

développer leur propre version, RoseTTAFold. Celle-ci utilise de nombreuses stratégies identiques à celles d'AlphaFold2, mais avec quelques particularités.

«Au moment où nous l'avons mis à disposition, c'était de loin le meilleur outil de prédiction de structure, mais il n'est pas aussi bon qu'AlphaFold2», déclare David Baker. Il souligne que DeepMind est une entité privée disposant, contrairement à la plupart des laboratoires universitaires, d'énormes ressources. Le succès d'AlphaFold2 «tient beaucoup à l'argent mis par Google, confirme Amelie Stein, de l'université de Copenhague, mais aussi aux compétences des nombreux experts réunis».

Depuis la sortie d'AlphaFold2 en juillet 2021, les laboratoires se sont rués dessus. Le logiciel est simple d'utilisation, bien qu'il nécessite plusieurs téraoctets d'espace disque pour les bases de données et plusieurs unités de traitement graphique (GPU) pour effectuer les analyses. «Les calculs sur une seule structure sont plutôt faciles», explique Arne Elofsson, de l'université de Stockholm. Mais les analyses de l'ensemble des protéines d'un organisme, ou protéome, sont encore hors de portée de la plupart des laboratoires universitaires.

Pour pallier cette difficulté, Martin Steinegger et ses collègues ont mis au point ColabFold, un système dématérialisé qui fait fonctionner

AlphaFold2 n'est pas encore adapté pour comprendre l'effet des mutations sur la structure des protéines

AlphaFold2 et RoseTTAFold en utilisant des bases de données distantes et la puissance de calcul fournie par Google. L'interface web est relativement simple et permet d'optimiser l'expérience.

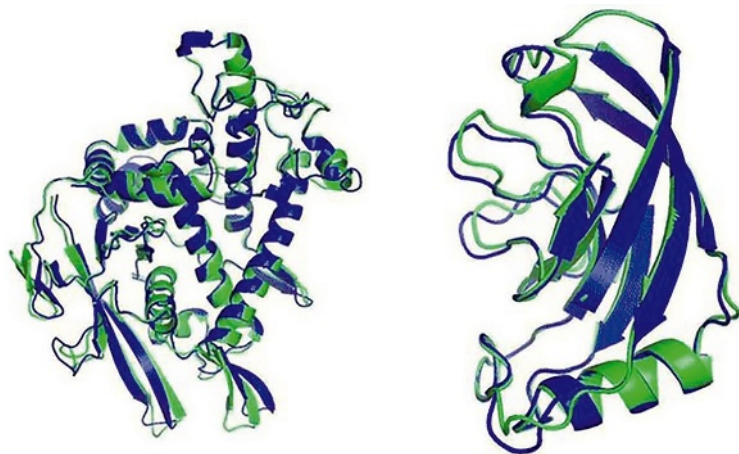
Même l'équipe de DeepMind a été surprise par les performances d'AlphaFold2 lors du CASP14.

Depuis, de nombreuses démonstrations des capacités du logiciel ont été apportées, mais également de ses limites. Dans une étude publiée parallèlement à l'article décrivant l'algorithme, l'équipe de Google a appliqué AlphaFold2 à un ensemble de données comprenant 98,5% du protéome humain. L'algorithme utilise un test de différence de distance locale prédite (pLDDT, pour *predicted local distance difference test*) pour quantifier sa confiance dans la position et l'orientation qu'il propose pour un acide aminé particulier dans la protéine réelle. Finalement, 36% de tous les résidus aminés du protéome ont pu être analysés avec un degré de confiance très élevé.

En août 2021, le groupe d'Alfonso Valencia, du Centre de supercalcul de Barcelone, en Espagne, a conclu de son côté qu'AlphaFold2 avait fait passer de 31 à 50% la proportion d'acides aminés des protéines humaines cartographiables avec précision.

SANS FAUX PLI

Yang Zhang s'attend à ce que le logiciel fasse un travail rapide sur les «fruits les plus faciles à cueillir» du protéome. Mais l'analyse de nombreuses autres protéines relève du défi, notamment celles qui sont composées de plusieurs unités fonctionnelles indépendantes (des domaines) reliées de façon flexible. Ici, les domaines individuels sont



Dans la détermination des structures de protéines, l'intelligence artificielle (*en bleu*) fait désormais presque jeu égal avec les études expérimentales (*en vert*).

accessibles au logiciel, mais pas leur orientation les uns par rapport aux autres.

