François-Xavier COUDERT

Introduction aux concepts de l'intelligence artificielle : les méthodes d'IA comme nouveau langage

D'après la conférence de François-Xavier Coudert, Directeur de recherche CNRS, Professeur attaché ENS – Université PSL.

Introduction

Compte tenu de la situation particulière de l'intelligence artificielle (IA) comme outil scientifique très jeune et porteur d'espoirs de progrès considérables pour de nombreux domaines scientifiques et techniques, je souhaite préciser mon positionnement personnel. C'est celui de chercheur, actif en recherche fondamentale et membre de projets de recherches partenariaux, mais c'est aussi celui d'un enseignant qui veut transmettre ces méthodes aux futurs chercheurs, aux étudiants et donc à la communauté tout entière. Mon point de vue est aussi informé par de nombreuses discussions avec des entreprises du domaine, sur le sujet « données numériques et apprentissage » en entreprise, dans le cadre d'une activité de consultance.

1 Apprentissage et intelligence artificielle

1.1. L'apprentissage dans le *machine learning*

1.1.1. *Machine learning* et programmation

Pour introduire l'intelligence artificielle (IA), il faut poser certaines définitions et la ***Figure 1*** est un bon point de départ : dans une définition axée sur la discipline « psychologie », on voit intervenir des éléments importants d'expérience, de pratique et d'étude ; on voit aussi apparaître un côté « processus itératif ».

Le *machine learning*, en français *apprentissage automatique* ou *apprentissage statistique*, est une sous-partie de l'intelligence artificielle à laquelle va se consacrer ce chapitre. Le but est de reproduire les éléments de l'apprentissage humain, donc de développer un modèle (ou algorithme) reposant sur l'utilisation de données qui représentent l'expérience. Le processus est lui-même algorithmique : on développe un algorithme via un algorithme, et l'apprentissage devient donc un problème d'optimisation. Dans un raccourci un peu simpliste, on dira que l'algorithme sous sa forme classique, c'est de dire « si j'ai des ingrédients et si j'ai une recette, je peux faire un gâteau ». Il peut être réussi, il peut être raté, là n'est pas la question, mais je peux faire un gâteau, et normalement je peux faire toujours le même gâteau si je prends toujours les mêmes ingrédients et toujours la même recette ; j'ai cette reproductibilité (***Figure 2***).

Dans le *machine learning*, on inverse le paradigme, puisque l'idée, ce n'est plus de se dire « je veux produire un gâteau », mais « je veux produire une recette » ; ou dans le langage de l'IA : « je veux produire un algorithme ». Je le produis via un autre procédé algorithmique, une optimisation, mais ce que j'obtiens à la fin est un algorithme. Cet algorithme sera appliqué nécessairement à de nouvelles données dans de nouvelles circonstances.

1.1.2. Parallèle avec l'apprentissage humain

Il faut faire le parallèle avec l'apprentissage humain. Si j'apprends aujourd'hui à faire des nœuds et que j'apprends sur une corde rouge de deux millimètres de diamètre, ce n'est pas pour toute ma vie faire des nœuds sur une corde rouge de deux millimètres de diamètre. C'est pour être capable d'utiliser ces nœuds

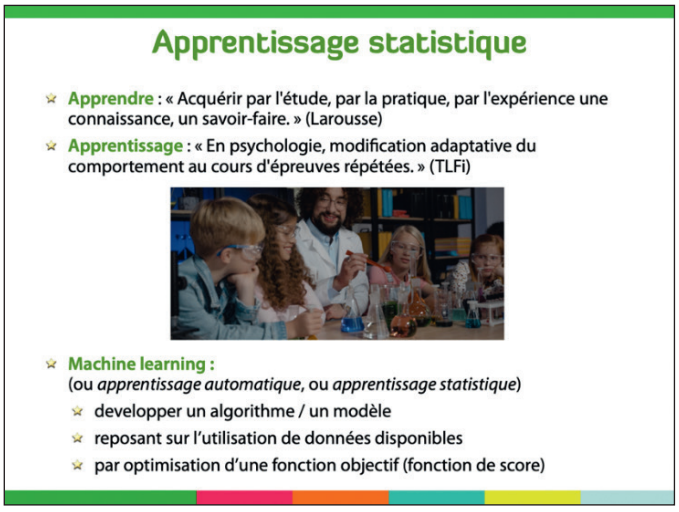


Figure 1

Définition de l'apprentissage et du machine learning *(ou en français, « apprentissage automatique » ou « apprentissage statistique »). Crédit photo : Tima Miroshnichenko, libre de droits.*

dans des circonstances qui vont différer de celles de mon apprentissage, en l'adaptant. Si je sais faire plusieurs nœuds, je serai, à terme, capable de les combiner, de faire de nouveaux nœuds dans des circonstances nouvelles, d'improviser. **C'est vraiment cela qui constitue le processus d'apprentissage, et c'est cela que cherche à reproduire l'intelligence artificielle et dans ce cas spécifique, le *machine learning*.**

1.2. Quand utiliser l'apprentissage ?

1.2.1. Cas où il n'est pas nécessaire

On va beaucoup parler d'exemples d'utilisation, de tâches où on veut recourir à des méthodes d'apprentissage faisant partie de l'intelligence artificielle, du *machine learning*. Mais il faut réaliser que l'on n'aura pas toujours besoin d'utiliser ces méthodes sophistiquées ; pour beaucoup de problèmes, l'utilisation du *machine learning* – l'utilisation d'une méthode d'apprentissage pour entraîner un algorithme – ne sera simplement pas nécessaire. Exemple : si je veux établir une fiche de paie, si je veux calculer mes impôts, la recette est connue, les règles sont connues, et il suffit de les appliquer. Si je veux établir un emploi du temps, j'ai un ensemble de contraintes, j'ai un ensemble de disponibilités des gens, j'ai un problème, certes difficile, mais qui peut très bien être résolu par des méthodes connues. Là, je n'ai pas besoin nécessairement d'utiliser des méthodes d'apprentissage.

