

# Modélisation 3D de protéines avec Libmol et l'IA (Swissmodel, Alphafold) et réflexion sur les enjeux

Niveau :  
1ere Spécialité

Introduction / Contexte

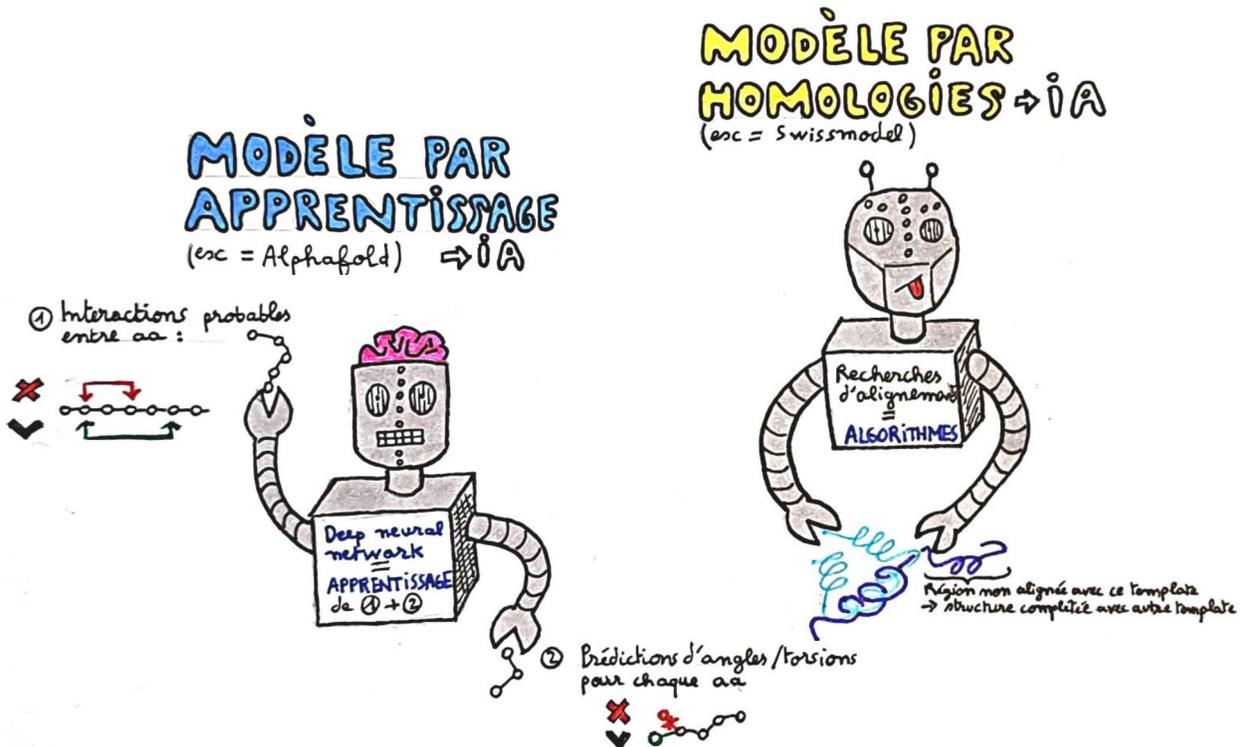
## THEME DE PROGRAMME :

Thème 1 : La Terre, la vie et l'évolution du vivant

Sous-thème 1A : Transmission, variation et expression du patrimoine génétique

Partie : Mutations de l'ADN et variabilité génétique

Les élèves évaluent la fiabilité de l'IA en comparant des structures 3D d'hémoglobines (sain vs drépanocytose) obtenues par modélisation classique et par intelligence artificielle (Swissmodel, Alphafold). Ils utilisent Libmol pour visualiser, mettre en forme et confronter les résultats. L'activité développe l'esprit critique sur l'IA, ses limites et ses enjeux en recherche et santé.





## Compétences travaillées



### Compétences disciplinaires :

- - **Mobiliser des connaissances en biologie moléculaire et en génétique**  
L'étude des structures de l'hémoglobine normale et drépanocytaire nécessite une compréhension des mutations génétiques et de leurs conséquences sur la structure et la fonction des protéines.
- - **Analyser et interpréter des résultats expérimentaux**  
Les élèves comparent les structures obtenues par différentes méthodes, évaluent la fiabilité des modèles et discutent des limites des approches utilisées.

