod, nemiralisib, netarsudil, ningetinib, parsaclisib, ravoxertinib, ripasudil, rogaratinib, ruboxistaurin, sotrastaurin, tomivosertib, umbralisib, vactosertib, verosudil, zanubrutinib).

Among those 32 compounds two were FDA-approved in 2017: netarsudil and midostaurin. 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 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 [21]). However, we also kept the stem 'cidib' to retrieve information on alvocidib, not yet referenced as alvociclib.

Besides those compounds, Table 1 gathers the 8 PKIs that reached phase 4 and were FDA-approved in 2018 and erdafitinib, a FGFR kinase inhibitor, approved in 2019. Those 9 PKIs were previously in a phase lower than 4 in our database, excepted baricitinib since it was approved by EMA in 2017. One should note that FDA recently approved alpelisib, a PI3Kα inhibitor, after the updated version of PKIDB and so not considered in this study.

This brings to 54 the total number of approved drugs on the market referenced in our database. As described in PKIDB, most of the PKIs are targeting more than one protein kinases and since the first version of PKIDB, new targets emerged such as the Wee1-like protein kinase inhibited by adavosertib.

Table 1. PKIs approved in 2018 and 2019 with their respective targets extracted from DrugBank

PKI	Unitprot ID	Gene name
Binimetinib	Q02750	MAP2K1
Dacomitinib	P00533	EGFR
Duvelisib	O00329	PI3KCD
	P48736	PI3KCG
Encorafenib	P15056	BRAF
Fostamatinib	P43405	SYK
Gilteritinib	P36888	FLT3
	P30530	AXL
	Q9UM73	ALK
Larotractinib	P04629	NTRK1
	Q16620	NTRK2
	Q16288	NTRK3
Lorlatinib	Q9UM73	ALK
	P08922	ROS1
Erdafitinib	P11362	FGFR1
Alpelisib*	P42336	PI3Kα

*Alpelisib was approved after the latest release. Uuniprot ID extracted from https://www.uniprot.org/.

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 [22,23]. During the search of a lead compound in a virtual or experimental screening campaign, such descriptors may serve as a filter to discard molecules and therefore decrease the size of the chemical library. The distribution of these descriptors calculated from inhibitors extracted from PKIDB is shown in Figure 2.

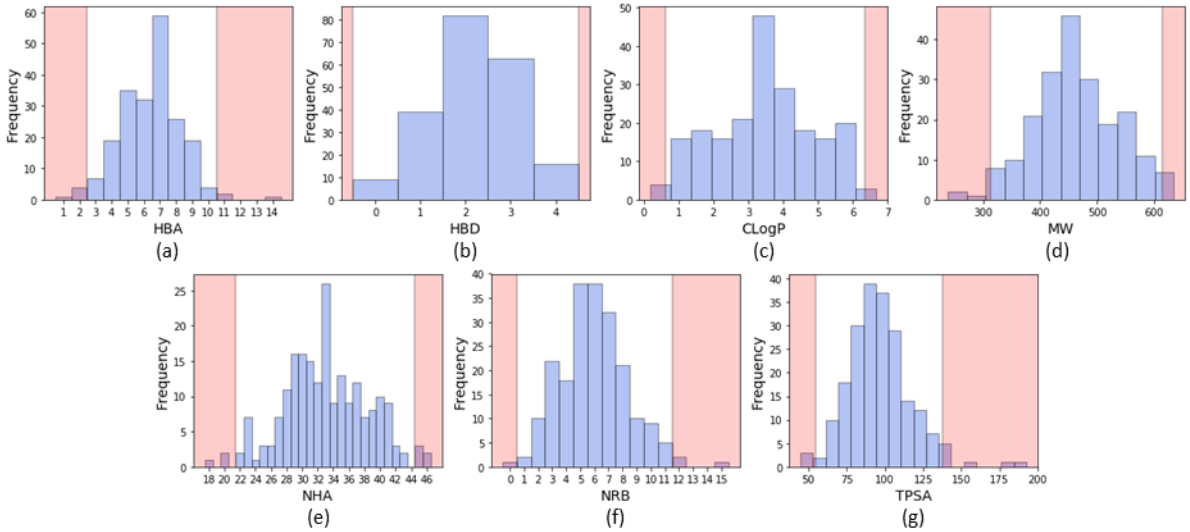


Figure 2. Distribution of physicochemical properties of PKIs: (a) Number of hydrogen bond acceptors (HBA); (b) Number of hydrogen bond donors (HBD); (c) Number of rotatable bonds (NRB); (d) Number of heavy atoms (NHA); (e) Molecular weight (MW); (f) ClogP (RDKit); (g) Topological polar surface area (TPSA). Pink areas represent values outside two standard deviations from the mean (95.4% confidence interval).

In a previous study [16], we analyzed the 'rule of five' descriptors detailed by Lipinski [22] for inhibitors in clinical trials and approved. Here, we updated the statistical analysis with new PKIs included in PKIDB and we compared them to PKIs included in ChEMBL (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	29/53 (55%)	17/53 (32%)	7/53 (13%)	0/53 (0%)
PKIs in clinical trials	99/156 (64%)	42/156 (27%)	15/156 (8%)	0/156 (0%)
All PKIs	128/209 (61%)	59/209 (28%)	22/209 (11%)	0/209 (0%)
PKIs ChEMBL	51,822/76,504 (68%)	18,613/76,504 (24%)	5,900/76,504 (8%)	169/76,504 (0.2%)

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

We found that 61% 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 [24]. Compared to the initial study, we removed the counter ion during the standardisation 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 209 in PKIDB) we reveal 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 clear difference between any kinase subsets for 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 rule.

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	18/53 (34%)	11/53 (21%)	2/53 (4%)	0/53 (0%)	2/53 (4%)	2/53 (4%)
PKIs in clinical trials	45/156 (29%)	26/156 (17%)	1/156 (1%)	0/156 (0%)	4/156 (3%)	6/156 (4%)
All PKIs	63/209 (30%)	37/209 (18%)	3/209 (1%)	0/209 (0%)	6/209 (3%)	8/209 (4%)
PKIs ChEMBL	18,892/76,504 (25%)	10,897/76,504 (14%)	924/76,504 (1%)	208/76,504 (0%)	3695/76,504 (5%)	2,051/76,504 (3%)

[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.

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

- A molecular weight (MW) between 312 and 614 Da (average of 463.3 Da)
- A ClogP (calculated with RDKit) between 0.6 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 96.5 Å^2)
- Between 1 and 11 rotatable bonds (NRB) (average of 6.0)
- Number of aromatic rings (NAR) between 1 and 5 (average of 3.5)
- 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 11 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).

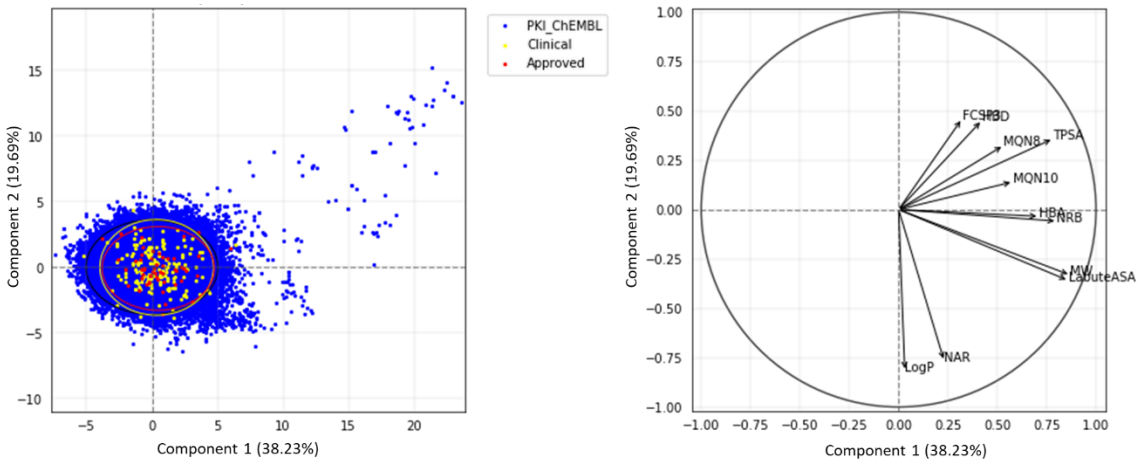


Figure 3. (a) PCA of PKIs from ChEMBL and PKIDB containing 76,504 and 209 compounds respectively. Black, yellow and red ellipses encompass 95% of the individuals from class "PKI_ChEMBL", "Clinical_PKI" and "Approved_PKI" respectively; **(b)** Correlation circle.

The two first principal components explain 38.2% and 20.0 % of the total variance respectively. PC3, not shown here, encompasses 12.2%. Thus, the 2D scatterplot illustrated here represents around 58% of the variance, an acceptable value avoiding an important loss of information.

Each dot on the graph (Figure 3a) represents a molecule. Few compounds from PKI_ChEMBL are projected in the upper right quadrant but none from PKIDB. Most of the compounds from PKIDB 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 close near 90° indicate 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 MW, LabuteASA, NRB and TPSA. These four variables contribute to PC1 at 17.6%, 18.0%, 15.3% and 14.5% respectively. The variables CLogP and NAR do not contribute to this axis and are negatively correlated with the second factorial axis (PC2). Their contribution of ClogP and NAR to PC2 are 30.8% and 27.2% respectively. To a lesser extent, this axis is also positively correlated with FCSP3 and HBD (contributions of 9.7% and 9.3% respectively). A weak angle between NAR and CLogP vectors is consistent with the fact that CLogP increases with the number of aromatic rings. In the same way, LabuteASA and MW are also strongly correlated.

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. Regarding 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 so explaining their position on the PCA map. Besides these few outliers, PKIDB remains very well encompassed within all PKIs found in the ChEMBL as shown in the figure with the ellipses surrounding 95% of individuals per category.

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 [25]. 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 [26].

The PMI plot (Figure 4) reveals a great majority of kinase inhibitors 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. Thus, type-II PKIs are much likely to be found close to the rod edge since they require an extended conformation for their binding [27]. The three molecules from PKIDB closest to the extreme vertices are mubritinib near the rod, mavelertinib near the disc and galunisertib 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 explore a new molecular space in PKIs. As shown in the Figure 4, the space covered by the PKI_ChEMBL dataset is much wider. The distribution of these PKI is concentrated to elongated and circular shapes. We also observe some compounds from PKI_ChEMBL 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.

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

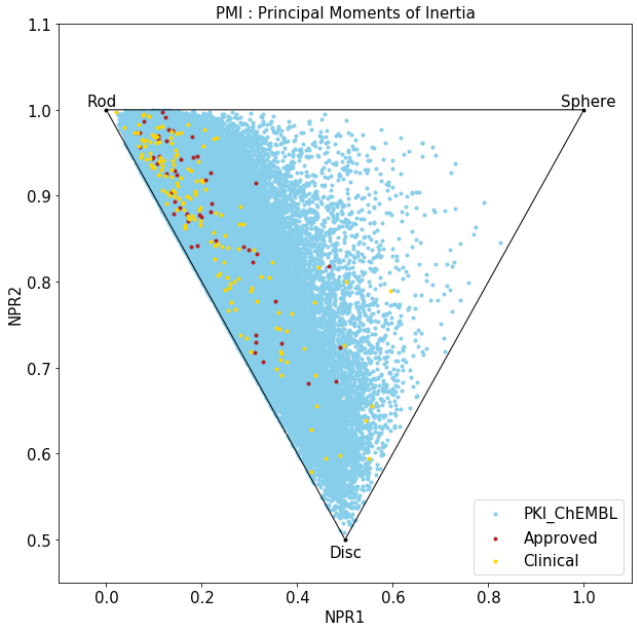


Figure 4. Principal Moments of Inertia (PMI) plot of PKIs in clinical trials (yellow), approved (red) and from ChEMBL database (light blue).

2.3. Scaffold diversity assessment

2.3.1. Generation of molecular scaffolds

To have a better 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 searched for the presence of macrocyclic molecules (rings > 12 atoms). In PKIDB, there are four macrocycles. Two of them are approved drugs: icotinib (CFDA approved) 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%). It is important to note that we excluded from PKIDB macrocycles containing the stems '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 [28].

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 graph framework issued from BM. As a reminder, Bemis and Murcko scaffold corresponds to the core of a molecule after removing side chains [19]. The graph framework corresponds to BM scaffold where each heteroatom was substituted by a carbon and each multiple bond by a single one. Therefore, such frameworks cover topologically equivalent BM scaffolds differentiated by heteroatom substitutions and bond types.

In PKIDB dataset, among 209 molecules, 198 present a unique BM scaffold and 186 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. In other words, in PKIDB almost each compound has its own scaffold (very high scaffold diversity). The molecular similarity mean, calculated with MACCS keys on all molecules indicates that both datasets are diverse with mean of Tanimoto similarity of about 0.5 (Table 4). However, in the ChEMBL dataset, the scaffold diversity is much lower with about a BM scaffold for about 2.7 molecules in average. Regarding the graph frameworks, they number tends to decrease compared to BM scaffolds: 1 GF for 1.1 and 5.7 molecules in PKIDB and PKI_ChEMBL respectively.

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.

Then, for each unique BM scaffold in PKIDB, we checked how many PKIs are obtained in PKI_ChEMBL. From the 198 unique BM scaffolds available in PKIDB, only 97 are present in PKI_ChEMBL which represent 2,402 molecules out of a total of 76,504 (3.1%). 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 7,597 compounds (10%) in PKI_ChEMBL.

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	209	4 (1.9%)	198 (94.7%)	186 (86.0%)	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.

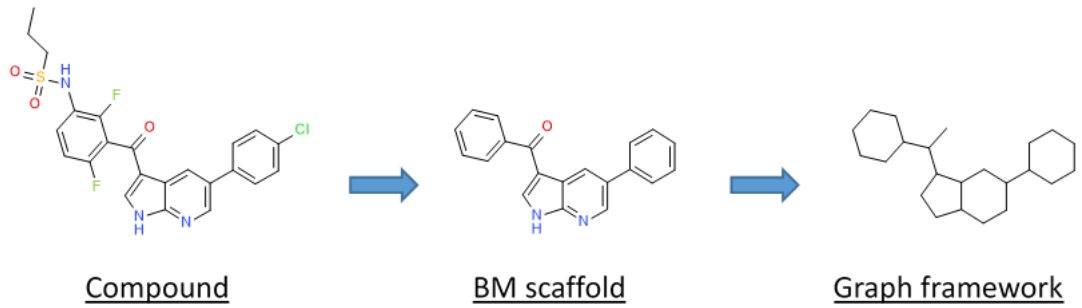


Figure 5. Representation of a molecular decomposition into scaffolds according to Bemis and Murcko (BM) and in graph framework.

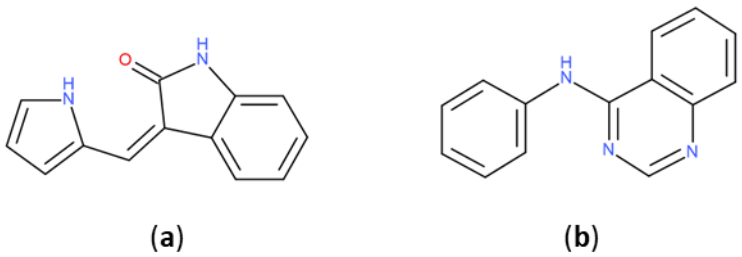


Figure 6. Most retrieved Bemis and Murcko scaffolds in PKIDB dataset (a): (3Z)-3-(1H-pyrrol-2-ylmethylene)indolin-2-one and in PKI_ChEMBL dataset (b): N-phenylquinazolin-4-amine.

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 [29]. 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 [30].

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 53 unique bicyclic scaffolds among the total of 168. More surprising, 28 out of these 53 bicycle 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 intensively studied [31]. Example of PKIs containing quinazoline scaffold are gefitinib, lapatinib, erlotinib, afatinib and more recently canertinib. Kinase inhibitors bearing this scaffold have mainly been designed to target EGFR. The second most represented bicyclic scaffold is the quinolone, another two-fused six-membered aromatic ring. It is worth noting that depending on the choice of the tautomeric form or the attached substituents, RDKit may have some issues in finding the aromaticity in the bicyclic scaffold and could return the indoline scaffold instead of the indole, 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. Curiously, unlike BM scaffolds where more than half scaffolds from PKIDB were not retrieved in PKI_ChEMBL, here only three bicycles (not shown) are not found in 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 3-substituted (4,6,7) 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-tris-quinoline in PKIDB and this is the third most represented scaffold in PKI_ChEMBL. Here again, the substituent in 7 position is always an ether. Interestingly, the second most retrieved substituted bicycles in PKI_ChEMBL is not found in top tenth of PKIDB. As shown 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.

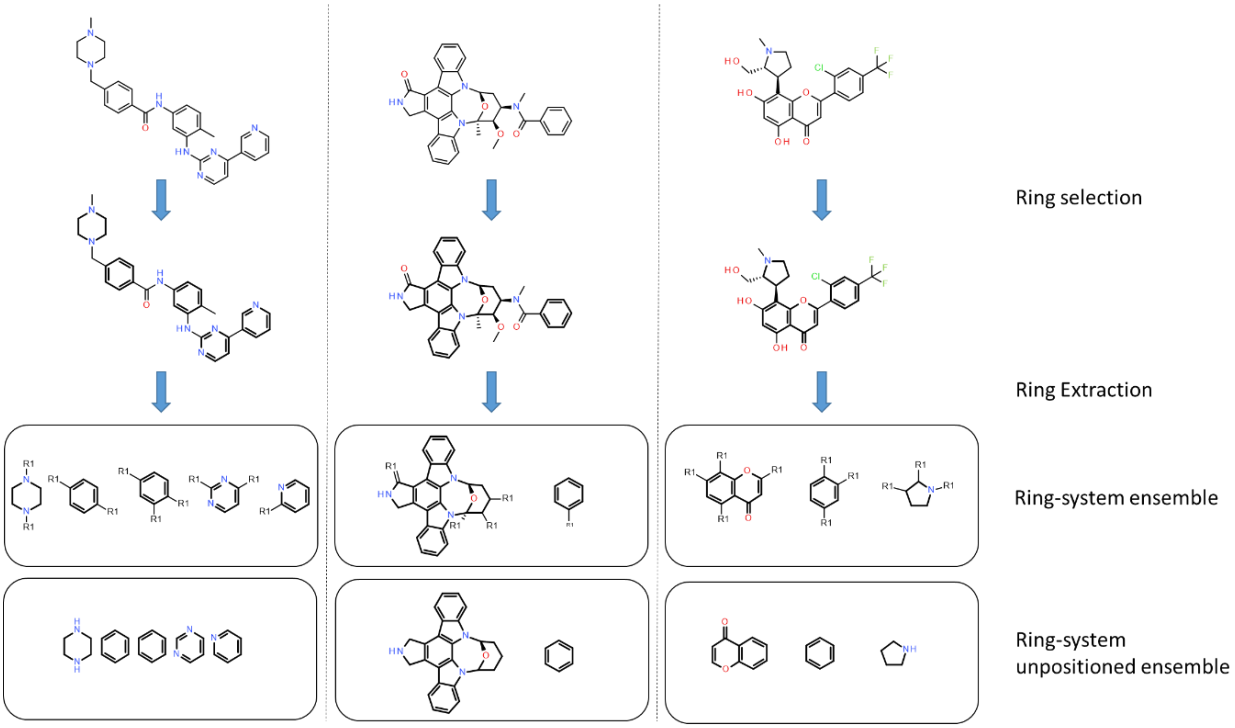


Figure 7. Application of the ring-system ensemble classification. Ring-system ensembles are obtained by removing substituents on acyclic bonds and by keeping attachment point (R1). The ring system unpositioned ensembles do not keep information on the attachment point. Rings are shown in bold.

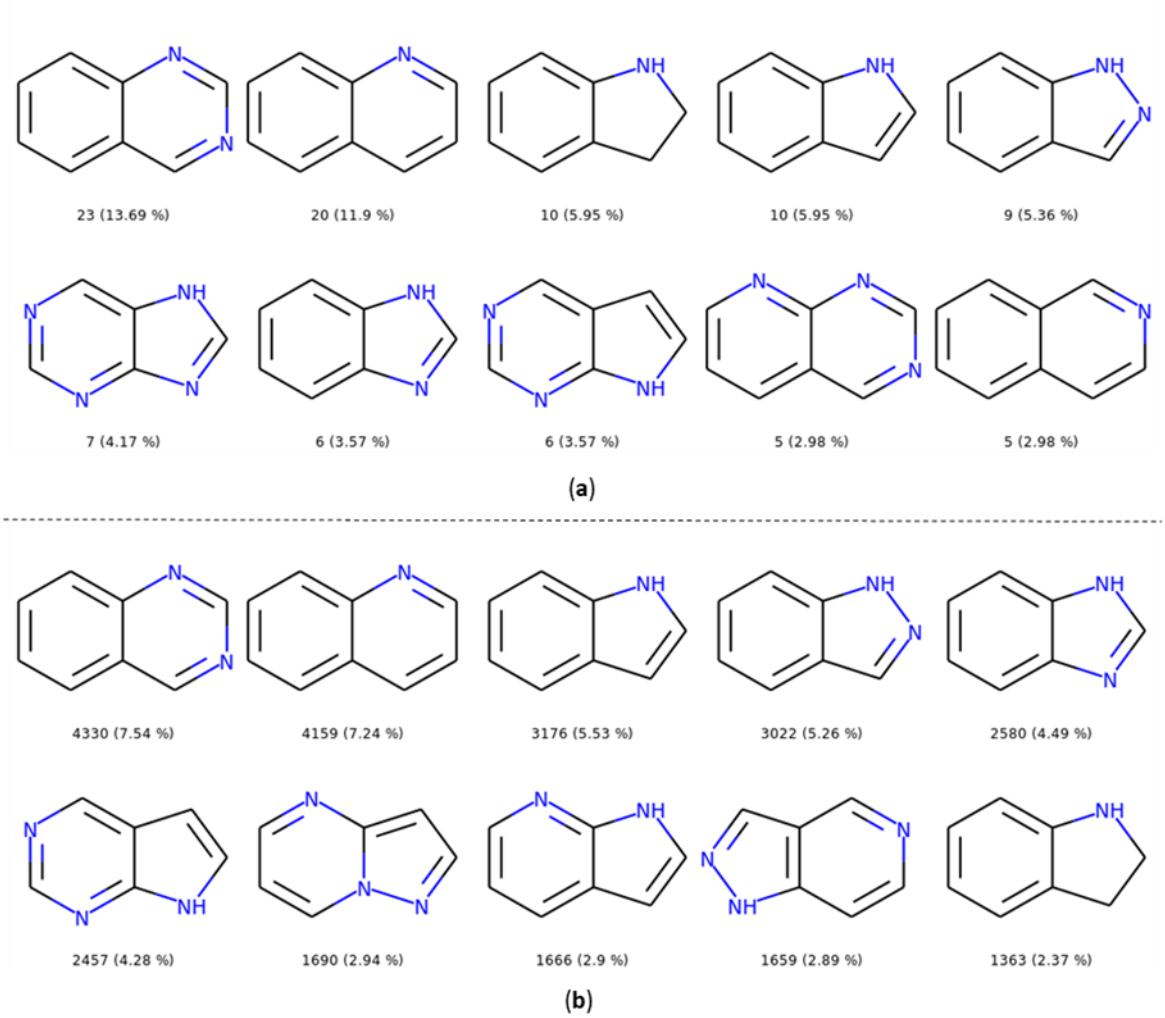
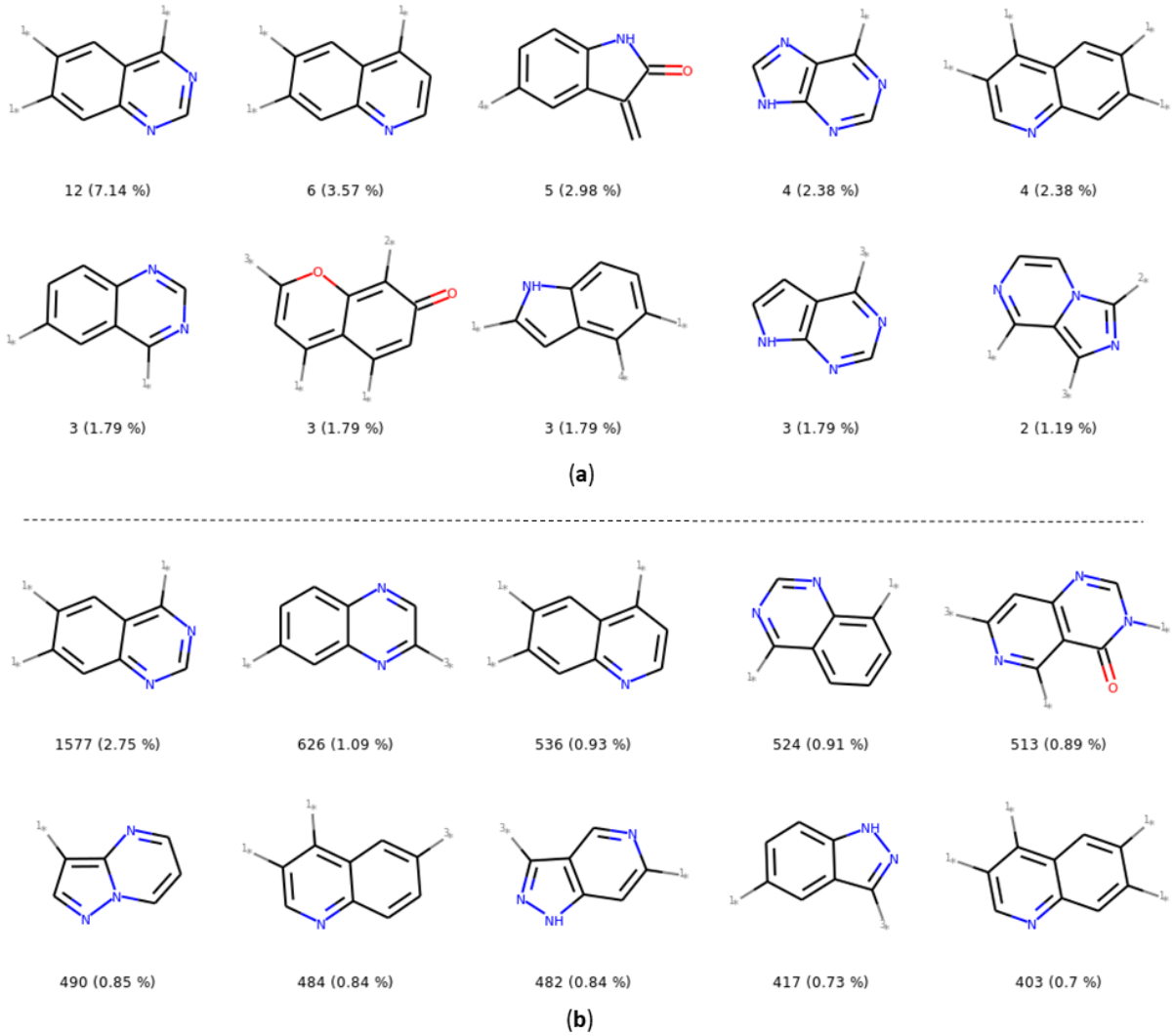


Figure 8. Top ten bicycles retrieved in PKIDB dataset (a) and in PKI_ChEMBL (b) with their occurrence and their frequency in brackets. In PKIDB there are 168 bicycles (53 unique) and in PKI_ChEMBL, there are 57,439 bicycles (697 unique).



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*

Figure 9. Top ten most retrieved bicycles with their substituents in PKIDB dataset (a) and in PKI_ChEMBL (b) with their occurrence and their frequency in brackets. In PKIDB there are 168 bicycles (125 unique) and in PKI_ChEMBL, there are 57,439 bicycles (4,480 unique).

3. Conclusion

PKIDB is a freely available database containing all kinase inhibitors on the market or in clinical trials gathered using their 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 part of protein kinase inhibitors, either approved

or in clinical trials, does not follow the recommended drug-like thresholds, especially regarding molecular weight and calculated LogP. Morever, 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 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, 209 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 [16]. 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'). Scaffolds analysis and clustering were performed with RDKIT and with Butina algorithm[32]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 [33] and Scikit-learn [34] packages. For PMI analysis, 3D conformations were generated using ETKDG method [35] followed with an optimization using the MMFF94 forcefield [36]. 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 [37] and seaborn [38] packages. Molecules were drawn with Biovia Draw 2018.

The PKI_ChEMBL dataset results from ChEMBL (version 'ChEMBL_24'). 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 [39] 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
LabuteASA	Labute's Approximate Surface Area
NAR	Number of aromatic rings
FCSP3	Fraction of C atoms that are SP3 hybridized
MQN8	Molecular Quantum Numbers
MQN10	Molecular Quantum Numbers

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. F.C, 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.

Author Contributions: C.B, F.C. and P.B. conceived and designed the experiments; C.B., G.P., S.B and F.C. performed the experiments; C.B, F.C., S.B., S.A.-S., C.M. and P.B. analyzed the data and wrote the paper.

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

Abbreviations

AMA	American Medical Association
ATP	adenosine triphosphate
BM	Bemis-Murcko scaffold
CFDA	China Food and Drug Administration
CLogP	calculated LogP
CML	chronic myelogenous leukemia
EMA	European Medicines Agency
HBA	number of hydrogen bond acceptors
HBD	number of hydrogen bond donors
FDA	Food and Drug Administration
GF	graph framework
INN	international nonproprietary name
IP	intellectual property
MW	molecular weight
NAR	number of aromatic rings
NCA	number of chiral atoms
NHA	number of heavy atoms
NME	new molecular entity
NRB	number of rotatable bonds
PCA	principal components analysis
PKI	protein kinase inhibitor

PMI	principal moments of inertia
SAR	structure–activity relationship
TPSA	topological polar surface area
USAM	United States Adopted Names

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.
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.
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.
4. Giamas, G.; Stebbing, J.; Vorgias, C.E.; Knippschild, U. Protein kinases as targets for cancer treatment. *Pharmacogenomics* **2007**, *8*, 1005–1016.
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.
6. Cohen, P. Immune diseases caused by mutations in kinases and components of the ubiquitin system. *Nat. Immunol.* **2014**, *15*, 521–529.
7. Fedorov, O.; Müller, S.; Knapp, S. The (un)targeted cancer kinome. *Nat. Chem. Biol.* **2010**, *6*, 166–169.
8. Roskoski, R. Properties of FDA-approved small molecule protein kinase inhibitors. *Pharmacol. Res.* **2019**.
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.
10. Maximiano, S.; Magalhães, P.; Guerreiro, M.P.; Morgado, M. Trastuzumab in the Treatment of Breast Cancer. *BioDrugs* **2016**, *30*, 75–86.
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. WHO | INN stems Available online: http://www.who.int/medicines/services/inn/stembook/en/ (accessed on Mar 20, 2019).
13. 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.
14. 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.
15. 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.
16. 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.
17. Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacol. Res.* **2016**, *103*, 26–48.
18. 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*.
19. Bemis, G.W.; Murcko, M.A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
20. Schuffenhauer, A.; Varin, T. Rule-Based Classification of Chemical Structures by Scaffold. *Mol. Inform.* **2011**, *30*, 646–664.

21. 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).
22. 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.
23. 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.
24. Wildman, S.A.; Crippen, G.M. Prediction of Physicochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868–873.
25. 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.
26. 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.
27. Mucs, D.; Bryce, R.A.; Bonnet, P. Application of shape-based and pharmacophore-based in silico screens for identification of Type II protein kinase inhibitors. *J. Comput. Aided Mol. Des.* **2011**, *25*, 569–581.
28. 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.
29. Zhang, J.; Yang, P.L.; Gray, N.S. Targeting cancer with small molecule kinase inhibitors. *Nat. Rev. Cancer* **2009**, *9*, 28–39.
30. Zhao, H.; Caflisch, A. Current kinase inhibitors cover a tiny fraction of fragment space. *Bioorg. Med. Chem. Lett.* **2015**, *25*, 2372–2376.
31. 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.
32. 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.
33. Halford, M. :*crown: Python factor analysis library (PCA, CA, MCA, MFA, FAMD): MaxHalford/prince*; 2019;
34. 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.
35. 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.
36. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.
37. 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;
38. 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;

39. 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.

Chapitre 2 : Développement d'une méthode de criblage virtuel appliquée à la cosmétique

2.1 Introduction et généralités sur le criblage virtuel

Comme son homologue expérimental (HTS), le criblage virtuel est une méthode visant à retrouver parmi des milliers, voire des millions, de molécules, celles possédant la meilleure affinité sur une cible donnée. Cependant dans ce cas, tout se passe *in silico*. Cette méthode est apparue à la fin des années 1970 et connait un succès croissant depuis les années 1990 (Figure 19). Dès cette période, quelques limites commençaient à apparaitre concernant le criblage à haut débit, notamment avec l'avènement de la chimie combinatoire et de l'explosion du nombre de molécules à tester, augmentant de fait son coût tout en diminuant son taux de touches originales[93]. L'idée du criblage virtuel est de permettre un premier tri afin de rapidement réduire le nombre de composés à tester expérimentalement, en éliminant les composés supposés inactifs et les molécules indésirables[94].

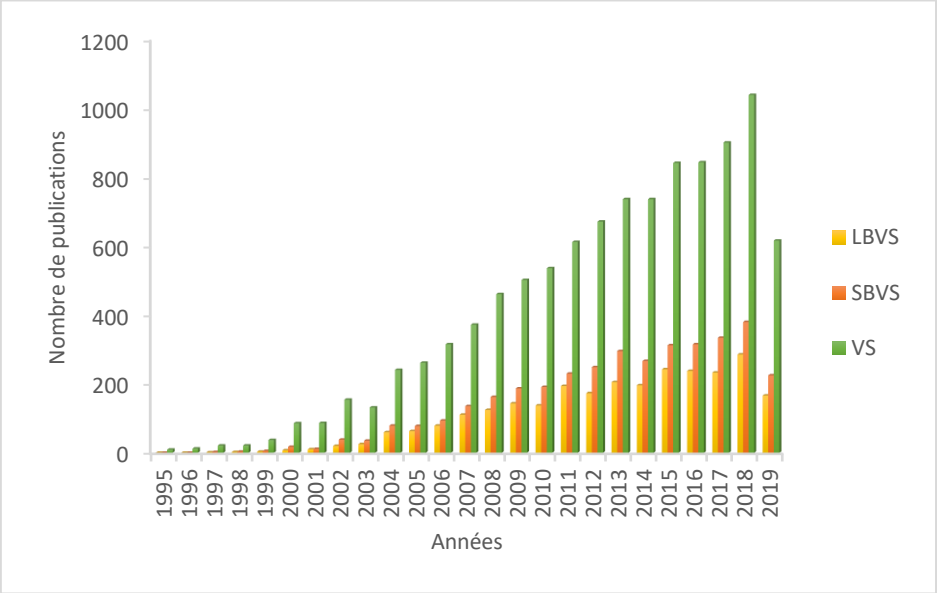


Figure 19 : Histogramme montrant le nombre d'occurences de papiers scientifiques contenant le terme « virtual screening » (vert), « structure based virtual screening » (orange) et « ligand based virtual screening » (jaune). *Recherche effectuée avec l'application Dimensions[95] la recherche des mots clés se faisant uniquement dans le titre et le résumé des publications.**

*Il faut néanmoins penser à mettre en perspective ces chiffres avec le fait que le nombre global de publications et de journaux scientifiques ne cessent également d'augmenter ces dernières années[96].

[93] Roger Lahana, « How many leads from HTS? », *Drug Discovery Today* 4, nº 10 (1 octobre 1999): 447-48, https://doi.org/10.1016/S1359-6446(99)01393-8.
[94] Yusuf Tanrikulu, Björn Krüger, et Ewgenij Proschak, « The holistic integration of virtual screening in drug discovery », *Drug Discovery Today* 18, nº 7 (1 avril 2013): 358-64, https://doi.org/10.1016/j.drudis.2013.01.007.
[95] « Dimensions », consulté le 28 juin 2019, https://app.dimensions.ai/discover/publication.
[96] « 21st Century Science Overload », *Canadian Science Publishing* (blog), consulté le 10 octobre 2019, http://blog.cdnsciencepub.com/21st-century-science-overload/.

2.1.1 Les différentes stratégies

Il existe deux catégories principales de méthodes lors d'un criblage virtuel : celles basées sur un ou plusieurs ligands reconnus comme étant actifs (« ligand-based ») et celles reposant sur la structure 3D de la cible (« structure-based »). Selon les informations disponibles au début du projet, on peut se diriger vers l'une, l'autre ou même les deux en parallèle[97]. Comme le montre la Figure 20, différentes techniques propres à chaque catégories peuvent ensuite être appliquées pour le criblage. Ici nous nous intéresserons uniquement à l'amarrage moléculaire (« docking ») de molécules dans une structure 3D, une méthode de la catégorie structure-based.

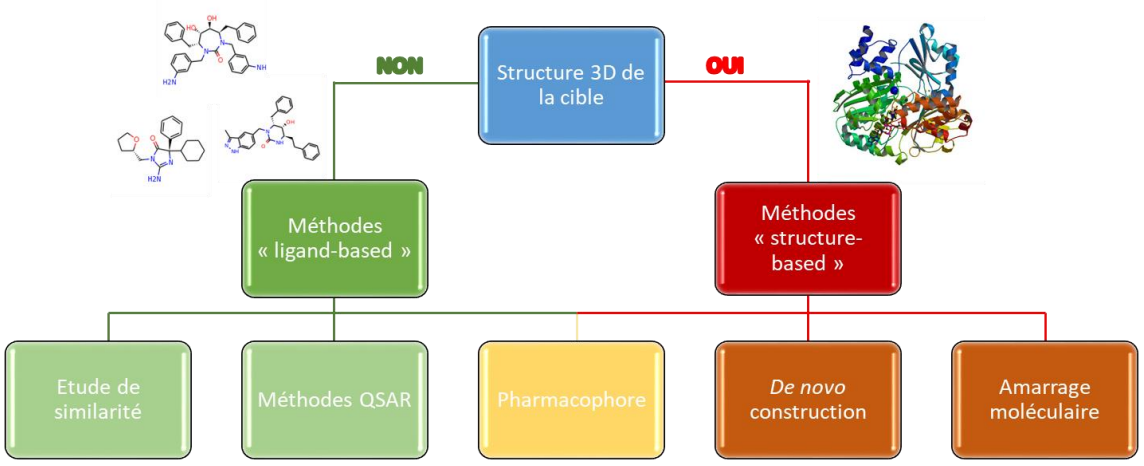


Figure 20 : Classification des méthodes de criblage virtuel selon les méthodes basés sur le(s) ligand(s) ou la structure 3D.

Dans ce cas de figure, le système considéré est composé de deux entités : la cible d'intérêt, plus particulièrement son site de liaison, et la molécule que l'on cherche à amarrer dans le dit site. Il s'agit là du cas classique, il existe d'autres systèmes d'amarrage moléculaire comme protéine-protéine, protéine-ADN, etc[98]. Le docking est composé de deux étapes, quel que soit le logiciel utilisé[99] :

- Une première étape d'exploration de l'espace conformationnel du site actif avec positionnement de la molécule à l'intérieur.

- Une seconde étape d'estimation énergétique (« scoring ») des différentes poses obtenues.

[97] Tudor I Oprea et Hans Matter, « Integrating virtual screening in lead discovery », *Current Opinion in Chemical Biology* 8, nº 4 (1 août 2004): 349-58, https://doi.org/10.1016/j.cbpa.2004.06.008.
[98] Yumeng Yan et al., « HDOCK: A Web Server for Protein–Protein and Protein–DNA/RNA Docking Based on a Hybrid Strategy », *Nucleic Acids Research* 45, nº W1 (3 juillet 2017): W365-73, https://doi.org/10.1093/nar/gkx407.
[99] Xuan-Yu Meng et al., « Molecular Docking: A powerful approach for structure-based drug discovery », *Current computer-aided drug design* 7, nº 2 (1 juin 2011): 146-57.

Il existe de nombreuses revues rendant compte des différentes méthodes d'explorations conformationnelles, des fonctions de scores ou encore des logiciels existants[100,101,102]. Cependant, un autre point important à considérer pour le choix du programme à utiliser est sa méthode utilisée, selon les degrés de liberté des liaisons (DLL) des éléments du système étudié[103] :

- Rigide, aucun DLL, le placement se fait uniquement par rotation et translation du ligand dans le site actif.

- Semi-flexible, DLL uniquement pour le ligand dont la conformation peut varier selon la fixation au site de liaison.

- Flexible, DLL pour le ligand et les acides aminés faisant partie du site actif de la cible.

Historiquement, le choix se faisait essentiellement par rapport aux contraintes de temps et de moyens, le docking flexible étant forcément beaucoup plus chronophage que le docking rigide. Désormais, avec l'amélioration des programmes et des ordinateurs, le docking rigide n'est pratiquement plus utilisé et chaque logiciel propose à minima une méthode semi-flexible[104].

Les différentes étapes à réaliser lors d'un projet de criblage virtuel sont illustrées par la Figure 21 et de nombreux exemples de succès sont disponibles dans la littérature scientifique[105,106,107].

[100] Nataraj S. Pagadala, Khajamohiddin Syed, et Jack Tuszynski, « Software for molecular docking: a review », *Biophysical Reviews* 9, nº 2 (16 janvier 2017): 91-102, https://doi.org/10.1007/s12551-016-0247-1.
[101] Elizabeth Yuriev, Jessica Holien, et Paul A. Ramsland, « Improvements, Trends, and New Ideas in Molecular Docking: 2012-2013 in Review », *Journal of Molecular Recognition: JMR* 28, nº 10 (octobre 2015): 581-604, https://doi.org/10.1002/jmr.2471.
[102] Leonardo G. Ferreira et al., « Molecular Docking and Structure-Based Drug Design Strategies », *Molecules* 20, nº 7 (juillet 2015): 13384-421, https://doi.org/10.3390/molecules200713384.
[103] Katrina W. Lexa et Heather A. Carlson, « Protein Flexibility in Docking and Surface Mapping », *Quarterly reviews of biophysics* 45, nº 3 (août 2012): 301-43, https://doi.org/10.1017/S0033583512000066.
[104] Elizabeth Yuriev, Mark Agostino, et Paul A. Ramsland, « Challenges and Advances in Computational Docking: 2009 in Review », *Journal of Molecular Recognition* 24, nº 2 (2011): 149-64, https://doi.org/10.1002/jmr.1077.
[105] Rognan, Didier, « Le criblage virtuel par docking moléculaire », in *Chemogénomique, des petites molécules pour explorer le vivant* (EDP Sciences, Collection Grenoble Sciences., 2007), 258.
[106] Alexander A. Alex et David S. Millan, « Chapter 5: Contribution of Structure-Based Drug Design to the Discovery of Marketed Drugs », in *Drug Design Strategies*, 2011, 108-63, https://doi.org/10.1039/9781849733410-00108.
[107] A. Lavecchia et C. Di Giovanni, « Virtual Screening Strategies in Drug Discovery: A Critical Review », *Current Medicinal Chemistry* 20, nº 23 (2013): 2839-60, https://doi.org/10.2174/09298673113209990001.

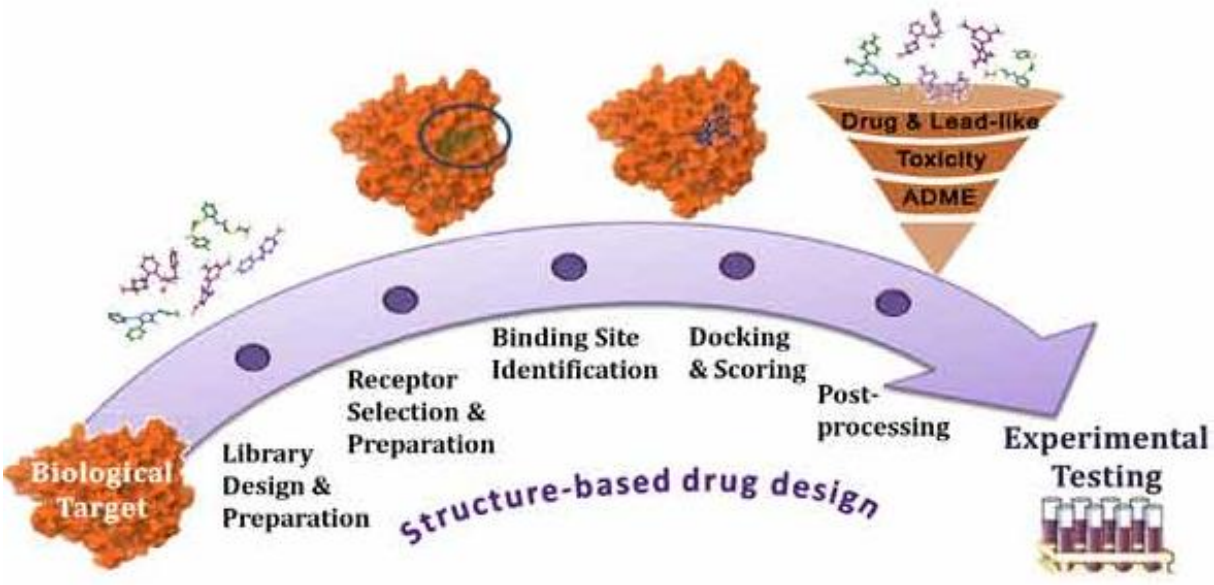


Figure 21 : Schéma des différentes étapes d'un projet de criblage virtuel. *D'après E. Lionta & al.*[108].

2.1.2 La création d'une structure protéique tridimensionnelle

Il existe trois principales techniques expérimentales pour obtenir un modèle 3D de protéine. La plus utilisée est la cristallographie aux rayons X, suivie par la résonance magnétique nucléaire (RMN). Cette dernière semble s'essouffler aux dépens de la microscopie électronique (EM) qui affiche une belle croissance ces dernières années (Figure 22). Malgré un total de 153 328 structures disponibles à ce jour (27/06/2019) dans la base de données PDB, toutes les protéines connues ne sont pas encore cristallisées. Dans le cas où la cible d'intérêt n'est pas disponible, la création d'un modèle par homologie est une solution permettant de se procurer une structure 3D[109].

[108] Evanthia Lionta et al., « Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances », *Current Topics in Medicinal Chemistry* 14, nº 16 (août 2014): 1923-38, https://doi.org/10.2174/1568026614666140929124445.

[109] Krzysztof Ginalski, « Comparative modeling for protein structure prediction », *Current Opinion in Structural Biology*, Theory and simulation/Macromolecular assemblages, 16, nº 2 (1 avril 2006): 172-77, https://doi.org/10.1016/j.sbi.2006.02.003.

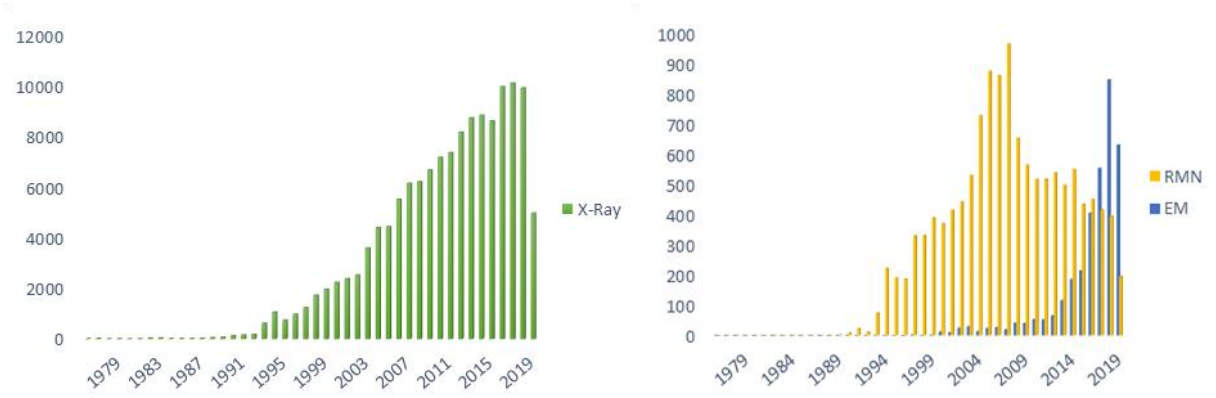


Figure 22 : Evolution du nombre de structures dans la PDB depuis 1976 en fonction de la méthode expérimentale utilisée. *A gauche, évolution du nombre de structures obtenues par cristallographie aux rayons X. A droite, en jaune évolution du nombre de structures obtenues par résonance magnétique nucléaire et en bleu par microscopie électronique. D'après les données de la PDB.*

Avant de lancer le criblage virtuel sur une cible, quelques étapes de préparation et de vérification sont nécessaires, que je vais détailler dans la partie suivante.

2.2 Paramétrage du criblage virtuel

2.2.1 Préparation de la structure 3D

La première chose à faire lors d'un amarrage moléculaire est de préparer la structure 3D de la cible[110]. En effet, les fichiers PDB bruts téléchargés ne sont pas utilisables en l'état pour du docking. Plusieurs paramètres sont à régler, par exemple la protonation des résidus à pH physiologique, car les structures issues d'expériences de cristallographie aux rayons X ne contiennent pas les atomes d'hydrogène. Dans le cadre de cette préparation, on peut aussi avoir besoin d'ajouter des atomes « lourds » s'ils sont manquants voir même certaines sous-parties entières de la protéine, qui peuvent être absentes et qu'il faut alors construire. Dans les fichiers PDB, il peut aussi y avoir présence d'agents de co-cristallisation comme le glycérol ou l'isopropanol, d'ions comme SO_4^{2-}, Cl^-, ou Na^+ qu'il faut supprimer car ils peuvent gêner le docking. De manière générale, on ne garde que les éléments essentiels de la structure téléchargée à savoir la protéine, le ligand co-cristallisé et les éventuelles molécules d'eau au sein du site actif si leurs présences sont nécessaires à la formation de liaison clés[111]. Enfin, une dernière étape consiste à vérifier les états d'ionisation et de tautomérisation des acides aminés du récepteur. La Figure 23 montre les différences entre une protéine « brute » et une protéine préparée à l'aide du module « Structure Preparation » du logiciel MOE (Chemical Computing Group, *version 2018.01*).

[110] G. Madhavi Sastry et al., « Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments », *Journal of Computer-Aided Molecular Design* 27, nº 3 (1 mars 2013): 221-34, https://doi.org/10.1007/s10822-013-9644-8.
[111] Benjamin C. Roberts et Ricardo L. Mancera, « Ligand−Protein Docking with Water Molecules », *Journal of Chemical Information and Modeling* 48, nº 2 (1 février 2008): 397-408, https://doi.org/10.1021/ci700285e.

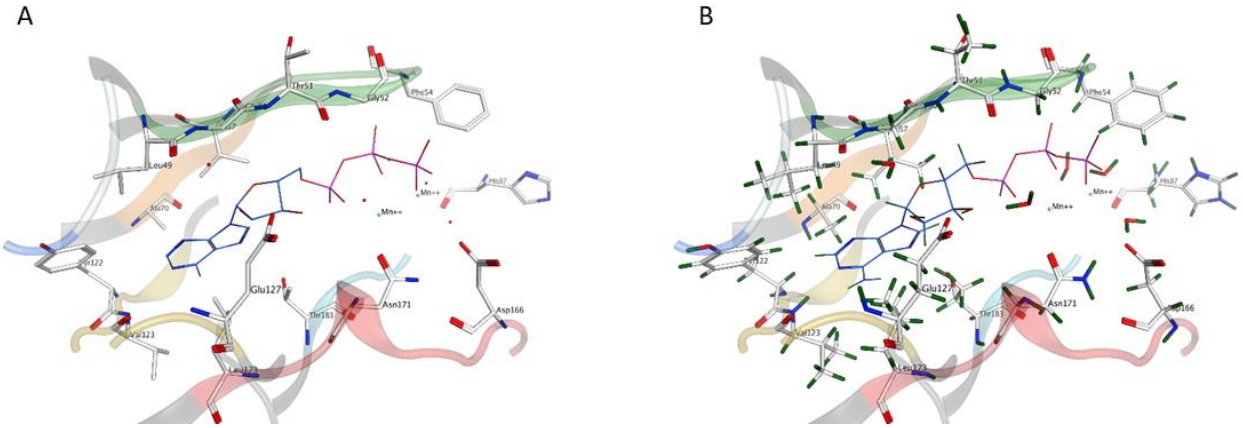


Figure 23 : Site actif brut (non préparé) (A) et site actif préparé pour l'amarrage moléculaire (B). *La structure 3D brute (à gauche) ne contient pas les hydrogènes (représentés en vert dans la structure préparée à droite). En plus de la protonation des résidus, l'état d'ionisation à pH physiologique est vérifié et les molécules d'eau en rouge sont optimisées et placées correctement.*

2.2.2 Caractérisation du site actif

Afin de positionner les molécules au bon endroit, les logiciels d'amarrage moléculaire ont besoin d'un repère leur permettant de créer et caractériser la cavité dans laquelle ils effectueront leurs opérations. La manière la plus simple - à condition que la structure 3D dont on dispose soit co-cristallisée avec un ligand dans son site actif - est de préciser au programme de se servir du ligand en précisant une certaine distance pour sélectionner tous les résidus de la future cavité. Si l'on ne dispose pas d'un ligand co-cristallisé mais que l'on connait la position du site actif, on peut indiquer au logiciel un repère géométrique précis (coordonnées XYZ du point de départ), là encore seulement les résidus à une distance donnée serviront pour la caractérisation de la cavité. Enfin, si aucune de ces informations n'est connue, on peut faire appel à des outils informatiques pour retrouver des cavités potentielles dans lesquelles effectuer le docking[112]. On retrouve dans la Figure 24 un exemple de cavité généré par le programme « rbcavity » du logiciel rDock[113] à partir d'un ligand.

[112] Daniel Barry Roche, Danielle Allison Brackenridge, et Liam James McGuffin, « Proteins and Their Interacting Partners: An Introduction to Protein–Ligand Binding Site Prediction Methods », *International Journal of Molecular Sciences* 16, nº 12 (décembre 2015): 29829-42, https://doi.org/10.3390/ijms161226202.
[113] Sergio Ruiz-Carmona et al., « RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids », *PLOS Computational Biology* 10, nº 4 (10 avril 2014): e1003571, https://doi.org/10.1371/journal.pcbi.1003571.

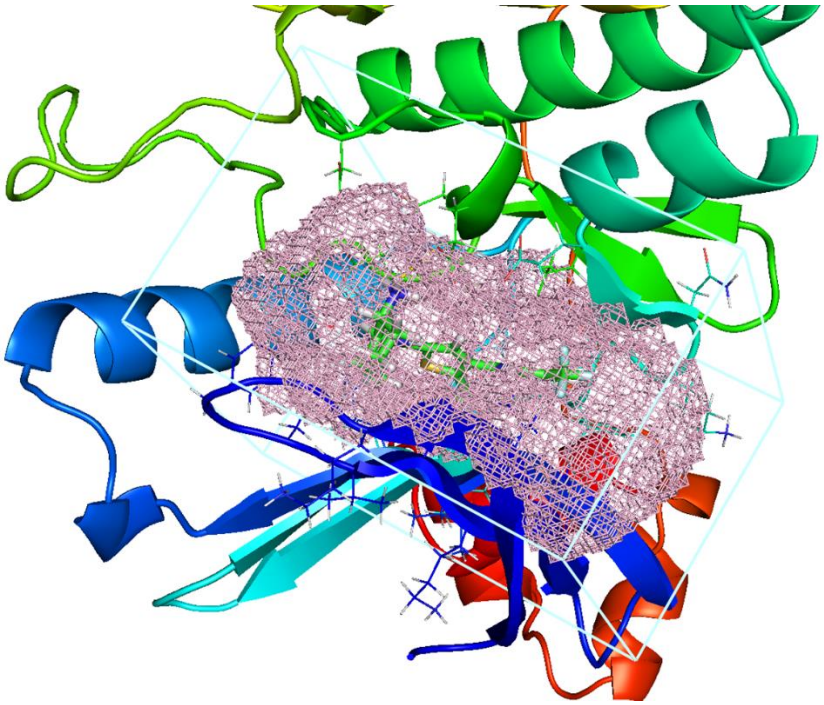


Figure 24 : Exemple d'une cavité utilisée pour effectuer un amarrage moléculaire. *La cavité a été construite par le logiciel rDock, le rectangle en bleu clair est la limite de distance par rapport au ligand (en vert) pour sa construction. La surface de la cavité est représentée en rose clair.*

2.3 Phase de validation

Une fois les réglages initiaux réalisés, avant de lancer le criblage virtuel sur la chimiothèque, il faut au préalable évaluer la méthode choisie afin de vérifier notamment si elle permet bien une séparation entre les composés connus comme actifs de ceux inactifs vrais ou présumés (les leurres).

2.3.1 L'amarrage du ligand initial

La précision du positionnement des molécules peut être évaluée en effectuant un amarrage moléculaire du ligand co-cristallisé afin de vérifier que le logiciel choisi est capable de le replacer à l'endroit initial (« redocking »). Pour cela, la métrique utilisée est l'écart quadratique moyen ou RMSD (« Root Mean Square Deviation », Équation 1). Le RMSD permet d'apprécier la distance moyenne entre atomes homologues (équivalents ou appariés) de deux conformations d'une même molécule dans un espace 3D. Plus sa valeur est basse, plus la pose prédite est similaire à la pose expérimentale (Figure 25). En général pour une validation d'un redocking la valeur du RMSD ne doit pas dépasser 2 Å[114].

[114] Esther Kellenberger et al., « Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy », *Proteins: Structure, Function, and Bioinformatics* 57, nº 2 (2004): 225-42, https://doi.org/10.1002/prot.20149.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(x_{ipred} - x_{iexp})^2 + (y_{ipred} - y_{iexp})^2 + (z_{ipred} - z_{iexp})^2]}$$

Équation 1 : Calcul du RMSD. *Avec N le nombre d'atomes des molécules, ipred et iexp correspondant aux atomes de la pose prédite et expérimentale respectivement et x, y, z aux coordonnées cartésiennes.*

Cette méthode étant intuitive, simple et rapide à mettre en place, la combinaison redocking et calcul de RMSD s'est imposée comme une référence pour la validation d'un protocole[115]. Cependant, afin de ne pas fausser le système, il faut respecter certaines règles et notamment fournir au logiciel une conformation différente du ligand co-cristallisé, ou au moins le déplacer, afin d'éviter de biaiser le placement de la molécule pendant le docking en lui donnant par avance une conformation bien adaptée au site de fixation. Il est toujours bon de s'intéresser aussi au rang en termes de scores de la pose présentant le plus faible RMSD. Par exemple dans le cas où l'on effectue un redocking en demandant 30 poses, si le meilleur RMSD est de 1 Å mais que la pose est au 29^{ième} rang en terme de score alors que les 10 poses présentant les meilleurs scores ont en moyenne un RMSD de 3 Å, il faut peut-être songer à modifier quelques paramètres du programme pour l'améliorer. Dans certains cas, le seuil de tolérance peut aussi être adapté selon la situation, comme dans le cas du redocking de fragments, le RMSD étant sensible à la taille des composés. Si une grande partie du ligand expérimental est exposé au solvant, là aussi on peut observer un RMSD élevé alors même que la partie du ligand située dans la cavité est fidèlement prédite. Au vu du faible nombre de poses généralement demandé, il vaut mieux toujours faire une inspection visuelle afin d'observer ce qu'il en est réellement et ainsi juger de la validation. Enfin, gardons en tête que le RMSD ne fournit aucune information sur la conservation des interactions originales entre le ligand co-cristallisé et le site actif. Pour cela, on peut se servir d'outils alternatifs comme SIFt ou SPLIF[116,117].

[115] Johannes Kirchmair et al., « Evaluation of the Performance of 3D Virtual Screening Protocols: RMSD Comparisons, Enrichment Assessments, and Decoy Selection—What Can We Learn from Earlier Mistakes? », *Journal of Computer-Aided Molecular Design* 22, nº 3 (1 mars 2008): 213-28, https://doi.org/10.1007/s10822-007-9163-6.

[116] Zhan Deng, Claudio Chuaqui, et Juswinder Singh, « Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions », *Journal of Medicinal Chemistry* 47, nº 2 (1 janvier 2004): 337-44, https://doi.org/10.1021/jm030331x.

[117] C. Da et D. Kireev, « Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study », *Journal of Chemical Information and Modeling* 54, nº 9 (22 septembre 2014): 2555-61, https://doi.org/10.1021/ci500319f.

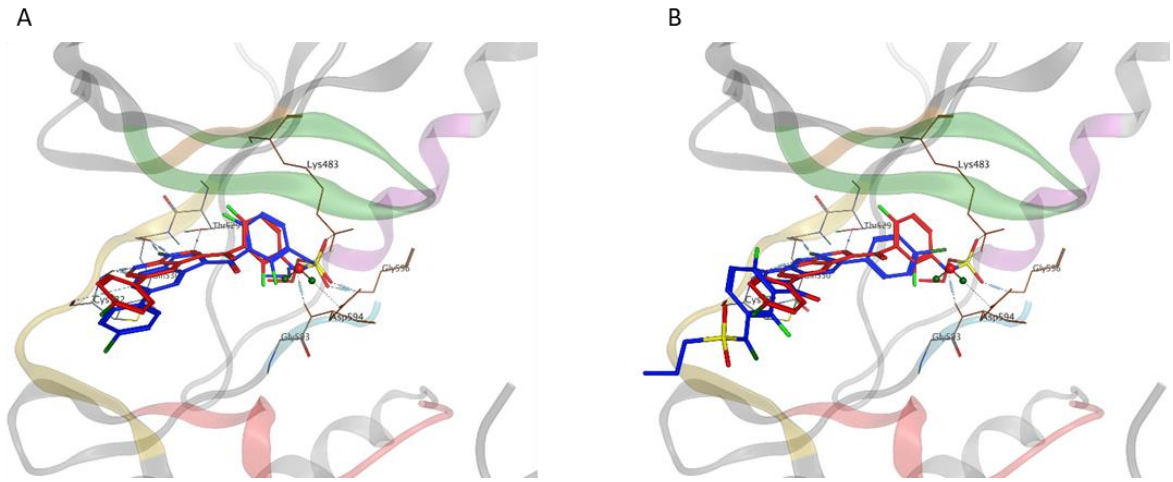


Figure 25 : Pose (en bleu) de redocking d'un ligand co-cristallisé (en rouge) avec un RMSD correct (A) et pose (en bleu) avec un mauvais RMSD (B). *A gauche le RMSD est de 0.91 Å, à droite de 11.09 Å. A droite, on peut observer que la molécule amarrée est totalement retournée par rapport au ligand d'origine.*

2.3.2 La discrimination des molécules

En plus de prédire un mode de liaison correct, une bonne méthode d'amarrage moléculaire doit être capable de différencier, à partir d'un jeu de données, les molécules actives des autres composés, les leurres (ou « decoys »). Ces jeux de données peuvent être téléchargés directement depuis une base de données spécifique (la DUD-E par exemple[118]), ou retrouvés dans les publications spécialisées dans la comparaison de logiciels de docking[119], ou enfin créées avec ses propres données ou celles disponibles dans les bases de données de bioactivité.

On parle d'enrichissement de chimiothèques pour décrire la capacité de la méthode à retrouver les composés actifs en premier par rapport aux leurres, autrement dit de leur attribuer les meilleurs scores. Un bon enrichissement est indispensable car le criblage virtuel a pour but de séparer les molécules actives des autres afin de ne tester expérimentalement que les meilleures. Deux méthodes principales sont utilisées pour évaluer l'enrichissement : le facteur d'enrichissement (EF) et la courbe ROC (« Receiver Operating Characteristic »).

2.3.3 Le facteur d'enrichissement

Le facteur d'enrichissement décrit la capacité de la méthode d'amarrage moléculaire à retrouver les composés actifs en tête de liste dans un sous-ensemble défini de la chimiothèque totale (Équation 2)[120].

[118] Michael M. Mysinger et al., « Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking », *Journal of Medicinal Chemistry* 55, nº 14 (26 juillet 2012): 6582-94, https://doi.org/10.1021/jm300687e.
[119] Nathalie Lagarde, Jean-François Zagury, et Matthieu Montes, « Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives », *Journal of Chemical Information and Modeling* 55, nº 7 (27 juillet 2015): 1297-1307, https://doi.org/10.1021/acs.jcim.5b00090.
[120] Niu Huang, Brian K. Shoichet, et John J. Irwin, « Benchmarking Sets for Molecular Docking », *Journal of Medicinal Chemistry* 49, nº 23 (16 novembre 2006): 6789-6801, https://doi.org/10.1021/jm0608356.

$$EF(x\%) = \frac{VP/n}{(VP + FN)/N}$$

Équation 2 : Calcul du facteur d'enrichissement pour la fraction des x premiers pourcents de la chimiothèque. *Avec VP : nombre de vrais positifs, n : nombre de composés dans la fraction sélectionnée, FN : nombre de faux négatifs et N : nombre total de composés dans la chimiothèque.*

Une bonne méthode de criblage retournera un EF élevé, très supérieur à 1. En pratique on le calcule pour des valeurs allant de (0,01 % à 10 %)[121] car c'est cette portion de molécules en tête de liste qui sera principalement envoyée en test expérimental. L'EF est une métrique dépendante du ratio de composés actifs présents dans le jeu de données. Ainsi, les jeux de données utilisés pour comparer plusieurs méthodes avec l'EF doivent présenter des ratios similaires. Il est toutefois à noter que l'EF est insensible à la distribution des actifs dans la portion définie : si on considère 15 molécules dont 7 actives, deux méthodes peuvent présenter un même EF alors que la méthode X retourne les actives en position 1 à 7 tandis que la méthode Y retourne les actives en position 8 à 15. Dans ce cas la méthode X est plus efficace que la méthode Y puisque les molécules sont retrouvées à un meilleur rang, bien que l'EF calculé soit identique. Ainsi, l'EF va permettre un aperçu ponctuel des performances, pour avoir une vision globale, il faut préférentiellement utiliser les courbes d'enrichissement, dont la courbe ROC.

2.3.4 La courbe ROC

La courbe ROC est une méthode d'évaluation de la performance d'un classifieur prédictif binaire (0 ou 1). Graphiquement, on la représente sous la forme d'une courbe représentant le taux de vrais positifs (TPR, « True Positive Rate »), aussi appelé Sensibilité (Équation 3), en fonction du taux de faux positifs (FPR, « False Positive Rate »), équivalent à la Spécificité (Équation 4)[122]. Plus simplement, une courbe ROC affiche le pourcentage de molécules actives en fonction des molécules leurres pour chaque fraction de la chimiothèque.

$$Sensibilité\ (Se) = \frac{VP}{VP + FN}$$

Équation 3 : Calcul de la sensibilité. *La sensibilité est définie comme le ratio du nombre d'actifs retrouvés dans une fraction donnée par rapport au nombre total d'actifs de la chimiothèque, avec VP : nombre de vrais positifs et FN : nombre de faux négatifs.*

$$Spécificité\ (Sp) = \frac{VN}{VN + FP}$$

Équation 4 : Calcul de la spécificité. *La spécificité est définie comme le ratio du nombre de leurres non retrouvés dans une fraction donnée par rapport au nombre total de leurres de la chimiothèque, avec VN : nombre de vrais négatifs et FP : nombre de faux positifs.*

[121] Hongming Chen et al., « On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors », *Journal of Chemical Information and Modeling* 46, nº 1 (1 janvier 2006): 401-15, https://doi.org/10.1021/ci0503255.
[122] Tom Fawcett, « An introduction to ROC analysis », *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27, nº 8 (1 juin 2006): 861-74, https://doi.org/10.1016/j.patrec.2005.10.010.

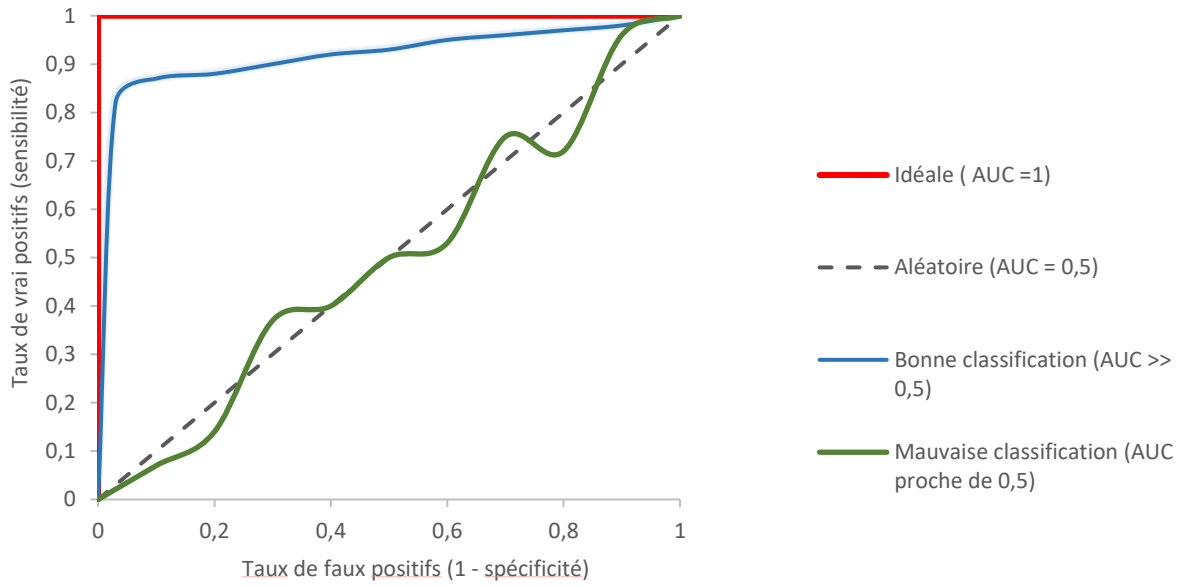


Figure 26 : Courbes ROC idéale (rouge), acceptable (bleue) et aléatoire (verte).

A la différence du facteur d'enrichissement, la courbe ROC est bornée par 0 et 1 à chaque axe et ne dépend pas du ratio d'actifs/leurres. La valeur de l'aire sous la courbe (AUC, « Area Under the Curve ») permet de quantifier la performance globale de la méthode et/ou d'en comparer plusieurs. Une valeur de 1 indique un classificateur parfait tandis qu'une valeur de 0,5 équivaut à un modèle prédictif aléatoire. De manière générale, plus la valeur de l'AUC est proche de 1, plus les paramètres choisis pour le criblage virtuel permettent de discriminer les composés actifs des leurres. Là aussi, comme l'EF, la valeur brute de l'AUC ne permet pas de refléter la différence de distribution des molécules actives, deux valeurs similaires peuvent donc provenir de courbes ROC très différentes. Ainsi, il demeure important de vérifier visuellement les courbes pour avoir une vision plus détaillée de chaque performance.

2.4 Post-traitement des résultats

Le score retourné lors d'un docking quantifie le degré avec lequel une molécule est liée au récepteur ciblé. Si les paramètres ont bien été optimisés et validés, le score doit être discriminant et donc permettre la sélection des molécules avec le plus haut potentiel d'affinité. Cependant, il existe de nombreuses autres techniques pour sélectionner les molécules après un criblage virtuel. Pour ne pas se fier à une unique fonction de score, on peut effectuer une nouvelle évaluation des poses obtenues avec une autre fonction de score et comparer les résultats. Cette stratégie est appelée « consensus scoring »[123], les molécules à sélectionner pour les tests expérimentaux seront ainsi celles présentant un très bon score pour chaque fonction différente.

En dehors du score, plusieurs autres éléments peuvent servir à la sélection des touches. Ainsi, on peut s'intéresser aux interactions entre la pose et le site actif et les comparer avec les interactions du ligand original à partir des empreintes d'interactions moléculaires[124]. On peut

[123] Paul S. Charifson et al., « Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins », *Journal of Medicinal Chemistry* 42, nº 25 (1 décembre 1999): 5100-5109, https://doi.org/10.1021/jm990352k.
[124] Gilles Marcou et Didier Rognan, « Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints », *Journal of Chemical Information and Modeling* 47, nº 1 (1 janvier 2007): 195-207, https://doi.org/10.1021/ci600342e.

aussi faire des études de diversité, de similarité ou pharmacophoriques. Un pharmacophore représente les groupements fonctionnels responsables d'une activité biologique d'un composé, il peut être en 2D ou 3D. Dans le cas d'un pharmacophore 3D, l'arrangement spatial de ces groupes dans le récepteur sera pris en compte. A l'aide de données expérimentales, on peut chercher un pharmacophore commun sur les molécules actives et sélectionner les composés issus du docking à la condition qu'ils possèdent ce même pharmacophore[125]. Enfin, on peut aussi utiliser des méthodes plus difficiles à mettre en œuvre avec l'utilisation de la mécanique moléculaire comme le MM-GBSA[126].

2.5 Présentation du projet

2.5.1 Contexte et but

La société Greenpharma, avec laquelle nous collaborons sur ce projet, est spécialisée dans la connaissance des substances naturelles et leurs valorisations notamment dans les domaines de la santé et la cosmétique. Dans le cadre de cette collaboration, un projet de développement de méthodes *in silico* pour l'identification de nouveaux produits naturels actifs a vu le jour. Il s'agit d'un projet appliqué au domaine de la cosmétique et plus particulièrement contre les marques du vieillissement de la peau : apparition de rides, de taches brunes et autres troubles pigmentaires. Après un état de l'art sur les cibles potentielles impliquées dans ces phénomènes, j'ai développé une méthode de criblage virtuel pour trouver parmi une base de données de produits naturels, les composés les plus prometteurs avant de les tester expérimentalement.

Un projet similaire s'est déjà déroulé dans l'équipe sur une cible différente, mené par un précédent doctorant : J.-M. Gally. De fait, la chimiothèque de produits naturels utilisée pour le criblage était déjà réalisée et je l'ai réutilisée telle quelle. Il s'agit d'une chimiothèque regroupant plusieurs bases de données de produits naturels contenant au total 4 377 molécules. Vous trouverez plus de détails sur sa préparation dans l'article à la fin de ce chapitre.

Les outils principaux utilisés lors de ce projet sont : GOLD[127] (*version 5.2.2*), avec la fonction de score ASP pour le criblage virtuel, MOE (Chemical Computing Group, *version 2018.01*) pour la visualisation des poses obtenues et le langage Python (*version 3.6.2)* pour le traitement post-docking et la sélection des composés à tester.

2.5.2 Le logiciel de docking GOLD

GOLD est un programme commercial développé par « The Cambridge Crystallographic Data Center » (http://www.ccdc.cam.ac.uk/). C'est un logiciel qui permet de réaliser du docking flexible. En effet, certains résidus du site actif peuvent subir des mouvements de rotation ou de translation. GOLD explore l'espace conformationnel du site actif à l'aide d'un algorithme génétique. Ce type d'algorithme améliore de façon itérative les solutions proposées en s'inspirant du processus naturel de l'évolution. Son nom fait référence à une analogie avec la théorie de l'évolution stipulant que les gènes conservés au sein d'une population donnée seront

[125] Thomas Seidel et al., « Strategies for 3D pharmacophore-based virtual screening », *Drug Discovery Today: Technologies*, 3D Pharmacophore Elucidation and Virtual Screening, 7, nº 4 (1 décembre 2010): e221-28, https://doi.org/10.1016/j.ddtec.2010.11.004.
[126] Paulette A. Greenidge, Richard A. Lewis, et Peter Ertl, « Boosting Pose Ranking Performance via Rescoring with MM-GBSA », *Chemical Biology & Drug Design* 88, nº 3 (2016): 317-28, https://doi.org/10.1111/cbdd.12763.
[127] G. Jones et al., « Development and Validation of a Genetic Algorithm for Flexible Docking », *Journal of Molecular Biology* 267, nº 3 (4 avril 1997): 727-48, https://doi.org/10.1006/jmbi.1996.0897.

ceux le plus adaptés aux besoins de l'espèce vis-à-vis de son environnement. Cela s'explique par le fait que certaines variations de gènes vont conférer aux individus les possédant un avantage compétitif par rapport aux autres. Cet avantage se traduit par une meilleure reproduction de ces individus permettant donc la transmission de ces gènes à l'ensemble de la population après plusieurs générations[128].

Dans le cas de GOLD, la population initiale est constituée de multiples poses obtenues aléatoirement. Ensuite, chaque pose va évoluer et subir des transformations arbitraires modifiant sa conformation et sa position dans le site actif. Les nouvelles conformations sont ensuite évaluées avec la fonction de score et remplacent les conformations initiales moins bonnes. Les meilleurs résultats sont ainsi favorisés en gardant uniquement les conformations présentant les meilleurs scores au fur et mesure du processus[165]. Comme les poses initiales et les transformations subies sont réalisées de manière aléatoire, on parle d'algorithme de type stochastique (à opposer aux types déterministes, pour lesquels des valeurs sont déterminées). GOLD possède 4 fonctions de score différentes pour évaluer les poses :

- « GoldScore », fonction historique, prenant en compte plusieurs paramètres dont l'énergie des liaisons hydrogène, l'énergie électrostatique, les liaisons de van der Waals…[165] En fonction de la torsion du ligand, une pénalité peut être attribuée.