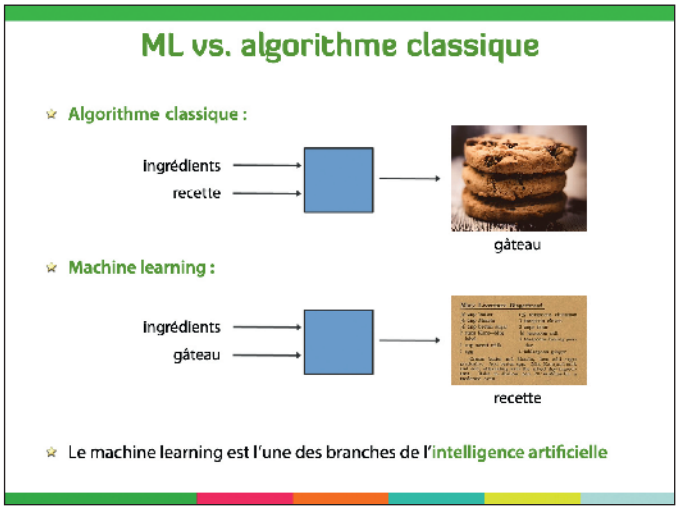


Figure 2

Machine learning *versus algorithme classique.*

1.2.2. Cas où il est nécessaire

Pour jouer aux échecs ou jouer au jeu de go, les règles sont connues, mais dans ce cas spécifique, vous savez peut-être – parce que cela a été très repris dans la presse – qu'il y a une plus-value dans l'utilisation des méthodes de *machine learning*.

En fait, on aura besoin de *machine learning* dans les cas où les données sont abondantes et où elles peuvent être produites assez facilement, puisque pour apprendre, il faudra entraîner un algorithme sur ces données. On pourra aussi utiliser le *machine learning* quand, à l'inverse, il y a peu de connaissances, ou quand elles sont rares et chères à obtenir, quand l'expertise humaine atteint ses limites, quand on n'arrive pas à trouver une compréhension fine du problème et qu'on cherche justement à la suppléer et à la complémenter

par une méthode basée sur les données.

On peut également utiliser le *machine learning* quand les humains ont une expertise mais sont peu capables de l'expliquer eux-mêmes ou de la rationaliser, ou encore pour aider à mieux comprendre, suppléer, complémenter des choses qui font plus appel à l'intuition, dans la reconnaissance vocale, dans la vision, ou si on veut un peu se rapprocher des exemples chimiques, puisque c'est ce qui nous intéresse dans le présent ouvrage, quand on va vouloir parler d'intuition chimique (***Figure 3***).

1.2.3. Lien avec l'intuition scientifique

Si je montre à un chimiste organicien expérimenté une centaine de molécules et que je lui demande : « J'ai pensé à utiliser ces produits-là, lesquels sont faisables ? », il va rapidement pouvoir me dire : « Cette molécule-là n'est pas du tout possible ; celle-ci je ne suis pas sûr, il faudrait que je regarde mais je pense qu'on peut y arriver. » Il y a une intuition chimique qui s'est développée chez lui au cours de sa carrière. Il s'agit de connaissances humaines mais qui sont difficiles à rationaliser, à expliquer. C'est là où les méthodes d'apprentissage par *machine learning* vont être intéressantes.

On pourra aussi utiliser ces méthodes d'apprentissage quand des solutions techniques existent aujourd'hui mais qu'elles sont coûteuses… Des exemples sont donnés plus bas.

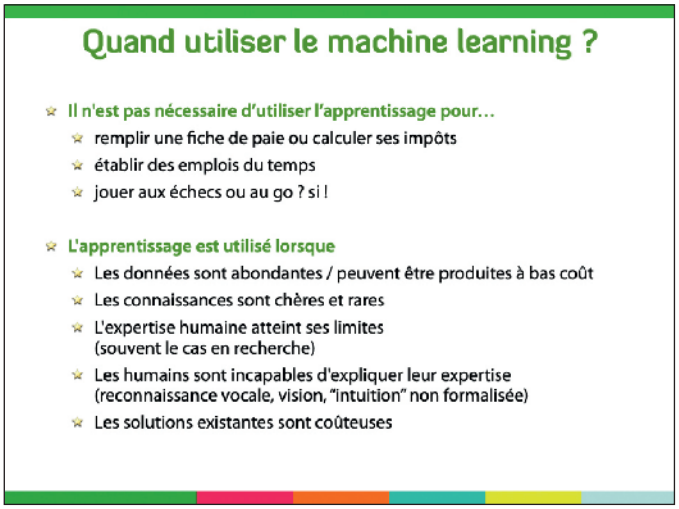


Figure 3

Quand utiliser le machine learning *?*

2 Application du *machine learning* à la chimie

2.1. L'intelligence artificielle en chimie : une révolution ?

2.1.1. L'innovation de l'intelligence artificielle

Compte tenu de la présence impressionnante dans les médias des propos et descriptions des méthodes d'apprentissage, on peut se demander si l'utilisation de l'intelligence artificielle est une « révolution obligatoire » dans tous les domaines. Regardons spécifiquement le cas de la chimie, car elle donne l'exemple d'un tournant dans certaines recherches.

