Université de Nantes

Departement de Bio-informatique

Mémoire de stage au CRCI2NA : Conception

de molécules de novo in silico à partir

d’interactions protéiques.

MENARD Théo

Tuteur : Maillasson Mike

Rapport soumis d’après les exigence de

l’Université de Nantes dans le

Master 1 de Bio-informatique Bio-statistique

2

2 mai 2025



Abstract

Ce projet s’inscrit dans un pipeline qui a pour objectif la génération in silico d’inhibiteurs,

dirigés contre une protéine cible, à partir d’une interaction protéine-protéine (iPP). Plus spé-

ciﬁquement, ce travail se situe en amont du pipeline et a pour but d’amener à la création d’un

programme qui identiﬁe les acides aminés impliqués dans une iPP. En l’absence de ligands

connus, nous proposons une stratégie de génération qui va utiliser la capacité générative d’un

autoencodeur existant Une première étape a été d’extraire les chaînes latérales qui consti-

tuent la ’point-chaud’ ou hotspot de l’interaction. Puis, à partir des chaînes latérales de ces

acides aminés, créer une pseudo-molécule capable de se ﬁxer à la protéine cible en reliant

les chaînes latérales via d’autres fragments que nous appellerons ici Linker - Nous nous ap-

puierons sur la banque de fragment Enamine comprehensive linker -. Cette liaison a pour

but de conserver la géométrie spatiale initiale entre les deux chaînes latérales. Par la suite,

une vériﬁcation par docking automatisé a été eﬀectué pour trouver les meilleures pseudo-

molécules parmi la génération d’une cinquantaine de celle-ci. Les meilleures pseudo-molécules

constituent alors une base d’entrée pour l’autoencodeur qui va apporter des modiﬁcations

aléatoires dans l’espoir d’optimiser encore plus la molécule. Les résultats sur deux interactions

connues IL2/IL2Ralpha et USP7/Ubiquitine en comparaison avec deux inhibiteurs connus,

respectivement FRH et EZF, ont permis de montrer la génération de molécules avec une plus

haute aﬃnité que le ligand de référence.

This project is part of a pipeline aimed at the in silico generation of inhibitors targe-

ting a speciﬁc protein, based on a protein-protein interaction (PPI). More speciﬁcally, this

work is positioned upstream in the pipeline and aims to create a program that identiﬁes the

amino acids involved in a PPI. In the absence of known ligands, we propose a generation

strategy that leverages the generative capability of an existing autoencoder. The ﬁrst step

involved extracting the side chains that form the "hotspot" of the interaction. Then, based

on the side chains of these amino acids, we generated a pseudo-molecule capable of bin-

ding to the target protein by connecting the side chains via additional fragments, referred

to here as linkers—speciﬁcally sourced from the Enamine comprehensive linker library. This

linking process is intended to preserve the original spatial geometry between the two side

chains. Subsequently, an automated docking veriﬁcation was performed to identify the best

pseudo-molecules among a set of around ﬁfty generated candidates. The top pseudo-molecules

then serve as input for the autoencoder, which introduces random modiﬁcations in the hope

of further optimizing the molecule. The results on two known interactions—IL2/IL2Ralpha

and USP7/Ubiquitin—compared with two known inhibitors, FRH and EZF respectively, de-

monstrated the generation of molecules with higher binding aﬃnity than the reference ligands.

Mots-clés : Bio-informatique, Inhibiteur, Protéines, autoencodeur, génération, de novo

Nombre total de mots : 3 832 mots.

i

Remerciements

Je tiens tout d’abord à remercier chaleureusement MAILLASSON Mike, mon tuteur de

stage, pour son encadrement bienveillant, sa disponibilité et ses conseils précieux tout au long

de ce projet. Son soutien constant et ses retours constructifs ont été essentiels à la progression

de mon travail.

Je remercie également GUITTENY Sarah, ma co-stagiaire, avec qui j’ai eu le plaisir de

collaborer tout au long de l’année. Nos échanges réguliers, tant sur le plan scientiﬁque que

technique, ont enrichi nos deux projets respectifs, dont les thématiques se complétaient de

manière naturelle.

Je souhaite exprimer ma reconnaissance à MORTIER Erwan, responsable scientiﬁque

du projet, pour son sourire et la compréhension dont il a fait preuve lorsque nous rencontrions

des diﬃcultés.

Un grand merci aussi à QUEMENER Agnès, spécialiste en modélisation moléculaire,

pour le temps qu’elle m’a consacré, en particulier dans l’utilisation du logiciel Discovery Stu-

dio, et pour ses explications toujours claires et pédagogiques.

Enﬁn, je remercie l’ensemble de l’équipe 12 pour leur accueil chaleureux et leur bonne

humeur au quotidien : Pierre, Coraly, Agath et Nina, merci pour vos échanges, vos conseils,

et l’ambiance agréable qui a régné tout au long du stage.

ii

Table des matières

[Table des ﬁgures](#br6)

[Liste des tableaux](#br7)

v

vi

[1](#br9)

[Introduction et Contexte](#br9)

1

1

2

2

2

3

3

3

[1](#br9)

[1](#br9)

[1](#br10)

[1](#br10)

[1](#br10)

[.1 Vision globale du projet](#br9) . . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.2 Contexte et environnement](#br9) . . . . . . . . . . . . . . . . . . . . . . . . . .

[.3 Problématique](#br10) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.4 Objectif et tâches](#br10) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.5 Solutions envisagées](#br10) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

[1](#br11)

[1](#br11)

[.5.1 Approche par deep learning](#br11) . . . . . . . . . . . . . . . . . . . . . .

[.5.2 Utilisation d’une banque de fragments](#br11) . . . . . . . . . . . . . . . . .

[2](#br12)

[3](#br14)

[Extraction des acides aminés clés](#br12)

4

4

5

[2](#br12)

[2](#br12)

[.1 Identiﬁcation des résidus en interaction à partir de ﬁchiers PDB](#br12) . . . . . . .

[.2 Comparaison avec Discovery Studio . . . . . . . . . . . . . . .](#br12) . . . . . . .

[Génération de pseudo-molécules](#br14)

6

6

6

7

7

[3](#br14)

[.1 Construction par assemblage de fragments](#br14) . . . . . . . . . . . . . . . . . .

[3](#br14)

[3](#br14)

[3](#br15)

[.1.1 Intégration d’une banque de linkers : Enamine](#br14) . . . . . . . . . . . .

[.1.2 RDKit et Openbabel . . . . . . . . . . . . .](#br14) . . . . . . . . . . . . .

[.1.3 Contraintes et ajustements techniques](#br15) . . . . . . . . . . . . . . . .

[4](#br16)

[Docking automatisé](#br16)

8

8

8

9

9

[4](#br16)

[4](#br16)

[.1 Automatisation de la procédure de docking](#br16) . . . . . . . . . . . . . . . . . .

[.2 Validation par comparaison avec un ligand de référence](#br16) . . . . . . . . . . . .

[4](#br16)

[.2.1 Extraction automatique du ligand et du récepteur](#br16) [à partir du PDB](#br17) .

[4](#br17)

[.3 Résultats et observations . . . . . . . . . . . . . . . . . . . . . . . . . .](#br17) . .