- « ChemScore », proposant une valeur d'énergie (**Δ**G) représentant le changement d'enthalpie libre qui se produit lorsque le ligand se fixe au récepteur. Il s'agit d'une fonction de type empirique dont les paramètres ont été calculés par régression linéaire à partir de 82 complexes protéines-ligands[129].

- « ASP », reposant sur une analyse de distance entre paires d'atomes issues de structures tridimensionnelles déterminées expérimentalement[130]. Elle est comparable aux fonctions du même type comme par exemple PMF[131] et Drugscore[132].

- « ChemPLP », aussi de type empirique, utilisant des termes de la fonction « ChemScore » ainsi que des termes de répulsion[133]. Il s'agit de la fonction la plus rapide à calculer, désormais utilisée par défaut dans les dernières versions de GOLD.

2.6 Etude d'un criblage virtuel sur la protéine kinase SIK2

[128] Melanie Mitchell, *An Introduction to Genetic Algorithms* (Cambridge, MA, USA: MIT Press, 1998).
[129] Marcel L. Verdonk et al., « Improved Protein-Ligand Docking Using GOLD », *Proteins* 52, nº 4 (1 septembre 2003): 609-23, https://doi.org/10.1002/prot.10465.
[130] Wijnand T. M. Mooij et Marcel L. Verdonk, « General and Targeted Statistical Potentials for Protein–Ligand Interactions », *Proteins: Structure, Function, and Bioinformatics* 61, nº 2 (2005): 272-87, https://doi.org/10.1002/prot.20588.
[131] I. Muegge et Y. C. Martin, « A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach », *Journal of Medicinal Chemistry* 42, nº 5 (11 mars 1999): 791-804, https://doi.org/10.1021/jm980536j.
[132] H. Gohlke, M. Hendlich, et G. Klebe, « Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions », *Journal of Molecular Biology* 295, nº 2 (14 janvier 2000): 337-56, https://doi.org/10.1006/jmbi.1999.3371.
[133] Oliver Korb, Thomas Stützle, et Thomas E. Exner, « Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS », *Journal of Chemical Information and Modeling* 49, nº 1 (janvier 2009): 84-96, https://doi.org/10.1021/ci800298z.

Article

Virtual screening of natural products to enhance melanogenesis

Colin Bournez [1], José-Manuel Gally [1], Samia Aci-Sèche [1], Philippe Bernard[2], Pascal Bonnet [1,*]

[1] Institut de Chimie Organique et Analytique (ICOA), UMR CNRS-Université d'Orléans 7311, Université d'Orléans BP 6759, 45067, Orléans Cedex 2, France
[2] Greenpharma SAS, 3 allée du Titane 45100 Orléans, France

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

Abstract: Natural products have long been an important source of inspiration for medicinal chemistry and drug discovery. In the cosmetic field, they remain the major elements of the composition and serve as marketing asset. Recent research showed the implication of salt-inducible kinases on the melanin production in skin via MITF regulation. Finding new potent modulators on such target could open the way to several cosmetic applications to attenuate visible signs of photoaging and ameliorate sunless tanning. Since virtual screening can be a powerful tool for detecting hit compounds in the early stages of a drug discovery process, we applied this approach on salt-inducible kinase 2 to discover potential interesting compounds. Here, we present the different steps from the construction of a database of natural products, to the validation of a docking protocol and the results of our virtual screening. Hits found from the screening were afterward tested *in vitro* to confirm their efficiency and results will be discussed here.

Keywords: chemoinformatics; natural product; virtual screening; docking; cosmetic; kinase

1. Introduction

Skin is the largest organ of the human body and acts as a waterproof physical barrier against the external environment [1]. The products employed for skin care and protection belong to cosmetics, they provide their effect by topical administration directly on the skin. The global cosmetic market is rapidly growing and is expecting to reach $429.8 billion by 2022 [2]. Among this global market, skin care segment is the most important one, expected at $163.5 billion by the same year [3].

Skin alteration can result from multiple intrinsic factors including aging, diet and hormonal influence. Moreover, due to its location at the body's surface, external factors, mainly ultraviolet radiation (UVR), also play an important role in skin degradation. Indeed, UVR effects on the skin, referred as photoaging, are quite significant. They might induce a variety of mutagenic and cytotoxic DNA lesions. UVR chronic exposition traits includes oxidative stress, pigmented spots, loss of skin tone, skin wrinkles, increased risks for skin cancer, etc [4]. As a defense, the organism secretes multiples pigments acting as absorbent filters to reduce the penetration through the epidermis of UV and thus reducing their damage [5]. In mammalians, these pigments are known as melanin, a generic term employed to group the three different types existing: eumelanin, pheomelanin, and neuromelanin. However, only eumelanin and pheomelanin are found in human epidermis [6]. The production of melanin (melanogenesis) results from a complex cascade pathway (Figure 1). It rises after the exposure of keratinocytes, the epidermis cells, to UVR. Damages done by UVR on keratinocytes DNA triggers p53-mediated transcription of the proopiomelanocortin (POMC) gene [7]. POMC peptide cleavage produces melanocyte-stimulating hormone (α-MSH), which is afterward secreted from the keratinocytes. Once α-MSH bounds to the melanocortin 1 receptor (MC1R), located on the membrane of melanocytes, cAMP level increases via the activation of the adenylate cyclase. High level of cAMP activates protein kinase A (PKA), which phosphorylates the cAMP-responsive-element-binding protein (CREB). This enhances the transcription of the microphthalmia-associated transcription factor (MITF)

gene [8]. This MITF induces the expression of tyrosinase catalysing the oxidation of L-tyrosine to L-DOPA and then to dopaquinone, serving as a common precursor to both pheomelanin and eumelanin [9]. Finally, once synthesized, the melanin is stored in melanosomes, which are transferred along melanocyte microtubules to basal keratinocytes [10]. The visible result of melanogenesis is the darkening of the skin, commonly called tanning.

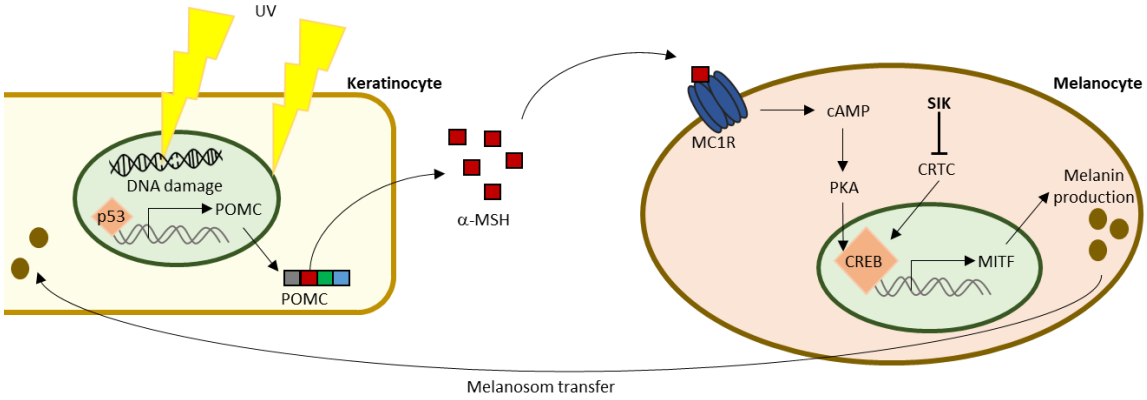


Figure 1. Melanin production induced by UV radiation and role of SIK kinase.

Skin pigmentation deregulation might have a great impact on the wellness of individuals [11]. Furthermore, with a rising demand in western culture for golden and uniform tan [12], being able to control (positively or negatively) the melanogenesis remains a great opportunity for the dermatological and cosmetic fields. Recent studies report that salt-inducible kinases (SIK) play an important role in melanin production by regulating the activity of the transcription factor CREB [13–15]. Thus, their inhibition can increase melanogenesis and induce skin tanning without UVR injuries. SIKs are serine/threonine kinases belonging to AMP-activated protein kinase (AMPK) family with three distinct isoforms: SIK1, 2, and 3 [16]. The predominant isoform in melanocytes is SIK2 [13].

In cosmetics, the trend observed nowadays is a rising demand for natural and organic products since consumers are more concerned about synthetic ingredients and chemical substances [17]. Although natural products were traditionally used since ancient time, they were gradually replaced by synthetic based ones, exhibiting similar properties, in the last one and half century [18]. However, thanks to the shift and the increasing request for natural, such substances are now becoming prevalent in modern cosmetic and cosmeceutic formulations [19]. Several natural compounds are already known to prevent UV damage with UV absorption property [20]. Some flavonoids are able to downregulate the melanogenesis or on the contrary to stimulate it [21,22]. In any event, these previous results prove that the melanogenesis can therefore be modulated with natural products. The goal of this work is to retrieve new interesting ones for cosmetic application. A topical application of such substances could be an interesting alternative to UV-tanning.

The application of structure-based virtual screening (VS) before experimental screening for discovering hit compounds on a target has been shown to be profitable [23]. To avoid high false-positive rate and compounds with low affinity, the methodology has to be gingerly validated before running on a whole library of compounds. In this study, a VS strategy applied on SIK2 kinase (Figure 2) were implemented. First, a group of several database of natural products were selected and prepared to serve as library to screen. Second, a rigorous creation of a 3D structure was performed out using homology modelling. Third, different software and methods were assessed on a small dataset to find the most efficient docking strategy before running it with the whole library. Finally, after a careful selection of hits, experimental tests were carried out to confirm them and quantify their binding affinities.

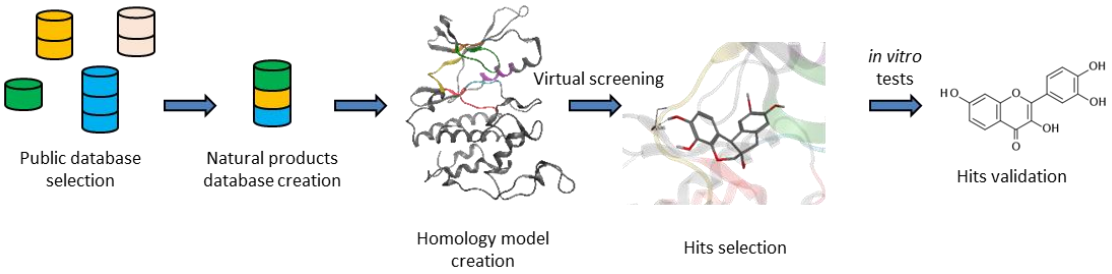


Figure 2. Workflow of the different steps implemented to retrieve interesting molecules.

2. Results

2.1. Creation of a natural products database

The primary goal was to obtain a database of natural products containing the information on the precise origin of the compounds and the different organisms from which they can be extracted. Nowadays, many databases of natural products are available and a non-exhaustive list may be found on the ZINC15 database catalog (http://zinc15.docking.org/catalogs/subsets/biogenic/) [24]. However, they may not be free or may not contain all necessary information desired on the origin of the compounds. Therefore, seven public curated databases were manually selected: BioPhytMol [25], KNApSAcK-3D [26], NuBBE [27], SANCDB [28], StreptomeDB [29], TCM [30] and TM-MC [31]. The characteristics of these databases are recapitulated in the Table 1. The majority of the compounds comes from plants, except for StreptomeDB that is exclusively composed of metabolites from bacteria. Three databases contain compounds from around the world (BioPhytMol, StreptomeDB and KNApSAcK-3D) while the other ones cover a restraint geographical area (Brasil for NuBBE, South Africa for SANCDB and Asian countries for the others). The number of molecules per database varies from 633 to 60,556 but all databases exhibit a good diversity with molecular similarity mean between all compounds around 0.40. TM-MC is the database showing the best compound diversity with a molecular similarity mean of 0.31.

Table 1. Characteristics of each database

Database	Localisation	Source	Year	Molecules	Molecular similarity mean (SD) [1]	URL
BioPhytMol	Worldwide	Plant	2014	633	0.39 (0.17)	http://ab-openlab.csir.res.in/biophytmol/
NuBBE	Brasil	Plant	2013	881	0.41 (0.19)	http://nubbe.iq.unesp.br/portal/nubbe-search.html
StreptomeDB	Worldwide	Bacteria	2013	4,040	0.4 (0.15)	http://132.230.56.4/streptomedb2/
SANCDB	South Africa	Plant, marine life	2015	712	0.43 (0.18)	https://sancdb.rubi.ru.ac.za/
KNApSAcK-3D	Worldwide	Plant	2012	51,179	0.44 (0.16)	http://knapsack3d.sakura.ne.jp/
TCM_Database @Taiwan	China	Plant, mineral, animal	2011	60,556	0.45 (0.16)	http://tcm.cmu.edu.tw/ [2]
TM-MC	Northeast Asian	Plant	2015	25,518	0.31 (0.18)	http://informatics.kiom.re.kr/compound/

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

[2]URL appears to be down, available via ZINC15: http://zinc15.docking.org/catalogs/tcmnp/

Each database's physico-chemical properties distribution is shown in Figure 3. According to all the descriptors calculated, no significant difference in the distribution of values by database is observable. A few databases contain heavy molecules (> 1,500 Da) but the average ranges from 327 Da for TM-MC to 556 Da for TCM. The extreme value of each descriptor corresponds to either compounds belonging to tannin family such as macabertin or ellagitanin, either complex glycosides as albizoside A (saponin family), or peptides, as siamycin II, mainly in StreptomeDB. An illustration of the greatest molecule from all database (MW: 3,738.30 Da) is provided in the supplemental information.

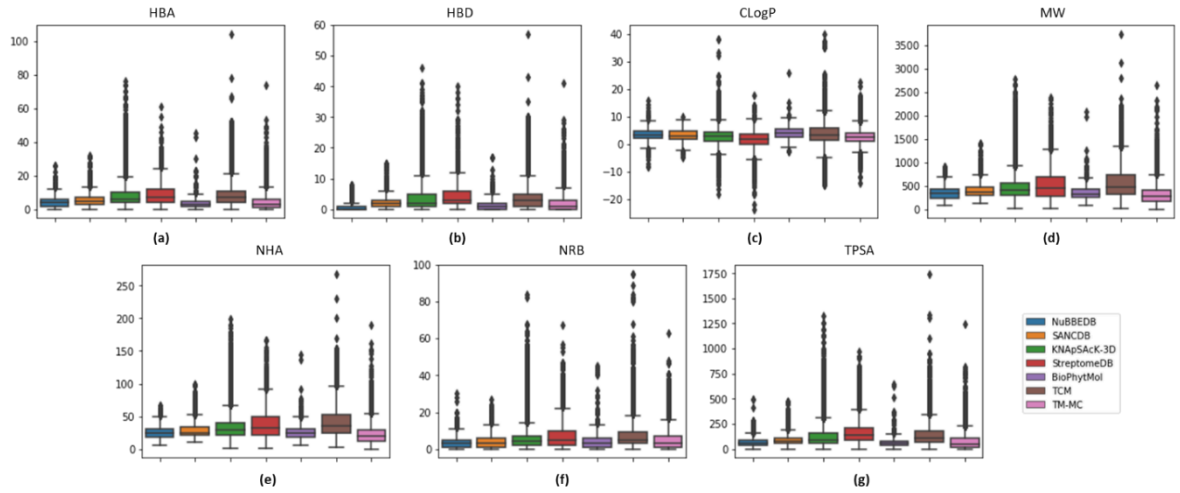


Figure 3. Distributions of physicochemical properties of gathered database: (a) Number of hydrogen bond acceptors (HBA); (b) Number of hydrogen bond donors (HBD); (c) Number of rotatable bonds (NRB); (d) Number of heavy atoms (NHA); (e) Molecular weight (MW); (f) ClogP (calculated with RDKit); (g) Topological polar surface area (TPSA).

The duplicates were eliminated based on the SMILES formula of molecules. Each database contained less than 5% of duplicates except TCM (16%). When compared two versus two (matrix available in the supplemental information), all databases contain at least one common identical molecule. Not surprisingly, since they both contain the greatest number of molecules, Knapsack-3D and TCM possess the highest rate of similar molecules (2,394 in common, 2.4%). On another side, SANCDB and NuBBE only share five molecules, making sense since these two databases contain few molecules, which, moreover, come from geographically distant areas.

Before preparing the molecules for the virtual screening, the last step carried out was to check their availability in Ambinter (http://www.ambinter.com/), a global chemical supplier, to ensure the possibility to purchase hits for further experimental testing. Ambinter also allowed us to incorporate its own natural product catalogue to our database. We ended with a database of 4,534 unique natural compounds. The Venn diagram on the Figure 4 represents the overlapping between the databases of origin of these natural compounds. Among the seven databases selected, four little ones gathered into the label "Others" in the Figure 4 provide 109 molecules. The biggest part of original molecules come from Ambinter natural products database, followed by TM-MC and TCM databases. The database providing the fewest molecules is SANCDB, with only two compounds purchasable.

A comparison based on the physico-chemical properties between our database and protein kinase inhibitors (PKI) from PKIDB [32] showed that the chemical space covered by our database is different from the one from PKI (see supplemental information).

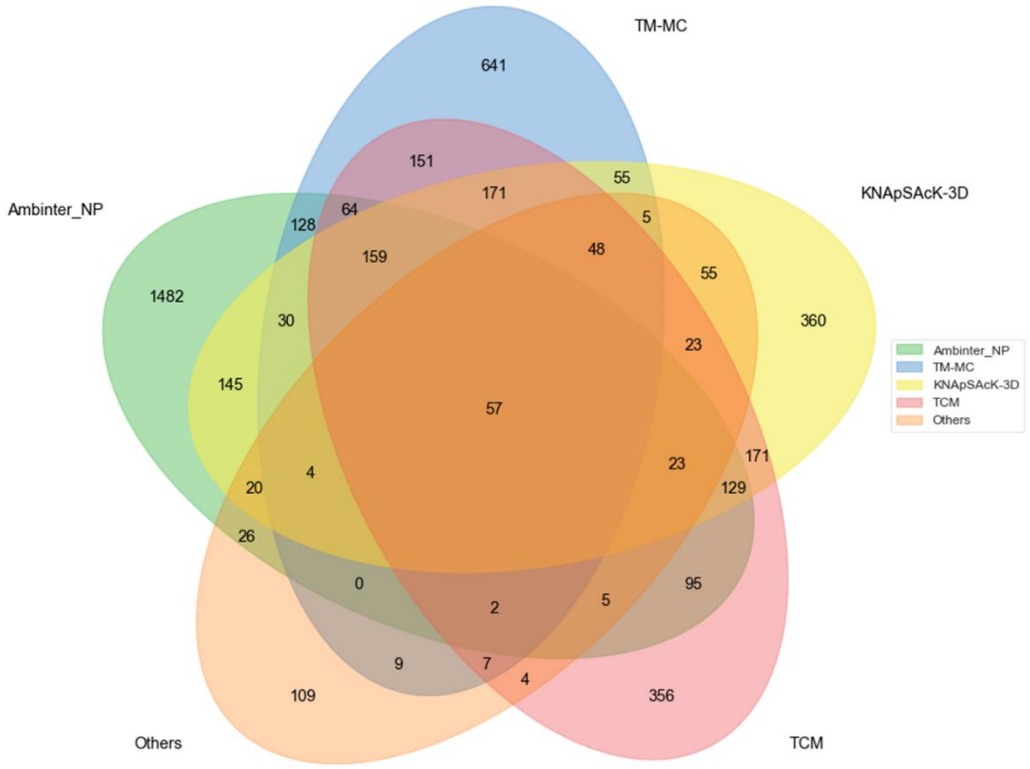


Figure 4. Venn diagram representing the origin of each molecule in our database of natural products. "Others" groups the databases containing less than 100 compounds (NuBBE, SANCDB, StreptomeDB and Biophytmol). "Ambinter_NP" refers to Ambinter's natural products.

2.2 Preparation of a SIK2 structure

In a VS project, two things are essential: a library of molecules and a 3D structure of the target. As no experimentally determined structures are available for SIK2 human kinase in the PDB [33], we created one by homology modelling. Homology modelling is a technique allowing the prediction of a protein 3D structure from its sequence thanks to a template. Two prerequisites are important for a good model construction: a high sequence identity between the target and the template and a correct superposition between them. Since it has been shown that 3D structure of proteins in a family are more conserved than sequence itself, a good template needs to be a protein evolutionarily related to the target [34].

Nowadays, several program are available for homology modelling [35]. Our model was built using Modeller [36] and the SIK2 sequence retrieved from UniProt database (ID: Q9H0K1) [37]. The reference template is a MAP/microtubule affinity-regulating kinase 4 (PDB code: 5ES1, resolution of 2.8 Å) [38], having a sequence similarity of 57% with SIK2. We independently evaluated our model with SwissModel "Structure Assessment" module. We thus obtained a MolProbity score of 2.97 and the Ramachandran plot returned 91.8% of favored position of amino acids, with three outliers (this plot is given in supplemental information) [39]. The global RMSD (root mean square deviation) of our model with the template calculated on backbone is 0.38 Å.

Protein kinases are flexible and may adopt several conformations. The consideration for kinase conformations is critical in the development of targeted kinase inhibitors, especially since there are different types for these kind of inhibitors [40]. To know the conformation of our model, an inspection of the configurations of the αC-helix and the conserved DFG motif into the ATP binding pocket was carried out. Depending on the positions of their amino acids, these motifs can either be in conformation "in" or in conformation "out". When both are in conformation "in", i.e. αC-in/DFG-in, the kinase is in

active state. All other possible configurations, i.e. αC-in/DFG-out, αC-out/DFG-in and αC-out/DFG-out represent the inactive states [41]. Each conformation exhibits a unique shape allowing or not the binding of ATP and the access to a back pocket. The main characteristics of an active conformation is the opening of the binding site, with the knockback of the activation loop, the orientation of the αC helix inward toward the active site and a ion-pair interaction between its conserved Glu and the Lys of the β3 strand. Another important feature is the orientation of the Phe from DFG motif inward toward an allosteric back pocket. In inactive conformation, the interaction between Lys and Glu disappears and Phe flips by ~180° opening the way to the allosteric pocket [42].

The catalysis of ATP requires the precise positioning of highly conserved motifs. Therefore, the kinase active state is highly conserved [43]. A comparison between our model and a reference structure (PDB code: 1ATP, chain E) [44] ensured that it was in active conformation. As seen in the Figure 5, our model presents a αC-in/DFG-in conformation. Kinase inhibitors targeting the active conformation of a kinase are categorized type I or $I^{1/2}$ [45].

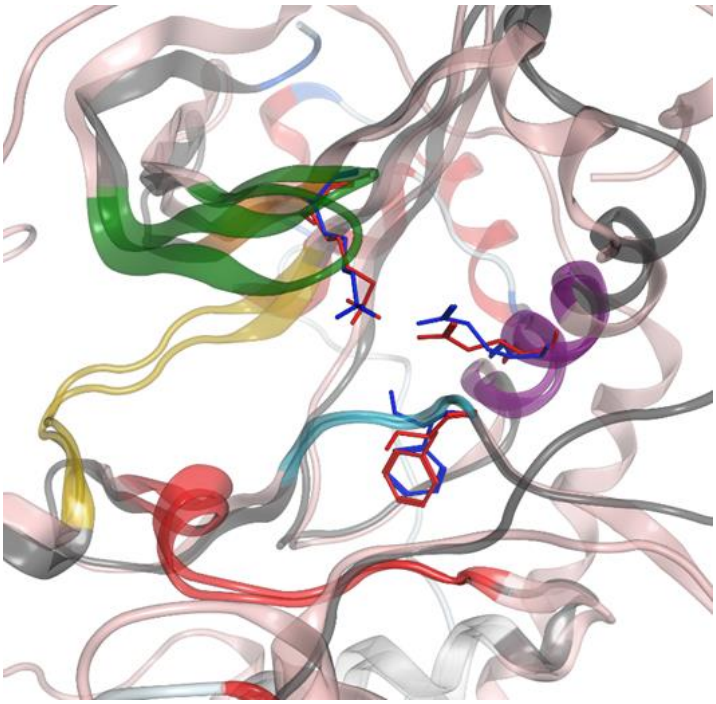


Figure 5. Superposition of the model and a protein kinase in active conformation (PDB: 1ATP, chain E). In blue: DFG motif, in purple: αC-helix, RMSD between conserved motifs is 1.1 Å. αC-helix's Glu interacts with catalytic lysine and Phe from DFG is oriented toward (DFG-in, αC-helix in).

2.3 Virtual screening

2.3.1 Validation of the method

The goal of this part was to find and select the best protocol to dock our natural product. After a visual inspection of the model superposed to the template, we observed that the original ligand, co-crystallized in the structure of PDB 5ES1, interacts similarly with our model and perfectly fits in the cavity. Thus, to choose a docking method for the VS, we first test which software could most faithfully reproduce the original inhibitor's experimental pose. Three different docking software were compared: Glide [46], GOLD [47] and rDock [48]. For consistency, the cavity was generated by the same way for all docking programs, by considering an area in a 6 Å spherical radius around the original co-crystallized ligand superposed on our model. Since no water molecule was involved in the binding mode, none were kept in the model. Once the cavity determined, we performed the docking of the original ligand in the three software. To evaluate performances two parameters were verified: the ability

to predict poses close to the experimental one (RMSD < 2 Å) and their rank according to the native score function. If not ranked first, RMSD value of the first ranked pose were also verified. The score function chosen for Glide and rDock, SP and S^{inter} respectively, are empirical ones, meaning they were determined via regression analysis of experimental dataset of protein-ligand complexes [46,48]. While for GOLD, the ASP fitness function score belongs to knowledge-based ones and is derived from an atom-atom potential also calculated from a database of protein-ligand complexes [49].

The docking results are summarized in the Table 2. The best pose, determined by lowest RMSD, and its rank by the docking program's scoring function is given in column 2. The column 3 lists the RMSD value of the top scored pose. For both column GOLD returned best results. Its best RMSD value over all poses is 1.02 Å, ranked third in score. Its first ranked pose has a RMSD value of 1.47 Å. Surprisingly, Glide failed to return an acceptable RMSD value of pose, contrary to rDock, which returned satisfying results. The Figure 6 illustrates the best pose according to RMSD value obtained with GOLD versus the original co-crystallized ligand from the template, the key interactions with the hinge, in yellow and the Lys from conserved motif AxK in β3-strand, in orange, are preserved.

Table 2. RMSD value and ranking results between the different software

Software (score function)	Best pose (Rank score)	1st ranked pose RMSD value
GLIDE (SP)	2.5 Å (5)	3.9 Å
rDock (S^{inter})	1.49 Å (7)	1.68 Å
Gold (ASP)	1.06 Å (3)	1.47 Å

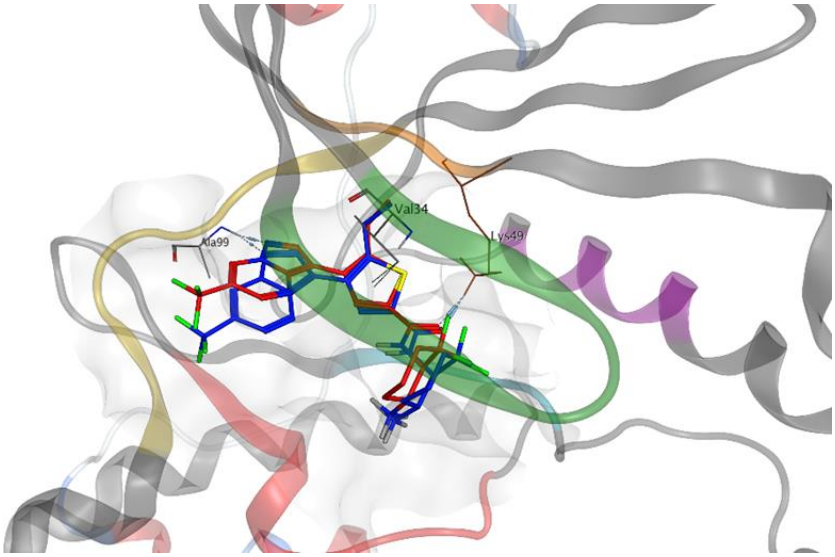


Figure 6. Original ligand co-crystallized from PDB 5ES1 in red and docking pose obtained with the best RMSD (1.06 Å) in blue.

Then, to assess the enrichment in active molecules, a dataset composed of 33 active compounds and 200 decoys (ratio: 0.17) was used. Here, the performance were judged based on the ability of the software to discriminate active compounds from inactive ones by using score value, illustrated by the value of area under the receiver operating characteristic curve (AUC). The higher the value of AUC, the better the enrichment of the true-positives compared to a random method where AUC would be equal to 0.5.

Several different protocols were tested to find out the one exhibiting the best enrichment (see supplemental information for more results). Once again, GOLD returned the best results with an AUC

of 0.81 (Figure 7) and an enrichment factor at 1% of 7.03. These results were obtained using VSPrep [50] to prepare the molecules for the virtual screening.

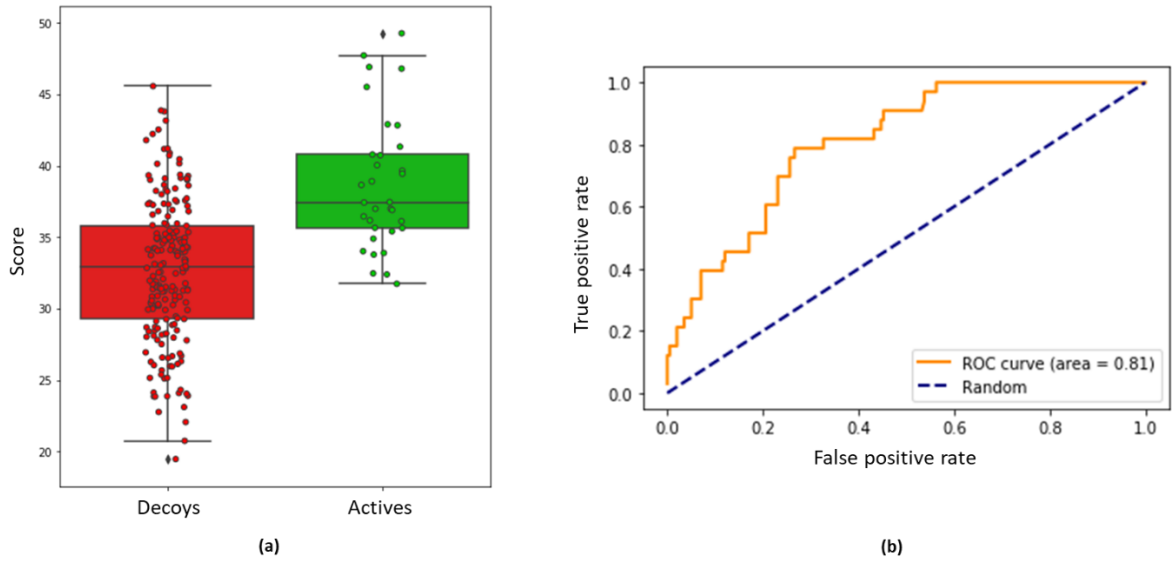


Figure 7. Docking's results between decoys (red) and actives molecules (green): (a) docking score dispersion; (b) ROC-curve plot.

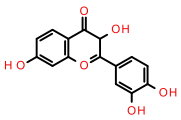
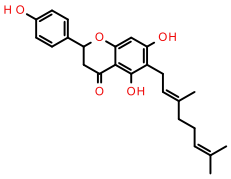
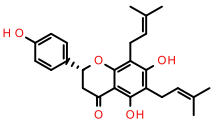
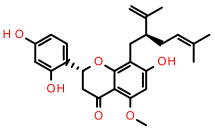
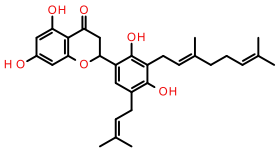
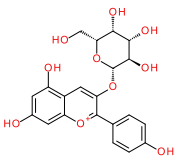
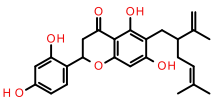
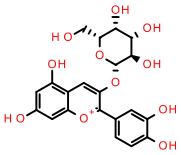
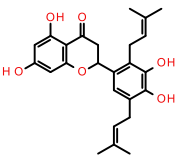
2.3.2 Docking of our natural product database

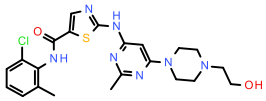
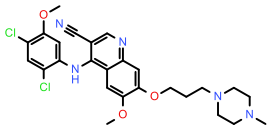
As the enrichment studies preconized, the molecules from the natural products database were prepared following the protocol with VSPrep. With the calculation of the different tautomeric forms, the total of molecules in our database roses to 10,939. The virtual screening of the entire database took about three days on a standard computer. At the rate of six poses returned per molecule, we ended with 65,586 poses meaning that GOLD was unable to perform calculation docking for eight compounds. The distribution range of score goes from -11.72 to 58.04 with an average of 22.94. The average difference of score between the six poses of a same molecule is 1.69, showing a consistency in GOLD scoring results.

Three known synthetic inhibitors were randomly added as witness for results: HG-9-91-01, YKL-06-061 and YKL-06-062 [15]. Their best poses were all ranked in top 100 with a score of 42.49 (47th), 42.04 (54th) and 41.78 (56th) respectively, proving the robustness of the docking method. Surprisingly, while inspecting the top 100th compound, a compound focused our attention. A synthetic kinase inhibitor not placed on purpose, the gefitinib, was retrieved on 35th position with a score of 43.43. After further investigation, it appeared that it came from the database TM-MC native from A. Radix (http://informatics.kiom.re.kr/compound/detail.do?id=gefitinib). Even if not a natural product, it came as a proof that our method was also able to score kinase inhibitors at the top.

Only the top 1,000 poses were considered for further analysis. For the hit selections, multiple parameters were checked including physico-chemical properties, interactions with the receptor but also marketing issue as the origin of the compound and the availability of intellectual property. Finally, after a rigorous visual inspection of remaining poses, 25 molecules were selected for experimental testing. Unfortunately, when purchasing the compounds, more than half of them were temporarily unavailable. All molecules sent for experimental test are recapitulated in Table 3. We also added two type I kinase inhibitors as positive test: bosutinib and dasatinib [32].

Table 3. Compounds selected for experimental test and their characteristics

ID	Molecule	CAS number	MW	CLogP	HBA	HBD	Score	Rank
Fisetin		528-48-3	286.05	2.28	6	4	44.25	23
6-geranylnaringenin		97126-57-3	408.19	5.74	5	3	37.17	148
6,8-diprenylnaringenin		68236-11-3	408.19	5.52	5	3	37.93	122
Kuraninone		34981-26-5	438.20	5.61	6	3	37.29	140
Sanggenol-P		1351931-30-0	492.25	6.96	6	4	41.33	62
Fragarin		574-57-2	468.08	-2.32	9	7	40.61	75
Kushenol F		34981-24-3	424.19	5.31	6	4	37.18	145
Cyanidin-3-O-galactoside		27661-36-5	484.08	-2.61	10	8	40.14	80
Sigmoidin A		176046-04-1	424.19	5.23	6	4	38.76	107

Dasatinib		302962-49-8	487.16	3.31	9	3	-	-
Bosutinib		1391063-17-4	529.17	5.19	8	1	-	-

The results of the experimental tests are presented in the Table 4. In addition, the molecules were not only tested versus SIK2 kinase but also against a panel of other kinases including BRAF and mutated BRAFV600E.

Table 4. Residual activity of tested kinases against natural products selected.

ID	SIK2	BRAF	BRAF V600E
Fisetin	24 (K_d: 270 nM)	9.6 (K_d: 150 nM)	11 (K_d: 110 nM)
6-geranylnaringenin	100	97	95
6,8-diprenylnaringenin	100	75	91
Kuraninone	100	88	86
Sanggenol- P	100	89	87
Fragarin	100	82	93
Kushenol F	100	88	86
Cyanidin-3-O-galactoside	100	90	97
Sigmoidin A	100	79	91
Dasatinib	0 (K_d: 3.9 nM)	0 (K_d 500 nM)	0.3 (Kd 570 nM)
Bosutinib	0 (K_d: 30 nM)	30 (K_d > 3000 nM)	35 (K_d > 3000 nM)

Unfortunately, despite a carefully validated docking protocol, experimental results were not as good as expected. Indeed, only the fisetin presented an acceptable rate of inhibition of 76% (K_d: 270 nM) against SIK2. No other compounds tested returned acceptable results (K_d > 10,000). The compounds used as positive tests, bosutinib and dasatinib, fulfilled their function with both 100% inhibition (K_d: 30 nM et 3.9 nM respectively).

Regarding the other targets, almost all molecules showed activity with around 10% of inhibition. Fisetin presents the highest inhibition rate with 90% on both wild-type (WT) and mutated BRAF, followed with the 6,8-diprenylnaringenin that inhibits WT BRAF at 25% and Kushenol F inhibiting mutated BRAFV600E at 15%.

3. Discussion

UVR is responsible for various physiological effects on the skin, resulting in pigmentation alterations. In response, the melanin secreted by the melanocytes acts as a photoprotective shield inducing skin tanning. Indoor tanning is based on UVR and therefore does not allow the avoiding of DNA damage contrarily to sunless tanning. Topical application of natural products enhancing the melanogenesis remains an opportunity for cosmetic products to reduce visible sign of photoaging and keep a golden and uniform tan. SIK kinase family are targets of importance to trigger melanogenesis, their inhibition can independently activate the tanning pathway without suffering from the damaging effects of UV.

The increase in the cost and time of the drug development process have led to greater usage of VS approaches. With advances in the field via computing power rising and better consideration of receptor flexibility, these methods tend to become interesting alternatives to traditional screening methods for compounds prioritization. Before running a VS campaign, some preparation steps are important to retrieve the methodology returning best enrichment results. In our case, several different protocols and multiple software were tested. Among the three software used, GOLD was the one giving the best results while Glide could not satisfy our validation request. The free software rDock showed interesting potential and proved it can be an interesting tool for screening campaign.

Nevertheless, even with a meticulous validated strategy and a manual selection of hits after the VS, experimental results did not meet our expectations. Indeed, our methodology failed to find new potential SIK2 inhibitors, the fisetin being already known as SIK inhibitor [51]. Multiple factors can explain it, starting with the 3D structure of the target obtained by homology modelling and not experimentally, that might not be as accurate as an experimental structure. Moreover, since no structural information were available, the implication of water molecules in the binding site could not be checked and the VS campaign was performed without taking them into account. Another improvement could be to construct several models with different templates and compare them to get a consensus. Although our goal was to find type 1 inhibitors, realizing a VS on other conformations of the kinase might confirm or reveal other hits. The other important feature that might explain the lack of inhibition is the skeleton of selected compounds for experimental test. Most of the flavonoids chosen scaffolds are flavan based and not flavone based as fisetin or quercetin. However it seems that flavon scaffold gives better results on SIK kinase family [51]. Yet, the poor number of experimentally tested molecules (9) does not permit to conclude that our strategy is wrong. As soon as other molecules will be available for purchase, they will be tested to validate or not their inhibition. If proven, this strategy could be used on other kinases implicated in melanogenesis, as for example KIT, to discover natural product inhibitors [9].

Natural products remain an interesting class of compounds for marketing asset in cosmetic and discovering ones able to upstream melanogenesis would be valuable. Moreover, as seen in the study, they can cover a different chemical space than kinase inhibitors, which is an important point since the intellectual property space of kinase inhibitors is crowded. However, the consideration for toxicity and safety for such compounds might raise interrogation. Previous results in mice over multiple months of treatment did not show any apparent associated toxicities [52]. Furthermore, the increase of melanin, especially for fair-skinned people, could act as an extra-protection again skin melanoma. Of note, even if some compounds exist for sunless tanning, they are not substitute to sunscreen, which remains one of the most efficient tools for optimal skin protection against skin photoaging and damage from UVR.

4. Materials and Methods

The natural products databases were all downloaded from their website or by request to their administrators. The molecules were standardized using VSPrep before being grouped. SMILES formulas were calculated using RDKit (version '2018-09-01').

All experiment and calculations on the molecules have been made with Python 3.6. We calculated the molecular descriptors using RDKit (version '2018-09-01'). The venn diagram was constructed using pyvenn (https://github.com/tctianchi/pyvenn).

The homology model was built using modeler version 9.16-1. It was prepared and refined for virtual screening with the "Structure Preparation" module from MOE (Chemical Computing Group, version 2018_01).

The three docking software versions are GOLD 5.2.2, Glide 7.5 and rDock 2013.1. rDock is freely available at http://rdock.sourceforge.net/, the two others are commercial and need a license. For the redocking of co-crystallized ligand, a new conformation was calculated using ETKDG method [53] followed by an optimization step using the MMFF94 forcefield [54]. The basic parameters of each program remained untouched. For rDock, we used the module "dock_solv" instead of the default setting "dock". The number of poses to return was ten in each case. For the enrichment studies, the active molecules dataset came from ZINC15 database and decoys from DUD-E [55] and a random

selection from dataset for JAK-2 kinase. For the virtual screening, all the molecules were fully prepared with VS-Prep. RDKit and Pandas were used to treat the results and select the best molecules. Poses visualization and interactions calculation were realized using MOE software (Chemical Computing Group, version 2018_01).

Experimental tests were conducted by the company DiscoverX with their KINOMEScan™ assay technology as competition binding assay at 10 μM concentration .

All the figures were made using matplotlib [56], seaborn [57] or MOE (Chemical Computing Group, version 2018_01). Molecules in 2D were drawn with Biovia Draw 2018.

Author Contributions: conceptualization, C.B., J.-M.G and P.B.; methodology, C.B. and J.-M.G.; formal analysis, C.B. and J.-M.G.; writing—original draft preparation, C.B.; writing—review and editing, S.A.-S. and P.B.; supervision, S.A.-S. and P.B.

Abbreviations

α-MSH	α-melanocyte-stimulating hormone
ATP	adenosine triphosphate
AUC	area under the curve
AMPK	AMP-activated protein kinase
BRAF	v-RAF murine sarcoma viral oncogene homolog B
cAMP	cyclic adenosine monophosphate
CREB	cAMP response element-binding protein
DFG	Asp-Phe-Gly
DNA	deoxyribonucleic acid
K_d	dissociation constant
KIT	mast/stem cell growth factor receptor
L-DOPA	L-3,4-dihydroxyphenylalanine
MC1R	melanocortin 1 receptor
MITF	microphtalmia-associated transcription factor
PDB	Protein Data Bank
RMSD	root-mean-square deviation
POMC	proopiomelanocortin
PKA	protein kinase A
SMILES	Simplified Molecular Input Line Entry Specification
SIK	salt-inducible kinase
UV	ultraviolet
UVR	ultraviolet radiation

References

1. Gilaberte, Y.; Prieto-Torres, L.; Pastushenko, I.; Juarranz, Á. Chapter 1 - Anatomy and Function of the Skin. In *Nanoscience in Dermatology*; Hamblin, M.R., Avci, P., Prow, T.W., Eds.; Academic Press: Boston, 2016; pp. 1–14 ISBN 978-0-12-802926-8.
2. Feetham, H.J.; Jeong, H.S.; McKesey, J.; Wickless, H.; Jacobe, H. Skin care and cosmeceuticals: Attitudes and trends among trainees and educators. *Journal of Cosmetic Dermatology* **2018**, *17*, 220–226.
3. Skin care industry: global skincare market size 2012-2024 Available online: https://www.statista.com/statistics/254612/global-skin-care-market-size/ (accessed on Jun 27, 2019).
4. Tobin, D.J. Introduction to skin aging. *Journal of Tissue Viability* **2017**, *26*, 37–46.
5. Brenner, M.; Hearing, V.J. The Protective Role of Melanin Against UV Damage in Human Skin. *Photochem Photobiol* **2008**, *84*, 539–549.
6. Thody, A.J.; Higgins, E.M.; Wakamatsu, K.; Ito, S.; Burchill, S.A.; Marks, J.M. Pheomelanin as well as eumelanin is present in human epidermis. *J. Invest. Dermatol.* **1991**, *97*, 340–344.

7. Cui, R.; Widlund, H.R.; Feige, E.; Lin, J.Y.; Wilensky, D.L.; Igras, V.E.; D'Orazio, J.; Fung, C.Y.; Schanbacher, C.F.; Granter, S.R.; et al. Central role of p53 in the suntan response and pathologic hyperpigmentation. *Cell* **2007**, *128*, 853–864.
8. Bertolotto, C.; Abbe, P.; Hemesath, T.J.; Bille, K.; Fisher, D.E.; Ortonne, J.P.; Ballotti, R. Microphthalmia gene product as a signal transducer in cAMP-induced differentiation of melanocytes. *J. Cell Biol.* **1998**, *142*, 827–835.
9. D'Mello, S.A.N.; Finlay, G.J.; Baguley, B.C.; Askarian-Amiri, M.E. Signaling Pathways in Melanogenesis. *Int J Mol Sci* **2016**, *17*.
10. Ando, H.; Niki, Y.; Ito, M.; Akiyama, K.; Matsui, M.S.; Yarosh, D.B.; Ichihashi, M. Melanosomes Are Transferred from Melanocytes to Keratinocytes through the Processes of Packaging, Release, Uptake, and Dispersion. *Journal of Investigative Dermatology* **2012**, *132*, 1222–1229.
11. Del Bino, S.; Duval, C.; Bernerd, F. Clinical and Biological Characterization of Skin Pigmentation Diversity and Its Consequences on UV Impact. *International Journal of Molecular Sciences* **2018**, *19*, 2668.
12. Chang, C.; Murzaku, E.C.; Penn, L.; Abbasi, N.R.; Davis, P.D.; Berwick, M.; Polsky, D. More Skin, More Sun, More Tan, More Melanoma. *Am J Public Health* **2014**, *104*, e92–e99.
13. Horike, N.; Kumagai, A.; Shimono, Y.; Onishi, T.; Itoh, Y.; Sasaki, T.; Kitagawa, K.; Hatano, O.; Takagi, H.; Susumu, T.; et al. Downregulation of SIK2 expression promotes the melanogenic program in mice. *Pigment Cell & Melanoma Research* **2010**, *23*, 809–819.
14. Clark, K.; MacKenzie, K.F.; Petkevicius, K.; Kristariyanto, Y.; Zhang, J.; Choi, H.G.; Peggie, M.; Plater, L.; Pedrioli, P.G.A.; McIver, E.; et al. Phosphorylation of CRTC3 by the salt-inducible kinases controls the interconversion of classically activated and regulatory macrophages. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 16986–16991.
15. Mujahid, N.; Liang, Y.; Murakami, R.; Choi, H.G.; Dobry, A.S.; Wang, J.; Suita, Y.; Weng, Q.Y.; Allouche, J.; Kemeny, L.V.; et al. A UV-Independent Topical Small-Molecule Approach for Melanin Production in Human Skin. *Cell Reports* **2017**, *19*, 2177–2184.
16. Wein, M.N.; Foretz, M.; Fisher, D.E.; Xavier, R.J.; Kronenberg, H.M. Salt-Inducible Kinases: Physiology, Regulation by cAMP, and Therapeutic Potential. *Trends Endocrinol. Metab.* **2018**, *29*, 723–735.
17. Kumar, V. Perspective of Natural Products in Skincare. *PPIJ* **2016**, *4*.
18. Fatima A, Shashi Alok, Agarwal P, Singh PP and Verma A: Benefits of Herbal extracts in Cosmetics. Int J Pharm Sci Res 2013: 4(10); 3746-3760. doi: 10.13040/IJPSR. 0975-8232.4(10).3746-60.
19. Fowler, J.F.; Woolery-Lloyd, H.; Waldorf, H.; Saini, R. Innovations in natural ingredients and their use in skin care. *J Drugs Dermatol* **2010**, *9*, S72-81; quiz s82-83.
20. Saewan, N.; Jimtaisong, A. Natural products as photoprotection. *Journal of Cosmetic Dermatology* **2015**, *14*, 47–63.
21. Choi, M.-H.; Shin, H.-J. Anti-Melanogenesis Effect of Quercetin. *Cosmetics* **2016**, *3*, 18.
22. Yoon, H.-S.; Lee, S.-R.; Ko, H.-C.; Choi, S.-Y.; Park, J.-G.; Kim, J.-K.; Kim, S.-J. Involvement of extracellular signal-regulated kinase in nobiletin-induced melanogenesis in murine B16/F10 melanoma cells. *Biosci. Biotechnol. Biochem.* **2007**, *71*, 1781–1784.
23. Lionta, E.; Spyrou, G.; Vassilatis, D.K.; Cournia, Z. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Curr Top Med Chem* **2014**, *14*, 1923–1938.
24. Sterling, T.; Irwin, J.J. ZINC 15 – Ligand Discovery for Everyone. *J. Chem. Inf. Model.* **2015**, *55*, 2324–2337.
25. Sharma, A.; Dutta, P.; Sharma, M.; Rajput, N.K.; Dodiya, B.; Georrge, J.J.; Kholia, T.; Bhardwaj, A.; OSDD Consortium BioPhytMol: a drug discovery community resource on anti-mycobacterial phytomolecules and plant extracts. *Journal of Cheminformatics* **2014**, *6*, 46.
26. Nakamura, K.; Shimura, N.; Otabe, Y.; Hirai-Morita, A.; Nakamura, Y.; Ono, N.; Ul-Amin, M.A.; Kanaya, S. KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites. *Plant Cell Physiol* **2013**, *54*, e4–e4.
27. Valli, M.; dos Santos, R.N.; Figueira, L.D.; Nakajima, C.H.; Castro-Gamboa, I.; Andricopulo, A.D.; Bolzani, V.S. Development of a natural products database from the biodiversity of Brazil. *J. Nat. Prod.* **2013**, *76*, 439–444.
28. Hatherley, R.; Brown, D.K.; Musyoka, T.M.; Penkler, D.L.; Faya, N.; Lobb, K.A.; Tastan Bishop, Ö. SANCDB: a South African natural compound database. *Journal of Cheminformatics* **2015**, *7*, 29.

29. Klementz, D.; Döring, K.; Lucas, X.; Telukunta, K.K.; Erxleben, A.; Deubel, D.; Erber, A.; Santillana, I.; Thomas, O.S.; Bechthold, A.; et al. StreptomeDB 2.0—an extended resource of natural products produced by streptomycetes. *Nucleic Acids Res* **2016**, *44*, D509–D514.
30. Chen, C.Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening In Silico. *PLOS ONE* **2011**, *6*, e15939.
31. Kim, S.-K.; Nam, S.; Jang, H.; Kim, A.; Lee, J.-J. TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine. *BMC Complement Altern Med* **2015**, *15*.
32. 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.
33. 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.
34. Kaczanowski, S.; Zielenkiewicz, P. Why similar protein sequences encode similar three-dimensional structures? *Theor Chem Acc* **2010**, *125*, 643–650.
35. Dalton, J.A.R.; Jackson, R.M. An evaluation of automated homology modelling methods at low target–template sequence similarity. *Bioinformatics* **2007**, *23*, 1901–1908.
36. Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr Protoc Bioinformatics* **2016**, *54*, 5.6.1-5.6.37.
37. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res* **2019**, *47*, D506–D515.
38. Sack, J.S.; Gao, M.; Kiefer, S.E.; Myers, J.E.; Newitt, J.A.; Wu, S.; Yan, C. Crystal structure of microtubule affinity-regulating kinase 4 catalytic domain in complex with a pyrazolopyrimidine inhibitor. *Acta Cryst F* **2016**, *72*, 129–134.
39. Chen, V.B.; Arendall, W.B.; Headd, J.J.; Keedy, D.A.; Immormino, R.M.; Kapral, G.J.; Murray, L.W.; Richardson, J.S.; Richardson, D.C. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* **2010**, *66*, 12–21.
40. Müller, S.; Chaikuad, A.; Gray, N.S.; Knapp, S. The ins and outs of selective kinase inhibitor development. *Nature Chemical Biology* **2015**, *11*, 818–821.
41. Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **2002**, *109*, 275–282.
42. Vijayan, R.S.K.; He, P.; Modi, V.; Duong-Ly, K.C.; Ma, H.; Peterson, J.R.; Dunbrack, R.L.; Levy, R.M. Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors. *J Med Chem* **2015**, *58*, 466–479.
43. Kornev, A.P.; Haste, N.M.; Taylor, S.S.; Ten Eyck, L.F. Surface comparison of active and inactive protein kinases identifies a conserved activation mechanism. *Proc Natl Acad Sci U S A* **2006**, *103*, 17783–17788.
44. 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*, 362–365.
45. Roskoski, R. Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes. *Pharmacological Research* **2016**, *103*, 26–48.
46. Friesner, R.A.; Banks, J.L.; Murphy, R.B.; Halgren, T.A.; Klicic, J.J.; Mainz, D.T.; Repasky, M.P.; Knoll, E.H.; Shelley, M.; Perry, J.K.; et al. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739–1749.
47. Jones, G.; Willett, P.; Glen, R.C.; Leach, A.R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727–748.
48. 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*, e1003571.
49. Mooij, W.T.M.; Verdonk, M.L. General and targeted statistical potentials for protein–ligand interactions. *Proteins: Structure, Function, and Bioinformatics* **2005**, *61*, 272–287.
50. 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*, 1700023.

51. Kumagai, A.; Horike, N.; Satoh, Y.; Uebi, T.; Sasaki, T.; Itoh, Y.; Hirata, Y.; Uchio-Yamada, K.; Kitagawa, K.; Uesato, S.; et al. A Potent Inhibitor of SIK2, 3, 3′, 7-Trihydroxy-4′-Methoxyflavon (4′-O-Methylfisetin), Promotes Melanogenesis in B16F10 Melanoma Cells. *PLoS One* **2011**, *6*.
52. D'Orazio, J.A.; Nobuhisa, T.; Cui, R.; Arya, M.; Spry, M.; Wakamatsu, K.; Igras, V.; Kunisada, T.; Granter, S.R.; Nishimura, E.K.; et al. Topical drug rescue strategy and skin protection based on the role of Mc1r in UV-induced tanning. *Nature* **2006**, *443*, 340–344.
53. 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.
54. Halgren, T.A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **1996**, *17*, 490–519.
55. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594.
56. 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;
57. 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;

Article

Virtual screening of natural products to enhance melanogenesis

Colin Bournez [1], José-Manuel Gally [1], Samia Aci-Sèche [1], Philippe Bernard[2], Pascal Bonnet [1,*]

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

[2] Greenpharma SAS, 3 allée du Titane 45100 Orléans, France

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

SUPPLEMENTARY INFORMATION

Preparation of natural products database

The Table 1 recapitulates the molecular descriptors used for PCA calculation, calculated with RDKit (*version 2018.01*).

Table 1. 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
LabuteASA	Labute's Approximate Surface Area
NAR	Number of aromatic rings
FCSP3	Fraction of C atoms that are SP3 hybridized
MQN8	Molecular Quantum Numbers
MQN10	Molecular Quantum Numbers

The PCA plot in Figure 1 illustrates the chemical space of the seven databases in a 2D reference frame represented by the two first principal components (PC1 and PC2). The two first principal components explain 55.6% and 14.9% of the total variance respectively. Thus, the 2D scatterplot showed here represents around 70.5% of the variance, an acceptable value avoiding an important loss of information. The graphical representation of normalized variables is shown in the correlation circle associated.

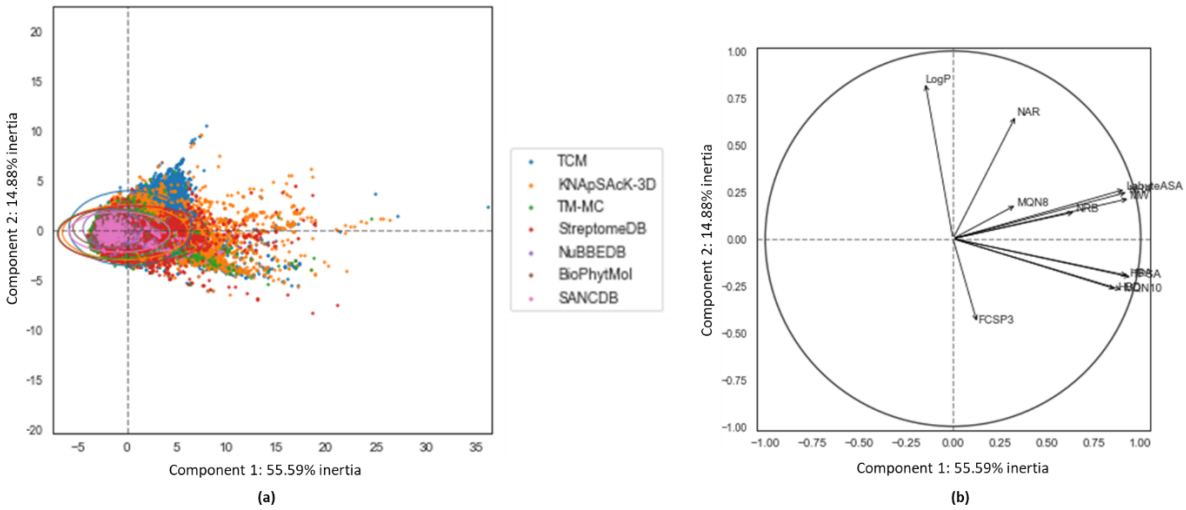


Figure 1. (a) PCA of molecules from the seven databases chosen. Colored ellipses encompass 95% of the individuals from their class respectively; (b) Correlation circle.

The Figure 2 represents the largest molecule in all database gathered, retrieved from ChEMBL database.

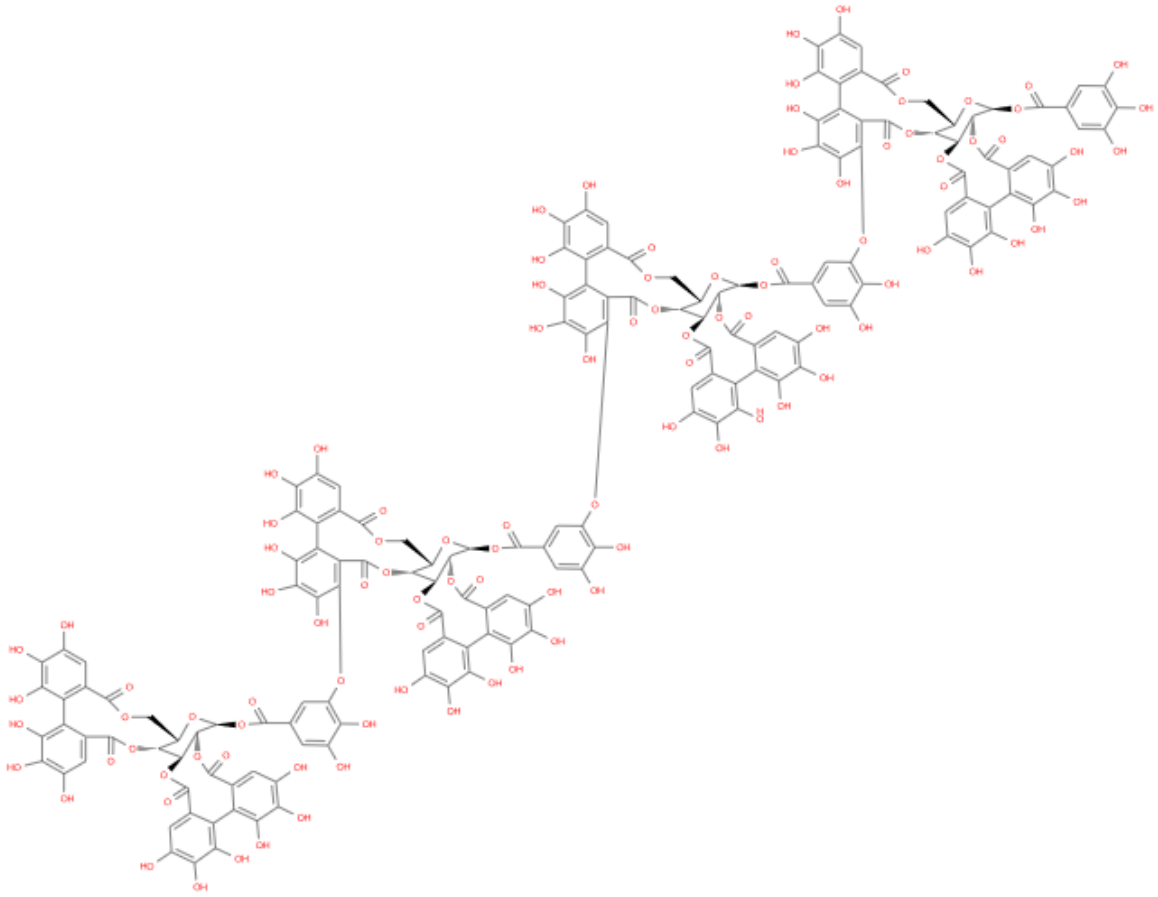


Figure 2. Biggest molecule retrieved from gathered databases (CHEMBL3215316, MW = 3,738.30 Da). It corresponds to a tetrameric ellagitanin.

The Figure 3 is the pairwise matrix of duplicates based on SMILES formula of the compounds.

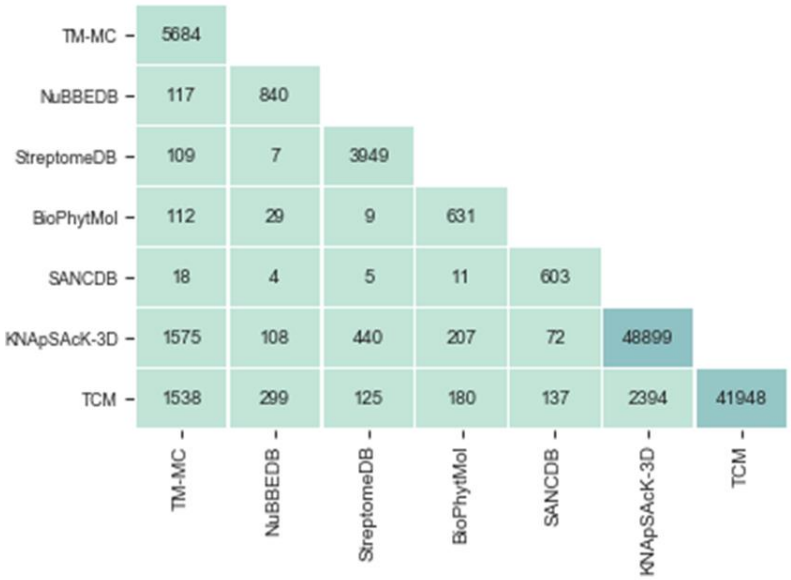


Figure 3. Pairwise duplicate molecules matrix

The PCA plot in Figure 4 illustrates the chemical space of the molecules from our database (green) versus the PKIs approved (blue) or in clinical trials (red) in a 2D reference frame represented by the two first principal components (PC1 and PC2). The two first principal components explain 42.5% and 20.9% of the total variance respectively. Thus, the 2D scatterplot showed here represents around 62.4% of the variance, an acceptable value avoiding an important loss of information. The graphical representation of normalized variables is shown in the correlation circle associated.

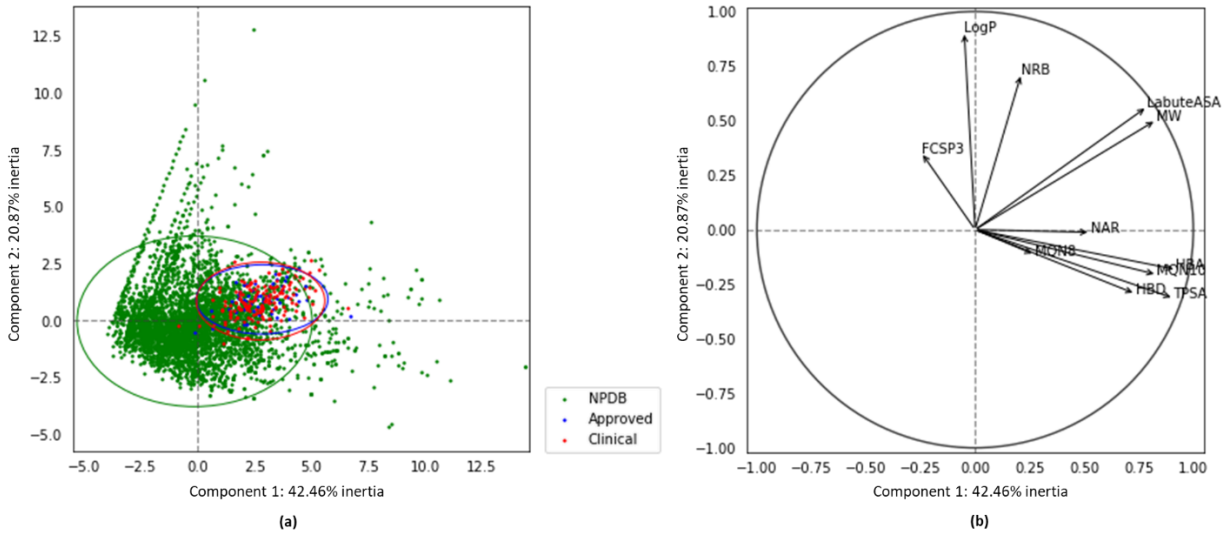


Figure 4. (a) PCA of PKIS in clinical trials (red), approved (blue) and natural products (green). Colored ellipses encompass 95% of the individuals from their class respectively; (b) Correlation circle.

The PMI plot in Figure 5 illustrates the shape diversity of the molecules in our database (green) versus PKIs approved (blue) or in clinical trials (red). The three corners represent distinctive shapes: rod (represented by diacetylene), disk (benzene) and sphere (adamantane).

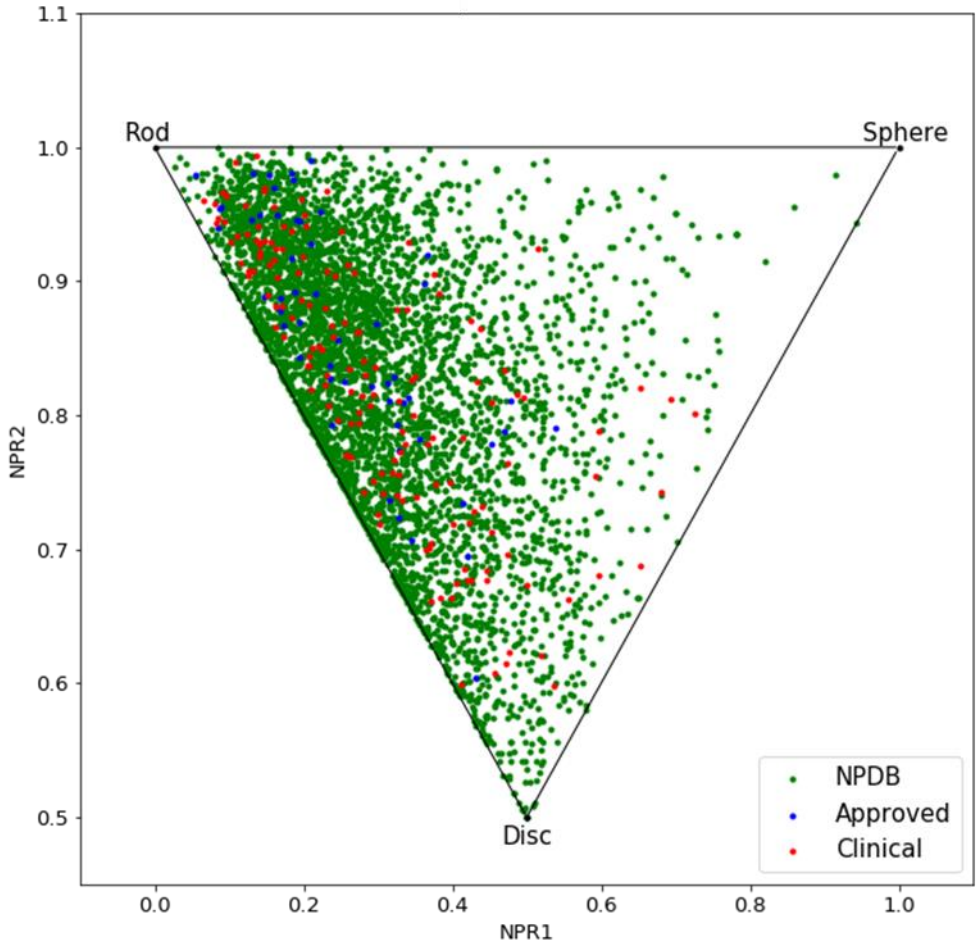


Figure 5. Principal Moments of Inertia (PMI) plot of PKIs in clinical trials (red), approved (blue) and molecules of NPDB (green).

Homology model validation

The Figure 6 shows the Ramachandran plot of our structure obtained by homology modeling with 91.8% favoured position of amino acids and three outliers.

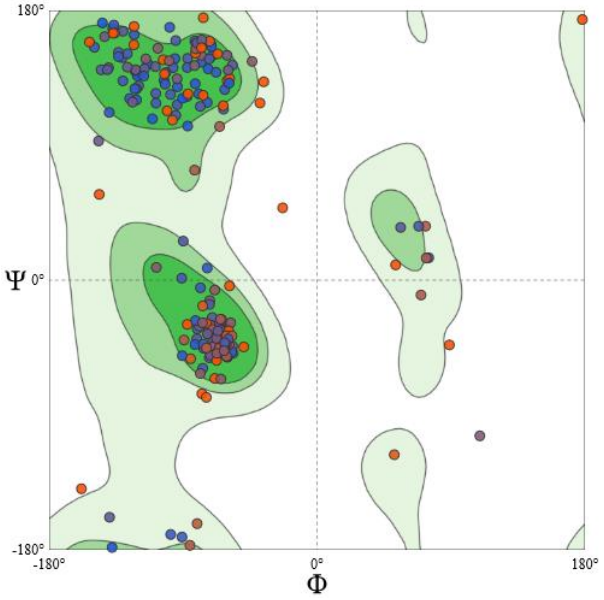


Figure 6. Ramachandran plot of SIK2 homology model from SWISS-MODEL "Structure Assessment" tool.

Docking enrichment studies

Docking with rDock

The Figure 7 shows the results obtained with the preparation and standardization of the molecules using RDKit and ETKDG method followed with an optimization using the MMFF94 forcefield. The score is turned in absolute (the S_{inter} score function of rDock returns a negative one).

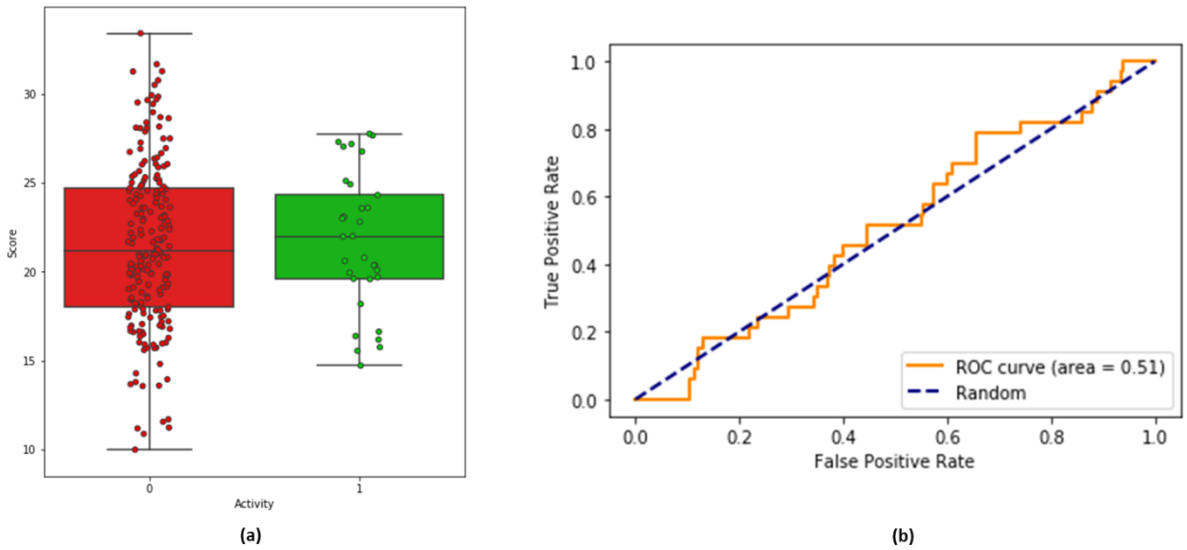


Figure 7. Docking results between decoys (red) and actives molecules (green) obtained with rDock from molecules prepared with RDKit: (a) absolute docking score dispersion; (b) ROC-curve plot.

The Figure 8 shows the results obtained with the preparation and standardization of the molecules using VS-Prep. The score is turned in absolute (the S_{inter} score function of rDock returns a negative one).

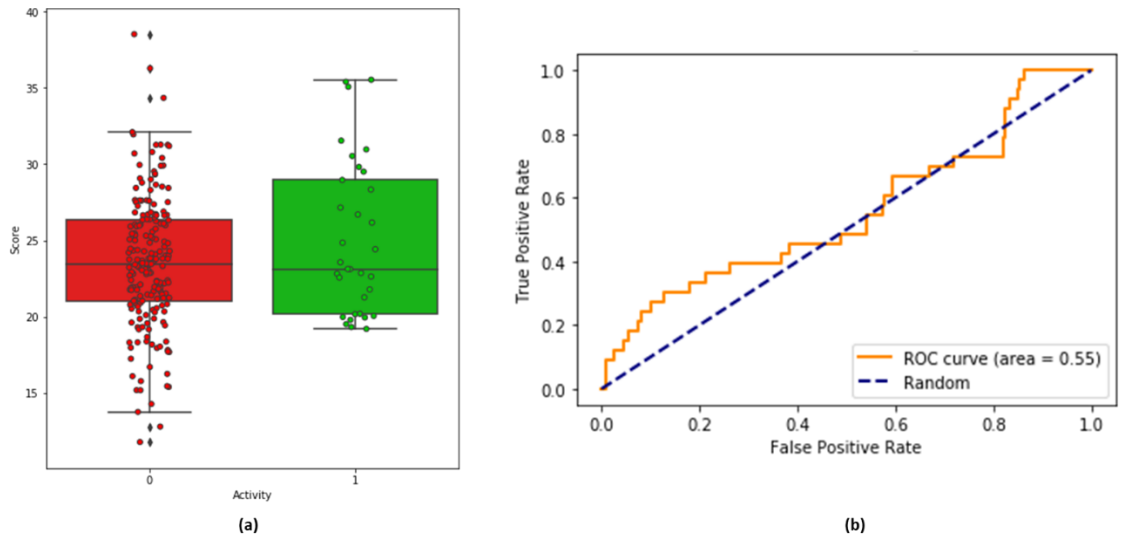


Figure 8. Docking results between decoys (red) and actives molecules (green) obtained with rDock from molecules prepared with VS-Prep: (a) absolute docking score dispersion; (b) ROC-curve plot.

Docking with Glide

The Figure 9 shows the results obtained with the preparation and standardization of the molecules using RDKit and ETKDG method followed with an optimization using the MMFF94 forcefield. The score is turned in absolute (the SP score function of Glide returns a negative one).

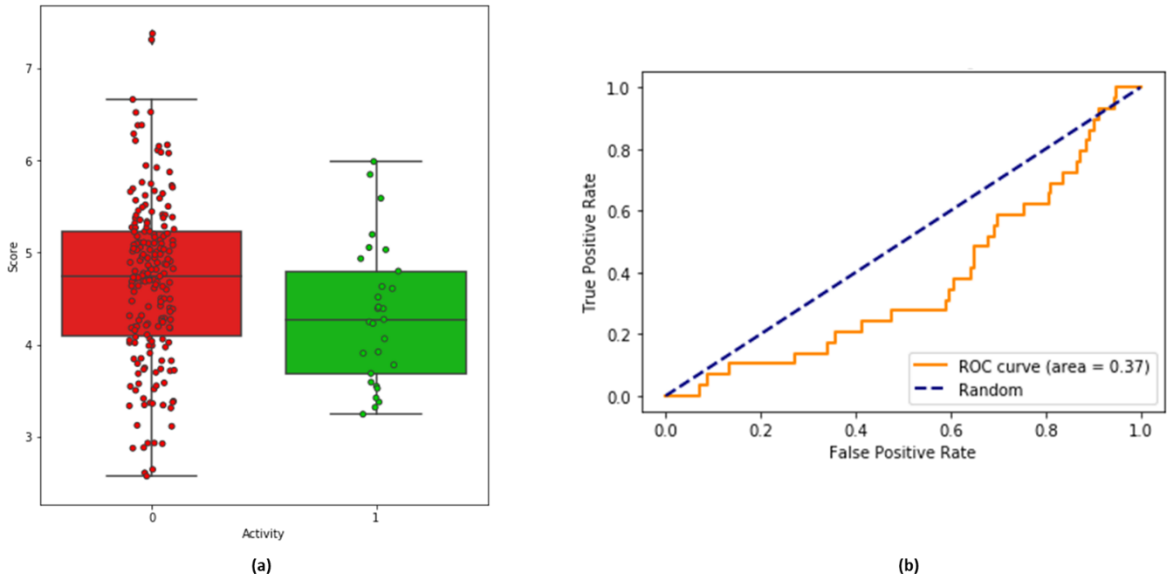


Figure 9. Docking results between decoys (red) and actives molecules (green) obtained with Glide from molecules prepared with RDKit: (a) absolute docking score dispersion; (b) ROC-curve plot.

The Figure 10 shows the results obtained with the preparation and standardization of the molecules using LigPrep from Maestro (Schrödinger). The score is turned in absolute (the SP score function of Glide returns a negative one).

