Prédiction de la structure des protéines par l'intelligence artificielle : état de l'art, validations et perspectives

Auteur : Grégoire Porteix — Mémoire de Master — Juillet 2025

Résumé

La prédiction de la structure tridimensionnelle des protéines à partir de leur séquence est un défi majeur en biologie structurale. Avec l'avènement d'AlphaFold, DeepMind a révolutionné ce champ, atteignant des précisions proches de celles de la cristallographie. Ce mémoire synthétise sept articles clés traitant de l'évolution des approches IA appliquées à la biologie structurale, incluant les progrès technologiques, les évaluations méthodologiques et les implications scientifiques. Nous présentons les structures algorithmiques des principaux modèles, leurs performances dans les compétitions telles que CASP, leurs limitations (protéines désordonnées, multimériques, MSA absents) et les perspectives à venir, notamment en conception de protéines et médecine de précision. Ce mémoire aborde des termes complexes. J'ai donc pris l'initiative d'y inclure un lexique à la fin.

Introduction

Depuis plus de cinquante ans, prédire la structure tridimensionnelle d'une protéine à partir de sa seule séquence d'acides aminés constitue un défi scientifique fondamental. Ce mémoire propose une revue critique des méthodes, résultats et perspectives fondés sur sept publications majeures sélectionnées pour leur impact sur le domaine entre 2019 et 2024.

La compréhension tridimensionnelle des protéines est considérée comme une clé essentielle pour décrypter le vivant. En biologie, la structure détermine la fonction. Cette équation est au cœur de la biologie structurale moderne. Toutefois, accéder à cette structure nécessite des techniques expérimentales lourdes, lentes et coûteuses, telles que la cristallographie aux rayons X, la cryo-microscopie électronique ou la résonance magnétique nucléaire. Bien que ces méthodes aient produit des avancées majeures, elles laissent encore un vaste pan de l'espace protéique inexploré.

C'est dans ce contexte qu'intervient une révolution conceptuelle et technologique : l'application de l'intelligence artificielle à la prédiction de la structure des protéines. Longtemps jugée inatteignable par approche purement computationnelle, cette tâche a été spectaculairement résolue en 2020 avec l'arrivée d'AlphaFold2, développé par DeepMind. Ce modèle a bouleversé les standards existants en atteignant des précisions comparables à celles des techniques expérimentales sur de nombreuses cibles. Son succès repose sur une combinaison novatrice de représentations évolutives (MSA)[1], de réseaux neuronaux de type Transformer [2], et d'architectures géométriques spécialisées.

Le bouleversement ne s'arrête pas à la performance technique. La mise à disposition publique du code source, des poids du modèle, et surtout d'une base de données complète (AlphaFold Protein Structure Database, AFDB), a permis une démocratisation sans précédent de l'accès à la structure protéique. Des millions de structures auparavant inconnues sont désormais accessibles librement, facilitant la recherche, l'enseignement, et la médecine.

Ce mémoire propose une synthèse critique des progrès récents dans ce domaine, à partir de sept articles scientifiques majeurs. Il s'intéresse à la fois aux fondements techniques du modèle AlphaFold et de ses dérivés, à leurs performances évaluées lors de benchmarks internationaux, à leurs multiples applications en biologie et médecine, ainsi qu'aux limites et perspectives d'avenir.

L'analyse portera également sur les bases de données nécessaires au fonctionnement de ces modèles, les scores de confiance associés à leurs prédictions, et les comparaisons avec des approches alternatives comme RoseTTAFold, ESMFold ou OmegaFold. Nous chercherons à comprendre dans quelle mesure ces outils peuvent s'intégrer dans les flux de travail cliniques, les laboratoires de recherche ou les initiatives de santé publique à grande échelle.

Au-delà des chiffres, il s'agit de poser les bases d'une réflexion sur l'émergence d'une nouvelle discipline : la biologie structurale assistée par l'intelligence artificielle. Cette discipline hybride conjugue les forces de la modélisation, de l'apprentissage automatique, de la bioinformatique et de la biologie moléculaire pour proposer un cadre unifié d'exploration des structures du vivant. Elle ouvre la voie à des avancées majeures en matière de compréhension fonctionnelle, de design thérapeutique, de classification évolutive, et de diagnostic de précision.

Ainsi, la prédiction structurale n'est plus seulement une prouesse algorithmique. Elle devient un pilier central de la biologie moderne, à l'interface entre les données, la machine, et l'humain.

1. Implémentation algorithmique et bases de données

Le socle technique de la prédiction structurale par intelligence artificielle repose sur des architectures neuronales profondes capables d'extraire, d'encoder puis de reconstituer la géométrie spatiale tridimensionnelle des protéines à partir d'une simple séquence d'acides aminés.