2.1.2. L'apport dans les compétences

Bien sûr les collègues, les experts sont clairs, on ne va pas remplacer aujourd'hui le chimiste par la machine, on ne va pas remplacer l'expert

ILLUSTRATION DE LA PUISSANCE DE L'IA EN CHIMIE

Voici une anecdote qui a beaucoup aidé mon laboratoire à réaliser la puissance de ces méthodes.

Cela se passe en 2018, lors d'une compétition CASP[1] au Mexique sur la prédiction du repliement des protéines. On donne à différentes équipes une séquence de protéines et on leur demande de prédire la structure tridimensionnelle (qu'on appelle le « repliement ») des protéines. C'est une question notablement difficile de la biochimie, parce que ce repliement, cette structure, dépend de beaucoup de petites interactions, de détails d'équilibre (***Figure 4***).

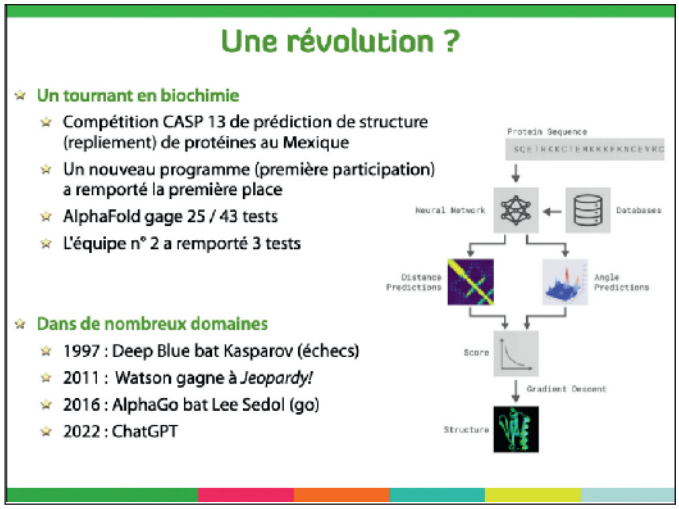


Figure 4

La révolution de l'intelligence artificielle ? À droite : exemple de la prédiction de structure de protéines (Gradient descent : descente du gradient[2]). Adapté de De novo structure prediction with deep-learning based scoring*, https://www.deepmind.com/blog/alphafold-using-ai-for-scientific-discovery*

Un nouvel entrant cette année-là – une équipe qui n'était pas présente les années précédentes – se place premier, gagne 25 tests sur 43, alors que l'équipe qui repart avec la médaille d'argent remporte 3 tests sur toute la série. Surprise de tous ! Miracle ? Non, simplement l'utilisation de l'IA ! C'est vraiment une méthode qui, quand elle a été introduite, a révolutionné le domaine de la prédiction de la structure des protéines.

On connaît d'autres exemples, que ce soit Watson à *Jeopardy*[3], AlphaGo[4], et puis plus récemment l'utilisation des **intelligences artificielles conversationnelles comme ChatGPT**[5].

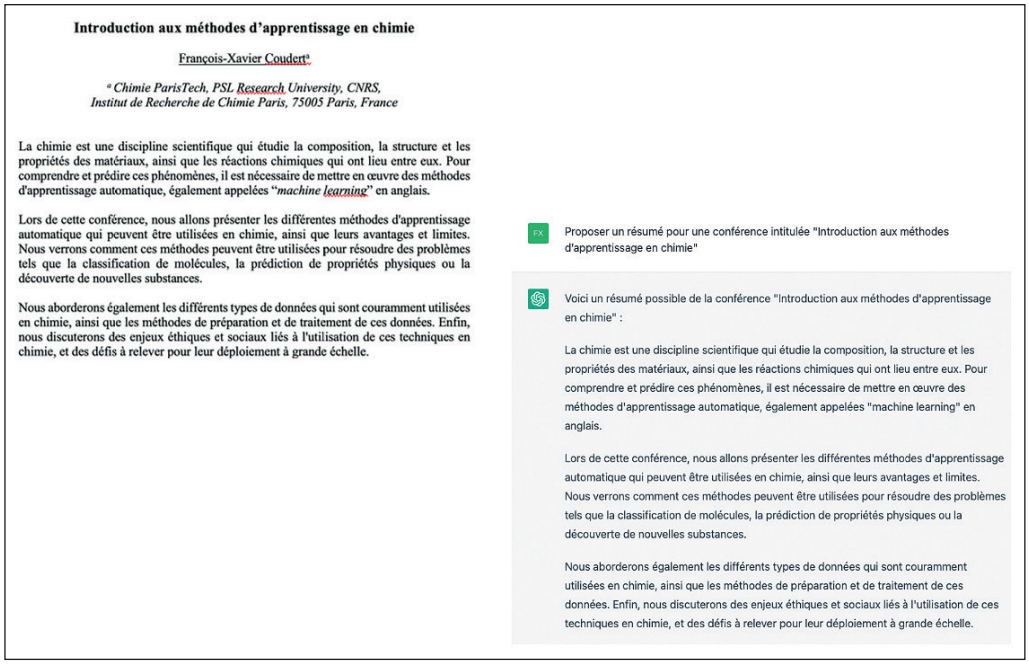
1. *Critical Assessment of protein Structure Prediction* (CASP) : littéralement, « évaluation critique de la structure des protéines », expérience visant à l'élaboration d'un état de l'art de la prédiction des structures (repliement) des protéines.
2. Algorithme d'optimisation.
3. *Jeopardy* est un jeu télévisé ressemblant à « Questions pour un champion ». Un superordinateur, Watson, développé par IBM y gagne 1 million d'euros en répondant correctement à la plupart des questions.
4. Programme informatique utilisant l'apprentissage capable de jouer au jeu de Go.
5. Agent conversationnel utilisant l'intelligence artificielle capable de tenir une conversation et de répondre à de multiples demandes sous forme d'échange de messages (chats) dans plusieurs langues.

humain par une intelligence artificielle c'est un faux débat. En revanche, ce qu'il peut se passer, c'est que les chimistes qui savent utiliser au sein de leur équipe – collaboration au sens large et non en tant qu'individu – les méthodes basées sur les données, les méthodes d'apprentissage statistique, auront un avantage compétitif sur ceux qui ne le savent pas. Les laboratoires, les équipes et l'industrie s'en rendent compte aujourd'hui : l'expert chimiste doit déjà maîtriser de nombreuses techniques différentes, et l'IA devient une corde de plus à son arc.