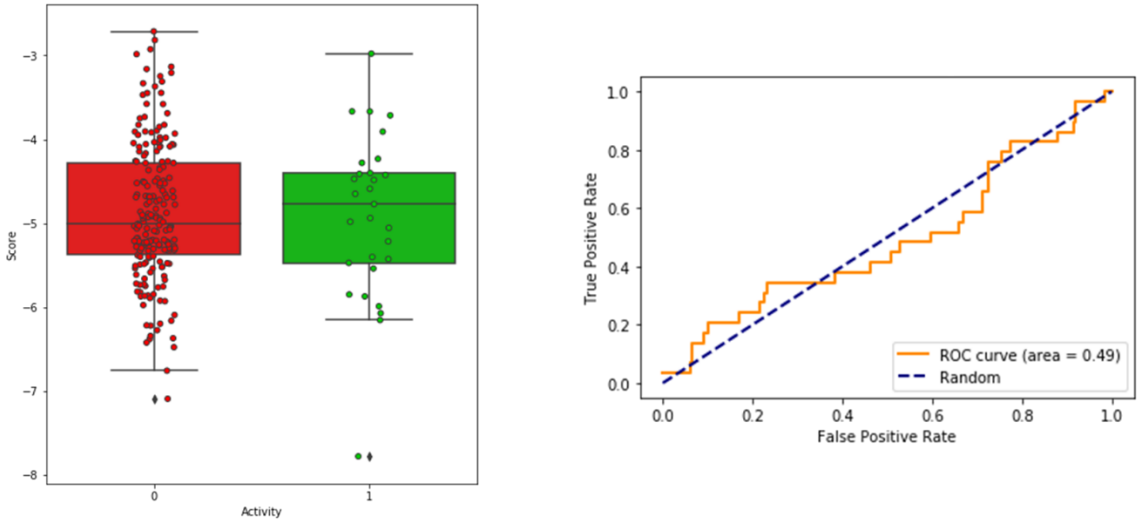


Figure 10. Docking results between decoys (red) and actives molecules (green) obtained with Glide from molecules prepared with LigPrep from Maestro: (a) absolute docking score dispersion; (b) ROC-curve plot.

The Figure 11 shows the results obtained with the preparation and standardization of the molecules using VS-Prep. The score is turned in absolute (the SP score function of Glide returns a negative one).

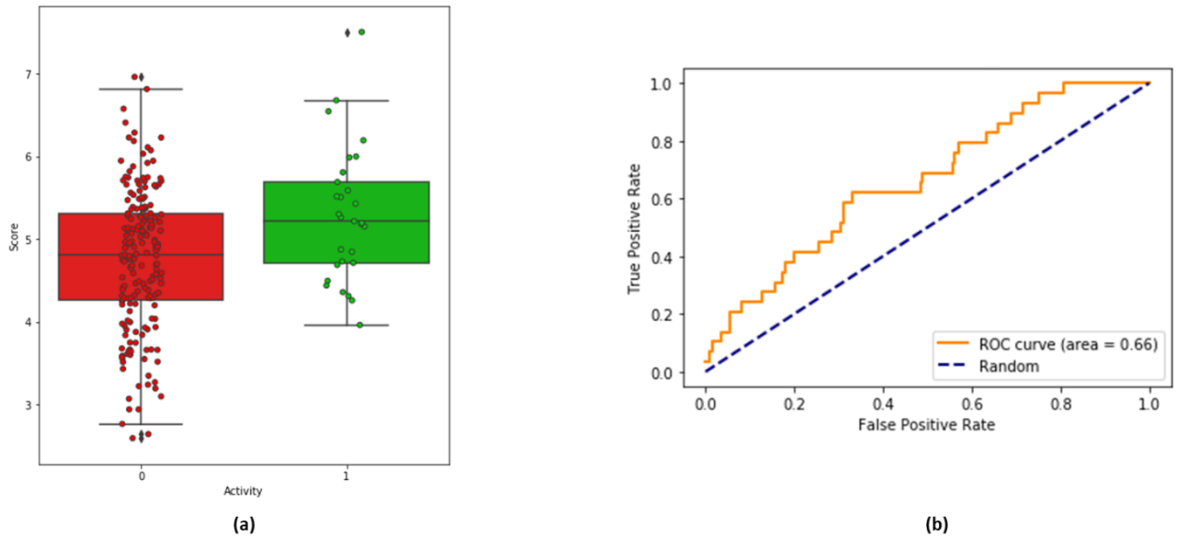
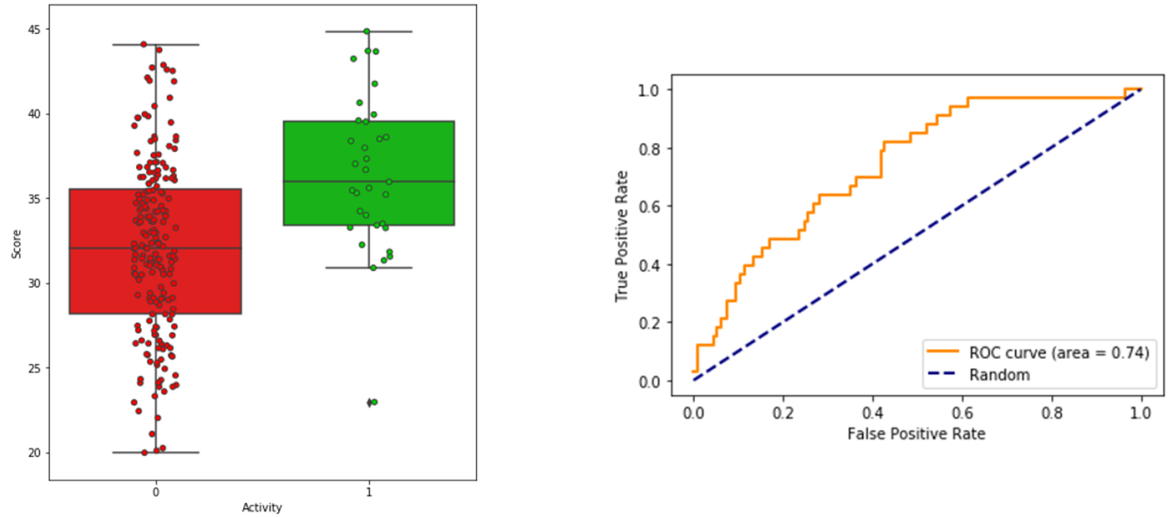


Figure 11. Docking results between decoys (red) and actives molecules (green) obtained with Glide from molecules prepared with VS-Prep: (a) absolute docking score dispersion; (b) ROC-curve plot.

Docking with GOLD

The Figure 12 shows the results obtained with the preparation and standardization of the molecules using RDKit and ETKDG method followed with an optimization using the MMFF94 forcefield. Score function is ASP.



(a) (b)

Figure 12. Docking results between decoys (red) and actives molecules (green) obtained with GOLD from molecules prepared with RDKit: (a) absolute docking score dispersion; (b) ROC-curve plot.

2.7 Bilan et conclusion

Ce projet s'inscrit parfaitement dans le changement de stratégie opéré par l'industrie pharmaceutique et cosmétique concernant le criblage à haut débit. Désormais, la tendance est de cribler des chimiothèques plus réduites, avec un ensemble de composés préalablement sélectionnés à l'aide d'outils de chémoinformatique[134]. De plus, de nos jours les ingrédients naturels sont devenus omniprésents dans les formulations cosmétiques et l'intérêt pour le naturel et le bio ne cesse de croître, devenant un très gros atout marketing[135]. La valorisation de produits naturels par le laboratoire est donc une démarche tout à fait dans l'air du temps, permettant de répondre à la demande sociétale croissante de ces produits. A l'avenir, la base de données pourra être améliorée avec des mises à jour régulières et en incorporant des composés d'origine marine, absents pour le moment mais qui ont déjà prouvés leur effets positifs dans d'autres études[136] .

Malgré des résultats en deçà de nos espérances, ce projet aura permis l'émergence de pratique standardisée dans l'équipe pour les projets d'amarrage moléculaire, notamment à l'aide des Jupyter notebooks (https://jupyter.org/) qui permettent de partager à la fois le code source, les commentaires et les résultats dans un format standard facilement lisible et exécutable. En effet, j'ai pu développer et partager un modèle de notebook utilisable par tous (à condition d'avoir au préalable installé toutes les bibliothèques nécessaires) permettant dans un premier temps de créer un jeu de données de molécules actives et inactives à partir des données expérimentales extraites de la ChEMBL. Puis, dans un deuxième temps, l'analyse des résultats après docking, par le logiciel du choix de l'utilisateur, à l'aide du score retourné pour chaque composé. Ainsi, ce notebook permet de calculer aisément la distribution du score des molécules, le facteur d'enrichissement sur une portion souhaitée, la courbe ROC et l'AUC associée, le meilleur rapport sensibilité/spécificité afin d'obtenir le score discriminant optimal… Il est utilisé pour la validation du protocole avant un criblage virtuel de librairie.

Enfin, dans le cadre de ce projet, d'autres cibles ont été criblées, notamment des enzymes impliquées dans le vieillissement de la peau comme les élastases ou les collagénases. De plus, les connaissances apprises grâce à ces recherches m'ont aussi permis d'aider des collègues travaillant sur des inhibiteurs de la hyaluronidase. J'ai ainsi pu leur proposer un mode de liaison validé à l'aide d'étude d'amarrage moléculaire leur permettant de mieux comprendre comment optimiser leurs composés.

---

[134] Caroline Barette et al., « Force et spécificité du criblage pour des molécules bioactives au CMBA-Grenoble - Une plate-forme dédiée à la découverte et à l'analyse de molécules bioactives et candidats médicaments », *médecine/sciences* 31, nº 4 (1 avril 2015): 423-31, https://doi.org/10.1051/medsci/20153104017.

[135] « Cosmétiques : le boom du bio ? », *IFOP* (blog), consulté le 17 juillet 2019, https://www.ifop.com/publication/cosmetiques-le-boom-du-bio/.

[136] Elena M. Balboa et al., « Cosmetics from Marine Sources », in *Springer Handbook of Marine Biotechnology*, éd. par Se-Kwon Kim, Springer Handbooks (Berlin, Heidelberg: Springer Berlin Heidelberg, 2015), 1015-42, https://doi.org/10.1007/978-3-642-53971-8_44.

Chapitre 3 : L'approche par fragments dans la conception de médicaments

3.1 Présentation générale

Une approche différente du criblage de molécules « drug-like » est apparue il y a une trentaine d'années : la conception de médicaments par la méthode des fragments, que l'on appelle aussi « Fragment-Based Drug Design » (FBDD). Le principe de cette méthode est de trouver des touches à partir de fragments moléculaires servant de points de départ à optimiser par différentes stratégies pour obtenir un candidat médicament. Cette technique est en plein essor ces dernières années depuis ses récents succès et depuis la remise en question des performances du criblage à haut débit qui a fourni parfois un très faible nombre de touches pour certaines cibles[137].

3.1.1 Apparition et essor

La naissance du concept d'approche par fragments remonte au début des années 1980, avec notamment les travaux de W.P. Jencks portant sur le principe d'additivité des énergies de liaison[138]. Il va ainsi proposer de décomposer les molécules en plusieurs sous-parties (les fragments) dont chacune contient au moins un groupe fonctionnel. L'affinité d'une molécule pour une cible pourra alors s'expliquer comme étant la somme des affinités de chacun de ses fragments avec cette même cible. Puis, P.R. Andrews va continuer à travailler sur le sujet en recherchant et caractérisant certains groupes fonctionnels importants pour la formation de liaisons et donc essentiels dans un fragment[139]. C'est finalement en 1992 que l'approche est expérimentalement démontrée par la découverte d'inhibiteurs de la protéine FK506 à partir de criblage de fragments par RMN[140]. Ce criblage a permis d'identifier deux fragments se liant dans des poches voisines du site actif de FK506. Chacun a ensuite été optimisé puis ils ont été reliés ensemble pour obtenir un composé d'une affinité de 19 nM. Cette méthode, appelée « *SAR by NMR* », a ouvert la voie à plusieurs techniques de génération de molécules à partir de fragments. Dès lors, il est devenu possible de généraliser la conception de composés à partir de fragments en suivant un protocole en trois étapes. Dans un premier temps, cribler des chimiothèques de fragments et détecter ceux se liant à la cible. Puis, dans un second temps, après optimisation de ces premières touches, effectuer un second criblage en présence d'une touche initiale déjà fixée afin de détecter un second site de liaison potentiel. Enfin, dans un troisième temps, relier chaque nouvelle touche obtenue lors du second criblage par synthèse organique avec une touche originale du premier criblage.

Parallèlement à ces approches expérimentales, de nombreuses initiatives *in silico* voient le jour. Le premier programme fonctionnel pour le FBDD est paru en 1985. Il s'agit de GRID, un logiciel qui permet de positionner des groupements fonctionnels virtuels dans une cible

[137] Ricardo Macarron, « Critical review of the role of HTS in drug discovery », *Drug Discovery Today* 11, nº 7 (1 avril 2006): 277-79, https://doi.org/10.1016/j.drudis.2006.02.001.
[138] William P. Jencks, « On the Attribution and Additivity of Binding Energies », *Proceedings of the National Academy of Sciences* 78, nº 7 (1 juillet 1981): 4046-50, https://doi.org/10.1073/pnas.78.7.4046.
[139] P. R. Andrews, D. J. Craik, et J. L. Martin, « Functional group contributions to drug-receptor interactions », *Journal of Medicinal Chemistry* 27, nº 12 (1 décembre 1984): 1648-57, https://doi.org/10.1021/jm00378a021.
[140] Suzanne B. Shuker et al., « Discovering High-Affinity Ligands for Proteins: SAR by NMR », *Science* 274, nº 5292 (29 novembre 1996): 1531-34, https://doi.org/10.1126/science.274.5292.1531.

donnée afin d'établir une carte des sites de liaison de haut potentiel, appelés « hot spots »[141]. GRID est considéré comme le précurseur du FBDD *in silico*, son analogue expérimental a même été développé après coup : la méthode MSCS[142]. Deux autres outils important sont développés par la suite : MCSS et LUDI. MCSS est une méthode utilisée pour cribler virtuellement les fragments et déceler leurs positions optimales dans le site de liaison à l'aide d'un calcul d'énergie minimale[143]. LUDI va encore plus loin en connectant les différents fragments placés pour créer une nouvelle molécule[144].

On retrouve les éléments majeurs ayant contribué à l'émergence du FBDD dans la Figure 27. Parmi ceux-ci, les plus importants sont les suivants :

- La création des premières entreprises de biotechnologies spécialisées dans le FBDD (Astex, Sunesis, Vernalis, …) à la fin des années 1990.

- L'étude de M. Hann en 2001 montrant que la complexité d'une molécule est inversement proportionnelle à son affinité avec une cible[145]. Autrement dit, une molécule a besoin de posséder suffisamment de différents groupes fonctionnels pour se lier efficacement à la cible. Cependant, un nombre trop important de tels groupements et une masse moléculaire trop élevée peuvent au contraire avoir des effets néfastes à sa fixation.

- La « règle de 3 » énoncée par M. Congreve en 2003 pour caractériser un fragment[146], connexe à la « règle de cinq » formulée par C.A. Lipinski six ans auparavant caractérisant les molécules « drug-like »[147]. Dans la foulée, les premières chimiothèques de fragments commerciales apparaissent.

- La notion de « ligand efficiency » (LE) comme métrique afin de quantifier la qualité des interactions formées entre le fragment et le site actif proposée par A.L. Hopkins en 2004[148]. Le LE désigne le rapport de l'énergie de liaison du complexe fragment-cible avec son nombre d'atomes lourds.

---

[141] P. J. Goodford, « A computational procedure for determining energetically favorable binding sites on biologically important macromolecules », *Journal of Medicinal Chemistry* 28, nº 7 (1 juillet 1985): 849-57, https://doi.org/10.1021/jm00145a002.

[142] C. Mattos et D. Ringe, « Locating and Characterizing Binding Sites on Proteins », *Nature Biotechnology* 14, nº 5 (mai 1996): 595-99, https://doi.org/10.1038/nbt0596-595.

[143] Andrew Miranker et Martin Karplus, « Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method », *Proteins: Structure, Function, and Bioinformatics* 11, nº 1 (1991): 29-34, https://doi.org/10.1002/prot.340110104.

[144] Hans-Joachim Böhm, « The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors », *Journal of Computer-Aided Molecular Design* 6, nº 1 (1 février 1992): 61-78, https://doi.org/10.1007/BF00124387.

[145] Michael M. Hann, Andrew R. Leach, et Gavin Harper, « Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery », *Journal of Chemical Information and Computer Sciences* 41, nº 3 (1 mai 2001): 856-64, https://doi.org/10.1021/ci000403i.

[146] Miles Congreve et al., « A 'Rule of Three' for fragment-based lead discovery? », *Drug Discovery Today* 8, nº 19 (1 octobre 2003): 876-77, https://doi.org/10.1016/S1359-6446(03)02831-9.

[147] Christopher A Lipinski et al., « Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings », *Advanced Drug Delivery Reviews*, Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998, 46, nº 1 (1 mars 2001): 3-26, https://doi.org/10.1016/S0169-409X(00)00129-0.

[148] Andrew L. Hopkins, Colin R. Groom, et Alexander Alex, « Ligand efficiency: a useful metric for lead selection », *Drug Discovery Today* 9, nº 10 (15 mai 2004): 430-31, https://doi.org/10.1016/S1359-6446(04)03069-7.

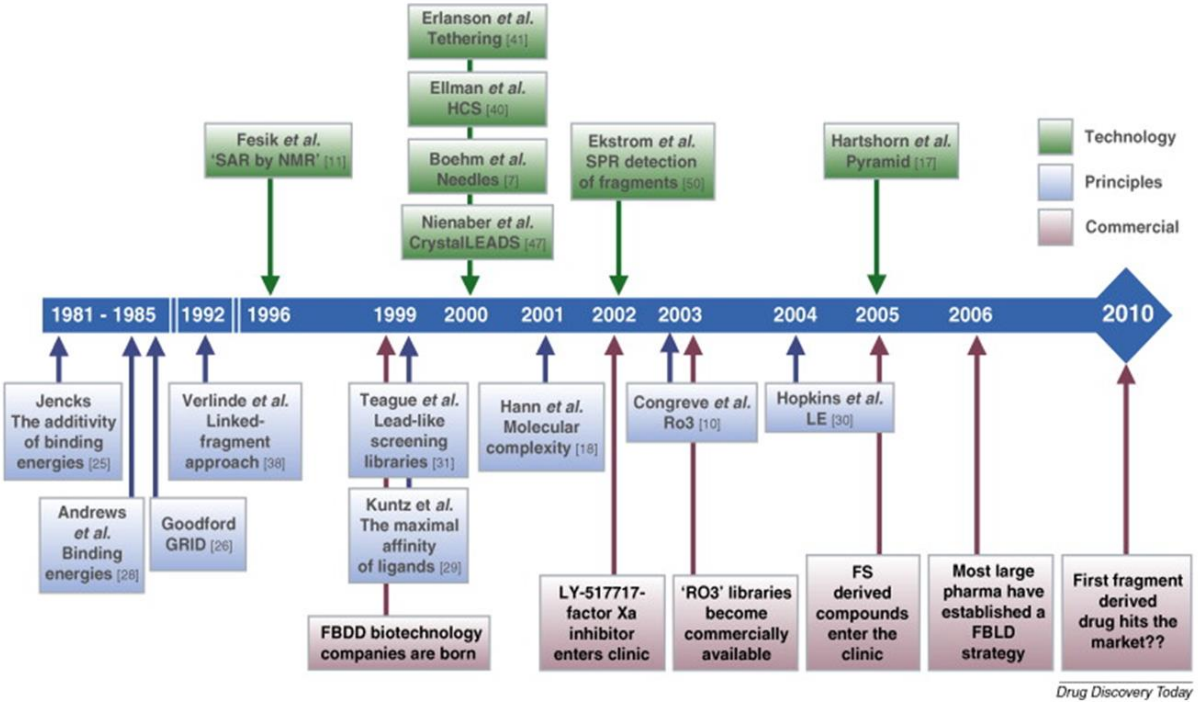


Figure 27 : Chronologie récapitulative des différents évènements dans le développement de l'approche par fragments pour trouver de nouveaux médicaments. *D'après G. Chessari & al.*[149].

Ainsi, en l'espace d'une quinzaine d'années, trois molécules* obtenues par FBDD ont été approuvées par la FDA, dont deux sont des inhibiteurs de protéines kinases :

- Le vemurafenib en 2011, développé par Plexxikon pour le traitement du mélanome métastatique chez les adultes. Il cible la kinase BRAF mutée (V600E)[150].

- Le venetoclax en 2016, développé par AbbVie pour le traitement de la leucémie lymphoïde chronique. Il cible la protéine Bcl-2[151].

- L'erdafitinib très récemment en 2019, développé par Astex Pharmaceuticals et Janssen pour le traitement du cancer avancé de la vessie. Il cible les kinases de la famille FGFR[152].

* Vous pourrez retrouver une représentation visuelle de ces trois composés ainsi que le(s) fragment(s) de départ utilisé(s) dans la revue à la fin de chapitre.

Le FBDD s'est désormais généralisé aussi bien au niveau des entreprises pharmaceutiques qu'au niveau du secteur académique[153]. On retrouve de nombreux livres et de plus en plus de publications associés à ce domaine (Figure 28). Enfin, il existe un blog dédié

[149] Gianni Chessari et Andrew J. Woodhead, « From fragment to clinical candidate—a historical perspective », *Drug Discovery Today* 14, nº 13 (1 juillet 2009): 668-75, https://doi.org/10.1016/j.drudis.2009.04.007.
[150] « Zelboraf (Vemurafenib) FDA Approval History », Drugs.com, consulté le 2 juillet 2019, https://www.drugs.com/history/zelboraf.html.
[151] « Venclexta (Venetoclax) FDA Approval History », Drugs.com, consulté le 2 juillet 2019, https://www.drugs.com/history/venclexta.html.
[152] « Balversa (Erdafitinib) FDA Approval History », Drugs.com, consulté le 2 juillet 2019, https://www.drugs.com/history/balversa.html.
[153] Monya Baker, « Fragment-Based Lead Discovery Grows Up », *Nature Reviews Drug Discovery* 12, nº 1 (janvier 2013): 5-7, https://doi.org/10.1038/nrd3926.

reconnu : http://practicalfragments.blogspot.com/, qui relate chaque semaine les développements importants dans le FBDD et propose une liste actualisée de tous les composés issus de cette approche (approuvés ou en essai clinique).

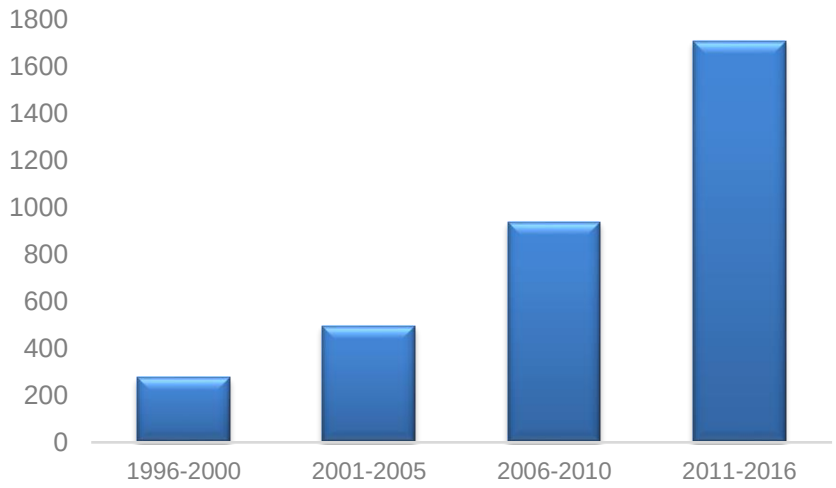


Figure 28 : Evolution du nombre d'articles scientifiques publiés traitant ou ayant un rapport avec l'approche par fragments. *D'après les données fournies par A. KS Romasanta et al.*[154].*

*Là encore il faut penser à mettre en perspective ces chiffres avec le fait que le nombre global de publications et de journaux scientifiques ne cessent également d'augmenter ces dernières années.

3.1.2 Les caractéristique des fragments

Les fragments se distinguent des molécules traditionnelles par leur faible complexité moléculaire et leur taille modérée (Figure 29). Il n'existe pas de définition officielle, cependant comme son nom l'indique il s'agit de morceaux de molécules pouvant aller de quelques atomes à un composé polycyclé. De manière conventionnelle, la majorité des chercheurs considère qu'un bon fragment devrait respecter la règle de 3[155] : un poids moléculaire inférieur à 300 Da, un coefficient de partage octanol/eau calculé (CLogP) inférieur à 3, un nombre de donneurs et d'accepteurs de liaisons hydrogène inférieur à 3. Comme le montre le Tableau 2, les valeurs des paramètres physico-chimiques des fragments peuvent varier selon s'ils sont expérimentaux ou virtuels. En effet, l'utilisation de fragments expérimentaux trop simples (moins de 10 atomes) pourrait soit empêcher de détecter une activité qui serait alors trop faible, soit empêcher de trouver le mode de liaison le plus spécifique car ce fragment pourrait se lier n'importe où dans le site actif. En plus d'un équilibre idéal entre simplicité et complexité à atteindre, pour un bon fragment, il faut aussi penser à sa future synthèse et envisager la possibilité de connecter plus ou moins facilement d'autres fragments dessus[156]. Enfin, il faut aussi penser qu'à cause de leur faible nombre de groupement polaire, les fragments peuvent présenter des problèmes de solubilité[157].

---

[154] Angelo K. S. Romasanta et al., « When Fragments Link: A Bibliometric Perspective on the Development of Fragment-Based Drug Discovery », *Drug Discovery Today* 23, nº 9 (1 septembre 2018): 1596-1609, https://doi.org/10.1016/j.drudis.2018.05.004.

[155] Miles Congreve et al., « A 'Rule of Three' for fragment-based lead discovery? », *Drug Discovery Today* 8, nº 19 (1 octobre 2003): 876-77, https://doi.org/10.1016/S1359-6446(03)02831-9.

[156] György M. Keserű et al., « Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia », *Journal of Medicinal Chemistry* 59, nº 18 (22 septembre 2016): 8189-8206, https://doi.org/10.1021/acs.jmedchem.6b00197.

[157] Alvin W. Hung et al., « Route to Three-Dimensional Fragments Using Diversity-Oriented Synthesis », *Proceedings of the National Academy of Sciences* 108, nº 17 (26 avril 2011): 6799-6804, https://doi.org/10.1073/pnas.1015271108.

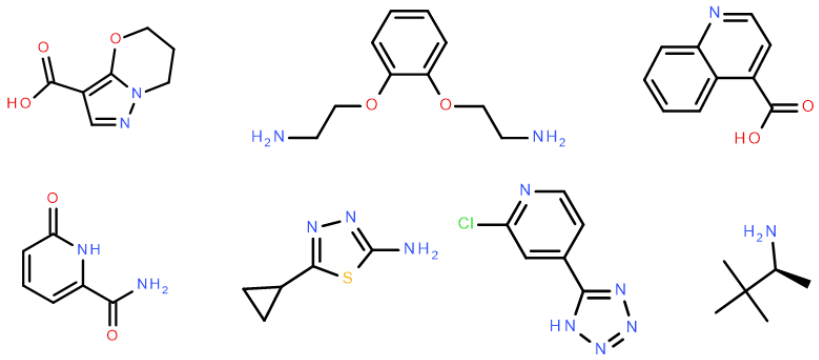


Figure 29 : Echantillon de différents fragments moléculaires.

Tableau 2 : Récapitulatif des différentes caractéristiques idéales pour une molécule drug-like, un fragment expérimental et un fragment virtuel. *Adapté de Z. D. Konteatis*[158].

Propriété	Molécules drug-likes	Fragments expérimentaux	Fragments virtuels
MW (Da)	<= 500	100 - 300	17 - 200
CLogP	<= 5	<= 3	<= 3
HBA	<= 10	<= 3	<= 2
HBD	<= 5	<= 3	<= 2
NRB	<= 7	<= 3	<= 3
TPSA (Å²)	<= 140	<= 60	<= 60
Solubilité*	> 50 µM	mM	-
Affinité attendue*	nM - µM	30 µM - nM	100 µM - nM

Avec MW la masse moléculaire, CLogP le coefficient octanol-eau calculé avec RDKit, HBA et HBD pour liaisons accepteuses et donneuses d'hydrogène respectivement, NRB le nombre de liaisons rotables et TPSA la surface polaire topologique.
* ordre de grandeur

Il est important de rappeler que les valeurs proposées par ce genre de règles demeurent des recommandations plutôt que des obligations, rien n'empêche de modifier les seuils si besoin. A titre d'exemple, dans le cas où un fragment contient un atome de brome ou d'iode (poids moléculaire de 79,90 Da et 126,90 Da respectivement), le poids moléculaire va rapidement augmenter et dépasser le seuil, contrairement au même fragment mais avec un atome de chlore à la place (poids moléculaire de 35,45 Da). D'autre paramètres physico-chimiques peuvent ainsi être pris en compte comme le nombre d'atomes lourds (en général limité à 17) ou des descripteurs basés sur la forme moléculaire [159].

[158] Zenon D. Konteatis, « In silico fragment-based drug design », *Expert Opinion on Drug Discovery* 5, nº 11 (1 novembre 2010): 1047-65, https://doi.org/10.1517/17460441.2010.523697.
[159] Harren Jhoti et al., « The "rule of Three" for Fragment-Based Drug Discovery: Where Are We Now? », *Nature Reviews Drug Discovery* 12, nº 8 (août 2013): 644, https://doi.org/10.1038/nrd3926-c1.

3.2 Avantages de l'approche par fragments

Les étapes pour aboutir *in fine* à un candidat médicament restent similaires entre l'approche classique de criblage à haut débit HTS et l'approche de criblage par fragments : la création d'une chimiothèque, le criblage sur la cible d'intérêt, l'identification, et pour finir, l'optimisation des touches. Cependant, l'approche par fragment démontre de nombreux avantages dont trois principaux :

- Une couverture de l'espace chimique beaucoup plus importante car un faible nombre de fragments peut représenter un espace chimique bien plus grand que celui couvert par le même nombre de molécules classiques[160].

- Une efficacité de liaison très élevée (LE > 0,3 kcal/mol) malgré une activité biologique faible (Figure 30). Contrairement à des molécules classiques, la grande majorité (voire tous) les atomes d'un fragment vont être impliqués dans une interaction avec la cible évitant des interactions défavorables que l'on peut retrouver avec des molécules plus grandes[161]. Qui plus est, les fragments sont en général moins sélectifs et peuvent donc se lier à de multiples cibles ce qui augmente les chances de trouver des touches lors d'un criblage[162].

- Une facilité à développer les touches jusqu'aux candidat médicament. De par sa petite taille, optimiser un fragment se fait plus facilement, de même que le contrôle des propriétés physico-chimiques et du profil ADME-Tox au fur et à mesure de l'agrandissement.

Ces avantages permettent l'usage d'une bibliothèque de criblage beaucoup plus réduite que lors d'un criblage HTS. Ainsi, les librairies de fragments comptent entre 500 et 2 000 composés[163], contrairement au criblage par HTS où ce chiffre peut grimper à un million. Dans le cas d'un criblage par fragments, le choix des touches ne se fait pas selon l'activité mais selon l'efficacité. Il existe plusieurs autres métriques pouvant être utilisés à la place du LE pour estimer l'efficacité d'une touche[164,165]. La marge de manœuvre pour optimiser les touches est beaucoup plus grande que lors d'un criblage HTS où les molécules présentent déjà des caractéristiques proches des limites des règles établies par Lipinski. On parle même d'« obésité moléculaire »[166] car leurs volumes moléculaires élevés rend difficile leur transformation en candidats médicaments. On peut résumer l'optimisation pour l'approche FBDD par la construction et l'ajout de groupement fonctionnels autour d'un noyau central, tandis que pour

[160] Philip J. Hajduk et Jonathan Greer, « A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned », *Nature Reviews Drug Discovery* 6, nº 3 (mars 2007): 211-19, https://doi.org/10.1038/nrd2220.
[161] I. D. Kuntz et al., « The Maximal Affinity of Ligands », *Proceedings of the National Academy of Sciences* 96, nº 18 (31 août 1999): 9997-10002, https://doi.org/10.1073/pnas.96.18.9997.
[162] Bas Lamoree et Roderick E. Hubbard, « Current perspectives in fragment-based lead discovery (FBLD) », *Essays in Biochemistry* 61, nº 5 (8 novembre 2017): 453-64, https://doi.org/10.1042/EBC20170028.
[163] György M. Keserű et al., « Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia », *Journal of Medicinal Chemistry* 59, nº 18 (22 septembre 2016): 8189-8206, https://doi.org/10.1021/acs.jmedchem.6b00197.
[164] Didier Rognan, « Fragment-Based Approaches and Computer-Aided Drug Discovery », in *Fragment-Based Drug Discovery and X-Ray Crystallography*, éd. par Thomas G. Davies et Marko Hyvönen, Topics in Current Chemistry (Berlin, Heidelberg: Springer Berlin Heidelberg, 2012), 201-22, https://doi.org/10.1007/128_2011_182.
[165] Maria Maddalena Cavalluzzi et al., « Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective », *Expert Opinion on Drug Discovery* 12, nº 11 (2 novembre 2017): 1087-1104, https://doi.org/10.1080/17460441.2017.1365056.
[166] Michael M. Hann, « Molecular Obesity, Potency and Other Addictions in Drug Discovery », *MedChemComm* 2, nº 5 (1 mai 2011): 349-55, https://doi.org/10.1039/C1MD00017A.

l'approche HTS il s'agit de transformations et modifications plus ou moins importantes sur une molécule proche du futur candidat médicament.

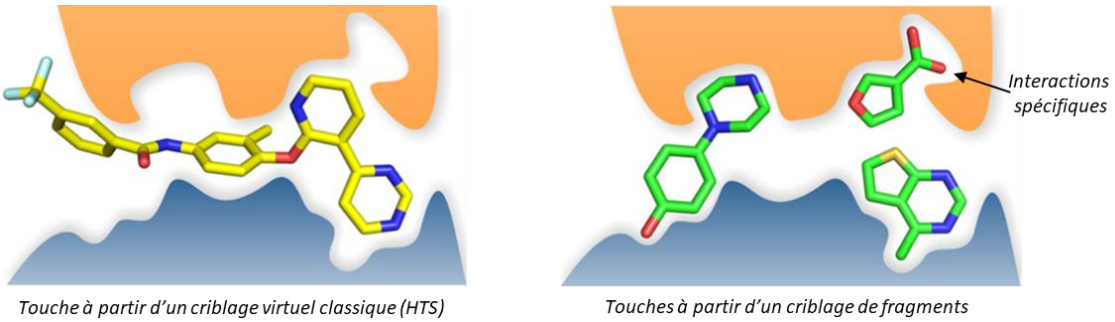


Figure 30 : Comparaison d'une touche entre un criblage HTS et un criblage de fragments. *D'après D.E. Scott & al.*[167].

De par ces avantages, le premier médicament découvert par approche FBDD, le vemurafenib, n'a mis que 6 ans pour être commercialisé à partir de sa découverte[168], contre 10 à 15 ans traditionnellement. Enfin, l'approche par fragments a permis à de petites entreprises et à des unités de recherche publiques d'obtenir des touches sans investir beaucoup d'argent dans l'achat de chimiothèques de milliers (millions) de molécules et de matériel automatisé pour le criblage à haut débit.

3.3 Méthodes expérimentales

Les méthodes expérimentales pour cribler des fragments doivent être en mesure de détecter des affinités faibles et de supporter des concentrations élevées car ceux-ci sont peu solubles[169]. Il s'agit de techniques biophysiques, parmi lesquelles on peut citer la résonance plasmonique de surface (SPR), la spectrométrie de masse (MS), la mesure de la température de dénaturation des protéines (TSA), la RMN et la cristallographie aux rayons X. Seules les deux dernières méthodes permettent de fournir des informations sur le mode de liaison entre le fragment et son récepteur[170]. Lors d'un criblage de fragments, la stratégie employée est souvent d'utiliser une première technique pour identifier des touches, puis de les confirmer par une seconde et différente approche dite orthogonale.

Un récent sondage réalisé par D. Erlanson pour le blog Practical Fragments montre les méthodes expérimentales les plus utilisées pour le criblage de fragments (Figure 31, A). Depuis 2016, la technique par RMN se classe première, devant la SPR et la cristallographie aux rayons X qui complète le podium. Cependant, on observe que l'utilisation de la cristallographie aux rayons X a doublé entre 2011 et 2016 (23 % vs 55 %) démontrant l'importance croissante de

[167] Duncan E. Scott et al., « Fragment-Based Approaches in Drug Discovery and Chemical Biology », *Biochemistry* 51, nº 25 (26 juin 2012): 4990-5003, https://doi.org/10.1021/bi3005126.
[168] Gideon Bollag et al., « Vemurafenib: The First Drug Approved for *BRAF*-Mutant Cancer », *Nature Reviews Drug Discovery* 11, nº 11 (novembre 2012): 873-86, https://doi.org/10.1038/nrd3847.
[169] Isabelle Krimm, « Le criblage de fragments: Une voie prometteuse pour la conception de médicaments », *médecine/sciences* 31, nº 2 (février 2015): 197-202, https://doi.org/10.1051/medsci/20153102017.
[170] Christopher W Murray et Tom L Blundell, « Structural biology in fragment-based drug design », *Current Opinion in Structural Biology*, Membranes / Engineering and design, 20, nº 4 (1 août 2010): 497-507, https://doi.org/10.1016/j.sbi.2010.04.003.

cette méthode pour le FBDD[171]. Ces résultats sont confirmés lors de l'analyse de la littérature scientifique, notamment l'avènement de la cristallographie ces dernières années (Figure 31, B).

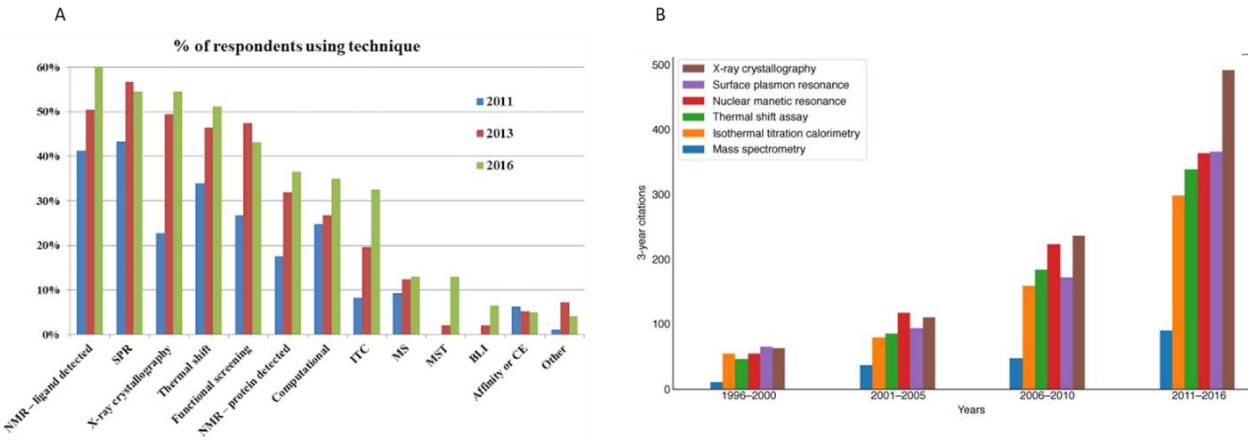


Figure 31 : Résultat d'un sondage auprès de la communauté scientifique pour connaitre les méthodes expérimentales de criblage de fragments les plus utilisées (A) et graphique montrant les citations de ces méthodes dans la littérature scientifique (B). *D'après le blog Practical Fragments*[172] *et A. KS. Romasanta et al.*[173].

Bien que ne donnant pas d'information sur l'affinité d'un fragment avec sa cible, la cristallographie aux rayons X est souvent un élément majeur voire indispensable à l'optimisation (comme le dit l'adage : une « image » vaut mille mots). Elle va servir à connaitre les sites de croissance potentiels d'une touche ainsi que la direction à emprunter pour guider sa transformation en candidat médicament (Figure 32). En général, cette méthode n'est pas utilisée à proprement dit pour le criblage de toute une chimiothèque, mais uniquement pour obtenir un surcroit d'informations sur des touches validées par une première méthode [174].

A noter tout de même que la cristallographie aux rayons X présente quelques limitations[175]. Cela reste une technique assez coûteuse et difficile à mettre en place dans des petites structures de recherche. De plus, les informations structurales obtenues par rayons X restent un modèle. Il s'agit d'une interprétation de la carte de densité des électrons, obtenue expérimentalement par le cristallographe, pouvant donc présenter des erreurs ou des imprécisions. Enfin, certaines protéines ne sont pas pour l'instant, ou alors très difficilement, cristallisables, comme les RCPGs, ce qui rend cette méthode inadéquate dans certains cas[176,177].

---

[171] Disha Patel, Joseph D. Bauman, et Eddy Arnold, « Advantages of Crystallographic Fragment Screening: Functional and Mechanistic Insights from a Powerful Platform for Efficient Drug Discovery », *Progress in biophysics and molecular biology* 116, nº 0 (2014): 92-100, https://doi.org/10.1016/j.pbiomolbio.2014.08.004.

[172] Dan Erlanson, « Practical Fragments: Fragments in the clinic: 2018 edition », *Practical Fragments* (blog), 6 octobre 2018, https://practicalfragments.blogspot.com/2018/10/fragments-in-clinic-2018-edition.html.

[173] Angelo K. S. Romasanta et al., « When Fragments Link: A Bibliometric Perspective on the Development of Fragment-Based Drug Discovery », *Drug Discovery Today* 23, nº 9 (1 septembre 2018): 1596-1609, https://doi.org/10.1016/j.drudis.2018.05.004.

[174] Douglas R. Davies, « Screening Ligands by X-Ray Crystallography », in *Structural Genomics and Drug Discovery: Methods and Protocols*, éd. par Wayne F. Anderson, Methods in Molecular Biology (New York, NY: Springer New York, 2014), 315-23, https://doi.org/10.1007/978-1-4939-0354-2_23.

[175] Andrew M. Davis, Stephen A. St-Gallay, et Gerard J. Kleywegt, « Limitations and lessons in the use of X-ray structural information in drug design », *Drug Discovery Today* 13, nº 19 (1 octobre 2008): 831-41, https://doi.org/10.1016/j.drudis.2008.06.006.

[176] Jean-Jacques Lacapère et al., « Determining Membrane Protein Structures: Still a Challenge! », *Trends in Biochemical Sciences* 32, nº 6 (1 juin 2007): 259-70, https://doi.org/10.1016/j.tibs.2007.04.001.

[177] Ilka Müller, « Guidelines for the successful generation of protein–ligand complex crystals », *Acta Crystallographica. Section D, Structural Biology* 73, nº Pt 2 (1 février 2017): 79-92, https://doi.org/10.1107/S2059798316020271.

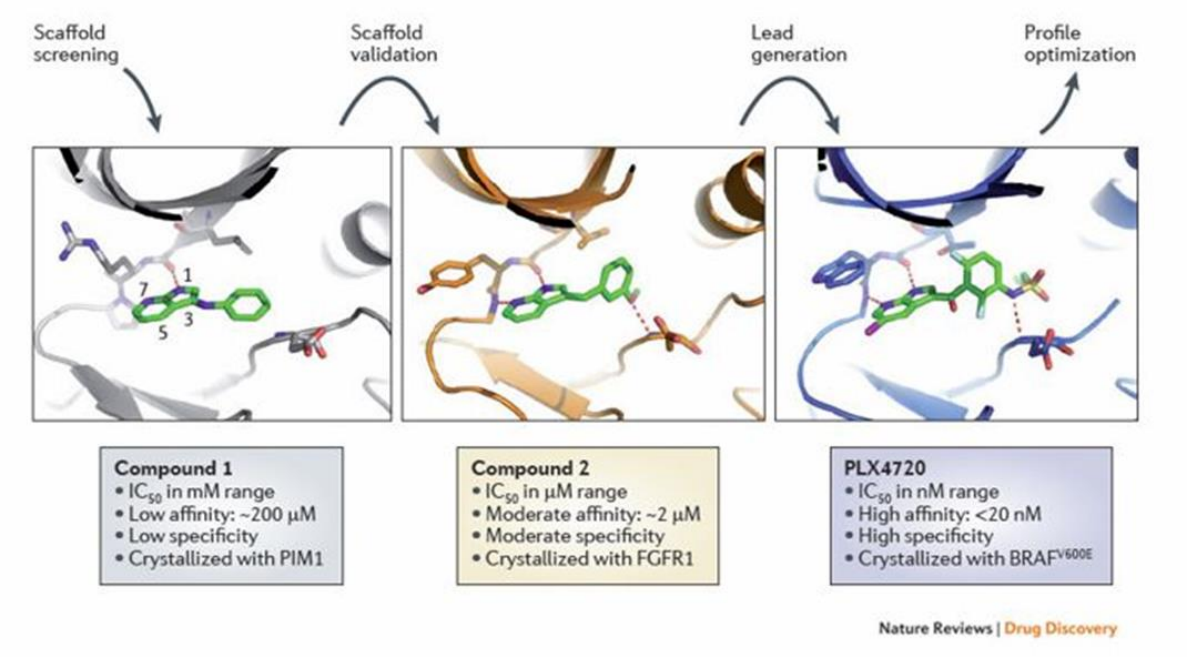


Figure 32 : Exemple de co-cristallisation d'une touche suivi de l'optimisation jusqu'à obtenir un lead. *En cristallisant la molécule en construction dans différentes kinases, les auteurs ont voulu montrer le mode d'interaction similaire d'un squelette moléculaire entre membres de cette famille de protéines. D'après G. Bollag & al.*[178].

Chaque technique expérimentale présente ses avantages et inconvénients que l'on retrouve dans plusieurs revues[179,180,181], mais en dehors de celles-ci, comme énoncé dans la présentation du FBDD, les méthodes *in silico* revêtent aussi une importance particulière, que ce soit pour la création d'une chimiothèque présentant une grande diversité ou pour la prédiction des interactions d'une touche avec la cible (amarrage moléculaire).

3.4 Revue sur les différents outils disponibles pour le FBDD *in silico*

Cette revue couvre l'étendue des différents aspects informatiques existant dans le domaine de l'approche par fragments et aborde les points suivants : la création d'une chimiothèque, les différentes stratégies d'optimisations d'une touche, les méthodes d'amarrages moléculaires et les différents logiciels existant, commerciaux ou non.

[178] Gideon Bollag et al., « Vemurafenib: The First Drug Approved for *BRAF*-Mutant Cancer », *Nature Reviews Drug Discovery* 11, nº 11 (novembre 2012): 873-86, https://doi.org/10.1038/nrd3847.
[179] Jacob Robson-Tull, « Biophysical Screening in Fragment-Based Drug Design: A Brief Overview », *Bioscience Horizons: The International Journal of Student Research* 11 (1 janvier 2018), https://doi.org/10.1093/biohorizons/hzy015.
[180] Daniel A. Erlanson et al., « Twenty Years on: The Impact of Fragments on Drug Discovery », *Nature Reviews Drug Discovery* 15, nº 9 (septembre 2016): 605-19, https://doi.org/10.1038/nrd.2016.109.
[181] Jean-Paul Renaud et al., « Biophysics in Drug Discovery: Impact, Challenges and Opportunities », *Nature Reviews Drug Discovery* 15, nº 10 (octobre 2016): 679-98, https://doi.org/10.1038/nrd.2016.123.

In silico Fragment-Based Drug Design: Current Status and Challenges

Colin Bournez[†], Nicolas Génin[†], Samia Aci-Sèche[†], Pascal Bonnet[†]*

[†]Institut de Chimie Organique et Analytique (ICOA UMR7311), Université d'Orléans - Pôle de chimie, rue de Chartres - BP 6759, 45067 Orléans Cedex 2, France

Keywords: Cheminformatics, Computational Fragment-Based Drug Design, fragment library, *de novo* design, docking.

ABSTRACT

Delivering successful outcomes beyond its early expectations, fragment-based drug design (FBDD) has become in the last decades a powerful method to guide the discovery of chemically active molecules. Of particular interest, computational FBDD can efficiently assist experimental investigations to improve the success rate of lead discovery and compound optimization. In the following article, we will review the recent advances in the field of *in silico* FBDD. The pros and cons of various computational methods for the design and selection of a fragment library, the virtual screening of fragments, and *de novo* design of lead compounds, will be discussed.

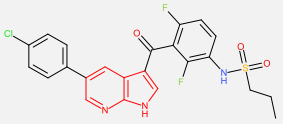
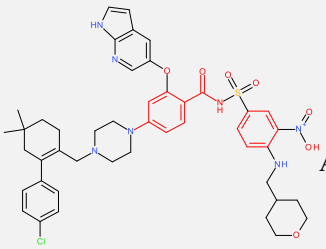
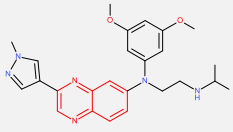
INTRODUCTION

Belonging to a competitive field in constant evolution, the drug discovery process is inherently innovation-oriented. The pharmaceutical industry, as well as the laboratories in academia, continuously prospect for new methods to find promising chemical entities. High-throughput screening (HTS) of compound libraries and their biological targets remains a dominant method to identify initial hits or lead compounds[1], and has proven relatively successful in the drug discovery field[2]. Yet the method also displays a series of limitations. First, the global success of an HTS screening campaign is evaluated at around 50%[3]. Next, despite a screened collection of millions of compounds, HTS only covers a very limited portion of the virtually infinite chemical space[4]. As a consequence, researchers are enticed to devise

novel strategies. One of such approaches consists in the following. Instead of screening large libraries of drug-size molecules, collections of smaller molecules (called fragments) can initially be considered. Those showing good potency can then be combined to form the final molecule. As such, the chemical space coverage can greatly be improved: a library of 10^3 fragments is as efficient, if not more, as one comprising 10^{5-6} of classical HTS compounds[5,6].

The latter approach was pioneered by Shuker et al.[7] in 1996. The researchers discovered a ligand with nanomolar affinity for the FK506 protein by linking together two fragments having each a micromolar affinity. Since then, Fragment-based drug design (FBDD) developed into a successful screening method and has been adopted by both the industry and academia[8,9]. A recent bibliometric analysis demonstrated a constant increase in the number of FBDD publications over the years[10]. At present, more than 45 drugs derived from FBDD have entered clinic trials[11] and 3 have been approved by the FDA: vemurafenib[12], venetoclax[13] and erdafitinib[14] (**Table 1**).

Table 1. FDA-approved drugs discovered from FBDD. *The starting fragments are colored in red in the Structure column.*

Name	Structure	Companies	Targets	Indication	Year of approval
Vemurafenib		Plexxikon/Roche	BRAF V600E	Metastatic melanoma with BRAF V600E mutation	2011
Venetoclax		Abbvie/Genentech	BCL-2	Chronic Lymphocytic Leukemia, Acute Myeloid Leukemia	2016
Erdafitinib		Astex/Janssen	FGFR1–4	Metastatic bladder cancer	2019

At first appreciation, the concept of fragment could appear unclear. According to its formal definition, "a fragment of something is a small piece or part of it"[15]. In the field of drug design, this translates to the assemblage of molecules to form a bigger one. A fragment can be

as simple as a single atom or as complex as a polycyclic ring system. In practice, fragments are about half the size of drug-like molecules and can be characterized by their chemical properties[16]. The routinely verified "Rule of Three" (RO3)[17] stipulates that most fragments will possess a molecular weight (MW) ≤ 300 Da, a LogP ≤ 3, and a number of H-bond donors/acceptors ≤ 3. It is to be noted that the latter rule serves primarily as a guideline rather than a strict limitation, one could quite use different parameters or adapt the threshold for a personal convenience. In general, due to their reduced spatial dimensions, fragments will display low target binding affinity: 100μM−10mM (traditional HTS molecules are up to the nM range[18]), but they are more likely to match the specific interaction requirements of the target.

However, Like HTS, FBDD is not exempt of limitations. The screening stage in particular can pose some hassle. Large quantities of high-purity protein targets are currently required, which can be challenging to compile. Otherwise, fragments can present solubility, degradation and reaction abnormalities[19], and targets such as GPCR membrane proteins may give rise to some difficulties[20]. Last but not least, whilst evolving fragment hits into lead compounds, key target interactions can be lost, thereby altering the primary potential.

Cheminformatics and *in silico* tools appear complementary to experimental FBDD. Use in combination, they may improve the efficiency and success of a pharmaceutical project. As instance, computational methods can deal with larger fragment libraries (up to millions of compounds) than those available for experimental screening, and for a significantly lower cost. As shown in Figure 1, *in silico* approaches can be integrated at each step of a FBDD project. We will review in the following sections the computational approaches that are available for the design of a fragment library, the finding and characterization of binding sites, the scoring of potential hits, and finally for the optimization of a hit to a lead compound. To complement existing literature on FBDD[21–25], we fill focus our attention on the state-of-the-art software tools, which are currently available and maintained (free of charges or commercial-licensed) and offer some transposable methodology in a rapidly evolving field.

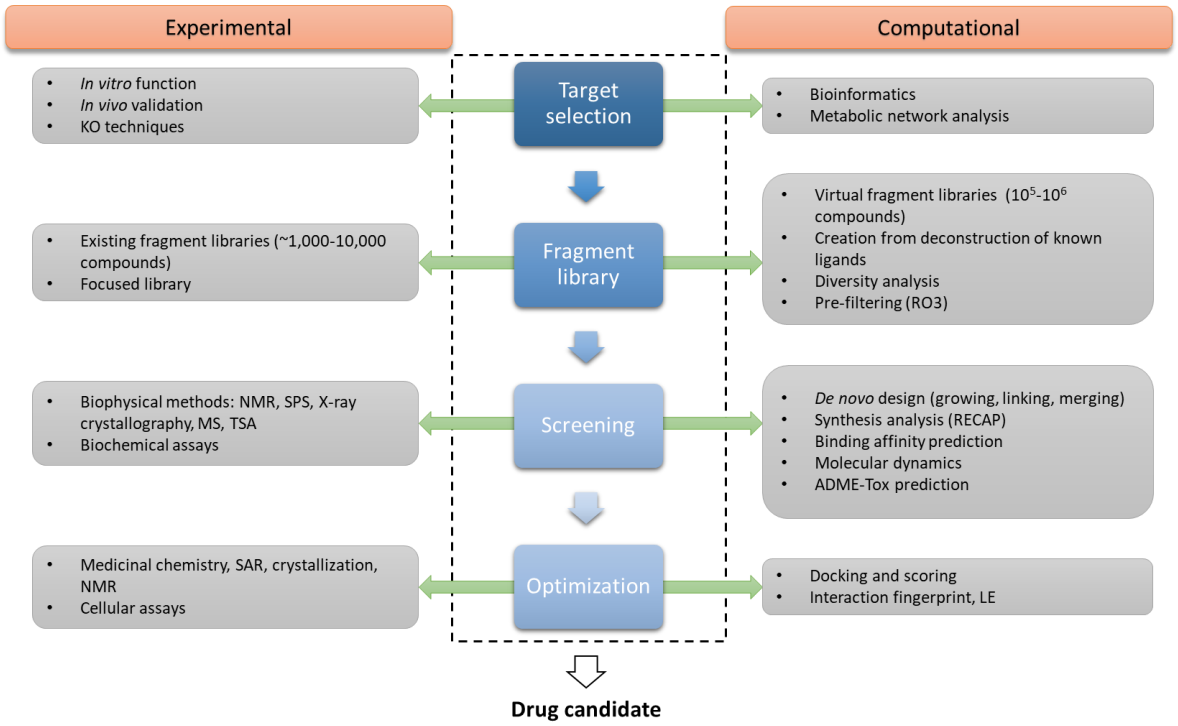


Figure 1. Computational and experimental complementary contributions to the FBDD process.

LIBRARY DESIGN

The first step in a FBDD campaign consists in the development of a fragment library to probe efficiently the chemical space and screen the target. One of the first choices to be addressed at the project inception is whether an existing library shall be put to contribution or a new one shall be custom-built. The majority of existing commercial libraries respects the Ro3 criteria. A non-exhaustive list with calculated physicochemical properties can be found on the Cambridge MedChem Consulting website[26] and in Keserü et al[27]. Most commercial libraries are downloadable in SDF format after online registration, and thus directly available for virtual screening.

Here, we will focus on the design of a custom library and the relevant computational tools to be leveraged. First, defining the size of the library is non-trivial, and will depend on the goal, the financial means and the experimental conditions of the screening campaign[28]. In any event, fragment libraries remain much smaller than those used in HTS campaigns. For an experimental screening, recent guidelines indicate an optimal size varying between 500 and 3000 units[27]. Nevertheless, the library should be as diverse as possible in terms of fragment properties, so as to optimize the chance of capturing target binding. In the first screening stage, fragment hits may not be fully involved in the final drug, but will serve as starting points for subsequent optimization stages. Second, the purity, stability and most importantly the solubility of the

fragments, are to be carefully taken into consideration. Because high fragment concentrations are needed in most screenings, the fragment must be very soluble in aqueous medium[9]. Third, if the library is stored internally, regular inspections should be performed to prevent precipitation/aggregation of the fragments and thus avoid false positives results. As far as virtual screening is concerned, the limitations are not the same, especially the size of the library which can go up to 10^6 compounds.

Two different strategies are generally applicable to build a fragment library: combining/editing existing databases or fragmenting existing compounds. In both cases the molecules can either be real or computer-generated. Numerous public databases are available to retrieve compounds with their associated bioactivity such as ChEMBL[29] and PubChem[30]. The ZINC[31] catalogue lists commercially-available compounds and can also serve as a starting point to gather molecules. Other databases are more specialized. For example, DrugBank[32] tackles either approved or experimental drugs, BindingDB[33] lists molecules with measured binding affinities, and e-Drug3D[34] compiles FDA-approved drugs. Otherwise, the following databases focus on specific targets. GLASS[35] is manually curated towards GPCR proteins. PKIDB[36] tracks protein kinase inhibitors that are approved or in clinical trials. Finally, eMolecules remains a commercial database gathering curated molecules from a variety of suppliers. The Figure 2 recapitulates the different steps to generate one's own fragment library.

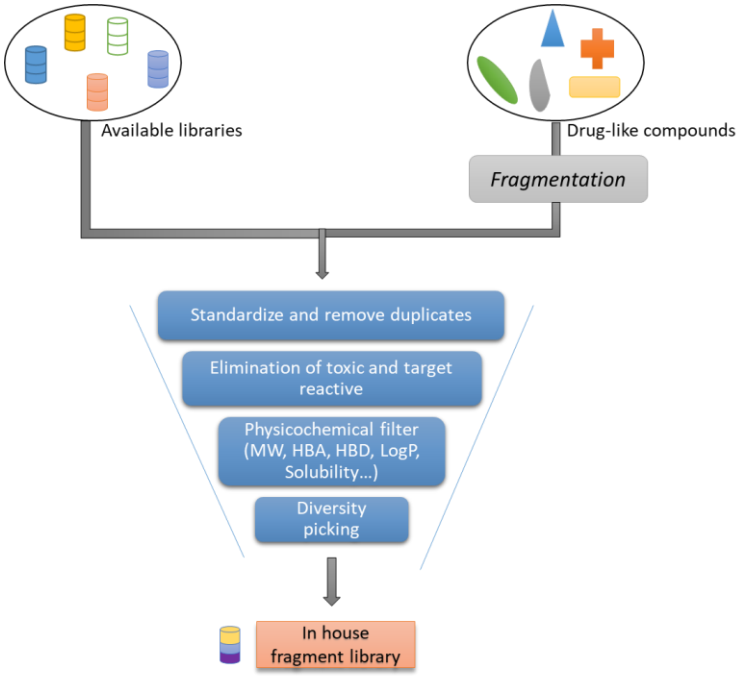


Figure 2. Custom fragment library generation steps.

When deciding to start from molecules, once they are compiled, multiple computational methods can be applied to perform their automatic decomposition into fragments. RECAP[37] is a tool of choice and uses common chemical reactions to cleave specific bonds. BRICS[38] algorithm is similar but leverages a more elaborate set of chemical rules and may thus generate more fragments. DAIM[39], molBLOCKS[40] and eMolFrag[41], can decompose compounds into chemically meaningful fragments and analyze the library outcome. Fragmentation tools are also included in the following Molecular Mechanics packages: MOE[42], Schrödinger[43], CHOMP[44] and Colibri[45]. Furthermore, fragmentation algorithms are implemented in open source development software such as Open Babel[46], CDK[47] and RDKit[48] or incorporated as nodes in KNIME workflows[49]. The latter development kits are particularly useful when devising personalized ways of fragmenting molecules. With a bit of knowledge and practice, cheminformatic suites can allow to fragment molecules according to one's chosen rules relative to SMARTS and SMILES patterns. It is possible for instance to extract the rotatable bonds in a molecule with the following SMARTS pattern in RDKit: *"[!$(*#*)&!D1]-&!@[!$(*#*)&!D1]"* or to split a molecule into its ring, linker and side-chain components as defined by Bemis and Murcko[50] (Figure 3).

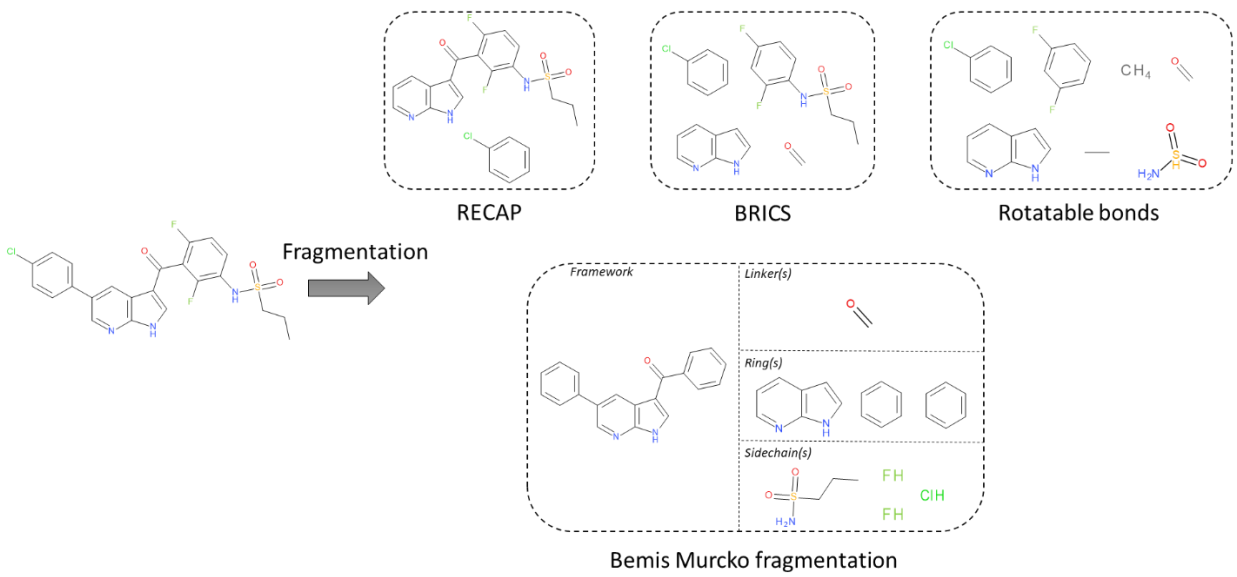


Figure 3. Compound fragmentation illustrated under different methods.

After the fragmentation step is complete, duplicates have to be removed. A good practice before proceeding to the removal, is to standardize the fragments as a preliminary step. Standardization is a multi-step process and includes the conversion of all molecules to a consistent format, tautomer enumeration and canonicalization, charge neutralization,

stereochemistry determination and the filtering-out of salt and solvent metabolites. Standardization becomes even more imperative when multiple database entries are aggregated, since each database may employ specific method to standardize their molecules and duplicates may go undetected. Open source suites such as RDKit now include the program MolVS[51] to standardize lists of molecules. Alternatively, VS-Prep[52] is a robust library preparation tool possessing many functions and filters. ChemAxon offers a standardizer and a structure checker to correct various structural issues. Similar to fragmentation tools, standardization kits are included in most Molecular Mechanics suites, such as the WASH function in MOE. After standardization, duplicates can be removed using canonical SMILES and/or InChiKey representations.

Further library expurgation can then be performed at this stage with the help of toxicity/reactivity and physicochemical filters to eliminate undesirable fragments. Regarding the former type of filtering, it is important to note that it does not constitute the final cleaning-up of the database, but that downstream filtering after the construction of molecules is required, since toxic or reactive effects may appear after the fragments are combined. In addition, due to the fact that deleterious effects may appear, but also disappear, when fragments are combined together, cautious procedure are to be taken with radical filters so as to not eliminate all the initially unfitting fragments. Systematically tagging the unfitting fragments in order to track them after their assembly may be seen as an alternative to their removal.

To determine toxicity/reactivity criteria, the following tools are available. For the prediction of ADME-Tox or reactive molecules, one can refer to ADMET[53], ADMET Predictor, and Percepta. Additional options include: SwissADME[54], Tox-Predict, PreADMET, FAF-Drugs4[55], eMolTox[56], and recently admetSAR 2.0[57]. There are also more specific tools tailored for cardiac toxicity (Pred-hERG[58]), skin sensitization (Pred-Skin[59]) or PAINS checking[60]. Some of the aforementioned computational solutions are available as webservers (see Supplementary Table 1), which can prove convenient for chemists with sparse knowledge in cheminformatics. It can also serve as a shortcut to skip the software installation and preparation steps, although confidentiality and data sharing agreements are potential issues to be considered prior to the submission of compounds.

Regarding the physicochemical assessment, one of the most used method is the compliance to the Ro3 (see *Introduction*). But alternative filtering rules, such as those described by Veber and al.[61], can be applied. Another possibility is to use tridimensional shape fitting with PMI[62] to discriminate flat and round molecules. Those thresholds should be adapted with great care to a library and its purpose (experimental or only virtual). The properties of *in silico* fragments

may indeed differ from biophysical ones[63]. Finally, one can chose among the myriad of existing descriptors[64] to create a personalized filter.

The representation method known as “molecular fingerprints” is a highly effective mean of evaluating the diversity of fragment library. The molecules are treated as bit strings, which allows rapid pairwise comparisons. The two major types of fingerprints are the structural fingerprints based on pre-existing substructure features, and the hash fingerprints that extract features of the molecule and use the hash to determine the bits. A variety of fingerprints are accessible to researchers, but MACCS keys[65] and EFCP[66], for structural and hash fingerprints respectively, are the most popular. More details on molecular fingerprint and similarity searching can be found in Cereto-Massagué et al.[67]. In addition, statistical calculations based on various descriptors, such as clustering[68] or PCA[69], permit both an analysis and an overview of the distribution of the molecules in the dataset.

When a virtual library is made, several formats are available for data storage. While opting for a personal or proprietary format can prevent confidentiality breaches, it may on the other hand induce compatibility issues with other software packages. The most common formats are SDF and CSV. They allow to store molecules with their coordinates together with all their properties and are usually read by most software. For the visualization and exploration of the chemical space of a library, DataWarrior[70] can be employed.

VIRTUAL SCREENING

Cell activity assays traditionally operated in HTS are not adapted for the screening of fragments, because their minute size renders the binding of fragment too weak to be detectable[71]. Therefore, alternative techniques have been developed or adapted to detect these interactions. Methods of choice comprise NMR spectroscopy, surface plasmon resonance (SPR) and X-ray crystallography[72]. They however present the drawback of requiring large quantities of protein and fragments for the production of readable raw data, which in turn necessitates that the fragments are soluble at high concentrations. The molecular species need also to be extremely pure which can be tricky to achieve. It follows that in practice experimental fragment screening is able to test hundreds to several thousand items but is not operational at higher scales. To screen the hundreds of thousands or even the millions of available fragments, virtual screening may constitute the sole solution at present. Fragment docking appears indeed especially useful for sorting out fragments in large libraries prior to experimental testing[73].

Pre-processing. To carry out a resilient virtual screening program and achieve satisfactory results, an essential step is the preparation, known as pre-processing, of both the fragments to screen and the target[74]. The aromatization of the molecules is standardized, titrable groups are ionized at the aimed pH, tautomeric states are enumerated, and if not done, a 3D conformer is generated which all docking programs require to run calculations. An efficient 2D to 3D conversion is a key step to most computational analyses. Depending on the library size, the accurate conformer generation can be a long step. For the target, the addition of hydrogen atoms and missing loops, the resolution of atom clashes, the protonation of the residues and particularly those of the binding site, which must be carefully verified[75]. Fortunately, programs can automate the procedure: the preparation of fragments for virtual screening can be achieved with VSPrep[52], and, together with the preparation of the target, SPORES[76]. RDKit can also generate 3D conformations using the ETKDG method[77]. Otherwise, the LigPrep[78] module (Schrödinger suite) allows the preparation of ligands and 3D conformations, and MOE's "Structure prepare" module enables the preparation of the target. Last but not least, the Protoss[79] webserver is also available to prepare proteins before virtual screening.

Detection of binding sites and hot spots. Once the fragment collection and the target are prepared, the following step is to identify and characterize important regions, termed "hot spots", which belong to the active site of the target and where fragments are prone to bind. Characterizing the binding site is straightforward when an X-ray/NMR/EM structure of the target including a ligand is available. In such a case, the binding site can simply be defined by the residues surrounding the ligand (usually within 5 Å). If no structure is accessible, one can resort to homology modelling, a prediction method allowing to obtain a 3D structure prediction of a target. The method notably relies on the concept that proteins possessing a medium to high degree of sequence conservation, also share similar secondary and tertiary structures[80]. Popular algorithms for modelling, reviewed by Liu et Capriotti[81], include MODELLER[82], SWISS-MODEL[83], HHpred[84] and I-TASSER[85]. An important element to the homology procedure is first to retrieve proteins with similar sequence through robust sequence alignment algorithms, such as BLAST[86]. Once identified (it is recommended that the 3D protein candidate has at least 40% sequence identity with the target), the model can be built, evaluated and refined. Recently, an AI-based prediction software, AlphaFold[87], has been published and demonstrates promising results in international protein folding competitions.

In the eventuality when no receptor-ligand structure is experimentally resolved, advanced computational methods can be put to contribution to elucidate the location of binding sites and hot spots. Two forerunner approaches were based on the GRID[88] and MCSS[89] tools and are

energy-oriented. An interaction energy between a sample of atoms/groups (e.g., water, methyl, hydroxyl) or a "probe", and the target, is calculated thanks to a force-field potential. The most favorable positions are identified and the surrounding residues are extracted. MCSS is based on a Monte Carlo simulation protocol where the probes are randomly placed into the binding site before undergoing an energy minimization, while GRID scans over the entire protein surface. Later, the CS-Map[90] method has been developed. A set of 14 organic solvent molecules are displaced around the protein surface, and the molecules are then clustered into consensus sites in which these probes are more likely to bind. More recently, the FTMap[91] technique takes advantage of 16 small molecular probes, with different degrees of hydrophobicity. It is based on the fast Fourier transform (FFT) correlation supporting the sampling of billions of probe positions. The last two methods are very similar to those employed experimentally such as Multiple Solvent Crystal Structures (MSCS). When confidentiality is not an issue, another possibility provided to researchers is Fragment Hotspot Maps[92], available via a web application. The potential binding sites interacting with a target, together with their localization and their qualitative assessment, can be executed in minutes. Alternatively, PLImap[93] combines the energy function of the PLIff[94] forcefield with a systematic grid-based search algorithm to identify favorable sites.

A central limitation with the aforementioned methods lies in the fact that they are structurally speaking "rigid". To cope with the lack of flexibility of the target, novel approaches coupled with Molecular Dynamics (MD) simulations were developed. SILCS[95] couples the binding of probes to all-atom explicit-solvent MD simulations, and then generates fragment binding probability regions, labeled "FragMaps". The MDmix[96] toolkit also employs MD simulations, but with different solvent parameters and with the addition of water displaceability predictions. Currently, MD-based methods still require significant computational running-power and user knowledge for the preparation and analysis of results. However, with the constant developments in the software and hardware industries, such methods should become more and more user-friendly and established as procedures for determining protein hot spots[97].

Finally, Machine Learning-based methods have also been developed, but are rather tailored to the description of protein-protein interactions (PPIs)[98,99]. The following deep learning approaches are otherwise at disposal. DeepSite[100], which is a knowledge-based approach with convolutional neural networks trained via the scPDB database (7,622 proteins). DeepDrug3D[101] on the other hand is also geared towards the characterization and classification of binding pockets, but with a convolutional neural network energy-based algorithm.

Docking. The computational method that predicts the position and conformation of a ligand within its binding site in a target, is known as docking. It stems from the combination of a search algorithm returning potential poses and a scoring function ranking them. The outmost challenge with the docking method lies in the exponential, virtually infinite, number of possible positions of a molecule within a target. Therefore, docking programs are compelled to be fast and effective in order to explore the conformational space whilst returning realistic outputs, including poses that are similar to the native one. The scoring function is applied to evaluate the fit of a pose within a binding site. Given that the docking process might return a large number of solutions, rapidity and efficiency of execution are also required for the scoring function. Ideally, poses close to the native ones should be ranked first and highly diverging positions last.

Docking procedures can generally be classified as rigid or flexible, depending on the degree of freedom allocated to the conformational flexibility of the ligand and the receptor. Rigid docking can draw analogy with the "Lock and Key Model", first presented by Emil Fisher. Essentially geometric and physiochemical complementarities between the ligand and the target are considered. The method is opted for when computation time is highly valued and/or when the docking library contains a large number of compounds. It is nevertheless commonly preferred to use flexible docking, as it allows to obtain better-fitting poses. It can include the systematic enumeration of conformations, MD simulations, Monte Carlo and genetic algorithms. If both the ligand and the target are computed as flexible, the calculation running time can become very costly, due to the exponential number of conformational sampling possibilities. Semi-flexible docking is an intermediary solution and consists in the following. The translational and rotational conformational degrees of freedom of the sole ligand are sampled, while the receptor is kept rigid. The latter method has been adopted by the great majority of docking programs[102].

Over the last three decades, more than 60 different docking tools have been developed[103] and have been exhaustively reviewed[103–106]. It is important to note that the bulk of docking programs were developed and optimized for drug-like molecules, raising question marks about the relevance of applying the force fields and scoring functions to fragments[107].

Scoring. Three typical scoring functions have been developed: force field-, empirical-, and knowledge-based. The force field is a concept derived from molecular mechanics and is employed to evaluate the potential energy of a system. In molecular docking, the potential energy is equivalent to the sum of the strength of the intermolecular components, which is broken down as van der Waals and electrostatic interactions. Additional terms such as the

intramolecular energy (bond stretching and torsional forces), conformational entropy and ligand dehydration, can be included in the final score. The empirical scoring functions are constructed on experimental binding free energy values. Energy terms are weighted as follows. The coefficients are determined through regression analysis of a training set of protein–ligand complexes and the estimation of binding affinities. It has been reported however that the scoring method may be suboptimal for fragments[21]. The third type of scoring function is knowledge-based. The reproduction of experimental structures rather than binding free energies is aimed for. They are based on a statistical analysis of interacting atom pairs in protein–ligand complexes from structural databases. The gathered data is then translated into a pseudopotential, describing the preferred geometries of the protein–ligand pairwise atoms. More recently, ingenious scoring functions (e.g., RF[108]) have been developed based on machine learning principles. They appear to outperform the traditional functions[109,110]. Deep learning scoring function are even being distributed now such as NNScore 2.0[111].

Each scoring function conveys with its criteria some limitations and is therefore suboptimal. To balance this intrinsic property, the consensus technique has been implemented, where multiple scoring functions are combined, and has proven very successful[112–114]. Another promising strategy is to update the score of a pose according to the similarity of the interaction fingerprint to a reference (IFP)[115]. As a matter of fact, binding mode interactions between ligand and protein complexes are routinely used to re-rank docking calculations[116]. Tools allowing to quickly retrieve noncovalent interactions between proteins and their ligands include SPLIF[117] and PLIP[118]. One more post processing approach to re-rank poses, based on the 3D-structure of a reference, is ROCS[119] toolkit. It performs direct shape comparisons between the poses and the reference. Finally, using instead energetic considerations, the Molecular Mechanics Poisson-Bolzmann/Generalized Born Surface Area (MM-PBSA, MM-GBSA) methods exhibit a strong rescoring ability[120,121].

The central qualitative landmark of a scoring function is its ability to assign known binders to high ranks. A healthy habit prior to running a virtual screening campaign is to verify that the program and parameters to be applied are able to discriminate known inhibitors from decoys. Widely spread quantitative metrics to assess a scoring function, reviewed in ref.[122], are enrichment factors (EF) and ROC curve analyses.

Remarkably, no significant difference in docking performance between fragments and druglike compounds has been reported[122]. Therefore, standard docking programs can be run for fragment docking. The most common options[123] are AutoDock[124], GOLD[125], and Glide[126].

Docking can also be performed on webservers via the e-LEA3D platform using PLANTS[127] or via MTiOpenScreen using Autodock 4.2.6[124]. Nevertheless, SEED[128] has been specifically developed for fragment docking. The optimal fragment positions within a rigid receptor are determined and are ranked according to their respective binding energies (sum of the van der Waals interaction and binding energy).