Le modèle AlphaFold2, présenté par Jumper et al. (2021), constitue une avancée majeure grâce à l'utilisation de trois blocs fondamentaux : un encodeur de multiples alignements de séquences (MSA), une représentation graphique des paires de résidus, et un module de repliement structurel itératif appelé « structure module ». Cette architecture permet une rétroaction continue entre les informations évolutives, les relations de distance et les prédictions spatiales.

L'article de Xu (2019) propose quant à lui une méthode antérieure fondée sur des réseaux de résidus à base de distances (distance maps)[3], ouvrant la voie à l'encodage topologique des chaînes protéiques via des contraintes géométriques. Bien que moins performant qu'AlphaFold2, ce type de représentation constitue une brique importante dans la genèse des modèles modernes.

L'architecture d'AlphaFold2 est décrite comme une chaîne pipeline[4] qui combine apprentissage auto-supervisé, attention bidirectionnelle multi-tête et intégration itérative des données. Selon Kovalevskiy et al. (2024), cette organisation algorithmique permet une prédiction dynamique, dans laquelle la structure 3D est affinée à chaque cycle en fonction des mises à jour internes du modèle.

Les modèles alternatifs tels que RoseTTAFold, OmegaFold et ESMFold, décrits dans Jänes et Beltrao (2024), diffèrent par leur traitement des alignements : certains s'appuient fortement sur les MSA, d'autres utilisent des représentations monomodales de type Transformer (ex. : ESM2). ESMFold, par exemple, élimine totalement le recours aux MSA et produit des prédictions directement à partir de la séquence brute en une seule passe, grâce à un langage moléculaire entraîné sur des milliards de protéines.

Du point de vue de l'infrastructure, la base de données AlphaFold Protein Structure Database (AFDB), détaillée dans Tunyasuvunakool et al. (2024), héberge plus de 214 millions de structures couvrant Uniprot, Ensembl, MGnify et d'autres bases biologiques. Chaque structure y est accompagnée d'un score pLDDT (predicted Local Distance Difference Test)[5], d'une estimation d'erreur (PAE)[6], et de métadonnées

fonctionnelles. La base s'appuie sur une architecture redondante haute performance avec hébergement chez EMBL-EBI, couplée à une API REST[7], une interface Web et une capacité d'interrogation en masse[8].

Enfin, Noé et al. (2020) rappellent que ces systèmes bénéficient de la convergence entre apprentissage profond, dynamique moléculaire et modélisation probabiliste. L'interopérabilité des données structurales, la compatibilité avec les banques PDB et l'intégration dans les pipelines de drug discovery constituent désormais des standards incontournables de l'architecture bioinformatique contemporaine.

2. Avancées techniques dans la prédiction structurale

Depuis la publication du modèle AlphaFold2, les avancées dans le domaine de la prédiction structurale ont connu une accélération remarquable. Le travail de Jumper et al. (2021) a introduit une architecture fondée sur un système de rétropropagation en boucle fermée[8], permettant une co-évolution entre les représentations de paires de résidus et les représentations issues des alignements multiples. Cette méthode a surpassé toutes les précédentes lors de la compétition CASP14[9], atteignant un score GDT_TS moyen supérieur à 90[10] pour les protéines cibles bien résolues.

Kovalevskiy et al. (2024) insistent sur l'importance des mécanismes de mise à jour itérative et du modèle structurel récurrent d'AlphaFold. Cette innovation repose sur une architecture de type Transformer multi-têtes combinée à un module de prédiction structurelle, le tout entraîné à l'aide d'un jeu massif de structures expérimentales (PDB) et de séquences alignées (MSA). Le processus de « recycling » permet d'itérer sur la prédiction jusqu'à obtention d'un état géométrique cohérent et stable, optimisé pour minimiser l'erreur PAE entre les paires de résidus[11].

Parallèlement, d'autres modèles ont émergé pour répondre à certaines limites d'AlphaFold. RoseTTAFold[12], combine trois flux d'information (séquence, MSA, distance) dans un même réseau attentionnel[13]. ESMFold, quant à lui, représente une rupture méthodologique : il ne dépend plus d'alignements multiples, mais utilise un modèle de langage de type Transformer pré-entraîné sur des centaines de millions de séquences protéiques (Chen et al., 2023). Cette approche permet une prédiction rapide, généralisable, mais légèrement moins précise sur les cibles complexes.

Jänes et Beltrao (2024) décrivent aussi les efforts visant à intégrer des éléments de dynamique moléculaire (folding pathway) ou d'énergie physique dans les architectures neuronales, afin de capturer non seulement l'état stable mais aussi les états intermédiaires. Des méthodes telles que RGN2 ou OmegaFold[14] tentent de combiner rapidité et précision sans dépendre des alignements MSA, en exploitant des représentations implicites de structure latente.

Ces évolutions techniques répondent aux critiques de Noé et al. (2020), qui soulignaient l'importance de relier les prédictions à des principes physico-chimiques et aux trajectoires dynamiques. Elles ouvrent la voie à une nouvelle génération de modèles hybrides, capables d'exploiter à la fois l'apprentissage statistique et les lois de la mécanique moléculaire.

