

ÉCOLE DOCTORALE
SANTÉ, SCIENCES BIOLOGIQUES ET CHIMIE DU VIVANT

Institut de Chimie Organique et Analytique

THÈSE présentée par :
José-Manuel GALLY

soutenue le : **19 Décembre 2017**

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

Discipline / Spécialité : Chimie / Chémoinformatique

**Développement d'outils de
chéminformatique pour
l'identification d'inhibiteurs de protéines
kinases à partir de fragments**

THÈSE dirigée par :

Pr. Pascal BONNET

ICOA, Université d'Orléans

RAPPORTEURS :

Dr. Isabelle KRIMM

ISA, Université Lyon 1

Pr. Matthieu MONTES

CNAM, Paris

JURY :

Dr. Samia ACI-SECHE

ICOA, CNRS Orléans

Dr. Philippe BERNARD

Greenpharma, Orléans

Pr. Pascal BONNET

ICOA, Université d'Orléans

Dr. Isabelle KRIMM

ISA, Université Lyon 1

Pr. Matthieu MONTES

CNAM, Paris

Dr. Xavier MORELLI

CRCM, Université de Marseille

" The secret to getting ahead is getting started."

Mark Twain

"Great ideas emerges from useless fragments of thoughts."

Michael Bassey Johnson

" Python is a truly wonderful language. When somebody comes up with a good idea it takes about 1 minute and five lines to program something that almost does what you want. Then it takes only an hour to extend the script to 300 lines, after which it still does almost what you want."

Jack Jansen

"Always code as if the guy who ends up maintaining your code will be a violent psychopath who knows where you live."

Martin Golding

Table des matières

1. INTRODUCTION	10
1.1 La conception de médicaments.....	10
1.2 L'essor des chimiothèques	11
1.2.1 La chémoinformatique	12
1.2.2 Les formats moléculaires	12
1.2.3 Les descripteurs moléculaires	13
1.3 Objectifs du projet de thèse.....	16
2. LES PROTEINES KINASES HUMAINES.....	18
2.1 Rôle physiologique.....	18
2.2 Classification des familles de protéines kinases humaines	21
2.3 Structure du domaine catalytique des protéines kinases	24
2.3.1 Description générale de la structure des domaines kinases	25
2.3.2 La conformation active	27
2.3.3 Les conformations inactives <i>DFG-out</i> et <i>αC-helix out</i>	30
2.4 Les inhibiteurs de protéines kinases	31
2.4.1 Les Inhibiteurs de type I.....	32
2.4.2 Les Inhibiteurs de type II.....	37
2.4.3 Les Inhibiteurs de type I ^{1/2}	40
2.4.4 Les Inhibiteurs de type III.....	45
2.4.5 Les Inhibiteurs de type IV	47
2.4.6 Les Inhibiteurs de type V	48
2.4.7 Les Inhibiteurs covalents	50
2.4.8 Bilan sur les inhibiteurs de protéines kinases.....	52
3. DEVELOPPEMENT DE VSPREP, UN OUTIL POUR LA PREPARATION DE PETITES MOLECULES POUR DU CRIBLAGE VIRTUEL.....	56
3.1 Introduction	56
3.1.1 Pourquoi un nouveau protocole ?	58
3.1.2 Le rôle de l' <i>open-source</i> en chémoinformatique	59
3.2 Conception de l'outil	60
3.2.1 Cahier des charges.....	60
3.2.2 Technologie de VSPrep	61
3.2.2.1 KNIME	62
3.2.2.2 RDKit	63

3.2.2.3	Logiciels gratuits alternatifs	64
3.2.3	Description de VSPrep	65
3.2.3.1	Lecture des molécules virtuelles.....	66
3.2.3.2	Standardisation des structures	68
3.2.3.3	Filtre des composés non désirés.....	71
3.2.3.4	Suppression des doublons.....	72
3.2.3.5	Enumération des tautomères.....	74
3.2.3.6	Création de sous-librairies.....	75
3.2.3.6.1	Filtres basés sur des descripteurs moléculaires.....	75
3.2.3.6.2	Filtres basés sur des sous-structures.....	76
3.2.3.6.3	Filtres basés sur des librairies de référence.....	77
3.2.3.7	Enumération des stéréoisomères	80
3.2.3.8	Génération de conformations	81
3.2.3.9	Analyse des résultats.....	84
3.3	Application de VSPrep dans un projet d'amarrage moléculaire à visée cosmétique	88
3.3.1	Préparation de la protéine.....	89
3.3.1.1	Identification de protéines kinases d'intérêt en cosmétique	89
3.3.1.2	Création de modèles par homologie de la protéine kinase KIT	91
3.3.2	Conception et préparation d'une base moléculaire de produits naturels	101
3.3.3	Amarrage moléculaire et résultats.....	104
3.3.3.1	Outil utilisé.....	104
3.3.3.2	Paramétrage de l'amarrage moléculaire.....	105
3.3.3.3	Résultats	106
3.4	Conclusion et perspectives.....	107
3.5	Annexes	109
3.5.1	Annexe 1. Capture d'écran du protocole VSPrep tel qu'implémenté dans KNIME	111
3.5.2	Annexe 2. Exemple de rapport graphique (1).....	112
3.5.3	Annexe 3. Exemple de rapport graphique (2).....	113
3.5.4	Annexe 4. Exemple de rapport graphique (3).....	114
3.5.5	Annexe 5. Exemple de rapport graphique (4).....	115
3.5.6	Annexe 6. Exemple de rapport graphique (6).....	116
4.	DEVELOPPEMENT D'UN NOUVEL OUTIL DE CONCEPTION DE LIGANDS A PARTIR DE FRAGMENTS	117
4.1	La conception de médicaments basée sur des fragments (FBDD)	117
4.1.1	Emergence du FBDD	117
4.1.2	Les propriétés des fragments.....	119
4.1.3	La conception des librairies de fragments.....	120
4.1.3.1	La taille de la librairie	121
4.1.3.2	La diversité chimique	121
4.1.3.3	La règle de 3.....	122
4.1.3.4	Les applications particulières.....	122
4.1.4	L'optimisation des fragments	123
4.1.4.1	Fragment Growing	124
4.1.4.2	Fragment Merging.....	125

4.1.4.3	Fragment Linking.....	126
4.1.4.4	Scaffold Hopping	127
4.1.4.5	Les outils mixtes	128
4.2	Développement de Frags2Drugs, un nouvel outil <i>in silico</i> FBDD	130
4.2.1	Pourquoi un nouvel outil ?.....	130
4.2.2	Création d'une base de données moléculaire de fragments issus de structures tridimensionnelles	131
4.2.2.1	Création d'un référentiel commun de domaines kinases	132
4.2.2.2	Fragmentation des ligands co-cristallisés.....	134
4.2.3	Conception et implémentation de Frags2Drugs	137
4.2.3.1	Configurator.....	139
4.2.3.2	Reader et Writer	139
4.2.3.3	Standardizer	139
4.2.3.4	Selector	140
4.2.3.5	Sorter	141
4.2.3.6	Clusterer	141
4.2.3.7	Mapper	142
4.2.3.8	Receptor	144
4.2.3.9	Linker	145
4.2.4	Validation de l'outil	147
4.2.4.1	Preuve de concept in silico	147
4.2.4.2	Validation in vitro	151
4.2.4.2.1	Application sur GSK3 β	152
4.2.4.2.2	Application sur BRAF.....	155
4.2.4.2.3	Conclusion sur la validation <i>in vitro</i>	157
4.3	Conclusion et perspectives.....	157
4.4	Annexes	160
4.4.1	Annexe 1	160
4.4.2	Annexe 2	163
4.4.3	Annexe 3	165
4.4.4	Annexe 4	167
5.	CONCLUSION GENERALE ET PERSPECTIVES	168
6.	COMMUNICATION SCIENTIFIQUE.....	169
6.1	Article.....	169
6.2	Communications orales.....	169
6.3	Communications par affiche	169
7.	BIBLIOGRAPHIE.....	172

Liste des abréviations

ABL	<i>Abelson murine leukemia viral oncogene homolog</i>
ADP	Adénosine diphosphate
aPK	Protéines kinases atypiques
ATP	Adénosine triphosphate
BTK	<i>Burton tyrosine kinase</i>
cAMP	Adénosine monophosphate cyclique
CCDC	<i>Cambridge Crystallographic Data Centre</i>
CDK	<i>Cyclin dependant kinase</i>
ChEMBL	<i>Chemical database of the European Molecular Biology Laboratory</i>
CN	Chimiothèque Nationale
CNE	Chimiothèque Nationale Essentielle
CSD	<i>Cambridge Structural Database</i>
DD	Conception de médicaments (<i>Drug Design</i>)
DFG	Asp - Phe - Gly
DOPE	Discrete Optimized Protein Energy
EGFR	Récepteur du facteur de croissance épidermique
ePK	Protéines kinases eucaryotes
FBDD	Conception de médicaments à partir de fragments (<i>Fragment-Based Drug Design</i>)
FDA	<i>Food and Drug Administration</i>
GA	<i>Genetic algorithm</i>
GLL	<i>GPCR Ligand Library</i>
GPCR	Récepteur couplé aux protéines G (<i>G protein-coupled receptor</i>)
GS	<i>Glycogen synthase</i>
GSK3	<i>Glycogen synthase kinase 3</i>
HDF	<i>Hierarchical Data Format</i>
HRD	His - Arg - Asp
HTS	<i>High-throughput screening</i>
Hyd	Groupeement hydrophobe
INSR	Récepteur à l'insuline (<i>Insulin receptor</i>)
IUPAC	<i>International Union of Pure and Applied Chemistry</i>
KDR	<i>Kinase insert domain receptor</i>
KID	<i>Kinase Insert Domain</i>
L-DOPA	L-3,4-dihydroxyphénylalanine
LE	<i>Ligand Efficiency</i>
mCRC	Cancer colorectal métastatique
MEK1	Protéine kinase MAP2K1
mNSCLC	Cancer du poumon non à petites cellules métastatiques
mTOR	<i>Mammalian target of rapamycin</i>
MW	Poids moléculaire (<i>Molecular Weight</i>)
NMR	Résonance Magnétique Nucléaire (<i>Nuclear Magnetic Resonance</i>)
NR	Récepteur Nucléaire (<i>Nuclear Receptor</i>)
Num HA	Nombre d'accepteurs de liaison hydrogène

Num HD	Nombre de donneurs de liaison hydrogène
Num RB	Nombre de liaisons rotatives
PAINS	<i>Pan Assay INterference compoundS</i>
PDB	<i>Protein data bank</i>
PEG	Polyéthylène glycol
PMI	Moment principal d'inertie (<i>Principal Moment of Inertia</i>)
PRKACA	<i>cAMP-dependent protein kinase catalytic subunit alpha</i>
QSAR	Relation quantitative structure à activité (<i>Quantitative Structure-Activity Relationship</i>)
RAF	<i>Rapidly Accelerated Fibrosarcoma</i>
RCSB	<i>Research Collaboratory for Structural Bioinformatics</i>
RDKit	<i>Rational Discovery Kit</i>
RMSD	<i>Root Mean Square Deviation</i>
R-spine	Epine régulatrice
SAR	Relation Structure à Activité (<i>Structure-Activity Relationship</i>)
SBDD	Conception de médicaments basée sur des structures (<i>Structure-Based Drug Design</i>)
SDF	<i>Structure Data File</i>
SH2	Src homology domain 2
SPR	Résonance Plasmonique de Surface (<i>Surface Plasmon resonance</i>)
Src	<i>Proto-oncogene tyrosine-protein kinase</i>
SYK	<i>Spleen tyrosine-protein kinase</i>
TPSA	Aire de la surface topologique polaire (<i>Topological Polar Surface Area</i>)

Liste des figures

Figure 1. Investissements en R&D des compagnies pharmaceutiques membres du PhRMA.....	11
Figure 2. Fiche informative sur une molécule présente dans la base ChEMBL.	14
Figure 3. Thématiques de recherche de l'équipe SB&C.....	16
Figure 4. Structure de l'adénosine triphosphate..	18
Figure 5. Schéma de phosphorylation et de déphosphorylation d'une protéine cible par une protéine kinase ou une phosphatase.....	20
Figure 6. Représentation schématisant les éléments principaux des voies de signalisation ERK1/2 et Akt en aval des récepteurs à l'insuline (INSR) et du facteur de croissance épidermique (EGFR)	20
Figure 7. Dendrogramme des 491 domaines ePK.....	21
Figure 8. Arbre phylogénétique représentant les relations de parenté entre les 491 domaines kinases eucaryotes et les 40 domaines kinases atypiques répertoriés à ce jour	22
Figure 9. Evolution du nombre de structures de domaines kinases humaines dans la PDB.....	24
Figure 10. Schéma des structures secondaires d'un domaine kinase typique	25
Figure 11. Représentation en ruban des conformations intermédiaires du résidu Phe du motif DFG par rapport aux configurations extrêmes DFG-out et DFG-in.....	27
Figure 12. Représentation en ruban du domaine kinase de PRKACA en conformation active.....	28
Figure 13. Représentation en surface des épines C et R sur une structure de Prkaca.....	29
Figure 14. Représentation en ruban des structures de ABL1 en conformation DFG-out et en conformation DFG-in.....	30
Figure 15. Représentation en ruban des structures de CDK2 en conformation α C-helix out et en conformation α C-helix in	31
Figure 16. Représentation du modèle d'interaction d'un inhibiteur de type I..	33
Figure 17. Représentation du modèle d'interaction d'un inhibiteur de type II, la protéine kinase est en conformation inactive DFG-out.....	38
Figure 18. Mode d'interaction de l'inhibiteur de type II quizartinib dans une structure de FLT3 en conformation inactive	38
Figure 19. Représentation de différentes poches situées dans le domaine kinase de ABL1	41
Figure 20. Représentation du modèle d'interaction d'un inhibiteur de type I $\frac{1}{2}$, la protéine kinase est en conformation inactive DFG-in/ α C-out	42
Figure 21. Mode de liaison de l'analogue Tak-733 en présence d'ATP dans MAP2K1.....	46
Figure 22. Projection des sites allostériques de différentes protéines kinases sur le même domaine kinase.....	47
Figure 23. Différence de distances mesurées sur c-Src entre les domaines kinase et SH2.....	49
Figure 24. Structure d'un inhibiteur de type V (bivalent).	49
Figure 25. Superposition des complexes d'imatinib/Abl1 et de nintedanib /KDR	54
Figure 26. Inhibiteurs de protéines kinases approuvés par la FDA, classés par type d'inhibition .	55
Figure 27. Différentes conformations du cyclohexane.....	58

Figure 28. Exemple d'utilisation d'un workflow KNIME.	63
Figure 29. Encapsulation du code source de RDKit.	64
Figure 30. Exemple d'incompatibilité du format SDF entre RDKit et rDock.	67
Figure 31. Exemple d'un mélange échouant à la conversion au format RDKit à cause d'un sel...67	
Figure 32. Tableau KNIME obtenu après l'étape de lecture d'un fichier SDF avec VSPrep.....	68
Figure 33. Exemple de la structure de l'InChIKey, déduit de l'InChI, avec la caféine.	70
Figure 34. Exemple d'un cas de figure où MolVS permet de détecter deux tautomères comme une molécule unique là où le programme InChI n'y arrive pas.	70
Figure 35. Traitement des doublons par regroupement de stéréochimie.	73
Figure 36. Prédiction de tautomères majoritaires d'une molécule à pH physiologique.	75
Figure 37. Exemple de reconnaissance d'une sous-structure considérée comme problématique par les filtres PAINS.	77
Figure 38. Nombre de molécules uniques indiquées comme appartenant à une sous-librairie selon un seuil de similarité	78
Figure 39. Diagrammes en moustache montrant le RMSD minimum entre les structures cristallographiques et 10 conformations générées avec RDKit pour chacune d'entre elles et partitionnées par nombre de liaisons qui tournent.	83
Figure 40. Représentation des liaisons qui tournent sur une molécule	84
Figure 41. Exemples de limites de représentations en chémoinformatique.....	85
Figure 42. Exemple de deux représentations différentes du groupement nitro.....	86
Figure 43. Exemples de molécules rencontrées dans le catalogue d'Ambinter	88
Figure 44. Apparition de taches brunes sur la peau lors du vieillissement (lentigo)	90
Figure 45 Représentation d'une coupe de peau humaine.	91
Figure 46. Principales voies de régulation de la mélanogénèse.	91
Figure 47. Alignement des molécules d'eau présentes dans les structures des domaines kinases de KIT disponibles dans la PDB.....	93
Figure 48. Estimation de la probabilité de présence de molécules d'eau déplaçables avec 3DRISM	94
Figure 49. Domaine kinase de KIT en conformation active	96
Figure 50. Alignement de la séquence de FGFR1 avec celle de KIT	97
Figure 51. Profile DOPE d'un modèle par homologie	98
Figure 52. Résultats obtenus par la plateforme QMEAN pour les clusters 1, 2 et 3.	99
Figure 53. Diagrammes de Ramachandran générés pour les clusters 1, 2 et 3.....	99
Figure 54. Position des deux résidus marginaux selon le diagramme de Ramachandran dans le cluster 3	100
Figure 55. Stratégie pour la préparation du jeu de données de ligands à cribler	103
Figure 56. Recouvrement des bases de données naturelles incluses dans la librairie NPDB3....	104
Figure 57. Validation du protocole d'amarrage moléculaire pour les clusters 1 et 2.	105
Figure 58. Courbes ROC obtenues pour les clusters 1, 2 et 3.....	106
Figure 59. Capture d'écran de VSPrep	111
Figure 60. Rapport graphique : nombre de molécules et d'entrées à chaque étape.....	112

Figure 61. Rapport graphique : distribution de descripteurs moléculaires courants.....	113
Figure 62. Rapport graphique : nombre de molécules par sous-librairie	114
Figure 63. Rapport graphique : nombre d’alertes par étape.....	115
Figure 64. Rapport graphique : temps de calculs de chaque étape	116
Figure 65. Les quatre approches principales du FBDD.	123
Figure 66. Exemple d’un problème de prise en charge de ligands avec des coordonnées alternatives.....	132
Figure 67. Référence de MOE pour la création d’un projet Kinase.	133
Figure 68. Résultats de la fragmentation de l’inhibiteur vemurafenib.	136
Figure 69. Représentation de fragments alignés dans le référentiel commun généré.	136
Figure 70. Schéma de la nomenclature retenue pour identifier les fragments.	137
Figure 71. Architecture de Frags2Drugs.....	138
Figure 72. Protocole typique du programme Frags2Drugs.....	138
Figure 73. Regroupement des poses de fragments selon deux méthodes différentes, supervisée et non supervisée.	142
Figure 74. Détection d’un encombrement stérique entre une structure DFG-in de CDK2 et un fragment issu de l’imatinib par mesure des rayons de van der Waals.	144
Figure 75. Algorithme du module Linker	146
Figure 76. Estimation de l’énergie conformationnelle des ligands co-cristallisés mesurée dans le champ de force MMFF94s.....	147
Figure 77. Exemple de la reconstruction itérative de vemurafenib co-cristallisé avec BRAF	148
Figure 78. Exemple de vérification du mode d’interaction protéine-ligand avec vemurafenib..	148
Figure 79. Quelques exemples d’échafaudages moléculaires utilisés dans le projet FBDD.....	151
Figure 80. Mode d’interaction du fragment initial au sein du site actif de GSK3 β	153
Figure 81. Positionnement du fragment initial au sein du site actif de GSK3 β représenté sous forme de surface moléculaire.	153
Figure 82. Pose du fragment initial dans l’application sur BRAF.	156
Figure 83. Exemple de résultat de l’approche Frags2Drugs dans une structure de BRAF.	157

Liste des tables

Table 1. Répartition des kinases et des domaines kinases chez l'Homme dans les différents groupes identifiés par G. Manning.....	23
Table 2. Liste des 10 inhibiteurs de protéines kinases de type I approuvés par la FDA	34
Table 3. Liste des 7 inhibiteurs de protéines kinases de type II approuvés par la FDA	39
Table 4. Liste des inhibiteurs de protéines kinases de type I½ approuvés par la FDA.	43
Table 5. Liste des inhibiteurs de protéines kinases de type III approuvés par la FDA.....	46
Table 6. Exemples d'inhibiteurs de protéines kinases de type IV.....	48
Table 7. Liste des inhibiteurs de protéines kinases covalents approuvés par la FDA	51
Table 8. Suites logicielles quotidiennement utilisées en chémoinformatique pour la préparation des petites molécules.....	57
Table 9. Principales bibliothèques de chémoinformatique implémentées dans le logiciel KNIME	61
Table 10. Liste des outils incorporés dans le protocole KNIME VSPrep	64
Table 11. Liste des filtres implémentés pour créer des sous-bibliothèques basées sur des descripteurs moléculaires.	76
Table 12. Liste des outils in silico testés pour l'énumération de stéréoisomères.....	81
Table 13. Liste des structures PDB identifiées comme pouvant résumer les conformations majoritaires de KIT	93
Table 14. Régions non résolues sur les séquences des domaines kinases de KIT et FGFR1	96
Table 15. Résultats de l'évaluation de la qualité des modèles par homologie générés	100
Table 16. Composition de NPDB2 avant préparation	102
Table 17. Liste des outils in silico FBDD capables d'effectuer plusieurs approches FBDD.	129
Table 18. Les méthodes de fragmentation appliquées aux ligands co-cristallisés.....	135
Table 19. Liste des méthodes actuellement implémentées dans Frags2Drugs pour sélectionner des sous-ensembles de fragments.....	140
Table 20. Conditions requises pour assembler deux fragments.	143
Table 21. Mesure du pourcentage d'activité résiduelle de la protéine kinase GSK3α/β.....	154
Table 22. Liste des outils in silico FBDD pour le fragment growing.....	160
Table 23. Liste des outils in silico FBDD pour le <i>fragment linking</i>	163
Table 24. Liste des outils in silico FBDD pour le <i>scaffold hopping</i>	165
Table 25. Liste des outils in silico FBDD combinant plusieurs approches	167

Remerciements

J'adresse avant tout ma plus profonde gratitude à tous les membres de mon jury pour le temps et l'énergie qu'ils auront consacrés à lire ce manuscrit, à venir assister à ma soutenance à Orléans et à évaluer mon travail : le Docteur Philippe Bernard, le Docteur Xavier Morelli, le Docteur Samia Aci-Sèche et le Professeur Pascal Bonnet. Je souhaite remercier tout particulièrement les rapporteurs de ma thèse : le Docteur Isabelle Krimm et le Professeur Matthieu Montes.

Je tiens également à exprimer mes plus sincères remerciements à la Région Centre-Val de Loire, la Fondation ARC pour la recherche sur le cancer et la société Greenpharma pour avoir financé ma thèse, ainsi qu'au programme Ambition Recherche et Développement 2020 Cosmétosciences pour le financement d'un projet de stage de Master 2.

Il aurait été bien difficile d'accomplir ce dernier sans le soutien inconditionnel, de loin ou de près, de nombreuses personnes.

Je remercie tout particulièrement le Professeur Pascal Bonnet pour m'avoir accepté au sein de son équipe, d'abord comme stagiaire de Master 2 puis ensuite en tant que doctorant. Pascal, merci pour ta grande patience, ta bienveillance ainsi que les conseils avisés que tu as su me fournir tout au long de ma thèse.

C'est avec bien trop peu de mots que j'adresse ma gratitude au Docteur Samia Aci-Sèche, pour avoir accepté de me co-encadrer avec une très grande disponibilité au cours de cette longue aventure, pas toujours tranquille, qu'est la thèse. Merci d'avoir toujours su trouver le temps pour m'encourager et me remonter le moral. J'ai conscience d'avoir eu beaucoup de chance de tomber sur toi !

De plus, je remercie également tous les autres membres de l'équipe Bioinformatique Structurale et Chémoinformatique, en commençant par une autre personne que j'ai également beaucoup mitraillée de questions scientifiques. Je fais bien sûr référence au Docteur Stéphane Bourg, avec qui j'ai eu de longues, longues discussions, en particulier sur la préparation des molécules et ces satanés tautomères. Merci à toi pour ton détecteur intégré de fautes d'orthographes ainsi que pour ta franchise, ton humour et ta bonne humeur au quotidien !

Je pense également aux autres membres de mon équipe avec qui j'ai passé tant de bons moments. Je remercie tout d'abord les post-docs pour leurs conseils avisés. Julien, merci pour ta patate quotidienne et les discussions toujours très scientifiques et sérieuses. J'en profite également pour saluer sa moitié, Mélissa pour ses coups de gueule à un certain jeu de société dont je tairai le nom. Encore félicitations à tous deux pour la naissance de ~~Co~~Raphaël ! Un merci tout particulier à Jade aussi, ma partenaire officielle de code de la route et de mots fléchés, pour son aide précieuse dans R, les rires et les chewing-gums ! Je salue également les anciens, en particulier Felipe et Rohit pour leurs déménagements mémorables, ainsi qu'Emilie qui a justement su se débrouiller toute seule.

Je remercie aussi les doctorants de l'équipe pour m'avoir accompagné tout au long de cette aventure. Merci Abdenmour, pour les instants de cinéma, parfois bien étranges mais toujours sympas,

malgré ton soutien inconditionnel à Adam Sandler. Je remercie également Baptiste pour sa banane de tous les jours et les moments « petite binouze » après de longues journées de travail. Je pense aussi à Fabrice pour avoir pris le temps de parler code et pour les visites de châteaux avec Typhaine ! Un grand merci également à Sonia, pour avoir toujours su m'écouter et me conseiller, ainsi que pour ta grande franchise. Je remercie aussi Colin, pour avoir su reprendre mon projet de thèse et épousé la cause des fragments. Merci également d'avoir su me décharger de travail pendant de précieuses heures durant la rédaction du manuscrit de thèse. Enfin, j'en profite également pour saluer Nicolas, que je m'appête à rejoindre dans le cercle des anciens doctorants de l'équipe. Merci pour les discussions, en français ou en anglais, sur le cinéma en général, sur les requins dans l'espace en particulier ! Merci à tous !

Je remercie également tous les étudiants de Master qui sont venus partager un temps le quotidien de notre équipe : Stéphanie, Mandy, Alexandre, Emeline, Mickaël, Dorian et tout particulièrement les stagiaires qui ont participé à mon projet de thèse, Alan et Jean-Thomas.

J'adresse également mes remerciements au Professeur Oliver Martin pour m'avoir accepté au sein de l'ICOA lorsqu'il était sous sa direction. Je remercie aussi le Professeur Gérard Guillaumet pour sa collaboration à mon projet de thèse, ainsi que les membres présents ou passés de son équipe : Sandrine Grosse, Mohammed Loubidi et Mohsine Driowya.

J'étends ces remerciements de manière générale à tous les autres membres du laboratoire. Je pense en particulier à ceux avec qui j'ai eu l'occasion de critiquer la nourriture du RU mais aussi de passer de bons moments en dehors du laboratoire, entre autres : Syntia, Chloé, Hélène, Giuliano et Laura.

Je tiens également à remercier l'équipe du partenaire industriel de ma thèse, Greenpharma. Un grand merci au Docteur Philippe Bernard pour avoir participé aux réunions malgré son emploi du temps chargé. Un grand merci aussi au Docteur Quoc-Tuan Do pour les discussions scientifiques et sa continuelle bienveillance. Je souhaite également exprimer ma gratitude au Docteur Sylvain Blondeau pour ses nombreuses explications. J'en profite également pour adresser un remerciement au reste des membres de la « Green Team » : Franck, Glorianne, Florence, Matthieu, Magali, Emilie, Véronique.

Je remercie également, sans pouvoir toutes les nommer, les personnes qui m'ont donné goût à la science, que ce soit en biologie, en bioinformatique ou en chémoinformatique, lors des différents cours, stages et congrès.

Je souhaite également exprimer mes remerciements à mes amis de Nantes et d'ailleurs. Je m'excuse déjà de ne pas tous les mentionner, je pense à vous mais aussi à tous les arbres qui seront sacrifiés pour imprimer ces pages. Plus particulièrement, je pense à mes amis du Master de Bioinformatique : Audrey, Adrien, Landry, Gildas, binôme (Charles-Henri) et Camille pour les séances intenses de programmation ou de jeux-vidéos, parfois les deux. Je pense également à mes « vieux amis » de Licence avec qui j'ai pu vérifier la bonne tenue de la cafétéria : Angélique, Bénédicte, Céline, Jeff, François, Cédric, Axel et Laëtitia. En remontant plus loin encore, je dis merci à mes amis de toujours d'avoir été là pour moi : Aymeric, Michael, Thomas B., Thomas C., Simon et tous les autres.

Enfin, je voudrais remercier ceux sans qui rien n'aurait été possible : mes parents pour avoir toujours cru en moi et m'avoir donné l'opportunité de poursuivre mes rêves. Un grand merci également à mes sœurs Karine et Joëlle pour leur soutien pendant toutes ces années.

1. Introduction

1.1 La conception de médicaments

La recherche et le développement d'un médicament, *Drug Design ou Drug Discovery* (DD), est un processus qui dure en général entre 8 et 15 ans, allant de l'idée originale (*hit*), puis de la tête de série (*lead*) jusqu'à la mise sur le marché du produit fini (médicament). Ce processus nécessite le savoir-faire de plusieurs disciplines scientifiques, telles que la chimie médicinale, la biochimie, la pharmacologie et la médecine, aboutissant à un coût total estimé à **plus** d'un milliard de dollars américains.¹ Malgré cette débauche de temps et d'argent, le taux de succès des projets de DD de l'industrie pharmaceutique reste faible et le nombre de nouveaux médicaments mis sur le marché stagne.²

Pourtant, le secteur pharmaceutique est l'un de ceux qui investit le plus au monde dans la recherche et le développement (R&D, 10-20% de son chiffre d'affaires, **Figure 1**) afin d'optimiser constamment chacune des étapes du développement d'un nouveau médicament.³

Ainsi, certaines avancées technologiques ont permis d'orienter le DD vers de nouvelles voies. On peut noter l'émergence des méthodes de résolution structurale de protéines, telles par exemple la spectroscopie de Résonance Magnétique Nucléaire (RMN), portée par les travaux de K. Wüthrich, récompensés par le Prix Nobel de Chimie en 2002⁴, ou la cristallographie aux rayons X. Ces techniques, maintenant employées dans la plupart des laboratoires de recherche, guident l'optimisation d'une touche (*hit*) ou d'une tête de série (*lead*) par les chercheurs en élucidant le mode d'interaction de la molécule avec sa cible. La détermination de ces structures 3D a permis l'essor d'un nouveau domaine de recherche : le *Structure-Based Drug Discovery* (SBDD).

Un autre bouleversement survient dans les années 1990 grâce au couplage de la chimie combinatoire, autorisant la synthèse rapide d'un grand nombre de molécules, avec le criblage à haut débit (HTS). La combinaison de ces techniques a ainsi permis de tester l'activité biologique d'un grand nombre de composés sur une ou plusieurs cibles données. Les touches (*hits*) obtenues sont alors confirmées par un second criblage dans lequel l'activité est mesurée de façon plus fine et les molécules présentant les meilleurs profils voient leur mode d'interaction validé pour une application SBDD.

¹ J.P. Hughes *et al.*, Principles of Early Drug Discovery, *British Journal of Pharmacology*, 162.6 (2011), 1239–49

² B. Munos, Lessons from 60 Years of Pharmaceutical Innovation, *Nature Reviews. Drug Discovery*, 8.12 (2009), 959–68

³ IRI - The EU Industrial R&D Investment Scoreboard, <http://iri.jrc.ec.europa.eu/scoreboard.html>, consulté le 15 Novembre 2017

⁴ Communiqué de Presse : Le Prix Nobel de Chimie 2002, <https://www.nobelprize.org>, consulté le 15 Novembre 2017

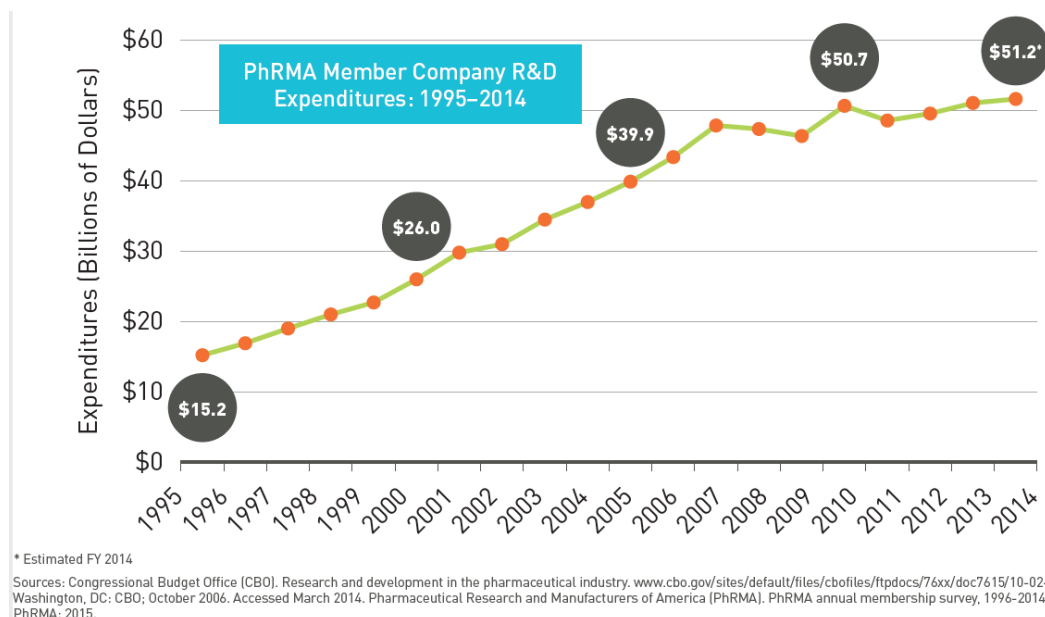


Figure 1. Investissements en R&D des compagnies pharmaceutiques membres du PhRMA.

Source : http://phrma-docs.phrma.org/sites/default/files/pdf/2015_phrma_profile.pdf

La démocratisation de la technique HTS dans les compagnies pharmaceutiques n'aurait jamais pu voir le jour sans les avancées survenues en parallèle dans le domaine de la biologie et l'informatique. En effet, le HTS entraîne la génération d'un volume de données sans précédent qu'il faut être capable d'acquérir, de traiter et d'archiver dans un temps rapide et avec un coût raisonnable.

1.2 L'essor des chimiothèques

L'avènement de la chimie combinatoire, dans un effort des compagnies pharmaceutiques d'identifier avec plus de chance une tête de série prometteuse (« cribler plus pour gagner plus »), s'est accompagné d'une multiplication de larges collections de composés chimiques, ou chimiothèques. Un véritable marché s'est alors développé pour la commercialisation de telles librairies de composés chimiques, pouvant contenir plusieurs millions de molécules.⁵

Parmi les catalogues des vendeurs, on distingue majoritairement deux types de librairies. Le premier concerne les molécules destinées au criblage tandis que le second renvoie aux librairies de réactifs (*building blocks*). D'autres types de chimiothèques plus ciblées existent, telles que les librairies de fragments, d'inhibiteurs de protéines kinases, d'inhibiteurs d'interaction protéine-protéine ou encore de produits naturels.

⁵ R. Macarron *et al.*, Impact of High-Throughput Screening in Biomedical Research, *Nature Reviews Drug Discovery*, 10.3 (2011)

Cependant, toutes les chimiothèques ne sont pas conçues pour la vente. Certaines se sont développées dans les laboratoires de recherche académique, que ce soit en interne ou disponibles en ligne (PubChem⁶).

Dans tous les cas, la nécessité de développer des outils pour interroger les données accumulées est très vite devenue essentielle. Ainsi, la première utilisation publiée d'un algorithme pour effectuer des recherches sous-structurales sur une base de données moléculaire informatisée remonte en effet en 1957.⁷ Ce travail découle de la volonté de la CAS *Chemical Abstracts Services*⁸ de décrire pour la première fois les structures chimiques sous forme de graphes afin de rechercher des molécules contenant certains motifs structuraux spécifiques.

1.2.1 La chimoinformatique

Définie en 2006 lors du congrès « *Workshop Chemoinformatics in Europe: Research and Teaching* » à Obernai sous l'égide des Pr Alexandre Varnek et Luc Morin-Allory (<http://www.chemoinformatique.fr>), la chimoinformatique est la science qui applique l'outil informatique à des problèmes relatifs à la chimie. La définition retenue est « la chimoinformatique est le domaine qui considère des molécules ou des réactions comme des objets (graphes, vecteurs) dans un espace chimique ». A l'instar du criblage réel, on pourra utiliser le terme criblage virtuel lors de la recherche de molécules bioactives sur une ou plusieurs cibles par des outils de chimoinformatique.

Pour réaliser de tels criblages, une question que l'on peut se poser est : comment représenter informatiquement une molécule ? Tout dépend de la manière dont on la considère. Ceci a donné naissance à une multitude de formats moléculaires.

1.2.2 Les formats moléculaires

Même s'il en existe un grand nombre, on peut classer les formats moléculaires en 3 grandes catégories : celles qui décrivent les molécules, celles qui décrivent des parties de molécules et celles qui représentent des réactions chimiques. Chaque catégorie présente une variabilité de dimension (1D, 2D, 3D) d'information qu'elle contient. Il est intéressant de noter qu'on pourrait également considérer un identifiant comme un format moléculaire « 0D » puisqu'il ne renseigne aucune information sur la structure chimique. C'est par exemple le cas des bases de données qui ne stockent qu'un identifiant pour ensuite retrouver la structure dans d'autres bases. Pour la suite de cette partie, nous nous focaliserons sur les formats de la première catégorie, représentant des molécules.

⁶ S. Kim *et al.*, PubChem Substance and Compound Databases, *Nucleic Acids Research*, 44. Database issue (2016), D1202–13

⁷ L. C. Ray and R. A. Kirsch, Finding Chemical Records by Digital Computers, *Science (New York, N.Y.)*, 126.3278 (1957), 814–19

⁸ CAS, Chemical Abstracts Service Home Page, <https://www.cas.org>, consulté le 15 Novembre 2017

Ainsi, les formats 1D représentent les molécules sous forme textuelle (SMILES, InChI/InChI-Key, nom IUPAC) et servent à retranscrire l'information topologique (atomes et liaisons) uniquement. Il s'agit de formats compacts et donc faciles à stocker.

Les formats 2D permettent quant à eux de représenter les molécules, sous forme de graphes. La principale différence avec le format 1D est que chaque atome possède des coordonnées (x, y). Le passage du format 1D au format 2D est très facile, c'est pourquoi beaucoup d'outils de chémoinformatique sont capables de lire un format 1D comme du SMILES et de le convertir en format 2D pour lui donner une représentation visuelle plus facile à comprendre.

Les formats 3D (Mol, SDF, Mol2, PDB, etc.) sont bien plus difficiles à obtenir que ceux décrits précédemment car ils doivent retranscrire le positionnement de chaque atome de la molécule dans l'espace de coordonnées cartésiennes (x, y et z) ou internes (distance, angle, dièdre). Afin d'obtenir des conformations raisonnables, la plupart des outils emploie un champ de force, où sont décrits des termes géométriques précis à respecter (distances interatomiques, angles de liaison et angles dièdres). Le passage du format 2D au format 3D est plus ou moins rapide selon la méthode suivie. Pour une molécule complexe (un macrocycle par exemple), elle peut aller de quelques secondes pour une distance géométrique classique (règles empiriques de termes géométriques) à plusieurs heures pour atteindre la convergence d'une minimisation dans un champ de force. Ces formats 3D sont employés pour les approches *in silico* SBDD et donc les plus couramment utilisés pour du criblage virtuel.

Un exemple de représentations d'une molécule est illustré par la **Figure 2**. Il existe d'autres façons de décrire une molécule, comme nous allons le voir dans la partie suivante.

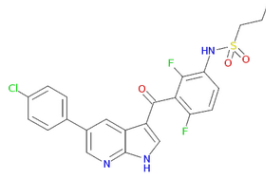
1.2.3 Les descripteurs moléculaires

Un descripteur moléculaire est une valeur, le plus souvent numérique (simple ou vectorielle), mesurée ou calculée à partir de la représentation d'une molécule, employée pour décrire ses propriétés physico-chimiques ou pharmacologiques. On distingue ainsi deux types de descripteurs moléculaires : les descripteurs expérimentaux et les descripteurs théoriques. Le travail réalisé au cours de cette thèse étant uniquement *in silico*, nous ne considérerons que les descripteurs moléculaires théoriques. A l'instar des formats moléculaires, les descripteurs peuvent être séparés en différentes catégories selon la dimension de l'information qu'ils portent.

Les descripteurs les plus simples sont les descripteurs 1D. Ils encodent une seule propriété physico-chimique. Parmi ces propriétés, on peut citer par exemple le poids moléculaire ou bien le nombre d'atomes lourds. Même si l'information qu'ils portent décrit une chimiothèque de manière grossière, ils sont très utilisés en chémoinformatique car très rapides à calculer.

Compound Report Card

Compound Name and Classification

Compound ID	CHEMBL1229517	
Compound Name	VEMURAFENIB	
ChEMBL Synonyms	ZELBORAF VEMURAFENIB RG 7204 PLX-4032 RO 5185426	
Max Phase	4 (Approved)	
Trade Names	ZELBORAF	
Molecular Formula	C23H18ClF2N3O3S	

CHEMBL1229517

Additional synonyms for CHEMBL1229517 found using [NCI Chemical Identifier Resolver](#)

Compound Representations

Molfile	Download MolFile
Canonical SMILES	<chem>CCCC(=O)(=O)Nc1ccc(F)c(C(=O)c2c[nH]c3ncc(cc23)c4ccc(Cl)cc4)c...</chem> Download SMILES
Standard InChI	InChI=1S/C23H18ClF2N3O3S/c1-2-9-33(31,32)29-19-8-7-18(25)20(... Download InChI
Standard InChI Key	GPXBXXGIAQBQNI-UHFFFAOYSA-N

Figure 2. Fiche informative sur une molécule présente dans la base ChEMBL.⁹ Plusieurs représentations sont accessibles à l'utilisateur.

Les descripteurs 2D prennent en compte la topologie des molécules sous forme de graphes. Ils peuvent être employés pour décrire la connectivité d'une molécule ou bien encore la présence ou l'absence de certains sous-graphes (fragments). Les descripteurs de cette catégorie sont rapides à calculer et les plus couramment utilisés sont les indices de Randic¹⁰ ou encore Hall & Kier¹¹. Il est à noter que la plupart de ces descripteurs ne prennent pas en compte la stéréochimie des composés mais sont très sensibles à l'état de protonation et à la forme tautomérique considérés. L'utilisateur peut les énumérer en amont du calcul des descripteurs afin d'en tenir compte.

Les descripteurs 3D utilisent l'information tridimensionnelle des molécules, c'est-à-dire la position relative de chacun des atomes dans l'espace. Ils permettent notamment d'évaluer des propriétés complexes telles que la forme de la molécule (Moment Principal d'Inertie, PMI¹²). Ces descripteurs demandent généralement des temps de calcul élevés, ce qui restreint leur application sur des

⁹ A. Gaulton *et al.*, ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery, *Nucleic Acids Research*, 40.Database issue (2012), D1100–1107

¹⁰ M. Randic, Characterization of Molecular Branching, *Journal of the American Chemical Society*, 97.23 (1975), 6609–15

¹¹ L. H. Hall *et al.*, Electropotential State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, *Journal of Chemical Information and Computer Sciences*, 35.6 (1995), 1039–45

¹² W. H. B. Sauer *et al.*, Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity, *Journal of Chemical Information and Computer Sciences*, 43.3 (2003), 987–1003

chimiothèques volumineuses. De plus, le choix de la conformation à utiliser n'est pas toujours simple, surtout pour des molécules avec de nombreuses liaisons qui tournent ou des cycles non plans. Il se révèle alors parfois nécessaire d'explorer l'espace conformationnel des structures afin de mieux prendre en compte la flexibilité réelle des molécules, ce qui peut grandement accroître la complexité du problème.

Au fil des années, de nombreux modèles prédictifs ont été développés pour décrire des propriétés physico-chimiques calculées à partir d'autres descripteurs moléculaires. Un bon exemple est le coefficient de répartition d'un composé entre l'eau et l'octanol, couramment appelé logP, qui permet d'évaluer l'hydrophobicité des molécules. De nombreux descripteurs ont été développés pour prédire cette propriété, les plus célèbres étant le CLogP, basé sur une approche par fragments, tandis que le SLogP est lui basé sur une approche par atome.¹³ Ces valeurs peuvent ensuite être employées comme des descripteurs classiques dans des études de diversité ou pour filtrer des chimiothèques.

Une autre façon de décrire les molécules est basée sur l'utilisation d'empreintes moléculaires, ou encore *fingerprints*. Ce sont des descripteurs très utilisés pour les études de similarité/diversité en chémoinformatique. Ils consistent typiquement en des chaînes de 0 et 1, dont chaque élément traduit de manière binaire la présence ou l'absence d'un ou plusieurs motifs structuraux. Cette représentation binaire permet de stocker un grand nombre d'informations de manière efficace et est particulièrement adaptée pour communiquer avec l'ordinateur. Il existe deux types principaux d'empreintes moléculaires : celles basées sur des motifs prédéfinis et celles construites à partir des motifs propres des molécules.

L'exemple le plus connu de la première catégorie est le *fingerprint* MACCS¹⁴, basé sur un ensemble de 166 motifs structuraux communément rencontrés en chimie thérapeutique et encodés sous forme de SMART¹⁵. L'inconvénient de ce genre de *fingerprints* est qu'ils contiennent un nombre de motifs prédéterminés, dont un grand nombre n'est pas forcément adapté au sujet de l'étude.

Le second type de *fingerprints* utilise une procédure de fragmentation de la molécule afin de la représenter par un ensemble de motifs inconnus à l'avance, car dépendants de chaque molécule. Parce que le volume d'information à décrire varie énormément, des algorithmes de hachage sont employés afin de compresser l'information en un nombre fixe de *bits*. Un exemple de tels fingerprints est ECFP.¹⁶

¹³ Y. Sakuratani *et al.*, Comparison of Predictivities of Log P Calculation Models Based on Experimental Data for 134 Simple Organic Compounds, *QSAR & Combinatorial Science*, 26.1 (2007), 109–16

¹⁴ J. L. Durant *et al.*, Reoptimization of MDL Keys for Use in Drug Discovery, *Journal of Chemical Information and Computer Sciences*, 42.6 (2002), 1273–80.

¹⁵ Daylight Theory: SMARTS - A Language for Describing Molecular Patterns, <http://www.daylight.com/dayhtml/doc/theory/theory.SMARTS.html>, consulté le 15 Novembre 2017.

¹⁶ D. Rogers *et al.*, Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, 50.5 (2010), 742–54

1.3 Objectifs du projet de thèse

J'ai effectué mon doctorat au sein de l'Institut de Chimie Organique et Analytique (ICOA) de l'Université d'Orléans. Il s'agit d'une unité mixte de recherche (UMR 7311) placée sous la tutelle de l'Université d'Orléans et du CNRS. Ses axes de recherche sont orientés vers la conception, la synthèse et l'analyse de molécules susceptibles de présenter une activité dans le domaine des sciences du vivant. Le laboratoire est notamment membre de la fédération de recherche « Physique et chimie du vivant » FR CNRS 2708 et des labex SynOrg et IRON.

J'ai plus particulièrement réalisé mes travaux au sein de l'équipe de Bioinformatique Structurale et Chémoinformatique, dirigée depuis 2012 par le Pr. Pascal Bonnet suite au départ en retraite du Pr. Luc Morin-Allory. L'équipe s'est, entre autres, spécialisée dans le domaine des protéines kinases, expertise qu'elle met à profit pour élaborer une plateforme *in silico* de criblage virtuel, prédiction d'activité et études polypharmacologiques et pharmacocinétiques (**Figure 3**).

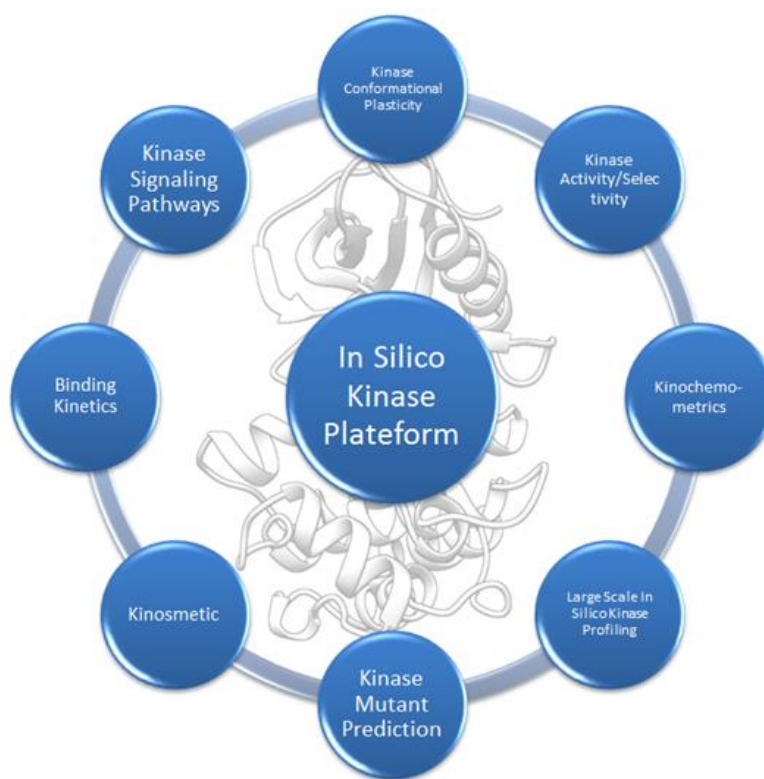


Figure 3. Thématiques de recherche de l'équipe SB&C.

Au cours de ma thèse j'ai collaboré avec l'entreprise Greenpharma (<http://www.greenpharma.com/>), une société de biotechnologie spécialisée dans l'identification de produits naturels en tant que molécules bioactives ou ingrédients en pharmacologie, cosmétique et agro-alimentaire. Elle a été fondée en 2000 par le docteur Philippe Bernard. En 2011, Greenpharma a acquis la

société Ambinter, ce qui lui procure l'un des plus gros catalogues de composés chimiques au monde, avec pas moins de 20 millions de structures (<http://www.ambinter.com/>).

Comme je l'ai mentionné, la chémoinformatique est très importante pour l'identification de nouvelles molécules bioactives et j'y ai donc eu recours pour développer deux outils *in silico* que j'ai pu appliquer pour identifier des molécules pouvant présenter des activités sur des protéines kinases.

Le chapitre 2 décrit ces protéines et met en avant leur rôle central dans la signalisation moléculaire. Leur dérégulation pouvant engendrer des pathologies graves, j'ai également décrit les différentes stratégies actuellement existantes pour inhiber leur activité.

J'ai également réalisé un amarrage moléculaire d'une base de données moléculaire de produits naturels assemblée depuis des chimiothèques publiques spécialisées. Pour ce faire, j'ai été confronté à un problème récurrent du criblage virtuel, voire du domaine de la chémoinformatique : la préparation des ligands. Constatant qu'il n'existait pas de protocole standard à base d'outils accessibles, j'ai décidé de concevoir un tel outil. Cette étape s'est révélée très chronophage. Son implémentation, ses forces, ses limites et son application sont présentées dans le chapitre 3.

Enfin, le chapitre 4 constitue l'objectif principal de ma thèse. J'ai développé un nouvel outil *in silico* de conceptions de molécules à partir de fragments, Frags2Drugs. Je présenterai dans un premier temps l'élaboration d'une base de données moléculaire de structures de protéines kinases, base qui va permettre de générer les fragments qui seront ensuite utilisés par le logiciel. Enfin, je présenterai les étapes de développement du logiciel et je terminerai par deux exemples d'application qui valident Frags2Drugs.

2. Les protéines kinases humaines

2.1 Rôle physiologique

Le terme kinase est issu de la jonction du mot « κινέω », mouvoir en grec ancien, et du suffixe « -ase », enzyme. Il regroupe la famille des protéines responsables de la catalyse enzymatique par transfert du groupement phosphate en position γ de l'adénosine triphosphate (ATP, **Figure 4**) vers le groupement hydroxyle d'un substrat. Ce dernier peut être un lipide, un sucre ou encore une protéine. Dans ce dernier cas, le transfert du groupement phosphate s'effectue sur la chaîne latérale d'un acide aminé spécifiquement reconnu par l'enzyme. Chez les eucaryotes, ce résidu peut être une sérine, une thréonine ou une tyrosine (Ser, Thr ou Tyr). Les kinases peuvent transiter entre une forme active, capable de phosphoryler leur substrat, et une forme inactive. Dans le cadre de ces travaux de thèse, seules les protéines kinases humaines seront abordées.

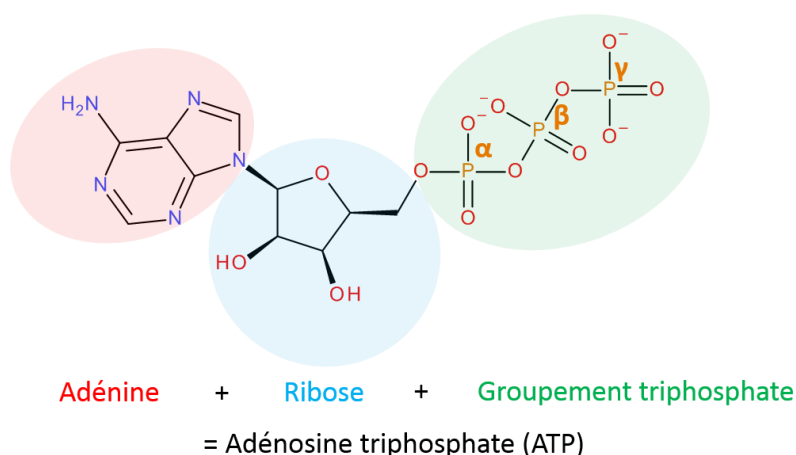


Figure 4. Structure de l'adénosine triphosphate. L'adénosine triphosphate (ATP) est composée d'une base azotée adénine associée à un ribose (adénosine), lui-même relié à un groupement triphosphate. Le phosphate γ est transféré par les kinases vers un substrat.

Le premier substrat phosphorylé à avoir été découvert est la caséine, protéine présente dans le lait (O. Hammarsten en 1883).¹⁷ Cette protéine a servi de modèle pour détecter l'activité de phosphorylation pendant de nombreuses années. La première protéine kinase a été découverte en 1954 par G. Burnett et E.P. Kennedy, en suivant le transfert d'un groupement phosphate radio-marqué (P32) vers la caséine dans les extraits mitochondriaux de foie de rat.¹⁸ Les travaux d'E.G. Krebs et E.H. Fischer ont permis par la suite de mettre en évidence l'activité d'une protéine kinase, alors appelée « phosphorylase kinase », dont le rôle est de phosphoryler le résidu Ser d'une protéine impliquée dans la

¹⁷ A. Sreelatha *et al.*, The Secretory Pathway Kinases, *Biochimica et Biophysica Acta*, 1854.10 0 0 (2015), 1687–93

¹⁸ G. Burnett *et al.*, The Enzymatic Phosphorylation of Proteins, *The Journal of Biological Chemistry*, 211.2 (1954), 969–80.

glycogénolyse.¹⁹ Presque dix ans plus tard, les travaux de D.A. Walsh et E.G. Krebs sur la protéine *cAMP-dependent protein kinase* (PRKACA), aussi connue sous le nom de PKA, ont permis de replacer cette fonction au sein d'un contexte de transduction de signaux moléculaires. Ainsi, la PKA est activée par l'adénosine monophosphate cyclique (cAMP) et active à son tour, par phosphorylation, la «phosphorylase kinase» mentionnée plus tôt, pour aboutir à l'activation de la caséine.²⁰

Cette modification post-traductionnelle de la structure d'une protéine kinase par une autre protéine kinase pose la base de nos connaissances actuelles sur la signalisation moléculaire (**Figure 5**). Ainsi, en présence d'ATP et de son substrat, une protéine kinase phosphoryle sa cible. Ceci induit un signal moléculaire comme l'activation d'une enzyme substrat. La réaction inverse est catalysée par une protéine phosphatase et permet l'inactivation de la cible phosphorylée en libérant le phosphate inorganique. La cible protéique peut être une autre protéine kinase.

Un équilibre physiologique subtil repose ainsi sur l'activation/inactivation de cascades de protéines (**Figure 6**). Si de nombreuses découvertes, qui se succèdent depuis les années 1990, suggèrent un vaste spectre d'implications de ces protéines dans les fonctions du vivant²¹, d'autres mécanismes de régulation coexistent dans la cellule. A titre d'exemple, certains parmi ceux-ci reposent sur d'autres types de modifications structurales par des transférases (glycosyltransférases, méthyltransférases, etc.) tandis que d'autres encore reposent sur la complexation de plusieurs protéines (ubiquitination pour la dégradation, sumoylation pour le transport nucléo-cytoplasmique, etc.).²²

Afin de mieux organiser la connaissance accumulée et de pouvoir comparer les différentes kinases entre elles, des approches ont été développées pour annoter et classer les kinases humaines.

¹⁹ E. G. Krebs *et al.*, Factors Affecting the Activity of Muscle Phosphorylase b Kinase, *The Journal of Biological Chemistry*, 234 (1959), 2867–73.

²⁰ D. A. Walsh *et al.*, An Adenosine 3,5-Monophosphate-Dependant Protein Kinase from Rabbit Skeletal Muscle, *The Journal of Biological Chemistry*, 243.13 (1968), 3763–65.

²¹ R. Roskoski, A Historical Overview of Protein Kinases and Their Targeted Small Molecule Inhibitors, *Pharmacological Research*, 100 (2015), 1–23

²² M. J. Lee *et al.*, Protein Regulation in Signal Transduction, *Cold Spring Harbor Perspectives in Biology*, 8.6 (2016)

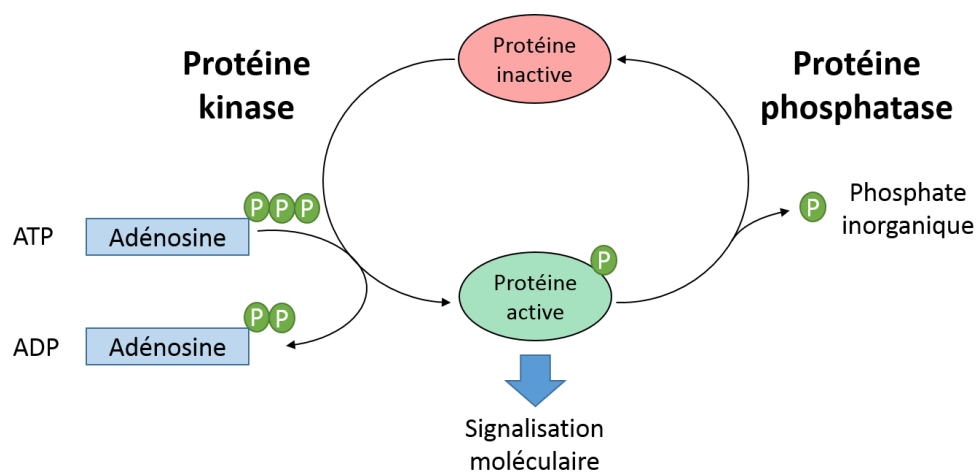


Figure 5. Schéma de phosphorylation et de déphosphorylation d'une protéine cible par une protéine kinase ou une phosphatase.

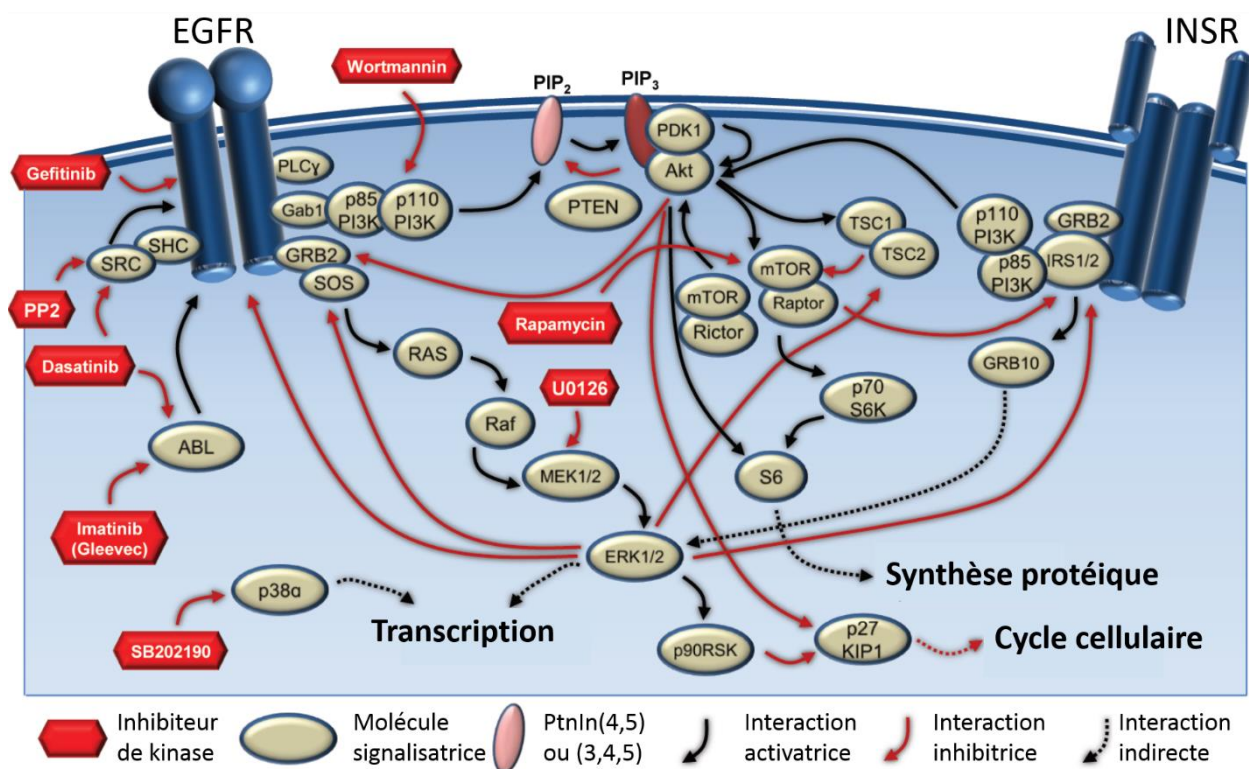


Figure 6. Représentation schématique montrant les éléments principaux des voies de signalisation ERK1/2 et Akt en aval des récepteurs à l'insuline (INSR) et du facteur de croissance épidermique (EGFR).²³

²³ S. M. Carlson et al., Using Small Molecules and Chemical Genetics to Interrogate Signaling Networks, ACS Chemical Biology, 6.1 (2011), 75–85

2.2 Classification des familles de protéines kinases humaines

La famille des kinases est vaste. Estimée à 1001 membres à la fin des années 1980²⁴, les progrès technologiques effectués dans le domaine du séquençage génomique ainsi que les travaux de l'équipe de G. Manning²⁵ ont permis, en 2002, de ramener ce nombre à 518, soit 1.7% du génome humain²⁶. Quelques erreurs de classification ont depuis été identifiées²⁷, sans que ce nombre ne soit remis en cause par la communauté scientifique.

En se basant sur les séquences du domaine catalytique des kinases, également appelé domaine kinase, les travaux de G. Manning cités précédemment ont permis de proposer une classification de l'ensemble des kinases humaines (kinome humain). Ce dernier regroupe 478 protéines kinases eucaryotes (ePK) typiques, c'est-à-dire comportant des motifs conservés dans les séquences responsables de leur activité catalytique. Les 40 kinases restantes sont classées séparément sous l'appellation de protéines kinases atypiques (aPK). Ces aPK peuvent utiliser d'autres types de substrats, comme des sucres ou des lipides, et divergent de manière importante des ePK par leur séquence. Environ 10% du total des kinases humaines sont appelées pseudokinases car elles ne présentent aucune ou très peu d'activité kinase malgré une forte similarité de séquence de leur domaine kinase.²⁸ Le degré de parenté entre les différents domaines kinases répertoriés dans cette classification est communément représenté sous la forme d'un dendrogramme (**Figure 7**) ou de manière plus détaillée sous la forme d'un arbre phylogénétique (**Figure 8**), qui peut être utilisé comme une cartographie du kinome et être annoté.

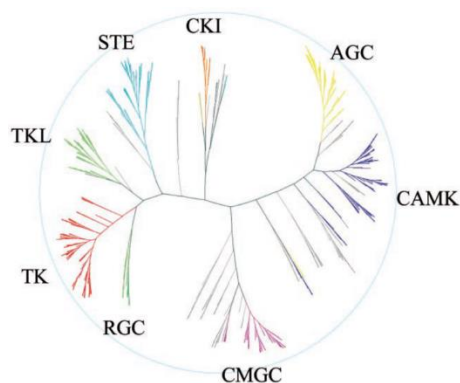


Figure 7. Dendrogramme des 491 domaines ePK.²⁵

Une même kinase pouvant comporter plusieurs domaines kinases, on dénombre actuellement non moins de 491 domaines ePK et 40 domaines aPK, respectivement regroupés en 9 et 8 groupes selon

²⁴ T. Hunter, A Thousand and One Protein Kinases, *Cell*, 50.6 (1987), 823–29.

²⁵ G. Manning *et al.*, The Protein Kinase Complement of the Human Genome, *Science (New York, N.Y.)*, 298.5600 (2002), 1912–34

²⁶ S. K Hanks, Genomic Analysis of the Eukaryotic Protein Kinase Superfamily: A Perspective, *Genome Biology*, 4.5 (2003), 111

²⁷ S. Braconi Quintaje and 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: MCP*, 7.8 (2008), 1409–19

²⁸ J. Boudeau *et al.*, Emerging Roles of Pseudokinases, *Trends in Cell Biology*, 16.9 (2006), 443–52

leur similarité (**Table 1**). Dans le cadre de cette étude, nous nous focaliserons uniquement sur les protéines kinases humaines typiques.

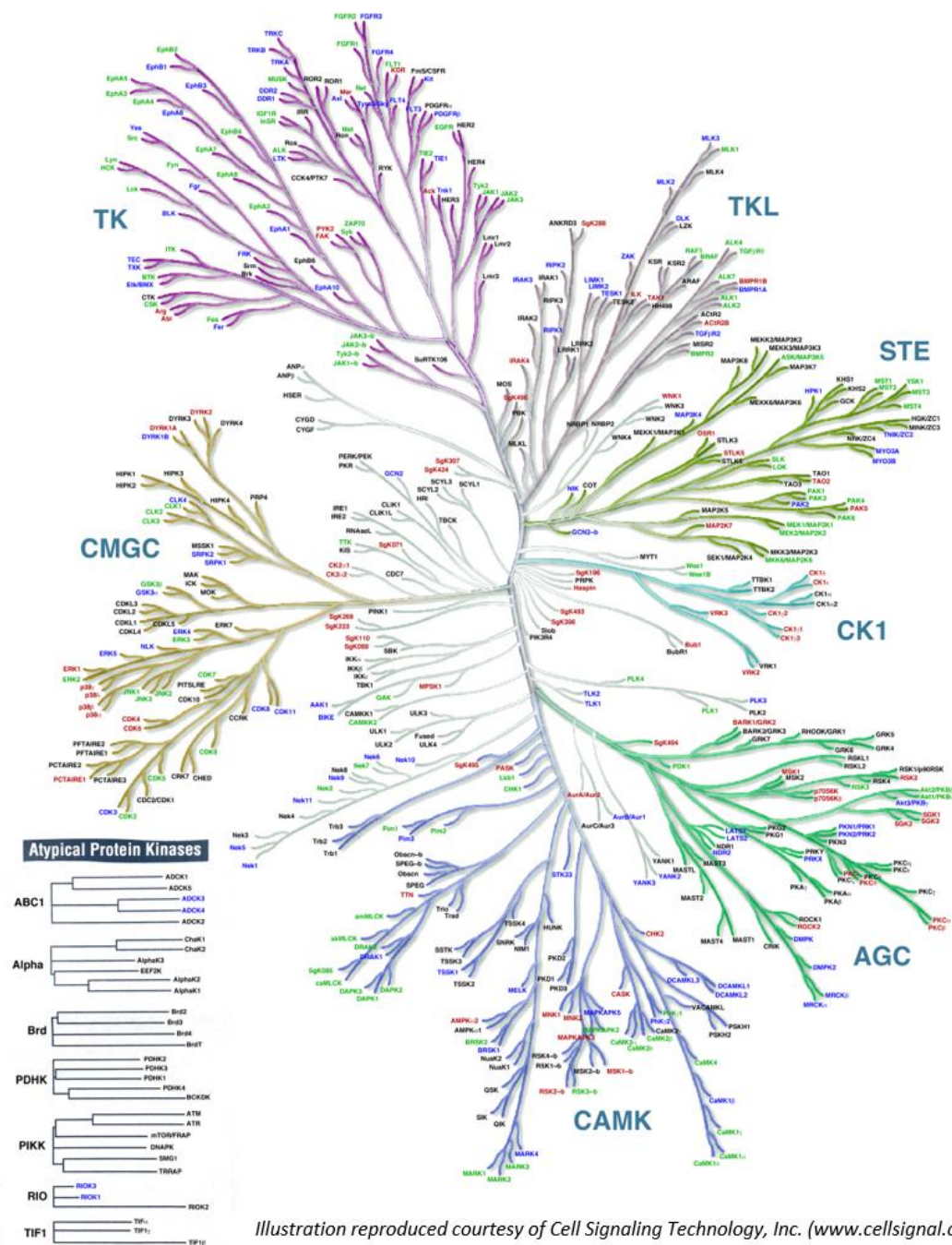


Figure 8. Arbre phylogénétique représentant les relations de parenté entre les 491 domaines kinases eucaryotes et les 40 domaines kinases atypiques répertoriés à ce jour. Les branches grises représentent le groupe « Autres ». Les protéines kinases atypiques sont représentées à part et se divisent elles-mêmes en 8 groupes, la partie « Autres » n'étant pas indiquée ici. Les PK dont le nom est affiché en rouge ont au moins une structure représentée dans la Protein Data Bank (PDB). Celles dont le nom est

en bleu sont présentes dans l'étude de M.W. Karaman et al.²⁹ Le nom des PK présentes dans les deux groupes est écrit en vert. Les travaux de Karaman ont servi, entre autres, à établir des cartes d'interaction pour 38 inhibiteurs testés sur 317 protéines kinases.

Table 1. Répartition des kinases et des domaines kinases chez l'Homme dans les différents groupes identifiés par G. Manning

Superfamille	Groupe	Nombre de protéines kinases	Nombre de domaines kinases
ePK	AGC	63	63
	CAMK	74	82
	CK1	12	12
	CMGC	61	61
	RGC	5	5
	STE	47	48
	TK	90	94
	TKL	43	43
	Autres	83	83
aPK	A6	2	2
	ABC1	5	5
	Alpha	6	6
	BRD	4	4
	PDHK	5	5
	PIKK	6	6
	RIO	3	3
	Autres	9	9
Total		518	531

²⁹ Mazen W. Karaman et al., A Quantitative Analysis of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 26.1 (2008), 127–32

2.3 Structure du domaine catalytique des protéines kinases

En 1991, la première structure d'un domaine catalytique de protéine kinase est déposée dans la *Protein Data Bank* (PDB), propriété de la *Research Collaboratory for Structural Bioinformatics* (RCSB)³⁰. Il s'agit de la forme apo de PRKACA (groupe AGC, code PDB : 2CPK) murine. Deux ans plus tard, la première structure holo de la même protéine est obtenue, cette fois-ci en présence d'ATP et d'un substrat peptidique (code PDB : 1ATP).

Grâce aux avancées de la cristallographie par rayons X et à l'engouement des compagnies pharmaceutiques, suite au succès fulgurant de l'imatinib, pour les protéines kinases, des structures de domaines kinases de plus en plus nombreuses deviennent accessibles à la communauté scientifique. Sur les dix dernières années, au moins 200 structures de domaines kinases ont ainsi été ajoutées chaque année dans la PDB (**Figure 9**). Certaines protéines kinases sont mieux représentées que d'autres, c'est par exemple le cas de CDK2. Cette dernière est impliquée dans la régulation du cycle cellulaire et présente donc un fort intérêt dans la recherche sur le cancer, ce qui se traduit par 377 structures recensées en septembre 2017, soit environ 10% de toutes les structures de kinases humaines disponibles. Il est à noter que l'ensemble des structures disponibles, très majoritairement obtenues par cristallographie aux rayons X, ne correspondent qu'à 261 kinases différentes, soit environ la moitié des 518 référencées par les travaux G. Manning.

