

EFREI Paris

Master en Bio-Informatique



Rapport Stage M2

**Pipeline pour l'amélioration
d'AlphaFold2**

I2BC

Nom Étudiant

Julie DANIEL

Numéro étudiant

20200365

Maître de Stage

Diego Zea



Date du stage : du 03/02/2025 au 31/07/2025

1 Remerciement	3
2 Introduction	4
3 Contexte du stage	5
3.1 Présentation de l'entreprise	5
3.2 Présentation de l'équipe	6
3.3 Présentation du projet	7
3.3.1 Contexte biologique	7
3.3.2 Le projet SPPICES	9
3.4 Environnement de travail	11
4 Méthodologie	12
4.1 Déroulement de la tâche 1A	12
4.2 Outils et technologies utilisés	12
4.2.1 AlphaFold2	13
4.2.2 Foldseek	15
4.3 Pipeline AlphaConformer	17
4.3.1 Présentation du fonctionnement	17
4.3.2 Résultats initiaux	20
5 Résultats	22
5.1 Evaluation de la pipeline	22
5.1.1 AF-Cluster	22
5.1.2 BioEmu	25
5.1.3 Comparaison avec AlphaConformer	27
5.2 Amélioration d'AlphaConformer	28
5.3 Construction de la base de données	30
5.3.1 Collecte des données	30
5.3.2 Filtrage des données	32
6 Retour d'expérience	35
6.1 Difficultés rencontrées et points clés	35
6.2 Aspects organisationnels et gestion de projet	36
7 Conclusion	38
8 Annexes	41
9 References	45

1 Remerciement

Je tiens tout d'abord à exprimer ma profonde gratitude à Diego Zea, mon encadrant de stage, pour sa disponibilité, ses conseils précieux et la confiance qu'il m'a accordée tout au long de ces six mois. Son accompagnement bienveillant et stimulant a grandement contribué à l'enrichissement de cette expérience.

Je remercie également Françoise Ochsenbein et Raphaël Guerois pour m'avoir accueilli au sein de leur équipe à l'I2BC, ainsi que pour les échanges scientifiques constructifs et l'ambiance de travail motivante.

Et je souhaite également adresser un merci particulier à l'ensemble des membres de l'équipe Guerois/Ochsenbein, pour leur accueil chaleureux, leur aide au quotidien, et les nombreuses discussions qui m'ont permis d'apprendre et de progresser.



Photo d'équipe du 30/06/2025

2 Introduction

Depuis toujours, la science cherche à comprendre le fonctionnement du corps humain comme celui du vivant. Au cœur de ces mécanismes se trouvent les protéines, des molécules quasi-maintenues indispensables à l'ensemble des processus biologiques.

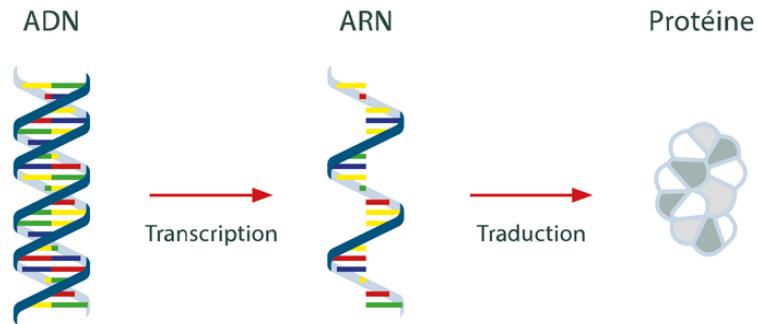


Figure 1: Schéma illustrant la synthèse d'une protéine. L'ADN est d'abord transcrit en ARN messager, qui est ensuite traduit en protéine.

Les protéines sont issues d'un processus biologique complexe qui commence par la transcription de l'ADN en ARN messager, ensuite celui-ci est traduit en une chaîne d'acides aminés créant la protéine. Une fois synthétisée, elle se replie spontanément en quelques millisecondes pour donner une forme spécifique, tridimensionnelle (voir **Figure 1**).

Ce repliement dépend d'interactions diverses à l'échelle atomique telles que les forces de Van der Waals, les liaisons hydrogènes, les interactions électrostatiques ou les interactions hydrophobes [1]. Une même séquence protéique peut conduire à diverses conformations qui peuvent être associées à différentes fonctions dans l'organisme [2].

C'est pour cela que connaître la structure 3D des protéines est un critère fondamental pour comprendre les mécanismes cellulaires et concevoir de nouveaux médicaments. Aujourd'hui, la détermination de la structure des protéines est réalisée soit par des méthodes expérimentales, soit par des méthodes de prédiction informatique.

Les méthodes de prédiction informatiques se sont particulièrement développées depuis 2012, et ont constitué une avancée majeure en bio-informatique en contribuant à la révolution de la biologie structurale. Nous avons l'exemple de l'algorithme de prédiction AlphaFold, développé par DeepMind, qui utilise un réseau de neurones afin d'obtenir une prédiction de la structure 3D des protéines [16]. Il arrive à obtenir une précision équivalente à celle des méthodes expérimentales, tout en étant moins coûteux et beaucoup plus rapide. Cependant, AlphaFold a quelques limites, il ne prédit qu'une forme de la protéine alors que celle-ci peut

exister sous plusieurs conformations [12]. Plusieurs études ont révélé qu'il avait tendance à prédire majoritairement la forme active des protéines [13]. La connaissance de la forme inactive d'une protéine est cependant une donnée cruciale, qui pourrait être considérée comme la forme, ou configuration, par "défaut" de la protéine avant d'interagir avec d'autres molécules. La détermination de cette structure permet d'analyser le comportement d'une protéine sans l'influence d'un ligand ou d'un cofacteur par exemple [2].

L'objectif de mon stage est donc d'améliorer les prédictions d'AlphaFold afin d'obtenir à la fois la forme inactive, appelée apo, et la forme active, appelée holo, d'une protéine. Connaître les deux conformations des protéines permettrait une meilleure compréhension de leur rôle et de leurs interactions dans l'organisme.

3 Contexte du stage

3.1 Présentation de l'entreprise

J'ai effectué mon stage de fin d'étude pour 26 semaines au sein d'une équipe de I2BC. L'institut de Biologie Intégrative de la Cellule (I2BC) est une Unité Mixte de Recherche (UMR) placée sous la tutelle conjointe du Commissariat à l'Énergie Atomique et aux Énergies Alternatives (CEA), du Centre National de la Recherche Scientifique (CNRS) et de l'Université Paris-Saclay.

Elle regroupe environ 60 équipes de recherche réunissant plus de 600 personnes, parmi lesquelles on trouve des chercheurs, des ingénieurs, des techniciens et des doctorants. Les recherches menées par I2BC visent à comprendre le fonctionnement de la cellule et notamment des processus moléculaires qui pilotent son architecture et ses propriétés afin de prédire les conséquences qu'entraînent les changements moléculaires dans les cellules normales ou pathologiques. [6]

L'I2BC fait partie du CNRS, qui est le principal organisme public de recherche en France, créé à la fin de l'année 1939. Le CNRS couvre de nombreux domaines scientifiques, notamment la physique, la biologie, ainsi que les sciences humaines et sociales. [5] Bien que placé sous la supervision du CNRS, les équipes de l'I2BC sont physiquement réparties sur différents sites. Notre équipe était implantée au sein du CEA, un organisme public de recherche fondé en 1945. Si le CEA était initialement centré sur les applications de l'énergie nucléaire, il a depuis diversifié ses activités vers le développement des énergies renouvelables, ainsi que vers la recherche dans les technologies de l'information et de la santé [4].

Étant implanté au sein d'un centre nucléaire, le site de l'I2BC fait l'objet de mesures de sécurité très strictes. L'accès au réseau informatique, aux infrastructures et aux laboratoires est très fortement protégé. Des travaux sont en cours afin de déménager et de regrouper toutes les équipes de l'I2BC sur le centre du CNRS à Gif-sur-Yvette, ce qui facilitera les relations scientifiques et les collaborations entre chercheurs.

3.2 Présentation de l'équipe

J'ai effectué mon stage au sein de l'équipe Ochsenbein/Guerois de l'I2BC, spécialisée en biologie structurale, et plus particulièrement dans l'étude des mécanismes d'assemblage et d'interaction dynamique entre certaines protéines [10]. L'équipe est organisée en deux pôles complémentaires : un pôle biologie expérimentale, encadré par Françoise Ochsenbein, et un pôle bioinformatique, dirigé par Raphaël Guerois. Ces deux pôles collaborent étroitement. Pour assurer un bon suivi et une cohésion d'équipe, des réunions sont organisées chaque semaine en alternance : une semaine en assemblée générale avec l'ensemble des membres, et la semaine suivante en sous-groupes, selon les spécialités.

Ces réunions sont l'occasion pour chacun de présenter l'avancée de son projet, d'échanger des idées, de recevoir des conseils et d'ajuster les orientations si nécessaire. Nous y intégrons également des séances de lecture d'articles scientifiques, permettant de rester informés des avancées récentes dans le domaine et d'identifier des méthodes ou approches innovantes que nous pourrions adapter à nos propres travaux.

Dans le milieu de la recherche, la lecture d'articles scientifiques constitue une étape fondamentale. Elle permet de s'informer sur les travaux déjà réalisés, d'identifier de nouvelles pistes de réflexion et d'orienter plus efficacement son propre projet. Toutes les hypothèses et stratégies d'analyse reposent en grande partie sur des études antérieures menées par d'autres équipes à travers le monde. Il est donc crucial d'y accorder une attention particulière afin de se démarquer et d'optimiser son travail.

On comprend ainsi rapidement l'importance de développer un esprit critique dans l'analyse des publications scientifiques. Les données peuvent être interprétées de manière biaisée, et il est possible d'utiliser certaines statistiques pour soutenir une hypothèse préétablie. Une évaluation rigoureuse des méthodes utilisées et des résultats obtenus est donc indispensable avant d'intégrer ces connaissances dans son propre projet de recherche.

Au quotidien, j'étais supervisé par mon maître de stage, Diego Zea. Nous avions des réunions hebdomadaires permettant de faire le point sur l'avancement du projet, de planifier les prochaines étapes, de définir les tests à réaliser et de résoudre ensemble d'éventuels blocages. Nous réalisions également des lectures d'articles scientifiques en lien avec notre sujet, suivies de discussions critiques afin d'évaluer leur impact potentiel sur notre projet.

Le projet, lancé par Diego Zea en 2023, a été retenu en 2024 dans le cadre d'un financement ANR *Agence Nationale de la Recherche* et s'étale sur une durée de quatre ans et demi à partir de 2025. Il se divise en plusieurs étapes. Au cours de mon stage, j'étais en charge de la phase 1A, tandis que Diego poursuivait en parallèle le développement d'une autre partie du projet.

3.3 Présentation du projet

Mon stage se déroule dans le cadre du projet de recherche SPPICES, *Scoring and Predicting Protein Interaction and Conformations based on Evolutionary Signals*, développé par mon maître de stage Diego Zea. Son principal objectif est de mieux comprendre les modes d'interaction entre les protéines ainsi que leurs dynamiques au sein de l'organisme.

3.3.1 Contexte biologique

Les protéines sont des molécules essentielles à la vie. Elles participent à presque tous les processus biologiques dans le corps humain : elles peuvent transporter des substances, réguler des réactions chimiques, défendre l'organisme, etc. Elles sont composées de chaînes d'acides aminés qui ont la capacité naturelle de se replier dans l'espace en structures complexes, leur fonction, et donc leur activité, sont directement liés à leur forme. Ce repliement peut prendre différentes formes et l'agencement précis de ces structures détermine les propriétés et les fonctions de la protéine. Une mauvaise forme peut entraîner une perte de fonction, voire causer des maladies.

À ce jour, plusieurs méthodes existent pour explorer la dynamique des protéines, mais elles présentent toutes des limites importantes. Les techniques de type Single Cell permettent d'observer des distributions d'états à l'équilibre, mais elles nécessitent des protocoles expérimentaux complexes et des temps de collecte de données très longs. La cryo-microscopie électronique peut révéler plusieurs états conformationnels et leurs probabilités d'apparition, mais elle est coûteuse et peu accessible. Les simulations de dynamique moléculaire offrent une exploration détaillée et précise des mouvements atomiques, mais leur coût computationnel est considérable et elles souffrent encore de limitations liées à l'approximation des forces intermoléculaires. Ces obstacles rendent difficile une analyse systématique et rapide de la flexibilité structurale des protéines à l'échelle du génome.

Ainsi prédire la forme qu'une protéine va adopter à partir de l'information génétique est un défi majeur en biologie. Ce n'est pas simple, car une même séquence d'acides aminés peut adopter différentes conformations dynamiques, souvent liées à des modes d'actions distincts. Ainsi, il ne suffit pas de connaître la séquence pour prédire de manière fiable la forme et la fonction d'une protéine. C'est dans ce contexte qu'AlphaFold a marqué un tournant

dans la prédition des structures protéiques : son approche repose non seulement sur la séquence des acides aminés, mais aussi sur les signaux évolutifs. En comparant les protéines homologues, présentes chez différentes espèces mais issues d'un ancêtre commun, il devient possible d'identifier les régions conservées au fil du temps. Ces régions conservées jouent un rôle clé car elles révèlent des contacts entre résidu, essentiels pour déterminer la structure d'une protéine. Ainsi, en exploitant ces signaux évolutifs, il devient possible de mieux prédire la structure et les interactions des protéines.

Cependant, AlphaFold ne prédit généralement qu'une seule conformation par protéine, et tend à favoriser la forme active, dite holo, correspondant à une protéine liée à un ligand (voir **Figure 2**). Or, la connaissance de la forme inactive, dite apo, est tout aussi essentielle. Cette conformation apo peut être considérée comme l'état de référence de la protéine, c'est-à-dire sa configuration en l'absence d'interaction avec d'autres molécules. Déterminer cette structure est crucial pour comprendre le comportement intrinsèque de la protéine, sans l'influence de ligands, de cofacteurs ou d'autres partenaires moléculaires. Par exemple, dans le cadre du développement de médicaments, il est parfois souhaitable de concevoir une molécule capable de stabiliser la forme apo d'une protéine afin d'en inhiber la fonction.

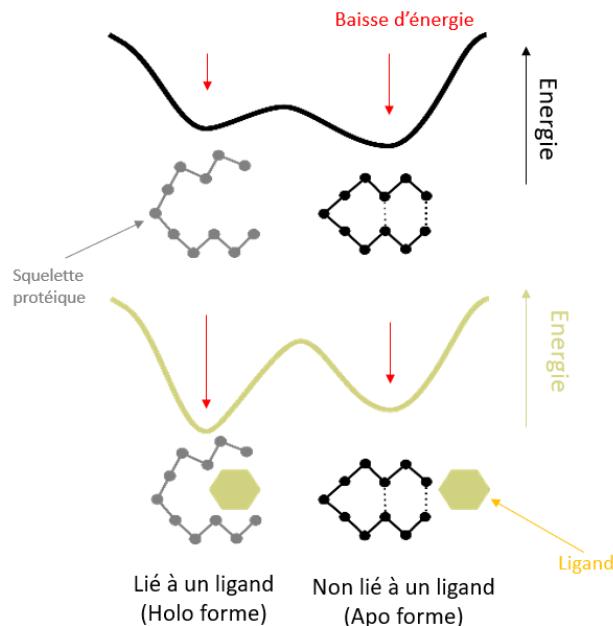


Figure 2: Schéma illustrant la diversité conformationnelle des protéines : une même protéine peut adopter différentes structures tridimensionnelles, représentées ici par son squelette protéique. La forme ouverte permet la liaison avec un ligand, contrairement à la forme fermée. Ces deux conformations présentent des niveaux d'énergie différents : la conformation apo correspond généralement à l'état de plus basse énergie de la protéine seule, tandis que la conformation holo représente l'état de plus basse énergie du complexe protéine-ligand.

