Cet ouvrage est issu du colloque « Chimie et intelligence artificielle »
qui s'est déroulé le 8 février 2023 à la Maison de la Chimie.

« COLLECTION CHIMIE ET ... »

Collection dirigée par Philippe Gœbel

Président de la Fondation internationale de la Maison de la Chimie

Chimie et intelligence artificielle

Carlo Adamo, Sylvie Bégin-Colin, François-Xavier Coudert, Guillaume Fayet, Loïc Jierry, Michel Lutz, Mario Maglione, Sébastien Preys, Gian Marco Rignanese, Julien Romestant, Laurent Schio, Cédric Villani

Coordonné par Danièle Olivier et Paul Rigny

edp sciences

Fondation de la Maison de la Chimie

Conception de la maquette intérieure et de la couverture :
Pascal Ferrari

Crédits couverture :
Images : © sdecoret – © Monopoly919 – © gen_A – © panuwat – © Gorodenkoff – © Alex / Adobe Stock

Mise en pages et couverture : Patrick Leleux PAO (Caen)
Conception graphique, visuel du colloque : CB Defretin

Imprimé en France

ISBN (papier) : 978-2-7598-3547-8
ISBN (ebook) : 978-2-7598-3548-5

Tous droits de traduction, d'adaptation et de reproduction par tous procédés, réservés pour tous pays. La loi du 11 mars 1957 n'autorisant, aux termes des alinéas 2 et 3 de l'article 41, d'une part, que les « copies ou reproductions strictement réservées à l'usage privé du copiste et non destinées à une utilisation collective », et d'autre part, que les analyses et les courtes citations dans un but d'exemple et d'illustration, « toute représentation intégrale, ou partielle, faite sans le consentement de l'auteur ou de ses ayants droit ou ayants cause est illicite » (alinéa 1er de l'article 40). Cette représentation ou reproduction, par quelque procédé que ce soit, constituerait donc une contrefaçon sanctionnée par les articles 425 et suivants du code pénal.

© Fondation de la Maison de la Chimie

Fondation de la Maison de la Chimie
28, rue Saint-Dominique
75007 Paris, France

Ont contribué à la rédaction de cet ouvrage :

Carlo Adamo
Directeur de l'Institut de Chimie des sciences de la vie et de la santé de l'École nationale supérieure de Chimie ParisTech et du CNRS, Membre de l'Institut Universitaire de France

Sylvie Bégin-Colin
Professeur et ancienne directrice ECPM-Université de Strasbourg (2014-2021), Institut de Physique et Chimie des Matériaux de Strasbourg, UMR 7504, CNRS-Unistra

François Xavier Coudert
Directeur de recherche CNRS, Professeur attaché ENS – Université PSL

Guillaume Fayet
Docteur de l'Université Paris VI, Responsable d'études et de recherche à l'Ineris

Loïc Jierry
Professeur à l'ECPM, Université de Strasbourg, Institut Charles Sadron, UPR CNRS 22

Michel Lutz
Chief Data Officer et Digital Factory Head of Data, Chercheur-associé au LIMOS (Laboratoire d'Informatique, de Modélisation et d'Optimisation des Systèmes), TotalEnergies

Mario Maglione
Directeur de recherche CNRS ICMCB Bordeaux
Co-pilote du PEPR DIADEM

Sébastien Preys
Chef de Projets Data Science – ONDALYS

Gian Marco Rignanese
Institute of Condensed Matter and Nanosciences (Université catholique de Louvain)
School of materials sciences and engineering Northwestern polytechnical University Xian (Chine)

Julien Romestant
Directeur de l'Intelligence économique de Cosmetic Valley

Laurent Schio
Responsable France de la plateforme de recherche Integrated Drug Discovery chez Sanofi

Cédric Villani
Université Lyon I et Institut des Hautes Études Scientifiques

Équipe éditoriale :
Danièle Olivier et Paul Rigny

Sommaire

Avant-propos, par **Paul RIGNY** 9
Préface, par **Danièle OLIVIER**..... 13

Partie 1 : Intelligence artificielle : recherche et formation en chimie

Chapitre 1 : Introduction aux concepts de l'intelligence artificielle ; les méthodes d'IA comme nouveau langage
par **François-Xavier COUDERT** 19

Chapitre 2 : L'intelligence artificielle comme moteur dans la recherche en chimie
par **Carlo ADAMO** 33

Chapitre 3 : Présentation de la Majeure Chimie&IA de l'ECPM. Description de l'apport de l'IA pour la préparation et la caractérisation de matériaux pour la santé
par **Sylvie BÉGIN-COLIN et Loïc JIERRY** 47

Chapitre 4 : L'expérience d'Ondalys dans la formation continue aux outils opérationnels de la chimiométrie et du machine learning
par **Sébastien PREYS** 65

Partie 2 : Intelligence artificielle dans la recherche en chimie, notamment dans la recherche de matériaux innovants

Chapitre 5 : Le projet DIADEM : accélérer la découverte de nouveaux matériaux grâce à l'intelligence artificielle
par **Mario MAGLIONE** 91

Chapitre 6 : Informatique des matériaux : comment combiner la puissance des calculs *ab initio* à haut débit et l'intelligence artificielle ?
par **Gian Marco RIGNANESE** 109

Chapitre 7 : Intelligence artificielle et nouvelles approches méthodologiques pour la maîtrise des risques industriels
par **Guillaume FAYET** 125

Partie 3 : Intelligence artificielle et industrie

Chapitre 8 : Intelligence artificielle et parfumerie cosmétique : nouvelles expériences client et réduction du *time to market*
par **Julien ROMESTANT** 149

Chapitre 9 : Transition énergétique et technologies numériques : comment la donnée est utilisée pour la stratégie multi-énergies de TotalEnergies
par **Michel LUTZ** 159

Chapitre 10 : De la sérendipité à l'intelligence artificielle en recherche pharmaceutique
par **Laurent SCHIO** 169

Chapitre 11 : Intelligence artificielle pour la science et l'industrie
par **Cédric VILLANI** 185

Paul RIGNY

Avant-propos

La Fondation de la Maison de la Chimie a créé en 2009, une collection de livres de vulgarisation pour public averti, la collection « Chimie et… » dont cet ouvrage est le vingt-neuvième exemplaire. Cette collection veut faire prendre conscience à quel point la chimie est essentielle à nos vies quotidiennes – un état de fait souvent sous-estimé. Chaque volume est en principe dévolu à un domaine d'activités particulier dont les acteurs, scientifiques et industriels, font ressortir les spécificités, les réalisations et les projets. On peut lister ainsi : les vêtements, l'habitation, les transports, les loisirs, les médicaments, l'environnement, la cosmétique, le sport ou l'art… D'autres volumes sont dévolus à des considérations spécifiques comme *Chimie et Notre-Dame de Paris* après l'incendie, l'archéologie d'Alexandrie ou plus génériques, comme « l'expertise scientifique » ou, précisément *Chimie et intelligence artificielle*.

Chacun des livres de la collection est précédé d'un colloque qui réunit un millier d'auditeurs du monde académique, de l'enseignement ou de la recherche industrielle, où les meilleurs spécialistes exposent leurs travaux. Leurs présentations au colloque sont prises comme bases des chapitres des ouvrages de la collection.

Il n'y a pas 5 ans, seuls quelques laboratoires actifs dans les sciences de l'informatique connaissaient l'intelligence artificielle (IA), puis une « explosion » a eu lieu. Des exemples frappants du succès des concepts associés ont créé une traînée de poudre et tous les domaines des sciences dures (physique, chimie, géologie, mais aussi sciences du vivant et même des sciences humaines malgré l'origine mathématique des concepts en jeu) ont réalisé les perspectives géantes qui s'ouvraient et ont entraîné d'autres mondes, comme la sociologie, grande consommatrice de statistiques, l'économie, les industriels, etc., à s'en rapprocher.

Aujourd'hui, il s'agit d'un enthousiasme (plutôt que d'un engouement car cela va durer) que toute la presse, même grand public, répand. Pas un journal qui ne cite l'intelligence artificielle comme s'il s'agissait d'une technique ancienne et complètement mûre. On

dirait qu'il s'agit d'une technique « factuelle » comme la multiplication des nombres ou la conduite d'une voiture ; on aurait oublié qu'elle est encore récente donc perfectible et limitée. Ces limitations appartiennent d'ailleurs plus au niveau de la psychologie du public des lecteurs que des acteurs des sciences dures : « elle pourrait nous abêtir » dit sévèrement Cédric Villani dans son chapitre !

Essayons-nous à quelques considérations sur la technique en réalisant que ce que font presque toujours les scientifiques et d'ailleurs chaque individu dans son champ, c'est de procéder par imitation. Le nœud de l'activité humaine, c'est « l'apprentissage » à un point qu'on ne réalise peut-être pas suffisamment. On le sait pour l'éducation du jeune enfant, on le sait pour l'élève qui est confronté aux règles de base, ou encore pour l'étudiant qui « apprend » un métier. On ne se l'avoue pas assez pour le professionnel, en particulier quand son métier est proche de l'invention (le chercheur) ; pourtant tout le monde commence sa réflexion en fouillant sa mémoire, son expérience ou celle de ses contemporains, puis en étudiant son contenu et en l'adaptant à son exigence personnelle du moment. Sans référence explicite à la « science des données », on voit que chacun travaille avec une « base de données », une base qu'il trouve dans sa mémoire et qu'il améliore en travaillant et en utilisant les techniques numériques. Tout est dit : cette phase de son activité intellectuelle est confiée d'abord partiellement puis de plus en plus complètement à l'outil informatique. Ce sont les prémices de l'intelligence artificielle !