Une avancée remarquable dans ces modèles est l'implémentation de scores de confiance intégrés. Jumper et al. ont introduit le pLDDT (Predicted Local Distance Difference Test), un score local indiquant la fiabilité de la prédiction pour chaque résidu. Ce score est désormais un standard dans les bases de données de structures prédictives, permettant aux chercheurs d'évaluer la pertinence locale d'un repliement. En complément, l'erreur PAE (Predicted Aligned Error) offre une vision globale des incertitudes entre domaines protéiques, essentielle dans l'analyse des interactions intramoléculaires et de

la flexibilité.

D'après Tunyasuvunakool et al. (2024), ces innovations ont été intégrées dans la base de données AFDB, dont l'interface propose une visualisation interactive colorée selon le pLDDT, ainsi que des cartes PAE exportables. Cette standardisation facilite le tri, l'analyse comparative et la génération automatisée de rapports structuraux, notamment pour les bioinformaticiens et les pharmacologistes.

Par ailleurs, les réseaux de neurones convolutionnels profonds[14] ont été progressivement remplacés par des architectures de type Transformer auto-attentives[15]. Cette transition technologique permet une capture plus fine des relations à longue distance au sein des chaînes protéiques, cruciale pour modéliser correctement les repliements complexes. Chen et al. (2023) ont montré que les performances de ces modèles dépendaient non seulement du volume de données d'entraînement, mais aussi de la diversité structurale lors de l'entrainement, soulignant la nécessité d'un apprentissage rigoureux.
Les approches sans MSA, comme ESMFold ou OmegaFold, sont particulièrement prometteuses dans des contextes où peu de données homologues existent. Elles offrent aussi une réduction drastique du temps de prédiction, passant de plusieurs minutes à quelques secondes par structure.

Enfin, l'intégration croissante de graphes, évoquée par Noé et al. (2020), ouvre des perspectives nouvelles. Des graph neural networks[15] permettent de simuler des interactions physiques ou de prédire des états intermédiaires dans des environnements contraints. Ces méthodes hybrides pourraient à terme rapprocher les prédictions purement statistiques de simulations atomiques réalistes, en vue d'une application clinique directe, notamment dans le domaine du repliement assisté par chaperonnes, des pathologies de conformation ou du design de protéines thérapeutiques sur mesure.

3. Évaluations et benchmarks

L'évaluation des performances des modèles de prédiction de structure protéique est essentielle pour valider leur fiabilité et leur utilité en recherche. Le principal cadre de validation est le concours international CASP (Critical Assessment of protein Structure Prediction), qui a lieu tous les deux ans. Lors de la 14^{e} édition (CASP14, 2020), AlphaFold2 a surpassé l'ensemble des participants, atteignant un score GDT_TS moyen de 92.4 pour les cibles difficiles, soit un niveau proche de la résolution expérimentale. Ce résultat, rapporté par Jumper et al. (2021), marque un tournant : la prédiction computationnelle est devenue compétitive face à la cristallographie ou la cryo-EM.

Le score GDT_TS (Global Distance Test Total Score) mesure la concordance entre la structure prédite et la structure réelle, en prenant en compte plusieurs seuils de distance entre atomes. Un GDT_TS > 90 indique généralement une très bonne prédiction. En complément, d'autres indicateurs sont utilisés : le RMSD (Root Mean Square Deviation)[16], le pLDDT (score de confiance local), et la PAE (Predicted Aligned Error). Kovalevskiy et al. (2024) insistent sur la robustesse du modèle AlphaFold dans la prédiction de structures de novo, incluant des protéines sans homologue connu.

Les modèles alternatifs comme RoseTTAFold ont également été testés dans le cadre de CASP. Ils obtiennent de bons scores sur certaines cibles, mais demeurent globalement en deçà d'AlphaFold2. Les approches basées sur les modèles de langage (ESMFold, OmegaFold) ont été évaluées plus récemment, notamment via la base CAMEO (Continuous Automated Model Evaluation)[17], un système de benchmarking continu en ligne. ESMFold, selon Chen et al. (2023), montre une précision compétitive sur les régions bien conservées mais chute pour les

structures atypiques ou sans alignement homologues. C'est par eilleurs un paradoxe étant donné qu'il n'utilise pas le MSA. Cela s'explique par son entrainement plus vaste que les autres modèles.

La base de données AFDB elle-même propose des indicateurs intégrés pour chaque structure : le score pLDDT (variant de 0 à 100) permet une estimation fine de la fiabilité locale. Une couleur bleue (> 90) indique une prédiction de haute confiance, le jaune et l'orange signalent des régions modérément prédictibles, et le rouge (< 50) reflète des structures peu fiables. Tunyasuvunakool et al. (2024) ont montré que ce score est corrélé avec les résidus rigides et bien structurés dans les expériences cristallographiques.