FRAGMENT EXPANSION

Once the binding mode of a fragment is identified, it becomes the starting point, or "seed", for the design of a ligand. The end goal is to optimize and evolve the seed into a drug lead, whilst increasing its potency and selectiveness. To do so, three alternatives are at disposal: growing, merging and linking (Figure 4).

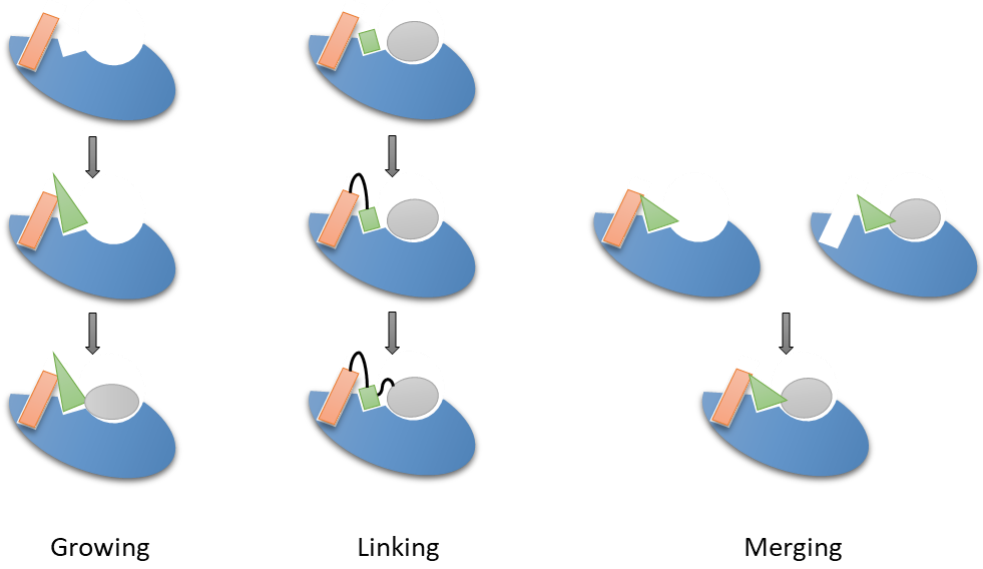


Figure 4. Fragment expansion methods.

Fragment growing is the most extensively applied strategy in FBDD[129]. It is indeed the less constraining method as only one predetermined seed is required, from which fragments are incrementally added. At each addition round, the ligand under construction is evaluated with docking attempts or more advanced criteria and further growth is either enabled or cancelled. Various programs implementing the method are the following. Early enactments include SPROUT[130] and GroupBuild[131]. As of recently, the AlleGrow application layered on GrowMol[132] can evaluate the interaction between the ligand and the binding site atoms, and

optimize it by energy minimization. LigBuilder 2[133] design ligands to best fit the shape of protein cavities. Drug-like fragments are prioritized according to ADME-Tox filters and atomic connection rules are enforced (e.g., O−O bonds are forbidden). The OpenGrowth[134] algorithm is trained with in-house databases and recommends drug-resembling structures. The modelled molecules tend to be more accessible to synthesis and display upgraded ADME properties in comparison to randomly generated molecules. Autogrow[135] randomly inserts fragments to the seed and performs docking positional refinement of the ligand being built. A genetic algorithm is applied to sort out promising drug-like fragments, which are then selected as reference points for the next generation. Custom filters are available, such as the possibility to discard any compound that does not contain key moieties. LEA3D[136] offers a collection of automated *de novo* drug design programs. Molecules are designed according to a library of molecular fragments. The best combinations of molecular fragments that fit predefined physicochemical constraints are determined. LEADOPT[137] generates molecules with respect to their Ligand Efficiency (LE) instead to scoring functions. In addition, ADME-Tox properties are evaluated thanks to the child-program SCADMET (including twelve important pharmacokinetic and toxic properties). Frags2Drugs is a recent suite allowing the design of new kinase inhibitors directly into the ATP binding site. It relies on an internal library of 3D fragments extracted from Protein Data Bank[138] (PDB). Kinase structures are stored as a graph representing the different binding configurations of the fragments. Kinase-specific filters[36] are employed for best outcomes.

Fragment linking corresponds to the original structure-activity relationships by nuclear magnetic resonance or "SAR by NMR" method[7]. Rather than initiating the protocol with a predetermined fragment, like the aforementioned procedure, multiples key interaction sites of the receptor are tested for. Then, linkers, consisting in one or more atoms, are used to connect the fragments together. The fragment linking has to be made without losing the initial binding mode, which constitutes a challenge. Furthermore, a methodological prerequisite is the fact that separate pockets must lie in the binding site of the receptor in order to support multiple fragments simultaneously. The following options are available. GANDI[139] can be accessed to automatically link pre-docked fragments. The computational tool exploits a genetic and minimization algorithm to check the energetic affinities of the modelled molecules. 2D and 3D-similarity mappings to known inhibitor(s) and binding mode(s) are also provided. PFVS, a method that allowed to discover picomolar inhibitors of membrane proteins, links fragments to a predetermined pharmacophore in the binding site. The pose is then optimized via minimization algorithms and MD simulations, together with binding free energy calculations. Finally, fragment linking can be done with the webserverACFIS[140].

Fragment merging combines the information of several reported hits and consists in the merging of fragments or ligands that contain a common chemical feature known to bind in the same target. The strategy can be used for the chemical modification of a ligand and derivative generation. The *in-silico* instrument identifies the largest substructure shared by two ligands, then superimposes the ligands, and 'mixes and matches' their chemical moieties to obtain different possible combinations. In medicinal chemistry, altering/swapping chemical groups of known ligands while keeping intact those responsible for receptor-binding is a common usage. The technique, analogous to fragment merging, is known as scaffold hopping[141], where fractions of the lead compound are methodologically swapped. Notable purposes are modifying affinity and selectivity characteristics, improving physicochemical and ADME-Tox properties, and finding patentable analogs. In a description of the fragment merging method[142], Hudson and colleagues provide an advisory checklist. It notably appears that the average distance between atoms shared by fragments is a key consideration for a successful merging. As a guideline, the authors estimate that an upper limit of 1 Å should be respected. Due to various implementation issues, fragment merging is least employed in FBDD projects[5]. It can nevertheless be run with the following computer packages. Ligmerge[143] aims to swap chemical moieties of known inhibitors in order to generate new molecules potentially displaying an improved potency. BREED[144] is provided as a module in MOE and Schrödinger. MED-SuMo[145] or Spark[146] aims to find equivalent replacements for key moieties in the lead compound. BROOD[44] can generate analogs of a lead by swapping select fragments sharing similar shapes and electrostatics (highest Tanimoto score). Recore[147]—performs fast replacements, seconded by a 3D database, of a given compound core, while letting the rest of the molecule unaffected. The tool provides the definition of custom pharmacophore constraints in order to restrict the number of output solutions.

The main challenge of *de novo* design and computational fragment expansion is to propose synthesizable compounds with balanced ADME-Tox properties. A number of software packages rigorously attempt to control for future synthesis potential: DOGS[148], LigBuilder2[133] and SYNOPSYS[149]. The PINGUI[150] program employs a "growing via merging" tactic. The computation is initiated with a seed bound in the active site, followed-up by its expansion with other fragments. The constraints imposed by the binding site of the target protein are taken into account and each fragment is added according to a compatible chemical reaction (a reference set of 58 robust organic reactions is used[151]). A different strategy is to assess the propensity of synthesis at a later stage. SYLVIA[152], the Sa_Score[153] RDKit module and the SCScore[154], can

all rapidly qualitatively asses the molecule to be constructed and can allow to rule it out if non-synthesizable. Finally, AI and deep neural network methods have been investigated for the prediction of compound synthesis feasibility[155]. Various AI-based retrosynthetic tools are reviewed in ref.[156]. Examples include the following. The RTSA method (LillyMol toolkit) is intended to retrieve potential synthetic avenues[157] and RXN[158] can predict chemical reactions.

CONCLUSION

The main advantages of FBDD are the ease with which to score a hit within a small library of fragments, thereby circumventing important investments in automation equipment, and the freedom, compared to traditional HTS, that is granted to optimize an initial hit into a lead compound thanks to its divided size. *In silico* methods have now become an essential complementary element to experimental FDBB. Great performance boosts together with cost-cuts are rendered possible. As described in this review, it appears that the computational tools can be applied all along or to independent stages of the lead discovery/optimization process. The promising development of specialized software for fragment docking and virtual screening brings about rising applications for *de novo* design of compounds. The computational methods are especially resilient in identifying hotspots in cavities and characterizing binding sites. The conjunction of these methods with homology modelling and binding pocket retrieval permits to mitigate the eventual lack of experimentally determined 3D structures.

A spectrum of challenges in the field of fragment-based design remains to be addressed. First, the scoring methods of the pose after fragment docking need to be improved. Fragments may be accommodated in the same binding site, as well as in multiple sub-regions belonging to different pockets lying on the protein surface. The fact that the correct location can be found, but at the same time not be ranked the highest because of the presence of alternative poses yielding similar scores, presently constitutes an issue. It appears therefore that accurate protein models and the definition of precise binding sites are required. Another issue with computational calculations being run on static models is that the dynamic biological environment of the target can be misrepresented. However, the simple rotation or torsion of one or more residues lying in the binding site may cause important overall conformational changes and even modify the binding modes of the fragments found in the static model. Due to low conformational flexibility and the lack of sufficient interactions with surrounding residues, many docking poses that are generated can be questionable and many computational results would therefore not be reproducible *in vivo.* Furthermore, the compounds modelled from *in*

silico FBDD run the risk of not being directly approved. They may lack suitable nanomolar activity, target selectivity, pharmacokinetic profile and ADME-Tox properties. Instead, the computer predicted molecules can often be seen as "concept compounds" requiring further optimization. Nevertheless, a good fragment screening campaign is directed to display a better hit rate than a the screening of a random compound collection.

Altogether, FBDD, focusing on efficiency rather than affinity, has led to a significant change in the drug discovery practice. While *in silico* FBDD methods are used primarily in the early drug discovery stages (hit identification and expansion), it becomes apparent that the usage of many of the computational tools being developed in the field, will inexorably accelerate for the later stages (e.g., ADME-Tox prediction and synthesis prediction).

ASSOCIATED CONTENT

Supplementary Table 1.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pascal.bonnet@univ-orleans.fr. Phone : +33

Author Contributions

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

ABBREVIATIONS

ADME-Tox: Absorption, Distribution, Metabolism, Excretion and Toxicity

AI: Artificial Intelligence

FDA: Food and Drug Administration

GPCR: G protein-coupled receptor

LE: Ligand Efficiency

MD: Molecular Dynamics

ML: Machine Learning

MM-GBSA: Molecular Mechanics Generalized Born Surface Area

PAINS: Pan Assay Interference Compounds

PCA: Principal Component Analysis

PMI: Principal Moments of Inertia

PPIs: Protein-Protein Interactions

REFERENCES

(1) Mayr, L. M.; Bojanic, D. Novel Trends in High-Throughput Screening. *Curr. Opin. Pharmacol.* **2009**, *9* (5), 580–588. https://doi.org/10.1016/j.coph.2009.08.004.

(2) 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.; et al. Impact of High-Throughput Screening in Biomedical Research. *Nat. Rev. Drug Discov.* **2011**, *10* (3), 188–195. https://doi.org/10.1038/nrd3368.

(3) Keserü, G. M.; Makara, G. M. The Influence of Lead Discovery Strategies on the Properties of Drug Candidates. *Nat. Rev. Drug Discov.* **2009**, *8* (3), 203–212. https://doi.org/10.1038/nrd2796.

(4) Kirkpatrick, P.; Ellis, C. Chemical space https://www.nature.com/articles/432823a (accessed Apr 23, 2019). https://doi.org/10.1038/432823a.

(5) 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.

(6) Hall, R. J.; Mortenson, P. N.; Murray, C. W. Efficient Exploration of Chemical Space by Fragment-Based Screening. *Prog. Biophys. Mol. Biol.* **2014**, *116* (2), 82–91. https://doi.org/10.1016/j.pbiomolbio.2014.09.007.

(7) 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.

(8) 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.

(9) 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.

(10) Romasanta, A. K.; van der Sijde, P.; Hellsten, I.; Hubbard, R. E.; Keseru, G. M.; van Muijlwijk-Koezen, J.; de Esch, I. J. When Fragments Link: A Bibliometric Perspective on the Development of Fragment-Based Drug Discovery. *Drug Discov. Today* **2018**.

(11) Erlanson, D. Practical Fragments: Fragments in the Clinic: 2018 Edition. *Practical Fragments*, 2018.

(12) 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.

(13) Souers, A. J.; Leverson, J. D.; Boghaert, E. R.; Ackler, S. L.; Catron, N. D.; Chen, J.; Dayton, B. D.; Ding, H.; Enschede, S. H.; Fairbrother, W. J.; et al. ABT-199, a Potent and Selective BCL-2 Inhibitor, Achieves Antitumor Activity While Sparing Platelets. *Nat. Med.* **2013**, *19* (2), 202–208. https://doi.org/10.1038/nm.3048.

(14) 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.; et al. 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.

(15) Fragment definition and meaning | Collins English Dictionary https://www.collinsdictionary.com/dictionary/english/fragment (accessed Apr 24, 2019).

(16) 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* (1), 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.

(17) 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.

(18) Zhu, Z.; Cuozzo, J. Review Article: High-Throughput Affinity-Based Technologies for Small-Molecule Drug Discovery. *J. Biomol. Screen.* **2009**, *14* (10), 1157–1164. https://doi.org/10.1177/1087057109350114.

(19) Davis, B. J.; Erlanson, D. A. Learning from Our Mistakes: The 'Unknown Knowns' in Fragment Screening. *Bioorg. Med. Chem. Lett.* **2013**, *23* (10), 2844–2852. https://doi.org/10.1016/j.bmcl.2013.03.028.

(20) Früh, V.; Zhou, Y.; Chen, D.; Loch, C.; Ab, E.; Grinkova, Y. N.; Verheij, H.; Sligar, S. G.; Bushweller, J. H.; Siegal, G. Application of Fragment-Based Drug Discovery to Membrane Proteins: Identification of Ligands of the Integral Membrane Enzyme DsbB. *Chem. Biol.* **2010**, *17* (8), 881–891. https://doi.org/10.1016/j.chembiol.2010.06.011.

(21) 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.

(22) Bian, Y.; Xie, X.-Q. (Sean). Computational Fragment-Based Drug Design: Current Trends, Strategies, and Applications. *AAPS J.* **2018**, *20* (3), 59. https://doi.org/10.1208/s12248-018-0216-7.

(23) DesJarlais, R. L. Chapter Six - Using Computational Techniques in Fragment-Based Drug Discovery. In *Methods in Enzymology*; Kuo, L. C., Ed.; Fragment-Based Drug Design; Academic Press, 2011; Vol. 493, pp 137–155. https://doi.org/10.1016/B978-0-12-381274-2.00006-6.

(24) Rognan, D. Fragment-Based Approaches and Computer-Aided Drug Discovery. In *Fragment-Based Drug Discovery and X-Ray Crystallography*; Davies, T. G., Hyvönen, M., Eds.; Topics in Current Chemistry; Springer Berlin Heidelberg: Berlin, Heidelberg, 2012; pp 201–222. https://doi.org/10.1007/128_2011_182.

(25) Sheng, C.; Zhang, W. Fragment Informatics and Computational Fragment-Based Drug Design: An Overview and Update. *Med. Res. Rev.* **2013**, *33* (3), 554–598. https://doi.org/10.1002/med.21255.

(26) Fragment Collections | Cambridge MedChem Consulting https://www.cambridgemedchemconsulting.com/resources/hit_identification/fragment_collections.html (accessed May 21, 2019).

(27) 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.

(28) Chen, I.-J.; Hubbard, R. E. Lessons for Fragment Library Design: Analysis of Output from Multiple Screening Campaigns. *J. Comput. Aided Mol. Des.* **2009**, *23* (8), 603–620. https://doi.org/10.1007/s10822-009-9280-5.

(29) 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* (D1), D945–D954. https://doi.org/10.1093/nar/gkw1074.

(30) 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* (D1), D1102–D1109. https://doi.org/10.1093/nar/gky1033.

(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) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; et al. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D1074–D1082. https://doi.org/10.1093/nar/gkx1037.

(33) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein–Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35* (Database issue), D198–D201. https://doi.org/10.1093/nar/gkl999.

(34) Pihan, E.; Colliandre, L.; Guichou, J.-F.; Douguet, D. E-Drug3D: 3D Structure Collections Dedicated to Drug Repurposing and Fragment-Based Drug Design. *Bioinforma. Oxf. Engl.* **2012**, *28* (11), 1540–1541. https://doi.org/10.1093/bioinformatics/bts186.

(35) Chan, W. K. B.; Zhang, H.; Yang, J.; Brender, J. R.; Hur, J.; Özgür, A.; Zhang, Y. GLASS: A Comprehensive Database for Experimentally Validated GPCR-Ligand Associations. *Bioinformatics* **2015**, *31* (18), 3035–3042. https://doi.org/10.1093/bioinformatics/btv302.

(36) 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.

(37) Lewell, X. Q.; Judd, D. B.; Watson, S. P.; Hann, M. M. RECAP Retrosynthetic 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.

(38) Degen Jörg; Wegscheid-Gerlach Christof; Zaliani Andrea; Rarey Matthias. 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.

(39) 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.

(40) Ghersi, D.; Singh, M. MolBLOCKS: Decomposing Small Molecule Sets and Uncovering Enriched Fragments. *Bioinforma. Oxf. Engl.* **2014**, *30* (14), 2081–2083. https://doi.org/10.1093/bioinformatics/btu173.

(41) Liu, T.; Naderi, M.; Alvin, C.; Mukhopadhyay, S.; Brylinski, M. Break Down in Order To Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with EMolFrag. *J. Chem. Inf. Model.* **2017**, *57* (4), 627–631. https://doi.org/10.1021/acs.jcim.6b00596.
(42) *Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group ULC, 1010 Sherbooke St.*
(43) *Schrödinger, LLC, New York, NY, 2019*.
(44) *BROOD 3.1.0.3: OpenEye Scientific Software, Santa Fe, NM. Http://Www.Eyesopen.Com.*
(45) *CoLibri Version 4.2, BioSolveIT, Http://Www.Biosolveit.de/CoLibri.*
(46) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.
(47) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminformatics* **2017**, *9* (1), 33. https://doi.org/10.1186/s13321-017-0220-4.
(48) *RDKit, Open-Source Cheminformatics. Http://Www.Rdkit.Org.*
(49) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*; Springer, 2007.
(50) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39* (15), 2887–2893. https://doi.org/10.1021/jm9602928.
(51) MolVS: Molecule Validation and Standardization — MolVS 0.1.1 documentation https://molvs.readthedocs.io/en/latest/ (accessed May 21, 2019).
(52) 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* (10), 1700023. https://doi.org/10.1002/minf.201700023.
(53) Directory of in silico Drug Design tools - ADMET https://www.click2drug.org/directory_ADMET.html (accessed May 3, 2019).
(54) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, *7*, 42717. https://doi.org/10.1038/srep42717.
(55) Lagorce, D.; Bouslama, L.; Becot, J.; Miteva, M. A.; Villoutreix, B. O. FAF-Drugs4: Free ADME-Tox Filtering Computations for Chemical Biology and Early Stages Drug Discovery. *Bioinforma. Oxf. Engl.* **2017**, *33* (22), 3658–3660. https://doi.org/10.1093/bioinformatics/btx491.
(56) Ji, C.; Svensson, F.; Zoufir, A.; Bender, A. EMolTox: Prediction of Molecular Toxicity with Confidence. *Bioinformatics* **2018**, *34* (14), 2508–2509. https://doi.org/10.1093/bioinformatics/bty135.
(57) Yang, H.; Lou, C.; Sun, L.; Li, J.; Cai, Y.; Wang, Z.; Li, W.; Liu, G.; Tang, Y. AdmetSAR 2.0: Web-Service for Prediction and Optimization of Chemical ADMET Properties. *Bioinformatics* **2019**, *35* (6), 1067–1069. https://doi.org/10.1093/bioinformatics/bty707.
(58) Braga, R. C.; Alves, V. M.; Silva, M. F. B.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-HERG: A Novel Web-Accessible Computational

Tool for Predicting Cardiac Toxicity. *Mol. Inform.* **2015**, *34* (10), 698–701. https://doi.org/10.1002/minf.201500040.
(59) Braga, R. C.; Alves, V. M.; Muratov, E. N.; Strickland, J.; Kleinstreuer, N.; Trospsha, A.; Andrade, C. H. Pred-Skin: A Fast and Reliable Web Application to Assess Skin Sensitization Effect of Chemicals. *J. Chem. Inf. Model.* **2017**, *57* (5), 1013–1017. https://doi.org/10.1021/acs.jcim.7b00194.
(60) 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.
(61) 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.
(62) 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* (3), 987–1003. https://doi.org/10.1021/ci025599w.
(63) Konteatis, Z. D. In Silico Fragment-Based Drug Design. *Expert Opin. Drug Discov.* **2010**, *5* (11), 1047–1065. https://doi.org/10.1517/17460441.2010.523697.
(64) Molecular Descriptors Software http://www.moleculardescriptors.eu/softwares/softwares.htm (accessed Apr 19, 2019).
(65) 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.
(66) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754. https://doi.org/10.1021/ci100050t.
(67) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular Fingerprint Similarity Search in Virtual Screening. *Methods* **2015**, *71*, 58–63. https://doi.org/10.1016/j.ymeth.2014.08.005.
(68) 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* (4), 747–750. https://doi.org/10.1021/ci9803381.
(69) Osolodkin, D. I.; Radchenko, E. V.; Orlov, A. A.; Voronkov, A. E.; Palyulin, V. A.; Zefirov, N. S. Progress in Visual Representations of Chemical Space. *Expert Opin. Drug Discov.* **2015**, *10* (9), 959–973. https://doi.org/10.1517/17460441.2015.1060216.
(70) Sander, T.; Freyss, J.; von Korff, M.; Rufener, C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. *J. Chem. Inf. Model.* **2015**, *55* (2), 460–473. https://doi.org/10.1021/ci500588j.
(71) Murray, C. W.; Rees, D. C. The Rise of Fragment-Based Drug Discovery. *Nat. Chem.* **2009**, *1* (3), 187–192. https://doi.org/10.1038/nchem.217.
(72) Robson-Tull, J. Biophysical Screening in Fragment-Based Drug Design: A Brief Overview. *Biosci. Horiz. Int. J. Stud. Res.* **2018**, *11*. https://doi.org/10.1093/biohorizons/hzy015.
(73) Chen, Y.; Shoichet, B. K. Molecular Docking and Ligand Specificity in Fragment-Based Inhibitor Discovery. *Nat. Chem. Biol.* **2009**, *5* (5), 358–364. https://doi.org/10.1038/nchembio.155.
(74) Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments | SpringerLink https://link.springer.com/article/10.1007/s10822-013-9644-8 (accessed May 6, 2019).

(75) Bax, B.; Chung, C.; Edge, C. Getting the Chemistry Right: Protonation, Tautomers and the Importance of H Atoms in Biological Chemistry. *Acta Crystallogr. Sect. Struct. Biol.* **2017**, *73* (2), 131–140. https://doi.org/10.1107/S2059798316020283.
(76) ten Brink, T.; Exner, T. E. PKa Based Protonation States and Microspecies for Protein–Ligand Docking. *J. Comput. Aided Mol. Des.* **2010**, *24* (11), 935–942. https://doi.org/10.1007/s10822-010-9385-x.
(77) 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.
(78) *LigPrep Schrödinger Release 2019-1: LigPrep, Schrödinger, LLC, New York, NY, 2019.*
(79) Bietz, S.; Urbaczek, S.; Schulz, B.; Rarey, M. Protoss: A Holistic Approach to Predict Tautomers and Protonation States in Protein-Ligand Complexes. *J. Cheminformatics* **2014**, *6*, 12. https://doi.org/10.1186/1758-2946-6-12.
(80) Kaczanowski, S.; Zielenkiewicz, P. Why Similar Protein Sequences Encode Similar Three-Dimensional Structures? *Theor. Chem. Acc.* **2010**, *125* (3), 643–650. https://doi.org/10.1007/s00214-009-0656-3.
(81) Liu, T.; Capriotti, G. W. T. and E. Comparative Modeling: The State of the Art and Protein Drug Target Structure Prediction http://www.eurekaselect.com/74238/article (accessed May 6, 2019).
(82) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinforma.* **2016**, *54*, 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.
(83) Waterhouse, A.; Bertoni, M.; Bienert, S.; Studer, G.; Tauriello, G.; Gumienny, R.; Heer, F. T.; de Beer, T. A. P.; Rempfer, C.; Bordoli, L.; et al. SWISS-MODEL: Homology Modelling of Protein Structures and Complexes. *Nucleic Acids Res.* **2018**, *46* (W1), W296–W303. https://doi.org/10.1093/nar/gky427.
(84) Söding, J. Protein Homology Detection by HMM-HMM Comparison. *Bioinforma. Oxf. Engl.* **2005**, *21* (7), 951–960. https://doi.org/10.1093/bioinformatics/bti125.
(85) I-TASSER server for protein 3D structure prediction | BMC Bioinformatics | Full Text https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-9-40 (accessed May 6, 2019).
(86) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215* (3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
(87) Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, S.; Bridgland, A.; Penedones, H.; et al. *De Novo Structure Prediction with Deep-Learning Based Scoring*; 2018.
(88) 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.
(89) 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.
(90) Dennis, S.; Kortvelyesi, T.; Vajda, S. Computational Mapping Identifies the Binding Sites of Organic Solvents on Proteins. *Proc. Natl. Acad. Sci.* **2002**, *99* (7), 4290–4295. https://doi.org/10.1073/pnas.062398499.
(91) Brenke, R.; Kozakov, D.; Chuang, G.-Y.; Beglov, D.; Hall, D.; Landon, M. R.; Mattos, C.; Vajda, S. Fragment-Based Identification of Druggable "hot Spots" of Proteins Using Fourier Domain Correlation Techniques. *Bioinforma. Oxf. Engl.* **2009**, *25* (5), 621–627. https://doi.org/10.1093/bioinformatics/btp036.

(92) Radoux, C. J.; Olsson, T. S. G.; Pitt, W. R.; Groom, C. R.; Blundell, T. L. Identifying Interactions That Determine Fragment Binding at Protein Hotspots. *J. Med. Chem.* **2016**, *59* (9), 4314–4325. https://doi.org/10.1021/acs.jmedchem.5b01980.
(93) Rathi, P. C.; Ludlow, R. F.; Hall, R. J.; Murray, C. W.; Mortenson, P. N.; Verdonk, M. L. Predicting "Hot" and "Warm" Spots for Fragment Binding. *J. Med. Chem.* **2017**, *60* (9), 4036–4046. https://doi.org/10.1021/acs.jmedchem.7b00366.
(94) Verdonk, M. L.; Ludlow, R. F.; Giangreco, I.; Rathi, P. C. Protein–Ligand Informatics Force Field (PLIff): Toward a Fully Knowledge Driven "Force Field" for Biomolecular Interactions. *J. Med. Chem.* **2016**, *59* (14), 6891–6902. https://doi.org/10.1021/acs.jmedchem.6b00716.
(95) Faller, C. E.; Raman, E. P.; MacKerell, A. D.; Guvench, O. Site Identification by Ligand Competitive Saturation (SILCS) Simulations for Fragment-Based Drug Design. *Methods Mol. Biol. Clifton NJ* **2015**, *1289*, 75–87. https://doi.org/10.1007/978-1-4939-2486-8_7.
(96) Alvarez-Garcia, D.; Barril, X. Molecular Simulations with Solvent Competition Quantify Water Displaceability and Provide Accurate Interaction Maps of Protein Binding Sites. *J. Med. Chem.* **2014**, *57* (20), 8530–8539. https://doi.org/10.1021/jm5010418.
(97) Arcon, J. P.; Defelipe, L. A.; Modenutti, C. P.; López, E. D.; Alvarez-Garcia, D.; Barril, X.; Turjanski, A. G.; Martí, M. A. Molecular Dynamics in Mixed Solvents Reveals Protein–Ligand Interactions, Improves Docking, and Allows Accurate Binding Free Energy Predictions. *J. Chem. Inf. Model.* **2017**, *57* (4), 846–863. https://doi.org/10.1021/acs.jcim.6b00678.
(98) Morrow, J. K.; Zhang, S. Computational Prediction of Hot Spot Residues. *Curr. Pharm. Des.* **2012**, *18* (9), 1255–1265.
(99) Melo, R.; Fieldhouse, R.; Melo, A.; Correia, J. D. G.; Cordeiro, M. N. D. S.; Gümüş, Z. H.; Costa, J.; Bonvin, A. M. J. J.; Moreira, I. S. A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces. *Int. J. Mol. Sci.* **2016**, *17* (8). https://doi.org/10.3390/ijms17081215.
(100) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: Protein-Binding Site Predictor Using 3D-Convolutional Neural Networks. *Bioinformatics* **2017**, *33* (19), 3036–3042. https://doi.org/10.1093/bioinformatics/btx350.
(101) Pu, L.; Govindaraj, R. G.; Lemoine, J. M.; Wu, H.-C.; Brylinski, M. DeepDrug3D: Classification of Ligand-Binding Pockets in Proteins with a Convolutional Neural Network. *PLOS Comput. Biol.* **2019**, *15* (2), e1006718. https://doi.org/10.1371/journal.pcbi.1006718.
(102) Meng, X.-Y.; Zhang, H.-X.; Mezei, M.; Cui, M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Curr. Comput. Aided Drug Des.* **2011**, *7* (2), 146–157.
(103) 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.
(104) de Ruyck, J.; Brysbaert, G.; Blossey, R.; Lensink, M. F. Molecular Docking as a Popular Tool in Drug Design, an in Silico Travel. *Adv. Appl. Bioinforma. Chem. AABC* **2016**, *9*, 1–11. https://doi.org/10.2147/AABC.S105289.
(105) Yuriev, E.; Ramsland, P. A. Latest Developments in Molecular Docking: 2010–2011 in Review. *J. Mol. Recognit.* **2013**, *26* (5), 215–239. https://doi.org/10.1002/jmr.2266.
(106) Ferreira, L. G.; Dos Santos, R. N.; Oliva, G.; Andricopulo, A. D. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules* **2015**, *20* (7), 13384–13421. https://doi.org/10.3390/molecules200713384.

(107) Joseph-McCarthy, D.; Campbell, A. J.; Kern, G.; Moustakas, D. Fragment-Based Lead Discovery and Design. *J. Chem. Inf. Model.* **2014**, *54* (3), 693–704. https://doi.org/10.1021/ci400731w.
(108) Wang, C.; Zhang, Y. Improving Scoring-Docking-Screening Powers of Protein-Ligand Scoring Functions Using Random Forest. *J. Comput. Chem.* **2017**, *38* (3), 169–177. https://doi.org/10.1002/jcc.24667.
(109) Ashtawy, H. M.; Mahapatra, N. R. A Comparative Assessment of Predictive Accuracies of Conventional and Machine Learning Scoring Functions for Protein-Ligand Binding Affinity Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **2015**, *12* (2), 335–347. https://doi.org/10.1109/TCBB.2014.2351824.
(110) Wójcikowski, M.; Ballester, P. J.; Siedlecki, P. Performance of Machine-Learning Scoring Functions in Structure-Based Virtual Screening. *Sci. Rep.* **2017**, *7*. https://doi.org/10.1038/srep46710.
(111) Durrant, J. D.; McCammon, J. A. NNScore 2.0: A Neural-Network Receptor–Ligand Scoring Function. *J. Chem. Inf. Model.* **2011**, *51* (11), 2897–2903. https://doi.org/10.1021/ci2003889.
(112) Palacio-Rodríguez, K.; Lans, I.; Cavasotto, C. N.; Cossio, P. Exponential Consensus Ranking Improves the Outcome in Docking and Receptor Ensemble Docking. *Sci. Rep.* **2019**, *9* (1), 5142. https://doi.org/10.1038/s41598-019-41594-3.
(113) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins. *J. Med. Chem.* **1999**, *42* (25), 5100–5109. https://doi.org/10.1021/jm990352k.
(114) Oda, A.; Tsuchida, K.; Takakura, T.; Yamaotsu, N.; Hirono, S. Comparison of Consensus Scoring Strategies for Evaluating Computational Models of Protein−Ligand Complexes. *J. Chem. Inf. Model.* **2006**, *46* (1), 380–391. https://doi.org/10.1021/ci050283k.
(115) Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints - Journal of Chemical Information and Modeling (ACS Publications) https://pubs.acs.org/doi/10.1021/ci600342e (accessed May 3, 2019).
(116) Jacquemard, C.; Drwal, M. N.; Desaphy, J.; Kellenberger, E. Binding Mode Information Improves Fragment Docking. *J. Cheminformatics* **2019**, *11* (1), 24. https://doi.org/10.1186/s13321-019-0346-7.
(117) Da, C.; Kireev, D. Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study. *J. Chem. Inf. Model.* **2014**, *54* (9), 2555–2561. https://doi.org/10.1021/ci500319f.
(118) Salentin, S.; Schreiber, S.; Haupt, V. J.; Adasme, M. F.; Schroeder, M. PLIP: Fully Automated Protein-Ligand Interaction Profiler. *Nucleic Acids Res.* **2015**, *43* (W1), W443-447. https://doi.org/10.1093/nar/gkv315.
(119) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50* (1), 74–82. https://doi.org/10.1021/jm0603365.
(120) Wichapong, K.; Rohe, A.; Platzer, C.; Slynko, I.; Erdmann, F.; Schmidt, M.; Sippl, W. Application of Docking and QM/MM-GBSA Rescoring to Screen for Novel Myt1 Kinase Inhibitors. *J. Chem. Inf. Model.* **2014**, *54* (3), 881–893. https://doi.org/10.1021/ci4007326.
(121) Greenidge, P. A.; Lewis, R. A.; Ertl, P. Boosting Pose Ranking Performance via Rescoring with MM-GBSA. *Chem. Biol. Drug Des.* **2016**, *88* (3), 317–328. https://doi.org/10.1111/cbdd.12763.

(122) 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.
(123) Chen, Y.-C. Beware of Docking! *Trends Pharmacol. Sci.* **2015**, *36* (2), 78–95. https://doi.org/10.1016/j.tips.2014.12.001.
(124) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30* (16), 2785–2791. https://doi.org/10.1002/jcc.21256.
(125) Verdonk, M. L.; Cole, J. C.; Hartshorn, M. J.; Murray, C. W.; Taylor, R. D. Improved Protein-Ligand Docking Using GOLD. *Proteins* **2003**, *52* (4), 609–623. https://doi.org/10.1002/prot.10465.
(126) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **2004**, *47* (7), 1739–1749. https://doi.org/10.1021/jm0306430.
(127) Korb, O.; Stützle, T.; Exner, T. E. PLANTS: Application of Ant Colony Optimization to Structure-Based Drug Design. In *Ant Colony Optimization and Swarm Intelligence*; Dorigo, M., Gambardella, L. M., Birattari, M., Martinoli, A., Poli, R., Stützle, T., Eds.; Lecture Notes in Computer Science; Springer Berlin Heidelberg, 2006; pp 247–258.
(128) Majeux, N.; Scarsi, M.; Apostolakis, J.; Ehrhardt, C.; Caflisch, A. Exhaustive docking of molecular fragments with electrostatic solvation. *Proteins Struct. Funct. Bioinforma.* **1999**, *37* (1), 88–105. https://doi.org/10.1002/(SICI)1097-0134(19991001)37:1<88::AID-PROT9>3.0.CO;2-O.
(129) 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.
(130) 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.
(131) 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.
(132) 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.
(133) 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.
(134) 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.
(135) 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.
(136) 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.
(137) 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.
(138) 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.
(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) 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.
(141) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening. *Angew. Chem. Int. Ed.* **1999**, *38* (19), 2894–2896. https://doi.org/10.1002/(SICI)1521-3773(19991004)38:19<2894::AID-ANIE2894>3.0.CO;2-F.
(142) Hudson, S. A.; Surade, S.; Coyne, A. G.; McLean, K. J.; Leys, D.; Munro, A. W.; Abell, C. Overcoming the Limitations of Fragment Merging: Rescuing a Strained Merged Fragment Series Targeting Mycobacterium Tuberculosis CYP121. *Chemmedchem* **2013**, *8* (9), 1451–1456. https://doi.org/10.1002/cmdc.201300219.
(143) 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.
(144) 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.
(145) Sperandio, O.; Andrieu, O.; Miteva, M. A.; Vo, M.-Q.; Souaille, M.; Delfaud, F.; Villoutreix, B. O. MED-SuMoLig: A New Ligand-Based Screening Tool for Efficient Scaffold Hopping. *J. Chem. Inf. Model.* **2007**, *47* (3), 1097–1110. https://doi.org/10.1021/ci700031v.
(146) *Spark, Version V10.5 , Cresset®, Litlington, Cambridgeshire, UK, ; Http://Www.Cresset-Group.Com/Spark/; Cheeseright, T.; Mackey, M.; Rose, S.; Vinter, A. Molecular Field Extrema as Descriptors of Biological Activity: Definition and Validation. J. Chem. Inf. Model. 2006, 46 (2), 665-676.*
(147) Patrick Maass, †; Tanja Schulz-Gasch, ‡; Martin Stahl, *; Matthias Rarey*, †. Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations https://pubs.acs.org/doi/abs/10.1021/ci060094h (accessed Apr 26, 2018). https://doi.org/10.1021/ci060094h.
(148) 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). https://doi.org/10.1371/journal.pcbi.1002380.
(149) 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.
(150) 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.

(151) Hartenfeller, M.; Eberle, M.; Meier, P.; Nieto-Oberhuber, C.; Altmann, K.-H.; Schneider, G.; Jacoby, E.; Renner, S. A Collection of Robust Organic Synthesis Reactions for in Silico Molecule Design. *J. Chem. Inf. Model.* **2011**, *51* (12), 3093–3098. https://doi.org/10.1021/ci200379p.
(152) Boda, K.; Seidel, T.; Gasteiger, J. Structure and Reaction Based Evaluation of Synthetic Accessibility. *J. Comput. Aided Mol. Des.* **2007**, *21* (6), 311–325. https://doi.org/10.1007/s10822-006-9099-2.
(153) 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.
(154) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.
(155) 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.
(156) Feng, F.; Lai, L.; Pei, J. Computational Chemical Synthesis Analysis and Pathway Design. *Front. Chem.* **2018**, *6*. https://doi.org/10.3389/fchem.2018.00199.
(157) Watson, I. A.; Wang, J.; Nicolaou, C. A. A Retrosynthetic Analysis Algorithm Implementation. *J. Cheminformatics* **2019**, *11* (1), 1. https://doi.org/10.1186/s13321-018-0323-6.
(158) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *ArXiv171104810 Cs Stat* **2017**.

Supplementary table 1. List of all tools referenced in the paper with their availability and website associated.

TOOL	USAGE	AVAILABILITY	WEBSITE
ChEMBL[29]	Library database	Free	https://www.ebi.ac.uk/chembl/
PubChem[30]	Library database	Free	https://pubchem.ncbi.nlm.nih.gov/
ZINC[31]	Library database	Free	http://zinc15.docking.org/
DrugBank[32]	Library database	Free	https://www.drugbank.ca/
BindingDB[33]	Library database	Free	https://www.bindingdb.org/bind/index.jsp
e-Drug3D[34]	Library database	Free	https://chemoinfo.ipmc.cnrs.fr/MOLDB/index.php
GLASS[35]	Library database	Free	https://zhanglab.ccmb.med.umich.edu/GLASS/
PKIDB[36]	Library database	Free	http://www.icoa.fr/pkidb/
eMolecules	Library database	Commercial	https://www.emolecules.com/
RECAP[37]	Library Fragmentation	Free	RDKit module
BRICS[38]	Library Fragmentation	Free	RDKit module
DAIM[39]	Library Fragmentation	Free	http://www.biochem-caflisch.uzh.ch/download
molBLOCKS[40]	Library Fragmentation	Free	http://compbio.cs.princeton.edu/molblocks/
eMolFrag[41]	Library Fragmentation	Free	https://github.com/liutairan/eMolFrag
MOE[42]	Library Fragmentation and Pre-processing	Commercial	https://www.chemcomp.com/Products.htm
Schrödinger[43]	Library Fragmentation	Commercial	https://www.schrodinger.com/
CHOMP[44]	Library Fragmentation	Commercial	https://www.eyesopen.com/brood
Colibri[45]	Library Fragmentation	Commercial	https://www.biosolveit.de/CoLibri/
Open Babel[46]	Library Fragmentation	Open-source	http://openbabel.org/wiki/Main_Page
CDK[47]	Library Fragmentation	Open-source	https://cdk.github.io/
RDKit[48]	Library Fragmentation and Pre-processing	Open-source	https://www.rdkit.org/
KNIME[49]	Library Fragmentation	Open-source	https://www.knime.com/
MolVS[51]	Library Standardization	Open-source	https://molvs.readthedocs.io/en/latest/
VS-Prep[52]	Library Standardization and Pre-processing	Free for academia	

ChemAxon	Library Standardization	Free fro academia	https://www.chemaxon.com
WASH	Library Standardization	Commercial	MOE module
ADMET[53]	Library Filtering		
PAINS[60]	Library Filtering	Webserver, free after registration	https://www.cbligand.org/PAINS/
SwissADME[54]	Library Filtering	Free webserver	http://www.swissadme.ch/
Tox-Predict	Library Filtering	Free webserver	https://apps.ideaconsult.net/ToxPredict
PreADMET	Library Filtering	Free webserver	https://preadmet.bmdrc.kr/
FAF-Drugs4[55]	Library Filtering	Free webserver	http://fafdrugs4.mti.univ-paris-diderot.fr/
eMolTox[56]	Library Filtering	Free webserver	http://xundrug.cn/moltox
admetSAR 20[57]	Library Filtering	Free webserver	http://lmmd.ecust.edu.cn/admetsar2/
Pred-hERG[58]	Library Filtering	Free webserver	http://predherg.labmol.com.br/
Pred-Skin[59]	Library Filtering	Free webserver	http://predskin.labmol.com.br/
ADMET Predictor	Library Filtering	Commercial	https://www.simulations-plus.com/software/admetpredictor/medehem-studio/
Percepta	Library Filtering	Commercial	https://www.acdlabs.com/products/percepta/
PMI[62]	Library Filtering		
MACCS keys[65]	Library Filtering		
EFCP[66]	Library Filtering		
DataWarrior[70]	Library Visualization	Free	http://www.openmolecules.org/datawarrior/
SPORES[76]	Pre-processing	Free for academia	https://uni-tuebingen.de/fakultaeten/mathematisch-naturwissenschaftliche-fakultaet/fachbereiche/pharmazie-und-biochemie/pharmazie/pharmazeutische-chemie/pd-dr-t-exner/research/spores/
LigPrep[78]	Pre-processing	Commercial	Schrödinger module
Protoss[79]	Pre-processing	Free webserver	
MODELLER[82]	Binding site/hot spot Elucidation	Free webserver	https://salilab.org/modeller/
SWISS-MODEL[83]	Binding site/hot spot Elucidation	Free webserver	https://swissmodel.expasy.org/
HHpred[84]	Binding site/hot spot Elucidation	Free webserver	https://toolkit.tuebingen.mpg.de/tools/hhpred
I-TASSER[85]	Binding site/hot spot Elucidation	Free webserver	https://zhanglab.ccmb.med.umich.edu/I-TASSER/
AlphaFold[87]	Binding site/hot spot Elucidation		
GRID[88]	Binding site/hot spot Elucidation		

MCSS[89]	Binding site/hot spot Elucidation		
CS-Map[90]	Binding site/hot spot Elucidation		
FTMap[91]	Binding site/hot spot Elucidation	Free after registration	http://ftmap.bu.edu/
Fragment Hotspot Maps[92]	Binding site/hot spot Elucidation	Free webserver	http://fragment-hotspot-maps.ccdc.cam.ac.uk/
PLImap[93]	Binding site/hot spot Elucidation	Open-source	https://bitbucket.org/AstexUK/pli
SILCS[95]	Binding site/hot spot Elucidation	Commercial	http://docs.silcsbio.com/2019.1/silcs/silcs.html
MDmix[96]	Binding site/hot spot Elucidation	Free	http://www.ub.edu/bl/software/
DeepSite[100]	Binding site/hot spot Elucidation	Free	www.playmolecule.org
DeepDrug3D[101]	Binding site/hot spot Elucidation	Open-source	https://github.com/pulimeng/DeepDrug3D
SPLIF[117]	Scoring		
PLIP[118]	Scoring	Free webserver	https://projects.biotec.tu-dresden.de/plip-web/plip/index
ROCS[119]	Scoring	Commercial	https://www.eyesopen.com/rocs
AutoDock[124]	Docking	Free, webserver	http://bioserv.rpbs.univ-paris-diderot.fr/services/MTiOpenScreen/
GOLD[125]	Docking	Commercial	https://www.ccdc.cam.ac.uk/solutions/csd-discovery/components/gold/
Glide[126]	Docking	Commercial	https://www.schrodinger.com/glide
PLANTS[127]	Docking	Free webserver	http://chemoinfo.ipmc.cnrs.fr/eDESIGN/index.html
SEED[128]	Docking		https://caflischlab-seed.readthedocs.io/en/latest/
SPROUT[130]	Expansion		http://www.keymodule.co.uk/products/sprout/sprout-classic.html
GroupBuild[131]	Expansion		
AlleGrow[132]	Expansion		
LigBuilder 2[133]	Expansion and Synthesis assessment	Free for academia	http://repharma.pku.edu.cn/ligbuilder/intro.html
OpenGrowth[134]	Expansion	Open-source	https://sourceforge.net/projects/opengrowth/
Autogrow[135]	Expansion	Open-source	https://autogrow.ucsd.edu/#download
LEA3D[136]	Expansion	Free for academia	https://chemoinfo.ipmc.cnrs.fr/LEA3D/index.html
LEADOPT[137]	Expansion	Free for non-profit institutions, available upon request	
Frags2Drugs	Expansion		http://sbc.icoa.fr/

GANDI[139]	Linking	Free for academia	http://www.biochem-caflisch.uzh.ch/download
PFVS	Linking		
ACFIS[140]	Linking	Free for academia	http://chemyang.ccnu.edu.cn/ccb/server/ACFIS/
Ligmerge[143]	Merging	Open-source	
BREED[144]	Merging	Commercial	MOE/Schrödinger module.
MED-SuMo[145]	Merging		
Recore[147]	Merging		
DOGS[148]	Synthesis assessment		
SYNOPSYS[149]	Synthesis assessment		
PINGUI[150]	Synthesis assessment	Free	http://kolblab.org/scubidoo/pingui/view/pg_customLib_getLib.php
SYLVIA[152]	Synthesis assessment	Commercial	https://www.mn-am.com/products/sylvia
Sa_Score[153]	Synthesis assessment	Open-source	https://github.com/rdkit/rdkit/tree/master/Contrib/SA_Score
SCScore[154]	Synthesis assessment		https://github.com/connorcoley/scscore
LillyMol	Synthesis assessment	Free	https://github.com/EliLillyCo/LillyMol
RXN[158]	Synthesis assessment	Free webserver	https://rxn.res.ibm.com/

« Dans le domaine scientifique, trouver la bonne formulation d'un problème permet souvent de le résoudre. »

L'univers dans une coquille de noix – Stephen Hawking

Chapitre 4 : Développement d'un logiciel de création de molécules via l'approche par fragments (Frags2Drugs)

Comme décrit dans le chapitre précédent, l'approche par fragments dans le cadre de la découverte et du développement de nouveaux médicaments est désormais hautement employée. Le programme présenté dans ce chapitre, Frags2Drugs (F2D), a été créé dans le but de concevoir des nouveaux inhibiteurs de protéines kinases directement dans le site actif de la cible à partir de fragments. Notre outil repose sur une chimiothèque interne de fragments en 3D enregistrée sous forme de graphe permettant l'usage de technologies récentes et proches de celles utilisées par les réseaux sociaux. A partir d'un fragment initial donné, F2D est capable de proposer rapidement des molécules spécifiques au site actif de la protéine kinase visée. Dans ce chapitre j'aborderai le contexte de la création de ce logiciel, les différentes étapes de sa validation et la nouvelle implémentation mise en place durant mon doctorat. Enfin, des résultats prometteurs sur de nouvelles molécules découvertes, synthétisées puis testées expérimentalement ainsi que différentes applications du programme seront aussi présentées.

N. B. : Une partie des illustrations dans ce chapitre sont réalisés avec des molécules et des fragments en 2D. Il s'agit là d'un but uniquement visuel afin de mieux percevoir et comprendre l'image, en réalité F2D traite exclusivement avec des composés en 3D.

4.1 Présentation de Frags2Drugs

4.1.1 Intérêt d'un nouveau logiciel

Au vu de l'abondance de programmes déjà développés en FBDD (partie 3.4), la question de la pertinence de créer un nouvel outil peut se poser. Plusieurs raisons ont poussé l'équipe à faire ce choix. Tout d'abord la vétusté de certains des programmes existants, développés dans les années 1990 et dont la maintenance et l'intégration aux outils actuels peut être compliquée, voire irréalisable. Cela mène à notre deuxième raison : la difficulté de s'approprier et de maitriser les logiciels déjà développés. En effet, de par leur nature certains d'entre eux peuvent être difficiles à comprendre et à prendre en main, notamment en raison du langage de programmation utilisé, d'éventuels formats propriétaires rendant les résultats impossibles à interfacer avec d'autres outils, d'une documentation incomplète ou obsolète, d'incompatibilité avec les ordinateurs récents ou certains systèmes d'exploitation… Une autre limitation soulevée est l'accessibilité des logiciels, notamment ceux sous licence, pouvant représenter un coût trop élevé pour l'équipe. Quant à ceux disponibles en ligne via un serveur web uniquement, des problèmes de confidentialité peuvent empêcher leur utilisation. A ces restrictions logistiques s'ajoutent celles de la performance avec un temps de calcul très long de certains outils. Cela est dû notamment à une exploration systématique des cavités des protéines et à des étapes de minimisation d'énergie coûteuses pour bien placer les fragments, nécessitant parfois plusieurs jours pour obtenir des résultats. De plus, nous nous sommes rendus compte que finalement, même en partant d'un fragment connu d'un ligand, par exemple la pyridine de l'imatinib, dans la plupart des cas, ce ligand initial ne figurait pas dans les solutions retournées par ces logiciels et ainsi ne permettant pas la validation de l'outil utilisé.

Il est dès lors apparue l'idée congrue de se lancer dans le développement d'un outil moderne, rapide et reposant sur les nombreuses données structurales disponibles dans la PDB. Grâce à ces données structurales, nous pouvons connaitre précisément l'emplacement des fragments et nous affranchir ainsi d'une exploration et d'une recherche systématique pour déterminer leurs positions spatiales. La combinaison de tous ces fragments placés dans un même référentiel spatial permet alors de créer rapidement de nouvelles molécules. Notre objectif final est d'obtenir un programme fonctionnel, facilement manipulable (même par un non chémoinformaticien), moderne et rapide. En outre, développer soi-même son logiciel procure une liberté totale sur le choix et la direction à emprunter. Cela va dans le sens de la volonté de l'équipe d'harmoniser ses différents programmes et projets vers une plateforme synchronisée (création d'un serveur web regroupant nos différents outils).

4.1.2 Introduction du projet

F2D est un programme initié et développé au sein du laboratoire avant mon arrivée par J.-M. Gally[182]. Je vais dans un premier temps revenir sur ses fonctionnalités et son implémentation originelle avant de parler des transformations que j'ai effectuées et des nouveautés implémentées.

4.1.2.1 Environnement de travail

Avec cette volonté d'harmoniser l'écosystème de l'équipe et d'éviter que chacun ne fasse ses scripts de son côté avec un langage différent, le langage Python a été choisi comme langage de référence. Le langage Python est devenu un standard en science des données (et science tout court) ces dernières années. Il bénéficie de nombreux modules déjà développés et d'une importante communauté très réactive[183]. De par l'existence et l'engouement de cette communauté, l'usage du langage Python épargne aux programmes de devenir rapidement obsolètes, voir inutilisables, à défaut d'un langage impopulaire, trop spécifique et/ou payant.

F2D est donc un programme développé en langage Python, dont les principales librairies utilisées sont spécifiées en Figure 33. Pour la gestion des librairies et des modules de programmation, le langage Python dispose d'un gestionnaire de paquets appelé Anaconda, qui permet de créer et gérer simplement des environnements virtuels propres à chaque projet (www.anaconda.com/). Anaconda a été largement plébiscité ces dernières années, notamment en science des données et dans la communauté scientifique pour simplifier la gestion des environnements de travail, la reproductibilité et le partage grâce à la simplicité de clonage d'environnements virtuels entre deux ordinateurs distants. En guise de cahier de laboratoire et pour le développement de F2D, nous avons opté pour les Jupyter notebooks. Pour gérer les différentes données du programme, nous utilisons les bibliothèques de science des données et de calculs numériques du langage Python, notamment celles développées dans le cadre de SciPy (https://www.scipy.org/). Les données sont ainsi manipulées sous forme de tableaux à doubles entrées (les « dataframes ») permettant des transformations et des calculs rapides directement sur les colonnes de nos choix. L'avantage de ces bibliothèques est l'optimisation et l'efficacité

[182] José-Manuel Gally, « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » (Orléans, 2017).
[183] « The Top Programming Languages 2019 - IEEE Spectrum », IEEE Spectrum: Technology, Engineering, and Science News, consulté le 18 septembre 2019, https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019.

puisque bien que l'interface soit en langage Python, les fonctionnalités sont développées en langage C ou langage Fortran pour une vitesse d'exécution accrue. Pour la partie chimie et la manipulation des molécules, nous utilisons la librairie RDKit (https://www.rdkit.org/). Là encore, il s'agit d'une librairie open-source, optimisée et suivie par une communauté réactive. Enfin, pour suivre le développement et les modifications apportés au code, un serveur Gitlab interne a été mis en place.

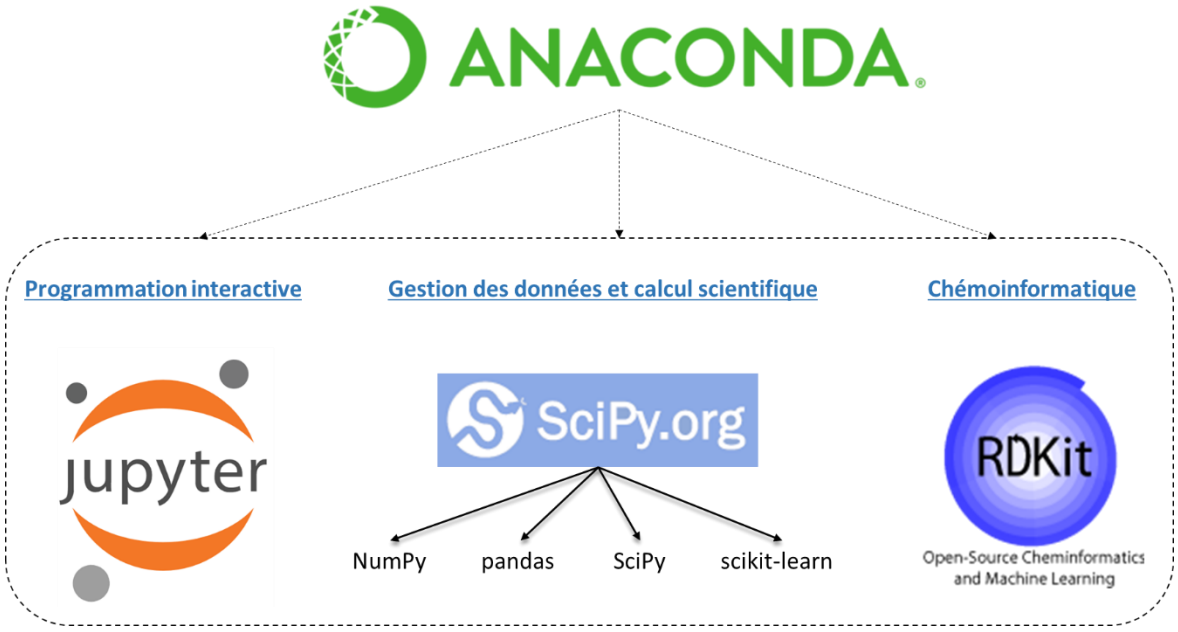


Figure 33 : Principaux outils en langage Python utilisés par F2D.

F2D est divisé en deux parties distinctes : dans un premier temps la création de la librairie de fragments et dans un deuxième temps la création de molécules par combinaisons de ces fragments. Ces deux parties sont indépendantes et F2D peut très bien être utilisé avec des données personnelles et une autre librairie de fragments que celle présentée ici. La prochaine partie de ce manuscrit résume donc le fonctionnement initial de F2D, pour plus de détails j'invite le lecteur à lire la thèse du Dr J.-M. Gally : « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments ».

4.1.2.2 Création de la librairie de fragments

Dans cette partie, notre finalité est d'acquérir une base de données de fragments 3D récupérés à partir de toutes les structures cristallographiques obtenues expérimentalement et disponibles. Les structures sont téléchargées depuis la PDB[184] à l'aide de deux requêtes PFAM[185] : PF07714 (pour « protein tyrosine kinase ») et PF00069 (pour « protein kinase domain »). Grâce à ces deux requêtes, 3809 structures de protéines kinases ont été collectées, toutes espèces confondues (*juillet 2016*). Chaque structure est ensuite séparée par chaine et seul le domaine kinase est conservé. Afin que les kinases soient toutes représentées dans un même

[184] Helen M. Berman et al., « The Protein Data Bank », *Nucleic Acids Research* 28, nº 1 (1 janvier 2000): 235-42, https://doi.org/10.1093/nar/28.1.235.
[185] Robert D. Finn et al., « The Pfam Protein Families Database: Towards a More Sustainable Future », *Nucleic Acids Research* 44, nº D1 (4 janvier 2016): D279-85, https://doi.org/10.1093/nar/gkv1344.

référentiel commun, chaque domaine est aligné sur la chaine E de la structure PDB 1ATP, la première protéine kinase cristallisée avec l'ATP[186]. Cette superposition se fait à l'aide du logiciel MOE (Chemical Computing Group, *version 2016.0802*) qui réalise un alignement optimal grâce aux motifs communs des kinases présentés dans la partie 1.5.2 du chapitre 1 de ce manuscrit. La Figure 34 ci-dessous illustre différentes kinases superposées avec, au sein de leurs sites actifs, leurs ligands respectifs.

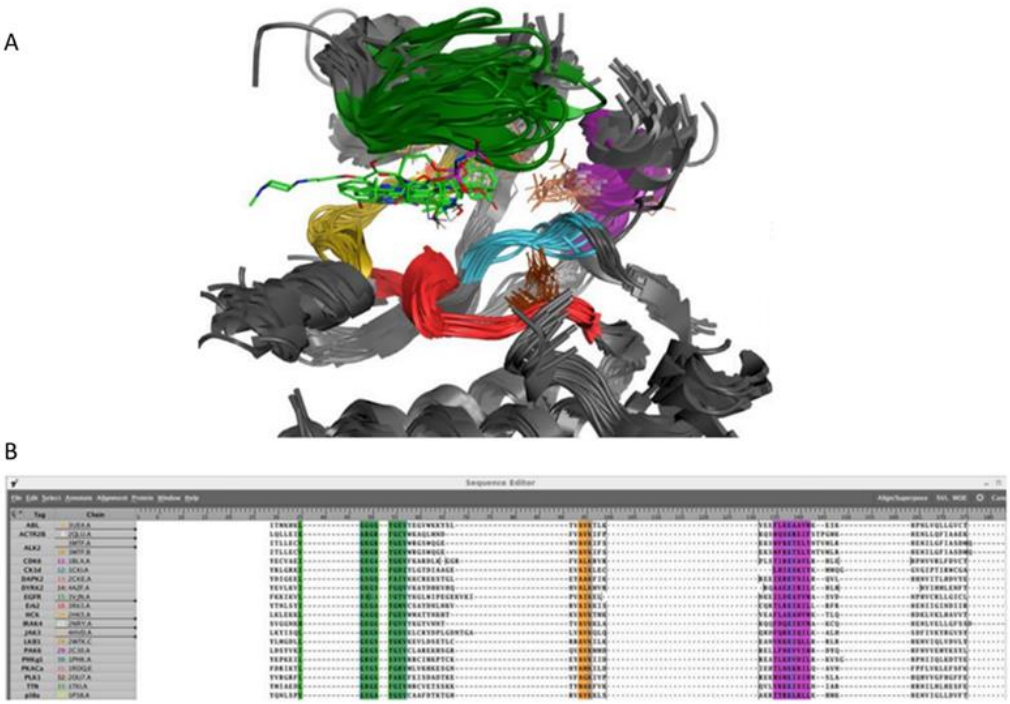


Figure 34 : Exemples de protéines kinases alignées avec leurs ligands (A) et leurs séquences alignées associées (B). *D'après J.-M. Gally[187]*.

Après la superposition s'ensuit une étape de fragmentation des ligands co-cristallisés. Avant de les fragmenter, ceux-ci doivent être isolés des autres molécules que l'on peut retrouver dans une structure cristallographique (co-facteur(s), solvant(s), contre-ion(s) ou agents de co-cristallisation comme le glycérol…). Ces molécules autres que le ligand pouvant potentiellement se retrouver aussi dans le site actif, on ne peut faire une sélection en se basant uniquement sur la cavité pour extraire les ligands seuls. Les composés indésirables sont donc retirés à l'aide d'une liste les regroupant[188]. Pour chaque structure, nous vérifions que nous récupérons bien un unique ligand en ouvrant le fichier et le cas échéant, après vérification manuelle, la liste des molécules indésirables est mise à jour si nécessaire. Les ligands sont ensuite tous sauvegardés dans un seul fichier au format standard SDF, sans conserver les coordonnées de leur cible protéique.

[186] J. Zheng et al., « 2.2 Å Refined Crystal Structure of the Catalytic Subunit of CAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor », *Acta Crystallographica Section D: Biological Crystallography* 49, nº 3 (1 mai 1993): 362-65, https://doi.org/10.1107/S0907444993000423.
[187] José-Manuel Gally, « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » (Orléans, 2017).
[188] Helena Strömbergsson et Gerard J Kleywegt, « A chemogenomics view on protein-ligand spaces », *BMC Bioinformatics* 10, nº Suppl 6 (16 juin 2009): S13, https://doi.org/10.1186/1471-2105-10-S6-S13.

Enfin, une fois les ligands alignés et isolés, pour obtenir des fragments, il faut de toute évidence les fragmenter. Cette étape est réalisée à l'aide de 8 différents algorithmes déjà implémentés dans des nœuds et assemblés dans un « workflow » KNIME[189]. Ainsi, à partir des 8 648 ligands de départ, on obtient 131 826 fragments, soit en moyenne 15 fragments par ligand.

4.1.2.3 Création des molécules

Une fois la bibliothèque de fragments 3D constituée, le programme F2D est divisé en plusieurs modules indépendants (Figure 35) dirigés à l'aide d'un fichier de configuration permettant de sélectionner les paramètres avant le lancement. Une utilisation typique de F2D commence par la lecture de la librairie de fragments afin de les charger en mémoire. Si besoin, ces fragments sont ensuite standardisés et regroupés entre eux par similarité pour identifier et retirer les doublons. Puis, à partir du fragment initial précisé en entrée, F2D recherche les possibilités de liaison avec ses voisins afin de trouver toutes les combinaisons possibles. Pour ne construire que des molécules spécifiques à la cible désirée, les fragments présentant un encombrement stérique avec celle-ci (ceux ne rentrant pas dans son site actif) sont retirés avant la recherche de combinaisons. Finalement, après la création de molécules en reliant les fragments entre eux, les résultats sont sauvegardés dans un fichier au format SDF pour une analyse ou un traitement ultérieur.

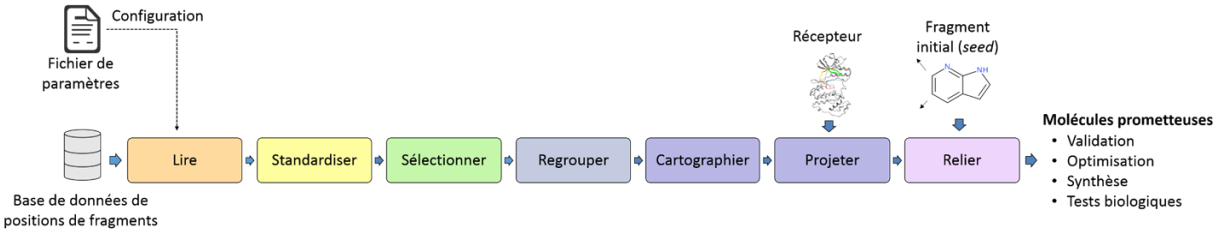


Figure 35 : Protocole classique d'utilisation de Frags2Drugs. *D'après J.-M. Gally*[190].

Je vais maintenant détailler les différents éléments du protocole suivi dans la prochaine partie du texte.

4.1.3 Description des différents modules de F2D

4.1.3.1 Lecture des paramètres

C'est le module initiateur du programme, il sert à lire à la fois le fichier de configuration pour paramétrer les options du programme avant son lancement, mais aussi à charger la base de données de fragments préparée en amont. Plusieurs formats d'entrées de fragments peuvent être pris en compte comme le format classique SDF ou le format binaire HDF5. Les fragments sont stockés sous la forme d'une table grâce à la librairie pandas (https://pandas.pydata.org/),

[189] Michael R. Berthold et al., « KNIME: The Konstanz Information Miner », in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)* (Springer, 2007).
[190] José-Manuel Gally, « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » (Orléans, 2017).

en conservant tous leurs attributs (descripteurs moléculaires et identifiants) pour permettre une application de filtre ultérieure aisée (Figure 36).

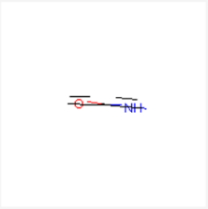
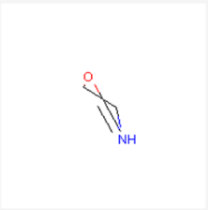
	ID	Molecule	SMILES	full_fragment_id	MW	LogP	TPSA	LabuteASA	HBA	NHA	HBD	NAR	NRB
135	4MEDA-ALI-L1F012		C1=COC=CN1	PIM1_HUMAN_4MED-A_alignment_L1F012	83.0	0.5	21.3	36.1	2	6	1	0	0
136	3U9NA-ALI-L1F009		C1=COCC1	CHK1_HUMAN_3U9N-A_alignment_L1F009	70.0	0.9	9.2	31.3	1	5	0	0	0
137	2RL5A-ALI-L1F008		C1=COCCN1	VGFR2_HUMAN_2RL5-A_alignment_L1F008	85.1	0.1	21.3	36.8	2	6	1	0	0
138	4KBAB-ALI-L1F008		C1=COCCNC1	KC1D_HUMAN_4KBA-B_alignment_L1F008	99.1	0.1	21.3	43.2	2	7	1	0	0

Figure 36 : Extrait de la table contenant tous les fragments utilisés par F2D.

4.1.3.2 Standardisation des fragments

Ce module est optionnel, il sert à s'assurer que les fragments soient tous préparés de manière similaire, notamment si ceux-ci proviennent d'un regroupement de différentes bases de données. Dans notre cas, tous les fragments proviennent de la PDB mais les utilisateurs peuvent rajouter leurs propres fragments issus de leurs données personnelles. Ce module inclut donc entre autres la conversion en tautomères canoniques, la représentation sous forme dite kekulé ou encore la neutralisation de la charge totale des fragments… Cette étape est particulièrement utile pour supprimer avec certitude les doublons (à l'étape Regroupement des doublons) de la bibliothèque de fragments. En effet, si deux fragments similaires présentent une charge différente par exemple, ils ne seront pas détectés comme doublons, les neutraliser résout ce problème.

4.1.3.3 Sélection sur critères

Ce module est lui aussi optionnel. Il s'agit d'une étape de sous-sélection de fragments selon des descripteurs moléculaires définis. Grâce à la bibliothèque pandas et à la gestion des tables contenant les fragments, on peut rapidement les filtrer selon plusieurs critères différents. Ces critères peuvent aussi bien être des comparaisons numériques (> ou < à un poids moléculaire donné ou une valeur de LogP…) ou des expressions régulières (texte pour retrouver des fragments précis à l'aide de leurs identifiants par exemple). Les annotations connues comme

la règle de 5[191], la règle de 3[192] ou le « lead-like »[193] sont directement implémentés et peuvent être utilisés facilement comme filtre si nécessaire. Enfin, si un utilisateur veut employer ses propres filtres ou d'autres descripteurs, il peut se servir des fonctionnalités de pandas pour les implémenter lui-même.

4.1.3.4 Regroupement des doublons

Ce module sert à éliminer les fragments doublons. Selon la façon dont la base de fragments a été préparée et du fait de la similarité de nombreux inhibiteurs de kinases, il se peut que certains fragments soient présents plusieurs fois. A noter que dans notre cas, comme nous travaillons avec des données en 3D, on appelle doublons, des fragments qui présentent à la fois une même structure chimique mais aussi des positions atomiques similaires. Pour les trouver et les supprimer, un premier regroupement par formule SMILES canonique est réalisé. Ensuite, les positions 3D de chaque membre du groupe sont comparées par calcul de RMSD entre chaque fragment. Si deux fragments présentent un RMSD inférieur au seuil arbitraire de 0,15 Å, seul le premier sera alors conservé. Un cas de figure est explicité en Figure 37. Après cette étape le nombre de fragments est passé de 131 826 à 81 667, soit 38 % de doublons supprimés.

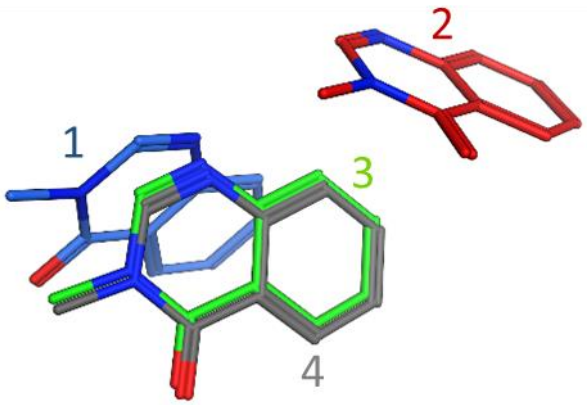


Figure 37 : Exemple d'élimination des fragment doublons. *Les fragments 1 (bleu), 2 (rouge), 3 (vert) et 4 (gris) sont des fragments ayant la même structure chimique et donc présentant la même formule SMILES. Après l'étape de regroupement, seul les fragments 1, 2, 4 seront conservés car les fragments 3 et 4 présentent une pose similaire (RMSD < 0,15 Å).*

4.1.3.5 Cartographie des fragments

Ce module initie la recherche de combinaisons possibles entre fragments. La règle pour que deux fragments puissent être reliés est que chacun présente au moins un atome dont la valence autorise une nouvelle liaison (à la place d'un atome d'hydrogène) et que ces atomes se situent à une position favorable l'un par rapport à l'autre, en vérifiant la longueur de la future liaison ainsi que les angles de liaison et dièdres avec les atomes voisins (Figure 38). Pour cela, dans un premier temps, pour chaque fragment nous annotons les atomes susceptibles de pouvoir

[191] Christopher A Lipinski et al., « Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings », Advanced Drug Delivery Reviews, Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998, 46, no 1 (1 mars 2001): 3-26, https://doi.org/10.1016/S0169-409X(00)00129-0.
[192] Harren Jhoti et al., « The "rule of Three" for Fragment-Based Drug Discovery: Where Are We Now? », *Nature Reviews Drug Discovery* 12, nº 8 (août 2013): 644, https://doi.org/10.1038/nrd3926-c1.
[193] Simon J. Teague et al., « The Design of Leadlike Combinatorial Libraries », *Angewandte Chemie International Edition* 38, nº 24 (1999): 3743-48, https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U.

former une nouvelle liaison (Figure 39, A). Ensuite, les fragments voisins sont déterminés à l'aide de leurs coordonnées spatiales. Le référentiel commun dans lequel se trouve tous les fragments suite à la superposition des protéines kinases est découpé en « voxels » (terme englobant le mot volume et élément, il s'agit d'un pixel en 3D servant à stocker des coordonnées spatiales). On peut assimiler un voxel à un cube d'un volume défini englobant une partie du référentiel spatial. Chaque fragment est donc défini par une liste de voxels résumant ses coordonnées spatiales. Deux fragments présentant un voxel commun sont considérés comme voisins et nous vérifions ensuite s'ils ont des atomes susceptibles de former une liaison réunissant les conditions énoncées plus haut (Figure 39, B). Si ces deux fragments ont des atomes trop proches ou superposés (encombrement stérique), ils seront automatiquement annotés comme incompatibles et inaptes à former une liaison. Si deux fragments sont incompatibles, ils ne pourront pas non plus se retrouver dans la même molécule finale par le biais d'une liaison avec un fragment tierce. Au final, chaque fragment possèdera donc une liste de fragments avec lesquels il peut se lier et une liste de fragments avec lesquels il ne peut être assemblé. Afin de simplifier la sélection des résultats et la future synthèse des molécules construites par F2D, des précautions ont été mises en places. Ainsi, nous empêchons les liaisons entre deux atomes d'azote, deux atomes d'oxygène ou un atome d'azote et un atome d'oxygène (liaisons N – N, N – O et O – O). Pour chaque liaison acceptée, un niveau de qualité est attribué en fonction des valeurs de la distance entre les deux atomes et des angles avec les atomes voisins comparés aux valeurs théoriques attendues fournies par le champ de force MMFF94[194]. Ce niveau de qualité reflète simplement le pourcentage de différence entre la valeur attendue et la valeur théorique (distorsion de la molécule).

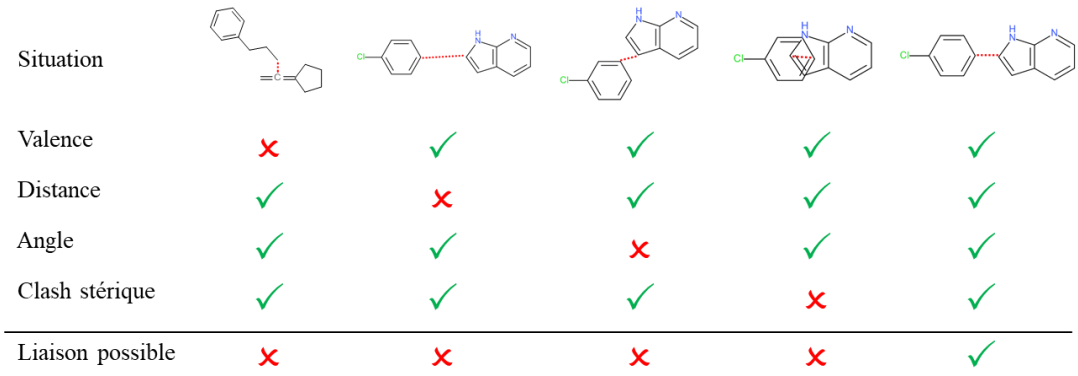


Figure 38 : Conditions à réunir pour relier deux fragments entre eux. *D'après J.-M. Gally*[195].

[194] Thomas A. Halgren, « Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94 », *Journal of Computational Chemistry* 17, nº 5-6 (1 avril 1996): 490-519, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.
[195] José-Manuel Gally, « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » (Orléans, 2017).

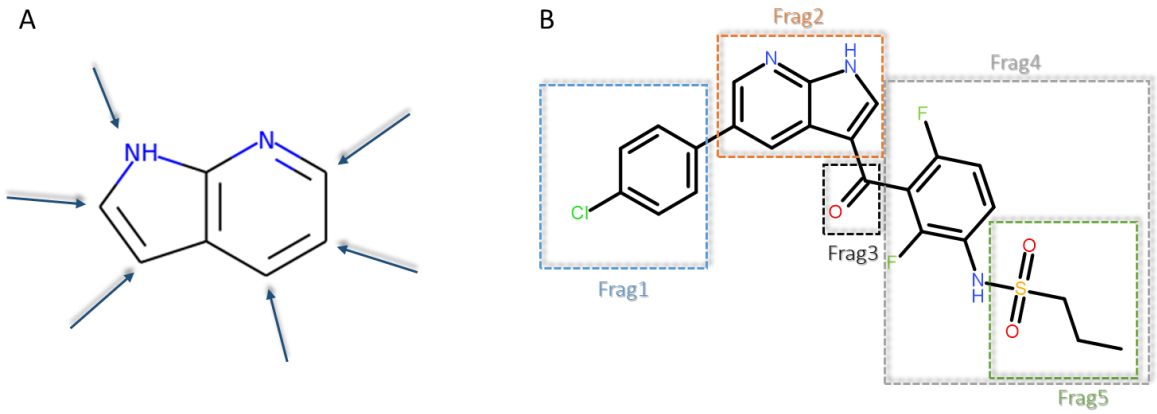


Figure 39 : Atomes aptes à créer une liaison sur un fragment (A) et exemples de fragments pouvant ou non être reliés (B). *Dans l'exemple (B), le fragment 1 et le fragment 2 réunissent toutes les conditions pour être liés tandis que le fragment 4 et le fragment 5 ne pourront être combinés car ils présentent un encombrement stérique.*

4.1.3.6 Projection dans le site actif

A partir de cette étape la protéine cible est prise en compte, les fragments entrant en collision avec son site actif sont écartés, évitant ainsi de générer des molécules non compatibles. Ce module permet aussi de prioriser la croissance des molécules vers des régions présentant le plus d'espace libre dans la cavité du site actif. En fonction du fragment initial et du ou des atomes choisis pour le départ de F2D, on pourra privilégier de grandir celui-ci en direction de la boucle d'activation ou vers les poches allostériques. Enfin, dans ce module il est aussi possible de considérer d'autres cavités provenant d'autre kinases (pour créer un inhibiteur double par exemple), ou bien des cofacteurs ou des molécules d'eau jugées importantes pour l'activité d'une molécule. Un fragment est considéré comme compatible avec une cavité si aucun de ses atomes ne se trouve à une distance inférieure à 1,3 Å des atomes des acides aminés formant le site actif de la cible. De plus, les atomes de ce fragment ne doivent pas non plus se situer à une distance supérieure à 3,5 Å des atomes du site actif. Cette précaution empêche la construction de molécules non compatibles avec la cavité mais aussi trop éloignées de cette même cavité. En éliminant les fragments trop distants du site actif, on empêche la création de molécules en direction du solvant.