Pour synthétiser telle molécule possédant telle propriété physique ou chimique, un bon chimiste aura une dizaine d'exemples rendus disponibles par son passé professionnel, un très bon chimiste en aura une centaine, une équipe de cribleurs comme on en trouve dans les entreprises pharmaceutiques va monter à un (ou quelques milliers) mais… s'il en faut quelques centaines de milliers ou quelques millions… il faut changer de monde… On atterrit alors dans celui de l'IA.

Nous ne décrirons certainement pas ici les techniques « inventées » ou mises ou point dans les laboratoires spécialisés, tout ce qu'on désigne par le « machine learning ». Elles existent et se perfectionnent, et en même temps s'associent à des techniques d'auto-contrôle de cohérence, nombreux, évolutifs et complexes. Le résultat en est que ces méthodes informatiques (ces algorithmes) sont mis à la portée des non-spécialistes, qui peuvent être chimistes, biologistes, journalistes ou médecins, ou etc. Après le côté conception de l'IA, voici son côté épreuve de son utilité réelle, qui semble convaincre de plus en plus les utilisateurs. Ces différentes phases sont abordées de façon pédagogique dans les chapitres de cet ouvrage.

On doit imaginer l'effet de ces techniques sur les métiers actuels : l'accès à une efficacité inatteignable par d'autres moyens que l'IA aura des

conséquences nécessairement énormes mais difficilement prévisibles. Retournons vers notre chimiste. Lui qui, naguère, cherchait le meilleur ligand pour sa protéine avec ses raisonnements et connaissances acquises, se flattait d'avoir fait passer quelques thèses de doctorat ; plus tard, avec les techniques de criblage, il démodait le passé avec des résultats dix fois plus nombreux et pertinents. Et bien l'IA va franchir plusieurs stades d'un coup et considérablement augmenter ces perspectives de résultats, par un facteur 10, par un facteur 100 ? Car la machinerie de l'IA aura pu fouiller dans d'autres champs de la chimie, dans celui des polymères s'il est minéraliste ou l'inverse, et examiner des comportements inconnus de lui.

Cet exemple n'est là que pour soutenir une discussion qualitative. Il montre que le champ des possibles accessibles à un laboratoire peut « exploser » ; il y a évidemment des cas où ceci est déjà acquis, mais cela va se répandre partout et tout révolutionner. Le monde du travail et l'appréciation des tâches dans un atelier, celui des laboratoires qui auront l'impression (réelle d'un certain sens) de gagner des générations et, si on s'aventure plus avant dans la société grand public, celui des médias, des juristes, des sociologues si consommateurs de statistiques, etc.

L'IA apporte-t-elle un Eldorado pour tous ces professionnels ? J'aurais tendance à dire « oui », si l'on arrive à gérer les changements sociaux associés au monde du travail (qu'il s'agisse de l'industrie ou des laboratoires). Apporte-t-elle un Eldorado à la Société ? Je serais beaucoup plus circonspect. Cédric Villani nous donne quelques avertissements qui font froid dans le dos : nos personnalités seront-elles résistantes aux risques de manipulation ? pourra-t-on échapper à l'avènement de changements sociaux qui toucheraient aux pires catastrophes humaines ?

Mais la réponse de chacun touche à sa culture ou à sa foi. L'homme saura-t-il se gérer ? pense-t-on « optimisme » ? ou « fatalisme » ? L'aventure de l'IA est partie. Il faut la creuser, la mener du mieux possible et… La Nave Va ! La chimie aussi !

Cet ouvrage, comme tous ceux de la collection « Chimie et… », est décliné en ressources pédagogiques facilement utilisables, dans notre médiathèque www.mediachimie.org

Paul Rigny
Conseiller du président
de la Fondation de la Maison
de la Chimie

Préface

Danièle OLIVIER

Ce 29e ouvrage de la collection « Chimie et… », issu du colloque du même nom, demeure fidèle aux objectifs de permettre, grâce à l'aide d'experts confirmés, l'accès des non-spécialistes, notamment des formateurs et leurs élèves, à une information scientifique rigoureuse et actualisée, dans des domaines sociétaux importants pour le présent et l'avenir.

L'IA, tout le monde en parle, tout le monde veut l'appliquer, certains l'appliquent déjà avec succès, mais en fait le développement en est très récent et sait-on vraiment de quoi on parle ?

Nous avons publié dans notre médiathèque www.mediachimie.org (dans l'espace colloque), un petit quizz pré-colloque (toujours en ligne) pour tester les connaissances des élèves sur ce point, et beaucoup d'adultes même diplômés, n'ont pas obtenu le maximum !

Bien que l'IA soit déjà présente dans la R&D, et que beaucoup de grands groupes industriels recherchent des compétences dans ce domaine, celle-ci est pratiquement ignorée de la majorité des chimistes et elle n'apparaît dans l'enseignement, au niveau supérieur, que depuis peu alors que tout le monde est convaincu de la place importante qu'elle pourra prendre dans un certain nombre de domaines d'applications, d'ailleurs à mieux définir.

Nous avons consacré la première partie de cet ouvrage à la Recherche et à la formation.

Les concepts de l'IA et du machine learning et l'utilisation de l'IA comme moteur dans la recherche en chimie sont présentés dans les deux premiers chapitres par François-Xavier Couderc et Carlo Adamo, tous deux professeurs responsables d'unités de recherches à l'ENSC ParisTech.

Le développement de l'énorme potentiel de l'IA par les chimistes dans leurs travaux de recherches entraîne une évolution des métiers de techniciens, d'ingénieurs et de chercheurs et requiert de nouvelles compétences pour

répondre aux besoins. C'est pourquoi **deux chapitres sont consacrés à la formation** :

– Le premier à l'École Européenne de Chimie Polymères et Matériaux de Strasbourg, pour la formation d'ingénieurs capables de parler un double langage : chimiste et à la fois compétent en IA pour la préparation et la caractérisation des matériaux pour la santé.

– Le deuxième dans le domaine de la formation continue présente l'expérience de la société Ondalys pour la formation continue aux outils opérationnels de la chimiométrie et du machine learning.

La deuxième partie est dédiée à la présentation de quelques exemples du développement de l'IA dans la recherche en chimie et notamment dans le domaine des matériaux innovants.

– Les matériaux sont au cœur des transitions énergétiques et numériques : ils permettent l'émergence de nouvelles technologies. Dans le premier chapitre, Mario Maglione montre sur des exemples du programme national Diadem comment on peut **accélérer la découverte de nouveaux matériaux grâce à l'IA.**

– Dans le deuxième chapitre, Gian Marco Rignanese passe en revue les progrès récents dans le domaine émergent **de l'informatique des matériaux qui combine la puissance des calculs *ab initio* à haut débit et l'IA.**

– Guillaume Fayet dans le dernier chapitre dresse un **panorama des travaux et perspectives de l'apport des nouvelles approches méthodologiques et de l'IA pour l'évaluation et la maîtrise des risques accidentels,** que ce soit dans un contexte réglementaire ou dans le développement de substances et de procédés plus sûrs.

La troisième partie est consacrée à l'émergence de l'IA, à la place qu'elle tient et aux projets de développement dans la R&D dans trois grands domaines d'applications industrielles.

– Les cosmétiques et la parfumerie représentés par Cosmetic Valley, le pôle d'activité de la filière.

– L'Énergie représentée par le groupe TotalEnergies.

– Les médicaments représentés par SANOFI.

Le dernier chapitre de cet ouvrage présente la conférence de clôture du colloque par Cédric Villani sous le titre **Intelligence artificielle pour la science et l'industrie**.

Cédric Villani est un mathématicien très connu, car il a eu la Médaille Field en 2010, qui est l'équivalent du prix Nobel pour cette discipline où le prix Nobel est inexistant. Professeur, chercheur, passionné de vulgarisation scientifique, homme politique, il a été chargé en 2017 par le Premier ministre Édouard Philippe d'une mission parlementaire sur l'IA, et en 2018 il a participé à la conférence « All for humanity », lors de laquelle il a présenté son rapport sur l'IA sous le titre « Donner un sens à l'IA ».

Nous espérons que cet ouvrage permettra aux chimistes et notamment aux plus jeunes, de mieux connaître ce sujet en plein développement dont la richesse des applications potentielles est encore difficile à prévoir.

Nous espérons qu'il apportera des informations utiles sur l'IA, non seulement pour le monde éducatif pour les programmes et l'actualisation des connaissances, mais aussi pour l'orientation des jeunes vers des métiers d'avenir.

Par ailleurs, cet ouvrage sera décliné en ressources pédagogiques facilement utilisables dans les différentes rubriques de notre médiathèque www.media-chimie.org.

L'aide à la formation et à l'orientation des jeunes, adaptée à l'évolution de notre Société et de ses problèmes est une mission fondamentale de la Fondation de la Maison de la Chimie.

Danièle OLIVIER
Vice-présidente de la Fondation internationale de la Maison de la Chimie

Partie 1

Intelligence artificielle : recherche et formation en chimie

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

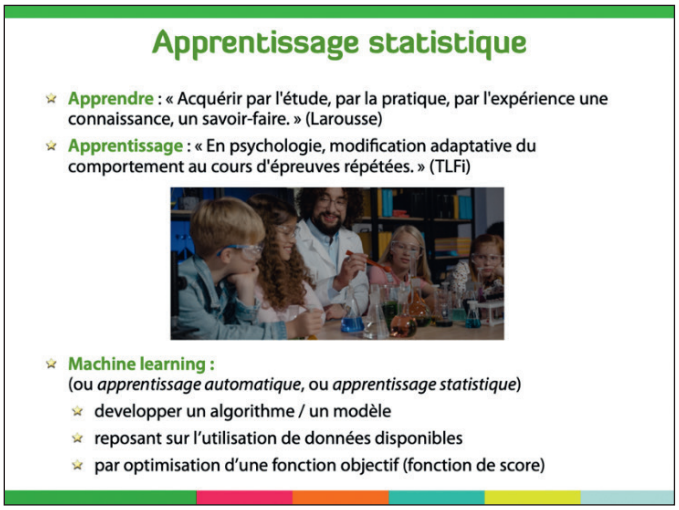


Figure 1

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.