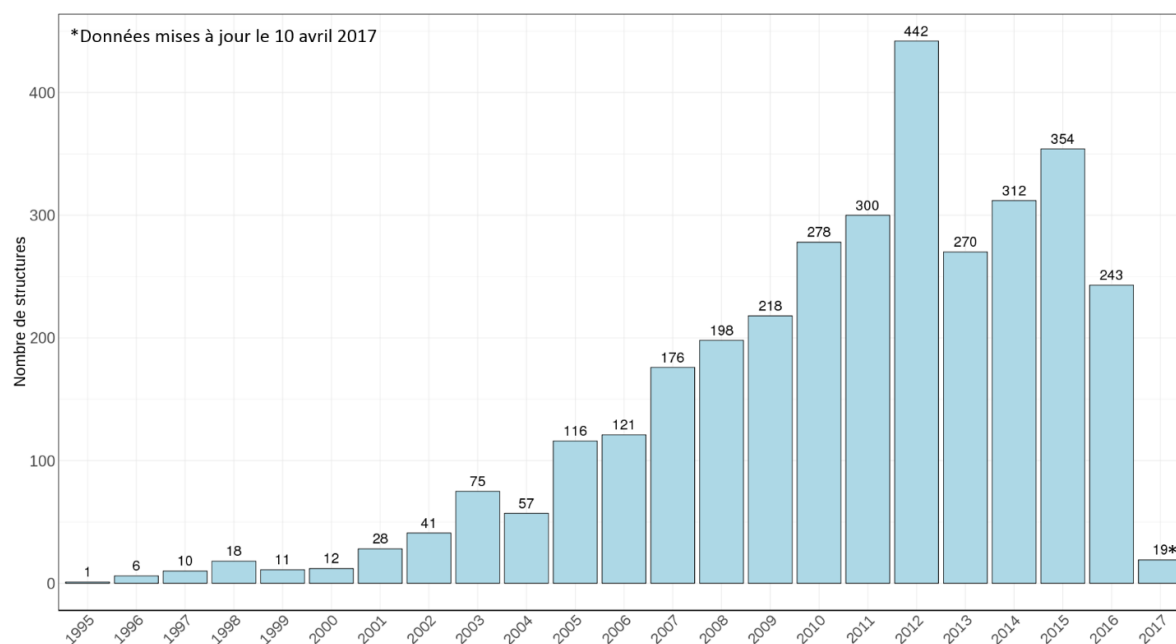


Figure 9. Evolution du nombre de structures de domaines kinases humaines dans la PDB. Cette figure représente les structures enregistrées de janvier 1995 à avril 2017. Le nombre de structures recensées comprend tous les types de kinases comportant les domaines PFAM PF07714 et PF00069.

³⁰ H. M. Berman *et al.*, The Protein Data Bank, *Acta Crystallographica. Section D, Biological Crystallography*, 58.Pt 6 No 1 (2002), 899–907.

De plus, au fur et à mesure de l'accumulation de structures de protéines kinases, il est très vite apparu que la grande similarité des séquences des domaines catalytiques se traduit par un repliement tridimensionnel commun à l'ensemble des protéines kinases. Nous verrons ainsi par la suite les éléments structuraux communs à ces protéines, avant de nous attarder plus en détails sur les deux conformations majoritaires qu'elles adoptent, selon si elles sont sous forme active ou inactive.

2.3.1 Description générale de la structure des domaines kinases

Le domaine kinase mesure entre 220 et 260 acides aminés et se compose de deux lobes réunis par une région charnière (**Figure 10**). Le plus petit lobe, le lobe N-terminal, se compose de cinq feuillets β antiparallèles ($\beta 1 - \beta 5$) et d'une hélice α (hélice C ou αC). Bien qu'elle appartienne au lobe N-terminal en termes de séquence, cette hélice occupe une position stratégique entre les deux lobes et permet de coupler l'ATP avec le feuillet $\beta 3$, au niveau du motif hautement conservé AxK (Ala - X - Lys, avec X un résidu quelconque) et plus particulièrement du résidu lysine appelé la lysine catalytique. La boucle reliant les deux premiers brins est appelée boucle P, ou boucle riche en glycines, et participe également à la bonne orientation de l'ATP au sein du site actif.

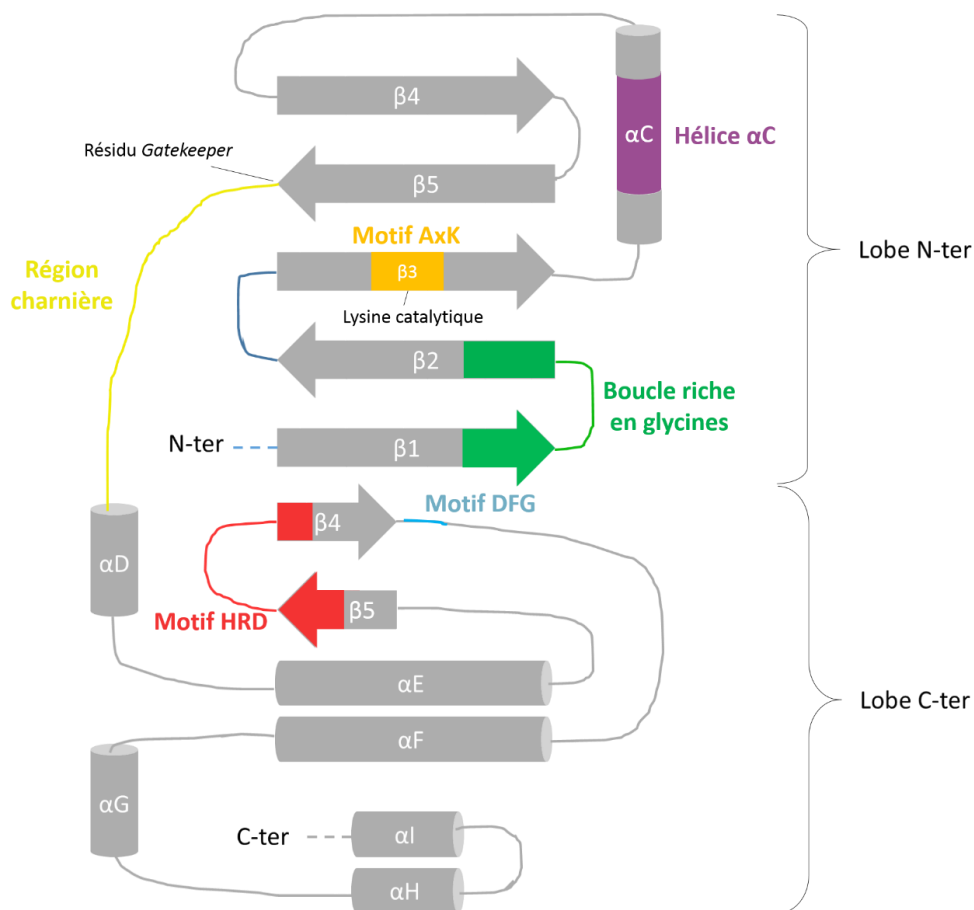


Figure 10. Schéma des structures secondaires d'un domaine kinase typique. Les principaux motifs structuraux importants pour la catalyse sont décrits en couleur tandis que le reste de la séquence est marqué en gris.

Le lobe C-terminal, qui est environ deux fois plus imposant que le premier, est composé de quatre brins β ($\beta 6 - \beta 9$) et de six à huit hélices α selon les structures. La boucle catalytique se trouve entre les brins $\beta 6$ et $\beta 7$ et contient le motif HRD (His - Arg - Asp) à son extrémité N-terminale. Ce motif serait responsable de l'orientation de l'hydroxyle du substrat, nécessaire à sa phosphorylation. La boucle d'activation, ou boucle A, se trouve quant à elle entre les brins $\beta 8$ et $\beta 9$. Elle comporte le motif DFG (Asp - Phe - Gly) dont le résidu Asp lie un ion métallique Mg^{2+} , ou Mn^{2+} , situé dans le site actif. Cet ion participe au positionnement des groupements phosphates β et γ , ce qui permet le transfert de ce dernier vers le substrat.³¹ La mutation d'au moins l'un des deux motifs cités (HRD et DFG) ou de la boucle riche en glycines (boucle P) aboutit à une perte de fonction de la protéine kinase. Les pseudokinases sont un bon exemple de ce cas de figure, à la différence près que cette perte de fonction n'est pas due à une mutation somatique et qu'elles remplissent d'autres rôles comme participer à la formation de complexes protéiques (*scaffolding function*).³² La région charnière, ou *hinge*, relie les deux sous-structures en connectant le brin $\beta 5$ du lobe C-terminal avec l'hélice αD du lobe N-terminal, ce qui confère une certaine flexibilité à la structure globale. Le premier résidu de ce motif en partant du brin $\beta 5$ est appelé *gatekeeper*, en raison de sa capacité à contrôler l'accès à deux poches hydrophobes adjacentes au site de fixation de l'adénine grâce à la taille de sa chaîne latérale.

Les protéines kinases jouant un rôle fondamental dans la signalisation moléculaire, elles font donc naturellement l'objet d'une régulation fine par la cellule (phosphorylation, dimérisation, translocation vers le noyau, etc.). Leur rôle d'activateurs/inactivateurs d'autres protéines, souvent en cascade, peut se traduire par un dysfonctionnement important des fonctions du vivant en cas de dérégulation, ne serait-ce que d'une seule protéine kinase. Pour cette raison, elles transitent régulièrement d'un état actif à un état inactif en fonction du mécanisme dans lequel elles sont impliquées. Plusieurs voies intrinsèques sont susceptibles de moduler l'activité phosphorylante des protéines kinases.

Deux différences conformationnelles principales ont été identifiées entre les formes actives et inactives des protéines kinases et portent sur les motifs DFG et hélice αC déjà mentionnés. Ceux-ci sont annotés *in* ou *out* selon leur orientation au sein du site actif et donc de leur participation à l'établissement d'interactions favorisant la liaison de l'ATP. Ainsi, si les protéines kinases actives ont besoin d'avoir leurs motifs structuraux en position *DFG-in* et *αC -helix in* pour assurer leur fonction, les protéines kinases inactives peuvent potentiellement adopter plusieurs combinaisons de conformations inactives : *DFG-in/ αC -helix out*, *DFG-out/ αC -helix in* et *DFG-out/ αC -helix out*. Cependant, il existe également toute une variété de positions intermédiaires pour ces motifs structuraux, ce qui rend compte de l'élasticité procurée par la structure bilobée du domaine kinase (**Figure 11**).

³¹ D. Fabbro *et al.*, Ten Things You Should Know about Protein Kinases: IUPHAR Review 14, *British Journal of Pharmacology*, 2015

³² J. Boudeau *et al.*, Emerging Roles of Pseudokinases, *Trends in Cell Biology* 16 (2006), no. 9:443–52

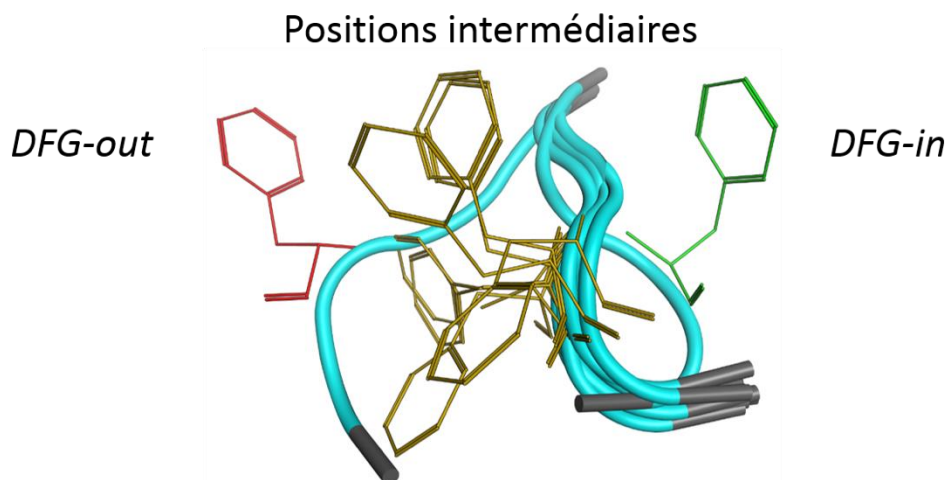


Figure 11. Représentation en ruban des conformations intermédiaires du résidu Phe du motif DFG par rapport aux configurations extrêmes DFG-out et DFG-in. Le motif DFG est coloré en fonction de son positionnement. En vert : DFG-in, en rouge : DFG-out, en jaune : positions intermédiaires.

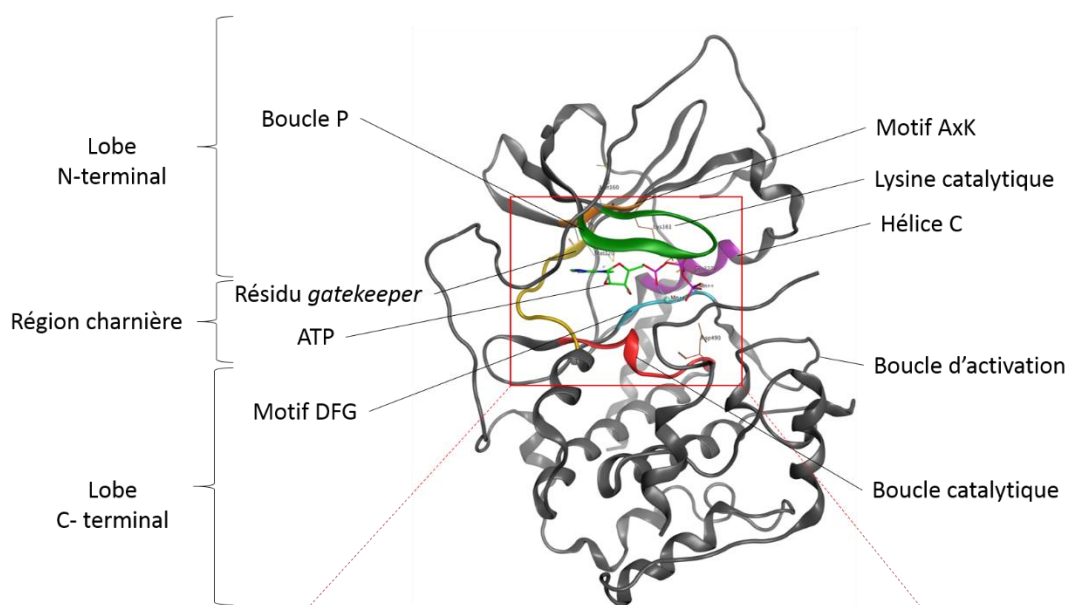
Puisque le rôle catalytique est le même chez toutes les protéines kinases, il est naturel de penser que les mécanismes de fixation de l'ATP et du substrat protéique sont également conservés. Afin de les décrire plus simplement sur l'ensemble des protéines kinases, nous allons prendre comme référence la numérotation des résidus présents dans le site actif de PRKACA (code PDB : 1ATP).

2.3.2 La conformation active

Comme son nom l'indique, la conformation active est le repliement tridimensionnel dans lequel la protéine kinase exerce une activité catalytique de phosphorylation (**Figure 12**).

Dans cette conformation, l'ATP se lie au domaine kinase au niveau de la cavité formée par les deux lobes et la région charnière. Le groupement adénine est entouré de résidus hydrophobes (Val57, Val123, Leu173, Phe327) et réalise des liaisons hydrogène avec des résidus de la région charnière (Met120, Glu121, Val123, Thr183). Le ribose est quant à lui stabilisé par des liaisons hydrogène avec des résidus du lobe C-terminal (Glu127, Glu170). La Lys72 (lysine catalytique) du motif AxK situé sur le brin β 3 forme un pont salin avec la Glu91 de l'hélice α C, ce qui contribue au bon positionnement des groupements phosphates α et β de l'ATP. Le groupement phosphate γ de l'ATP est positionné de manière adéquate pour son transfert vers le substrat protéique grâce à l'importante flexibilité de la boucle P. Le résidu Asp166 du motif HRD, localisé sur la boucle catalytique, joue le rôle d'accepteur pour le transfert du groupement phosphate γ vers la protéine cible. L'Asp184 du motif DFG lie un ion Mg^{2+} ou Mn^{2+} situé dans la cavité et qui participe au bon positionnement des groupements phosphates β et γ . La boucle d'activation permet de positionner le groupement hydroxyle du peptide substrat. L'hélice F, sur le lobe C-terminal, permet la fixation du substrat une fois la boucle A correctement positionnée. Une fois la phosphorylation effectuée, les produits de la catalyse, l'ADP et le substrat phosphorylé sont libérés.

A)



B)

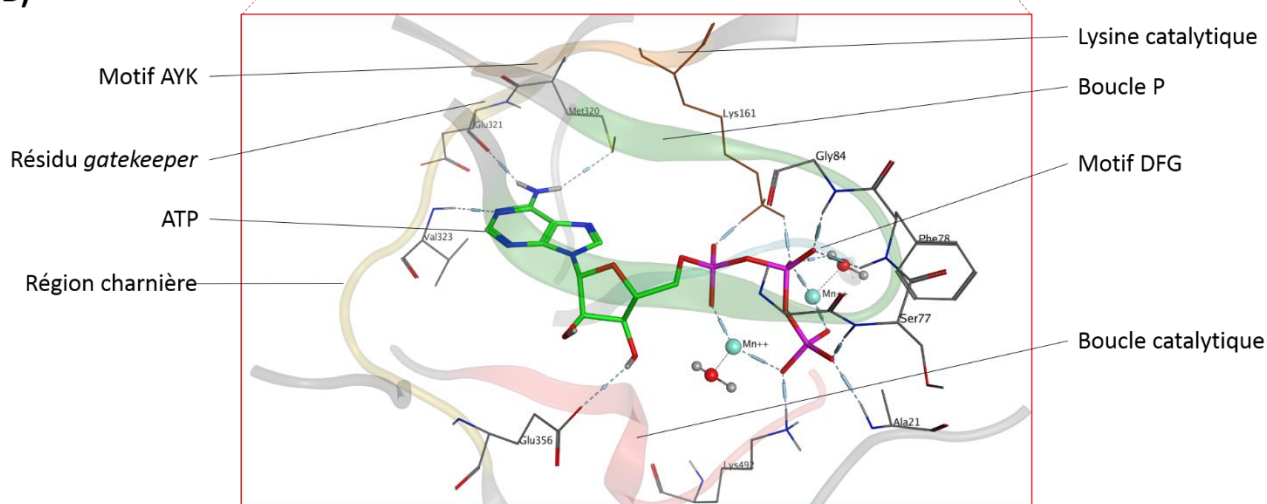


Figure 12. Représentation en ruban du domaine kinase de PRKACA en conformation active. Code PDB : 1ATP. A) Structure complète avec motifs conservés coloriés : boucle P (vert), région charnière (jaune), motif AxK (orange), motif DFG (bleu), boucle catalytique (rouge), hélice C (violet). B) Agrandissement du site catalytique (encadré rouge).

Nous avons vu précédemment que les différents motifs présents sur les deux lobes du domaine kinase contribuent chacun à la catalyse de manière différente. Cependant, il existe d'autres motifs dont les résidus sont non consécutifs mais assemblés dynamiquement lorsque le domaine kinase adopte une conformation active. Ces structures éphémères ont été identifiées par analyse de motifs spatiaux plutôt que de séquences protéiques.³³ Elles sont appelées épines hydrophobes (*hydrophobic spines*) en raison de

³³ S. S. Taylor *et al.*, Protein Kinases: Evolution of Dynamic Regulatory Proteins, *Trends Biochem. Sci.*, 36.2 (2011), 65–77

leur nature de motifs structuraux alignés dans l'espace et composés de constituants hydrophobes (**Figure 13**).

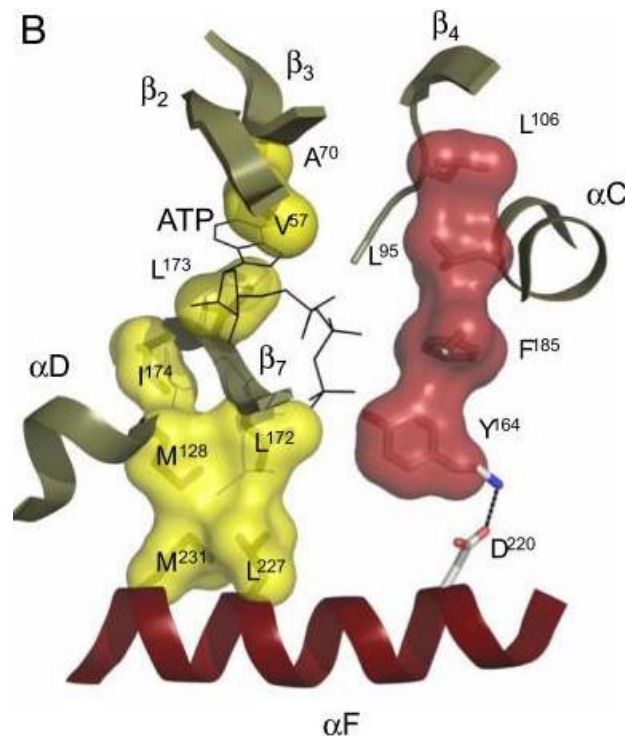


Figure 13. Représentation en surface des épines C et R sur une structure de Prkaca. En jaune : épine C, en rouge : épine R. Le groupement adénine de l'ATP s'intercale entre les deux lobes de la protéine pour stabiliser l'épine C.³⁴

La première, l'épine régulatrice (*Regulatory spine* ou *R spine*), se compose de quatre résidus (deux sur chaque lobe) dont l'alignement dépend de la position de la boucle d'activation. Lorsque la boucle est en conformation inactive, il n'y a pas de formation de l'épine régulatrice. Le second motif, l'épine hydrophobe catalytique (*Catalytic spine* ou *C spine*), rassemble également des résidus des deux lobes (2 pour le N-ter et un pour le C-ter) mais se distingue de la première en alignant également le noyau adénine de l'ATP, ce qui semble stabiliser sa position au sein du site actif en vue de la catalyse.

L'activation du domaine kinase peut être constitutive, mais elle passe en général par la phosphorylation de certains résidus clés. Par exemple, PRKACA est phosphorylée sur la Thr197 tandis qu'INSR est activée par la phosphorylation de trois acides aminés Tyr présents sur la boucle d'activation. D'autres mécanismes peuvent également participer à la modulation de l'activité kinase. Ainsi, la phosphorylation de certains sous-domaines régulateurs, comme par exemple le domaine SH2, permet l'activation des protéines kinases. Les mécanismes connus de régulation par d'autres domaines de la protéine kinase sont divers : repliement structural permettant une conformation active du domaine

³⁴ A. P. Kornev *et al.*, A Helix Scaffold for the Assembly of Active Protein Kinases, *Proceedings of the National Academy of Sciences of the United States of America*, 105.38 (2008), 14377–82

kinase, localisation de la protéine dans le même compartiment cellulaire que son substrat, marquage pour dégradation, etc.²²

2.3.3 Les conformations inactives *DFG-out* et α C-helix out

En 1995 est déposée dans la PDB la structure du domaine kinase du récepteur à l'insuline (INSR), présentant un repliement tridimensionnel atypique (code PDB : 1IRK).³⁵ Dans cette conformation, plus tard appelée *DFG-out*, la boucle d'activation, et notamment le motif DFG, sont considérablement déplacés par rapport à leurs conformations dans les autres structures disponibles à l'époque. Le résidu Phe du motif DFG subit une rotation de l'ordre de 180° et se retrouve dans une poche située au niveau du site de liaison de l'adénosine de l'ATP. Ce dernier, ainsi que le substrat protéique, ne peuvent alors plus se fixer (**Figure 14**). D'autres structures présentant le même repliement seront par la suite résolues pour d'autres protéines kinases, avec ou sans ligand, présentant la conformation *DFG-out* comme une conformation conservée des protéines kinases inactives.

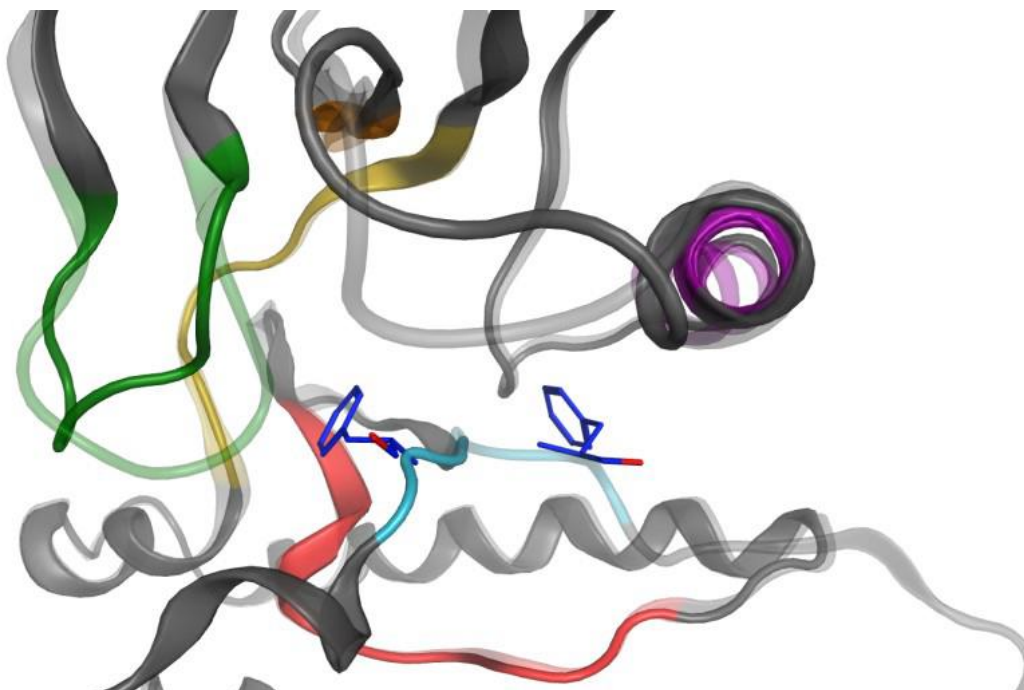


Figure 14. Représentation en ruban des structures de ABL1 en conformation *DFG-out* et en conformation *DFG-in*.³⁶
Codes PDB : *DFG-out* (2HIW), *DFG-in* (2GQG). Cette dernière est représentée en transparent pour souligner le déplacement de la Phe (en bleu foncé) du motif DFG et de la boucle d'activation.

En 1996, une structure résolue de CDK2 présente elle aussi une conformation atypique (code PDB : 1HCL), plus tard définie comme α C-helix out. Cet état se manifeste par la présence d'une cavité inter-lobale

³⁵ S. R. Hubbard *et al.*, Crystal Structure of the Tyrosine Kinase Domain of the Human Insulin Receptor, *Nature*, 372.6508 (1994), 746–54

³⁶ N. Bosc, Développement de nouvelles approches protéo-chimométriques appliquées à l'étude des interactions et de la sélectivité des inhibiteurs de kinases (Université d'Orléans, 2015)

plus étroite due à un glissement relatif des deux lobes, l'un par rapport à l'autre. De plus, la partie N-terminale de l'hélice C se trouve en retrait par rapport au site catalytique. Il en résulte la rupture du pont salin entre le résidu Glu51 de l'hélice C et la lysine catalytique (Lys33) du brin β 3. Enfin, la boucle d'activation se replie sur elle-même, son extrémité N-terminale formant une petite hélice tournée vers l'hélice α C (**Figure 15**). Dans cette conformation, seul l'ATP, et non le substrat protéique, peut se fixer dans le site actif, prévenant toute activité phosphorylante (code PDB : 1HCK). Dans le cas de cette protéine kinase, la transition en conformation active s'effectue par la complexation avec une autre protéine de la famille des cyclines.

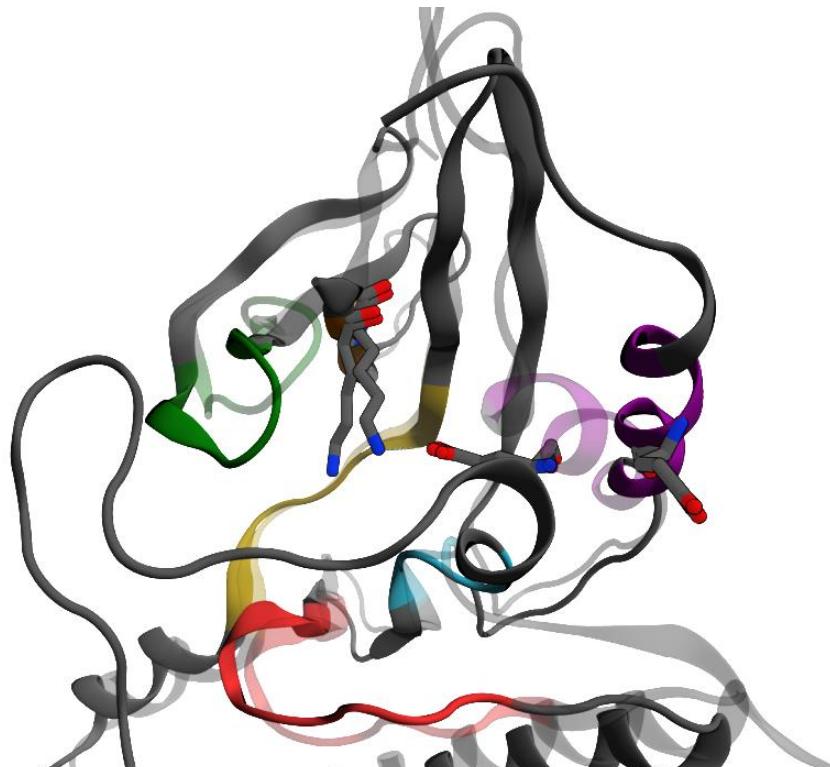


Figure 15. Représentation en ruban des structures de CDK2 en conformation α C-helix out et en conformation α C-helix in.³⁶ Codes PDB : α C-helix out (1HCL), α C-helix in (1JST). Cette dernière est représentée en transparent pour souligner le déplacement de l'hélice C et la modification de la boucle d'activation.

2.4 Les inhibiteurs de protéines kinases

Jusqu'à la fin des années 1990, les compagnies pharmaceutiques se montraient réticentes à l'idée de travailler sur le développement d'inhibiteurs de protéines kinases. Tout d'abord, l'ATP étant présent à des concentrations élevées dans la cellule (de 1 à 5mM), il paraissait en effet difficile de développer des inhibiteurs qui en seraient compétitifs.³⁷ D'autre part, la grande similarité de séquence entre les protéines kinases laissait présager des problèmes de sélectivité et donc une toxicité et des effets secondaires importants.²⁹

³⁷ Z. A. Knight *et al.*, Features of Selective Kinase Inhibitors, *Chemistry & Biology*, 12.6 (2005), 621–37

Les résultats très positifs obtenus durant les phases de développement clinique avec l'imatinib, développé par Novartis pour lutter contre la leucémie myéloïde chronique (CML), ont encouragé les autres compagnies pharmaceutiques à démarrer leurs propres projets portant sur des protéines kinases.³⁸ Ainsi, de nombreuses petites molécules ont été développées avec, comme objectif, l'occupation du site actif du domaine kinase pour empêcher la fixation de l'ATP. L'imatinib a ensuite été approuvé pour la commercialisation par l'agence américaine des produits alimentaires et médicamenteux (*Food and Drug Administration* ou FDA) en 2001 sous l'appellation Gleevec® et a largement contribué au chiffre d'affaire de Novartis, jusqu'à la perte d'exclusivité de son exploitation en 2016.³⁹

Les protéines kinases représentent aujourd'hui l'une des familles thérapeutiques les plus étudiées par l'industrie pharmaceutique, avec environ 7% des parts du marché, soit 240 milliards de dollars de recettes par an.⁴⁰ On dénombre actuellement non moins de 30 petites molécules approuvées par la FDA et 150 autres molécules en phases cliniques ou précliniques.

L'accumulation au cours du temps de structures cristallographiques de complexes domaine kinase/inhibiteur a permis de mettre en lumière des mécanismes d'inhibition communs. Les inhibiteurs se classent ainsi en fonction de leur mode d'interaction avec leur récepteur.⁴¹ Parmi les inhibiteurs non covalents, on distingue trois types de mode d'interaction compétitif de l'ATP (types I, I½ et II) et deux types de mode d'interaction allostérique non compétitif de l'ATP (types III et IV). De plus, les inhibiteurs de type V, également appelés inhibiteurs bivalents, sont capables de se lier à la fois au site actif et à un autre domaine de la kinase. Enfin, les inhibiteurs de type VI se lient de manière covalente à leur cible.

2.4.1 Les Inhibiteurs de type I

Alors que leur mode d'interaction était encore inconnu, les inhibiteurs des protéines kinases étaient supposés entrer en compétition avec l'ATP en occupant le site actif à la place du substrat endogène, empêchant alors toute activité phosphorylante. La petite molécule est donc censée mimer les interactions formées normalement par l'ATP dans le domaine kinase en conformation active.⁴² C'est par ce concept que sont aujourd'hui définis les inhibiteurs de type I.

Ces derniers comportent généralement un noyau de type adénine capable de former des interactions favorables avec les motifs conservés du domaine catalytique. En règle générale, les molécules

³⁸ L. N. Johnson, Protein Kinase Inhibitors: Contributions from Structure to Clinical Compounds, *Quarterly Reviews of Biophysics*, 42.1 (2009), 1–40

³⁹ Novartis Delivered Solid 2016 Performance, with Growth Products Absorbing Gleevec US LOE; Innovation Momentum Continued; Announces Share Buyback, *Novartis*, <https://www.novartis.com/news/media-releases/novartis-delivered-solid-2016-performance-growth-products1-absorbing-gleevec-us>, consulté le 2 Octobre 2017.

⁴⁰ Poster : Unexplored Opportunities in the Druggable Human Genome, <http://www.nature.com/nrd/posters/druggablegenome/index.html>, consulté le 3 Octobre 2017.

⁴¹ R. Roskoski, Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes, *Pharmacological Research*, 103 (2016), 26–48

⁴² F. Zuccotto *et al.*, Through the "Gatekeeper Door": Exploiting the Active Kinase Conformation, *Journal of Medicinal Chemistry*, 53.7 (2010), 2681–94

de cette classe peuvent établir jusqu'à trois liaisons hydrogène avec la région charnière⁴² tout en reproduisant les interactions stabilisantes de type π - π avec les résidus hydrophobes du feuillet β 2 présent sur le lobe C-terminal et du feuillet β 6 présent sur le lobe N-terminal (**Figure 16**).⁴¹

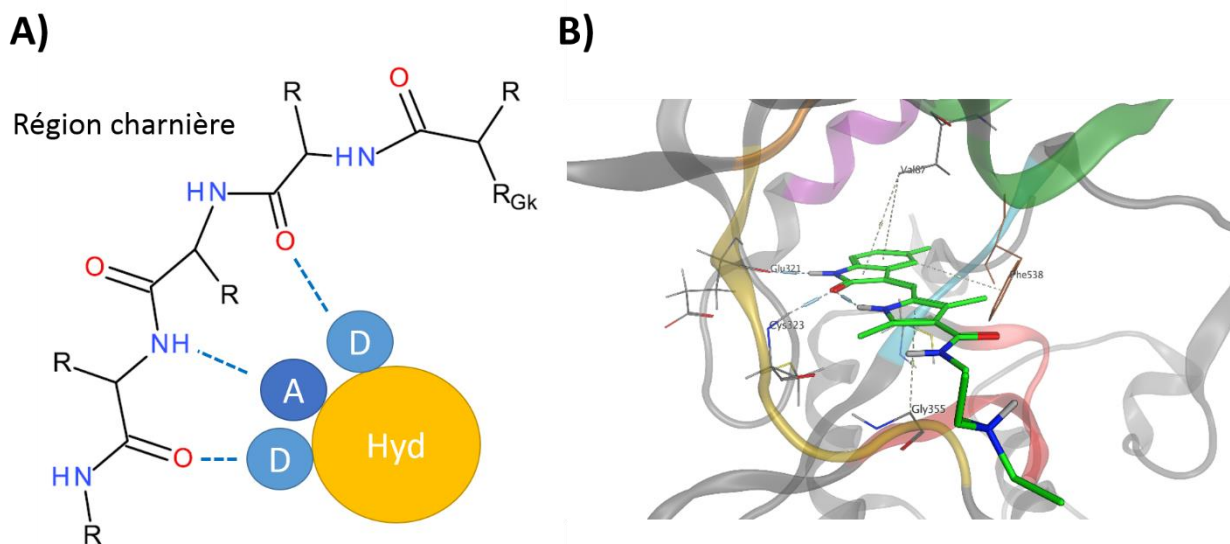
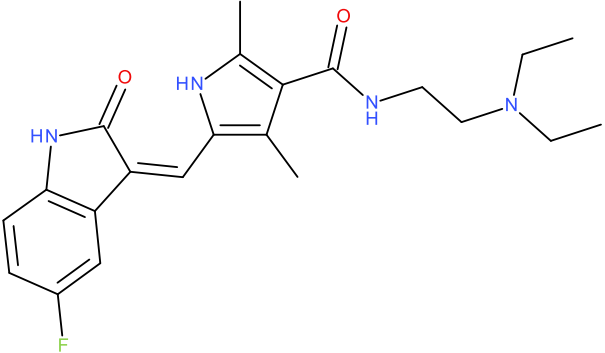
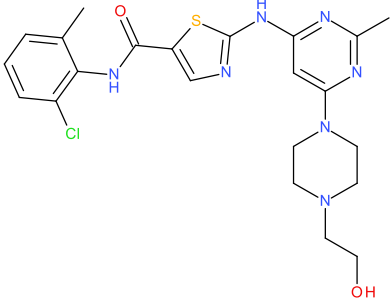
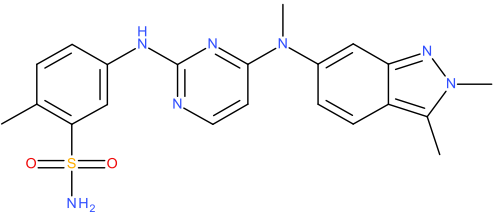
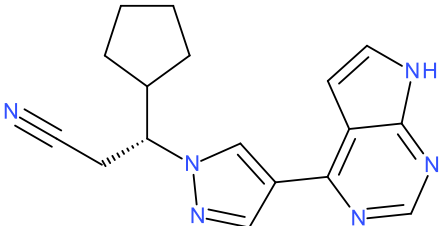


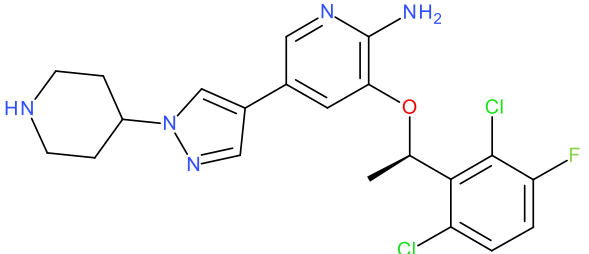
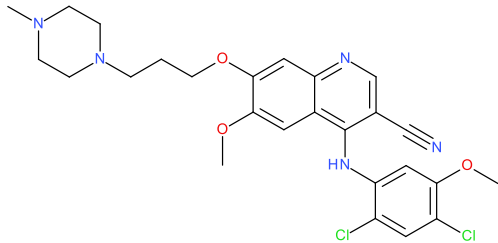
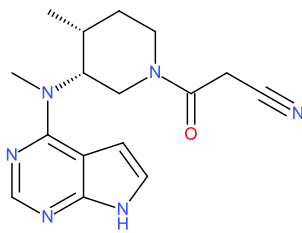
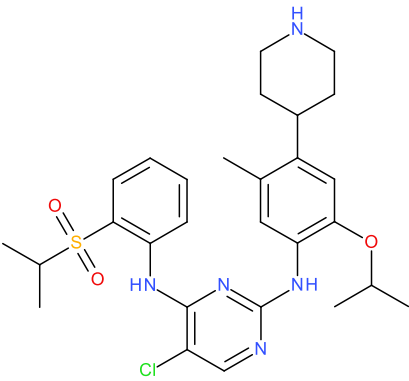
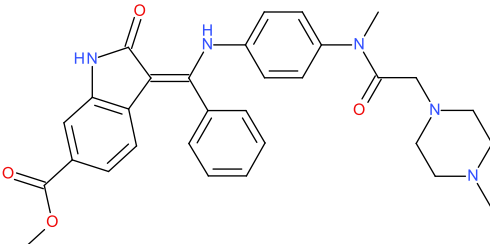
Figure 16. Représentation du modèle d'interaction d'un inhibiteur de type I. A) Modèle pharmacophorique. Hyd : région hydrophobe, A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés. B) Exemple de l'inhibiteur sunitinib interagissant avec la protéine kinase KDR (code PDB : 4AGD).

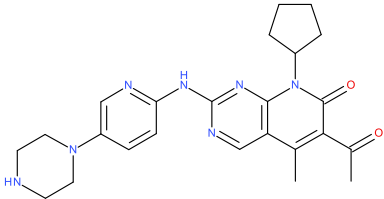
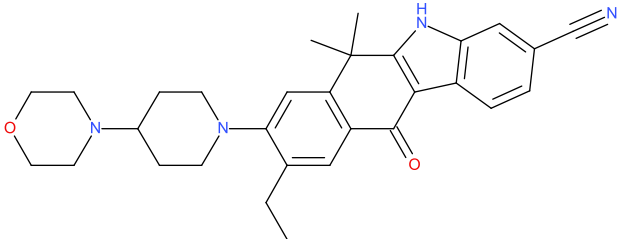
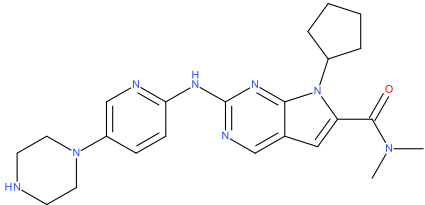
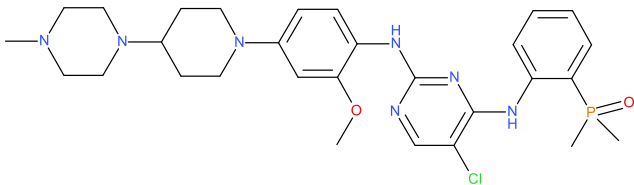
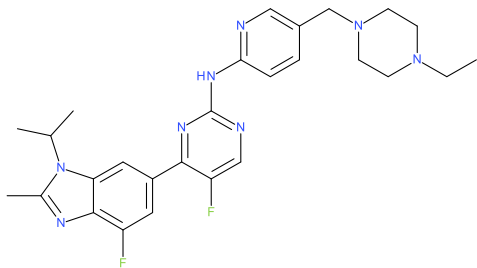
Cette classe d'inhibiteurs est la première à avoir été développée car, initialement, seules les formes phosphorylées des protéines kinases étaient employées dans les essais biochimiques.⁴² Elle peut donc se révéler particulièrement intéressante dans le cas où la protéine ciblée adopte majoritairement une conformation active. Cependant, du fait de la structure hautement conservée des domaines kinases en conformation active, il peut se révéler difficile d'accéder à une bonne sélectivité de ces composés. De plus, cette classe de molécules, présentant un noyau aromatique de type adénine, a déjà été hautement étudiée par l'industrie pharmaceutique, ce qui peut rendre compliquée la découverte de nouveaux chémotypes encore libres d'exploitation.

Certaines molécules telles que l'imatinib ont été initialement conçues comme des inhibiteurs de type I, mais les structures obtenues par co-cristallisation avec leurs cibles ont révélé un autre mode de fixation : celui d'un inhibiteur de type II. Les modes d'interaction protéine-ligand sont donc maintenant systématiquement validés par cristallographie aux rayons X afin de confirmer les mécanismes d'action des molécules développées. Il existe actuellement 13 inhibiteurs de type I approuvés par la FDA (**Table 2**), sunitinib étant le premier vrai inhibiteur de type I approuvé par la FDA en 2006.

Table 2. Liste des 10 inhibiteurs de protéines kinases de type I approuvés par la FDA

Structure chimique	Nom générique Entreprise Année d'approbation	Cibles principales	Indications	Codes PDB
	Sunitinib Pfizer 2006	PDGFR α/β , VEGFR1/2/3, Kit, Flt3, CSF- 1R, RET	Tumeur stromale gastro- intestinale, cancer du rein	4QMZ, 4KS8, 4AGD, 3TI1, 2Y7J, 3MIY, 3G0E, 3G0F
	Dasatinib BMS 2006	BCR-Abl, Src, Lck, Yes, Fyn, Kit, EphA2, PDGFR β	Leucémie myéloïde chronique (CML)	4QMS, 4XEY, 4XLI, 3QLG, 2Y6O, 3SXR, 3OHT, 3LFA, 3OCT, 3K54, 3G5D, 2ZVA, 2GQG
	Pazopanib GSK 2009	VEGFR1/2/3, PDGFR α/β , FGFR1/3, Kit, Lck, Fms, Itk	Cancer du rein	
	Ruxolitinib Incyte 2011	JAK1/2, Src	Myélofibrose	4U5J

	Crizotinib Pfizer 2011	ALK, c-Met (HGFR), ROS, MST1R	Cancer métastatique non à petites cellules du poumon (mNSCLC)	4C9W, 3ZBF, 4ANQ, 4ANS, 2YFX, 2XP2, 2WGJ
	Bosutinib Wyeth 2012	BCR-Abl, Src, Lyn, Hck	CML	4QMN, 5AJQ, 4OTW, 4MXY, 4MXO, 4MXX, 4MXZ, 3UE4, 3SOA
	Tofacitinib Pfizer 2012	JAK1/3	Arthrite rhumatoïde (RA)	4OTI, 3LXK, 3LXN, 3FUP, 3EYG
	Ceritinib Novartis 2014	ALK, IGF-1R, InsR, ROS1	mNSCLC	4MKC
	Nintedanib Boehringer Ingelheim 2014	FGFR1/2/3, PDGFRα/β, VEGFR1/2/3, Flt3	Fibrose pulmonaire idiopathique (IPF)	3C7Q

	Palbociclib Park Davis 2015	CDK4/6	Cancer du sein	2EUF
	Alectinib Hoffmann-La Roche Inc 2015	ALK	Cancer bronchique non à petites cellules	3AOX
	Ribociclib Novartis / Astex Pharmaceuticals 2017	CDK4/CDK6	Cancer du sein	5L2T
	Brigatinib ARIAD Pharmaceuticals, Inc. 2017	ALK, EGFR	NSCLC	5J7H
	Abemaciclib Eli Lilly 2017	CDK4, CDK6	Cancer du sein	5L2S

2.4.2 Les Inhibiteurs de type II

Les inhibiteurs de type II ciblent et stabilisent la conformation inactive *DFG-out* des protéines kinases. Ce type d'inhibition a été découvert par hasard. En effet, comme évoqué au paragraphe précédent, le premier inhibiteur de type II, imatinib, a été initialement développé pour être un inhibiteur de type I.⁴³ C'est en observant son mode de liaison avec Abl1 par cristallographie (code PDB : 1IEP) que les chercheurs ont pu identifier ce nouveau mécanisme.⁴⁴ L'utilisation d'une protéine non phosphorylée, et donc en conformation inactive, a en effet permis d'observer un déplacement important de la boucle d'activation, dont la Phe du motif DFG effectue une rotation de 180° vers l'intérieur du site de liaison de l'ATP. Après cette transition *DFG-in* vers *DFG-out*, l'ATP ne peut plus se fixer mais une poche hydrophobe est dégagée en contrepartie. Le ligand est ainsi stabilisé par des liaisons hydrogène avec la région charnière, la poche de l'ATP et la poche allostérique juxtaposée. On peut également diviser les inhibiteurs de type II en deux sous-catégories selon qu'ils forment des interactions avec la poche hydrophobe allostérique ou non (respectivement II A et II B).⁴² Bien qu'ils effectuent des interactions avec la région charnière de la protéine kinase, les inhibiteurs de type II ne sont pas des inhibiteurs compétitifs de l'ATP puisque ce dernier ne se fixe pas en conformation inactive.

L'inhibition de type II est considérée possible dans le cas de protéines kinases possédant un résidu *gatekeeper* de petite taille⁴², afin de permettre au ligand l'accès à la poche hydrophobe adjacente. Un exemple typique d'inhibiteur de type II est imatinib (**Figure 17**). Il existe cependant une structure du domaine kinase de FLT3, dont le résidu *gatekeeper* est un acide aminé volumineux (Phe), cristallisée en présence de l'inhibiteur de type II quizartinib (code PDB : 4XUF). Ce type d'inhibition est donc tout de même possible dans certains cas (**Figure 18**).

Les conformations inactives des protéines kinases présentant un repliement tridimensionnel moins conservé que la conformation active, les inhibiteurs de type II sont en général considérés plus sélectifs et plus affins que les inhibiteurs de type I.⁴⁵ Cependant, des études récentes de sélectivité sur un panel de 300 protéines kinases ont montré que certains inhibiteurs de type I et II pouvaient partager des profils de sélectivité similaires, allant de peu à très sélectifs.⁴⁶

⁴³ R. Capdeville *et al.*, Glivec (STI571, Imatinib), a Rationally Developed, Targeted Anticancer Drug, *Nature Reviews. Drug Discovery*, 1.7 (2002), 493–502

⁴⁴ T. Schindler *et al.*, Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase, *Science (New York, N.Y.)*, 289.5486 (2000), 1938–42

⁴⁵ D. K. Treiber *et al.*, Ins and Outs of Kinase DFG Motifs, *Chemistry & Biology*, 20.6 (2013), 745–46

⁴⁶ M. I. Davis *et al.*, Comprehensive Analysis of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 29.11 (2011), 1046–51

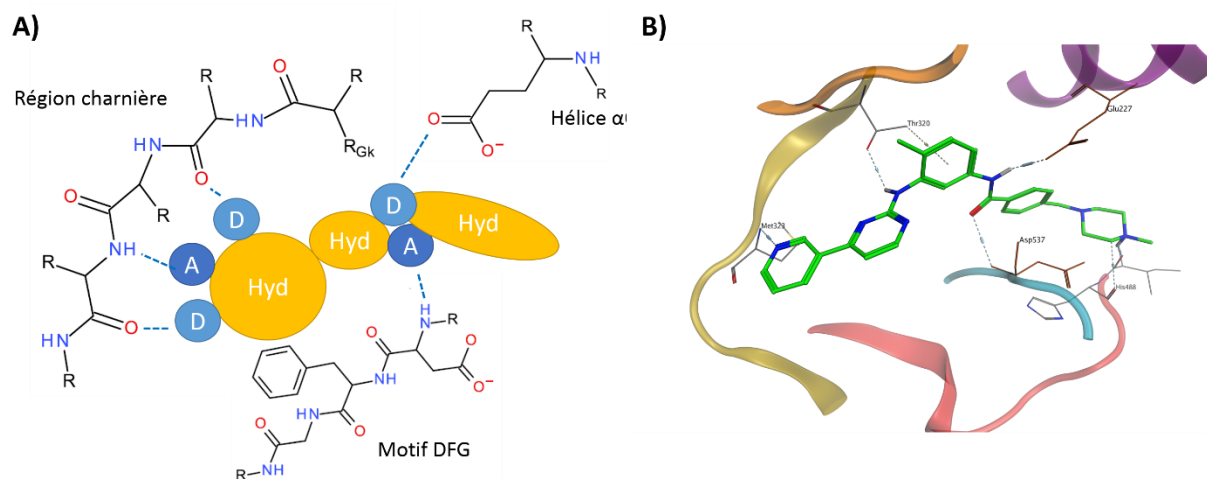


Figure 17. Représentation du modèle d'interaction d'un inhibiteur de type II, la protéine kinase est en conformation inactive DFG-out. A) Représentation du modèle pharmacophorique. Hyd : région hydrophobe, A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés. B) Exemple de l'inhibiteur imatinib interagissant avec la protéine kinase Abl1 (code PDB : 2HYY).

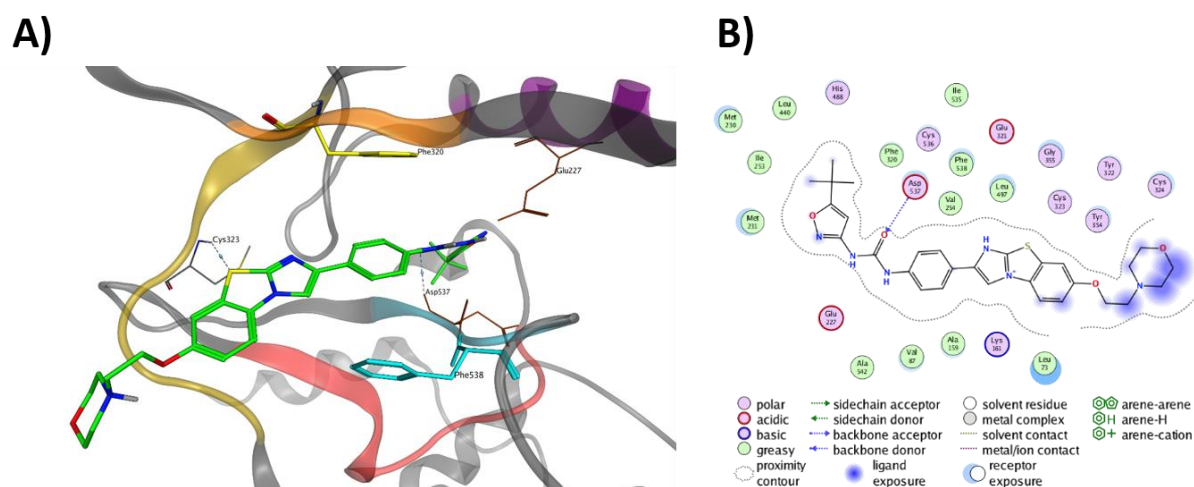
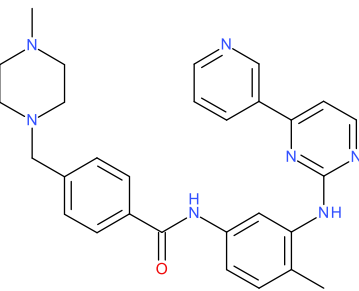
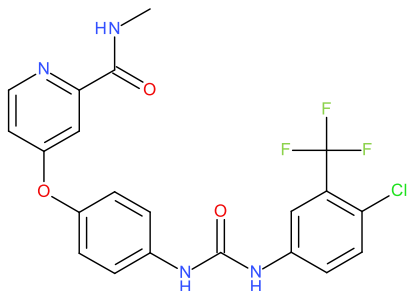
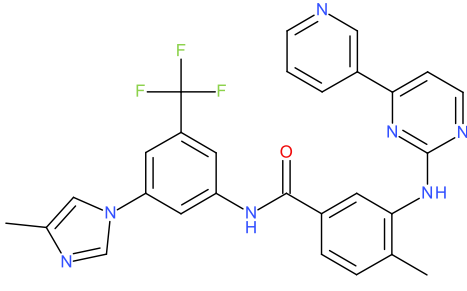
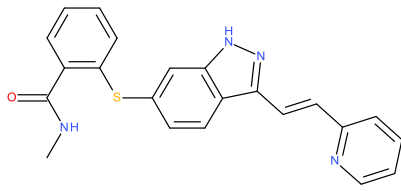
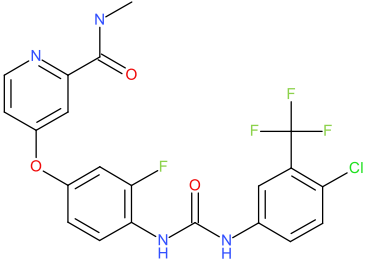
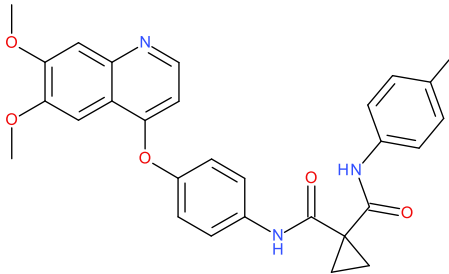
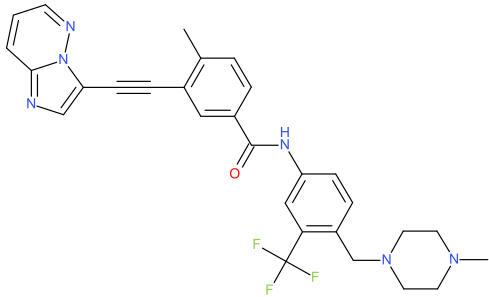


Figure 18. Mode d'interaction de l'inhibiteur de type II quizartinib dans une structure de FLT3 en conformation inactive (code PDB : 4XUF). En jaune est représenté le résidu gatekeeper (Phe) et en bleu le résidu Phe du motif DFG.

Il existe actuellement 7 inhibiteurs de type II approuvés par la FDA dont 4 ont été approuvés la même année, en 2012 (Table 3). Il n'y a cependant plus eu d'autre inhibiteur de protéine kinase approuvé par la FDA depuis.

Table 3. Liste des 7 inhibiteurs de protéines kinases de type II approuvés par la FDA

Structure chimique	Nom générique Entreprise Année d'approbation	Cibles principales	Indications	Codes PDB
	Imatinib Novartis 2001	BCR-Abl, Kit, PDGFR	CML	4CSV, 4BKJ, 3OEZ, 3PYY, 3MS9, 3MSS, 3K5V, 3HEC, 3GVU, 3FW1, 2PL0, 2OIQ, 2HYY, 1XBB, 1T46, 1OPJ, 1IEP
	Sorafenib Bayer 2005	B/C-Raf, B- Raf (V600E), Kit, Flt3, RET, VEGFR1/2/3, PDGFRβ	Carcinome rénal	3WZE, 4ASD, 3RGF, 3HEG, 3GCS, 1UWH, 1UWJ
	Nilotinib Novartis 2007	BCR-Abl, PDGFR, DDR1	CML	3CS9, 3GP0
	Axitinib Pfizer 2012	VEGFR1/2/3, PDGFRβ	Carcinome rénal	4TWP, 4WA9, 4AG8, 4AGC

	Regorafenib Bayer 2012	VEGFR1/2/3, BCR-Abl, B-Raf, B-Raf (V600E), Kit, PDGFRα/β, RET, FGFR1/2, Tie2, Eph2A	Cancer colorectal métastatique (mCRC)	-
	Cabozantinib Exelixis 2012	RET, Met, VEGFR1/2/3, Kit, TrkB, Flt3, Axl, Tie2	MTC	3B8R
	Ponatinib ARIAD Pharmaceuticals, Inc. 2012	BCR-Abl, BCR-Abl T315I, VEGFR, PDGFR, FGFR, EphR, Src, Kit, RET, Tie2, Flt3	CML	4V01, 4V04, 4UXQ, 4QRC, 4U0I, 4TYJ, 4C8B, 3ZOS, 3OXZ, 3IK3

2.4.3 Les Inhibiteurs de type I^{1/2}

Nous avons vu que les inhibiteurs de type II, en ciblant la conformation inactive *DFG-out* du domaine kinase, permettent d'éviter certains problèmes rencontrés avec les inhibiteurs de type I, notamment au niveau de la sélectivité et de la propriété intellectuelle. Ils ne sont donc en général pas utilisables avec des protéines kinases adoptant majoritairement une conformation active *DFG-in*.³¹ De plus, les techniques de criblage actuelles se focalisent essentiellement sur l'identification de molécules compétitives de l'ATP en utilisant des domaines phosphorylés et donc en conformation active.

Les inhibiteurs de type I^{1/2}, également appelés types I sélectifs, tentent de combiner les avantages des classes I et II décrites précédemment. Ils peuvent cibler les conformations active *DFG-in* ou inactive *DFG-in* et *αC-out*. Ils exploitent un site additionnel par rapport aux inhibiteurs de type I, proche du site de fixation de l'ATP, où ce dernier ne forme aucune interaction. Ce site est une petite poche hydrophobe adjacente à celle formée par les acides aminés hydrophobes du brin β2 et se situe derrière le site de liaison de l'adénine (**Figure 19**). La forme et le volume de cette poche dépendent principalement du résidu *gatekeeper*.

Le fait que ce résidu ne soit que très peu conservé parmi les différentes familles de protéines kinases donne l'opportunité d'une plus grande sélectivité des inhibiteurs de type I½. En effet, les inhibiteurs de cette classe ne peuvent pas se fixer dans les protéines kinases présentant un résidu *gatekeeper* volumineux, ce qui réduit les possibilités d'inhiber la mauvaise cible.

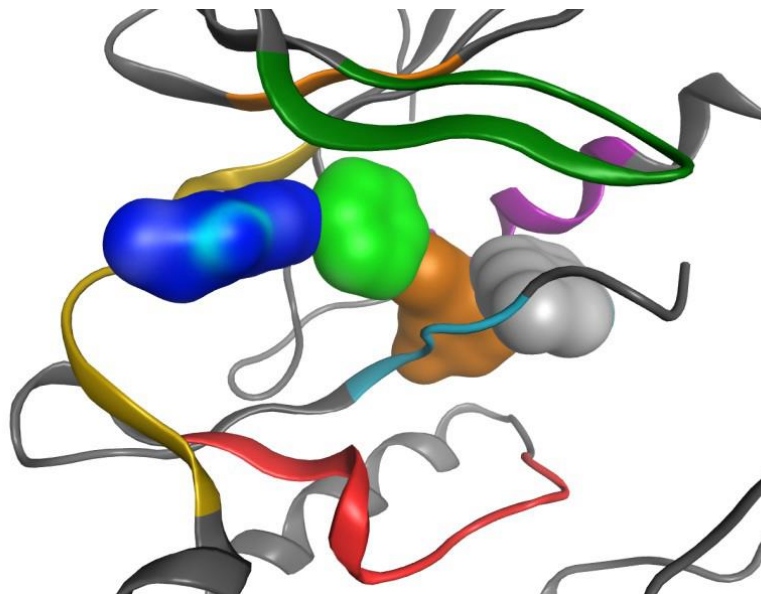
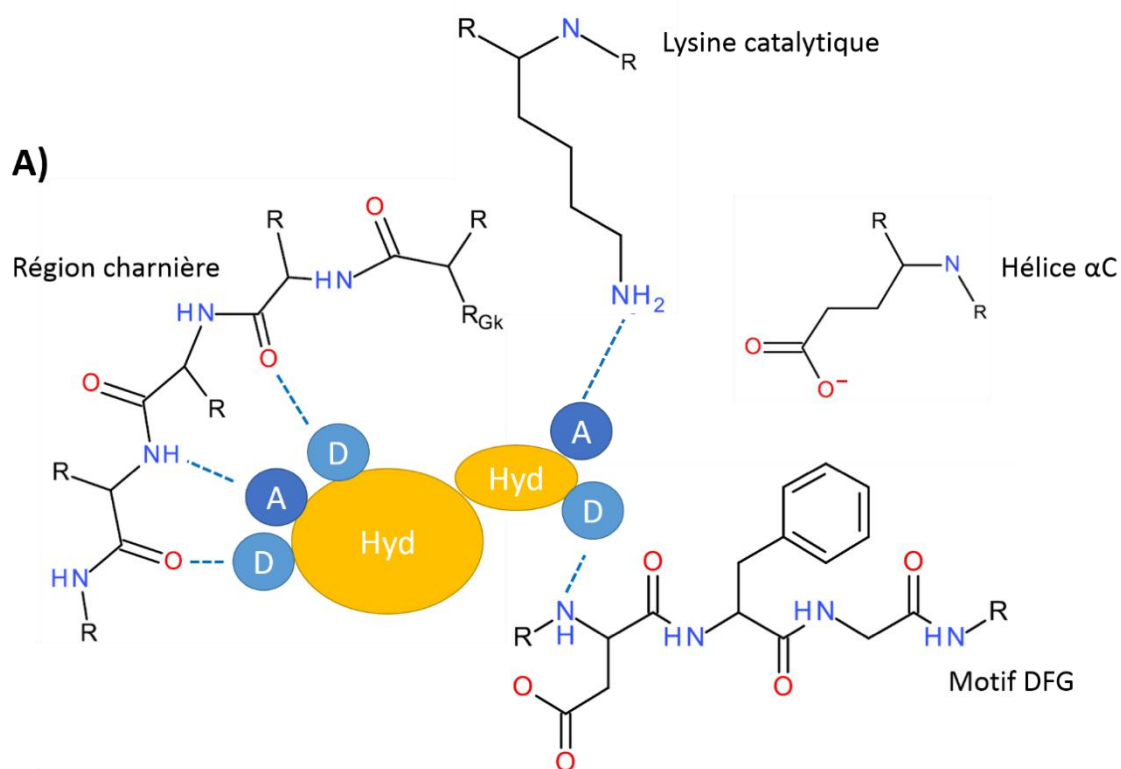


Figure 19. Représentation de différentes poches situées dans le domaine kinase de ABL1.³⁶ Code PDB : 1IEP. Le site de liaison de l'adénine est représenté en bleu, la poche hydrophobe additionnelle en vert, la poche hydrophobe adjacente en orange et la poche allostérique en gris.

Un bon exemple de ce type de molécules est vemurafenib (**Figure 20**) qui inhibe la protéine kinase BRAF. Dans sa forme mutée V600E, l'acide aminé Phe du motif DFG en position *in* bloque l'accès à une grande partie de la poche allostérique. Le groupement propylsulfonamide de l'inhibiteur permet de contourner ce problème et pénétrer profondément dans le site actif pour réaliser une liaison hydrogène avec le résidu Asp du motif DFG.⁴⁷

Il existe actuellement 7 inhibiteurs de type I½ de protéines kinases dont l'exploitation commerciale a été approuvée par la FDA (**Table 4**).

⁴⁷ G. Bollag *et. al.*, Vemurafenib: The First Drug Approved for BRAF-Mutant Cancer, *Nature Reviews. Drug Discovery*, 11.11 (2012), 873–86



B)

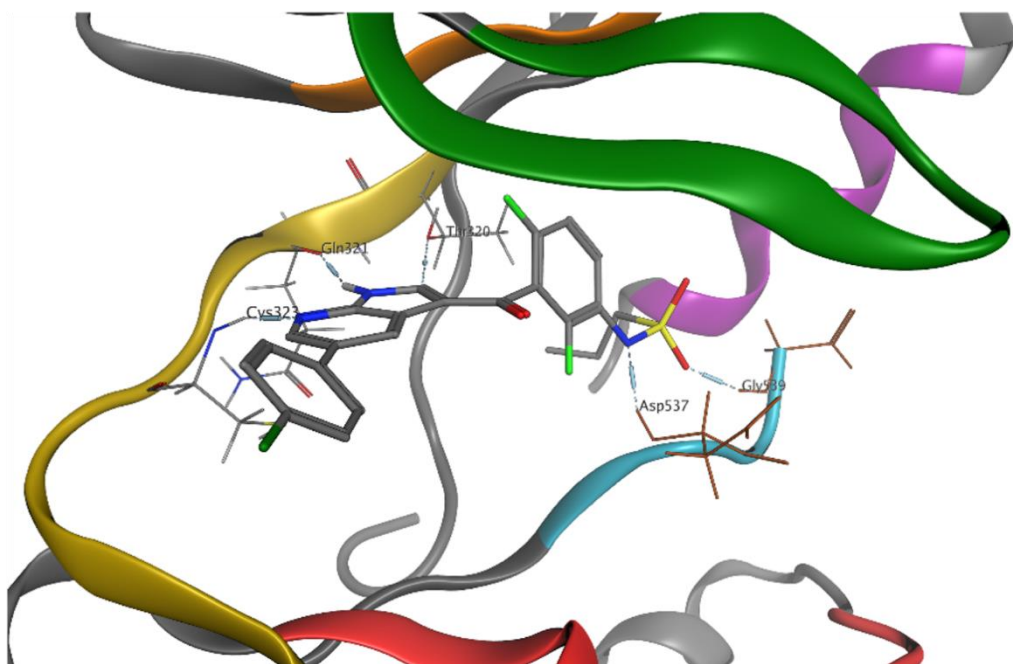
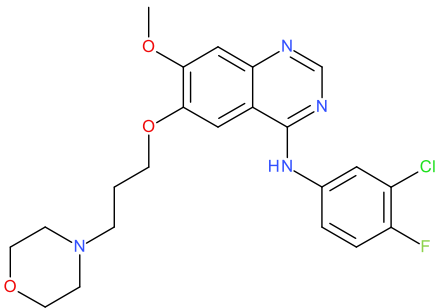
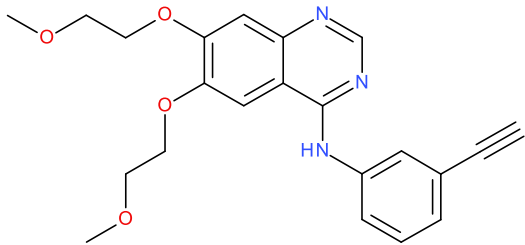
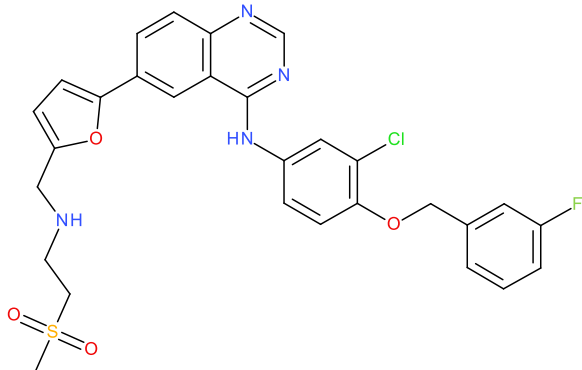
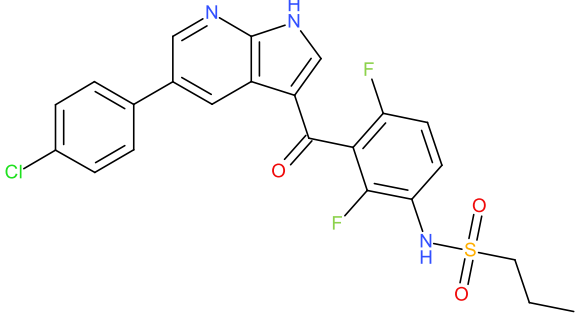
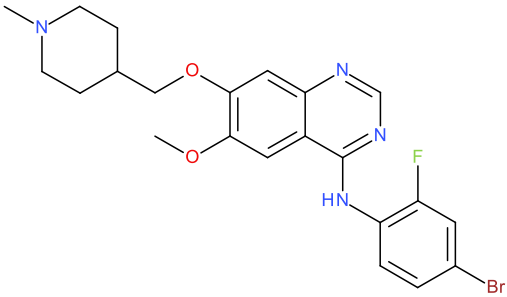
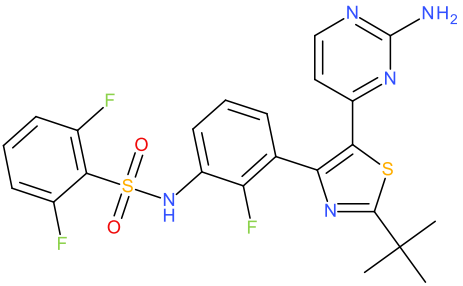
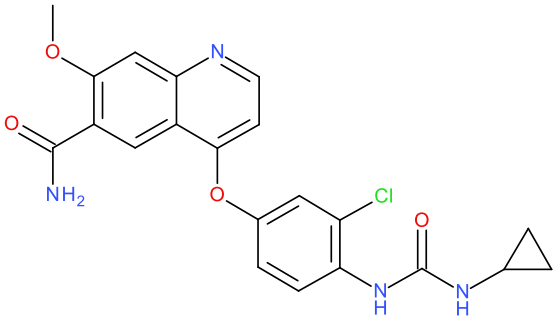


Figure 20. Représentation du modèle d'interaction d'un inhibiteur de type I½, la protéine kinase est en conformation inactive DFG-in/ αC -out. A) Modèle pharmacophorique, Hyd : région hydrophobe, A/D : accepteur/donneur de liaison hydrogène. Les liaisons hydrogène sont représentées en pointillés. B) Exemple de l'inhibiteur vemurafenib interagissant avec la protéine kinase BRAF (code PDB : 30G7).

Table 4. Liste des inhibiteurs de protéines kinases de type I½ approuvés par la FDA.