4.1.3.7 Recherche de combinaisons et construction des molécules

Ce module est le cœur même du projet. Grâce à toutes les informations préalablement recueillies, il va lancer la création des molécules par combinaison des fragments entre eux et en les assemblant à partir du fragment d'origine. L'algorithme, pour construire toutes les molécules possibles, repose sur un parcours de graphe en largeur (cf. partie 4.2.2). Les molécules sont donc créées niveau par niveau : d'abord on ajoute au fragment initial tous ses fragments voisins au premier degré, puis on ajoute les voisins des voisins, etc. Cette approche nécessite de garder en mémoire toutes les molécules à chaque couche. Les résultats sont ensuite retournés à l'utilisateur sous forme d'un fichier SDF contenant toutes les molécules construites.

4.2 Reprise du projet

A mon arrivée dans l'équipe, F2D n'était encore ni pleinement fonctionnel, ni terminé. Mon objectif étant de le finir et de l'optimiser, j'ai commencé par me familiariser avec le code

source pour comprendre, puis maitriser les différentes bibliothèques utilisées. Une fois le code approprié, j'ai alors retouché les modules « Cartographie des fragments » et « Projection dans le site actif » qui présentaient des bogues, et j'ai terminé le module « Recherche de combinaisons et construction des molécules ». Une fois tous les modules terminés, j'ai mis en place de nombreux tests unitaires afin de m'assurer que tout fonctionnait correctement et permettre un meilleur suivi du code, notamment lors de retouches ponctuelles. Ainsi, à chaque retouche il suffit de lancer les tests pour s'apercevoir d'une possible erreur créée. J'ai aussi mis en place un système d'enregistrement (appelé « log ») qui permet de suivre le programme et ses avancées. Chaque action est ainsi répertoriée au sein d'un journal (fichier texte) pour la suivre précisément et comprendre exactement quand et où le programme rencontre des difficultés lorsqu'il ne retourne pas de résultats. Les logs permettent d'analyser pas à pas l'activité d'un outil. Pour finir, j'ai aussi transformé tout ce code en un package Anaconda facilement transposable et installable sur un autre ordinateur.

4.2.1 Premières améliorations

Une fois F2D achevé, j'ai commencé à optimiser le code afin de le rendre plus performants. Parmi ces premières transformations, on peut citer principalement :

- La mise à jour des différentes bibliothèques utilisées et le passage à la dernière version de Python (version *3.6.2, début 2017*).

- L'arrêt du système de voxels pour déterminer les fragments voisins, fastidieux et coûteux en temps de calcul, pour passer directement par les vraies coordonnées atomiques (x, y, z). En effet, la bibliothèque Numpy est suffisamment puissante pour retourner très rapidement une distance minimale entre deux fragments grâce à l'usage des matrices de coordonnées. Cette distance minimale entre deux fragments nous renseigne tout de suite sur leur statut de voisins ou non et ensuite continuer les investigations entre eux.

- La maximisation des performances du langage Python. Cela inclut entre autres l'usage de types de données vectorisées, l'utilisation d'objets de type « set » au lieu de « list » pour comparer des données, la transformation des fonctions pour l'emploi de la bibliothèque Numba (https://numba.pydata.org/) qui optimise la compilation du code et accélère les fonctions, etc.

- La parallélisation des fonctions pour exploiter tous les processeurs de l'ordinateur et accélérer les processus de calculs.

Malgré ces premières améliorations et l'optimisation du code, les premiers essais en condition réelles (avec la librairie complète de fragments) se sont révélés plutôt infructueux et très rapidement deux problèmes sont apparus. D'abord un problème de temps de calcul très long dû à la croissance exponentielle (« explosion combinatoire ») des opérations à effectuer au fur et à mesure de l'avancement du programme (Figure 40). Ensuite, une consommation de mémoire vive (« RAM ») trop importante pouvant amener l'ordinateur à se bloquer, forçant l'utilisateur à le redémarrer. Cette saturation de la RAM provient de l'architecture même du programme et de son parcours en largeur (niveau par niveau) pour créer les molécules, ce qui

oblige l'ordinateur à stocker en mémoire tous les composés en cours de fabrication, celle-ci ne pouvant être libérée qu'à la toute fin de l'exécution du programme.

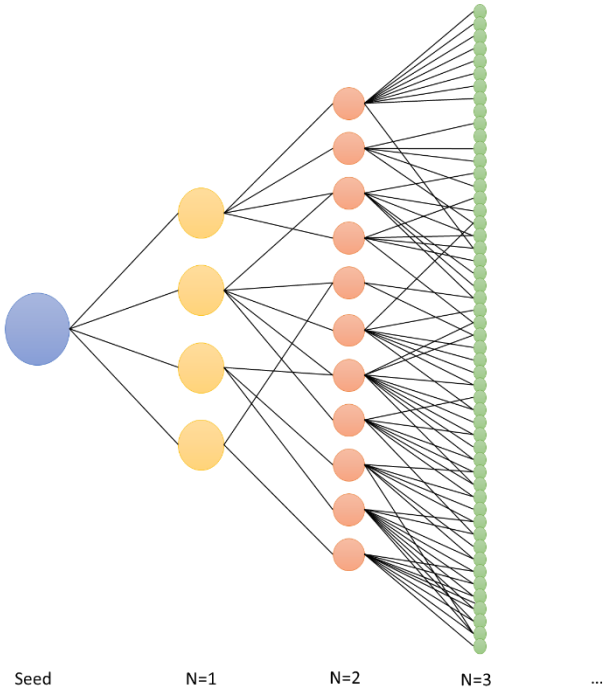


Figure 40 : Représentation abstraite de la complexité au fur et à mesure de l'ajout de fragments sur le fragment initial.

Finalement, suite à ces premiers essais non-concluants et conscient que la méthode développée à ce stade n'est pas la plus adaptée, une refonte partielle du programme a été décidée. De par la nature des données et les relations entre elles (comparables à un réseau), nous avons implémenté une approche basée sur les graphes et non plus sur les tables. Je vais détailler le développement de cette nouvelle méthode dans la prochaine partie après une courte introduction sur les graphes.

4.2.2 Introduction aux graphes

Un graphe G fait référence à un système composé d'objets dans lequel certains d'entre eux (ou tous) peuvent être en relation. Il est défini par deux ensembles V et E tel que G = (V, E) avec V son ensemble de sommets (ou nœuds) et E son ensemble d'arêtes (ou relations)[196]. La Figure 41 indique les multiples représentations possibles d'un même graphe. Les nœuds représentent les objets constituant le graphe et une arête équivaut à une paire de sommets reliés entre eux, elle est donc associée à deux nœuds distincts. Les graphes sont des outils puissants permettant de modéliser de nombreux problèmes : un réseau de transport (un plan de métro, où les sommets sont les stations et les rails les arêtes), un réseau social (chaque sommet représente une personne et une arête les relie si elles se connaissent), des voies de signalisation en biochimie…. De manière plus générale, un graphe permet de représenter les connexions d'un ensemble complexe en exprimant les différentes relations entre les éléments le constituant. Quand on utilise des graphes, on fait référence à la théorie des graphes qui est devenue une

[196] Richard J. Trudeau, *Introduction to Graph Theory* (Courier Corporation, 2013).

branche des mathématiques et est désormais employée dans de multiples disciplines comme la chimie, la biologie, les sciences sociales et l'informatique.

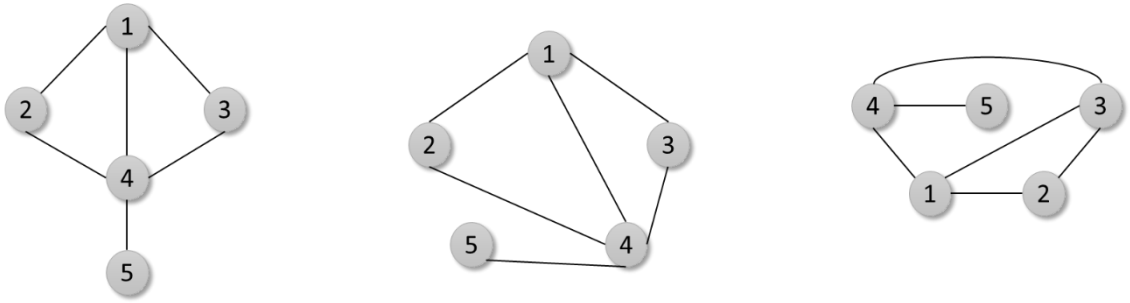


Figure 41 : Différentes représentations d'un même graphe. *Dans ce cas le graphe comporte 5 sommets et 6 arêtes, G = (V, E) avec V = {1, 2, 3, 4, 5} et E = {{1, 2}, {{1, 3}, {{1, 4}, {{2, 4}, {{3, 4}, {{4, 5}}.*

4.2.2.1 Historique

L'origine des graphes est associée au problème des ponts de Königsberg (qui aujourd'hui se nomme Kaliningrad) introduit en 1735 par L. Euler, considéré comme le fondateur de la théorie des graphes[197]. Ce problème est le suivant : Königsberg possède sept ponts qui enjambent la rivière Pregel et l'on s'interroge sur l'existence d'un chemin permettant de passer par tous les ponts de la ville tout en traversant chacun d'entre eux uniquement une fois, puis de revenir à son point de départ*. L. Euler a alors modélisé ce problème par un graphe : chaque parcelle de terre délimitée par la rivière est associée à un sommet et chaque pont définit une arête (Figure 42).

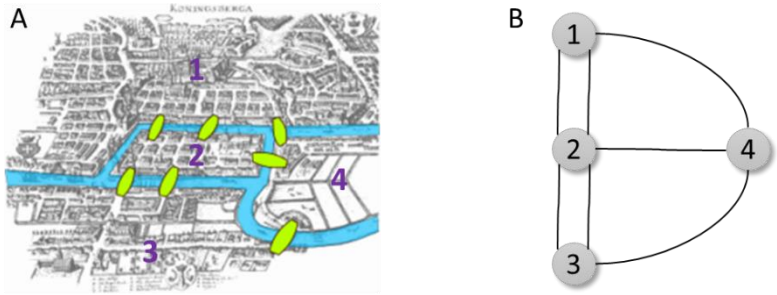


Figure 42 : Représentation de la ville de Königsberg et ses sept ponts (A) et modélisation par un graphe (B). *Pour une meilleure visibilité les ponts sont colorés en vert et les différentes zones de la ville numérotées.*

* Concernant la réponse au problème, je vous laisse le soin de chercher par vous-même.

4.2.2.2 Les différentes catégories

Il existe plusieurs types de graphes et plusieurs méthodes pour les classifier[198]. Ici, je ne rentrerais pas dans les détails et me focaliserais sur le principal. Un graphe peut être non-orienté : ses arêtes sont parcourues dans un sens ou dans l'autre sans incidence, elles sont donc bidirectionnelles. Dans le cas contraire on parle de graphe orienté, les deux nœuds sont

[197] Gerald L. Alexanderson, « About the Cover: Euler and Königsberg's Bridges: A Historical View », *Bulletin of the American Mathematical Society* 43, nº 04 (18 juillet 2006): 567-74, https://doi.org/10.1090/S0273-0979-06-01130-X.
[198] Claude Flament, *Théorie des graphes et structures sociales* (Walter de Gruyter GmbH & Co KG, 2017).

connectés de manière spécifique et les arêtes sont unidirectionnelles. Le sommet de départ est alors l'origine et celui d'arrivée la destination. A titre d'exemple le réseau social Facebook est ainsi un graphe non-orienté (une relation « ami » se fait systématiquement réciproquement), tandis que celui de Twitter est pour sa part un graphe orienté (une personne peut « suivre » une autre sans que cela ne soit réciproque).

4.2.2.3 Le parcours d'un graphe

Le parcours d'un graphe consiste à explorer les différents sommets de proche en proche à partir d'un nœud initial. Il existe de nombreux algorithmes spécifiques à cette problématique. Les plus connus sont le parcours en profondeur et le parcours en largeur. On peut aussi citer l'algorithme de Dijkstra, aussi connu sous le nom d'algorithme du plus court chemin[199]. Le but d'un tel algorithme est de sélectionner à partir des sommets déjà visités le prochain sommet à aller voir, il doit déterminer un ordre de visite tout en évitant de revenir sur ses pas.

4.2.2.3.1 Le parcours en profondeur

L'algorithme de parcours en profondeur (appelé DFS, « Depth-First Search ») va explorer un graphe en empruntant les chemins un par un jusqu'au bout : pour chaque nœud il visite son premier sommet voisin et continue jusqu'au nœud terminal de la branche, puis il repart alors du dernier croisement pour aller sur un autre chemin[200]. On peut assimiler le DFS à la méthode intuitive que l'on utiliserait pour trouver la sortie d'un labyrinthe sans tourner en rond. En effet, dans ce cas tant que l'on ne rencontre pas d'impasse on continue sur le même chemin, tandis que si l'on rencontre une impasse, on revient sur ses pas jusqu'à trouver une autre voie et la poursuivre, là encore, jusqu'au bout.

4.2.2.3.2 Le parcours en largeur

L'algorithme de parcours en largeur (appelé BFS, « Breadth-First Search ») explore un graphe niveau par niveau : l'exploration commence par le nœud initial, puis tous les nœuds à une arête de ce nœud initial (les voisins du premier degré), puis les voisins non explorés des voisins (second degré) et ainsi de suite[201]. Le BFS diffère du DFS par sa façon de parcourir le graphe, non pas en explorant un chemin jusqu'au bout mais en listant d'abord tous les voisins d'un nœud pour ensuite les explorer un par un avant de passer aux voisins suivants.

Ces deux méthodes principales pour explorer un graphe présentent donc des différences notoires (Figure 43). L'une n'est pas forcément meilleure que l'autre, tout dépend du contexte et du but recherché. L'algorithme DFS se concentre sur une seule voie et la poursuit jusqu'à sa fin tandis que l'approche BFS évalue tous les chemins possibles à partir d'un nœud et les explore tous simultanément couche par couche. On utilisera donc le BFS dans le cas de la recherche du plus court chemin entre deux nœuds. En revanche, on utilisera le DFS pour rechercher l'existence même d'un chemin entre deux nœuds.

[199] E. W. Dijkstra, « A Note on Two Problems in Connexion with Graphs », *Numerische Mathematik* 1, nº 1 (1 décembre 1959): 269-71, https://doi.org/10.1007/BF01386390.
[200] Edouard Lucas, *Récréations mathématiques (2ème éd.)*, 1891, https://gallica.bnf.fr/ark:/12148/bpt6k3943s.
[201] « Parcours d'un graphe », consulté le 18 septembre 2019, https://haltode.fr/algo/structure/graphe/parcours.html.

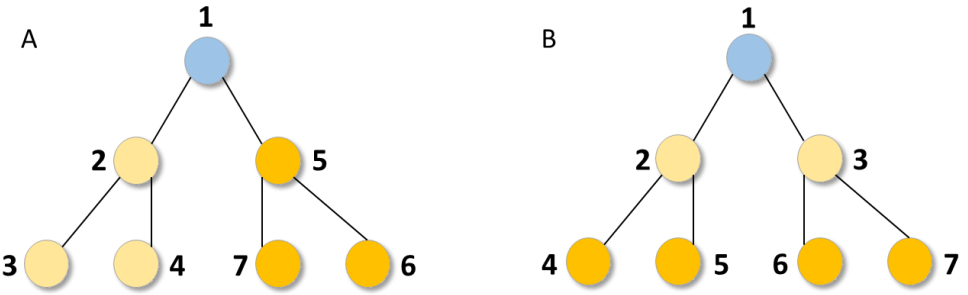


Figure 43 : Exploration d'un graphe selon les algorithmes DFS (A) et BFS (B). *Les chiffres correspondent à l'ordre des nœuds visités lors du parcours du graphe.*

Le point vraiment important à prendre en compte est la différence d'implémentation informatique. Le parcours en largeur fonctionne avec un système de queue pour stocker les informations sur les nœuds voisins à visiter, il nécessite donc de la mémoire pour fonctionner, proportionnellement à la taille du graphe. Le parcours en profondeur repose sur un système d'empilement et n'a pas besoin de garder en mémoire le graphe entier, il est fondé sur la récursivité (boucle auto-génératrice)[202]. Cette différence a une incidence de taille sur le programme. Dans notre cas, comme je l'ai énoncé dans la description de la première version de F2D, celle-ci reposait sur une exploration en largeur et cela a causé nos problèmes de mémoire saturée. En effet, avec le grand nombre de combinaisons possibles, au fur et à mesure de l'avancement dans les couches du graphe, celle-ci ne pouvait plus suivre.

4.2.2.4 Les bases de données orientées graphes

Une base de données (BDD) est un ensemble de données organisées dans le but de faciliter une recherche d'information rapide. Historiquement, pour stocker les données, le modèle relationnel a dominé depuis les années 80, avec des systèmes tels qu'Oracle ou MySQL. Néanmoins, ces dernières années, suite à des problèmes rencontrés avec ce type de BDD (de performances sur les gros volumes de données notamment), la tendance du « NOSQL » a peu à peu émergé. Le NOSQL (pour « Not Only SQL ») reflète une vaste catégorie de systèmes ne suivant pas la modélisation relationnelle, et donc n'utilisant pas le langage SQL comme langage de requête[203]. Parmi ces systèmes NOSQL, on retrouve les bases de données orientées graphe. Comme son nom l'indique, une BDD orientée graphe est une BDD qui va utiliser la théorie des graphes, pour représenter et stocker les données comme illustré Figure 44. L'implémentation repose sur trois blocs de bases : le nœud, la relation (avec une orientation et un type) et les attributs (sur un nœud et/ou une relation)[204].

[202] Axel Chambily et Pétrut Constantine, « Cours sur la récursivité », Developpez.com, consulté le 24 septembre 2019, http://recursivite.developpez.com/.
[203] LearnAnalytics, « What Is NoSQL and Is It the next Big Trend in Databases? », *Big Data Path* (blog), 27 mars 2018, https://bigdatapath.wordpress.com/2018/03/27/what-is-nosql-and-is-it-the-next-big-trend-in-databases/.
[204] « Les Bases Orientées Graphes, NoSQL et Neo4j », InfoQ, consulté le 19 septembre 2019, https://www.infoq.com/fr/articles/graph-nosql-neo4j/.

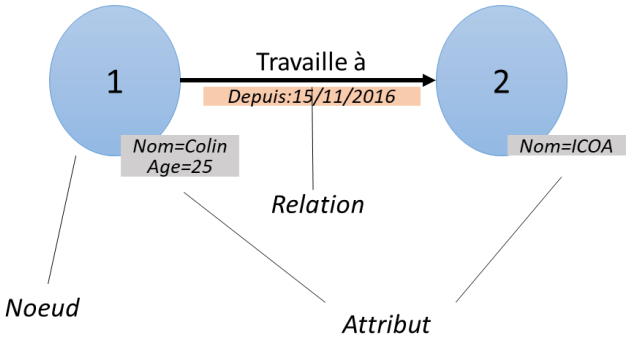
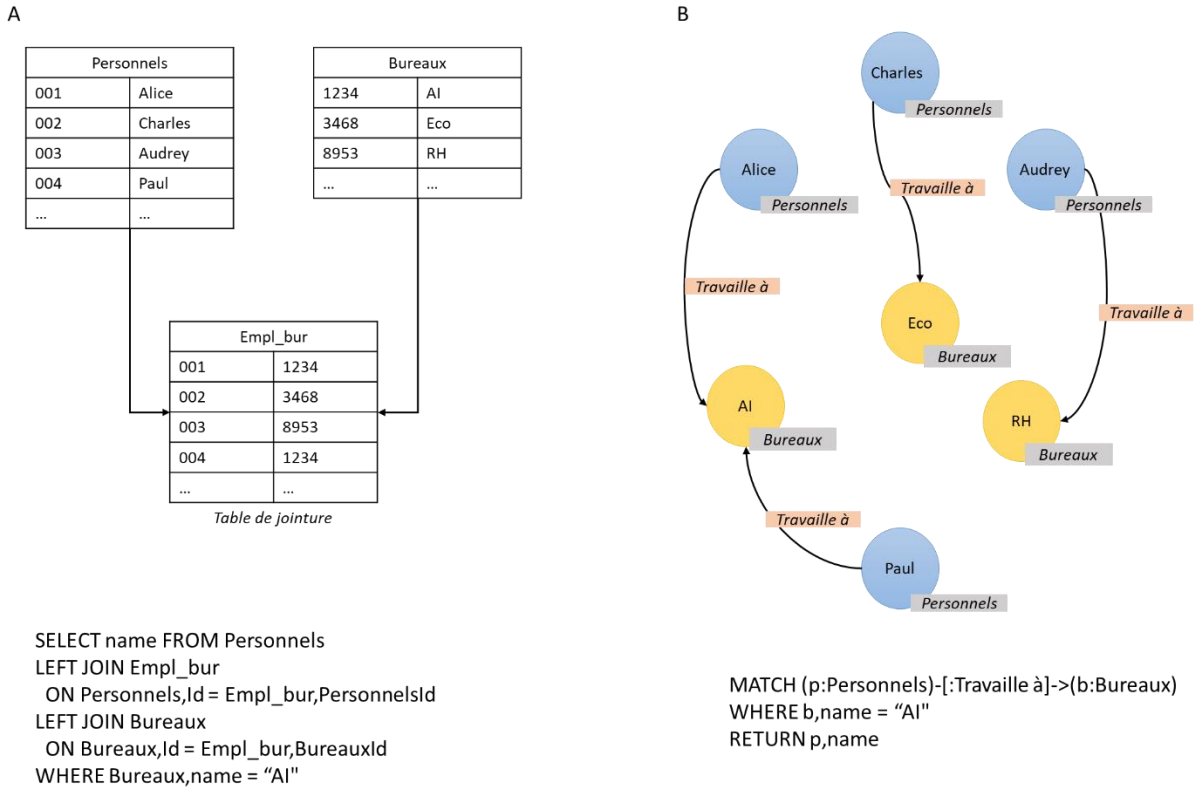


Figure 44 : Représentation graphique d'une base de données orientée graphe.

Une telle structure est beaucoup plus appropriée et optimisée lorsque que l'on souhaite exploiter et explorer les relations existantes entre nos données. Comme le montre la Figure 45, avec une BDD orientée graphe, les recherches se font beaucoup plus facilement et intuitivement qu'avec une base de données relationnelle traditionnelle.




```
SELECT name FROM Personnels
LEFT JOIN Empl_bur
  ON Personnels,Id = Empl_bur,PersonnelsId
LEFT JOIN Bureaux
  ON Bureaux,Id = Empl_bur,BureauxId
WHERE Bureaux,name = "AI"
```

```
MATCH (p:Personnels)-[:Travaille à]->(b:Bureaux)
WHERE b,name = "AI"
RETURN p,name
```

Figure 45 : Implémentation d'une BDD relationnelle (A) et d'une BDD orientée graphe (B). *Comme le montre le code nécessaire pour retrouver les employés travaillant au bureau AI, les requêtes sur une BDD orientée graphe sont plus simples à réaliser et plus intuitives, ne nécessitant pas de jointure entre les tableaux de données.*

Les BDD orientées graphe sont utilisées par exemple dans la modélisation des réseaux sociaux et plus généralement pour tout ce qui touche aux recommandations (achats internet, « amis », page internet…). Mais elles le sont aussi désormais dans de nombreux domaines scientifiques comme la biologie pour explorer les réseaux métaboliques ou moléculaires[205].

[205] Wei Gao et al., « Study of biological networks using graph theory », *Saudi Journal of Biological Sciences* 25, nº 6 (1 septembre 2018): 1212-19, https://doi.org/10.1016/j.sjbs.2017.11.022.

Ainsi, l'usage de telles BDD est fortement conseillé pour traiter les données connectées, cela permet d'éviter de passer par de multiples jointures très coûteuses en temps de calcul et longues à écrire. Les relations entre les nœuds sont directs, on parle dans ce cas de « contiguïté des données sans index » (ou « index-free adjacency »)[206]. Grâce aux réponses rapides, elles peuvent être utilisées directement dans des applications web par exemple. Un autre avantage des BDD orientées graphes est la simplicité des requêtes, facilitant leur emploi et leur développement. Enfin, comme on le voit Figure 45, les données sont représentées naturellement et la visualisation et la lecture sont beaucoup plus fluides.

4.2.3 Développement d'une nouvelle architecture

4.2.3.1 Nouvelle forme de stockage des fragments

Pour pallier les problèmes de performance et de mémoire de F2D, nous avons donc décidé de nous tourner vers une nouvelle implémentation à l'aide d'une base de données orientée graphe. En effet, nos données et notre problématique sont parfaitement adaptées à l'usage de cette technologie car nous cherchons à étudier les relations entre les fragments. Pour cela, nous avons opté pour Neo4J (https://neo4j.com/, *version community 3.3.0)*, un logiciel pour créer et gérer sa BDD orientée graphe. L'avantage de Neo4J est sa communauté d'utilisateurs très développée et réactive et sa version gratuite. Neo4J dispose de son propre langage de requête : Cypher.

La manière de stocker les données a donc été radicalement modifiée, désormais les fragments sont stockés directement sous forme d'un graphe après l'étape de calcul des relations entre eux. Pour rappel, dans la première version de F2D, ces relations étaient stockées sous forme de listes intégrées dans une nouvelle colonne du tableau regroupant tous les fragments que l'on voit Figure 36. Or, c'est particulièrement la recherche et les jointures entre ces listes qui posaient problème. Comme le montre la Figure 46, cette nouvelle forme de stockage permet de repérer et de connaitre directement les fragments capables de se lier (relations I pour inclusion) et les fragments incompatibles (relations E pour exclusion).

[206] « Bases de données graphes - Les modèles de données », Versusmind, consulté le 27 septembre 2019, https://versusmind.eu/blog/bases-de-donnees-graphes-les-modeles-de-donnees.

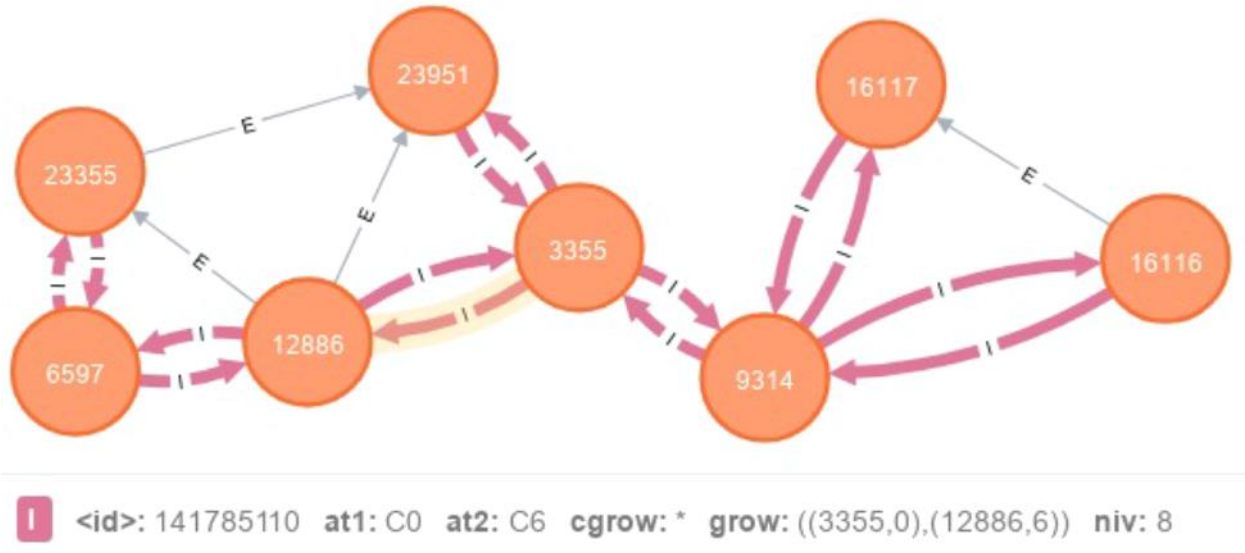


Figure 46 : Extrait de la BDD orientée graphe contenant les fragments de la librairie de F2D. *Les propriétés comme la masse moléculaire des fragments ou les atomes impliqués dans la liaison sont aussi stockés dans la BDD sous forme d'attributs. Ici, sur la relation I surlignée, on voit en attribut les deux types d'atomes impliqués dans la liaison, leur numérotation dans le fragment (carbone 0 du fragment 3355 et carbone 6 du fragment 12886) et le niveau de qualité globale de la relation (8).*

Etant donné que le résultat de l'étape « Projection dans le site actif » peut aussi être modélisé par des relations entre les fragments et les structures cristallisées, celles-ci ont aussi été ajoutées à la BDD. Chaque chaine de structures cristallographiques est représentée par un nœud ayant pour identité son code PDB permettant de la retrouver facilement. Un fragment ne présentant pas d'encombrement stérique avec une structure 3D sera ainsi relié à elle par une relation de type « C » pour compatible. En revanche, un fragment en collision avec un atome de la structure ou au contraire beaucoup trop éloigné des résidus du site actif ne sera pas relié à cette cavité, comme précisé lors de la présentation du module plus haut (partie 4.1.3.6).

En résumé, notre BDD orientée graphe est constituée de deux types de nœuds : fragment ou résidus des protéines. A ces nœuds s'ajoutent trois types de relations : E pour exclusion entre deux fragments car trop proches spatialement, I pour inclusion signifiant que les deux fragments peuvent être reliés entre eux et C pour compatibilité entre le fragment et une protéine. La BDD initiale de F2D prend environ 40 Go de mémoire, elle est constituée de plus de 80 000 nœuds et de plusieurs centaines de millions de relations. Grâce à l'optimisation du code pour calculer les relations entre fragments, elle est construite en moins de 24 h en partant des ligands alignés (fragmentation + calcul des relations et stockage).

4.2.3.2 Nouvel algorithme de construction des molécules

Le stockage des fragments ayant été modifié, il a ensuite fallu revoir la méthodologie pour les relier entre eux. De plus, comme je l'ai précisé à la fin du paragraphe présentant F2D (partie 4.1.2), la méthode initialement employée (dérivée du parcours en largeur) n'était pas adaptée et posait des problèmes de saturation de mémoire. A présent, nous allons tirer profit de la puissance des graphes et interroger directement la BDD.

Pour commencer, le lancement de F2D n'est pas révisé par rapport à la version initiale, il nécessite toujours un fragment de départ spécifié par l'utilisateur, avec optionnellement un ou plusieurs atomes précis, et une cible distincte (une structure cristallographique). La Figure

47 montre un exemple de fragment initial dans le site actif de la cible choisie pour démarrer F2D, l'atome sur lequel l'agrandissement sera effectué est entouré d'une ellipse noire.

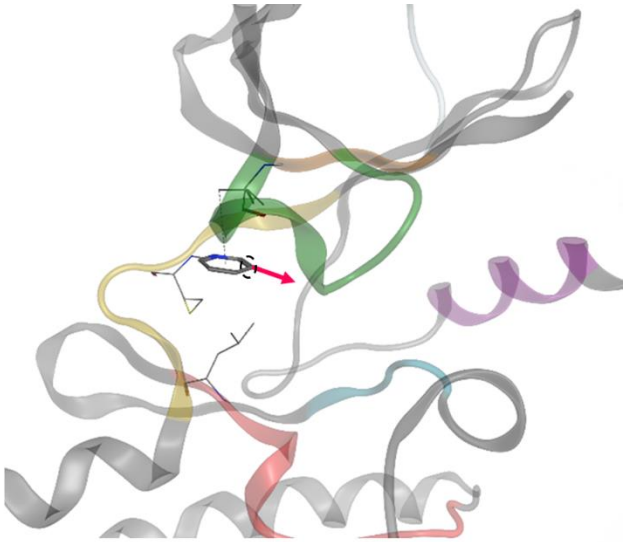


Figure 47 : Fragment de départ choisi pour lancer F2D dans sa cible. *Ici l'atome de départ est précisé, entourée en noir, pour diriger la croissance vers l'intérieur du site actif (code PDB : 2HYY).*

Une fois ces divers paramètres réglés, la première étape est la sélection des fragments compatibles avec la cible spécifiée. A l'aide d'une simple requête Cypher, les fragments ayant une relation de type C avec la structure sont conservés tandis que les autres sont écartés. Ensuite une autre requête Cypher permet d'éliminer tous les fragments ayant une relation de type E, donc présentant une gêne stérique avec le fragment initial. Cette première étape de sélection du sous-graphe de fragments adéquats pour les futures combinaisons est illustrée par la Figure 48.

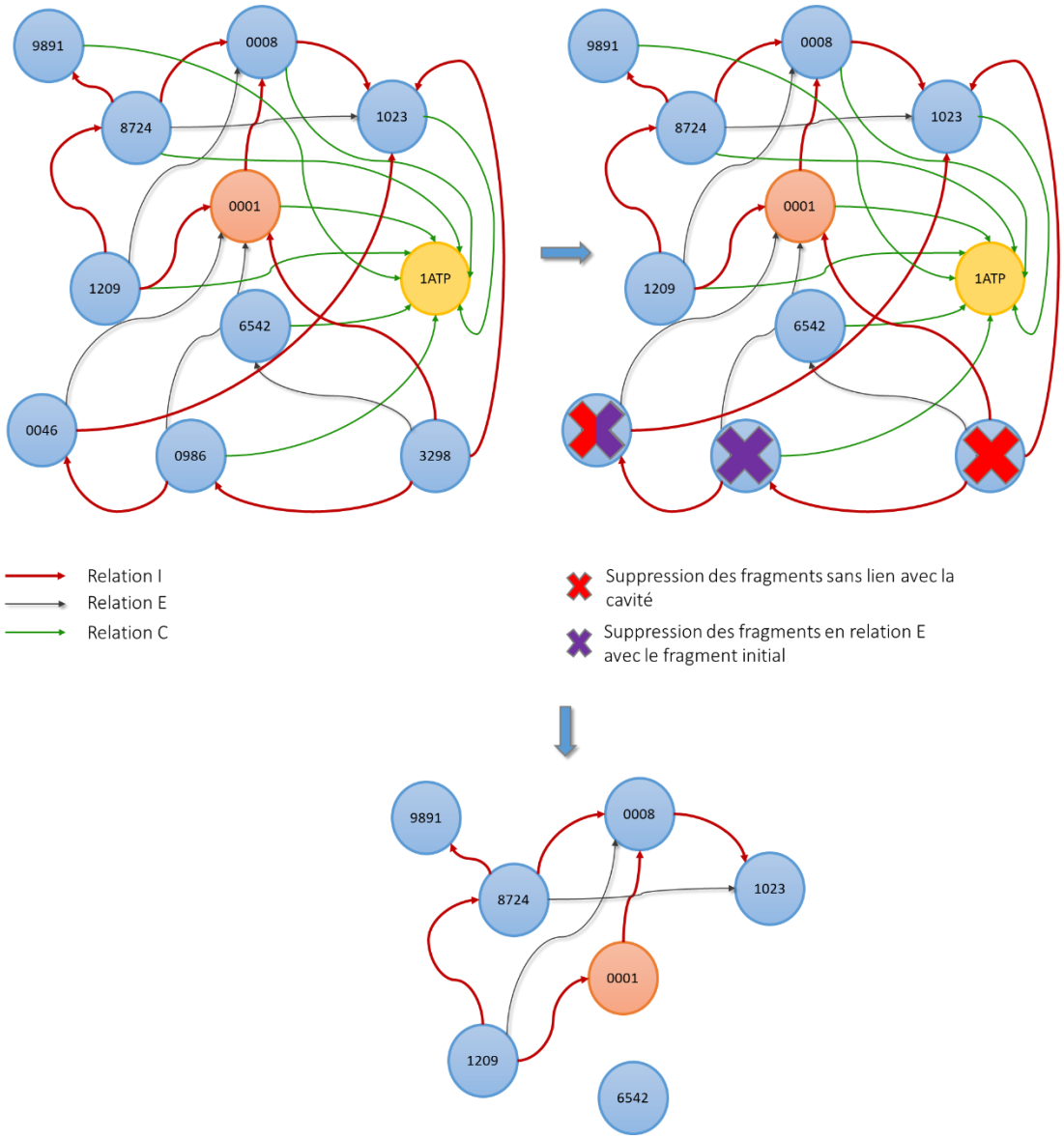


Figure 48 : Première sous-sélection des fragments avec élimination des fragments incompatibles avec la cavité précisée et ceux en exclusion avec le fragment initial. *Ici, le fragment initial serait le fragment 0001 (en rouge clair), la structure cristallographique 1ATP (en jaune). Après la sélection, nous obtenons un graphe soustrait des nœuds représentant les structures cristallographiques inutiles pour la suite.*

De ce sous-graphe, toujours à partir du fragment initial défini, on recherche tous les chemins possibles en passant par les relations I pour extraire à nouveau le sous-graphe correspondant. L'intérêt de cette deuxième sous-sélection permet de réduire encore le graphe de départ à parcourir en éliminant les fragments qui n'ont aucun lien direct ou indirect avec le fragment initial mais sont compatibles avec la cavité et ont donc passé le premier filtre. Dans la Figure 48, ce cas de figure concerne le fragment 6542. Mais surtout, pour cette deuxième sous-sélection nous prenons aussi en compte soit un seuil limite de fragments à ajouter au fragment de départ, soit un poids moléculaire maximal à ne pas dépasser. Le cas échéant F2D pourrait tourner presque indéfiniment au vu du nombre de possibilités de chemins. Dans notre cas, nous avons opté pour un seuil de poids moléculaire de 650 Da maximum. Cette opération est réalisable du fait que chaque nœud fragment possède l'attribut de son poids moléculaire. Ainsi, en additionnant au fur et à mesure du parcours les poids moléculaires, on connait le futur poids de la molécule et une fois le seuil atteint, le parcours sur la branche s'arrête. Après cette

deuxième étape de sélection, le graphe obtenu devient le graphe sur lequel F2D va travailler pour construire les molécules. Cette deuxième sous-sélection est présentée Figure 49.

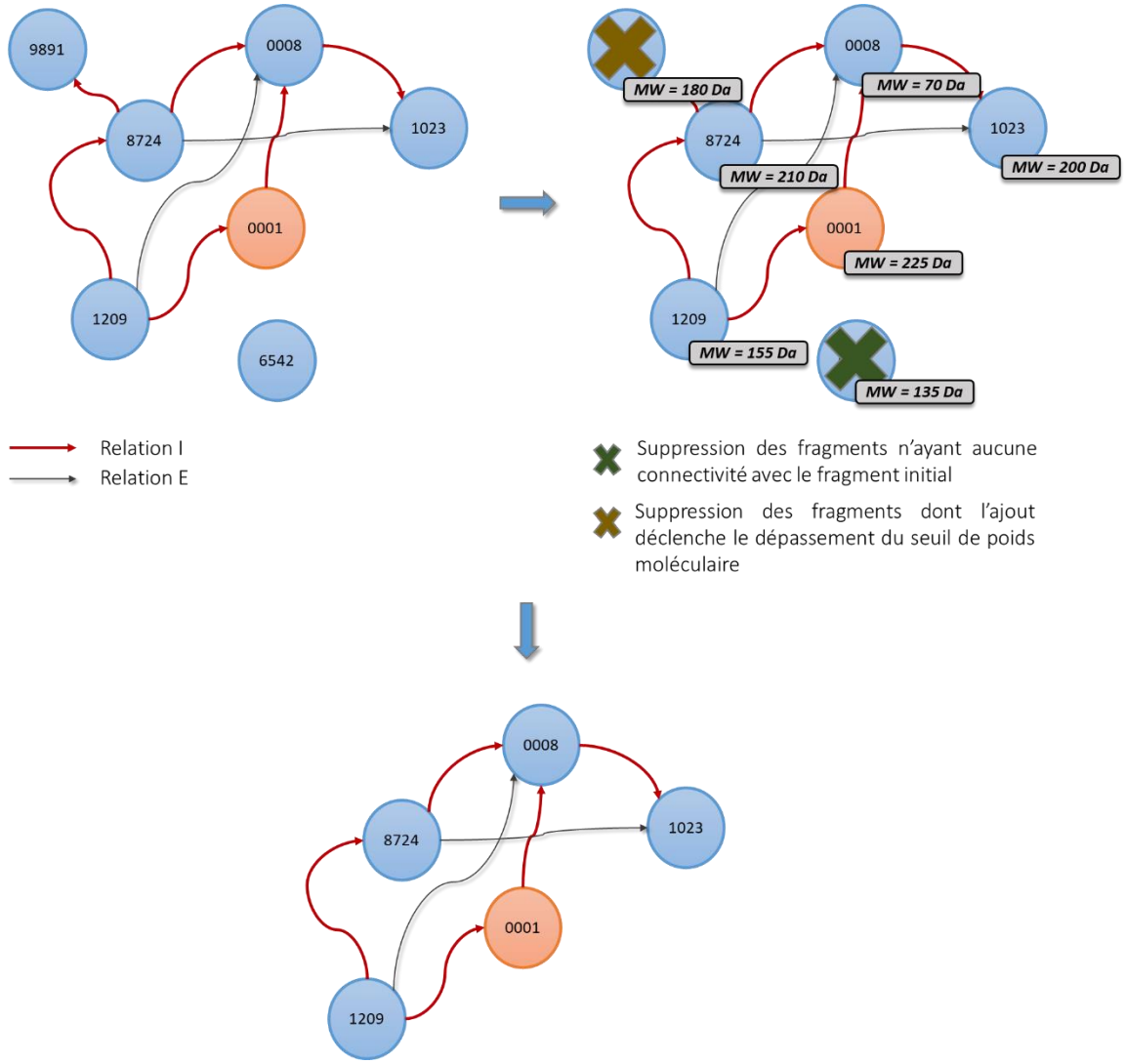


Figure 49 : Deuxième sous-sélection de fragments avec élimination des fragments sans lien avec le fragment initial et ceux responsables d'un dépassement du seuil de poids moléculaire limite spécifié. *Ici, le fragment 6542 n'est relié ni directement ni indirectement au fragment initial 0001. Le fragment 9891 est soustrait car il ne peut être ajouté à une molécule en construction (225 + 155 + 210 + 180 = 770 >> 650, le seuil limite).*

A ce stade, on peut légitimement penser qu'une fois ce travail exécuté, le plus important est accompli et qu'il ne reste plus qu'à suivre les relations I en partant du fragment initial pour créer les composés possibles. Il n'en est pourtant rien, en réalité c'est à partir de cet instant que la partie la plus complexe du programme débute. Le lecteur assidu aura remarqué que nous avons bien réglé le problème des relations E mais seulement avec le fragment initial. En effet, si l'on se réfère à la Figure 49, on constate qu'en parcourant les relations I, on peut construire, entre autres, les molécules suivantes : *0001-1209*, *0001-1209-8724*, *0001-1209-8724-0008*. Or, comme je l'ai déjà précisé, une relation E entre deux fragments signifie qu'ils ne peuvent se retrouver dans une même molécule car ils sont en collision. La molécule contenant les fragments *0001-1209-8724-0008* ne peut donc être retournée par F2D, du fait de la relation E entre les fragments *1209* et *0008*, sous peine de se retrouver avec un composé présentant une conformation impossible. Pour chaque visite de sommet, il faut ainsi vérifier si celui-ci ne

présente aucune relation E avec tous les autres fragments déjà ajoutés au fragment initial. D'autant que les relations E à vérifier lors de la construction ne sont que la partie émergée de l'iceberg. En effet, il y a beaucoup d'autres paramètres à vérifier lors du parcours du graphe à la recherche de toutes les combinaisons possibles et pour ajouter un fragment sur une molécule en construction. Il faut ainsi toujours regarder les attributs des relations I, notamment les atomes impliqués dans les liaisons afin d'éviter de rajouter un fragment sur le même atome par lequel le précédent est déjà branché. Au fur et à mesure du parcours d'une branche, il est donc nécessaire de constamment remettre à jour les valences atomiques et les possibilités de liaisons des atomes et d'adapter celles-ci en conséquence. Enfin, il faut aussi continuer à prendre en compte le poids moléculaire du fragment à ajouter sur la molécule en construction pour ne pas dépasser le seuil. Un exemple complet est fourni par la Figure 50 dans lequel à partir d'un petit graphe, toutes les combinaisons possibles de création de molécules sont retournées.

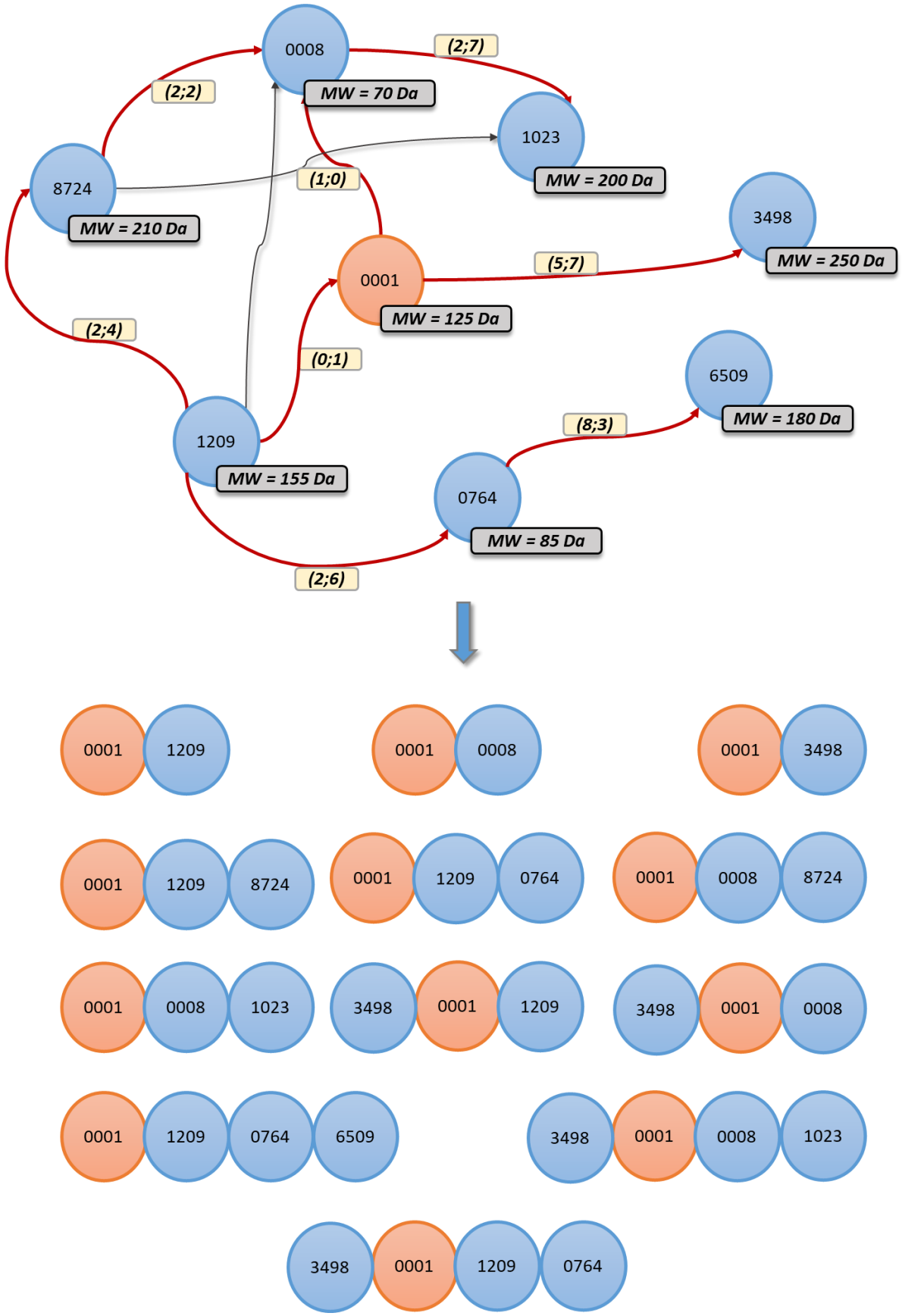


Figure 50 : Recherche de toutes les combinaisons possibles à partir d'un graphe déjà filtré. *Exemples de molécules que l'on ne reconstruit pas : 0001-1209-8724-0008 (relation E entre 1209 et 0008), 0001-1209-8724-0764 (le fragment 8724 et 0764 sont tous les deux reliés à l'atome 2 du fragment 1209, or celui-ci n'a qu'une possibilité de liaison), 3498-0001-1209-0764-6509 (dépassement du poids limite autorisé), etc.*

Une première difficulté nous est alors apparue : Neo4J demeure certes un outil très puissant pour un parcours de graphe simple mais présente quelques complications lorsqu'on lui demande de réaliser des retours en arrière entre les nœuds pour vérifier les relations E. Il faut en outre garder en tête les chiffres de la BDD et le nombre de relations à parcourir de plusieurs centaines de millions contrairement à l'exemple fourni très simplifié. Qui plus est, une BDD fonctionnant avec Neo4J est une BDD physique, elle est donc directement stockée sur le disque dur de l'ordinateur et fonctionne sous forme de serveur local. Il faut donc penser qu'à chaque requête, on interroge ce serveur. Il y a alors un laps de temps entre l'envoi de la requête, son exécution et le retour de la réponse. Même si ce laps de temps est insignifiant pour une requête, sur des milliers de requêtes d'affilée, il commence à peser dans le temps de calcul. Pour pallier ces difficultés, deux solutions ont été trouvées et mises en places :

- Le chargement du sous-graphe des fragments sélectionnés pour la construction des molécules directement dans la RAM avec la bibliothèque optimisée pour les graphes Networkx (https://networkx.github.io/). Si la BDD toute entière est beaucoup trop massive pour être chargée directement, le sous-graphe de sélection est quant à lui gérable. Le fait de travailler directement dans la mémoire vive permet d'éviter d'interroger un serveur distant et d'accéder plus vite aux informations. De plus, en chargeant le graphe avec Networkx, celui-ci devient un objet Python que l'on peut manipuler bien plus aisément, car comme je l'ai précisé, Neo4J fonctionne avec son propre langage qui n'est pas facile à appréhender quand il faut faire des requêtes compliquées avec beaucoup de conditions (notre cas) au contraire d'un simple parcours de graphe.

- La séparation du graphe en deux sous-graphes distincts : le graphe d'inclusion (GxI) rassemblant toutes les relations de types I et le graphe d'exclusion (GxE) rassemblant toutes les relations de types E. Cette séparation évite les retours en arrière lors d'un ajout de fragment en vérifiant s'il ne présente bien aucune relation E avec les autres. Il est en fait plus rapide de rechercher directement dans le GxE s'il existe des relations entre tous les fragments de la molécule en construction. Le parcours se fait ainsi sur le GxI tout en interrogeant à chaque nœud le GxE pour avancer.

Une fois que tous les chemins possibles ont été parcourus en respectant les conditions imposées, les molécules sont construites à l'aide des fragments qui se situent toujours dans une table extérieure à la BDD. En effet, Neo4J ne peut pas stocker directement les objets en tant que fragments. La BDD orientée graphe répertorie uniquement les identifiants propres à chaque fragment servant à les retrouver ensuite à partir du fichier SDF qui les contient tous. La Figure 51 illustre un exemple concret de la création de toutes les molécules possibles en prenant uniquement les 8 fragments provenant du vemurafenib avec le GxI et le GxE respectivement.

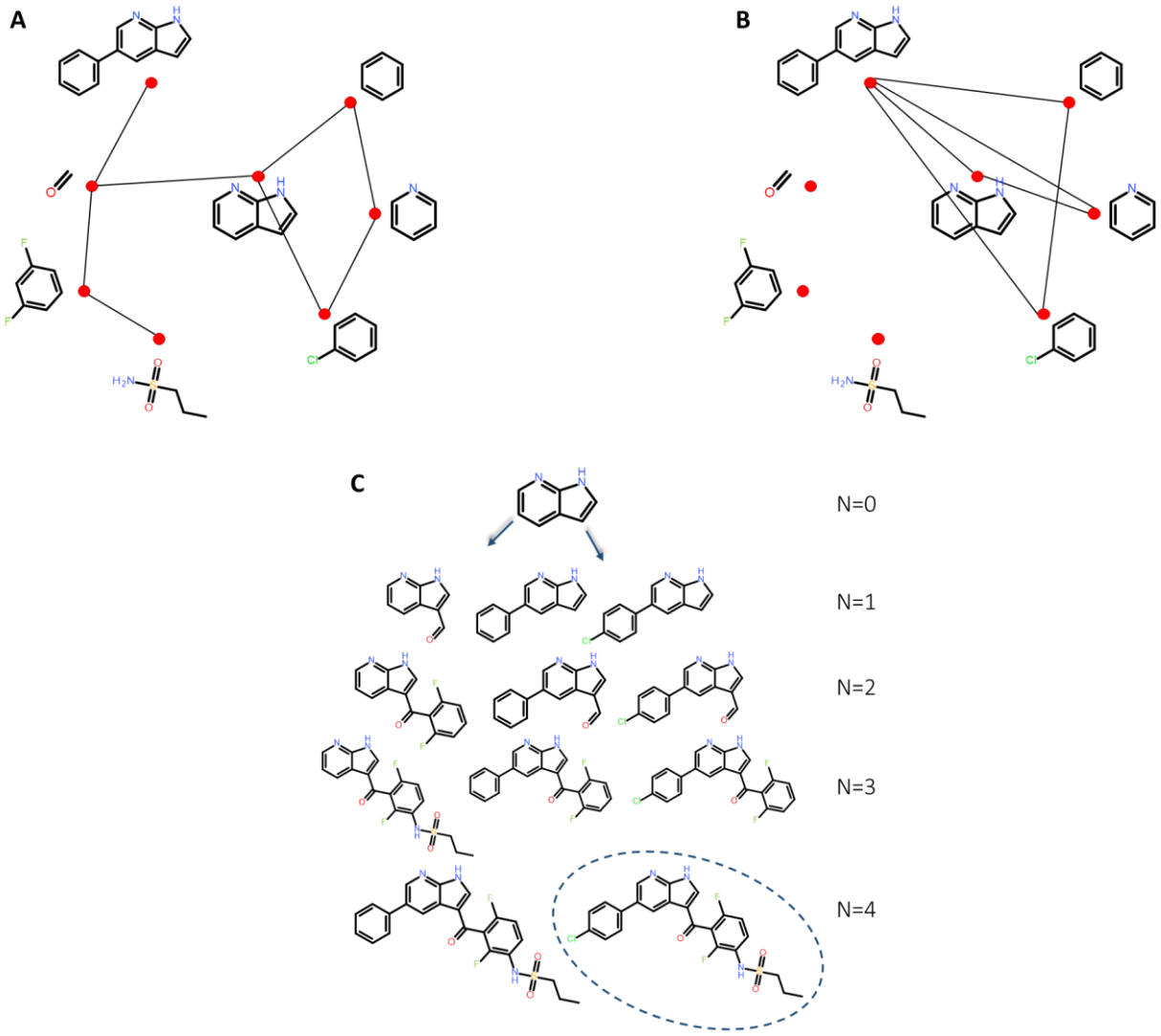


Figure 51 : Exemple d'un graphe d'inclusion (A), d'un graphe d'exclusion (B) et de la reconstruction de toutes les molécules possibles en partant du fragment azaindole (C). *Les résultats sont classés par nombre d'ajouts de fragments sur le fragment initial dont les flèches indiquent les atomes de départ.*

Pour finir, avant de retourner les résultats à l'utilisateur, une étape de regroupement est réitérée avec les molécules construites. En effet, un même chemin peut être traversé de plusieurs manières différentes par F2D et retourner la même molécule. Cette étape de regroupement est similaire à celle présentée dans le module « Regroupement des doublons » (regroupement par SMILES puis calcul du RMSD pour éliminer les poses < 1 Å). Les molécules sont ensuite retournées à l'utilisateur en format standard SDF.

4.2.3.3 Nouvelle méthode de fragmentation

La deuxième transformation conséquente du programme que j'ai effectuée concerne la fragmentation des ligands. Ce travail a été motivé par deux facteurs : s'affranchir du logiciel KNIME pour ne travailler qu'avec du langage Python et éviter d'avoir trop de fragments inutiles (un fragment faisant partie d'un autre). La nouvelle méthode de fragmentation a été développée avec les bibliothèques RDKit et Networkx, en transformant la molécule en graphe pour travailler dessus plus facilement. Comme pour Neo4J, Networkx permet de stocker des attributs sur les nœuds ou sur les relations dans un graphe. Ici, chaque nœud est un atome de la molécule

avec son hybridation, sa valence, etc., et chaque relation représente une liaison avec comme attribut son type. Les principes de notre méthode de fragmentation sont les suivants :

- La molécule est découpée en cycles et lieurs de cycles.
- Pour éviter les atomes solitaires, si un unique atome est relié à un cycle, cet atome reste avec le cycle.
- Les cycles fusionnés ne sont pas fragmentés.
- Les doubles/triples liaisons ne sont pas fragmentées.
- Lorsque deux cycles (fusionnés ou simples) sont reliés par un unique atome, 4 fragments sont formés : chaque cycle seul et chaque cycle avec cet atome en plus (Figure 52).

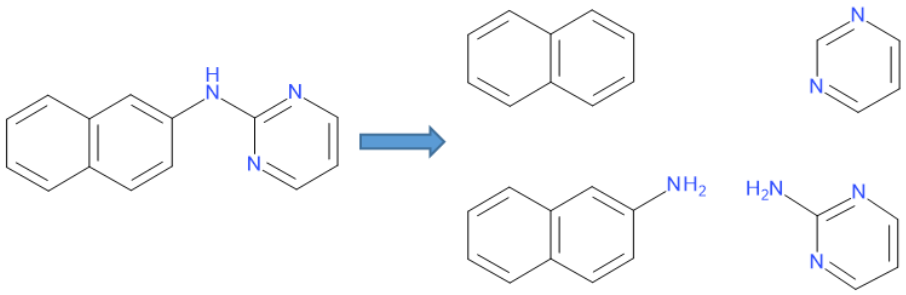


Figure 52 : Cas particulier de la fragmentation de deux cycles reliés par un seul atome.

La Figure 53 montre les différences entre l'ancienne méthode de fragmentation réalisée avec KNIME et la nouvelle développée entièrement en langage Python.

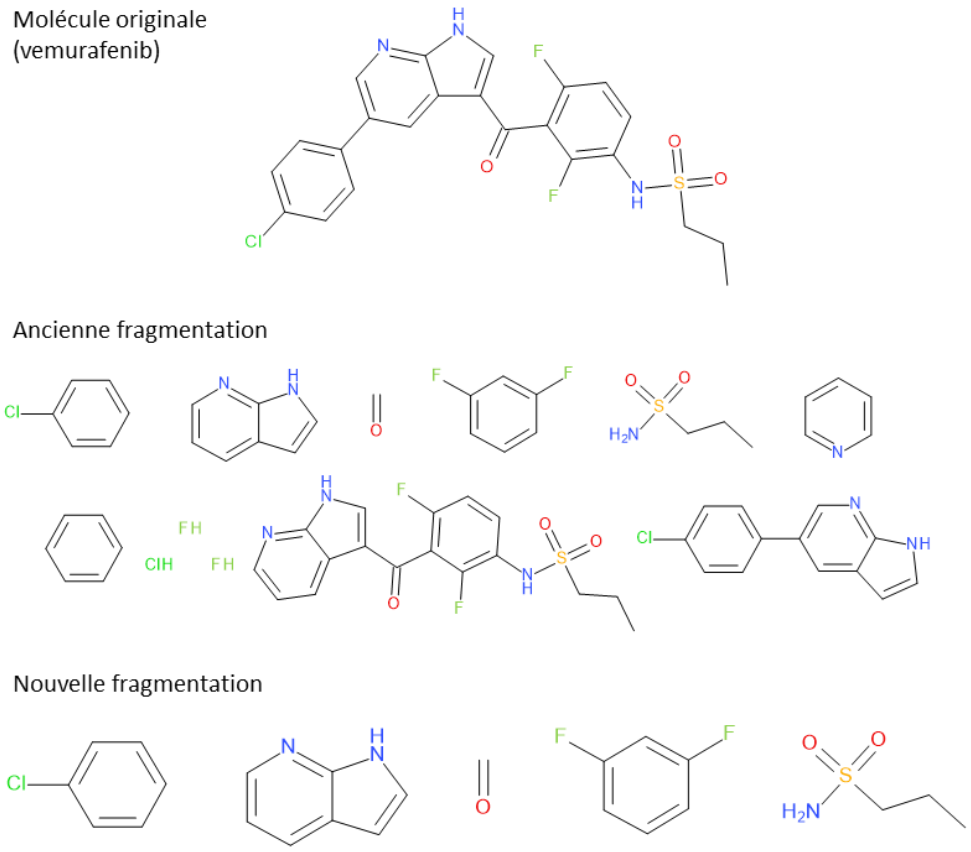


Figure 53 : Différence entre ancienne et nouvelle méthode de fragmentation visualisée sur le vemurafenib.

Comme constaté sur cette figure, avec la nouvelle méthode de fragmentation, il n'y a pas de redondance de fragments (fragment contenant un autre fragment). Un filtre a ensuite été mis en place pour supprimer les fragments ne respectant pas le poids de la règle des 3 (> 300 Da). Nous sommes ainsi passé de 81 667 fragments à 21 835. A noter que ces chiffres représentent le nombre de fragments après l'étape de regroupement pour supprimer les doublons. Les caractéristiques de ces deux BDD sont résumées Tableau 3, Tableau 4 et par la Figure 54.

Tableau 3 : Paramètres statistiques des différents descripteurs physico-chimiques de l'ancienne librairie de fragments.

	MW	CLogP	HBA	HBD	NRB	TPSA	NHA	NAR	NCA
Minimum	26,00	-6,14	0	0	0	0	2	0	0
Maximum	300,00	5,50	9	6	13	177,60	23	5	7
Moyenne	149,27	1,09	2,32	1,10	1,03	42,13	10,77	1,36	0,09
Médiane	136,00	1,14	2,00	1,00	1,00	40,50	10,00	1,00	0,00
Ecart-type	72,60	1,25	1,69	0,93	1,34	27,52	5,37	0,99	0,39

Avec MW la masse moléculaire, CLogP le coefficient octanol-eau calculé avec RDKit, HBA et HBD pour nombre d'accepteurs et de donneurs de liaisons hydrogène respectivement, NRB le nombre de liaisons à rotation libre, TPSA la surface polaire topologique, NHA le nombre d'atomes lourds, NAR le nombre de cycles aromatiques et NCA le nombre d'atomes chiraux.

Tableau 4 : Paramètres statistiques des différents descripteurs physico-chimiques de la nouvelle librairie de fragments.

	MW	CLogP	HBA	HBD	NRB	TPSA	NHA	NAR	NCA
Minimum	26,00	-3,76	0	0	0	0,00	2	0	0
Maximum	281,10	5,03	7	4	9	118,40	21	5	6
Moyenne	87,49	0,65	1,20	0,68	0,11	24,94	6,34	0,80	0,02
Médiane	80,00	0,80	1,00	1,00	0,00	25,80	6,00	1,00	0,00
Ecart-type	34,47	1,07	1,06	0,68	0,54	20,03	2,68	0,76	0,21

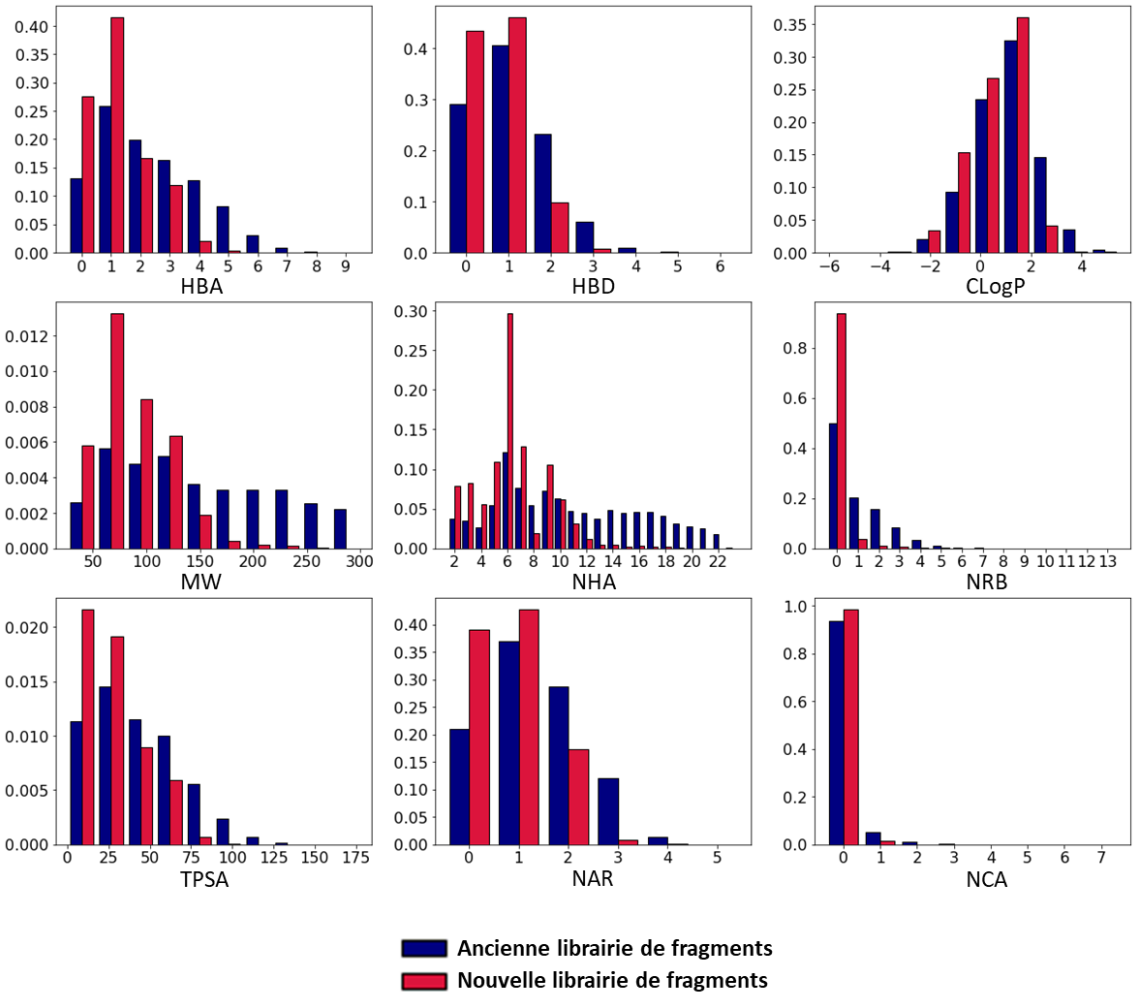


Figure 54 : Histogrammes comparant les fréquences de distribution de plusieurs descripteurs moléculaires entre l'ancienne librairie de fragments et la nouvelle.

La baisse importante du nombre de fragments permet de diminuer le nombre de combinaisons et la taille de stockage de la BDD, rendant F2D plus optimal. La différence majeure entre les deux librairies se situe au niveau de la taille des fragments les constituant. Dans la nouvelle librairie il n'y a plus de fragments imposants grâce à la nouvelle méthode qui fragmente systématiquement entre les cycles ce qui n'était pas le cas avant (moyenne du poids moléculaire 149 Da avant contre 87 Da désormais). Logiquement, les descripteurs corrélés au poids moléculaire montrent aussi une différence comme le nombre d'atomes lourds (NHA) avec une moyenne qui passe de 10 à 6. En règle générale, ceux-ci vont baisser par rapport aux valeurs de l'ancienne librairie de fragments. Cependant, en regardant l'histogramme du CLogP,

on constate que sa distribution, ainsi que celle du nombre d'atomes chiraux reste similaire dans les deux librairies. La majorité des fragments présentent un caractère à tendance lipophile (CLogP > 0) corroborant les données des librairies commerciales de fragments[207,208,209]. En ce qui concerne les atomes chiraux (NCA), seuls quelques fragments en possèdent mais le plus grand nombre en est dépourvu.

De toutes les façons, il faut bien comprendre que le but de ce travail n'est pas de trouver des nouveaux fragments ou de modifier intrinsèquement l'ancienne librairie de fragments, mais bien d'harmoniser l'outil et de passer uniquement par Python tout en simplifiant la méthode originale. Il eut été vain de vouloir transformer totalement la librairie de fragments puisque les données de départ (les ligands co-cristallisés) restent les mêmes dans les deux cas. On ne peut donc pas « inventer » de nouveaux fragments.

4.3 Validation de F2D

Pour valider notre outil et notre méthode de fragmentation nous avons décidé de vérifier si celui-ci était bien capable de reconstruire les ligands co-cristallisés dans les structures de kinases de la PDB. Le but étant de prouver que F2D est apte à retrouver les données déjà existantes avant de fournir d'autres solutions de composés pour une cible donnée. Pour cela, un protocole automatique de reconstruction a été mis en place. Ainsi, pour chaque ligand ses fragments sont isolés du reste de la BDD et le fragment relié à la charnière centrale de la protéine kinase est choisi comme fragment initial. Si aucun fragment n'est relié à la charnière, un autre fragment est sélectionné aléatoirement car dans le cadre de cette validation le choix du fragment initial importe peu. Ensuite, à partir de ce fragment initial, toutes les molécules possibles sont construites et on vérifie la présence du ligand co-cristallisé dans la liste de celles-ci. Lors du premier essai, environ 65 % des ligands co-cristallisés ont pu être reconstruits. Cette valeur n'étant pas suffisante, des investigations manuelles ont été menées sur les ligands ayant échoués afin de comprendre pourquoi. Je vais décrire plus en détail ci-dessous plusieurs problèmes rencontrés et les solutions apportées.

4.3.1 Cas des distance inter-atomiques

La distance inter-atomique concerne la distance entre atomes d'une même molécule. Ce premier problème soulevé survient notamment lorsque le ligand se replie beaucoup sur lui-même et que ses atomes sont de ce fait très (voire trop) rapprochés. Un tel cas de figure est illustré Figure 55. Comme on le voit, le groupement morpholine terminal, de par le repliement du ligand dans la cavité, se rapproche du groupement benzène central.

[207] « Fragment Libraries | FBDD screening compounds | Life Chemicals », consulté le 18 septembre 2019, https://lifechemicals.com/screening-libraries/fragment-libraries.
[208] « Library for Fragment-Based Drug Discovery – Prestwick Chemical », consulté le 18 septembre 2019, http://www.prestwickchemical.com/libraries-screening-lib-drug-frag.html.
[209] « Fragment Library-MedchemExpress », MedchemExpress.com, consulté le 18 septembre 2019, https://www.medchemexpress.com/screening/Fragment_Library.html.

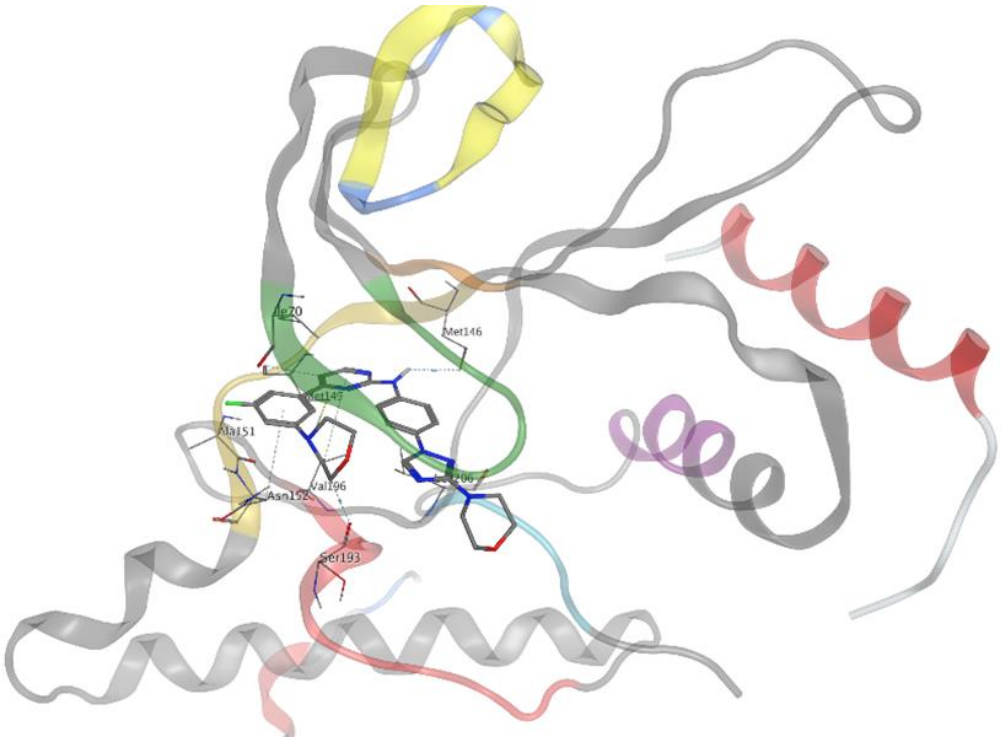


Figure 55 : Ligand SR-3562 lié à MAPK10. *Code PDB : 3KVX.*

La Figure 56 représente le ligand en 2D et sa fragmentation par F2D en 8 fragments uniques. L'analyse de la fragmentation seule montre qu'il est tout à fait possible de le reconstruire en suivant l'enchainement suivant : *4->2->5->6->1->3* ou *4->2->7->0->1->3*. La fragmentation n'est donc pas mise en cause dans l'échec de la reconstruction du ligand.

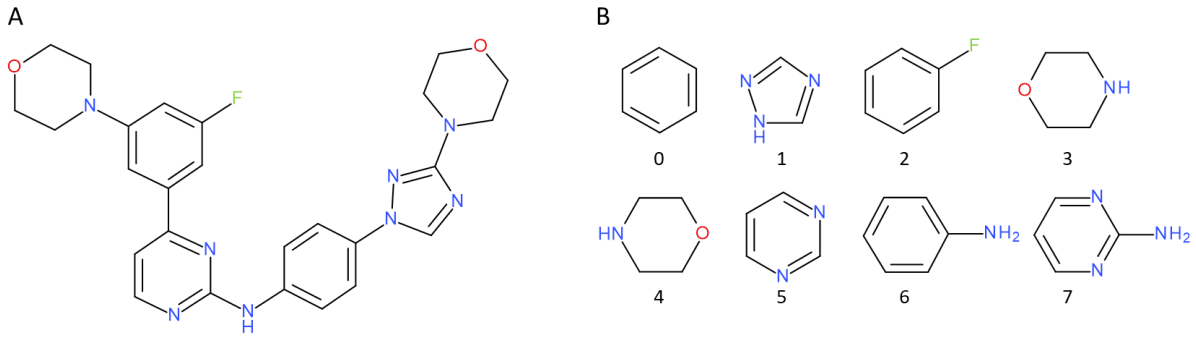


Figure 56 : Représentation en 2D du ligand SR-3562 (A) et sa fragmentation (B).

Pour trouver la cause, nous nous sommes ensuite intéressés au graphe des connexions possibles entre les fragments (Figure 57). Sur la figure les relations sont représentées par défaut avec une direction (flèche orientée), mais dans la pratique on peut ignorer le sens de ces relations (en le précisant dans les requêtes) et considérer ce graphe comme non-orienté.