3.3.2 Le projet SPPICES

SPPICES est un projet de recherche qui vise à approfondir notre compréhension de la dynamique et des interactions des protéines. Le projet va ainsi se baser sur les connaissances évolutives, sur les avancées révolutionnaires d'AlphaFold2 et les récentes approches d'apprentissage profond. SPPICES a pour objectif de surmonter les limites actuelles en intégrant les signaux évolutifs à la bioinformatique structurale afin de mettre en lumière les comportements complexes essentiels à la fonction des protéines.

Les protéines adoptent des comportements structuraux complexes qui sont essentiels à leur bon fonctionnement. Comprendre et prédire ces comportements est un enjeu fondamental en biologie structurale. En raison de son fort potentiel scientifique, SPPICES a été sélectionné pour un financement de 360 000 € sur 54 mois par l'Agence nationale de la recherche (ANR).

Le projet SPPICES se divise en trois grands axes : la diversité conformationnelle des protéines, les protéines intrinsèquement désordonnées et les interactions entre protéines. Mon stage s'inscrit dans le premier axe, centré sur la diversité conformationnelle, et plus précisément dans la tâche 1A. Cette tâche porte sur l'utilisation de modèles structuraux pour générer différentes conformations d'une même protéine à l'aide d'AlphaFold2.

AlphaFold2 tend à prédire une seule conformation "dominante" pour une protéine.(voir **Figure 3**) Or, dans la réalité biologique, certaines protéines peuvent exister sous plusieurs conformations, forme apo, non liée à un ligand, ou forme holo, liée à un ligand.

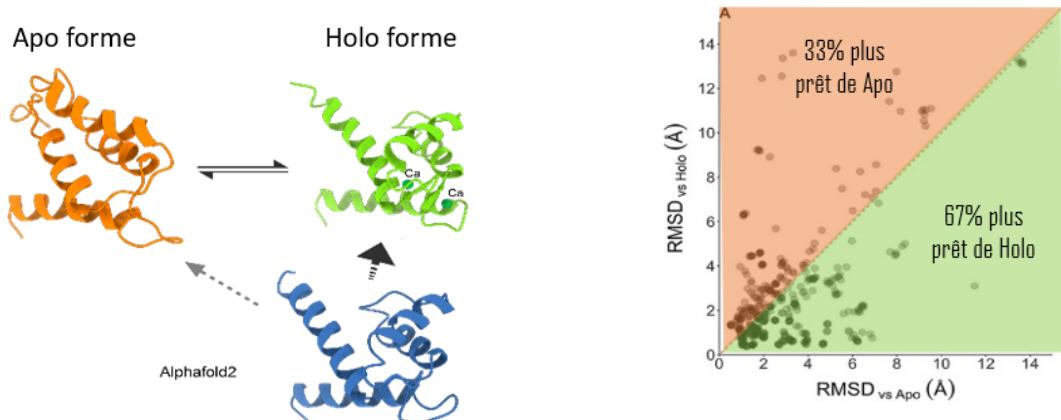


Figure 3: Évaluation des performances d'AlphaFold2 : le modèle a tendance à favoriser la conformation holo, par rapport à la forme apo. En lançant AlphaFold2 plusieurs fois sur la même protéine, dans 67% des cas, la conformation prédictive était plus proche de la structure holo, contre 33% des cas où elle se rapprochait davantage de la forme apo.

La tâche 1A explore une nouvelle approche qui consiste à guider AlphaFold2 vers une structure alternative. Plusieurs équipes de recherche travaillent déjà sur cette problématique et ont développé des outils tels que AF-Cluster[30] ou BioEmu[32]. Cependant, Diego Zea souhaite se démarquer de ces approches et obtenir de meilleurs résultats. Pour cela, il s'est inspiré des réflexions présentées dans le chapitre "Easy Not Easy: Comparative Modeling with High-Sequence Identity Templates" [11], ainsi que de l'hypothèse ConTemplate [12], qui propose que si deux protéines partagent une conformation, elles pourraient en partager d'autres. À partir de cette idée, il a développé un nouveau pipeline nommé AlphaConformer, qui repose sur l'utilisation de modèles structuraux, appelés templates, pour influencer directement la prédiction d'AlphaFold.

Diego Zea a initié ce travail en mai 2023. Les premiers résultats sont prometteurs : dans 4 cas sur 10, cette stratégie a permis d'éviter le biais d'AlphaFold2 en faveur d'une seule conformation. Par exemple, dans le cas de l'Uracile-ADN glycosylase, *protéine avec l'identifiant Uniprot 1AKZ*, le pipeline a permis de modéliser correctement les deux états connus (voir **Figure 4**). Le pipeline AlphaConformer arrive à prédire une conformation proche de holo, avec une RMSD de 1,2 Ångströms, et une conformation proche de apo, avec une RMSD de 1,1 Ångströms.

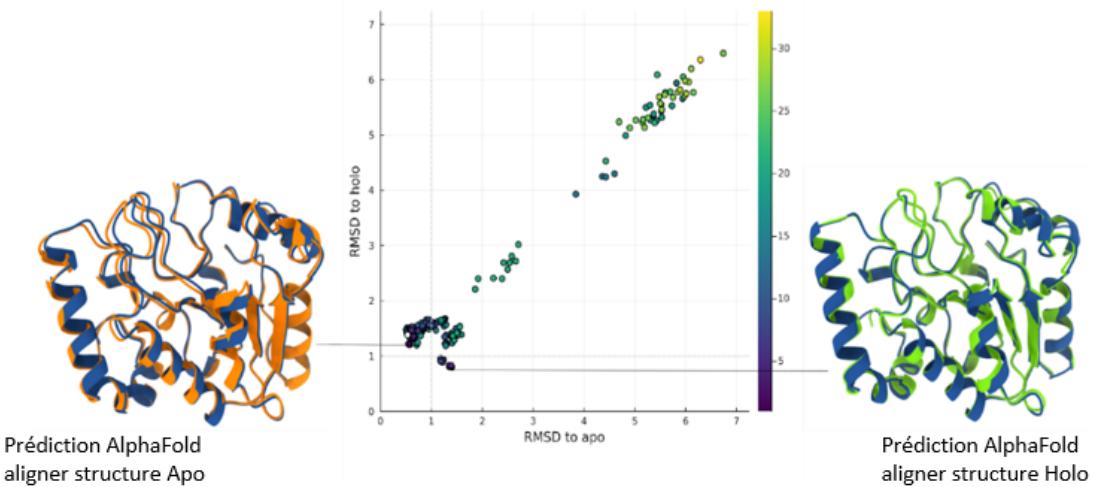


Figure 4: Premiers résultats obtenus avec le pipeline AlphaConformer. Chaque point représente une prédiction générée par AlphaFold2. Nous observons des structures proches à la fois de la conformation holo et apo, ce qui est encourageant et constitue l'objectif visé pour l'ensemble des protéines testées. Une prédiction est considérée comme excellente si la RMSD (Root Mean Square Deviation) par rapport à la structure de référence (apo ou holo) est inférieure à 1 Å. La RMSD mesure la distance moyenne entre les carbones alpha de la chaîne principale des protéines : plus cette valeur est faible, plus les structures sont similaires.

L'objectif principal de mon stage est d'améliorer le pipeline AlphaConformer afin qu'il puisse prédire efficacement plusieurs conformations pour un plus grand nombre de protéines. Pour cela, une première étape essentielle consistait à comparer les premiers résultats du pipeline AlphaConformer avec ceux d'autres approches existantes, telles qu'AlphaCluster ou BioEmu. Ensuite, une grande partie de mon travail s'est concentrée sur l'optimisation de notre pipeline : ajustement des paramètres, choix des meilleures stratégies d'alignement, amélioration des étapes de sélection des modèles, etc. Enfin, j'ai construit un jeu de données cohérent, bien structuré et représentatif, qui servira à l'entraînement final du pipeline.

3.4 Environnement de travail

Pour mener à bien ce projet de recherche, nous disposons d'un environnement riche en ressources biologiques, informatiques et logicielles.

Nous exploitons dans notre domaine un large éventail de données biologiques disponibles pour le public, comprenant en particulier des séquences de protéines, leurs structures tridimensionnelles, expérimentales ou prédictives, leurs classifications en familles et diverses annotations fonctionnelles et évolutives dans des bases de données reconnues telles que UniProt, PDB, AlphaFold DB, ou SIFT. Dès lors, nous avons une vision intégrée des différents aspects de la structure, de la fonction et de l'évolution des protéines.

D'un point de vue technique, le développement du pipeline de prédiction s'appuie en grande partie sur le langage de programmation Julia, un langage de hautes performances particulièrement adapté au calcul scientifique, qui permet une syntaxe astucieusement proche de celle du langage Python mais dont la force réside dans une rapidité d'exécution et dans la possibilité d'effectuer des manipulations mathématiques ou sur des données biologiques à l'aide d'outils dédiés tels que BioStructures.jl[14] ou encore MIToS.jl [15]. Pour ce qui est de l'exécution des calculs et des prédictions structurelles, celle-ci est effectuée sur un cluster de calcul à distance, géré par un gestionnaire de tâches SLURM, qui fournit à tous les membres de l'équipe l'accès à un environnement de travail de hautes performances. Dans le cadre de mon stage, j'ai travaillé sur le nœud 48, équipé de plus de 10 To de stockage, de 80 coeurs CPU pour le calcul parallèle, et de 3 GPU, des éléments essentiels pour l'exécution d'outils utilisant des technologies d'apprentissage profond, comme AlphaFold2.

Nous avons installé et configuré AlphaFold2 sur cette infrastructure, ce qui nous permet de générer des prédictions de structures protéiques en haute qualité. En complément, nous utilisons également Foldseek, un outil performant pour la comparaison rapide de structures protéiques, dont le fonctionnement sera détaillé dans une section ultérieure du rapport.

Cet environnement technique robuste nous permet de traiter efficacement de grands volumes de données et de mener des expérimentations complexes dans des délais raisonnables.

4 Méthodologie

4.1 Déroulement de la tâche 1A

L'objectif principal de ce projet, et plus spécifiquement de la tâche 1A, est d'améliorer les capacités de prédiction d'AlphaFold2 afin qu'il puisse modéliser non seulement la conformation holo, forme liée à un ligand mais aussi la conformation apo, forme sans ligand, d'une même protéine. Pour cela, Diego Zea a initié le développement d'un pipeline en langage Julia, dont le rôle est de préparer les données d'entrée de manière à guider AlphaFold2 vers la prédiction de ces deux états conformationnels distincts. Durant mes six mois de stage, j'ai ainsi contribué à plusieurs étapes clés de la tâche 1A, avec pour objectif d'optimiser les performances globales du pipeline.

La première étape consistait à comparer AlphaConformer avec d'autres méthodes publiées qui tentent également de résoudre le problème de la prédiction multi-conformationnelle. Cette phase a impliqué un travail de veille scientifique important, comprenant la lecture et l'analyse critique d'articles récents. Elle a permis d'évaluer la pertinence et l'originalité du pipeline par rapport aux méthodes existantes, et ainsi de mesurer sa valeur ajoutée.

Ensuite, la deuxième étape a porté sur l'amélioration des performances du pipeline AlphaConformer, en l'appliquant au jeu de données présenté par Saldaño et al. (2022)[13]. L'objectif était d'identifier les limites du pipeline existant, puis de proposer des améliorations. Cette phase a demandé un important travail de compréhension du code déjà en place, ainsi qu'une analyse approfondie des techniques utilisées. Reprendre un code existant peut s'avérer plus complexe que de partir de zéro, notamment lorsqu'il faut en modifier la logique tout en conservant sa stabilité.

Enfin, la troisième étape était de créer une base de données structurée. Ce travail a nécessité une attention particulière et une réflexion approfondie. Nous nous sommes appuyés sur la littérature scientifique récente pour définir des critères de sélection pertinents (cutoffs, évaluations logiques, normes actuelles, etc.). Cette étape a été continue tout au long du stage, évoluant au fur et à mesure des besoins du projet.

4.2 Outils et technologies utilisés

Avant de présenter en détail le travail réalisé au cours de ces six mois, il est essentiel de comprendre le fonctionnement et l'utilité d'AlphaFold et de Foldseek, deux outils qui ont joué un rôle central dans les différentes étapes du projet.

4.2.1 AlphaFold2

Comme expliquer précédemment, les protéines se replient spontanément dans l'organisme pour former des structures complexes, et ce repliement détermine directement leur fonction biologique. Il est donc essentiel de pouvoir identifier et analyser ces structures. Parmi les approches classiques utilisées pour cela, on trouve la cristallographie aux rayons X, la cryo-microscopie électronique *Cryo-EM* et la résonance magnétique nucléaire *RMN*. Il est également possible de prédire les structures protéiques à l'aide d'algorithmes fondés sur l'homologie, sur la reconnaissance de repliement, ou plus récemment grâce au deep learning, avec AlphaFold2, devenu la méthode la plus fiable à ce jour.

AlphaFold, développé par DeepMind, est une intelligence artificielle capable de prédire la structure tridimensionnelle des protéines à partir de leur séquence d'acides aminés. Ce développement a constitué une avancée majeure en biologie structurale, permettant de réduire un processus qui prenait des semaines en laboratoire à quelques heures de calcul. AlphaFold a ainsi transformé divers domaines de la biologie et de la médecine, notamment en recherche biomédicale, en industrie pharmaceutique et en biologie fondamentale. Il utilise des méthodes d'apprentissage profond ainsi que des informations évolutives issues des alignements multiples de séquence *MSA* pour prédire les structures protéiques avec une précision comparable à celle des techniques expérimentales. Les prédictions sont évaluées à l'aide de métriques telles que le pLDDT et pTM.

Le pLDDT, *Predicted Local Distance Difference Test*, est un score de confiance qui indique la fiabilité de la prédiction à l'échelle locale, permettant d'identifier les régions où la structure est la plus fiable pour l'analyse. Le pTM, *Predicted Template Modeling Score*, évalue la précision globale du repliement d'une protéine individuelle. Il permet d'estimer la confiance dans la structure prédite, notamment en ce qui concerne l'organisation des domaines et les relations spatiales à longue distance.

AlphaFold a évolué à travers plusieurs versions. La première, lancée en 2016, exploitait des réseaux neuronaux convolutifs et l'alignement multiple pour prédire les structures protéiques. La version 2, lancée en 2018 et utilisée dans notre projet, repose sur un modèle de transformateur graphique et un processus de recyclage, ce qui améliore la précision des prédictions. Lors du CASP14, *Critical Assessment of Protein Structure Prediction* en 2020[16], AlphaFold2 a surpassé toutes les autres méthodes en atteignant une précision proche de 90 % pour la majorité des protéines testées. (voir **Figure 5-a**) Cette version améliore considérablement la modélisation des interactions à longue distance, gère mieux les régions désordonnées et flexibles, et permet la prédiction de complexes multiprotéiques. AlphaFold3, publié en 2021, s'est orienté vers la prédiction des interactions entre protéines et diverses molécules, telles que des ligands, de l'ADN ou encore de l'ARN. Dans le cadre de notre projet, nous nous concentrerons sur AlphaFold2, qui offre une large gamme d'analyses.