L'analyse comparative des performances repose aussi sur les temps de calcul et la consommation de ressources. ESMFold, utilisant un passage unique de Transformer préentraîné, peut prédire des structures en quelques secondes, contre plusieurs minutes pour AlphaFold2, notamment sur les architectures GPU non parallélisées[18]. Noé et al. (2020) soulignent la nécessité d'une évaluation coût-efficacité intégrée, essentielle dans des pipelines à haut débit (ex : criblage structural, annotation fonctionnelle à large échelle).

Xu (2019), avec son approche fondée sur les cartes de distances, propose une autre méthode d'évaluation : comparer la topologie prédite aux conformations expérimentales via des métriques géométriques (contacts natifs, score TM, distance Cα)[19]. Ce type d'analyse structurelle est particulièrement utile pour évaluer la préservation des domaines fonctionnels dans des prédictions longues ou flexibles.

Enfin, la communauté scientifique propose désormais des métabenchmarks intégrant plusieurs dimensions : précision atomique, fidélité topologique, conservation évolutive, cohérence stérique, temps de prédiction, et utilisabilité biologique. Ces outils favorisent une évaluation complète et multi-échelle, indispensable pour guider le choix du modèle en fonction de l'application visée (diagnostic, design, annotation, exploration phylogénétique). Jänes et Beltrao (2024) proposent une normalisation des métriques, notamment dans l'évaluation des états multi-chaînes et des complexes protéiques, encore peu couverts par les modèles actuels.

En conclusion, les benchmarks révèlent non seulement la supériorité d'AlphaFold2 sur des cibles classiques, mais mettent aussi en lumière les zones grises : régions désordonnées, interfaces protéine-protéine, prédictions sans MSA. L'évaluation est donc un enjeu méthodologique central, appelé à évoluer avec la sophistication croissante des modèles et la diversité des tâches biologiques adressées.

Outre les évaluations standardisées dans CASP et CAMEO, plusieurs initiatives communautaires ont été mises en place pour évaluer l'applicabilité fonctionnelle des prédictions IA. Des tests croisés sur des protéines de fonction connue mais structure inconnue ont été réalisés, comparant la compatibilité entre prédiction structurale et annotations biologiques (sites actifs, sites de liaison, domaines fonctionnels). Les résultats ont montré une bonne corrélation, notamment pour les enzymes et les récepteurs membranaires, bien que des écarts subsistent dans les régions périphériques flexibles. Cela met en lumière la nécessité de combiner évaluation géométrique et évaluation fonctionnelle dans les futurs benchmarks intégrés.

Enfin, un enjeu central est l'harmonisation des benchmarks. Les bases utilisées, les types de cibles, la méthodologie de comparaison, le traitement des régions disordonnées ou des boucles non résolues varient encore beaucoup entre les articles. Une proposition, selon Chen et al. (2023), serait de développer un standard communautaire (Benchmark Structural Ontology) avec des niveaux d'exigence (basic, intermediate, expert) permettant de situer objectivement les performances des modèles en fonction de leur usage envisagé. Cela permettrait une meilleure comparabilité entre outils et une intégration plus cohérente dans les pipelines industriels ou cliniques.

4. Applications en recherche biomédicale

Les applications biomédicales des modèles de prédiction structurale par intelligence artificielle sont nombreuses. Elles touchent à la fois la recherche fondamentale, le diagnostic moléculaire, la pharmacologie, et la médecine personnalisée. Grâce à la précision des prédictions d'AlphaFold2 et de ses successeurs, il est désormais possible d'anticiper la structure tridimensionnelle de protéines humaines inconnues, de cartographier les effets structuraux des mutations génétiques, et de concevoir des agents thérapeutiques sur mesure.

Selon Jumper et al. (2021), AlphaFold2 a permis de modéliser la quasi-totalité du protéome humain, avec un haut niveau de confiance (pLDDT > 90) pour plus de 36 % des résidus. Cette avancée représente un bouleversement dans l'annotation fonctionnelle des protéines, notamment pour les familles peu étudiées. En parallèle, la base AFDB offre un accès libre aux structures de milliers d'agents pathogènes, ouvrant la voie à des stratégies vaccinales et antivirales rapides (Tunyasuvunakool et al., 2024).

Une des premières retombées biomédicales majeures a été observée durant la pandémie de COVID-19. Les structures des protéines du SARS-CoV-2, comme la NSP6 ou l'ORF3a[20], ont été rapidement modélisées grâce à AlphaFold, permettant d'anticiper leurs mécanismes d'ancrage membranaire et leurs interactions. Ce travail a été réutilisé pour le criblage virtuel de molécules inhibitrices ciblant ces régions transmembranaires, comme le rappellent Kovalevskiy et al. (2024).