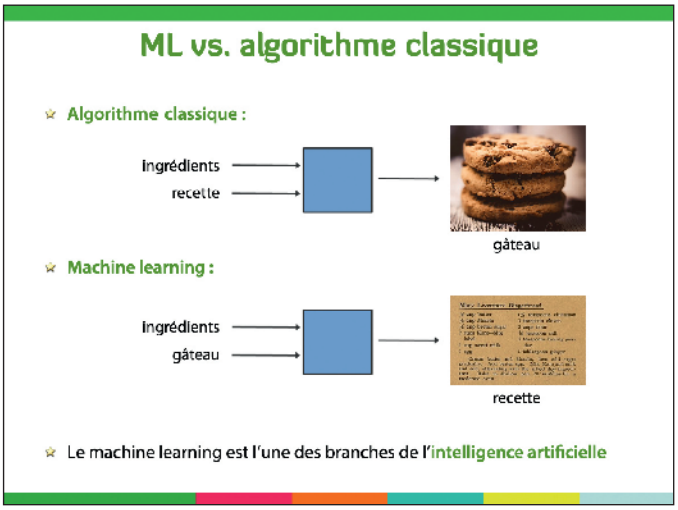


Figure 2

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.

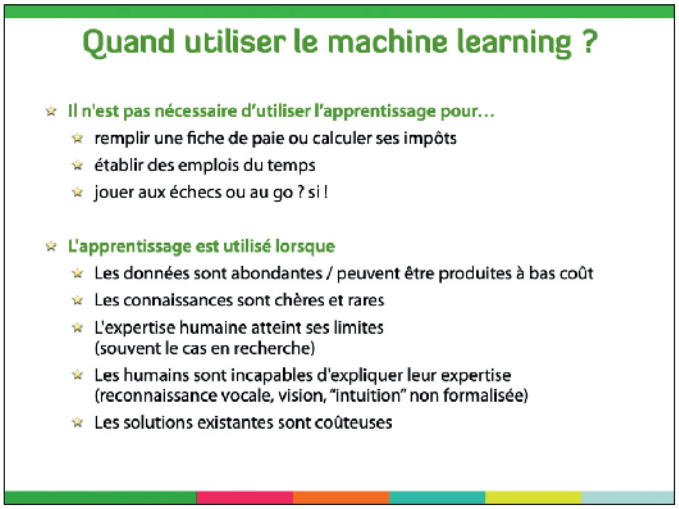


Figure 3

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***).

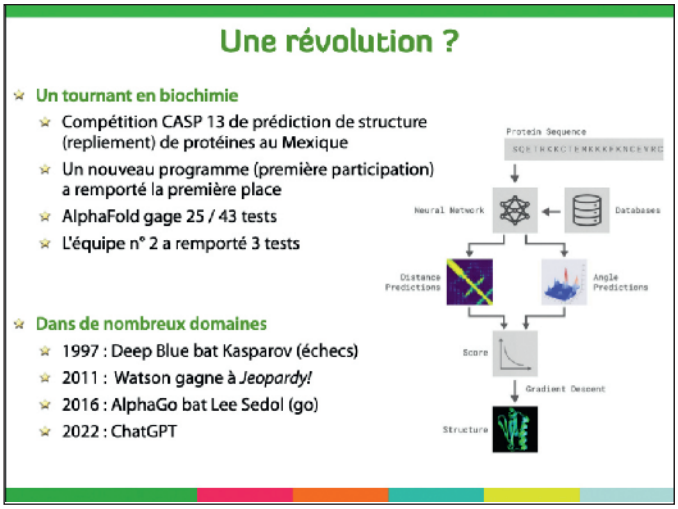


Figure 4

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].

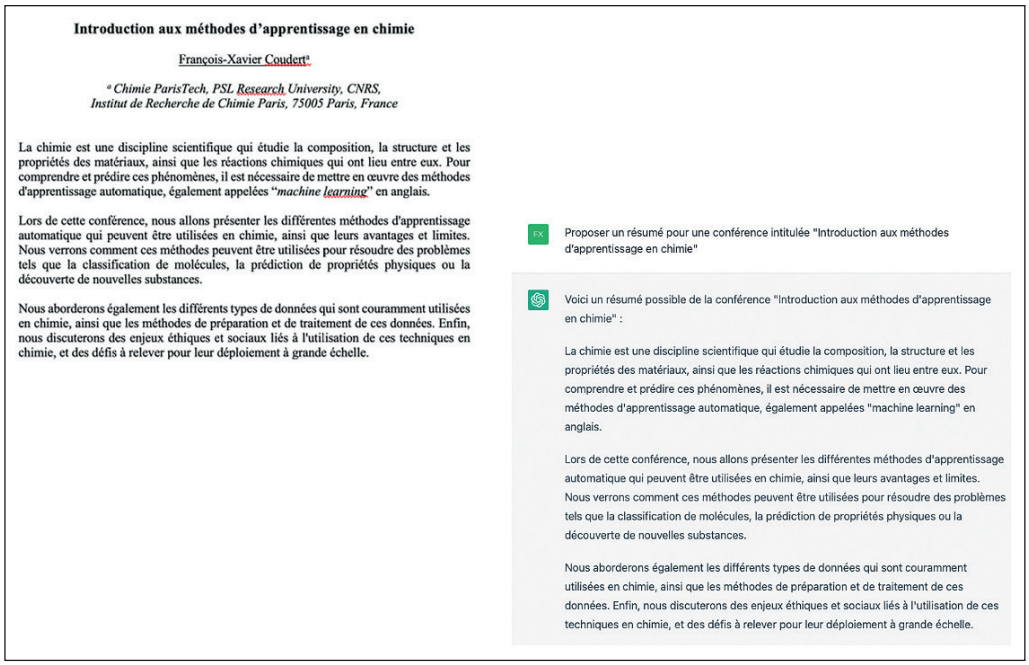
1. *Critical Assessment of protein Structure Prediction* (CASP) : littéralement, « évaluation critique de la structure des protéines », expérience visant à l'élaboration d'un état de l'art de la prédiction des structures (repliement) des protéines.
2. Algorithme d'optimisation.
3. *Jeopardy* est un jeu 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].

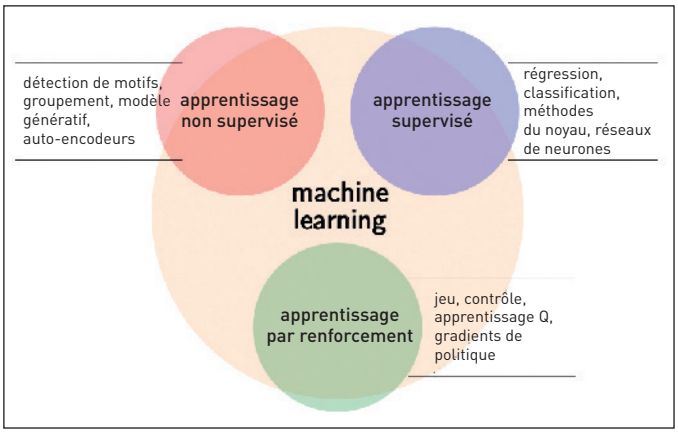


Figure 6

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

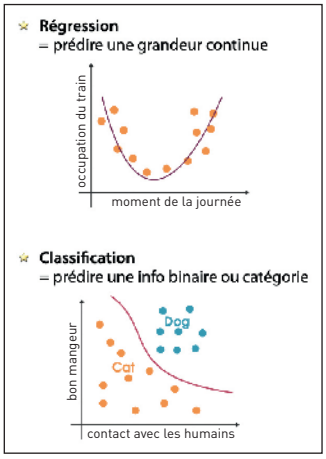


Figure 7

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.

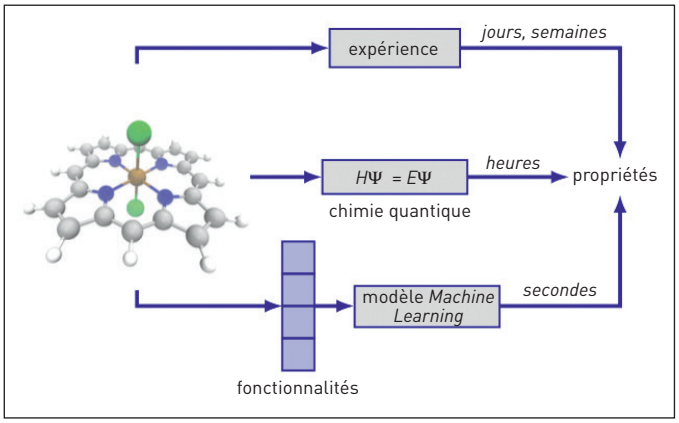


Figure 8

Différents méthodes pour obtenir les relations propriétés/structure.

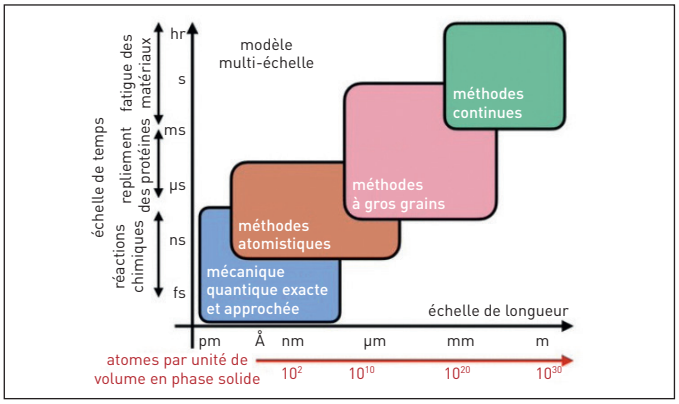


Figure 9

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.