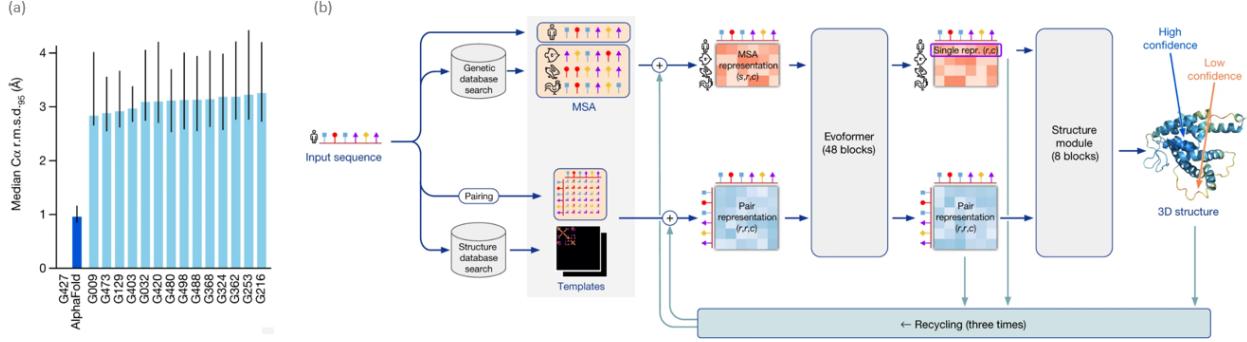


Figure 5: Performances d’AlphaFold2 au CASP14 et schéma de son fonctionnement publié par Jumper et al[16] . (a) Résultat du CASP14 avec comparaison des performances prédictives d’AlphaFold2 avec d’autres modèles testées, mesurées par la RMSD moyenne entre la conformation prédite et la conformation cible. AlphaFold2 montre une meilleure précision avec une RMSD médiane inférieure à 1Å. (b) Schéma simplifié du fonctionnement d’AlphaFold2. À partir d’une séquence d’acides aminés en entrée, l’algorithme utilise deux types de bases de données : génétiques pour générer des alignements multiples de séquences (MSA), et structurelles pour identifier des modèles structuraux (templates). Ces informations sont fusionnées et transformées en représentations internes par l’Evoformer, un réseau de neurones profond composé de 48 blocs. Ces représentations sont ensuite exploitées par le module structurel, composé de 8 blocs, pour générer une structure 3D finale. Le processus est itératif avec trois cycles de recyclage permettant de raffiner les prédictions. Une structure 3D est sortie par le réseau, avec une confiance associée à chaque région de la structure qui est codée par couleur (bleu = haute confiance, orange = faible confiance).

Algorithme d’AlphaFold2 (voir Figure 5-b)

L’algorithme commence par la saisie d’une séquence protéique sous forme d’une chaîne d’acides aminés. AlphaFold2 recherche ensuite des séquences homologues dans de vastes bases de données protéiques afin de générer un alignement multiple. Cette étape permet d’identifier des contraintes évolutives, notamment les co-mutations, qui révèlent des interactions stabilisant la structure finale de la protéine.

L’Evoformer transforme ensuite les données d’entrée en une représentation structurale riche, exploitable par la suite du modèle. Il reçoit en entrée la représentation MSA, qui capture les informations évolutives, ainsi qu’une matrice de paires de résidu, qui encode les relations spatiales potentielles entre acides aminés. Cette dernière est construite à partir de structures modèles *templates* identifiées dans les bases de données. L’architecture de l’Evoformer repose sur un réseau profond composé de 12 à 48 couches, intégrant plusieurs mécanismes essentiels à la prédiction de la structure finale.

Une fois les informations traitées par l’Evoformer, AlphaFold2 utilise un réseau de neurones spécialisé pour convertir ces représentations en coordonnées atomiques. Contrairement aux approches traditionnelles, AlphaFold2 effectue un raffinement cyclique en réinjectant ses prédictions précédentes comme nouvelles entrées, ce qui améliore progressivement la précision. L’optimisation s’arrête lorsque le modèle atteint un seuil de convergence et ne peut plus améliorer la prédiction.

AlphaFold2 produit ensuite une structure tridimensionnelle accompagnée de deux métriques d’évaluation, la première, le pLDDT, la seconde, l’pTM. Grâce à ces métriques, il est possible d’interpréter et de valider les prédictions produites par AlphaFold2, offrant ainsi une solution robuste pour la modélisation des structures protéiques.

Limite d’AlphaFold2

Bien qu’AlphaFold2 représente une avancée majeure, il présente encore certaines limitations. Sa précision dépend fortement des données utilisées pour l’alignement multiple des séquences *MSA*. AlphaFold2 éprouve également des difficultés à modéliser les protéines intrinsèquement désordonnées ou celles dont la conformation varie en fonction de l’environnement. Son approche produit une structure statique, alors que de nombreuses protéines adoptent des formes multiples pour remplir leur fonction biologique.

Ceci représente l’entièvre problématique de mon sujet de stage qui est d’améliorer les prédictions de AlphaFold2 afin qu’il puisse nous donner la structure apo et holo d’une protéine. AlphaFold2 parvient à prédire correctement la conformation holo dans environ 70 % des cas (voir **Figure 3**).

L’objectif du pipeline AlphaConformer est donc de construire des MSA, alignements multiples de séquences, et de sélectionner des modèles structuraux, templates, appropriés, afin de guider AlphaFold2 vers la prédiction des deux conformations : apo et holo. Le pipeline permet ainsi de lancer directement AlphaFold2 à l’étape de l’Evoformer de son processus, en contournant certaines étapes initiales pour un contrôle plus précis de la prédiction. Dans le cadre de notre projet, nous utilisons ColabFold[18] afin de lancer AlphaFold2 plus facilement, en y intégrant nos propres MSA et templates.

4.2.2 Foldseek

Le projet SPPICES s’appuie sur l’hypothèse ConTemplate[12], qui suppose que si deux protéines partagent une conformation, elles ont probablement d’autres conformations en commun. À partir de cette idée, on cherche à identifier des protéines similaires, appelées homologues, pour aider AlphaFold2 à prédire plusieurs formes possibles d’une même protéine.

Ces protéines homologues servent à la fois de templates et de sources de séquences dans les MSA, ce qui permet de mieux orienter les prédictions d'AlphaFold2 vers les différentes conformations possibles. Pour identifier ces homologues, deux approches principales existent, la comparaison par séquence, qui est sensible aux variations locales et moins efficace dans les cas de faible conservation, et la comparaison par structure 3D, qui permet de détecter des relations plus lointaines entre protéines, même lorsque leur séquence a divergé.

Dans le cadre de notre projet, nous cherchons à exploiter les structures 3D pour améliorer la qualité des modèles. Cependant, les méthodes classiques d'alignement structurel, comme TM-align, bien qu'efficaces, sont particulièrement lentes et donc peu adaptées à l'analyse de grands jeux de données.

C'est dans ce contexte que l'outil Foldseek s'impose comme une solution de choix. Il a profondément révolutionné l'alignement structurel en combinant rapidité, robustesse et précision. Contrairement aux approches traditionnelles, Foldseek transforme les structures 3D en une représentation simplifiée sous forme d'alphabet structurel : par exemple, les hélices sont codées par "H", les feuillets par "E", les coudes par "C", etc. Cette abstraction permet de réduire la complexité des structures, tout en capturant des motifs pertinents, facilitant ainsi leur comparaison.

Foldseek utilise ensuite un algorithme dérivé de MMseqs2[20], un moteur de recherche rapide et efficace initialement conçu pour la comparaison de séquences d'acide aminé. Grâce à cette combinaison, Foldseek peut effectuer des alignements locaux sur les représentations structurelles, tout en conservant des performances très élevées, notamment pour détecter des domaines homologues multiples. Foldseek s'impose ainsi comme la solution pour les alignements structurels, qui offre le parfait compromis entre performance et coût d'exécution (voir **Figure 6**)

Pour évaluer les résultats des alignements entre protéines, plusieurs scores sont utilisés. Le premier est le bit score, qui mesure la qualité globale de l'alignement : plus il est élevé, meilleur est l'alignement. Ensuite on utilise l'e-value, qui permet de savoir si l'alignement observé pourrait être obtenu par hasard : plus cette valeur est petite, plus le résultat est considéré comme fiable. Enfin, on utilise le TM-score qui sert à comparer la forme globale des deux protéines, peu importe leur taille.

Dans notre projet, Foldseek est utilisé à de nombreuses reprises, notamment pour identifier des protéines homologues à partir de grandes bases de données structurales, pour sélectionner les templates à fournir à AlphaFold2 en entrée, et pour constituer les MSA qui guideront la prédiction conformationnelle.

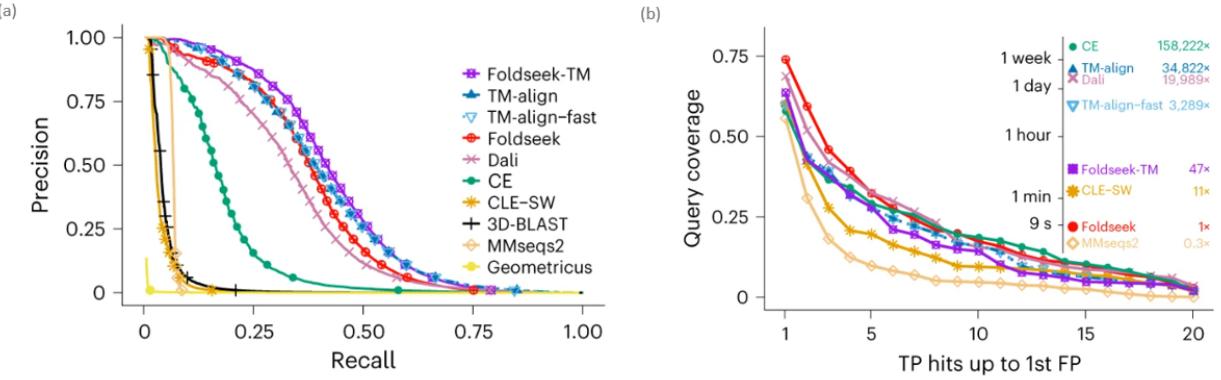


Figure 6: Performances de Foldseek publié par Van Kempen et al. [19]. (a) Distribution des sensibilités pour la détection d’homologies dans la base de données SCOPe40, évaluée via l’aire sous la courbe ROC. Les outils Foldseek-TM, TM-align et Foldseek présentent les meilleures sensibilités, indiquant des performances supérieures pour l’identification d’homologues. (b) Évaluation de la sensibilité des recherches sur des modèles protéiques AlphaFold2, incluant des structures multi-domaines et de pleine longueur. Foldseek et Dali se distinguent par leurs hautes performances, mais Foldseek offre un avantage majeur en vitesse : il est environ 20 000 fois plus rapide que Dali tout en ayant des performances comparables.

4.3 Pipeline AlphaConformer

Avant d’aborder précisément les tâches effectuées durant le stage, il était également nécessaire d’acquérir une compréhension approfondie de la pipeline AlphaConformer. Cette étape de compréhension a constitué l’une des toutes premières de mon travail. En effet, il est difficile de comparer AlphaConformer à d’autres méthodes, de préparer efficacement les données nécessaires à son évaluation ou encore d’en optimiser les performances, sans en maîtriser les principes de base.

Comprendre son architecture, son mode de fonctionnement, et les choix technologiques qui la composent était donc un prérequis fondamental pour mener à bien l’ensemble des missions qui m’ont été confiées par la suite.

4.3.1 Présentation du fonctionnement

AlphaConformer prend en entrée une protéine au format .pdb et vise à générer plusieurs modèles structuraux, dont certains doivent se rapprocher de la forme holo et d’autres de la forme apo, afin de reproduire la diversité conformationnelle naturelle de la protéine. Les étapes du pipeline sont schématisées dans la **Figure 7**

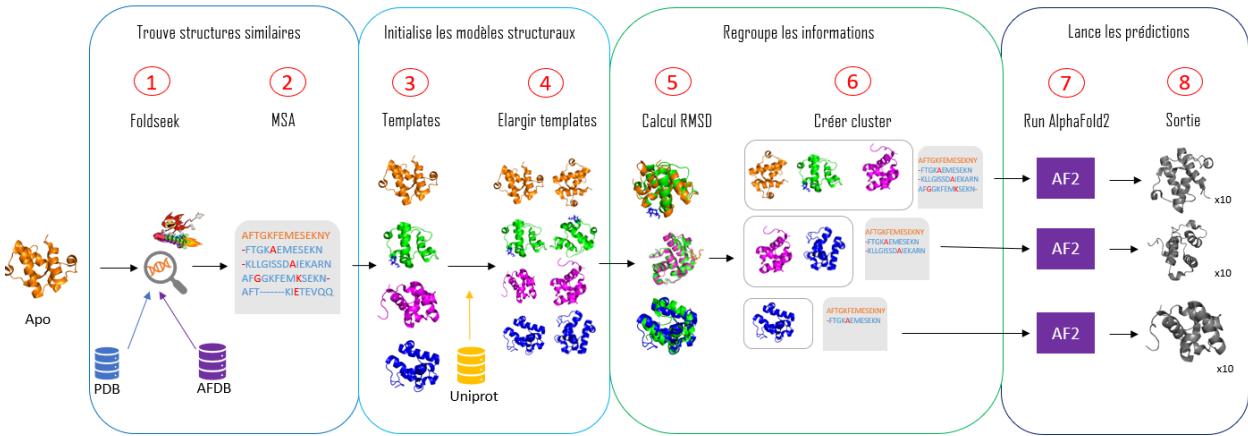


Figure 7: Pipeline AlphaConformer. Le pipeline prend en entrée un fichier .pdb correspondant à une structure protéique. (1) Foldseek est utilisé pour rechercher, dans les bases de données PDB[21] et AFDB[22], des structures similaires. (2) Les séquences associées aux résultats de Foldseek sont extraites et alignées pour former un alignement de séquence multiple (MSA). (3) Les structures correspondantes sont également récupérées et alignées afin d'être utilisées comme templates pour AlphaFold2. (4) Pour enrichir cet ensemble de templates, la base UniProt est interrogée afin de récupérer toutes les structures partageant le même identifiant d'accèsion que celles détectées par Foldseek. (5) Toutes les structures collectées sont comparées deux à deux à l'aide de la RMSD, afin de mesurer leurs différences conformationnelles. (6) À partir de cette matrice de similarité, des clusters sont générés selon un cutoff RMSD donné. (7) Chaque cluster est ensuite utilisé indépendamment comme entrée pour AlphaFold2, avec son propre MSA et ses propres templates. AlphaFold2 va prédire 10 structures pour chaque clusters L'objectif est que, parmi les 10 structures prédites, au moins l'une corresponde à une conformation holo et une autre à une conformation apo.

La première étape de la pipeline repose sur l'utilisation de Foldseek qui permet d'identifier des structures protéiques similaires à celle donnée en entrée, à partir d'une ou plusieurs bases de données structurelles (voir **Figure 7-1**) . Les protéines homologues détectées servent ensuite à générer un MSA, utilisé pour fournir un contexte évolutif à AlphaFold2, afin d'orienter ses prédictions vers des états conformationnels alternatifs (voir **Figure 7-2**).

Les alignements issus de Foldseek sont triés en fonction de leur e-value, plus l'e-value est faible, plus la similarité entre les structures est significative. Les séquences retenues sont ensuite alignées et sauvegardées dans un fichier au format .a3m, utilisé en entrée d'AlphaFold2 en MSA.

Pour développer ce pipeline, nous utilisons deux bases de données comme sources pour Foldseek. La première est la base PDB, *Protein Data Bank*[21], c'est la base de données de référence mondiale pour les structures 3D de macromolécules biologiques. La seconde est AFDB *AlphaFold Protein Structure Database*[22], cette base contient les prédictions de structures effectuées par AlphaFold pour des millions de protéines, accompagnées d'un score pLDDT élevé. L'AFDB constitue une ressource essentielle, notamment pour les protéines sans structure expérimentale connue.

Dans un second temps, les structures associées aux séquences identifiées par Foldseek sont récupérées. Ces structures sont ensuite alignées et utilisées comme modèles structuraux *templates* pour guider la prédiction d'AlphaFold2 (voir **Figure 7-3**). L'intégration de ces templates fournit une information tridimensionnelle précieuse, complémentaire au MSA, et permet d'orienter AlphaFold2 vers des conformations alternatives spécifiques.

Cette étape constitue la principale valeur ajoutée du pipeline AlphaConformer par rapport aux autres approches existantes d'exploration conformationnelle. En combinant données évolutives avec les MSA et structurales avec les templates, AlphaConformer a pour objectif d'améliorer significativement la diversité et la précision des structures prédites.

La troisième étape vise à élargir encore davantage la diversité des templates. Pour cela, toutes les protéines partageant le même identifiant UniProt[23] que celles identifiées par Foldseek sont intégrées, même si leurs séquences exactes ou leurs structures diffèrent légèrement (voir **Figure 7-4**). Cette stratégie permet d'englober plusieurs conformations naturelles de la même protéine, favorisant ainsi l'exploration d'états conformationnels alternatifs.