[5](#br18)

[6](#br21)

[Vériﬁcation et comparaison](#br18)

10

10

12

12

[5](#br18)

[5](#br18)

[5](#br20)

[.1 Analyse structurale et similarité chimique](#br18) . . . . . . . . . . . . . . . . . . .

[.2 Candidats potentiels . . . . . . . . . . .](#br18) . . . . . . . . . . . . . . . . . . .

[.3 Amélioration de la banque de fragments](#br20) . . . . . . . . . . . . . . . . . . . .

[Limites, perspectives et conclusion](#br21)

13

13

13

13

13

[6](#br21)

[6](#br21)

[.1 Limites . . . . . . . . . . . . . .](#br21) . . . . . . . . . . . . . . . . . . . . . . .

[6](#br21)

[6](#br21)

[.1.1 Obstacles rencontrés](#br21) . . . . . . . . . . . . . . . . . . . . . . . . . .

[.1.2 Limite de la méthode utilisé](#br21) . . . . . . . . . . . . . . . . . . . . . .

[.2 Perspectives d’amélioration](#br21) . . . . . . . . . . . . . . . . . . . . . . . . . .

iii

TABLE DES MATIÈRES

iv

[6](#br22)

[6](#br22)

[.3 Avantages et inconvénient de la méthodes](#br22) . . . . . . . . . . . . . . . . . .

[.4 Conclusion générale . . . . . . . . . . .](#br22) . . . . . . . . . . . . . . . . . . .

14

14

[Bibliographie](#br23)

15

16

16

17

[Appendices](#br24)

[A An Appendix](#br24)

[B An Appendix Chapter (Optional)](#br25)

Table des ﬁgures

[1](#br9)

[.1 Vu globale du pipeline](#br9) . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

1

[5](#br19)

[5](#br19)

[5](#br19)

[5](#br19)

[.1 Histogrammes descriptifs](#br19) . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.2 Régression linéaire . . .](#br19) . . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.3 (a) QED-IL2 (b) SAS-IL2](#br19) . . . . . . . . . . . . . . . . . . . . . . . . . . .

[.4 Meilleurs candidats . . . .](#br19) . . . . . . . . . . . . . . . . . . . . . . . . . . .

11

11

11

12

v

Liste des tableaux

[2](#br13)

[4](#br17)

[.1 Tableau de la comparaison des acides aminés retenues pour l’interaction IL2/IL2Ralpha.](#br13) 5

[.1 Meilleurs SMILES](#br17) . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .

9

vi

Liste des Abbreviations

CRCI2NA

DS

Centre de Recherche en Cancérologie et Immunologie Nantes-Angers

Discovery Studio

FMO

GLiCID

HBD

Fragment Molecular Orbital

Groupement Ligérien pour le Calcul Intensif Distribué

Hydrogen Bond Donors

HBA

IMPACT

IL2

Hydrogen Bond Acceptors

Interactions Moléculaires Puces Activités

Interleukine 2

iPP

LP

Interaction Protéine-Protéine

LogP

MW

Molecular Weight

PDB

Protein Data Bank

PROTAC

QED

SAS

Proteolysis Targeting Chimeras

Quantitative Estimate of Drug-likeness

Synthetic Accessibility Score

Standard Data Format

SDF

SMILES

TPSA

Simpliﬁed Molecular-Input Line-Entry System

Topological Polar Surface Area

vii

Chapitre 1

Introduction et Contexte

L’objectif principal du travail a été la mise en place d’un pipeline bio-informatique semi-

automatisée, allant de l’analyse d’une interaction protéique jusqu’à la génération et l’évalua-

tion de pseudo-ligands. Ce mémoire présente les diﬀérentes étapes de cette démarche, les

outils mobilisés, les diﬃcultés rencontrées, ainsi que les résultats obtenus et les perspectives

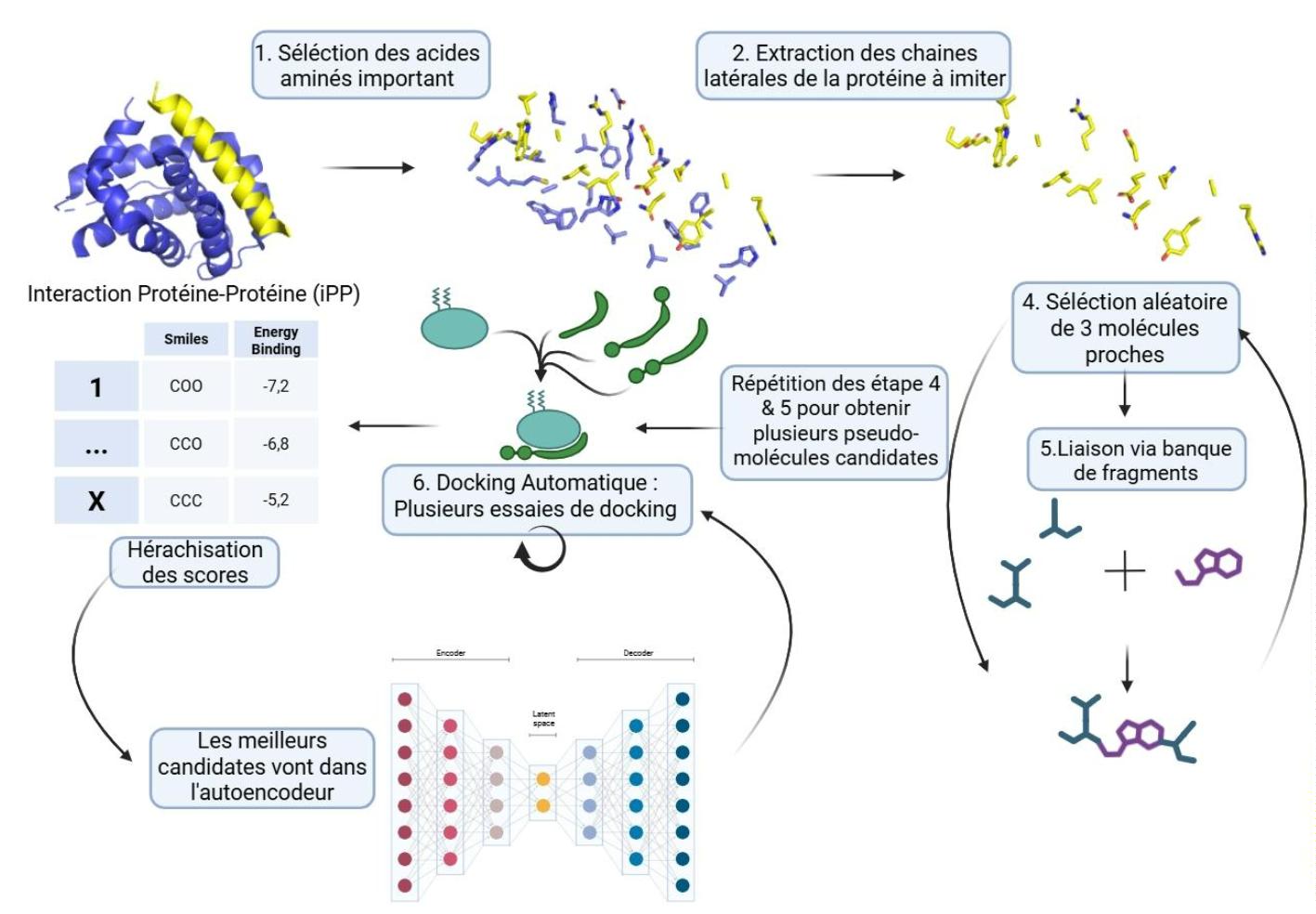
envisagées.