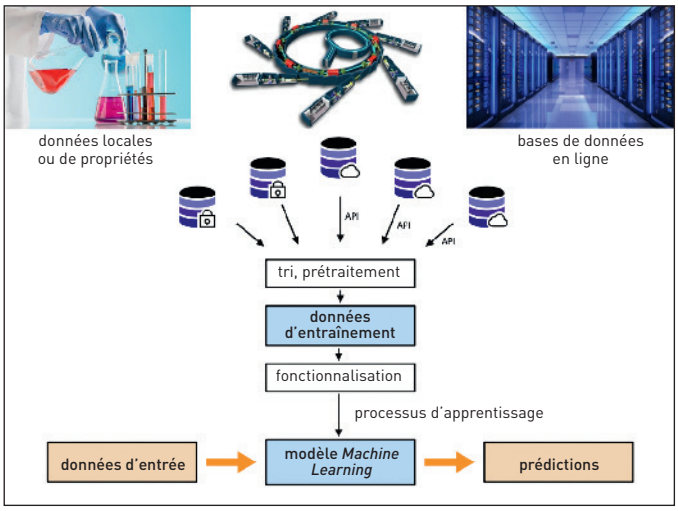


Figure 10

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

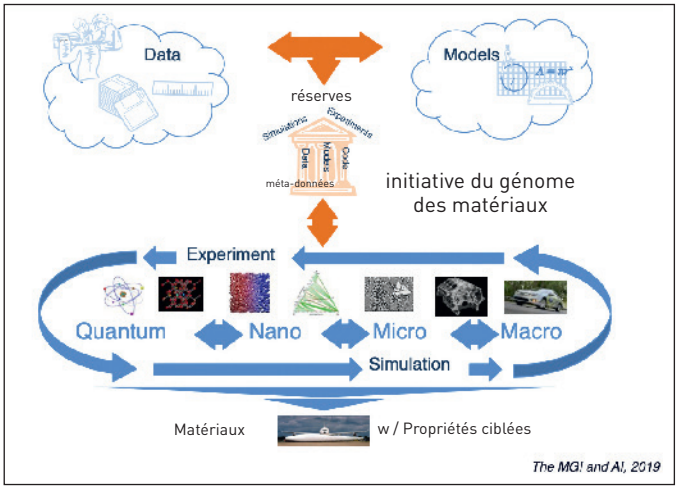


Figure 11

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.**

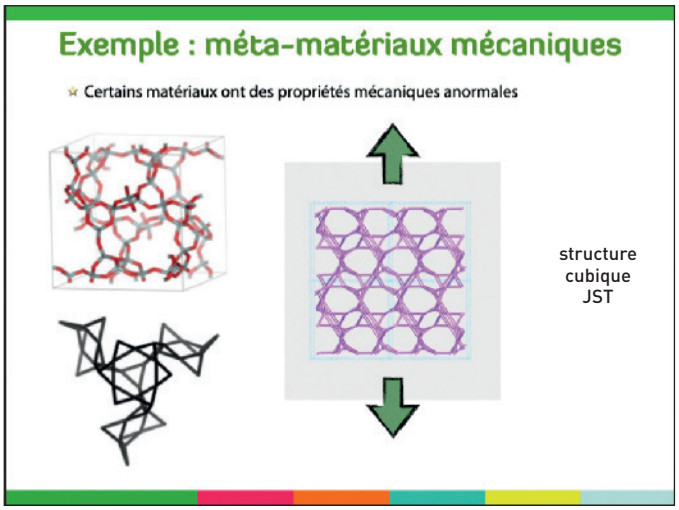


Figure 12

Des matériaux aux propriétés inhabituelles.

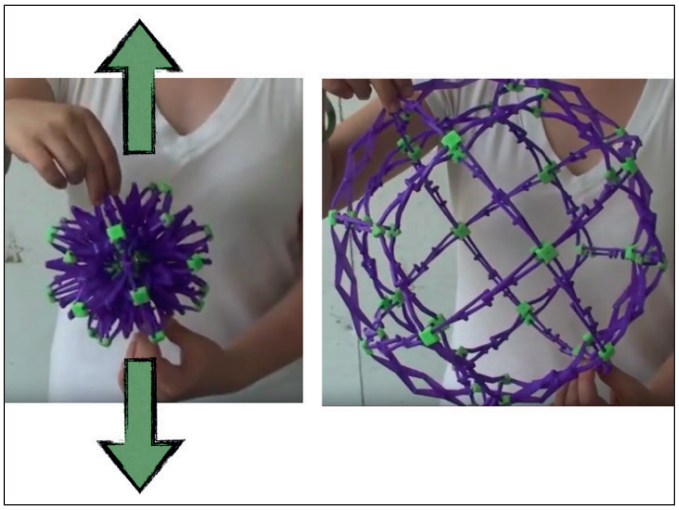


Figure 13

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

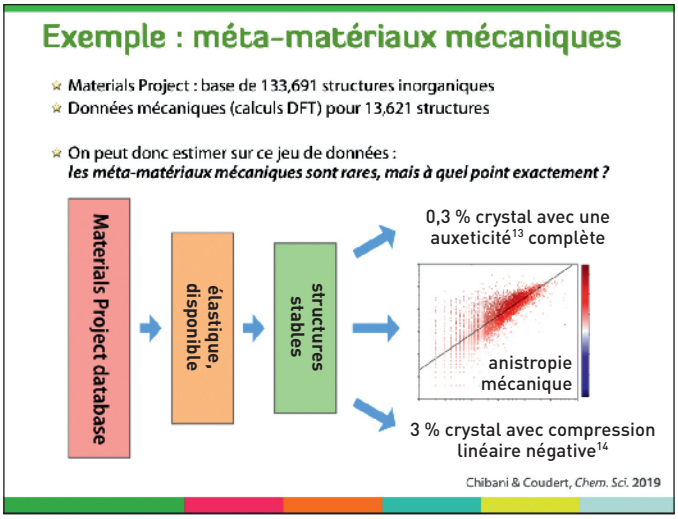


Figure 14

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.

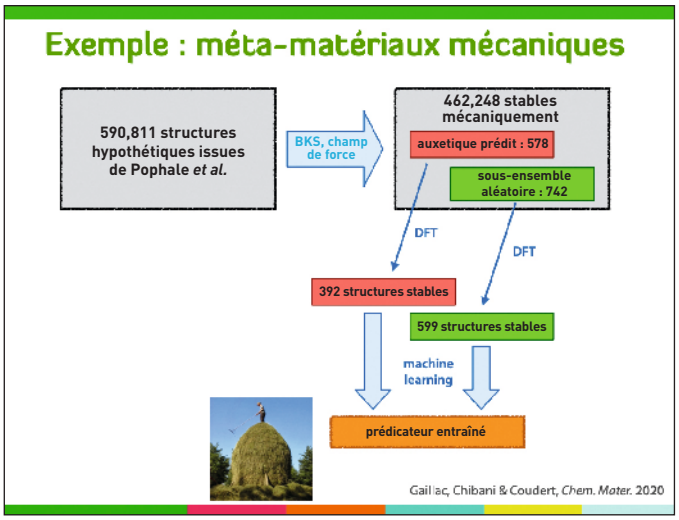


Figure 15

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 !

Carlo ADAMO

L'intelligence artificielle comme moteur dans la recherche en chimie

Professeur Carlo Adamo, Directeur de l'Institut de chimie des sciences de la vie et de la santé de l'École nationale supérieure de Chimie ParisTech et du CNRS, membre de l'Institut Universitaire de France, Paris.

1 Introduction : la chimie, un espace encore à explorer

Il existe encore un espace chimique énorme que l'on voudrait explorer de façon efficace. Actuellement, environ 350 000 produits chimiques sont commercialisés, mais on connaît au total 20 000 000 molécules organiques. Si on prend une seule de ces molécules, comme l'hexane, et que l'on considère les 150 substituants possibles à la place des hydrogènes, on arrive à 1 030 molécules que l'on pourrait créer.

L'exploration de cet espace chimique peut être faite dans deux sens (***Figure 1***) :

– L'apprentissage déductif[1] : la connaissance des lois par modélisation permet de générer l'information qui peut générer des datas[2].

1. Déductif : qui procède par déduction.
2. Datas : données.

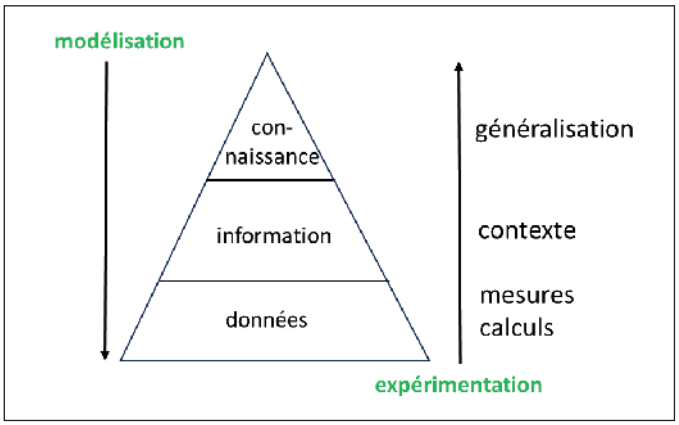


Figure 1

Schéma des types d'apprentissage.