Une fois tous les modèles structuraux collectés, nous évaluons leur similarité avec la structure de départ en calculant la RMSD, une mesure quantitative de la distance moyenne entre les atomes de deux structures alignées (voir **Figure 7-5**). Ces valeurs de RMSD permettent ensuite de regrouper les protéines identifiées en clusters, selon différents seuils (cutoffs), afin de constituer des sous-groupes de modèles plus ou moins proches de la structure initiale. Chaque cluster possède son propre ensemble de MSA et de templates (voir **Figure 7-6**).

AlphaFold2 est alors lancé indépendamment sur chacun de ces clusters pour générer une série de prédictions, avec 10 modèles par cluster (voir **Figure 7-7**). Ces prédictions sont évaluées à l'aide des scores internes d'AlphaFold2, notamment le pLDDT, qui mesure la confiance locale du modèle.

Pour évaluer les performances du pipeline, les modèles générés sont comparés aux structures expérimentales apo et holo via le calcul du RMSD entre les prédictions et les structures cibles. Cette étape permet d'identifier, parmi tous les modèles produits, ceux qui s'approchent le plus des états conformationnels réels.

L'objectif global est donc de guider AlphaFold2 pour qu'il explore de manière contrôlée l'espace structural d'une protéine, en tirant parti à la fois de l'information évolutive avec le MSA, des structuraux avec les templates et de la variabilité fonctionnelle naturelle via les clusters RMSD.

4.3.2 Résultats initiaux

Les premiers résultats obtenus avec le pipeline AlphaConformer sont encourageants, bien qu'ils ne soient actuellement satisfaisants que pour 4 protéines sur 12. Parmi elles, la protéine 1AKZ se distingue particulièrement : nous avons réussi à générer des prédictions dont la distance RMSD avec les structures expérimentales apo et holo est inférieure à 1 Ångströms, ce qui constitue un excellent niveau de précision (voir **Figure 8**). Ce type de résultat représente l'objectif cible à atteindre pour l'ensemble des protéines de notre jeu de données.

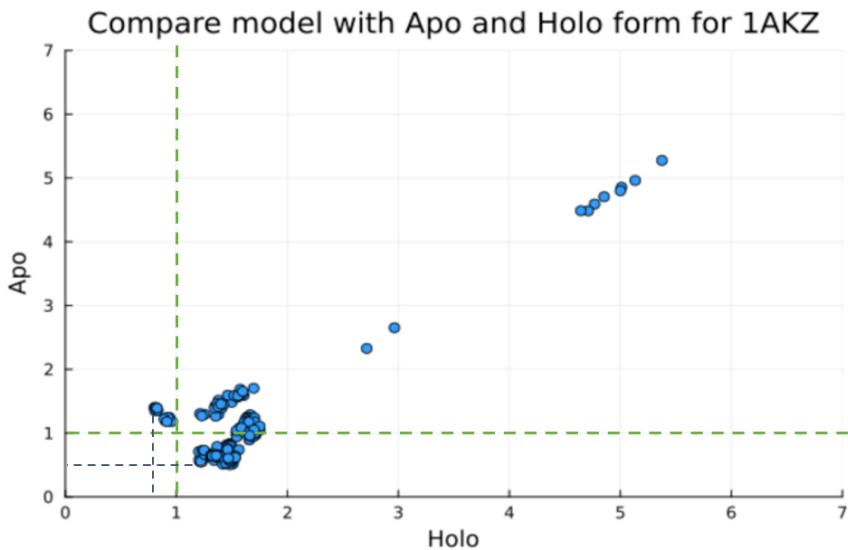


Figure 8: Premier résultat d'AlphaConformer sur la protéine 1AKZ. Chaque point bleu correspond à une structure sortie du pipeline. Nous avons calculé la RMSD avec les structures apo et holo afin de visualiser la dispersion des prédictions. Un résultat est considéré comme très satisfaisant si la RMSD avec la structure apo ou holo est inférieure à 1 Å.

Après la publication des premiers résultats, Diego Zea a entrepris plusieurs modifications importantes pour optimiser le pipeline. Initialement, la recherche de structures similaires était réalisée à l'aide de US-align[26], un algorithme d'alignement de structure. Toutefois, cette méthode s'est révélée trop lente pour les volumes de données manipulés. Elle a été remplacée par Foldseek[19], un outil beaucoup plus rapide et performant pour effectuer des recherches de similarité structurale.

Par ailleurs, Diego Zea a également modifié la méthode de clustering des structures. auparavant, l'algorithme utilisé était Hobohm[27], qui cherche à constituer des clusters en comparant les structures déjà regroupées avec les nouvelles structures candidates, ce qui permet de gagner du temps. Cependant, cette méthode générait des clusters plus larges mais potentiellement moins homogènes. Aujourd'hui, la méthode mise en place consiste à comparer directement chaque structure .pdb avec toutes les autres, afin de former davantage de clusters, mais plus homogènes, contenant des protéines plus similaires. Cela réduit la taille des MSA associés, mais améliore la précision des templates utilisés par AlphaFold2.

À mon arrivée, ma première mission a consisté à tester le pipeline et à corriger les éventuelles erreurs d'exécution. Les premiers mois ont ainsi été consacrés à la résolution de bugs techniques, jusqu'à parvenir à un fonctionnement stable du pipeline sur un ensemble de 12 protéines.

Suite aux modifications apportées, notamment l'intégration de Foldseek et l'ajout de nouveaux algorithmes de clustering, nous avons constaté un changement notable dans la nature des résultats générés. Le pipeline produit désormais un plus grand nombre de prédictions structurales. Cependant, les prédictions obtenues sont majoritairement plus proches de la conformation apo que de la conformation holo, ce qui est cohérent puisque la structure donnée en entrée du pipeline est la structure apo. Dans 75% des cas testés, les prédictions se rapprochent davantage de la structure d'entrée que de la conformation alternative. Malgré cela, la qualité des prédictions pour la conformation holo reste globalement correcte : nous considérons une prédition comme excellente si la RMSD est inférieure à 1 \AA , acceptable entre 1 \AA et 2 \AA , et mauvaise au-delà de 3 \AA . Les mauvaises prédictions pour la conformation holo ne représentent que 15% des cas, ce qui reste un taux encourageant.

Par exemple, pour la protéine 1AKZ, on obtient une RMSD inférieure à 1 \AA pour la structure apo. En revanche, les résultats obtenus pour les conformations holo restent en retrait, avec des RMSD supérieures à 1 \AA , ce qui reste acceptable mais encore perfectible (voir **Figure 9-a**). Un autre exemple, avec la protéine 2F63, montre quelques prédictions prometteuses, mais aucune ne parvient à atteindre une RMSD inférieure à 2 \AA pour la conformation holo (voir **Figure 9-b**). Cela montre que le pipeline est sur la bonne voie, mais qu'un travail supplémentaire est nécessaire pour équilibrer la qualité des prédictions entre les deux états conformationnels.

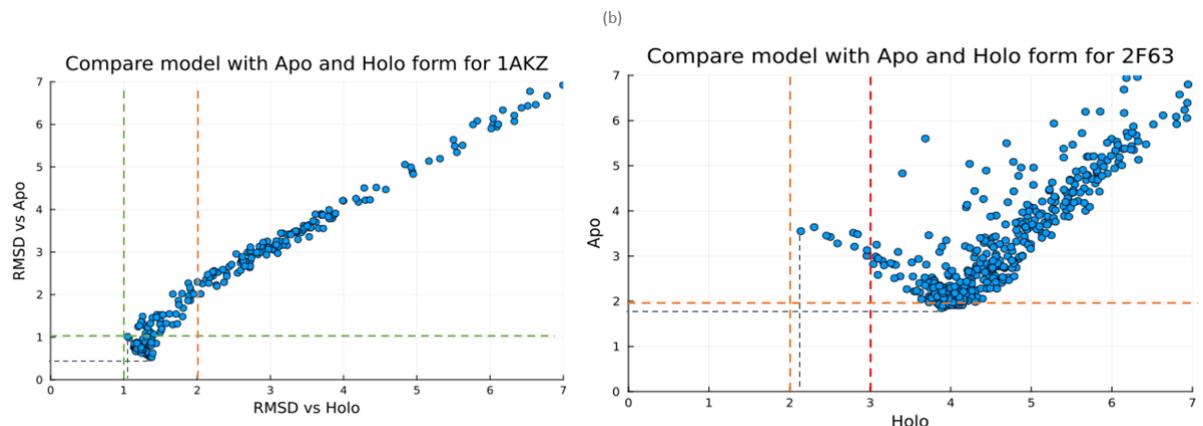


Figure 9: Résultats d'AlphaConformer sur les protéines 1AKZ et 2F63 avant optimisation. (a) Pour 1AKZ, les prédictions présentent une RMSD comprise entre 1 et 2 \AA par rapport aux structures de référence. (b) Pour 2F63, les RMSD se situent entre 2 et 3 \AA . Ces résultats sont satisfaisants mais restent perfectibles, en particulier pour 2F63. L'objectif est d'atteindre une précision comprise entre 1 et 2 \AA pour un maximum de protéines testées.

5 Résultats

5.1 Evaluation de la pipeline

Pour commencer, nous devions évaluer objectivement les performances de notre pipeline AlphaConformer, il était essentiel de le comparer aux méthodes actuelles qui cherchent également à explorer ou améliorer la diversité des prédictions générées par AlphaFold2. Plusieurs approches récentes, telles que AF-Cluster et BioEmu, ont proposé des stratégies innovantes.

5.1.1 AF-Cluster

Les protéines ont souvent la capacité d'adopter différentes conformations pour accomplir diverses fonctions biologiques. Toutefois, les méthodes de prédiction de structure, comme AlphaFold2, se concentrent généralement sur une seule conformation, ce qui ne reflète pas la réelle diversité fonctionnelle des protéines.

Pour aborder ce défi, l'équipe de Kern[29] a développé en 2024 AF-cluster[30], proposant une méthode combinant le clustering des MSA avec DBSCAN[28] et ensuite utilisé AlphaFold2. L'équipe suggère que les différentes conformations fonctionnelles d'une protéine sont conservées au sein de familles de séquences évolutivement similaires. En regroupant ces séquences par similarité, il est possible de capturer des signaux évolutifs spécifiques à chaque conformation.

AF-cluster effectue un clustering des séquences MSA, puis utilise chaque cluster comme entrée distincte pour AlphaFold2. Cela permet de générer plusieurs modèles de structure, chacun correspondant à une conformation potentielle de la protéine. L'équipe a validé cette méthode en l'appliquant à des protéines connues pour adopter plusieurs conformations, notamment sur KaiB, démontrant ainsi la capacité de AF-cluster à prédire efficacement les états alternatifs.

La méthode AF-Cluster adopte une approche similaire à celle du pipeline AlphaConformer, mais avec une stratégie centrée exclusivement sur la manipulation des MSA, contrairement à AlphaConformer qui combine également l'usage de modèles structuraux avec les templates. AF-Cluster cherche uniquement à orienter les prédictions d'AlphaFold2 en regroupant les séquences du MSA selon leur similarité.

Dans leur étude, les auteurs montrent que cette approche permet d'obtenir, avec un bon taux de réussite, des prédictions des structures apo et holo avec une confiance élevée (voir **Figure 10**). Pour évaluer l'efficacité de leur méthode, ils génèrent également des MSA aléatoires qui servent de contrôles, permettant ainsi de démontrer que le regroupement ciblé améliore significativement les résultats.

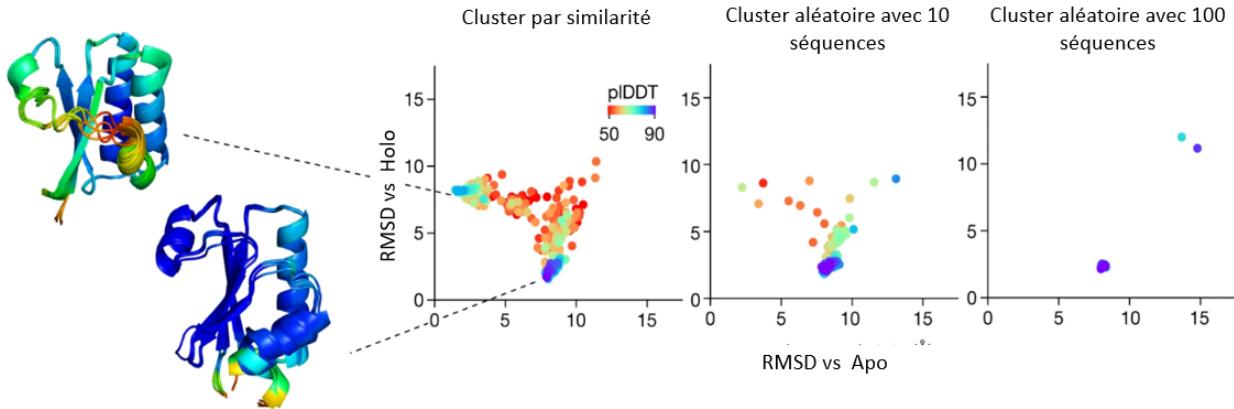


Figure 10: Résultats publiés par Wayment-Steele et al. pour AF-Cluster [30]. Le pipeline regroupe les séquences du MSA selon leur similarité, ce qui permet de générer des conformations proches de la forme apo et de la forme holo, tout en conservant un score de confiance plDDT élevé. Les auteurs comparent ces résultats à ceux obtenus à partir de MSA aléatoires, qui ne parviennent à prédire que la conformation holo.

Les conclusions de l’article suggèrent que plus les séquences présentes dans le MSA sont homogènes en termes de conformation, plus AlphaFold2 est susceptible de prédire correctement les différentes structures adoptées par une même protéine. Cette méthode met donc en lumière le rôle central de la composition du MSA dans l’orientation des prédictions structurelles d’AlphaFold2.

Nous avons appliqué la méthode AF-Cluster à notre jeu de données test, issu de l’étude de Saldaño[13], afin d’évaluer sa capacité à prédire les conformations apo et holo. L’exécution du pipeline sur 12 protéines a montré que le regroupement des séquences MSA par similarité permettait de générer une plus grande diversité structurale, augmentant ainsi le nombre total de prédictions par rapport à l’utilisation de MSA aléatoires. Toutefois, une analyse plus fine de la qualité des conformations obtenues révèle que, pour 9 protéines sur 12, cette structuration par similarité n’améliore pas la précision des prédictions. Au final, AF-Cluster n’a permis de retrouver à la fois les conformations apo et holo, avec des RMSD inférieurs à 3Å, que dans 25% des cas.

Par exemple, pour la protéine 2F63 (voir **Figure 11-a**), nous avons constaté une légère amélioration avec AF-Cluster par rapport aux MSA aléatoires (U100, correspondant à 100 séquences sélectionnées aléatoirement de manière uniforme). Toutefois, ces gains restent modestes : la RMSD minimale par rapport à la structure apo passe ainsi de 1,9Å avec un MSA aléatoire à 1,8Å avec AF-Cluster.