Les segments de protéines qui sont naturellement désordonnés dans leur état, qui correspondraient à plus d'un tiers de tous les acides aminés du protéome humain, constituent un défi encore plus grand. Aucun algorithme ne peut actuellement prédire comment ils se replient, mais John Jumper note que la méthode des pLDDT peut aider à délimiter ces segments dans une structure.

Une caractéristique inattendue d'AlphaFold2 et de RoseTTAFold est leur capacité à prédire les structures des homodimères, ces complexes formés de deux protéines identiques, et des hétérodimères, quand les protéines assemblées diffèrent. Arne Elofsson et son équipe ont réussi à modéliser jusqu'à 59% des complexes à deux protéines qu'ils ont analysés avec AlphaFold2. Cependant, ce processus est plus difficile sur le plan informatique lorsque l'on tente d'identifier des complexes probables à partir de zéro, plutôt qu'en modélisant des paires d'interactions connues. Mais en appliquant plusieurs algorithmes d'apprentissage profond en parallèle, David Baker et son équipe ont identifié et modélisé des centaines de complexes multiprotéiques à partir de millions de paires d'interactions possibles dans le protéome de la levure *Saccharomyces cerevisiae*. Ici, pour filtrer les paires d'acides aminés qui interagissent le plus probablement, RoseTTAFold était environ cent fois plus rapide qu'AlphaFold2, celui-ci ayant ensuite été utilisé sur le sous-ensemble obtenu.

Sentant l'enthousiasme pour ce domaine, DeepMind a publié en octobre 2021 AlphaFold-Multimer, spécifiquement entraîné pour s'attaquer aux complexes protéiques. Le programme a prédit des interactions avec une haute précision pour 34% des complexes homodimériques testés, et pour 23% des complexes hétérodimériques.

FRONTIÈRES FONCTIONNELLES

De nombreuses questions restent encore hors de portée, note Debora Marks. «Si votre technologie copie la cristallographie, c'est formidable, mais de tels instantanés structuraux statiques ne disent rien du comportement dynamique d'une protéine donnée.» Ainsi, AlphaFold2 produit généralement une seule réponse «correcte» pour chaque séquence. Toutefois, de nombreuses protéines ont plusieurs conformations, toutes pertinentes pour la fonction, par exemple l'état activé ou inhibé d'une enzyme. Pour résoudre ces problèmes, des

techniques expérimentales qui montrent la structure de la protéine réelle dans de multiples états, comme la cryo-EM, sont encore nécessaires.

AlphaFold2 n'est pas non plus adapté pour prédire comment les changements d'acides aminés individuels modifient la structure des protéines, un facteur pourtant crucial pour comprendre l'effet des mutations. Une raison en est que l'algorithme converge vers une solution correcte à partir de nombreuses séquences légèrement différentes, explique Amelie Stein. Néanmoins, son équipe a obtenu des prédictions plus précises en couplant l'apprentissage profond à d'autres algorithmes d'analyse des mutations.

Les biologistes structurels ne sont pas près de perdre leur emploi. En fait, ils sont même désormais en mesure de consacrer plus de temps à d'autres questions fondamentales. Randy Read, de l'université de Cambridge, au Royaume-Uni, note, par exemple, que les prédictions d'AlphaFold2 aident déjà beaucoup les cristallographes dans l'interprétation de leurs données lorsqu'elles sont incomplètes. Les concepteurs de protéines *de novo* tireront également profit d'AlphaFold2 en simplifiant la phase de calcul des modèles qui précède les tests en laboratoire.

Pour Mohammed AlQuraishi, la biologie structurale est à l'aube d'une nouvelle ère qui privilégiera la fonction des protéines plutôt que leur forme. Cette dernière ne sera plus une fin en soi, et la discipline va mériter la composante «biologie» de son nom.

— L'auteur —

> **Michael Eisenstein**
est journaliste scientifique,
basé à Philadelphie.

Cet article a paru sous le titre
«Artificial intelligence powers
protein-folding predictions»
sur le site Nature.com,
le 23 novembre 2021.

— À lire —

> **J. Jumper et al.**, Highly accurate
protein structure prediction with
AlphaFold, *Nature*, 2021.

> **K. Tunyasuvunakool et al.**,
Highly accurate protein structure
prediction for the human
proteome, *Nature*, 2021.

> **W. Zheng et al.**, Protein
structure prediction using deep
learning distance and hydrogen-
bonding restraints in CASP14,
Proteins, 2021.

> **I. Humphreys et al.**, Computed
structures of core eukaryotic
protein complexes, *Science*, 2021.