Structure chimique	Nom générique Entreprise Année d'approbation	Cibles principales	Indications	Codes PDB
	Gefitinib Astrazeneca 2003-2005, 2015	EGFR	mNSCLC	4WKQ, 4I22, 3UG2, 2ITO, 2ITY, 2ITZ
	Erlotinib OSI 2004	EGFR	mNSCLC	1M17, 4HJO
	Lapatinib GSK 2007	EGFR, ErbB2	Cancer du sein métastatique	1XKK, 3BBT

	Vemurafenib Plexxikon/Genentech 2011	A/B/C-Raf, B-Raf (V600E), SRMS, ACK1, MAP4K5, FGR	Mélanome	3OG7
	Vandetanib IPR Pharms 2011	EGFRs, VEGFRs, RET, Brk, Tie2, EphRs, Src	Cancer médullaire de la thyroïde (MTC)	2IVU
	Dabrafenib GSK 2013	BRAF	Mélanome	4XV2, 5CSW, 5HIE
	Lenvatinib Easai 2015	VEGFRs, FGFRs, PDGFR, Kit, RET	Cancer différencié de la thyroïde	3WZD

2.4.4 Les Inhibiteurs de type III

Découverts au milieu des années 2000 grâce à des campagnes de criblage à haute débit, les inhibiteurs de type III constituent une classe d'inhibiteurs non covalents, non compétitifs de l'ATP, qui se fixent dans une poche adjacente au site de liaison de l'ATP.⁴⁸⁻⁴⁹ Cela leur procure un avantage sur les autres types d'inhibiteurs présentés jusqu'ici (types I, II et I½) car l'accès au site actif peut être difficile pour l'inhibiteur lorsque l'ATP est en concentration intracellulaire importante (jusqu'à 10mM).⁵⁰

Puisqu'ils ciblent un site peu conservé, les inhibiteurs de type III sont capables de former des interactions plus spécifiques avec leur cible et bénéficient donc d'une sélectivité accrue par rapport aux inhibiteurs de type I, I½ et II déjà décrits. La possibilité de développer des inhibiteurs plus sélectifs, donc moins susceptibles de provoquer des effets secondaires chez les patients, est prometteuse. Cependant, les tests de criblage actuels ne sont pas optimisés pour identifier ce type d'interaction car il n'y a pas de moyen direct de détecter les inhibiteurs de type III.⁵¹ Ainsi, les touches d'un criblage primaire doivent être systématiquement confirmées par un criblage secondaire et dans le cas où l'inhibition est validée, le mécanisme sous-jacent doit être déterminé par résolution de la structure tridimensionnelle.

A ce jour, plusieurs inhibiteurs de type III ont été recensés dont deux ont obtenu une autorisation de mise sur le marché par la FDA en 2013 (**Table 5**).⁵² Aujourd'hui encore, aucune structure tridimensionnelle de trametinib en présence de sa cible n'est disponible dans la PDB. Cette molécule cible la protéine kinase MAP2K1 (MEK1) tout en autorisant la fixation de l'ATP. Cependant, la structure co-cristallisée d'un analogue Tak-733 (code PDB : 3PP1) avec la même protéine kinase nous permet de spéculer sur un mode d'interaction similaire (**Figure 21**). La rotation du résidu Phe du motif DFG libérerait la poche allostérique adjacente à la poche de fixation de l'ATP, autorisant la fixation du ligand. Celui-ci permettrait alors de bloquer la protéine kinase dans cette conformation, empêchant la reconnaissance du substrat protéique tout en autorisant la fixation de l'ATP. Ce mécanisme reposerait principalement sur la formation d'un réseau de liaisons hydrogène entre la lysine catalytique, l'inhibiteur et l'ATP.⁵³ Dans la structure en présence de l'analogue (code PDB : 3PP1), le résidu Lys entretient respectivement une et deux liaisons hydrogène avec l'ATP et Tak-733, tandis que ces deux derniers forment eux-mêmes une liaison hydrogène entre eux. Une autre liaison hydrogène avec un résidu Ser en position C-terminal de la boucle d'activation permet de stabiliser la protéine kinase dans une conformation inactive, rapprochant la boucle d'activation du lobe N-terminal.

⁴⁸ M. A. Bogoyevitch *et al.*, A New Paradigm for Protein Kinase Inhibition: Blocking Phosphorylation without Directly Targeting ATP Binding, *Drug Discovery Today*, 12.15–16 (2007), 622–33

⁴⁹ L. Garuti *et al.*, Non-ATP Competitive Protein Kinase Inhibitors, *Current Medicinal Chemistry*, 17.25 (2010), 2804–21

⁵⁰ L. O. Kirkland *et al.*, Non-ATP Competitive Protein Kinase Inhibitors as Anti-Tumor Therapeutics, *Biochemical Pharmacology*, 77.10 (2009), 1561–71

⁵¹ L. Krim Gavrin *et al.*, Approaches to Discover Non-ATP Site Kinase Inhibitors, *MedChemComm*, 4.1 (2012), 41–51

⁵² C. J. M. Wright *et al.*, Trametinib: First Global Approval, *Drugs*, 73.11 (2013), 1245–54

⁵³ Q. Dong *et al.*, Discovery of TAK-733, a Potent and Selective MEK Allosteric Site Inhibitor for the Treatment of Cancer, *Bioorganic & Medicinal Chemistry Letters*, 21.5 (2011), 1315–19

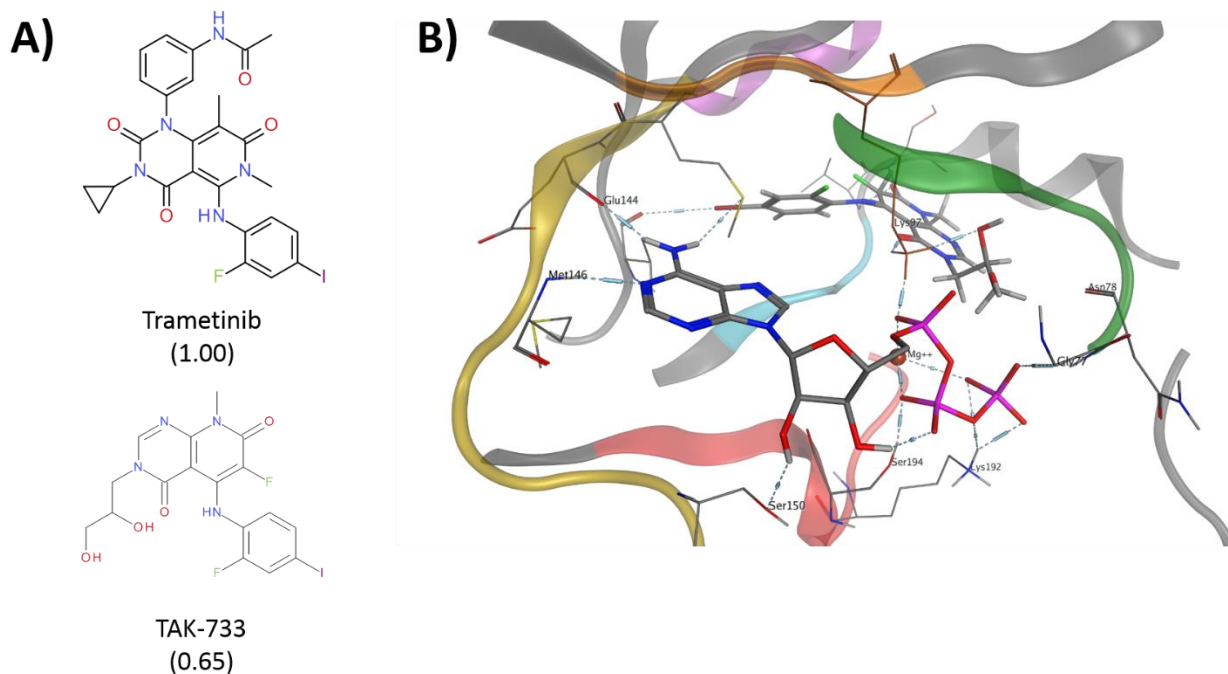
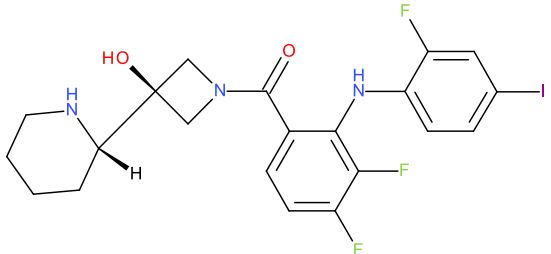


Figure 21. Mode de liaison de l'analogue Tak-733 en présence d'ATP dans MAP2K1. Code PDB : 3PP1. A) Structures de l'inhibiteur de type III trametinib et de son analogue TAK-733. Les valeurs entre parenthèses correspondent à la similarité de Tanimoto par rapport à trametinib. B) Représentation sous forme de ruban de la fixation de l'analogue Tak-733 en présence d'ATP dans MAP2K1 (code PDB : 3PP1).

Table 5. Liste des inhibiteurs de protéines kinases de type III approuvés par la FDA

Structure chimique	Nom générique Entreprise Année d'approbation	Cibles principales	Indications	Codes PDB
	Trametinib GSK 2013	MAP2K1/2, BRAF	Mélanome	-

	Cobimetinib Genentech Inc. 2015	MEK1	Mélanome (mutation BRAF V600E)	4AN2, 4LMN
---	---	------	---	------------

2.4.5 Les Inhibiteurs de type IV

Les inhibiteurs de type IV sont une autre classe d'inhibiteurs allostériques. Ils se différencient des inhibiteurs de type III par une fixation 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.⁵⁴ La fixation d'un inhibiteur de type IV dans la poche allostérique entraîne un changement conformationnel du domaine kinase, conduisant à son inhibition. Puisque chaque poche est potentiellement très différente en séquence, le chémotype des inhibiteurs de type IV est très divers (**Figure 22**).

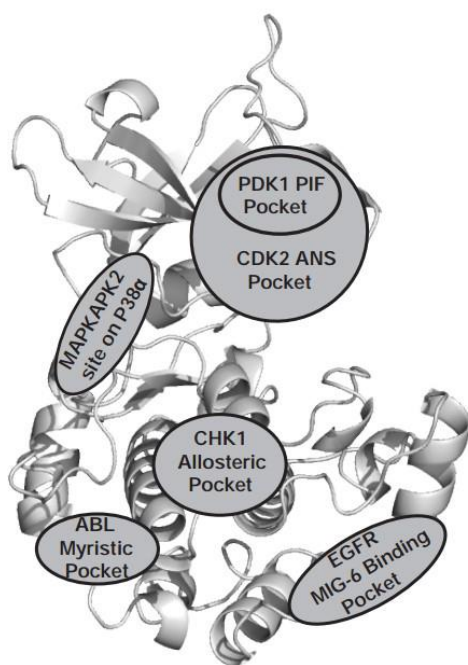
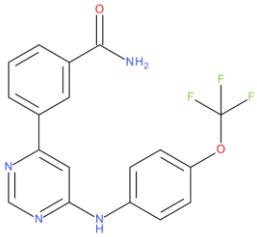
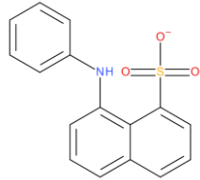


Figure 22. Projection des sites allostériques de différentes protéines kinases sur le même domaine kinase.⁵⁴

⁵⁴ V. Lamba *et al.*, New Directions in Targeting Protein Kinases: Focusing upon True Allosteric and Bivalent Inhibitors, *Current Pharmaceutical Design*, 18.20 (2012), 2936–45

A notre connaissance, aucun inhibiteur de type IV n'a encore bénéficié d'une autorisation de mise sur le marché approuvée par la FDA (**Table 6**).

Table 6. Exemples d'inhibiteurs de protéines kinases de type IV

Structure chimique	Nom de la molécule	Cible	Codes PDB
	GNF-2	ABL	3K5V
	ANS	CDK2	3PXZ

2.4.6 Les Inhibiteurs de type V

Les inhibiteurs de type V constituent la plus récente addition à la famille des inhibiteurs de protéines kinases.⁵⁵ Ils allient les caractéristiques des inhibiteurs de type I à IV et 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.

Une étude récente⁵⁵ portant sur la Tyr protéine kinase c-Src a permis d'identifier des inhibiteurs bivalents pouvant se lier au site actif de son domaine kinase et à son domaine régulateur SH2 simultanément. La distance entre ces deux domaines varie selon si la protéine kinase adopte une conformation ouverte (de 27 à 67 Å) ou fermée (37 Å, **Figure 23**). Le groupe de B. Soellner a utilisé cette information pour concevoir un inhibiteur pouvant cibler spécifiquement la conformation ouverte de c-Src.

La structure des inhibiteurs bivalents identifiés peut se diviser en trois constituants. Le premier, un inhibiteur de protéine kinase compétitif, est employé pour se lier sur le site actif. Le deuxième est un peptide qui peut se fixer sur le domaine SH2. Enfin, ces deux parties sont reliées entre elles par un polymère de polyéthylène glycol (PEG) et le tout est assemblé directement *in situ* par « chimie click ».⁵⁶ D'une part les résultats montrent que le composé est bien sélectif de la forme ouverte de c-Src par rapport

⁵⁵ T. K. Johnson *et al.*, Bivalent Inhibitors of C-Src Tyrosine Kinase That Bind a Regulatory Domain, *Bioconjugate Chemistry*, 27.7 (2016), 1745–49 <<https://doi.org/10.1021/acs.bioconjchem.6b00243>>.

⁵⁶ F. E. Kwarcinski *et al.*, Discovery of Bivalent Kinase Inhibitors via Enzyme-Templated Fragment Elaboration, *ACS Medicinal Chemistry Letters*, 6.8 (2015), 898–901

à la forme fermée. D'autre part, ils montrent que l'inhibiteur est également sélectif de c-Src par rapport à d'autres protéines kinases homologues.

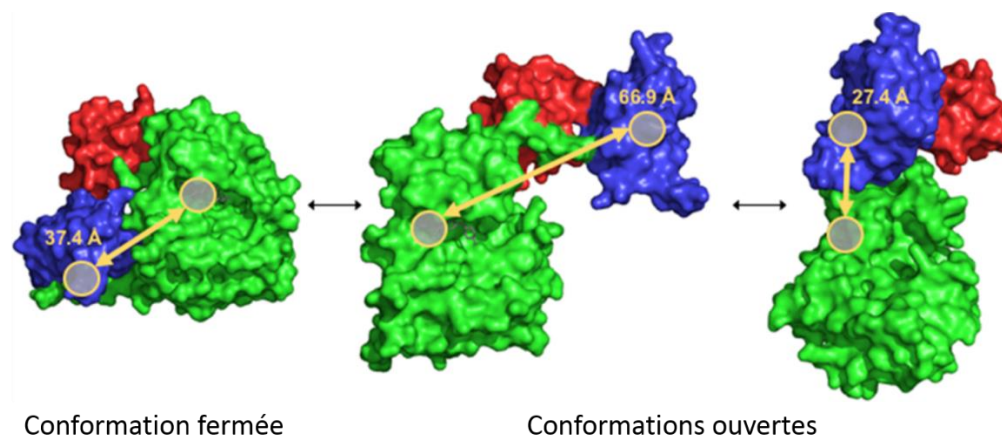


Figure 23. Différence de distances mesurées sur c-Src entre le domaine kinase et le domaine SH2. La protéine comporte trois domaines : kinase (vert), SH2 (vert), SH3 (rouge). Les codes PDB sont : 2SRC (gauche), 1Y57 (milieu) et 1OPL (droite).⁵⁵

Il est intéressant de noter que cette sélectivité proviendrait non pas de la variation de motifs au sein du site actif mais de la flexibilité de la protéine kinase, permettant de bien positionner les substituants pour une inhibition efficace. Ainsi, l'inhibiteur bivalent illustré **Figure 24** serait même capable d'inhiber la forme mutée T338M (*gatekeeper*) de c-Src ($IC_{50} = 19 \mu M$). D'autres études récentes portent notamment sur le développement d'inhibiteurs bivalents de BCR-Abl⁵⁷ et ERK.⁵⁸

Pour le moment aucun inhibiteur de type V n'a été approuvé par la FDA.

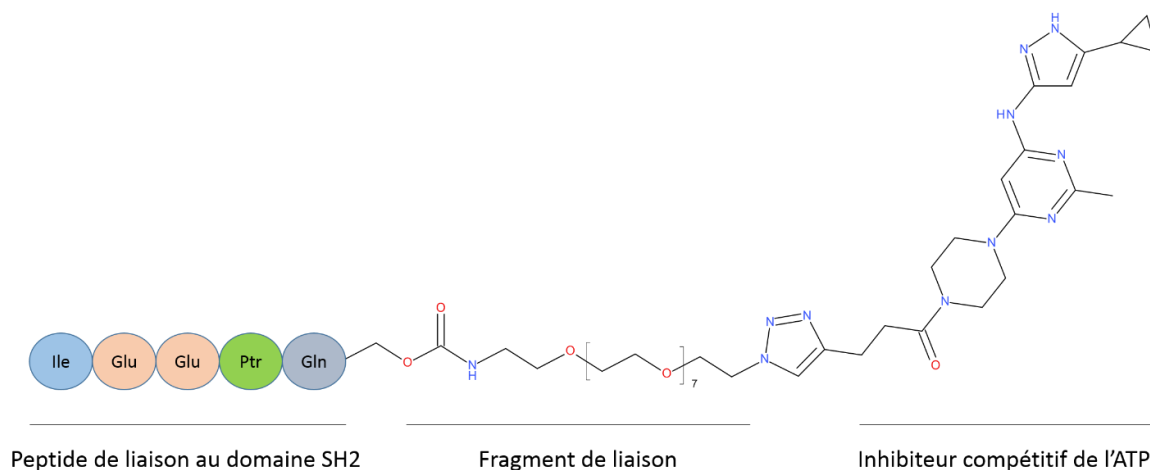


Figure 24. Structure d'un inhibiteur de type V (bivalent).

⁵⁷ C. M. Gower *et al.*, Conversion of a Single Polypharmacological Agent into Selective Bivalent Inhibitors of Intracellular Kinase Activity, *ACS Chemical Biology*, 11.1 (2016), 121–31

⁵⁸ B. C. Lechtenberg *et al.*, Structure-Guided Strategy for the Development of Potent Bivalent ERK Inhibitors, *ACS Medicinal Chemistry Letters*, 8.7 (2017), 726–31

2.4.7 Les Inhibiteurs covalents

Les inhibiteurs covalents, parfois également appelés type VI⁴¹, utilisent un mécanisme différent de tous ceux mentionnés auparavant. Ils portent en effet un groupement capable de former une liaison covalente avec un acide aminé nucléophile au sein du site actif (Cys, Lys ou Tyr), d'où leur appellation d'inhibiteurs covalents.⁵⁹

Le premier inhibiteur covalent approuvé par la FDA, en 2013, est afatinib. Tout comme les inhibiteurs de type I½ erlotinib et gefitinib, il cible la protéine kinase EGFR. Le résidu *gatekeeper* de cette dernière est capable de subir une mutation T790M qui rend la protéine mutée résistante au traitement chez certains patients.⁶⁰ Elle n'empêche pas la fixation de ces inhibiteurs mais couplée avec une autre mutation L858R, elle augmente sensiblement l'affinité de EGFR pour l'ATP par rapport à sa forme sauvage ($K_{M,ATP}$ respectivement de 8,4μM et 5,2μM). Contrairement aux inhibiteurs non covalents, afatinib bloque définitivement l'activité de la protéine kinase une fois fixé dans le site actif dans sa forme sauvage ou mutée (codes PDB : 4G5J, 4G5P).

La même année un autre inhibiteur covalent a été approuvé par la FDA. Il s'agit d'ibrutinib, qui cible la protéine kinase BTK. Le groupement N-acryloylpiperidine se comporte comme un accepteur de Michaël dans une réaction avec un résidu Cys (C418) du site actif (code PDB : 5P9J).

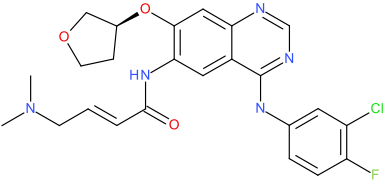
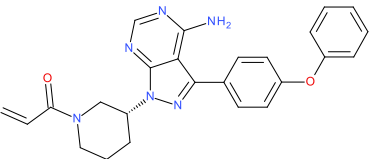
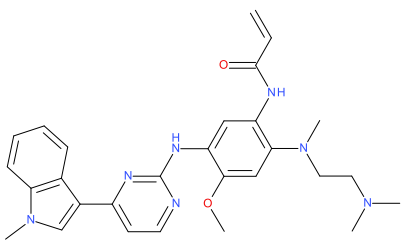
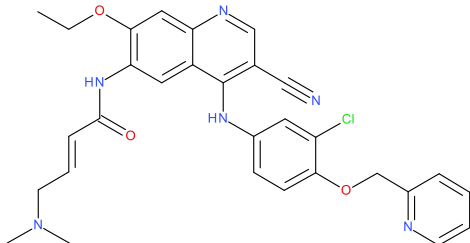
L'utilisation d'inhibiteurs covalents semble être un bon moyen d'augmenter le temps de résidence de l'inhibiteur dans sa cible et ainsi d'inhiber efficacement son activité phosphorylante. Cependant, le résidu formant la liaison covalente avec l'inhibiteur est lui aussi susceptible de muter, ce qui rend l'inhibiteur inefficace. Ainsi, les patients traités à l'ibrutinib développent parfois une mutation C418S qui provoque une résistance au traitement.⁶¹ Depuis, trois autres inhibiteurs de type covalent ont été approuvés par la FDA, donnant un total de 5 inhibiteurs. (**Table 7**).

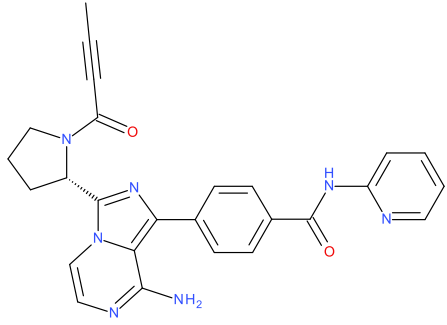
⁵⁹ Q. Liu *et al.*, Developing Irreversible Inhibitors of the Protein Kinase Cysteine, *Chemistry & Biology*, 20.2 (2013), 146–59

⁶⁰ C.-H. Yun *et al.*, The T790M Mutation in EGFR Kinase Causes Drug Resistance by Increasing the Affinity for ATP, *Proceedings of the National Academy of Sciences of the United States of America*, 105.6 (2008), 2070–75

⁶¹ J. A. Woyach *et al.*, Resistance Mechanisms for the Brutons Tyrosine Kinase Inhibitor Ibrutinib, *The New England Journal of Medicine*, 370.24 (2014), 2286–94

Table 7. Liste des inhibiteurs de protéines kinases covalents approuvés par la FDA

Structure chimique	Nom générique Entreprise Année d'approbation	Cibles principales	Indications	Codes PDB
	Afatinib Boehringer Ingelheim 2013	EGFR, ErbB2, ErbB4	mNSCLC	4G5J, 4G5P
	Ibrutinib Pharmacyclics / JNJ 2013	BTK	Lymphome à cellules du manteau	5P9J
	Osimertinib AstraZeneca 2015	EGFR	Cancer du poumon	4ZAU
	Neratinib Puma Biotechnology, Inc. 2017	HER2, EGFR	Cancer du sein	3W2Q, 2JIV

	Acalabrutinib AstraZeneca 2017	BTK	Lymphome à cellules du manteau	-
---	--------------------------------------	-----	--------------------------------------	---

2.4.8 Bilan sur les inhibiteurs de protéines kinases

Le marché des inhibiteurs de protéines kinases se répartit actuellement majoritairement entre les types I, I½ et II qui sont des molécules compétitrices de l'ATP. Cependant, les patients traités avec ces molécules sont susceptibles de développer des résistances dues à des mutations des protéines kinases ciblées. Ceci peut en effet résulter en une augmentation de l'affinité de la protéine mutée pour l'ATP et donc empêcher l'inhibiteur d'accéder au site actif.

Pour contourner ce problème, des inhibiteurs covalents ont été développés. Ils ne nécessitent d'accéder au site de liaison de l'ATP qu'une seule fois puisqu'ils forment une liaison covalente et bloquent ainsi définitivement l'accès du site actif à l'ATP. Cependant, des mutations peuvent apparaître chez les patients traités par ces molécules thérapeutiques et rendre les inhibiteurs inefficaces. C'est notamment le cas quand une mutation de la séquence protéique entraîne un encombrement stérique dans le site actif et empêche complètement la fixation du ligand.

C'est par exemple le cas de figure rencontré par les inhibiteurs de type II imatinib et nilotinib qui ciblent la protéine kinase Abl1. La mutation du résidu *gatekeeper* Thr de cette dernière, en un acide aminé Ile (T315I), bloque l'accès à la poche hydrophobe allostérique et prévient la formation d'une liaison hydrogène. L'inhibiteur de type II ponatinib a alors été développé spécifiquement pour pouvoir inhiber la forme mutée de la protéine Abl1 en remplaçant le groupement aminopyrimidine de l'imatinib par un groupement alcyne, plus petit.⁶²

Un aspect, autre que l'efficacité, à prendre très tôt en compte lors du développement de nouveaux inhibiteurs est la sélectivité. Les inhibiteurs de type II sont conçus pour se lier avec la poche hydrophobe allostérique, peu conservée parmi les différentes familles de protéines kinases par rapport aux autres motifs structuraux déjà présentés.⁴² Ils sont ainsi censés établir des interactions beaucoup plus spécifiques que ce que peuvent proposer les inhibiteurs de type I et I½.

⁶² T. Zhou *et. al.*, Structural Mechanism of the Pan-BCR-ABL Inhibitor Ponatinib (AP24534): Lessons for Overcoming Kinase Inhibitor Resistance, *Chemical Biology & Drug Design*, 77.1 (2011), 1–11

Une autre approche a donc été explorée avec le développement d'inhibiteurs de type III et IV. Ceux-ci présentent les avantages significatifs de ne pas être en compétition avec l'ATP et de cibler des poches allostériques peu conservées à l'échelle du kinome. Cependant, la première génération d'inhibiteurs de type IV présente l'inconvénient de cibler des sites allostériques plutôt plats, ce qui ne donne pas beaucoup de possibilités d'interaction et se traduit par une affinité faible.

Les inhibiteurs de type V, dits bivalents, sont capable d'inhiber deux sites différents en même temps, ce qui semble leur conférer une meilleure sélectivité. Il s'agit d'une approche encore récente mais qui semble fournir des résultats prometteurs. Cependant, les modes d'interaction peuvent se révéler complexes à élucider car cette approche requiert d'utiliser la séquence entière de la protéine kinase étudiée, ce qui peut poser des difficultés pour obtenir des structures tridimensionnelles.

Enfin, une autre tendance émergente sont les inhibiteurs de type VI, également appelés inhibiteurs covalents, qui peuvent se fixer sur leur récepteur de manière à les inhiber durablement.

La classification des inhibiteurs de protéines kinases peut parfois se révéler une tâche ardue, puisqu'elle nécessite la résolution d'une structure tridimensionnelle du complexe protéine-ligand. On se souvient par exemple du cas de l'imatinib dont le vrai mode de fixation dans le site actif a surpris la communauté scientifique au début des années 2000, résultant en la découverte des inhibiteurs de type II. Cependant, une structure de cette même molécule dans le site actif de la protéine kinase SYK montre un mode d'inhibition de type I, du fait de la présence d'un résidu Met comme *gatekeeper*, empêchant l'accès à la poche hydrophobe adjacente au site de fixation de l'ATP à l'inhibiteur⁶³ (code PDB : 1XBB). Un autre exemple est le cas de l'inhibiteur de type II nintedanib, co-cristallisé dans la protéine KDR (code PDB : 3C7Q). Nintedanib présente en effet une forte similarité avec imatinib, avec un groupement piperazine qui pourrait se lier dans la poche hydrophobe allostérique, tandis que le groupement amide pourrait agir en tant que connecteur. Les groupements oxindole et ester pourraient quant à eux interagir avec la région charnière de la protéine kinase. Cependant, la structure cristallographique montre bien une inhibition de type I, malgré une conformation en *DFG-out* de la cible. Les groupements piperazine et amide de l'inhibiteur sont exposés au solvant tandis que le groupement ester effectue une liaison hydrogène avec la lysine catalytique (**Figure 25**). Ainsi, il ne fait aucun doute que les travaux futurs sur les inhibiteurs de protéines kinases réservent encore bien des surprises.

⁶³ N. Bosc *et al.*, A Proteomic Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues, *ACS Chemical Biology*, 10.12 (2015), 2827–40

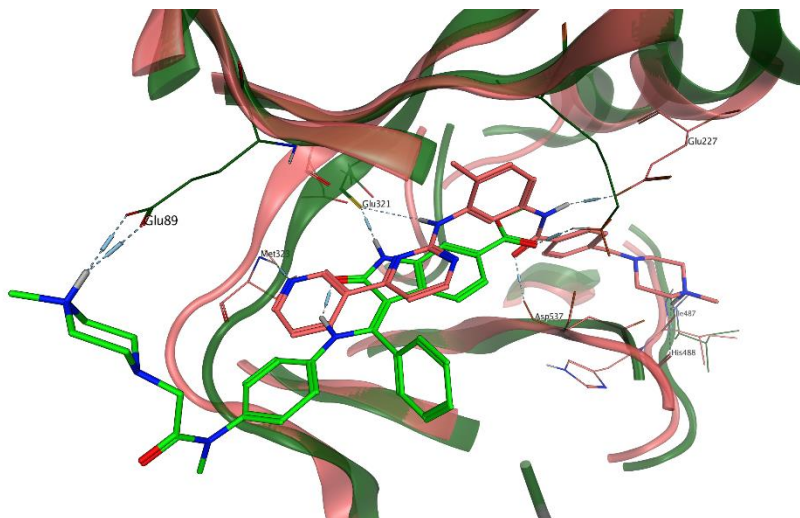


Figure 25. Superposition des complexes d'imatinib/Abl1 et de nintedanib /KDR. Codes PDB des complexes : imatinib/Abl1 (2HYY), nintedanib/KDR (3C7Q). En vert foncé : KDR, en vert clair : nintedanib, en rose foncé : Abl1, en rose clair : imatinib.

En conclusion de cette partie, la recherche et le développement de nouveaux inhibiteurs de protéines kinases a grandement progressé ces 20 dernières années. Pendant ce laps de temps, la famille des protéines kinases s'est imposée comme un axe de recherche thérapeutique majeur, notamment en oncologie. Ceci se traduit actuellement par 36 inhibiteurs de protéines kinases (et 2 inhibiteurs de lipides kinases, 2014 : idealisib, 2017 : copanlisib) bénéficiant d'une autorisation de mise sur le marché par la FDA, nombre en augmentation constante (**Figure 26**). Ces inhibiteurs ne ciblent cependant qu'environ 10% du kinome humain, laissant de côté un grand nombre de protéines potentiellement impliquées dans diverses maladies.⁶⁴ Qu'elle se base sur des approches phénotypiques ou bien qu'elle soit guidée par une cible, la recherche de nouveaux inhibiteurs de protéines kinases est grandement aidée par les outils *in silico*. Que ce soit pour l'identification de molécules potentiellement bioactives extraites de chimiothèques, par des approches de criblage virtuel telles que l'amarrage moléculaire ou encore pour la conception d'inhibiteurs *de novo*, il est nécessaire d'avoir à disposition des structures tridimensionnelles des protéines kinases d'intérêt (cristallographie aux rayons X, spectroscopie de RMN, cryo-microscopie électronique) afin d'identifier et d'optimiser les modes d'interaction des molécules développées.

⁶⁴ P. Cohen, Protein Kinases--the Major Drug Targets of the Twenty-First Century?, *Nature Reviews. Drug Discovery*, 1.4 (2002), 309–15

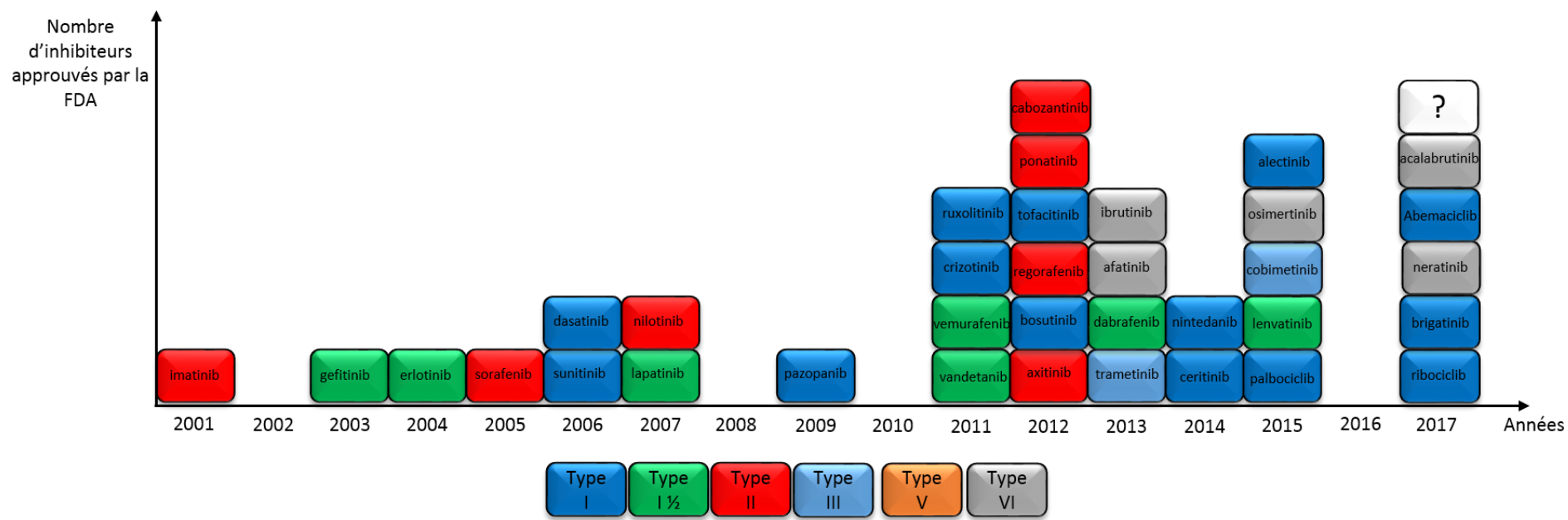


Figure 26. Inhibiteurs de protéines kinases approuvés par la FDA, classés par type d'inhibition. Le type VI correspond ici à aux inhibiteurs covalents.

3. Développement de VSPrep, un outil pour la préparation de petites molécules pour du criblage virtuel

3.1 Introduction

Dans le domaine de la chimoinformatique, contrairement à celui de la bioinformatique, le développement d'outils performants de criblage virtuel est longtemps demeuré l'apanage des sociétés privées. Quelques compagnies, dont les principales sont citées dans la **Table 8**, proposent des suites logicielles très performantes qui sont quotidiennement employées dans les travaux de recherche afin de produire des résultats le plus rapidement possible. Ce concept est particulièrement adapté aux grands groupes pharmaceutiques, pour lesquels la possibilité de trouver une nouvelle molécule thérapeutique brevetable justifie l'investissement de moyens financiers importants. Cependant, même si ces sociétés privées prennent en compte les limitations budgétaires des laboratoires académiques en leur proposant des tarifs réduits, le coût des licences annuelles peut vite se révéler pesant.

De ce fait les laboratoires académiques ne peuvent utiliser qu'une partie de ces suites logicielles. Or, elles ont chacune leurs forces et faiblesses, et couvrent un large champ de fonctionnalités allant de la visualisation (Discovery Studio⁶⁵, MOE⁶⁶, etc.), à la préparation des molécules (Sybyl⁶⁷, MOE⁶⁸, Ligprep⁶⁹, Moka⁷⁰, etc.) comme au criblage virtuel (GLIDE⁷¹, GOLD⁷², etc.). Ces suites logicielles sont donc parfois complémentaires. Malheureusement il est fréquent de ne pas pouvoir reproduire des travaux publiés car ils ont été obtenus à l'aide d'un logiciel propriétaire dont la licence n'est pas accessible⁷³. En partie pour répondre à ce problème, de nombreux outils académiques ont vu le jour. Ces derniers présentent cependant l'inconvénient d'avoir été développés de manière indépendante pour des tâches spécifiques, sans cohérence entre eux, ce qui complique l'élaboration de protocoles complets. Ce travail a pour but de proposer une alternative gratuite au processus de préparation des petites molécules pour du criblage

⁶⁵ BIOVIA Discovery Studio | Predictive Modeling & Science Simulation Software App, <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/>, consulté le 28 Août 2017.

⁶⁶ Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2017., https://www.chemcomp.com/Research-Citing_MOE.htm, consulté le 31 Janvier 2017.

⁶⁷ SYBYL-X Suite, Certara, 2016, <https://www.certara.com/software/molecular-modeling-and-simulation/sybyl-x-suite>, consulté le 10 Novembre 2016

⁶⁸ SD File Processing with MOE Pipeline Tools, <https://www.chemcomp.com/journal/sdtools.htm>, consulté le 29 Novembre 2017.

⁶⁹ Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.

⁷⁰ MoKa - PKa Modelling, <http://www.moldiscovery.com/software/moka>, consulté le 29 Août 2017

⁷¹ Citations | Schrödinger | GLIDE, <https://www.schrodinger.com/citations>, consulté le 29 Août 2017

⁷² G. Jones *et al.*, Development and Validation of a Genetic Algorithm for Flexible Docking, *Journal of Molecular Biology*, 267.3 (1997), 727–48

⁷³ J. C. Cole *et al.*, Comparing Protein–ligand Docking Programs Is Difficult, *Proteins: Structure, Function, and Bioinformatics*, 60.3 (2005), 325–332

virtuel, en proposant un programme regroupant diverses fonctionnalités et accessibles à l'utilisateur dans une même plate-forme.

Table 8. Suites logicielles quotidiennement utilisées en chémoinformatique pour la préparation des petites molécules.

Compagnie actuelle	Suite logicielle	Date de fondation
Certara*	Sybyl ⁷⁴	1978
Molecular Discovery	MoKa ⁷⁵	1984
Computational Chemistry Group	MOE ⁷⁶	1997
OpenEye Scientific Software, Inc.	OpenEye ⁷⁷	1997
ChemAxon	JChem ⁷⁸	1998
BIOVIA**	Discovery Studio ⁷⁹	1999
AMBER	AmberTools ⁸⁰	2002
Schrödinger, LLC	Maestro ⁸¹ , LigPrep ⁸²	2005

*Sybyl a été originalement développé par TRIPOS, qui a fusionné en 2008 avec Pharsight pour devenir Certara.

**Pipeline Pilot a été initialement développé par SciTegic, société qui a été acquise par Accelrys en 2004, à son tour absorbée par Dassault Systèmes en 2014 sous le nom de BIOVIA.

De nombreux exemples, récents et plus anciens, ont montré l'importance de la préparation des molécules pour les expériences de criblage virtuel⁸³. En effet, comme nous l'avons vu au premier chapitre, les molécules sont décrites de manière statique en informatique. Il n'existe par exemple à notre connaissance aucun logiciel capable de faire varier la position des protons sur le ligand lors d'un amarrage moléculaire. Ainsi, le choix du tautomère à cribler est déterminant pour l'identification de liaisons hydrogène avec les résidus du site actif. Le même raisonnement peut être appliqué à l'état de protonation de la protéine et du ligand, même si la « bonne » protonation des molécules à l'intérieur du site actif de la

⁷⁴ SYBYL-X Suite.

⁷⁵ Molecular Discovery Ltd., <http://www.moldiscovery.com/>, consulté le 30 Août 2017.

⁷⁶ MOE: Molecular Operating Environment, https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm, consulté le 24 Novembre 2016

⁷⁷ OpenEye Scientific Software, Cheminformatics and Molecular Modeling Software (OpenEye), <http://www.eyesopen.com>, consulté le 24 Novembre 2016

⁷⁸ ChemAxon – Software for Chemistry and Biology, <https://www.chemaxon.com>, consulté le 9 Novembre 2016

⁷⁹ BIOVIA Discovery Studio | Predictive Modeling & Science Simulation Software App

⁸⁰ The Amber Molecular Dynamics Package, <http://ambermd.org/>, consulté le 30 Août 2017

⁸¹ Citations | Schrödinger | Maestro, <https://www.schrodinger.com/citations>, consulté le 29 Août 2017

⁸² Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.

⁸³ S. Forli, Charting a Path to Success in Virtual Screening, *Molecules (Basel, Switzerland)*, 20.10 (2015), 18732–58

cible est difficile à prédire. Une autre limitation couramment rencontrée est la description de la stéréochimie, souvent manquante, soit par erreur, soit du fait de la présence d'un mélange racémique. Dans les deux cas, il faut impérativement énumérer les stéréoisomères possibles afin de lever toute ambiguïté structurale sur les ligands d'intérêt. Une autre limitation du criblage virtuel vient de ce que la plupart des outils disponibles ne permettent pas l'échantillonnage conformationnel des cycles à la volée. Ainsi, il n'est pas possible pour un cycle à six atomes de passer dynamiquement d'une conformation chaise à une conformation bateau (**Figure 27**) car seules les liaisons non impliquées dans un cycle peuvent tourner.

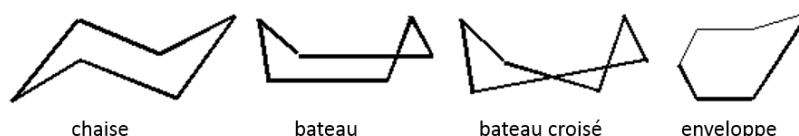


Figure 27. Différentes conformations du cyclohexane.

Il est donc nécessaire de générer en amont les conformations des cycles présents dans les molécules afin de ne pas manquer de touche potentielle. De nombreuses études montrent l'impact que la préparation des molécules a sur la qualité des résultats obtenus. Il est important de noter qu'elle peut également jouer un grand rôle dans la performance du protocole. Par exemple, tout chimiste computationnel est tôt ou tard confronté au problème de la présence de doublons dans les chimiothèques. Il est primordial de les identifier et de les filtrer pour ne pas accroître inutilement la charge de calculs et biaiser les résultats. Nous verrons également que suivant les projets, il peut être souhaitable de cibler un espace chimique donné, pour rester par exemple dans l'espace des molécules « *drug-like* ». La préparation des molécules est donc une tâche importante avec un nombre de facteurs importants à considérer pour obtenir une chimiothèque prête à l'emploi pour du criblage virtuel.

Toutes les étapes de préparation mentionnées, sans compter d'autres tâches récurrentes, telles que la standardisation des structures ou encore la suppression des sels, peuvent s'effectuer de diverses manières. Chaque outil a en effet sa propre implémentation, voire son propre algorithme, ce qui résulte en une multitude de protocoles très hétérogènes. Les logiciels utilisés peuvent être sous licence commerciale ou bien ne plus être disponibles (non maintien du projet, perte du code source, etc.), ce qui empêche la comparaison avec des résultats publiés ou tout simplement leur reproduction.

Ces raisons expliquent en partie pourquoi il n'y a, à l'heure de la rédaction de ce manuscrit, aucun protocole standard de préparation de petites molécules largement accepté au sein de la communauté scientifique. Ceci empêche une comparaison efficace des résultats de criblages virtuels.

3.1.1 Pourquoi un nouveau protocole ?

Afin de valider des protocoles expérimentaux proposés par des pairs, il est important que les chercheurs aient accès à l'ensemble des outils cités dans les publications. Cependant, les outils les plus couramment mentionnés pour la préparation des petites molécules pour du criblage virtuel sont commerciaux.

Pour cette raison, nous nous sommes basés dans la mesure du possible sur des outils gratuits pour les laboratoires académiques, pour réaliser un nouveau protocole de préparation de petites molécules,

VSPrep.⁸⁴ Il permettra notamment de pouvoir reproduire plus facilement les résultats publiés de criblages virtuels mais pourra également s'accompagner d'une réduction des coûts liés à l'achat de licences annuelles d'outils commerciaux.

Pour ces raisons, nous avons décidé d'établir notre propre protocole de préparation de petites molécules pour du criblage virtuel, VSPrep, essentiellement à partir d'outils dits « *open source* » et disposant d'une communauté scientifique réactive.

3.1.2 Le rôle de l'*open-source* en chémoinformatique

Le terme *open source*, ou « code source ouvert », réfère à n'importe quel programme dont le code source est disponible pour l'utilisation, l'étude, la copie, le partage ou la modification au gré des besoins des utilisateurs ou autres développeurs.⁸⁵

Ce concept est loin d'être nouveau. Dans les années 1960, lorsque personne n'était prêt à acheter un ordinateur à cause de son prix prohibitif, la commercialisation de logiciels était impensable.⁸⁶ Au contraire, les producteurs délivraient gratuitement leur code source qu'il était même possible de modifier afin de pouvoir vendre plus rapidement leur matériel. La situation a changé dès les années 1970, lorsque la législation de nombreux pays a classifié cette approche de concurrence déloyale envers de nouvelles entreprises qui se basaient sur la vente de logiciels pour croître.

Peu à peu popularisé auprès du public par le développement d'outils informatiques très visibles, dont l'exemple le plus célèbre est certainement le système d'exploitation Linux développé au début des années 1990, le modèle économique de l'*open source* a finalement su s'imposer avec le temps comme une alternative viable aux logiciels payants. La principale raison de son succès repose sur son modèle collaboratif d'utilisateurs et de développeurs qui s'entraident pour améliorer les outils, ce qui en théorie en réduit le temps de développement et en facilite la maintenance. Cependant, cette organisation ne bénéficie en général pas du même niveau de services fournis par des outils payants. En effet, il suffit que les principaux développeurs soient indisponibles pour que la correction d'éventuelles erreurs dans le code ou l'ajout de fonctionnalités désirées prennent plusieurs mois voire plusieurs années. Il peut alors se révéler plus fructueux de directement contribuer à l'outil afin de l'améliorer.

Cette idée de de partage et de collaboration au cœur de l'*open source* se marie tout naturellement avec la recherche scientifique. Cette nouvelle tendance, l'*open science*⁸⁷, va même plus loin en partageant également les données. C'est par exemple le cas de résultats de HTS provenant de grands groupes

⁸⁴ J.-M. Gally *et al.*, VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening, *Molecular Informatics*, 2017

⁸⁵ What Is Open Source?, *Opensource.Com*, <https://opensource.com/resources/what-open-source>, consulté le 11 Novembre 2017

⁸⁶ IBM Redbooks | Linux Handbook A Guide to IBM Linux Solutions and Resources, 2016, <http://www.redbooks.ibm.com/abstracts/sg247000.html?Open>, consulté le 11 Novembre 2017.

⁸⁷ D. L. Shaw, Is Open Science the Future of Drug Development?, *The Yale Journal of Biology and Medicine*, 90.1 (2017), 147–51.

pharmaceutiques.⁸⁸ Quelques-unes de ces données ont d'ailleurs été utilisées pour développer VSPrep décrit dans ce chapitre de manuscrit de thèse.

L'un des plus célèbres exemples du succès de l'*open science* est la plate-forme foldit⁸⁹. Cette dernière, sous les apparences d'un jeu de puzzle en ligne massivement multijoueur et accessible au public, permet de faire des prédictions de repliement tridimensionnel de protéines en s'aidant de la contribution des joueurs. Il est intéressant de noter que cette approche a permis de proposer une structure tridimensionnelle de la protéase de HIV en seulement 3 semaines, là où des équipes de recherches avaient échoué pendant 10 ans.⁹⁰

Au cours de mes travaux de thèse, et plus particulièrement lors du développement de l'outil VSPrep, j'ai moi aussi eu l'occasion de me familiariser avec plusieurs outils *open source*, dont les principaux sont décrits dans la partie 3.2.2.

3.2 Conception de l'outil

Dans cette partie, nous aborderons la conception de VSPrep, un nouveau protocole de préparation de petites molécules pour du criblage virtuel. Nous définirons tout d'abord les attentes du projet, puis les outils sur lesquels il repose. Enfin, après un rapide descriptif nous verrons son implémentation, étape par étape, au sein d'un outil de *pipelining*, ce qui nous permettra de mieux exposer les raisons derrière certains choix de conception.

3.2.1 Cahier des charges

Afin de convaincre le plus grand nombre de chimoinformaticiens, nous avons établi en quelques points les critères auxquels devra répondre le nouvel outil développé.

Tout d'abord, comme exposé précédemment, il ne devra comporter que des outils gratuits pour les académiques, voire si possible libres de droit afin que les laboratoires privés puissent également l'employer. Cela permettra notamment de grandement améliorer les échanges entre partenaires académiques et industriels.

Afin de simplifier à la fois le développement, l'utilisation et la maintenance, un outil principal sera utilisé comme échafaudage du protocole. Les outils de *pipelining* graphiques tels que Pipeline Pilot⁹¹ ou

⁸⁸ Y. Gao *et al.*, A Broad Activity Screen in Support of a Chemogenomic Map for Kinase Signalling Research and Drug Discovery, *Biochemical Journal*, 451.2 (2013), 313–28

⁸⁹ S. Cooper *et al.*, Predicting Protein Structures with a Multiplayer Online Game, *Nature*, 466.7307 (2010), 756–60

⁹⁰ M. J. Coren Company Fast, Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks, *Scientific American*, <https://www.scientificamerican.com/article/foldit-gamers-solve-riddle>, consulté le 11 Novembre 2017

⁹¹ BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis, <http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot>, consulté le 9 Novembre 2016

encore KNIME⁹² sont actuellement très populaires dans la communauté scientifique car tous les chercheurs qui réalisent des criblages virtuels n'ont pas forcément de formation informatique avancée.

Enfin, l'outil devra bien sûr répondre à tous les besoins des utilisateurs. Un grand nombre d'options seront implémentées afin de permettre à l'utilisateur le paramétrage le plus fin possible et prévoir le maximum de scénarii. De plus, des rapports sous forme de graphiques seront produits automatiquement afin de simplifier l'analyse des résultats de l'outil.

3.2.2 Technologie de VSPrep

La plupart des librairies disponibles librement étant développées depuis plusieurs années, elles disposent déjà d'un nombre important de fonctionnalités. Cependant, je n'ai pas trouvé de protocole complet dans un seul outil qui puisse effectuer l'ensemble des principales étapes de préparation des petites molécules pour du criblage virtuel. J'ai donc décidé d'y remédier. Afin de mener ce projet à bien, j'ai besoin de deux outils principaux.

Le premier est un outil de pipelining (tel que KNIME, Orange⁹³, Pipeline Pilot, etc.) disposant de fonctionnalités chémoinformatiques. Le second est une librairie de chémoinformatique qui me permettra de manipuler à ma guise des objets moléculaires même en dehors de KNIME.

J'ai donc d'abord identifié les librairies *open source* les plus couramment utilisées (**Table 9**), afin de faire mon choix.

Table 9. Principales librairies de chémoinformatique implémentées dans le logiciel KNIME

Librairie	Année de parution	Besoin d'une licence ?
CDK ⁹⁴	2003	non
Open Babel ⁹⁵	2005	non
Indigo ⁹⁶	2009	non

⁹² M. R. Berthold *et al.*, KNIME: The Konstanz Information Miner, in *Data Analysis, Machine Learning and Applications*, ed. by Christine Preisach *et al.*, Studies in Classification, Data Analysis, and Knowledge Organization (Springer Berlin Heidelberg, 2008), pp. 319–26

⁹³ Orange – Visual Programming, <https://orange.biolab.si/features/visual-programming/>, consulté le 15 Novembre 2017

⁹⁴ C. Steinbeck *et al.*, The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics, *Journal of Chemical Information and Computer Sciences*, 43.2 (2003), 493–500

⁹⁵ R. Guha *et al.*, The Blue Obelisk—Interoperability in Chemical Informatics, *Journal of Chemical Information and Modeling*, 46.3 (2006), 991–98

⁹⁶ D. Pavlov *et al.*, Indigo: Universal Cheminformatics API, *Journal of Cheminformatics*, 3.1 (2011), P4

RDKit ⁹⁷	2006	non
Erl Wood ⁹⁸	2011	non
Vernalis ⁹⁹	2014	non
EMBL-EBI	2013	non
Enalos	2012	oui
LHASA	2015	oui
OCHEM	2014	oui

3.2.2.1 KNIME

KNIME⁹² est un outil de fouille de données (*datamining*) codé en Java et permettant l'enchaînement d'instructions plus ou moins complexes afin de réaliser des protocoles automatisés (*pipelines*). Notre choix s'est porté sur ce logiciel car il propose des fonctionnalités avancées à la fois en *pipelining* et en chémoinformatique.

Il se présente sous la forme d'une interface graphique comprenant un espace de travail ainsi qu'une liste de fonctions prêtes à l'emploi (nœuds). Plusieurs nœuds peuvent être reliés entre eux pour réaliser un protocole (*workflow*).

De nombreux outils de chémoinformatiques sont implémentés dans KNIME, y compris RDKit. Les nœuds développés avec cette librairie sont nativement parallélisés et lui confère l'avantage de présenter de bonnes performances par rapport à ses concurrents.

Un exemple simple de *workflow* est illustré **Figure 28**. Il commence par la lecture d'un fichier de molécules au format SDF, qui sont ensuite converties au format RDKit. Des descripteurs moléculaires peuvent alors être calculés pour filtrer les molécules, par exemple, en fonction de leur poids moléculaire (PM). Enfin, les résultats sont conservés dans un autre fichier SDF.

⁹⁷ RDKit, <http://www.rdkit.org>, consulté le 9 Novembre 2016

⁹⁸ Erl Wood Cheminformatics Nodes for KNIME (Trusted Extension), <https://www.knime.com/community/erlwood>, consulté le 29 Novembre 2017

⁹⁹ Vernalis Nodes for KNIME (Trusted Extension), <https://www.knime.com/book/vernal-nodes-for-knime-trusted-extension>, consulté le 30 Août 2017

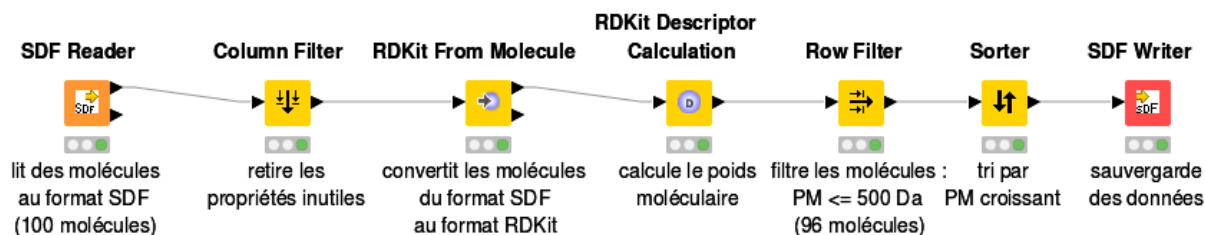


Figure 28. Exemple d'utilisation d'un workflow KNIME. Le nombre de molécules est indiqué entre parenthèses après chaque changement.

La réalisation de workflows plus complexes est également possible (boucles, conditions, parallélisation, graphes) et ceci sans écrire une ligne de code. Ce côté « facile à apprendre, dur à maîtriser » a séduit plus d'un membre de la communauté de chimoinformatique. Plusieurs langages de script (Python, Jython, Perl, Bash) ainsi que les formats moléculaires les plus couramment rencontrés tels que SDF, MOL2 ou encore SMILES sont supportés dans KNIME, ce qui permet aux utilisateurs plus avancés de développer des nœuds chimoinformatiques, soit gratuitement accessibles à tous (RDKit, Indigo, Marvin, CDK), soit nécessitant une licence (MOE, Schrödinger, Infocom¹⁰⁰). Il s'agit aujourd'hui de l'une des plateformes de *pipelining* les plus populaires avec une communauté d'utilisateurs très actifs et des mises à jour régulières.

3.2.2.2 RDKit

Le *Rational Discovery Kit*⁹⁷ (RDKit) est une collection d'outils de chimoinformatique et d'apprentissage automatique (*machine learning*) initialement développée au début des années 2000 par la société Rational Discovery. Son domaine d'application était alors la construction de modèles prédictifs de propriétés ADMET ainsi que d'activités biologiques.¹⁰¹ La fermeture de la société en 2006 a poussé ses créateurs à diffuser l'outil publiquement avec une licence *open source*. Depuis, le développement de l'outil se poursuit essentiellement au sein de la société Novartis et l'outil gagne ainsi, chaque année, en fonctionnalités. Parmi celles-ci, on peut par exemple citer le calcul de descripteurs 3D, d'empreintes moléculaires (*fingerprints*) ou encore la minimisation énergétique de conformations 3D dans un champ de force.

Le noyau du code est écrit en C++ et l'outil est encapsulé en Java, C# et Python, donnant de nombreuses possibilités de développement d'algorithmes pour pallier d'éventuelles fonctionnalités manquantes. De plus, cet outil permet l'administration de bases de données moléculaires grâce à une interface avec le langage PostgreSQL (Figure 29).

Enfin, la communauté du projet RDKit est extrêmement active et réactive, espoir de longévité et de maintenabilité des outils développés à partir de cette librairie.

¹⁰⁰ ChemAxon Node for KNIME JChem Extensions,, <http://infocom-science.jp>, consulté le 12 Novembre 2017

¹⁰¹ An Overview of the RDKit — The RDKit 2017.09.1 Documentation, <http://www.rdkit.org/docs/Overview.html>, consulté le 11 Novembre 2017

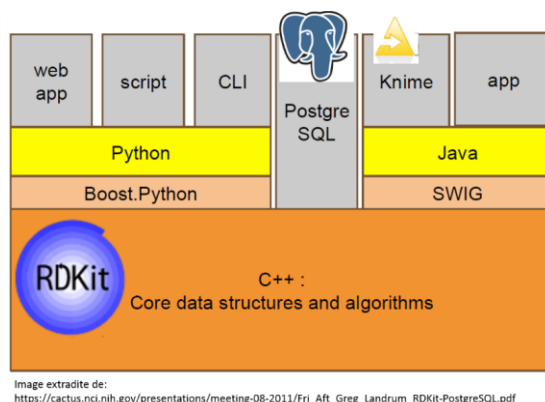


Figure 29. Encapsulation du code source de RDKit. Extraite de :
https://cactus.nci.nih.gov/presentations/meeting-08-2011/Fri_Aft_Greg_Landrum_RDKit-PostgreSQL.pdf

3.2.2.3 Logiciels gratuits alternatifs

Pour pallier les fonctionnalités actuellement manquantes de RDKit, nous nous sommes systématiquement tournés vers des logiciels gratuits alternatifs. Lorsqu'aucun équivalent satisfaisant n'a pu être identifié, nous avons incorporé des fonctionnalités de Marvin (logiciel commercial sous licence académique) au sein du *workflow* (**Table 10**).

Table 10. Liste des outils incorporés dans le protocole KNIME VSPrep

Outil	Type de licence	Tâche(s)	Implémentation
RDKit	<i>open source</i>	outil principal	nœuds KNIME et scripts Python ou Java
Programme cxcalc ¹⁰²	commerciale*	tautomères, stéréoisomères	appel au programme cxcalc depuis KNIME
Noeuds de Vernalis ⁹⁹	<i>open source</i>	stéréoisomères	nœud KNIME
Ambit-tautomers ¹⁰³	<i>open source</i>	tautomères	nœud KNIME
MolVS ¹⁰⁴	<i>open source</i>	standardisation	scripts Python

* gratuit pour les laboratoires académiques selon des conditions particulières¹⁰⁵

¹⁰² *Calculator Plugins Were Used for Structure Property Prediction and Calculation, Marvin 15.1.19, 2016, ChemAxon*

¹⁰³ N. T. Kochev *et al.*, *Ambit-Tautomer: An Open Source Tool for Tautomer Generation, Molecular Informatics*, 32.5–6 (2013), 481–504

¹⁰⁴ MolVS: Molecule Validation and Standardization — MolVS 0.0.9 Documentation, <http://molvs.readthedocs.io/en/latest>, consulté le 30 Août 2017

¹⁰⁵ Academic License Conditions « ChemAxon – Software for Chemistry and Biology, *ChemAxon - Software for Chemistry and Biology R&D*, <https://www.chemaxon.com/my-chemaxon/academic-license-conditions>, consulté le 12 Novembre 2017

Marvin est une collection d'outils permettant le dessin et la visualisation de structures chimiques. Elle permet également le calcul de prédictions basées sur la structure de la molécule grâce notamment au programme *cxcalc* (état de protonation majoritaire, énumération de tautomères, etc.) Elle est entièrement incluse dans la suite logicielle JChem, dont le but principal est l'administration de bases de données moléculaires relationnelles, incluant la standardisation de structures chimiques, ou encore la fragmentation de molécules. Ces deux outils sont écrits en Java.

Pour le moment 4 programmes de Marvin sont disponibles gratuitement sous forme de nœuds KNIME:

- *molconverter* : conversion de molécules en d'autres formats (i.e. SDF en SMILES) ou modification de structures (ajout d'atomes d'hydrogènes explicites)
- *marvinsketch* : dessin de structures chimiques
- *marvinview* : visualisation basique de structures chimiques (plutôt des petites molécules)
- *marvinspace* : visualisation avancée de structures chimiques (macromolécules, surfaces)

Il existe également d'autres nœuds KNIME, commerciaux, développés par la société Infocom, qui permettent d'accéder au reste des fonctionnalités offertes par JChem (standardisation des structures chimiques, etc.). Mon approche a été d'employer les nœuds KNIME quand ils étaient disponibles (notamment *molconverter* et *marvinview*) et d'exécuter le programme *cxcalc* de Marvin directement dans KNIME dans le cas contraire.

Cependant, cette implémentation suppose que l'utilisateur ait accès à une licence Marvin, ce qui n'est pas forcément le cas. Les deux principales étapes concernées par cette limitation sont l'énumération de tautomères et stéréoisomères. Des alternatives ont donc été implémentées afin que le programme puisse s'exécuter pour le plus grand nombre. Ces dernières sont décrites dans les paragraphes respectives 3.2.3.5 et 3.2.3.7.

3.2.3 Description de VSPrep

VSPrep est un protocole (*workflow*) implémenté dans la plateforme KNIME. Il contient plus de 1800 nœuds dont une vingtaine de scripts développés en Python, en R et en Java, selon les besoins. Le protocole proposé se découpe en huit étapes et peut s'exécuter soit en mode interactif dans l'interface KNIME, soit directement en ligne de commandes. Les formats moléculaires les plus courants (notamment SDF, Mol2 et SMILES) sont supportés et plus d'une centaine d'options ont été implémentées, ce qui permet à l'utilisateur de paramétrer le protocole de la manière la plus adaptée à ses besoins. Une illustration de l'outil est disponible dans l'**Annexe 1**.

Pour des raisons de confidentialité, seules les molécules identifiées sur le site Ambinter (www.ambinter.com) seront utilisées à titre d'exemples pour illustrer certains cas rencontrés dans les différentes figures de ce chapitre. Parmi les quelques 20 millions de structures que contient cette base de données moléculaire, on peut en effet trouver de nombreux cas particuliers pour lesquels les codes Ambinter sont indiqués.

3.2.3.1 Lecture des molécules virtuelles

La représentation informatique des molécules est une tâche extrêmement complexe. Plusieurs formats de fichiers moléculaires ont été développés et sont communément utilisés pour représenter les petites molécules. Tout d'abord les formats dits « 1D », tels que le SMILES¹⁰⁶, l'InChI¹⁰⁷ ou encore le nom IUPAC encodent uniquement la topologie des molécules sous la forme de représentation textuelle. A l'opposé, les tables de connectivité, notamment SDF¹⁰⁸, Mol et Mol2¹⁰⁹, renseignent également sur les coordonnées atomiques, ce qui permet une représentation de la molécule soit schématique (2D), soit dans l'espace (3D).

Chacun de ces formats a ses propres règles pour définir les propriétés des molécules, ce qui se traduit par des difficultés pour comparer les molécules avec les différents outils. Pour compliquer les choses, certains formats ont une flexibilité au niveau de leur écriture, ce qui autorise de multiples syntaxes équivalentes pour représenter une même caractéristique. Un exemple parmi d'autres est illustré **Figure 30**. En effet, la charge formelle des atomes peut être indiquée de deux manières différentes dans un fichier SDF. Ceci peut être à l'origine d'erreurs de post-traitement (par exemple dans le cas d'un amarrage moléculaire) dans le cas où l'outil de criblage virtuel employé recherche l'information au mauvais endroit.¹¹⁰

La première étape du protocole VSPrep se déroule en trois temps. Tout d'abord, il convient de configurer le travail à réaliser en déterminant à la fois quelles étapes de préparation doivent être réalisées et de quelle manière. Pour cela, un fichier de configuration par défaut propose à l'utilisateur une centaine de paramètres modulables. Si l'utilisateur ne fournit aucun fichier de configuration, alors toutes les options ont une valeur par défaut et le format moléculaire est déduit de l'extension du fichier.

La seconde tâche à effectuer est la lecture du fichier moléculaire. Comme indiqué précédemment, plusieurs formats sont supportés et lus à l'aide des nœuds disponibles. Les entrées de ces fichiers qui présentent des problèmes de format sont enregistrées à part pour laisser à l'utilisateur le soin de les inspecter. Les molécules qui ont pu être lues sont alors converties au format RDKit. Parce que la représentation des molécules en chémoinformatique a des limites, la conversion de certaines entrées peut échouer, pour différentes raisons. Par exemple, l'hypervalence de certains atomes dans des mélanges empêche la conversion de l'ensemble des entités présentes (**Figure 31**). Afin de prévenir ce cas de figure, le mode de détection des erreurs de RDKit est désactivé et le soin de filtrer les erreurs revient à l'étape de standardisation, dont la première tâche est justement de séparer les mélanges.

¹⁰⁶ Daylight Theory: SMILES, <http://www.daylight.com/dayhtml/doc/theory/theory.SMILES.html>, consulté le 31 Août 2017

¹⁰⁷ S. R. Heller *et al.*, InChI, the IUPAC International Chemical Identifier, *Journal of Cheminformatics*, 7 (2015), 23

¹⁰⁸ Ctfles.Book - MDL_SDF.pdf, http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/MDL_SDF.pdf, consulté le 31 Août 2017

¹⁰⁹ Tripos Mol2 File Format, <http://chemyang.ccnu.edu.cn/ccb/server/AIMMS/mol2.pdf>, consulté le 31 Août 2017

¹¹⁰ RDKit / Mailing Lists, <https://sourceforge.net/p/rdkit/mailman/message/34740156>, consulté le 31 Août 2017

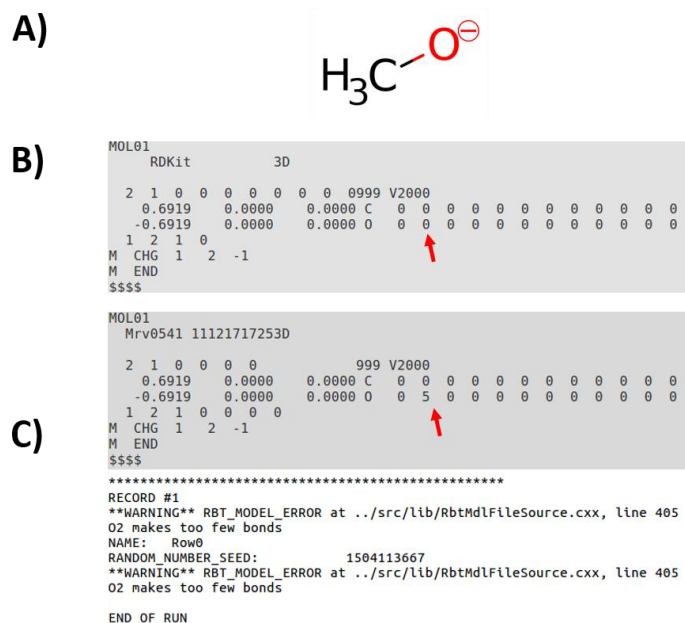
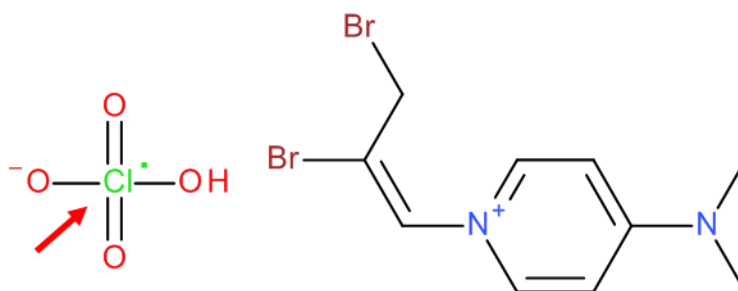


Figure 30. Exemple d'incompatibilité du format SDF entre RDKit et rDock. rDock est un outil d'amarrage moléculaire détaillé dans la partie 4.2.4.1. A) Molécule de méthanol portant une charge négative sur l'atome d'oxygène. B) Format SDF généré avec RDKit (haut) et Marvin (bas). Des flèches rouges indiquent la différence entre les deux fichiers. C) Erreur obtenue lors de l'amarrage moléculaire de la structure générée par RDKit avec l'outil rDock¹¹¹ à cause de la non détection de la charge négative (problème de valence).



Amb10553997

Figure 31. Exemple d'un mélange échouant à la conversion au format RDKit à cause d'un sel. L'erreur est causée par l'hypervalence indiquée en rouge.

Enfin, la dernière étape de la lecture du fichier moléculaire est l'identification claire et sans ambiguïté des molécules. Pour ce faire, un numéro séquentiel est ajouté à chaque identifiant. Afin de faciliter l'automatisation du protocole, le champ renseignant l'identifiant des molécules doit être spécifié

¹¹¹ S. Ruiz-Carmona *et al.*, RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids, *PLoS Computational Biology*, 10.4 (2014), e1003571

par l'utilisateur. Le résultat obtenu est une table KNIME contenant l'information structurale ainsi que l'identifiant d'origine, auquel un suffixe est ajouté pour le rendre unique (**Figure 32**).

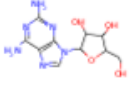
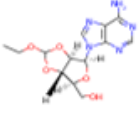
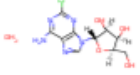
Row ID	S Molecule_ID	RDKit	S database	S dupl_Molecul...	S Entry_ID
Row0	Amb1001785_1		adenosine-similar	Amb1001785_1	Amb1001785_1
Row1	Amb1006965...		adenosine-similar	Amb10069657_1	Amb10069657_1
Row2	Amb1007059...		adenosine-similar	Amb10070594_1	Amb10070594_1

Figure 32. Tableau KNIME obtenu après l'étape de lecture d'un fichier SDF avec VSPrep.

3.2.3.2 Standardisation des structures

La "bonne" représentation des molécules est une condition essentielle non seulement à leur traitement dans des projets chémoinformatiques mais également à leur préparation. Cependant, toutes les bases de données moléculaires ne sont pas conçues avec le même objectif, ce qui se traduit par une grande hétérogénéité aussi bien au niveau des données que de leur représentation.¹¹² Par exemple, un commerçant de produits chimiques a besoin de renseigner les sels associés à un composé car ceux-ci déterminent son aspect (poudre, huile, etc.) ainsi que sa solubilité. Ceci représente néanmoins peu d'intérêt pour le chémoinformaticien qui exécute un criblage virtuel pour identifier des touches. Une autre différence concerne l'utilisation de mélanges dans les expériences de criblages expérimentaux (afin de réduire les coûts des expériences). En effet, avoir plusieurs entités dans une seule entrée d'un format moléculaire donné complique grandement la tâche de détecter et de retirer les erreurs dans les structures lors d'un criblage virtuel.

Ainsi, la première tâche réalisée au cours de l'étape de standardisation est la suppression des sels afin de ne garder que la molécule avec le plus haut poids moléculaire. Dans le cas où l'utilisateur souhaite conserver d'éventuels cofacteurs présents dans les mélanges, il est également possible de séparer la structure en plusieurs entrées différentes. La suite logique consiste alors à appliquer les règles strictes de conversion RDKit (valence, stéréochimie, etc.) afin de retirer les molécules erronées dès que possible. Nous aurions pu décider d'ajouter une option pour corriger les erreurs structurales à la volée, mais nous avons choisi de laisser ce soin à l'utilisateur. Il est en effet difficile de considérer tous les cas possibles mais cela permet de rester vigilant quant aux structures enregistrées. Une fois ces opérations menées à bien, les

¹¹² A. J. Williams *et al.*, Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation, *Drug Discovery Today*, 17.13–14 (2012), 685–701

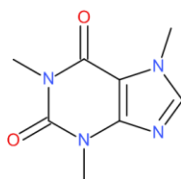
molécules sont représentées sous forme Kekulé afin de limiter les erreurs potentielles dues à une détection variable de l'aromaticité en fonction des outils. De plus, les molécules sont neutralisées puis leur tautomère canonique est calculé grâce à l'outil MolVS, basé lui aussi sur RDKit, afin d'avoir une seule et même représentation pour chaque molécule.

Cette démarche nous assure l'élimination d'un maximum de doublons, plus importante que le recours à des approches plus simples comme l'utilisation directe de l'InChIKey (voir **Figure 33** et **Figure 34**). En effet, les règles de standardisation implémentées dans l'InChI sont pour le moment seulement réduites à des cas très courants (une nouvelle version avec plus de règles est en cours de développement).