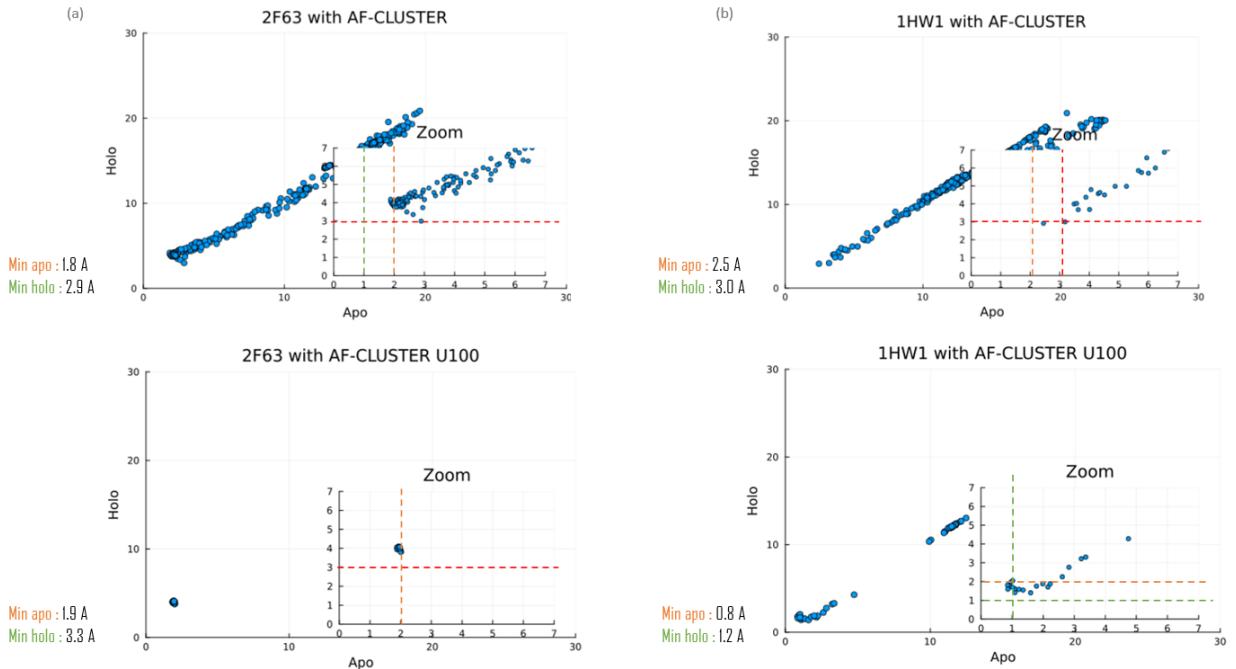


Figure 11: Résultats d’AF-Cluster appliqués à la base de données de Saldaño [13]. Nous avons comparé les prédictions d’AlphaFold2 générées à partir de deux types de MSA : ceux construits par similarité avec AF-Cluster et des MSA aléatoires avec 100 séquences. Pour chaque configuration, nous avons analysé la RMSD minimale obtenue par rapport aux conformations apo et holo. (a) AF-Cluster permet d’obtenir de meilleures RMSD que les MSA aléatoires, avec des différences modérées allant de 0,1 à 0,4 Å. (b) Illustrent un exemple contraire où les MSA aléatoires donnent de meilleurs résultats, avec des écarts plus marqués allant de 1,5 à 1,8 Å. On retrouve ces résultats pour 9 protéines sur 12.

En revanche, pour les protéines 1HWI (voir **Figure 11-b**), les MSA générés par AF-Cluster ne conduisent pas à de meilleures prédictions. Au contraire, les différences de performance sont marquées. AF-Cluster génère une structure avec une RMSD de 2,5 Å par rapport à la conformation apo, tandis que les MSA aléatoires permettent d’atteindre une prédition bien plus précise, avec une RMSD de seulement 0,8 Å. Ce type d’écart est significatif, puisqu’il correspond à un passage d’une approximation globale (entre 2 et 3 Å) à une quasi-superposition atomique (entre 0 et 1 Å).

Ces résultats suggèrent que, bien que les MSA construits par AF-Cluster soient plus homogènes, cette homogénéité tend davantage à éloigner les prédictions des conformations cibles apo ou holo, plutôt qu’à les en rapprocher. En définitive, les MSA issus d’AF-Cluster n’apportent pas de bénéfices clairs en termes de diversité conformationnelle, et ne permettent pas, sur la majorité des protéines testées, de dépasser les limitations connues d’AlphaFold2.

Ces résultats mettent en évidence un point fondamental pour notre pipeline AlphaConformer, la diversité des MSA est essentielle. Plus les MSA sont variés, plus ils intègrent une richesse d’informations évolutives susceptibles de guider AlphaFold2 vers les conformations apo et holo.

5.1.2 BioEmu

La méthode BioEmu[32] a également été développée dans le but de prédire efficacement la diversité conformationnelle des protéines tout en réduisant drastiquement les coûts en temps et en calcul. BioEmu vise à atteindre une précision équivalente à celle des simulations de dynamique moléculaire bien convergées ou des expériences cryo-EM à résolution multiple, tout en nécessitant seulement quelques heures de calcul, et pour un coût très réduit. L'objectif de cette approche est d'entraîner un modèle capable de reproduire fidèlement les comportements expérimentaux des protéines.

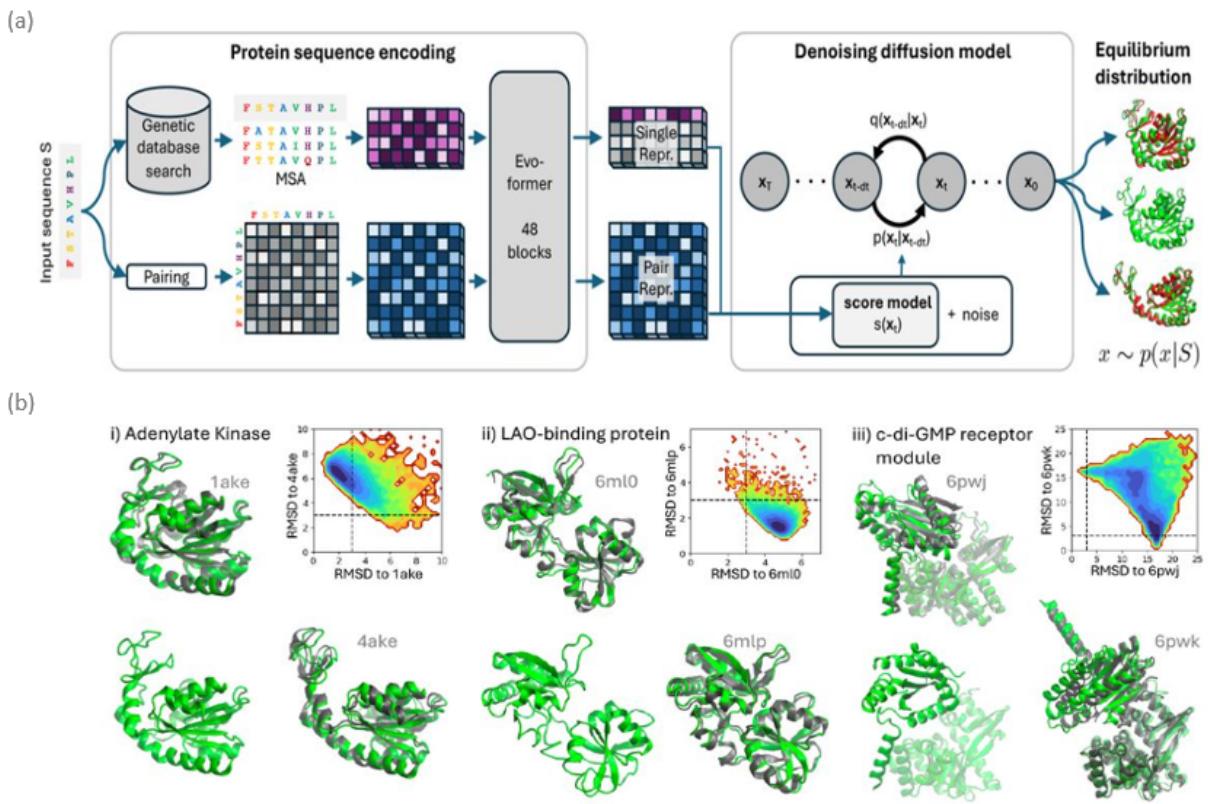


Figure 12: Résultats publiés par Lewis et al. sur BioEmu[32]. (a) Schéma de l'architecture du pipeline BioEmu. Le modèle reçoit en entrée une séquence protéique, applique le pipeline d'AlphaFold2, puis récupère les sorties de l'Evoformer pour lancer un modèle de diffusion. Ce dernier introduit un bruit contrôlé afin d'explorer différentes conformations structurales. (b) Performances de BioEmu sur trois protéines disposant de structures apo et holo expérimentales. Le modèle parvient à prédire ces deux états avec une RMSD inférieure à 3Å, illustrant sa capacité à générer des conformations alternatives réalisistes.

Contrairement à AlphaFold2, qui prédit une seule structure stable, BioEmu génère un ensemble de conformations possibles à travers un processus de diffusion, simulant la transition progressive d'un bruit aléatoire vers une structure stable. BioEmu reprend le pipeline d'AlphaFold2, puis exploite les sorties de l'Evoformer pour appliquer son propre processus de diffusion (voir **Figure 12-a**). Ce mécanisme permet d'explorer l'espace conformationnel de manière contrôlée et efficace, en seulement une centaine d'étapes de génération.

Pour construire un modèle robuste, BioEmu est d'abord pré-entraîné sur une version compressée de la base de données AFDB[22], enrichie par des techniques d'augmentation de données favorisant la diversité conformationnelle. Il est ensuite affiné à l'aide de simulations de dynamique moléculaire et de mesures expérimentales de stabilité protéique.

Les résultats publiés par BioEmu sont très encourageants (voir **Figure 12-b**). Le modèle est capable de prédire avec justesse les états fonctionnels de nombreuses protéines. Malgré quelques limitations dans la prédiction précise de certains états apo complexes, BioEmu parvient à reproduire avec une bonne fidélité des conformations expérimentales. Il atteint un taux de prédiction correct de 85 % avec une précision atomique inférieure à 3 Å de RMSD. Cela démontre sa capacité à explorer l'espace conformationnel de manière réaliste et à prédire des états fonctionnels pertinents avec une efficacité inégalée par les méthodes actuelles.

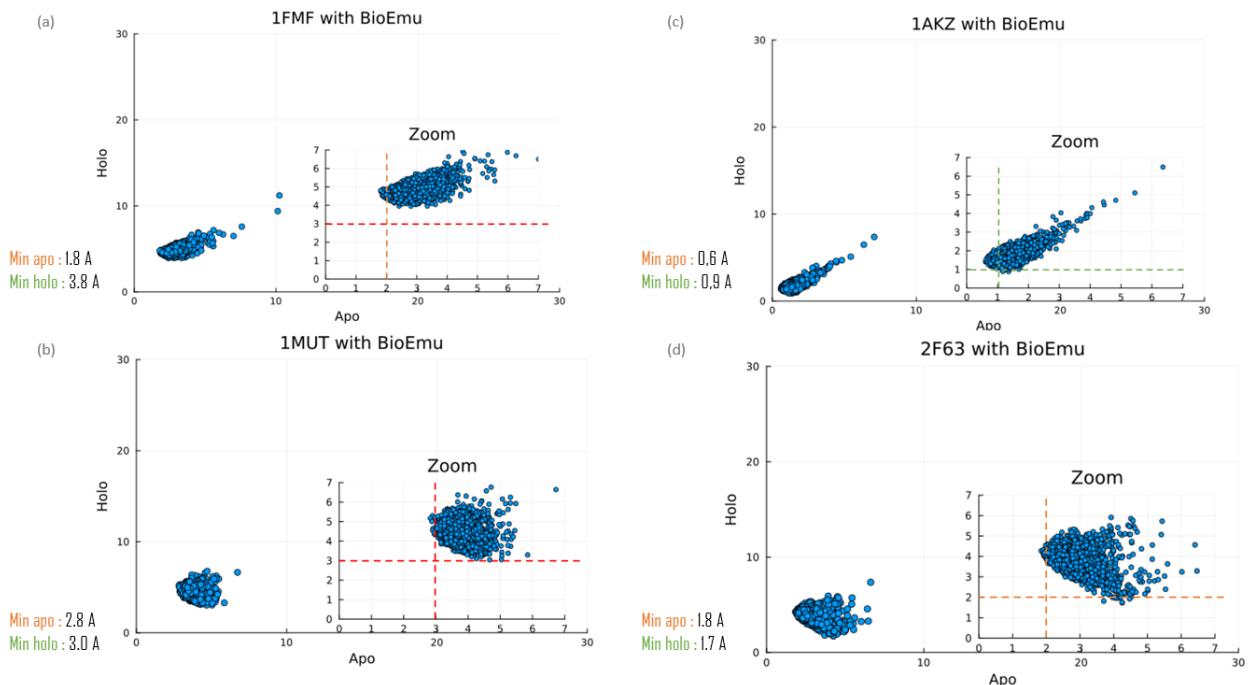


Figure 13: Résultats de BioEmu appliqués au jeu de données de Saldaño [13]. (a) et (b) Les prédictions ne sont pas optimales, avec des RMSD supérieures à 3 Å pour les structures holo. (c) et (d) Excellentes performances de BioEmu, avec des RMSD inférieures à 2 Å pour les deux conformations, apo et holo. On retrouve ses résultats pour 8 protéines sur 12.

Nous avons appliqué le modèle BioEmu à notre jeu de données test, basé sur l'étude de Saldaño[13], afin d'évaluer sa capacité à prédire les conformations apo et holo. Après exécution du pipeline sur 12 protéines, les résultats ont démontré des performances très prometteuses : pour 8 protéines, BioEmu parvient à prédire à la fois les conformations apo et holo avec une RMSD inférieure à 2Å (voir **Figure 13-c** et **d**). Pour les 4 protéines restantes, la conformation apo ou holo est prédictive avec une RMSD inférieure à 4Å (voir **Figure 13-a** et **b**).

Ces résultats mettent en évidence la robustesse de BioEmu dans l'exploration de la diversité conformationnelle. En injectant un bruit contrôlé sur les sorties d'AlphaFold2 via un modèle de diffusion, BioEmu parvient à générer des structures très proches des conformations apo et holo. Ce modèle représente donc, à ce jour, le principal concurrent de notre pipeline AlphaConformer. Reste à savoir si l'utilisation de templates structuraux, sans modifier les sorties internes d'AlphaFold2, permettra d'atteindre une performance comparable. C'est précisément ce que nous allons explorer dans la section suivante.

5.1.3 Comparaison avec AlphaConformer

Afin d'évaluer de manière rigoureuse et objective les performances du pipeline AlphaConformer, nous avons comparé ses résultats à ceux de deux autres méthodes récentes : AF-cluster et BioEmu.

Nous avons sélectionné un ensemble de 4 protéines issues du jeu de données de Saldaño[13], chacune de ces protéines a ensuite été analysée à l'aide des trois pipelines, afin de comparer leurs capacités respectives à prédire différentes conformations.

L'analyse des résultats montre que les trois pipelines présentent des performances globalement satisfaisantes (voir **Figure 14-a**). BioEmu se distingue particulièrement par sa capacité à prédire avec une grande précision les deux conformations, apo et holo, en atteignant une RMSD inférieure à 2Ångströms pour l'ensemble des protéines testées. AlphaConformer affiche des performances similaires, tandis qu'AF-Cluster ne parvient à atteindre ce niveau de précision pour les deux conformations que dans deux cas sur quatre.