1

.1 Vision globale du projet

Figure 1.1 : Vue globale du pipeline

1



CHAPITRE 1. INTRODUCTION ET CONTEXTE

2

1

.2 Contexte et environnement

Ce stage a été réalisé au sein de l’équipe 12 du CRCI2NA (Centre de Recherche en Can-

cérologie et Immunologie Nantes-Angers), pour la plateforme IMPACT (Interactions Molé-

culaires Puces Activités) et avec l’utilisation du service GLiCID (Groupement Ligérien pour

le Calcul Intensif Distribué).

IMPACT : Une plateforme de protéomique regroupant plusieurs outils d’analyse et tech-

nologies dédiées à l’étude globale des interactions biomoléculaires.

L’environnement de travail reposait sur l’utilisation du langage Python et de bibliothèques

spécialisées telles que RDKit(Chémio-informatique), TensorFlow(Réseaux neuronaux) ou

encore Biopython, avec une manipulation régulière de ﬁchiers structuraux au format PDB

(Proteine Data Bank) ou de structures moléculaires SMILES (Simpliﬁed Molecular-Input

Line-Entry System). Plusieurs programmes étaient déjà mis en place pour permettre la gé-

nération d’un modèle de deep learning via un réseau de neurones (cf. chapitre Autoencodeur).

SMILES : Notation linéaire spéciﬁque traduisant la conformation 2D d’une molécule en

utilisant les caractères ASCII.

1

.3 Problématique

Les interactions protéine-protéine (iPP) jouent un rôle central dans de nombreux pro-

cessus biologiques, qu’il s’agisse de signalisation cellulaire, de régulation enzymatique ou d’as-

semblage de complexes multi-protéiques.

Par exemple, l’interleukine 2 (IL2), molécule pro-inﬂammatoire sur laquelle il existe des

ligands inhibiteurs connus et que nous utiliserons pour vériﬁer la pertinence du programme

mis en place. Il est possible de trouver un grand nombre d’interactions protéiques et leur

inhibiteur associé sur la plateforme Inhibitors of Protein-Protein Interaction Database [[Labbé](#br23)

[et al.](#br23) ([2015](#br23))].

[La modu](#br23)lation de ces interactions constitue un enjeu majeur pour la découverte de nou-

veaux médicaments, notamment dans le cadre de pathologies impliquant des dérèglements de

réseaux protéiques. Toutefois, la conception de petites molécules capables d’interférer eﬃcace-

ment avec une interface protéique reste un déﬁ important, en raison de la surface généralement

large et peu creusée de ces interfaces, ce qui amène à un glissement du ligand sur la molécule

cible.

1

.4 Objectif et tâches

Objectif : Ce stage s’est inscrit dans cette problématique, avec pour objectif de concevoir

des pseudo-molécules capables de perturber une protéine cible à partir d’une interface protéine-

protéine. Les informations structurales proviennent de complexes cristallisés enregistrés sur la

plateforme PDB.

La démarche retenue consiste à identiﬁer les résidus d’acides aminés impliqués dans l’in-

teraction, à les transformer en fragments chimiques, puis à les relier par des linkers pour

générer des composés synthétiquement accessibles. Ces composés sont ensuite évalués par

docking moléculaire, avant d’être optimisés par des approches d’apprentissage automatique,

CHAPITRE 1. INTRODUCTION ET CONTEXTE

3

notamment un autoencodeur moléculaire.

Tâches :Analyse structurelle de l’interaction protéique à partir d’un ﬁchier PDB

—

Identiﬁcation des acides aminés importants dans une iPP en s’appuyant sur des

paramètres structuraux et fonctionnels (hydrophobicité, interactions non covalentes,

complémentarité électrostatique, etc.)(Cf.chapitre [2](#br12))

—

—

Extraction des chaînes latérales.(Cf.chapitre [2)](#br12)

Liaison des chaînes latérales entre elles à [l’aide](#br12) de l’outil RDKit, en cherchant à

respecter les distances et orientations initiales des résidus dans l’espace.(Cf.chapitre [3](#br14))

Automatisation du docking pour identiﬁer les meilleurs candidats.(Cf.chapitre [4](#br16))

—

1

.5 Solutions envisagées

L’utilisation de scripts Python pour automatiser la gestion des ﬁchiers Protein Data Bank

(PDB). La liaison des chaînes latérales se fait via une banque de fragments moléculaires

capable de lier deux molécules (linker).

1

.5.1 Approche par deep learning

Une approche par Deep Learning a été envisagée. Cette dernière s’inspirait de Github

de références (Exemple : [[Imrie et al.](#br23) ([2020](#br23))]). Cependant, par souci de temps et à cause

de problèmes de versions [ou de licences, elle](#br23) a dû être écartée. Cette solution représente

maintenant une piste d’amélioration.

1

.5.2 Utilisation d’une banque de fragments

La seconde approche a été d’utiliser une banque de fragments pour lier les chaînes latérales

entre elles. Pour ce projet, nous nous sommes basés sur la banque de données Enamine

comprehensive linker comprenant 18 604 composants. Cette banque a été nettoyée et triée

au préalable par un autre scripte.

Chapitre 2

Extraction des acides aminés clés

La première étape du pipeline de conception a consisté à analyser les structures de com-

plexes protéine-protéine pour en extraire les acides aminés impliqués de manière signiﬁca-

tive dans l’interface d’interaction. Cette extraction structurelle, essentielle pour générer des

pseudo-ligands pertinents, repose sur des critères géométriques et physico-chimiques.

2

.1 Identiﬁcation des résidus en interaction à partir de

ﬁchiers PDB

Voici la série d’étapes mise en place :

—

Séparer les chaînes protéiques du complexe dans des ﬁchiers PDB diﬀérents pour

en faciliter l’utilisation plus tard ;

—

Filtrer les chaînes similaires aﬁn d’éviter les artefacts dus à la symétrie cristalline ou

à la présence de multimères. Un alignement de séquence des chaînes entre elles, avec

un seuil de 90% de similarité toléré, permet d’identiﬁer quelles chaînes sont à écarter

de l’analyse. Une reconnaissance du collagène est incluse au programme pour éviter

d’omettre l’analyse des brins d’un même multimère collagénique ;

—

—

Calculer les distances inter-résidus entre les atomes Ca(Carbone alpha) ou les

atomes latéraux. Les résidus situés à une distance inférieure à 5.2 Å de la chaîne parte-

naire ont été retenus comme candidats à l’interaction. Ce seuil est aussi dépendant du

type d’interaction (cf. plus bas). Il a été observé que certains ﬁchiers PDB d’interaction

protéique avaient des distances entre les 2 protéines supérieures à 8 Å, dans ce cas, le

programme n’est pas en capacité d’identiﬁer des acides aminés candidats.