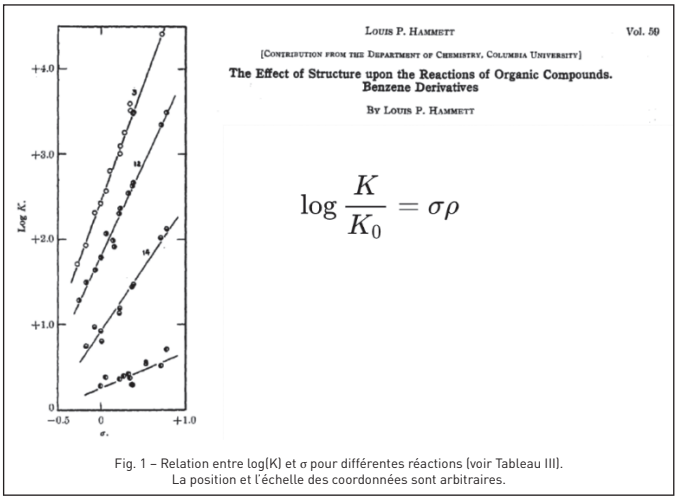


Figure 2

Extrait d'un article de Louis P. Hammett, introduisant la loi de Hammett (de J. Am. Chem ; Soc. 59, 1937, 96), concernant l'effet de la structure sur les réactions de composés organiques dérivés du benzène.
K = constante de réaction pour les benzènes substitués
K_0 = constante de réaction de référence
σ = constante du substituant
ρ = constante de réaction.
La constante σ dépend uniquement du substituant et la constante ρ dépend uniquement du type de réaction. Une centaine de valeurs de σ tabulées pour les substituants para et méta, mais pas pour ortho (à cause des effets stériques).

– Ou vice-versa l'apprentissage inductif[3] : à partir des données expérimentales, on génère des informations pour arriver à des lois générales.

Voyons quelques exemples de questions que peuvent se poser les chimistes.

1. Quelles sont les structures que je peux générer pour améliorer les propriétés d'une molécule déjà connue ? Puis-je synthétiser et produire ces molécules ? L'intuition chimique, donc l'expérience qu'on a comme expérimentateur ou comme théoricien, est un filtre important qui nous aide beaucoup dans notre recherche, mais la recherche des relations entre structure et propriétés est fondamentale en chimie, notamment en chimie physique pour répondre à ce type de question.

La première relation entre structure et propriété date de 1868, réalisée par Crum-Brown et Fraser qui ont eu l'idée de relier la structure d'une série de dérivées de la strychnine[4] à leur activité physiologique[5]. Naturellement, ils imaginaient qu'il existait une structure mère, mais ils n'étaient pas capables à l'époque de prouver

3. Inductif : qui procède par induction (opération mentale qui consiste à remonter des faits à la loi, de cas particuliers à une proposition plus générale), opposé à déduction.
4. La strychnine est une molécule chimique, un alcaloïde toxique extrait de la noix vomique. Cette substance est un poison.
5. La physiologie est la science étudiant le fonctionnement d'un organe ou d'un organisme vivant.

cette relation entre structure et propriété.

2. Chercher la raison pour pouvoir prédire. Un exemple important de ce type de démarche pour les chimistes est la loi de Hammet (***Figure 2***) qui date de 1937 et qui relie par une fonction linéaire la constante de vitesse[6] de réactions de benzènes substitués à la nature des substituants. Cependant si la méthode fonctionne très bien pour des substituants en méta[7] et en para, elle ne fonctionne pas pour des substituants en ortho, car dans ce dernier cas les mécanismes de réaction sont plus complexes.

C'est un exemple des limitations que l'on peut avoir dans ce type de relation et dans les bases des données qui y sont associées.

3. Un autre exemple qui commence à entrer dans le monde de l'IA est le projet Dendral. C'est une recherche des années 1960, conçue par deux informaticiens, un biologiste et un chimiste assez connus. L'objectif de ce projet était d'analyser avec l'aide de l'ordinateur des spectres de masses[8]. C'est le premier projet connu d'intelligence artificielle où le travail du chercheur est remplacé par le travail de la machine, avec une rapidité majeure mais tout en gardant l'expérience et la précision que peut avoir l'expérimentateur. Le but était de créer un système expert pour la connaissance d'un cas spécifique : les spectres de masse. À l'époque, on ne disposait pas de beaucoup de mémoire vive, 512 kilobytes, et le coût des systèmes était autour de 30 000 $ par mois.

Le système de fonctionnement est présenté sur la ***Figure 3***.

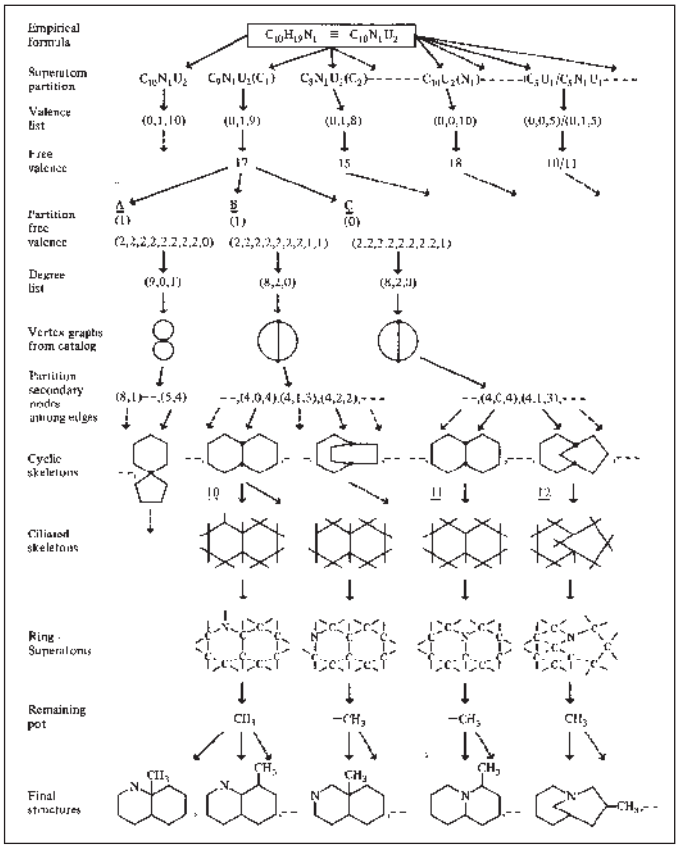


Figure 3

Principes du système Dendral.

6. Grandeur caractérisant la vitesse de réaction, dépendante uniquement de la température.
7. Ortho, méta, para : désignation de la position des substituants secondaires par rapport à un substituant principal sur un cycle benzénique.
8. La spectrométrie de masse est une technique d'analyse qui permet la détermination des masses moléculaires des composés analysés ainsi que leur identification et leur quantification.

Le terme DENDRAL vient de dendron[9], qui est un mot grec. Ils utilisaient des motifs de fragmentation[10] qui venaient de l'expérience pour l'input et ils faisaient des suppositions sous forme d'arbres de décision (***Figure 3***) pour arriver, à la fin, à interpréter le spectre. L'étape importante était la codification[11] de la structure moléculaire, et cela reste toujours un problème : comment expliquer à l'ordinateur la structure de nos systèmes ?

2 L'intelligence artificielle au service de la chimie

Le chapitre de François-Xavier Couderc introduit le concept d'intelligence artificielle. Le *machine learning*[12] est la technologie qui concerne les algorithmes et les méthodes qui utilisent une intelligence artificielle. Le *deep learning*[13] est un sous-domaine du *machine learning*. On peut espérer que l'IA, et les méthodes associées de *machine learning*, soient une opportunité pour trouver de nouvelles relations qui ont jusqu'à maintenant échappé à notre chère intuition chimique.

Des méthodes de *machine learning* sont présentées dans le chapitre de François-Xavier Couderc, elles sont nombreuses et les plus simples sont des régressions linéaires ou multilinéaires[14] ou de l'analyse comprenant principalement des arbres de décision. Beaucoup de chercheurs chimistes ont utilisé au moins une fois l'une de ces méthodes.

On peut utiliser l'IA pour beaucoup de choses en chimie, par exemple pour prédire des propriétés moléculaires, pour faire du design moléculaire[15], du *drug discovery*[16] pour mettre au point des nouveaux médicaments, ou encore pour prédire des produits ou faire de la rétrosynthèse[17]. Elle est aussi utilisée en chimie physique pour prévoir des propriétés thermodynamiques[18], des ph, des coefficients de partition[19],

9. Dendron : arbre en grec.
10. Input : représente l'ensemble des données fournies en entrée d'un programme informatique.
11. Codification : encoder l'information afin qu'un ordinateur puisse la déchiffrer.
12. Le *machine learning* est un sous-ensemble de l'intelligence artificielle (IA). Cette technologie vise à apprendre aux machines à tirer des enseignements des données et à s'améliorer avec l'expérience, au lieu d'être explicitement programmées pour le faire.
13. Le *deep learning* (ou « apprentissage profond ») est un sous-domaine du *machine learning*, lui-même faisant partie de la grande famille de l'intelligence artificielle. Il correspond à toutes les techniques de réseaux de neurones artificiels.
14. Régressions multilinéaires : enchaînement de plusieurs régressions linéaires.
15. Design moléculaire : confection de molécules.
16. *Drug discovery* : conception, découverte de médicaments.
17. Rétrosynthèse : synthèse à partir du produit pour obtenir les réactifs (méthode inverse d'une synthèse directe).
18. La thermodynamique est une branche de la physique qui étudie les propriétés des systèmes où interviennent les notions de température et de chaleur.
19. Coefficients de partition : coefficients de probabilité relatifs à une fonction et décrivant la probabilité d'avoir la molécule dans la configuration définie par ladite fonction.