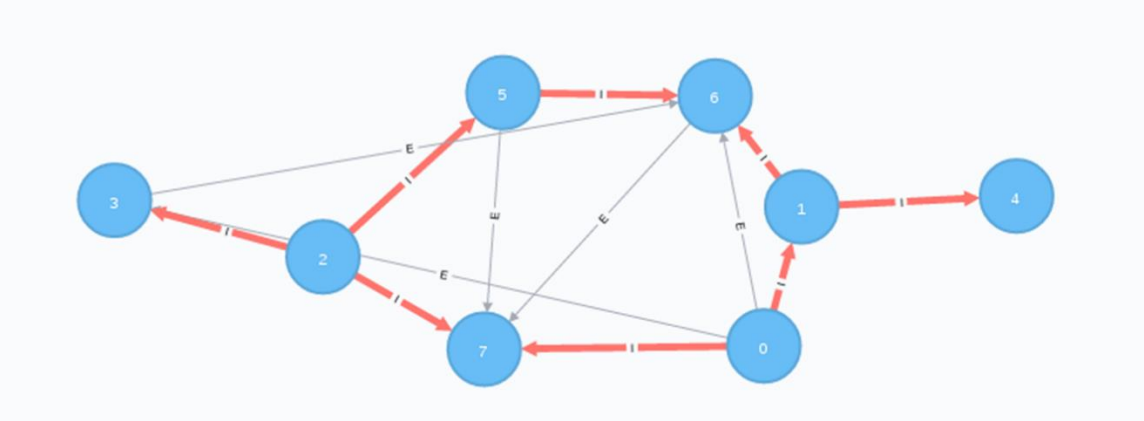


Figure 57 : Graphe représentant les possibilités de connexion entre les fragments du ligand SR-3562. *Les relations I signifient que deux fragments peuvent être reliés, les relations E signifient qu'ils sont incompatibles.*

En se fiant aux relations I (en rouge), on constate bien que les deux chemins énoncés au-dessus sont possibles et que les fragments sont bien capables d'être reliés entre eux pour recréer le ligand initial. Là encore, cela valide la méthode de F2D pour calculer les liaisons potentielles entre deux fragments. Dans ce cas, pour discerner le problème il faut regarder les relations E (en gris). Sur le graphe, il y a 5 relations E : entre les fragments 5 et 7, 6 et 7, 0 et 6, 3 et 6 et 3 et 0. Les 3 premières relations E découlent de la méthode de fragmentation et sont tout à fait justifiées, la Figure 56 (B) indique bien qu'il s'agit de fragments ayant des atomes en commun et présentant donc une gêne stérique (soit le cycle entier pour le couple de fragments 5 et 7 et 0 et 6, soit l'atome de liaison entre les cycles pour le couple 6 et 7). Cependant, ce ne sont pas ces relations E qui empêchent la reconstruction du ligand car comme on l'a vu les deux chemins sont possibles pour le recréer. Ce qui bloque la reconstruction du ligand c'est la relation E entre le fragment 6 et 3 d'une part et 3 et 0 d'autre part car quand il existe une relation E entre deux fragments, ces deux fragments ne peuvent se retrouver dans la même molécule. Ainsi, si l'on refait le chemin de F2D en commençant par le fragment 4, on obtient :

- Itération 1 : *4*
- Itération 2 : *4->1*
- Itération 3 : *4->1->6* ou *4->1->0*
- Itération 4 : *4->1->6->5* ou *4->1->0->7*
- Itération 5 *: 4->1->6->5->2* ou *4->1->0->7->2*
- Itération 6 : *stop*, bien que les fragments 2 et 6 soient reliés par une relation I, F2D vérifie avant que le fragment à ajouter (3) ne soit incompatible avec aucun des autres fragments sur la molécule. Dans ce cas, le fragment 3 est incompatible avec le 6 et avec le 0, les deux chemins s'arrêtent et le ligand n'est donc pas reconstruit.

La question qui se pose est pourquoi il existe une relation E entre le fragment 3 et les fragments 0 et 6 ? Cette relation provient du repliement de la molécule sur elle-même dans la cavité, causant une promiscuité entre un atome du fragment 3 et un atome du cycle benzénique commun aux fragments 0 et 6, comme le révèle la Figure 58 et le calcul de distance entre ces

deux atomes. La distance de 2,28 Å entre eux se situe sous la limite préalablement paramétrée dans F2D pour définir un encombrement stérique et ce afin d'éviter de construire des molécules énergiquement défavorables de conformation incorrecte. Notre outil détecte donc ces deux atomes comme en collision et les fragments concernés se retrouvent en situation d'exclusion.

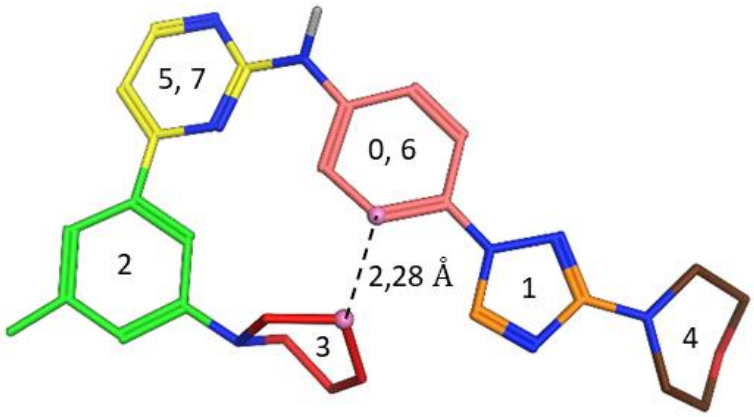


Figure 58 : Calcul de la distance entre deux atomes du ligand SR-3562. *Les numéros dans la molécule correspondent aux numéros des fragments.*

Cette situation se retrouve dans plusieurs autres ligands, la solution apportée a été de baisser le seuil de tolérance de l'encombrement stérique à 2,25 Å. En revanche, lorsque l'abaissement de cette limite n'a pas suffi, le choix a été pris d'écarter les ligands du set de validation en estimant que ceux-ci présentaient des erreurs au niveau des positions atomiques et une conformation trop défavorable. Nous n'avons pas réalisé de filtre des structures de la PDB concernant la résolution, néanmoins les molécules qui présentent ce type de conformation ont majoritairement des résolutions moyennes ou mauvaises (>> 2 Å).

4.3.2 Cas de la distorsion des angles

Cette deuxième situation problématique rencontrée nous empêchant de reconstruire les ligands co-cristallisés est aussi dû à l'encombrement stérique intramoléculaire. Il s'agit cette fois-ci d'une gêne stérique amenant à des déformations importantes au niveau des angles dihédraux et de la planarité moléculaire, particulièrement quand deux cycles aromatiques sont liés à un troisième sur des atomes adjacents (Figure 59). Cette proximité de cycles « force » ceux-ci à s'éloigner du plan parallèle et à se positionner plus ou moins perpendiculairement au cycle central.

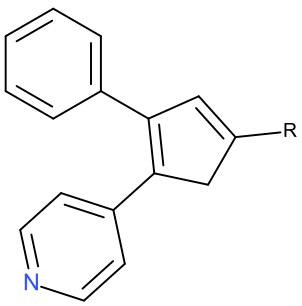


Figure 59 : Exemple de molécule présentant une gêne stérique intramoléculaire modifiant la planarité.

On retrouve cette configuration dans de nombreuses structures cristallographiques, comme illustrés Figure 60.

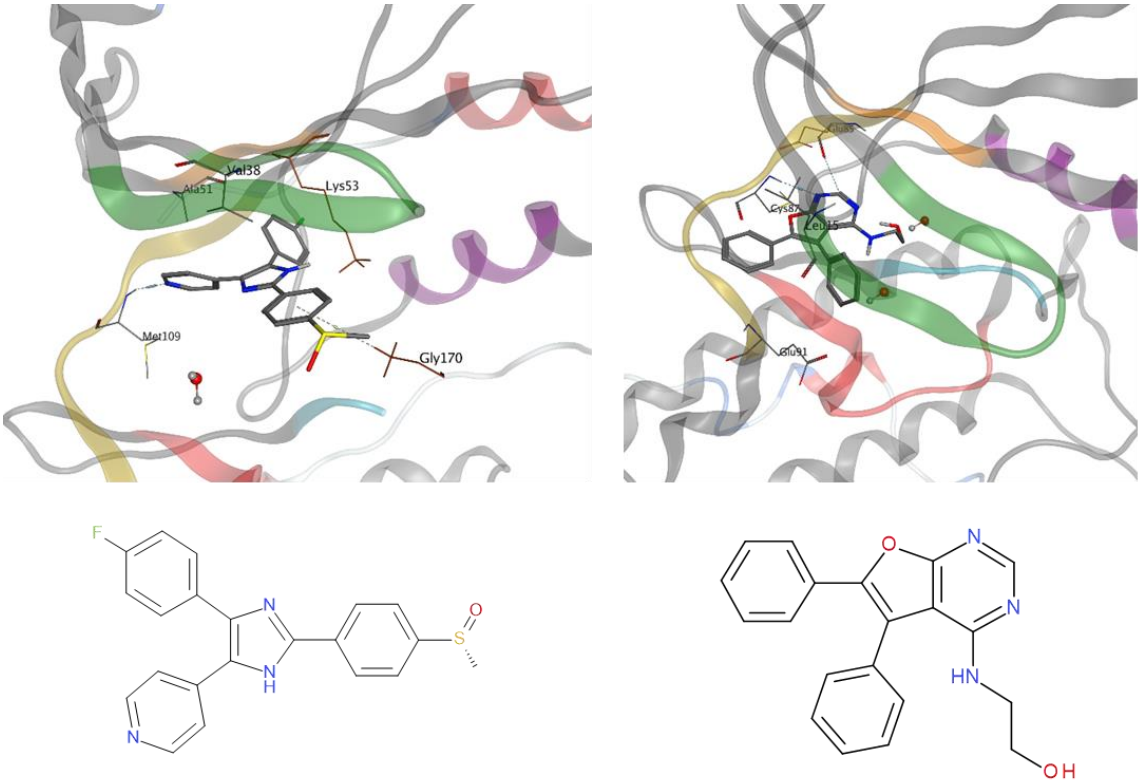


Figure 60 : Représentation dans leurs sites actifs de deux ligands co-cristallisés avec leurs structures en 2D correspondantes. *Codes PBD : 2EWA (gauche) et 2BRB (droite).*

Ces deux exemples représentent deux situations différentes : dans le cas de la structure 2EWA les deux cycles attachés au groupement imidazole sont tous les deux tournés d'environ 60° par rapport à celui-ci à cause de leurs gênes stériques mutuelles, tandis que pour la structure 2BRB, un des groupements benzène attaché au noyau furopyrimidine est bien dans le plan défini par ce dernier, alors que l'autre lui est quasiment perpendiculaire. La Figure 61 permet de mieux apprécier l'arrangement spatial des cycles que je viens de décrire.

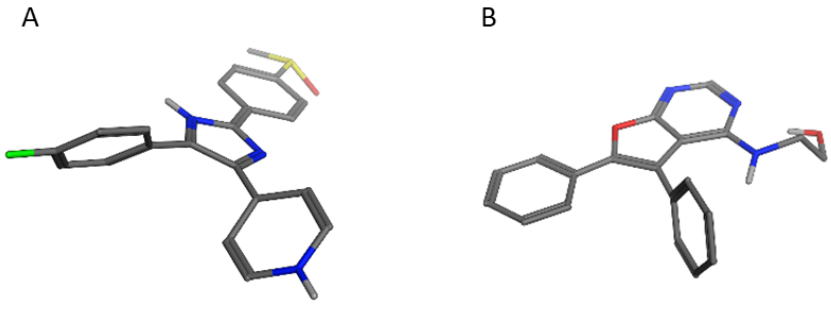


Figure 61 : Extraction de ligands co-cristallisés. *Provenance des ligands : PDB 2EWA (A) et PDB 2BRB (B).*

Les fragments de chacun de ces ligands ainsi que les relations entre eux sont présentés Figure 62. Lors du calcul des possibilités de liaison entre les fragments, l'étape de vérification des angles de liaisons et des angles dièdres échoue car l'outil va considérer que la déformation est trop importante et donc que la liaison n'est pas possible. Or, comme expliqué dans

l'introduction de F2D, comme ces deux fragments sont suffisamment proches pour créer une liaison mais que les conditions ne sont pas réunies, ils sont alors considérés en exclusion. Il faut bien comprendre que F2D calcule les relations deux à deux sans prendre en compte l'environnement extérieur, c'est pour cela qu'il estime que les molécules seraient difformes et met ces fragments en exclusion. En effet, s'il n'y avait qu'un seul cycle aromatique relié sur le cycle central, il serait bien plan. On constate cela en regardant la Figure 62, lorsque les deux cycles sont à environ 60°, les deux présentent une exclusion avec le fragment central, mais si l'un des deux est plan et l'autre perpendiculaire, seul le cycle en position perpendiculaire sera en situation d'exclusion.

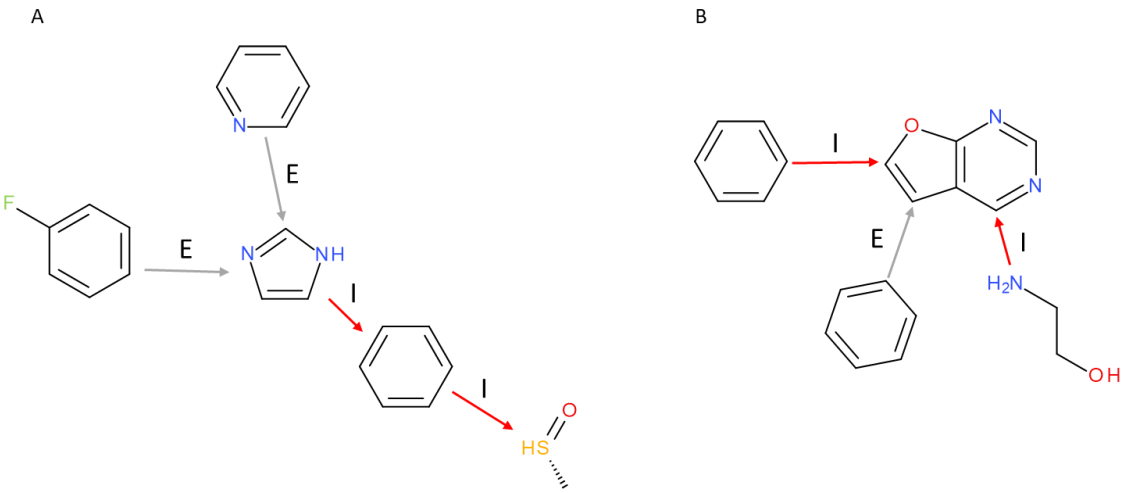


Figure 62 : Représentation des relations entre les différents fragments d'un ligand. *Provenance des ligands : PDB 2EWA (A) et PDB 2BRB (B).*

Pour remédier à ce problème et reconstruire les ligands co-cristallisés présentant cette structure chimique, la règle de fragmentation a été adaptée. Désormais, lorsqu'un ligand présente ce cas de figure, F2D conserve en plus un fragment contenant le squelette aromatique central et les deux cycles reliés dessus (Figure 63). Cette nouvelle règle nous permet de valider le logiciel en reconstruisant bien les ligands co-cristallisés tout en évitant un laxisme trop important au niveau des angles dièdres de liaison nouvellement créées et d'éviter de construire un nombre trop important de molécules avec une géométrie incorrecte.

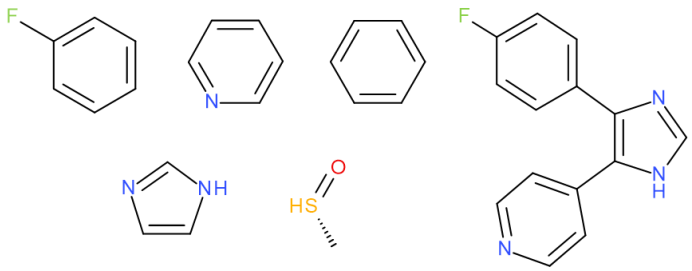


Figure 63 : Exemple de la nouvelle fragmentation sur un ligand co-cristallisé. *Provenance du ligand : PDB 2EWA.*

4.3.3 Cas des valences atomiques

Cet autre problème récurrent est cette fois rencontré lors de la fragmentation. Il concerne la valence des atomes, notamment des atomes d'azote et de soufre. Par exemple, comme le

montre la Figure 64, lorsqu'un ligand présente un groupement sulfone relié à des cycles de part et d'autre. Dans ce cas, l'atome de soufre est en hybridation sp3 hexavalent. Or, une fois fragmenté ce même soufre se retrouve amputé de 2 liaisons car les cycles sont isolés. Comme le soufre peut aussi adopter une hybridation sp2 tetravalent, RDKit considère qu'il est en configuration correcte et le laisse tel quel, au lieu de lui ajouter deux atomes d'hydrogène à la place des cycles. Seulement, lors de l'étape de recherche des atomes potentiels pouvant accepter une liaison dans le fragment, comme ce soufre est passé en hybridation sp2 et une valence comblée, ce dernier n'est pas considéré comme pouvant accepter de liaison supplémentaire. Ainsi, F2D ne le prendra pas en compte pour la suite des opérations. Il sera alors impossible de reconstruire le ligand initial.

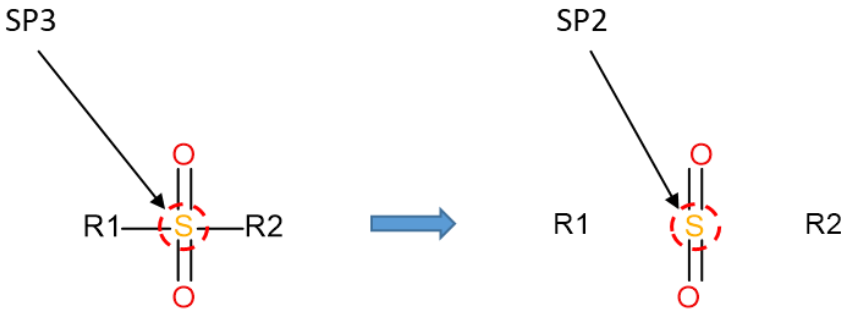


Figure 64 : Fragmentation d'un groupement sulfone relié à deux cycles. *Ici R1 et R2 sont des cycles.*

Pour pouvoir travailler avec des objets moléculaires, RDKit nécessite de vérifier qu'une molécule est correcte (fonction « Sanitize »), sinon la librairie va boguer et faire arrêter le programme. Dans certains cas, la fragmentation modifie les propriétés atomiques de la molécule considérée et RDKit ne parvient plus à comprendre la molécule. Il s'agit d'une situation rencontrée pour plusieurs ligands, qui dépend notamment de la forme tautomérique adoptée pour le représenter. Dans ce manuscrit je ne vais pas m'appesantir plus sur ce sujet qui pourrait faire l'objet d'une autre thèse à lui seul… Pour y remédier, nous gardons en mémoire les attributs des atomes du ligand initial (valence, hybridation…), et nous les réassignons manuellement si besoin lors de bogues suite à la fragmentation, du fait des modifications réalisées par RDKit. L'avantage de RDKit est d'avoir la possibilité de modifier manuellement ce genre de propriétés et ainsi forcer la valence ou l'hybridation à la valeur souhaitée, à condition que la structure chimique reste cohérente bien sûr.

4.3.4 Cas de la fragmentation

Là encore, il s'agit d'un problème dû à la méthode de fragmentation qui concerne certains macrocycles, notamment la staurosporine et ses dérivés. La non reconstruction de tels ligands s'explique par notre choix délibéré de ne pas casser les cycles fusionnés. Or, dans le cas de la staurosporine ou de ses dérivés, cela amène intrinsèquement à un très gros fragment dont la taille excède la limite fixée à 300 Da, rendant la reconstruction impossible par manque de fragments essentiels (Figure 65).

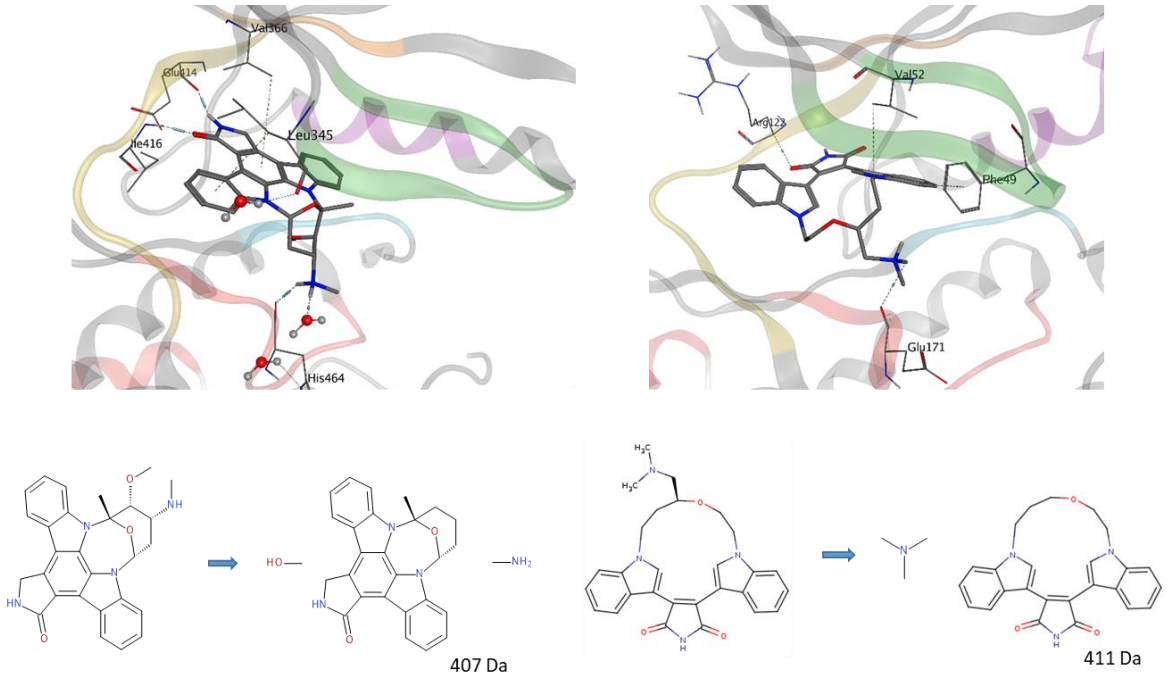


Figure 65 : Exemple de deux macrocycles dans leurs sites actifs avec leurs structures en 2D et leurs fragmentations. *Ici, comme le filtre de poids moléculaire des fragments est paramétré à 300 Da, les fragments de 407 et 411 Da ne seront pas conservés, empêchant la reconstruction du ligand. Codes PDB : 3S95 (gauche) et 2J2I (droite).*

La solution apportée est d'écarter ces composés du set de validation, nous savons bien que F2D est capable de les retrouver si besoin mais nous ne souhaitons pas avoir des fragments trop imposants qui vont tout de suite limiter l'agrandissement du fragment initial du fait de notre limite de poids imposés.

4.3.5 Cas de la structure 3D

Certaines données dans la PDB sont clairement erronées et le ligand est trop déformé pour que nous puissions le reconstruire. La Figure 66 illustre deux de ces ligands sans la protéine pour mieux les observer. Les images démontrent clairement que ces ligands présentent des erreurs et ne peuvent réellement avoir cette conformation, (il ne faut pas oublier que la structure cristallographique reste un modèle pouvant présenter des bévues)[210].

[210] Andrew M. Davis, Stephen A. St-Gallay, et Gerard J. Kleywegt, « Limitations and lessons in the use of X-ray structural information in drug design », *Drug Discovery Today* 13, nº 19 (1 octobre 2008): 831-41, https://doi.org/10.1016/j.drudis.2008.06.006.

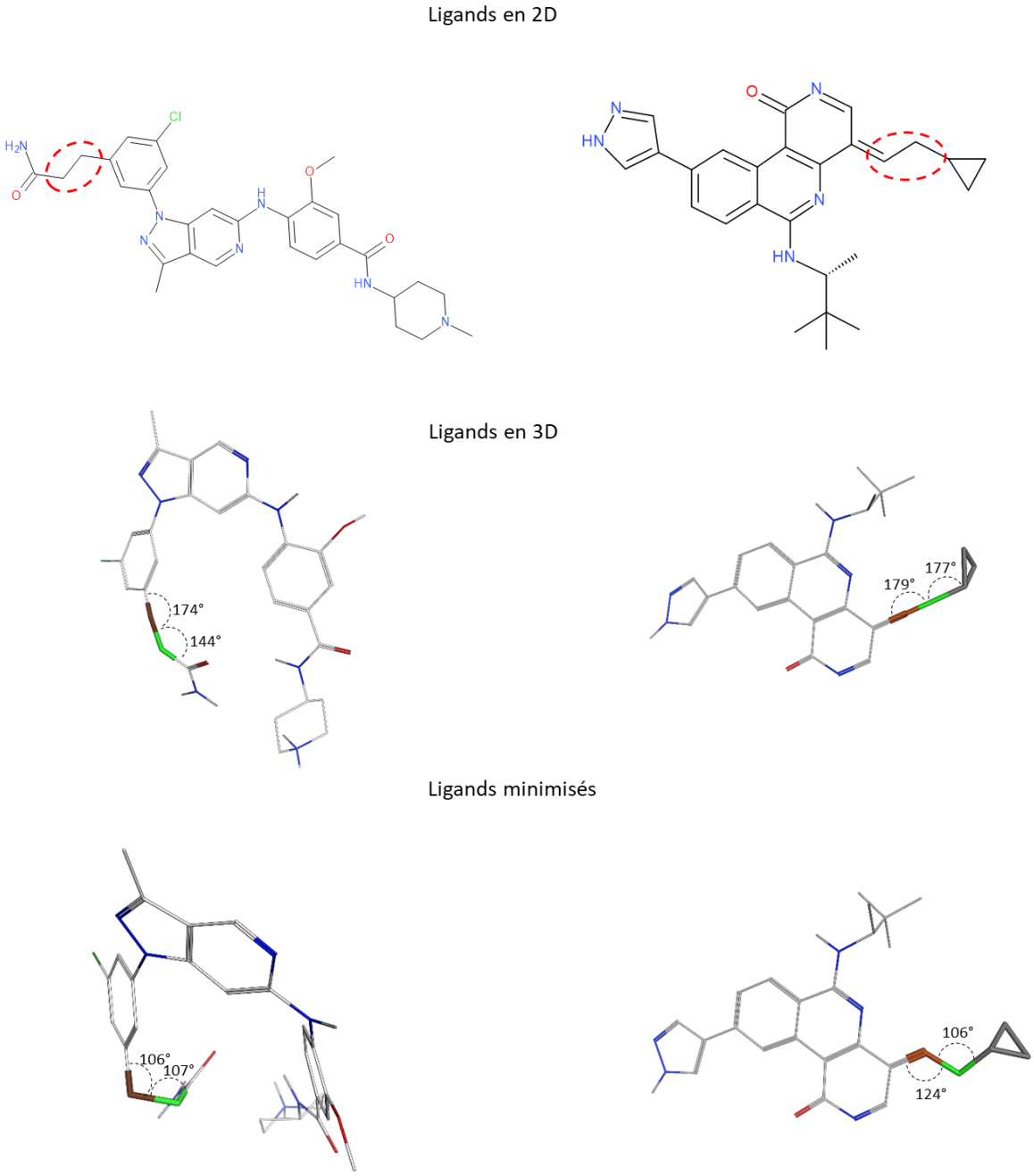


Figure 66 : Exemple de ligands co-cristallisés présentant une incohérence structurale. *L'ellipse rouge sur les ligands en 2D montrent les atomes impliqués dans les liaisons incorrectes. Les atomes impliqués dans ces liaisons sont coloriés dans les ligands en 3D et les valeurs des angles de liaisons indiquées. La minimisation est effectuée avec la fonction « Energy Minimize » de MOE avec les paramètres définis par défaut. Provenance des ligands : PDB 3DBF (gauche) et PDB 3NAY (droite).*

Dans ces deux exemples, l'aberration se situe au niveau d'une liaison simple aliphatique. Les valeurs d'angles sont beaucoup trop élevées par rapport aux valeurs théoriques attendues (autour de 109° pour un carbone tétrahédrique)[211,212]. Ici aussi, ce type de structure

[211] « Illustrated Glossary of Organic Chemistry - Bond angle », consulté le 19 septembre 2019, http://www.chem.ucla.edu/~harding/IGOC/B/bond_angle.html.
[212] Thomas A. Halgren, « Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94 », *Journal of Computational Chemistry* 17, nº 5-6 (1 avril 1996): 490-519, https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

présente des valeurs de métriques de validation très moyennes (que l'on retrouve sur le graphique de la page d'accueil d'une structure sur le site de la PDB : « *la qualité globale en un coup d'œi*l »). Malgré un seuil de tolérance, notre outil n'est pas paramétré pour recréer ce genre de liaison, et ce encore une fois afin de ne pas créer de molécules avec une conformation aberrante. Toutes les structures présentant une conformation clairement erronée sont donc délibérément retirées du jeu de validation.

Une fois toutes les investigations manuelles terminées, après avoir réglé les problèmes rencontrés et ajustés les paramètres de F2D, le nouveau bilan fait état de plus de 99 % des ligands co-cristallisés reconstruits (il subsiste encore quelques problèmes marginaux de géométrie et de tautomérie non-résolus pour l'instant). Cette première étape de validation de notre outil achevée et concluante, j'ai ensuite essayé de mettre en place une méthode rapide pour connaitre la sélectivité des molécules construites par rapport aux différentes familles de protéines kinases.

4.4 Sélectivité des molécules construites

Le but de cette partie est de trouver une méthode rapide pour connaitre la sélectivité des molécules construites par F2D dans une cible par rapport aux autres kinases. Pour cela, la technique retenue est de rechercher dans quelles structures 3D on peut reconstruire les molécules tout en gardant leur même mode d'intraction. Ainsi, pour chaque composé obtenu, deux résultats sont possibles par structure, soit on parvient à le reconstruire (= actif), soit on n'y parvient pas (= inactif). Pour valider ce modèle simpliste, je me suis d'abord focalisé sur l'imatinib en partant de son groupement pyridine. J'ai ainsi analysé si F2D était capable de reconstruire l'imatinib dans les 3 092 structures humaines de protéines kinases à ma disposition. La condition pour que la reconstruction soit effective dans une structure 3D est qu'aucun atome lourd de l'imatinib ne se retrouve à une distance inférieure à 1,1 Å de tous les atomes du récepteur (cette distance est délibérément plus courte qu'une liaison hydrogène forte pour laisser une marge de tolérance)[213].

Avant d'expliquer plus en détail la méthodologie, je tiens à préciser un point particulier entre deux termes employés : la structure 3D d'une protéine kinase et sa famille. Ce que j'appelle structure dans ce paragraphe correspond bien à une représentation 3D d'une protéine obtenue expérimentalement (une structure = un code PDB). Une famille quant à elle fait référence à la sous-famille d'une protéine kinase selon la classification présentée dans l'introduction de ce manuscrit. Une famille peut donc avoir plusieurs structures la représentant. Par exemple, on trouve 59 structures obtenues par cristallographie aux rayons X de la protéine kinase appartenant à la sous-famille ABL1 chez les humains (août 2019) dans la PDB.

4.4.1 Analyse de la reconstruction de l'imatinib

L'approche a débuté par un premier regard sur le nombre de structures différentes dans lesquelles F2D est capable de construire l'imatinib : 102 (3 %). Avant même de m'intéresser

[213] George A. Jeffrey, *An Introduction to Hydrogen Bonding* (Oxford University Press, 1997).

aux familles, j'ai analysé la conformation de la région charnière et de l'hélice αC de ces 102 structures. Cette étape revêt une grande importance car l'imatinib est un inhibiteur de type II, il est donc censé être actif uniquement sur des protéines kinases en conformation DFG-out[214]. Les données pour connaitre la conformation des kinases ont été récupérées à partir de la base de données KLIFS (« Kinase-Ligand Interaction Fingerprints and Structures database »)[215]. Cette base utilise les données structurales et les positions spatiales des résidus pour classifier les conformations des structures de kinases. Le Tableau 5 récapitule les résultats de la construction de l'imatinib selon les conformations des structures.

Tableau 5 : Répartition des structures de protéines kinases dans lesquelles F2D a pu reconstruire l'imatinib selon leurs conformations. *D'après les données de la BDD KLIFS.*

	Hélice αC	Inter	Out	Out-like	Total
DFG					
In		0	9	0	9
Inter		9	0	2	11
Out		0	78	1	79
Out-like		0	3	0	3
Total		9	90	3	102

Cette première observation concorde bien avec les valeurs attendues, F2D reconstruit en grande majorité l'imatinib dans des structures en conformation DFG-out (dans 79 % des cas). Pour les 9 cas de reconstruction en conformation DFG-in, une inspection visuelle montre que l'alignement de ces structures dans le référentiel commun permet de laisser de la place à l'imatinib dans la cavité (Figure 67). En effet, comme on l'aperçoit, le motif DFG même en conformation in est en retrait par rapport à la référence en jaune. Ce recul est dû à la conformation out de l'hélice αC (aucune reconstruction de l'imatinib n'a été possible dans une protéine présentant une conformation de l'hélice αC en in).

[214] R. S.K. Vijayan et al., « Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors », *Journal of Medicinal Chemistry* 58, nº 1 (8 janvier 2015): 466-79, https://doi.org/10.1021/jm501603h.
[215] Albert J. Kooistra et al., « KLIFS: A Structural Kinase-Ligand Interaction Database », *Nucleic Acids Research* 44, nº D1 (4 janvier 2016): D365-71, https://doi.org/10.1093/nar/gkv1082.

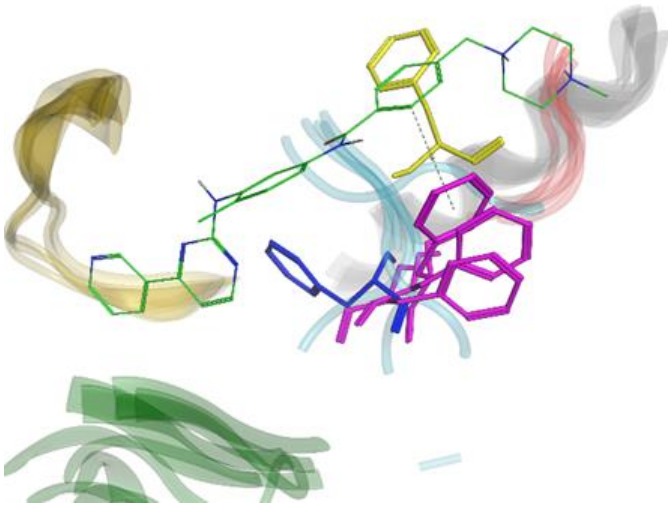


Figure 67 : Reconstruction de l'imatinib par F2D dans plusieurs protéines kinases. *En bleu la phénylalanine du motif DFG d'une kinase ABL1 en conformation DFG-out (code PDB : 2HYY), en jaune celle d'une kinase en conformation DFG-in dans laquelle l'imatinib ne peut être reconstruit (code PDB : 1ATP) et en violet celles de kinases DFG-in dans lesquelles il est reconstruit (codes PDB : 3BV3, 4EH8 et 4UMU). L'imatinib est montré en vert.*

Ces données m'ont aussi permis de vérifier dans un deuxième temps que nous reconstruisons bien l'imatinib dans toutes les structures de sa cible initiale en conformation DFG-out. Hors mutation T315I que l'on sait synonyme de résistance à ce médicament[216], il y a 8 structures d'ABL1 cristallisées sous cette conformation. Sur ces 8, F2D reconstruit l'imatinib dans 4 d'entre elles et échoue pour les autres. Pour comprendre cet échec, il faut bien réaliser que pour ce modèle de sélectivité, on programme F2D comme pour un lancement normal. Le fragment initial est le groupement pyridine provenant de la structure 2HYY, il est donc fixe. Or, il suffit d'un infime décalage dans l'alignement pour que les atomes de l'imatinib se retrouvent trop proches des atomes d'une autre kinase, empêchant la reconstruction du ligand. C'est ce qu'il se passe pour ces 4 structures d'ABL1, le groupement pipérazine terminal ne rentre pas en collision directe avec la protéine mais un de ses atomes d'azotes se rapproche à moins de 1,1 Å des atomes de ces structures. Nous verrons plus loin dans l'évaluation du modèle pourquoi ce revers n'est pas forcément problématique pour la suite. En revanche, concernant la mutation T315I, qui se situe au niveau du « gatekeeper » et bloque l'accès de l'imatinib au site actif (obstacle allostérique)[217], il y a 3 structures cristallisées la portant : 2V7A, 3QRJ et 4TWP (seule 3QRJ présente une conformation DFG-out). Dans aucune de ces trois structures F2D n'a été capable de reconstruire l'imatinib, nous pouvons donc poursuivre les investigations.

4.4.2 Recherche d'un modèle de sélectivité

L'analyse de la sélectivité se fait en étudiant la reconstitution de l'imatinib en fonction des familles de protéines kinases. Pour déterminer dans quelles familles il est actif ou non, je me suis reposé sur les données expérimentales accessibles (Davis & al.[218]). Dans cette étude les

[216] Mercedes E. Gorre et al., « Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification », *Science* 293, nº 5531 (3 août 2001): 876-80, https://doi.org/10.1126/science.1062538.
[217] Catherine Roche-Lestienne et Claude Preudhomme, « Résistance au Glivec® : actualités », *Hématologie* 13, nº 1 (1 janvier 2007): 43-53, https://doi.org/10.1684/hma.2007.0087.
[218] Mindy I. Davis et al., « Comprehensive Analysis of Kinase Inhibitor Selectivity », *Nature Biotechnology* 29, nº 11 (novembre 2011): 1046-51, https://doi.org/10.1038/nbt.1990.

auteurs ont testé l'activité de 72 inhibiteurs de protéines kinases connus, dont l'imatinib, sur un panel de 442 kinases. Ces 442 kinases représentent 379 familles uniques (comme différentes mutations sont testées pour une même protéine, elle peut être représentée plusieurs fois). Les résultats des expériences sont des valeurs de K_d en nM et nous avons fixé le seuil d'activité à 1 000 nM (ou 1 µM) pour discriminer les cibles dans lesquelles l'imatinib est actif des autres. Avec ce seuil, l'imatinib inhibe 32 kinases du panel (soit 14 familles) et est inactif sur les 410 restantes (365 familles uniques). Comme plusieurs kinases appartiennent à la même famille, elles peuvent potentiellement apparaitre dans chacun des deux sous-ensembles : actifs ou inactifs. Dans le cas de l'imatinib, seule la kinase ABL1 apparait dans chacune des catégories, notamment à cause de la présence de la forme mutée T315I dans le panel des protéines testées. Pour la suite des opérations, le choix a été pris de retirer la forme T315I pour simplifier les opérations, nous avons de toute manière vérifié auparavant que F2D ne reconstruisait pas l'imatinib dans une structure porteuse de cette mutation.

Toutes les protéines du panel de l'étude ne sont pas encore cristallisées et disponibles dans la PDB. Pour ne pas trop biaiser les métriques d'évaluation de notre méthode, nous avons décidé de travailler uniquement avec les structures présentant une conformation DFG-out, conformément au mode d'interaction de l'imatinib. Sur les 3 092 structures à notre disposition, seules 237 présentent cette conformation tout en faisant partie du panel des kinases testées expérimentalement. Le diagramme de Venn présenté Figure 68 représente de façon graphique les différentes données que je viens d'expliciter.

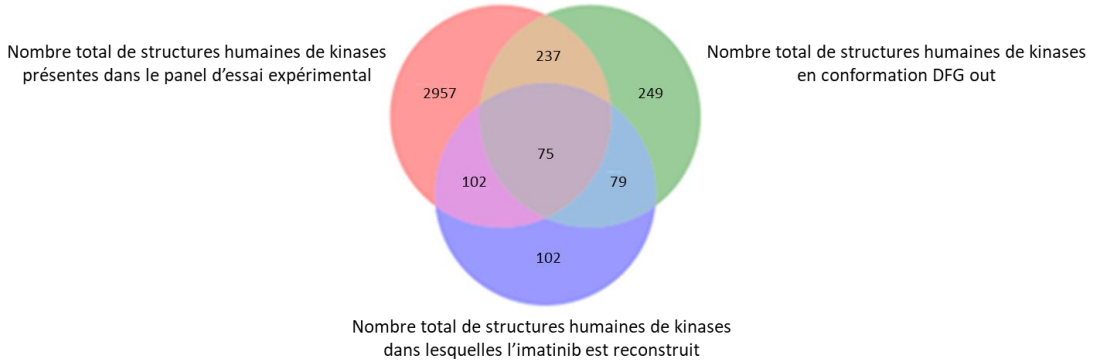


Figure 68 : Diagramme de Venn récapitulant les différentes structures de protéines kinases humaines selon leurs conformations et la reconstruction de l'imatinib.

Au total, les 237 structures 3D en conformation DFG-out représentent 50 familles de kinases du panel d'essai expérimental. Plus précisément, 30 structures correspondent à 6 familles inhibés par l'imatinib (catégorie actif) et les 207 autres à 44 familles non-inhibées (catégorie inactif). Ces données sont récapitulées dans le Tableau 6.

Tableau 6 : Récapitulatif de l'activité de l'imatinib par famille.

	Familles testées expérimentalement	Familles possédant une structure cristallisée en conformation DFG-out
Imatinib actif (K_d <= 1 µM)	14	6
Imatinib inactif (K_d > 1 µM)	365	44
Total	379	50

Les résultats de la reconstitution de l'imatinib dans les 237 structures 3D sont montrés Figure 69, les structures ayant été regroupées par catégories de familles (actif ou inactif). Pour les 6 familles appartenant à la catégorie actif, F2D parvient toujours à reconstruire l'imatinib dans au moins une structure. Pour les structures de la catégorie inactif, l'imatinib n'est jamais reconstruit dans 30 des 44 familles. Parmi les 14 familles de catégorie inactif restantes, il est reconstruit dans toutes les structures appartenant à 6 de ces familles. Pour les 8 dernières familles, il est reconstruit ou non selon les structures. On retrouve bien dans cette figure le fait que pour la moitié des structures ABL1, F2D n'est pas capable de retourner l'imatinib parmi ces résultats, comme je l'ai indiqué plus haut.

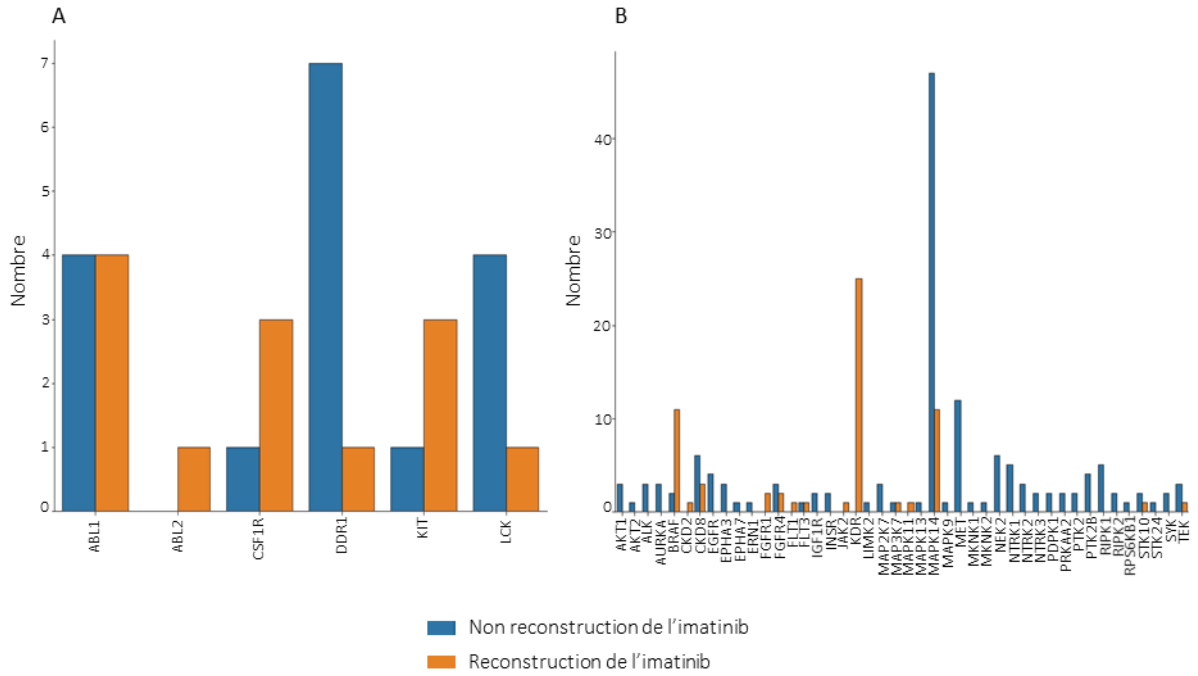


Figure 69 : Nombre de structures dans lesquelles F2D parvient à reconstruire l'imatinib ou échoue par familles de kinases. *Le graphe est séparé en deux, en familles dans lesquelles l'imatinib est actif (A) ou inactif (B) d'après les données expérimentales de l'étude de Davis & al. En bleu, F2D échoue à reconstruire l'imatinib, en orange il y parvient.*

Pour finir et évaluer ce modèle de prédiction, j'ai construit deux matrices de confusion différentes. La première matrice de confusion (Figure 70, A) reflète simplement les résultats bruts :

- Vrai positif (VP) : imatinib reconstruit dans une structure de famille active.
- Faux positif (FP) : imatinib reconstruit dans une structure de famille inactive.
- Vrai négatif (VN) : imatinib non-reconstruit dans une structure de famille inactive.
- Faux négatif (FN) : imatinib non-reconstruit dans une structure de famille active.

La seconde matrice de confusion (Figure 70, B) prend en compte une seule prédiction par famille de kinases. Pour cela, on considère comme prédit actif les familles de kinases dans lesquelles F2D a reconstruit l'imatinib dans au moins une structure. Au contraire, si F2D ne le reconstruit dans aucune des structures d'une même famille, elle sera alors prédite dans la catégorie inactif :

- Vrai positif (VP) : imatinib reconstruit dans au moins une structure de famille active.
- Faux positif (FP) : imatinib reconstruit dans au moins une structure de famille inactive.
- Vrai négatif (VN) : imatinib non-reconstruit dans aucune structure de famille inactive.
- Faux négatif (FN) : imatinib non-reconstruit dans aucune structure de famille active.

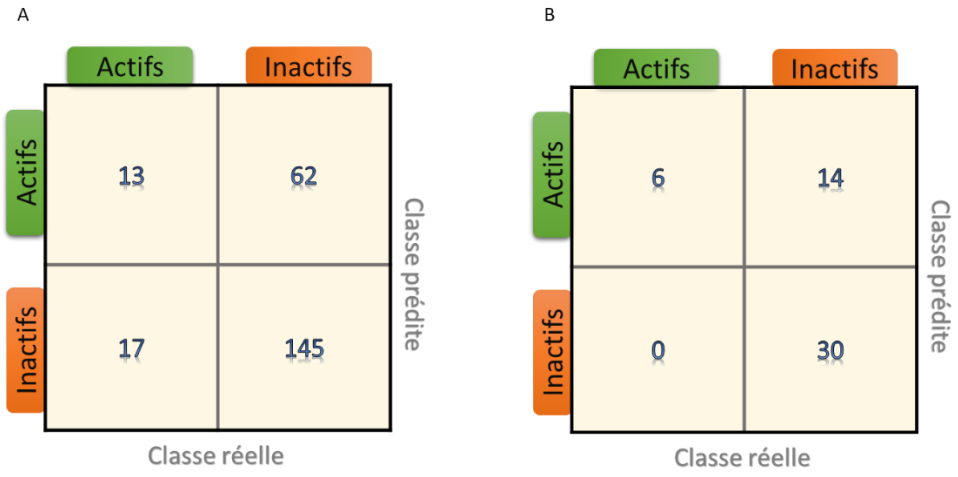


Figure 70 : Matrices de confusions des résultats de reconstruction de l'imatinib dans les protéines kinases humaines. *La matrice de confusion A reflète la reconstruction de l'imatinib dans toutes les protéines kinases humaines en conformation DFG-out, la matrice de confusion B est obtenue en regroupant les résultats par familles de kinase.*

Le Tableau 7 récapitule les différentes métriques calculées à partir de ces matrices de confusion. J'ai déjà expliqué comment calculer la sensibilité (Équation 3) et la spécificité (Équation 4) dans la partie 2.3.4 de ce manuscrit. La précision (ou valeur prédictive positive) est la proportion des vrais positifs par rapports aux positifs totaux, elle est calculée selon l'Équation 5. La précision est donc une mesure de la capacité à ne pas prédire des solutions

fausses. Le F1-score (ou F-mesure) est la moyenne harmonique entre la précision et la sensibilité. Il sert à mesurer la capacité à bien prédire toutes les vraies solutions et à refuser les autres. Il est calculé selon l'Équation 6. Enfin, l'exactitude (ou justesse) reflète la proportion des prédictions correctes par rapport au prédiction totales (Équation 7)[219].

$$Précision = \frac{VP}{VP + FP}$$

Équation 5 : Calcul de la précision. *La précision est définie comme le ratio du nombre de vrais positifs par rapport à la totalité des composants prédits positifs, avec VP : nombre de vrais positifs et FP : nombre de faux positifs.*

$$F1\text{-}score = 2 * \frac{(précison * sensibilité)}{(précision + sensibilité)}$$

Équation 6 : Calcul du F1-score. *Le F1-score est défini comme la moyenne harmonique entre la précision et la sensibilité.*

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

Équation 7 : Calcul de l'exactitude. *L'exactitude est définie comme le ratio du nombre de vraies prédictions par rapports au nombre total de prédictions, avec VP : nombre de vrais positifs, VN : nombre de vrais négatifs, FP : nombre de faux positifs et FN : nombre de faux négatifs.*

Tableau 7 : Calcul des différentes métriques associées au modèle de spécificité A et B.

	Sensibilité	Spécificité	Précision	F1-score	Exactitude
Modèle A	0,43	0,70	0,17	0,25	0,67
Modèle B	1,00	0,68	0,30	0,46	0,72

Les résultats finaux indiquent que le modèle A, qui prend en compte toutes les structures, est meilleur pour discriminer les vrais négatifs (spécificité supérieure) et donc repérer les potentielles familles où une molécule ne serait pas active. Cependant, dans toutes les autres métriques, et particulièrement la sensibilité, ce modèle présente des résultats trop faibles pour être accepté. En revanche, le modèle B (en ne prenant un compte qu'une prédiction par famille) est 100 % exact pour prédire les familles dans lesquelles un composé sera actif (vrais positifs) et présente une spécificité tolérable. Globalement, le modèle B est supérieur au modèle A avec une exactitude de 0,72. Cependant ce modèle peut encore être amélioré. En effet, cette métrique ne peut suffire à refléter les résultats car dans notre cas les classes sont déséquilibrées (il y a beaucoup plus de familles dans la catégorie des inactifs que dans celle des actifs)[220]. La précision et le F1-score présentent des valeurs trop basses pour les deux modèles. Pour en être sûr, j'ai réalisé ce même protocole à tous les inhibiteurs de kinase co-cristallisés testés expérimentalement sur le panel de kinases (sur les 72 il y en a 34). Toutefois, les résultats

[219] Aditya Mishra, « Metrics to Evaluate Your Machine Learning Algorithm », Medium, 1 novembre 2018, https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.
[220] « Classification : justesse | Cours d'initiation au machine learning », Google Developers, consulté le 24 septembre 2019, https://developers.google.com/machine-learning/crash-course/classification/accuracy.

obtenus avec les autres ligands sont similaires à ceux présentés ici avec l'imatinib. A l'heure actuelle, ce modèle pour calculer la sélectivité potentielle des molécules construites avec F2D a donc encore besoin d'être affiné avant d'être intégré définitivement. Il existe encore cependant de nombreuses difficultés avant d'y parvenir car les familles de protéines kinases ne disposent pas toutes de structures cristallisées, empêchant la vérification par reconstruction de F2D dans de telles familles. A titre informatif, l'option permettant de connaitre rapidement pour chaque composé retourné par F2D, dans quelle structure disponible, l'outil est capable de le reconstruire ou non est déjà implémentée. Ce résultat se présente sous la forme d'une carte de chaleur avec en abscisse les molécules retournées par F2D et en ordonnées toutes les structures humaines des kinases. La case croisant une structure de protéine kinase avec une molécule est alors coloriée en bleu si cette molécule peut être reconstruite dans le site actif de la protéine kinase, en rouge le cas échéant.

4.5 Traitement post-résultats

J'ai implémenté plusieurs méthodes afin de trier les résultats et de ne conserver que les molécules présentant les meilleures caractéristiques pour procéder à leur future synthèse en vue de tests expérimentaux. Comme F2D peut retourner plusieurs dizaines de milliers de propositions, il est nécessaire de filtrer ces résultats avec des méthodes efficaces et rapides que je vais maintenant présenter.

4.5.1 Filtrage physico-chimique

Comme annoncé dans l'avant-propos de ce manuscrit, l'équipe SB&C de l'ICOA est spécialisée dans les protéines kinases. De fait, notre expérience dans le domaine et nos études sur les inhibiteurs de protéines kinases nous ont amenés à la création de nouveaux filtres spécifiques pour cette famille de protéines. Les filtres dénommés « kinase-like » sont issus de la publication de F. Carles et al.[221], remis à jour avec les nouveaux inhibiteurs de kinase découverts entre temps et que l'on retrouve dans la publication du chapitre 1 (partie 1.7). Ces filtres sont récapitulés dans le Tableau 8.

Tableau 8 : Filtres physico-chimiques « kinase-like »

	Minimum	Maximum
MW (Da)	312	614
CLogP	0,6	6,3
HBA	3	10
HBD	0	4
NRB	1	11
TPSA ($Å^2$)	55	138

Ces différents descripteurs sont calculés après construction des molécules à l'aide la bibliothèque RDKit. Une molécule ne sera pas conservée si elle viole un de ces critères. A noter

[221] Fabrice Carles et al., « PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials », *Molecules* 23, nº 4 (15 avril 2018): 908, https://doi.org/10.3390/molecules23040908.

que d'autres filtres peuvent être implémentés par l'utilisateur et que les seuils peuvent être aisément modifiés au besoin.

4.5.2 Filtre par sous-structures

Parmi les molécules créées par F2D et passant les premiers filtres physico-chimiques mis en place, certains groupements chimiques peuvent s'avérer indésirables (notamment pour la synthèse ou la stabilité). Les composés présentant ces groupements chimiques peuvent être filtrés à l'aide d'une recherche par sous-structure avec la notation SMARTS[222]. Le langage SMARTS (« SMILES arbitrary target specification ») a été développé dans le but de retrouver des sous-structures précises dans une molécule. Sa notation permet un encodage très clair définissant exactement le groupement ou les atomes recherchés spécifiquement. Les atomes peuvent ainsi faire l'objet d'une description de leur valence, leur hybridation, mais aussi de leurs atomes voisins. On peut donc aisément créer une requête pour trouver tous les atomes carbones aliphatiques reliés à un cycle d'un côté et à un atome d'azote sp2 de l'autre par exemple. Le nombre d'atomes d'hydrogène reliés à un atome lourd peut lui aussi être explicité. Les requêtes SMARTS sont utilisées notamment dans les règles de rétrosynthèse RECAP pour retrouver des types de liaisons précis[223]. Des exemples de groupements chimiques indésirables sont illustrés Figure 71 avec les requêtes SMARTS associées.

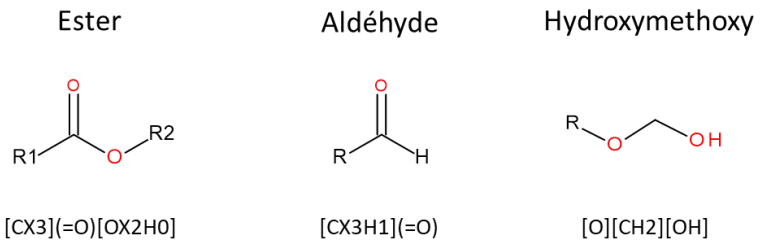


Figure 71 : Exemple de requêtes SMARTS pour identifier des groupements chimiques.

Les requêtes SMARTS sont des outils puissants et très rapides, chaque utilisateur peut créer sa propre requête personnalisée. Dans notre cas on se sert des requêtes SMARTS pour éliminer des composés indésirables mais bien évidemment on peut aussi les utiliser pour au contraire localiser un composé spécifique portant le groupement chimique recherché.

4.5.3 Amarrage moléculaire

Comme F2D n'applique pas de minimisation ou d'optimisation de conformation des molécules dans la cavité, l'amarrage moléculaire va nous servir à vérifier et valider le mode d'interaction du ligand dans le site actif. J'ai expliqué en détail en quoi consiste l'amarrage moléculaire dans le chapitre précédent, je ne vais donc pas revenir sur le procédé ici. Chaque molécule créée est ainsi re-amarrée dans le site actif de la protéine et les poses obtenues sont comparées à la pose originale donnée par F2D.

---

[222] « Daylight Theory: SMARTS - A Language for Describing Molecular Patterns », consulté le 12 juin 2019, https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35.

[223] Xiao Qing Lewell et al., « RECAP Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry », *Journal of Chemical Information and Computer Sciences* 38, nº 3 (18 mai 1998): 511-22, https://doi.org/10.1021/ci970429i.

Le choix du logiciel pour réaliser l'amarrage s'est porté sur rDock (*version 2013.1)* car il s'agit d'un outil libre d'accès et facile d'utilisation. De plus, rDock peut s'installer aisément à l'aide d'Anaconda, comme une bibliothèque du langage Python classique et nous pouvons facilement le combiner avec nos données et le lancer depuis un notebook. Il s'intègre donc très bien dans le projet F2D. A noter que cet outil est aussi utilisé en routine dans notre laboratoire pour de nombreux autres projets. rDock a été développé par l'entreprise Vernalis, il est écrit majoritairement en C++ et Fortran. Il repose sur un algorithme génétique et est capable de faire du docking semi-flexible (le récepteur ne subit aucune transformation mais le ligand fait l'objet de changement de conformation)[224]. Pour fonctionner rDock a besoin de la protéine cristallisée préparée (protonation des acides aminés, vérification des tautomères…), dans notre cas avec le logiciel MOE (Chemical Computing Group, *version 2016.0802*), en format MOL2. Puis il calcule lui-même la cavité d'un site actif, soit à l'aide du ligand original, soit à l'aide de coordonnées spatiales. Les composés à amarrer doivent eux aussi être fournis préparés, en format standard SDF. Ici, pour ne pas biaiser l'amarrage moléculaire, nous réinitialisons les ligands obtenus avec F2D en leur donnant une nouvelle conformation et de nouvelles coordonnées aléatoires. Vu le nombre de résultats possibles retournés par F2D (plusieurs milliers ou plus), nous avons paramétré 6 poses par ligand, les auteurs estimant que 5 poses suffisent à filtrer rapidement les molécules grâce au score fourni. Le score choisi est le S_{inter}, qui traduit les interactions récepteur-ligands. Une fois l'amarrage moléculaire réalisé, le RMSD (Équation 1) entre les poses obtenues et la pose originale de F2D est calculé. Si aucune des 6 poses ne possède un RMSD inférieur à 2 Å avec la pose originale, la molécule est alors signalée à l'utilisateur. Deux exemples de molécules créées avec F2D (en rouge) et amarrées (en bleu) sont présentés Figure 72.

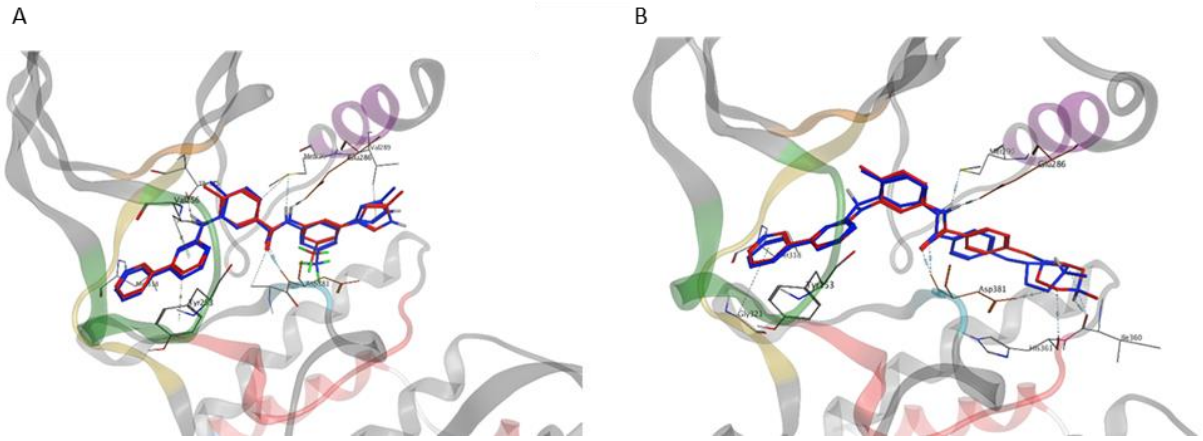


Figure 72 : Comparaison entre les molécules construites avec F2D (en rouge) et les mêmes molécules amarrées avec rDock (en bleu). *Le composé A présente un RMSD de 0,59 Å avec son homologue amarré et le B un RMSD de 0,89 Å.*

4.5.4 Calcul du SA_Score

Le SA_Score (« Synthetic Accessibility Score ») est un score de 0 à 10 ayant pour but d'évaluer la difficulté de synthèse de composés. Plus le score est faible, plus la molécule sera facile à synthétiser, plus il s'approche de 10, plus cette synthèse sera compliquée. Il a été

[224] Sergio Ruiz-Carmona et al., « RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids », *PLOS Computational Biology* 10, nº 4 (10 avril 2014): e1003571, https://doi.org/10.1371/journal.pcbi.1003571.

développé par P. Ertl & A. Schuffenhauer[225]. Le calcul se fait à partir d'une combinaison entre plusieurs descripteurs et des fragments de la molécule plus ou moins pénalisants : les groupements chimiques non-standard, la stéréocomplexité, ou encore le poids moléculaire vont être des facteurs importants. Un point intéressant de l'étude est la comparaison entre leur SA_Score et un score donné par des chimistes médicinaux et de synthèse sur un ensemble de 40 molécules. Les résultats montrent une très bonne concordance avec un coefficient de détermination r^2 de 0,89.

Le SA_Score est donc une option qui peut être prise en compte pour sélectionner des molécules parmi les résultats de F2D. Il est aussi implémenté en langage Python et est très rapide à calculer (une à deux minutes pour plusieurs milliers de molécules). C'est un critère important car il permet de donner rapidement une idée de la difficulté ou non d'un composé à être synthétisé, chose dont on ne se rend pas forcément compte à l'œil nu sans examiner manuellement toutes les molécules. La Figure 73 présente des composés issus de résultats obtenus en lançant F2D à partir du groupement pyridine de l'imatinib. La molécule présentant le SA_Score le plus haut parmi ces résultats est l'avant dernière. La dernière est quant à elle la molécule présentant le plus haut score mais ce, avant filtration par les filtres kinases-like. La moyenne des résultats du SA_Score pour toutes les molécules construites et filtrées à partir de la pyridine dans la structure PDB 2HYY est de 2,92, les multiples restrictions et filtres mis en place permettent donc bien d'éviter au maximum de retourner des molécules impossibles à synthétiser. A titre de comparaison, sans les filtres, la moyenne du SA_Score est de 3,52 et la moyenne du SA_Score des inhibiteurs de kinase présent dans la BDD PKIDB est de 3,27.

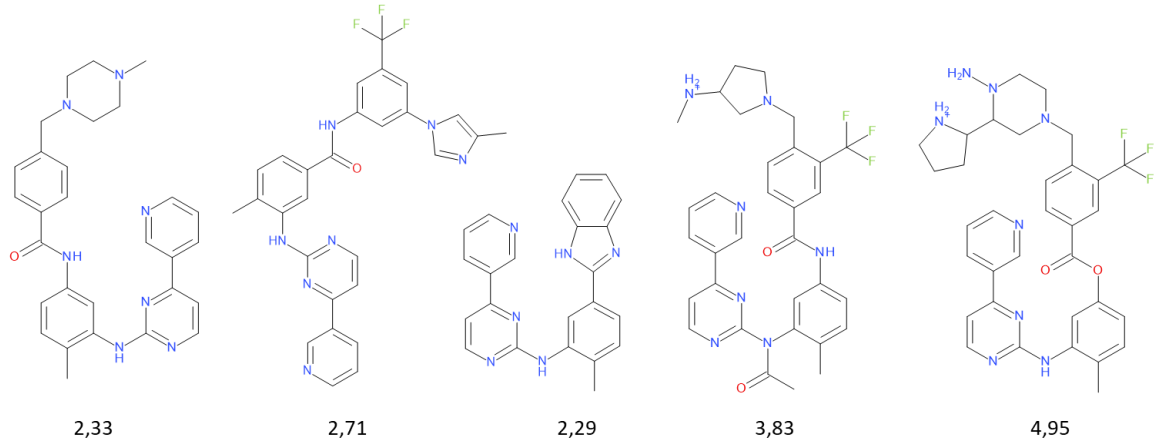


Figure 73 : Exemples de composés obtenus avec F2D avec les SA_Score correspondants. *La cinquième molécule, à droite, est une molécule trouvée avant filtration des résultats.*

4.5.5 Projections ACP et PMI

L'Analyse en Composantes Principales (ACP) et celles des Principaux Moments d'Inertie (PMI) permettent d'apprécier l'espace chimique couvert par les composés obtenus. Plus précisément, l'ACP va se focaliser sur des descripteurs moléculaires classiques, tandis que le PMI est exclusivement indicateur de la forme 3D des molécules (cf. publication du chapitre 1, partie 1.7). Grâce à ces projections, l'utilisateur peut rapidement découvrir si les molécules fournies par F2D s'éloignent ou non de l'espace déjà couvert par les inhibiteurs de protéines

[225] Peter Ertl et Ansgar Schuffenhauer, « Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions », *Journal of Cheminformatics* 1, nº 1 (10 juin 2009): 8, https://doi.org/10.1186/1758-2946-1-8.

kinases connus. En d'autres termes, l'ACP et le PMI sont utiles pour juger rapidement de l'originalité des composés. Il s'agit d'un indice important, particulièrement lors de la recherche de nouveaux composés intéressants car la plupart des squelettes moléculaires sont déjà protégés par des brevets[226] et il est difficile de pouvoir valoriser une molécule si celle-ci fait déjà l'objet d'une protection intellectuelle.

Grâce à l'usage des notebooks et de leur affichage interactif, il suffit de passer la souris par-dessus un point projeté correspondant à une molécule pour avoir directement le visuel de celle-ci. Cela évite ainsi une perte de temps pour la retrouver en fouillant manuellement le fichier SDF. L'ACP et le PMI sont calculés comme décrit dans le papier « PKIDB2 » (partie 1.7). La Figure 74 présente un exemple obtenu avec les 49 857 molécules retournées par F2D en lançant le programme à partir d'un groupement tétrahydroisoquinoline reliée à la structure charnière dans une structure MELK (code PDB : 4D2P). Sur cette figure, les composés retournés par F2D en violet sont comparés aux inhibiteurs de kinase de PKIDB en bleu. Avec l'ACP nous constatons qu'une part conséquente de ceux-ci sont projetés hors de l'ellipse de confiance des PKI englobant 95 % des individus, présentant ainsi potentiellement un intérêt pour en faire un inhibiteur original. Les résultats sont moins impressionnants sur la représentation du PMI où les molécules de F2D sont majoritairement projetées sur l'axe « barre-disque », autrement dit des formes plutôt plates et allongées. Si F2D n'a pu construire aucun composé de forme sphérique (coin droit vide), cela s'explique par le choix du fragment initial lié à la région charnière forçant l'agrandissement en longueur vers l'intérieur du site. De plus, comme on le voit sur la figure, aucun inhibiteur de PKIDB ne possède non plus de forme sphérique car la cavité du site actif des protéines kinases ne présente pas cette forme.

[226] Irini Akritopoulou-Zanze et Philip J. Hajduk, « Kinase-targeted libraries: The design and synthesis of novel, potent, and selective kinase inhibitors », *Drug Discovery Today* 14, nº 5 (1 mars 2009): 291-97, https://doi.org/10.1016/j.drudis.2008.12.002.

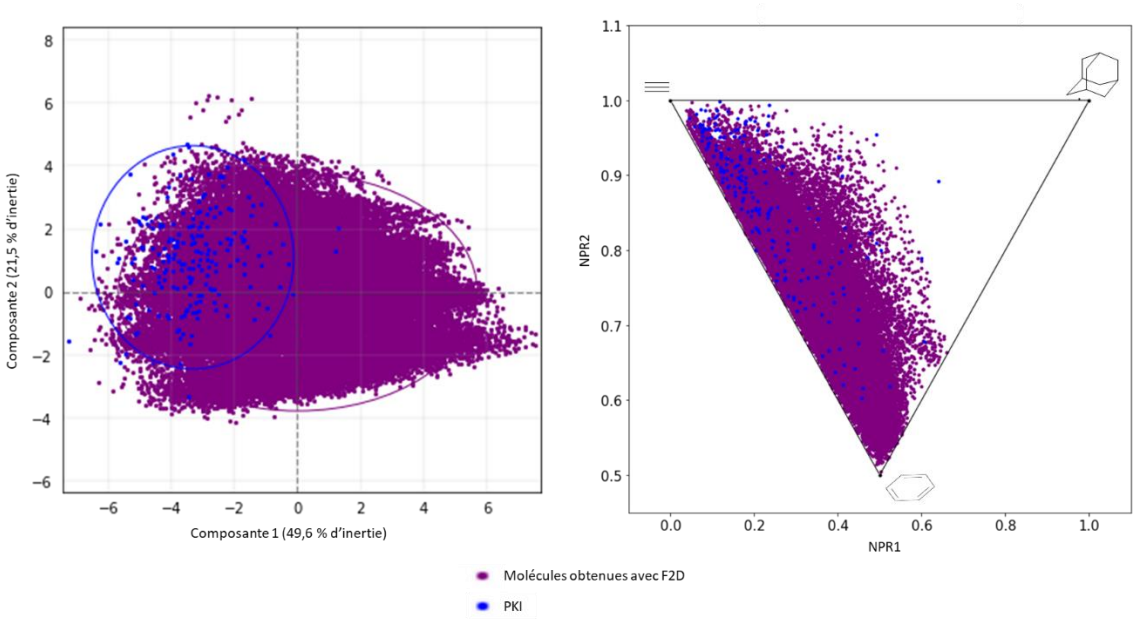


Figure 74 : Exemple d'une projection d'ACP (gauche) et d'une projection des PMI (droite) calculés à partir des résultats de F2D.

4.5.6 Couplage avec la ChEMBL

Un des réflexes courant d'un chémoinformaticien lorsqu'il dispose d'une molécule est de chercher dans une BDD de bioactivité si celle-ci a déjà été synthétisée et testée (dans notre cas il s'agit de la ChEMBL). Quand on ne dispose que d'une dizaine de molécules, on peut se permettre d'aller vérifier cela manuellement directement sur le site de la BDD, soit en dessinant la molécule, soit à l'aide de sa représentation SMILES. Cependant, à partir de plusieurs centaines de composés ce travail devient fastidieux et insensé et il faut bien entendu automatiser le processus. Pour cela, la ChEMBL a développé une interface de programmation (API) sous la forme d'un module en langage Python. Cette initiative permet de faire des requêtes directement depuis le notebook en langage Python permettant à l'utilisateur de se dérober à l'apprentissage du langage SQL, mais aussi de facilement l'intégrer dans ses projets. La prise en main est très facile et les possibilités similaires à celles que l'on a en allant directement sur le site. Seulement, en passant par cette API, les requêtes s'effectuent toujours sur le serveur distant de la ChEMBL il y a donc un laps de temps non-négligeable à notre niveau pour recevoir les résultats. De plus, ces API n'apprécient guère les boucles de répétition de requêtes et finissent souvent par les rejeter, bannissant l'utilisateur pour un certain temps et stoppant le programme. Cette option a donc été abandonnée et une autre solution a été mise en place.

La ChEMBL est une BDD qui propose son téléchargement intégral sur un ordinateur personnel afin de pouvoir travailler dessus en version locale et s'affranchir des difficultés énoncées juste au-dessus. J'ai donc téléchargé la version courante de la ChEMBL (*ChEMBL24.1, septembre 2018*) et implémenté le « cartridge » de RDKit me permettant de travailler sur cette BDD locale à l'aide de requêtes en langage Python directement depuis le notebook (https://www.rdkit.org/docs/Cartridge.html).

Ainsi, pour chaque molécule construite avec F2D, une recherche par similarité est effectuée dans la ChEMBL. Cette recherche se fait à l'aide des empreintes moléculaires

circulaires de Morgan[227] et de l'indice de Tanimoto. Elle permet de retrouver pour chaque molécule si celle-ci, ou la molécule la plus similaire le cas échéant, possède des données de bioactivité. Si oui, les cibles connues pour la molécule considérée et les résultats des tests sont directement fournis à l'utilisateur sous forme de tables et de graphiques (Figure 75).

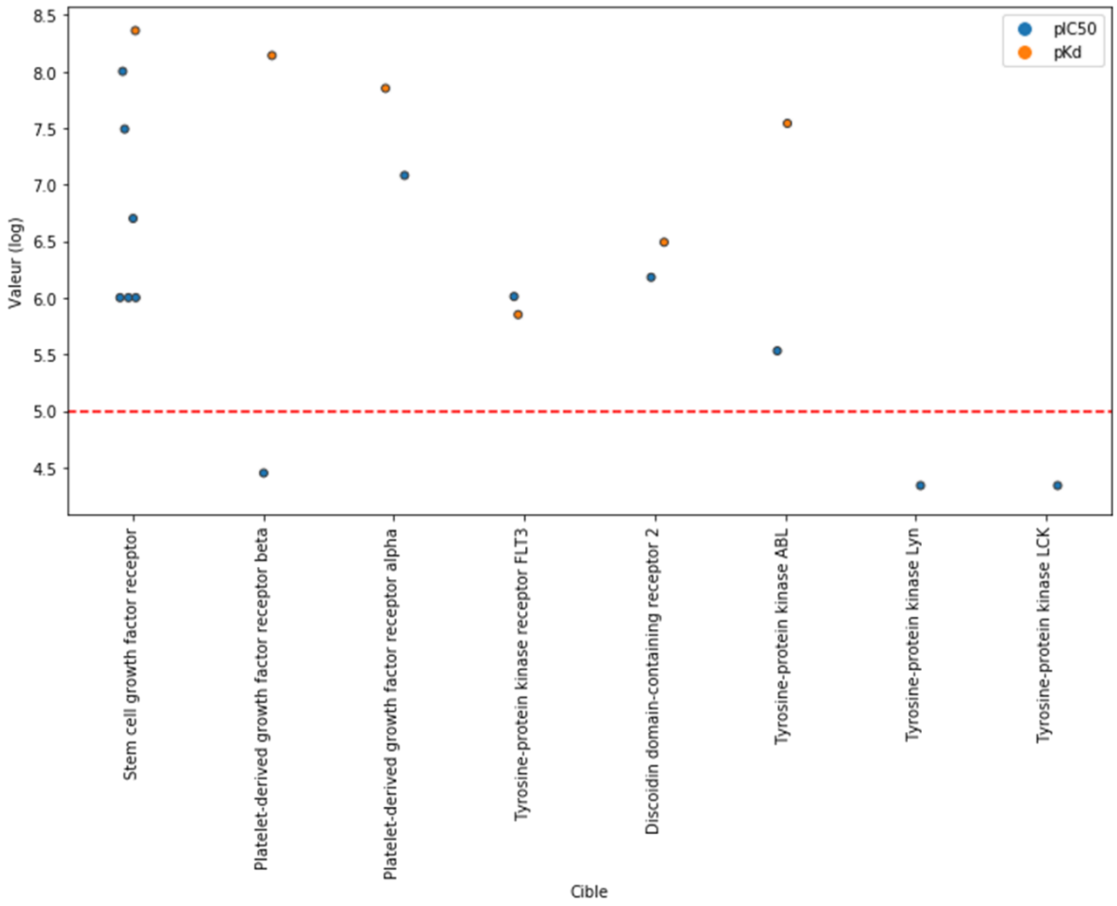


Figure 75 : Graphique résumant les différents tests d'une molécule obtenue avec F2D et répertoriée dans la ChEMBL.

De manière générale concernant le post-traitement des résultats de F2D, comme le fichier rendu est en format standard SDF, l'utilisateur peut tout à fait utiliser les principaux outils de chémoinformatique pour rechercher des molécules intéressantes par ses propres moyens. Les options présentées dans ce manuscrit sont implémentées pour faciliter l'usage à un non-initié et accélérer l'analyse des molécules. Un utilisateur averti pourra disposer des résultats à sa guise et les traiter avec un autre logiciel ou un autre langage s'il le souhaite.

4.6 Dernières améliorations

Lors de ma dernière année de thèse, j'ai eu l'occasion de recruter et d'encadrer un stagiaire en troisième année de licence informatique (T. Misiek) pour m'aider à optimiser encore F2D, notamment grâce à sa formation en programmation et en algorithmique. Son apport a permis d'améliorer la vitesse d'exécution du programme grâce à plusieurs modifications.

[227] David Rogers et Mathew Hahn, « Extended-Connectivity Fingerprints », *Journal of Chemical Information and Modeling* 50, nº 5 (24 mai 2010): 742-54, https://doi.org/10.1021/ci100050t.