Dans le domaine de la génétique humaine, les prédictions de structure permettent aujourd'hui d'analyser les effets pathogènes de mutations ponctuelles. Un changement d'acide aminé dans une région à forte contrainte stérique ou à haute densité de contact peut être interprété comme délétère, même en l'absence de données fonctionnelles. Cette approche est notamment explorée par les auteurs comme Jänes et Beltrao (2024) pour des maladies rares, le cancer, ou des pathologies neurodégénératives. Les cartes PAE[21] et les régions désordonnées identifiées par AlphaFold permettent aussi de mieux interpréter les effets d'interactions faibles ou transitoires entre partenaires moléculaires.

Les techniques de drug design bénéficient également de ces avancées. La connaissance précise de la cavité de liaison, de la flexibilité locale, et des interactions hydrophobes offre un avantage considérable pour la conception de petites molécules ou de peptides thérapeutiques. Chen et al. (2023) rapportent que les outils dérivés d'AlphaFold sont désormais intégrés à de nombreux pipelines industriels, notamment pour la redécouverte de médicaments (repurposing), l'optimisation de l'affinité ligand-protéine, et la conception de protéines thérapeutiques de novo (ex. : mini-protéines, anticorps, protéines de fusion).

Des projets pilotes ont utilisé ESMFold pour annoter en temps réel des mutations issues de séquençages cliniques. Grâce à la rapidité de prédiction de ce modèle, il est possible d'intégrer la structure attendue d'une protéine mutée dans une plateforme de diagnostic de pathologies héréditaires. Xu (2019) avait déjà souligné que les approches à base de cartes de distances pouvaient détecter des altérations topologiques locales critiques, y compris dans des domaines apparemment conservés.

Les approches de dynamique moléculaire intégrée, comme décrites par Noé et al. (2020), ouvrent de nouvelles perspectives dans la prédiction des effets allostériques, la flexibilité domainale, ou encore la réponse à des modifications post-traductionnelles. En simulant plusieurs états structuraux possibles, ces modèles pourraient être utilisés pour prédire les effets d'un environnement cellulaire donné (pH, ions, partenaires) sur la structure adoptée.

À terme, la combinaison des modèles de prédiction structurelle, des bases de données omiques, et de l'IA explicable pourrait déboucher sur des outils d'aide à la décision clinique. Dans un contexte de médecine de précision, les médecins pourraient s'appuyer sur une interface structurale enrichie pour mieux classer les variants génétiques, orienter un traitement ciblé, ou prédire l'émergence de résistances. Le potentiel d'impact sur le système de santé est considérable, comme le soulignent les auteurs de tous les articles analysés dans ce mémoire.

Une application en plein essor est la cartographie structurale des interactomes[22]. Grâce aux prédictions IA, il devient possible d'anticiper les interfaces potentielles entre protéines au sein de réseaux d'interactions cellulaires. Cette approche, soutenue par les données de prédiction d'AlphaFold-Multimer, permet de générer des modèles complexes à grande échelle. Jänes et Beltrao (2024) mentionnent que ces modèles sont utilisés pour reconstituer des complexes entiers, comme le ribosome, le spliceosome ou des machineries de réparation de l'ADN, avec un niveau de détail auparavant réservé à la cryo-microscopie électronique.

Dans le domaine de l'oncologie, la capacité à prédire la structure des variants somatiques retrouvés dans les tumeurs représente une avancée majeure. De nombreux cancers présentent des mutations affectant la conformation, la stabilité ou la dynamique des protéines clés (TP53, KRAS, EGFR)[23]. Les modèles AlphaFold peuvent simuler les effets structurels de ces mutations et aider à distinguer les mutations silencieuses des variants fonctionnellement perturbateurs. Kovalevskiy et al. (2024) suggèrent que cette approche pourrait devenir un outil de tri clinique des variants d'importance inconnue (VUS), aujourd'hui problématique en oncogénétique.

En infectiologie, les protéines virales, bactériennes ou parasitaires peuvent être modélisées pour identifier des épitopes vaccinaux ou des inhibiteurs d'enzymes clés. AlphaFold a permis de prédire des structures virales de pathogènes émergents comme Nipah, Lassa ou Marburg[24], facilitant la conception de vaccins à base de protéines recombinantes. Tunyasuvunakool et al. (2024) notent que plus de 100 000 structures de pathogènes ont été modélisées en 2023, dont beaucoup issues de familles encore non structurés expérimentalement.

La neurologie bénéficie aussi des prédictions IA, en particulier pour les protéines associées aux maladies neurodégénératives comme Alzheimer (APP, Tau), Parkinson (alpha-synucléine), ou SLA (TDP-43). Bien que certaines soient désordonnées ou en agrégation, les segments structurés modélisables permettent d'étudier les régions d'initiation de la pathologie, ou les interfaces avec les chaperonnes. Noé et al. (2020) évoquent des simulations IA couplées à la dynamique moléculaire pour mieux comprendre l'amorçage de l'agrégation amyloïde.