La définition des centres stéréochimiques est une autre source d'embarras lors du nettoyage des bases de données moléculaires. Il est en effet très fréquent que des centres chiraux sur une molécule soient laissés indéfinis par manque d'information structurale, par oubli ou bien parce que les mauvais atomes sont considérés comme chiraux. Ceci peut alors résulter dans la génération de conformations d'un seul stéréoisomère et donc biaiser les résultats de criblage virtuel. De plus, pour les molécules en trois dimensions, cette information est superflue et peut être donc omise, ce qui peut poser problème avec les outils qui recherchent des informations pré-calculées. Dans ce cas, nous retirons toute indication de stéréochimie des molécules puis définissons les centres stéréochimiques (stéréochimie tétraédrique et double liaison) uniquement à partir des coordonnées tridimensionnelles. Pour les molécules en deux dimensions, RDKit effectue lui-même une standardisation des étiquettes chirales lors de l'étape de conversion.

La bonne spécification des stéréochimies permet également d'identifier des doublons, notamment dans le cas où une même molécule se retrouve plusieurs fois avec des combinaisons différentes de centres chiraux définis et indéfinis. Chaque atome est renuméroté de manière canonique afin de pouvoir comparer les différentes spécifications de chaque centre stéréochimique, défini ou non. Cette comparaison repose également sur le calcul de l'InChIKey en tâche finale pour identifier les doublons, comme expliqué dans la sous-partie correspondante.

Caféine
(Amb19079926)



InChI InChI=1S/C8H10N4O2/c1-10-4-9-6-5(10)7(13)12(3)8(14)11(6)2/h4H,1-3H3

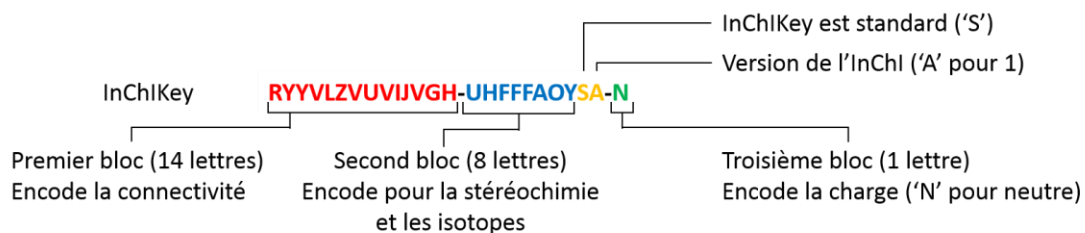


Figure 33. Exemple de la structure de l'InChIKey, déduit de l'InChI, avec la caféine.

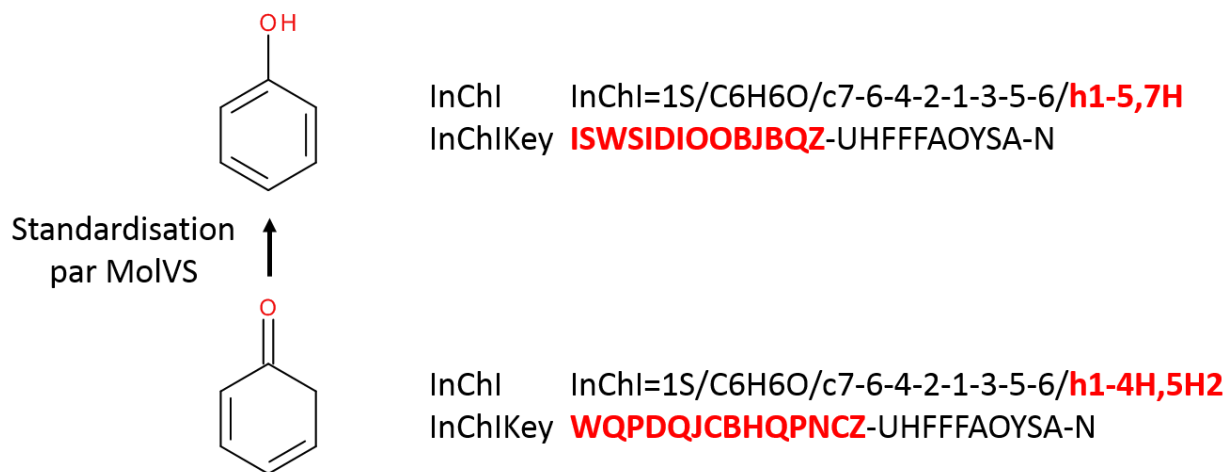


Figure 34. Exemple d'un cas de figure où MolVS permet de détecter deux tautomères comme une seule molécule là où le programme InChI n'y arrive pas.

3.2.3.3 Filtre des composés non désirés

Cette étape vise à réduire le plus possible le volume de la base de données moléculaire étudiée. En effet, certaines des opérations décrites-ci après s'avérant particulièrement chronophages, il est important de ne pas surcharger le *workflow* inutilement. Une série de filtres optionnels a donc été développée de manière à filtrer les molécules le plus rapidement possible.

Les filtres les plus rapides identifient la présence de caractéristiques structurales par expressions régulières déjà répertoriées directement dans le SMILES. Cela permet par exemple de retirer de l'ensemble des molécules de solvants qui ne présentent aucun intérêt thérapeutique. RDKit possède un nœud déjà implémenté dans KNIME permettant de retirer des sels. Cependant, ce programme ne peut les retirer qu'à la condition qu'ils soient en présence d'autres molécules. Les molécules restantes sont récupérées sous forme aromatique, nous obligeant à les repasser sous forme Kekulé. Enfin un bogue du programme modifie l'étiquette 3D des molécules enregistrées en deux dimensions de « faux » à « vrai », véritable source d'erreur comme nous le verrons par la suite. Nous avons donc opté pour une autre approche, qui consiste à utiliser une base de solvants rassemblée à partir des informations du nœud RDKit déjà mentionné à laquelle est ajoutée une liste d'agents de cristallisation communément trouvés dans la littérature.¹¹³ Un *workflow* annexe a ainsi été développé pour permettre à l'utilisateur de moduler plus facilement le contenu de cette liste de solvants, qui doivent eux-aussi être standardisés. Les composés identifiés comme des solvants sont alors retirés du jeu de données par simple reconnaissance d'InChIKey. Les molécules comportant des isotopes, des carbones chargés ou ne possédant aucun atome de carbone sont ensuite filtrées en parcourant la chaîne de caractères composant le SMILES (associé nativement au format RDKit) de chaque entrée.

La seconde partie des filtres repose sur le calcul de descripteurs moléculaires. Ainsi, les molécules contenant trop de centres chiraux non définis (2 par défaut) sont retirées à partir d'un certain seuil afin de ne pas énumérer des centaines, voire des milliers, de stéréoisomères à partir d'une même molécule. Les molécules possédant un grand nombre de liaisons rotatives (50 par défaut), donc très flexibles, sont également filtrées car cela se traduit par une recherche conformationnelle extrêmement longue. Les molécules comportant des atomes autres que H, B, C, N, O, P, S, F, Cl, I et Br sont considérées inorganiques et retirées du jeu de données car rarement employées en recherche thérapeutique. Enfin, les molécules comportant des cycles non fusionnés de plus de 7 atomes restent à l'heure actuelle un défi pour la recherche conformationnelle avec RDKit¹¹⁴ et peuvent donc également être retirées.

¹¹³ H. Strömbergsson *et al.*, A Chemogenomics View on Protein-Ligand Spaces, *BMC Bioinformatics*, 10.6 (2009), 1–11

¹¹⁴ J.-P. Ebejer *et al.*, Freely Available Conformer Generation Methods: How Good Are They?, *Journal of Chemical Information and Modeling*, 52.5 (2012), 1146–58

3.2.3.4 Suppression des doublons

On appelle doublons des molécules présentes dans des bases de données moléculaires sous des formes dupliquées ou sous différents états de protonation ou de tautomérie. S'il est intéressant en soi d'avoir plusieurs formes de chaque molécule pour en exploiter au maximum le potentiel, cela peut se traduire par un surcoût de temps de calcul. Dans la suite de ce chapitre, on dénommera « entrées » les différentes lignes d'une table KNIME correspondant à une seule et même molécule, par exemple après énumération des tautomères. Comme les molécules ont été standardisées sous forme de tautomère canonique, il suffit de comparer les InChIKey correspondants afin de retirer l'essentiel des doublons, la limite étant le nombre de réactions de standardisation actuellement prévues par l'outil MolVS.

Une autre forme de doublons peut être considérée lors de la comparaison des centres chiraux définis et non définis. Ainsi, si une molécule a un centre chiral non défini, elle n'a pas le même InChIKey^{115,116} que l'une de ses configurations définies. Si une même molécule est présente à la fois avec un centre chiral défini et indéfini, alors des doublons seront générés lors de l'étape d'énumération des stéréoisomères. La stratégie adoptée ici est donc de résumer le maximum de molécules par le minimum d'entrées. Par exemple, si une molécule avec un centre chiral est présente sous une configuration définie dans une entrée et sous une forme indéfinie dans une autre, seule la forme indéfinie sera conservée. Les deux énantiomères seront énumérés ultérieurement pour chacun des tautomères identifiés pour l'entrée retenue. Plusieurs cas de figures sont prévus dans le protocole et sont décrits dans la . Il peut sembler au premier abord contre-intuitif de retirer des énantiomères puis les énumérer ensuite, mais cela relève en fait d'une optimisation du temps de calcul. En effet, ne pas regrouper les entrées d'une même molécule par stéréochimie peut résulter en un nombre considérable de configurations différentes potentiellement partiellement définies. Seuls les centres chiraux définis étant protégés lors de l'énumération des tautomères, cela peut résulter non seulement en une surcharge de travail pour le *workflow* mais également en la création de nouveaux doublons qu'il faut ensuite retirer.

¹¹⁵ I. Pletnev *et. al.*, InChIKey Collision Resistance: An Experimental Testing, *Journal of Cheminformatics*, 4 (2012), 39

¹¹⁶ J. Chambers *et. al.*, UniChem: Extension of InChI-Based Compound Mapping to Salt, Connectivity and Stereochemistry Layers, *Journal of Cheminformatics*, 6.1 (2014)

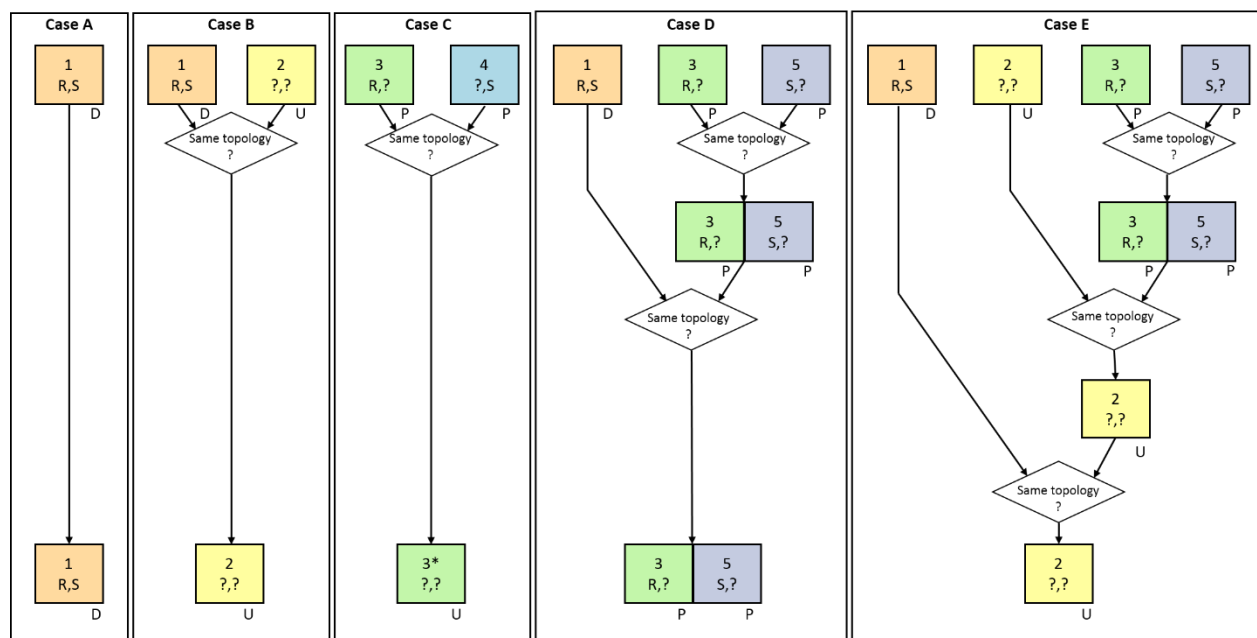


Figure 35. Traitement des doublons par regroupement de stéréochimie. Cas A : pas de doublon. Cas B : la stéréochimie indéfinie est préférée à la stéréochimie définie. Cas C : l'union des centres stéréochimiques non définis est effectuée pour les molécules avec des stéréochimies partiellement définies. Cas D et E : dans le cas de plusieurs définitions de la stéréochimie d'une même molécule, la comparaison est d'abord faite entre les centres stéréochimiques partiellement définis pour prendre en compte le cas où leur union résulte en une stéréochimie complètement indéfinie. Le résultat est ensuite comparé à une stéréochimie non définie si présente, puis à une stéréochimie définie, toujours si présente.

3.2.3.5 Enumération des tautomères

Une fois la base de données moléculaire nettoyée, il peut se révéler intéressant, voire nécessaire, d'énumérer différents états de protonation et de tautomérie pour chaque molécule.^{117,118,119,120,121,122,123}

Deux approches principales permettent l'énumération de tautomères. La première ne considère que la topologie de la molécule et consiste à appliquer des règles de transformation simples pour varier la position des hydrogènes près ou sur des groupements fonctionnels déterminés.

La seconde approche, plus complexe, prend en compte le pH de l'environnement pour déduire les états de protonation possibles des composés et ainsi prédire une distribution des états de protonation et de tautomérie à un pH donné (**Figure 36**). Cette dernière méthode permet d'énumérer des tautomères plus pertinents mais nécessite d'avoir une idée du pH dans la cavité d'une protéine, information difficile à appréhender.

Comme mentionné précédemment, ce *workflow* est principalement basé sur la librairie de chémoinformatique RDKit. Malgré une communauté croissante et très active, il n'y a pas encore dans cette librairie, à notre connaissance, d'implémentation disponible d'énumération de tautomères prenant en considération le pH. Pour cette raison, et après comparaison des performances de plusieurs outils gratuits pour les laboratoires académiques (Ambit-Tautomers¹⁰³, TauThor¹²¹, SPORES¹¹⁷, OpenBabel¹²⁴) nous avons retenu l'utilisation du module cxcalc de la suite JChem. Ce dernier doit être installé séparément du *workflow* et doit être configuré avec une licence valide. Pour ces raisons, nous avons également utilisé Ambit-Tautomers, qui repose sur des critères énergétiques, même si ses performances sont en deçà de celles du module de ChemAxon, avec beaucoup plus de tautomères générés souvent peu plausibles par rapport au pH souhaité. Dans les deux cas, l'utilisation d'un outil externe nous contraint à convertir les molécules du format RDKit vers un format universel, lisible pour les deux outils mentionnés. Le format SDF a été préféré au format SMILES car il permet de conserver les coordonnées initiales des structures.

¹¹⁷ T. ten Brink *et al.*, Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein–Ligand Docking Results, *Journal of Chemical Information and Modeling*, 49.6 (2009), 1535–46

¹¹⁸ F. Oellien *et al.*, The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening†, *Journal of Chemical Information and Modeling*, 46.6 (2006), 2342–54

¹¹⁹ L. Guasch *et al.*, Experimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples, *Journal of Chemical Information and Modeling*, 56.11 (2016), 2149–61

¹²⁰ T. Kalliokoski *et al.*, The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening, *Journal of Chemical Information and Modeling*, 49.12 (2009), 2742–48

¹²¹ F. Milletti *et al.*, Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases, *Journal of Chemical Information and Modeling*, 49.1 (2009), 68–75

¹²² P. Pospisil *et al.*, Tautomerism in Computer-Aided Drug Design, 2003, <http://informahealthcare.com/doi/abs/10.1081/RRS-120026975%20>, consulté le 17 March 2014

¹²³ M. Sitzmann *et al.*, Tautomerism in Large Databases, *Journal of Computer-Aided Molecular Design*, 24.6–7 (2010), 521–51

¹²⁴ Open Babel, http://openbabel.org/wiki/Main_Page, consulté le 24 Novembre 2016

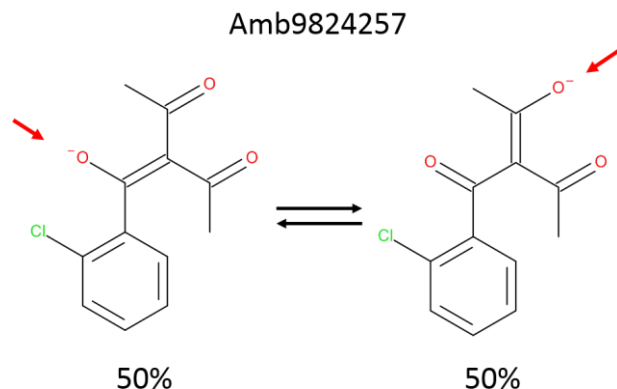


Figure 36. Tautomères majoritaires à pH 7.4 d'une molécule. Les deux formes sont prédites à 50% de présence par cxcalc (Marvin). Le déplacement de la charge négative indiquée par la flèche rouge peut influencer significativement les résultats de criblages virtuels.

3.2.3.6 Création de sous-librairies

Selon les projets, il peut être intéressant de se limiter à un espace chimique donné, par exemple pour rester dans le domaine des molécules dites « drug-like ». Différents filtres ont donc été implémentés permettant, selon le choix de l'utilisateur, d'annoter ou de retirer des molécules.

3.2.3.6.1 Filtres basés sur des descripteurs moléculaires

La première catégorie concerne des filtres reconnus par la communauté scientifique et basés sur des descripteurs moléculaires. Un premier exemple de ce genre de filtres provient de la règle de cinq de Lipinski¹²⁵, qui est issue d'une étude statistique visant à estimer la biodisponibilité orale des composés.

Les paramètres pris en compte sont le poids moléculaire (PM), l'hydrophobicité (logP) ainsi que le nombre de donneurs et d'accepteurs de liaison hydrogène (respectivement Num HD et Num HA). De par la nature de ces critères, ce filtre est très rapide à calculer. Il a également été optimisé par la suite (règle de Veber¹²⁶) et a inspiré nombre d'autres études, notamment la règle de trois pour les fragments¹²⁷ et la règle de quatre¹²⁸ pour les inhibiteurs de PPI (*protein-protein interactions*). Il est à noter que ces filtres constituent plus un guide qu'une règle absolue. Ainsi, un certain nombre de médicaments ne suivent pas la règle de cinq et ne sont donc pas « drug-like ». L'appartenance des composés à ces sous-librairies est déterminée à partir d'un nombre de violations autorisées que l'utilisateur peut spécifier. Ces filtres sont décrits ci-dessous dans la **Table 11**. L'ensemble des descripteurs moléculaires sont calculés avec RDKit.

¹²⁵ C. A. Lipinski *et. al.*, Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, *Advanced Drug Delivery Reviews*, 23.1–3 (1997), 3–25

¹²⁶ D. F. Veber *et. al.*, Molecular Properties That Influence the Oral Bioavailability of Drug Candidates, *Journal of Medicinal Chemistry*, 45.12 (2002), 2615–23.

¹²⁷ M. Congreve *et. al.*, A "Rule of Three" for Fragment-Based Lead Discovery?, *Drug Discovery Today*, 8.19 (2003), 876–77

¹²⁸ V. Hamon *et. al.*, 2P2IHUNTER: A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators via a Dedicated Support Vector Machine, *Journal of The Royal Society Interface*, 11.90 (2013), 20130860–20130860

Table 11. Liste des filtres implémentés pour créer des sous-librairies basées sur des descripteurs moléculaires.

Nom	Description
<i>Drug-like</i>	$150 \leq PM \leq 500$ Da $Slogp^* \leq 5$ $Num\ HD \leq 5$ $Num\ HA \leq 10$
<i>Extended drug-like</i>	Drug-like ET $Num\ RB \leq 7$ $TPSA < 150$
<i>Lead-like</i>	$250 \leq MW \leq 350$ Da $Slogp^* \leq 3.5$ $Num\ RB \leq 7$
<i>Fragment-like</i>	$MW < 300$ Da $Num\ HD \leq 3$ $Num\ HA \leq 3$ $Slogp^{**} \leq 3$
<i>Extended fragment-like</i>	Fragment-like ET $Num\ RB \leq 3$ $TPSA \leq 60$
<i>PPI-like</i>	$TPSA \leq 60$ $MW > 400$ Da $Num\ R \geq 4$ $Num\ HA > 4$ $Slogp > 4^{***}$

Num HD : Nombre d'hydrogènes donateurs, Num HA : Nombre d'hydrogènes accepteurs, MW : Poids moléculaire, TPSA : Aire de surface polaire topologique, Num RB : nombre de liaisons rotatives, Num R : nombre de cycles. * filtre d'origine avec xlogp, ** filtre d'origine avec clogp, *** filtre d'origine avec alogp.

3.2.3.6.2 Filtres basés sur des sous-structures

En 2010, Jonathan Baell a publié une liste de motifs structuraux souvent associés à des faux positifs dans les campagnes de criblage à haut débit. L'activité mesurée de ces molécules proviendrait non pas d'un quelconque effet de la molécule sur la cible d'intérêt, mais plutôt de suites de réactions dues à certains groupements fonctionnels. Sous l'acronyme de PAINS (*Pan Assay INterference compounds*), on regroupe des composés pouvant se lier de manière covalente au récepteur ou encore former des agrégats dans les tubes expérimentaux. Cette liste de motifs structuraux a d'abord été publiée au format

propriétaire SLN (Tripos) puis convertie au format public SMARTS avec l'aide de l'auteur original.¹²⁹ Cependant, du fait du problème des représentations différentes des molécules selon les outils, il n'était pas possible de reproduire les résultats initiaux et donc une partie des motifs PAINS n'étaient pas identifiés dans le jeu de données initial. C'est pourquoi de nouvelles requêtes SMARTS, plus spécifiques de l'outil RDKit, ont été redéveloppées par G. Landrum pour reproduire avec succès l'intégralité des touches.¹³⁰ Ces filtres sont implémentés tels quels dans le *workflow* par catégories A, B et C (correspondant respectivement aux filtres ayant provoqué plus de 150 touches, moins de 150 touches et moins de 15 touches, voir **Figure 37** pour un exemple de filtre). En 2017, une nouvelle série de filtres, cette fois-ci basés sur plusieurs études différentes, a été publiée et pourrait être incluse dans le *workflow*.

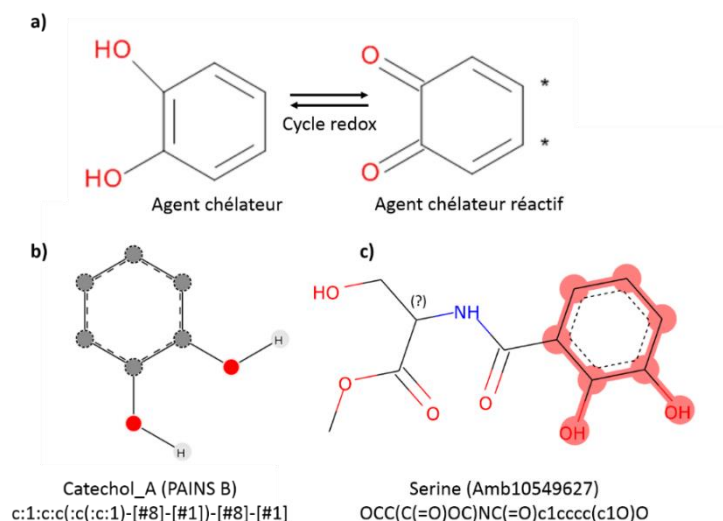


Figure 37. Exemple de reconnaissance d'une sous-structure considérée comme problématique par les filtres PAINS. a) Le cycle redox du catéchol provoque diverses interférences dans les tests biochimiques. b) Le filtre PAINS identifiant le catéchol au format SMARTS. c) Une molécule filtrée avec en rouge la sous-structure recherchée.

3.2.3.6.3 Filtres basés sur des librairies de référence

La troisième et dernière catégorie de filtres actuellement implémentée se base sur des recherches par similarité des composés par rapport à des librairies ciblées. Des empreintes moléculaires (*fingerprints*) de type clés MACCS¹³¹ sont calculées pour l'ensemble des entrées du jeu de données à préparer ainsi que pour les librairies de référence. La similarité de Tanimoto est alors calculée pour chaque combinaison entrée - référence et le maximum est comparé avec un seuil donné par l'utilisateur, déterminant l'appartenance à la sous-librairie correspondante.

¹²⁹ S. Saubern, Rajarshi Guha, and Jonathan B. Baell, KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries, *Molecular Informatics*, 30.10 (2011), 847–50

¹³⁰ G. Landrum, Curating the PAINS Filters, <http://rdkit.blogspot.com/2015/08/curating-pains-filters.html>, consulté le 10 Novembre 2016

¹³¹ J. Duan *et al.*, Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods, *Journal of Molecular Graphics & Modelling*, 29.2 (2010), 157–70

L'élaboration de ces sous-librairies a suivi une démarche uniformisée consistant à recenser dans la littérature des bases publiques ciblant des familles protéiques (Kinases, RCPG, récepteurs nucléaires) ou recensant des produits naturels (NP). Les fichiers moléculaires ont ensuite été standardisés avec VSPrep puis les empreintes moléculaires ont été calculées et sauvegardées pour chaque molécule. Pour chaque base de données finale, une sélection par diversité a été réalisée avec RDKit autorisant 1000 structures au maximum pour limiter les temps de calculs. Des *workflows* utilitaires permettent de moduler le contenu de ces librairies en modifiant les bases de données moléculaires sources.

Le seuil de similarité appliqué, habituellement décrit entre 0,6 et 0,7, a été déterminé en faisant varier sa valeur entre 0 et 1, comme indiqué sur la **Figure 38**.

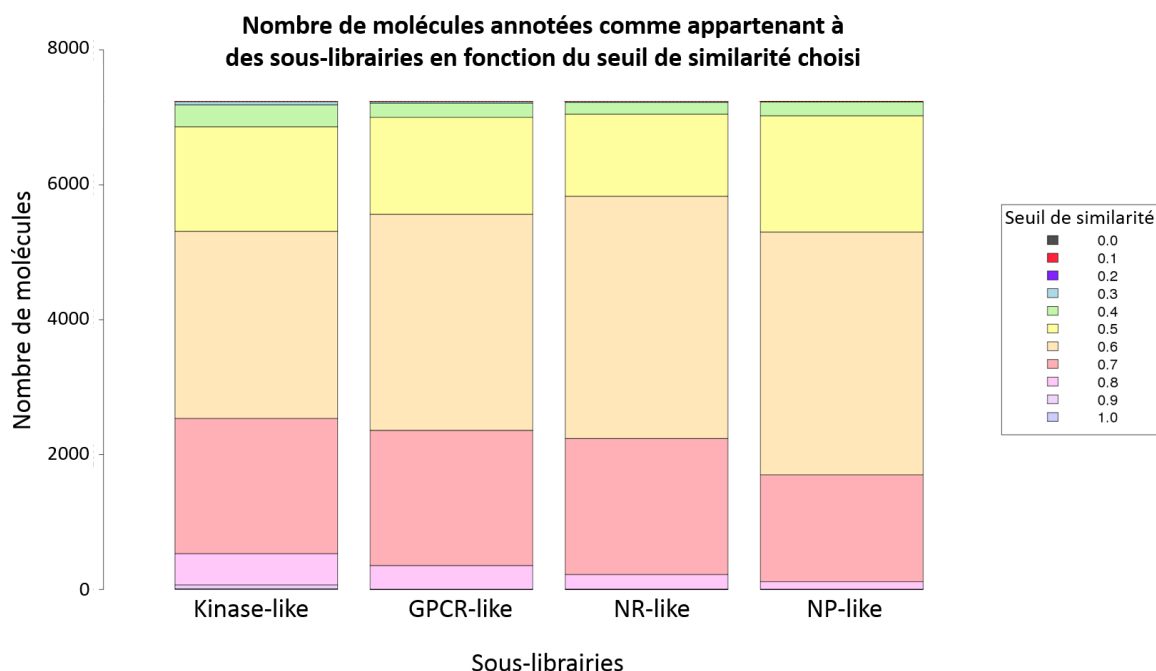


Figure 38. Nombre de molécules uniques indiquées comme appartenant à une sous-librairie selon un seuil de similarité. La base de données utilisée comme échantillon est la base interne de l'ICOA (8000 molécules).

Pour le moment quatre librairies de référence sont incluses dans le *workflow*.

3.2.3.6.3.1 Kinase-like

Comme il a été décrit au chapitre précédent, les protéines kinases représentent un enjeu majeur pour la recherche thérapeutique. De nombreuses études relatent les campagnes de criblages à haut débit (HTS) réalisées pour identifier de nouvelles molécules bioactives. Le jeu de données ici utilisé comme librairie de référence provient de la compilation de plusieurs publications par le docteur N. Bosc³⁶, données

récemment actualisées par le doctorant F. Carles. Pas moins de six études^{55,102,132,133,134,135} ont été analysées et les données recoupées avec la base ChEMBL⁹, ce qui a abouti à environ 2000 molécules ayant une activité avec un IC₅₀ inférieur à 10µM sur 200 protéines kinases. Les structures ont été extraites au format SMILES depuis la base ChEMBL et standardisées avec VSPrep.

3.2.3.6.3.2 GPCR-like

Les RCPG, récepteurs couplés aux protéines G, représentent un domaine d'étude très important dans la recherche de nouveaux composés thérapeutiques, avec 30 à 40% de part du marché des médicaments commercialisés. Les travaux de l'équipe de C. N. Cavasotto¹³⁶ ont permis de mettre à disposition de la communauté scientifique des données d'activité, extraites de la base GLIDA¹³⁷ (elle-même recensant l'information présente dans différentes bases de données d'activité) reliées à des structures issues de la base PubChem.⁶ Le jeu de données représente pas moins de 25 145 ligands classés comme agonistes ou antagonistes sur un total de 147 RCPG. Ce jeu de données est essentiellement employé comme *benchmark* pour du criblage virtuel, à l'instar de la base DUD-E¹³⁸, et comporte ainsi environ un million de structures leurres (*decoys*), c'est-à-dire des molécules proches d'un point de vue physico-chimique mais différent d'un point de vue structural de composés actifs connus mais dont l'activité n'a pas été mesurée expérimentalement.

Les auteurs ont préparé les structures par l'outil LigPrep¹³⁹ de Schrödinger (prédiction du tautomère le plus probable à pH physiologique et identification des stéréoisomères) et les empreintes moléculaires MACCS ainsi que les descripteurs moléculaires ont été calculés avec la librairie de chémoinformatique OpenBabel. Le jeu de données utilisé comme référence pour VSPrep est une sélection par diversité des molécules actives uniquement (agonistes et antagonistes), c'est-à-dire présentes dans la GPCR Ligand Library (GLL).

3.2.3.6.3.3 NR-like

Les récepteurs nucléaires représentent un autre domaine de recherche thérapeutique intense. Les travaux de l'équipe de M. Montes ont permis la création de NRList BDB¹⁴⁰, une base de données établie manuellement et regroupant un ensemble de 9905 actifs et leurres sur un total de 27 récepteurs

¹³² J. T. Metz *et al.*, Navigating the Kinome, *Nature Chemical Biology*, 7.4 (2011), 200–202

¹³³ J. Bain *et al.*, The Selectivity of Protein Kinase Inhibitors: A Further Update, *The Biochemical Journal*, 408.Pt 3 (2007), 297–315

¹³⁴ T. Anastasiadis *et al.*, Comprehensive Assay of Kinase Catalytic Activity Reveals Features of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 29.11 (2011), 1039–45

¹³⁵ O. Fedorov *et al.*, A Systematic Interaction Map of Validated Kinase Inhibitors with Ser/Thr Kinases, *Proceedings of the National Academy of Sciences of the United States of America*, 104.51 (2007), 20523–28

¹³⁶ E. A. Gatica and Claudio N. Cavasotto, Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors, *Journal of Chemical Information and Modeling*, 52.1 (2011), 1–6

¹³⁷ Y. Okuno *et al.*, GLIDA: GPCR-Ligand Database for Chemical Genomic Drug Discovery, *Nucleic Acids Research*, 34.Database issue (2006), D673-677

¹³⁸ M. M. Mysinger *et al.*, Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, *Journal of Medicinal Chemistry*, 55.14 (2012), 6582–94

¹³⁹ Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.

¹⁴⁰ N. Lagarde *et al.*, NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database, *Journal of Medicinal Chemistry*, 57.7 (2014), 3117–25

nucléaires. Les agonistes et antagonistes inverses, partiels ou faibles ont été retirés du jeu de données par étude bibliographique (ChEMBL⁹) pour ne garder que les « vrais » agonistes et antagonistes. Les structures des actifs ont été extraites de la base ZINC¹⁴¹ aux formats Mol2 et SMILES quand disponibles, sinon de la base ChEMBL au format SMILES. Les leurres ont été générés par l'outil de la base DUD-E.¹³⁸ La plateforme proposée facilite la réalisation de criblages virtuels tels que l'amarrage moléculaire en proposant des structures de protéines corrigées et disposant de charges partielles calculées dans le champ de force AMBER.⁸¹

Les structures des ligands présents dans la base ZINC sont disponibles soit en Mol2, soit en SMILES. Pour les autres, seul le SMILES est extrait de la base ChEMBL ou de la base PubChem et les coordonnées tridimensionnelles sont générées par l'outil Corina (version démo en ligne¹⁴²). Les charges partielles ont été calculées avec OpenBabel. Nous avons retiré les leurres de notre jeu de données pour nous concentrer uniquement sur les molécules actives (7853 agonistes et 2053 antagonistes). Nous avons standardisé les structures en partant du format SMILES, les charges partielles n'étant de toute façon actuellement pas prises en compte dans notre outil de préparation des petites molécules.

3.2.3.6.3.4 NP-like

Une base de données de produits naturels a également été développée. Elle provient de l'intégration de huit bases de données publiques spécifiques des produits naturels, décrites dans le chapitre suivant. Plus de bases avaient été identifiées mais n'ont pas été retenues car il n'était pas possible d'identifier l'organisme source, ce qui laissait le doute quant à leur véritable provenance. Au final 4000 structures uniques sont disponibles pour effectuer la recherche par similarité.

L'assemblage et la préparation de cette base de produits naturels sont décrits en détails dans la partie **3.3.2**. Un recensement des bases de produits naturels publiques a permis de sélectionner 8 bases différentes affichant les relations molécule - organisme source.

3.2.3.7 Enumération des stéréoisomères

Comme mentionné précédemment, la présence de marqueurs adéquats sur les atomes et les liaisons des molécules est nécessaire pour indiquer la stéréochimie des composés. Sans cette information, la génération de conformations résulte en un mélange indéterminé de stéréoisomères, ce qui complique grandement la traçabilité des entrées. Parmi les différents outils disponibles (**Table 12**), nous avons retenu ChemAxon car il permet d'énumérer les centres stéréochimiques non définis. Dans les nœuds KNIME disponibles, le *SMILES Stereoisomer Enumerator* (Vernalis) ne varie que les centres chiraux définis. Nous ne sommes pas parvenus à faire fonctionner le nœud Indigo correspondant (Enantiomer Generator) sur notre installation. Nous avons songé à développer un script RDKit pour contourner ce problème, mais

¹⁴¹ J. J. Irwin *et al.*, ZINC: A Free Tool to Discover Chemistry for Biology, *Journal of Chemical Information and Modeling*, 52.7 (2012), 1757–68

¹⁴² Online Demo - Fast 3D Structure Generation with CORINA Classic | MN-AM, https://www.mn-am.com/online_demos/corina_demo, consulté le 22 Août 2017

d'après les retours de la communauté RDKit, ce dernier est déjà en cours de développement et devrait être prochainement disponible.

Table 12. Liste des outils in silico testés pour l'énumération de stéréoisomères

Outil	Distributeur	Disponibilité	Enumération des centres stéréochimiques
cxcalc	ChemAxon	commercial	définis, non définis ou les deux pour centres chiraux et doubles liaisons
Speedy SMILES Stereoisomer Enumerator	Vernalis	<i>open source</i> , inclus dans KNIME	définis, centres chiraux
RDKit	RDKit	open source	définis, non définis ou les deux pour centres chiraux et double liaisons*
OpenBabel	OpenBabel	<i>open source</i>	non définis, centres chiraux ¹⁴³
Indigo Enantiomer Generator (KNIME)	EPAM	<i>open source</i>	**
SPORES	Université de Tübingen	<i>open source</i>	erreurs rencontrées lors des tests***

*à notre connaissance l'outil n'est pas encore accessible à la communauté scientifique

** Problème décrit sur le forum de KNIME depuis plusieurs années, pas de correction de la part des auteurs

*** Nécessite la conversion de la molécule au format mol2

3.2.3.8 Génération de conformations

Dans le cas où la structure tridimensionnelle de la cible thérapeutique est disponible, de nombreuses techniques de criblage virtuel peuvent être employées afin de prédire le mode d'interaction entre le ligand et le récepteur. Parmi ces dernières, on retiendra surtout l'amarrage moléculaire (*docking*) et la recherche pharmacophorique mentionnés au chapitre précédent. De manière moins directe, des études se basant uniquement sur les coordonnées tridimensionnelles des ligands peuvent également être menées afin de sélectionner les molécules présentant la forme désirée (globulaire, plane, ponctuelle), toujours d'après l'analogie clé-serrure. Tous ces travaux nécessitent des coordonnées tridimensionnelles réalistes pour les ligands afin d'améliorer les résultats. La grande majorité des algorithmes d'amarrage moléculaire ne font varier que les angles dièdres non impliqués dans des cycles, il peut alors se révéler nécessaire de générer

¹⁴³ General Discussion - How to Enumerate All Possible Stereoisomers of a Molecule with a Not Defined Stereocenter?, <http://forums.openbabel.org/How-to-enumerate-all-possible-stereoisomers-of-a-molecule-with-a-not-defined-stereocenter-td4659459.html>, consulté le 30 Août 2017

plusieurs conformations par ligand afin d'échantillonner convenablement l'espace conformationnel des molécules.

Les avancées en cristallographie des petites molécules ont permis aux chimioinformaticiens d'accéder à de plus en plus de structures tridimensionnelles, de mieux en mieux résolues. Des jeux de données tels que la *Cambridge Structural Database*¹⁴⁴ (CSD) et Astex ont été établis pour valider les outils de génération de conformations. La validation d'un outil repose sur plusieurs critères. Tout d'abord, la qualité des conformations est validée par comparaison de déviation de la racine de la moyenne des carrés (RMSD pour *Root Mean Square Deviation*). Dans notre cas, les valeurs à comparer sont les coordonnées atomiques entre une structure expérimentale, obtenue par cristallographie, et une structure théorique, générée par ordinateur. Plus le RMSD est faible, plus les structures sont superposables, une valeur nulle indique une superposition totale des deux structures et indique donc que l'outil employé a pu reproduire la conformation de référence. Cependant, une structure cristallographique ne donne qu'une vision statique d'une molécule et ne reflète que partiellement la réalité. La qualité de l'échantillonnage conformationnel, c'est-à-dire l'aptitude de l'outil à s'écarter des solutions connues tout en respectant les contraintes géométriques et énergétiques propres au champ de force employé pour trouver des conformations stables pertinentes, est également cruciale. Des mesures reposant sur le RMSD, la forme des composés (PMI pour Moment Principal d'Inertie¹²) ou encore le rayon de giration sont couramment employées afin d'évaluer la qualité de l'exploration conformationnelle. Les deux critères suivants découlent des contraintes techniques liées aux temps de calculs : la vitesse de génération et l'identification des conformations redondantes (toujours selon le RMSD). Du fait de la nature extrêmement complexe de l'opération effectuée, la recherche conformationnelle, surtout lorsqu'elle emploie des étapes de minimisation dans un champ de force, est l'opération la plus chronophage de la préparation des ligands. Générer des conformations redondantes peut se traduire par une multiplication de solutions identiques obtenues lors du criblage virtuel, d'où un gâchis des ressources pour une étape déjà très chronophage.

C'est pourquoi nous nous sommes appliqués à optimiser cette étape autant que possible. Plusieurs comparatifs d'outils de génération de conformations ont déjà été publiés. Nous nous en sommes inspirés pour identifier l'outil le plus apte à répondre à nos besoins. L'une des raisons majeures pour laquelle RDKit a été retenu comme l'outil principal de ce protocole de préparation de molécules est sa capacité à proposer des conformations de bonne qualité en un temps relativement court. La méthode employée repose sur la création d'une conformation tridimensionnelle initiale à partir de distances géométriques basées sur des mesures cristallographiques afin de partir de modèles plus réalistes. Cette conformation est alors minimisée dans un champ de force jusqu'à atteindre un critère de convergence ou un nombre maximum d'itérations. Dans le cas où plusieurs conformations sont nécessaires, plusieurs conformations initiales sont générées puis minimisées indépendamment. Cette approche est appropriée quand la molécule est très flexible mais elle pose deux problèmes principaux. Tout d'abord, le nombre de conformations initiales nécessaires pour retranscrire correctement l'échantillonnage de la molécule varie selon sa flexibilité (**Figure 39**). Il faut donc adapter le nombre de conformations initiales demandées par molécule pour limiter les risques de redondance. La seconde question à considérer est comment comparer les conformations entre elles. En effet, la majorité des algorithmes d'amarrage moléculaire débutent leur

¹⁴⁴ The Cambridge Crystallographic Data Centre (CCDC), <https://www.ccdc.cam.ac.uk>, consulté le 31 Août 2017

processus en variant aléatoirement les angles dièdres des liaisons non impliquées dans des cycles et laissent les angles dièdres impliqués dans les cycles inchangés (**Figure 40**). Ceci signifie que l'exploration conformationnelle n'est assurée que pour les angles dièdres non impliqués dans un cycle. Il convient alors d'explorer convenablement l'espace conformationnel des cycles avant le criblage virtuel. RDKit permet de générer beaucoup de conformations par molécule, y compris des conformations de cycle. Un script Python a été développé pour classifier les molécules selon la présence possible ou non de conformations de cycle. Il permet notamment de calculer le RMSD entre les atomes appartenant à des cycles seulement, ce qui permet de réduire considérablement le nombre de conformations générées à considérer lors du criblage virtuel.

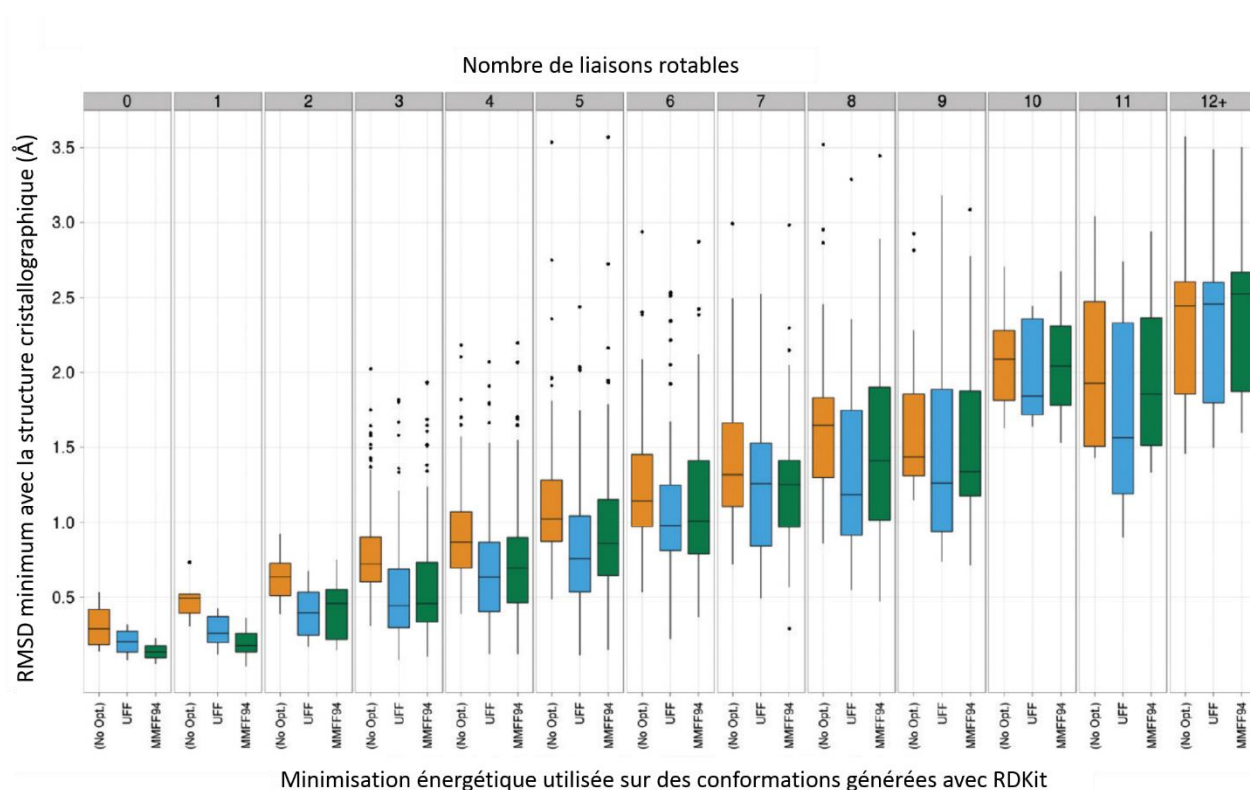


*Figure 39. Diagrammes en moustache montrant le RMSD minimum entre les structures cristallographiques et 10 conformations générées avec RDKit pour chacune d'entre elles et partitionnées par nombre de liaisons qui tournent.*¹¹⁵

Amb1031787

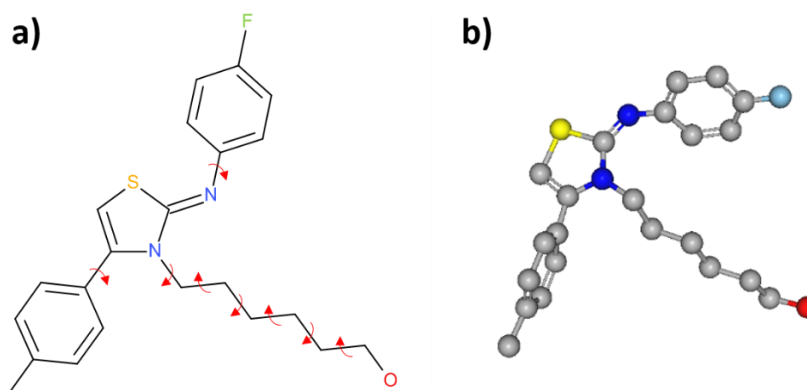


Figure 40. Représentation des liaisons qui tournent sur une molécule. a) Les flèches rouges indiquent les liaisons susceptibles d'effectuer des rotations sur elles-mêmes et les angles dièdres qu'elles forment sont changés lors des amarrages moléculaires. b) Représentation d'une conformation correspondante en trois dimensions.

Néanmoins, l'outil mis en place présente des limites significatives. Il aurait été certainement plus pertinent de développer un algorithme générant des conformations de cycles, mais nous nous sommes limités à mettre au point un script permettant de filtrer les conformations par calcul de RMSD uniquement sur les cycles pouvant présenter différentes conformations. Nous avons également mentionné le problème du nombre de conformations à générer selon le type de molécule. Il a été montré que la méthode décrite ne permet pas de reproduire de manière fiable des structures cristallographiques de molécules présentant des cycles de plus de six atomes¹⁴⁵ (considérées ici comme des macrocycles). Un nouvel algorithme¹⁴⁵, implémenté dans l'outil Surflex de Certara, a été proposé pour générer des conformations de macrocycles et semble répondre à nos attentes. Comme cet outil n'est pas gratuit, l'une des perspectives de ce travail sera de ré-implémenter cette approche avec RDKit afin de pouvoir générer des conformations de molécules plus complexes.

3.2.3.9 Analyse des résultats

En sortie du protocole, l'utilisateur souhaite obtenir une base de données de ses molécules préparées. Il peut être cependant difficile d'évaluer la qualité du travail accompli, ou simplement de la base de données moléculaire fournie en entrée. Pour permettre à l'utilisateur d'analyser les résultats rapidement, nous avons mis en place un système de suivi des molécules, auxquels s'ajoutent des rapports graphiques, générés de façon automatique. Ces graphes sont décrits en annexe de ce chapitre. Ils peuvent se révéler particulièrement précieux lors de la préparation des bases de données moléculaires, dans lesquelles le nombre de molécules peut dissuader la vérification manuelle des structures. Cela a notamment été le cas pour la chimiothèque de l'ICOA, dans laquelle nous avons pu détecter des erreurs dans des structures et sonder les limites de la représentation des molécules pour du criblage virtuel.

¹⁴⁵ A. E. Cleves *et al.*, ForceGen 3D Structure and Conformer Generation: From Small Lead-like Molecules to Macrocyclic Drugs, *Journal of Computer-Aided Molecular Design*, 31.5 (2017), 419–39

La **Figure 41** illustre trois exemples de cas rencontrés où les molécules fournies en entrée ne peuvent être utilisées pour du criblage virtuel. Le premier (a) concerne les liaisons covalentes de coordination (liaisons datives). En effet, telle que la molécule est représentée sur les deux premières structures, elle ne peut être interprétée par la majorité des outils chimioinformatiques. Par exemple, RDKit utilise le symbole « ~ » pour ces liaisons dans le format SMILES alors que ce symbole signifie un type de liaison inconnu d'après la nomenclature officielle.¹⁰⁷ Une représentation approximative est parfois utilisée pour pouvoir inclure ces molécules dans des criblages virtuels, affichée en dernière position, avec la formation d'une liaison covalente entre l'atome de brome et l'atome de phosphore. Les charges formelles de ces derniers ne sont alors plus neutres, ce qui impacte grandement les résultats de criblages virtuels.

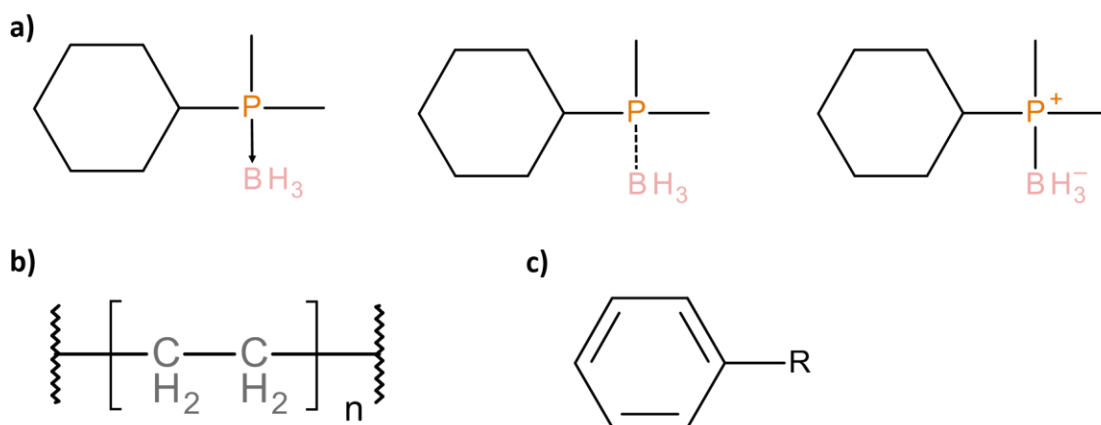


Figure 41. Exemples de limites de représentations en chimioinformatique. a) Les liaisons covalentes b) Les polymères c) Les structures de Markush

L'exemple (b) illustre le cas où une structure chimique (polyéthylène) n'est pas définie exactement. Ceci implique qu'elle pourrait être potentiellement infinie, il n'est alors pas possible de calculer des descripteurs moléculaires ou d'attribuer des coordonnées à l'ensemble des atomes. Certains outils se contentent de ne modéliser que le monomère, ce qui ne correspond pas à la réalité chimique du composé.

Le même raisonnement peut être appliqué à l'exemple (c), où un groupement R est employé pour signifier la présence d'un substituant donné sur cette position du cycle. Il est à noter qu'il s'agit d'une représentation extrêmement employée, notamment pour la couverture de brevets (structures de Markush). Ainsi, lorsque l'information de la nature du substituant est perdue, la structure de la molécule n'est plus définie. Certains outils proposent de remplacer ces groupements par des atomes de carbone ou bien simplement de les supprimer.

VSPrep a été notamment validé sur plusieurs bases de données moléculaires :

a) Chimiothèque de l'ICOA

Cette chimiothèque a pour but de répertorier les molécules ayant été synthétisées à l'Institut de Chimie Organique et Analytique d'Orléans (ICOA) et de pouvoir rapidement les localiser. Il n'y a donc initialement

pas vocation à réaliser des criblages virtuels sur ces molécules. Les quelques 8000 molécules recensées dans une base de données moléculaire Instant JChem de ChemAxon ont été dessinées manuellement par des chimistes organiciens non experts des contraintes de représentation propres à la chimoinformatique. Il s'agit pour cette raison d'un jeu de données intéressant car relativement restreint mais compliqué à gérer du fait de la multitude de cas particuliers à prendre en compte. Il s'agit également du premier jeu de données (format SDF) avec lequel l'outil VSPrep a été paramétré. Un exemple de cas particulier également rencontré dans la base Ambinter est indiqué sur la **Figure 42** ci-dessous.

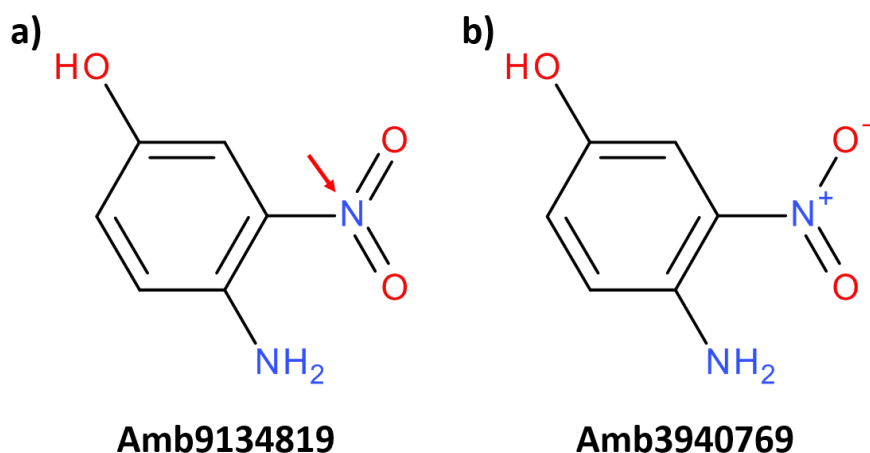


Figure 42. Exemple de deux représentations différentes du groupement nitro. a) La molécule est détectée avec un problème de valence sur l'azote indiqué par une flèche rouge et peut donc causer des erreurs. b) Le groupement nitro est maintenant chargé ce qui permet aux outils chimoinformatiques de traiter la molécule. Vu que la liaison portant le groupe peut tourner, le choix de l'oxygène qui doit porter la charge négative n'est pas important.

b) Chimiothèque Nationale (CN) et Chimiothèque Nationale Essentielle (CNE)

La Chimiothèque Nationale¹⁴⁶ résulte de l'initiative de rassembler les composés issus des laboratoires académiques de chimie en France dans une plateforme permettant de centraliser le stockage et facilitant la distribution pour les tests de criblage. Elle comporte environ 60 000 entrées administrées par l'outil Instant JChem (ChemAxon). Les travaux réalisés sur cette base de données ont permis d'identifier les parties de mon *workflow* qui nécessitaient les optimisations les plus urgentes afin d'obtenir les résultats dans des délais raisonnables et donc d'améliorer les performances globales de l'outil.

¹⁴⁶ F. Mahuteau-Betzer, The French National Compound Library: advances and future prospects, *Medecine Sciences: M/S*, 31.4 (2015), 417–22

c) Catalogue 2014 d'Ambinter

Ambinter¹⁴⁷ est une compagnie de vente de composés chimiques à l'échelle mondiale. Elle a été en activité 25 ans avant d'être rachetée en 2012 par la société Greenpharma¹⁴⁸, qui l'administre depuis lors. Elle constitue l'une des plus grandes bases de données de produits chimiques au monde, avec plus de 20 millions de références assemblées depuis les catalogues de nombreux fournisseurs.

Nous avons nettoyé la base de données Ambinter avec notre outil et nous avons transmis les résultats obtenus à notre partenaire industriel, Greenpharma qui a mis les fichiers préparés à disposition sur leur site internet (www.ambinter.com) pour des applications de criblage virtuel. Le point de départ a été le fichier au format SDF téléchargeable sur le site. Ce dernier avait déjà été partiellement nettoyé en interne grâce au logiciel Sybyl pour retirer les sels et les doublons strictes. Les difficultés rencontrées lors de la préparation de cette base de données moléculaire proviennent non seulement du volume de données, qui a permis de mettre en évidence des failles dans le logiciel KNIME (120 GB de mémoire vive consommés dans un nœud Python qui ne faisait que rajouter un numéro séquentiel à chaque identifiant !), mais également de l'exploration d'un espace chimique bien plus vaste que celui des chimiothèques explorées habituellement par la médecine thérapeutique traditionnelle.

Deux exemples illustrés **Figure 43** donnent une idée des cas rencontrés lors de sa préparation. Dans le premier cas (a), cinq molécules de méthane constituent un mélange fourni par un vendeur de produits chimiques. Ce type de structure n'est absolument pas adapté à du criblage virtuel et est retirée automatiquement du *workflow*.

Dans le second cas se trouve une molécule possédant un atome de ruthénium, qui est donc inorganique. Or, la majorité des outils de chémoinformatique ont été conçus pour des applications de recherche de nouveaux médicaments et supportent donc principalement des molécules organiques. Par exemple, le ruthénium n'est pas supporté dans le champ de force MMFF94 et il n'est donc pas possible de générer des conformations en l'utilisant avec RDKit. Par défaut, seules les molécules organiques sont préservées par le *workflow*.

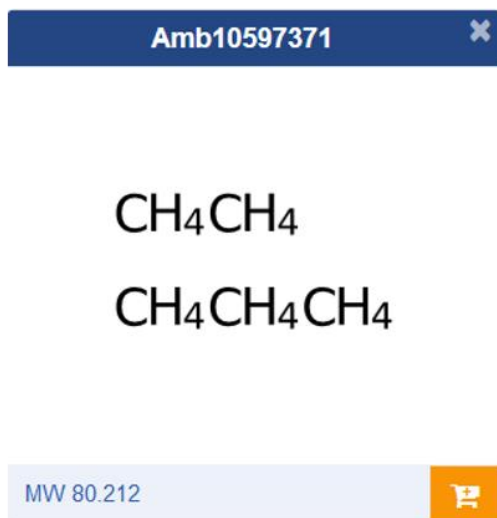
Les structures chimiques de cette base moléculaire n'étant pas confidentielles, le catalogue préparé (10 millions de structures tridimensionnelles en incluant les tautomères et les stéréoisomères) est maintenant disponible gratuitement sur le site Ambinter pour du criblage virtuel.¹⁴⁹

¹⁴⁷ Chemical Supplier of Screening Compounds, Building Blocks - Ambinter, <http://www.ambinter.com/>, consulté le 13 Novembre 2017

¹⁴⁸ GreenPharma | R&D Products and Services to Pharmaceutical, Cosmetic and Nutraceutical Research, <http://www.greenpharma.com/>, consulté le 13 Novembre 2017

¹⁴⁹ Ambinter Catalogue Request, <http://www.ambinter.com/catalog/SDF>, consulté le 12 Novembre 2017

a)



b)

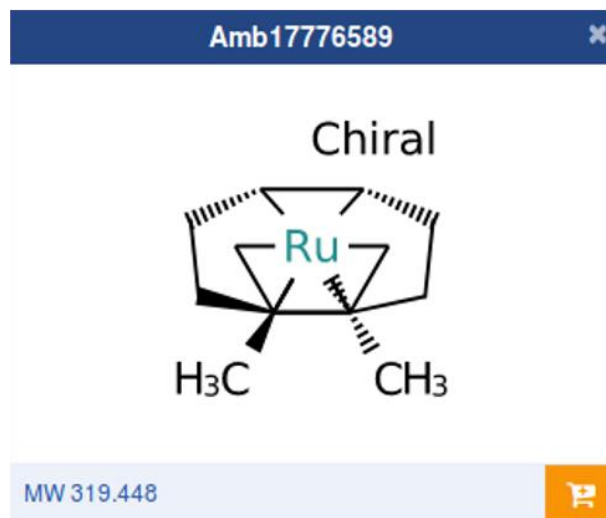


Figure 43. Exemples de molécules rencontrées dans le catalogue d'Ambinter. a) Mélange de molécules de méthane telles que proposées par un fournisseur. b) Molécule inorganique.

3.3 Application de VSPrep dans un projet d'amarrage moléculaire à visée cosmétique

Les produits naturels constituent depuis l'aube des temps une grande source d'inspiration pour l'Homme. L'exemple le plus ancien jamais trouvé concerne certainement la découverte en 1960 d'un squelette de Néandertalien dans la caverne de Shanidar IV¹⁵⁰ (Iraq). La particularité de cet endroit est qu'une importante concentration de pollens y a été trouvée, ce qui lui a d'ailleurs valu le sobriquet de « tombe aux fleurs ». Cependant, une analyse des pollens présents sur le sol a permis de mettre en évidence que toutes les plantes desquelles ils étaient issus possédaient des propriétés médicinales (diurétiques, stimulantes, astringentes). Ceci pourrait laisser entendre qu'elles étaient en fait employées comme remède, faisant alors remonter les origines de la médecine traditionnelle à plus de 40 000 ans avant J. C.

De plus, la découverte de morceaux d'ocre mélangés à des coquillages broyés dans la caverne de Blombos¹⁵¹ (Afrique du Sud), laisserait entendre que ces ingrédients pourraient avoir servi comme peintures pour le corps. Ceci placerait donc l'élaboration des premiers produits cosmétiques à pas moins de 100 000 ans avant J. C.

Il semble que l'Homme, et plus particulièrement le chimiste médicinal, ait toujours aujourd'hui un attrait marqué pour les produits naturels, comme en attestent les nombreuses études portant sur

¹⁵⁰ J. Lietava, Medicinal Plants in a Middle Paleolithic Grave Shanidar IV?, *Journal of Ethnopharmacology*, 35.3 (1992), 263–66.

¹⁵¹ I. Watts, Red Ochre, Body Painting and Language: Interpreting the Blombos Ochre, *The Cradle of Language*, https://www.academia.edu/1560509/Red_ochre_body_painting_and_language_interpreting_the_Blombos_ochre, consulté le 14 Novembre 2017.

l'identification de nouveaux inhibiteurs de protéines kinases issus de la Nature.^{152,153} Il faut dire que ces molécules présentent deux avantages importants. Tout d'abord, ces composés résultent d'une longue histoire de coévolution avec leur cible, ce qui leur confère souvent des éléments structuraux propres leur assurant une meilleure sélectivité.¹⁵⁴ Enfin, ils constituent une grande source de diversité. Ainsi, il a été estimé que 45% des médicaments *best sellers* dérivent en fait de produits naturels. De plus, il resterait encore aujourd'hui 40% d'échafaudages moléculaires présents dans la nature mais absents de la chimie médicinale.¹⁵⁵ Pour ces raisons, de nombreux outils ont été développés afin de mieux explorer la biodiversité¹⁵⁶ et de nombreuses sociétés s'intéressent à l'exploitation des produits naturels en cosmétique.

Parmi celles-ci, la société Greenpharma, basée sur le campus de l'Université d'Orléans, est impliquée dans la valorisation des produits naturels par leur utilisation dans les domaines de la santé et de la cosmétique. C'est grâce à une collaboration entre cette société et l'équipe de Bioinformatique Structurale et Chémoinformatique (SB&C), dirigée par le Pr. Pascal Bonnet, qu'a été initiée cette partie de mon projet de thèse. L'objectif de ce travail est d'identifier, par l'utilisation de méthodes *in silico*, des produits naturels présentant des activités contre le vieillissement de la peau en nous focalisant sur la réduction des taches brunes et des troubles pigmentaires en général.

Trois tâches m'ont été confiées afin de répondre à cette problématique.

Premièrement j'ai effectué un état de l'art sur diverses protéines impliquées dans la régulation de la mélanogenèse, dont l'une des cibles validées avec notre partenaire commercial, une protéine kinase, sera présentée brièvement. J'ai ensuite recensé des bases moléculaires de produits naturels publiques que j'ai dû préparer pour du criblage virtuel, préparé avec VSprep. Finalement, j'ai utilisé les résultats précédents pour réaliser l'amarrage moléculaire des molécules préparées dans les cibles identifiées.

3.3.1 Préparation de la protéine

3.3.1.1 Identification de protéines kinases d'intérêt en cosmétique

Dans le cadre de ce projet de recherche en cosmétique, nous nous intéressons au phénomène de vieillissement de la peau, et plus particulièrement au phénomène d'apparition de taches brunes. Celles-ci peuvent survenir après une exposition prolongée aux rayons UV (lentigo) ou des toxines bactériennes,

¹⁵² A. Marston, Natural Products as a Source of Protein Kinase Activators and Inhibitors, *Current Topics in Medicinal Chemistry*, 11.11 (2011), 1333–39

¹⁵³ X. Luo *et al.*, The Role of Targeting Kinase Activity by Natural Products in Cancer Chemoprevention and Chemotherapy, *Oncology Reports*, 34.2 (2015), 547–54

¹⁵⁴ T. Rodrigues *et al.*, Counting on Natural Products for Drug Design, *Nature Chemistry*, 8.6 (2016), nchem.2479

¹⁵⁵ L. Müller-Kuhrt, Putting Nature Back into Drug Discovery, *Nature Biotechnology*, 21.6 (2003), nbt0603-602–602

¹⁵⁶ Q.-T. Do *et al.*, How to Valorize Biodiversity? Lets Go Hashing, Extracting, Filtering, Mining, Fishing, *Planta Medica*, 81.6 (2015), 436–49

après des troubles hormonaux (masque de grossesse) ou naturellement en vieillissant (**Figure 44**) et peuvent s'accompagner d'un mal-être chez les personnes atteintes.



Figure 44. Apparition de taches brunes sur la peau lors du vieillissement (lentigo)¹⁵⁷

Cette pigmentation brune de la peau résulte d'un processus complexe¹⁵⁸ qui, dans l'épiderme comme dans les follicules pileux, débute avec la synthèse de la mélanine (pigment) par les mélanosomes (organites) au sein des mélanocytes (cellules). Les mélanosomes sont ensuite transférés aux kératinocytes environnants (estimés à 36 par mélanocyte en moyenne¹⁵⁹), qui vont transporter le pigment et éventuellement le dégrader. Chez l'Homme, les mélanocytes se situent au niveau des follicules pileux et dans l'assise basale de l'épiderme (**Figure 45**). La mélanine se décompose en deux pigments différents : l'eumélanine (pigment brun-noir) et la phéomélanine (pigment jaune-rouge), produits respectivement par des eumélanosomes et des phéomélanosomes. Son rôle biologique consiste à protéger la peau contre les effets néfastes des rayons UV, empêchant ainsi le développement de cancers de la peau.

Sa synthèse repose sur une régulation physiologique fine de la cellule et fait intervenir de nombreux acteurs (**Figure 46**). Parmi ceux-ci, les plus importants sont la protéine Tyr kinase KIT, l'hormone α -Melanocyte-Stimulating-Hormone (α -MSH), le facteur de transcription MITF (*Microphthalmia-Associated Transcription Factor*) ainsi que la Protéine Kinase C (PKC).¹⁶⁰ La mutation de gènes clés dans la mélanogenèse mène à sa dérégulation, aboutissant à différentes catégories de génodermatoses (maladies rares de la peau). Ainsi la dérégulation de l'activité de KIT conduit à la prolifération de mélanoblastes (précurseurs de mélanocytes) aboutissant à du piébalisme.^{161,162}

¹⁵⁷ V. Romano, Comment se déroulent les séances médicales de traitement par laser des tâches ?, *Dr Valeria Romano*, <https://www.dr-romano-esthetique.com/fr/laser/taches-pigmentaire-peelings>, consulté le 13 Novembre 2017

¹⁵⁸ M. Démarchez, Le mélanocyte et la pigmentation, <https://biologiedelapeau.fr>, 2017, <https://biologiedelapeau.fr/spip.php?article12>, consulté le 14 Novembre 2017

¹⁵⁹ M. Seiberg, Keratinocyte-Melanocyte Interactions during Melanosome Transfer, *Pigment Cell Research*, 14.4 (2001), 236–42

¹⁶⁰ S. A. N. D'Mello *et al.*, Signaling Pathways in Melanogenesis, *International Journal of Molecular Sciences*, 17.7 (2016)

¹⁶¹ I. Ferreira dos Santos Videira *et al.*, Mechanisms Regulating Melanogenesis, *Anais Brasileiros De Dermatologia*, 88.1 (2013), 76–83

¹⁶² M. Picardo *et al.*, The Genetic Determination of Skin Pigmentation: KITLG and the KITLG/c-Kit Pathway as Key Players in the Onset of Human Familial Pigmentary Diseases, *Journal of Investigative Dermatology*, 131.6 (2011), 1182–85

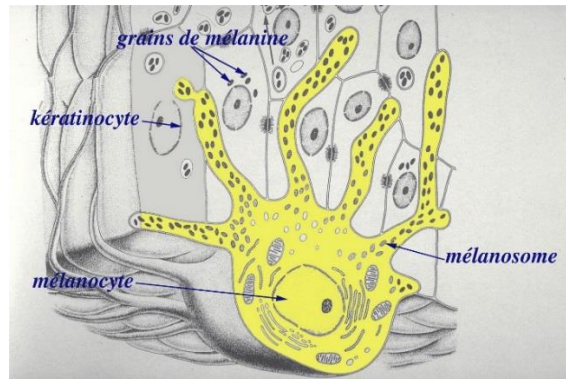


Figure 45 Représentation d'une coupe de peau humaine.¹⁶³

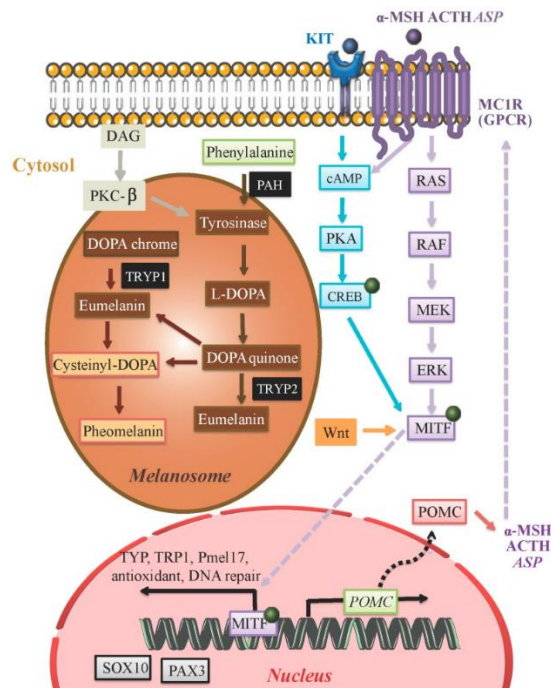


Figure 46. Principales voies de régulation de la mélanogénèse.¹⁶¹

Comme nous venons de le voir, la protéine kinase KIT joue un rôle primordial dans la mélanogénèse. Une discussion avec notre partenaire industriel nous a amené à considérer cette protéine kinase pour réaliser du criblage virtuel.

3.3.1.2 Création de modèles par homologie de la protéine kinase KIT

Ce travail a été accompli avec un étudiant de Master 2 que j'ai encadré, Jean-Thomas Heinrich.

¹⁶³ Melanoma Treatment - General Information About Melanoma, <https://my.pearlpoint.org/resources/melanoma-treatment-general-information-about-melanoma>, consulté le 13 Novembre 2017

Nous avons vu au chapitre 2 que la structure bilobée des domaines kinases leur confère une certaine flexibilité. Deux approches sont possibles afin de prendre cette dernière en compte lors des calculs d'amarrage moléculaire. Une première approche consiste à utiliser un grand nombre de structures tridimensionnelles d'une même protéine¹⁶⁴ (par exemple CDK2) comme cibles de l'amarrage. Cependant, cette approche peut être biaisée par les conditions de résolution de l'expérience (résolution, pH, etc.). Une autre méthode consiste à effectuer en amont une simulation de dynamique moléculaire de la protéine en forme apo et de sélectionner des conformations issues de la trajectoire obtenue.¹⁶⁵ Cependant ce type de simulation peut atteindre des temps de calculs considérables, surtout dans le cas où le modélisateur cherche à explorer l'espace conformationnel de la protéine en capturant un évènement rare comme le repliement d'une boucle.