Par exemple, pour la protéine 1AKZ, BioEmu prédit les conformations apo et holo avec une RMSD inférieure à 1 Ångströms. AlphaConformer suit de près avec une RMSD de 1,1 Ångströms pour la conformation holo. En revanche, AF-cluster reste au-dessus de 2 Ångströms pour les deux conformations, indiquant une moins bonne précision. (voir **Figure 14-b**)

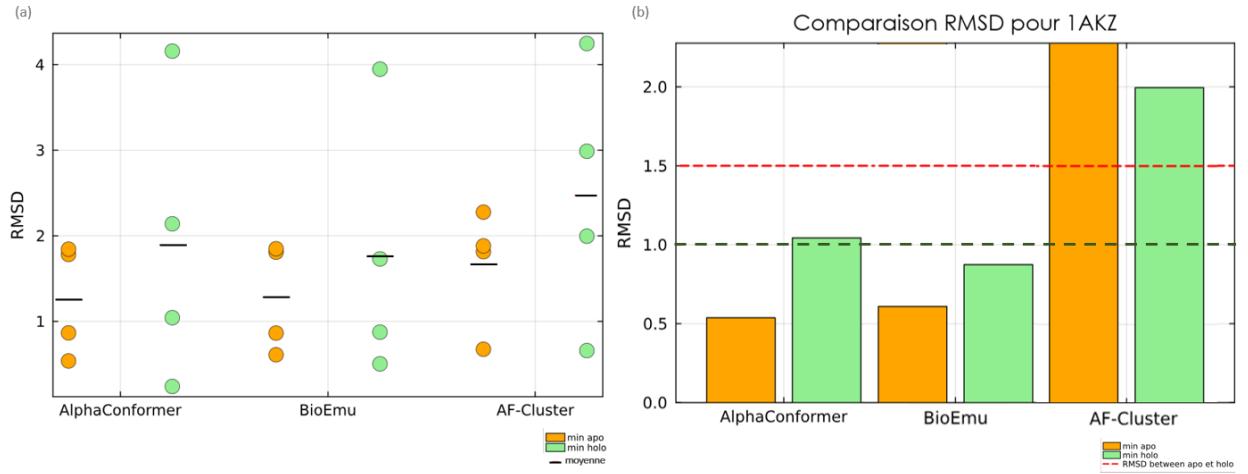


Figure 14: Comparaison des performances d’AlphaConformer, AF-cluster et BioEmu.(a) Évaluation des trois pipelines sur quatre protéines issues du jeu de données de Saldaño[13]. AlphaConformer et BioEmu présentent des performances comparables, tandis qu’AF-cluster montre une précision moindre dans la prédiction des conformations apo et holo. (b) Pour la protéine 1AKZ, BioEmu se démarque avec les meilleures prédictions, atteignant une RMSD inférieure à 1Å (ligne verte) pour les deux conformations.

Ces résultats montrent qu’AlphaConformer, grâce à l’utilisation de templates structuraux, améliore significativement les performances par rapport à AF-Cluster, qui n’utilise que les alignements multiples (MSA). Bien qu’il ne modifie pas le modèle AlphaFold2, AlphaConformer parvient à s’approcher des résultats de BioEmu, qui, elle, adapte directement AlphaFold2 par fine-tuning et ajout de bruit sur les sorties. L’utilisation des templates comme données d’entrée s’avère donc être une stratégie très prometteuse.

5.2 Amélioration d’AlphaConformer

Après avoir réussi à faire fonctionner le pipeline et l’avoir comparé aux méthodes déjà existantes, l’objectif suivant était d’améliorer les performances d’AlphaConformer afin de réussir à prédire les conformations apo et holo avec une RMSD inférieure à 2Å pour l’ensemble des protéines.

Après de nombreux tests et réflexions, nous avons décidé de revenir à l’utilisation de l’algorithme de Hobohm[27], sans subdiviser les résultats en plusieurs sous-groupes. En effet, nos diverses expérimentations ont montré que plus les MSA contiennent de séquences, plus AlphaFold2 parvient à prédire des structures proches des conformations apo et holo.

Dans le but d'améliorer davantage les performances du pipeline, nous avons ensuite mené une série d'expérimentations sur différents paramètres. Nous avons notamment testé :

- L'utilisation d'une ou deux bases de données (PDB seule, ou PDB + AFDB) comme entrée pour Foldseek.
- Différents seuils sur l'e-value des résultats de Foldseek pour filtrer les structures homologues : $1e-5$, $1e-1$ ou encore sans seuil (NaN).
- Plusieurs valeurs de cutoff pour l'algorithme de Hobohm 0.5 , 1 , 1.5 , et 2 \AA afin de définir les clusters structuraux.

Ces tests ont pour but d'identifier les configurations permettant d'obtenir les meilleures performances, tant en termes de diversité conformationnelle que de précision des prédictions.

L'exécution complète de ces tests étant particulièrement long, nous ne disposons pour l'instant que des résultats pour 2 protéines sur les 12 prévues (voir **Figure 15**). Malgré ce nombre limité, une tendance claire se dessine : les meilleures prédictions, celles qui se rapprochent le plus des conformations apo et holo, sont obtenues en combinant les bases de données PDB et AFDB.

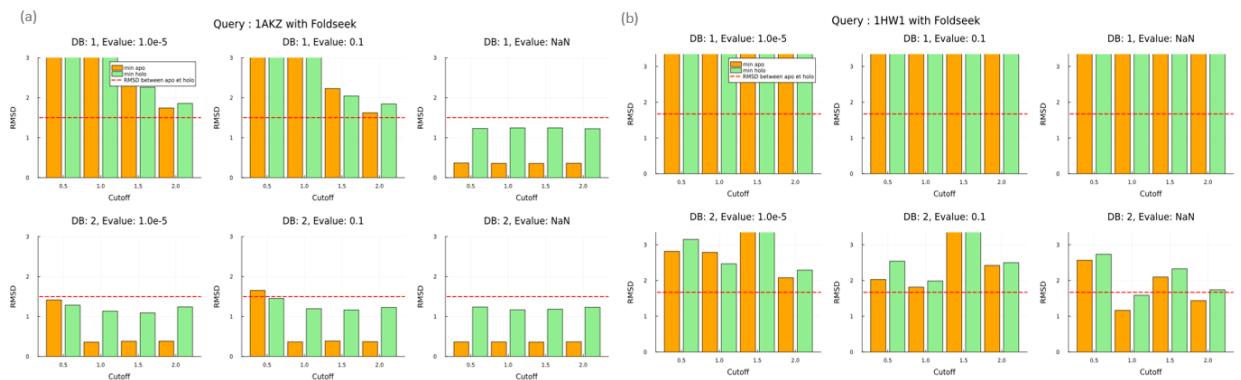


Figure 15: Test des différents paramètres d'AlphaConformer. Évaluation des performances du pipeline en fonction de plusieurs paramètres : utilisation d'une seule base de données (ligne 1, DB1) ou de deux bases combinées (ligne 2, DB2), application de seuils variés sur l'e-value des résultats de Foldseek, et modification du seuil de regroupement (cutoff) pour le clustering. L'objectif est d'obtenir des prédictions dont la RMSD est inférieure à 2\AA , idéalement plus faible que la RMSD entre les conformations apo et holo. (a) Les meilleurs résultats sont obtenus sans filtre sur Foldseek ; les différents cutoffs de clustering produisent des performances comparables. (b) L'utilisation combinée de deux bases de données améliore nettement les performances. En l'absence de filtre sur Foldseek et avec un cutoff de 1\AA , la RMSD prédite passe sous celle séparant les conformations apo et holo, ce qui constitue la configuration la plus performante testée.

Pour la protéine 1AKZ, toutes les prédictions présentent une RMSD inférieure à 1,5 Å qui correspond à la RMSD entre la structure apo et holo (voir **Figure 15-a**). Pour 1HW1, bien que les RMSD soient légèrement supérieures, elles restent nettement meilleures que celles obtenues en utilisant uniquement PDB (voir **Figure 15-b**). Dans les deux cas, la configuration la plus performante repose sur l'utilisation des deux bases de données, sans filtre sur les résultats de Foldseek, et avec un cutoff de 1 Å pour le clustering.

La prochaine étape consiste à étendre ces tests aux 10 autres protéines du jeu de données afin de vérifier la robustesse des paramètres identifiés. Une fois cette validation complétée, nous aurons optimisé les performances du pipeline AlphaConformer. Nous pourrons alors l'appliquer à notre propre base de données, dans le but d'évaluer la fiabilité et la généralisabilité des résultats obtenus.

5.3 Construction de la base de données

La dernière étape essentielle de mon stage a été la constitution d'une base de données regroupant des protéines ayant au moins deux conformations distinctes : une forme apo, sans ligand, et une forme holo, liée à un ligand. Cette base de données est indispensable pour l'entraînement et l'évaluation du pipeline AlphaConformer, car elle fournit des exemples concrets permettant de tester la capacité du modèle à distinguer et prédire différentes conformations. L'objectif était de rassembler un jeu de données varié, couvrant des types de protéines et de conformations aussi divers que possible, afin de garantir la robustesse et le pouvoir de généralisation du pipeline.

5.3.1 Collecte des données

La préparation des données est une étape cruciale dans tout projet de deep learning. Pour qu'AlphaFold2 puisse prédire efficacement les états apo et holo, il doit être alimenté par des données fiables, précises et bien annotées. Nous avons donc collecté et filtré des données à partir de plusieurs bases de données de référence, en nous concentrant principalement sur trois sources majeures.

La première, la Protein Data Bank *PDB*[21], est la base de données de référence mondiale pour les structures 3D de macromolécules biologiques. Elle regroupe des structures expérimentales obtenues par des techniques telles que la cristallographie aux rayons X, la cryo-microscopie électronique et la résonance magnétique nucléaire *RMN*. Les fichiers disponibles, au format .pdb, contiennent des informations détaillées sur la position des atomes et les interactions intra- et intermoléculaires.

La deuxième, UniProt[23], est une base de données centrale pour les protéines, regroupant des informations sur leurs séquences, fonctions, structures, annotations biologiques et interactions. Chaque protéine est identifiée par un identifiant unique appelé numéro d'accession UniProt, ce qui facilite le croisement des données entre différentes sources.

Et enfin, la base SIFTS *Structure Integration with Function, Taxonomy and Sequence*[24] nous a permis d'effectuer le lien entre les identifiants PDB et UniProt. Cette base, mise à jour chaque semaine, permet d'associer avec précision les structures 3D expérimentales aux séquences de référence. Une même protéine peut exister sous plusieurs conformations, chacune étant stockée dans la PDB avec un identifiant différent, mais toutes étant liées à un seul identifiant UniProt. Grâce à SIFTS, nous avons pu établir des correspondances fiables entre ces différentes structures pour identifier, pour chaque protéine, au moins une conformation apo et une conformation holo.

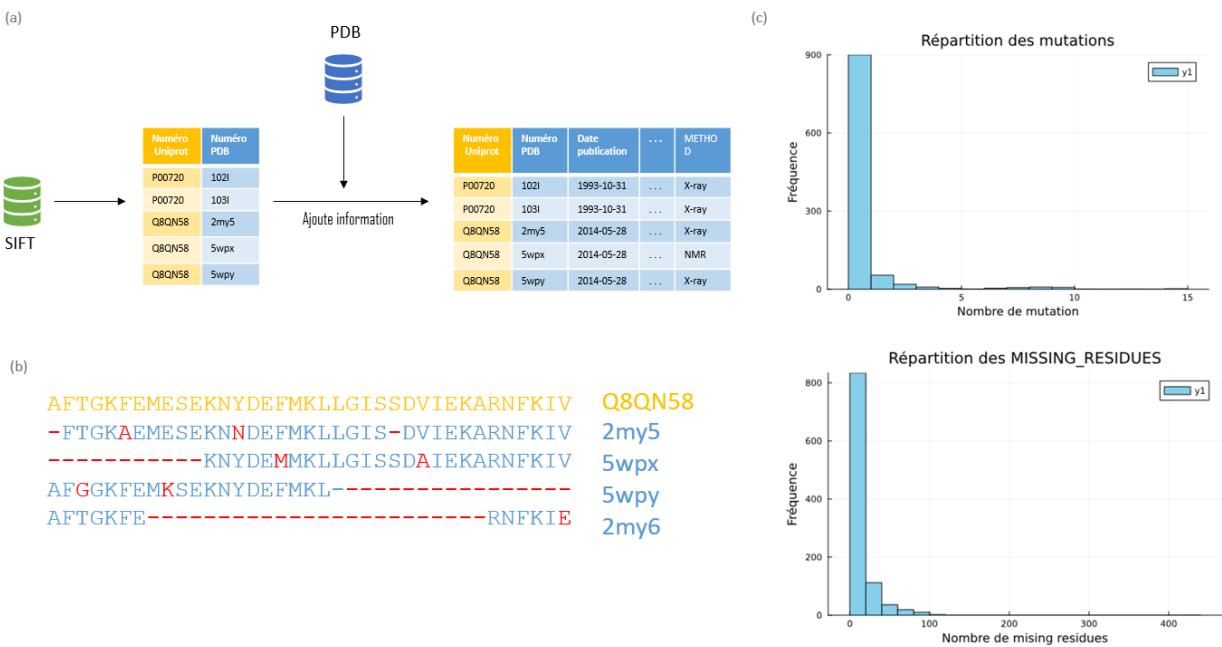


Figure 16: Collecte des données à partir des bases SIFTS[24], PDB[21] et UniProt[23]. (a) À partir de la base SIFTS, récupération des identifiants UniProt et PDB pour chaque protéine, puis extraction d'informations supplémentaires via la base de données PDB. (b) Récupération de la séquence de référence depuis UniProt, puis comparaison avec toutes les structures associées au même identifiant UniProt afin de détecter les variations. (c) Enregistrement du nombre de mutations observées ainsi que des résidus manquants pour chaque structure.

Notre premier objectif était de télécharger l'ensemble des fichiers PDB répertoriés dans la base SIFTS, puis d'en extraire les informations clés disponibles dans les fichiers PDB, notamment la date de publication, la méthode d'extraction, etc. Grâce à SIFTS, nous avons également pu établir la correspondance entre les numéros d'accession UniProt et les identifiants PDB.(voir **Figure 16-a**). Ensuite, nous avons aligné chaque séquence protéique issue des fichiers PDB avec la séquence de référence fournie par UniProt, nous avons fait cela en utilisant les informations d'appariement disponibles dans SIFTS, (voir **Figure 16-b**).

Cette étape est essentielle pour détecter d'éventuelles variations de séquence, telles que des mutations ou des résidu manquants, susceptibles d'avoir un impact sur la structure tridimensionnelle de la protéine. La majorité des protéines présentent une séquence identique à la référence UniProt, mais certaines montrent des divergences, notamment des mutations ponctuelles ou des régions incomplètes. Ces informations ont été intégrées dans notre base de données structurée afin de permettre des analyses plus complètes et précises (voir **Figure 16-c**).

Nous avons ensuite ajouté à notre base de données une colonne permettant d'identifier les structures apo et holo. Pour cela, nous avons utilisé les informations issues de la base de données BioLiP Biological Ligand–Protein Database [25], spécialisée dans l'annotation des interactions entre protéines et ligands. Cette base de données nous a permis d'ajouter une colonne indiquant les ligands présents dans chaque structure, ce qui nous a permis de distinguer clairement les conformations apo, sans ligand, et holo, avec ligand.

5.3.2 Filtrage des données

Une fois notre base de données construite, nous disposions de plus de 800 000 structures protéiques. Afin de garantir la qualité et la pertinence des données pour notre étude, nous avons appliqué plusieurs étapes de préfiltrage.

Tout d'abord, nous avons conservé uniquement les structures déterminées avec une résolution inférieure à 3 Ångströms . Cette limite correspond à un seuil de qualité communément admis dans la littérature, permettant de s'assurer que les structures obtenues par des méthodes expérimentales telles que la cristallographie aux rayons X *X-ray* ou la cryo-microscopie électronique *Cryo-EM* sont suffisamment précises et fiables. En dessous de ce seuil, les modèles atomiques sont généralement considérés comme représentatifs de la réalité biologique.

Nous avons ensuite appliqué un second filtre en ne conservant que les structures publiées après avril 2018. Cette date correspond à la période d’entraînement du modèle AlphaFold2. Ce filtre permet d’éviter que notre pipeline analyse des structures qui auraient pu être déjà apprises par AlphaFold2 lors de son développement. Ainsi, nous nous assurons que les performances d’AlphaConformer ne reposent pas sur une simple capacité de mémorisation, mais bien sur sa capacité à prédire de nouvelles conformations à partir d’informations issues des alignements multiples de séquences *MSA* et des modèles structuraux fournis *template*.

Enfin, un troisième filtre a été appliqué pour ne conserver que les numéros UniProt associés à au moins une structure apo et une structure holo, toutes deux satisfaisant les critères précédents. Cette étape était essentielle pour garantir que nous puissions comparer différentes conformations pour une même protéine.

Après application de l’ensemble de ces filtres, notre base de données a été réduite à environ 100 000 structures, soit un sous-ensemble plus restreint mais beaucoup plus pertinent et exploitable pour les objectifs de notre projet.