De plus en plus, la recherche et l'innovation impliquent la génération de très larges quantités de données et permettent d'explorer et de valoriser celles qui étaient déjà présentes. Au titre de l'anecdote : je me suis demandé ce qui se passerait si j'envoyais un résumé écrit par le fameux bot ChatGPT (***Figure 5***)... et c'est ce que j'ai fait. À ce jour, je n'ai pas eu de retour, donc j'espère qu'il était très bien, je n'ai pas changé une virgule.

2.2. Le langage des méthodes d'apprentissage

Pour entrer un peu plus dans le vif du sujet et parler des types de problèmes que nous allons voir aujourd'hui, je veux décrire un peu le langage qui est utilisé dans les méthodes d'apprentissage. C'est important quand on rentre dans un nouveau domaine d'en maîtriser les codes et le langage.



Introduction aux méthodes d'apprentissage en chimie

François-Xavier Coudert[a]

[a] *Chimie ParisTech, PSL Research University, CNRS, Institut de Recherche de Chimie Paris, 75005 Paris, France*

La chimie est une discipline scientifique qui étudie la composition, la structure et les propriétés des matériaux, ainsi que les réactions chimiques qui ont lieu entre eux. Pour comprendre et prédire ces phénomènes, il est nécessaire de mettre en œuvre des méthodes d'apprentissage automatique, également appelées "*machine learning*" en anglais.

Lors de cette conférence, nous allons présenter les différentes méthodes d'apprentissage automatique qui peuvent être utilisées en chimie, ainsi que leurs avantages et limites. Nous verrons comment ces méthodes peuvent être utilisées pour résoudre des problèmes tels que la classification de molécules, la prédiction de propriétés physiques ou la découverte de nouvelles substances.

Nous aborderons également les différents types de données qui sont couramment utilisées en chimie, ainsi que les méthodes de préparation et de traitement de ces données. Enfin, nous discuterons des enjeux éthiques et sociaux liés à l'utilisation de ces techniques en chimie, et des défis à relever pour leur déploiement à grande échelle.

Proposer un résumé pour une conférence intitulée "Introduction aux méthodes d'apprentissage en chimie"

Voici un résumé possible de la conférence "Introduction aux méthodes d'apprentissage en chimie" :

La chimie est une discipline scientifique qui étudie la composition, la structure et les propriétés des matériaux, ainsi que les réactions chimiques qui ont lieu entre eux. Pour comprendre et prédire ces phénomènes, il est nécessaire de mettre en œuvre des méthodes d'apprentissage automatique, également appelées "machine learning" en anglais.

Lors de cette conférence, nous allons présenter les différentes méthodes d'apprentissage automatique qui peuvent être utilisées en chimie, ainsi que leurs avantages et limites. Nous verrons comment ces méthodes peuvent être utilisées pour résoudre des problèmes tels que la classification de molécules, la prédiction de propriétés physiques ou la découverte de nouvelles substances.

Nous aborderons également les différents types de données qui sont couramment utilisées en chimie, ainsi que les méthodes de préparation et de traitement de ces données. Enfin, nous discuterons des enjeux éthiques et sociaux liés à l'utilisation de ces techniques en chimie, et des défis à relever pour leur déploiement à grande échelle.

Figure 5

Exemple de résumé d'une conférence écrit par ChatGPT.

Pour parler de méthodes d'apprentissage, on va utiliser plusieurs types de définitions de méthodes (***Figure 6***) : les méthodes « supervisées » pour ce qui est de l'apprentissage, les méthodes « non supervisées », et puis les méthodes « par renforcement ».

Les **méthodes « supervisées »** permettent de faire des prédictions basées sur des données existantes, les **méthodes « non supervisées »** de détecter des tendances, des patterns[6], de l'information, de valoriser ou de mieux visualiser des données disponibles. Les **méthodes « par renforcement »**, quant à elles, sont utilisées par exemple dans tout ce qui est apprentissage pour les jeux (pas uniquement, mais entre autres) ; ce sont aujourd'hui les moins utilisées dans notre domaine[7].

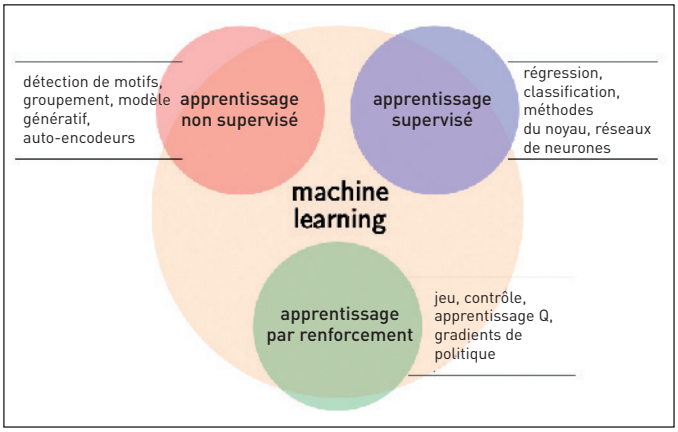


Figure 6

Les différents langages du machine learning. *Figure reprise, avec autorisation, des cours disponibles en ligne de Chloé-Agathe Azencott (https://cazencott.info). Elle est aussi l'autrice d'un livre en français très accessible,* Introduction au Machine Learning, *chez Dunod InfoSup (*https://cazencott.info/index.php/pages/Introduction-au-Machine-Learning*).*

2.3. L'apprentissage « supervisé »

2.3.1. Explications

L'apprentissage « supervisé » est l'une des tâches les plus couramment pratiquées en chimie. Il permet d'utiliser un jeu de données pour créer un algorithme qui fera de la prédiction (***Figure 7***), typiquement d'une grandeur continue (on parle de régression) ou d'une catégorie (on parle de classification).