Aﬃnement physico-chimique. Une étape d’aﬃnement physico-chimique a ensuite

permis de ﬁltrer les résidus sur la base de leurs propriétés chimiques, les paramètres pris

en compte sont ceux qu’utilise Discovery Studio :

•

•

•

Interaction hydrophobe entre les résidus de type Leu, Ile, Phe etc. Le seuil reste

,2 Å.

Interactions électrostatiques (présence de Lys, Arg, Asp, Glu), pour un seuil de

,0 Å.

Capacité de liaison hydrogène (atomes donneurs et accepteurs) avec un seuil de

.4 Å.

5

5

3

Cette approche s’inspire des critères employés par le logiciel Discovery Studio, qui a

également été utilisé ponctuellement à des ﬁns de validation visuelle. Le script en Python a

4

CHAPITRE 2. EXTRACTION DES ACIDES AMINÉS CLÉS

5

été conçu pour prendre en compte ces propriétés à partir des séquences et des coordonnées

atomiques extraites.

Au terme de cette étape, un sous-ensemble de résidus d’acides aminés a été identiﬁé

comme constituant l’essentiel de l’interface. Chaque résidu a été isolé dans un ﬁchier PDB in-

dividuel, en vue de sa recombinaison ultérieure dans la génération de pseudo-molécules. Cette

structuration a permis une automatisation plus simple du traitement par la suite.

Cependant, le programme reste améliorable. En eﬀet, les interactions conservées ne prennent

pas en compte l’angulation des acides aminés entre eux, ce qui peut porter préjudice quant à

la quantité d’acides aminés considérés comme importants dans cette interaction, notamment

pour les interactions de type hydrophobe et hydrogène.

2

.2 Comparaison avec Discovery Studio

Chaîne A

Chaîne B

~~Programme~~

~~Programme~~

~~DS~~

~~-~~

~~DS~~

~~PRO34(1)~~

LYS35(3)

ARG38(6)

~~GLU1(1)~~

LEU2(2)

CYS3(1)

ASP4(3)

ASP5(1)

ASP6(1)

~~-~~

LYS35(2)

ARG38(3)

LEU2(1)

-

THR41(2) THR41(1)

ASP4(1)

-

PHE42(5)

LYS43(2)

PHE44(2)

PHE42(2)

LYS43(1)

PHE44(1)

ASP6(1)

MET25(2) MET25(2)

TYR45(3) TYR45(2)

ASN27(2)

GLU29(1)

PHE34(3)

ARG35(3)

ARG36(5)

LYS38(4)

SER41(1)

LEU42(4)

TYR43(4)

ASN57(1)

-

ASN27(2)

GLU61(1)

GLU62(2)

-

GLU29(1)

GLU62(1)

-

PRO65(3) PRO65(1)

-

LEU66(1)

GLU68(4)

VAL69(2)

LEU72(3)

-

ARG36(1)

LYS38(1)

-

GLU68(2)

-

LEU72(2)

LEU42(6)

-

CYS105(1) CYS105(1)

GLU106(1) GLU106(1)

GLU62(1)

GLU113(1)

-

-

-

-

GLU107(2)

-

TYR119(1) TYR119(1)

HIS120(3)

HIS120(1)

Table 2.1 : Tableau de la comparaison des acides aminés retenues pour l’interac-

tion IL2/IL2Ralpha.

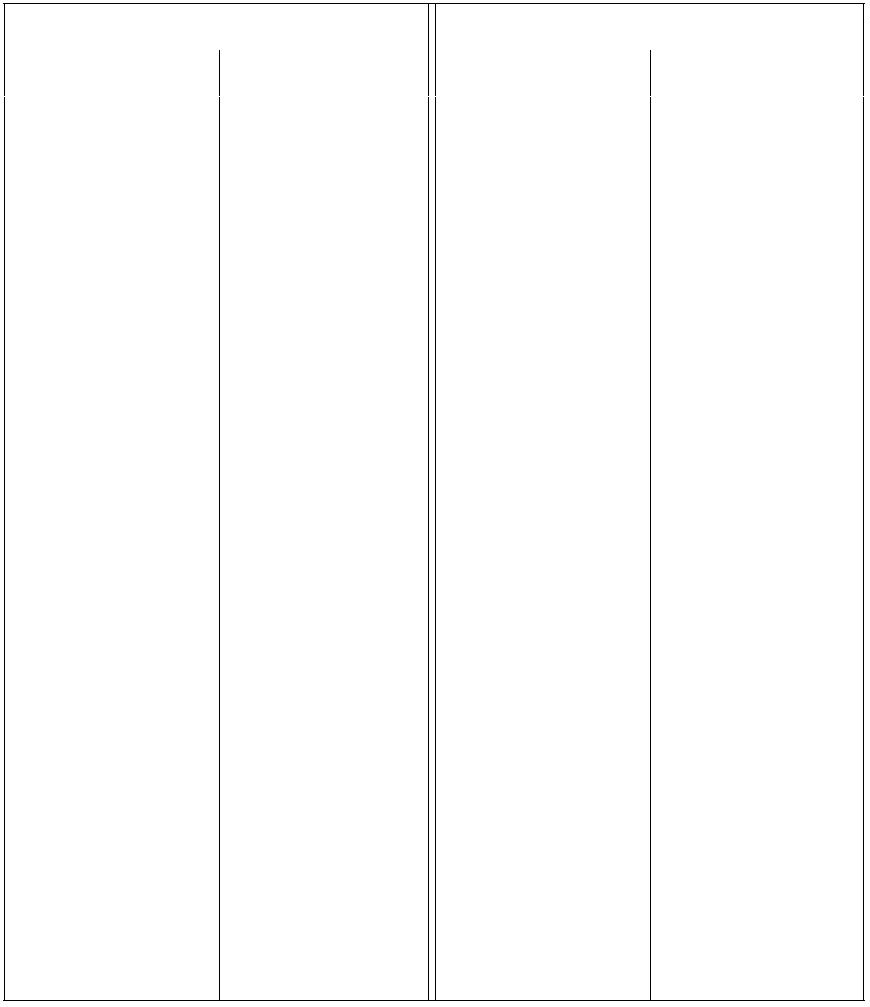
Ce tableau compare les acides aminés considérés comme importants pour l’interaction par

le script créé par rapport aux interactions proposées par Discovery Studio(DS). Les chiﬀres

entre parenthèses représentent le nombre d’interactions pour lesquelles l’acide aminé est im-

pliqué. Par rapport à DS, le programme mis en place surestime la quantité d’interactions et

le nombre d’acides aminés importants.

ooxWord://word/media/image4.jpeg

Chapitre 3

Génération de pseudo-molécules

L’objectif a été de transformer les chaînes latérales extraites en une molécule unique ca-

pable de mimer leur organisation spatiale et leurs interactions. Pour cela, les chaînes latérales

des résidus sélectionnés ont été reliées entre elles au moyen de linkers chimiques, aﬁn de former

ce que l’on nomme ici des pseudo-molécules. Ces entités ont vocation à reproduire l’empreinte

structurale de l’interaction protéique tout en restant compatibles avec les contraintes de la

chimie de synthèse.