La bio-ingénierie est un autre domaine révolutionné par la modélisation structurale IA. Les biotechnologistes peuvent aujourd'hui concevoir des enzymes modifiées, des protéines de fusion ou des senseurs moléculaires avec des contraintes structurales intégrées. L'utilisation combinée de la prédiction AlphaFold et de design inverse (via des outils comme ProteinMPNN ou RFdiffusion)[25] permet d'optimiser la thermostabilité, la solubilité ou la sélectivité enzymatique. Chen et al. (2023) illustrent cela par le design de nouvelles ligases ubiquitine E3 synthétiques à fonctions ciblées.

Dans le domaine des maladies rares, les prédictions AlphaFold permettent d'explorer des mutations sur des protéines sans structure résolue, ce qui est fréquent dans ces pathologies. Les médecins et chercheurs peuvent visualiser les effets structurels d'un changement d'acide aminé, identifier des régions critiques ou proposer une reclassification des variants. L'intégration de ces données structurales dans les bases de données cliniques comme ClinVar ou LOVD est en cours, et des algorithmes de classification automatisée émergent, alimentés par ces prédictions tridimensionnelles.

Enfin, l'éducation et la formation bénéficient

également de ces avancées. Les modèles structurels prédits sont désormais utilisés comme supports pédagogiques dans les formations en biochimie, biologie moléculaire, pharmacologie ou médecine. Grâce à leur accessibilité via des interfaces comme Mol*, AlphaFold DB ou iCn3D, les étudiants peuvent manipuler, comprendre et explorer des structures complexes dès les premières années d'études biomédicales.

En résumé, les domaines d'application biomédicale des prédictions IA couvrent un spectre très large, depuis la recherche fondamentale jusqu'au soin clinique. Leur intégration dans les flux de travail quotidiens en laboratoire, hôpital, industrie ou salle de classe transforme profondément la manière dont la structure des protéines est perçue, utilisée et valorisée.

5. Limites actuelles et perspectives

Malgré les progrès spectaculaires réalisés ces dernières années, les modèles de prédiction structurale basés sur l'intelligence artificielle présentent encore plusieurs limites majeures, tant sur le plan algorithmique que biologique. La compréhension de ces limites est essentielle pour orienter les futurs développements, affiner leur utilisation dans des contextes cliniques, et mieux anticiper leurs zones d'échec.

Une des limitations les plus largement reconnues est la difficulté à prédire correctement les structures des protéines dites « désordonnées » ou intrinsèquement désorganisées (IDPs). Ces protéines ou segments protéiques, qui représentent une part non négligeable du protéome eucaryote, ne possèdent pas de conformation stable unique dans des conditions physiologiques. AlphaFold2, comme le notent Kovalevskiy et al. (2024), tend à générer des structures rigides même pour des régions désordonnées, en attribuant des scores de confiance faibles mais sans capturer la nature dynamique et contextuelle de ces conformations. Or, ces régions sont souvent cruciales dans les mécanismes d'interaction transitoire, les voies de signalisation cellulaire, ou la reconnaissance moléculaire flexible.

Une autre limite concerne les complexes multi-protéiques et les structures oligomériques. Si des extensions telles qu'AlphaFold-Multimer ont été développées, les performances restent inégales. Jumper et al. (2021) soulignent que la prédiction d'interfaces protéine-protéine, en particulier dans des cas asymétriques ou allostériques, reste une tâche difficile. Les interfaces faibles ou régulées contextuellement (phosphorylation, présence de cofacteurs) sont souvent mal modélisées. De même, les structures de très grands complexes (>5 chaînes) sont hors de portée pratique, faute de mémoire ou de données d'entraînement pertinentes.

Le besoin d'un multiple sequence alignment (MSA) de qualité pour les modèles classiques constitue aussi un facteur limitant. Les protéines peu conservées, récemment apparues, ou issues de clades sous-représentés dans les bases génomiques, génèrent des prédictions de moins bonne qualité. Jänes et Beltrao (2024) insistent sur le biais de couverture : les régions du vivant sur-représentées dans les bases d'entraînement bénéficient de meilleures prédictions, tandis que les protéines de virus rares, d'organismes extrêmophiles ou de microbiotes peu explorés restent mal prédites.

Des efforts ont été menés pour pallier cette dépendance, notamment avec les modèles sans MSA comme ESMFold ou OmegaFold. Toutefois, comme le notent Chen et al. (2023), ces modèles, bien que plus rapides et indépendants des bases d'alignement, souffrent encore d'une perte de précision sur les cibles complexes, notamment les architectures non globulaires ou les protéines à repliement couplé à la liaison (coupled folding and binding).

Au-delà des limites techniques, se pose la question de l'interprétabilité. Les modèles actuels sont souvent décrits comme des « boîtes noires »

dont les prédictions, aussi précises soient-elles, ne sont pas toujours accompagnées d'une justification mécanistique. Cela soulève des enjeux méthodologiques et éthiques, notamment dans un contexte clinique où une mauvaise interprétation peut induire un biais diagnostic ou thérapeutique. Noé et al. (2020) proposent de coupler les prédictions neuronales à des systèmes de modélisation physique ou probabiliste, afin d'offrir des prédictions plus explicables et intégrables dans des workflows validés.