### Compétences transversales :

- - **Collaborer en équipe pour mener une investigation scientifique**  
La réalisation de l'activité en groupe favorise la coopération, la répartition des tâches et la communication entre les membres de l'équipe.
- - **Développer l'esprit critique et l'argumentation scientifique**  
Les élèves évaluent la fiabilité des modèles d'IA, identifient les limites des outils utilisés et argumentent leurs conclusions de manière rigoureuse.
- - **Communiquer des résultats scientifiques à l'oral et à l'écrit**  
Les élèves présentent leurs analyses, interprétations et conclusions, développant ainsi leurs compétences en communication scientifique.

### Compétences du CRCN (Cadre de Référence des Compétences Numériques) :

- - **Domaine 5 : Environnement numérique**  
5.3 Utiliser des outils numériques pour simuler ou modéliser :  
modélisation 3D de structures protéiques avec des logiciels spécialisés (Libmol, SwissModel, AlphaFold).
- - **Domaine 3 : Création de contenus**  
3.1 Développer des documents multimédia :  
=> présentation et confrontation des résultats sous forme de visualisations comparées

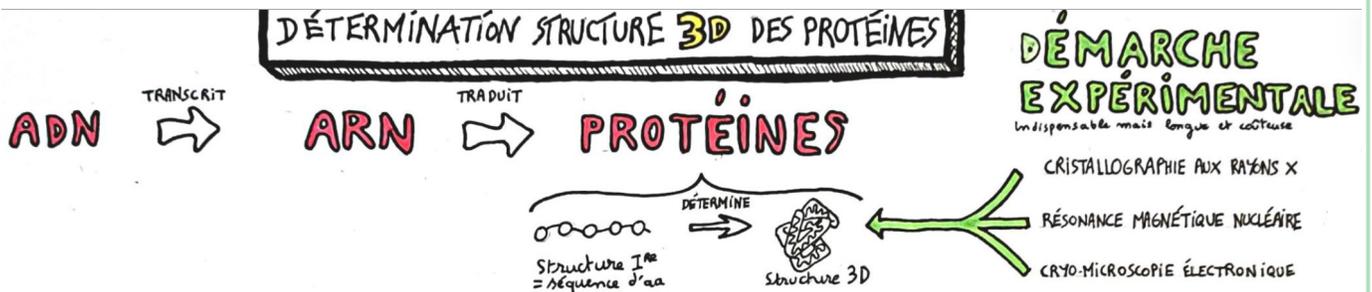
## Déroulement de la séance

### Mise en contexte (15 minutes)

Dans un contexte de classe dialoguée, l'enseignant aura pris soin de recueillir les représentations initiales des élèves concernant l'utilisation de l'IA, ses principes et ses enjeux.

Le sujet de la modélisation moléculaire par IA, largement méconnu des élèves, est alors introduit en s'appuyant sur les connaissances des cours précédents concernant la synthèse des protéines. On évoquera les points suivant :

- Le **prix Nobel de chimie 2024** a été décerné à Demis Hassabis et John M. Jumper (société Google DeepMind) pour leurs travaux sur la prédiction de la structure tridimensionnelle des protéines, assistée par l'IA, à partir de leur séquence d'acides aminés.
- Présentation, à l'aide du document annexe, des **techniques expérimentales de détermination** de la structure tridimensionnelle des protéines



- Les techniques expérimentales longues et coûteuses auront permis, en **plusieurs décennies**, de déterminer  $250 \cdot 10^3$  structures tridimensionnelles de protéines sur les  $250 \cdot 10^6$  protéines connues soit **1/1000<sup>ème</sup>**.
- Concernant la **modélisation 3D par IA**, en revanche, il aura suffi de **1 ans** à AlphaFold pour déterminer la **structure 3D de la totalité des protéines connues !**  
Article de la Recherche « La quasi-totalité des structures de protéines désormais prédites »

Ces résultats impressionnants conduisent assez naturellement à se poser la question de la