3

.1 Construction par assemblage de fragments

La génération des pseudo-molécules a été réalisée à l’aide de la bibliothèque RDKit, un

outil open source spécialisé dans la chimie computationnelle. Chaque chaîne latérale d’acide

aminé a été représentée comme un fragment chimique autonome, et les coordonnées ato-

miques ont été conservées pour préserver la géométrie initiale.

3

.1.1 Intégration d’une banque de linkers : Enamine

Pour améliorer la pertinence chimique des liaisons, une banque de fragments moléculaires

(linkers) issue de la base de données Enamine a été utilisée. Ces fragments ont été importés

au format SDF(Standard Data Format) et intégrés dans le processus d’assemblage en tant

que blocs de liaison. Cette approche permet d’explorer un espace chimique plus réaliste et de

concevoir des structures potentiellement synthétisables.

Avant utilisation, un programme permet de ’nettoyer’ la banque de données. En eﬀet, dans

les banques Enamine, il existe des atomes parasites, ne faisant pas partie de la structure,

tels que des chlorures (Cl) ou bromures (Br). Ces atomes parasites sont identiﬁés, car ils sont

séparés de la structure par des points (Exemple : ’Cl.CCO’).

Ensuite, la banque est triée par ordre de grandeur en Ångström. La taille des fragments

est calculée en faisant la distance euclidienne entre le 1er atome et le dernier.

Note : certain fragment comporte des cycles, ce qui fausse la numérotation des atomes. Il

est donc possible que certains linkers ne se trouvent pas dans la bonne catégorie de taille.

6

CHAPITRE 3. GÉNÉRATION DE PSEUDO-MOLÉCULES

7

L’utilisateur peut aussi traiter sa propre banque de données et utiliser celle-ci pour la gé-

nération de ses pseudo-molécules.

L’inspiration de cette stratégie provient des approches de type PROTAC (Proteolysis

Targeting Chimeras), dans lesquelles deux unités fonctionnelles sont reliées par un linker op-

timisé pour l’interaction biologique. Une analogie peut être faite ici, avec comme diﬀérence

que les fragments initiaux ne sont pas des ligands existants, mais des motifs issus de résidus

d’interface protéique.

3

.1.2 RDKit et Openbabel

RdKit et Openbabel sont les deux outils principaux qui ont permis de mettre en place ces

liaisons. RDKit est au cœur de la gestion des molécules et permet la liaison de deux molécules

entre elles. Openbabel est un outil open source qui permet la génération de ﬁchiers sous

format PDB pour passer d’une molécule en 2D à une molécule en 3D, étape capitale pour la

suite des analyses et le docking.

3

.1.3 Contraintes et ajustements techniques

La génération d’une pseudo-molécule se fait en prenant 3 chaînes latérales qui se suivent

de manière aléatoire. La distance entre ces molécules est calculée aﬁn de sélectionner le lin-

ker parmi ceux ayant une taille appropriée pour que les deux molécules d’origine gardent une

distance relative similaire à leur position dans la protéine d’origine. Ce procédé est répété

plusieurs fois.

L’assemblage des fragments a présenté plusieurs déﬁs :

1

2

3

. La détermination automatique des points de liaison sur chaque fragment. Pour

cela, les atomes à valence libre sont repérés, c’est-à-dire, ceux qui sont reliés à au moins

un hydrogène. ce sont deux atomes à valence libre qui sont reliés entre eux.

. La génération correcte de la conformation 3D du composé ﬁnal. La fonctionnalité

intégrée de RDkit ne fonctionnant pas, c’est openbabel qui permet de générer des

molécules 3D.

. Le contrôle de la validité chimique des structures obtenues (absence de cycles

instables, saturation des valences).

Finalement, cette étape a permis de produire automatiquement des ensembles de pseudo-

molécules candidates, reproduisant les positions clés de l’interface PPI ciblée, tout en ouvrant

la voie à des évaluations in silico de leur potentiel d’inhibition.

Chapitre 4

Docking automatisé

Une fois les pseudo-molécules générées, l’étape suivante a consisté à évaluer leur capacité

à interagir avec la protéine cible. Cette évaluation a été réalisée à l’aide de techniques de

docking moléculaire, qui permettent d’estimer l’aﬃnité d’un ligand pour un site de liaison

donné sur une protéine. L’objectif était de comparer les pseudo-molécules générées à un

ligand de référence, aﬁn d’identiﬁer celles présentant les meilleurs scores d’interaction.

4

.1 Automatisation de la procédure de docking

L’environnement initial basé sur Jupyter Lab a été transformé en un script Python auto-

matisé, aﬁn de faciliter le traitement de grands ensembles de molécules. Ce script réalise les

étapes suivantes :

—

Préparation du ligand et du récepteur : Les charges de Gasteiger sont ajoutées.

Méthode d’ajout des charges partielles itérative qui ajuste les charges entre atomes

voisins en fonction de leur électronégativité.

—

—

Déﬁnition de la zone de docking : centrée sur les acides aminés identiﬁés comme

importants dans l’interaction.

Docking des pseudo-molécules : chaque molécule est soumise au programme de

docking, et le score d’énergie d’interaction est enregistré.

Ce processus a été optimisé pour fonctionner sur la plateforme GLiCID, ce qui a permis

d’exécuter les calculs en parallèle sur des ressources adaptées.

4

.2 Validation par comparaison avec un ligand de réfé-

rence

Pour évaluer la pertinence des pseudo-molécules générées, une molécule de référence

connue pour interagir avec la cible a été utilisée comme point de comparaison. Dans le cas du

complexe IL2/IL2Ralpha, le ligand SP4206 ou FRH (PDB ID : 1PY2) a servi de référence. Les

scores de docking obtenus avec les pseudo-molécules ont été comparés à celui de SP4206, et

les structures présentant un score supérieur ou équivalent ont été retenues comme candidates

prioritaires.

8

CHAPITRE 4. DOCKING AUTOMATISÉ

9

4

.2.1 Extraction automatique du ligand et du récepteur à partir du

PDB

Un programme permet l’extraction du ligand et du récepteur dans deux ﬁchiers PDB

séparés pour pouvoir être utilisés dans le programme de docking et ainsi servir de références.

Gestion des cas ambigus

Certains cas ont soulevé des diﬃcultés particulières, notamment lorsque les structures cris-

tallines représentaient des dimères sans interaction réelle, ou incluaient des molécules parasites

(e.g. NAG, HOH, FUC, ZN) faussant l’analyse. Un ﬁltrage basé sur la présence de résidus

biologiquement pertinents a été mis en place pour éliminer ces entités.

4

.3 Résultats et observations

Cette étape a permis d’identiﬁer plusieurs pseudo-molécules avec des scores d’interaction

prometteurs. Toutefois, le docking étant une approximation basée sur des modèles de scoring,

les résultats ont été interprétés avec prudence, et complétés par des analyses supplémentaires,

notamment une comparaison des structures chimiques (cf. section suivante).