Sur le plan computationnel, les ressources nécessaires à l'entraînement et à l'exécution des modèles comme AlphaFold2 restent très élevées. Même si des versions optimisées ont été proposées (OpenFold, ColabFold), la généralisation de ces outils à des contextes à bas budget (pays en développement, laboratoires hospitaliers) demeure un défi. Les exigences matérielles limitent aussi les possibilités de personnalisation ou de fine-tuning local sur des familles de protéines spécifiques.

Par ailleurs, peu de modèles actuels prennent en compte la dynamique temporelle des protéines. Or, de nombreuses fonctions biologiques dépendent de transitions conformationnelles induites par des ligands, des modifications post-traductionnelles ou des gradients physico-chimiques. L'absence de représentation des états intermédiaires ou d'ensembles conformationnels rend la prédiction incomplète pour ces cas. Xu (2019) et Noé et al. (2020) évoquent la possibilité d'intégrer des approches de dynamique moléculaire ou des modèles markoviens dans les futurs modèles hybrides.

Enfin, l'exploitation des prédictions reste partiellement limitée par le manque d'interfaces intuitives, de normes d'annotation universelles, et de pipelines intégrés. Bien que des bases comme AFDB proposent des interfaces riches, leur intégration dans les plateformes hospitalières, les outils de génomique clinique ou les applications mobiles reste marginale. La transformation de ces prédictions en outils décisionnels robustes nécessite encore un travail d'ingénierie logicielle, d'interopérabilité, et de validation terrain.

Les perspectives sont néanmoins nombreuses et enthousiasmantes. L'évolution vers des modèles auto-explicatifs et adaptables en temps réel est en cours. Des travaux intègrent déjà l'apprentissage par renforcement, les bases multimodales (séquence + fonction + localisation). L'émergence de l'apprentissage contrastif, des pré-entraînements multimodaux (image + séquence), ou des interfaces vocales intelligentes pour la modélisation structurale pourrait rendre ces outils encore plus accessibles et puissants.

En conclusion, malgré ses limites actuelles, la prédiction structurale par IA marque une rupture en biologie. Les efforts pour dépasser les verrous identifiés sont nombreux, coordonnés, et soutenus par une communauté scientifique active. Les prochaines générations de modèles, enrichies, distribuées, interprétables et dynamiques, promettent d'ouvrir de nouveaux horizons pour la compréhension du vivant et l'innovation thérapeutique.

Disponibilité des données

Toutes les structures prédites sont accessibles sur https://alphafold.ebi.ac.uk/.

Remerciements

Je remercie l'équipe pédagogique et mes encadrants pour leur accompagnement tout au long de l'année.

Financements

Ce travail n'a bénéficié d'aucun financement externe.

Conflits d'intérêts

L'auteur déclare n'avoir aucun conflit d'intérêts.

Références

[1] Jumper J., Evans R., Pritzel A., et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589.
[2] Tunyasuvunakool K., Damas J., Adzhubei A., et al. AlphaFold Protein Structure Database in 2024. Nucleic Acids Research, 2024, Vol. 52: D368–D375.
[3] Kovalevskiy O., Nicholls R. A., Shabalin I. G., et al. AlphaFold two years on: validation and impact. Acta Crystallographica Section D, 2024, 80: 197–213.
[4] Chen L., Zheng W., Zhang Y. Protein folds vs. protein folding: differing questions, different challenges. Trends in Biochemical Sciences, 2023, 48(2): 104–116.
[5] Xu J. Distance-based protein folding powered by deep learning. Proceedings of the National Academy of Sciences, 2019, 116(34): 16856–16865.
[6] Noé F., Tkatchenko A., Müller K. R., Clementi C. Machine learning for molecular simulation. Annual Review of Physical Chemistry, 2020, 71: 361–390.
[7] Jänes J., Beltrao P. From structure to function through deep learning: recent advances in protein structure prediction. Molecular Systems Biology, 2024, 20(1): e11315.

Lexique

1. MSA (Multiple Sequence Alignment)

Alignement multiple de séquences permettant d'identifier des positions évolutivement conservées ou co-variables entre différentes protéines homologues. Cette information est utilisée pour déduire des contraintes structurales implicites.

2. Réseau neuronal Transformer

Architecture de réseau de neurones fondée sur le mécanisme d'attention. Elle permet de modéliser efficacement les relations entre les éléments d'une séquence, indépendamment de leur distance. Utilisée dans des modèles comme AlphaFold2 ou ESMFold.

3. Carte de distances (distance map)

Représentation bidimensionnelle où chaque case correspond à la distance prédite ou observée entre deux résidus d'une protéine. Cette carte sert de support pour entraîner les modèles à reconnaître des motifs structuraux.