**fiabilité de la modélisation tridimensionnelle par IA.**

## Activité en autonomie guidée (1h30)

Durée	Activité / Consigne	Modalité	Matériel	Objectif pédagogique
10 min	<b>Générer</b> un modèle 3D d'hémoglobine S avec Swissmodel via sa séquence d'acides aminés. <b>Télécharger</b> modèle avec + grand indice de fiabilité (QMEAN).	Copier-Coller séquence aa du .edi dans Swissmodel Récupération du modèle généré d' <b>HbS au format pdb</b>	<a href="#">Swissmodel</a> <a href="#">Séquence aa d'HbS (.edi)</a>	Utiliser une IA fonctionnant par <b>homologie</b>
15 min	<b>Visualiser et traiter</b> votre modèle d'HbS dans Libmol. <b>Comparer</b> avec le modèle d'HbA issu de Swissmodel fourni	Charger les modèles HbS et HbA issus de Swissmodel dans la visionneuse	<a href="#">Visionneuse Libmol HbA via Swissmodel (.pdb)</a>	Utiliser un outil de traitement et visualisation Incidence mutation
20 min	<b>Comparer</b> les HbA/HbS <b>Swissmodel</b> (modèles par homologie) aux HbA/HbS <b>Alphafold</b> (modèles par apprentissage) et aux modèles issus des données expérimentales	Charger les modèles HbS et HbA issus d'Alphafold et de Libmol dans la visionneuse	<a href="#">HbA (A.Fold)</a> <a href="#">HbS (A.Fold)</a>  Modèles intégrés HbA/HbS à Libmol	Extrême similitude des structures 3D quel que soit le modèle qui met en évidence une grande fiabilité des résultats

**La fiabilité surprenante des modèles tridimensionnelles de protéines générés par IA interroge alors la **nécessité de poursuivre les recherches expérimentales, qui apparaissent nettement moins efficaces.****

Durée	Activité / Consigne	Modalité	Matériel	Objectif pédagogique
5 min	Expliquez pourquoi l' <b>IA n'est pas suffisante</b> pour déterminer seule la structure 3D des protéines.		<a href="#">Fiche Annexe</a>	<b>Interroger le fonctionnement de la modélisation par IA.</b> <b>Comprendre la notion de modèle.</b>  <b>Exercer son esprit critique</b>  <b>Extraire et mettre en relation des informations issues d'un schéma</b>
10 min	Quelles <b>différences</b> majeures y'a-t-il entre les IA générant des <b>modèles par apprentissage</b> (ex : Alphafold) et les IA générant des <b>modèles par homologie</b> (ex : Swissmodel) ?			
30 min	4) Expliquez les <b>enjeux</b> concernant la détermination de la structure 3D des protéines par IA.  5) D'après vous, en vous appuyant seulement sur les documents et sur votre raisonnement, quels pourraient être les <b>risques</b> associés à accorder trop de confiance à l'IA concernant la détermination des structures protéiques ?  6) Quelles <b>mesures</b> pourraient être prises pour diminuer ces risques ?  7) Question ouverte : quelles sont vos <b>impressions</b> concernant l'utilisation de l'IA dans les sciences ? Ce TP/TD vous a-t-il fait évoluer dans votre vision de l'IA (Vous a-t-il rendu plus enthousiaste ? Plus méfiant(e) ? Ou encore plus mitigé(e) ?)			Travail en groupe (2 ou 3 élèves) à partir de documents

### Conclusion : (5/10 minutes)

- 1) Les **modèles IA** sont alimentés des résultats des travaux expérimentaux. Leur fiabilité en dépend. Les modèles IA peuvent en retour guider la **recherche expérimentale** : ces travaux **se complètent plus qu'ils ne s'opposent**
- 2) L'**ampleur des enjeux** et la **rapidité des découvertes** peuvent conduire à considérer le développement de la modélisation tridimensionnelle des protéines par IA comme une **question socialement vive** (enjeux, inégalités possibles...).



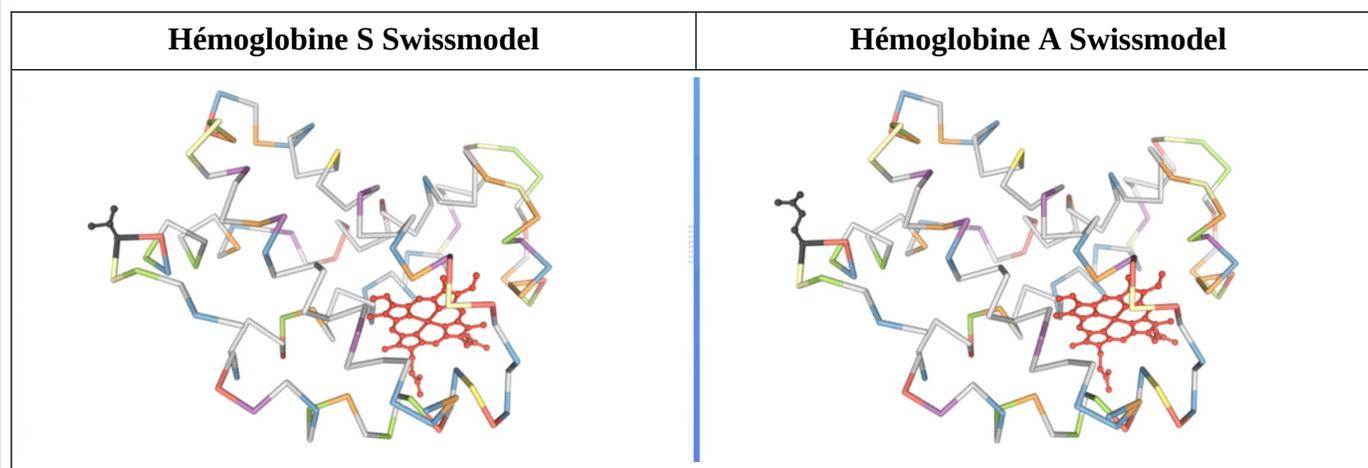
## Analyse de la séance / Retour d'expérience