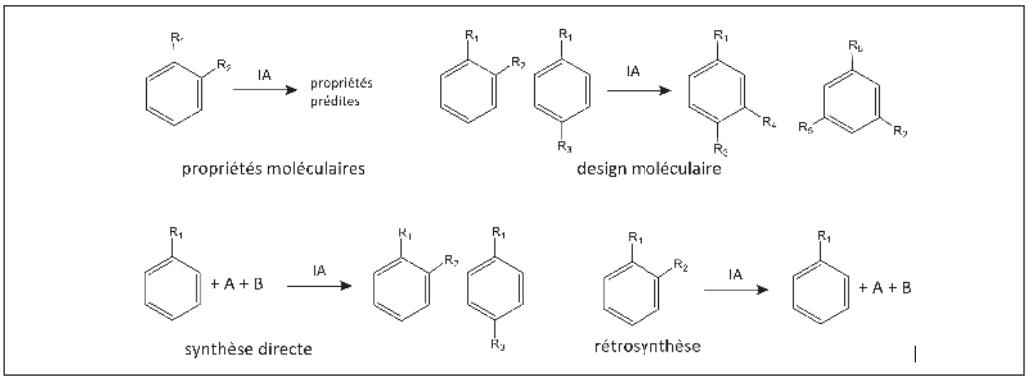


Figure 4

Applications de l'IA en chimie moléculaire.

et en spectroscopies IR[20] ou UV[21].

L'IA permet aussi de prévoir des propriétés liées au risque ou à la toxicité, ou enfin des propriétés de types biologiques bien spécifiques comme des concentrations dans des corps humains.

Ces dernières années, le nombre de publications qui associent le mot chimie aux mots « *machine learning* », « intelligence artificielle » ou « *deep learning* », a rapidement augmenté. Cela concerne tous les domaines de la chimie (beaucoup en chimie analytique, chimie physique, un peu moins en chimie organique) et sur des systèmes qui vont des petites molécules, aux alliages, ou aux polymères, pour citer quelques exemples.

2.1. Exemples d'application de l'IA en chimie moléculaire

L'IA permet de prédire des propriétés, le produit d'une synthèse, et de faire du design moléculaire ou de la rétrosynthèse (***Figure 4***).

En pharmacie l'IA est utilisée depuis de nombreuses années pour découvrir de nouveaux principes actifs.

C'est le cas par exemple dans le domaine des antibiotiques où l'on utilise des méthodes de *machine learning* avec des descripteurs[22] qui fournissent l'information chimique sur la molécule. La première difficulté est de trouver les descripteurs les plus adaptés à la propriété recherchée. Ensuite le *machine learning* est utilisé pour prévoir les activités.

On peut faire la même chose avec d'autres méthodes, comme celle des réseaux de neurones, en passant par une représentation codifiée de la molécule.

Pour illustrer l'intérêt de ce type d'application, prenons l'exemple (***Figure 5***) de la découverte d'un nouvel

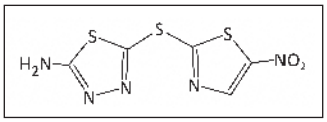


Figure 5

Exemple de machine learning pour la découverte de nouveaux antibiotiques : molécule.

20. IR : infrarouges.
21. UV : ultraviolets.

22. Descripteur : caractéristique spécifique à une molécule.

antibiotique, l'halicine, dans lequel les méthodes de *machine learning* ont été appliquées à une propriété spécifique qui était l'inhibition de la croissance d'*E. coli*[23]. Ensuite, les équipes de chercheurs ont fait du *screening*[24] sur plus de 107 millions de molécules, ce qui leur a permis de découvrir une nouvelle molécule qu'ils ont appelée halicine, en l'honneur de HAL qui était l'ordinateur intelligent dans l'*Odyssée de l'Espace*. Ce qui était intéressant, c'est qu'il n'y avait, dans leur approche, aucune supposition sur le mode de fonctionnement des médicaments ainsi que sur l'attribution des groupements chimiques. Ils ont finalement trouvé une structure différente de celles des antibiotiques traditionnels. Cette nouvelle molécule *a priori* fonctionne bien pour différentes bactéries et il semble que son mécanisme d'action soit lui aussi complètement différent de celui des autres antibiotiques : c'est une molécule qui va agir sur le flux de protons à travers une membrane cellulaire. L'halicine affiche une activité bactéricide contre une large gamme de pathogènes dont la tuberculose mycobactérienne et les bactéries résistantes aux antibiotiques.

L'intelligence artificielle apparaît dans cet exemple comme instrument de rupture par rapport à l'expérience et à l'intuition chimique et permet d'étendre l'espace chimique.

Mais comme cela a été rapporté dans le chapitre de François-Xavier Couderc, le choix du dispositif sur lequel on va appliquer la méthode de *machine learning* est important, notamment la définition et la qualité de la propriété que l'on va rechercher.

Applications de l'IA en synthèse organique

L'IA a aussi été utilisée en synthèse et en rétrosynthèse organique afin de prédire les produits de réaction selon les réactifs ou agents de réaction utilisés. Des méthodes d'apprentissage ont été mises au point par exemple par le MIT[25], qui sont toujours des méthodes de *machine learning*. On part des réactifs que l'on doit codifier en termes de graphes moléculaires, et on leur associe une propriété avec des réseaux de neurones. Cela permet ensuite de faire du screening avec de bons résultats pour de nombreuses réactions : des réactions organomagnésiennes, des halogénations, etc. La qualité du résultat varie un peu avec la réaction. C'est très précis pour les couplages de Suzuki[26], un peu moins sur l'addition de Grignard[27], mais ces méthodes

23. *E. coli* : *Escherichia coli* est une bactérie naturellement présente dans la microflore digestive de l'Homme et des animaux à sang chaud.
24. *Screening* : criblage.
25. MIT : Massachusetts Institute of Technology : Université prestigieuse américaine.
26. Réaction de Suzuki : réaction chimique entre deux groupements aryle (composés aromatiques).
27. Réaction de Grignard : réaction d'addition entre un halogénure organomagnésien et un composé organique porteur d'un groupe carbonyle, typiquement un aldéhyde ou une cétone, pour donner respectivement un alcool secondaire ou un alcool tertiaire.

restent tout de même des outils très puissants (*Figure 6*).

On peut aussi faire des prédictions de régiosélectivité[28]**,** dans lesquelles on couple par exemple des méthodes de *machine learning* avec des méthodes de chimie computationnelle[29] plus traditionnelles. On peut citer des méthodes qui sont basées sur la théorie de la fonctionnelle de la densité[30], donc sur les propriétés de la structure électronique, ou encore basées sur la régiosélectivité. On obtient alors des prédictions plus précises, autour de 90 %.

2.2. L'intelligence artificielle au service de l'industrie

Régulation Reach : ces méthodes de *machine learning* introduites à différents niveaux ont dépassé le contexte académique mais aussi le contexte industriel pour arriver au niveau de la réglementation européenne. Si on prend par exemple la réglementation REACH[31], on y trouve la promotion de méthodes alternatives à l'expérimentation animale. L'idée est de mettre sur le

28. Régiosélectivité : une réaction chimique est dite régiosélective si l'un des réactifs ou des intermédiaires réactionnels réagit préférentiellement avec certains sites d'un autre réactif parmi plusieurs possibilités. On obtient ainsi plusieurs isomères d'une molécule.
29. Chimie computationnelle : étude de la chimie à l'aide de l'outil informatique.
30. Théorie de la fonctionnelle de la densité : méthode de calcul quantique permettant l'étude de la structure électronique, en principe de manière exacte.

31. REACH : REACH est un règlement européen entré en vigueur en 2007 pour sécuriser la fabrication et l'utilisation des substances chimiques dans l'industrie européenne.

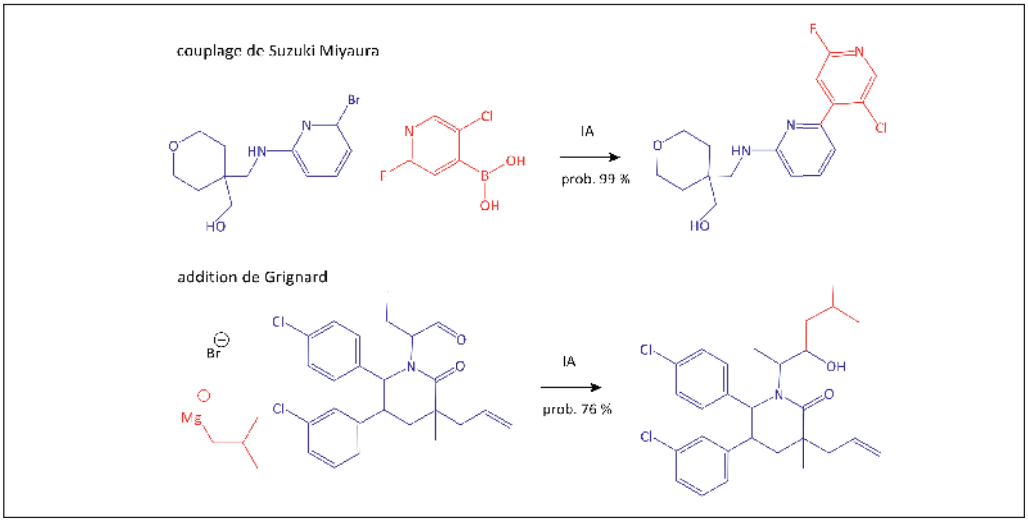


Figure 6

Application du machine learning sur deux types de réaction : le couplage de Suzuki et l'addition de Grignard.

même niveau les résultats expérimentaux et ceux obtenus avec des méthodes théoriques. L'une de ces méthodes est la méthode QSAR[32] (*Figure 7*) de modélisation semi-empirique qui traite les prédictions entre structure et activité. Une seconde méthode est la méthode SAR[33] qui est choisie avec des critères scientifiques incluant une demande d'applicabilité définie, une certaine robustesse ainsi que des objectifs finaux définis. Il est intéressant de remarquer que les critères scientifiques ont été définis par l'OECD[34] il y a des années.

La boîte à outil QSAR a été financée par l'OECD et l'ECHA[35] et consiste en une compilation de toutes ces méthodes et fait le lien avec l'approche plus traditionnelle de la chimie computationnelle qui utilise les méthodes développées en chimie théorique.