Une solution hybride a été développée au sein de notre laboratoire. Elle se base sur une classification non supervisée (faite en interne) des structures cristallographiques des protéines kinases (publication en cours). Elle a notamment permis de mettre en évidence 3 groupes majoritaires (*clusters*) de conformations des domaines kinases en se basant sur des descripteurs tridimensionnels. Un quatrième cluster a également été identifié mais ce dernier étant sous-représenté, il a été ignoré dans le reste de l'étude. Chaque protéine kinase est donc représentée par 3 structures, issues de chacun des 3 *clusters*. Lorsqu'il n'y a pas de structure dans le *cluster* de la séquence correspondante, un modèle par homologie est construit à partir de la structure du *cluster* la plus proche en séquence. Nous espérons ainsi proposer un échantillonnage efficace de l'espace conformationnel du domaine kinase. C'est cette approche qui a été employée pour réaliser le criblage virtuel décrit dans la partie **3.3.3**.

Pour prendre en compte les 3 conformations représentatives de la protéine kinase KIT, nous avons utilisé les structures PDB 3G0E (chaîne A), 1PKG (chaîne B) et 1FGK (chaîne B, **Table 13**). Cette dernière n'est pas une structure de KIT mais de FGFR1. Elle a été identifiée dans l'approche de *clustering* structure/séquence car c'est une protéine homologue de KIT (40,7% d'identité de séquence¹⁶⁶ d'après l'outil Clustal Omega¹⁶⁷). L'objectif est donc de remplacer la séquence de FGFR1 par celle de KIT_(WT) tout en conservant la conformation de la structure de FGFR1.

La prise en compte des éventuelles molécules d'eau au sein du site actif est un autre paramètre à considérer pour le succès de l'amarrage moléculaire.^{168,169} Afin de déterminer si leur présence est nécessaire ou non, deux approches ont été employées. Premièrement, l'ensemble des structures de KIT

¹⁶⁴ S.-Y. Huang *et al.*, Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking, *Proteins: Structure, Function, and Bioinformatics*, 66.2 (2007), 399–421

¹⁶⁵ H. Alonso, *et al.*, Combining Docking and Molecular Dynamic Simulations in Drug Design, *Medicinal Research Reviews*, 26.5 (2006), 531–68

¹⁶⁶ D. J. Diller *et al.*, Homology Models, and High Throughput Docking, *J. Med. Chem.*, 46.22 (2003), 4638–47

¹⁶⁷ F. Sievers *et al.*, Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega, *Molecular Systems Biology*, 7 (2011), 539

¹⁶⁸ J. M. Hayes *et al.*, Kinetics, in Silico Docking, Molecular Dynamics, and MM-GBSA Binding Studies on Prototype Indirubins, KT5720, and Staurosporine as Phosphorylase Kinase ATP-Binding Site Inhibitors: The Role of Water Molecules Examined, *Proteins*, 79.3 (2011), 703–19

¹⁶⁹ M. Kaur *et al.*, Exploring the Role of Water Molecules for Docking and Receptor Guided 3D-QSAR Analysis of Naphthyridine Derivatives as Spleen Tyrosine Kinase (Syk) Inhibitors, *Journal of Chemical Information and Modeling*, 52.10 (2012), 2619–30

ont été superposées pour une inspection visuelle. Aucune molécule d'eau ne semblait fortement conservée au sein du site actif (**Figure 47**).

Table 13. Liste des structures PDB identifiées comme pouvant résumer les conformations majoritaires de KIT

Numéro de <i>cluster</i>	Code PDB	Résolution (Å)	Année	Classification DFG	Classification hélice α C
1	3G0E	1.60	2009	<i>out</i>	<i>in</i>
2	1PKG	2.90	2003	<i>in</i>	<i>in</i>
3	1FGK*	2.00	1997	**	<i>in</i>

* Structure de FGFR1 ; ** pas de classification possible pour MOE

Compte tenu du faible nombre de structures à ma disposition, j'ai alors employé l'algorithme 3DRISM¹⁷⁰, implémenté dans MOE⁷⁷ (version 2015), afin de prédire la présence de molécules d'eau prédites « non déplaçables » car elles contribueraient fortement à la stabilité du complexe. Si des molécules sont prédites comme fortement stabilisantes au niveau du site de fixation de l'ATP, il faut les inclure dans notre étude. Les résultats obtenus par l'outil de prédiction semblent indiquer qu'il n'y avait pas de molécules d'eau stabilisante (**Figure 48**) et nous avons pris la décision ne pas les conserver lors de la génération des modèles par homologie.

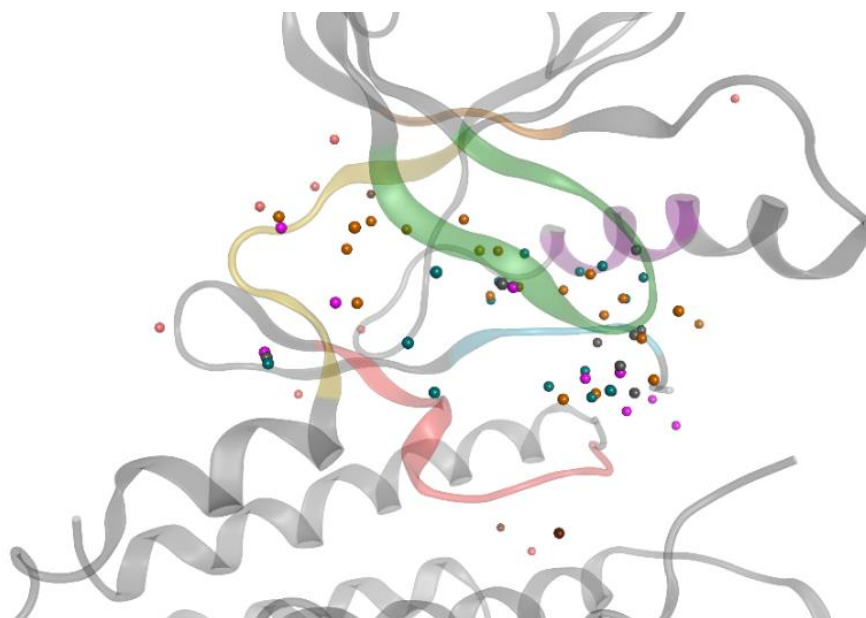


Figure 47. Alignement des molécules d'eau présentes dans les structures des domaines kinases de KIT disponibles dans la PDB. Codes PDB : 1PKG (chaîne A en jaune), 1T45 (orange), 3G0E (bleu), 3GOF (chaîne A en rose, chaîne B en marron), 4HVS (gris), 4UOI (violet).

¹⁷⁰ National Center for Biotechnology Information (NCBI), Internet Bethesda (MD): National Library of Medicine (US), 1988, <https://www.ncbi.nlm.nih.gov>

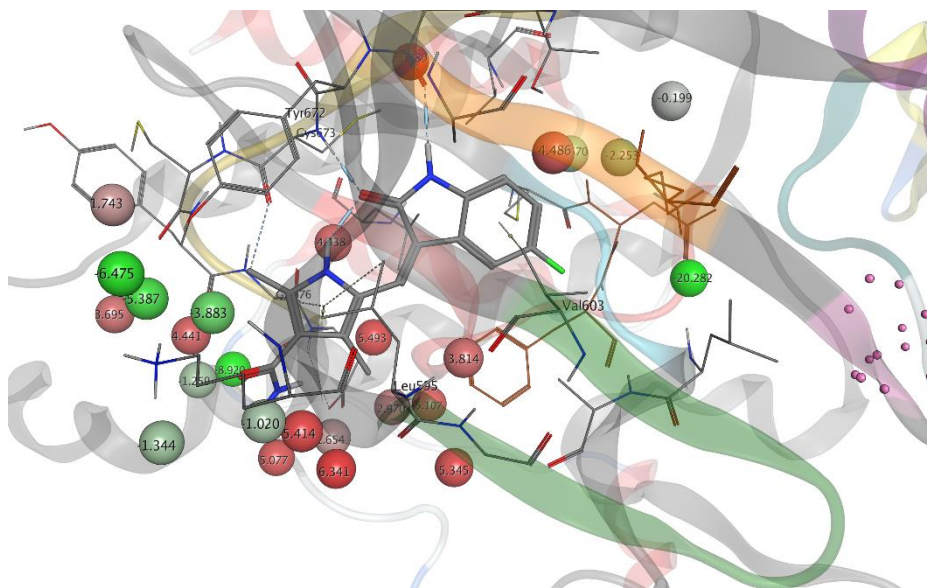


Figure 48. Estimation de la probabilité de présence de molécules d'eau déplaçables avec 3DRISM (code PDB : 3G0E). Les sphères indiquent l'emplacement possible des molécules d'eau. Celles-ci sont coloriées par un gradient en fonction de leur contribution à la stabilisation du système (rouge : déstabilisantes, vert : stabilisantes).

Un autre facteur important à considérer pour la préparation du récepteur est l'intégrité de la structure tridimensionnelle. En effet, plus il y a de régions non résolues et plus il sera difficile de combler les « trous » (

Table 14) et d'obtenir une structure complète. Malheureusement, KIT possède un domaine *Kinase Insert Domain* (KID) d'une longueur de 77 acides aminés qui n'est résolu ni dans 1PKG (non résolu, **Figure 49**), ni dans 3G0E (mutation délétère), ni dans aucune structure connue de protéine. Aucune séquence homologue n'ayant pu être identifiée par *blastp*¹⁷¹ sur le site NCBI¹⁷², nous avons omis ce domaine de notre étude. En effet, la reconstruction d'une si grande séquence sans modèle sur lequel se baser est plus susceptible de fausser les résultats que de les améliorer. Les autres parties non résolues, bien plus petites, sont quant à elles comblées lors de la génération des modèles par homologie.

¹⁷¹ S. F. Altschul *et al.*, Basic Local Alignment Search Tool, *Journal of Molecular Biology*, 215.3 (1990), 403–10

¹⁷² National Center for Biotechnology Information (NCBI), Internet Bethesda (MD): National Library of Medicine (US), 1988, <https://www.ncbi.nlm.nih.gov>

Table 14. Régions non résolues sur les séquences des domaines kinases de KIT (589 – 937) et FGFR1 (478-767).

Code PDB (Chaîne)	3GOE (A)	1PKG (B)	1FGK (B)
Région non résolue*	*	691-764	501-504
	931-937	817-830**	581-592
		929-937	

* Mutation délétère du domaine KID; ** Mutation proche du site actif

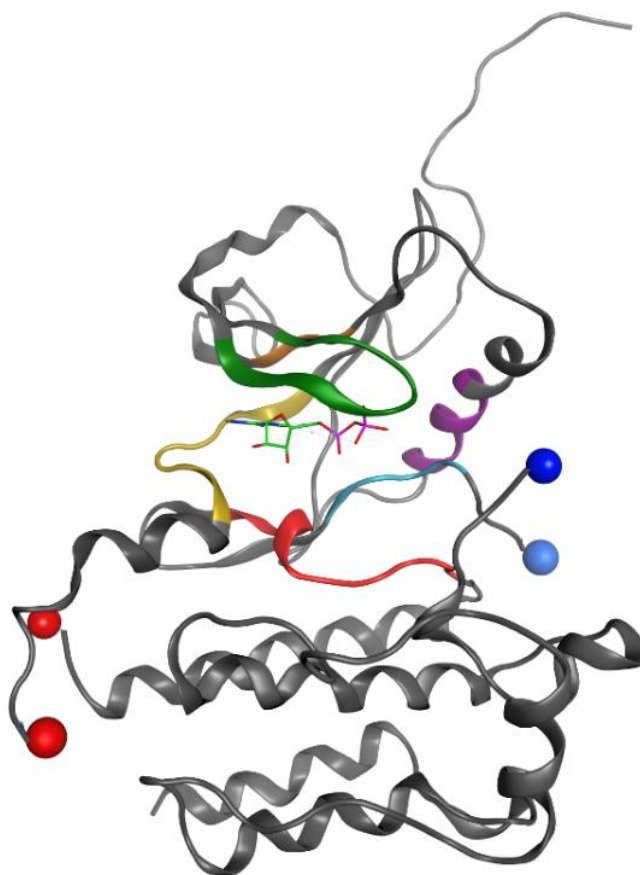


Figure 49. Domaine kinase de KIT en conformation active (code PDB : 1PKG). Les paires colorées de sphères indiquent les limites des deux parties non résolues de la structure (en bleu : KID).

Le logiciel Modeller¹⁷³ (version 9.16) a été employé afin de construire les modèles par homologie. Pour ce faire, il se base sur une séquence de référence et sur une structure de départ. Cette dernière est alors modifiée (en général substitutions et additions) afin d'arborer la même séquence d'acides aminés

¹⁷³ A. Sali *et al.*, Comparative Protein Modelling by Satisfaction of Spatial Restraints, *Journal of Molecular Biology*, 234.3 (1993), 779–815

que la séquence de référence, tout en conservant la conformation de départ. L'alignement entre la séquence de référence et la séquence de la structure est donc primordial. La séquence native de KIT (P10721) a été extraite du site Uniprot¹⁷⁴ et a servi de séquence de référence afin de générer des modèles sauvages de KIT (**Figure 50**).

```
>P1;P10721
sequence:P10721::::::::::
LSFGKTLGAGAFGKVV EATAYGLIKSDAAMTVAVKMLKPSAHLTEREALMSELKVLSYLG
NHMNIVNLLGACTIGGPTLVITEYCCYGDLLNFLRRKR/ALDLEDLLSFYSYQVAKGMAFL
ASKNCIHRDLAARNILLTHGRITKICDFGLARDIKNDSNYVVKG NARLPVKWMAPE SIFN
CVYTFESDVWSYGI FLWELFSLGSSPYPGMPVDSKFYKMIKEGFRMLSPEHAPAEMYDIM
KTCWDADPLKRPTFKQIVQLIEKQISESTNHIYS*

>P1;FGFR1_HUMAN_1FGK-B_aligned.pdb
structureX:FGFR1_HUMAN_1FGK-B_aligned.pdb:478:B:761:B::-1.00:-1.00
LVLGKPLGEGAFGQVVLAE AIGLPNR--VTKVAVKMLKSDATEKDLSDLISEMEMMKMIG
KHKNIINLLGACTQDGPLYVIVEYASKGNLREYLQARR/QLSSKDLVSCAYQVARGMEYL
ASKKCIHRDLAARNVLTEDNVMKIADFG LADYYK-----KGRLPVKWM APEALFD
RIYTHQSDVWSFGVLLWEIFTLGGSPYPGPVPE-ELFKLLKEGHRMDKPSNCTNELYMM
RDCWHAVPSQRPTFKQLVEDLDRIVALT-----*
```

Figure 50. Alignement de la séquence de FGFR1 (1FGK) avec celle de KIT (code Uniprot: P10721)

L'utilisateur fixe lui-même les parties fixes du modèle, c'est-à-dire qui ne seront pas échantillonnées pour minimiser les structures tridimensionnelles formées au cours des calculs. J'ai laissé un tampon de 5 résidus de part et d'autre de chaque partie manquante afin de permettre à l'outil d'optimiser la géométrie des attaches de la partie reconstituée et d'ainsi générer des conformations basses en énergie. De plus, sa très grande taille le place hors de portée des approches de reconstruction *de novo* de Modeller, limité en général au maximum à 10-15 acides aminés manquants.

Plusieurs modèles sont générés par l'outil Modeller à partir des informations fournies par l'utilisateur. Le logiciel les trie selon le *DOPE Score* (*Discrete Optimized Protein Energy Score*), censé évaluer la qualité locale des modèles générés (**Figure 51**). Plus ce score est faible, meilleur est le modèle. Ce score étant système dépendant, nous sélectionnons le modèle avec le score le plus bas pour inspection visuelle. Si la structure paraît raisonnable, alors nous faisons appel à des métriques proposées par d'autres outils disponibles en ligne afin de compléter notre analyse.

¹⁷⁴ UniProt: The Universal Protein Knowledgebase, *Nucleic Acids Research*, 45.D1 (2017), D158–69

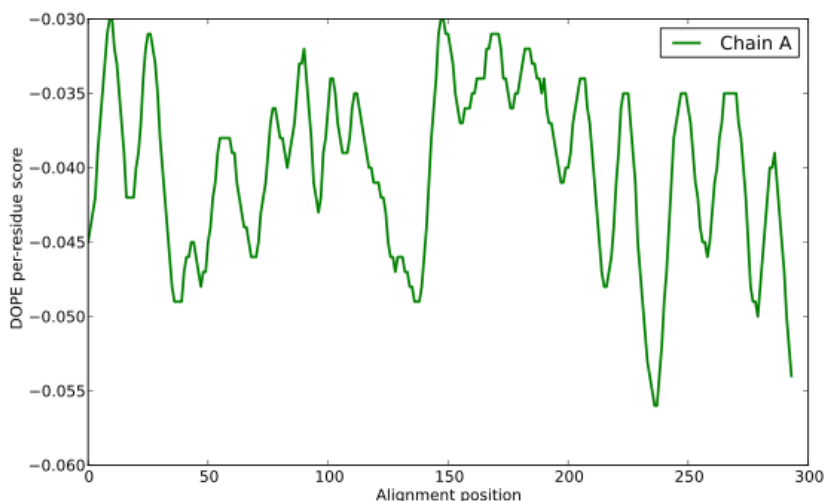


Figure 51. Profile DOPE d'un modèle par homologie (1PKG).

Parmi ces derniers, QMEAN¹⁷⁵ compare le modèle généré avec plusieurs centaines de structures cristallographiques de référence pour lui attribuer un score (QMEAN Score) en prenant en compte plusieurs caractéristiques structurales globales et locales. Il fournit également un z-score (écart par rapport à la valeur moyenne) normalisé. Plus le QMEAN Score est grand, meilleur est le modèle, tandis que plus le z-score augmente, plus on s'écarte des bonnes structures de référence (**Figure 52**). RAMPAGE¹⁷⁶ permet quant à lui de produire un diagramme de Ramachandran, qui permet de mettre en évidence les résidus présentant une conformation erronée (**Figure 53**).

Dans la suite de l'étude, les modèles par homologie effectués à partir des structures 3G0E, 1PKG et 1FGK seront respectivement dénommés cluster 1, cluster 2 et cluster 3.

¹⁷⁵ P. Benkert *et al.*, QMEAN: A Comprehensive Scoring Function for Model Quality Assessment, *Proteins*, 71.1 (2008), 261–77

¹⁷⁶ RAMPAGE: Ramachandran Plot Assessment, <http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>, consulté le 13 Novembre 2017.

Les résultats de ces évaluations sont résumés dans la **Table 15**. Le QMEAN Score varie entre 0 et 1 et plus sa valeur est grande, meilleur est le modèle. Ainsi, la qualité des modèles obtenus semble satisfaisante pour notre étude. De plus, la **Figure 52 C** nous permet de vérifier quel aspect des modèles s'écarte des structures de référence. Dans notre cas, la poche apparaît nettement moins solvatée que des poches semblables dans les structures de référence, ce qui est en accord avec notre décision de ne pas considérer les molécules d'eau. De plus, la projection de deux acides aminés en dehors des zones attendues sur le diagramme de Ramachandran indique un problème structural. Cependant, ces deux acides aminés sont éloignés du site actif (**Figure 54**) et ne devraient pas avoir d'impact sur les résultats (distance > 18 Å).

Table 15. Résultats de l'évaluation de la qualité des modèles par homologie générés

Cluster	1	2	3
QMEAN score	0.85	0.82	0.76
QMEAN z-score	0,80	0,54	- 0,15
Validité de la géométrie du modèle (Ramachandran)	oui	oui	oui*

* 2 acides aminés présentent des valeurs marginales : Leu611 et Ala629

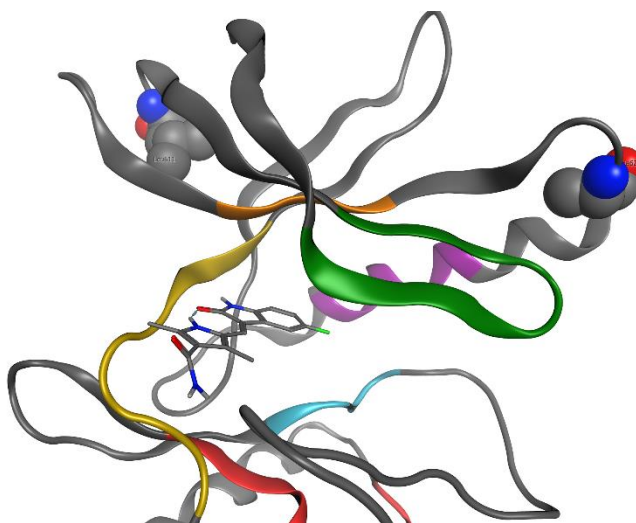


Figure 54. Position des deux résidus marginaux selon le diagramme de Ramachandran dans le cluster 3. Les acides aminés concernés sont représentés en sphères (rayons de van Der Waals). Le ligand co-cristallisé du cluster 1 a été aligné avec cette structure pour mieux visualiser à quelle distance ils se trouvent du site actif (Leu611 : 18 Å, Ala629 : 23 Å)

3.3.2 Conception et préparation d'une base moléculaire de produits naturels

Au début de mes travaux, je me suis contenté de récupérer le plus grand nombre de catalogues possibles. Je me suis cependant vite rendu compte que mon jeu de données contenait plusieurs millions de molécules et que je n'avais pas la moindre idée de quelles molécules étaient vraiment naturelles. En effet, il arrive fréquemment que des catalogues soient qualifiés de naturels alors qu'ils contiennent également des produits dérivés (hémi-synthèse) ou bien synthétiques.

J'ai alors pris le parti de recenser les bases de données dont l'information disponible permettait de relier la molécule à l'organisme source, aboutissant à 7 bases moléculaires publiques : BioPhytMol¹⁷⁷, KNApSack-3D¹⁷⁸, NuBBE¹⁷⁹, SANCDB¹⁸⁰, StreptomeDB^{181,182}, TCM¹⁸³, TM-TC¹⁸⁴). J'ai également pu bénéficier de la collaboration de Greenpharma, qui m'a donné la liste des produits naturels inclus dans la base Ambinter que j'ai pu ainsi extraire de leur site (Ambinter-NP). La base moléculaire NuBBE détaillant l'origine de ses composés, j'ai pu en retirer quelques molécules synthétiques. J'ai ensuite rassemblé les structures dans une seule base moléculaire au format SDF, « NPDB2 » (**Table 16**). Afin d'éviter tout risque de confusion, on appellera NPDB2 la base fournie en entrée à l'outil et « NPDB2prep » le résultat produit.

Comme la base NPDB2 a été assemblée depuis plusieurs origines différentes, il y a un fort risque de présence de doublons. J'ai donc utilisé les fonctionnalités de VSPrep pour supprimer les entrées redondantes tout en gardant l'information des doublons (voir partie 3.2.3.4). Pour ce faire, seules les étapes de lecture, standardisation, filtre et suppression des doublons sont effectuées. De plus, l'étape d'énumération du tautomère canonique n'as pas été effectuée en raison du temps de calculs exponentiel pour certains composés. Une étude plus minutieuse sur le comportement de MolVS semble donc nécessaire pour pouvoir prendre en compte ce genre d'erreurs qui sont très pénalisantes sur des bases moléculaires volumineuses.

¹⁷⁷ A. Sharma *et al.*, BioPhytMol: A Drug Discovery Community Resource on Anti-Mycobacterial Phytomolecules and Plant Extracts, *Journal of Cheminformatics*, 6.1 (2014), 46

¹⁷⁸ K. Nakamura *et al.*, KNApSack-3D: A Three-Dimensional Structure Database of Plant Metabolites, *Plant & Cell Physiology*, 54.2 (2013), e4

¹⁷⁹ M. Valli *et al.*, Development of a Natural Products Database from the Biodiversity of Brazil, *Journal of Natural Products*, 76.3 (2013), 439–44

¹⁸⁰ R. Hatherley *et al.*, SANCDB: A South African Natural Compound Database, *Journal of Cheminformatics*, 7 (2015), 29

¹⁸¹ X. Lucas *et al.*, StreptomeDB: A Resource for Natural Compounds Isolated from Streptomyces Species, *Nucleic Acids Research*, 41.Database issue (2013), D1130-1136

¹⁸² D. Klementz *et al.*, StreptomeDB 2.0-an Extended Resource of Natural Products Produced by Streptomyces, *Nucleic Acids Research*, 44.D1 (2016), D509-514

¹⁸³ C. Y.-C. Chen, TCM Database@Taiwan: The Worlds Largest Traditional Chinese Medicine Database for Drug Screening In Silico, *PLoS ONE*, 6.1 (2011), e15939

¹⁸⁴ R. Xue *et al.*, TCMID: Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis, *Nucleic Acids Research*, 41.Database issue (2013), D1089-1095

En effet, ce projet étant une collaboration avec l'entreprise Greenpharma, nous avons choisi de réduire l'espace chimique des composés à ceux qui sont présents dans le catalogue d'Ambinter afin de pouvoir commander les éventuelles touches du criblage plus facilement. Nous avons donc suivi le même protocole de préparation que pour NPDB2, ce qui n'a demandé que 24h de temps machine (5,7M de molécules, soit environ $5,7.106 / (24 \times 3600) = 65,9$ entrées/s).

Table 16. Composition de NPDB2 avant préparation

Localisation	Année	Nom de la base de données	Organismes sources	Nombre d'entrées*
Mondial	2014	BioPhytMol	plantes	633
Brésil	2013	NuBBE-NP	plantes, champignons	881
Mondial	2013	StreptomeDB	bactéries	3991
Afrique du Sud	2015	SANCDDB	plantes, microorganismes marins	623
Mondial	2012	KNAPSAck-3D	plantes, champignons, microorganismes	50036
Chine	2011	<u>TCM Database@Taiwan</u>	plantes, animaux, minéraux	60556
Asie du Nord-Est	2015	TM-MC	plantes	5697
Mondial	2011*	Ambinter-NP	plantes, champignons, microorganismes	2294

*Date de téléchargement : Avril 2016

**Date de rachat d'Ambinter par Greenpharma

Enfin, les deux bases préparées ont été regroupées en utilisant l'InChI-Key et l'intersection obtenue, NPDB3, représente 0,07% du catalogue d'Ambinter. Elle est soumise au reste du protocole VSPrep (énumération des tautomères, création de sous-librairies, énumération des stéréoisomères, génération de coordonnées 3D).

A la fin de cette préparation, la base moléculaire la plus représentée est bien entendu Ambinter- NP mais on observe que toutes les bases de données moléculaires d'origine sont représentées dans les résultats (**Figure 56**).

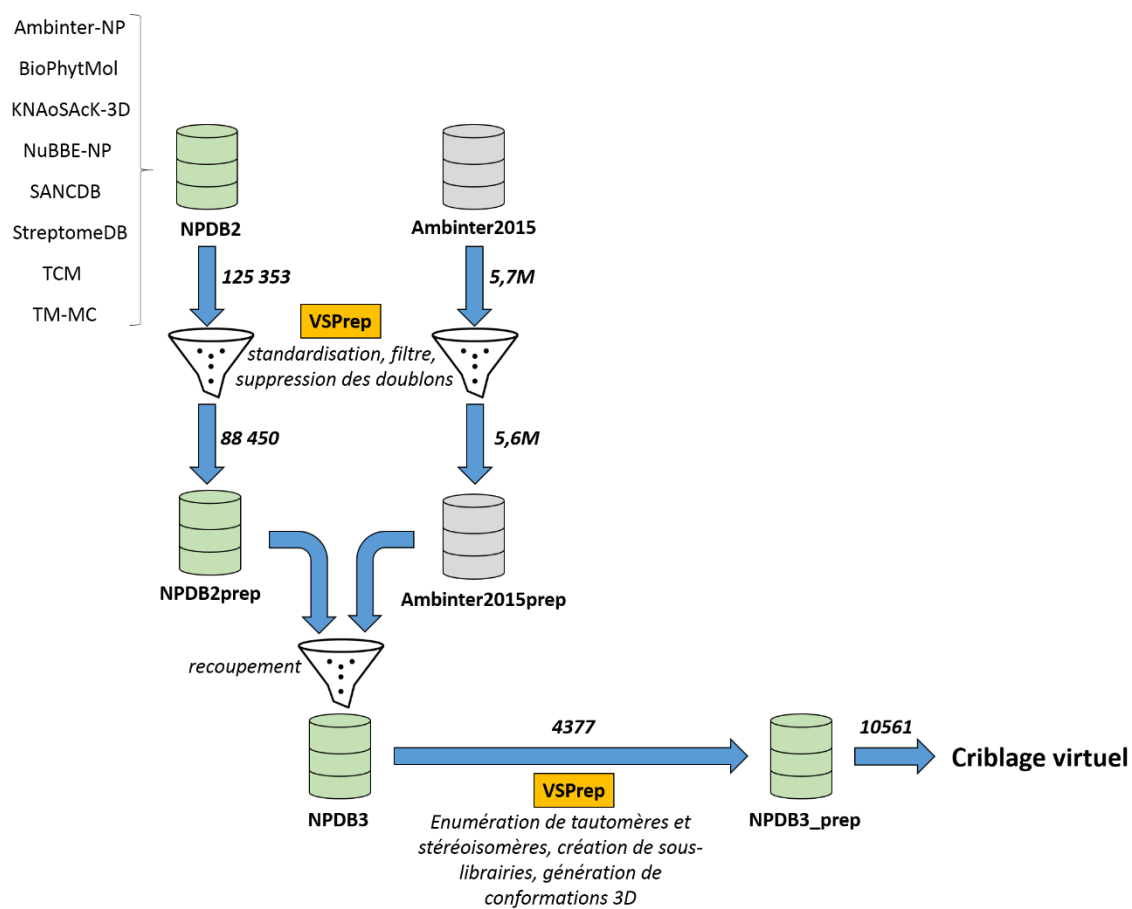


Figure 55. Stratégie pour la préparation du jeu de données de ligands à cribler

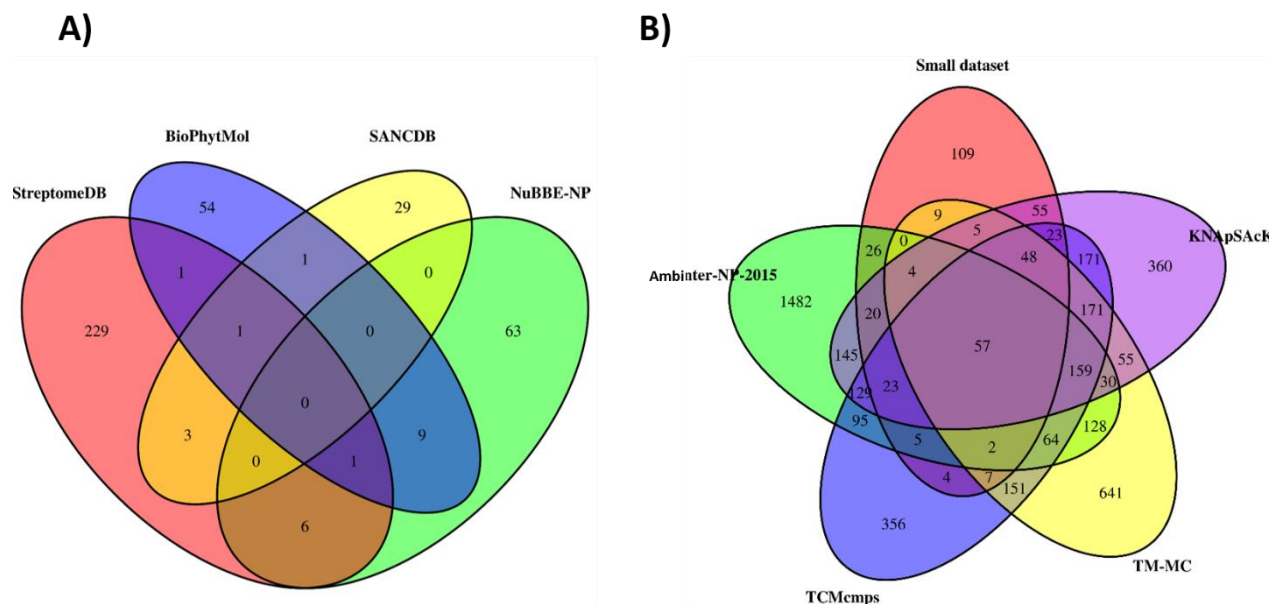


Figure 56. Recouvrement des bases de données naturelles incluses dans la librairie NPDB3. Pour un souci de clarté, l'information a été découpée en deux diagrammes de Venn. A) Recouvrement des bases moléculaires présentant le moins d'entrées. B) Recouvrement des bases moléculaires restantes et le sous-ensemble de la figure (A).

3.3.3 Amarrage moléculaire et résultats

3.3.3.1 Outil utilisé

L'amarrage de la base moléculaire NPDB3 a été réalisé avec l'outil GOLD.⁷³ Cet outil repose sur un algorithme génétique (GA), c'est-à-dire qu'il s'inspire de la théorie de l'évolution afin de générer des poses de ligands dans le site actif du récepteur considéré. Ici, un ligand est placé de manière aléatoire en de nombreuses occurrences au sein du site actif afin de le peupler de solutions possibles de poses. Chaque individu contient un chromosome, qui encode des informations sur les liaisons hydrogène et interactions hydrophobes possibles. A chaque chromosome est assigné un score de complémentarité (*fitness*) avec le récepteur. Une nouvelle génération d'individus est ensuite formée de manière itérative, afin de privilégier les solutions filles offrant les meilleurs scores. Un chromosome peut également subir substitution au lieu de se reproduire.

L'outil propose plusieurs scores pour estimer l'énergie de liaison protéine-ligand. Le GoldScore est le score historique de GOLD et consiste à prédire les poses du ligand en prenant en compte différents facteurs énergétiques (liaisons hydrogène, interaction métallique, etc.) Le ChemScore consiste quant à lui à incorporer un terme supplémentaire (ΔG) pour estimer le changement d'énergie libre qui se produit lorsque le ligand se fixe au récepteur (modèle d'apprentissage). Le score ASP est empirique et repose sur un potentiel de distance atome-atome dérivé de structures tridimensionnelles de complexes protéine-ligand. Le score ChemPLP utilise quant à lui des termes de van der Waals et de répulsion.

3.3.3.2 Paramétrage de l'amarrage moléculaire

Compte tenu des paramètres des fonctions de score de GOLD décrits plus haut, nous avons choisi d'employer le score ASP dans notre étude, combiné à un protocole utilisant les paramètres par défaut de GOLD. La cavité est définie par un ligand co-cristallisé. Quand ce dernier n'est pas disponible (comme par exemple dans les structures du cluster 3), le ligand d'une autre structure de la protéine kinase est aligné pour le positionner dans le site actif (par exemple, le ligand du cluster 1 positionné dans la structure du cluster 3). Les résultats montrent un excellent alignement des poses cristallographiques avec les poses amarrées du ligand, nous permettant donc de valider notre protocole (**Figure 57**).

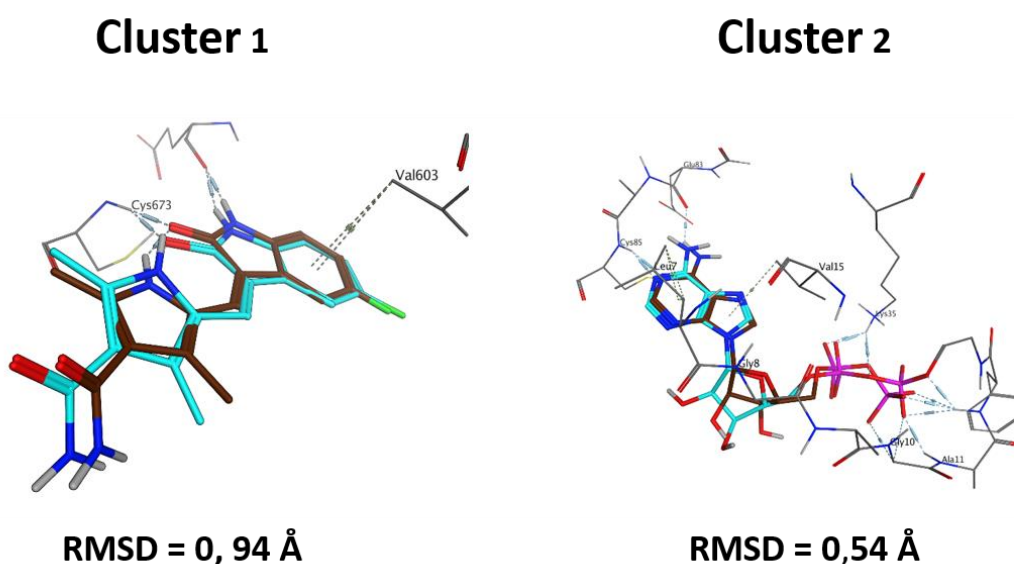


Figure 57. Validation du protocole d'amarrage moléculaire pour les clusters 1 et 2.

Maintenant que l'on sait que l'outil est capable de générer des poses de bonne qualité sur notre système, il peut se révéler judicieux d'étudier comment il parvient à discriminer des molécules actives de molécules inactives.

La courbe ROC est une méthode visuelle d'évaluation de la performance d'un classificateur prédictif binaire (oui/non). Elle représente le taux de bonnes prédictions (Vrai Positifs ou VP) en fonction du taux de mauvaises prédictions (Faux Positifs ou FP). L'aire sous la courbe (AUC) est une mesure comprise entre 0 et 1 qui permet de quantifier cette performance. Ainsi, une valeur de 1 indique un classificateur parfait bon (aucune erreur) tandis qu'une valeur de 0,5 indique un modèle prédictif aléatoire et qu'une valeur de 0 détermine un modèle parfaitement mauvais (que des erreurs).

Il s'agit d'une métrique très appliquée dans le cadre du criblage virtuel, où le chimoinformaticien essaie de prédire les molécules actives en fonction d'un score. Pour ce faire, il doit choisir un seuil de score qui permet d'identifier un ratio maximum de VP/FP.

Dans notre cas, nous devons évaluer si le score fourni par GOLD nous permet d'identifier correctement des molécules bioactives sur notre cible, la protéine kinase KIT, parmi un groupe de molécules inactives. Pour ce faire, nous recherchons sur la base ChEMBL⁹ les données d'activité disponibles sur KIT (ChEMBL1936). Les quelques 723 mesures de constante d'inhibition K_i disponibles sont filtrées par un seuil d'activité couramment employé en *kinase research*, 1 μ M, ce qui nous donne 163 molécules que nous qualifierons de composés actifs (VP).

Il nous faut maintenant nous procurer des structures de molécules inactives sur KIT. Pour cela, nous avons utilisé l'outil en ligne de la base DUD-E¹⁸⁵ qui recense plus de 10 000 molécules considérées inactives sur KIT. Parce que les structures de ces molécules sont proches des actifs d'un point de vue physico-chimique mais dissimilaires d'un point de vue structural, elles sont appelées leurres (*decoys*). Il sera ainsi plus dur pour l'outil de criblage de séparer les deux ensembles, ce qui permettra de mettre en avant les bons modèles de classification. Nous avons choisi de ne sélectionner que 1000 *decoys* de manière aléatoire afin de réduire les temps de calculs, obtenant ainsi un ratio *decoys*/actifs de 6 inférieur à celui des études publiées récemment.¹⁸⁵

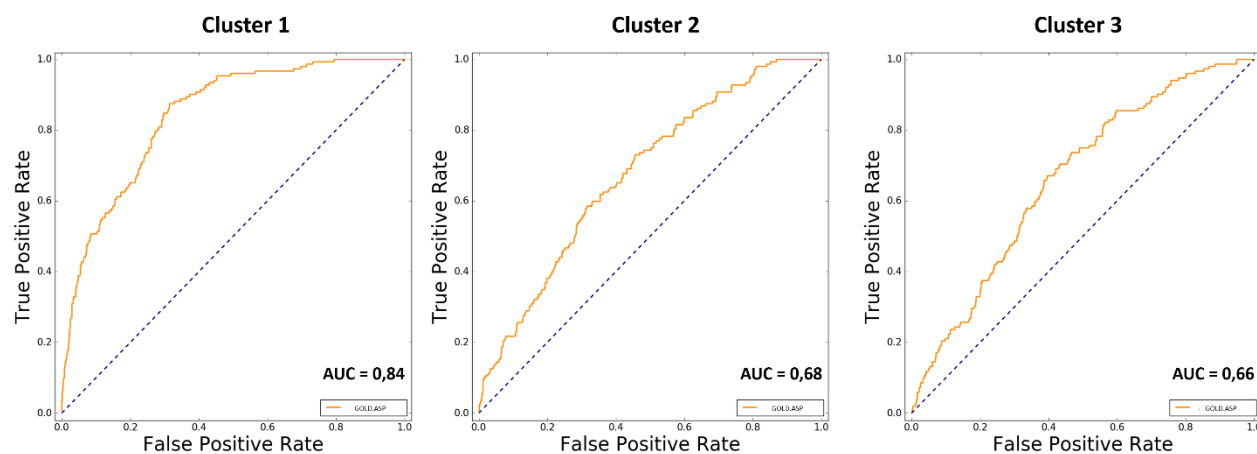


Figure 58. Courbes ROC obtenues pour les clusters 1, 2 et 3.

3.3.3.3 Résultats

La discrimination actifs/inactifs ne semble bien fonctionner que pour le cluster 1 (AUC = 0,84), ce qui semble montrer que le protocole ne permet pas de séparer correctement les molécules actives des leurres dans 2 conformations sur 3 (*clusters* 2 et 3). Ces résultats ne semblent pas provenir d'erreurs dans les modèles par homologie des clusters 2 et 3, ces modèles ayant été validés de la même manière que le modèle du cluster 1.

Une raison pouvant expliquer ce résultat provient du mode d'interaction des molécules actives utilisées dans ce criblage. En effet, il faudrait évaluer à quel type d'inhibiteurs de protéine kinase

¹⁸⁵ J. Xia *et al.*, Benchmarking Methods and Data Sets for Ligand Enrichment Assessment in Virtual Screening, *Methods (San Diego, Calif.)*, 0 (2015), 146–57

appartiennent les molécules actives référencées dans la base DUD-E. Si ce sont majoritairement des types I, alors ces molécules ciblent uniquement la conformation active de KIT (cluster 1, *DFG-in*, *α C-helix in*), ce qui expliquerait pourquoi nous obtenons une bonne valeur d'AUC dans ce cas et pas dans les autres.

3.4 Conclusion et perspectives

Un protocole de préparation de petites molécules a été développé tout au long de mes travaux de thèse. Le temps considérable consacré à l'élaboration de cet outil s'explique en partie par la plateforme de développement KNIME, dont l'apprentissage et les nombreuses mises à jour ont plus d'une fois nécessité d'actualiser le *workflow* entier, certains nœuds devenant obsolètes et leur remplacement ne constituant pas forcément une tâche anodine. L'autre raison qui justifie le temps alloué à ce travail est la nature même de ces travaux. En effet, le développement d'un tel protocole, composé uniquement d'outils gratuits pour les académiques, a soulevé des difficultés aussi nombreuses qu'imprévisibles, telles que le traitement d'un fichier moléculaire présentant un problème nouveau ou encore l'apparition de bogues au fur des mises à jour de nœuds par la communauté scientifique. La plupart de ces bogues ont pu être résolus en faisant remonter le problème auprès de la communauté scientifique mais certains demeurent, d'où la nécessité de développer rapidement des mises à jour. Il n'a en effet pas toujours été possible d'attendre que le bogue soit corrigé, surtout dans le cadre de travaux ayant un impact sur d'autres projets de recherche. En outre, ce protocole est déjà employé au sein de l'équipe SB&C de l'ICOA depuis plus de deux ans, ce qui a permis de le perfectionner au gré des retours des membres de l'équipe impliqués dans le projet.

L'outil présente néanmoins des certaines limites. Tout d'abord, l'utilisation de ChemAxon, faute d'alternative pertinente lors du développement initial, prévient son application dans le monde industriel pour les entreprises n'ayant pas de licence valide et remet donc en cause l'aspect universel recherché. Il est cependant important de noter que de nombreux outils académiques sont toujours développés grâce aux efforts de la communauté scientifique, cette situation est donc amenée à changer au cours des prochaines années.

Une autre limite de l'outil provient de son implémentation dans le logiciel KNIME. De prime abord estimé rapide, son développement s'est vu ralentir considérablement lorsque des mises à jour ont nécessité la réécriture de parties entières du *workflow*. Des nœuds dédiés à la maintenance, jusqu'à alors commerciaux, ont récemment été mis librement à la disposition des utilisateurs, ce qui permettra de simplifier le travail requis pour maintenir le bon fonctionnement du protocole.

Pour pallier ces limitations, nous avons étudié la possibilité d'implémenter ce *workflow* dans du code Python. Cela représenterait en effet des avantages notables, malgré le coût en temps de redéveloppement que cela représenterait. Reconnu par la communauté informatique comme l'un des langages informatiques les plus adaptés au calcul scientifique, grâce au développement de bibliothèques extrêmement optimisées, le langage Python dispose en effet de nombreux atouts. Considéré comme facile d'écriture, par rapport à d'autres langages comme le C ou le JAVA, il permet de réaliser rapidement des prototypes, dont il est ensuite possible d'optimiser significativement les performances grâce à des bibliothèques optimisées pour le *Big Data*. De plus, il est parfaitement interfacé avec RDKit et avec des systèmes

de bases de données tels que le postgresql, qui permet notamment l'indexation de l'information chimique et accroîtrait grandement la rapidité des requêtes SQL.

Enfin, ce que nous avons perçu comme l'une des limitations de l'outil KNIME provient de sa conception même. En effet, à l'inverse d'outils de *pipelining* tels que PipelinePilot, les protocoles sont réalisés au fur et à mesure, ce qui veut dire que chaque nœud à exécuter doit parcourir l'ensemble de la table donnée en entrée. Ceci se traduit non seulement par une consommation de mémoire importante, vu que toute la table doit être chargée en même temps, mais également par une perte de rapidité puisque la table est sans cesse parcourue par chacun des nœuds.

Nous avons en outre identifié une librairie Python, LUIGI, développé en 2013, qui nous permettrait de résoudre ces deux problèmes d'un coup. Cette librairie a été développée par la société Spotify, qui propose l'écoute de musiques en ligne, afin de générer des modèles d'apprentissage en temps réel sur les goûts musicaux des auditeurs. Cette librairie a donc été conçue pour traiter du big data et nous permettrait de nous focaliser sur trois points qui nous semblent primordiaux :

1. Traitement de chaque donnée (molécule) de manière individuelle, donc parallélisation optimale possible, voire développement d'une architecture pour tirer profit des clusters haute performance
2. Réalisation automatique de fichiers de log avec plusieurs niveaux d'information mais également possibilité de reprise d'une étape interrompue, par exemple en cas d'erreur ou bien de panne
3. Unification du code source vers le Python, ce qui se traduit par un véritable gain en lisibilité et a fortiori en maintenabilité

De plus, le protocole d'amarrage moléculaire présenté pourrait être affiné, par exemple en vérifiant l'impact de la prise en compte des molécules d'eau sur les solutions trouvées. De même, une étude comparative de différents scores pourrait permettre d'améliorer les résultats en favorisant les meilleures solutions lors du tri des poses (par score). Enfin, il semble nécessaire de classifier les inhibiteurs par chémotype afin de sélectionner des actifs qui puissent se lier sur la conformation ciblée. Il est à noter que le cluster 3 présente un motif DFG intermédiaire entre les positions *in* et *out*, ce qui soulève la question du type d'inhibiteur à employer. En effet, ce cas de figure n'étant pas décrit dans les modes d'interaction des inhibiteurs exposés dans le chapitre 2.

3.5 Annexes

Annexe 1. Capture d'écran de VSPrep

L'outil VSPrep est un workflow KNIME qui se découpe en trois parties principales.

Tout d'abord, en rouge, se trouve l'étape de configuration du protocole. Trois saisies sont nécessaires de la part de l'utilisateur. Le premier nœud sert de chemin d'entrée au fichier moléculaire à préparer, le second au fichier de configuration fourni par l'utilisateur et le dernier permet de spécifier s'il s'agit d'un nouveau travail ou la reprise d'un précédent qui aurait été interrompu. La seconde partie du workflow représente le cœur du projet et contient toutes les opérations chémoinformatiques appliquées aux molécules. Chaque nœud correspond à une étape et le premier permet la lecture des molécules sous forme de tableau KNIME. Les autres nœuds sont optionnels mais il est vivement recommandé de tous les employer. Chacun de ces nœuds possède une entrée et trois sorties. L'entrée correspond à la première sortie du nœud précédent et contient les molécules à traiter. La seconde sortie contient les alertes concernant les entrées et la dernière sortie les alertes concernant les molécules. Lorsqu'une sortie ne contient aucune donnée, elle est marquée par une croix rouge, ce qui permet de rapidement identifier les étapes ayant provoqué des alertes.

La dernière étape du protocole consiste en l'analyse et l'archivage des résultats. Elle permet la génération des graphes présentés dans l'annexe I et peut compresser le ou les fichiers de sortie du protocole.

Annexe 2. Rapport graphique : nombre de molécules et d'entrées à chaque étape

Une même molécule peut être représentée par plusieurs entrées au sein d'une même base de données moléculaire en tant tautomères, états de protonation, stéréoisomères et conformations. Ce graphe aide à évaluer la proportion entrées/molécules (respectivement en jaune et en orange). Avant l'étape de suppression des doublons, le protocole ne fait pas la différence entre les entrées et les molécules.

Annexe 3. Rapport graphique : distribution de descripteurs moléculaires

Ces descripteurs sont utilisés par des règles de filtre de bases de données moléculaires largement employées telles que la règle des 5 de Lipinski et permettent donc d'obtenir un regard plus fin sur le contenu de l'ensemble préparé. Ce graphe n'est généré que si l'étape de création de sous-librairies est exécutée.

Annexe 4. Rapport graphique : nombre de molécules par sous-librairie

Chaque entrée est passée dans une série de filtres successifs prédéfinis et si au moins une des entrées correspondant à une même molécule est marquée comme appartenant à un sous-ensemble, alors la molécule est considérée comme appartenant à ce même sous-ensemble. Pour les molécules qui ne rentrent dans aucune catégorie, l'étiquette « Autre » est appliquée.

Annexe 5. Rapport graphique : nombre d'alertes par étape.

Des alertes peuvent être déclenchées à chaque étape du protocole selon des filtres prédéfinis et les paramètres rentrés par l'utilisateur. Elles se décomposent en trois catégories.

En orange sont indiqués des avertissements provoqués pour attirer l'attention de l'utilisateur sur des cas particuliers (suppression de sels, détection de PAINS, etc.).

En rouge sont indiquées les erreurs détectées par le protocole et qui entraînent la suppression des entrées correspondantes.

En vert sont indiqués les filtres qui retirent des entrées d'après un scénario prévu par le protocole.

Les molécules et les entrées se distinguent par la nuance de couleur : en foncé les molécules et en clair les entrées. Cela permet notamment d'évaluer combien de molécules uniques provoquent des alertes à chaque étape.

Annexe 6. Rapport graphique : temps de calculs par étape

Ce graphe permet d'identifier les étapes les plus chronophages du protocole et peut servir à l'utilisateur pour mieux paramétrer le protocole au vu de ses besoins. Le temps est indiqué au format HH:MM:SS.

3.5.1 Annexe 1. Capture d'écran du protocole VSPrep tel qu'implémenté dans KNIME

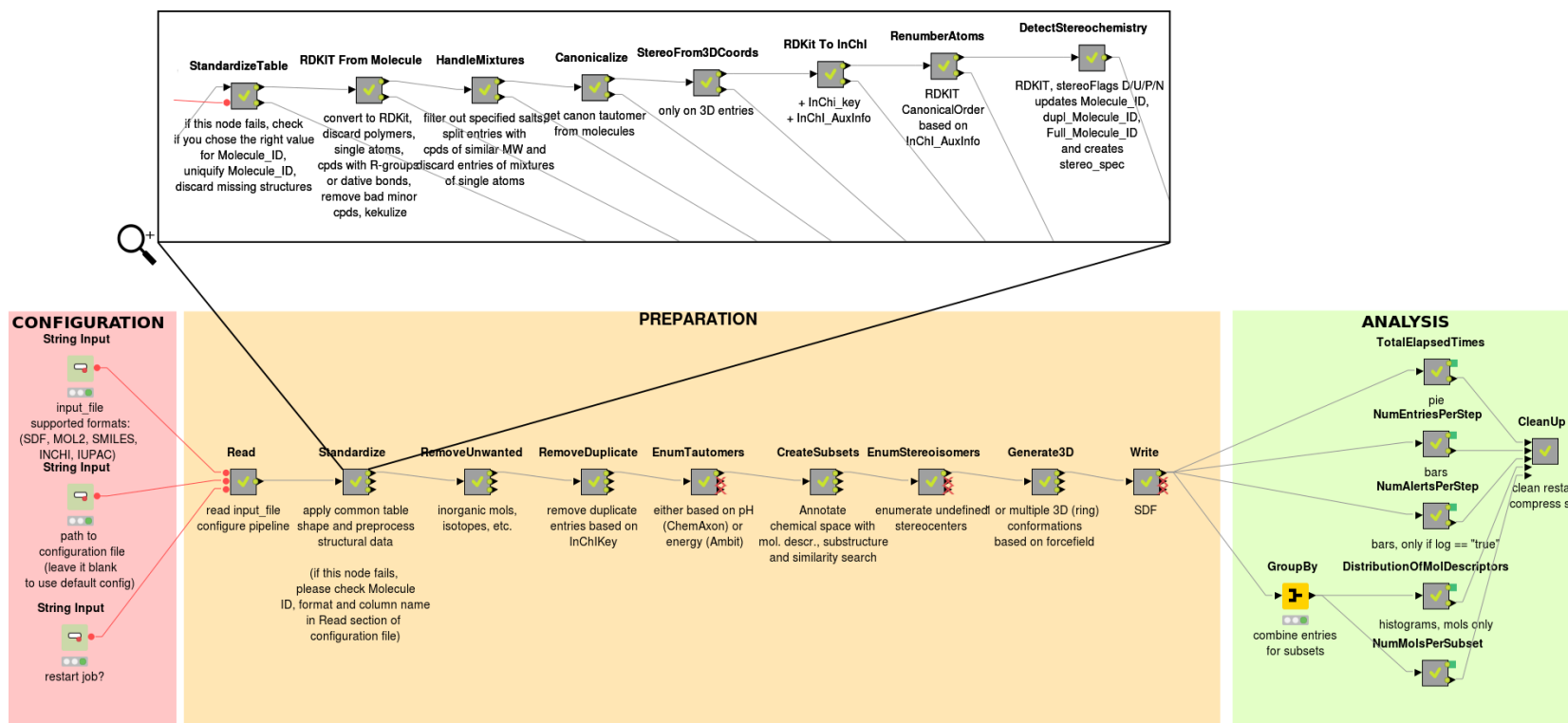


Figure 59. Capture d'écran de VSPrep. Version lors de la rédaction du manuscrit : 0.96.6.

3.5.2 Annexe 2. Exemple de rapport graphique (1)

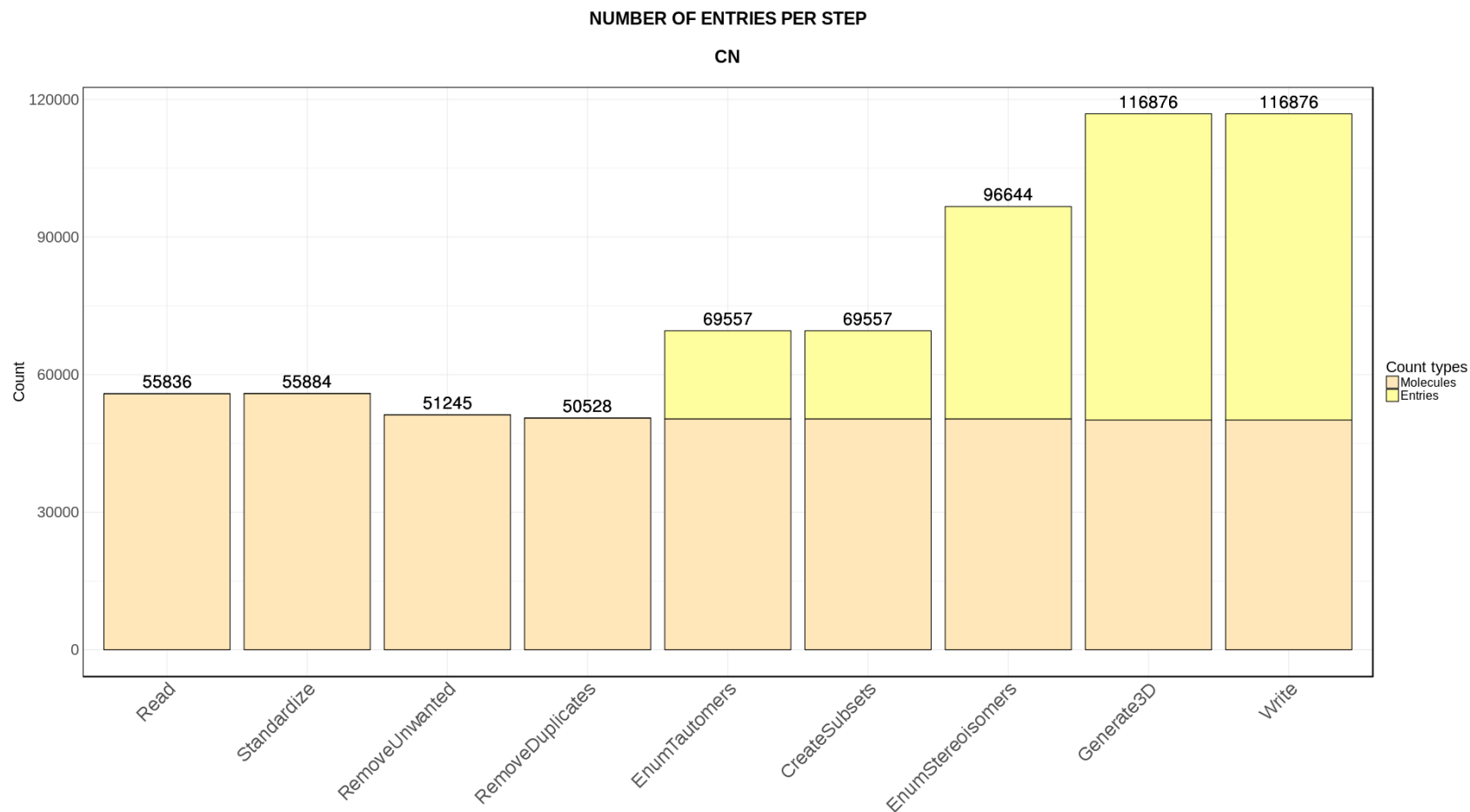


Figure 60. Rapport graphique : nombre de molécules et d'entrées à chaque étape.

3.5.3 Annexe 3. Exemple de rapport graphique (2)

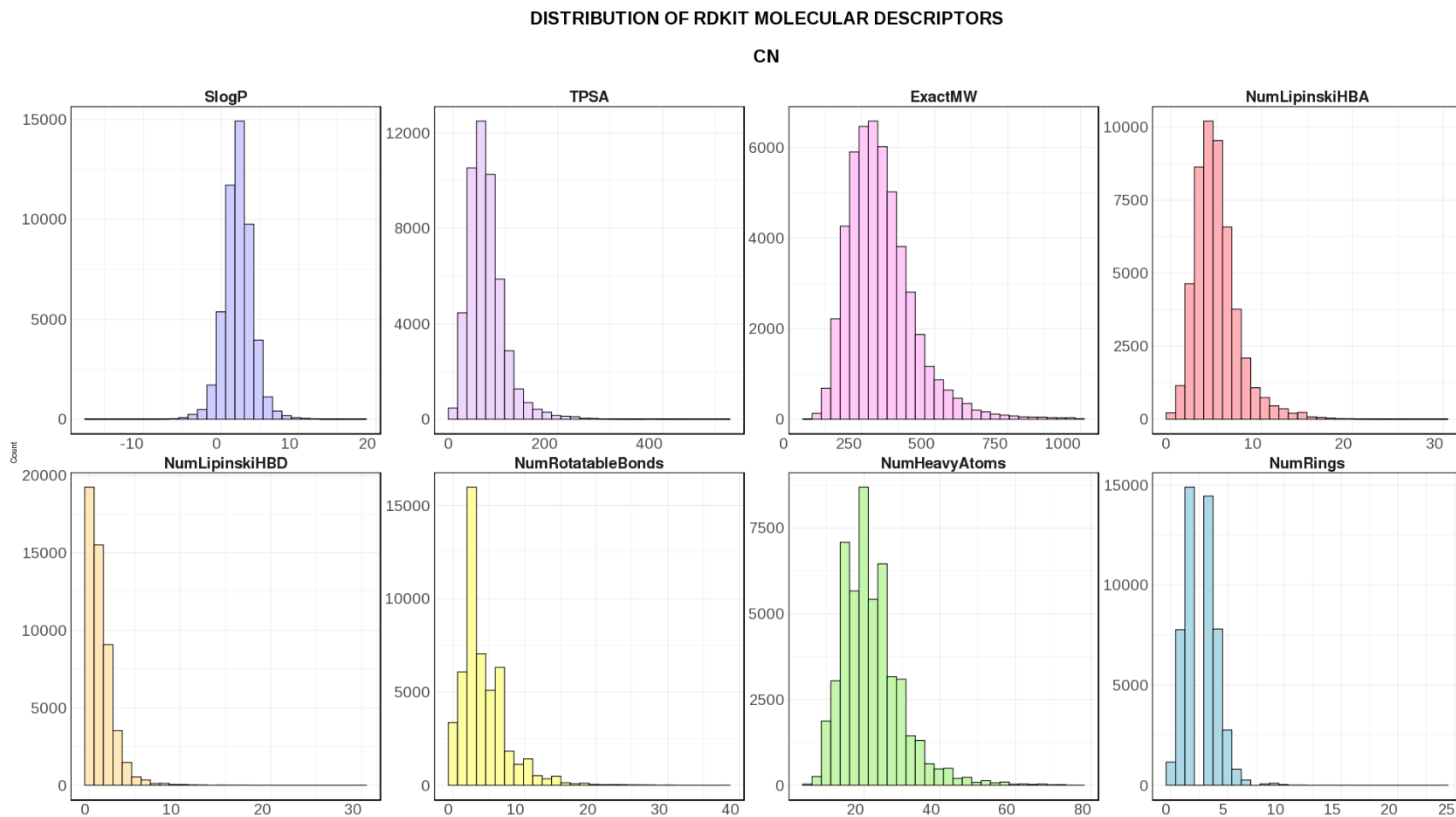


Figure 61. Rapport graphique : distribution de descripteurs moléculaires courants. Ces derniers sont calculés avec RDKit.

3.5.4 Annexe 4. Exemple de rapport graphique (3)

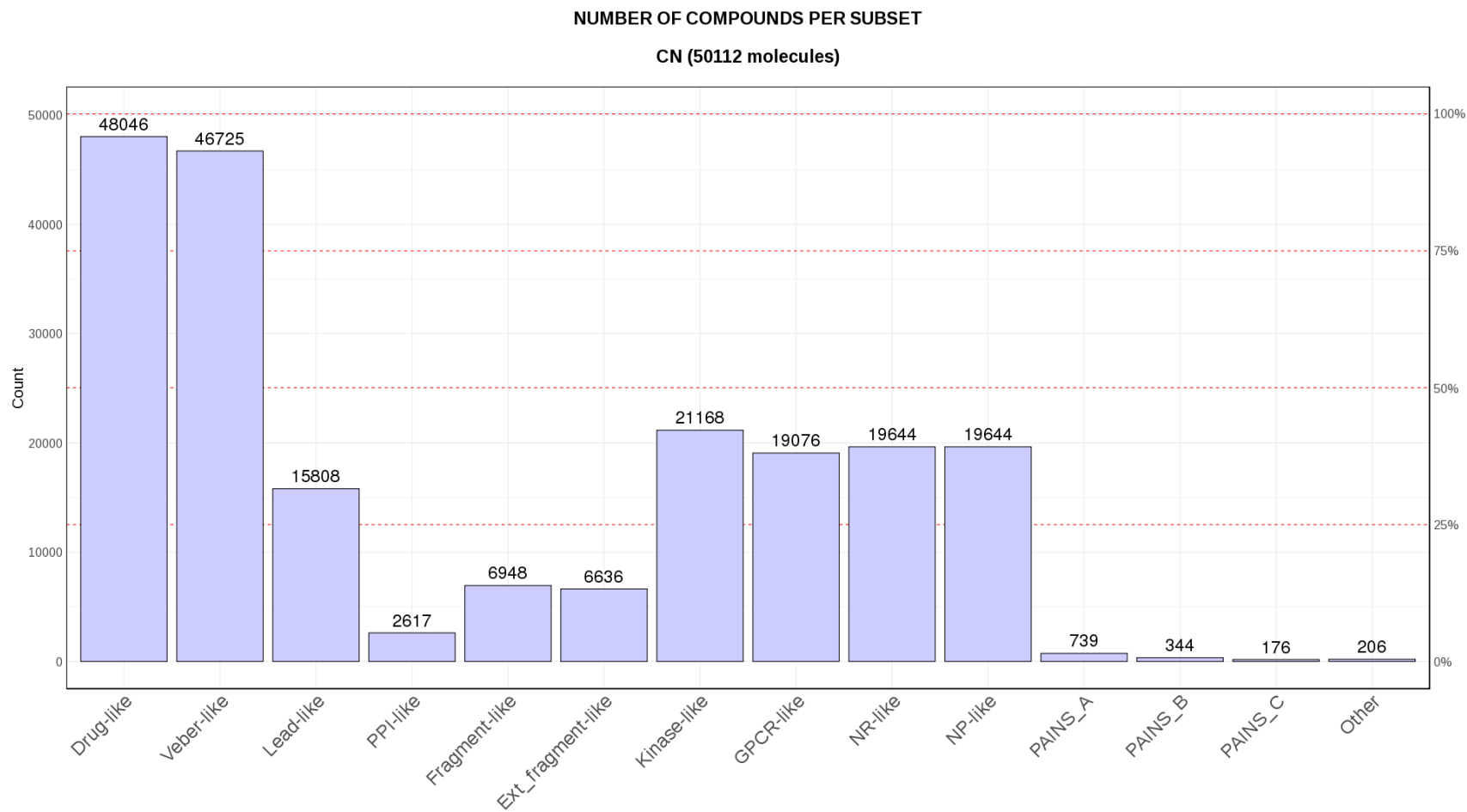


Figure 62. Rapport graphique : nombre de molécules par sous-librairie

3.5.5 Annexe 5. Exemple de rapport graphique (4)

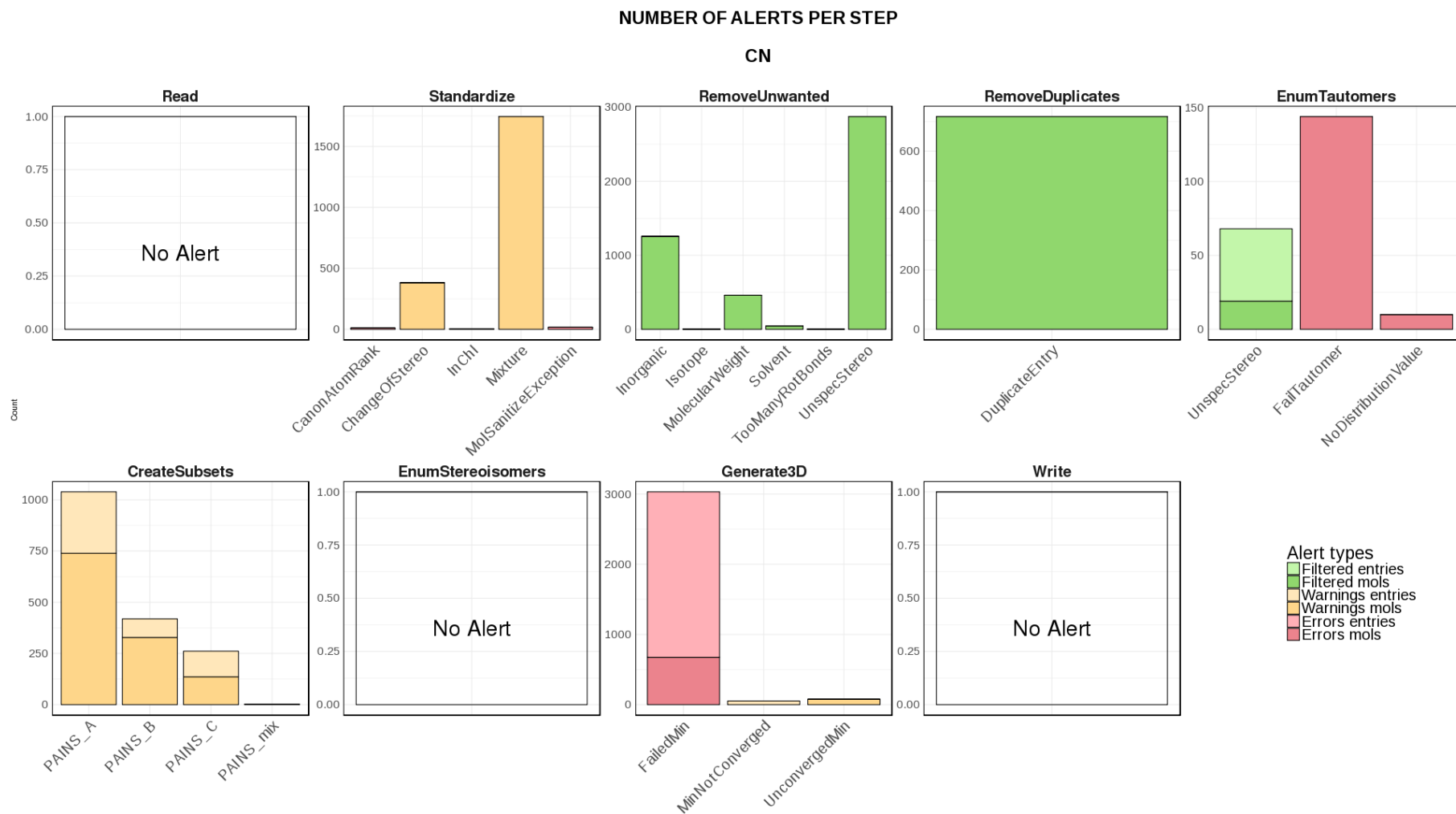


Figure 63. Rapport graphique : nombre d'alertes par étape.

3.5.6 Annexe 6. Exemple de rapport graphique (6)

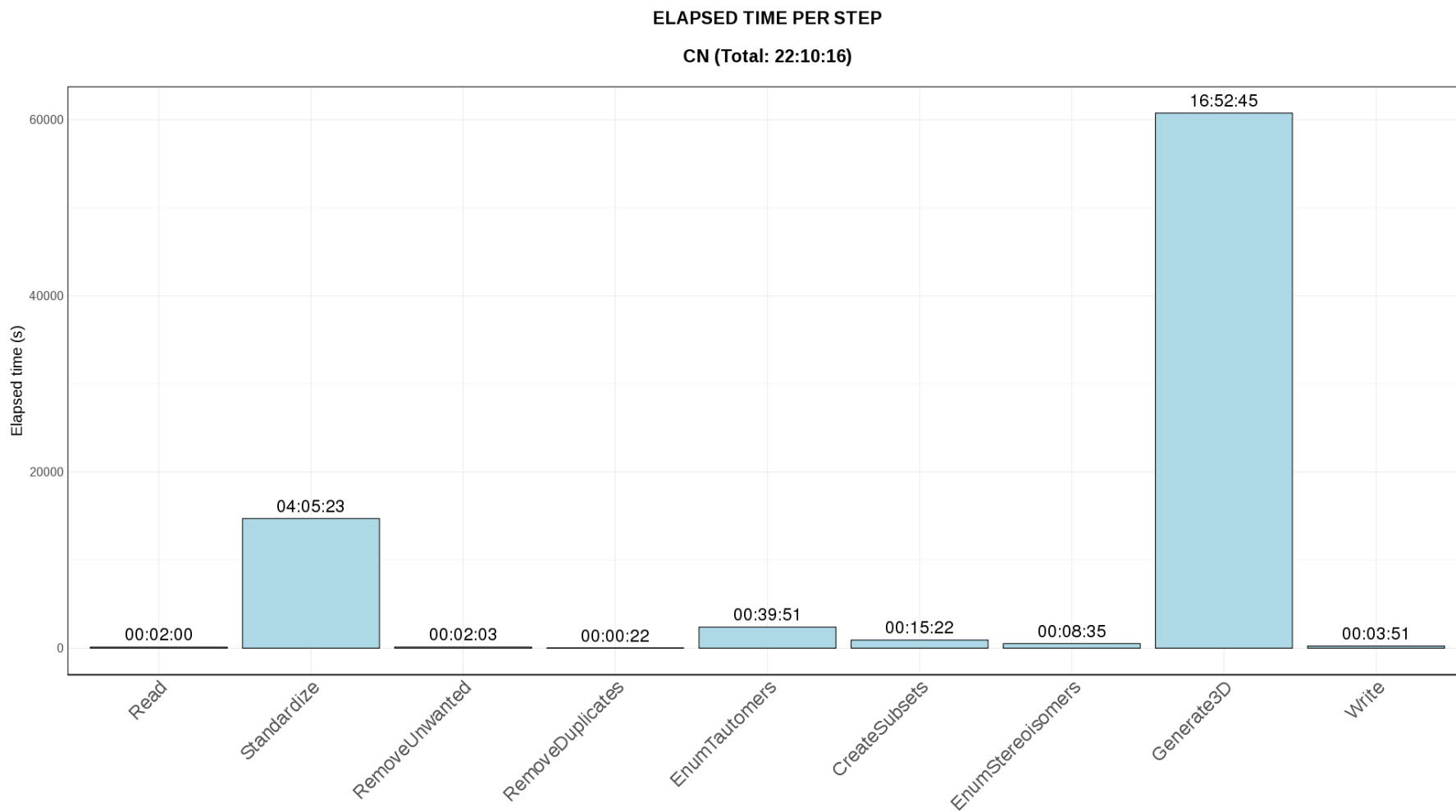


Figure 64. Rapport graphique : temps de calculs par étape

4. Développement d'un nouvel outil de conception de ligands à partir de fragments

4.1 La conception de médicaments basée sur des fragments (FBDD)

La conception de médicaments basée sur des fragments, ou *Fragment-Based Drug Design* (FBDD), est une approche efficace permettant d'identifier des molécules bioactives (touches ou *hits*) qu'il sera possible d'optimiser par la suite (têtes de série ou *leads*) dans le processus de conception de médicaments. Comme son nom l'indique, elle repose sur l'utilisation de petites briques moléculaires, appelées fragments, qui peuvent être assemblées dans le but de concevoir des molécules avec une meilleure affinité pour la protéine ciblée. Pour ce faire, elle s'appuie sur l'information contenue dans des structures tridimensionnelles et s'inscrit aussi dans l'approche *Structure-Based Drug Design* (SBDD).