SMILES

Energie de liaison

-7.0

CC(C)CC(Nc1ccc(CN2CCC(O)CC2)cc1

CCc1ccc(O)cc1)c1cc(NC(=O)OCC2c3ccccc3-

c3ccccc32)ccc1C(CCCO)C(=O)N(C)CC(=O)O

Cc1cc(CCC(=O)O)cc(CC(CCCNC(=N)N)C2CN

(Cc3ccccc3)CCC2NCCc2ccccc2)c1OCCCCNC(=N)N

N=C(N)NCCCCC1CN(Cc2ccccc2)CCC1NC

(CCCNC(=N)N)c1c(OCCc2ccccc2)ccc([C@H]

(NC(=O)OCC2c3ccccc3-c3ccccc32)C(F)(F)F)c1F

-

7.0

-6.9

Référence

CC(C)C[C@@H](NC(N)N)C(O)NCC(O)N1CCC(C2CC(C3CCC

(OCC4CCC(C(O)O)O4)C(Cl)C3Cl)NN2C)CC1

-

5.7

Table 4.1 : Meilleurs SMILES

Ce tableau présente les trois meilleurs SMILES des molécules avec l’énergie de liaison la

plus basse.

Cette procédure automatisée constitue une base robuste pour le criblage in silico de grandes

bibliothèques de composés, et peut être facilement enrichie par l’intégration d’autres moteurs

de docking ou de fonctions de scoring plus sophistiquées.



Chapitre 5

Vériﬁcation et comparaison

Les diﬀérentes étapes du pipeline — de l’analyse des interfaces PPI à l’optimisation par

apprentissage automatique — ont permis de générer un ensemble de pseudo-molécules candi-

dates et de les évaluer quantitativement et qualitativement. Cette section présente les étapes

de validation et d’évaluation des résultats obtenus.

5

.1 Analyse structurale et similarité chimique

Les molécules générées ont également été comparées à des ligands connus sur le plan

structurel :

—

Calcul de similarité Tanimoto entre empreintes moléculaires (ﬁngerprints), Plus le

score est proche de 1, plus les molécules sont similaires.

—

—

Analyse du respect de la géométrie initiale (positions relatives des groupes fonctionnels),

Facteur descriptif moléculaire comparé au ligand de référence :

•

•

•

LP(LogP) : Hydrophobicité, mesure de la solubilité dans les lipides par rapport à

l’eau. Plus la valeur est élevée, plus la molécule est lipophile.

MW (Molecular Weight) : Poids moléculaire en Daltons (g/mol). Une valeur trop

élevée (> 500) peut poser un problème pour l’absorption.

TPSA (Topological Polar Surface Area) : Surface polaire totale (en Å2). Liée

à la capacité de la molécule à faire des liaisons hydrogène. Important pour la

perméabilité cellulaire.

•

•

HBD : Nombre d’atomes donneurs de liaisons hydrogène (souvent des -OH ou

-

NH). Une valeur trop élevée amène à une mauvaise perméabilité.

HBA : Nombre d’atomes accepteurs de liaisons hydrogène (ex : O, N). Une valeur

trop élevée peut aussi diminuer la biodisponibilité.

—

—

Estimation de l’accessibilité synthétique (SAS) via des scores de complexité chi-

mique [[Ertl and Schuﬀenhauer](#br23) ([2009](#br23))]. Plus ce score est élevé, plus la molécule est

diﬃcile [à synthétiser.](#br23)

Estimation de la ressemblance à des médicaments(QED) par rapport à une base

de données de médicament.

Les molécules issues du pipeline présentaient en général un bon alignement des fonctions

chimiques principales, bien que certaines présentaient des défauts de saturation ou des confor-

mations non réalistes dues à la ﬂexibilité excessive des linkers.

1

0

CHAPITRE 5. VÉRIFICATION ET COMPARAISON

11

Figure 5.1 : Histogrammes descriptifs

Figure 5.2 : Régression linéaire

Ces deux ﬁgures descriptives nous permettent de situer le ligand de référence (barre verti-

cale rouge) par rapport aux pseudo-molécules générées. On y observe que le poids moléculaire

et la surface polaire sont relativement similaires. Par contre, le ligand de référence présente une

hydrophobicité largement inférieure et des atomes donneur/accepteur de liaisons hydrogènes

plus nombreux. Tous ces paramètres, sauf les atomes accepteurs, qui semblent être légèrement

corrélés avec l’énergie de liaison. Une optimisation de ces paramètres pourrait constituer une

piste d’amélioration.

(a)

(b)

Figure 5.3 : (a) QED-IL2 (b) SAS-IL2

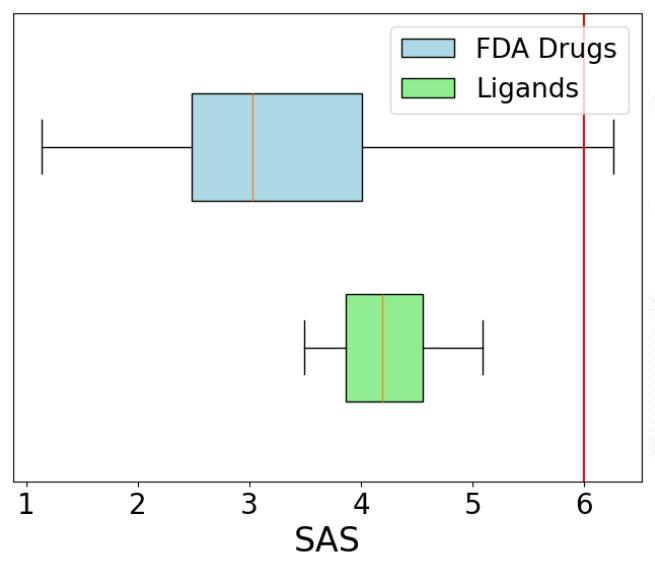
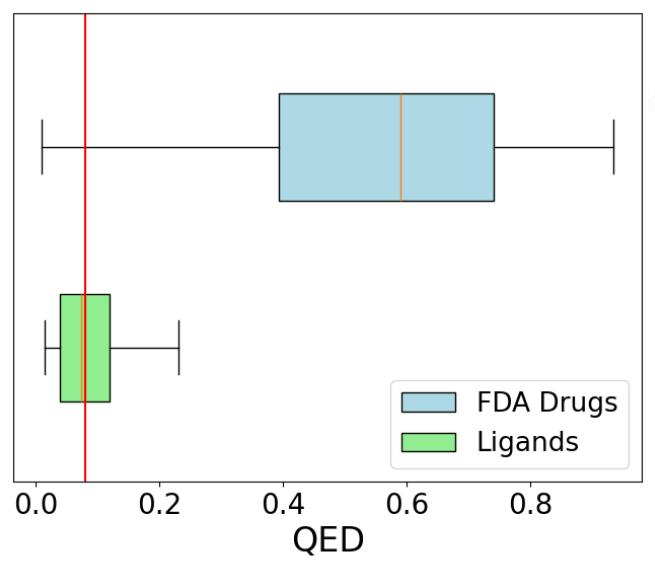
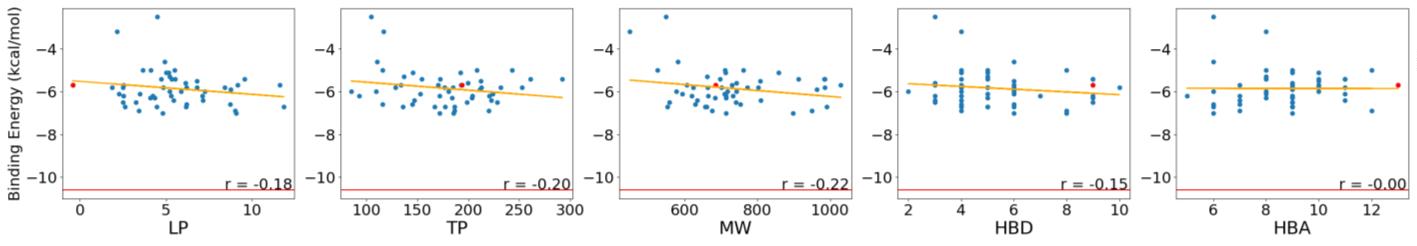
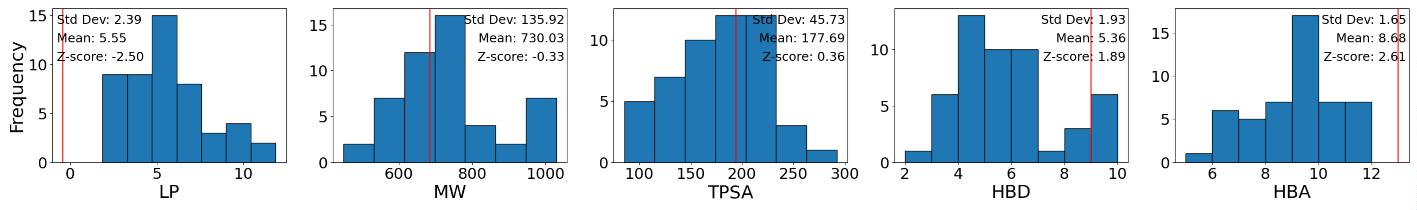
Le QED, qui nous donne la ressemblance à une base de données de molécules pharmaceu-