Il est apparu au cours de cette séance que les élèves, bien qu'utilisateurs réguliers d'IA (principalement ChatGPT), connaissaient peu leurs fonctionnements et avaient une **connaissances assez limitées** des enjeux (éthiques, scientifiques, climatiques...) liés à l'utilisation de l'IA. Ces difficultés ont pu être paliées par une **appétance** des élèves pour le sujet qui a été accueilli avec un certain enthousiasme. Néanmoins, à cela s'ajoutait quelques **difficultés techniques liées à l'utilisation des logiciels** (notamment concernant libmol et l'utilisation de l'interface de visualisation comparative) qui ont fait que les élèves ont manqués de temps pour finaliser leurs travaux en une seule séance lors de ma première tentative. J'ai repensé alors la séance (que j'ai pu tester à une 2e occasion) pour la rendre plus efficace et plus balisée. Les modèles alphafold ont été directement fournis aux élèves par soucis de gain de temps et aussi du fait que l'utilisation d'Alfafold nécessite l'utilisation d'un compte google qui ne respecte donc pas les normes RGPD.

Le **choix de l'hémoglobine n'était pas le plus pertinent** pour tester l'incidence sur la forme d'une protéine d'une mutation du gène codant la protéine mais il a été contraint par la progression de la collègues remplacée, contexte dans lequel a été effectuée la séance. Par ailleurs, le fait de traiter à la fois du sujet de la forme des protéines, de la modélisation par IA et d'y ajouter la question des mutations était plutôt ambitieux.

Ci-dessous : **Comparaison des 2 modélisations d'hémoglobines générées par Swissmodel**

=> la parfaite adéquation des structures (au delà de la différence des acides aminés) s'explique par la **modélisation par homologie**  
=> Swissmodel compare les mêmes séquences aux mêmes templates et propose le template qui correspond le mieux à la séquence, il n'y aura pas d'incertitude puisque Swissmodel va proposer **toujours le même résultat** :



=> Remarque : à la différence de Swissmodel, Alphafold fonctionne par prédiction de la position spatiale des acides aminés par rapport qu contexte de la séquence. Ce dernier fonctionnement rappelle le modèle de langage probabiliste utilisé pour les IA pour construire une phrase mot par mot selon le contexte qui précède le mot.

Le TP conduisait à afficher et traiter **6 molécules**, ce qui devenait rapide une fois Libmol maîtrisé, mais restait un peu **fastidieux**. Dans les faits, une fois les HbS et HbA générées par Swissmodel et comparées à l'aide de Libmol, il a été demandé à certains groupes de travailler sur HbS et d'autres sur HbA. Chaque groupe n'a donc traité que **4 modélisations d'hémoglobines**.

Il pourrait néanmoins être intéressant de **s'affranchir** pour cette séance **de la question des mutations** et de se limiter aux seules questions de formes des protéines et des modélisations par IA **pour éprouver plus rigoureusement la fiabilité des modèles moléculaires générés** et mettre d'avantage l'accent sur les divergences entre les modèles générés par homologie et ceux générés par apprentissage.

De ce point de vue, l'étude des **protéines prions** pourraient s'avérer intéressante. En effet, la protéine prion normale (PrPC) et pathologique (PrPSc) ont la même séquence primaire d'acides aminés, mais elles **diffèrent considérablement par leur structure tridimensionnelle**. Ces différences structurelles ne sont **pas codées par des séquences de nucléotides différentes** mais résultent de leur repliement et de leur conformation 3D.