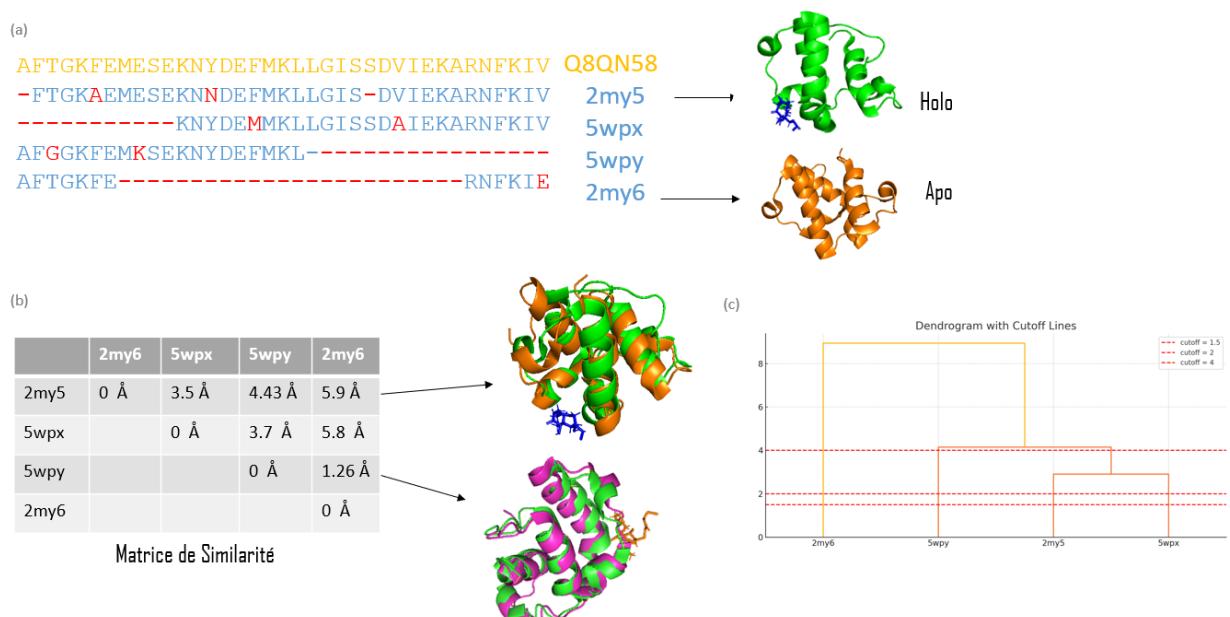


Figure 17: Analyse structurale des protéines associées à un même identifiant UniProt. (a) Alignement des séquences protéiques en fonction de la séquence de référence issue d’UniProt. (b) Comparaison des structures par calcul des distances RMSD, permettant de générer une matrice de similarité. (c) Construction d’un dendrogramme basé sur les valeurs de RMSD, suivi d’une segmentation en clusters selon différents seuils (cutoffs). L’objectif est d’identifier des conformations distinctes, en séparant les formes apo et holo dans des clusters différents.

Nous avons ensuite mené une analyse structurale approfondie pour chaque protéine retenue. Cette étape consistait à aligner l'ensemble des structures associées à un même identifiant UniProt, afin d'évaluer leurs différences conformationnelles en calculant la RMSD *Root Mean Square Deviation* pour chaque paire de structures (voir **Figure 17-a et b**). À partir de la matrice de similarité ainsi obtenue, nous avons généré un dendrogramme et appliqué différents seuils de coupure *cutoffs* : 0.5, 1, 1.5 et 2 Å pour créer des clusters de structures similaires.

L'objectif était d'identifier que les formes apo et holo appartiennent à des clusters séparés. Dans l'exemple présenté en **Figure 17-c**, la forme apo correspond à la structure 2my6, tandis que les autres sont des conformations holo. Les lignes rouges indiquent les différents cutoffs testés. Dans chaque cas, la forme apo est bien séparée des formes holo, ce qui valide la diversité conformationnelle. Nous avons ainsi conservé uniquement les identifiants UniProt satisfaisant ce critère.

L'objectif de cette étape était de sélectionner, pour chaque identifiant UniProt, une paire de structures apo et holo présentant la plus grande différence conformationnelle, afin de maximiser la pertinence de l'analyse et de garantir une distance minimale entre les deux conformations.

Dans un second temps, nous avons utilisé Foldseek pour écarter les protéines dont des structures similaires, appelé homologues, auraient pu être observés lors de l'entraînement d'AlphaFold2. Pour cela, nous avons recherché, pour chaque protéine de notre jeu de données, l'existence de structures homologues publiées avant 2018. Si un homologue était identifié par Foldseek, la protéine concernée était exclue. Foldseek retourne, pour chaque comparaison, une e-value, reflétant la probabilité que la similarité soit due au hasard, ainsi qu'un TM-score, mesurant la similarité structurelle globale.

Nous avons considéré deux protéines comme homologues si le TM-score était supérieur à 0.5 et l'e-value inférieure à 1e-5. Ce double critère permet d'assurer que seules les structures véritablement similaires sont exclues, tout en maintenant un niveau de rigueur élevé dans la sélection. Cette étape critique a permis d'éliminer les biais potentiels liés à l'entraînement d'AlphaFold2, et d'obtenir un jeu de données propre, indépendant et représentatif de cas non vus par le modèle.

À l'issue du processus de filtrage, notre base de données initiale est passée de plus de 100 000 structures à un ensemble réduit de 200 structures, correspondant à 40 identifiants UniProt uniques. Ce jeu de données final constitue désormais la référence pour la poursuite de l'axe 1 du projet SPPICES, dédié à l'étude de la diversité conformationnelle des protéines.

Dans les prochaines étapes, nous prévoyons de tester le pipeline AlphaConformer sur un échantillon issu de cette base, puis de réévaluer ses performances sur ces données spécifiques. Cela nous permettra de mesurer l’efficacité réelle de la méthode lorsque les protéines ne font partie ni de l’entraînement initial d’AlphaFold2, ni d’un ensemble de structures homologues. Ce test est donc essentiel pour évaluer les capacités de généralisation du modèle, et non simplement ses facultés de mémorisation. Nous comparerons également ses performances à celles d’AF-cluster et de BioEmu. Ces expérimentations constituent les objectifs des dernières semaines de mon stage de fin d’études, et sont déterminantes pour valider la robustesse de notre approche.

6 Retour d’expérience

6.1 Difficultés rencontrées et points clés

Ce stage de six mois a été particulièrement riche en apprentissage, tant sur le plan scientifique que technique. Les enseignements suivis au CNAM durant le premier semestre, notamment ceux portant sur le drug design, la dynamique moléculaire et le repliement des protéines, m’ont fourni des bases solides pour aborder ce stage. Ces connaissances m’ont permis de mieux saisir les enjeux liés à la prédiction des structures protéiques, en particulier la complexité de leurs changements de conformation.

L’un des principaux défis de ce stage a été de reprendre le pipeline existante, AlphaConformer, et de l’améliorer. Cela impliquait de maîtriser rapidement : un nouveau langage de programmation, Julia, assimiler les objectifs biologiques du projet, analyser la logique algorithmique du code, puis identifier des pistes concrètes d’optimisation. Ce processus nécessitait à la fois rigueur, autonomie, esprit de synthèse et une forte capacité d’adaptation.

Sur le plan technique, l’environnement de travail ajoutait une contrainte supplémentaire. Le pipeline était exécuté sur un cluster distant mis à disposition par l’i2BC. Travailler sur un cluster implique une dépendance à sa disponibilité : pannes, maintenances, files d’attente, limite de mémoire et GPU ou autres imprévus pouvaient ralentir considérablement l’avancée du projet. Il a donc fallu apprendre à s’organiser efficacement, planifier les tâches à forte consommation de ressources en dehors des heures de pointe, et surveiller en continu l’exécution des scripts.

Le traitement des données représentait également un défi majeur. Les structures protéiques manipulées sont volumineuses et leur traitement nécessite une importante puissance de calcul. Par exemple, l’exécution d’AlphaConformer sur une seule protéine prend environ 40 minutes. Entre erreurs de codage, dépassements de mémoire, et interruptions imprévues, il a été essentiel de faire preuve de réactivité et d’optimisation pour limiter les pertes de temps.

Ce stage m'a également permis de développer un ensemble de compétences variées et plus complémentaires. Sur le plan scientifique, j'ai consolidé mes connaissances en dynamique moléculaire, en prédiction de structures protéiques et en analyse fonctionnelle. Sur le plan technique, j'ai renforcé ma maîtrise de l'environnement shell via Bash, amélioré la gestion des ressources mémoire, et appris à concevoir des scripts robustes et reproductibles. J'ai également découvert et utilisé les gestionnaires de ressources SLURM et PBS. Enfin, ce stage a surtout été l'occasion de développer mon esprit critique, une compétence essentielle en recherche scientifique.

En effet, dans un projet de recherche, il ne s'agit pas seulement d'utiliser des outils existants ou d'appliquer des méthodes connues. Il faut également savoir évaluer la fiabilité des sources, questionner les résultats publiés et identifier les éventuels biais dans les approches d'autres équipes. Pour renforcer cette compétence chez les doctorants, les stagiaires et les ingénieurs de l'équipe, les chefs ont mis en place, deux fois par mois, des séances d'analyse critique d'articles scientifiques. Pendant deux heures, nous décortiquions des publications pour juger de leur robustesse : l'étude est-elle biaisée ? Les résultats sont-ils généralisables ? Les cas présentés sont-ils représentatifs ou seulement choisis parce qu'ils fonctionnent ? L'objectif de ces séances était d'apprendre à sélectionner les informations pertinentes pour nos propres travaux, sans perdre de temps à suivre des pistes fondées sur des résultats douteux.

En conclusion, ce stage m'a permis d'acquérir une expérience de terrain précieuse, entre biologie structurale, bioinformatique et développement de logiciel. J'ai appris à conduire un projet complexe dans un environnement technique exigeant, tout en développant des compétences analytiques, critiques et opérationnelles qui me seront utiles pour la suite de mon parcours professionnel.

6.2 Aspects organisationnels et gestion de projet

Pendant toute la durée de mon stage, j'ai travaillé de manière autonome sur la tâche 1A, tout en bénéficiant du soutien et des conseils de mon encadrant, Diego Zea.

Dans le domaine de la recherche, chacun est généralement responsable de son propre projet. D'un point de vue hiérarchique, Raphaël Guerois est le chef de l'équipe de recherche et supervise donc directement Diego Zea. Il est ainsi mon N+2. Toutefois, dans la pratique, le projet SPPICES est celui de Diego : il en est à l'origine, en a formulé l'hypothèse, obtenu les financements, et assure la direction scientifique. Nous avons donc collaboré en binôme tout au long du stage, sans interactions avec d'autres membres de l'équipe sur cette tâche précise.

Afin d'assurer un bon suivi du projet et une coordination efficace, nous avons mis en place une organisation inspirée de la méthode agile. Concrètement, nous fonctionnions en "sprints" de deux semaines, à l'issue desquels nous faisions une rétrospective pour identifier ce qui avait bien fonctionné, ce qui pouvait être amélioré, et définir les actions à mettre en place pour la suite. Cette approche a permis d'adapter l'organisation du travail à nos besoins respectifs, tout en assurant une progression régulière et harmonieuse du projet.

Le premier mois de stage s'est déroulé entièrement en télétravail, le temps d'obtenir mon badge d'accès au CEA. Cette période a été mise à profit pour approfondir la bibliographie, m'approprier les besoins et objectifs du projet, et me former au langage de programmation Julia. Durant cette phase, nous faisions un point quotidien d'environ 15 minutes afin de répondre à mes questions et de suivre mes avancées. Une fois les bases posées, j'ai commencé à développer le pipeline de création de la base de données, ce qui m'a rapidement confronté aux contraintes techniques du projet : traitement de données massives, gestion de la mémoire, dépendance au bon fonctionnement du cluster distant de l'I2BC.

Au fil des semaines, nous avons espacé les points de suivi à une fréquence d'environ tous les deux jours. Cela me permettait de gagner en autonomie tout en recevant un accompagnement régulier pour orienter le travail et optimiser l'utilisation des ressources. Après environ trois mois, je maîtrisais suffisamment bien les outils et le sujet pour proposer mes propres méthodes d'analyse, représentations graphiques et interprétations des résultats, ce qui a contribué à faire progresser le projet.

D'un point de vue organisationnel, le fait de commencer le stage sans planification précise des six mois m'a quelque peu déstabilisé au départ, car j'aime avoir un cadre clair d'organisation. J'ai toutefois appris à développer ma capacité d'adaptation et à faire preuve de plus de flexibilité. Avec le recul, je considère que l'approche agile s'est révélée très pertinente dans ce contexte de recherche, car elle favorise l'autonomie, l'épanouissement dans les tâches assignées, et permet d'ajuster les objectifs au fur et à mesure, en fonction des résultats, des imprévus techniques ou des disponibilités.

7 Conclusion

Le projet SPPICES *Structure Prediction and Protein Interaction by Conformational Exploration and Simulation* vise à mieux comprendre les mécanismes d’interaction entre protéines ainsi que leur repliement en structures tridimensionnelles. Mon stage de Master 2 s’est inscrit dans la sous-tâche 1A, qui a pour objectif d’améliorer les performances d’AlphaFold2[16] dans la prédiction de différentes conformations structurales des protéines, en particulier les formes apo et holo.

Dans ce cadre, mon maître de stage, Diego Zea, a développé un pipeline nommé AlphaConformer. Ce dernier repose sur l’utilisation de modèles structuraux, appelés templates, qui servent de guide pour orienter les prédictions d’AlphaFold2 vers des conformations alternatives pertinentes. Pour contribuer à ce projet, j’ai mené trois tâches principales au cours des six mois de stage.

Tout d’abord, la maîtrise du pipeline AlphaConformer, puis la comparaison de ses performances avec d’autres méthodes existantes. L’objectif était de situer AlphaConformer par rapport à l’état de l’art. Nous l’avons ainsi comparé à deux méthodes récentes développées en 2024 : AF-Cluster[30] et BioEmu[32], toutes deux capables de générer plusieurs conformations, mais sans recours aux templates. Nous avons appliqué ces trois pipelines sur un ensemble de 4 protéines sélectionnées dans la base de données de Saldaño[13], puis comparé les structures prédites aux conformations de référence, apo et holo, via des mesures de RMSD.

Ensuite, j’ai travaillé à l’amélioration du pipeline AlphaConformer. Cette étape a débuté par une phase intensive de débogage, suivie de l’optimisation de plusieurs paramètres afin de maximiser la qualité des prédictions. Après de nombreux tests, nous avons choisi d’utiliser les bases de données PDB[21] et AFDB[22] pour augmenter la diversité des templates, de ne pas filtrer les résultats produits par Foldseek[19] afin de préserver cette diversité structurale, et d’appliquer un seuil de 1 Ångströms pour le clustering des modèles structuraux.

Et enfin j’ai effectué la création d’une base de données de protéines présentant à la fois des conformations apo et holo. Un soin particulier a été apporté pour s’assurer que ces structures ne soient pas présentes dans les données d’entraînement d’AlphaFold2, afin d’évaluer les performances réelles du pipeline et non des effets de mémorisation. Le processus de sélection a nécessité l’analyse des homologues, des dates de publication et des résolutions structurales pour garantir des données pertinentes et sans biais.

La tâche 1A touche désormais à sa fin, bien qu'il reste encore quelques étapes importantes à finaliser avant sa clôture.

Dans un premier temps, il reste à finaliser la constitution de notre base de données en la divisant en trois sous-ensembles : entraînement, validation et test. Cette organisation facilitera son exploitation lors des prochaines étapes du projet SPPICES, notamment pour l'évaluation finale d'AlphaConformer. La répartition se fera en tenant compte des similarités de séquence et de structure entre les protéines, de manière à maximiser la distance entre l'ensemble test et les ensembles d'entraînement et de validation. Ce choix méthodologique vise à évaluer de façon rigoureuse la capacité de généralisation du pipeline, indépendamment de tout effet de mémorisation.