tiques, nous apprend que nos molécules ainsi que le ligand de référence sont très diﬀérentes

de la structure globale d’un médicament. Le SAS quant à lui nous apprend que nos pseudo-

molécules sont plus diﬃciles à synthétiser que les médicaments de référence, mais moins que

le ligand de référence.



CHAPITRE 5. VÉRIFICATION ET COMPARAISON

12

5

.2 Candidats potentiels

En prenant compte de la QED et du SAS ainsi que de l’énergie de liaison, nous obtenons

les molécules suivantes :

Figure 5.4 : Meilleurs candidats

On retrouve huit candidats potentiels sont identiﬁés parmi les molécules présentant une

haute aﬃnité dans la moitié supérieur du QED et la moitié inférieur pour le SAS. Les candidats

présentent un score de Tanimoto très faible, indiquant ainsi qu’aucun d’entre eux ne présente

vraiment de similarité avec le ligand d’origine. Il est donc diﬃcile de valider complètement

cette méthode.

5

.3 Amélioration de la banque de fragments

Ces résultats sont très dépendants de la banque de fragments utilisés pour relier les chaînes

latérales entre elles. Ainsi, un changement de cette base de données pourrait faire varier les

résultats, il serait donc intéressant d’explorer d’autres banques de fragments, et éventuellement

les combiner pour obtenir de meilleurs résultats. Il serait aussi possible d’améliorer la sélection

de ces fragments pour qu’ils soient plus adaptés au récepteur. Pour cela, on peut faire deux

choses :

—

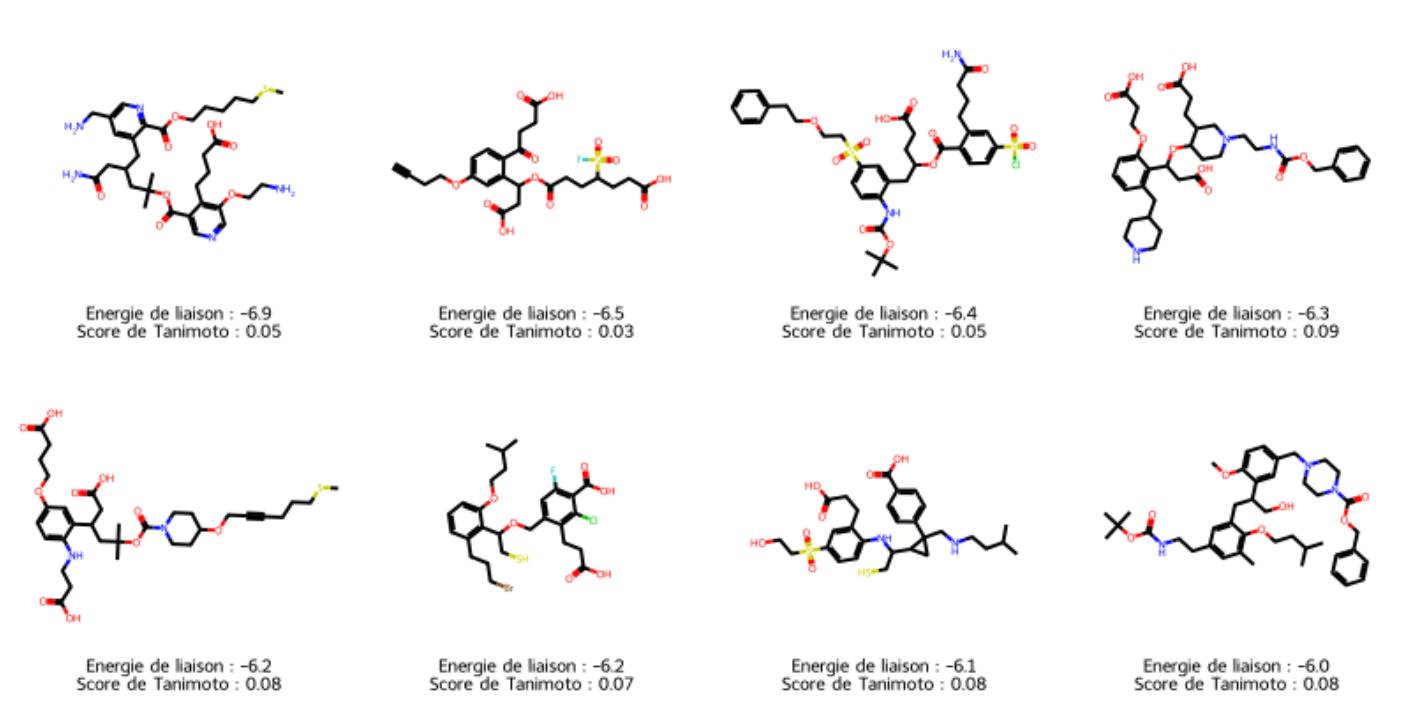
Améliorer la classiﬁcation des banques de données en indiquant les propriétés physico-

chimiques de la molécule, autre que la taille.

—

Utilisation de machine learning pour optimiser la sélection du ligand en fonction du

score de docking.