L'étude des **conséquences des mutations à un TP préalable** rendrait le travail plus fluide et permettrait donc d'approfondir le travail critique lors du TP sur la modélisation par IA.

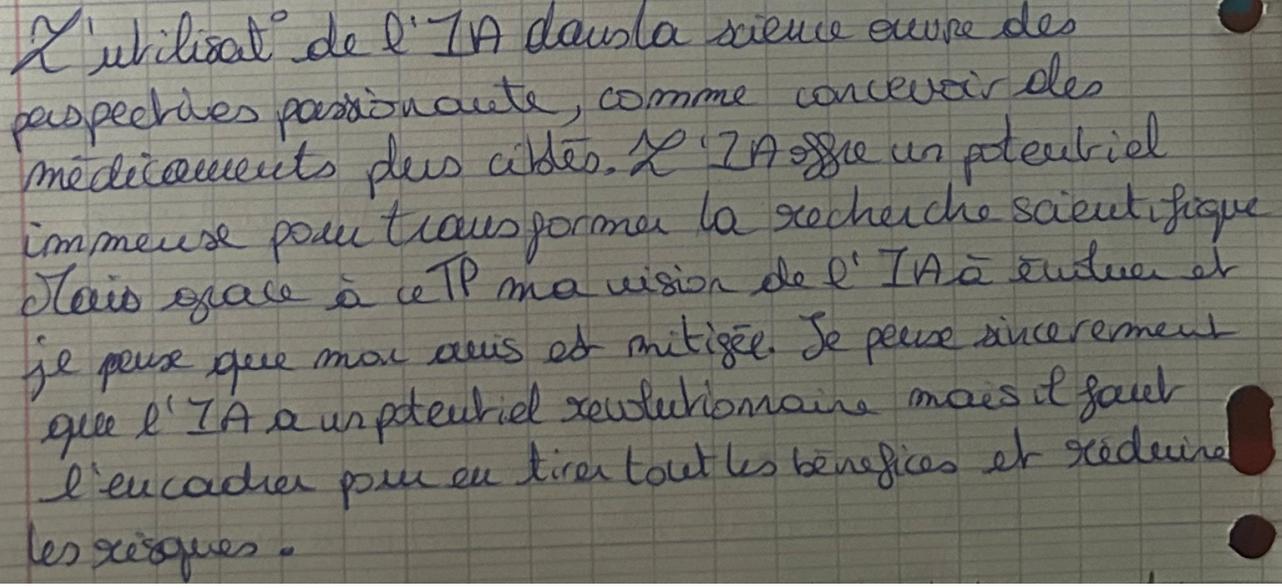
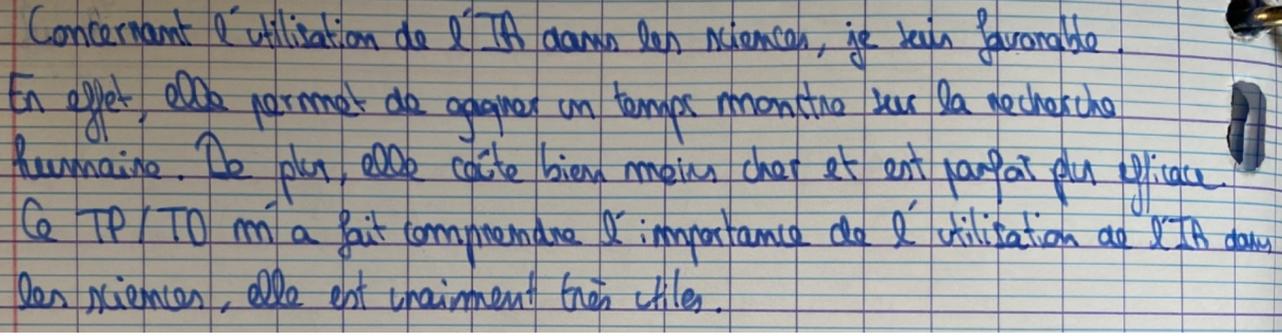
**Remarque :** Pour comparer les modèles générés par IA et les modèles expérimentaux présents sur Libmol il est intéressant de cliquer sur le code pdb (entouré ci-dessous) pour obtenir des précisions. Le modèle d'hémoglobine HbS sur Libmol est ainsi issu des techniques de cristallographie aux rayons X.