4.1.1 Emergence du FBDD

La naissance de ce concept FBDD remonte au début des années 1980, et plus particulièrement aux travaux de W.P. Jencks sur le principe d'additivité des énergies de liaison.¹⁸⁶ Le concept proposé est qu'une molécule peut être décomposée en plusieurs sous-unités (fragments qui contiennent au moins un groupement fonctionnel) et que son affinité pour une cible donnée s'explique par la combinaison des interactions protéine-ligand formées par les fragments. Ce concept est par la suite affiné par les travaux de P.R. Andrews¹⁸⁷ pour considérer certains groupes fonctionnels comme importants pour la formation de liaisons. En 1985, le développement du programme GRID¹⁸⁸ permet de positionner des groupements fonctionnels virtuels afin d'établir une carte des sites de liaison (*hot spots*) dans les structures de macromolécules. L'approche in silico MCSS¹⁸⁹ étend alors son application à des groupes fonctionnels plus grands tandis que l'outil LUDI¹⁹⁰ permet entre autres de prédire des interactions hydrogène entre le récepteur et les fragments utilisés.

¹⁸⁶ W. P. Jencks, On the Attribution and Additivity of Binding Energies, *Proceedings of the National Academy of Sciences of the United States of America*, 78.7 (1981), 4046–50

¹⁸⁷ P. R. Andrews *et al.*, Functional Group Contributions to Drug-Receptor Interactions, *Journal of Medicinal Chemistry*, 27.12 (1984), 1648–57

¹⁸⁸ P. J. Goodford, A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules, *Journal of Medicinal Chemistry*, 28.7 (1985), 849–57

¹⁸⁹ A. Miranker *et al.*, Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method, *Proteins*, 11.1 (1991), 29–34

¹⁹⁰ H. J. Böhm, LUDI: Rule-Based Automatic Design of New Substituents for Enzyme Inhibitor Leads, *Journal of Computer-Aided Molecular Design*, 6.6 (1992), 593–606

L'approche FBDD a évolué au gré des d'avancées technologiques expérimentales qui ont servi à la fois de validation des concepts décrits plus tôt et de sources d'inspiration pour de nouvelles théories. Ainsi, en 1992, une méthode est développée pour relier deux molécules de faible poids moléculaire par synthèse organique afin d'optimiser des composés contre la trypanosomiase.¹⁹¹ Cette méthodologie a pu être approfondie par la suite avec le développement de plusieurs techniques de génération de structures de complexes protéine-ligand. Parmi celles-ci, la Relation Structure-Activité (RSA) par spectroscopie de Résonance Magnétique Nucléaire (RMN), ou plus simplement « SAR par RMN¹⁹² », est appliquée pour générer des structures de protéines cibles. Il devient alors possible de cribler des chimiothèques de fragments et de détecter ceux qui ont réussi à se lier à la cible. Un second criblage est alors effectué en présence du premier fragment, qui a été optimisé entre temps, afin de détecter un second site de fixation potentiel. Chaque touche obtenue par ce criblage secondaire est alors reliée par synthèse organique au fragment optimisé. Cette approche, dénommée *fragment linking*, a montré son utilité, notamment par la découverte d'inhibiteurs très actifs de la protéine FKBP. De plus, l'élaboration des premières plateformes de cristallographie par rayons X à haut débit (Astex, SGX)^{193,194} ou encore le développement d'autres techniques expérimentales¹⁹⁵ comme par exemple la Résonance Plasmique de Surface (RPS), ont permis, par la multiplication des structures tridimensionnelles disponibles, l'essor des approches SBDD en général et tout particulièrement du FBDD.^{196,197,198,199,200} Toutes ces découvertes successives, ainsi que les premiers succès du FBDD, marquent un tournant dans le monde pharmaceutique au début des années 2000. En l'espace d'une quinzaine d'années, deux molécules basées sur des fragments ont été approuvées par la FDA (vemurafenib en 2011 et venetoclax en 2016), montrant le succès de cette approche.^{201,202}

¹⁹¹ C. L. Verlinde *et al.*, In Search of New Lead Compounds for Trypanosomiasis Drug Design: A Protein Structure-Based Linked-Fragment Approach, *Journal of Computer-Aided Molecular Design*, 6.2 (1992), 131–47

¹⁹² S. B. Shuker *et al.*, Discovering High-Affinity Ligands for Proteins: SAR by NMR, *Science (New York, N.Y.)*, 274.5292 (1996), 1531–34

¹⁹³ I. Tickle *et al.*, High-Throughput Protein Crystallography and Drug Discovery, *Chemical Society Reviews*, 33.8 (2004), 558–65

¹⁹⁴ T. L. Blundell *et al.*, High-Throughput X-Ray Crystallography for Drug Discovery, *Current Opinion in Pharmacology*, 4.5 (2004), 490–96

¹⁹⁵ G. Chessari *et al.*, From Fragment to Clinical Candidate—a Historical Perspective, *Drug Discovery Today*, 14.13–14 (2009), 668–75

¹⁹⁶ M. Whittaker, Picking up the Pieces with FBDD or FADD: Invest Early for Future Success, *Drug Discovery Today*, 14.13–14 (2009), 623–24

¹⁹⁷ R. J. Bienstock, Overview: Fragment-Based Drug Design, in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, ACS Symposium Series, 1076, 0 vols (American Chemical Society, 2011), MLXXVI, 1–26, consulté le 18 Mai 2016

¹⁹⁸ H. Chen *et al.*, Evolutions in Fragment-Based Drug Design: The Deconstruction–reconstruction Approach, *Drug Discovery Today*, 20.1 (2015), 105–13

¹⁹⁹ R. A. E. Carr *et al.*, Fragment-Based Lead Discovery: Leads by Design, *Drug Discovery Today*, 10.14 (2005), 987–92

²⁰⁰ D. E. Scott *et al.*, Fragment-Based Approaches in Drug Discovery and Chemical Biology, *Biochemistry*, 51.25 (2012), 4990–5003

²⁰¹ D. Erlanson, Fragments in the Clinic: 2016 Edition, <http://practicalfragments.blogspot.com/2016/07/fragments-in-clinic-2016-edition.html>, consulté le 12 Novembre 2017

²⁰² D. A. Erlanson *et al.*, Twenty Years on: The Impact of Fragments on Drug Discovery, *Nature Reviews. Drug Discovery*, 15.9 (2016), 605–19

4.1.2 Les propriétés des fragments

Les fragments sont par nature des petites molécules. A l'instar de la règle de 5 proposée par C.A. Lipinski¹²⁵, une règle de 3 a été proposée par M. Congreve¹²⁷ de chez Astex Pharmaceuticals afin de concevoir des librairies de fragments optimisées. Cette dernière considère qu'une molécule est un bon fragment si elle respecte ces quatre critères : poids moléculaire < 300 Da, nombre d'accepteurs d'atomes d'hydrogène ≤ 3, nombre de donneurs d'atomes d'hydrogène ≤ 3 et logP ≤ 3. Ce dernier est estimé par le calcul du ClogP.¹³ Deux critères optionnels sont également recommandés : NRot ≤ 3 et TPSA ≤ 60 Å².

Les fragments offrent beaucoup d'avantages, comparés à la grande majorité des composés issus des chimiothèques de HTS. Comme ils sont petits, ils peuvent accéder à des poches plus étroites situées dans des cavités ou à la surface de la protéine et donc former plus facilement des interactions que des molécules plus volumineuses.^{203,204} De plus, partir d'une plus petite structure autorise un meilleur contrôle des propriétés physico-chimiques lors de l'optimisation et donc potentiellement un accès à un bon profil ADME.

Cependant l'optimisation des fragments se doit d'être particulièrement minutieuse afin d'obtenir une molécule finale qui se lie efficacement à son récepteur. De plus, le fragment initial n'ayant que peu d'atomes lourds, il ne se lie donc que faiblement à son récepteur. Il est d'autant plus nécessaire de recourir à des techniques très sensibles pour détecter les touches que les meilleurs points de départ ne sont pas forcément les fragments initiaux qui se lient de la manière la plus affine à leur cible.

Afin de pouvoir mieux quantifier la qualité des interactions formées, plusieurs métriques ont été développées au cours du temps.²⁰⁵ La première d'entre elle, le *Ligand Efficiency* (LE) a été proposée par A.L. Hopkins en 2004²⁰⁶ et propose de diviser l'énergie de liaison formée entre la protéine et le ligand par le nombre d'atomes lourds du ligand (**Equation 1**). Encore aujourd'hui, il s'agit de l'une des métriques les plus couramment employées dans le FBDD bien qu'elle ait déjà remise en cause dans plusieurs études.^{207,208}

²⁰³ S. Barelier *et al.*, Ligand Specificity in Fragment-Based Drug Design, *Journal of Medicinal Chemistry*, 53.14 (2010), 5256–66

²⁰⁴ Y. Chen *et al.*, Molecular Docking and Ligand Specificity in Fragment-Based Inhibitor Discovery, *Nature Chemical Biology*, 5.5 (2009), 358–64

²⁰⁵ A. Kumar *et al.*, Fragment Based Drug Design: From Experimental to Computational Approaches, *Current Medicinal Chemistry*, 19.30 (2012), 5128–47

²⁰⁶ A. L. Hopkins *et al.*, Ligand Efficiency: A Useful Metric for Lead Selection, *Drug Discovery Today*, 9.10 (2004), 430–31

²⁰⁷ M. D. Shultz, Setting Expectations in Molecular Optimizations: Strengths and Limitations of Commonly Used Composite Parameters, *Bioorganic & Medicinal Chemistry Letters*, 23.21 (2013), 5980–91

²⁰⁸ J. Polanski *et al.*, Beware of Ligand Efficiency (LE): Understanding LE Data in Modeling Structure-Activity and Structure-Economy Relationships, *Journal of Cheminformatics*, 9 (2017)

Equation 1

$$LE = \frac{\Delta G}{HAC}$$

avec ΔG l'énergie de liaison en kcal.mol⁻¹ et HAC le nombre d'atomes lourds du ligand.

Appliquée sur des données biologiques comprenant des mesures d'efficacité d'inhibition comme l'IC₅₀ (concentration de l'inhibiteur en M à 50% d'inhibition), on obtient l'Equation 2²⁰⁷ :

Equation 2

$$LE = 1.4 \cdot \frac{-\log(IC_{50})}{HAC}$$

En conditions standard (300 K, pH neutre et concentration à 1 M), la valeur de -2,303 .RT s'approxime à -1,37 kcal.mol⁻¹.

Cette formule peut être davantage simplifiée par l'Equation 3 pour obtenir l'Equation 4, utilisée lors de certains de mes travaux de thèse :

Equation 3

$$pIC_{50} = -\log_{10}(IC_{50})$$

Equation 4

$$LE = 1.4 \cdot \frac{pIC_{50}}{HAC}$$

Par ailleurs, des études statistiques menées par P.J. Hadjuk²⁰⁹ montrent que lors des optimisations réussies de *hits* en *lead*, l'énergie libre de la molécule grandit de 0,3 kcal.mol⁻¹ à chaque ajout de nouvel atome lourd, jusqu'à atteindre une limite de 12 kcal.mol⁻¹. Cette valeur limite de 0,3 kcal.mol⁻¹ par atome lourd avait déjà été recommandée par A.L. Hopkins deux années auparavant et est couramment utilisée comme référence pour sélectionner les meilleures touches dans les études de FBDD.

4.1.3 La conception des librairies de fragments

La conception de chimiothèques de fragments constitue la première étape de tout criblage (réel ou virtuel). Les librairies de fragments peuvent être filtrées, afin de cibler certains domaines d'application, selon des critères comme des propriétés moléculaires (mesurées ou calculées) ou bien la présence de groupements fonctionnels particuliers. Les principales considérations à avoir lors de la conception d'une chimiothèque pour du FBDD sont décrites ci-après.

²⁰⁹ P. J. Hadjuk, Fragment-Based Drug Design: How Big Is Too Big?, *Journal of Medicinal Chemistry*, 49.24 (2006), 6972–76

4.1.3.1 La taille de la librairie

La taille d'une chimiothèque, qu'elle contienne des fragments ou d'autres types de molécules, doit dépendre avant tout de l'utilisation qui en est faite. Ainsi, les grands groupes pharmaceutiques ont permis l'émergence de techniques de criblage à haut débit (chimie combinatoire, HTS) afin de rechercher des principes actifs dans l'espace chimique le plus vaste possible. Malgré la découverte de nombreuses molécules bioactives par le passé, on peut considérer que cette approche n'est pas efficace. En effet, l'espace chimique, estimé à 10^{60} molécules *drug-like* (voir chapitre 1), est bien trop vaste pour pouvoir espérer l'explorer exhaustivement. Le groupe de C. Murray a d'ailleurs estimé, en considérant une taille **moyenne de 400-500 Da par composé, que sa masse ferait 10^{11} fois celle de la Terre.**²¹⁰

En contrepartie, si on se limite à des composés de moins de 160 Da, alors l'espace chimique se réduit considérablement (14 millions de composés). Dès lors, cribler 10 000 fragments se révèle bien plus efficace que cribler un million de composés *drug-like*.²¹¹

Une bonne estimation de l'espace chimique actuellement exploré peut être faite en rassemblant les catalogues de tous les vendeurs de composés chimiques. La somme de tous ces composés a été estimée à seulement 10^8 molécules en 2011 par le groupe d'E. Jacoby.²¹²

La taille de la librairie dépend avant tout de l'utilisation qui en sera faite, et plus particulièrement de la rapidité des méthodes de criblages utilisées.²¹³ Ainsi, les petites librairies sont aujourd'hui appropriées pour des expériences de cristallographie par rayons X, de RPS ou de RMN. Ces techniques sont très efficaces pour identifier des *hits* lors de criblage de fragmentothèques.

4.1.3.2 La diversité chimique

Concevoir une librairie diverse est un important facteur de succès de l'approche FBDD. Cela permet notamment d'explorer l'espace chimique *drug-like* plus efficacement et donc d'obtenir des touches plus diverses en terme de structure et d'originalité. De plus, différents groupements chimiques signifient différents modes de liaison avec le récepteur et différentes voies de synthèse pour accéder aux composés optimisés.

La création d'une librairie diverse s'appuie sur des calculs de distance de similarité (par exemple distance de Tanimoto, voir chapitre 1) et sur la représentation des molécules par des descripteurs

²¹⁰ R. J. Hall *et al.*, Efficient Exploration of Chemical Space by Fragment-Based Screening, *Progress in Biophysics and Molecular Biology*, 116.2–3 (2014), 82–91

²¹¹ P. J. Hajduk *et al.*, A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned, *Nat Rev Drug Discov*, 6.3 (2007), 211–19

²¹² S. Renner *et al.*, Recent Trends and Observations in the Design of High-Quality Screening Collections, *Future Medicinal Chemistry*, 3.6 (2011), 751–66

²¹³ G. 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*, 59.18 (2016), 8189–8206

topologiques adaptés ou des empreintes moléculaires appelées aussi *fingerprints*. Cette approche a notamment été employée lors du développement de sous-librairies avec VSPrep comme indiqué au chapitre précédent. Il est également possible de retirer les fragments présentant des groupements estimés trop réactifs et pouvant former des agrégats (PAINS, voir chapitre 3).

4.1.3.3 La règle de 3

Même s'il a été estimé que des molécules plus petites et donc moins complexes devraient fournir un meilleur taux de touches que des molécules *drug-like*²¹¹, il ne faut pas oublier que la règle de 3 a été conçue pour servir de repère plutôt que de filtre strict. Retirer tous les composés ne respectant pas les critères définis plus haut équivaut à prendre le risque de perdre des touches intéressantes. Le cas du groupe de G. Klebe en est d'ailleurs un bon exemple. D'une librairie de seulement 364 fragments²¹⁴, dont la plupart ne respectent par la règle de 3, 11 touches ont été identifiées par cristallographie aux rayons X pour inhiber l'endothiapepsine, une protéase (une enzyme catalysant la dégradation de protéines) isolée chez le champignon *Cryphonectria parasitica*.

4.1.3.4 Les applications particulières

Une librairie de fragments ou plus largement de composés peut être conçue pour une application spécifique. Cela peut par exemple concerner un criblage sur une famille de protéines donnée. Dans ce cas, une possibilité est d'extraire les fragments présents dans des molécules publiées comme bioactives, voire si possible des médicaments. Le postulat de départ est, dans ce cas, que l'activité ou la spécificité des ligands brevetés provient de certains fragments spécifiques qui les composent et que ces propriétés sont transférées aux molécules assemblées.

Il est également possible de prendre en considération dès le départ une technique de criblage en particulier. Parmi ces dernières, on notera le développement récent du criblage par RMN basé sur l'isotope ¹⁹F.²¹⁵ Cette technique récente nécessitant la présence d'atomes de fluor dans les fragments, de nombreux catalogues commerciaux se sont rapidement dotés de librairies de fragments comportant du fluor. A titre d'exemple, Otava propose un tel catalogue, composé de 1077 molécules.²¹⁶

Toutes les bases de données moléculaires de fragments ne proviennent pas forcément de vendeurs chimiques. Par exemple, la GDB-17²¹⁷ résulte de l'énumération systématique de toutes les possibilités de connectivités de molécules organiques jusqu'à 17 atomes, atteignant le nombre impressionnant de 166 milliards de molécules virtuelles. Cependant, un tel volume de données n'est pas

²¹⁴ H. Köster *et al.*, A Small Nonrule of 3 Compatible Fragment Library Provides High Hit Rate of Endothiapepsin Crystal Structures with Various Fragment Chemotypes, *Journal of Medicinal Chemistry*, 54.22 (2011), 7784–96

²¹⁵ J. B. Jordan *et al.*, Fragment Based Drug Discovery: Practical Implementation Based on ¹⁹F NMR Spectroscopy, *Journal of Medicinal Chemistry*, 55.2 (2012), 678–87

²¹⁶ Fluorine Fragment Library, <http://www.otavachemicals.com/products/fragment-libraries/19f-nmr-fluorine-containing-fragment-library>, consulté le 9 Novembre 2017

²¹⁷ L. Ruddigkeit *et al.*, Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17, *Journal of Chemical Information and Modeling*, 52.11 (2012), 2864–75

facile à traiter, même avec les machines d'aujourd'hui et les auteurs ont par la suite filtré cette base de fragments par limite de complexité (nombre maximum de centres chiraux, de cycles, etc.) pour publier la FDB-17²¹⁸ (10 millions de fragments).

4.1.4 L'optimisation des fragments

Le mode de liaison d'un fragment peut être soit prédit par approche informatique (amarrage moléculaire), soit observé expérimentalement (cristallographie aux rayons X, spectroscopie de RMN ou cryo-microscopie électronique). L'information ainsi obtenue peut ensuite être employée dans des projets FBDD afin d'optimiser la touche obtenue. Cette optimisation consiste à transformer le premier fragment, présentant une interaction importante avec la cible, en une molécule plus grande qui gagne alors en efficacité et en spécificité.

A ce jour, on dénombre au moins 35 approches principales pour optimiser un fragment obtenu depuis la structure tridimensionnelle d'un complexe, respectivement nommées selon les anglicismes suivants : « *fragment growing* », « *fragment merging* », « *fragment linking* » et « *scaffold hopping* » (Figure 65).²¹⁹

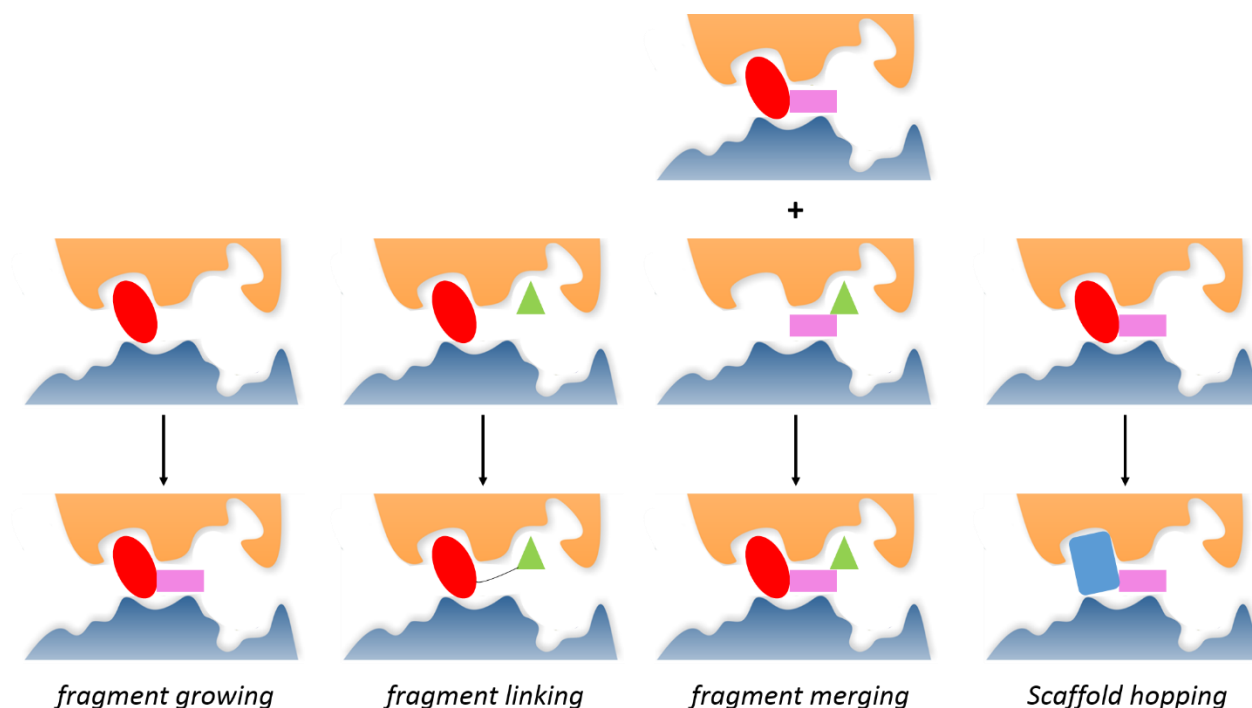


Figure 65. Les quatre approches principales du FBDD.

²¹⁸ R. Visini *et al.*, Fragment Database FDB-17, *Journal of Chemical Information and Modeling*, 57.4 (2017), 700–709

²¹⁹ D. Joseph-McCarthy *et al.*, Fragment-Based Lead Discovery and Design, *Journal of Chemical Information and Modeling*, 54.3 (2014), 693–704

Ce dernier est ici considéré comme une extension de l'approche classique *Structure-Based Drug Design* (SBDD), définie au chapitre 1. Comme nous allons le voir, de nombreux outils *in silico* ont par ailleurs été développés pour accélérer la recherche de molécules bioactives par FBDD. Cependant, les molécules ainsi identifiées restent virtuelles et le chimoinformaticien doit prendre en compte le fait que leur synthèse n'est pas nécessairement facile d'accès. Des méthodes ont donc été incorporées à certains outils pour prendre en compte la faisabilité synthétique des molécules soit pendant, soit après la création des liaisons entre fragments.

Dans les sous-parties suivantes, nous nous concentrerons sur les concepts de chacune de ces méthodes en mentionnant quelques outils *in silico* déjà développés.

4.1.4.1 Fragment Growing

Le *fragment growing* consiste à étendre en plusieurs étapes un fragment initial présentant une interaction importante avec la cible. Les fragments ajoutés permettent d'établir de nouvelles liaisons avec le récepteur. Cette approche est itérative et il est donc possible de vérifier que la croissance du ligand permet bien de former de nouvelles interactions stabilisantes au sein du site actif en se basant sur des données structurales ou de mesure d'activité (par exemple RMN ou SPR). Pour cette raison, il s'agit de la méthode *in silico* la plus couramment employée dans les projets FBDD.

De nombreux outils informatiques ont déjà été développés pour réaliser une approche par *fragment growing*¹⁹⁷ (**Annexe 1**). Parmi ces outils, certains ajoutent itérativement des fragments à un fragment initial en prenant en considération la complémentarité avec le récepteur (GroupBuild²²⁰, GrowMol²²¹/Allegrow²²², SMOG²²³/CombiSMOG²²⁴, AutoGrow²²⁵). Certains de ces outils poussent le concept encore plus loin en agrandissant les molécules non plus fragment par fragment mais atome par

²²⁰ S. H. Rotstein and M. A. Murcko, GroupBuild: A Fragment-Based Method for de Novo Drug Design, *Journal of Medicinal Chemistry*, 36.12 (1993), 1700–1710

²²¹ R. S. Bohacek *et al.*, Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth, *Journal of the American Chemical Society*, 116.13 (1994), 5560–71

²²² K. Biggadike *et al.*, Design and X-Ray Crystal Structures of High-Potency Nonsteroidal Glucocorticoid Agonists Exploiting a Novel Binding Site on the Receptor, *Proceedings of the National Academy of Sciences of the United States of America*, 106.43 (2009), 18114–19

²²³ R. S. DeWitte *et al.*, SMOG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 2. Case Studies in Molecular Design, *Journal of the American Chemical Society*, 119.20 (1997), 4608–17

²²⁴ B. A. Grzybowski *et al.*, From Knowledge-Based Potentials to Combinatorial Lead Design in Silico, *Accounts of Chemical Research*, 35.5 (2002), 261–69

²²⁵ J. D. Durrant *et al.*, AutoGrow: A Novel Algorithm for Protein Inhibitor Design, *Chemical Biology & Drug Design*, 73.2 (2009), 168–78

atome (Legend²²⁶, GenStar²²⁷) tandis que d'autres sont conçus pour l'assemblage de peptides (GROW²²⁸). Il peut également être pertinent d'utiliser des pharmacophores pour concevoir des molécules reproduisant des interactions ligand-protéine clés et cette approche a été implémentée au cœur d'algorithmes FBDD tels que Skelgen²²⁹, PRO_SELECT²³⁰ et NovoFlap²³¹.

Pour répondre au problème de la faisabilité synthétique des solutions trouvées, certains outils incorporent des règles de rétro-synthèse au cœur de leur algorithme pour ne former que des molécules synthétiquement faisables (FOG²³², PINGUI²³³) ou ajoutent cette fonctionnalité post construction (AutoGrow 3.0²³⁴).

4.1.4.2 Fragment Merging

Le *fragment merging* consiste à fusionner deux fragments par un motif commun (que l'on appellera par la suite fragment commun) qui se lie à la même position et de la même manière sur une cible donnée. La perte de degrés de liberté de la molécule associée à la jonction des trois fragments considérés est en général énergétiquement favorable. Le succès de cette technique dépend de la superposition du fragment commun, la distance moyenne maximale de tous les atomes considérés ne devant pas dépasser 1 Å.²³⁵

Une étude récente²³⁶ démontre de manière expérimentale l'utilité de cette approche en l'appliquant à la conception d'un composé antibactérien pour lutter contre *Mycobacterium tuberculosis*. Dans ces travaux, l'analyse des structures cristallographiques issues d'un criblage de fragments, a montré que deux d'entre eux en particulier, étaient capables de se lier chacun deux fois au site de fixation de la protéine cible EthR (facteur de transcription). La superposition des structures des complexes a permis

²²⁶ Y. Nishibata *et al.*, Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation., *Tetrahedron*, 47.43 (1991), 8985–90

²²⁷ S. H. Rotstein *et al.*, GenStar: A Method for de Novo Drug Design, *Journal of Computer-Aided Molecular Design*, 7.1 (1993), 23–43.

²²⁸ J. B. Moon *et al.*, Computer Design of Bioactive Molecules: A Method for Receptor-Based de Novo Ligand Design, *Proteins: Structure, Function, and Bioinformatics*, 11.4 (1991), 314–28

²²⁹ N. P. Todorov *et al.*, Evaluation of a Method for Controlling Molecular Scaffold Diversity in de Novo Ligand Design, *Journal of Computer-Aided Molecular Design*, 11.2 (1997), 175–92

²³⁰ C. W. Murray *et al.*, PRO_SELECT: Combining Structure-Based Drug Design and Combinatorial Chemistry for Rapid Lead Discovery. 1. Technology, *Journal of Computer-Aided Molecular Design*, 11.2 (1997), 193–207

²³¹ J. R. Damewood *et al.*, NovoFLAP: A Ligand-Based de Novo Design Approach for the Generation of Medicinally Relevant Ideas, *Journal of Chemical Information and Modeling*, 50.7 (2010), 1296–1303

²³² P. S. Kutchukian *et al.*, FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space, *J. Chem. Inf. Model.*, 49.7 (2009), 1630–42

²³³ F. Chevillard *et al.*, De Novo Design of Selective Compounds: A Fragment-Based Pipeline Applied to the B2 Adrenergic Receptor, *Journal of Cheminformatics*, 6.Suppl 1 (2014), P25

²³⁴ J. D. Durrant *et al.*, AutoGrow 3.0: An Improved Algorithm for Chemically Tractable, Semi-Automated Protein Inhibitor Design, *Journal of Molecular Graphics & Modelling*, 44 (2013)

²³⁵ S. A. Hudson *et al.*, Overcoming the Limitations of Fragment Merging: Rescuing a Strained Merged Fragment Series Targeting Mycobacterium Tuberculosis CYP121, *Chemmedchem*, 8.9 (2013), 1451–56

²³⁶ P. O. Nikiforov *et al.*, A Fragment Merging Approach towards the Development of Small Molecule Inhibitors of Mycobacterium Tuberculosis EthR for Use as Ethionamide Boosters, *Organic & Biomolecular Chemistry*, 14.7 (2016), 2318–26

d'envisager trois manières différentes de relier les 4 positions de fragments observées. Parmi les solutions obtenues, deux d'entre elles conservaient le mode d'interaction des fragments initiaux. Parmi ces deux molécules formées, l'une d'entre elle a montré une inhibition 100 fois plus importante que les fragments initiaux, atteignant une valeur d'IC₅₀ de 35µM.

L'approche de *fragment merging* a déjà été employée avec succès sur des kinases telles que Jun N-terminal kinase 3 (JNK3)²³⁷ et Pi3γ kinase²⁰⁰. Il n'existe pas, à notre connaissance, d'outil *in silico* spécifique pour le *fragment merging* puisque conceptuellement cette approche peut se découper en deux croissances successives (A vers B puis A-B vers C).

4.1.4.3 Fragment Linking

L'approche *fragment linking* consiste à relier deux fragments, se liant sur une même cible de manière non-compétitive, à l'aide d'un fragment de liaison (*linker*). Théoriquement, la molécule ainsi formée est supposée avoir une énergie de liaison plus favorable que celles des deux fragments isolés. Cependant, il n'est pas nécessairement vrai que le mode de liaison de chaque fragment soit conservé une fois reliés entre eux. En effet, un *linker* trop flexible, ou au contraire trop rigide, peut par exemple empêcher la molécule de reproduire les interactions protéine-ligand clés. De plus, le choix du fragment de liaison est très important pour la modulation des propriétés physico-chimiques de la molécule, comme par exemple ses propriétés ADMET ou encore sa solubilité. En outre, il peut avoir des conséquences sur l'accessibilité synthétique du composé final.

De nombreux outils *in silico* de liaison de fragments ont été développés ces 30 dernières années (Annexe 2). LUDI¹⁹², Builder²³⁸, HOOK²³⁹, PROLIGAND²⁴⁰, DLD²⁴¹, LEA3D²⁴², CONFIRM²⁴³, GANDI²⁴⁴ et l'approche Fragment Shuffling²⁴⁵ décrite par B. Nisius et U. Rester (Bayer) sont des exemples d'outils permettant la liaison de fragments pré-ancrés dans le site actif à l'aide d'un fragment de liaison. D'autres outils combinent à ce concept l'utilisation de pharmacophores, comme c'est par exemple le cas de

²³⁷ S. J. Hughes *et al.*, Fragment Based Discovery of a Novel and Selective PI3 Kinase Inhibitor, *Bioorganic & Medicinal Chemistry Letters*, 21.21 (2011), 6586–90

²³⁸ R. A. Lewis *et al.*, Automated Site-Directed Drug Design Using Molecular Lattices, *Journal of Molecular Graphics*, 10.2 (1992), 66–78, 106

²³⁹ M. B. Eisen *et al.*, HOOK: A Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site, *Proteins*, 19.3 (1994), 199–221

²⁴⁰ D. E. Clark *et al.*, PRO-LIGAND: An Approach to de Novo Molecular Design. 1. Application to the Design of Organic Molecules, *Journal of Computer-Aided Molecular Design*, 9.1 (1995), 13–32.

²⁴¹ A. Miranker *et al.*, An Automated Method for Dynamic Ligand Design, *Proteins*, 23.4 (1995), 472–90

²⁴² Dominique Douguet *et al.*, LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design, *Journal of Medicinal Chemistry*, 48.7 (2005), 2457–68

²⁴³ D. C. Thompson *et al.*, CONFIRM: Connecting Fragments Found in Receptor Molecules, *Journal of Computer-Aided Molecular Design*, 22.10 (2008), 761–72

²⁴⁴ F. Dey *et al.*, Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization, *Journal of Chemical Information and Modeling*, 48.3 (2008), 679–90

²⁴⁵ B. Nisius *et al.*, Fragment Shuffling: An Automated Workflow for Three-Dimensional Fragment-Based Ligand Design, *J. Chem. Inf. Model.*, 49.5 (2009), 1211–22

NEWLEAD²⁴⁶, SPLICE²⁴⁷, FlexNovo²⁴⁸ et du script « combine_fragments » commercialisé par Schrödinger²⁴⁹. Afin de mieux estimer la qualité des interactions protéine-ligand formées, certains outils vont jusqu'à employer des techniques de dynamique moléculaire, comme par exemple les programmes CONCEPTS²⁵⁰ (liaison d'atomes) ou CONCERTS²⁵¹ (liaison de fragments). Le logiciel SPROUT^{252,253} intègre quant à lui des règles de rétrosynthèse directement dans son algorithme de liaison.

4.1.4.4 Scaffold Hopping

Le principe du remplacement d'échafaudage moléculaire (*scaffold-²⁵⁴ ou lead hopping²⁵⁵*) est de moduler les propriétés physico-chimiques d'une molécule bioactive tout en conservant son mode d'interaction avec sa cible et en produisant de l'innovation. La principale contrainte de cette approche est que le fragment substituant doit présenter une forme et un potentiel électrostatique compatibles avec le fragment initial, sans quoi la molécule finale ne pourra pas se lier de la même manière à son récepteur.

Il s'agit en fait d'une approche SBDD classique complémentaire au FBDD et est largement appliquée par la communauté scientifique de chimie médicinale puisqu'un grand nombre de médicaments développés au cours de ces dernières années résultent de la modification de structures déjà existantes.²⁵⁶ Pour cette raison, il est logique qu'un grand nombre d'outils *in silico* aient été développés ces 30 dernières

²⁴⁶ V. Tschinke *et al.*, The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses, *Journal of Medicinal Chemistry*, 36.24 (1993), 3863–70

²⁴⁷ C. M. W. Ho *et al.*, SPLICE: A Program to Assemble Partial Query Solutions from Three-Dimensional Database Searches into Novel Ligands, *Journal of Computer-Aided Molecular Design*, 7.6 (1993), 623–47

²⁴⁸ J. Degen *et al.*, FlexNovo: Structure-Based Searching in Large Fragment Spaces, *ChemMedChem*, 1.8 (2006), 854–68

²⁴⁹ Schrödinger Release 2017-3: Canvas, Schrödinger, LLC, New York, NY, 2017., *Combine_fragments.Py*.

²⁵⁰ D. A. Pearlman *et al.*, CONCEPTS: New Dynamic Algorithm for de Novo Drug Suggestion, *Journal of Computational Chemistry*, 14.10 (1993), 1184–93

²⁵¹ D. A. Pearlman *et al.*, CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design, *J. Med. Chem.*, 39.8 (1996), 1651–63

²⁵² V. J. Gillet *et al.*, SPROUT: Recent Developments in the de Novo Design of Molecules, *Journal of Chemical Information and Computer Sciences*, 34.1 (1994), 207–17

²⁵³ V. J. Gillet *et al.*, SPROUT, HIPPO and CAESA: Tools for de Novo Structure Generation and Estimation of Synthetic Accessibility, *Perspectives in Drug Discovery and Design*, 3.1 (1995), 34–50

²⁵⁴ S. G. *et al.*, "Scaffold-Hopping" by Topological Pharmacophore Search: A Contribution to Virtual Screening, *Angewandte Chemie (International Ed. in English)*, 38.19 (1999), 2894–96

²⁵⁵ R. D. Cramer *et al.*, "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities, *Journal of Medicinal Chemistry*, 47.27 (2004), 6777–91

²⁵⁶ H.-J. Böhm *et al.*, Scaffold Hopping, *Drug Discovery Today. Technologies*, 1.3 (2004), 217–24

années (Annexe 3). Les principaux sont LeapFrog²⁵⁷, CAVEAT²⁵⁸, BREED²⁵⁹, SHOP²⁶⁰, ParaFrag²⁶¹, CombiGlide²⁶², Med-SuMo²⁶³, Scaffold Replacement²⁶⁴, LigMerge²⁶⁵, Core Hopping²⁶⁶, Scaffold Hopping²⁶⁷ et BROOD²⁶⁸.

L'approche *in silico* FBDD développée dans le cadre de ces travaux de thèse repose sur l'assemblage de fragments directement dans la structure du site actif d'une protéine kinase. Ces fragments peuvent provenir de différentes origines. Ils peuvent par exemple correspondre à des produits chimiques réels, proposés par exemple dans les catalogues de vendeurs de composés chimiques. Une autre possibilité est d'utiliser des molécules uniquement virtuelles, comme c'est notamment le cas pour la base de données moléculaire FDB-17, déjà mentionnée dans la partie 4.1.3. Certaines de ces bases de données moléculaires comme la ZINC²⁶⁹ proposent des molécules avec des coordonnées tridimensionnelles et sont donc utilisables telles quelles pour réaliser du criblage virtuel comme de l'amarrage moléculaire.

4.1.4.5 Les outils mixtes

Les différentes approches décrites précédemment ont chacune des avantages et des inconvénients. Par exemple, le *fragment growing* est conceptuellement clair et permet une vérification minutieuse des résultats mais est souvent très chronophage du fait du grand nombre de solutions potentiellement possibles. La méthode de *fragment linking* est en comparaison plus rapide mais elle requiert le bon positionnement d'un fragment de liaison adéquat.

²⁵⁷ LEAPFROG TRIPOS ASSOCIATES, INC. Trademark Information, *Trademarkia.Com*,

<https://www.trademarkia.com/leapfrog-tripos-associates-inc-74380604.html>, consulté le 3 Novembre 2017

²⁵⁸ G. Lauri *et al.*, CAVEAT: A Program to Facilitate the Design of Organic Molecules, *Journal of Computer-Aided Molecular Design*, 8.1 (1994), 51–66

²⁵⁹ A. C. Pierce *et al.*, BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease, *Journal of Medicinal Chemistry*, 47.11 (2004), 2768–75

²⁶⁰ F. Fontaine *et al.*, SHOP: A Method for Structure-Based Fragment and Scaffold Hopping, *ChemMedChem*, 4.3 (2009), 427–39

²⁶¹ A.-J. Jakobi *et al.*, ParaFrag--an Approach for Surface-Based Similarity Comparison of Molecular Fragments, *Journal of Molecular Modeling*, 14.7 (2008), 547–58

²⁶² Schrödinger Release 2016-4: CombiGlide, Schrödinger, LLC, New York, NY, 2016

²⁶³ O. Sperandio *et al.*, MED-SuMoLig: A New Ligand-Based Screening Tool for Efficient Scaffold Hopping, *Journal of Chemical Information and Modeling*, 47.3 (2007), 1097–1110

²⁶⁴ Scaffold Replacement in MOE, <https://www.chemcomp.com/journal/newscaffold.htm>, consulté le 3 Novembre 2017

²⁶⁵ S. Lindert *et al.*, LigMerge: A Fast Algorithm to Generate Models of Novel Potential Ligands from Sets of Known Binders, *Chemical Biology & Drug Design*, 80.3 (2012), 358–65

²⁶⁶ Core Hopping | Schrödinger, <https://www.schrodinger.com/core-hopping>, consulté le 3 Novembre 2017

²⁶⁷ M. J. Vainio *et al.*, Scaffold Hopping by Fragment Replacement, *Journal of Chemical Information and Modeling*, 53.7 (2013), 1825–35

²⁶⁸ OpenEye Scientific Software, BROOD | Lead Optimization, <https://www.eyesopen.com/brood>, consulté le 19 Octobre 2017

²⁶⁹ J. J. Irwin *et al.*, ZINC--a Free Database of Commercially Available Compounds for Virtual Screening, *Journal of Chemical Information and Modeling*, 45.1 (2005), 177–82

Certains outils rassemblent les différentes approches sous une seule interface, maximisant ainsi leur fonctionnalité (**Table 17**).

Table 17. Liste des outils in silico FBDD capables d'effectuer plusieurs approches FBDD.

Outil	<i>fragment growing</i>	<i>fragment linking</i>	<i>scaffold hopping</i>
LigBuilder, LigBuilder 2	✓	✓	✓
LEADOPT	✓	✗	✓
RECORE	✓	✓	✓
Spark	✓	✓	✓
BROOD v3.0	✗	✓	✓
ACFIS	✓	✗	✓

Parmi ces derniers, LigBuilder²⁷⁰ ajoute itérativement des fragments d'après un pharmacophore sur le récepteur. En 2011, l'outil a bénéficié d'une importante mise à jour (LigBuilder 2²⁷¹) qui permet d'intégrer plus de paramètres, comme un score de complémentarité entre le récepteur et le ligand et la prédiction de la faisabilité synthétique. LEADOPT²⁷² effectue quant à lui une cartographie de la cavité afin d'orienter la croissance du ligand de manière à prévenir les encombrements stériques entre la protéine et le ligand. Il est capable de substituer un fragment par un autre afin d'améliorer l'affinité de la molécule assemblée. ReCore²⁷³ est similaire à CAVEAT dans son approche (remplacement de fragments en conservant les interactions avec la protéine) mais se distingue par son optimisation pour être appliqué sur de larges bases de données moléculaires de fragments. Il est capable d'effectuer le *fragment growing*, *fragment linking* et *scaffold hopping*. Une mise à jour en 2015 de l'outil BROOD, qui intègre plusieurs niveaux de données (structurales, diversité, modèles statistiques), lui confère la capacité d'effectuer du *scaffold hopping*. Spark²⁷⁴ effectue du remplacement de fragments en se basant sur leur forme et leurs propriétés électrostatiques. Il peut également être appliqué à des études de *fragment growing* et *linking*.

²⁷⁰ R. Wang *et al.*, LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design, *Molecular Modeling Annual*, 6.7–8 (2000), 498–516

²⁷¹ Y. Yuan *et al.*, LigBuilder 2: A Practical de Novo Drug Design Approach, *Journal of Chemical Information and Modeling*, 51.5 (2011), 1083–91

²⁷² G.-B. Li *et al.*, LEADOPT: An Automatic Tool for Structure-Based Lead Optimization, and Its Application in Structural Optimizations of VEGFR2 and SYK Inhibitors, *European Journal of Medicinal Chemistry*, 93 (2015), 523–38

²⁷³ P. Maass *et al.*, Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations, *Journal of Chemical Information and Modeling*, 47.2 (2007), 390–99

²⁷⁴ Spark: Find New Chemistry Ideas Using Bioisosteres, Cresset, 2012, <http://www.cresset-group.com/products/spark/>, consulté le 3 Novembre 2017

Enfin, ACFIS²⁷⁵ est un outil hébergé sur un serveur en ligne.²⁷⁶ Il permet la déconstruction d'un ligand co-cristallisé afin d'en substituer un fragment depuis un point d'ancrage. Sa méthode de liaison repose sur l'outil de *fragment growing* AutoGrow 3 et s'effectue donc par réaction chimique virtuelle. Il est à remarquer que le développement de ce genre d'outils FBDD simples à utiliser, car délocalisés sur un serveur, pourrait encourager plus de scientifiques à essayer ces approches et ainsi participer à la visibilité de cette méthode.

4.2 Développement de Frags2Drugs, un nouvel outil *in silico* FBDD

Comme nous l'avons vu, de nombreux outils d'optimisation *in silico* de fragments ont été développés au cours des 30 dernières années. L'outil Frags2Drugs, décrit dans ce chapitre de manuscrit de thèse, a été conçu selon une approche de *fragment growing* car nous ne développons pas de fragment de liaison entre deux fragments spécifiques. J'aborderai tout d'abord les raisons qui ont conduit à l'initiation de ce projet ainsi que le développement de l'outil en lui-même et sa validation dans des projets de recherche en interne.

4.2.1 Pourquoi un nouvel outil ?

L'une des premières tâches que j'ai eu à effectuer au cours de ma thèse a été de faire un état de l'art des outils *in silico* FBDD. J'ai pu recenser un grand nombre d'outils (Annexes 1-3), pour la plupart commerciaux, qui varient autant par leur approche (combinatoire exhaustive ou stochastique, dynamique moléculaire, pharmacophores, etc.) que par leur implémentation (outils en C/C++/Fortran, Python ou bien service accessible directement en ligne).

Une question légitime se pose alors. Que puis-je apporter de plus par rapport à ce qui existe déjà ?

Les outils des années 1990 sont pour la plupart codés dans un mélange de C/C++/Fortran et, par nature, leur maintenance est plus difficile que celle de programmes de langages plus récents et orientés objets comme le Python. De plus, il n'est pas trivial de les interfacer avec d'autres applications sans fournir d'importants efforts, comme cela a pu être fait pour l'élaboration de Cinfony (interface commune à plusieurs bibliothèques de chémoinformatique).²⁷⁷

²⁷⁵ G.-F. Hao *et al.*, ACFIS: A Web Server for Fragment-Based Drug Discovery, *Nucleic Acids Research*, 2016

²⁷⁶ ACFIS Server, <http://chemyang.ccnu.edu.cn/ccb/server/ACFIS/>, consulté le 20 Octobre 2017

²⁷⁷ N. M. OBoyle *et al.*, Cinfony--Combining Open Source Cheminformatics Toolkits behind a Common Interface, *Chemistry Central Journal*, 2 (2008), 24

De plus, les molécules ainsi virtuellement conçues ne sont pas forcément accessibles par chimie de synthèse, l'outil doit donc prendre ce paramètre en compte. Trois approches peuvent être employées pour pallier ce problème :

1. Les fragments peuvent être obtenus avec des algorithmes prenant en compte des règles de rétrosynthèse.
2. Les fragments peuvent être connectés selon des réactions chimiques simples (« *click-chemistry* »)
3. La faisabilité synthétique des molécules finales peut être évaluée par un score d'accessibilité synthétique.

Une autre limitation est que la plupart de ces outils réalisent une exploration systématique des groupements ajoutés au fragment initial au fur et à mesure de son grandissement, ce qui résulte en des temps de calcul considérables (même si cette estimation se base sur les données produites sur des machines bien moins puissantes que ceux disponibles de nos jours). Le nombre de structures tridimensionnelles disponibles publiquement étant en continuelle expansion, nous avons accès à de plus en plus d'information structurale. Il nous est alors apparu pertinent de développer une approche statique, donc rapide, où des fragments *in situ* sont combinés en conservant leurs coordonnées initiales. Nous verrons par la suite que cela donne un grand nombre de possibilités et a suffi comme point de départ à ce projet.

Finalement, un grand nombre des outils existants ne sont pas gratuits, pas *open source* ou sont accessibles en ligne uniquement, ce qui peut provoquer des problèmes de confidentialité.

4.2.2 Création d'une base de données moléculaire de fragments issus de structures tridimensionnelles

L'approche *in silico* FBDD développée dans le cadre de ces travaux de thèse repose sur l'assemblage de fragments directement dans la structure du site actif d'une protéine kinase. Pour cette raison, il est nécessaire d'employer des structures tridimensionnelles de fragments. Plusieurs alternatives sont possibles pour accéder à ces structures. Tout d'abord, certaines structures cristallographiques permettent de visualiser le mode de liaison de fragments au sein du domaine kinase. Il est alors possible d'extraire ces fragments afin de les utiliser dans notre approche. Il est également envisageable de considérer des molécules plus complexes, qui ne respectent donc pas la règle de 3. Les ligands extraits de cette manière peuvent ensuite être découpés par différents algorithmes pour obtenir des fragments, tout en conservant les coordonnées tridimensionnelles de la molécule d'origine. Enfin, une autre approche est de réaliser l'amarrage moléculaire de bases de données moléculaires de fragments. Les molécules peuvent alors être en deux dimensions, comme c'est le cas dans la plupart des catalogues de vendeurs de composés chimiques. Une étape supplémentaire de préparation, par exemple avec VSPrep, est alors requise avant de pouvoir les utiliser. Elles peuvent aussi être disponibles en trois dimensions (Ambinter, base de données que nous avons préparée avec VSPrep et qui est disponible en ligne¹⁴⁹, ZINC²⁶⁹), ce qui en permet une utilisation plus directe.

Dans le cadre de mes travaux de thèse, je me suis surtout focalisé sur l'extraction puis la fragmentation des ligands co-cristallisés avec des protéines kinases en utilisant des outils dédiés. Pour ce faire, j'ai d'abord dû mettre au point un référentiel commun de structures de domaines kinases afin de pouvoir en combiner l'information. Nous verrons donc dans un premier temps comment les structures de domaines kinases ont été extraites puis alignées et enfin comment les ligands ont été fragmentés pour générer notre base de données de fragments.

4.2.2.1 Création d'un référentiel commun de domaines kinases

Les quelques 3809 structures de protéines kinases disponibles en juillet 2016 ont été extraites directement de la PDB sur le site de la RCSB à l'aide d'une requête portant sur les numéros d'accès PFAM PF07714, « protéine tyrosine kinase », et PF00069, « domaine de protéine kinase ». Ces numéros d'accès permettent d'identifier à la fois les structures de tyrosine kinases et de sérine-thréonine kinases, même si quelques erreurs ont pu être détectées (**Figure 66**) et corrigées par la suite en utilisant la librairie Biopython.²⁷⁸

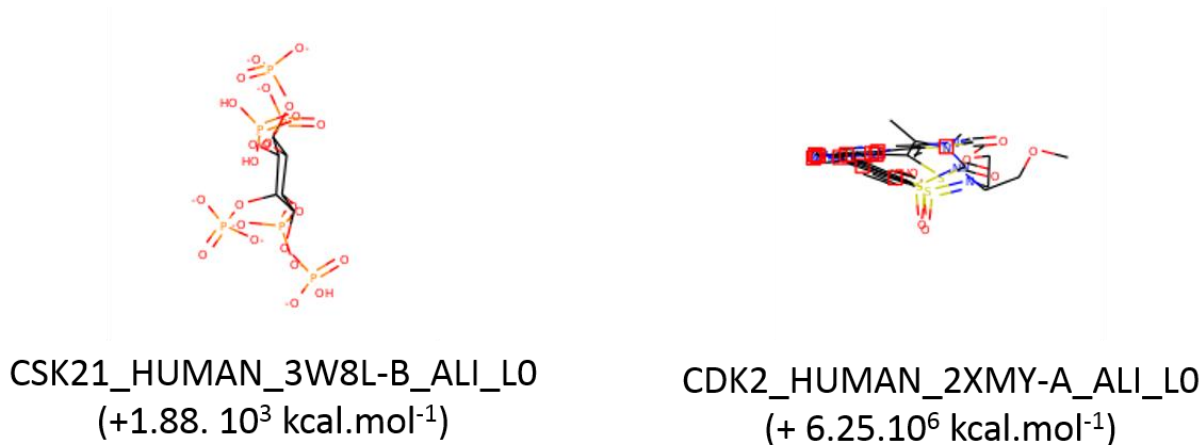


Figure 66. Exemple d'un problème de prise en charge de ligands avec des coordonnées alternatives. Les valeurs énergétiques des conformations calculées par RDKit (MMFF94s) sont indiquées entre parenthèses. L'explication des noms est décrite ci-dessous.

J'ai ainsi conservé les chaînes de chaque complexe dans un fichier séparé en préservant les molécules de solvant et les éventuels cofacteurs. De même, lorsque des atomes présentaient deux coordonnées alternatives possibles, la structure entière a été divisée en deux modèles traités de manière indépendante par la suite.

²⁷⁸ P. J. A. Cock *et al.*, Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics, *Bioinformatics (Oxford, England)*, 25.11 (2009), 1422–23

A)



B)

[illegible]

Une inspection visuelle des ligands extraits a alors permis de mettre rapidement en évidence des problèmes topologiques pour certains d'entre eux, essentiellement issus d'une mauvaise lecture du format PDB pour les hétéroatomes. Pour contourner le problème, les ligands ont tous été téléchargés

séparément depuis la PDB, au format SDF, et inclus lors de l'alignement des structures de protéines kinases avec MOE. Enfin, les molécules identifiées comme des ligands par MOE sont extraites au format SDF pour post-traitement. Un dernier nettoyage est encore requis avant de pouvoir passer à l'étape de fragmentation des ligands.

En effet, dans une structure cristallographique, on trouve typiquement plusieurs types de molécules en plus du récepteur et du ligand. Il y a par exemple des cofacteurs, des molécules d'eau, des contre-ions ou encore d'autres molécules nécessaires lors des expériences (molécules de glycérol, etc.). L'information permettant de savoir quelle molécule est le « bon » ligand dans une structure n'est pas toujours évidente. En effet, on ne peut se contenter de faire une sélection par cavité du fait de la présence d'inhibiteurs allostériques dans le jeu de données. Nous avons donc pris le parti de retirer les composés indésirables en se basant sur une liste publiée dans une étude récente.¹¹³ Ceci permet notamment de retirer la majorité des ions et des molécules de solvant et organiques comme le glycérol. Les petites molécules restantes sont ensuite fragmentées comme décrit ci-dessous.

La grande majorité des structures de complexes présente des parties non résolues, majoritairement dans la boucle d'activation qui est très flexible. Nous n'avons pas cherché à combler les parties manquantes et à corriger manuellement les divers autres problèmes récurrents dans les structures cristallographiques (atomes lourds manquants sur certains résidus comme des lysines, tautomère des histidines, etc.) puisque nous nous intéressons surtout aux ligands. Nous nous sommes donc contentés de préparer les complexes en appliquant le module « Structure Preparation » de MOE en mode automatique, sauf pour les parties manquantes du domaine, que nous n'avons pas modifiées .

4.2.2.2 Fragmentation des ligands co-cristallisés

J'ai développé un *workflow* KNIME pour pouvoir découper de manière aisée les ligands en fragments. Ce dernier permet de tirer profit de différents algorithmes implémentés dans deux outils commerciaux, MOE et JChem. Le premier propose en effet des nœuds KNIME dédiés, et notamment un nœud fragmentant directement les molécules. Le second dispose également de nœuds KNIME susceptibles de nous intéresser mais ceux-ci sont développés par une autre société, Infocom Corporation, et requièrent une licence supplémentaire pour fonctionner. La commande fragment de JChem a donc été implémentée manuellement dans le *workflow* KNIME. Au total, 8 algorithmes différents (**Table 18**) ont été employés pour générer autant de fragments que possibles à partir de chaque molécule (**Figure 68**). Les fragments doublons (identiques en topologie et en coordonnées 3D) ont été supprimés après standardisation. Ainsi, des 8648 ligands de départ, on obtient 131 826 fragments, soit en moyenne 15 fragments par ligand (**Figure 69**).

Table 18. Les méthodes de fragmentation appliquées aux ligands co-cristallisés

Logiciel	Nom de la méthode	Description
	RECAP	Découpe les molécules selon les règles de rétrosynthèse RECAP ²⁷⁹ (les molécules fragmentées pourraient être synthétisées à partir des fragments obtenus)
	RingAtoms	Rrompt les liaisons entre un cycle et un atome n'appartenant à aucun cycle (sauf les doubles liaisons non impliquées dans des cycles)
	RingBlock	Rrompt les liaisons non cycliques (sauf les doubles liaisons non impliquées dans des cycles)
	Scaffold Tree ²⁸⁰	Rrompt les liaisons entre l'échafaudage moléculaire et ses substituants
JChem	FragmenterAll	Découpe la molécule selon un ensemble de 10 fonctions chimiques prédéterminées (ester, amine, urée, etc.)
	FragmenterRecap	Agit pareil que FragmenterAll mais se restreint aux voies de rétrosynthèses définies par RECAP
	RingChain	Rrompt les liaisons non impliquées dans des cycles (sauf les doubles liaisons)
	RingChainRecap	Agit pareil que RingChain mais se restreint aux voies de rétrosynthèses définies par RECAP

²⁷⁹ X. Q. 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*, 38.3 (1998), 511–22

²⁸⁰ A. Schuffenhauer *et al.*, The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification, *Journal of Chemical Information and Modeling*, 47.1 (2007), 47–58

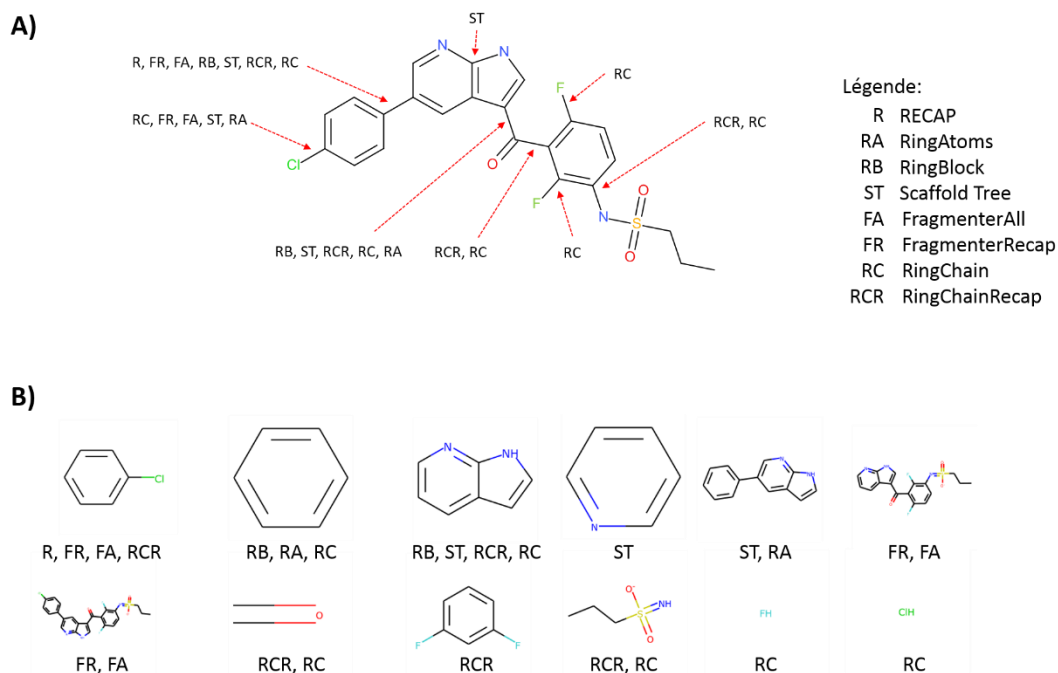


Figure 68. Résultats de la fragmentation de l'inhibiteur vemurafenib par les 8 méthodes utilisées. A) Déconstruction de la molécule. Les flèches rouges en pointillés indiquent les endroits où elle est découpée. B) Fragments obtenus. La molécule FH est obtenue deux fois.

Finalement, chaque fragment résultant de cette approche est annoté par un identifiant unique grâce à un *workflow* capable de récupérer des données sur la RCSB. Le nom des protéines n'étant pas standard dans le titre des structures PDB (par exemple : BRAF, B-RAF), le nom Uniprot (*BRAF_HUMAN*) est employé pour regrouper toutes les structures d'une même protéine kinase (**Figure 70**). La méthode de fragmentation est également sauvegardée pour faciliter l'analyse des résultats.

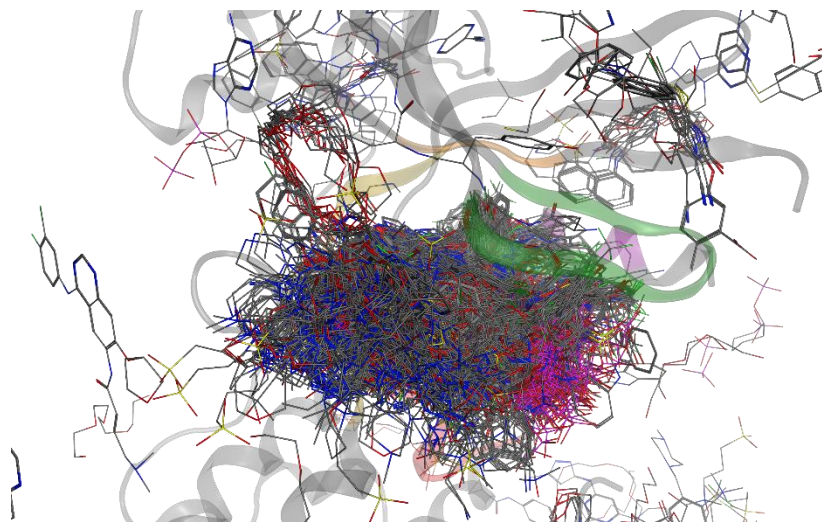


Figure 69. Représentation de fragments alignés dans le référentiel commun généré (exemple de la protéine kinase MELK. Code PDB : 4D2W. Pour des raisons de visibilité, seuls les 30 000 premiers fragments sur un total de 131 826 sont représentés.

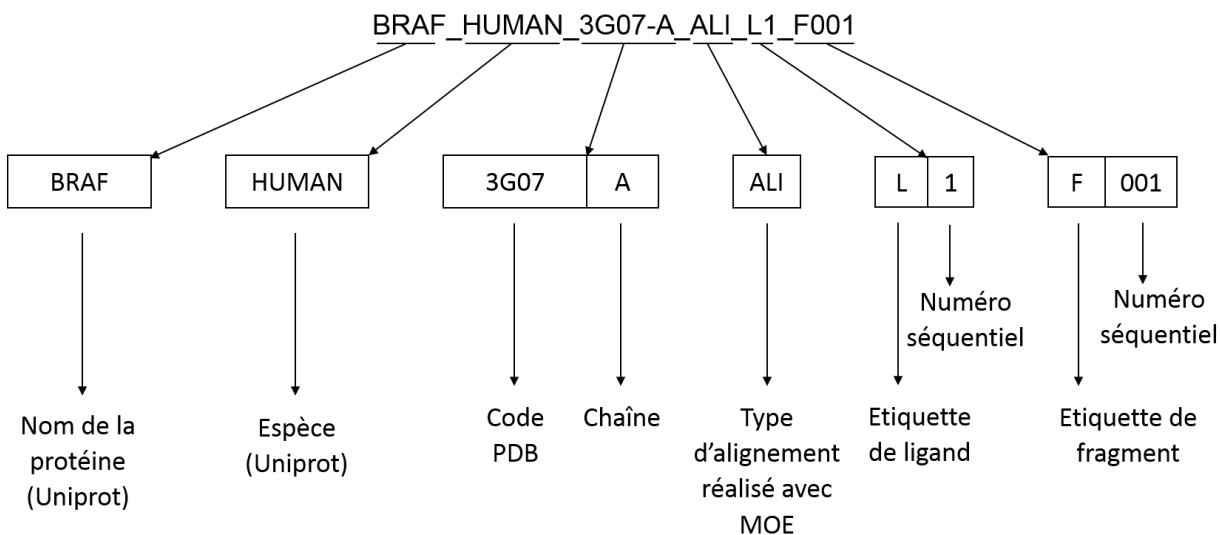


Figure 70. Schéma de la nomenclature retenue pour identifier les fragments.

4.2.3 Conception et implémentation de Frags2Drugs

Frag2Drugs est un outil de conception *de novo* de molécules *in silico*. Comme son nom l'indique, il utilise des fragments qu'il combine ensuite en des nouvelles molécules potentiellement bioactives. Cet assemblage s'effectue de manière itérative en combinant un fragment unique avec une succession d'autres fragments. Il est implémenté dans le langage informatique Python (version 3) et tire profit des bibliothèques RDKit pour le traitement de l'information chimique et Pandas pour la gestion optimisée des données sous forme de table (*dataframe*). Deux tables sont ainsi utilisées lors de l'utilisation de l'outil, une table contenant le fragment à considérer comme point de départ et une autre table contenant tous les autres fragments qui seront ainsi ajoutés au fur et à mesure lors du *fragment growing*. L'outil est structuré comme un paquet Python standard et peut donc être installé avec aisance. Des tests automatisés assurent la maintenabilité du code en permettant de vérifier qu'une mise à jour n'introduit pas de nouveaux bogues. L'outil est utilisable soit sous forme d'exécutable, en ligne de commande, soit sous forme de bibliothèque Python pour une utilisation plus poussée. L'architecture globale se décompose en 12 modules permettant d'effectuer les différentes tâches nécessaires au bon fonctionnement du programme (Figure 75).

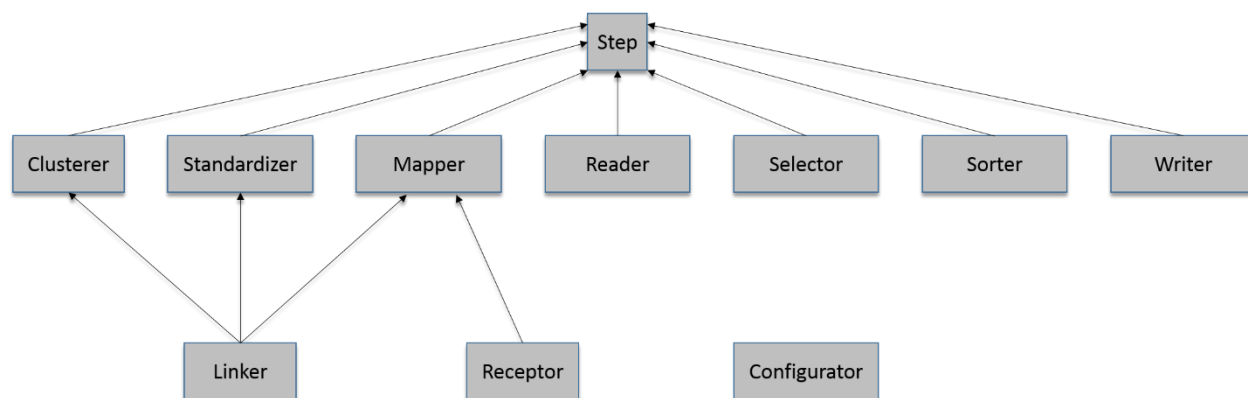


Figure 71. Architecture de Frags2Drugs.

Un bloc correspond à un module tandis que les flèches indiquent les relations d'héritage (réutilisation du code) entre les modules.

Chaque module représente une unité logique et indépendante dans la chaîne d'instructions nécessaires au bon déroulement des calculs. Ainsi, dans une utilisation typique de l'outil (**Figure 72**), la première étape consiste à lire le fichier moléculaire contenant les fragments afin de les charger en mémoire. La représentation de ces fragments est ensuite standardisée pour en assurer une bonne prise en charge. Optionnellement, il est possible de sélectionner un sous-ensemble de molécules selon divers critères. Les fragments résultants pouvant comporter des positions redondantes dans l'espace, une étape de classification (*clustering*) est effectuée pour identifier et retirer les doublons. Après quoi est réalisée une cartographie des fragments, ce qui permet d'annoter chacun d'entre eux en fonction de leur emplacement respectif et donc de déterminer les combinaisons possibles à considérer pour la suite. Les fragments peuvent être projetés dans le site actif d'une structure de protéine kinase afin de ne préserver uniquement ceux qui ne provoquent pas d'encombrement stérique avec des acides aminés du récepteur ou des molécules co-cristallisées (eau, cofacteurs, substrat, etc.). Toutes les conditions sont alors remplies pour effectuer l'opération de génération de nouvelles molécules en reliant les fragments. Finalement, les résultats sont sauvegardés pour une analyse ou un traitement ultérieur. Chacune de ces étapes est implémentée sous la forme d'un module qui exécute la tâche requise, comme nous allons le détailler dans les sous-parties suivantes.

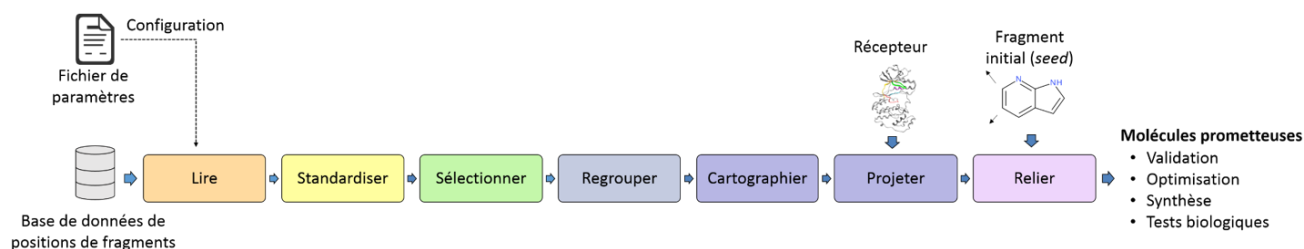


Figure 72. Protocole typique du programme Frags2Drugs

4.2.3.1 Configurator

Le premier module de l'outil permet de configurer le travail à exécuter en lisant un fichier de paramètres. Chaque option est définie pour une étape précise du protocole et le travail est effectué exactement dans l'ordre qui est spécifié.

Chaque étape a un nombre défini de paramètres attendus qui ont une valeur par défaut. Une vérification de chaque paramètre est effectuée afin de prévenir toute erreur de saisie et d'assurer le bon fonctionnement de l'outil. Il s'agit du seul module qui n'apparaît pas dans le fichier de configuration de manière explicite.

4.2.3.2 Reader et Writer

Ces modules permettent respectivement la lecture (*Reader*) et l'écriture (*Writer*) des molécules par le programme. Le format moléculaire utilisé est le SDF, mais il est également possible d'employer des fichiers binaires (HDF5). Ces derniers permettent la sauvegarde des molécules au format RDKit sous forme de table, ce qui permet de les relire très vite ultérieurement. Grâce à une option, l'utilisateur peut spécifier s'il souhaite traiter des fragments initiaux ou voisins.

Le chargement des fragments en mémoire se fait sous la forme d'une table grâce à la librairie Pandas telle qu'incorporée dans RDKit. Ce type de table permet de réunir à la fois l'objet moléculaire et les éventuelles informations qui y sont associées dans une structure de données intuitive à interroger et optimisée en accès.

Initialement, les molécules étaient lues uniquement au format SDF mais cela se traduisait par une somme de travail superflue puisqu'il fallait convertir les molécules au format RDKit à chaque lecture. Pour cette raison, les porteurs du projet RDKit ont développé des fonctions permettant de convertir directement un SDF en table Pandas. Cependant, l'implémentation officielle présente l'inconvénient de filtrer les molécules échouant à la conversion sans en avertir l'utilisateur. J'ai donc redéfini cette fonction directement dans le code source du module *Reader* en la modifiant de manière à pouvoir être alerté des erreurs de conversion des molécules au format RDKit.