Chapitre 6

Limites, perspectives et conclusion

Ce stage a permis de développer et d’explorer une méthode innovante de conception de

pseudo-ligands à partir d’interfaces protéine-protéine, en combinant des approches d’analyse

structurale, de génération moléculaire et d’intelligence artiﬁcielle. Bien que les résultats obte-

nus soient prometteurs, plusieurs limites ont été identiﬁées tout au long du projet.

6

.1 Limites

6

.1.1 Obstacles rencontrés

La première limite est d’ordre technique. Bien que prometteur, l’intégration de certains ou-

tils issus de la communauté open source (tels que DeLinker [[Imrie et al. (2020)],](#br23) Protac-invent

ou Reinvent) s’est révélée impossible, en raison de la [nécessité de conﬁgurations](#br23) matérielles

spéciﬁques (GPU, bibliothèques incompatibles, dépendances obsolètes ou restreintes par des

licences).

Ensuite, les performances du modèle d’autoencodeur ont été aﬀectées par la taille du jeu

de données d’entraînement et par les limitations en mémoire. La génération de SMILES valides

n’a pas toujours été assurée, et les variations dans l’espace latent conduisaient parfois à des

structures chimiquement peu plausibles.

6

.1.2 Limite de la méthode utilisé

Enﬁn, le choix des résidus d’interaction, bien qu’aﬃné par des critères physicochimiques,

reste partiellement dépendant de seuils arbitraires (distance, hydrophobicité. . .), et pourrait

bénéﬁcier d’une intégration de méthodes de détéction de hotspot plus robustes, comme celles

basées sur la méthode Fragment Molecular Orbital (FMO) [[Monteleone et al. (2022](#br23))] ou

des approches d’apprentissage supervisé. Note : Un hotspot [(ou point chaud en français)](#br23)

ensemble des résidus essentiels à l’interaction.

6

.2 Perspectives d’amélioration

Plusieurs pistes d’amélioration ont été identiﬁées :

1

3

CHAPITRE 6. LIMITES, PERSPECTIVES ET CONCLUSION

14

1

. Enrichissement de la bibliothèque de fragments (linkers), par l’ajout de bases de données

plus larges (ZINC, ChEMBL fragments), et une meilleure catégorisation des liaisons

chimiques autorisées.

2

3

4

. Identiﬁcation automatique des hotspots via des modèles pré-entraînés ou des méthodes

de mécanique quantique, pour améliorer la sélection initiale des résidus.

. Optimisation des modèles de génération moléculaire, notamment par l’usage de trans-

formeurs moléculaires (e.g. SyntaLinker) ou de générateurs conditionnels.

. Intégration de critères multi-objectifs, comme la toxicité, la solubilité ou la biodisponi-

bilité, pour rendre le pipeline plus réaliste en contexte pharmaceutique.

6

.3 Avantages et inconvénient de la méthodes

La création de petites molécules inhibitrices se présente comme une alternative à l’utilisa-

tion d’anticorps, qui est un moyen coûteux avec peu de biodisponibilité oral et avec un poids

moléculaire élevé. En eﬀet, comme décrit dans la revue At The Interface : Small-Molecule

Inhibitors of Soluble Cytokines [[Raavi et al. (2025](#br23))], les petites molécules inhibitrices sont

un sujet de recherche actif [dans le cas des](#br23) c[ytokines.](#br23) Elles ont pour avantages d’avoir un

poids moléculaire bas, une bonne biodisponibilité et une absorption rapide, pouvant oﬀrir une

alternative thérapeutique avec plus de contrôle.

Cependant, les interface protéine-protéine se présente comme des surfaces plate, large et

rigide plutôt que des cavités dans lesquelles se logent les ligands. Ainsi, le manque de hotspot

spéciﬁque de haute aﬃnité rend diﬃcile la conception de petits d’inhibiteurs sans risque qu’ils

interfèrent d’autres interactions.

6

.4 Conclusion générale

Ce travail a permis de concevoir et d’implémenter un pipeline de génération de pseudo-

ligands ciblant les interfaces protéine-protéine (iPP), en combinant des outils de chimio-

informatique, de génération moléculaire, et de criblage virtuel. À partir de triplets de résidus

sélectionnés sur une interface cible, le pipeline génère des fragments connectés chimiquement

plausibles, les ﬁltre selon des critères de "drug-likeness" (QED, SAS), puis les évalue via des

calculs de docking.

L’approche a montré sa capacité à produire des candidats moléculaires respectant les

contraintes physico-chimiques classiques et présentant des scores de docking comparables,

voire supérieurs, à ceux d’un ligand de référence. Bien que certaines étapes puissent être af-

ﬁnées, notamment le choix des linkers ou l’exploration plus systématique de l’espace latent,

les résultats obtenus sont encourageants pour une première version du pipeline.

À terme, ce type d’approche pourrait être étendu à d’autres interfaces iPP, ou couplé à

des modèles d’apprentissage automatique pour optimiser la sélection des fragments et des

linkers. Le pipeline développé constitue ainsi une base ﬂexible pour l’exploration rationalisée

d’inhibiteurs d’interfaces protéiques.

Bibliographie

Ertl, P. and Schuﬀenhauer, A. (2009), ‘Estimation of synthetic accessibility score of drug-like

molecules based on molecular complexity and fragment contributions’, Journal of Chemin-

formatics .

URL: https ://doi.org/10.1186/1758-2946-1-8

Imrie, F., Bradley, A. R., van der Schaar, M. and Deane, C. M. (2020), ‘Deep generative

models for 3d linker design’, Journal of Chemical Information and Modeling .

URL: https ://doi.org/10.1021/acs.jcim.9b01120

Labbé, C. M., Kuenemann, M. A., Zarzycka, B., Vriend, G., Nicolaes, G. A., Lagorce, D.,

Miteva, M. A., Villoutreix, B. O. and Sperandio, O. (2015), ‘ippi-db : an online database of

modulators of protein–protein interactions’, Nucleic Acids Research 44(D1), D542–D547.

URL: https ://doi.org/10.1093/nar/gkv982

Monteleone, S., Fedorov, D. G., Townsend-Nicholson, A., Southey, M., Bodkin, M. and Hei-

fetz, A. (2022), ‘Hotspot identiﬁcation and drug design of protein–protein interaction mo-

dulators using the fragment molecular orbital method’, Journal of Chemical Information

and Modeling 62(16), 3784–3799. PMID : 35939049.

URL: https ://doi.org/10.1021/acs.jcim.2c00457

Raavi, Koehler, A. N. and Vegas, A. J. (2025), ‘At the interface : Small-molecule inhibitors

of soluble cytokines’, Chemical Reviews 125(9), 4528–4568. PMID : 40233276.

URL: https ://doi.org/10.1021/acs.chemrev.4c00469

1

5

Annexe A

An Appendix

Some lengthy tables, codes, raw data, length proofs, etc. which are very important but

not essential part of the project report goes into an Appendix. An appendix is something a

reader would consult if he/she needs extra information and a more comprehensive understating

of the report. Also, note that you should use one appendix for one idea.

An appendix is optional. If you feel you do not need to include an appendix in your report,

avoid including it. Sometime including irrelevant and unnecessary materials in the Appendices

may unreasonably increase the total number of pages in your report and distract the reader.

1

6

Annexe B

An Appendix Chapter (Optional)

.

..

1

7