Par exemple, déterminer la constante d'équilibre[8] ou l'enthalpie de formation[9] de tel composé, c'est une tâche de régression. On peut avoir aussi des méthodes dites « de classification », qui permettent de prédire cette fois-ci non plus une information continue, mais soit une information binaire, soit une information catégorielle. Est-ce que telle molécule est soluble dans l'eau, dans l'éthanol, dans les deux ou dans aucun ? C'est ce que l'on appellera une tâche de classification.

2.3.2. Son utilité

L'apprentissage supervisé fait partie des méthodes le plus largement utilisées en chimie, simplement parce qu'elles s'apparentent aux méthodes « relation structure/propriétés », que ce soit sur des molécules ou sur les matériaux. On complémente cette branche de la chimie ainsi que la chimie théorique, qui est déjà largement utilisée, par des méthodes de *machine learning*. Pourquoi ? Si l'on s'intéresse à

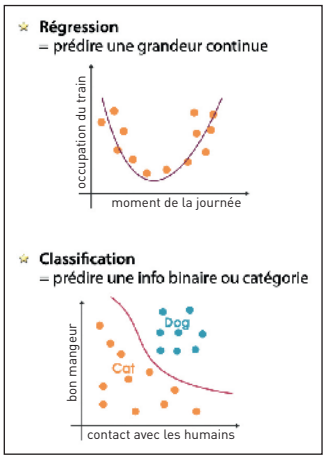


Figure 7

Exemples de régression et de classification. Figure adaptée, avec autorisation, des cours disponibles en ligne de Chloé-Agathe Azencott (https://cazencott.info).

6. Pattern : motif.
7. Différentes méthodes d'analyse de données.
8. Constante décrivant l'état d'équilibre d'un système chimique.
9. Énergie à fournir pour former une mole d'un composé.

une nouvelle molécule et que l'on veut calculer une propriété chimique ou physique, un point de solubilité, une activité catalytique[10], on peut utiliser plusieurs techniques (***Figure 8***).

10. Pour une enzyme, correspond à la quantité d'enzyme pour catalyser une réaction dans des conditions données.

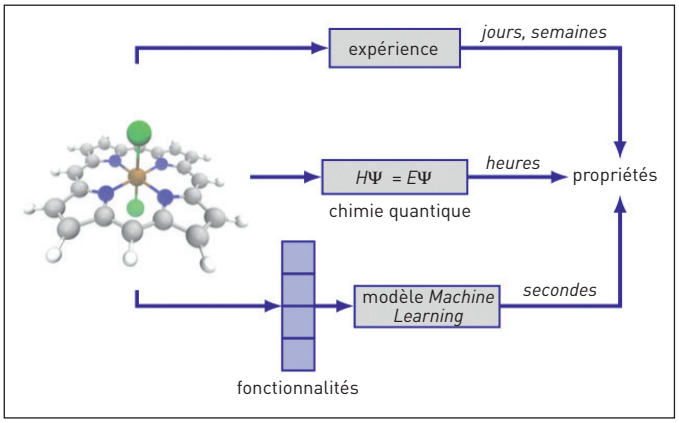


Figure 8

Différents méthodes pour obtenir les relations propriétés/structure.

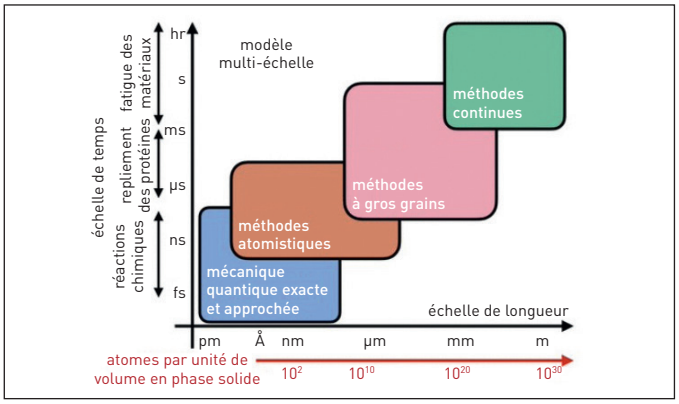


Figure 9

Échelle de temps et de distance utilisées en chimie théorique.
Tiré de Combining Machine Learning and Computational Chemistry for Predictive Insights Into Chemical Systems
(https ://doi.org/10.1021/acs.chemrev.1c00107)

La première peut être de synthétiser la molécule, le matériau, de faire une mesure expérimentale ou plusieurs, afin de mesurer la propriété. La deuxième est d'utiliser les méthodes de chimie théorique (ou chimie computationnelle), qui sont nombreuses, bien développées, et ont des décennies d'utilisation, de validation (présentées sur la ***Figure 9***), avec différents niveaux de précision, différentes échelles d'espace et de temps qui leur sont accessibles.

2.3.3. La prédiction de données

La troisième voie consiste à se dire : « si j'ai assez de données, je peux entraîner un algorithme sur ces données, un modèle de prédiction qui me permettra ensuite de prédire des relations de structure/propriétés ». Cela se fait en prenant initialement des bases de données existantes, ou en en créant pour l'occasion.