4. Pipeline algorithmique

Enchaînement structuré de traitements informatiques successifs, de l'entrée (séquence protéique brute) à la sortie (structure 3D prédite), incluant l'encodage, la prédiction, l'affinage et l'évaluation.

5. pLDDT (Predicted Local Distance Difference Test)

Score de confiance local attribué à chaque résidu dans une structure prédite. Il reflète la fiabilité de la prédiction au niveau local, sur une échelle de 0 à 100.

6. PAE (Predicted Aligned Error)

Matrice indiquant l'erreur attendue dans la position relative de deux résidus lorsque leur alignement est supposé correct. Utilisée pour estimer l'incertitude structurale entre domaines ou régions distantes.

7. API REST

Interface informatique permettant à un programme d'interagir avec une base de données ou un service web via des requêtes HTTP. Utilisée

pour automatiser l'accès aux données structurales dans des bases comme AlphaFold DB.

8. Rétropropagation en boucle fermée (Recycling)

Technique d'optimisation utilisée dans AlphaFold2 consistant à réinjecter la prédiction intermédiaire dans le modèle pour améliorer la prédiction finale au cours d'itérations successives.

9. CASP (Critical Assessment of protein Structure Prediction)

Compétition internationale biennale qui évalue de manière indépendante les performances des méthodes de prédiction de structures protéiques sur des cibles dont la structure expérimentale est connue mais non publiée.

10. GDT_TS (Global Distance Test Total Score)

Score de similarité structurelle entre une prédiction et une structure de référence. Il mesure la proportion de résidus correctement positionnés dans un seuil de tolérance donné.

11. Erreur PAE entre résidus

Valeur extraite de la matrice PAE qui quantifie l'incertitude relative entre deux résidus spécifiques dans une prédiction structurale.

12. RoseTTAFold

Modèle de prédiction structurale développé par l'équipe du Rosetta Institute. Il combine des informations issues de la séquence, des alignements multiples et des représentations de distance dans un réseau à attention intégrée.

13. Attention multi-tête (Multi-head attention)

Composant du Transformer qui permet de traiter simultanément différentes représentations internes de la séquence, capturant ainsi des relations multiples à différentes échelles.

14. Réseau neuronal convolutionnel (CNN)

Type de réseau neuronal historiquement utilisé pour le traitement d'images et de cartes 2D, incluant les cartes de distances. Progressivement remplacé par les Transformers pour la modélisation de séquences longues.

15. Graph Neural Network (GNN)

Modèle de réseau neuronal conçu pour traiter des graphes, structures de données composées de nœuds (résidus) et d'arêtes (interactions). Utilisé pour représenter les relations structurales dans les protéines.

16. RMSD (Root Mean Square Deviation)

Mesure standard de la déviation quadratique moyenne entre les positions atomiques d'une structure prédite et d'une structure de référence expérimentale.

17. CAMEO (Continuous Automated Model Evaluation)

Plateforme d'évaluation automatisée qui teste en continu les performances des méthodes de prédiction sur de nouvelles structures expérimentales disponibles dans la PDB.

18. GPU non parallélisé

Unité de traitement graphique utilisée pour l'inférence des modèles de deep learning. Un GPU non parallélisé est un GPU dont les calculs ne sont

pas répartis sur plusieurs cœurs ou cartes, ce qui limite la vitesse de prédiction.

19. TM-score (Template Modeling Score)

Indice de similarité structurelle moins sensible à la taille des protéines que le RMSD. Utilisé pour comparer globalement deux structures protéiques.

20. Protéines virales (ex. NSP6, ORF3a)

Protéines codées par le génome d'un virus, jouant un rôle clé dans son cycle de vie. Leur modélisation structurale permet de mieux comprendre les mécanismes d'infection et d'identifier des cibles thérapeutiques.

21. Cartes PAE

Représentations visuelles codées en couleur de la matrice PAE, permettant d'identifier les régions d'une structure prédite pour lesquelles les positions relatives sont incertaines.

22. Interactome

Ensemble des interactions physiques ou fonctionnelles entre protéines dans un organisme. La modélisation structurale permet de prédire les interfaces et de reconstruire des réseaux d'interaction à grande échelle.

23. Variants somatiques (TP53, KRAS, EGFR, etc.)

Mutations acquises au cours de la vie cellulaire, souvent associées à des processus tumoraux. Leur impact peut être évalué en modélisant la structure de la protéine mutée.

24. Pathogènes émergents (ex. Nipah, Lassa, Marburg)

Agents infectieux nouvellement identifiés ou en forte progression, souvent associés à des risques épidémiques. La prédiction structurale de leurs protéines permet une réponse plus rapide en termes de diagnostic ou de vaccin.

25. ProteinMPNN, RFdiffusion

Outils de design de protéines basés sur des réseaux neuronaux, permettant de générer des séquences susceptibles de replier vers une structure cible définie. Utilisés pour le design de novo en biotechnologie et pharmacologie.