En parallèle, plusieurs axes d'amélioration restent à explorer pour optimiser les performances de la méthode. Nous poursuivrons l'analyse des paramètres restants sur l'ensemble des protéines du jeu de données. Une évolution prometteuse consisterait à substituer Foldseek par MAFFT[34] pour la génération des alignements multiples de séquences (MSA). Contrairement à Foldseek, qui aligne les structures en se basant sur leur conformation 3D, MAFFT se fonde uniquement sur les séquences d'acides aminés.

Nous poursuivrons également notre veille scientifique afin d'identifier et d'évaluer d'autres pipelines développés récemment pour la prédiction de conformations multiples.

Enfin, nous visons la rédaction et la publication d'un article scientifique synthétisant l'ensemble de notre travail : méthodologie, résultats comparatifs et conclusions sur l'impact des templates structuraux dans la prédiction de conformations alternatives via AlphaFold2. Cette contribution pourrait s'avérer précieuse pour la communauté, en apportant un éclairage nouveau sur l'intégration d'informations structurelles dans les modèles de prédiction de la flexibilité conformationnelle des protéines.

Ce stage a constitué une expérience enrichissante, tant sur le plan technique que personnel. Il m'a offert l'opportunité de consolider mes acquis tout en explorant de nouvelles compétences, dans un environnement de recherche stimulant.

Sur le plan technique, j'ai approfondi mes connaissances en biologie structurale, notamment à travers l'analyse des structures protéiques et l'étude des conformations apo et holo. J'ai renforcé ma maîtrise de plusieurs outils bioinformatiques, tels que la génération de MSA, l'utilisation d'AlphaFold2, BioEmu et AF-Cluster, ainsi que le développement d'un pipeline sur mesure. J'ai également appris à exploiter un cluster de calcul distant de manière autonome, et à concevoir des scripts efficaces pour automatiser des processus complexes, notamment dans le cadre de la création d'une base de données structurales.

Sur le plan méthodologique et personnel, ce stage m'a permis de développer des compétences transversales essentielles à toute démarche de recherche. J'ai gagné en rigueur scientifique, en capacité d'adaptation, en organisation multitâche, ainsi qu'en esprit critique. La lecture régulière d'articles scientifiques m'a aidée à mieux appréhender la littérature du domaine, à intégrer de nouveaux concepts rapidement, et à évaluer la pertinence des approches proposées dans d'autres travaux.

Cette expérience a également renforcé mon intérêt pour la biologie structurale, et plus particulièrement pour l'étude de la dynamique conformationnelle des protéines, un champ de recherche à la fois vaste, encore largement inexploré, et crucial pour la compréhension des fonctions biologiques. La place croissante de la bioinformatique dans ce domaine m'a convaincue de son potentiel, tant en termes de gain de temps que de réduction des coûts expérimentaux.

Par ailleurs, ce stage a confirmé mon désir de poursuivre dans le domaine de la recherche. Le fait de partir d'une problématique ouverte, d'avancer pas à pas grâce aux connaissances disponibles, et de construire des solutions progressivement, est une démarche qui me stimule profondément. Contribuer, même indirectement, à des avancées médicales via la recherche fondamentale représente une source de motivation forte et durable.

À la suite de ce stage, j'ai eu la chance de me voir proposer un contrat à durée déterminée de deux ans avec Diego Zea, mon encadrant, afin de poursuivre le travail engagé. Dans un premier temps, je participerai à la finalisation de la tâche 1A du projet SPPICES, en vue de la rédaction et la publication d'un article scientifique. Par la suite, je m'impliquerai dans le second axe du projet, consacré aux protéines intrinsèquement désordonnées (IDP), un domaine émergent et encore peu exploré.

Ce contrat représente une opportunité précieuse pour continuer à développer mes compétences en biologie structurale et en deep learning, tout en m'ouvrant à de nouvelles dimensions comme la gestion de projet, le management collaboratif et la documentation rigoureuse de code pour une diffusion open source.

En somme, ce stage de Master 2 a marqué une étape déterminante dans mon parcours. Il a confirmé mon orientation vers la recherche et m'a donné envie de m'investir pleinement dans des projets ambitieux à l'interface entre biologie et informatique. Je suis impatiente de poursuivre cette aventure scientifique et de découvrir les défis et les apprentissages que les mois à venir me réservent.

8 Annexes

Glossaire

acides aminés sont les petites molécules qui servent de base pour construire les protéines.

Il en existe 20 différents dans le corps humain, et ils jouent un rôle essentiel dans presque toutes les fonctions biologiques, comme la croissance, la réparation des tissus et le bon fonctionnement des cellules.

ADN (acide désoxyribonucléique) est une molécule qui contient toutes les instructions nécessaires pour construire et faire fonctionner un organisme vivant.

apo est la forme inactive d'une protéine, lorsqu'elle est dépourvue de son ligand. Dans cet état, la protéine peut avoir une conformation légèrement différente de sa forme active, car l'absence du ligand peut influencer son repliement ou sa flexibilité.

ARN (acide ribonucléique) est une molécule qui aide à transmettre et utiliser l'information contenue dans l'ADN pour fabriquer des protéines. C'est un messager essentiel dans le fonctionnement des cellules.

biologie structurale est une branche de la biologie qui étudie la forme des molécules dans les êtres vivants, en particulier les protéines et l'ADN, pour comprendre comment leur structure influence leur fonction.

conformation protéique est la forme ou la structure qu'une protéine adopte dans l'espace.

Cette forme est importante car elle détermine comment la protéine fonctionne dans le corps.

dynamique des protéines fait référence aux mouvements et changements de forme que peuvent adopter les protéines dans le temps. Ces mouvements sont souvent essentiels pour qu'elles remplissent correctement leurs fonctions biologiques.

e-value indique le nombre de résultats similaires qu'on pourrait obtenir par hasard dans une base de données. Plus l'e-value est faible, plus la similarité est significative.

familles de séquences regroupe des séquences d'ADN ou de protéines qui sont similaires entre elles parce qu'elles partagent une origine évolutive commune. Les séquences d'une même famille ont souvent des fonctions biologiques similaires. Elles peuvent avoir des variations, mais conservent des régions importantes communes.

fine-tuning est une méthode d'entraînement qui consiste à reprendre un modèle déjà pré-entraîné et à le réentraîner sur un jeu de données plus spécifique ou plus restreint pour l'adapter à une tâche précise.

génome est l'ensemble de l'information génétique d'un organisme, c'est-à-dire toutes ses instructions codées dans l'ADN, qui permettent de construire et faire fonctionner ses cellules.

holo est la forme active de la protéine, lorsqu'elle est associée à son ligand. La liaison du ligand peut induire un changement de conformation, stabilisant ainsi la structure fonctionnelle nécessaire à son activité biologique.

homologue est une protéine ou une séquence génétique qui partage une origine commune avec une autre. Deux protéines sont dites homologues si elles viennent d'un ancêtre commun, ce qui se reflète souvent par des similitudes dans leur séquence ou leur structure.

information génétique est l'ensemble des instructions contenues dans l'ADN d'un organisme, qui déterminent ses caractéristiques et permettent le bon fonctionnement de ses cellules.

interaction protéique désigne le fait que deux ou plusieurs protéines se lient ou communiquent entre elles pour accomplir une fonction dans la cellule, comme transmettre un signal ou construire une structure.

ligand est une molécule, souvent petite, qui se lie spécifiquement à une protéine, en général à un site précis, comme une clé dans une serrure.

modèles structuraux sont des représentations en trois dimensions de la forme d'une molécule, comme une protéine. Il permet de visualiser comment ses différentes parties sont organisées dans l'espace et d'étudier son fonctionnement ou ses interactions avec d'autres molécules.

mécanismes cellulaires est un processus ou une série d'actions qui se produisent à l'intérieur d'une cellule pour assurer son bon fonctionnement, comme la production d'énergie, la communication avec d'autres cellules, ou la division cellulaire.

pipeline est une chaîne ou une séquence organisée d'étapes ou de traitements automatisés, où la sortie d'une étape sert d'entrée à la suivante. Il permet de traiter des données ou d'exécuter des opérations de manière efficace, reproductible et souvent en grande quantité, en automatisant un processus complexe.

protéine est une molécule essentielle à la vie, composée d'une chaîne d'acides aminés. Elle remplit de nombreuses fonctions dans le corps, comme construire les tissus, catalyser des réactions chimiques, ou transmettre des signaux.

RMSD est une mesure utilisée pour comparer deux structures, par exemple deux protéines. c'est la distance moyenne entre les atomes de deux structures après les avoir superposées. Plus la RMSD est petite, plus les structures sont similaires.

résidu désigne un acide aminé qui fait partie de la chaîne polypeptidique. Chaque résidu est composé d'un groupe amino, d'un groupe carboxyle, et d'une chaîne latérale (R) qui est spécifique à chaque acide aminé.

signaux évolutifs sont des indices issus de la comparaison de séquences de gènes ou de protéines entre différentes espèces. Ils permettent d'identifier les parties conservées au cours de l'évolution, souvent essentielles pour la structure ou la fonction des molécules biologiques.

Ångströms Å est une unité de mesure utilisée à l'échelle atomique, équivalente à 10.e-10 mètres.

Sommaire des Figures

1	Schéma illustrant la synthèse d'une protéine.	4
2	Schéma illustrant la diversité conformationnelle des protéines	8
3	Évaluation des performances d'AlphaFold2	9
4	Premiers résultats obtenus avec le pipeline AlphaConformer.	10
5	Performances d'AlphaFold2 au CASP14 et schéma de son fonctionnement publié par Jumper et al[16]	14
6	Performances de Foldseek publié par Van Kempen et al. [19]	17
7	Pipeline AlphaConformer.	18
8	Premier résultat d'AlphaConformer sur la protéine 1AKZ.	20
9	Résultat d'AlphaConformer sur la protéine 1AKZ et 2F63 avant amélioration.	21
10	Résultats publiés par Wayment-Steele et al. pour AF-Cluster [30].	23
11	Résultats d'AF-Cluster appliqués à la base de données de Saldaño [13].	24
12	Résultats publiés par Lewis et al. sur BioEmu[32].	25
13	Résultats de BioEmu appliqués au jeu de données de Saldaño [13].	26
14	Comparaison des performances d'AlphaConformer, AF-cluster et BioEmu .	28
15	Test des différents paramètres d'AlphaConformer	29
16	Collecte des données à partir des bases SIFTS, PDB et UniProt.	31
17	Analyse structurale des protéines associées à un même identifiant UniProt. .	33

9 References

- [1] The Realm of Unconventional Noncovalent Interactions in Proteins: Their Significance in Structure and Function - 2023 - Adhav <https://pubs.acs.org/doi/10.1021/acsomega.3c00205>.
- [2] Conformational selection in protein binding and function - 2014 - Weik <https://arxiv.org/pdf/1409.2584.pdf>.
- [3] I2BC. Institut de Biologie Intégrative de la Cellule <https://www.i2bc.paris-saclay.fr/>.
- [4] CEA Joliot. I2BC https://joliot.cea.fr/drive/joliot/en/Pages/research_entities/I2BC_saclay.aspx.
- [5] CNRS - Centre National de la Recherche Scientifique <https://www.cnrs.fr/fr/nos-recherches/disciplines>.
- [6] Université Paris-Saclay. Institute of Integrative Cell Biology (I2BC) <https://www.universite-paris-saclay.fr/en/laboratories/institut-de-biologie-integrative-de-la-cellule-i2bc-0>.
- [7] PluginLabs Paris-Saclay. Institute of Integrative Cell Biology (I2BC) <https://www.pluginlabs-universiteparissaclay.fr/en/fiche/institute-of-integrative-cell-biology-i2bc/>.
- [8] Nature Index. Institute of Integrative Biology of the Cell (I2BC) <https://www.nature.com/nature-index/institution-outputs/france/institute-of-integrative-biology-of-the-cell-i2bc/549cf3ca140ba07f7e8b4567>.
- [9] Société Française de Virologie. I2BC <https://sfv-virologie.org/the-institute-for-integrative-biology-of-the-cell-i2bc-is-seeking-to-recruit-jun>
- [10] I2BC- Equipe OCHSENBEIN/Guerois <https://www.i2bc.paris-saclay.fr/molecular-assemblies-and-genome-integrity/>.
- [11] Easy Not Easy: Comparative Modeling with High-Sequence Identity Templates – 2023 - Zea <https://sciwheel.com/work/citation?ids=16225439&pre=&suf=&sa=0>.
- [12] ConTemplate Suggests Possible Alternative Conformations for a Query Protein of Known Structure – 2015 - Narunsky [https://www.cell.com/structure/fulltext/S0969-2126\(15\)00368-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0969212615003688%3Fshowall%3Dtrue](https://www.cell.com/structure/fulltext/S0969-2126(15)00368-8?_returnURL=https%3A%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0969212615003688%3Fshowall%3Dtrue).

- [13] Impact of protein conformational diversity on AlphaFold predictions – 2022 - Saldaño
<https://academic.oup.com/bioinformatics/article/38/10/2742/6563595>.
- [14] BioStructures.jl: read, write and manipulate macromolecular structures in Julia - 2020 - Greener
<https://pubmed.ncbi.nlm.nih.gov/32407511/>.
- [15] MIToS.jl: mutual information tools for protein sequence analysis in the Julia language - 2017 - Zea
<https://pubmed.ncbi.nlm.nih.gov/27797756/>.
- [16] Highly accurate protein structure prediction with AlphaFold – 2021 - Jumper
<https://www.nature.com/articles/s41586-021-03819-2>.
- [17] From interaction network to interfaces, scanning intrinsically disordered regions using AlphaFold2 – 2024 - Zea
<https://www.nature.com/articles/s41467-023-44288-7>.
- [18] Easy and accurate protein structure prediction using ColabFold - 2023 - Gyuri
<https://par.nsf.gov/biblio/10537802-easy-accurate-protein-structure-prediction-us>
- [19] Fast and accurate protein structure search with Foldseek – 2024 – Van Kempen
<https://www.nature.com/articles/s41587-023-01773-0>.
- [20] MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets – 2017 – Martin Steinegger & Johannes Söding
<https://www.nature.com/articles/nbt.3988>.
- [21] Base de donnée PDB
<https://www.rcsb.org/>.
- [22] Base de donnée de AlphaFold nommé AFDB
<https://alphafold.ebi.ac.uk/>.
- [23] Base de donnée Uniprot
<https://www.uniprot.org/>.
- [24] Base de donnée SIFTS
<https://www.ebi.ac.uk/pdbe/docs/sifts/>.
- [25] Base de donnée BioLiP
<https://zhanggroup.org/BioLiP/index.cgi>.
- [26] Alignement de structure protéique : US-align
<https://colab.ws/articles/10.1038/s41592-022-01585-1>.

- [27] Algorithme de clustering Hobohm
<https://metacpan.org/pod/String::Cluster::Hobohm>
- [28] Algorithme de clustering DBSCAN
<https://datascientest.com/machine-learning-clustering-dbscan>
- [29] Kern équipe
<https://kernlab-scripps.github.io/>
- [30] Predicting multiple conformations via sequence clustering and AlphaFold2 - 2023 - Wayment-Steele
<https://www.nature.com/articles/s41586-023-06832-9>
- [31] AlphaFold predictions of fold-switched conformations are driven by structure memo-
rization – 2024 – Chakravarty
<https://www.nature.com/articles/s41467-024-51801-z#Abs1>
- [32] Scalable emulation of protein equilibrium ensembles with generative deep learning -
2024 - Sarah Lewis
<https://www.biorxiv.org/content/10.1101/2024.12.05.626885v2.full>
- [33] AlphaFold predictions of fold-switched conformations are driven by structure memo-
rization – 2024 – Chakravarty
<https://www.nature.com/articles/s41467-024-51801-z#Abs1>
- [34] Algorithme d'alignement de séquence MAFFT
<https://mafft.cbrc.jp/alignment/server/index.html>