4.2.3.3 Standardizer

Les divers travaux réalisés dans le cadre du projet VSPrep m'ont permis de mieux me rendre compte des difficultés à considérer lors de la prise en charge de molécules en chimoinformatique. Il m'a alors apparu pertinent d'inclure un module durant le processus pour standardiser les molécules de manière à s'assurer qu'elles soient toutes représentées de la même manière.

Ainsi des fonctions déjà mentionnées dans le chapitre précédent, telles que la conversion en tautomères canoniques et la représentation sous forme Kekulé, ont été implémentées pour les raisons décrites dans la partie 3.5.2.2. D'autres fonctions répondent quant à elles au besoin spécifique de former

de nouvelles liaisons sur une molécule avec la librairie RDKit. Ainsi, les hydrogènes sont indiqués de manière implicite pour éviter les erreurs de valence. Pour la même raison, les étiquettes sur les atomes chiraux sont supprimées, sans perte d'information puisque de toute façon, les fragments étant en trois dimensions, elles peuvent être recalculées facilement, et que l'ajout de nouveaux fragments peut les rendre caduc.

4.2.3.4 Selector

Selon l'étude menée, il peut se révéler intéressant de se focaliser sur un type de molécules donné. Une étape optionnelle de sélection a donc été incorporée dans le protocole de Frags2Drugs. Les molécules peuvent ainsi être filtrées selon plusieurs critères différents en tirant profit de la puissance des structures de données des tables Pandas. Il est par exemple possible de sélectionner des entrées en utilisant des comparaisons algébriques ou bien des expressions régulières. De plus, un adaptateur simple mais efficace a été développé pour permettre le calcul à la volée des descripteurs RDKit, ce qui permet de filtrer les molécules par poids moléculaire par exemple. Des filtres plus poussés ont été mis en place en combinant plusieurs descripteurs, pour accéder par exemple aux molécules *fragment-like* en autorisant un nombre maximum de violations des critères spécifiés. D'autres types de sélection envisageables n'ont pas encore été implémentés faute de temps mais présenteraient un intérêt certain, comme la recherche par sous-structure.

Table 19. Liste des méthodes actuellement implémentées dans Frags2Drugs pour sélectionner des sous-ensembles de fragments

Méthode	Description	Exemple
SelectByNumber	Extraie les fragments selon la valeur d'une propriété numérique	« SCORE <= -10 » pour extraire tous les fragments avec un score inférieur ou égal à -10.
SelectByRegex	Extraie les fragments selon la valeur d'une propriété textuelle	« BRAF.* » pour extraire tous les fragments issus de ligands co-cristallisés avec BRAF.
SelectByMolDescr	Calcule un descripteur moléculaire et extraie les fragments respectant la condition demandée	« NumHeavyAtoms >=4 AND NumHeavyAtoms <= 19 » pour extraire tous les fragments comportant entre 4 et 19 atomes lourds.
SelectByFilter	Calcule un ensemble de descripteurs moléculaires définissant un sous-ensemble et au choix annote ou filtre les fragments selon le nombre maximum de violations demandé (Lipinski, Veber, Drug-like, Lead-like,	« organic AND Fragment-like » pour extraire tous les fragments organiques et respectant la règle de 3.

	Fragment-like, Ext-fragment-like, PPI-like comme définis dans VSPrep*)	
--	--	--

* Voir partie 3.2.3

4.2.3.5 Sorter

Comme son nom l'indique, cette étape optionnelle permet de trier les fragments contenus selon une propriété donnée. Elle ne représente pas elle-même de grand intérêt sauf quand elle est combinée avec une autre étape. Un bon exemple d'application est décrit dans la sous-partie suivante.

4.2.3.6 Clusterer

Selon la façon dont la base de fragments a été assemblée, il se peut que certains d'entre eux soient présents plusieurs fois. Cela peut représenter une information complémentaire sur les positions possibles du fragment considéré mais cela peut aussi résulter en l'apparition d'information redondante dans le cas où ces positions se concentrent au même endroit.

Pour résoudre ce problème, j'ai incorporé une étape de classification ou *clustering* des positions 3D des fragments pour identifier les doublons par mesure de RMSD. Deux algorithmes ont été implémentés pour ne garder qu'un seul représentant par ensemble de poses (**Figure 73**). Le premier prend en référence le premier fragment lu par groupe et filtre chaque pose présentant une valeur de RMSD inférieure à un seuil défini. Il s'agit donc d'une approche supervisée, dont le biais est introduit par le classement des poses des fragments, qu'il faut donc trier auparavant selon la valeur à considérer, par exemple un score donné. Pour pallier ce biais, une seconde méthode, non supervisée, a été développée pour calculer toutes les combinaisons possibles de RMSD et regrouper les poses de fragments selon le seuil défini par l'utilisateur. Cette seconde approche entraîne cependant des temps de calculs plus longs, d'autant plus conséquents lorsqu'un fragment est représenté de nombreuses fois.

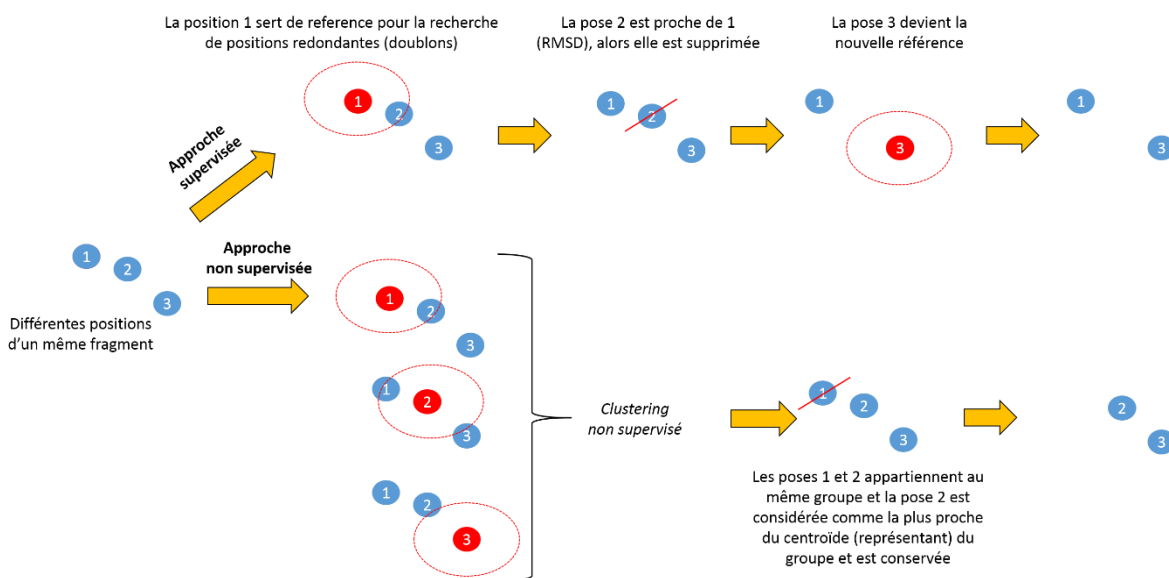


Figure 73. Regroupement des poses de fragments selon deux méthodes différentes, supervisée et non supervisée.

4.2.3.7 Mapper

L'une des principales conditions sur lesquelles repose le succès de cet outil est sa capacité à identifier rapidement quels fragments peuvent être assemblés. Dans la version prototype de Frags2Drugs, cette étape n'existait pas et chaque combinaison possible de fragment était testée, ce qui résultait en un temps de calculs exponentiellement long, dû notamment à de nombreux tests inutiles. En effet, mêmes s'ils sont dans un référentiel commun, tous les fragments, du fait de leur topologie ou de leur position respective, ne peuvent former des liaisons.

Une étape de cartographie a donc été développée afin de permettre d'identifier dans quel cas un fragment pouvait être relié à un autre. Ainsi, deux fragments peuvent être assemblés si chacun présente au moins un atome dont la valence autorise une nouvelle liaison (en retirant un atome d'hydrogène de part et d'autre) et si ces atomes se situent à une position favorable l'un par rapport à l'autre (distance de liaison, angles, etc.). Chaque atome pouvant former une nouvelle liaison est alors annoté en préparation de l'étape suivante.

Cas rencontré					
Valence	✗	✓	✓	✓	✓
Distance	✓	✗	✓	✓	✓
Angle de liaison	✓	✓	✗	✓	✓
Clash stérique	✓	✓	✓	✗	✓
Formation d'une liaison	✓	✗	✗	✗	✓

Table 20. Conditions requises pour assembler deux fragments. La liaison considérée est indiquée par des pointillés reliant les atomes candidats

Afin de déterminer plus rapidement les combinaisons de fragments possibles, un système de voxels a été mis en place. Ce terme, provenant de la fusion des mots « volume » et « élément », sert à indiquer une unité de volume dans l'espace, comme un pixel en 3D. Ainsi, le référentiel commun aux fragments est découpé en voxels et chaque fragment est annoté avec ses coordonnées en voxels, en arrondissement à la valeur supérieure de leurs coordonnées respectives en x, y et z. Cette approche emploie les fonctionnalités de requête de la librairie Pandas afin d'accéder très rapidement aux fragments présentant des voxels communs.

Dans le cas où deux fragments présentent au moins un voxel commun, un test est effectué pour vérifier s'ils n'ont pas des atomes trop proches voire superposés. Si tel est le cas, alors ils sont mutuellement exclus de l'étape de liaison des fragments. Dans le cas contraire, ils sont identifiés comme voisins et pourront potentiellement être combinés ultérieurement.

L'étape de cartographie de tous les fragments entre eux peut se révéler coûteuse en temps calculs. Pour cette raison elle n'est effectuée qu'une seule fois au début de l'importation de la base de fragments, puis à chaque ajout de positions de fragments dans la base de données.

4.2.3.8 Receptor

Lors de cette étape, la structure de la protéine kinase cible va être prise en compte dans les calculs. Ainsi, il est possible de retirer tous les fragments présentant un mauvais alignement avec la structure utilisée et donc d'éviter de générer des molécules ne rentrant pas dans le site actif d'intérêt. Cette vérification s'effectue en utilisant la même démarche de voxels que présentée dans la partie précédente, cette fois-ci appliquée pour identifier d'éventuels encombrements stériques entre les acides aminés du récepteur et les fragments alignés. On utilise pour cela la mesure du recouvrement des rayons de van der Waals (**Figure 74**). De manière plus générale, un paramètre permet de déterminer une distance minimale entre les atomes des fragments et ceux des acides aminés afin de prioriser la croissance des molécules assemblées vers des régions présentant plus d'espace libre dans le site actif.

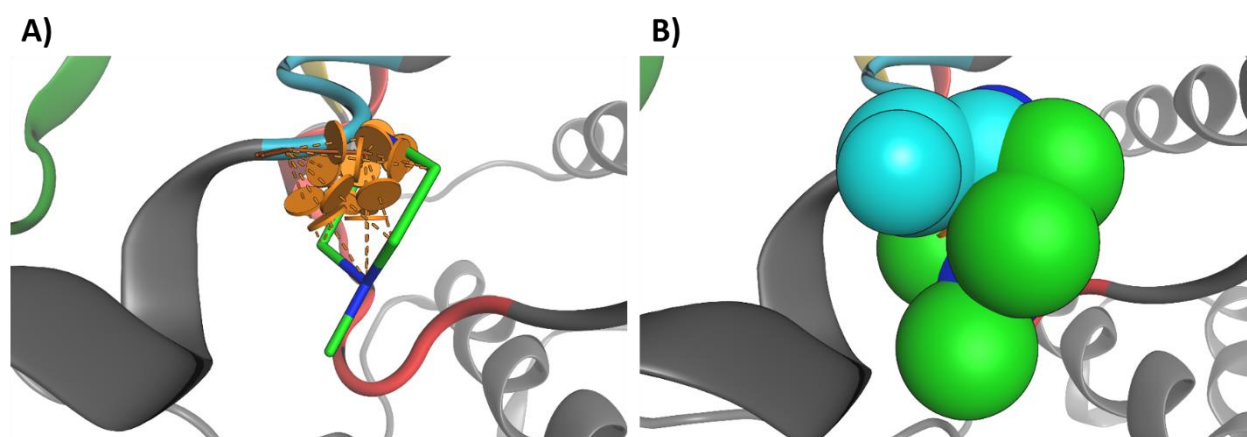


Figure 74. Détection d'un encombrement stérique entre une structure DFG-in de CDK2 et un fragment issu de l'imatinib par mesure des rayons de van der Waals. Codes PDB : structure DFG-in de CDK2 (2XMY), fragment issu de l'imatinib (2HYY). A) Représentation des encombrements dans MOE. B) Représentation du recouvrement des rayons de van der Waals entre le récepteur (bleu) et le fragment (vert).

Le module *Receptor* permet de considérer dans les calculs aussi bien la séquence d'acides aminés que n'importe quel cofacteur ou molécule de solvants jugés pertinents par l'utilisateur. Enfin, plusieurs structures peuvent être lues en même temps pour réaliser cette étape, afin de prendre en compte la flexibilité du domaine kinase d'une même protéine ou bien de concevoir une molécule capable d'inhiber plusieurs protéines kinases (« *dual inhibitors* »).

Ce module emploie la librairie Biopandas²⁸¹, préférée ici à la librairie Biopython plus largement diffusée et mentionnée précédemment dans la partie 4.2.2.1., car elle permet de travailler avec une table contenant directement les atomes de chaque acide aminé. Cela se traduit par une simplification importante de l'algorithme et évite de créer des boucles imbriquées pour itérer chacun des atomes de chaque résidu à comparer avec chaque atome de chaque fragment.

²⁸¹ S. Raschka, Biopandas: Working with Molecular Structures in Pandas DataFrames, *The Journal of Open Source Software*, 2.14 (2017), 279

4.2.3.9 Linker

L'étape d'assemblage des fragments constitue le cœur de ce projet *in silico* FBDD. Elle consiste à utiliser l'information calculée par le module « *Mapper* », contenant les informations des fragments voisins avec leurs atomes aptes à former de nouvelles liaisons (« atomes candidats »), afin de générer de nouvelles molécules de manière itérative. Les solutions trouvées en assemblant des fragments initiaux (FI ou *seeds*) avec des fragments voisins (FV) lors d'une itération constituent le point de départ de la suivante (Figure 75).

Cette étape comporte actuellement encore plusieurs limitations. Premièrement, il faut tester chaque combinaison d'atomes candidats de chaque combinaison de fragments initiaux et de fragments voisins. L'algorithme décrit résulte en effet en une boucle de complexité 4 (« fragment initial (1)-atome candidat (2) avec fragment voisin (3)-atome candidat(4) »), ce qui est extrêmement coûteux en temps de calculs. Deuxièmement, une contrainte technique est à imputer à la librairie RDKit. En effet, le code source actuel de l'outil ne permet la comparaison avec des valeurs de distances et d'angles dans un champ de force que si la liaison à tester existe déjà. Il est donc nécessaire de construire d'abord les molécules en se basant uniquement sur les informations de proximité (fragments voisins) et d'atomes candidats (apte à former de nouvelles liaisons), ce qui peut aboutir à des géométries peu vraisemblables, puis de les filtrer ultérieurement. Une alternative aurait pu être d'établir un dictionnaire regroupant toutes les valeurs propres au champ de force, c'est-à-dire chaque combinaison possible de distances et d'angles de liaison et d'angles dièdres mais ces valeurs dépendant de plusieurs paramètres (type d'atome, état d'hybridation, charge). Ce dictionnaire aurait donc été fastidieux à écrire et complexe à interroger. Troisièmement, la valeur énergétique utilisée pour filtrer les molécules assemblées de conformation défavorable est arbitraire, même si une étude a été réalisée pour estimer une valeur appropriée (Figure 76). Nous avons préféré laisser un seuil énergétique bas (500 kcal.mol⁻¹) car même si cela signifie que quelques ligands cristallographiques ne pourront pas être reconstruits (car présentant des conformations contraintes), la qualité globale des molécules assemblées en sera renforcée.

Une solution pour contourner la première limitation décrite est proposée dans la partie 4.4, dans les perspectives de développement de Frags2Drugs, tandis que des alternatives sont encore en cours d'étude pour les deux autres.

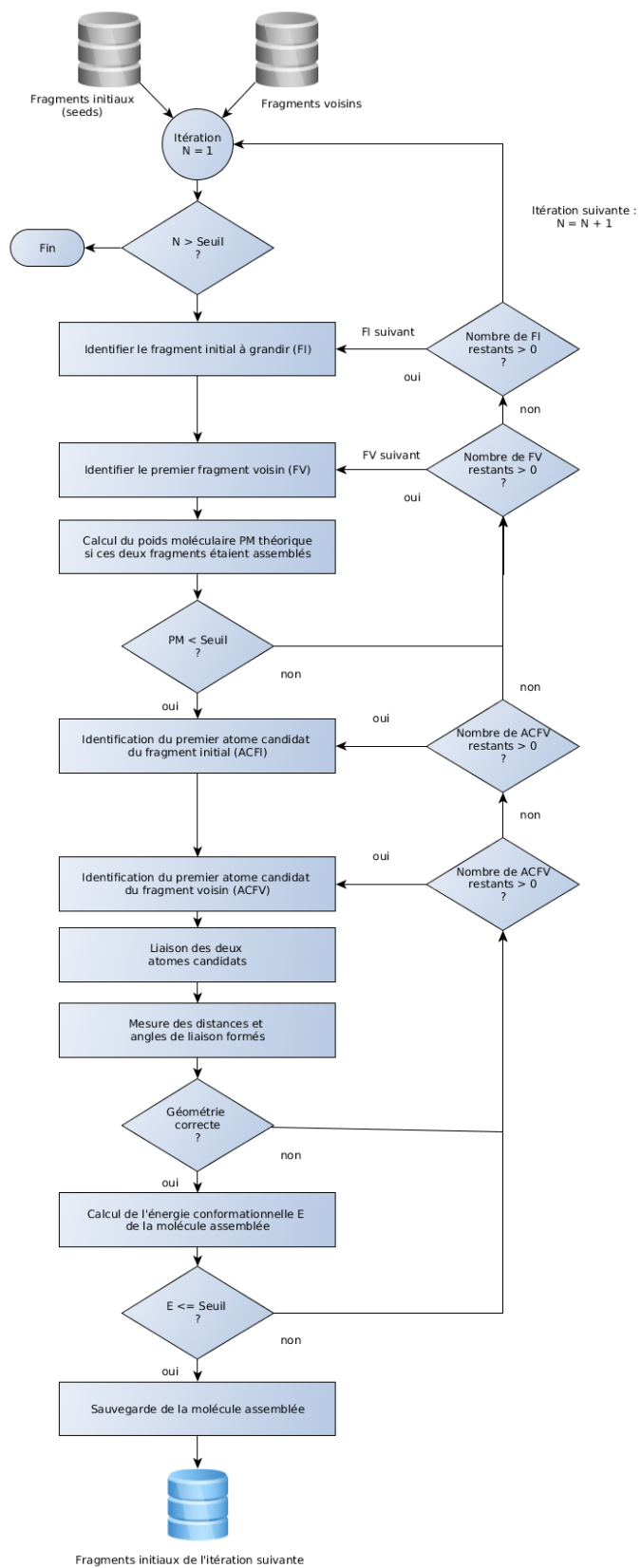


Figure 75. Algorithme du module Linker

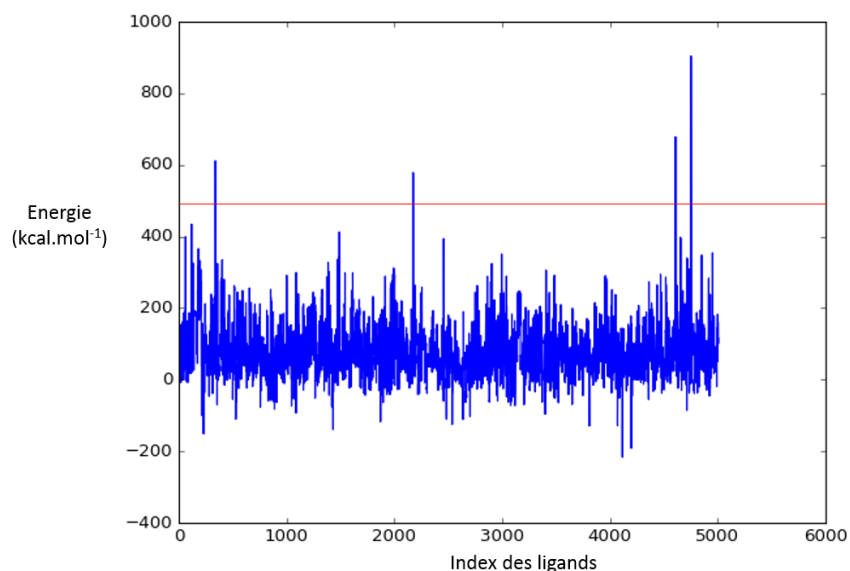


Figure 76. Estimation de l'énergie conformationnelle des ligands co-cristallisés mesurée dans le champ de force MMFF94s. La droite rouge indique la valeur du seuil choisi (500kcal.mol⁻¹)

4.2.4 Validation de l'outil

L'approche a été validée de deux manières. Tout d'abord, des inhibiteurs de protéines kinases co-cristallisés ont été fragmentés en amont et les fragments résultants ont été utilisés par l'outil pour reconstruire la molécule initiale.

Enfin, notre collaboration avec une équipe de chimie de synthèse experte dans la chimie des hétérocycles nous permet d'identifier des molécules présentant un certain potentiel comme inhibiteurs de protéines kinases et démontre l'utilité de l'outil dans des projets de recherche en cours.

4.2.4.1 Preuve de concept *in silico*

La première forme de validation de l'outil a consisté à vérifier que les fragments issus d'une molécule co-cristallisée pouvaient être réassemblés en la molécule d'origine. Pour ce faire différents inhibiteurs de 4 protéines kinases (BRAF, MELK, p38, Abl) co-cristallisés ont été fragmentés puis recomposés avec succès en n'utilisant que les fragments correspondants. (**Figure 77**).

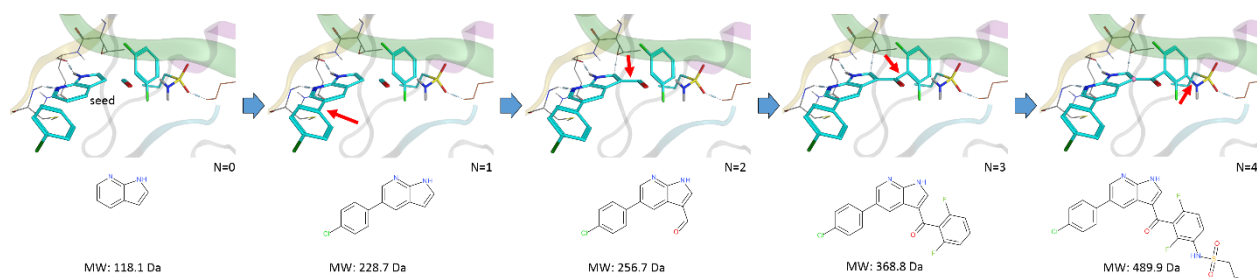
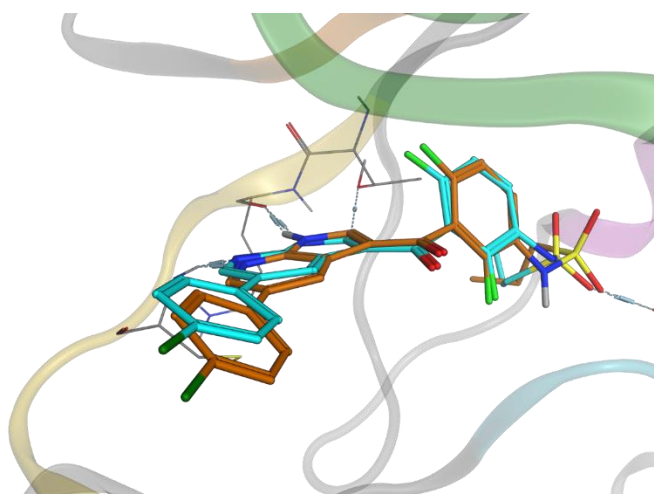


Figure 77. Exemple de la reconstruction itérative de vemurafenib co-cristallisé avec BRAF. Code PDB : 3OG7. Les flèches rouges indiquent la liaison effectuée à chaque étape N. Les flèches bleues indiquent le passage à une nouvelle itération.

Cette vérification a notamment permis de mettre en évidence des limitations dans la version prototype de l'outil, où même des fragments pourtant en position idéale pour être reliés ne produisaient pas toujours les molécules initiales. Tout d'abord, la conformation de certains ligands co-cristallisés présente parfois des valeurs de distances et/ou d'angles de liaisons s'écartant sensiblement du champ de force employé, d'où la nécessité d'inclure une certaine tolérance lors de la vérification de ces paramètres. De plus, certains choix de conception de la librairie RDKit aboutissent à des cas de figure inattendus. Cette étape de validation et de réglage minutieux et d'optimisation de l'outil a représenté un travail conséquent lors de mes trois années de thèse.

La superposition du ligand cristallographique et de la molécule assemblée a été évaluée par une mesure de RMSD et une inspection visuelle a été réalisée pour contrôler les poses obtenues (**Figure 78**).



BRAF (vemurafenib)
 pIC_{50} : 7.7
 MW: 489.9 Da

Figure 78. Exemple de vérification du mode d'interaction protéine-ligand avec vemurafenib. Code PDB : 3OG7. Ligand brun : référence cristallographique, ligand bleu : ligand réassemblé et amarré.

Afin de valider le mode d'interaction de la conformation des molécules obtenues par Frags2Drugs, chaque molécule est amarrée (« *dockée* ») ultérieurement dans le site actif de la protéine et le RMSD entre les deux conformations (issues de Frags2Drugs et du docking) est comparé. Pour réaliser cet amarrage moléculaire (*docking*), nous avons employé un outil libre et *open-source* d'amarrage moléculaire déjà utilisé en routine au sein de notre laboratoire. Ce logiciel, appelé rDock, regroupe des programmes en C++ et fortran ainsi que des scripts en Python et Perl.¹¹¹

L'outil prend en entrée le récepteur au format Mol2 et les ligands à amarrer au format SDF. Il est recommandé par les développeurs de l'outil de préparer le récepteur et les ligands avec des outils dédiés en amont de l'amarrage moléculaire. Nous avons dans notre cas employé les structures des récepteurs préparés avec MOE comme décrit précédemment et réaliser un amarrage moléculaire sans contrainte. Les ligands correspondent quant à eux aux fragments issus des inhibiteurs co-cristallisés. Nous avons dans notre cas défini la cavité d'un site actif en utilisant son ligand initial.

La génération des poses repose sur l'utilisation d'un algorithme génétique. Ce genre d'approche est directement inspiré du vivant et plus particulièrement des phénomènes aléatoires de mutations et de recombinaison des chromosomes. Ici le chromosome consiste en le centre de masse du ligand, l'orientation de ce dernier, ainsi que les angles dièdres des liaisons qui tournent d'une part sur le ligand et d'autre part sur le récepteur. Chaque pose est le résultat d'un processus complet mélangeant alternativement algorithme génétique (GA) et Monte Carlo.

Tout d'abord, la position (centre de masse) et les angles dièdres du ligand sont transformés de manière aléatoire afin de le positionner au sein du site actif. Ensuite, les composants du chromosome sont mutés de manière aléatoire (centre de masse, orientation et angles dièdres du ligand et du récepteur), ce qui permet de varier les interactions protéine-ligand. Une génération est considérée avoir réussi lorsque le nombre de nouveaux individus créés. Il n'y a pas de limite de génération, le GA continue donc jusqu'à convergence, c'est-à-dire lorsque le meilleur score d'interaction ne varie que très peu ($0,1\text{kcal.mol}^{-1}$) pendant trois générations consécutives.

Il s'ensuit alors une étape de Monte Carlo à basse température, méthode de dynamique moléculaire simplifiée permettant de modéliser des sauts aléatoires d'un système initial à des systèmes finaux possibles. Cette étape a pour but perturber le ligand et de lui permettre de sortir de son puit énergétique. Enfin, une nouvelle étape de GA est réalisée, cette fois-ci pour minimiser la pose formée avec le récepteur. Les auteurs suggèrent 100 poses par petite molécule pour réaliser un criblage exhaustif, c'est à dire censé couvrir l'ensemble des possibilités d'interaction du ligand. Pour le criblage de données plus volumineuses, le temps de calculs d'un amarrage exhaustif serait rédhibitoire. De plus, les auteurs estiment qu'au bout de 5 poses, on a au moins une valeur raisonnable pour le score du ligand. Il est donc possible de réaliser le criblage virtuel en plusieurs étapes, avec seulement 5 poses, pour pouvoir rapidement filtrer les molécules ne pouvant pas se lier avec un bon score au sein du site actif. Les molécules restantes sont très minoritaires et peuvent donc être criblées de manière exhaustive.²⁸² Il est à noter que l'outil ne varie pas les angles dièdres des liaisons impliquées dans des cycles, il est donc

²⁸² Multi-Step Protocol for HTVS | rDock <<http://rdock.sourceforge.net/multistep-protocol-for-htvs/>> [consulté le 18 Octobre 2017]

nécessaire de générer de multiples conformations en amont pour les molécules contenant des cycles flexibles.

Afin d'orienter la génération de la pose vers des solutions raisonnables, l'affinité des ligands avec la cavité est estimée à l'aide d'une fonction de score. Cette dernière comprend des termes intermoléculaires (cavité et ligand) ou intramoléculaires (cavité ou ligand). Il est également possible de biaiser le score en ajoutant des contraintes extérieures (**Equation 5**).

Equation 5

$$S^{total} = S^{inter} + S^{intra} + S^{site} + S^{restraint}$$

S^{inter} est le score représentant le plus d'intérêt pour notre étude car il traduit les interactions récepteur-ligand. S^{intra} représente l'énergie interne du ligand (liaison et angles dièdres). S^{site} représente quant à lui l'énergie interne des chaînes latérales des acides aminés des régions flexibles du site actif et permet d'appliquer une pénalité sur les poses dont des atomes lourds sortent hors de la cavité. Enfin, $S^{restraint}$ est utilisé pour introduire des points d'ancrage du ligand ou encore des pharmacophores dans le calcul du score total.

L'outil rDock a été employé pour reproduire les interactions protéines-ligands cristallographiques. J'ai tenté dans un premier temps de réaliser un amarrage moléculaire sans contraintes afin de replacer chaque fragment à son emplacement initial dans le site actif. Cependant, cela n'a pas été possible pour deux raisons évidentes.

Tout d'abord, comme expliqué dans la partie 4.1.2, les fragments ne constituent que des liaisons faibles avec leur récepteur. Les outils d'amarrage moléculaire ont donc en général du mal à les positionner de manière adéquate dans le récepteur.

rDock n'est pas non plus spécifiquement paramétré pour effectuer de l'amarrage moléculaire de fragments contrairement à d'autres outils d'amarrage moléculaire tels que SEED^{283,284} ou S4MPLE²⁸⁵. Pour cette raison, il n'est pas possible de trier les poses en fonction du score pour ne garder que les « meilleures » poses comme cela se fait habituellement dans les expériences d'amarrage moléculaire. D'autre part, tous les fragments inclus dans un ligand ne forment pas forcément une interaction avec la

²⁸³ N. Majeux *et al.*, Exhaustive Docking of Molecular Fragments with Electrostatic Solvation, *Proteins*, 37.1 (1999), 88–105

²⁸⁴ N Majeux, M Scarsi, and A Caflisch, Efficient Electrostatic Solvation Model for Protein-Fragment Docking, *Proteins*, 42.2 (2001), 256–68

²⁸⁵ L. Hoffer, Jean-Paul Renaud, and Dragos Horvath, In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol Using S4MPLE, *Journal of Chemical Information and Modeling*, 53.4 (2013), 836–51

protéine, en particulier lorsqu'il s'agit de fragments de liaison du ligand. L'outil n'est alors pas capable de retrouver la position cristallographique du fragment.

L'approche retenue a donc été d'utiliser directement les molécules assemblées pour la réalisation de l'amarrage moléculaire. Afin de ne pas biaiser les résultats, nous avons généré des conformations tridimensionnelles « aléatoires » des ligands avant chaque *docking*.

4.2.4.2 Validation in vitro

La preuve de concept *in silico* de l'outil permet de s'assurer qu'il fonctionne correctement, mais il reste encore à prouver son utilité. J'ai eu pour cela l'opportunité de collaborer avec l'équipe de synthèse organique du Pr G. Guillaumet, et plus particulièrement avec le doctorant M. Loubidi et le chercheur post-doctorant M. Driowya, au cours de mes travaux de thèse. J'ai effectué l'amarrage d'échafaudages moléculaires originaux (aucune structure renseignée dans Scifinder²⁸⁶), préparés avec VSPrep, ayant une forte similarité avec le noyau adénine de l'ATP (**Figure 79**) et présentant un intérêt en synthèse organique.²⁸⁷ Les poses les plus intéressantes ont alors été retenues comme points de départ du projet FBDD. L'outil a été employé avec la base de fragments, issus d'inhibiteurs de kinases co-cristallisés et fragmentés, décrite plus tôt (4.2.2.2) afin de proposer des structures aux chimistes organiciens qui ont ensuite priorisé les voies de synthèse correspondant le mieux à leur expertise. La synthèse de certains de ces composés s'est révélée très complexe, comme indiqué d'ailleurs dans le manuscrit de thèse de M. Loubidi.

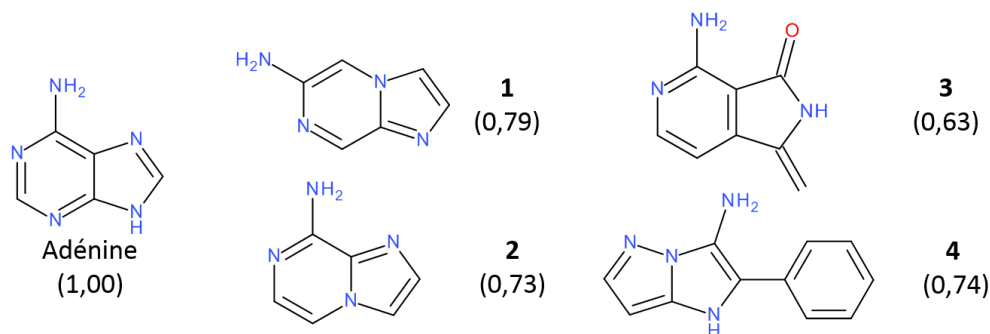


Figure 79. Quelques exemples d'échafaudages moléculaires utilisés dans le projet FBDD. La similarité du coefficient de Tanimoto avec le noyau adénine de l'ATP est indiquée entre parenthèses.

L'amarrage moléculaire a été réalisé avec le logiciel rDock selon le protocole décrit précédemment. Des poses formant des interactions intéressantes ont été sélectionnées en inspectant les résultats visuellement. La direction de la croissance du fragment, définie par le choix des positions à grandir, a été décidée en fonction de l'orientation de la pose de départ considérée.

²⁸⁶ D. D. Ridley, Strategies for Chemical Reaction Searching in SciFinder, *Journal of Chemical Information and Computer Sciences*, 40.5 (2000), 1077–84

²⁸⁷ M. Loubidi, Synthèse et Réactivité de Bicycles Imidazo[1,2- α]imidazo[1,5- α]imidazoles à Visée Thérapeutique (Université d'Orléans, 2017)

Des travaux internes précédents ont orienté le choix des premières protéines kinases cibles considérées, GSK3 β et BRAF, et d'autres travaux sont actuellement en cours sur KIT, CDK5 et p38 α .

4.2.4.2.1 Application sur GSK3 β

Initialement découverte pour son rôle dans la régulation de la synthèse de glycogène, la *glycogen synthase kinase 3* (GSK3) est une protéine kinase à activité Ser/Thr (groupe CMGC). Malgré ce que laisse entendre son nom, son rôle ne se limite pas à la phosphorylation de la *glycogen synthase* (GS) puisqu'on lui compte plus de 100 substrats connus à ce jour.²⁸⁸ Ainsi, la dérégulation de son activité est associée à de nombreux dysfonctionnements physiologiques et psychiatriques graves tels que la dépression bipolaire²⁸⁹, les maladies neurodégénératives d'Alzheimer²⁹⁰ et de Parkinson²⁹¹, le diabète²⁹², des maladies auto-immunes²⁹³ ou encore différents types de cancer^{294,295} (sein, colon, etc.), ce qui en fait une cible thérapeutique majeure. Cette protéine se dérive majoritairement en deux isoformes, GSK3 α et GSK3 β , qui se distinguent principalement par leur région N- et C-terminales. Leur site actif ne diffère que par un seul acide aminé (Glu196 pour GSK3 α et Asp133 pour GSK3 β), ce qui rend la tâche de développer un inhibiteur spécifique de l'une des isoformes extrêmement difficile²⁹⁶, même si certains inhibiteurs spécifiques de l'isoforme α ont déjà pu être développés par le passé.²⁹⁷

A notre connaissance, il n'existe actuellement aucune structure de GSK3 β en conformation inactive (*DFG-out* ou *α C-helix out*). J'ai donc utilisé une structure en conformation active de cette protéine kinase, d'une résolution faible (2.1 Å) et déjà employée dans un projet interne (code PDB : 1Q41). Plusieurs poses intéressantes ont été retenues et à titre d'exemple, nous nous attarderons sur l'une d'elles en particulier. La pose représentée sur la **Figure 80** forme deux liaisons hydrogène avec les résidus Asp99 et Val101 de la région charnière de la protéine kinase. De plus, il est possible de la grandir de façon à explorer la poche hydrophobe allostérique adjacente à la poche de fixation de l'ATP. Ainsi cette pose constitue un bon point de départ pour orienter la conception *de novo* vers un inhibiteur de type I½ (**Figure 81**).

²⁸⁸ E. Beurel *et al.*, Glycogen Synthase Kinase-3 (GSK3): Regulation, Actions, and Diseases, *Pharmacology & Therapeutics*, 0 (2015), 114–31

²⁸⁹ A. Luca *et al.*, Gsk3 Signalling and Redox Status in Bipolar Disorder: Evidence from Lithium Efficacy, *Oxidative Medicine and Cellular Longevity*, 2016 (2016)

²⁹⁰ M. Llorens-Marín *et al.*, GSK-3 β , a Pivotal Kinase in Alzheimer Disease, *Frontiers in Molecular Neuroscience*, 7 (2014)

²⁹¹ M. Golpich *et al.*, Glycogen Synthase Kinase-3 Beta (GSK-3 β) Signaling: Implications for Parkinsons Disease, *Pharmacological Research*, 97 (2015), 16–26

²⁹² S. AlFadhli *et al.*, De-Regulation of Diabetic Regulatory Genes in Psoriasis: Deciphering the Unsolved Riddle, *Gene*, 593.1 (2016), 110–16

²⁹³ I. Nasri *et al.*, Anti-Inflammatory and Anticancer Effects of Flavonol Glycosides from *Diplotaxis Harra* through GSK3 β Regulation in Intestinal Cells, *Pharmaceutical Biology*, 55.1 (2017), 124–31

²⁹⁴ A. Felipe Aristizabal-Pachon *et al.*, Role of GSK3 β in Breast Cancer Susceptibility, *Cancer Biomarkers: Section A of Disease Markers*, 18.2 (2017), 169–75

²⁹⁵ J. A. McCubrey *et al.*, GSK-3 as Potential Target for Therapeutic Intervention in Cancer, *Oncotarget*, 5.10 (2014), 2881–2911

²⁹⁶ R. V. Bhat *et al.*, Glycogen Synthase Kinase 3: A Drug Target for CNS Therapies, *Journal of Neurochemistry*, 89.6 (2004), 1313–17

²⁹⁷ F. Lo Monte *et al.*, Identification of Glycogen Synthase Kinase-3 Inhibitors with a Selective Sting for Glycogen Synthase Kinase-3 α , *Journal of Medicinal Chemistry*, 55.9 (2012), 4407–24

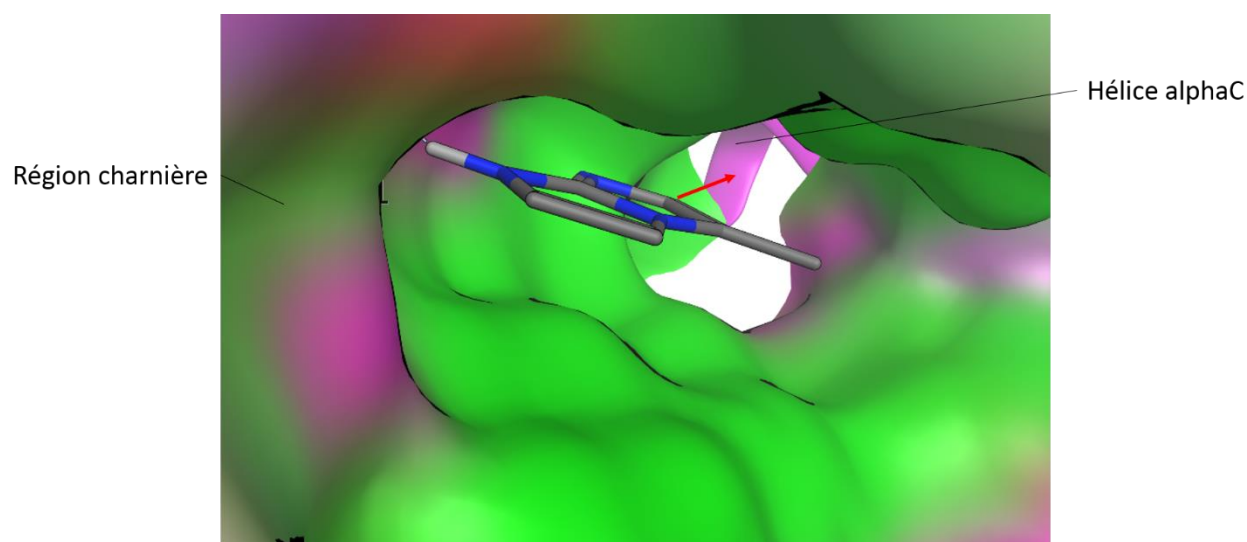
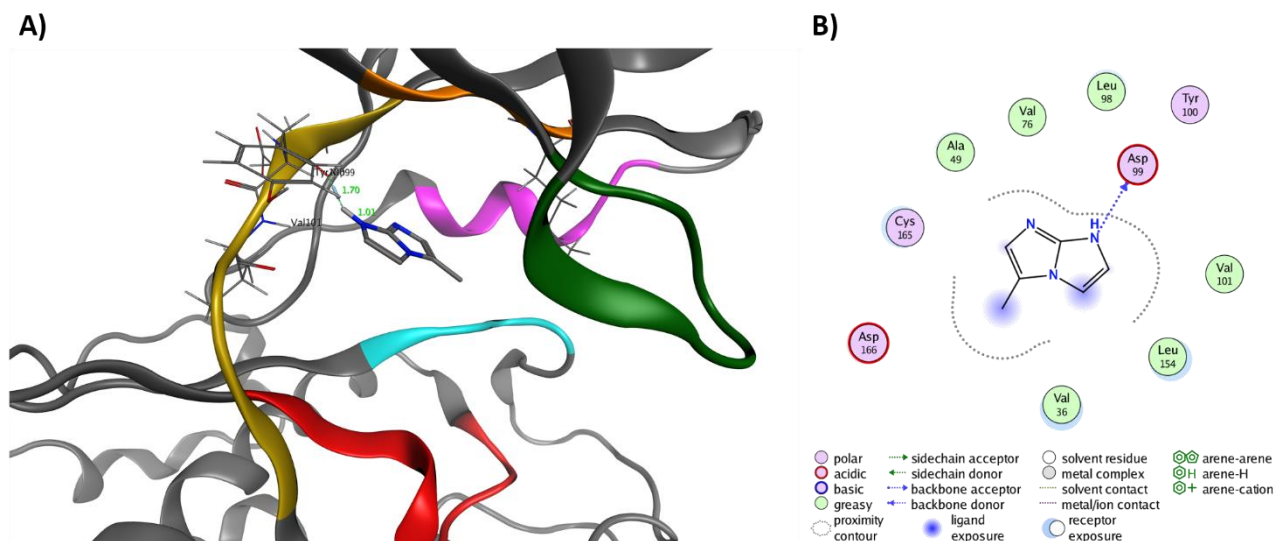


Figure 81. Positionnement du fragment initial au sein du site actif de GSK3β représenté sous forme de surface moléculaire. Code PDB : 1Q41. Une flèche rouge indique dans quelle direction la croissance de la molécule va s'orienter. Code de couleur de la surface : bleu : polaire, gris : hydrophobe, rouge : exposé au solvant.

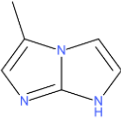
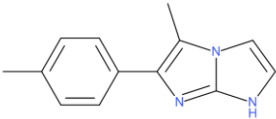
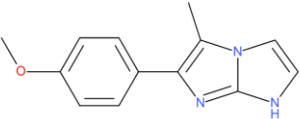
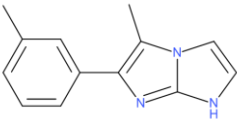
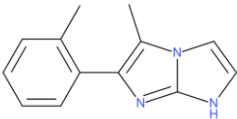
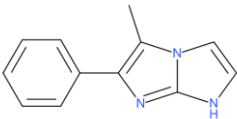
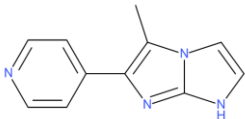
Les travaux du Dr. Loubidi ont abouti à la synthèse de six molécules d'une série imidazo[1,2- α]imidazoliques, identifiées au bout d'une étape de grandissement. Leur efficacité ainsi que leur sélectivité *in vitro* a été mesurée par le Dr. S. Ruchaud de la Station Biologique de Roscoff²⁹⁸ sur GSK3 (mélange de GSK3 β et GSK3 α). Le test biologique développé ne permet pas de tester spécifiquement l'isoforme β . L'activité résiduelle a été mesurée par criblage primaire à deux concentrations d'inhibiteurs différentes, 10 et 30 μ M. La relation dose-réponse (8 concentrations) observée pour les composés 46 et

²⁹⁸ B. Baratte *et al.*, Le criblage à Roscoff - Une recherche d'inhibiteurs de kinases tournée vers la mer, *médecine/sciences*, 31.5, 538-45

48 a permis d'obtenir leur valeur d'IC₅₀ et donc de calculer leur LE selon la formule décrite précédemment (Equation 4).

Table 21. Mesure du pourcentage d'activité résiduelle de la protéine kinase GSK3 α/β .

Les symboles « - » indiquent des valeurs non mesurées. * HA : Atome lourd

Structure	Référence	Concentrations (μ M)	Activités résiduelles (%)	IC ₅₀ (μ M)	pIC ₅₀	LE (kcal.mol ⁻¹ .HA ⁻¹)*
	Fragment initial	10 30	- -	- -	- -	-
	25	10 30	119 107	- -	- -	-
	46	10 30	107 31	0.63	6.2	0.51
	47	10 30	121 81	- -	- -	-
	48	10 30	74 23	1.69	5.8	0.51
	49	10 30	126 110	- -	- -	-
	50	10 30	139 65	- -	- -	-

Les composés 46 et 48 ont d'excellentes valeurs de LE (0.51 dans les deux cas soit bien supérieur au seuil recommandé de 0.3), ce qui en font donc de très bons points de départ pour poursuivre la conception d'inhibiteurs efficaces de GSK3 α/β . Une synthèse est en cours afin d'améliorer l'activité biologique de ces deux composés. De plus, l'activité du fragment initial sera également mesurée puisque cette donnée manque actuellement. Cependant, il est possible que des problèmes de solubilité émergent pour ce fragment.

4.2.4.2.2 Application sur BRAF

BRAF est un membre de la famille de Ser/Thr protéines kinases Raf (groupe TKL). Elle joue un rôle central dans la signalisation moléculaire et plus particulièrement dans la cascade de signalisation MAPK, également appelée voie Ras-Raf-MEK-ERK. Son activité participe à la prolifération et la survie cellulaires et la dérégulation de son activité à cause de mutations peut engendrer plusieurs types de cancers (colorectal, thyroïdien papillaire, etc.) et plus particulièrement du mélanome. En effet, environ 50% des patients atteints de cette maladie présentent une mutation du domaine kinase de BRAF, très majoritairement V600E.^{299,300}

La chimiothérapie la plus utilisée aujourd'hui pour ces patients est l'utilisation de l'inhibiteur de protéine kinase vemurafenib⁴⁷ (type 1 ½), le premier inhibiteur de protéines kinases conçu à partir de fragments et approuvé par la FDA en 2011. Cette molécule est assez spécifique de la forme mutée et est moins efficace contre la forme sauvage. D'autres inhibiteurs de protéines kinases sont alors utilisés, comme sorafenib ou une combinaison de trametinib (MEK) et dabrafenib (BRAF) mais ils présentent un profil sélectif bien moins intéressant et, comme pour vemurafenib, leur application conduit donc à des effets secondaires chez les patients traités.

La recherche de nouveaux inhibiteurs de BRAF sauvage constitue donc un projet intéressant pour une application de mon outil. Nous sommes cette fois-ci partis d'une proposition d'échafaudage moléculaire proposée par le Pr. P. Bonnet suite à des travaux précédents³⁰¹. Nous avons tout d'abord amarré l'échafaudage moléculaire pour déterminer son mode de fixation sur une structure interne de BRAF sauvage. Une pose en particulier nous a paru intéressante car elle se situe au niveau de la région charnière et son orientation dans l'espace permet un grandissement vers le fond de la poche hydrophobe (**Figure 82**).

²⁹⁹ H. Davies *et al.*, Mutations of the BRAF Gene in Human Cancer, *Nature*, 417.6892 (2002), 949–54

³⁰⁰ T. K. Eigentler *et al.*, Protein Kinase Inhibitors in Melanoma, *Expert Opinion on Pharmacotherapy*, 14.16 (2013), 2195–2201

³⁰¹ R. Arora *et al.*, Structural Investigation of B-Raf Paradox Breaker and Inducer Inhibitors, *Journal of Medicinal Chemistry*, 58.4 (2015), 1818–31

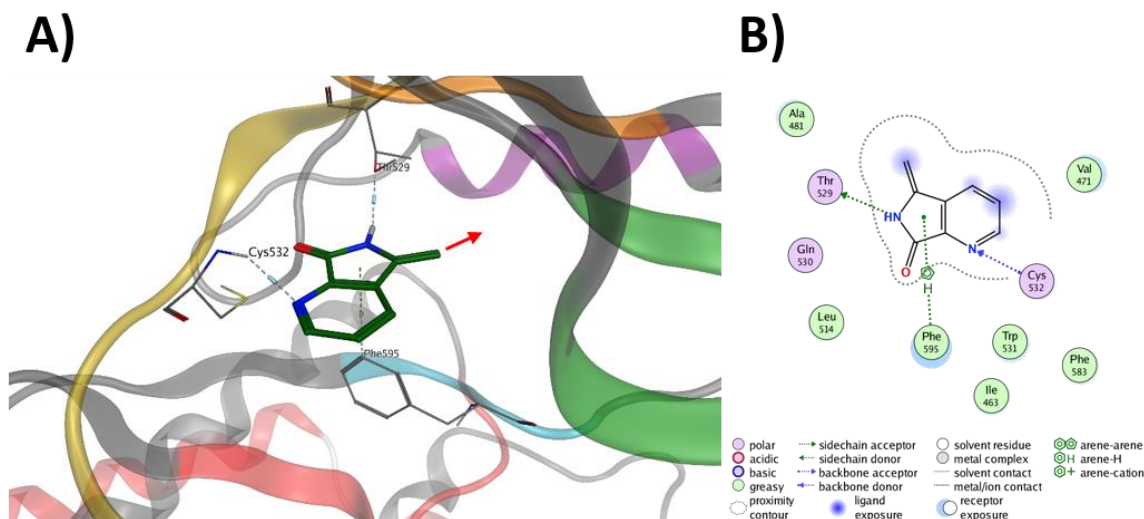


Figure 82. Pose du fragment initial dans l'application sur BRAF. A) Représentation du fragment lié à la région charnière. Les flèches en rouges indiquent la direction donnée à la croissance de la molécule. B) Diagramme d'interaction du fragment initial.

Nous avons analysé les résultats au bout d'une seule étape de grandissement. Il se trouve que parmi les fragments voisins se trouvent des fragments issus de l'inhibiteur de BRAF AZ-628 ($IC_{50(V600E)} = 30$ nM et $IC_{50(WT)} = 100$ nM).³⁰² Notre fragment initial occupant une position similaire à l'échafaudage moléculaire de AZ-628 (code PDB : 5HID), des molécules hybrides ont pu être identifiées parmi les solutions proposées (**Figure 83 A**), comme cela aurait pu être découvert dans une approche de *scaffold hopping*.

Cependant, un paramétrage fin des différentes tolérances géométriques est encore nécessaire pour affiner les résultats. En effet, dans certains cas la géométrie de la molécule assemblée *in situ* est suffisamment contrainte pour qu'il ne soit pas possible de reproduire le mode d'interaction attendu (**Figure 83 B**). Pour le moment, nous avons dépriorisé ces composés virtuels pour nous focaliser uniquement sur les solutions les plus prometteuses à partir d'autres échafaudages moléculaires. Des travaux actuels sont en cours au sein de notre équipe pour mieux définir les contraintes géométriques requises pour sauvegarder une solution.

³⁰² C. Montagut *et al.*, Elevated CRAF as a Potential Mechanism of Acquired Resistance to BRAF Inhibition in Melanoma, *Cancer Research*, 68.12 (2008), 4853–61

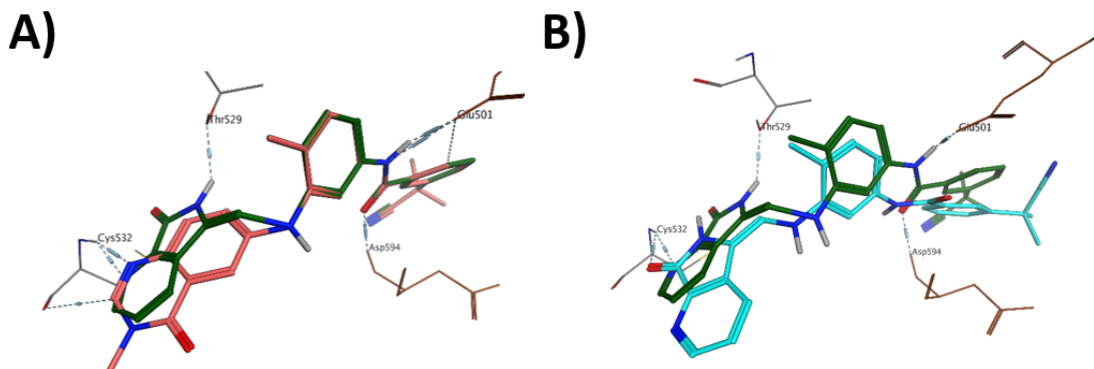


Figure 83. Exemple de résultat de l'approche Frags2Drugs dans une structure de BRAF. A) Superposition entre la molécule assemblée (vert) et l'inhibiteur AZ-628 (code PDB : 5HID). B) Validation du mode d'interaction entre la molécule assemblée telle qu'identifiée par l'outil (vert) et la solution amarrée (blue). La valeur du RMSD entre ces dernières est 2,7 Å.

4.2.4.2.3 Conclusion sur la validation *in vitro*

D'autres travaux n'ont pas été présentés dans ce manuscrit de thèse. Au total, treize échafaudages moléculaires ont été amarrés dans différentes protéines kinases, dont BRAF (cible biologique étudiée dans l'équipe SB&C de l'ICOA en innovation thérapeutique), GSK3 β et CDK5 (tests biologiques accessibles et donc possibilité de valider l'outil Frags2Drugs) ainsi que KIT (cible biologique étudiée dans l'équipe SB&C de l'ICOA en cosmétique). Les composés les plus intéressants seront testés biologiquement une fois leur synthèse terminée, ce qui servira de preuve de concept et permettra de démontrer l'utilité de l'outil Frags2Drugs dans des projets de FBDD.

4.3 Conclusion et perspectives

Ce travail de thèse a permis l'initiation du développement d'un outil de conception *in silico* d'inhibiteurs de protéines kinases à partir de fragments. J'ai bénéficié de la collaboration établie avec l'équipe de synthèse du Professeur G. Guillaumet pour appliquer mon travail au développement d'inhibiteurs et ainsi faire la preuve de concept de Frags2Drugs.

De plus, les connaissances acquises lors du développement de l'autre outil conçu au cours de cette thèse, VSprep, m'ont permis de mieux appréhender les limitations de la représentation des molécules en chémoinformatique et donc de mieux tirer profit de la librairie RDKit. J'ai également eu l'occasion de me familiariser avec la librairie Pandas et son approche par table, permettant d'optimiser beaucoup d'aspects limitants rencontrés lors de l'élaboration du premier prototype de Frags2Drugs, non présenté dans ce manuscrit. Toutes les fonctionnalités alors développées n'ont pas pu être incorporées dans la nouvelle mouture faute de temps, notamment l'implémentation d'un outil d'amarrage moléculaire gratuit pour pouvoir attribuer un score à chaque pose de fragment dans le site actif (*rescoring*) et ainsi pouvoir suivre son évolution au cours des itérations de grandissement.

Deux stagiaires de Master 2 ont également contribué à ces travaux de thèse. Le premier, Alan Oblé (2014), a participé à l'élaboration d'un workflow KNIME pour fragmenter les ligands extraits des structures cristallographiques. Le second, Jean-Thomas Heinrich (2015), a réalisé une série de tests de

l'outil, ce qui m'a permis d'identifier et de corriger des erreurs. Ce projet a également pu bénéficier récemment d'un nouveau financement par la Région Val Centre de Loire et d'une collaboration avec la société Greenpharma, ce qui a permis de recruter un nouveau doctorant. Ainsi, Colin Bournez est désormais le porteur de ce projet et participe déjà activement à l'optimisation du code pour le rendre à la fois plus rapide et plus robuste (parallélisation, nettoyage de code, etc.), en plus d'apporter de nouvelles fonctionnalités, notamment en ce qui concerne la faisabilité synthétique des molécules conçues. Pour ce faire, plusieurs règles empiriques peuvent être implémentées pour autoriser la jonction d'atome dans un contexte de groupements chimiques réactifs.^{303,304} Parmi celles-ci, BRICS³⁰⁵ et RECAP²⁷⁹ sont couramment employées et permettraient d'accroître la pertinence des résultats proposés. De plus, cela permettrait également de réduire le nombre de possibilités à chaque étape et donc d'améliorer les performances de l'outil.

Ce projet bénéficie également du support ponctuel d'un autre membre actuel de l'équipe, le chercheur postdoctorant Pascal Krezel. Ce dernier contribue actuellement à l'élaboration d'un réseau (approche des graphes) optimisé des fragments à relier en annotant non seulement les fragments voisins mais également les atomes qui pourraient les relier entre eux. Ce travail est destiné à remplacer la cartographie actuellement en place. Cette optimisation importante devrait permettre de remplacer la boucle de complexité 4 décrite dans l'algorithme du module *Linker* en un parcours de graphe interfacé avec une base de données à l'aide d'une librairie Python dédiée, et ainsi résulter en une économie de temps de calculs considérable.

Enfin, lors de l'initiation de ce projet aucun outil FBDD publié n'était capable d'explorer l'espace chimique de manière rationnelle pour s'écarter de solutions connues. Une approche de recherche automatisée reposant sur des données biologiques publiques (Open Phacts³⁰⁶, ChEMBL⁹), comme en est capable OpenGrowth (publication en 2015), constituerait un atout important pour prioriser les composés les plus innovants.

Le pari de l'approche de Frags2Drugs est l'utilisation de fragments issus d'inhibiteurs actifs pour générer de nouvelles idées de molécules bioactives. Sa démarche se veut rapide en combinant des fragments tridimensionnels et donc déjà positionnés dans le site actif de la protéine d'intérêt. Cependant, l'alignement des domaines kinases est également primordial pour le succès de cette technique. Il suffit en effet qu'une structure de domaine kinase soit mal alignée pour que des combinaisons potentiellement intéressantes ne soient pas explorées. Selon le même principe, des fragments sont susceptibles d'être retirés du jeu de données car ils se superposent à certains acides aminés du récepteur, ce qui peut favoriser la population de fragments dans des zones du site actif présentant plus d'espace. Une alternative pourrait

³⁰³ P. Ertl *et al.*, Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions, *Journal of Cheminformatics*, 1.1 (2009), 8

³⁰⁴ P. Bonnet, Is Chemical Synthetic Accessibility Computationally Predictable for Drug and Lead-like Molecules? A Comparative Assessment between Medicinal and Computational Chemists, *European Journal of Medicinal Chemistry*, 54 (2012), 679–89

³⁰⁵ J. Degen *et al.*, On the Art of Compiling and Using "drug-like" Chemical Fragment Spaces, *ChemMedChem*, 3.10 (2008), 1503–7

³⁰⁶ Antony J. Williams *et al.*, Open PHACTS: Semantic Interoperability for Drug Discovery, *Drug Discovery Today*, 17.21–22 (2012), 1188–98

être de permettre leur déplacement dans le site actif, par exemple en autorisant une légère minimisation de la structure formée en fixant le fragment initial pour conserver l'interaction clé recherchée. Cependant, cela se traduirait par un temps de calculs sensiblement plus long. De plus, le nombre de solutions générées par l'outil croît exponentiellement en fonction du nombre d'étapes ce qui est actuellement une limitation de l'outil. La solution pour le moment retenue est de rajouter des positions de fragments dans la base de données, soit en rajoutant d'autres méthodes d'alignement, soit en réalisant des expériences d'arrimage moléculaire (*docking*) de ligands ou de fragments directement. Il est cependant à noter que la plupart des algorithmes d'arrimage moléculaire ne sont pas paramétrés pour gérer des molécules de si petite taille et formant des interactions si faibles avec le récepteur, ce qui peut se traduire par un score imprécis.³⁰⁷ De plus, un fragment seul est susceptible de se lier de toute autre manière qu'un même fragment incorporé dans un ligand de plus grande taille.^{308,309}

De même, puisque la plupart des inhibiteurs de protéines kinases co-cristallisés, utilisés dans notre étude pour générer des fragments, forment des interactions avec la région charnière, on peut également se demander si on n'introduit pas un biais dans l'étude en orientant la conception de molécules vers des inhibiteurs de type I, I ½ ou II. En effet, la conception d'inhibiteurs allostériques serait plus compliquée du fait de la nature peu conservée des poches allostériques, surtout dans le cadre d'inhibiteurs de type IV qui se lient loin du site actif. Dans ces cas-là, l'approche Frags2Drugs pourrait tout de même être réalisée mais en aval d'un amarrage moléculaire d'une base de fragments dans la cavité d'intérêt.

De manière plus générale, l'outil développé est potentiellement applicable sur toutes les familles de protéines présentant assez d'informations structurales, comme par exemple pour les RCPG ou les récepteurs nucléaires. Nous avons fait le choix dans notre étude de nous focaliser d'abord sur les protéines kinases parce qu'elles présentent un site actif bien conservé, ce qui en autorise l'alignement dans un même référentiel et que de nombreuses structures cristallographiques existent dans la RCSB.

Pour conclure cette partie de chapitre de thèse, Frags2Drugs est un outil déjà opérationnel et appliqué à des projets de recherche au sein du laboratoire. Il présente cependant actuellement des limitations techniques, notamment au niveau du temps de calculs, que d'autres membres de l'équipe sont déjà en train d'améliorer. L'outil en l'état a néanmoins déjà permis d'identifier des composés présentant de très bonnes valeurs de LE sur la protéine GSK3 et a permis une approche de *scaffold hopping* sur la protéine BRAF. Le projet Frags2Drugs va continuer de s'améliorer au cours des prochaines années et permettra de concevoir de nouvelles molécules bioactives pour des projets de recherche en innovation thérapeutique et en cosmétique et peut-être de déposer de nouveaux brevets.

³⁰⁷ M. L. Verdonk *et al.*, Docking Performance of Fragments and Druglike Compounds, *Journal of Medicinal Chemistry*, 54.15 (2011), 5422–31

³⁰⁸ D. Kozakov *et al.*, 'Ligand Deconstruction: Why Some Fragment Binding Positions Are Conserved *et al.* Are Not', *Proceedings of the National Academy of Sciences of the United States of America*, 112.20 (2015), E2585-2594

³⁰⁹ M. N. Drwal *et al.*, Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands?, *Journal of Chemical Information and Modeling*, 57.5 (2017), 1197–1209

4.4 Annexes

4.4.1 Annexe 1

Table 22. Liste des outils in silico FBDD pour le fragment growing

Identifiant	Outil	Année	Type	Compagnie	Descriptif succinct
1	Legend	1991	Université	Tokyo	Construction de ligand atome par atome en partant d'un point de départ dans la cavité, croissance guidée par calcul d'énergie de vdW
2	FOUNDATION	1993	Université	Washington	Ajout de fragments sur un fragment initial selon un pharmacophore 3D
3	GroupBuild	1993	Industrie	Vertex	Ajout de fragments sur un fragment initial avec minimisation à chaque étape au sein du site actif
4	SPROUT	1993	Industrie	Simbiosys (avant: Université de Leeds)	Liaison entre plusieurs fragments in situ puis amarrage moléculaire de la solution trouvée en conservant les interactions initiales, la version SynSPROUT permet d'effectuer les liaisons par chimie de synthèse virtuelle
5	GrowMol	1994	Industrie	Ciba-Geigy Corporation	Ajout itératif d'atomes ou de fragments d'une librairie 3D dans le site actif à partir d'un point de départ défini dans l'espace et guidé par un score de complémentarité entre le ligand formé et le récepteur
6	GenStar	1995	Université	Harvard	Construction de ligand atome par atome soit depuis un inhibiteur initial soit de novo selon un modèle d'interaction avec le récepteur
7	PRO_SELECT	1997	Industrie	Proteus Molecular Design Ltd.	Ajout de substituants sur un fragment initial selon des contraintes pharmacophoriques et exploration conformationnelle de la partie grandie pour optimiser les interactions protéine-ligand
8	Skelgen	1997	Université	Cambridge	Définition d'un pharmacophore par rapport à un ligand initial puis ajout/remplacement de fragments sur un

					échafaudage moléculaire selon un algorithme de Monte Carlo
9	SMoG	1997	Université	Harvard	Remplissage du site actif avec des fragments aléatoires pour déterminer quels types de fragments donnent les meilleures interactions, puis utilisation de cette information pour étendre un fragment initial avec les groupes fonctionnels les plus efficaces en appliquant un algorithme de Monte Carlo
10	LigBuilder	2000	Université	Pékin	Ajout itératif de fragments selon pharmacophore sur récepteur, peut également faire du linking ou du hopping
11	Allegrow	2001	Industrie	Ciba-Geigy Corporation	Amélioration de GrowMol
12	CombiSMoG	2002	Université	Harvard	Amélioration de SMoG avec une fonction de score plus performante et un algorithme de Monte Carlo combinatoire
13	AutoGrow	2009	Université	San Diego	Algorithme génétique utilisant un fragment initial auquel est ajouté un autre fragment de manière aléatoire, les molécules assemblées sont amarrées et les meilleures solutions utilisées pour la génération suivante, mise à jour de 2013 permet l'ajout de fragments selon les réactions de "click-chemistry" et le filtre des solutions non "drug-like".
14	FOG	2009	Université	Harvard	Ajout de fragments sur un fragment initial selon un modèle de chaînes de Markov de liaisons fragment-fragment possibles, avec classification des résultats en "drug-like" et "non-drug-like".
15	NovoFlap	2010	Industrie	AstraZeneca	Création de nouvelles molécules à partir d'une base de fragments en 3D, les solutions sont itérativement filtrées selon un score obtenu par comparaison avec la forme et le pharmacophore de molécules de référence
16	LigBuilder 2	2011	Université	Pékin	Evolution du LigBuilder avec score de complémentarité cavité-ligand et prédiction de la faisabilité synthétique
17	LeadOpt	2015	Université	Sichuan	Quadrillage de la cavité d'un complexe protéine-ligand pour déterminer l'orientation à la croissance de la

					molécule, estimation du LE et d'un score ADMET pour filtrer les solutions, peut être employé pour retirer des fragments présents dans la molécule initiale pour enlever des encombrements stériques avec le récepteur (fragment hopping)
18	OpenGrowth	2015	Université	Harvard	Ajout itératif de fragments depuis un fragment initial placé dans le site actif. Chaque fragment est ensuite ajoutée sur le précédent par incrémentation de rotamères, la meilleure orientation est déterminée par un champ de force puis est minimisée de manière à optimiser davantage les interactions protéine-ligand. Les solutions trouvées sont ensuite filtrées selon leur innovation (recherche de présence dans la base de données ChEMBL), par score (SMoG2001 ou SMoG2015). Le processus global est biaisé pour identifier des molécules innovantes (FOG)
19	ACFIS	2016	Web-service	Wuhan	Minimisation du complexe fourni en entrée puis découpage du ligand en fragment principal et substituents, ces derniers sont alors remplacés par des fragments amarrés sur les points d'ancrage définis, les solutions sont filtrées par énergie d'interaction; l'outil repose sur Autogrow pour l'attachement des nouveaux fragments sur le fragment principal
20	PINGUI	2016	Université	Marbourg	Amarrage moléculaire de fragments avec SEED puis formation de réactions chimiques au sein du site actif pour former de nouveaux ligands, les 500 meilleurs produits sont alors amarrés avec DOCK et leur position minimisée
21	Frag2Drugs	2018?	Université	Orléans	Voir partie 4.2.3

4.4.2 Annexe 2

Table 23. Liste des outils in silico FBDD pour le fragment linking

Identifiant	Outil	Année	Type	Compagnie	Descriptif succinct
1	LUDI	1992	Industrie	BASF	Identification des sites de liaisons probables par règles empiriques de formation de liaisons hydrogènes et pour positionner des fragments dans le site actif puis génération de solutions en ajoutant des fragments de liaison entre les poses de fragments
2	Builder	1992	Université	San Diego	Génération de solutions combinatoires par recombinaison de molécules amarrées dans le site actif
3	CONCEPTS	1993	Industrie	Vertex	Diffusion d'atomes dans le site actif au cours d'une dynamique moléculaire, puis formation/rupture de liaisons à intervalles réguliers pour former des interactions plus stabilisantes avec le récepteur
4	NEWLEAD	1993	Industrie	CIBA-GEIGY	Ajout de linkers entre plusieurs fragments selon un modèle pharmacophorique 3D
5	SPLICE	1993	Université	Washington	Formation d'un pharmacophore à partir d'un complexe récepteur-ligand formé des molécules combinées à partir de fragments issus d'une base de données et répondant à ces critères
6	SPROUT	1993	Industrie	Simbiosys (avant: Université de Leeds)	Liaison entre plusieurs fragments in situ puis amarrage moléculaire de la solution trouvée en conservant les interactions initiales, la version SynSPROUT permet d'effectuer les liaisons par chimie de synthèse virtuelle
7	HOOK	1994	Université	York	Recherche dans une base de données moléculaire un squelette moléculaire qui puisse rentrer dans le site actif, puis fixation à deux poses de fragments obtenus par MCSS (crochets)

8	PROLIGAND	1994	Industrie	Proteus Molecular Design Ltd.	Ajout de linkers/fragments entre des fragments in situ de manière à compléter le site actif
9	Builder v2.0	1995	Université	San Diego	Amélioration de Builder avec plus de réactions chimiques et plus de linkers
10	DLD	1995	Université	Harvard	Echantillonnage de solutions de ligands possibles par liaisons d'atomes de carbone avec des groupements chimiques pré-ancrés dans le site actif
11	CONCERTS	1996	Industrie	Vertex	Diffusion de fragments dans le site actif au cours d'une dynamique moléculaire, puis formation/rupture de liaisons à intervalles réguliers pour former des interactions plus stabilisantes avec le récepteur
12	FlexNovo	2004	Université	Hambourg	Amarrage moléculaire flexible de fragments avec FlexX en deux temps, une fois sans puis une fois avec contraintes pharmacophoriques, les solutions trouvées par la dernière approche servent de fragments initiaux qui sont étendus avec les meilleures poses de fragments de la première étape d'amarrage moléculaire, les solutions trouvées privilégient les combinaisons de fragments avec les meilleurs scores d'amarrage moléculaire
13	LEA3D	2005	Académie	Institut Pasteur	Amarrage moléculaire (FlexX) de fragments issus de molécules bioactives au sein d'un site actif et utilisation du score obtenu dans algorithme génétique pour les combiner les meilleurs modes d'interaction entre eux
14	CONFIRM	2008	Industrie	BIOVIA	Connexion de fragments trouvés dans une structure de récepteur
15	GANDI	2008	Université	Zurich	Utilisation de propriétés structurales d'inhibiteurs connus en addition d'un champ de force énergétique pour relier des fragments pré-amarrés et de nouveaux fragments
16	combine_fragments	2009	Industrie	Schrödinger	Liaison de fragments dans le récepteur selon pharmacophore
17	Fragment Shuffling	2009	Industrie	Bayer	Fragmentation de ligands issus de complexes alignés et construction incrémentielle des combinaisons possibles depuis un fragment initial

4.4.3 Annexe 3

Table 24. Liste des outils in silico FBDD pour le scaffold hopping

Identifiant	Outil	Année	Type	Compagnie	Descriptif succinct
1	LeapFrog	1993	Industrie	Tripos (Certara)	Modifications successives d'un ligand tout en évaluant son énergie de liaison avec le récepteur, le changement sont conservés ou ignorés selon le résultat
2	CAVEAT	1994	Université	Californie	Recherche dans une base de données de fragments pouvant remplacer des substituants du ligand selon leur orientation dans l'espace pour conserver les interactions avec le récepteur
3	MCDNLG	1995	Industrie	Agouron Pharmaceuticals	Algorithme de Monte Carlo pour faire apparaître et disparaître les atomes ou fragments d'un ligand initial, ceux-ci sont alors soumis à un mouvement (rotation, translation, etc.) ou à une substitution
4	BREED	2004	Industrie	Vertex, implementé dans MOE et Schrodinger	Hybridation de poses de ligands sur des liaisons superposables
5	SHOP	2007	Université	Copenhagen	Remplacement de fragments selon différents types de contraintes (géométrie, forme et descripteurs d'interactions)
6	ParaFrag	2008	Industrie	F. Hoffmann - La Roche AG	Remplacement de fragments par génération de pharmacophore et recherche de similarité
7	CombiGlide	2009	Industrie	Schrödinger	Remplacement d'un fragment par réactions chimiques, évaluation des solutions avec scores Glide
8	Med-SuMo	2009	Industrie	MEDIT	Alignement de complexes protéine-ligand dans un même référentiel puis hybridation des ligands in situ selon leur forme et un pharmacophore

9	Scaffold Replacement	2010	Industrie	MOE	Remplacement de fragments par approche de pharmacophore, les solutions trouvées peuvent être filtrées (modèles QSAR, pharmacophores, score d'accessibilité synthétique, score d'interaction protéine-ligand, etc.)
10	LigMerge	2012	Université	Californie	Superposition de deux fragments (PDB seulement) sur un motif commun puis génération des combinaisons possibles, les solutions sont ensuite amarrées sur le récepteur pour en estimer l'affinité
11	Core Hopping	2012	Industrie	Schrödinger	Remplacement du fragment principal en conservant les substituants, peut prendre en compte les interactions formées avec le récepteur
12	Scaffold Hopping	2013	Industrie	AstraZeneca	Remplacement combinatoire d'un fragment initial selon son orientation 3D, résultats triés en fonction de leur score de recouvrement de volume et similarité électrostatique

4.4.4 Annexe 4

Table 25. Liste des outils FBDD in silico combinant plusieurs approches.