On va avoir sur le schéma de la ***Figure 10***, deux axes principaux. L'axe vertical, c'est celui de l'entraînement du modèle, de ce que l'on va appeler « l'apprentissage », et l'axe horizontal, celui de l'exploitation du modèle, donc de la réalisation des prédictions. La première chose à faire est l'entraînement dans l'axe vertical, c'est-à-dire rassembler des données.

Pour prédire la solubilité d'un grand nombre de molécules organiques, il faut une base de données initiale. Donc je la construis, je la trouve, je la crée, elle peut être expérimentale, elle peut être issue de calculs théoriques. Je dois d'abord la rassembler et vérifier que toutes les données sont pertinentes. Ce sont ces

données d'entraînement qui seront utilisées pour l'entraînement du modèle de *machine learning*. Et c'est l'optimisation des prédictions sur ces données-là qui va me donner **le prédicteur, donc l'algorithme issu de l'entraînement**. C'est le *ML-trained model*[11] (***Figure 10***).

2.3.4. Utilisation de l'algorithme

Cet algorithme entraîné peut être utilisé pour prédire les propriétés de nouvelles molécules. Évidemment, si les molécules ne ressemblent pas suffisamment aux données initiales, je vais avoir une faible précision, puisque le modèle ne s'est pas entraîné sur cette gamme de molécule. L'idée est que le modèle a une certaine capacité de généralisation, qu'il est capable de traiter des molécules qui ne sont pas celles qu'il aurait vues lors de son entraînement. Sinon, d'ailleurs, il serait globalement inutile, ce serait juste de la mémorisation et non de l'apprentissage. Ceci dit, la capacité de généralisation du modèle reste limitée. S'il n'a vu que des petites molécules organiques et que je lui présente demain une grosse protéine ou un matériau hybride organique/inorganique, il ne sera pas capable de correctement prédire ses propriétés.

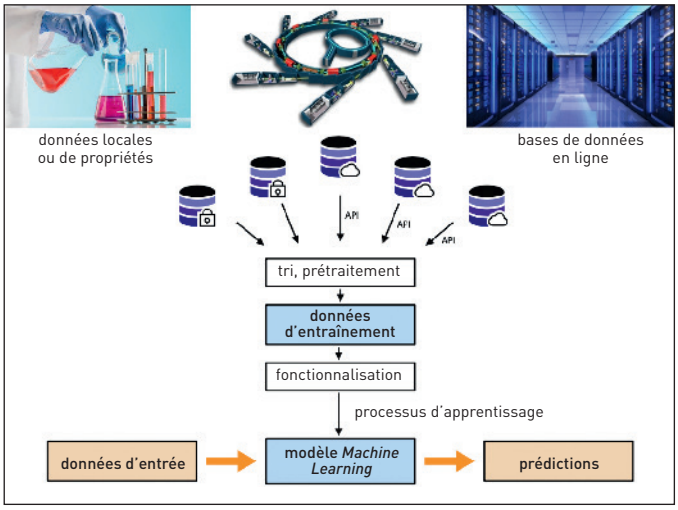


Figure 10

Processus d'entraînement d'un algorithme. Reproduit de Machine learning approaches for the prediction of materials properties *(https://doi.org/10.1063/5.0018384)*

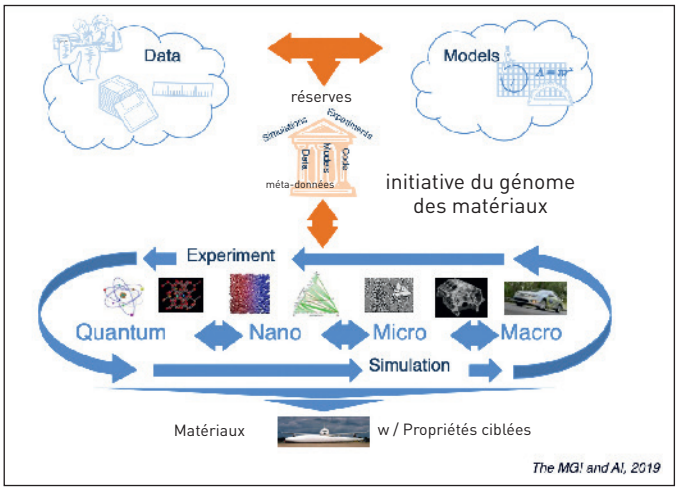


Figure 11

Boucle d'obtention des propriétés des matériaux. Tiré de Materials Genome Initiative & Artificial Intelligence @ NIST, James Warren.

3. Un exemple personnel d'intelligence artificielle en chimie

3.1. Présentation de l'exemple

L'utilisation de ces modèles s'insère dans l'écosystème plus large de la découverte en chimie.