Ainsi, l'architecture globale a été transformée par la création de plusieurs objets propres là où auparavant nous nous contentions d'utiliser des listes ou des dictionnaires. L'algorithme de parcours du graphe a subi des changements avec l'implémentation d'appels récursifs plus poussés au fur et à mesure de la construction des molécules. Dorénavant, F2D est fondé sur notre propre objet graphe et non plus sur celui du module Networkx. Dans cet objet chaque nœud fragment possède en attribut sa liste de voisins, son poids moléculaire et la liste des fragments en exclusion avec lui. Le nœud de départ reste le centre névralgique qui va recevoir toutes les informations des autres nœuds pour créer les molécules dans un attribut résultat que lui seul porte. Cette nouvelle méthode fonctionne avec des paquets d'information échangés par les nœuds. Ce que nous appelons un paquet est ici un objet spécifiquement créé contenant toutes les informations nécessaires : le chemin emprunté pour retourner à la racine, les nœuds déjà visités, l'addition du poids moléculaire de tous les fragments à l'instant t et les valences actualisées des différents atomes impliqués dans les liaisons. A chaque fois qu'un nœud reçoit le paquet, selon les cas il renvoie le paquet actualisé à l'envoyeur ou à ses voisins. Ce nouvel algorithme est donc entièrement fondé sur une communication poussée entre nœuds avec trois actions principales réalisées par chaque nœud : recevoir, actualiser, retourner. Ainsi, là où avant nous nous contentions d'un dictionnaire actualisé au fur et à mesure du parcours et de vérification systématique des relations avec le GxE, désormais ces étapes se font directement par l'échange d'informations constant entre chaque fragment voisin.

En plus de l'algorithme général, nous avons aussi été en mesure de créer notre propre BDD orientée graphe directement en langage Python, évitant ainsi d'avoir recours à l'outil extérieur Neo4J. Cette fois encore, cela va dans la volonté d'harmoniser nos programmes et d'éviter au maximum l'emploi d'outils extérieurs plus compliqués à installer et prendre en main. Dorénavant, notre librairie de fragments est stockée sous forme de fichiers binaires (format « pickle ») permettant une ouverture et une lecture très rapide des relations et ainsi une sélection du sous-graphe pour lancer F2D beaucoup plus efficace qu'auparavant. Par exemple, le temps de recherche de la liste des voisins d'un fragment dans la BDD, de l'ordre de la seconde avec Neo4J est maintenant de l'ordre de la milliseconde avec nos fichiers pickles. Le Tableau 9, extrait du rapport de stage de T. Misiek, récapitule les différences de temps d'exécution de la nouvelle version développée comparée à l'ancienne, pour chaque étape clé de F2D. Comme on le voit dans ce tableau, l'amélioration de F2D est ressentie à chaque phase, de la création du GxI à la sauvegarde des résultats. Le nombre total de molécules retournées est le total brut (avant l'utilisation de filtres et l'étape de regroupement). La différence de nœuds dans le GxI observée à partir de l'itération 3 résulte de la recherche plus précise de relations E dans une branche, non plus uniquement avec le fragment initial mais aussi avec les autres fragments de cette branche. Par exemple, sur une branche de 4 fragments (1-2-3-4) si le fragment 4 présente une relation E avec le 2, dans la nouvelle version il ne sera pas conservé car il est inutile puisqu'il ne pourra être ajouté à la molécule finale. Tandis que dans l'ancienne version, comme nous vérifiions uniquement pour la création du GxI, les relations E avec le fragment initial, le fragment 4 est conservé dans le GxI. Cependant, lors de la recherche des combinaisons possibles, comme la condition qu'il n'y ait pas de relation E entre fragments pour créer la molécule est évidemment vérifiée, ce fragment n'était pas utilisé. La différence de molécules créées observée est-elle due à un bogue dans l'ancienne version : dans certains cas l'hybridation des atomes au fur et à mesure de la construction d'une molécule n'était pas bien actualisée, deux fragments se retrouvant sur un même atome qui ne pouvait supporter qu'une liaison. Or, lors de la construction d'une telle molécule, comme RDKit ne parvenait pas à ajouter deux fragments sur un même atome, il ne prenait pas en compte le dernier fragment ajouté. Par exemple, ces 3 combinaisons de fragments représentant une molécule (1-2-3, 1-2-4 et 1-2-5) donnaient en réalité les 3 molécules suivantes : 1-2, 1-2, 1-2, car les fragments 3, 4 et 5 sont reliés au fragment 2 sur le même atome que celui-ci est relié au fragment 1, rendant impossible

leur ajout (Figure 76). A noter que bien que ce bogue est problématique car il augmente artificiellement les molécules créées et donc le temps de calcul, dans les résultats finaux il ne se voit pas car tous les doublons sont éliminés avec l'étape de regroupement avant de retourner les résultats à l'utilisateur.

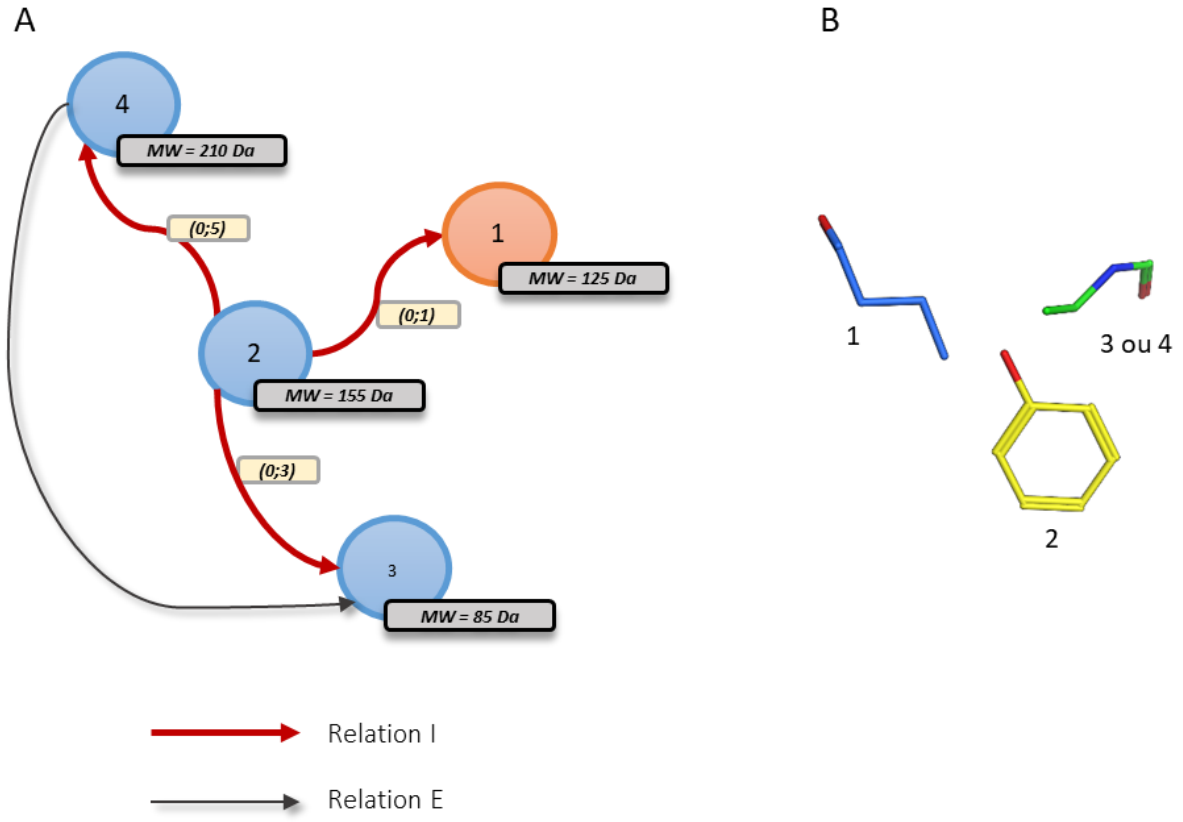


Figure 76 : Graphe montrant un exemple de problème d'actualisation des hybridations atomiques au fur et à mesure de la construction dans F2D (A) et représentation en 3D des fragments (B). *Dans le cas présent, F2D proposait comme combinaison (1, 1-2, 1-2-3 et 1-2-4). Or, l'atome 0 du fragment 2 ne peut supporter qu'un seul ajout de fragment, les molécules 1-2-3 et 1-2-4 sont donc impossibles à construire par RDKit.*

Toutefois, ces différences de temps d'exécution du programme entre les versions ne sont pas précisément quantifiables car cela dépend de beaucoup de paramètres d'entrées spécifiés lors du lancement de F2D (fragment initial, cavité et atomes sur lesquelles grandir). Pour l'exemple donné dans le Tableau 9, le fragment initial est un groupement adénine originaire de la structure PDB 4NST, sans précision d'atomes sur lesquels grandir.

Tableau 9 : Différences de performance entre l'ancienne version de F2D et la dernière version.

Itération*	Version avec Neo4J	Version avec Python
2	• Création GxI : 21 secondes (94 nœuds) • Création GxE : 0,41 seconde • Parcours : 0,91 seconde • Création des molécules : 0,1 seconde • Sauvegarde : 0,3 seconde **Total : 22 secondes (340 molécules créées)**	• Création GxI : 0,02 seconde (94 nœuds) • Création GxE : 0,04 seconde • Parcours : 0,03 seconde • Création des molécules : 0,03 seconde • Sauvegarde : 0,75 seconde **Total : 1 seconde (137 molécules créées)**
3	• Création GxI : 78 secondes (568 nœuds) • Création GxE : 6,6 secondes • Parcours : 27 secondes • Création des molécules : 4,4 secondes • Sauvegarde : 4,4 secondes **Total : 121 secondes (4 409 molécules créées)**	• Création GxI : 0,07 seconde (288 nœuds) • Création GxE : 0,12 seconde • Parcours : 0,19 seconde • Création des molécules : 0,6 seconde • Sauvegarde : 1,27 seconde **Total : 2 secondes (1 784 molécules créées)**
4	• Création GxI : ≈ 6 minutes (1 065 nœuds) • Création GxE : 16 secondes • Parcours : ≈ 2 minutes • Création des molécules : 8 secondes • Sauvegarde : 7 secondes **Total : ≈ 9 minutes (24 723 molécules créées)**	• Création GxI : 0,24 seconde (458 nœuds) • Création GxE : 0,22 seconde • Parcours : 1,26 seconde • Création des molécules : 4,96 secondes • Sauvegarde : 4,2 secondes **Total : 11 secondes (12 907 molécules créées)**
Limite de poids (< 650 Da)	• Création GxI : ≈ 17 minutes (1224 nœuds) • Création GxE : 33 secondes • Parcours : ≈ 3 minutes • Création des molécules : 21 secondes • Sauvegarde : 18 secondes **Total : ≈ 22 minutes (33 500 molécules créées)**	• Création GxI : 1,32 seconde (557 nœuds) • Création GxE : 0,29 seconde • Parcours : 2,30 secondes • Création des molécules : 9,08 secondes • Sauvegarde : 6,81 secondes **Total : 20 secondes (21 292 molécules créées)**

*Nombre de fragments ajoutés au fragment initial

Finalement, pour obtenir plus de nouveautés, et grâce aux améliorations du programme, la fragmentation des ligands co-cristallisés a aussi été mise à jour : les cycles fusionnés sont maintenant fragmentés et les substituants d'un cycle (halogènes, groupement méthane, …) sont aussi gérés seuls. Les possibilités de création de molécules sont donc encore plus grandes qu'avant avec cette librairie de fragments augmentée (environ 55 000 fragments) et les atomes seuls qui peuvent se nicher dans des recoins de la cavité pour concevoir ainsi un inhibiteur « sur mesure ».

4.7 Exemples d'applications et de résultats

Durant ma thèse, F2D a été utilisé dans de nombreux projets en interne ou en collaboration avec d'autres laboratoires. Parmi les cibles visées lors de ces applications, nous pouvons citer BRAF (cf. thèse de J.-M. Gally), Pim1, DYRK[228] ou encore ABL1. Dans ce manuscrit je me focaliserai sur une application possible spécifique de F2D, la construction de macrocycles, et sur les résultats obtenus et validés expérimentalement sur la protéine kinase ABL1.

4.7.1 Création de macrocycles

Les macrocycles sont des structures moléculaires qui font déjà l'objet de plusieurs applications thérapeutiques dans de nombreux domaines (ADN, ARN, RCPGs, interactions protéine-protéine, etc)[229]. De par leur taille et leur forme, ils présentent des propriétés intéressantes, notamment pour cibler des sites actifs nécessitant des inhibiteurs imposants[230]. Pour les protéines kinases, comme je l'ai écrit dans le premier article (partie 1.7), plusieurs macrocycles sont en phases cliniques ou approuvés comme par exemple le lorlatinib. Le lorlatinib est le premier macrocycle approuvé comme inhibiteur de kinase, en 2018. Il est utilisé pour traiter le cancer du poumon non à petites cellules, ALK-positif et cible les kinases ALK et ROS1[231]. Les macrocycles représentent donc une famille de structures potentiellement prometteuses et l'objectif ici est de voir si F2D demeure capable de créer de tels composés dans le site actif à partir d'un fragment initial précis. Un filtre spécifique a été mis en place pour ne retrouver que les macrocycles parmi l'ensemble de composés retournés (nous avons défini un macrocycle comme un cycle composé de plus de 12 atomes[232,233]). Les macrocycles sont ensuite filtrés par les filtres kinase-like comme n'importe quel autre résultat de F2D. Comme illustré Figure 77, plusieurs fragments initiaux ont été testés, en précisant un ou plusieurs atomes

[228] Florence Couly et al., « Development of Kinase Inhibitors via Metal-Catalyzed C–H Arylation of 8-Alkyl-Thiazolo[5,4-f]-Quinazolin-9-Ones Designed by Fragment-Growing Studies », *Molecules* 23, nº 9 (29 août 2018): 2181, https://doi.org/10.3390/molecules23092181.
[229] Fabrizio Giordanetto et Jan Kihlberg, « Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? », *Journal of Medicinal Chemistry* 57, nº 2 (23 janvier 2014): 278-95, https://doi.org/10.1021/jm400887j.
[230] Philipp Ermert, « Design, Properties and Recent Application of Macrocycles in Medicinal Chemistry », Text, octobre 2017, https://doi.org/info:doi/10.2533/chimia.2017.678.
[231] Center for Drug Evaluation and Research, « FDA Approves Lorlatinib for Second- or Third-Line Treatment of ALK-Positive Metastatic NSCLC », *FDA*, 2 septembre 2019, http://www.fda.gov/drugs/fda-approves-lorlatinib-second-or-third-line-treatment-alk-positive-metastatic-nsclc.
[232] Andrei K. Yudin, « Macrocycles: Lessons from the Distant Past, Recent Developments, and Future Directions », *Chemical Science* 6, nº 1 (2015): 30-49, https://doi.org/10.1039/C4SC03089C.
[233] Christian Heinis, « Drug Discovery: Tools and Rules for Macrocycles », *Nature Chemical Biology* 10, nº 9 (septembre 2014): 696-98, https://doi.org/10.1038/nchembio.1605.

possibles de départ. Dans la majorité des situations de départ, F2D est capable de construire des macrocycles (d'une dizaine de composés à plusieurs milliers, selon les cas).

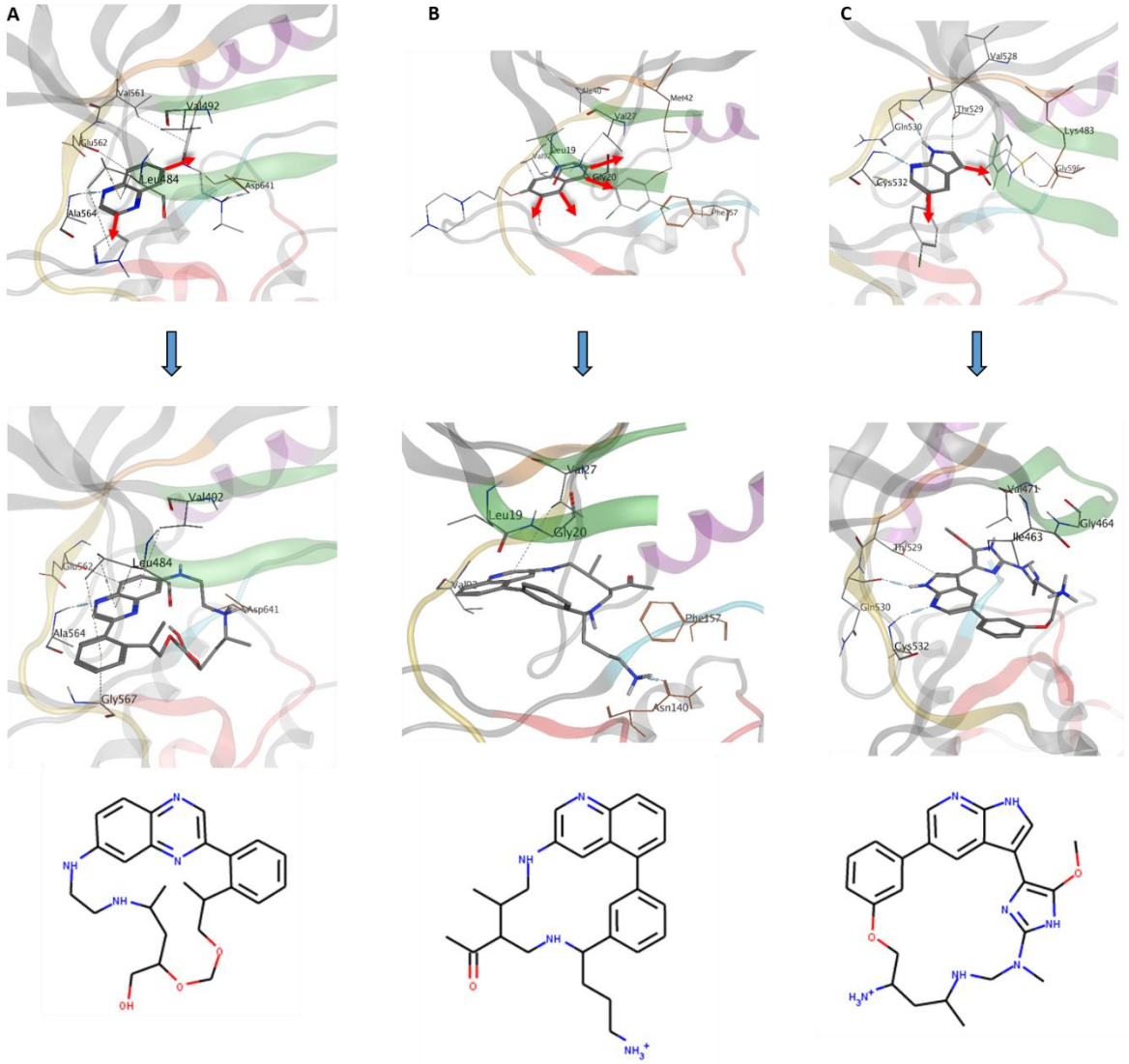


Figure 77 : Exemple de constructions de macrocycles à partir de différents fragment initiaux. *Dans chaque cas, la première figure montre le ligand dans son site actif avec le fragment de départ en gras. Le fragment A correspond à un groupement quinoxaline (code PDB : 5EW8), le fragment B un groupement 7-azaindole (code PDB : 3OG7) et le fragment C un groupement quinoline (code PDB : 3SOA), les flèches rouges indiquent les différents atomes de départ pour chaque fragment.*

Comme on l'observe sur les figures, les macrocycles construits par F2D sont très bien ancrés dans le site actif réalisant de multiples interactions avec les résidus de la cavité. Les valeurs des différents descripteurs moléculaires calculées sur ces 3 macrocycles sont exposés dans le Tableau 10. Les colonnes « Catégories » et « PAINS » sont les résultat donnés par l'outil de calcul de descripteurs moléculaires MolDesc, développé par l'équipe (http://moldesc.icoa.fr/). L'étiquette « Drug-like » signifie que la molécule satisfait aux règles de Lipinski. Les PAINS (« Pan Assay Interference Compounds ») sont des molécules qui vont réagir non-spécifiquement avec un grand panel de cibles et donner des résultats faux positifs dans la plupart des criblages HTS (les PAINS dépendent toutefois de la méthode de criblage

utilisée). Nous vérifions que nos composés ne présentent pas de caractéristiques semblables aux PAINS en recherchant des groupements chimiques communs à la plupart des PAINS connus[234].

Tableau 10 : Résumé des caractéristiques physico-chimiques des macrocycles construits par F2D.

ID	MW	CLogP	HBA	HBD	NRB	TPSA	NHA	NAR	NCA	Catégorie	PAINS	SA_Score
A	490,3	2,2	7	4	1	128	36	4	0	Drug-like, Kinase-like	Non	6,2
B	417,3	3,8	4	3	4	81,7	31	3	0	Drug-like, Kinase-like	Non	5,8
C	436,2	3,5	7	3	1	88,5	32	3	0	Drug-like, Kinase-like	Non	5,4

Les macrocycles créés par F2D sont conformes au filtre kinase-like mais aussi aux règles de Lipinski et ne présentent pas de sous-structure PAINS. Cependant, le SA_Score indique que ces composés sont plus difficiles à synthétiser. En comparaison, les macrocycles présents dans PKIDB ont tous un SA_Score plus faible (< 5, Figure 78). A noter que les macrocycles sont connus comme étant une classe de molécules chimiques pouvant présenter des difficultés de synthèse[235].

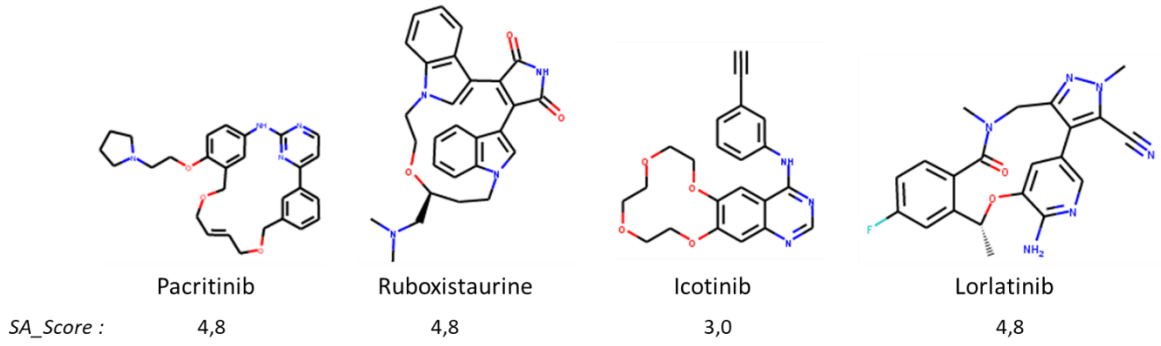


Figure 78 : Macrocycles inhibiteurs de kinase en essais cliniques ou approuvés avec leur SA_Score.

L'amarrage moléculaire de ces composés, réalisé avec rDock, montre une difficulté générale à retrouver la pose originale. En revanche, sur ces 3 exemples nous retrouvons bien parmi les 6 poses retournées, au moins une conformation présentant un RMSD < 2 Å avec le macrocycle original construit par F2D. En ce qui concerne les projections ACP et PMI, les résultats montrent quelques molécules en dehors de l'espace chimique des inhibiteurs de kinase et la grande majorité des macrocycles construits se retrouvent sur l'axe « barre-disque » dans la projection du PMI. Enfin, la recherche dans la ChEMBL de composés similaires avec un seuil de Tanimoto à 0,7 n'a rien donné.

Pour l'instant, ce travail a seulement fait l'objet d'un poster présenté lors du congrès RICT 2019 à Nantes, mais aucune de ces molécules n'a été synthétisée. Les investigations continuent et des projets de collaborations sont en discussion. Néanmoins, nous avons démontré

[234] Jonathan B. Baell et Georgina A. Holloway, « New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays », *Journal of Medicinal Chemistry* 53, nº 7 (8 avril 2010): 2719-40, https://doi.org/10.1021/jm901137j.
[235] James R. Donald et William P. Unsworth, « Ring-Expansion Reactions in the Synthesis of Macrocycles and Medium-Sized Rings », *Chemistry – A European Journal* 23, nº 37 (3 juillet 2017): 8780-99, https://doi.org/10.1002/chem.201700467.

que notre outil est capable de construire des macrocycles et un nouveau filtre a été implémenté pour les retrouver spécifiquement.

4.7.2 Application sur ABL1

4.7.2.1 Introduction du projet

La protéine kinase ABL1 fait partie du groupe des tyrosines-kinases. Elle est impliquée dans le développement de la leucémie myéloïde chronique (LMC). Historiquement, avant la découverte de son implication dans cette pathologie, des traitements classiques contre le cancer étaient distribués au patient comme le busulfan ou l'hydroxyurée. Du fait de la forte cytotoxicité de ces composés, de l'interféron alfa recombinant (rIFN-α) a aussi été employé. Grâce à l'évolution des connaissances de cette maladie et l'apparition de données structurales sur les cibles impliquées pour aider au développement d'un nouveau médicament, l'imatinib, le premier inhibiteur spécifique d'ABL1 a été approuvé en 2001. L'imatinib est rapidement devenu le traitement de référence pour la LMC. Depuis, d'autres inhibiteurs ciblant ABL1 ont été mis sur le marché comme le dasatinib et, avec l'apparition de résistance des patients à ces molécules, le nilotinib ou le bosutinib. ABL1 est à ce jour une des protéines kinases les plus étudiées et nous avons décidé d'appliquer F2D en utilisant un fragment initial de l'imatinib. Ce fragment, dans le cadre de cette application, est le groupement pyridine relié à la charnière centrale (hinge), on le retrouve dans la Figure 47 (partie 4.2.3.2). L'agrandissement s'est fait à partir d'un atome de carbone en position meta, par rapport à l'atome d'azote, et ce, afin de privilégier l'ajout de fragments en direction de la poche allostérique.

4.7.2.2 Analyse des résultats

Les résultats ont été retournés par F2D en moins de 10 minutes et nous avons obtenu 2 334 molécules uniques avant d'appliquer nos filtres kinase-like et 1 469 après. Un échantillon de ces résultats en 2D est montré Figure 79. Parmi les résultats, nous retrouvons bien l'imatinib, le ligand originel d'où provient le fragment initial. De plus, comme nous pouvons le constater dans l'échantillon présenté Figure 79, nous sommes aussi capable de reconstruire le nilotinib, un autre inhibiteur de la protéine kinase ABL1. La présence de ces deux inhibiteurs dans les molécules construites permet de valider la méthode et de continuer l'analyse.

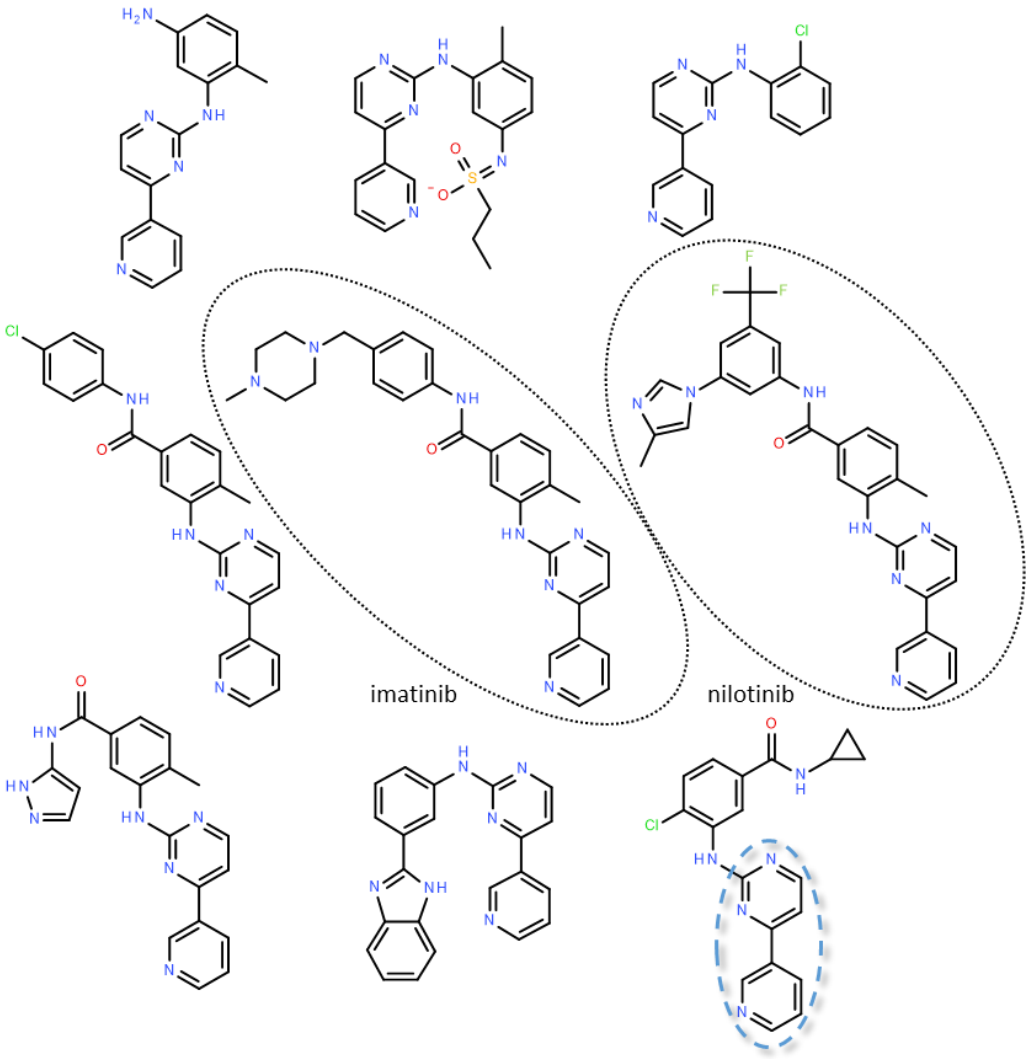


Figure 79 : Echantillon de molécules créées par F2D à partir du groupement pyridine de l'imatinib relié à la charnière centrale dans une structure 3D d'ABL1. *Les molécules sont montrées en 2D pour une meilleure visualisation et partagent un groupement 4-(3-pyridyl)pyrimidine entouré d'une ellipse bleue. Agrandissement effectué dans la structure PDB 2HYY.*

Au vu du faible nombre de composés créés par F2D, nous avons pu nous permettre dans un premier temps une première inspection visuelle. Dès ce premier regard, nous avons remarqué des dizaines de molécules présentant un squelette moléculaire intéressant. Ce châssis est représenté Figure 80, néanmoins pour des raisons de confidentialités, je ne peux divulguer le noyau central car il est encore en cours d'investigation et un dépôt de brevet est envisagé. En effet, après quelques recherches par sous-structure dans la ChEMBL et SciFinder (https://scifinder.cas.org), nous n'avons trouvé aucun composé ne présentant ce squelette moléculaire, nous laissant ainsi le champ libre afin de l'exploiter.

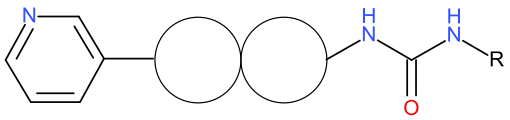


Figure 80 : Châssis moléculaire d'intérêt construit par F2D. *Les cercles représentent le noyau central non divulgué pour des raisons de confidentialité.*

La Figure 81 illustre des exemples de molécule construites arborant le squelette moléculaire d'importance (avec le noyau central représenté par sa surface moléculaire), dans le

site actif, en précisant leurs interactions avec les résidus du site actif d'ABL1. Logiquement, ces molécules forment des liaisons hydrogène avec la charnière centrale (en jaune) grâce au groupement pyridine initial. Néanmoins, les composés sont aussi capables d'interagir avec des résidus appartenant à la boucle P (en vert), à l'Asp du motif DFG (en bleu) et avec un Glu appartenant à l'hélice αC plus profond dans le site actif (en violet). F2D a bien reconstruit des ligands de type II allant cibler la poche hydrophobe accessible grâce à la conformation en DFG-out de la protéine kinase.

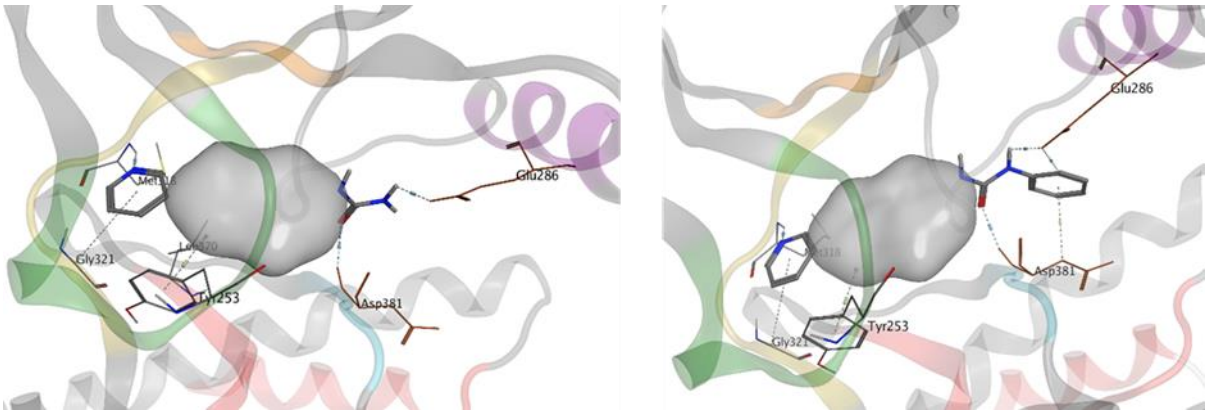


Figure 81 : Molécules construites avec F2D dans le site actif d'une structure 3D d'ABL1. *Pour des raisons de confidentialité, le noyau central des molécules a été dissimulé et remplacé par sa surface en gris clair. A gauche, le squelette principal avec un groupement urée seul, à droite, le groupement urée paré d'un groupement benzène. Code PDB : 2HYY.*

Afin de vérifier les positions dans le site actif, les molécules, avec une nouvelle conformation, ont été amarrées avec rDock. La cavité pour l'amarrage moléculaire a été calculée à partir de l'imatinib co-cristallisé dans le code PDB 2HYY. L'imatinib et le nilotinib ont été très bien replacés par le logiciel (RMSD de 0,89 Å et de 0,59 Å avec la pose obtenue par F2D respectivement, Figure 72). Quant à nos molécules d'intérêt, on retrouve pour la grande majorité d'entre elles au moins une pose sur les 6 présentant un RMSD < 1 Å avec son homologue construite par F2D. La Figure 82 illustre un exemple d'une telle molécule (en rouge) avec la meilleure pose obtenue par amarrage moléculaire (en bleu), pour un RMSD de 0,76 Å. Sur cette figure, on observe également que les interactions avec les résidus du site actif sont conservées avec la meilleure pose de l'amarrage moléculaire.

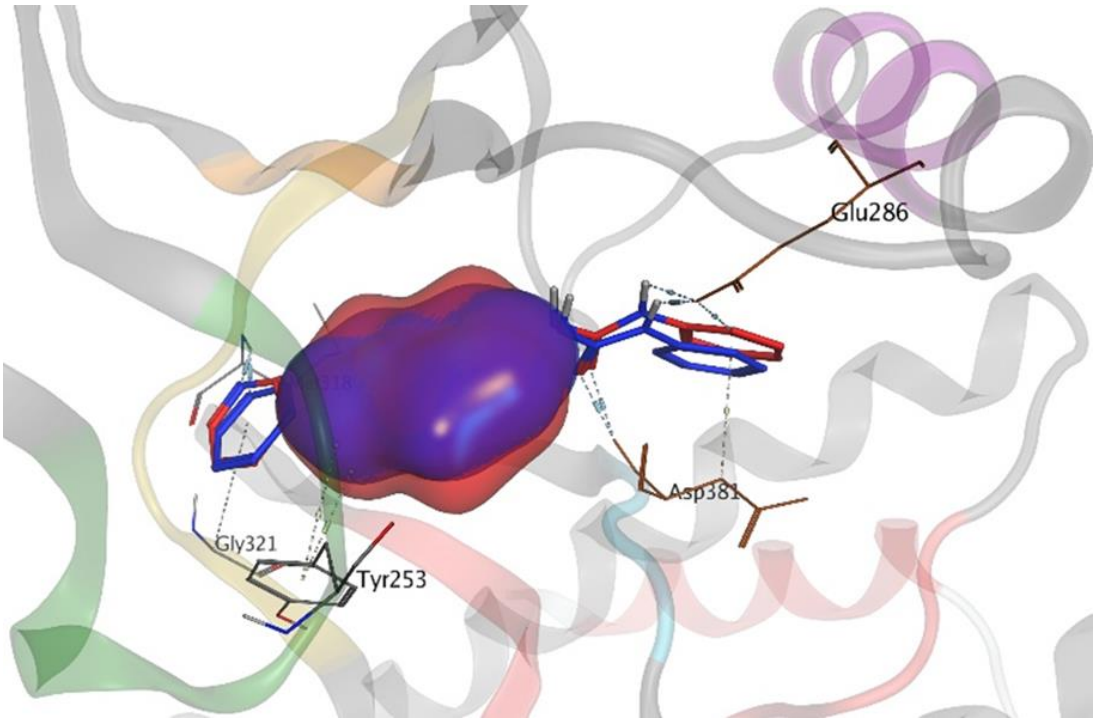


Figure 82 : Comparaison entre une molécule construite avec F2D (en rouge) et cette molécule amarrée avec rDock (en bleu). *Pour des raisons de confidentialité, le noyau central des molécules a été dissimulé et remplacé par sa surface en rouge et bleu clair respectivement. La molécule originale présente un RMSD de 0,76 Å avec son homologue amarré.*

L'étude des projections ACP et PMI (Figure 83) révèle que nos molécules (en rouge) sont englobés dans le même espace chimique que les PKI (en bleu). Les composés synthétisés présentent donc des paramètres physicochimiques similaires aux inhibiteurs de kinase en phase clinique ou approuvé. Concernant la forme 3D, nos molécules sont massées au niveau du coin supérieur gauche de la projection des PMI. Elles présent donc une forme type « barre », ou longiligne, caractéristique des inhibiteurs de kinase de type II, permettant notamment d'atteindre la poche allostérique dévoilée par le retournement du motif DFG. La majorité des autres molécules construites par F2D (en gris) présente également des caractéristiques physicochimiques similaires aux PKI et est englobée dans le même espace chimique. De plus, la forme 3D dominante de ces autres molécules est aussi la forme type « barre ». Cette prévalence s'explique par le choix du fragment initial et de la cavité en conformation DFG-out dans laquelle nous avons exécuté F2D.

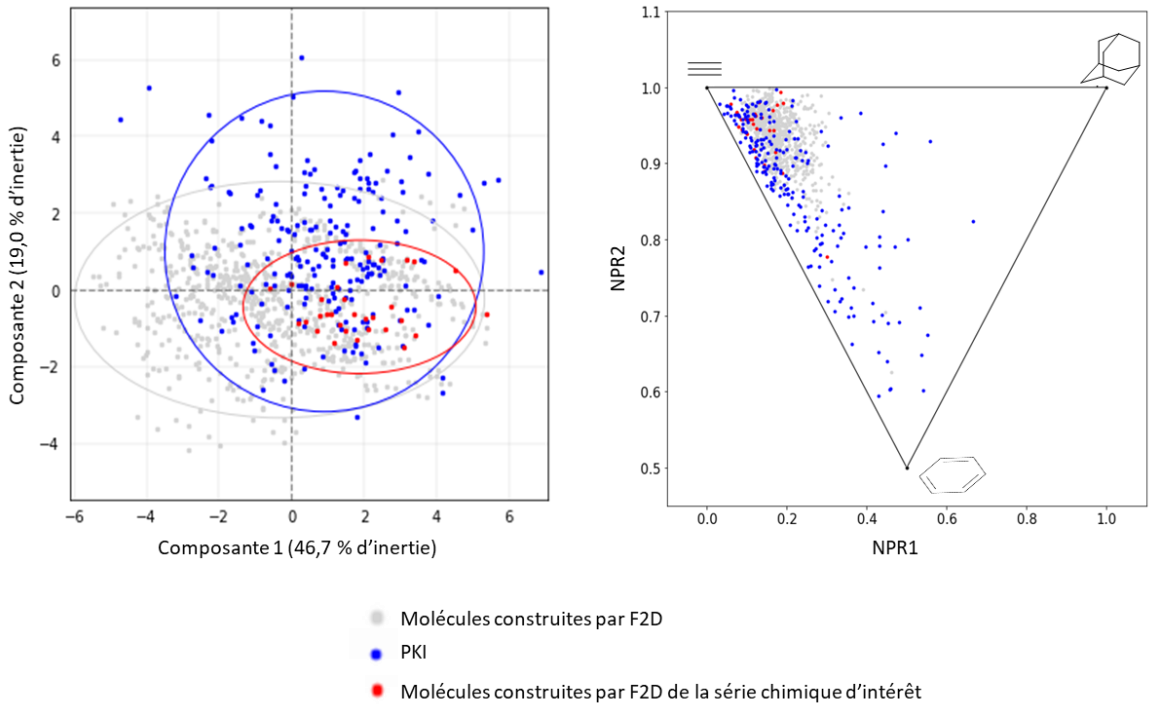


Figure 83 : Projection de l'ACP (gauche) et des PMI (droite) des différentes molécules construites par F2D dans une structure 3D d'ABL1, à partir du groupement pyridine de l'imatinib. *En gris, les molécules construites par F2D, en bleu les PKI de PKIDB et en rouge les molécules portant le squelette d'intérêt. Agrandissement effectué dans la structure PDB 2HYY.*

A la suite de ces premières observations, nous avons décidé de synthétiser quelques composés portant le châssis moléculaire prometteur afin de les tester expérimentalement sur la cible d'intérêt.

4.7.2.3 Synthèse de molécules construites par F2D

D'après les réactifs disponibles, un panel de molécules trouvées par F2D a été synthétisé au sein de notre laboratoire et est présenté Figure 84.

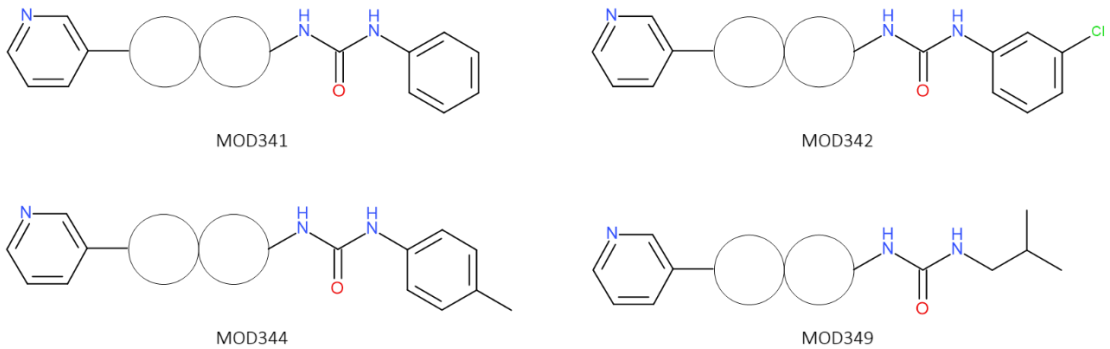
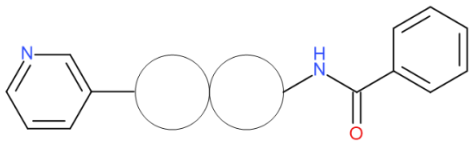


Figure 84 : Echantillon de molécules construites par F2D et synthétisées. *Les cercles représentent le noyau central non divulgué pour des raisons de confidentialité.*

De plus, une autre série chimique a été synthétisée en transformant le groupement urée relié au noyau central par un groupement amide, et ce, afin de mimer cette fonction souvent présente dans les inhibiteurs de type II. Cette fois les molécules n'ont donc pas été trouvées

directement par F2D, mais en sont des dérivés. Un exemple d'un tel composé est montré Figure 85.



MOD475

Figure 85 : Exemple d'une molécule dérivée des résultats de F2D et synthétisée. *Les cercles représentent le noyau central non divulgué pour des raisons de confidentialité.*

Le Tableau 11 résume les caractéristiques physico-chimiques des molécules présentées dans ce paragraphe, ainsi que divers filtres calculés avec MolDesc. Comme pour les macrocycles, les composés que nous avons synthétisés sont aussi drug-like et respectent les règles de Lipinski. En outre, ils ne présentent pas de structures PAINS pouvant fausser les résultats des tests de criblage. La plus légère molécule synthétisée (MOD349) est même conforme au filtre lead-like, faisant d'elle un très bon candidat pour une optimisation plus poussée par un chimiste médicinal si les premiers résultats s'avèrent prometteurs. De manière générale, tous ces composés présentent un poids moléculaire relativement faible (< 400 Da), laissant des possibilités pour de la pharmacomodulation comme des modifications ou des ajouts de groupement si nécessaire. Enfin, tous ces composés présentent un SA_Score faible et ne devraient donc pas présenter de difficultés à la synthèse.

Tableau 11 : Résumé des caractéristiques physico-chimiques des molécules synthétisées.

ID	MW	CLogP	HBA	HBD	NRB	TPSA	NHA	NAR	NCA	Catégories	PAINS	SA_Score
MOD341	329,1	4,3	3	3	3	82,7	25	4	0	Drug-like, Kinase-like	Non	2,22
MOD344	343,1	4,6	3	3	3	82,7	26	4	0	Drug-like, Kinase-like	Non	2,26
MOD349	309,2	3,4	3	3	4	82,7	23	3	0	Lead-like, Drug-like, Kinase-like	Non	2,43
MOD342	363,1	4,9	3	3	3	82,7	26	4	0	Drug-like, Kinase-like	Non	2,32
MOD475	314,1	3,9	3	2	3	70,7	24	4	0	Drug-like, Kinase-like	Non	2,15

La synthèse des molécules a été réalisé par le Pr G. Guillaumet et le post-doctorant M. Driowya. Pour les molécules contenant le groupement urée, la synthèse se fait en 5 étapes pour arriver au précurseur montré Figure 86 possédant une amine libérée (rendement global de 27 %). La libération de cette amine sur le groupement central permet de faire réagir notre précurseur avec une fonction isocyanate portant le groupement de notre choix (réaction de type : amine + isocyanate = urée, soit $R-NH2 + R'-NCO \rightarrow RNHC(O)NH-R'$). La réaction est illustrée Figure 86 et nous obtenons des rendements allant de 30 à 86 %.

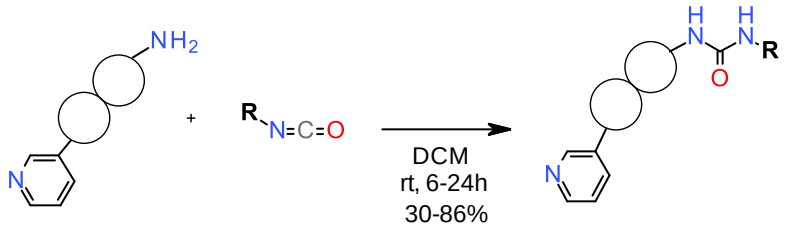


Figure 86 : Extrait de la synthèse des molécules obtenues par F2D. *Avec DCM pour dichlorométhane, rt pour « room temperature » soit température ambiante. Les cercles représentent le noyau central non divulgué pour des raisons de confidentialité.*

Concernant les composés avec un groupement amide au lieu d'un groupement urée, malgré l'accès en une étape aux composés à partir du précurseur, la synthèse est plus compliquée et nous n'avons malheureusement pas réussi à créer suffisamment de produit pour finir les tests expérimentaux. La synthèse de plus de composés est actuellement reprise par S. Front-Deschamps de la plateforme de synthèse de l'ICOA.

4.7.2.4 Tests expérimentaux

4.7.2.4.1 Test d'affinité sur des kinases présélectionnées

Une fois les composés synthétisés, nous les avons envoyés à la compagnie Eurofins DiscoverX (https://www.discoverx.com/home) pour qu'ils effectuent des tests d'affinité. Nous avons d'abord ciblé la protéine kinase ABL1 sous forme inactive, donc non-phosphorylée, mutée et non mutée. Puis, pour comparer les résultats nous avons aussi testés sur la forme active d'ABL1. Ce premier test expérimental, dit « single point », est un test compétitif de liaison (« binding test »), mesurant l'activité résiduelle des kinases avec la présence du composé comparé à un échantillon contrôle sans le composé. Plus d'informations sur ce test sont disponibles sur le site de l'entreprise : KINOMEscan® Assay Platform. Un premier test d'affinité est réalisé à la concentration de 10 µM ou 1 µM puis, selon les résultats une confirmation en dose réponse. Les résultats de ces premiers tests sont présentés Tableau 12. Puisque la majorité des molécules testées présentaient une très bonne inhibition à 10 µM, les tests suivants ont été effectués à 1 µM.

Tableau 12 : Résultats en pourcentage d'activité résiduelle des tests d'affinités, à la concentration de 1 µM, des molécules synthétisées contre différentes kinases ABL1.

ID	ABL1 (P)	ABL1 (NP)	ABL1 T315I (P)	ABL1 T315I (NP)
MOD341	14	37	86	97*
MOD344	26	64	98	99*
MOD349	*ND*	0,65	96*	*ND*
MOD342	*ND*	16*	*ND*	93*
MOD475	47	78	100	ND

Avec P pour phosphorylé, NP pour non-phosphorylé et ND pour non-déterminé.
*Test réalisé à la concentration de 10 µM.

Tout d'abord comme on le voit dans le tableau, pour certains tests la valeur n'a pu être déterminée faute soit de matière suffisante, soit d'échec du test. Malgré cela, nous pouvons constater que nos molécules ne sont absolument pas actives sur la forme mutée (T315I) d'ABL1. Comme pour l'imatinib, nous supposons que cette mutation du gatekeeper empêche nos molécules de se positionner dans le site actif. Cependant, ce n'est pas ce que nous avions

observé lors de l'analyse des résultats de Frags2Drugs, où nous distinguons clairement que nos molécules ont de la place pour se positionner dans le site actif, même avec la mutation T315I (Figure 87). En outre, nous constatons aussi que contrairement à ce que l'on attendait, il semblerait que nos molécules soient légèrement plus actives sur la forme active d'ABL1 (14 % d'activité résiduelle vs 37 % ou 26 % vs 64 % pour les molécules MOD341 et MOD344 respectivement). Ces premiers résultats pourraient indiquer que nos molécules ne se lient pas uniquement à la forme inactive d'ABL1, mais aussi à sa forme active car le groupement R (Figure 86) n'est peut-être pas suffisamment volumineux pour engendrer le déplacement « in » en « out » de la phénylalanine du motif DFG.

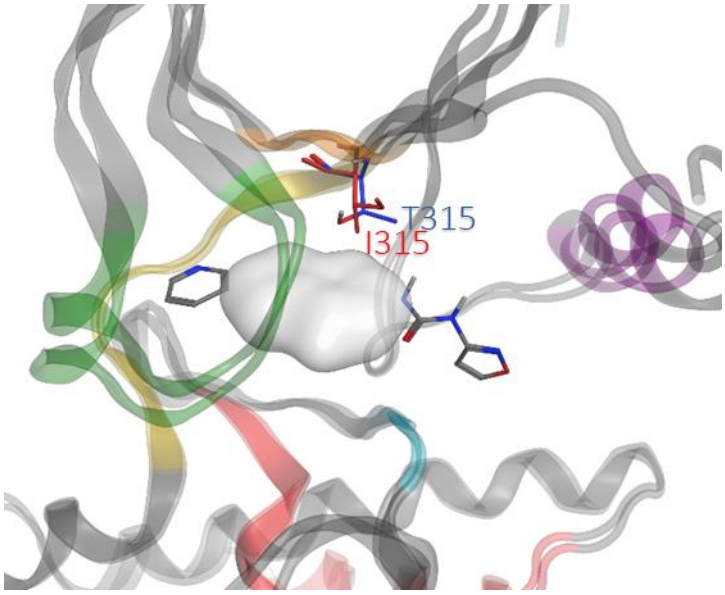


Figure 87 : Exemple d'une molécule construite par F2D dans le site actif d'une structure 3D ABL1 WT et ABL1 T315I. *La mutation n'engendre pas d'encombrement stérique et n'empêche pas la molécule de se positionner correctement. Codes PDB : 2HYY (WT) et 3QRJ (T315I).*

Pour continuer l'exploration et confirmer ces premiers résultats single point, nous avons demandé une détermination de la constante de dissociation K_d sur ces cibles afin de connaitre la confirmation dose-réponse de l'activité des composés. La Figure 88 montre des exemples de courbes dose-réponses pour déterminer le K_d sur la protéine kinase ABL1 non-phosphorylée, tandis que le Tableau 13 récapitule les différentes valeurs de pK_d obtenues sur toutes les cibles.

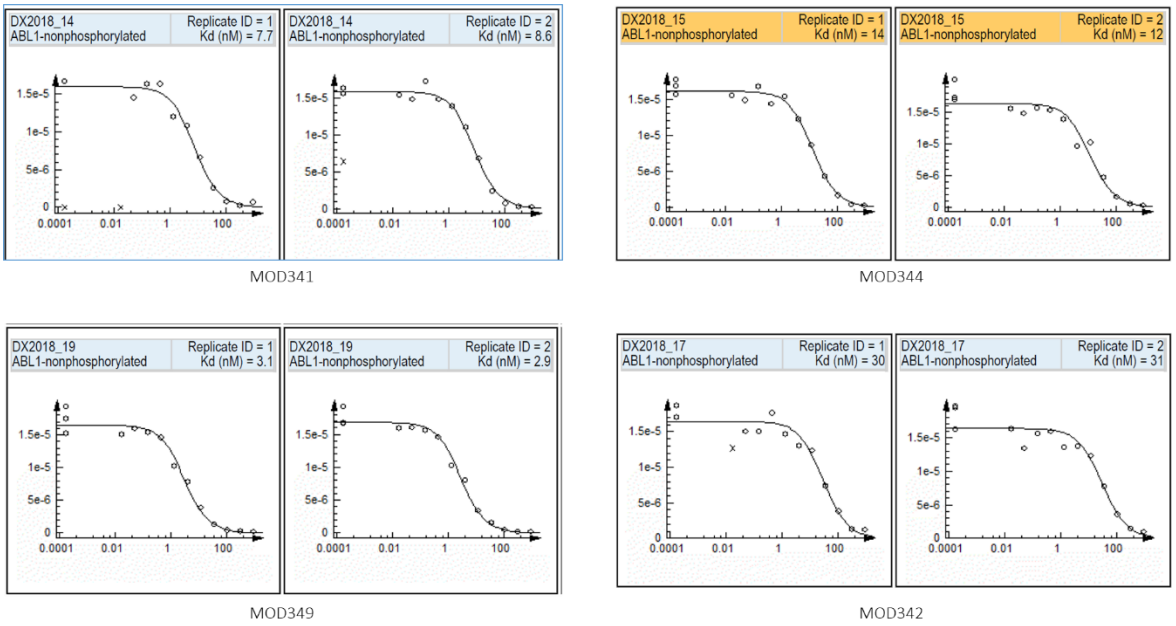


Figure 88 : Courbes de détermination de K_d sur la protéine kinase ABL1 non-phosphorylée. *Les résultats proviennent de la compagnie DiscoverX, pour chaque molécule deux essais ont été réalisés pour confirmer la valeur obtenue.*

Tableau 13 : Résultats de la détermination des pK_d des molécules synthétisées contre différentes kinases ABL1.

ID	ABL1 (P)	ABL1 (NP)	ABL1 T315I (P)	ABL1 T315I (NP)
MOD341	6,6	8,1	< 6	< 5
MOD344	6,6	7,9	< 6	< 5
MOD349	*ND*	8,5	< 5	*ND*
MOD342	*ND*	7,5	*ND*	< 5
MOD475	6,2	< 5	< 5	ND

Les valeurs des pK_d confirment bien l'inactivité de nos molécules sur la forme mutée (T315I) d'ABL1 active ou inactive. Cependant, cette fois, les valeurs des pK_d semblent clairement indiquer des inhibiteurs de type II avec plus d'un log de différence de pK_d constaté entre les formes ABL1 phosphorylées ou non (excepté pour la molécule MOD475, qui possède un groupement amide à la place de l'urée pour rappel).

Pour l'instant nous ne pouvons qu'émettre des hypothèses quant au mode exact d'inhibition de nos molécules, seule une structure cristallographique nous permettrait de les valider ou non et celle-ci est encore en cours de préparation. Toutefois, fort de ces premiers résultats, nous avons décidé de poursuivre les investigations expérimentales à travers un test de sélectivité.

4.7.2.4.2 Test de sélectivité

Comme je l'ai indiqué précédemment, la difficulté avec la famille des protéines kinases est de trouver un inhibiteur sélectif d'une seule famille à cause de la ressemblance des sites actifs de ces membres. Afin de vérifier la sélectivité de nos composés, nous sommes encore passé par l'entreprise Eurofins DiscoverX et leur plateforme KINOMEscan®. Il s'agit des mêmes essais par compétition, les molécules ont été testées à une concentration de 1 µM. Mais seulement, cette fois il ne s'agit plus uniquement de tester sur la kinase ABL1 mais sur un panel de plus de 450 (en incluant les formes phosphorylées ou non et les mutations pertinentes). Ce test de sélectivité nous permet ainsi de juger si une de nos molécules va être spécifique d'une

ou peu de protéines kinases ou au contraire très exhaustive et donc active sur toutes celles testées. Une molécule non-spécifique n'est d'aucun intérêt (pour un candidat médicament) car l'inhibition d'autres kinases que celle impliquée dans la maladie visée va être source d'effets secondaires et de toxicité. Les résultats de deux profils de sélectivité sur un panel de 100 kinases sont représentés Figure 89. Attention, cette fois les résultats sont exprimés en pourcentage d'inhibition

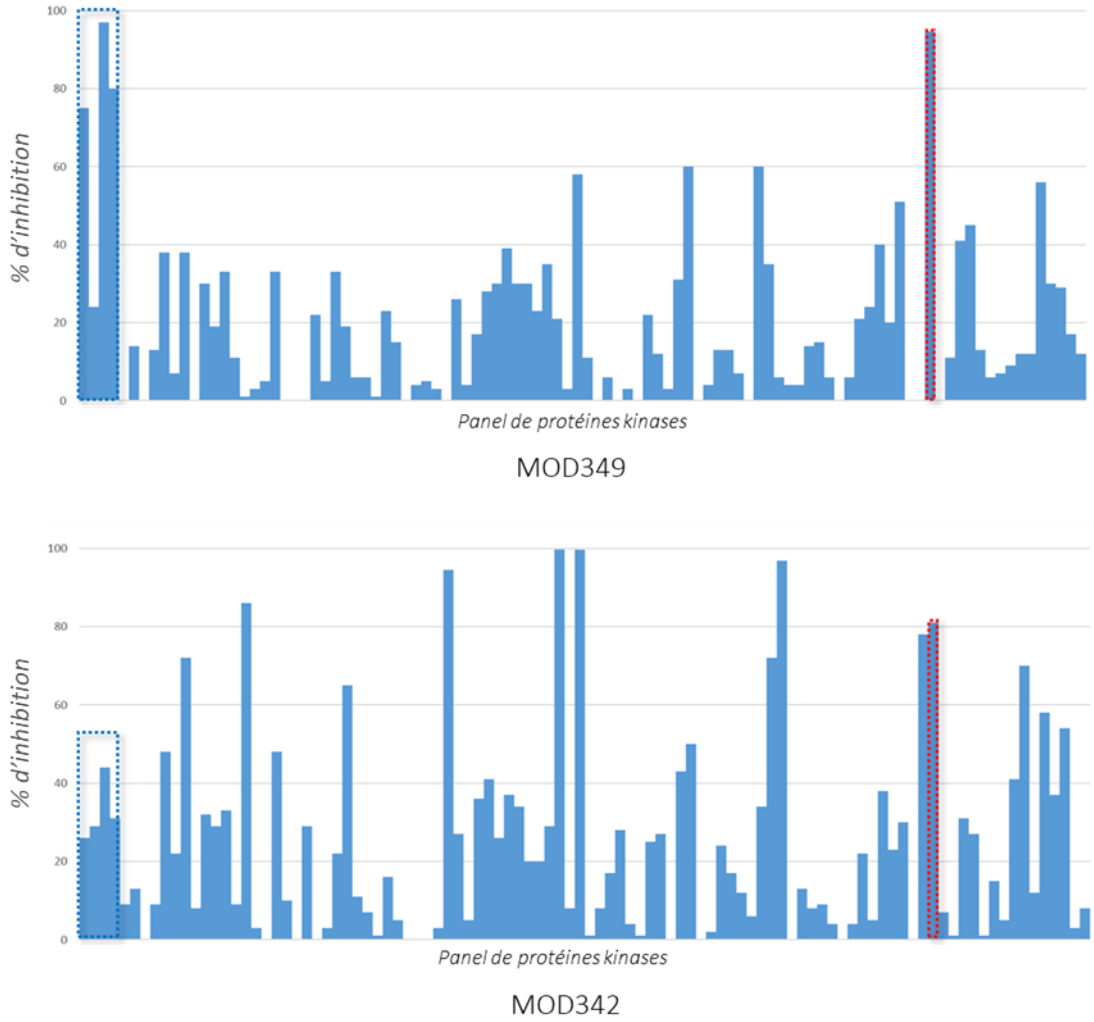


Figure 89 : Profil de sélectivité obtenus avec deux molécules sur un panel de protéines kinases. *Les résultats sont fournis par l'entreprise DiscoverX et sont exprimés en pourcentage d'inhibition. En encadré bleu, les kinases ABL1 (phosphorylée ou non, mutée ou non) et en encadré rouge, une protéine kinase d'intérêt révélée.*

Les deux profils de sélectivité, des molécules MOD349 et MOD342 respectivement, présentent des différences d'activité sur le panel de protéines kinases. En effet, la molécule MOD342 inhibe plusieurs protéines kinases autres qu'ABL1 (qui d'ailleurs n'est que faiblement inhibée par rapport aux autres). Tandis que la molécule MOD349 présente un très bon profil de sélectivité en inhibant seulement les kinases ABL1 et une autre kinase à plus de 60 %. Une confirmation sur cette autre protéine kinase qui apparait dans les deux profils a alors été demandé avec une détermination du K_d, dont les résultats sont donnés dans le Tableau 14. La molécule MOD475 est celle possédant le meilleur pK_d sur toutes les molécules testées. Cette autre protéine kinase inhibée est une cible très intéressante pour un candidat-médicament car il n'existe pas encore de molécules inhibitrices spécifiques. Au vu du profil de sélectivité, du manque d'activité sur ABL T315I et des résultats des courbes dose-réponses, et du peu d'intérêt

aujourd'hui de cibler ABL1 puisque de nombreux médicaments sont disponibles, nous avons choisi de repositionner nos molécules d'intérêt sur cette autre kinase impliquée dans le cancer du cerveau notamment. Une collaboration est en cours pour des essais expérimentaux avec l'Institut du Cerveau et de la Moelle Epinière de l'Hôpital Salpêtrière (ICM) sur des modèles cellulaires plus poussés.

Tableau 14 : Résultats de la détermination du pKd des molécules synthétisées contre une autre protéine kinase d'intérêt.

ID	Autre cible
MOD341	6,7
MOD344	6,9
MOD349	6,9
MOD342	6,4
MOD475	7,0

Après ces résultats satisfaisants de tests d'affinités seulement sur des kinases isolées, nous avons décidé de continuer les tests biologiques sur des cellules.

4.7.2.4.3 Criblage cellulaire

Ces tests cellulaires ont pour but de déterminer la cytotoxicité de nos molécules à travers des expériences de viabilité et de croissance cellulaire. Le criblage cellulaire a été réalisé par la plate-forme ImPACcell (Imagerie Pour Analyse et Criblage cellulaire, http://imagerie-puces-a-cellules.univ-rennes1.fr/index.html) sur sept lignées cellulaires différentes :

- HuH7 : cellules provenant du carcinome hépatocellulaire (cancer primitif du foie).
- CaCo-2 et HCT-116 : cellules provenant du cancer du côlon.
- MDA-MDB-231, MDA-MB-468 et MCF7 : cellules provenant du cancer du sein.
- PC3 : cellules provenant du cancer de la prostate.

Les données sur l'origine des cellules proviennent du catalogue ECACC (« European Collection of Authenticated Cell Cultures »)[236]. Les lignées cellulaires provenant de même cancers sont différentes car elles proviennent de différents patients/tissus et de cancers à différents stages. Comme pour les tests d'affinités, les premiers tests cellulaires commencent par une première détermination à une dose unique (25 µM) puis une détermination de la concentration inhibitrice médiane (IC_{50}). Le test de viabilité et de croissance cellulaire consiste en un comptage des noyaux colorés au Hoechst, qui permet de distinguer les différentes phases du cycle cellulaire, par comparaison à un témoin contrôle. Les résultats du test à dose unique sont récapitulés dans le Tableau 15.

[236] « European Collection of Authenticated Cell Cultures (ECACC) », consulté le 10 octobre 2019, https://www.phe-culturecollections.org.uk/collections/ecacc.aspx.

Tableau 15 : Résultats des tests cellulaires effectués à une concentration de 25 µM. *Les chiffres représentés correspondent aux nombres de cellules en % du contrôle (solvant DMSO) à 100 % de viabilité. Les résultats en rouge indiquent une diminution de plus de 30 % des cellules.*

	HuH7	CaCo-2	MDA-MB-231	HCT116	PC3	MDA-MB-468	MCF7
DMSO*	100	100	100	100	100	100	100
ROSCO*	15	21	23	8	27	8	20
DOXO*	61	64	37	22	41	26	37
TAXOL*	44	62	32	7	35	22	23
MOD341	93	100	107	93	95	89	93
MOD342	89	79	98	73	86	97	71
MOD343	77	71	105	73	85	71	62
MOD344	71	75	94	67	85	82	66
MOD475	47	51	28	5	31	26	14

Avec DMSO pour diméthylsulfoxyde (témoin négatif), ROSCO pour roscovitine, DOXO pour doxorubicine et TAXOL pour taxol, les trois témoins positifs respectivement.
*Composés contrôles.

Par comparaison aux témoins positifs cytotoxiques, nos molécules ne présentent pas d'activité sur les lignées cellulaires testées (moins de 30 % de mort cellulaire), excepté pour la molécule MOD475. Il est toutefois important de ne pas oublier que les molécules n'ont pas encore été optimisées pour leur propriétés drug-like. Les molécules synthétisées n'ont été utilisées que pour valider notre outil dans un premier temps. Les résultats de l'expérience plus précise de la détermination des IC_{50} (en µM) sont répertoriés dans le Tableau 16 avec des exemples de courbes de détermination montrés Figure 90.

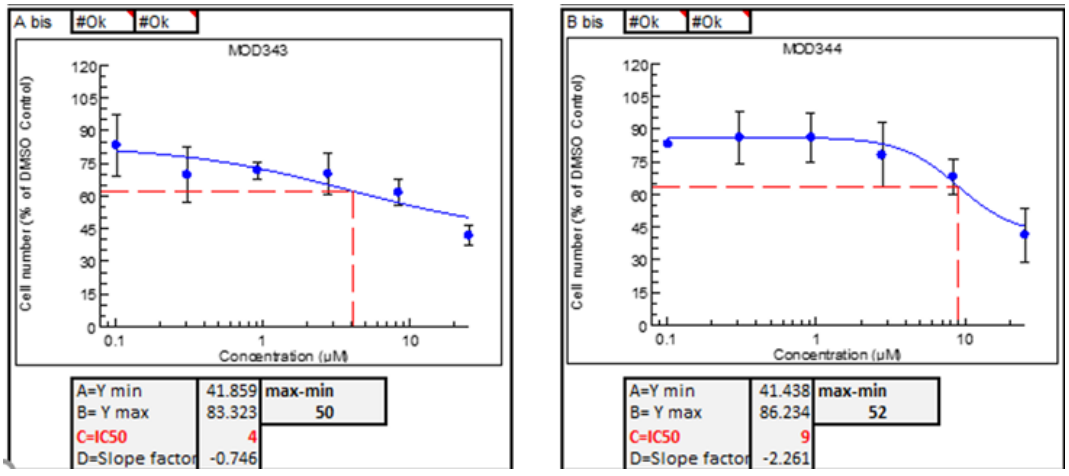


Figure 90 : Courbes d'IC_{50} obtenues avec la molécule MOD343 (gauche) et MOD344 (droite) sur les lignées cellulaires MDA-MB468. *L'amplitude correspond à la différence entre le premier et le dernier point de la courbe (max - min).*

Tableau 16 : Résultats des IC_{50}, en µM, sur les lignées cellulaires testées avec leurs amplitudes correspondantes.

	HuH7	CaCo-2	MDA-MB-231	HCT116	PC3	MDA-MB-468	MCF7	Fibro*
DMSO	> 25	> 25	> 25	> 25	> 25	> 25	> 25	> 25
ROSCO	11 *(76)*	13 *(84)*	16 *(81)*	12 *(96)*	15 *(74)*	19 *(91)*	9 *(93)*	9 *(43)*
DOXO	0,04 *(51)*	0,05 *(47)*	0,02 *(65)*	0,06 *(81)*	0,05 *(54)*	0,05 *(78)*	0,04 *(70)*	0,01 *(40)*
TAXOL	0,009 *(62)*	0,02 *(56)*	0,02 *(66)*	0,002 *(93)*	0,004 *(71)*	0,004 *(86)*	0,005 *(80)*	0,013 *(35)*
MOD341	2,7 *(38)*	> 25 *(49)*	> 25 *(73)*	> 25 *(71)*	> 25 *(73)*	7,2 *(41)*	> 25 *(31)*	> 25 *(72)*
MOD342	3,3 *(30)*	> 25 *(43)*	30 *(73)*	> 25 *(11)*	> 25 *(72)*	5.3 *(38)*	> 25 *(49)*	> 25 *(10)*
MOD343	8,6 *(44)*	6,3 *(55)*	> 25 *(20)*	68 *(71)*	> 25 *(22)*	4 *(50)*	4 *(38)*	> 25 *(5)*
MOD344	6,4 *(38)*	6,2 *(59)*	31 *(70)*	29 *(73)*	> 25 *(24)*	9 *(52)*	22 *(68)*	> 25 *(10)*
MOD475	0,9 *(66)*	1,8 *(75)*	2,2 *(77)*	1,0 *(96)*	2,7 *(72)*	0,8 *(88)*	1,2 *(88)*	3 *(32)*

*Fibroblastes (molécules saines)

Le Tableau 17 donne la valeur des pIC_{50} calculées à partir des IC_{50} avec l'outil « A pIC50 value calculator » (https://www.sanjeevslab.org/tools.html).

Tableau 17 : Résultats des pIC_{50} sur les lignées cellulaires testées avec leurs amplitudes correspondantes.

	HuH7	CaCo-2	MDA-MB-231	HCT116	PC3	MDA-MB-468	MCF7	Fibro
DMSO	< 4,6	< 4,6	< 4,6	< 4,6	< 4,6	< 4,6	< 4,6	< 4,6
ROSCO	4,96 *(76)*	4,89 *(84)*	4,80 *(81)*	4,92 *(96)*	4,82 *(74)*	4,72 *(91)*	5,04 *(93)*	5,04 *(43)*
DOXO	7,40 *(51)*	7,30 *(47)*	7,70 *(65)*	7,22 *(81)*	7,30 *(54)*	7,30 *(78)*	7,40 *(70)*	8 *(40)*
TAXOL	8,05 *(62)*	7,70 *(56)*	7,70 *(66)*	8,70 *(93)*	8,40 *(71)*	8,40 *(86)*	8,30 *(80)*	7,89 *(35)*
MOD341	5,57 *(38)*	< 4,6 *(49)*	< 4,6 *(73)*	< 4,6 *(71)*	< 4,6 *(73)*	5,14 *(41)*	< 4,6 *(31)*	< 4,6 *(72)*
MOD342	5,48 *(30)*	< 4,6 *(43)*	4,52 *(73)*	< 4,6 *(4,96)*	< 4,6 *(72)*	5.28 *(38)*	< 4,6 *(49)*	< 4,6 *(10)*
MOD343	5,07 *(44)*	5,20 *(55)*	< 4,6 *(20)*	4,16 *(71)*	< 4,6 *(22)*	5,40 *(50)*	5,40 *(38)*	< 4,6 *(5)*
MOD344	5,19 *(38)*	5,21 *(59)*	4,51 *(70)*	4,54 *(73)*	< 4,6 *(24)*	5,04 *(52)*	4,66 *(68)*	< 4,6 *(10)*
MOD475	6,05 *(66)*	5,74 *(75)*	5,66 *(77)*	6 *(96)*	5,57 *(72)*	6,10 *(88)*	5,92 *(88)*	5,52 *(32)*

Plusieurs points ressortent de ce tableau. Tout d'abord, nous constatons que les molécules avec le groupement urée (MOD341, MOD342, MOD343 et MOD344) sont inactives

sur les fibroblastes et ne présentent donc pas de cytotoxicité sur des cellules saines contrairement à notre molécule avec le groupement amide remplaçant le groupement urée (MOD475). Toutefois, la molécule roscovicine (aussi appelé seliciclib), qui présente également une cytotoxicité sur les fibroblastes, est un inhibiteur de kinase en phase clinique, ciblant les kinases CDK[237]. Bien que non-idéal, il ne s'agit pas donc pas d'un critère purement éliminatoire. D'autant que, encore une fois, nos molécules n'ont pas encore été optimisées pour aboutir au candidat-médicament. De plus, en regardant les IC_{50} des molécules MOD343 et MOD344, nous remarquons que dans les lignées cellulaires où elles sont actives, elles présentent des affinités similaires ou meilleures que la roscovicine. Cependant, elles ne sont pas actives sur toutes les lignées cellulaires, nous pouvons donc nous attendre à moins d'effets secondaires de la part de telles molécules. En effet, ces molécules ne sont pas actives (ou faiblement) sur MDA-MB-231, HCT116 et PC3. Les molécules MOD341 et MOD342 sont faiblement actives, et ce, uniquement sur les lignées cellulaires HuH7 et MDA-MB-468, qui plus est, avec une amplitude faible, ce qui en fait les molécules les moins intéressantes. En revanche, la molécule MOD475 présente une très bonne activité, avec des amplitudes élevées, sur toutes les lignées cellulaires, il s'agit donc d'une très bonne molécule de départ pour le développement d'un candidat médicament.

Un travail intéressant, faisant appel à la bioinformatique, serait de rechercher les différences de profil protéique entre les différentes lignées cellulaires testées afin de mettre en évidence certaines cibles potentielles de nos molécules. En effet, en comparant le profil protéique des lignées cellulaires et le profil d'inhibition de nos molécules sur ces lignées (comparaison d'empreintes) nous pourrions identifier les cibles inhibées ou non. Pour l'instant, faute de temps, ce travail en est encore à l'étape préliminaire de recherche de profil protéique à l'aide de l'outil CCLE (« Cancer Cell Line Encyclopedia »)[238].

4.7.2.5 Perspectives du projet

Aujourd'hui les chimistes de synthèse du laboratoire sont en train de re-synthétiser plus de composés et de continuer à exemplifier les deux séries grâce au soutien d'un chimiste médicinal afin de pouvoir continuer les tests biologiques et obtenir les valeurs manquantes. Nous attendons aussi les résultats de structure obtenue par diffraction aux rayons X qui vont nous permettre de visualiser nos molécules dans le site actif et ainsi comparer le mode d'interaction de la structure cristallographique avec nos poses obtenues par Frags2Drugs. Une collaboration est en cours, avec l'ICM de l'Hôpital Salpêtrière, pour repositionner nos molécules sur la protéine kinase d'intérêt trouvée lors du profil de sélectivité et un brevet est en cours d'écriture.

Concernant les molécules présentées dans la thèse de J.-M. Gally ciblant BRAF et en utilisant cette fois-ci une approche de « scaffold hopping »[239], elles ont également été synthétisées et montrent des résultats très encourageants sur les tests expérimentaux effectués.

---

[237] « Search of: CYC065 - List Results - ClinicalTrials.Gov », consulté le 11 octobre 2019, https://www.clinicaltrials.gov/ct2/results?term=CYC065&Search=Search.

[238] Mahmoud Ghandi et al., « Next-Generation Characterization of the Cancer Cell Line Encyclopedia », *Nature* 569, nº 7757 (mai 2019): 503-8, https://doi.org/10.1038/s41586-019-1186-3.

[239] José-Manuel Gally, « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » (Orléans, 2017).

4.8 Bilan et perspectives

Frags2Drugs est un outil déjà impliqué dans des projets de recherche au sein du laboratoire afin de concevoir de nouveaux inhibiteurs de protéine kinase. Les problèmes initiaux de temps de calcul et de gestion de la RAM sont désormais réglés. Notre programme a été validé par la reconstruction des inhibiteurs de kinases co-cristallisés et paramétré manuellement pour concevoir des molécules avec une géométrie la plus correcte possible. La précision sur la sélectivité des résultats demeure à améliorer, notamment du fait qu'il existe encore de nombreuses difficultés pour la calculer. Une limitation majeure à l'heure actuelle est que toutes les protéines kinases n'ont pas encore été structuralement résolues. De plus, les données expérimentales fournies par les BDD publiques sont parfois erronées ou contradictoires entre elles. Comme pour la majorité des projets en science des données, que ce soit pour la collecte, le paramétrage ou la validation de notre outil, il faut toujours faire attention aux données employées et en réaliser un nettoyage méticuleux en amont.