L'utilisation conjointe de la chimie théorique et de l'IA est efficace dans trois domaines :

– Donner des informations sur les mécanismes qui sont sous-jacents à des relations que l'on trouve à un niveau de l'IA.

32. QSAR : la modélisation semi-empirique QSAR (*Quantitative Structure Activity Relationship*) a comme objectif la prédiction des effets d'une variation de la structure moléculaire sur l'activité biologique.
33. SAR : la méthode est la même que la méthode QSAR sans l'aspect quantitatif.
34. OECD : l'Organisation de coopération et de développement économiques (OCDE) est une organisation internationale qui œuvre pour la mise en place de politiques meilleures pour une vie meilleure.
35. ECHA : Agence européenne des produits chimiques.

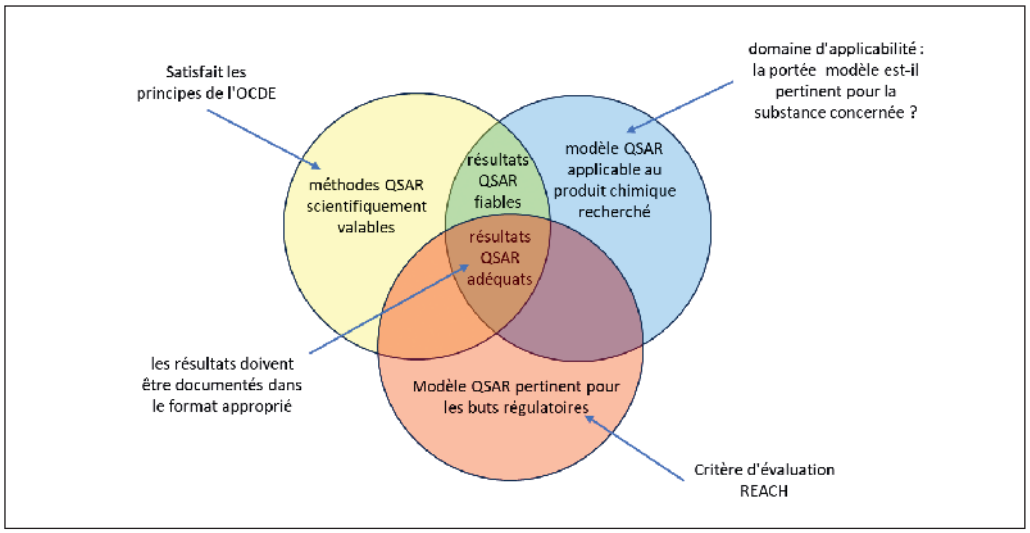


Figure 7

Utilisation des méthodes QSAR.

– Alimenter l'IA à travers des bases de données des descripteurs.

– Utiliser l'IA pour des nouvelles méthodes de chimie théorique.

Prenons quelques exemples.

Information sur les mécanismes réactionnels

La ***Figure 8*** est l'exemple d'un travail réalisé il y a plusieurs années sur les propriétés explosives, ayant pour objectif la recherche d'un mécanisme de réaction, en utilisant conjointement la chimie computationnelle et l'IA.

Les propriétés explosives sont des propriétés macroscopiques importantes pour REACH. Par exemple, dans le cas des composés nitrés du benzène, on peut avoir des relations simples linéaires du même type que la loi de Hammett entre les propriétés explosives et la nature des substituants. Ces relations sont valables pour des substitutions méta et para des cycles du benzène, mais en ortho on n'a pas la même précision parce que le mécanisme réactionnel est différent.

Une analyse avec les outils de la chimie computationnelle va permettre de comprendre les relations que l'on trouve avec des méthodes d'intelligence artificielle.

Couplage de l'IA et de la chimie computationnelle classique

Les méthodes de la chimie computationnelle classique (chimie théorique et chimie quantique) peuvent

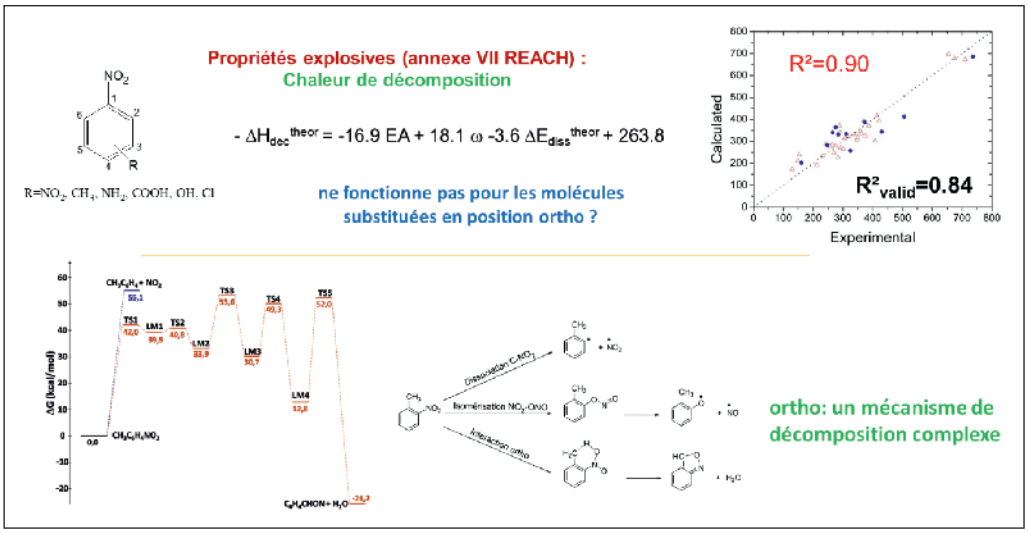


Figure 8

Mécanisme de réaction par l'association de la chimie computationnelle et de l'intelligence artificielle dans le cas des propriétés explosives de composés nitrés du benzène :
– haut : un modèle QSAR capable de prédire la chaleur de décomposition pour les composés substitués en position méta et para, mais qui ne fonctionne pas pour les molécules substituées en position ortho ;
– bas : mécanisme de décomposition de composés substitués en ortho (ortho-nitrotoluène).

être utilisées pour créer des bases de données à partir de descripteurs complètement théoriques, liés à la structure électronique des molécules. Naturellement, la rapidité et l'homogénéité de la base de données ainsi créée seront l'intérêt principal. Mais la limite de cette méthode arrive quand on se demande si le résultat obtenu permet de synthétiser ou non une nouvelle molécule. Il faudrait introduire des critères de synthétisabilité, mais c'est quelque chose qui n'est pas évident à réaliser.

La théorie de la fonctionnelle de la densité est la méthode la plus utilisée en chimie théorique dans laquelle on cherche à relier les propriétés des molécules à leurs structures électroniques, comme l'énergie, qui est une fonction de la densité électronique. Ce sont des méthodes qui ont été introduites par Kohn et Sham[36] (***Figure 9***) il y a quelques années. Cependant, dans tous les travaux réalisés, il nous manque un petit morceau pour obtenir l'énergie exacte du système et pour avoir la relation exacte entre l'énergie et les autres propriétés de la densité électronique.

Pour combler cette lacune, on peut imaginer d'utiliser l'intelligence artificielle au niveau de la modélisation. Le nombre d'articles publiés où le terme « théorie de la fonctionnelle de la densité » et « *machine learning* » sont associés, a augmenté rapidement ces dernières années.

L'utilisation des méthodes de l'IA pour définir le petit morceau d'Énergie qui manque au niveau de la théorie de la fonctionnelle de la densité (le terme Exc dans l'équation dans la ***Figure 9***) a été réalisée par une équipe de DeeperMind (une société de Google), l'université de Madrid et le Max Planck Institute. La ***Figure 10*** est un graphique qu'on aime bien montrer entre théoriciens : en ordonnée figure l'erreur sur l'énergie de réaction pour une série de systèmes (base de données GMTKN55), en kcal/mol. Les acronymes GGA, méta-GGA et hybrides représentent des méthodes « traditionnelles » de la chimie computationelle (méthodes DFT) et le label DM21 représente les résultats obtenus

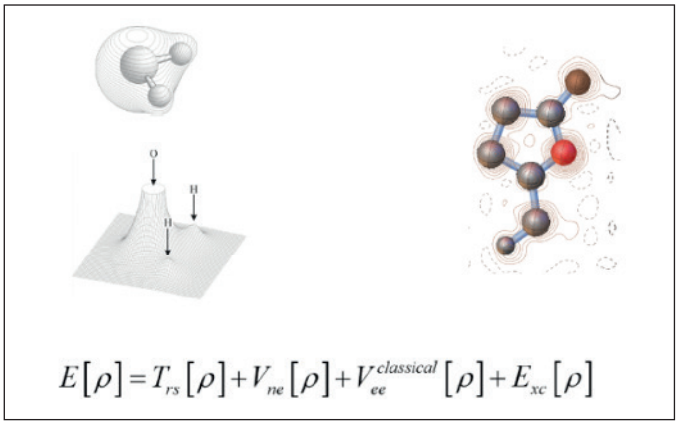


Figure 9

Brève introduction à la théorie de la fonctionnelle de la densité : la théorie de la fonctionnelle de la densité (DFT) est une méthode de chimie théorique utilisée en physique et en chimie pour étudier la structure électronique des atomes, des molécules et des phases condensées.
La DFT en tant qu'approche chimique : la connaissance de la densité électronique (ρ) est tout ce qui est nécessaire pour une détermination complète de l'énergie et de toutes les propriétés moléculaires. La DFT moderne est fondée sur l'approche proposée par Kohn, prix Nobel en Chimie en 1998.
Note : fonctionnelle en mathématique est une « fonction d'une fonction ».