On présente ici un exemple dans le domaine des matériaux (***Figure 11***). Le travail s'insère

11. *ML-trained model* : modèle de *Machine Learning* entraîné.

dans un écosystème plus large d'innovation, dans le cadre des matériaux où l'on a déjà des données existantes qui viennent soit de calculs antérieurs, soit d'expériences qui ont été faites, de bases de données structurelles, de bases de données de propriétés. Ces bases de données s'intègrent dans une boucle de découvertes, expériences, rationalisation, théories, suppositions, hypothèses, nouvelles expériences et on recommence. L'idée est d'**utiliser les méthodes d'apprentissage pour pouvoir accélérer un peu cette boucle.**

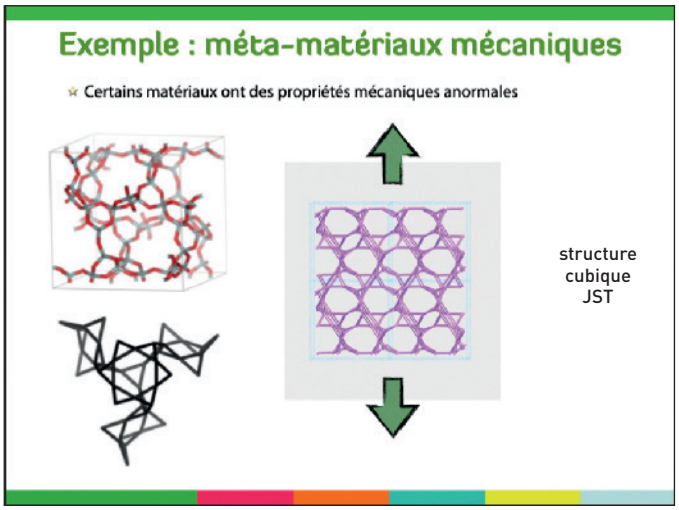


Figure 12

Des matériaux aux propriétés inhabituelles.

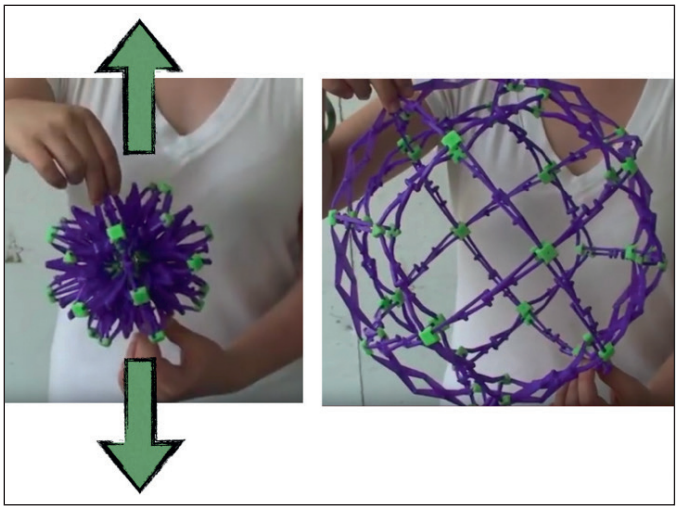


Figure 13

Exemple de la sphère avant et après élongation. Hoberman Sphere, *images tirées du site du revendeur https://www.educationstation.ca/*

Voici un exemple présenté avec un peu plus de détails, qui s'intéresse à la découverte de l'utilisation de *machine learning* pour la prédiction de matériaux qu'on appelle **méta-matériaux mécaniques** (***Figure 12***). Ce sont des matériaux qui ont une propriété mécanique anormale, inhabituelle. Typiquement, quand vous prenez un matériau et que vous l'étirez, il est censé devenir un peu plus fin dans les autres directions par compensation. Mais ce n'est pas le cas de tous les matériaux. Les enfants jouent pas mal avec des petits modèles comme celui de la ***Figure 13*** ; celui-ci est appelé un « *Hoberman Sphere* » et lorsque l'on tire dans une direction, ils s'étendent dans toutes les autres directions[12].

3.2. Rareté de cette matière cristalline

Cette propriété est rare dans la matière cristalline, et si l'on regarde les matériaux connus, on ne trouve que cinq cristaux qui ont cette propriété sur des centaines de milliers de structures cristallines connues. On s'est donc posé la question

12. JST : Type de structure d'une zéolite, un type de minéral.

suivante : comment identifier ces matériaux aux propriétés dites « rares » ? Et puis peut-être en découvrir d'autres. C'est une approche qui se prête assez naturellement à l'utilisation des méthodes basées sur les données, pour des bases de structures de matériaux inorganiques simples. Avec la base qui s'appelle « *Materials Project* » (***Figure 14***), on peut trouver facilement et en accès libre environ 133 000 structures inorganiques. Pour un certain nombre de ces structures – 13 000, soit à peu près 10 % de la base de données – des propriétés mécaniques ont été calculées.

3.3. Utilisation des données

Une des questions de base, pour comprendre la méthode est : peut-on utiliser ces données-là pour prédire des propriétés mécaniques d'autres matériaux ? Pour y répondre, on prend toute la base de données, on considère les matériaux pour lesquels on a des informations mécaniques, on effectue un entraînement et on crée par *machine learning* un prédicteur que l'on réapplique à toute la base de données complète. On peut alors quantifier ces matériaux.

Sans aller dans le détail, on peut intégrer ces méthodes dans une approche à plus large échelle. Plus précisément, on se place entre la modélisation classique, qui va être un peu imprécise pour des propriétés qui dépendent vraiment du détail de l'organisation microscopique, et une approche de type chimie quantique qui est très précise et peut prédire les propriétés des matériaux, mais qui demande une quantité de calculs importante et ne peut donc pas être utilisée sur des milliers, des dizaines de milliers, des centaines de milliers de matériaux.