Actuellement F2D présente encore quelques limitations. On peut notamment citer l'étroitesse de son domaine d'application : uniquement les protéines kinases. Plusieurs tentatives non présentées dans ce manuscrit ont eu lieu pour essayer de le rendre utilisable sur tout type de cible avec notamment de l'amarrage moléculaire de fragments dans le site actif pour ensuite créer le réseau de relations entre eux. Malheureusement, ces essais se sont montrés infructueux et non-concluants, l'amarrage moléculaire de fragments demeurant un défi en chémoinformatique[240]. Une autre limitation provient du programme tel qu'il a été pensé : utiliser des fragments issus d'inhibiteurs co-cristallisés pour générer de nouvelles molécules. Cette méthode est certes rapide car les positions des fragments sont connues à l'avance, cependant les résultats restent très dépendants du choix du fragment initial, qui doit donc se faire minutieusement. L'alignement des domaines kinases est également un point primordial pour le succès de la technique employée. En effet, si une structure est mal intégrée au référentiel, ses fragments peuvent ne jamais servir et donc nous pouvons perdre de potentielles combinaisons intéressantes, tout comme les fragments en collision stérique avec la cible sélectionnée qui ne seront pas pris en compte. Cela nous ramène au point discuté juste au-dessus concernant l'amarrage moléculaire. En effet, une telle méthode nous permettrait de replacer tous les fragments dans le site actif sans collisions stériques. Cependant, en plus d'être imprécise, cette solution serait aussi très chronophage et nous perdrions le bénéfice de toutes les optimisations apportées. Pour l'instant nous avons donc décidé de pas intégrer cette technique et de faire avec les fragments dont nous disposons, les résultats préliminaires étant suffisamment satisfaisants. Il n'est toutefois pas exclu qu'avec les progrès en la matière, l'amarrage moléculaire soit intégré à F2D. Cela permettrait aussi d'agrandir l'espace chimique des fragments en rajoutant ceux que l'on souhaite. Enfin, comme la plupart des inhibiteurs de protéines kinases co-cristallisés sont des inhibiteurs de type I, I ½ ou II (réalisant une interaction avec la charnière centrale de la kinase), la majorité des molécules créées par F2D va aussi refléter cette tendance. Un attribut porté par chaque fragment présentant ses différentes interactions avec les motifs structuraux des kinases permettrait de facilement se débarrasser de ceux que l'on ne souhaite pas afin de construire d'autre types d'inhibiteurs. Il est à noter qu'il ne sera pas encore possible de créer des inhibiteurs allostériques car par définition ces

[240] Marcel L. Verdonk et al., « Docking Performance of Fragments and Druglike Compounds », *Journal of Medicinal Chemistry* 54, nº 15 (11 août 2011): 5422-31, https://doi.org/10.1021/jm200558u.

inhibiteurs se lient dans une poche extérieure au site de l'ATP et ces poches sont peu conservées parmi les kinases, nous ne disposons donc pas de suffisamment de fragments. Là encore, un amarrage moléculaire pourrait débloquer cette option.

Un point parfois soulevé par mes collègues lors des présentations en congrès ou en colloque de F2D concerne l'originalité des molécules. La remarque est justifiée, il est assez logique de s'interroger sur notre capacité à créer des composés innovant à partir de composés connus et documentés. En réalité, comme je l'ai montré dans ce chapitre, nous sommes tout à fait capables de sortir de l'espace chimique déjà exploré et de créer des molécules intéressantes, dont des macrocycles, grâce au nombre de fragments disponibles de toutes les tailles et à leurs multiples combinaisons qui permettent de sortir des sentiers battus.

Plusieurs extensions à ce projet sont déjà envisagées et plus ou moins avancées. Tout d'abord le couplage des résultats avec un autre programme développé dans l'équipe par F. Carles fondé sur la protéochémométrie pour prédire la sélectivité et l'activité d'une molécule sur les différentes familles de protéines kinases[241]. Un réel outil de rétro-synthèse pourrait aussi être incorporé au lieu du simple score de synthèse actuellement implémenté. Je peux aussi citer la continuité des essais d'amarrage moléculaire, particulièrement pour agrémenter notre librairie de fragments moins classiques et ainsi ouvrir encore plus l'espace chimique. D'ores et déjà, mon successeur G. Peyrat travaille sur une adaptation de F2D aux bromodomaines, qui comme les protéines kinases possèdent des motifs en commun et peuvent facilement être tous superposés dans un même référentiel. Un projet de création de ligands covalents est aussi en cours, en lançant F2D à partir d'un fragment accroché à la cible sur une cystéine par exemple. Les premiers résultats expérimentaux suivent leurs phases de développement mais sont très prometteurs comme en attestent les tests biologiques. Des dépôts de brevets de molécules trouvées avec F2D ne devraient pas tarder à se concrétiser. Un nouveau membre de l'équipe a d'ailleurs été embauché dans le but d'exemplifier ces composés, étape requise pour le dépôt. Enfin, un serveur web est actuellement en cours de finalisation pour permettre l'utilisation de notre programme à tout utilisateur via la plateforme d'outils proposée par l'équipe (http://sbc.icoa.fr/).

Deux points importants sur lesquels j'aimerais revenir pour terminer ce chapitre. Tout d'abord, je tiens à insister sur le fait que malgré des exemples et des illustrations relativement simplistes dans ce manuscrit, il faut bien garder à l'esprit que ce projet repose sur des données conséquentes (plusieurs centaines de millions de relations) et que l'algorithmique développée ne se résume pas juste à de simples parcours de graphes mais fait appel à de nombreux modules en langage Python et à une ingénierie bien plus avancée. Enfin, je tiens à souligner l'importance de plusieurs personnes impliquées dans le projet Frags2Drugs : José-Manuel Gally en premier lieu pour m'avoir aiguillé au début de ma thèse et légué ce projet en s'assurant que je l'avais bien pris en main. Ensuite, Pascal Krezel, qui a grandement contribué à l'optimisation du code et à la maitrise des graphes, Thomas Misiek qui a apporté ses connaissances et sa touche d'informaticien – ce qui nous a manqué parfois afin de parfaire le programme - et Gautier Peyrat

[241] Fabrice Carles, « Développement d'une approche protéo-chimiométrique tridimensionnelle pour l'identification d'inhibiteurs de protéines kinases. » (Orléans, 2019).

pour la mise à jour de la base de données, la validation des paramètres et la continuité du projet. Pour ma part, mon histoire avec F2D s'achève avec ce chapitre mais je laisse désormais cet outil en sachant qu'il est entre de bonnes mains.

…, *« et continuez jusqu'à ce que vous arriviez à la fin ; là, vous vous arrêterez. »*

Alice au pays des merveilles – Lewis Caroll

Chapitre 5 : Conclusion générale

Afin de prendre toute la mesure de mon arrivée jusqu'ici, je débuterai cette conclusion générale par quelques mots sur mon parcours personnel. A l'origine, mes premières intentions étaient de passer par un Institut Universitaire de Technologie (IUT) pour rejoindre une école d'ingénieur en biologie, sans passer par la case « prépas » (au grand dam de mes parents). Au vue de mes premiers résultats, certes à la hauteur de mon investissement personnel je dois l'avouer (il est tout de même fort regrettable de ne pas avoir une évaluation sur la participation et l'organisation aux fêtes), j'ai dû me rendre à l'évidence : mon plan 1 était un peu compromis. De DUT je suis donc passé à Licence puis Master et une chose en entrainant une autre, me voilà à terminer la rédaction de ma thèse en vue d'obtenir un Doctorat. L'ironie dans cette épopée c'est que durant toutes mes études supérieures j'ai cherché en vain à échapper aux cours de chimie (organique particulièrement). Cette matière se refusait à moi et malgré toute la meilleure volonté du monde mes travaux pratiques se soldaient toujours par un mauvais rendement, quand rendement il y avait, ou par un produit impur voire carrément le mauvais produit. C'est finalement grâce aux enseignements en bioinformatique et chémoinformatique durant mon cursus que j'ai fini par trouver ma voie et ainsi me retrouver dans cette position.

A présent trois années se sont écoulées, plus ou moins rapidement, depuis que j'ai débuté ce travail en 2016. A mon arrivée dans l'équipe, il y avait 33 inhibiteurs de kinases approuvés par différents organismes à travers le monde. Aujourd'hui en 2019, il en existe 54 ce qui démontre l'importance de cette famille de protéines comme cibles, particulièrement dans le domaine de l'oncologie. D'aucuns souhaiteraient réduire leurs efforts dans cette lutte, d'autant que cette même année, hasard du calendrier, on apprend que le cancer est devenue la cause principale de décès dans les pays riches[242]. Je ne sais pas s'il faut s'en réjouir mais en tout cas le travail ne devrait donc pas manquer à l'avenir. Pourtant, il devient délicat de mettre sur le marché un nouveau médicament, l'industrie pharmaceutique ayant de plus en plus de difficultés à innover et sortir de nouveaux blockbusters. Désormais confrontées à des nouveaux défis, les méthodes employées pour la conception de médicament doivent désormais évoluer. En effet, nous nous dirigeons de plus en plus vers une médecine personnalisée, particulièrement en cancérologie[243]. Cette médecine a pour but de diagnostiquer le plus précisément les pathologies et leurs nuances afin d'orienter le patient vers le traitement le plus optimal. Cela passe notamment par le traitement et la compréhension des données massives fournies par la génomique ou la métabolomique. On cherche maintenant à traiter chaque cancer en fonction de son profil moléculaire et génomique plutôt que par rapport à sa localisation et son analyse histologique. Deux disciplines tirent leurs épingles du jeu et prennent de plus en plus d'importance dans ce processus : la bioinformatique et la chémoinformatique. La première pour étudier les données issues des séquençages génomiques et la seconde pour étudier la structure des protéines et comprendre l'effet que les mutations peuvent engendrer sur celle-ci.

Les travaux de recherche réalisés durant cette thèse tentent d'apporter des solutions à ces challenges à l'aide de l'informatique et de la modélisation moléculaire. C'est notamment le cas de Frags2Drugs, le logiciel que nous avons développé pour la conception d'inhibiteurs de

[242] « Le cancer devient la première cause de décès dans les pays riches », *Le Monde.fr*, 3 septembre 2019, https://www.lemonde.fr/societe/article/2019/09/03/le-cancer-devient-la-premiere-cause-de-deces-dans-les-pays-riches_5505928_3224.html.

[243] « Cancer : la médecine personnalisée est possible », Inserm - La science pour la santé, consulté le 26 septembre 2019, https://www.inserm.fr/cancer-medecine-personnalisee-est-possible.

protéines kinases à partir de fragments moléculaires. Ce programme est fondé sur les données expérimentales disponibles et permet de construire des composés directement dans une cavité afin de les adapter au site actif de la cible et prendre en compte sa spécificité. Grâce à son optimisation et aux nombreuses possibilités de combinaisons entre tous nos fragments, Frags2Drugs est capable de proposer des solutions innovantes rapidement. De plus, se voulant facile à prendre en main, il peut être utilisé par toutes personnes sans besoin de connaissances avancées en informatique ou chémoinformatique. Pour guider l'utilisateur, plusieurs méthodes de filtrage des résultats ont aussi été implémentées permettant de ne conserver que les molécules présentant le plus haut potentiel d'activité. Ce programme est dorénavant associé à de multiples projets dans l'équipe, mais aussi en collaboration avec d'autres laboratoires et entreprises. Les premiers résultats sont très prometteurs et les molécules que nous avons créées sont actuellement en cours de tests biologiques avancés (cellulaires puis *in vivo* si les retours sont satisfaisants) avec à la clé plusieurs dépôts de brevets possibles. A terme ce logiciel devrait être rendu accessible à tous via un serveur web et ainsi, je l'espère, bénéficier à toute la communauté (chémoinformaticiens comme chimistes de synthèses ou médicinaux) pour les assister dans leurs projets. Frags2Drugs peut notamment servir à donner des idées de création de molécules, d'agrandissement d'un fragment prometteur ou encore d'exemplification de séries chimiques.

Durant cette thèse, j'ai aussi mené un projet d'amarrage moléculaire virtuel à but cosmétique en utilisant uniquement des produits naturels. L'amarrage moléculaire virtuel reste une alternative intéressante à son homologue le criblage à haut débit car il représente un coût beaucoup plus faible et reste beaucoup plus rapide à exécuter. Cependant, comme nous l'avons vu, il se peut que quelque fois les résultats ne soient pas à la hauteur de nos espérances malgré une sélection minutieuse des composés à tester expérimentalement. L'amarrage moléculaire demeure pour l'instant un outil de plus dans la mallette du chémoinformaticien, il ne doit pas être considéré comme la panacée mais utilisé en association avec d'autres méthodes de sélection de molécules afin de maximiser les chances de ne retenir que les composés les plus intéressants.

Enfin, à titre personnel, cette thèse m'a permis de compléter ma formation en chémoinformatique mais plus globalement en tant que scientifique. Je pense notamment à la manière de conduire un projet à bien, de se creuser les méninges (seul ou à plusieurs) afin d'apporter des solutions aux problèmes rencontrés, de ne pas chercher à s'éparpiller et de vouloir en faire trop mais plutôt se focaliser sur une tâche précise avant de passer à la suite, etc. J'ai eu le privilège de présenter ces travaux dans de multiples congrès nationaux et internationaux. Ces réunions sont toujours intéressantes et regorgent d'opportunités pour avancer dans nos projets car la discussion avec des collègues permet d'avoir un autre point de vue et peut débloquer une situation. J'ai aussi eu l'opportunité de recruter, former et encadrer un stagiaire et ainsi m'ouvrir au management. Et pour finir, j'ai réveillé l'âme d'enseignant qui somnolait en moi à travers différents cours prodigués aussi bien à des étudiants en deuxième année de Licence qu'à d'autres en Master et ainsi faire découvrir à mon tour la chémoinformatique.

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 international

- **Bournez, C.** ; Krezel, P. ; Gally, J.-M. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs: Discovery of new kinase inhibitors from 3D fragment network
  Chemoinformatics Strasbourg Summer School 2018
  juin 2018 - Strasbourg.

- **Diharce, J.** ; Bournez, C. ; Krezel, P. ; Fruit, C. ; Besson, T. ; Bonnet, P.
  Fragment-based drug design approach using a novel in silico tools to inhibit DYRK kinase family
  26èmes Rencontres Internationales des Pharmacochimistes de l'Arc Atlantique et 32èmes Journées Franco-Belges de Chimie Thérapeutique (GP2A-JFB 2018)
  juin 2018 - Asnelles sur Mer.

- **Bournez, C.** ; Gally, J.-M. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs, a novel in silico FBDD tool
  4,6th Young Research Fellow Meeting
  mars 2018 - Orléans.

Communications orales dans un congrès national

- **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
  nov. 2019 - Paris.

- **Bournez, C.** ; Krezel, P. ; Gally, J.-M. ; Driowya, M. ; Aci-Sèche, S. ; Guillaumet, G. ; Bonnet, P.
  D'un réseau de fragments 3D à la conception de nouveaux inhibiteurs de kinases
  30ème colloque Biotechnocentre
  octobre 2018 - Seillac.

- **Bournez, C.** ; Gally, J.-M. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs, a novel in silico FBDD tool
  17èmes REncontres en Chimie Organique Biologique (RECOB17)
  mars 2018 - Aussois.

Communications flashs dans un congrès international

- **Bournez, C.** ; Gally, J.-M. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs: A 3D fragment network to discover new compounds
  4,66th ACS National Meeting
  août 2018 - Boston (USA).

Communications par affiches

Communications par poster dans un congrès international

- **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.

- **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.

- **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.

- **Bournez, C**. ; Gally, J.-M. ; Krezel, P. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs: A 3D fragment network to discover new compounds
  4,66th ACS National Meeting
  août 2018 - Boston (USA).

- **Bournez, C.** ; Krezel, P. ; Gally, J.-M. ; Aci-Sèche, S. ; Bonnet, P.
  Frags2Drugs: Finding new kinase inhibitors from graph-based fragments linking approach
  26èmes Rencontres Internationales des Pharmacochimistes de l'Arc Atlantique et 32èmes Journées Franco-Belges de Chimie Thérapeutique (GP2A-JFB 2018)
  juin 2018 - Asnelles sur Mer.

Communications par poster dans un congrès national

- **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
  nov. 2019 - Paris.

- **Bournez, C.** ; Gally, J.-M. ; Do Q.T. ; Aci-Sèche, S. ; Bernard, P. ; Bonnet, P.
  Identification d'ingrédients actifs naturels par criblage virtuel et validation expérimentale
  Cosmétosciences
  février 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

Akritopoulou-Zanze, Irini, et Philip J. Hajduk. « Kinase-targeted libraries: The design and synthesis of novel, potent, and selective kinase inhibitors ». *Drug Discovery Today* 14, nº 5 (1 mars 2009): 291-97. https://doi.org/10.1016/j.drudis.2008.12.002.

Alex, Alexander A., et David S. Millan. « Chapter 5: Contribution of Structure-Based Drug Design to the Discovery of Marketed Drugs ». *Drug Design Strategies*, 108-63, 2011. https://doi.org/10.1039/9781849733410-00108.

Alexanderson, Gerald L. « About the Cover: Euler and Königsberg's Bridges: A Historical View ». *Bulletin of the American Mathematical Society* 43, nº 04 (18 juillet 2006): 567-74. https://doi.org/10.1090/S0273-0979-06-01130-X.

Andrews, P. R., D. J. Craik, et J. L. Martin. « Functional group contributions to drug-receptor interactions ». *Journal of Medicinal Chemistry* 27, nº 12 (1 décembre 1984): 1648-57. https://doi.org/10.1021/jm00378a021.

Atwell, Shane, Jason M. Adams, John Badger, Michelle D. Buchanan, Ingeborg K. Feil, Karen J. Froning, Xia Gao, et al. « A Novel Mode of Gleevec Binding Is Revealed by the Structure of Spleen Tyrosine Kinase ». *The Journal of Biological Chemistry* 279, nº 53 (31 décembre 2004): 55827-32. https://doi.org/10.1074/jbc.M409792200.

Atzori, Alessio, Neil J. Bruce, Kepa K. Burusco, Berthold Wroblowski, Pascal Bonnet, et Richard A. Bryce. « Exploring Protein Kinase Conformation Using Swarm-Enhanced Sampling Molecular Dynamics ». *Journal of Chemical Information and Modeling* 54, nº 10 (27 octobre 2014): 2764-75. https://doi.org/10.1021/ci5003334.

Axel Chambily et Pétrut Constantine. « Cours sur la récursivité ». Developpez.com. Consulté le 24 septembre 2019. http://recursivite.developpez.com/.

Baell, Jonathan B., et Georgina A. Holloway. « New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for Their Exclusion in Bioassays ». *Journal of Medicinal Chemistry* 53, nº 7 (8 avril 2010): 2719-40. https://doi.org/10.1021/jm901137j.

Baker, Monya. « Fragment-Based Lead Discovery Grows Up ». *Nature Reviews Drug Discovery* 12, nº 1 (janvier 2013): 5-7. https://doi.org/10.1038/nrd3926.

Balboa, Elena M., Enma Conde, M. Luisa Soto, Lorena Pérez-Armada, et Herminia Domínguez. « Cosmetics from Marine Sources ». *Springer Handbook of Marine Biotechnology*, édité par Se-Kwon Kim, 1015-42. Springer Handbooks. Berlin, Heidelberg: Springer Berlin Heidelberg, 2015. https://doi.org/10.1007/978-3-642-53971-8_44.

Barette, Caroline, Emmanuelle Soleilhac, Céline Charavay, Claude Cochet, et Marie-Odile Fauvarque. « Force et spécificité du criblage pour des molécules bioactives au CMBA-Grenoble - Une plate-forme dédiée à la découverte et à l'analyse de molécules bioactives et candidats médicaments ». *médecine/sciences* 31, nº 4 (1 avril 2015): 423-31. https://doi.org/10.1051/medsci/20153104017.

Berman, Helen M., John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, et Philip E. Bourne. « The Protein Data Bank ». *Nucleic Acids Research* 28, nº 1 (1 janvier 2000): 235-42. https://doi.org/10.1093/nar/28.1.235.

Berthold, Michael R., Nicolas Cebron, Fabian Dill, Thomas R. Gabriel, Tobias Kötter, Thorsten Meinl, Peter Ohl, Christoph Sieb, Kilian Thiel, et Bernd Wiswedel. « KNIME: The Konstanz Information Miner ». *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

Böhm, Hans-Joachim. « The Computer Program LUDI: A New Method for the de Novo Design of Enzyme Inhibitors ». *Journal of Computer-Aided Molecular Design* 6, nº 1 (1 février 1992): 61-78. https://doi.org/10.1007/BF00124387.

Bollag, Gideon, James Tsai, Jiazhong Zhang, Chao Zhang, Prabha Ibrahim, Keith Nolop, et Peter Hirth. « Vemurafenib: The First Drug Approved for *BRAF*-Mutant Cancer ». *Nature Reviews Drug Discovery* 11, nº 11 (novembre 2012): 873-86. https://doi.org/10.1038/nrd3847.

Brown, Frank K. « Chapter 35 - Chemoinformatics: What is it and How does it Impact Drug Discovery. » In *Annual Reports in Medicinal Chemistry*, édité par James A. Bristol, 33:375-84. Academic Press, 1998. https://doi.org/10.1016/S0065-7743(08)61100-8.

Burnett, George, et Eugene P. Kennedy. « The Enzymatic Phosphorylation of Proteins ». *Journal of Biological Chemistry* 211, nº 2 (12 janvier 1954): 969-80.

Canadian Science Publishing. « 21st Century Science Overload ». Consulté le 10 octobre 2019. http://blog.cdnsciencepub.com/21st-century-science-overload/.

Carles, Fabrice. « Développement d'une approche protéo-chimiométrique tridimensionnelle pour l'identification d'inhibiteurs de protéines kinases. » Orléans, 2019.

Carles, Fabrice, Stéphane Bourg, Christophe Meyer, et Pascal Bonnet. « PKIDB: A Curated, Annotated and Updated Database of Protein Kinase Inhibitors in Clinical Trials ». *Molecules* 23, nº 4 (15 avril 2018): 908. https://doi.org/10.3390/molecules23040908.

Carter, Paul J. « Potent Antibody Therapeutics by Design ». *Nature Reviews Immunology* 6, nº 5 (mai 2006): 343. https://doi.org/10.1038/nri1837.

Castagna, M., Y. Takai, K. Kaibuchi, K. Sano, U. Kikkawa, et Y. Nishizuka. « Direct Activation of Calcium-Activated, Phospholipid-Dependent Protein Kinase by Tumor-Promoting Phorbol Esters ». *The Journal of Biological Chemistry* 257, nº 13 (10 juillet 1982): 7847-51.

Cavalluzzi, Maria Maddalena, Giuseppe Felice Mangiatordi, Orazio Nicolotti, et Giovanni Lentini. « Ligand efficiency metrics in drug discovery: the pros and cons from a practical perspective ». *Expert Opinion on Drug Discovery* 12, nº 11 (2 novembre 2017): 1087-1104. https://doi.org/10.1080/17460441.2017.1365056.

Charifson, Paul S., Joseph J. Corkery, Mark A. Murcko, et W. Patrick Walters. « Consensus Scoring: A Method for Obtaining Improved Hit Rates from Docking Databases of Three-Dimensional Structures into Proteins ». *Journal of Medicinal Chemistry* 42, nº 25 (1 décembre 1999): 5100-5109. https://doi.org/10.1021/jm990352k.

Chen, Hongming, Paul D. Lyne, Fabrizio Giordanetto, Timothy Lovell, et Jin Li. « On Evaluating Molecular-Docking Methods for Pose Prediction and Enrichment Factors ». *Journal of Chemical Information and Modeling* 46, nº 1 (1 janvier 2006): 401-15. https://doi.org/10.1021/ci0503255.

Chessari, Gianni, et Andrew J. Woodhead. « From fragment to clinical candidate—a historical perspective ». *Drug Discovery Today* 14, nº 13 (1 juillet 2009): 668-75. https://doi.org/10.1016/j.drudis.2009.04.007.

Code de la santé publique - Article L5111-1, L5111-1 Code de la santé publique § (s. d.). Consulté le 5 juin 2019.

Cohen, Philip. « Protein Kinases — the Major Drug Targets of the Twenty-First Century? » *Nature Reviews Drug Discovery* 1, nº 4 (avril 2002): 309-15. https://doi.org/10.1038/nrd773.

Collett, M. S., et R. L. Erikson. « Protein Kinase Activity Associated with the Avian Sarcoma Virus Src Gene Product ». *Proceedings of the National Academy of Sciences of the United States of America* 75, nº 4 (avril 1978): 2021-24. https://doi.org/10.1073/pnas.75.4.2021.

Commissioner, Office of the. « The Drug Development Process - Step 3: Clinical Research ». WebContent. Consulté le 20 mars 2019. https://www.fda.gov/forpatients/approvals/drugs/ucm405622.htm.

Congreve, Miles, Robin Carr, Chris Murray, et Harren Jhoti. « A 'Rule of Three' for fragment-based lead discovery? » *Drug Discovery Today* 8, nº 19 (1 octobre 2003): 876-77. https://doi.org/10.1016/S1359-6446(03)02831-9.

Couly, Florence, Marine Harari, Carole Dubouilh-Benard, Laetitia Bailly, Emilie Petit, Julien Diharce, Pascal Bonnet, et al. « Development of Kinase Inhibitors via Metal-Catalyzed C–H Arylation of 8-Alkyl-Thiazolo[5,4-f]-Quinazolin-9-Ones Designed by Fragment-Growing Studies ». *Molecules* 23, nº 9 (29 août 2018): 2181. https://doi.org/10.3390/molecules23092181.

« Coût des nouveaux traitements de lutte contre l'hépatite C - Sénat ». Consulté le 11 juin 2019. https://www.senat.fr/questions/base/2014/qSEQ140712580.html.

Crino, Peter B. « The MTOR Signalling Cascade: Paving New Roads to Cure Neurological Disease ». *Nature Reviews Neurology* 12, nº 7 (juillet 2016): 379-92. https://doi.org/10.1038/nrneurol.2016.81.

Cyphers, Soreen, Emily F. Ruff, Julie M. Behr, John D. Chodera, et Nicholas M. Levinson. « A Water-Mediated Allosteric Network Governs Activation of Aurora Kinase A ». *Nature Chemical Biology* 13, nº 4 (avril 2017): 402-8. https://doi.org/10.1038/nchembio.2296.

Da, C., et D. Kireev. « Structural Protein–Ligand Interaction Fingerprints (SPLIF) for Structure-Based Virtual Screening: Method and Benchmark Study ». *Journal of Chemical Information and Modeling* 54, nº 9 (22 septembre 2014): 2555-61. https://doi.org/10.1021/ci500319f.

Davies, Douglas R. « Screening Ligands by X-Ray Crystallography ». *Structural Genomics and Drug Discovery: Methods and Protocols*, édité par Wayne F. Anderson, 315-23. Methods in Molecular Biology. New York, NY: Springer New York, 2014. https://doi.org/10.1007/978-1-4939-0354-2_23.

Davis, Andrew M., Stephen A. St-Gallay, et Gerard J. Kleywegt. « Limitations and lessons in the use of X-ray structural information in drug design ». *Drug Discovery Today* 13, nº 19 (1 octobre 2008): 831-41. https://doi.org/10.1016/j.drudis.2008.06.006.

Davis, Mindy I., Jeremy P. Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M. Wodicka, Gabriel Pallares, Michael Hocker, Daniel K. Treiber, et Patrick P. Zarrinkar. « Comprehensive Analysis of Kinase Inhibitor Selectivity ». *Nature Biotechnology* 29, nº 11 (novembre 2011): 1046-51. https://doi.org/10.1038/nbt.1990.

« Daylight Theory: SMARTS - A Language for Describing Molecular Patterns ». Consulté le 12 juin 2019. https://www.daylight.com/dayhtml/doc/theory/theory.smarts.html#RTFToC35.

De Verdier, C. H. « Isolation of Phosphothreonine from Bovine Casein ». *Nature* 170, nº 4332 (8 novembre 1952): 804-5.

« Demande initiale d'AMM - ANSM : Agence nationale de sécurité du médicament et des produits de santé ». Consulté le 21 mars 2019. https://www.ansm.sante.fr/Activites/Autorisations-de-Mise-sur-le-Marche-AMM/Demande-initiale-d-AMM/(offset)/1.

Deng, Zhan, Claudio Chuaqui, et Juswinder Singh. « Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions ». *Journal of Medicinal Chemistry* 47, nº 2 (1 janvier 2004): 337-44. https://doi.org/10.1021/jm030331x.

Dijkstra, E. W. « A Note on Two Problems in Connexion with Graphs ». *Numerische Mathematik* 1, nº 1 (1 décembre 1959): 269-71. https://doi.org/10.1007/BF01386390.

DiMasi, Joseph A., Henry G. Grabowski, et Ronald W. Hansen. « Innovation in the pharmaceutical industry: New estimates of R&D costs ». *Journal of Health Economics* 47 (1 mai 2016): 20-33. https://doi.org/10.1016/j.jhealeco.2016.01.012.

DiMasi, Joseph A, Ronald W Hansen, et Henry G Grabowski. « The price of innovation: new estimates of drug development costs ». *Journal of Health Economics* 22, nº 2 (1 mars 2003): 151-85. https://doi.org/10.1016/S0167-6296(02)00126-1.

« Dimensions ». Consulté le 28 juin 2019. https://app.dimensions.ai/discover/publication.

Donald, James R., et William P. Unsworth. « Ring-Expansion Reactions in the Synthesis of Macrocycles and Medium-Sized Rings ». *Chemistry – A European Journal* 23, nº 37 (3 juillet 2017): 8780-99. https://doi.org/10.1002/chem.201700467.

Durant, Joseph L., Burton A. Leland, Douglas R. Henry, et James G. Nourse. « Reoptimization of MDL Keys for Use in Drug Discovery ». *Journal of Chemical Information and Computer Sciences* 42, nº 6 (1 novembre 2002): 1273-80. https://doi.org/10.1021/ci010132r.

Drugs.com. « Balversa (Erdafitinib) FDA Approval History ». Consulté le 2 juillet 2019. https://www.drugs.com/history/balversa.html.

Drugs.com. « Venclexta (Venetoclax) FDA Approval History ». Consulté le 2 juillet 2019. https://www.drugs.com/history/venclexta.html.

Drugs.com. « Viagra: How a Little Blue Pill Changed the World ». Consulté le 17 avril 2019. https://www.drugs.com/slideshow/viagra-little-blue-pill-1043.

Drugs.com. « Zelboraf (Vemurafenib) FDA Approval History ». Consulté le 2 juillet 2019. https://www.drugs.com/history/zelboraf.html.

Erlanson, Dan. « Practical Fragments: Fragments in the clinic: 2018 edition ». *Practical Fragments* (blog), 6 octobre 2018. https://practicalfragments.blogspot.com/2018/10/fragments-in-clinic-2018-edition.html.

Erlanson, Daniel A., Stephen W. Fesik, Roderick E. Hubbard, Wolfgang Jahnke, et Harren Jhoti. « Twenty Years on: The Impact of Fragments on Drug Discovery ». *Nature Reviews Drug Discovery* 15, nº 9 (septembre 2016): 605-19. https://doi.org/10.1038/nrd.2016.109.

Ermert, Philipp. « Design, Properties and Recent Application of Macrocycles in Medicinal Chemistry ». Text, octobre 2017. https://doi.org/info:doi/10.2533/chimia.2017.678.

Ertl, Peter, et Ansgar Schuffenhauer. « Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions ». *Journal of Cheminformatics* 1, nº 1 (10 juin 2009): 8. https://doi.org/10.1186/1758-2946-1-8.

« European Collection of Authenticated Cell Cultures (ECACC) ». Consulté le 10 octobre 2019. https://www.phe-culturecollections.org.uk/collections/ecacc.aspx.

European Medicines Agency. « Fotivda ». Text, 17 septembre 2018. https://www.ema.europa.eu/en/medicines/human/EPAR/fotivda.

Fabbro, Doriano. « 25 Years of Small Molecular Weight Kinase Inhibitors: Potentials and Limitations ». *Molecular Pharmacology* 87, nº 5 (1 mai 2015): 766-75. https://doi.org/10.1124/mol.114.095489.

Fabbro, Doriano, Sandra W Cowan-Jacob, et Henrik Moebitz. « Ten things you should know about protein kinases: IUPHAR Review 14 ». *British Journal of Pharmacology* 172, nº 11 (juin 2015): 2675-2700. https://doi.org/10.1111/bph.13096.

Fawcett, Tom. « An introduction to ROC analysis ». *Pattern Recognition Letters*, ROC Analysis in Pattern Recognition, 27, nº 8 (1 juin 2006): 861-74. https://doi.org/10.1016/j.patrec.2005.10.010.

Ferguson, Fleur M., et Nathanael S. Gray. « Kinase Inhibitors: The Road Ahead ». *Nature Reviews Drug Discovery* 17, nº 5 (16 mars 2018): 353-77. https://doi.org/10.1038/nrd.2018.21.

Ferreira, Leonardo G., Ricardo N. Dos Santos, Glaucius Oliva, et Adriano D. Andricopulo. « Molecular Docking and Structure-Based Drug Design Strategies ». *Molecules* 20, nº 7 (juillet 2015): 13384-421. https://doi.org/10.3390/molecules200713384.

Finn, Robert D., Penelope Coggill, Ruth Y. Eberhardt, Sean R. Eddy, Jaina Mistry, Alex L. Mitchell, Simon C. Potter, et al. « The Pfam Protein Families Database: Towards a More Sustainable Future ». *Nucleic Acids Research* 44, nº D1 (4 janvier 2016): D279-85. https://doi.org/10.1093/nar/gkv1344.

Flament, Claude. *Théorie des graphes et structures sociales*. Walter de Gruyter GmbH & Co KG, 2017.

Fosgerau, Keld, et Torsten Hoffmann. « Peptide therapeutics: current status and future directions ». *Drug Discovery Today* 20, nº 1 (1 janvier 2015): 122-28. https://doi.org/10.1016/j.drudis.2014.10.003.

« Fragment Libraries | FBDD screening compounds | Life Chemicals ». Consulté le 18 septembre 2019. https://lifechemicals.com/screening-libraries/fragment-libraries.

G. de la Torre, Beatriz, et Fernando Albericio. « The Pharmaceutical Industry in 2018. An Analysis of FDA Drug Approvals from the Perspective of Molecules ». *Molecules* 24, nº 4 (janvier 2019): 809. https://doi.org/10.3390/molecules24040809.

Gally, José-Manuel. « Développement d'outils de chémoinformatique pour l'identification d'inhibiteurs de protéines kinases à partir de fragments. » Orléans, 2017.

Gao, Wei, Hualong Wu, Muhammad Kamran Siddiqui, et Abdul Qudair Baig. « Study of biological networks using graph theory ». *Saudi Journal of Biological Sciences* 25, nº 6 (1 septembre 2018): 1212-19. https://doi.org/10.1016/j.sjbs.2017.11.022.

Gaulton, Anna, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, et al. « The ChEMBL Database in 2017 ». *Nucleic Acids Research* 45, nº D1 (4 janvier 2017): D945-54. https://doi.org/10.1093/nar/gkw1074.

Ghandi, Mahmoud, Franklin W. Huang, Judit Jané-Valbuena, Gregory V. Kryukov, Christopher C. Lo, E. Robert McDonald, Jordi Barretina, et al. « Next-Generation Characterization of the Cancer Cell Line Encyclopedia ». *Nature* 569, nº 7757 (mai 2019): 503-8. https://doi.org/10.1038/s41586-019-1186-3.

Ginalski, Krzysztof. « Comparative modeling for protein structure prediction ». *Current Opinion in Structural Biology*, Theory and simulation/Macromolecular assemblages, 16, nº 2 (1 avril 2006): 172-77. https://doi.org/10.1016/j.sbi.2006.02.003.

Giordanetto, Fabrizio, et Jan Kihlberg. « Macrocyclic Drugs and Clinical Candidates: What Can Medicinal Chemists Learn from Their Properties? » *Journal of Medicinal Chemistry* 57, nº 2 (23 janvier 2014): 278-95. https://doi.org/10.1021/jm400887j.

Gohlke, H., M. Hendlich, et G. Klebe. « Knowledge-Based Scoring Function to Predict Protein-Ligand Interactions ». *Journal of Molecular Biology* 295, nº 2 (14 janvier 2000): 337-56. https://doi.org/10.1006/jmbi.1999.3371.

Goodford, P. J. « A computational procedure for determining energetically favorable binding sites on biologically important macromolecules ». *Journal of Medicinal Chemistry* 28, nº 7 (1 juillet 1985): 849-57. https://doi.org/10.1021/jm00145a002.

Goodsell, David S., Teresa A. Larsen, et T. J. O'Donnell. « 1994 Molecular Graphics Art Show and Video Show ». *Journal of Molecular Graphics* 13, nº 4 (1 août 1995): 223-34. https://doi.org/10.1016/0263-7855(95)00036-6.

Google Developers. « Classification : justesse | Cours d'initiation au machine learning ». Consulté le 24 septembre 2019. https://developers.google.com/machine-learning/crash-course/classification/accuracy.

Gorre, Mercedes E., Mansoor Mohammed, Katharine Ellwood, Nicholas Hsu, Ron Paquette, P. Nagesh Rao, et Charles L. Sawyers. « Clinical Resistance to STI-571 Cancer Therapy Caused by BCR-ABL Gene Mutation or Amplification ». *Science* 293, nº 5531 (3 août 2001): 876-80. https://doi.org/10.1126/science.1062538.

Greenidge, Paulette A., Richard A. Lewis, et Peter Ertl. « Boosting Pose Ranking Performance via Rescoring with MM-GBSA ». *Chemical Biology & Drug Design* 88, nº 3 (2016): 317-28. https://doi.org/10.1111/cbdd.12763.

Gross, Stefan, Rami Rahal, Nicolas Stransky, Christoph Lengauer, et Klaus P. Hoeflich. « Targeting Cancer with Kinase Inhibitors ». *The Journal of Clinical Investigation* 125, nº 5 (1 mai 2015): 1780-89. https://doi.org/10.1172/JCI76094.

« Guidelines : ICH ». Consulté le 18 mars 2019. http://www.ich.org/products/guidelines.

Hajduk, Philip J., et Jonathan Greer. « A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned ». *Nature Reviews Drug Discovery* 6, nº 3 (mars 2007): 211-19. https://doi.org/10.1038/nrd2220.

Halgren, Thomas A. « Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94 ». *Journal of Computational Chemistry* 17, nº 5-6 (1 avril 1996): 490-519. https://doi.org/10.1002/(SICI)1096-987X(199604)17:5/6<490::AID-JCC1>3.0.CO;2-P.

Hammarsten, Olof. « Zur Frage, ob das Caseïn ein einheitlicher Stoff sei. » *Zeitschrift für physiologische Chemie* 7, nº 3 (1883): 227–273. https://doi.org/10.1515/bchm1.1883.7.3.227.

Hann, Michael M. « Molecular Obesity, Potency and Other Addictions in Drug Discovery ». *MedChemComm* 2, nº 5 (1 mai 2011): 349-55. https://doi.org/10.1039/C1MD00017A.

Hann, Michael M., Andrew R. Leach, et Gavin Harper. « Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery ». *Journal of Chemical Information and Computer Sciences* 41, nº 3 (1 mai 2001): 856-64. https://doi.org/10.1021/ci000403i.

Heinis, Christian. « Drug Discovery: Tools and Rules for Macrocycles ». *Nature Chemical Biology* 10, nº 9 (septembre 2014): 696-98. https://doi.org/10.1038/nchembio.1605.

Hidaka, Hiroyoshi, Masaki Inagaki, Sachiyo Kawamoto, et Yasuharu Sasaki. « Isoquinolinesulfonamides, novel and potent inhibitors of cyclic nucleotide-dependent protein kinase and protein kinase C ». *Biochemistry* 23, nº 21 (9 octobre 1984): 5036-41. https://doi.org/10.1021/bi00316a032.

Hill, Stephen J. « G-protein-coupled receptors: past, present and future ». *British Journal of Pharmacology* 147, nº Suppl 1 (janvier 2006): S27-37. https://doi.org/10.1038/sj.bjp.0706455.

Hillisch, Alexander, Nikolaus Heinrich, et Hanno Wild. « Computational Chemistry in the Pharmaceutical Industry: From Childhood to Adolescence ». *ChemMedChem* 10, nº 12 (2015): 1958-62. https://doi.org/10.1002/cmdc.201500346.

Hopkins, Andrew L., Colin R. Groom, et Alexander Alex. « Ligand efficiency: a useful metric for lead selection ». *Drug Discovery Today* 9, nº 10 (15 mai 2004): 430-31. https://doi.org/10.1016/S1359-6446(04)03069-7.

Huang, Niu, Brian K. Shoichet, et John J. Irwin. « Benchmarking Sets for Molecular Docking ». *Journal of Medicinal Chemistry* 49, nº 23 (16 novembre 2006): 6789-6801. https://doi.org/10.1021/jm0608356.

Hubbard, Stevan R. « The Insulin Receptor: Both a Prototypical and Atypical Receptor Tyrosine Kinase ». *Cold Spring Harbor Perspectives in Biology* 5, nº 3 (mars 2013). https://doi.org/10.1101/cshperspect.a008946.

Hung, Alvin W., Alex Ramek, Yikai Wang, Taner Kaya, J. Anthony Wilson, Paul A. Clemons, et Damian W. Young. « Route to Three-Dimensional Fragments Using Diversity-Oriented Synthesis ». *Proceedings of the National Academy of Sciences* 108, nº 17 (26 avril 2011): 6799-6804. https://doi.org/10.1073/pnas.1015271108.

Hunter, T. « A Thousand and One Protein Kinases ». *Cell* 50, nº 6 (11 septembre 1987): 823-29.

Huse, Morgan, et John Kuriyan. « The Conformational Plasticity of Protein Kinases ». *Cell* 109, nº 3 (3 mai 2002): 275-82. https://doi.org/10.1016/S0092-8674(02)00741-9.

IFOP. « Cosmétiques : le boom du bio ? » Consulté le 17 juillet 2019. https://www.ifop.com/publication/cosmetiques-le-boom-du-bio/.

InfoQ. « Les Bases Orientées Graphes, NoSQL et Neo4j ». Consulté le 19 septembre 2019. https://www.infoq.com/fr/articles/graph-nosql-neo4j/.

Inserm - La science pour la santé. « Cancer : la médecine personnalisée est possible ». Consulté le 26 septembre 2019. https://www.inserm.fr/cancer-medecine-personnalisee-est-possible.

Illergård, Kristoffer, David H. Ardell, et Arne Elofsson. « Structure Is Three to Ten Times More Conserved than Sequence--a Study of Structural Response in Protein Cores ». *Proteins* 77, nº 3 (15 novembre 2009): 499-508. https://doi.org/10.1002/prot.22458.

« Illustrated Glossary of Organic Chemistry - Bond angle ». Consulté le 19 septembre 2019. http://www.chem.ucla.edu/~harding/IGOC/B/bond_angle.html.

Imai, Kohzoh, et Akinori Takaoka. « Comparing Antibody and Small-Molecule Therapies for Cancer ». *Nature Reviews Cancer* 6, nº 9 (septembre 2006): 714. https://doi.org/10.1038/nrc1913.

Jeffrey, George A. *An Introduction to Hydrogen Bonding*. Oxford University Press, 1997.

Jencks, William P. « On the Attribution and Additivity of Binding Energies ». *Proceedings of the National Academy of Sciences* 78, nº 7 (1 juillet 1981): 4046-50. https://doi.org/10.1073/pnas.78.7.4046.

Jhoti, Harren, Glyn Williams, David C. Rees, et Christopher W. Murray. « The “rule of Three” for Fragment-Based Drug Discovery: Where Are We Now? » *Nature Reviews Drug Discovery* 12, nº 8 (août 2013): 644. https://doi.org/10.1038/nrd3926-c1.

Jones, G., P. Willett, R. C. Glen, A. R. Leach, et R. Taylor. « Development and Validation of a Genetic Algorithm for Flexible Docking ». *Journal of Molecular Biology* 267, nº 3 (4 avril 1997): 727-48. https://doi.org/10.1006/jmbi.1996.0897.

Kellenberger, Esther, Jordi Rodrigo, Pascal Muller, et Didier Rognan. « Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy ». *Proteins: Structure, Function, and Bioinformatics* 57, nº 2 (2004): 225-42. https://doi.org/10.1002/prot.20149.

Keserű, György M., Daniel A. Erlanson, György G. Ferenczy, Michael M. Hann, Christopher W. Murray, et Stephen D. Pickett. « Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia ». *Journal of Medicinal Chemistry* 59, nº 18 (22 septembre 2016): 8189-8206. https://doi.org/10.1021/acs.jmedchem.6b00197.

Khanna, Ish. « Drug discovery in pharmaceutical industry: productivity challenges and trends ». *Drug Discovery Today* 17, nº 19 (1 octobre 2012): 1088-1102. https://doi.org/10.1016/j.drudis.2012.05.007.

Kim, Sunghwan, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, et al. « PubChem 2019 Update: Improved Access to Chemical Data ». *Nucleic Acids Research* 47, nº D1 (8 janvier 2019): D1102-9. https://doi.org/10.1093/nar/gky1033.

Kirchmair, Johannes, Patrick Markt, Simona Distinto, Gerhard Wolber, et Thierry Langer. « Evaluation of the Performance of 3D Virtual Screening Protocols: RMSD Comparisons, Enrichment Assessments, and Decoy Selection—What Can We Learn from Earlier Mistakes? » *Journal of Computer-Aided Molecular Design* 22, nº 3 (1 mars 2008): 213-28. https://doi.org/10.1007/s10822-007-9163-6.

Knight, Zachary A., et Kevan M. Shokat. « Features of Selective Kinase Inhibitors ». *Chemistry & Biology* 12, nº 6 (1 juin 2005): 621-37. https://doi.org/10.1016/j.chembiol.2005.04.011.

Konteatis, Zenon D. « In silico fragment-based drug design ». *Expert Opinion on Drug Discovery* 5, nº 11 (1 novembre 2010): 1047-65. https://doi.org/10.1517/17460441.2010.523697.

Kooistra, Albert J., Georgi K. Kanev, Oscar P. J. van Linden, Rob Leurs, Iwan J. P. de Esch, et Chris de Graaf. « KLIFS: A Structural Kinase-Ligand Interaction Database ». *Nucleic Acids Research* 44, nº D1 (4 janvier 2016): D365-71. https://doi.org/10.1093/nar/gkv1082.

Korb, Oliver, Thomas Stützle, et Thomas E. Exner. « Empirical Scoring Functions for Advanced Protein-Ligand Docking with PLANTS ». *Journal of Chemical Information and Modeling* 49, nº 1 (janvier 2009): 84-96. https://doi.org/10.1021/ci800298z.

Kornev, Alexandr P., Susan S. Taylor, et Lynn F. Ten Eyck. « A Helix Scaffold for the Assembly of Active Protein Kinases ». *Proceedings of the National Academy of Sciences of the United States of America* 105, nº 38 (23 septembre 2008): 14377-82. https://doi.org/10.1073/pnas.0807988105.

Krebs, Edwin G., Donald J. Graves, et Edmond H. Fischer. « Factors Affecting the Activity of Muscle Phosphorylase b Kinase ». *Journal of Biological Chemistry* 234, nº 11 (11 janvier 1959): 2867-73.

Krimm, Isabelle. « Le criblage de fragments: Une voie prometteuse pour la conception de médicaments ». *médecine/sciences* 31, nº 2 (février 2015): 197-202. https://doi.org/10.1051/medsci/20153102017.

Kristen, Arnt V, Senda Ajroud-Driss, Isabel Conceição, Peter Gorevic, Theodoros Kyriakides, et Laura Obici. « Patisiran, an RNAi therapeutic for the treatment of hereditary transthyretin-mediated amyloidosis ». *Neurodegenerative Disease Management* 9, nº 1 (27 novembre 2018): 5-23. https://doi.org/10.2217/nmt-2018-0033.

Kuntz, I. D., K. Chen, K. A. Sharp, et P. A. Kollman. « The Maximal Affinity of Ligands ». *Proceedings of the National Academy of Sciences* 96, nº 18 (31 août 1999): 9997-10002. https://doi.org/10.1073/pnas.96.18.9997.

Kuriyan, John, et David Eisenberg. « The Origin of Protein Interactions and Allostery in Colocalization ». *Nature* 450 (12 décembre 2007): 983-90. https://doi.org/10.1038/nature06524.

K. Yudin, Andrei. « Macrocycles: Lessons from the Distant Past, Recent Developments, and Future Directions ». *Chemical Science* 6, nº 1 (2015): 30-49. https://doi.org/10.1039/C4SC03089C.

Lacapère, Jean-Jacques, Eva Pebay-Peyroula, Jean-Michel Neumann, et Catherine Etchebest. « Determining Membrane Protein Structures: Still a Challenge! » *Trends in Biochemical Sciences* 32, nº 6 (1 juin 2007): 259-70. https://doi.org/10.1016/j.tibs.2007.04.001.

Lagarde, Nathalie, Jean-François Zagury, et Matthieu Montes. « Benchmarking Data Sets for the Evaluation of Virtual Ligand Screening Methods: Review and Perspectives ». *Journal of Chemical Information and Modeling* 55, nº 7 (27 juillet 2015): 1297-1307. https://doi.org/10.1021/acs.jcim.5b00090.

Lahana, Roger. « How many leads from HTS? » *Drug Discovery Today* 4, nº 10 (1 octobre 1999): 447-48. https://doi.org/10.1016/S1359-6446(99)01393-8.

Lamoree, Bas, et Roderick E. Hubbard. « Current perspectives in fragment-based lead discovery (FBLD) ». *Essays in Biochemistry* 61, nº 5 (8 novembre 2017): 453-64. https://doi.org/10.1042/EBC20170028.

Lavecchia, A., et C. Di Giovanni. « Virtual Screening Strategies in Drug Discovery: A Critical Review ». *Current Medicinal Chemistry* 20, nº 23 (2013): 2839-60. https://doi.org/10.2174/09298673113209990001.

« Le cancer devient la première cause de décès dans les pays riches ». *Le Monde.fr*, 3 septembre 2019. https://www.lemonde.fr/societe/article/2019/09/03/le-cancer-devient-la-premiere-cause-de-deces-dans-les-pays-riches_5505928_3224.html.

LearnAnalytics. « What Is NoSQL and Is It the next Big Trend in Databases? » *Big Data Path* (blog), 27 mars 2018. https://bigdatapath.wordpress.com/2018/03/27/what-is-nosql-and-is-it-the-next-big-trend-in-databases/.

Levene, P. A., et C. L. Alsberg. « The Cleavage Products of Vitellin ». *Journal of Biological Chemistry* 2, nº 1 (8 janvier 1906): 127-33.

Lewell, Xiao Qing, Duncan B. Judd, Stephen P. Watson, et Michael M. Hann. « RECAP Retrosynthetic Combinatorial Analysis Procedure: A Powerful New Technique for Identifying Privileged Molecular Fragments with Useful Applications in Combinatorial Chemistry ». *Journal of Chemical Information and Computer Sciences* 38, nº 3 (18 mai 1998): 511-22. https://doi.org/10.1021/ci970429i.

Lexa, Katrina W., et Heather A. Carlson. « Protein Flexibility in Docking and Surface Mapping ». *Quarterly reviews of biophysics* 45, nº 3 (août 2012): 301-43. https://doi.org/10.1017/S0033583512000066.

« Library for Fragment-Based Drug Discovery – Prestwick Chemical ». Consulté le 18 septembre 2019. http://www.prestwickchemical.com/libraries-screening-lib-drug-frag.html.

Lindsley, Craig W. « New Statistics on the Cost of New Drug Development and the Trouble with CNS Drugs ». *ACS Chemical Neuroscience* 5, nº 12 (17 décembre 2014): 1142-1142. https://doi.org/10.1021/cn500298z.

« L'innovation thérapeutique, un processus long et coûteux ». Consulté le 5 mars 2019. https://www.leem.org/linnovation-therapeutique-un-processus-long-et-couteux-0.

Lionta, Evanthia, George Spyrou, Demetrios K. Vassilatis, et Zoe Cournia. « Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances ». *Current Topics in Medicinal Chemistry* 14, nº 16 (août 2014): 1923-38. https://doi.org/10.2174/1568026614666140929124445.

Lipinski, Christopher A, Franco Lombardo, Beryl W Dominy, et Paul J Feeney. « Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings ». *Advanced Drug Delivery Reviews*, Special issue dedicated to Dr. Eric Tomlinson, Advanced Drug Delivery Reviews, A Selection of the Most Highly Cited Articles, 1991-1998, 46, nº 1 (1 mars 2001): 3-26. https://doi.org/10.1016/S0169-409X(00)00129-0.

Lipmann, Fritz A., et P. A. Levene. « Serinephosphoric Acid Obtained on Hydrolysis of Vitellinic Acid ». *Journal of Biological Chemistry* 98, nº 1 (10 janvier 1932): 109-14.

Lucas, Edouard. *Récréations mathématiques (2ème éd.)*, 1891. https://gallica.bnf.fr/ark:/12148/bpt6k3943s.

Macarron, Ricardo. « Critical review of the role of HTS in drug discovery ». *Drug Discovery Today* 11, nº 7 (1 avril 2006): 277-79. https://doi.org/10.1016/j.drudis.2006.02.001.

Madhavi Sastry, G., Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, et Woody Sherman. « Protein and Ligand Preparation: Parameters, Protocols, and Influence on Virtual Screening Enrichments ». *Journal of Computer-Aided Molecular Design* 27, nº 3 (1 mars 2013): 221-34. https://doi.org/10.1007/s10822-013-9644-8.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter, et S. Sudarsanam. « The Protein Kinase Complement of the Human Genome ». *Science* 298, nº 5600 (6 décembre 2002): 1912-34. https://doi.org/10.1126/science.1075762.

« Marché mondial ». Consulté le 17 avril 2019. https://www.leem.org/marche-mondial.

Marcou, Gilles, et Didier Rognan. « Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints ». *Journal of Chemical Information and Modeling* 47, nº 1 (1 janvier 2007): 195-207. https://doi.org/10.1021/ci600342e.

Mattos, C., et D. Ringe. « Locating and Characterizing Binding Sites on Proteins ». *Nature Biotechnology* 14, nº 5 (mai 1996): 595-99. https://doi.org/10.1038/nbt0596-595.

Mayer, B. J., H. Hirai, et R. Sakai. « Evidence That SH2 Domains Promote Processive Phosphorylation by Protein-Tyrosine Kinases ». *Current Biology: CB* 5, nº 3 (1 mars 1995): 296-305.

MedchemExpress.com. « Fragment Library-MedchemExpress ». Consulté le 18 septembre 2019. https://www.medchemexpress.com/screening/Fragment_Library.html.

Medscape. « Tivozanib for Kidney Cancer Rejected by FDA ». Consulté le 21 mars 2019. http://www.medscape.com/viewarticle/805578.

Meng, Xuan-Yu, Hong-Xing Zhang, Mihaly Mezei, et Meng Cui. « Molecular Docking: A powerful approach for structure-based drug discovery ». *Current computer-aided drug design* 7, nº 2 (1 juin 2011): 146-57.

Miranker, Andrew, et Martin Karplus. « Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method ». *Proteins: Structure, Function, and Bioinformatics* 11, nº 1 (1991): 29-34. https://doi.org/10.1002/prot.340110104.

Mishra, Aditya. « Metrics to Evaluate Your Machine Learning Algorithm ». Medium, 1 novembre 2018. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234.

Mitchell, Melanie. *An Introduction to Genetic Algorithms*. Cambridge, MA, USA: MIT Press, 1998.

« Molecular Descriptors for Chemoinformatics | Methods and Principles in Medicinal Chemistry ». Consulté le 19 avril 2019. https://onlinelibrary.wiley.com/doi/book/10.1002/9783527628766.

« Molecular Descriptors Software ». Consulté le 19 avril 2019. http://www.moleculardescriptors.eu/softwares/softwares.htm.

Mooij, Wijnand T. M., et Marcel L. Verdonk. « General and Targeted Statistical Potentials for Protein–Ligand Interactions ». *Proteins: Structure, Function, and Bioinformatics* 61, nº 2 (2005): 272-87. https://doi.org/10.1002/prot.20588.

Mucs, Daniel, Richard A. Bryce, et Pascal Bonnet. « Application of Shape-Based and Pharmacophore-Based in Silico Screens for Identification of Type II Protein Kinase Inhibitors ». *Journal of Computer-Aided Molecular Design* 25, nº 6 (1 juin 2011): 569-81. https://doi.org/10.1007/s10822-011-9442-0.

Muegge, I., et Y. C. Martin. « A General and Fast Scoring Function for Protein-Ligand Interactions: A Simplified Potential Approach ». *Journal of Medicinal Chemistry* 42, nº 5 (11 mars 1999): 791-804. https://doi.org/10.1021/jm980536j.

Mullard, Asher. « 2018 FDA Drug Approvals ». *Nature Reviews Drug Discovery* 18 (15 janvier 2019): 85. https://doi.org/10.1038/d41573-019-00014-x.

Müller, Ilka. « Guidelines for the successful generation of protein–ligand complex crystals ». *Acta Crystallographica. Section D, Structural Biology* 73, nº Pt 2 (1 février 2017): 79-92. https://doi.org/10.1107/S2059798316020271.

Muller, P. « Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994) ». *Pure and Applied Chemistry* 66, nº 5 (2009): 1077–1184. https://doi.org/10.1351/pac199466051077.

Murray, Christopher W, et Tom L Blundell. « Structural biology in fragment-based drug design ». *Current Opinion in Structural Biology*, Membranes / Engineering and design, 20, nº 4 (1 août 2010): 497-507. https://doi.org/10.1016/j.sbi.2010.04.003.

Mysinger, Michael M., Michael Carchia, John. J. Irwin, et Brian K. Shoichet. « Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking ». *Journal of Medicinal Chemistry* 55, nº 14 (26 juillet 2012): 6582-94. https://doi.org/10.1021/jm300687e.

Nagar, Bhushan, Oliver Hantschel, Markus Seeliger, Jason M. Davies, William I. Weis, Giulio Superti-Furga, et John Kuriyan. « Organization of the SH3-SH2 Unit in Active and Inactive Forms of the c-Abl Tyrosine Kinase ». *Molecular Cell* 21, nº 6 (17 mars 2006): 787-98. https://doi.org/10.1016/j.molcel.2006.01.035.

NobelPrize.org. « The Nobel Prize in Physiology or Medicine 2018 ». Consulté le 5 juin 2019. https://www.nobelprize.org/prizes/medicine/2018/summary/.

NobelPrize.org. « All Nobel Prizes in Physiology or Medicine ». Consulté le 12 avril 2019. https://www.nobelprize.org/prizes/lists/all-nobel-laureates-in-physiology-or-medicine/.

Oprea, Tudor I, et Hans Matter. « Integrating virtual screening in lead discovery ». *Current Opinion in Chemical Biology* 8, nº 4 (1 août 2004): 349-58. https://doi.org/10.1016/j.cbpa.2004.06.008.

Overington, John P., Bissan Al-Lazikani, et Andrew L. Hopkins. « How Many Drug Targets Are There? » *Nature Reviews Drug Discovery* 5, nº 12 (décembre 2006): 993. https://doi.org/10.1038/nrd2199.

Pagadala, Nataraj S., Khajamohiddin Syed, et Jack Tuszynski. « Software for molecular docking: a review ». *Biophysical Reviews* 9, nº 2 (16 janvier 2017): 91-102. https://doi.org/10.1007/s12551-016-0247-1.

« Parcours d'un graphe ». Consulté le 18 septembre 2019. https://haltode.fr/algo/structure/graphe/parcours.html.

Patel, Disha, Joseph D. Bauman, et Eddy Arnold. « Advantages of Crystallographic Fragment Screening: Functional and Mechanistic Insights from a Powerful Platform for Efficient Drug Discovery ». *Progress in biophysics and molecular biology* 116, nº 0 (2014): 92-100. https://doi.org/10.1016/j.pbiomolbio.2014.08.004.

Ponomarenko, Elena A., Ekaterina V. Poverennaya, Ekaterina V. Ilgisonis, Mikhail A. Pyatnitskiy, Arthur T. Kopylov, Victor G. Zgoda, Andrey V. Lisitsa, et Alexander I. Archakov. « The Size of the Human Proteome: The Width and Depth ». *International Journal of Analytical Chemistry* 2016 (2016). https://doi.org/10.1155/2016/7436849.

PricewaterhouseCoopers. « Global Top 100 Companies 2019 ». PwC. Consulté le 17 octobre 2019. https://www.pwc.com/gx/en/services/audit-assurance/publications/global-top-100-companies-2019.html.

« Protein Kinases: Human Protein Kinases Overview | CST ». Consulté le 5 juin 2019. https://www.cellsignal.com/contents/science-protein-kinases/protein-kinases-human-protein-kinases-overview/kinases-human-protein.

Pushpakom, Sudeep, Francesco Iorio, Patrick A. Eyers, K. Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, et al. « Drug Repurposing: Progress, Challenges and Recommendations ». *Nature Reviews Drug Discovery* 18, nº 1 (janvier 2019): 41-58. https://doi.org/10.1038/nrd.2018.168.

Quintaje, Silvia Braconi, et Sandra Orchard. « The Annotation of Both Human and Mouse Kinomes in UniProtKB/Swiss-Prot: One Small Step in Manual Annotation, One Giant Leap for Full Comprehension of Genomes ». *Molecular & Cellular Proteomics* 7, nº 8 (1 août 2008): 1409-19. https://doi.org/10.1074/mcp.R700001-MCP200.

Ray, Louis C., et Russell A. Kirsch. « Finding Chemical Records by Digital Computers ». *Science* 126, nº 3278 (25 octobre 1957): 814-19. https://doi.org/10.1126/science.126.3278.814.

« Recherche et développement ». Consulté le 21 mars 2019. https://www.leem.org/recherche-et-developpement.

Renaud, Jean-Paul, Chun-wa Chung, U. Helena Danielson, Ursula Egner, Michael Hennig, Roderick E. Hubbard, et Herbert Nar. « Biophysics in Drug Discovery: Impact, Challenges and Opportunities ». *Nature Reviews Drug Discovery* 15, nº 10 (octobre 2016): 679-98. https://doi.org/10.1038/nrd.2016.123.

Research, Center for Drug Evaluation and. « FDA Approves Lorlatinib for Second- or Third-Line Treatment of ALK-Positive Metastatic NSCLC ». *FDA*, 2 septembre 2019. http://www.fda.gov/drugs/fda-approves-lorlatinib-second-or-third-line-treatment-alk-positive-metastatic-nsclc.

Riniker, Sereina, et Gregory A. Landrum. « Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation ». *Journal of Chemical Information and Modeling* 55, nº 12 (28 décembre 2015): 2562-74. https://doi.org/10.1021/acs.jcim.5b00654.

Roberts, Benjamin C., et Ricardo L. Mancera. « Ligand−Protein Docking with Water Molecules ». *Journal of Chemical Information and Modeling* 48, nº 2 (1 février 2008): 397-408. https://doi.org/10.1021/ci700285e.

Robson-Tull, Jacob. « Biophysical Screening in Fragment-Based Drug Design: A Brief Overview ». *Bioscience Horizons: The International Journal of Student Research* 11 (1 janvier 2018). https://doi.org/10.1093/biohorizons/hzy015.

Roche, Daniel Barry, Danielle Allison Brackenridge, et Liam James McGuffin. « Proteins and Their Interacting Partners: An Introduction to Protein–Ligand Binding Site Prediction Methods ». *International Journal of Molecular Sciences* 16, nº 12 (décembre 2015): 29829-42. https://doi.org/10.3390/ijms161226202.

Roche-Lestienne, Catherine, et Claude Preudhomme. « Résistance au Glivec® : actualités ». *Hématologie* 13, nº 1 (1 janvier 2007): 43-53. https://doi.org/10.1684/hma.2007.0087.

Rogers, David, et Mathew Hahn. « Extended-Connectivity Fingerprints ». *Journal of Chemical Information and Modeling* 50, nº 5 (24 mai 2010): 742-54. https://doi.org/10.1021/ci100050t.

Rognan, Didier. « Fragment-Based Approaches and Computer-Aided Drug Discovery ». *Fragment-Based Drug Discovery and X-Ray Crystallography*, édité par Thomas G. Davies et Marko Hyvönen, 201-22. Topics in Current Chemistry. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. https://doi.org/10.1007/128_2011_182.

Rognan, Didier. « Le criblage virtuel par docking moléculaire ». *Chemogénomique, des petites molécules pour explorer le vivant*, 258. EDP Sciences, Collection Grenoble Sciences., 2007.

Romasanta, Angelo K. S., Peter van der Sijde, Iina Hellsten, Roderick E. Hubbard, Gyorgy M. Keseru, Jacqueline van Muijlwijk-Koezen, et Iwan J. P. de Esch. « When Fragments Link: A Bibliometric Perspective on the Development of Fragment-Based Drug Discovery ». *Drug Discovery Today* 23, nº 9 (1 septembre 2018): 1596-1609. https://doi.org/10.1016/j.drudis.2018.05.004.

Roskoski, Robert. « A historical overview of protein kinases and their targeted small molecule inhibitors ». *Pharmacological Research* 100 (1 octobre 2015): 1-23. https://doi.org/10.1016/j.phrs.2015.07.010.

Roskoski, Robert. « Classification of small molecule protein kinase inhibitors based upon the structures of their drug-enzyme complexes ». *Pharmacological Research* 103 (1 janvier 2016): 26-48. https://doi.org/10.1016/j.phrs.2015.10.021.

Ruiz-Carmona, Sergio, Daniel Alvarez-Garcia, Nicolas Foloppe, A. Beatriz Garmendia-Doval, Szilveszter Juhos, Peter Schmidtke, Xavier Barril, Roderick E. Hubbard, et S. David Morley. « RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids ». *PLOS Computational Biology* 10, nº 4 (10 avril 2014): e1003571. https://doi.org/10.1371/journal.pcbi.1003571.

Santos, Rita, Oleg Ursu, Anna Gaulton, A. Patrícia Bento, Ramesh S. Donadi, Cristian G. Bologa, Anneli Karlsson, et al. « A Comprehensive Map of Molecular Drug Targets ». *Nature Reviews Drug Discovery* 16, nº 1 (janvier 2017): 19-34. https://doi.org/10.1038/nrd.2016.230.

Scott, Duncan E., Anthony G. Coyne, Sean A. Hudson, et Chris Abell. « Fragment-Based Approaches in Drug Discovery and Chemical Biology ». *Biochemistry* 51, nº 25 (26 juin 2012): 4990-5003. https://doi.org/10.1021/bi3005126.

« Search of: CYC065 - List Results - ClinicalTrials.Gov ». Consulté le 11 octobre 2019. https://www.clinicaltrials.gov/ct2/results?term=CYC065&Search=Search.

Seidel, Thomas, Gökhan Ibis, Fabian Bendix, et Gerhard Wolber. « Strategies for 3D pharmacophore-based virtual screening ». *Drug Discovery Today: Technologies*, 3D Pharmacophore Elucidation and Virtual Screening, 7, nº 4 (1 décembre 2010): e221-28. https://doi.org/10.1016/j.ddtec.2010.11.004.

Shuker, Suzanne B., Philip J. Hajduk, Robert P. Meadows, et Stephen W. Fesik. « Discovering High-Affinity Ligands for Proteins: SAR by NMR ». *Science* 274, nº 5292 (29 novembre 1996): 1531-34. https://doi.org/10.1126/science.274.5292.1531.

Siramshetty, Vishal B., Janette Nickel, Christian Omieczynski, Bjoern-Oliver Gohlke, Malgorzata N. Drwal, et Robert Preissner. « WITHDRAWN—a Resource for Withdrawn and Discontinued Drugs ». *Nucleic Acids Research* 44, nº Database issue (4 janvier 2016): D1080. https://doi.org/10.1093/nar/gkv1192.

« Société Française de Chémoinformatique ». Consulté le 3 avril 2019. http://www.chemoinformatique.fr/.

Strömbergsson, Helena, et Gerard J Kleywegt. « A chemogenomics view on protein-ligand spaces ». *BMC Bioinformatics* 10, nº Suppl 6 (16 juin 2009): S13. https://doi.org/10.1186/1471-2105-10-S6-S13.

Tamaoki, Tatsuya, Hisayo Nomoto, Isami Takahashi, Yuzuru Kato, Makoto Morimoto, et Fusao Tomita. « Staurosporine, a potent inhibitor of phospholipidCa++dependent protein kinase ». *Biochemical and Biophysical Research Communications* 135, nº 2 (13 mars 1986): 397-402. https://doi.org/10.1016/0006-291X(86)90008-2.

Tanrikulu, Yusuf, Björn Krüger, et Ewgenij Proschak. « The holistic integration of virtual screening in drug discovery ». *Drug Discovery Today* 18, nº 7 (1 avril 2013): 358-64. https://doi.org/10.1016/j.drudis.2013.01.007.

Teague, Simon J., Andrew M. Davis, Paul D. Leeson, et Tudor Oprea. « The Design of Leadlike Combinatorial Libraries ». *Angewandte Chemie International Edition* 38, nº 24 (1999): 3743-48. https://doi.org/10.1002/(SICI)1521-3773(19991216)38:24<3743::AID-ANIE3743>3.0.CO;2-U.

« The Obernai Declaration », 2006, 2.

« The Top Programming Languages 2019 - IEEE Spectrum ». IEEE Spectrum: Technology, Engineering, and Science News. Consulté le 18 septembre 2019. https://spectrum.ieee.org/computing/software/the-top-programming-languages-2019.

Trudeau, Richard J. *Introduction to Graph Theory*. Courier Corporation, 2013.

Vane, J. R., et R. M. Botting. « The Mechanism of Action of Aspirin ». *Thrombosis Research* 110, nº 5 (15 juin 2003): 255-58. https://doi.org/10.1016/S0049-3848(03)00379-7.

Verdonk, Marcel L., Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, et Richard D. Taylor. « Improved Protein-Ligand Docking Using GOLD ». *Proteins* 52, nº 4 (1 septembre 2003): 609-23. https://doi.org/10.1002/prot.10465.

Verdonk, Marcel L., Ilenia Giangreco, Richard J. Hall, Oliver Korb, Paul N. Mortenson, et Christopher W. Murray. « Docking Performance of Fragments and Druglike Compounds ». *Journal of Medicinal Chemistry* 54, nº 15 (11 août 2011): 5422-31. https://doi.org/10.1021/jm200558u.

Versusmind. « Bases de données graphes - Les modèles de données ». Consulté le 27 septembre 2019. https://versusmind.eu/blog/bases-de-donnees-graphes-les-modeles-de-donnees.

Vijayan, R. S.K., Peng He, Vivek Modi, KrisnaC. Duong-Ly, Haiching Ma, Jeffrey R. Peterson, Roland L. Dunbrack, et Ronald M. Levy. « Conformational Analysis of the DFG-Out Kinase Motif and Biochemical Profiling of Structurally Validated Type II Inhibitors ». *Journal of Medicinal Chemistry* 58, nº 1 (8 janvier 2015): 466-79. https://doi.org/10.1021/jm501603h.

Wilson, Leah J., Adam Linley, Dean E. Hammond, Fiona E. Hood, Judy M. Coulson, David J. MacEwan, Sarah J. Ross, et al. « New Perspectives, Opportunities, and Challenges in Exploring the Human Protein Kinome ». *Cancer Research* 78, nº 1 (1 janvier 2018): 15-29. https://doi.org/10.1158/0008-5472.CAN-17-2291.

Wybenga-Groot, L. E., B. Baskin, S. H. Ong, J. Tong, T. Pawson, et F. Sicheri. « Structural Basis for Autoinhibition of the Ephb2 Receptor Tyrosine Kinase by the Unphosphorylated Juxtamembrane Region ». *Cell* 106, nº 6 (21 septembre 2001): 745-57.

Yan, Yumeng, Di Zhang, Pei Zhou, Botong Li, et Sheng-You Huang. « HDOCK: A Web Server for Protein–Protein and Protein–DNA/RNA Docking Based on a Hybrid Strategy ». *Nucleic Acids Research* 45, nº W1 (3 juillet 2017): W365-73. https://doi.org/10.1093/nar/gkx407.

Yuriev, Elizabeth, Mark Agostino, et Paul A. Ramsland. « Challenges and Advances in Computational Docking: 2009 in Review ». *Journal of Molecular Recognition* 24, nº 2 (2011): 149-64. https://doi.org/10.1002/jmr.1077.

Yuriev, Elizabeth, Jessica Holien, et Paul A. Ramsland. « Improvements, Trends, and New Ideas in Molecular Docking: 2012-2013 in Review ». *Journal of Molecular Recognition: JMR* 28, nº 10 (octobre 2015): 581-604. https://doi.org/10.1002/jmr.2471.

Zhao, Zheng, Lei Xie, et Philip E. Bourne. « Insights into the binding mode of MEK type-III inhibitors. A step towards discovering and designing allosteric kinase inhibitors across the human kinome ». *PLoS ONE* 12, nº 6 (19 juin 2017). https://doi.org/10.1371/journal.pone.0179936.

Zheng, J., E. A. Trafny, D. R. Knighton, N. Xuong, S. S. Taylor, L. F. Ten Eyck, et J. M. Sowadski. « 2.2 Å Refined Crystal Structure of the Catalytic Subunit of CAMP-Dependent Protein Kinase Complexed with MnATP and a Peptide Inhibitor ». *Acta Crystallographica Section D: Biological Crystallography* 49, nº 3 (1 mai 1993): 362-65. https://doi.org/10.1107/S0907444993000423.

Zhou, Qing, Lin Wu, Luo Feng, Tongtong An, Ying Cheng, Jianying Zhou, Junling Li, et al. « Safety and efficacy of abivertinib (AC0010), a third-generation EGFR tyrosine kinase inhibitor, in Chinese patients with EGFR-T790M positive non-small cell lung cancer (NCSLC). » *Journal of Clinical Oncology* 37, nº 15_suppl (20 mai 2019): 9091-9091. https://doi.org/10.1200/JCO.2019.37.15_suppl.9091.

Voilà papa, j'ai écouté tes conseils, j'ai fait de mon mieux pour être un « trouveur » plutôt qu'un chercheur !

Colin BOURNEZ

Conception d'un logiciel pour la recherche de nouvelles molécules bioactives

Résumé :

La famille des protéines kinases est impliquée dans plusieurs processus de contrôle des cellules, comme la division ou la signalisation cellulaire. Elle est souvent associée à des pathologies graves, dont le cancer, et représente ainsi une famille de cibles thérapeutiques importantes en chimie médicinale. A l'heure actuelle, il est difficile de concevoir des inhibiteurs de protéines kinases novateurs, notamment par manque de sélectivité du fait de la grande similarité existant entre les sites actifs de ces protéines. Une méthode expérimentale ayant fait ses preuves et aujourd'hui largement utilisée dans la conception de composés innovants est l'approche par fragments. Nous avons donc développé notre propre logiciel, Frags2Drugs, qui utilise cette approche pour construire des molécules bioactives. Frags2Drugs repose sur les données expérimentales disponibles publiquement, plus particulièrement sur les structures des ligands co-cristallisés avec des protéines kinases. Nous avons tout d'abord élaboré une méthode de fragmentation de ces ligands afin d'obtenir une librairie de plusieurs milliers de fragments tridimensionnels. Cette librairie est alors stockée sous la forme d'un graphe où chaque fragment est modélisé par un nœud et chaque relation entre deux fragments, représentant une liaison chimique possible entre eux, par une arête. Nous avons ensuite développé un algorithme permettant de calculer toutes les combinaisons possibles de tous les fragments disponibles, et ce directement dans le site actif de la cible. Notre programme Frags2Drugs peut créer rapidement des milliers de molécules à partir d'un fragment initial défini par l'utilisateur. De plus, de nombreuses méthodes ont été implémentées pour filtrer les résultats afin de ne conserver que les composés les plus prometteurs. Le logiciel a été validé sur trois protéines kinases impliquées dans différents cancers. Les molécules proposées ont ensuite été synthétisées et ont montré d'excellentes activités in vitro.

Mots clés : Chémoinformatique, Recherche pharmaceutiques, Fragment, Protéine kinase, Fouille de données, Parcours de graphe.

Design of a research tool for the discovery of new bioactive molecules

Summary:

Kinases belong to a family of proteins greatly involved in several aspects of cell control including division or signaling. They are often associated with serious pathologies such as cancer. Therefore, they represent important therapeutic targets in medicinal chemistry. Currently, it has become difficult to design new innovative kinase inhibitors, particularly since the active site of these proteins share a great similarity causing selectivity issues. One of the main used experimental method is fragment-based drug design. Thus, we developed our own software, Frags2Drugs, which uses this approach to build bioactive molecules. Frags2Drugs relies on publicly available experimental data, especially co-crystallized ligands bound to protein kinase structure. We first developed a new fragmentation method to acquire our library composed of thousands of three-dimensional fragments. Our library is then stored as a graph object where each fragment corresponds to a node and each relation, representing a possible chemical bond between fragments, to a link between the two concerned nodes. We have afterwards developed an algorithm to calculate all possible combinations between each available fragment, directly in the binding site of the target. Our program Frags2Drugs can quickly create thousands of molecules from an initial user-defined fragment (the seed). In addition, many methods for filtering the results, in order to retain only the most promising compounds, were also implemented. The software were validated on three protein kinases involved in different cancers. The proposed molecules were then synthesized and show excellent in vitro activity.

Keywords: Chemoinformatics, Pharmaceutical research, Fragment, Protein kinase, Data mining, Graph theory.

Institut de Chimie Organique et Analytique UMR CNRS 7311
Université d'Orléans
Rue de Chartres
45067 Orléans