The screenshot shows the LibMol interface. At the top, the PDB ID '2HBS' is highlighted with a red box. A blue arrow points from this box to a popup window titled 'Dimère d'hémoglobine drépanocytaire désoxygénée'. The popup contains the following information:

- Source du modèle:** Protein Data Bank, 2HBS THE HIGH RESOLUTION CRYSTAL STRUCTURE OF DEOXYHEMOGLOBIN S, Auteurs : Harrington, D.J., Adachi, K., Royer Junior, W.E.
- Référence:** PDB ID: 2HBS, Harrington, D.J., Adachi, K., Royer Jr., W.E. (1997) "The high resolution crystal structure of deoxyhemoglobin S.", Journal of molecular biology 272: 398-407 DOI: 10.1006/jmbi.1997.1253
- Résumé:** We have refined the crystal structure of deoxyhemoglobin S (beta Glu6->Val) at 2.05 Å resolution to an R-factor of 16.5% (free R=21.5%) using crystals isomorphous to those originally grown by Wishner and Love. A predominant feature of this crystal form is a double strand of hemoglobin tetramers that has been shown by a variety of techniques to be the fundamental building block of the intracellular sickle cell fiber. The double strand is stabilized by lateral contacts involving the mutant valine interacting with a pocket between the E and F helices on another tetramer. The new structure reveals some marked differences from the previously refined 3.0 Å resolution structure, including several residues in the lateral contact which have shifted by as much as 3.5 Å. The lateral contact includes, in addition to the hydrophobic interactions involving the mutant valine, hydrophilic interactions and bridging water molecules at the periphery of the contact. This structure provides further insights into hemoglobin polymerization and may be useful for the structure-based design of therapeutic agents to treat sickle cell disease.

**Retours d'élèves sur la question ouverte :**

**Quelles sont vos impressions concernant l'utilisation de l'IA dans les sciences ?**

1	<p>L'utilisation de l'IA dans les sciences m'a laissé une impression mitigée. D'une part j'ai été impressionné par la capacité des outils d'IA (comme Alphafold ou Swissmodel) à traiter rapidement une immense quantité de données. Cela ouvre des perspectives intéressantes dans des domaines comme la médecine, où l'IA peut simplifier les tâches de l'humain. Cependant, ce TP m'a également rendu plus conscient des limites et des risques liés à l'IA. En effet, l'IA peut commettre des erreurs et cela dépend aussi de l'évolution de l'IA. L'IA est un progrès dans les sciences.</p>
2	 <p>L'utilisation de l'IA dans la science ouvre des perspectives passionnantes, comme concevoir des médicaments plus ciblés. L'IA offre un potentiel immense pour transformer la recherche scientifique. Mais grâce à ce TP ma vision de l'IA a évolué et je pense que mon avis est mitigé. Je pense sincèrement que l'IA a un potentiel révolutionnaire mais il faut l'encadrer pour en tirer tout les bénéfices et réduire les risques.</p>
3	 <p>Concernant l'utilisation de l'IA dans les sciences, je suis favorable. En effet, elle permet de gagner un temps monstrueux sur la recherche humaine. De plus, elle coûte bien moins cher et est parfois plus efficace. Ce TP/TO m'a fait comprendre l'importance de l'utilisation de l'IA dans les sciences, elle est vraiment très utile.</p>
4	<p>Mes principales impressions concernant l'IA dans les Sciences sont que l'IA est un moyen d'accélérer des processus, de détermination par exemple. Mais l'IA ne doit pas être utilisée pour créer des nouvelles techniques de recherche où autre, en effet il peut y avoir des erreurs car l'IA n'est pas encore totalement fiable. Ce TP m'a fait découvrir que l'on utilisait l'IA dans les Sciences, j'en suis impressionné mais je pense qu'il ne faut pas trop compter sur l'IA, au risque de faire des erreurs.</p>



## Pistes d'amélioration



=> Rendre la séance moins linéaire en repensant les supports. Difficultés néanmoins liées à l'originalité du sujet.

Cette séance est très propice à une ludification (enquête, jeu de rôle...) qui pourrait limiter cette linéarité en concevant, par exemple, diverses missions pouvant varier d'un groupe à l'autre (par exemple en travaillant sur des protéines différentes, ou encore en donnant à certains groupes avec des modèles Alphafold et à d'autres des modèles Swissmodel).

=> Différenciation, au delà de l'accompagnement par l'enseignant, sous forme de tutos vidéos (éventuellement en amont de la séance) pour les élèves plus en difficulté notamment concernant l'utilisation de logiciels comme Libmol.