Essayons d'abord de contourner une limitation du problème posé qui est que l'on cherche à voir un phénomène rare puisque j'ai très peu de résultats positifs pour beaucoup de calculs ; c'est d'ailleurs caractéristique du fait que ces matériaux-là sont très rares. Pour contourner la limitation, on peut partir d'une base de données très large d'un

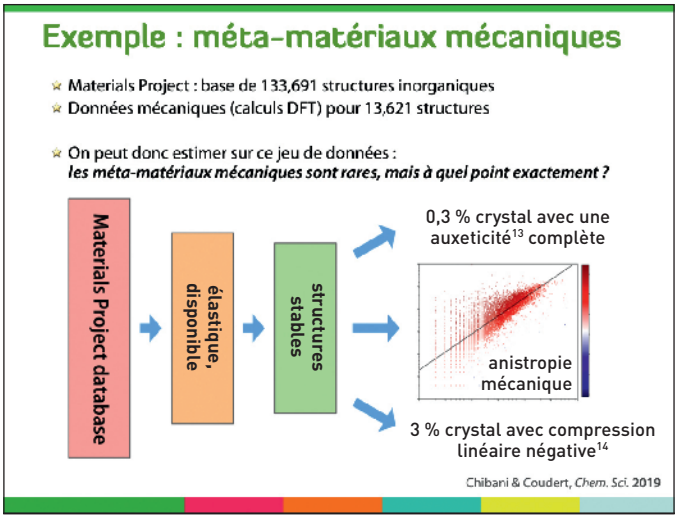


Figure 14

Rareté des méta-matériaux mécaniques issus d'une base de données.

13. Matériau s'élargissant perpendiculairement à la direction de l'étirement.
14. NLC ou compressibilité linéaire négative, est la réaction d'un matériau dont l'une au moins des directions présente une expansion sous une compression mécanique isotrope.

demi-million de structures (***Figure 15***). Une première exploration avec une modélisation classique permet de réduire ce nombre à 462 000, ce qui reste important. On en choisit aléatoirement qui ont cette propriété ou pas cette propriété prédite, puis on fait des calculs très précis sur ce millier de matériaux. On entraîne ainsi un prédicteur puis on recommence.

3.4. Création de la boucle

La suite logique du travail est de réappliquer le prédicteur à toute la base, de prédire à nouveau un certain nombre de structures et de recommencer. Cette technique de traitement permet d'accélérer la boucle de découverte de ces matériaux « rares ». Cependant, la précision finale du modèle sur ce problème spécifique n'est pas forcément très élevée.

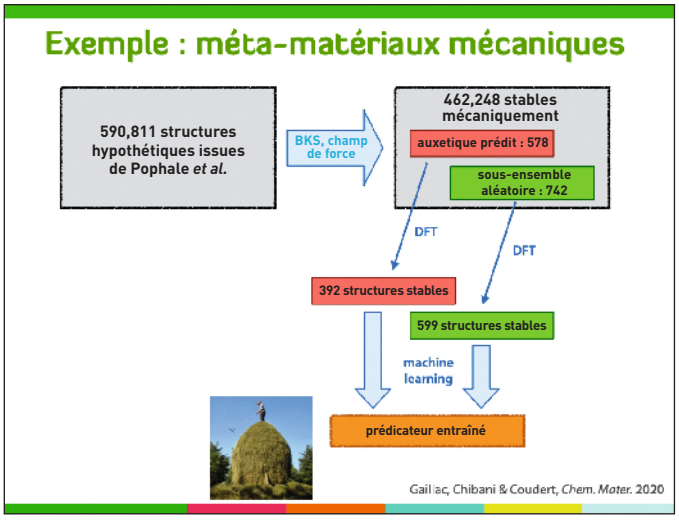


Figure 15

Étapes de l'élaboration d'un prédicteur pour les méta-matériaux mécaniques, paysan sur une botte de foin.

Le prédicteur qu'on a produit au final a encore 50 % de faux positifs, ce qui peut sembler énorme et qui pour beaucoup d'applications serait un « *deal-breaker*[15] » total. Mais ici, comme on a des données initialement très mal équilibrées avec très peu de matériaux, avoir 50 % de faux positifs n'est pas si grave, et certainement moins grave que d'avoir de nombreux faux négatifs. Ça veut déjà dire qu'on peut ensuite recaractériser la moitié des matériaux qu'on trouve par une autre méthode, qu'on a déjà bien accéléré la vitesse à laquelle on a trouvé des matériaux et réduit le nombre de calculs qu'on veut faire.

3 Bases de données disponibles et leurs défauts

Base de données

Il y aura beaucoup de composants moléculaires à discuter à l'avenir, donc pour s'intéresser spécifiquement aux matériaux, on disposera de larges bases de données. C'est pour ça que ces méthodes basées sur les données sont très utilisées aujourd'hui, et justifient de grands espoirs dans la recherche.

Le grand nombre de bases de données qui existe vient du processus académique de publication qui exige depuis longtemps que les matériaux cristallins découverts et rapportés aient leurs structures publiées.

On a donc des bases de données de grandes quantités existantes,

15. *Deal-breaker* : rédhibitoire.

comme la *Cambridge Structural Database*[16] qui est l'une des plus grandes, ayant dépassé le million de structures. Un processus actuellement en œuvre est de les augmenter avec des données issues de la chimie théorique par calculs, pour pouvoir à l'avenir améliorer les méthodes prédites.

16. *Cambridge Structural Database* : Base de données structurelle de Cambridge.

Conclusion
Pour l'efficacité de l'apprentissage : tous contre le biais de publication !

En conclusion, il est approprié de citer l'importance d'un « biais » qui apporte une limitation intrinsèque à la qualité des bases de données et limite donc la performance d'ensemble de l'approche des propriétés des matériaux par *machine learning*.

Ce « biais » très important est le biais de publication, venant du fait que, dans les bases de données, on ne trouve par nature que des choses qui ont marché. Ceci vient du choix des auteurs, mais a pour conséquence de diminuer considérablement la puissance de la recherche comme on l'a vu plus haut. Certains collègues essaient de contourner ce biais en faisant comprendre que l'on n'a pas, en chimie, assez de données négatives publiées, accessibles, disponibles aujourd'hui en chimie.

Si vous faites de la chimie qui rate, parlez-en autour de vous et publiez-le. Vous ferez faire un gros progrès aux méthodes d'intelligence artificielle !