Identifiant	Outil	Année	Type	Compagnie	Descriptif succinct
1	LigBuilder	2000	Université	Pékin	Ajout itératif de fragments selon pharmacophore sur récepteur, peut également faire du <i>fragment linking</i> ou du <i>hopping</i>
2	LigBuilder 2	2011	Université	Pékin	Evolution du LigBuilder avec score de complémentarité cavité-ligand et prédiction de la faisabilité synthétique
3	LEADOPT	2015	Université	Sichuan	Quadrillage de la cavité d'un complexe protéine-ligand pour déterminer l'orientation à la croissance de la molécule, estimation du LE et d'un score ADMET pour filtrer les solutions, peut être employé pour retirer des fragments présents dans la molécule initiale pour enlever des encombrements stériques avec le récepteur (<i>fragment replacement</i>)
4	ReCore	2007	Industrie	BioSolvelt	Similaire à CAVEAT mais optimisé pour de larges bibliothèques de fragments, employable à la fois pour des recherches conformationnelles, <i>fragment growing</i> , <i>linking</i> et <i>hopping</i>
5	Spark	2012	Industrie	Cresset	Remplacement de fragments en considérant la forme et leurs propriétés électrostatiques, peut également faire du <i>fragment growing</i> et <i>linking</i>
6	BROOD v3.0	2015	Industrie	OpenEye	Substitution de fragments présents sur un ligand selon divers paramètres (pharmacophores, sélectivité, diversité, etc.), dernière version permet également de faire du <i>fragment linking</i>
7	ACFIS	2016	Web-service	Wuhan	Minimisation du complexe fourni en entrée puis découpage du ligand en fragment principal et substituent, ces derniers sont alors remplacés par des fragments amarrés sur les points d'ancrage définis, les solutions sont filtrées par énergie d'interaction; l'outil repose sur Autogrow pour l'attachement des nouveaux fragments sur le fragment principal

5. Conclusion générale et perspectives

Venant au départ d'une formation de biologiste, puis de bioinformaticien, ces années de thèse m'ont donné l'opportunité de me familiariser avec le monde de la chimoinformatique et de la modélisation moléculaire, ses bons comme ses mauvais côtés.

Je pense notamment au développement d'un certain protocole général de préparation de petites molécules, travail nécessaire mais épuisant compte tenu du nombre astronomique de cas particuliers à prendre en compte. Ce projet s'est révélé bien plus chronophage qu'initialement prévu. Le développement de cet outil m'a grandement aidé à approfondir mes connaissances dans le domaine tout en me montrant une nouvelle façon de concevoir des programmes grâce à la plateforme KNIME et à son approche de *pipelining*. Il est maintenant systématiquement utilisé dans l'ensemble des projets de criblage virtuel de l'équipe SB&C du Pr. Pascal Bonnet. Ce projet m'a par ailleurs apporté la joie des récompenses avec le prix poster du public lors d'une présentation aux Journées *Chemoinformatics Strasbourg Summer School (C3S)* en 2014. La récente publication de l'outil dans le numéro spécial « Chemoinformatics in France » du journal *Molecular Informatics* (Gally *et al.* Mol. Inf. 2017, 36(10)) a concrétisé cette satisfaction.

Une fois fonctionnel, cet outil de préparation de molécules a pu être appliqué à un certain nombre de chimiothèques, dont la Chimiothèque Nationale et Ambinter (www.ambinter.com), prêtes pour du criblage virtuel suite à leur préparation avec VSPrep. Il reste cependant encore beaucoup à faire pour améliorer l'outil et le rendre vraiment polyvalent. Une nouvelle implémentation en Python pourrait résoudre de nombreux problèmes auxquels j'ai été confronté tout en octroyant à l'outil beaucoup de nouvelles fonctionnalités intéressantes. Afin de mener tout cela à bien, le Dr Stéphane Bourg de l'équipe SB&C de l'ICOA est maintenant à la tête de ce projet pour développer plus loin l'approche VSPrep.

J'ai eu moi aussi l'opportunité d'appliquer cet outil sur un projet de criblage de produits naturels intégrant les données structurales de trois conformations différentes d'une protéine kinase. J'ai hélas manqué de temps pour exploiter entièrement cet aspect multi-structures afin de mener ce projet à terme. Un biais possible que je n'avais pas pris en compte initialement est le mode de fixation des inhibiteurs de protéines kinases sur des structures de conformations différentes. Des travaux sont en cours pour mieux paramétrer l'expérience et ainsi réussir à intégrer plus de données.

Enfin, j'ai pu également initier un projet de développement d'outil *in silico* FBDD. Par manque de temps, je n'ai pu apporter qu'une première brique à cet outil et toutes les fonctionnalités clés de l'outil ne sont pas encore implémentées, ce qui fait qu'il n'a pu paraître aussi original dans ce manuscrit de thèse qu'initialement prévu. Cependant, ce projet a fait l'objet d'un nouveau financement de thèse et mon successeur, Colin Bournez, est déjà en train de modifier le code source de l'outil pour le rendre plus performant.

6. Communication scientifique

6.1 Article

1. Gally, J.-M. ; Bourg, S. ; Do Q.T. ; Aci-Sèche, S. ; Bonnet, P. VSPrep: A general KNIME workflow for the preparation of molecules for virtual screening.
Molecular Informatics 2017, 36, 1700023.

6.2 Communications orales

1. Gally, J.-M. ; Heinrich, J. T. ; Obled, A. ; Aci-Sèche, S. ; Bonnet, P.
Frag2Drugs: a novel in silico tool for Fragment-Based Drug Discovery
30^{èmes} Journées Franco-Belges de Pharmacochimie (JFB2016)
mai 2016 - Amboise.
2. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Canault, B. ; Arora, R. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
A freely accessible workflow for the preparation of molecules for virtual screening
7^{èmes} journées de la Société Française de Chemoinformatique (SFCi 2015)
oct. 2015 - Nice.

6.3 Communications par affiche

1. Gally, J.-M. ; Bourg, S. ; Do Q.T. ; Aci-Sèche, S. ; Bonnet, P.
VSPrep : A General KNIME Workflow for the Preparation of Molecules for Virtual Screening
8es journées de la Société Française de Chemoinformatique - SFCi2017
oct. 2017 - Orléans.
2. Gally, J.-M. ; Bourg, S. ; Do Q.T. ; Aci-Sèche, S. ; Bonnet, P.
VSPrep : A General KNIME Workflow for the Preparation of Molecules for Virtual Screening
6th RDKit UGM 2017
sept. 2017 - Berlin (Allemagne).
3. Gally, J.-M. ; Bourg, S. ; Do Q.T. ; Aci-Sèche, S. ; Bonnet, P.
VSPrep: a freely accessible workflow for the preparation of molecules for virtual screening
20^{ème} congrès du Groupe Graphisme et Modélisation Moléculaire - GGMM 2017
mai 2017 - Reims.
4. Gally, J.-M. ; Heinrich, J. T. ; Obled, A. ; Aci-Sèche, S. ; Bonnet, P.
Frag2Drugs: a novel in silico tool for Fragment-Based Drug Discovery
XXIV^{ème} Journées Jeunes Chercheurs (SCT)
févr. 2017 - Châtenay-Malabry.
5. Gally, J.-M. ; Heinrich, J. T. ; Obled, A. ; Aci-Sèche, S. ; Bonnet, P.
Frag2Drugs: A new tool to design new kinase inhibitors from fragments
Prix de la meilleure affiche
29^{ème} Colloque Biotechnocentre
oct. 2016 - Seillac.
6. Gally, J.-M. ; Heinrich, J. T. ; Obled, A. ; Aci-Sèche, S. ; Bonnet, P.
Frag2Drugs: identification of kinase inhibitors using a novel in silico FBLD tool
Seventh Joint Sheffield Conference on Chemoinformatics (MGMS 2016)
juil. 2016 - Sheffield (Royaume-Uni).

7. Gally, J.-M. ; Heinrich, J. T. ; Obled, A. ; Aci-Sèche, S. ; Bonnet, P.
Frag2Drugs: a novel in silico tool for Fragment-Based Drug Discovery
 30^{èmes} Journées Franco-Belges de Pharmacochimie (JFB2016)
mai 2016 - Amboise.

8. Canault, B. ; Gally, J.-M. ; Braka, A. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
Evaluation of rDock docking tool for kinase drug discovery
 Journée Jeunes Chercheurs SCT
févr. 2016 - Lille.

9. Aci-Sèche, S. ; Bonnet, P. ; Bourg, S. ; Braka, A. ; Canault, B. ; Carles, F. ; Diharce, J. ; Fogha, J. ; Gally, J.-M. ; Lesellier, E. ; West, C. ; Ziada, S.
Current Development of in silico tools for Drug Discovery
 Journée Fédération de Recherche Physique et Chimie du Vivant – FR2708
janv. 2016 - Orléans.

10. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Canault, B. ; Arora, R. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
A freely accessible and automated workflow for the preparation of molecules for virtual screening
 Journée Fédération de Recherche Physique et Chimie du Vivant – FR2708
janv. 2016 - Orléans.

11. Bourg, S. ; Pihan, E. ; Canault, B. ; Arora, R. ; Golib, F. ; Gally, J.-M. ; Bosc, N. ; Aci-Sèche, S. ; Bonnet, P.
In silico kinase platform (ISKP): a tool for the identification of selective kinase inhibitors
 51th International Conference on Medicinal Chemistry (RICT 2015)
juil. 2015 - Avignon.

12. Aci-Sèche, S. ; Bonnet, P. ; Bosc, N. ; Bourg, S. ; Braka, A. ; Canault, B. ; Cristofaro, E. ; Fogha, J. ; Gally, J.-M. ; Morin-Allory, L. ; Lesellier, E. ; Obled, A. ; West, C. ; Ziada, S.
Group of Structural Bioinformatics & Chemoinformatics
 1st ICOA symposium
juin 2015 - Orléans.

13. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Arora, R. ; Canault, B. ; Bourg, S. ; Aci-Sèche, S. ; Bonnet, P.
Combining the strengths of freely accessible toolkits: a universal workflow for the preparation of molecules for virtual screening
 Journées de la Section Régionale Centre Ouest de la SCF
févr. 2015 - Orléans.

14. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Arora, R. ; Bourg, S. ; Aci-Sèche, S. ; Canault, B. ; Bonnet, P.
A free accessible workflow for preprocessing virtual screening experiments
 27^{ème} Colloque Biotechnocentre
oct. 2014 - Seillac.

15. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Arora, R. ; Bourg, S. ; Canault, B. ; Bonnet, P.
Development of a universal workflow for the preparation of molecular databases for virtual screening
 prix du public

Chemoinformatics Strasbourg Summer School (CS3)
juin 2014 - Strasbourg.

16. Gally, J.-M. ; Bosc, N. ; Pihan, E. ; Arora, R. ; Bourg, S. ; Canault, B. ; Bonnet, P.
Development of a universal workflow for the preparation of molecular databases for virtual screening
Journée scientifique de la Fédération de Recherche FR2708
juin 2014 - Orléans.

7. Bibliographie

- [1] J.P. Hughes et al., Principles of Early Drug Discovery, *British Journal of Pharmacology*, 162.6 (2011), 1239–49
- [2] B. Munos, Lessons from 60 Years of Pharmaceutical Innovation, *Nature Reviews. Drug Discovery*, 8.12 (2009), 959–68
- [3] IRI - The EU Industrial R&D Investment Scoreboard, <http://iri.jrc.ec.europa.eu/scoreboard.html>, consulté le 15 Novembre 2017
- [4] Communiqué de Presse : Le Prix Nobel de Chimie 2002, <https://www.nobelprize.org>, consulté le 15 Novembre 2017
- [5] R. Macarron et al., Impact of High-Throughput Screening in Biomedical Research, *Nature Reviews Drug Discovery*, 10.3 (2011)
- [6] S. Kim et al., PubChem Substance and Compound Databases, *Nucleic Acids Research*, 44. Database issue (2016), D1202–13
- [7] L. C. Ray and R. A. Kirsch, Finding Chemical Records by Digital Computers, *Science (New York, N.Y.)*, 126.3278 (1957), 814–19
- [8] CAS, Chemical Abstracts Service Home Page, <https://www.cas.org/>, consulté le 15 Novembre 2017
- [9] A. Gaulton et al., ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery, *Nucleic Acids Research*, 40.Database issue (2012), D1100–1107
- [10] M. Randic, Characterization of Molecular Branching, *Journal of the American Chemical Society*, 97.23 (1975), 6609–15
- [11] L. H. Hall et al., Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information, *Journal of Chemical Information and Computer Sciences*, 35.6 (1995), 1039–45
- [12] W. H. B. Sauer et al., Molecular Shape Diversity of Combinatorial Libraries: A Prerequisite for Broad Bioactivity, *Journal of Chemical Information and Computer Sciences*, 43.3 (2003), 987–1003
- [13] Y. Sakuratani et al., Comparison of Predictivities of Log P Calculation Models Based on Experimental Data for 134 Simple Organic Compounds, *QSAR & Combinatorial Science*, 26.1 (2007), 109–16
- [14] J. L. Durant et al., Reoptimization of MDL Keys for Use in Drug Discovery, *Journal of Chemical Information and Computer Sciences*, 42.6 (2002), 1273–80.
- [15] Daylight Theory: SMARTS - A Language for Describing Molecular Patterns, <http://www.daylight.com/dayhtml/doc/theory/theory.SMARTS.html>, consulté le 15 Novembre 2017.
- [16] D. Rogers et al., Extended-Connectivity Fingerprints, *Journal of Chemical Information and Modeling*, 50.5 (2010), 742–54
- [17] A. Sreelatha et al., The Secretory Pathway Kinases, *Biochimica et Biophysica Acta*, 1854.10 0 0 (2015), 1687–93
- [18] G. Burnett et al., The Enzymatic Phosphorylation of Proteins, *The Journal of Biological Chemistry*, 211.2 (1954), 969–80.
- [19] E. G. Krebs et al., Factors Affecting the Activity of Muscle Phosphorylase b Kinase, *The Journal of Biological Chemistry*, 234 (1959), 2867–73.
- [20] D. A. Walsh et al., An Adenosine 3,5-Monophosphate-Dependant Protein Kinase from Rabbit Skeletal Muscle, *The Journal of Biological Chemistry*, 243.13 (1968), 3763–65.
- [21] R. Roskoski, A Historical Overview of Protein Kinases and Their Targeted Small Molecule Inhibitors, *Pharmacological Research*, 100 (2015), 1–23
- [22] M. J. Lee et al., Protein Regulation in Signal Transduction, *Cold Spring Harbor Perspectives in Biology*, 8.6 (2016)
- [23] S. M. Carlson et al., Using Small Molecules and Chemical Genetics to Interrogate Signaling Networks, *ACS Chemical Biology*, 6.1 (2011), 75–85

- [24] T. Hunter, A Thousand and One Protein Kinases, *Cell*, 50.6 (1987), 823–29.
- [25] G. Manning et al., The Protein Kinase Complement of the Human Genome, *Science* (New York, N.Y.), 298.5600 (2002), 1912–34
- [26] S. K Hanks, Genomic Analysis of the Eukaryotic Protein Kinase Superfamily: A Perspective, *Genome Biology*, 4.5 (2003), 111
- [27] S. Braconi Quintaje and 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: MCP*, 7.8 (2008), 1409–19
- [28] J. Boudeau et al., Emerging Roles of Pseudokinases, *Trends in Cell Biology*, 16.9 (2006), 443–52
- [29] Mazen W. Karaman et. al., A Quantitative Analysis of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 26.1 (2008), 127–32
- [30] H. M. Berman et al., The Protein Data Bank, *Acta Crystallographica. Section D, Biological Crystallography*, 58.Pt 6 No 1 (2002), 899–907.
- [31] D. Fabbro et al., Ten Things You Should Know about Protein Kinases: IUPHAR Review 14, *British Journal of Pharmacology*, 2015
- [32] J. Boudeau et al., Emerging Roles of Pseudokinases, *Trends in Cell Biology* 16 (2006), no. 9:443–52
- [33] S. S. Taylor et al., Protein Kinases: Evolution of Dynamic Regulatory Proteins, *Trends Biochem. Sci.*, 36.2 (2011), 65–77
- [34] A. P. Kornev et al., A Helix Scaffold for the Assembly of Active Protein Kinases, *Proceedings of the National Academy of Sciences of the United States of America*, 105.38 (2008), 14377–82
- [35] S. R. Hubbard et al., Crystal Structure of the Tyrosine Kinase Domain of the Human Insulin Receptor, *Nature*, 372.6508 (1994), 746–54
- [36] N. Bosc, Développement de nouvelles approches protéo-chimométriques appliquées à l'étude des interactions et de la sélectivité des inhibiteurs de kinases (Université d'Orléans, 2015)
- [37] Z. A. Knight et al., Features of Selective Kinase Inhibitors, *Chemistry & Biology*, 12.6 (2005), 621–37
- [38] L. N. Johnson, Protein Kinase Inhibitors: Contributions from Structure to Clinical Compounds, *Quarterly Reviews of Biophysics*, 42.1 (2009), 1–40
- [39] Novartis Delivered Solid 2016 Performance, with Growth Products Absorbing Gleevec US LOE; Innovation Momentum Continued; Announces Share Buyback, Novartis, <https://www.novartis.com/news/media-releases/novartis-delivered-solid-2016-performance-growth-products1-absorbing-gleevec-us>, consulté le 2 Octobre 2017.
- [40] Poster : Unexplored Opportunities in the Druggable Human Genome, <http://www.nature.com/nrd/posters/druggablegenome/index.html>, consulté le 3 Octobre 2017.
- [41] R. Roskoski, Classification of Small Molecule Protein Kinase Inhibitors Based upon the Structures of Their Drug-Enzyme Complexes, *Pharmacological Research*, 103 (2016), 26–48
- [42] F. Zuccotto et al., Through the “Gatekeeper Door”: Exploiting the Active Kinase Conformation, *Journal of Medicinal Chemistry*, 53.7 (2010), 2681–94
- [43] R. Capdeville et al., Glivec (STI571, Imatinib), a Rationally Developed, Targeted Anticancer Drug, *Nature Reviews. Drug Discovery*, 1.7 (2002), 493–502
- [44] T. Schindler et al., Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase, *Science* (New York, N.Y.), 289.5486 (2000), 1938–42
- [45] D. K. Treiber et al., Ins and Outs of Kinase DFG Motifs, *Chemistry & Biology*, 20.6 (2013), 745–46
- [46] M. I. Davis et. al., Comprehensive Analysis of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 29.11 (2011), 1046–51
- [47] G. Bollag et. al., Vemurafenib: The First Drug Approved for BRAF-Mutant Cancer, *Nature Reviews. Drug Discovery*, 11.11 (2012), 873–86
- [48] M. A. Bogoyevitch et al., A New Paradigm for Protein Kinase Inhibition: Blocking Phosphorylation without Directly Targeting ATP Binding, *Drug Discovery Today*, 12.15–16 (2007), 622–33

- [49] L. Garuti et al., Non-ATP Competitive Protein Kinase Inhibitors, *Current Medicinal Chemistry*, 17.25 (2010), 2804–21
- [50] L. O. Kirkland et al., Non-ATP Competitive Protein Kinase Inhibitors as Anti-Tumor Therapeutics, *Biochemical Pharmacology*, 77.10 (2009), 1561–71
- [51] L. Krim Gavrin et al., Approaches to Discover Non-ATP Site Kinase Inhibitors, *MedChemComm*, 4.1 (2012), 41–51
- [52] C. J. M. Wright et al., Trametinib: First Global Approval, *Drugs*, 73.11 (2013), 1245–54
- [53] Q. Dong et. al., Discovery of TAK-733, a Potent and Selective MEK Allosteric Site Inhibitor for the Treatment of Cancer, *Bioorganic & Medicinal Chemistry Letters*, 21.5 (2011), 1315–19
- [54] V. Lamba et al., New Directions in Targeting Protein Kinases: Focusing upon True Allosteric and Bivalent Inhibitors, *Current Pharmaceutical Design*, 18.20 (2012), 2936–45
- [55] T. K. Johnson et al., Bivalent Inhibitors of C-Src Tyrosine Kinase That Bind a Regulatory Domain, *Bioconjugate Chemistry*, 27.7 (2016), 1745–49
<<https://doi.org/10.1021/acs.bioconjchem.6b00243>>.
- [56] F. E. Kwarcinski et. al., Discovery of Bivalent Kinase Inhibitors via Enzyme-Templated Fragment Elaboration, *ACS Medicinal Chemistry Letters*, 6.8 (2015), 898–901
- [57] C. M. Gower et. al., Conversion of a Single Polypharmacological Agent into Selective Bivalent Inhibitors of Intracellular Kinase Activity, *ACS Chemical Biology*, 11.1 (2016), 121–31
- [58] B. C. Lechtenberg et. al., Structure-Guided Strategy for the Development of Potent Bivalent ERK Inhibitors, *ACS Medicinal Chemistry Letters*, 8.7 (2017), 726–31
- [59] Q. Liu et. al., Developing Irreversible Inhibitors of the Protein Kinase Cysteine, *Chemistry & Biology*, 20.2 (2013), 146–59
- [60] C.-H. Yun et. al., The T790M Mutation in EGFR Kinase Causes Drug Resistance by Increasing the Affinity for ATP, *Proceedings of the National Academy of Sciences of the United States of America*, 105.6 (2008), 2070–75
- [61] J. A. Woyach et. al., Resistance Mechanisms for the Brutons Tyrosine Kinase Inhibitor Ibrutinib, *The New England Journal of Medicine*, 370.24 (2014), 2286–94
- [62] T. Zhou et. al., Structural Mechanism of the Pan-BCR-ABL Inhibitor Ponatinib (AP24534): Lessons for Overcoming Kinase Inhibitor Resistance, *Chemical Biology & Drug Design*, 77.1 (2011), 1–11
- [63] N. Bosc et. al., A Proteomic Analysis of Human Kinome: Insight into Discriminant Conformation-Dependent Residues, *ACS Chemical Biology*, 10.12 (2015), 2827–40
- [64] P. Cohen, Protein Kinases--the Major Drug Targets of the Twenty-First Century?, *Nature Reviews. Drug Discovery*, 1.4 (2002), 309–15
- [65] BIOVIA Discovery Studio | Predictive Modeling & Science Simulation Software App, <http://accelrys.com/products/collaborative-science/biovia-discovery-studio/>, consulté le 28 Août 2017.
- [66] Molecular Operating Environment (MOE), 2013.08; Chemical Computing Group Inc., 1010 Sherbooke St. West, Suite #910, Montreal, QC, Canada, H3A 2R7, 2017., https://www.chemcomp.com/Research-Citing_MOE.htm, consulté le 31 Janvier 2017.
- [67] SYBYL-X Suite, Certara, 2016, <https://www.certara.com/software/molecular-modeling-and-simulation/sybyl-x-suite>, consulté le 10 Novembre 2016
- [68] SD File Processing with MOE Pipeline Tools, <https://www.chemcomp.com/journal/sdtools.htm>, consulté le 29 Novembre 2017.
- [69] Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.
- [70] MoKa - PKa Modelling, <http://www.moldiscovery.com/software/moka>, consulté le 29 Août 2017
- [71] Citations | Schrödinger | GLIDE, <https://www.schrodinger.com/citations>, consulté le 29 Août 2017
- [72] G. Jones et. al., Development and Validation of a Genetic Algorithm for Flexible Docking, *Journal of Molecular Biology*, 267.3 (1997), 727–48
- [73] J. C. Cole et. al., Comparing Protein–ligand Docking Programs Is Difficult, *Proteins: Structure, Function, and Bioinformatics*, 60.3 (2005), 325–332

- [74] SYBYL-X Suite.
- [75] Molecular Discovery Ltd., <http://www.moldiscovery.com/>, consulté le 30 Août 2017.
- [76] MOE: Molecular Operating Environment, https://www.chemcomp.com/MOE-Molecular_Operating_Environment.htm, consulté le 24 Novembre 2016
- [77] OpenEye Scientific Software, Cheminformatics and Molecular Modeling Software (OpenEye), <http://www.eyesopen.com>, consulté le 24 Novembre 2016
- [78] ChemAxon – Software for Chemistry and Biology, <https://www.chemaxon.com>, consulté le 9 Novembre 2016
- [79] BIOVIA Discovery Studio | Predictive Modeling & Science Simulation Software App
- [80] The Amber Molecular Dynamics Package, <http://ambermd.org/>, consulté le 30 Août 2017
- [81] Citations | Schrödinger | Maestro, <https://www.schrodinger.com/citations>, consulté le 29 Août 2017
- [82] Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.
- [83] S. Forli, Charting a Path to Success in Virtual Screening, *Molecules* (Basel, Switzerland), 20.10 (2015), 18732–58
- [84] J.-M. Gally et. al., VSPrep: A General KNIME Workflow for the Preparation of Molecules for Virtual Screening, *Molecular Informatics*, 2017
- [85] What Is Open Source?, *Opensource.Com*, <https://opensource.com/resources/what-open-source>, consulté le 11 Novembre 2017
- [86] IBM Redbooks | Linux Handbook A Guide to IBM Linux Solutions and Resources, 2016, <http://www.redbooks.ibm.com/abstracts/sg247000.html?Open>, consulté le 11 Novembre 2017.
- [87] D. L. Shaw, Is Open Science the Future of Drug Development?, *The Yale Journal of Biology and Medicine*, 90.1 (2017), 147–51.
- [88] Y. Gao et. al., A Broad Activity Screen in Support of a Chemogenomic Map for Kinase Signalling Research and Drug Discovery, *Biochemical Journal*, 451.2 (2013), 313–28
- [89] S. Cooper et al., Predicting Protein Structures with a Multiplayer Online Game, *Nature*, 466.7307 (2010), 756–60
- [90] M. J. Coren Company Fast, Foldit Gamers Solve Riddle of HIV Enzyme within 3 Weeks, *Scientific American*, <https://www.scientificamerican.com/article/foldit-gamers-solve-riddle>, consulté le 11 Novembre 2017
- [91] BIOVIA Pipeline Pilot | Scientific Workflow Authoring Application for Data Analysis, <http://accelrys.com/products/collaborative-science/biovia-pipeline-pilot>, consulté le 9 Novembre 2016
- [92] M. R. Berthold et. al., KNIME: The Konstanz Information Miner, in *Data Analysis, Machine Learning and Applications*, ed. by Christine Preisach et al., *Studies in Classification, Data Analysis, and Knowledge Organization* (Springer Berlin Heidelberg, 2008), pp. 319–26
- [93] Orange – Visual Programming, <https://orange.biolab.si/features/visual-programming/>, consulté le 15 Novembre 2017
- [94] Christoph Steinbeck et. al., The Chemistry Development Kit (CDK): An Open-Source Java Library for Chemo- and Bioinformatics, *Journal of Chemical Information and Computer Sciences*, 43.2 (2003), 493–500
- [95] Rajarshi Guha et. al., The Blue Obelisk—Interoperability in Chemical Informatics, *Journal of Chemical Information and Modeling*, 46.3 (2006), 991–98
- [96] D. Pavlov et. al., Indigo: Universal Cheminformatics API, *Journal of Cheminformatics*, 3.1 (2011), P4
- [97] RDKit, <http://www.rdkit.org>, consulté le 9 Novembre 2016
- [98] Erl Wood Cheminformatics Nodes for KNIME (Trusted Extension), <https://www.knime.com/community/erlwood>, consulté le 29 Novembre 2017

- [99] Vernalis Nodes for KNIME (Trusted Extension), <https://www.knime.com/book/vernalis-nodes-for-knime-trusted-extension>, consulté le 30 Août 2017
- [100] ChemAxon Node for KNIME JChem Extensions,, <http://infocom-science.jp>, consulté le 12 Novembre 2017
- [101] An Overview of the RDKit — The RDKit 2017.09.1 Documentation, <http://www.rdkit.org/docs/Overview.html>, consulté le 11 Novembre 2017
- [102] Calculator Plugins Were Used for Structure Property Prediction and Calculation, Marvin 15.1.19, 2016, ChemAxon
- [103] N. T. Kochev et al., Ambit-Tautomer: An Open Source Tool for Tautomer Generation, *Molecular Informatics*, 32.5–6 (2013), 481–504
- [104] MolVS: Molecule Validation and Standardization — MolVS 0.0.9 Documentation, <http://molvs.readthedocs.io/en/latest>, consulté le 30 Août 2017
- [105] Academic License Conditions « ChemAxon – Software for Chemistry and Biology, ChemAxon - Software for Chemistry and Biology R&D, <https://www.chemaxon.com/my-chemaxon/academic-license-conditions>, consulté le 12 Novembre 2017
- [106] Daylight Theory: SMILES, <http://www.daylight.com/dayhtml/doc/theory/theory.SMILES.html>, consulté le 31 Août 2017
- [107] S. R. Heller et. al., InChI, the IUPAC International Chemical Identifier, *Journal of Cheminformatics*, 7 (2015), 23
- [108] Ctfles.Book - MDL_SDF.pdf, http://infochim.u-strasbg.fr/recherche/Download/Fragmentor/MDL_SDF.pdf, consulté le 31 Août 2017
- [109] Tripos Mol2 File Format, <http://chemyang.ccnuc.edu.cn/ccb/server/AIMMS/mol2.pdf>, consulté le 31 Août 2017
- [110] RDKit / Mailing Lists, <https://sourceforge.net/p/rdkit/mailman/message/34740156>, consulté le 31 Août 2017
- [111] S. Ruiz-Carmona et al., RDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids, *PLoS Computational Biology*, 10.4 (2014), e1003571
- [112] A. J. Williams et al., Towards a Gold Standard: Regarding Quality in Public Domain Chemistry Databases and Approaches to Improving the Situation, *Drug Discovery Today*, 17.13–14 (2012), 685–701
- [113] H. Strömbergsson et al., A Chemogenomics View on Protein-Ligand Spaces, *BMC Bioinformatics*, 10.6 (2009), 1–11
- [114] J.-P. Ebejer et al., Freely Available Conformer Generation Methods: How Good Are They?, *Journal of Chemical Information and Modeling*, 52.5 (2012), 1146–58
- [115] I. Pletnev et. al., InChIKey Collision Resistance: An Experimental Testing, *Journal of Cheminformatics*, 4 (2012), 39
- [116] J. Chambers et. al., UniChem: Extension of InChI-Based Compound Mapping to Salt, Connectivity and Stereochemistry Layers, *Journal of Cheminformatics*, 6.1 (2014)
- [117] T. ten Brink et al., Influence of Protonation, Tautomeric, and Stereoisomeric States on Protein–Ligand Docking Results, *Journal of Chemical Information and Modeling*, 49.6 (2009), 1535–46
- [118] F. Oellien et. al., The Impact of Tautomer Forms on Pharmacophore-Based Virtual Screening†, *Journal of Chemical Information and Modeling*, 46.6 (2006), 2342–54
- [119] L. Guasch et. al., Experimental and Chemoinformatics Study of Tautomerism in a Database of Commercially Available Screening Samples, *Journal of Chemical Information and Modeling*, 56.11 (2016), 2149–61
- [120] T. Kalliokoski et. al., The Effect of Ligand-Based Tautomer and Protomer Prediction on Structure-Based Virtual Screening, *Journal of Chemical Information and Modeling*, 49.12 (2009), 2742–48
- [121] F. Milletti et. al., Tautomer Enumeration and Stability Prediction for Virtual Screening on Large Chemical Databases, *Journal of Chemical Information and Modeling*, 49.1 (2009), 68–75
- [122] P. Pospisil et. al., Tautomerism in Computer-Aided Drug Design, 2003, <http://informahealthcare.com/doi/abs/10.1081/RRS-120026975%20>, consulté le 17 March 2014
- [123] M. Sitzmann et al., Tautomerism in Large Databases, *Journal of Computer-Aided Molecular Design*, 24.6–7 (2010), 521–51

- [124] Open Babel, http://openbabel.org/wiki/Main_Page, consulté le 24 Novembre 2016
- [125] C. A. Lipinski et. al., Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings, *Advanced Drug Delivery Reviews*, 23.1–3 (1997), 3–25
- [126] D. F. Veber et. al., Molecular Properties That Influence the Oral Bioavailability of Drug Candidates, *Journal of Medicinal Chemistry*, 45.12 (2002), 2615–23.
- [127] M. Congreve et. al., A “Rule of Three” for Fragment-Based Lead Discovery?, *Drug Discovery Today*, 8.19 (2003), 876–77
- [128] V. Hamon et. al., 2P2IHUNTER: A Tool for Filtering Orthosteric Protein-Protein Interaction Modulators via a Dedicated Support Vector Machine, *Journal of The Royal Society Interface*, 11.90 (2013), 20130860–20130860
- [129] S. Saubern, Rajarshi Guha, and Jonathan B. Baell, KNIME Workflow to Assess PAINS Filters in SMARTS Format. Comparison of RDKit and Indigo Cheminformatics Libraries, *Molecular Informatics*, 30.10 (2011), 847–50
- [130] G. Landrum, Curating the PAINS Filters, <http://rdkit.blogspot.com/2015/08/curating-pains-filters.html>, consulté le 10 Novembre 2016
- [131] J. Duan et. al., Analysis and Comparison of 2D Fingerprints: Insights into Database Screening Performance Using Eight Fingerprint Methods, *Journal of Molecular Graphics & Modelling*, 29.2 (2010), 157–70
- [132] J. T. Metz et al., Navigating the Kinome, *Nature Chemical Biology*, 7.4 (2011), 200–202
- [133] J. Bain et al., The Selectivity of Protein Kinase Inhibitors: A Further Update, *The Biochemical Journal*, 408.Pt 3 (2007), 297–315
- [134] T. Anastassiadis et al., Comprehensive Assay of Kinase Catalytic Activity Reveals Features of Kinase Inhibitor Selectivity, *Nature Biotechnology*, 29.11 (2011), 1039–45
- [135] O. Fedorov et al., A Systematic Interaction Map of Validated Kinase Inhibitors with Ser/Thr Kinases, *Proceedings of the National Academy of Sciences of the United States of America*, 104.51 (2007), 20523–28
- [136] E. A. Gatica and Claudio N. Cavasotto, Ligand and Decoy Sets for Docking to G Protein-Coupled Receptors, *Journal of Chemical Information and Modeling*, 52.1 (2011), 1–6
- [137] Y. Okuno et al., GLIDA: GPCR-Ligand Database for Chemical Genomic Drug Discovery, *Nucleic Acids Research*, 34.Database issue (2006), D673–677
- [138] M. M. Mysinger et al., Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking, *Journal of Medicinal Chemistry*, 55.14 (2012), 6582–94
- [139] Schrödinger Release 2017-3: LigPrep, Schrödinger, LLC, New York, NY, 2017.
- [140] N. Lagarde et al., NRLiSt BDB, the Manually Curated Nuclear Receptors Ligands and Structures Benchmarking Database, *Journal of Medicinal Chemistry*, 57.7 (2014), 3117–25
- [141] J. J. Irwin et al., ZINC: A Free Tool to Discover Chemistry for Biology, *Journal of Chemical Information and Modeling*, 52.7 (2012), 1757–68
- [142] Online Demo - Fast 3D Structure Generation with CORINA Classic | MN-AM, https://www.mn-am.com/online_demos/corina_demo, consulté le 22 Août 2017
- [143] General Discussion - How to Enumerate All Possible Stereoisomers of a Molecule with a Not Defined Stereocenter?, <http://forums.openbabel.org/How-to-enumerate-all-possible-stereoisomers-of-a-molecule-with-a-not-defined-stereocenter-td4659459.html>, consulté le 30 Août 2017
- [144] The Cambridge Crystallographic Data Centre (CCDC), <https://www.ccdc.cam.ac.uk>, consulté le 31 Août 2017
- [145] A. E. Cleves et al., ForceGen 3D Structure and Conformer Generation: From Small Lead-like Molecules to Macrocyclic Drugs, *Journal of Computer-Aided Molecular Design*, 31.5 (2017), 419–39
- [146] F. Mahuteau-Betzer, The French National Compound Library: advances and future prospects, *Medecine Sciences: M/S*, 31.4 (2015), 417–22

- [147] Chemical Supplier of Screening Compounds, Building Blocks - Ambinter, <http://www.ambinter.com/>, consulté le 13 Novembre 2017
- [148] GreenPharma | R&D Products and Services to Pharmaceutical, Cosmetic and Nutraceutical Research, <http://www.greenpharma.com/>, consulté le 13 Novembre 2017
- [149] Ambinter Catalogue Request, <http://www.ambinter.com/catalog/SDF>, consulté le 12 Novembre 2017
- [150] J. Lietava, Medicinal Plants in a Middle Paleolithic Grave Shanidar IV?, *Journal of Ethnopharmacology*, 35.3 (1992), 263–66.
- [151] I. Watts, Red Ochre, Body Painting and Language: Interpreting the Blombos Ochre, *The Cradle of Language*, https://www.academia.edu/1560509/Red_ochre_body_painting_and_language_interpreting_the_Blombos_ochre, consulté le 14 Novembre 2017.
- [152] A. Marston, Natural Products as a Source of Protein Kinase Activators and Inhibitors, *Current Topics in Medicinal Chemistry*, 11.11 (2011), 1333–39
- [153] X. Luo et al., The Role of Targeting Kinase Activity by Natural Products in Cancer Chemoprevention and Chemotherapy, *Oncology Reports*, 34.2 (2015), 547–54
- [154] T. Rodrigues et al., Counting on Natural Products for Drug Design, *Nature Chemistry*, 8.6 (2016), nchem.2479
- [155] L. Müller-Kuhrt, Putting Nature Back into Drug Discovery, *Nature Biotechnology*, 21.6 (2003), nbt0603-602–602
- [156] Q.-T. Do et al., How to Valorize Biodiversity? Lets Go Hashing, Extracting, Filtering, Mining, Fishing, *Planta Medica*, 81.6 (2015), 436–49
- [157] V. Romano, Comment se déroulent les séances médicales de traitement par laser des tâches ?, Dr Valeria Romano, <https://www.dr-romano-esthetique.com/fr/laser/taches-pigmentaire-peelings>, consulté le 13 Novembre 2017
- [158] M. Démarchez, Le mélanocyte et la pigmentation, <https://biologiedelapeau.fr>, 2017, <https://biologiedelapeau.fr/spip.php?article12>, consulté le 14 Novembre 2017
- [159] M. Seiberg, Keratinocyte-Melanocyte Interactions during Melanosome Transfer, *Pigment Cell Research*, 14.4 (2001), 236–42
- [160] S. A. N. D'Mello et al., Signaling Pathways in Melanogenesis, *International Journal of Molecular Sciences*, 17.7 (2016)
- [161] I. Ferreira dos Santos Videira et al., Mechanisms Regulating Melanogenesis, *Anais Brasileiros De Dermatologia*, 88.1 (2013), 76–83
- [162] M. Picardo et al., The Genetic Determination of Skin Pigmentation: KITLG and the KITLG/c-Kit Pathway as Key Players in the Onset of Human Familial Pigmentary Diseases, *Journal of Investigative Dermatology*, 131.6 (2011), 1182–85
- [163] Melanoma Treatment - General Information About Melanoma, <https://my.pearlpoint.org/resources/melanoma-treatment-general-information-about-melanoma>, consulté le 13 Novembre 2017
- [164] S.-Y. Huang et al., Ensemble Docking of Multiple Protein Structures: Considering Protein Structural Variations in Molecular Docking, *Proteins: Structure, Function, and Bioinformatics*, 66.2 (2007), 399–421
- [165] H. Alonso, et al., Combining Docking and Molecular Dynamic Simulations in Drug Design, *Medicinal Research Reviews*, 26.5 (2006), 531–68
- [166] D. J. Diller et al., Homology Models, and High Throughput Docking, *J. Med. Chem.*, 46.22 (2003), 4638–47
- [167] F. Sievers et al., Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega, *Molecular Systems Biology*, 7 (2011), 539
- [168] J. M. Hayes et al., Kinetics, in Silico Docking, Molecular Dynamics, and MM-GBSA Binding Studies on Prototype Indirubins, KT5720, and Staurosporine as Phosphorylase Kinase ATP-Binding Site Inhibitors: The Role of Water Molecules Examined, *Proteins*, 79.3 (2011), 703–19
- [169] M. Kaur et al., Exploring the Role of Water Molecules for Docking and Receptor Guided 3D-QSAR Analysis of Naphthyridine Derivatives as Spleen Tyrosine Kinase (Syk) Inhibitors, *Journal of Chemical Information and Modeling*, 52.10 (2012), 2619–30

- [170] National Center for Biotechnology Information (NCBI), Internet Bethesda (MD): National Library of Medicine (US), 1988, <https://www.ncbi.nlm.nih.gov>
- [171] S. F. Altschul et al., Basic Local Alignment Search Tool, *Journal of Molecular Biology*, 215.3 (1990), 403–10
- [172] National Center for Biotechnology Information (NCBI), Internet Bethesda (MD): National Library of Medicine (US), 1988, <https://www.ncbi.nlm.nih.gov>
- [173] A. Sali et al., Comparative Protein Modelling by Satisfaction of Spatial Restraints, *Journal of Molecular Biology*, 234.3 (1993), 779–815
- [174] UniProt: The Universal Protein Knowledgebase, *Nucleic Acids Research*, 45.D1 (2017), D158–69
- [175] P. Benkert et al., QMEAN: A Comprehensive Scoring Function for Model Quality Assessment, *Proteins*, 71.1 (2008), 261–77
- [176] RAMPAGE: Ramachandran Plot Assessment, <http://mordred.bioc.cam.ac.uk/~rapper/rampage.php>, consulté le 13 Novembre 2017.
- [177] A. Sharma et al., BioPhytMol: A Drug Discovery Community Resource on Anti-Mycobacterial Phytomolecules and Plant Extracts, *Journal of Cheminformatics*, 6.1 (2014), 46
- [178] K. Nakamura et al., KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites, *Plant & Cell Physiology*, 54.2 (2013), e4
- [179] M. Valli et al., Development of a Natural Products Database from the Biodiversity of Brazil, *Journal of Natural Products*, 76.3 (2013), 439–44
- [180] R. Hatherley et al., SANCDB: A South African Natural Compound Database, *Journal of Cheminformatics*, 7 (2015), 29
- [181] X. Lucas et al., StreptomeDB: A Resource for Natural Compounds Isolated from *Streptomyces* Species, *Nucleic Acids Research*, 41.Database issue (2013), D1130-1136
- [182] D. Klementz et al., StreptomeDB 2.0-an Extended Resource of Natural Products Produced by Streptomyces, *Nucleic Acids Research*, 44.D1 (2016), D509-514
- [183] C. Y.-C. Chen, TCM Database@Taiwan: The Worlds Largest Traditional Chinese Medicine Database for Drug Screening In Silico, *PLoS ONE*, 6.1 (2011), e15939
- [184] R. Xue et al., TCMID: Traditional Chinese Medicine Integrative Database for Herb Molecular Mechanism Analysis, *Nucleic Acids Research*, 41.Database issue (2013), D1089-1095
- [185] J. Xia et al., Benchmarking Methods and Data Sets for Ligand Enrichment Assessment in Virtual Screening, *Methods (San Diego, Calif.)*, 0 (2015), 146–57
- [186] W. P. Jencks, On the Attribution and Additivity of Binding Energies, *Proceedings of the National Academy of Sciences of the United States of America*, 78.7 (1981), 4046–50
- [187] P. R. Andrews et al., Functional Group Contributions to Drug-Receptor Interactions, *Journal of Medicinal Chemistry*, 27.12 (1984), 1648–57
- [188] P. J. Goodford, A Computational Procedure for Determining Energetically Favorable Binding Sites on Biologically Important Macromolecules, *Journal of Medicinal Chemistry*, 28.7 (1985), 849–57
- [189] A. Miranker et al., Functionality Maps of Binding Sites: A Multiple Copy Simultaneous Search Method, *Proteins*, 11.1 (1991), 29–34
- [190] H. J. Böhm, LUDI: Rule-Based Automatic Design of New Substituents for Enzyme Inhibitor Leads, *Journal of Computer-Aided Molecular Design*, 6.6 (1992), 593–606
- [191] C. L. Verlinde et al., In Search of New Lead Compounds for Trypanosomiasis Drug Design: A Protein Structure-Based Linked-Fragment Approach, *Journal of Computer-Aided Molecular Design*, 6.2 (1992), 131–47
- [192] S. B. Shuker et al., Discovering High-Affinity Ligands for Proteins: SAR by NMR, *Science (New York, N.Y.)*, 274.5292 (1996), 1531–34
- [193] I. Tickle et al., High-Throughput Protein Crystallography and Drug Discovery, *Chemical Society Reviews*, 33.8 (2004), 558–65
- [194] T. L. Blundell et al., High-Throughput X-Ray Crystallography for Drug Discovery, *Current Opinion in Pharmacology*, 4.5 (2004), 490–96
- [195] G. Chessari et al., From Fragment to Clinical Candidate—a Historical Perspective, *Drug Discovery Today*, 14.13–14 (2009), 668–75

- [196] M. Whittaker, Picking up the Pieces with FBDD or FADD: Invest Early for Future Success, *Drug Discovery Today*, 14.13–14 (2009), 623–24
- [197] R. J. Bienstock, Overview: Fragment-Based Drug Design, in *Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, ACS Symposium Series, 1076, 0 vols (American Chemical Society, 2011), mlxxvi, 1–26, consulté le 18 Mai 2016
- [198] H. Chen et al., Evolutions in Fragment-Based Drug Design: The Deconstruction–reconstruction Approach, *Drug Discovery Today*, 20.1 (2015), 105–13
- [199] R. A. E. Carr et al., Fragment-Based Lead Discovery: Leads by Design, *Drug Discovery Today*, 10.14 (2005), 987–92
- [200] D. E. Scott et al., Fragment-Based Approaches in Drug Discovery and Chemical Biology, *Biochemistry*, 51.25 (2012), 4990–5003
- [201] D. Erlanson, Fragments in the Clinic: 2016 Edition, <http://practicalfragments.blogspot.com/2016/07/fragments-in-clinic-2016-edition.html>, consulté le 12 Novembre 2017
- [202] D. A. Erlanson et al., Twenty Years on: The Impact of Fragments on Drug Discovery, *Nature Reviews. Drug Discovery*, 15.9 (2016), 605–19
- [203] S. Barelier et al., Ligand Specificity in Fragment-Based Drug Design, *Journal of Medicinal Chemistry*, 53.14 (2010), 5256–66
- [204] Y. Chen et al., Molecular Docking and Ligand Specificity in Fragment-Based Inhibitor Discovery, *Nature Chemical Biology*, 5.5 (2009), 358–64
- [205] A. Kumar et al., Fragment Based Drug Design: From Experimental to Computational Approaches, *Current Medicinal Chemistry*, 19.30 (2012), 5128–47
- [206] A. L. Hopkins et al., Ligand Efficiency: A Useful Metric for Lead Selection, *Drug Discovery Today*, 9.10 (2004), 430–31
- [207] M. D. Shultz, Setting Expectations in Molecular Optimizations: Strengths and Limitations of Commonly Used Composite Parameters, *Bioorganic & Medicinal Chemistry Letters*, 23.21 (2013), 5980–91
- [208] J. Polanski et al., Beware of Ligand Efficiency (LE): Understanding LE Data in Modeling Structure-Activity and Structure-Economy Relationships, *Journal of Cheminformatics*, 9 (2017)
- [209] P. J. Hajduk, Fragment-Based Drug Design: How Big Is Too Big?, *Journal of Medicinal Chemistry*, 49.24 (2006), 6972–76
- [210] R. J. Hall et al., Efficient Exploration of Chemical Space by Fragment-Based Screening, *Progress in Biophysics and Molecular Biology*, 116.2–3 (2014), 82–91
- [211] P. J. Hajduk et al., A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned, *Nat Rev Drug Discov*, 6.3 (2007), 211–19
- [212] S. Renner et al., Recent Trends and Observations in the Design of High-Quality Screening Collections, *Future Medicinal Chemistry*, 3.6 (2011), 751–66
- [213] G. M. Keserü 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*, 59.18 (2016), 8189–8206
- [214] H. Köster et al., A Small Nonrule of 3 Compatible Fragment Library Provides High Hit Rate of Endothiapepsin Crystal Structures with Various Fragment Chemotypes, *Journal of Medicinal Chemistry*, 54.22 (2011), 7784–96
- [215] J. B. Jordan et al., Fragment Based Drug Discovery: Practical Implementation Based on ¹⁹F NMR Spectroscopy, *Journal of Medicinal Chemistry*, 55.2 (2012), 678–87
- [216] Fluorine Fragment Library, <http://www.otavachemicals.com/products/fragment-libraries/19f-nmr-fluorine-containing-fragment-library>, consulté le 9 Novembre 2017
- [217] L. Ruddigkeit et al., Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17, *Journal of Chemical Information and Modeling*, 52.11 (2012), 2864–75
- [218] R. Visini et al., Fragment Database FDB-17, *Journal of Chemical Information and Modeling*, 57.4 (2017), 700–709
- [219] D. Joseph-McCarthy et al., Fragment-Based Lead Discovery and Design, *Journal of Chemical Information and Modeling*, 54.3 (2014), 693–704

- [220] S. H. Rotstein and M. A. Murcko, GroupBuild: A Fragment-Based Method for de Novo Drug Design, *Journal of Medicinal Chemistry*, 36.12 (1993), 1700–1710
- [221] R. S. Bohacek et al., Multiple Highly Diverse Structures Complementary to Enzyme Binding Sites: Results of Extensive Application of a de Novo Design Method Incorporating Combinatorial Growth, *Journal of the American Chemical Society*, 116.13 (1994), 5560–71
- [222] K. Biggadike et al., Design and X-Ray Crystal Structures of High-Potency Nonsteroidal Glucocorticoid Agonists Exploiting a Novel Binding Site on the Receptor, *Proceedings of the National Academy of Sciences of the United States of America*, 106.43 (2009), 18114–19
- [223] R. S. DeWitte et al., SMOG: De Novo Design Method Based on Simple, Fast, and Accurate Free Energy Estimates. 2. Case Studies in Molecular Design, *Journal of the American Chemical Society*, 119.20 (1997), 4608–17
- [224] B. A. Grzybowski et al., From Knowledge-Based Potentials to Combinatorial Lead Design in Silico, *Accounts of Chemical Research*, 35.5 (2002), 261–69
- [225] J. D. Durrant et al., AutoGrow: A Novel Algorithm for Protein Inhibitor Design, *Chemical Biology & Drug Design*, 73.2 (2009), 168–78
- [226] Y. Nishibata et al., Automatic Creation of Drug Candidate Structures Based on Receptor Structure. Starting Point for Artificial Lead Generation., *Tetrahedron*, 47.43 (1991), 8985–90
- [227] S. H. Rotstein et al., GenStar: A Method for de Novo Drug Design, *Journal of Computer-Aided Molecular Design*, 7.1 (1993), 23–43.
- [228] J. B. Moon et al., Computer Design of Bioactive Molecules: A Method for Receptor-Based de Novo Ligand Design, *Proteins: Structure, Function, and Bioinformatics*, 11.4 (1991), 314–28
- [229] N. P. Todorov et al., Evaluation of a Method for Controlling Molecular Scaffold Diversity in de Novo Ligand Design, *Journal of Computer-Aided Molecular Design*, 11.2 (1997), 175–92
- [230] C. W. Murray et al., PRO_SELECT: Combining Structure-Based Drug Design and Combinatorial Chemistry for Rapid Lead Discovery. 1. Technology, *Journal of Computer-Aided Molecular Design*, 11.2 (1997), 193–207
- [231] J. R. Damewood et al., NovoFLAP: A Ligand-Based de Novo Design Approach for the Generation of Medicinally Relevant Ideas, *Journal of Chemical Information and Modeling*, 50.7 (2010), 1296–1303
- [232] P. S. Kutchukian et al., FOG: Fragment Optimized Growth Algorithm for the de Novo Generation of Molecules Occupying Druglike Chemical Space, *J. Chem. Inf. Model.*, 49.7 (2009), 1630–42
- [233] F. Chevillard et al., De Novo Design of Selective Compounds: A Fragment-Based Pipeline Applied to the B2 Adrenergic Receptor, *Journal of Cheminformatics*, 6.Suppl 1 (2014), P25
- [234] J. D. Durrant et al., AutoGrow 3.0: An Improved Algorithm for Chemically Tractable, Semi-Automated Protein Inhibitor Design, *Journal of Molecular Graphics & Modelling*, 44 (2013)
- [235] S. A. Hudson et al., Overcoming the Limitations of Fragment Merging: Rescuing a Strained Merged Fragment Series Targeting Mycobacterium Tuberculosis CYP121, *Chemmedchem*, 8.9 (2013), 1451–56
- [236] P. O. Nikiforov et al., A Fragment Merging Approach towards the Development of Small Molecule Inhibitors of Mycobacterium Tuberculosis EthR for Use as Ethionamide Boosters, *Organic & Biomolecular Chemistry*, 14.7 (2016), 2318–26
- [237] S. J. Hughes et al., Fragment Based Discovery of a Novel and Selective PI3 Kinase Inhibitor, *Bioorganic & Medicinal Chemistry Letters*, 21.21 (2011), 6586–90
- [238] R. A. Lewis et al., Automated Site-Directed Drug Design Using Molecular Lattices, *Journal of Molecular Graphics*, 10.2 (1992), 66–78, 106
- [239] M. B. Eisen et al., HOOK: A Program for Finding Novel Molecular Architectures That Satisfy the Chemical and Steric Requirements of a Macromolecule Binding Site, *Proteins*, 19.3 (1994), 199–221
- [240] D. E. Clark et al., PRO-LIGAND: An Approach to de Novo Molecular Design. 1. Application to the Design of Organic Molecules, *Journal of Computer-Aided Molecular Design*, 9.1 (1995), 13–32.
- [241] A. Miranker et al., An Automated Method for Dynamic Ligand Design, *Proteins*, 23.4 (1995), 472–90
- [242] Dominique Douguet et al., LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design, *Journal of Medicinal Chemistry*, 48.7 (2005), 2457–68

- [243] D. C. Thompson et al., CONFIRM: Connecting Fragments Found in Receptor Molecules, *Journal of Computer-Aided Molecular Design*, 22.10 (2008), 761–72
- [244] F. Dey et al., Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization, *Journal of Chemical Information and Modeling*, 48.3 (2008), 679–90
- [245] B. Nisius et al., Fragment Shuffling: An Automated Workflow for Three-Dimensional Fragment-Based Ligand Design, *J. Chem. Inf. Model.*, 49.5 (2009), 1211–22
- [246] V. Tschinke et al., The NEWLEAD Program: A New Method for the Design of Candidate Structures from Pharmacophoric Hypotheses, *Journal of Medicinal Chemistry*, 36.24 (1993), 3863–70
- [247] C. M. W. Ho et al., SPLICE: A Program to Assemble Partial Query Solutions from Three-Dimensional Database Searches into Novel Ligands, *Journal of Computer-Aided Molecular Design*, 7.6 (1993), 623–47
- [248] J. Degen et al., FlexNovo: Structure-Based Searching in Large Fragment Spaces, *ChemMedChem*, 1.8 (2006), 854–68
- [249] Schrödinger Release 2017-3: Canvas, Schrödinger, LLC, New York, NY, 2017., `Combine_fragments.Py`.
- [250] D. A. Pearlman et al., CONCEPTS: New Dynamic Algorithm for de Novo Drug Suggestion, *Journal of Computational Chemistry*, 14.10 (1993), 1184–93
- [251] D. A. Pearlman et al., CONCERTS: Dynamic Connection of Fragments as an Approach to de Novo Ligand Design, *J. Med. Chem.*, 39.8 (1996), 1651–63
- [252] V. J. Gillet et al., SPROUT: Recent Developments in the de Novo Design of Molecules, *Journal of Chemical Information and Computer Sciences*, 34.1 (1994), 207–17
- [253] V. J. Gillet et al., SPROUT, HIPPO and CAESA: Tools for de Novo Structure Generation and Estimation of Synthetic Accessibility, *Perspectives in Drug Discovery and Design*, 3.1 (1995), 34–50
- [254] S. G. et al., “Scaffold-Hopping” by Topological Pharmacophore Search: A Contribution to Virtual Screening, *Angewandte Chemie (International Ed. in English)*, 38.19 (1999), 2894–96
- [255] R. D. Cramer et al., “Lead Hopping”. Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities, *Journal of Medicinal Chemistry*, 47.27 (2004), 6777–91
- [256] H.-J. Böhm et al., Scaffold Hopping, *Drug Discovery Today. Technologies*, 1.3 (2004), 217–24
- [257] LEAPFROG TRIPOS ASSOCIATES, INC. Trademark Information, [Trademarkia.Com](https://www.trademarkia.com/leapfrog-tripos-associates-inc-74380604.html), <https://www.trademarkia.com/leapfrog-tripos-associates-inc-74380604.html>, consulté le 3 Novembre 2017
- [258] G. Lauri et al., CAVEAT: A Program to Facilitate the Design of Organic Molecules, *Journal of Computer-Aided Molecular Design*, 8.1 (1994), 51–66
- [259] A. C. Pierce et al., BREED: Generating Novel Inhibitors through Hybridization of Known Ligands. Application to CDK2, P38, and HIV Protease, *Journal of Medicinal Chemistry*, 47.11 (2004), 2768–75
- [260] F. Fontaine et al., SHOP: A Method for Structure-Based Fragment and Scaffold Hopping, *ChemMedChem*, 4.3 (2009), 427–39
- [261] A.-J. Jakobi et al., ParaFrag--an Approach for Surface-Based Similarity Comparison of Molecular Fragments, *Journal of Molecular Modeling*, 14.7 (2008), 547–58
- [262] Schrödinger Release 2016-4: CombiGlide, Schrödinger, LLC, New York, NY, 2016
- [263] O. Sperandio et al., MED-SuMoLig: A New Ligand-Based Screening Tool for Efficient Scaffold Hopping, *Journal of Chemical Information and Modeling*, 47.3 (2007), 1097–1110
- [264] Scaffold Replacement in MOE, <https://www.chemcomp.com/journal/newscaffold.htm>, consulté le 3 Novembre 2017
- [265] S. Lindert et al., LigMerge: A Fast Algorithm to Generate Models of Novel Potential Ligands from Sets of Known Binders, *Chemical Biology & Drug Design*, 80.3 (2012), 358–65
- [266] Core Hopping | Schrödinger, <https://www.schrodinger.com/core-hopping>, consulté le 3 Novembre 2017
- [267] M. J. Vainio et al., Scaffold Hopping by Fragment Replacement, *Journal of Chemical Information and Modeling*, 53.7 (2013), 1825–35

- [268] OpenEye Scientific Software, BROOD | Lead Optimization, <https://www.eyesopen.com/brood>, consulté le 19 Octobre 2017
- [269] J. J Irwin et al., ZINC--a Free Database of Commercially Available Compounds for Virtual Screening, *Journal of Chemical Information and Modeling*, 45.1 (2005), 177–82
- [270] R. Wang et al., LigBuilder: A Multi-Purpose Program for Structure-Based Drug Design, *Molecular Modeling Annual*, 6.7–8 (2000), 498–516
- [271] Y. Yuan et al., LigBuilder 2: A Practical de Novo Drug Design Approach, *Journal of Chemical Information and Modeling*, 51.5 (2011), 1083–91
- [272] G.-B. Li et al., LEADOPT: An Automatic Tool for Structure-Based Lead Optimization, and Its Application in Structural Optimizations of VEGFR2 and SYK Inhibitors, *European Journal of Medicinal Chemistry*, 93 (2015), 523–38
- [273] P. Maass et al., Recore: A Fast and Versatile Method for Scaffold Hopping Based on Small Molecule Crystal Structure Conformations, *Journal of Chemical Information and Modeling*, 47.2 (2007), 390–99
- [274] Spark: Find New Chemistry Ideas Using Bioisosteres, Cresset, 2012, <http://www.cresset-group.com/products/spark/>, consulté le 3 Novembre 2017
- [275] G.-F. Hao et al., ACFIS: A Web Server for Fragment-Based Drug Discovery, *Nucleic Acids Research*, 2016
- [276] ACFIS Server, <http://chemyang.ccnu.edu.cn/ccb/server/ACFIS/>, consulté le 20 Octobre 2017
- [277] N. M. OBoyle et al., Cinfony--Combining Open Source Cheminformatics Toolkits behind a Common Interface, *Chemistry Central Journal*, 2 (2008), 24
- [278] P. J. A. Cock et al., Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics, *Bioinformatics (Oxford, England)*, 25.11 (2009), 1422–23
- [279] X. Q. 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*, 38.3 (1998), 511–22
- [280] A. Schuffenhauer et al., The Scaffold Tree – Visualization of the Scaffold Universe by Hierarchical Scaffold Classification, *Journal of Chemical Information and Modeling*, 47.1 (2007), 47–58
- [281] S. Raschka, BioPandas: Working with Molecular Structures in Pandas DataFrames, *The Journal of Open Source Software*, 2.14 (2017), 279
- [282] Multi-Step Protocol for HTVS | RDock <<http://rdock.sourceforge.net/multistep-protocol-for-htvs/>> [consulté le 18 Octobre 2017]
- [283] N. Majeux et al., Exhaustive Docking of Molecular Fragments with Electrostatic Solvation, *Proteins*, 37.1 (1999), 88–105
- [284] N Majeux, M Scarsi, and A Caflisch, Efficient Electrostatic Solvation Model for Protein-Fragment Docking, *Proteins*, 42.2 (2001), 256–68
- [285] L. Hoffer, Jean-Paul Renaud, and Dragos Horvath, In Silico Fragment-Based Drug Discovery: Setup and Validation of a Fragment-to-Lead Computational Protocol Using S4MPLE, *Journal of Chemical Information and Modeling*, 53.4 (2013), 836–51
- [286] D. D. Ridley, Strategies for Chemical Reaction Searching in SciFinder, *Journal of Chemical Information and Computer Sciences*, 40.5 (2000), 1077–84
- [287] M. Loubidi, Synthèse et Réactivité de Bicycles Imidazo[1,2- α]Imidazo[1,5- α]Imidazoles à Visée Thérapeutique (Université d'Orléans, 2017)
- [288] E. Beurel et al., Glycogen Synthase Kinase-3 (GSK3): Regulation, Actions, and Diseases, *Pharmacology & Therapeutics*, 0 (2015), 114–31
- [289] A. Luca et al., Gsk3 Signalling and Redox Status in Bipolar Disorder: Evidence from Lithium Efficacy, *Oxidative Medicine and Cellular Longevity*, 2016 (2016)
- [290] M. Llorens-Marín et al., GSK-3 β , a Pivotal Kinase in Alzheimer Disease, *Frontiers in Molecular Neuroscience*, 7 (2014)
- [291] M. Golpich et al., Glycogen Synthase Kinase-3 Beta (GSK-3 β) Signaling: Implications for Parkinsons Disease, *Pharmacological Research*, 97 (2015), 16–26

- [292] S. AlFadhli et al., De-Regulation of Diabetic Regulatory Genes in Psoriasis: Deciphering the Unsolved Riddle, *Gene*, 593.1 (2016), 110–16
- [293] I. Nasri et al., Anti-Inflammatory and Anticancer Effects of Flavonol Glycosides from *Diplotaxis Harra* through GSK3 β Regulation in Intestinal Cells, *Pharmaceutical Biology*, 55.1 (2017), 124–31
- [294] A. Felipe Aristizabal-Pachon et al., Role of GSK3 β in Breast Cancer Susceptibility, *Cancer Biomarkers: Section A of Disease Markers*, 18.2 (2017), 169–75
- [295] J. A. McCubrey et al., GSK-3 as Potential Target for Therapeutic Intervention in Cancer, *Oncotarget*, 5.10 (2014), 2881–2911
- [296] R. V. Bhat et al., Glycogen Synthase Kinase 3: A Drug Target for CNS Therapies, *Journal of Neurochemistry*, 89.6 (2004), 1313–17
- [297] F. Lo Monte et al., Identification of Glycogen Synthase Kinase-3 Inhibitors with a Selective Sting for Glycogen Synthase Kinase-3 α , *Journal of Medicinal Chemistry*, 55.9 (2012), 4407–24
- [298] B. Baratte et al., Le criblage à Roscoff - Une recherche d'inhibiteurs de kinases tournée vers la mer, *médecine/sciences*, 31.5, 538–45
- [299] H. Davies et al., Mutations of the BRAF Gene in Human Cancer, *Nature*, 417.6892 (2002), 949–54
- [300] T. K. Eigentler et al., Protein Kinase Inhibitors in Melanoma, *Expert Opinion on Pharmacotherapy*, 14.16 (2013), 2195–2201
- [301] R. Arora et al., Structural Investigation of B-Raf Paradox Breaker and Inducer Inhibitors, *Journal of Medicinal Chemistry*, 58.4 (2015), 1818–31
- [302] C. Montagut et al., Elevated CRAF as a Potential Mechanism of Acquired Resistance to BRAF Inhibition in Melanoma, *Cancer Research*, 68.12 (2008), 4853–61
- [303] P. Ertl et al., Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions, *Journal of Cheminformatics*, 1.1 (2009), 8
- [304] P. Bonnet, Is Chemical Synthetic Accessibility Computationally Predictable for Drug and Lead-like Molecules? A Comparative Assessment between Medicinal and Computational Chemists, *European Journal of Medicinal Chemistry*, 54 (2012), 679–89
- [305] J. Degen et al., On the Art of Compiling and Using “drug-like” Chemical Fragment Spaces, *ChemMedChem*, 3.10 (2008), 1503–7
- [306] Antony J. Williams et al., Open PHACTS: Semantic Interoperability for Drug Discovery, *Drug Discovery Today*, 17.21–22 (2012), 1188–98
- [307] M. L. Verdonk et al., Docking Performance of Fragments and Druglike Compounds, *Journal of Medicinal Chemistry*, 54.15 (2011), 5422–31
- [308] D. Kozakov et al., ‘Ligand Deconstruction: Why Some Fragment Binding Positions Are Conserved et al. Are Not’, *Proceedings of the National Academy of Sciences of the United States of America*, 112.20 (2015), E2585-2594
- [309] M. N. Drwal et al., Do Fragments and Crystallization Additives Bind Similarly to Drug-like Ligands?, *Journal of Chemical Information and Modeling*, 57.5 (2017), 1197–1209

José-Manuel GALLY

DEVELOPPEMENT D'OUTILS DE CHEMOINFORMATIQUE POUR L'IDENTIFICATION D'INHIBITEURS DE PROTEINES KINASES A PARTIR DE FRAGMENTS

Résumé :

La conception de médicaments est un long processus complexe impliquant de nombreuses disciplines différentes. Entre la découverte de la touche (molécule active initiale) puis de la tête de série (molécule optimisée), jusqu'à la mise sur le marché du produit fini (médicament), il s'écoule en général une dizaine d'année, pour un coût avoisinant le milliard d'euros. Afin d'optimiser ce processus, de nombreuses approches sont développées pour identifier plus rapidement les molécules les plus prometteuses pour une cible thérapeutique donnée. Les protéines kinases (PK) jouent un rôle majeur dans la signalisation moléculaire et dans les mécanismes cellulaires. La dérégulation d'une PK peut engendrer des pathologies graves telles que des maladies neurodégénératives ou le cancer ; la conception d'inhibiteurs de PK (PKI) est ainsi un domaine de recherche thérapeutique extrêmement actif. Dans ce manuscrit de thèse, deux approches de chémoinformatique complémentaires sont explorées pour l'identification de nouveaux PKI. Tout d'abord, un criblage virtuel de produits naturels a été réalisé dans un contexte de recherche de molécules bioactives à finalité cosmétique ou thérapeutique. Pour ce travail, il a d'abord été nécessaire de développer un outil *in silico* de préparation de molécules, *VSPrep*, qui s'appuie sur des outils gratuits pour les chercheurs académiques. Enfin, un nouvel outil de conception de molécules bioactives à partir de fragments moléculaires (FBDD) a également été développé. L'implémentation de cet outil, *Frag2Drugs*, ainsi que sa validation dans des projets FBDD sont décrites dans ce manuscrit.

Mots : protéines kinases, préparation de ligands, *VSPrep*, produits naturels, fragments, *Frag2Drugs*, conception de médicaments

DEVELOPEMENT OF CHEMOINFORMATICS TOOLS FOR THE IDENTIFICATION OF PROTEIN KINASE INHIBITORS FROM FRAGMENTS

Summary :

Drug design is a long and complex multidisciplinary process. From the discovery of the initial hit (bioactive molecule), to the lead (optimized compound), and finally to the commercialization of the end product (drug), almost 10 years are necessary and this process costs an average of \$1 billion dollars. In order to optimize this process, new methods are relentlessly developed so that novel promising molecules might be identified faster for a given target. Protein kinases (PK) play a central role in most molecular pathways and are essential to control cellular mechanisms. The mutation of one PK can lead to severe pathologies such as neurodegenerative diseases or cancer, making the research of PK inhibitors (PKI) an intense area of therapeutic research. In this manuscript, two complementary chemoinformatics approaches are discussed for identifying PKI, and both depend on the preparation of small molecules (ligands). For this specific task, a new workflow protocol, *VSPrep*, was developed using only freely available tools for academics. First, a virtual screening approach of natural products was performed in order to identify bioactive molecules for cosmetics or therapeutic applications. Second, a novel *in silico* Fragment-Based Drug Discovery (FBDD) tool, *Frag2Drugs*, was developed specifically for kinase research. It combines molecular fragments derived from PKI into novel inhibitors bound to specific protein kinases. The tool was validated on several internal kinase projects leading to novel protein kinase inhibitors with acceptable drug-like properties.

Keywords : protein kinases, ligand preparation, *VSPrep*, natural products, fragments, *Frag2Drugs*, Drug Design

Institut de Chimie Organique et Analytique

**UMR CNRS – Université d'Orléans 7311
rue de Chartres, 45067 Orléans**