36. Kohn et Sham : deux scientifiques ayant effectué des recherches sur la théorie de la fonctionnelle de la densité.

avec la méthode produite avec cette approche de l'IA (méthode DeeperMind21). On voit bien qu'on obtient un joli complément de la modélisation un peu plus traditionnelle.

Cependant, il y a donc encore un peu de travail à faire mais ces modèles sont assez hautement compétitifs dans leur domaine d'applicabilité. Comme exemple, la ***Figure 11*** présente l'application de cette méthode à des systèmes chimiques très simples, des agrégats de molécules d'eau plus ou moins protonées. En bleu, est représentée l'erreur sur l'énergie de l'interaction des molécules dans des petits agrégats obtenue avec les méthodes DeeperMind21, et, en rouge et vert, celles issues des méthodes plus traditionnelles de chimie computationnelle. Les grandes variations que l'on peut observer pour les données obtenues avec la méthode DM212 (ligne verte) indiquent qu'il y a encore un peu de travail à faire, mais ces modèles basées sur l'IA sont assez hautement compétitifs.

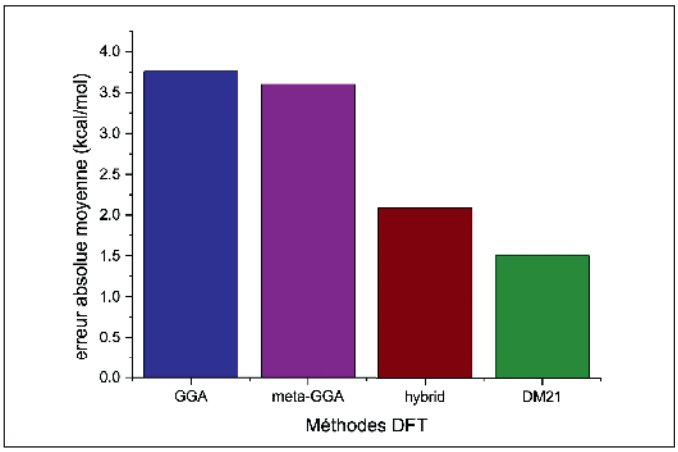


Figure 10

Performances de différentes méthodes DFT : le label DM21 fait référence à la méthode obtenue avec AI, tandis que les autres résultats ont été obtenus avec des méthodes plus traditionnelles. Données extraites de : Science 2021, 374, 1385-1389.

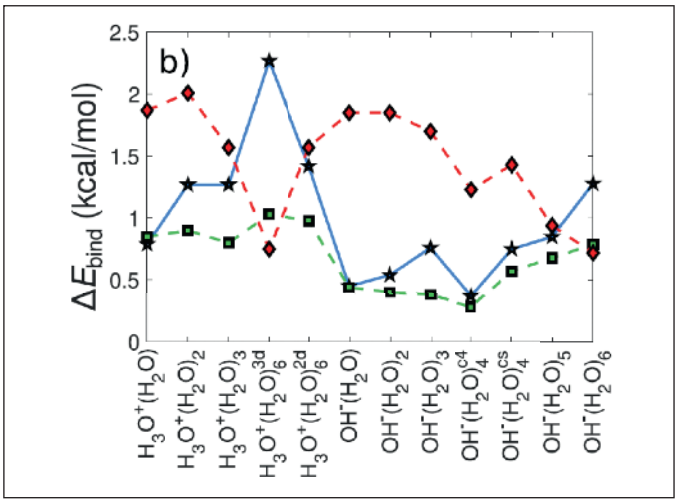


Figure 11

Comparaison de la méthode DeeperMind avec d'autres méthodes pour le calcul de l'erreur sur la valeur de l'Énergie d'interaction entre molécules d'eau plus ou moins protonées. Figure extraite de J. Chem. Phys. 156 (2022) 161103.

Conclusion

Des méthodes appropriées d'intelligence artificielle, au sens bien créées et bien pensées, sont donc une belle opportunité et même un défi en chimie. Elles vont nous permettre d'accélérer le processus de recherche (***Figure 1***), d'explorer une nouvelle voie. Elles sont assez complémentaires à d'autres types d'approche *in silico*[37] plus traditionnelles, mais, en même temps, il ne faut pas oublier que ces méthodes d'IA demandent des compétences spécifiques.

QUELQUES LECTURES SUPPLÉMENTAIRES

Lederberg, J., Feigenbaum, E.A., Lindsay, R.K. Applications of Artificial Intelligence for Organic Chemistry: The DENDRAL Project, McGraw-Hill, 1980.

Nieto-Draghi, C., Fayet, G., Creton, B., Rozanska, X.; Rotureau, P., De Hemptinne J.-C., Ungerer P., Rousseau B., Adamo C. A General Guidebook for the Theoretical Prediction of Physicochemical Properties of Chemicals for Regulatory Purposes. *Chem. Rev.* 2015, *115*, 13093-13164.

Strieth-Kalthoff, F., Sandfort, F., Segler, M.H.S., Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* 2020, *49*, 6154-6168.

Stokes, J.M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N.M., MacNair, C.R., French, S., Carfrae, L.A., Bloom-Ackermann, Z., Tran, V.M. Chiappino-Pepe, A., Badran, A.H., Andrews, I.W., Chory, E.J., Church, G.M., Brown, E.D., Jaakkola, T.S., Barzilay, R., Collins, J.J. A Deep Learning Approach to Antibiotic Discovery. *Cell* 2020, *180*, 688-702.e13.

Tkatchenko, A. Machine Learning for Chemical Discovery. *Nat Commun* 2020, *11*, 4125.

Baum, Z.J., Yu, X., Ayala, P.Y., Zhao, Y., Watkins, S.P., Zhou, Q. Artificial Intelligence in Chemistry: Current Trends and Future Directions. *J. Chem. Inf. Model.* 2021, *61*, 3197-3212.

37. *In silico* : se dit d'une méthode d'étude effectuée au moyen d'ordinateurs (dont les puces sont principalement composées de silicium), permettant d'analyser des données et de modéliser des phénomènes.

Guan, Y., Coley, C.W., Wu, H., Ranasinghe, D., Heid, E., Struble, T.J., Pattanaik, L., Green, W.H., Jensen, K.F. Regio-Selectivity Prediction with a Machine-Learned Reaction Representation and on-the-Fly Quantum Mechanical Descriptors. *Chem. Sci.* 2021, *12*, 2198-2208.

Kirkpatrick, J., McMorrow, B., Turban, D.H.P., Gaunt, A.L., Spencer, J. S., Matthews, A.G.D.G., Obika, A., Thiry, L., Fortunato, M., Pfau, D., Castellanos, L.R., Petersen, S., Nelson, A. W. R., Kohli, P., Mori-Sánchez, P., Hassabis, D., Cohen, A.J. Pushing the Frontiers of Density Functionals by Solving the Fractional Electron Problem. *Science* 2021, *374*, 1385-1389.

Boiko, D.A., MacKnight, R., Kline, B., Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature* 2023, *624*, 570-578.

Présentation de la Majeure Chimie&IA de l'ECPM

Description de l'apport de l'IA pour la préparation et la caractérisation de matériaux pour la santé

Sylvie Bégin-Colin, Professeur et ancienne directrice de l'ECPM, Université de Strasbourg (activités de recherche à l'Institut de Physique et Chimie des Matériaux de Strasbourg, CNRS-UNISTRA UMR7504).
Loïc Jierry, Professeur à l'ECPM, Université de Strasbourg (activités de recherche Institut Charles Sadron, CNRS-UPR22).

Sylvie BÉGIN-COLIN a obtenu en 1992 un doctorat à l'Université de Nancy dans le domaine de la science du génie des matériaux, puis a été 11 ans chargée de recherche au CNRS dans cette même discipline à l'École des Mines de Nancy. Elle devient en 2003 Professeur à l'École européenne de Chimie Polymères et Matériaux (ECPM) de l'Université de Strasbourg dont elle a assuré la direction entre 2014 et 2021. Elle effectue son activité de recherche au sein de l'Institut de Physique et de Chimie des Matériaux de Strasbourg. Son activité porte sur la synthèse, la fonctionnalisation et la structuration de nanoparticules d'oxydes pour des applications en santé, énergie et environnement.
Loïc JIERRY devient docteur de l'Université de Strasbourg en chimie organique, en 2003. Il a été attaché temporaire d'enseignement et recherche au sein du laboratoire du Professeur

Sylvie BÉGIN-COLIN *et* Loïc JIERRY

Jean-Marie Lehn à l'ISIS[1] en 2004, puis à l'École Normale Supérieure de Lyon en 2007. Entre 2005 et 2007, il a été chef de projets dans les entreprises MENARINI et ALSACHIM. Il rejoint l'ECPM en 2009 en tant que Maître de Conférences et devient Professeur en 2018 où il a mis en place et codirige actuellement avec Sylvie BÉGIN-COLIN la Majeure Chimie et Intelligence Artificielle (Chimie&IA).

Introduction au chapitre

La première partie du chapitre (auteur Loïc Jierry) traite de la mise en place d'une formation pédagogique unique, conduite par l'ECPM à l'Université de Strasbourg sur la formation d'une promotion d'ingénieurs chimistes spécialisés dans l'utilisation de l'intelligence artificielle en chimie. La structure pédagogique en cause s'intitule « Majeure Chimie&IA ».

La seconde partie (autrice Sylvie Begin-Colin) présente de façon détaillée les travaux de son équipe utilisant l'intelligence artificielle pour contrôler la synthèse de nanoparticules de différentes formes développées comme agents de diagnostic et thérapie pour le traitement de cancer. La réputation ancienne de « problèmes impossibles » a été battue en brèche par le recours judicieux à l'IA.

Première partie
Présentation de la Majeure Chimie&IA de l'ECPM

Introduction

Cette première partie présente la Majeure Chimie&IA, qui a été mise en place en 2019 à l'École européenne de Chimie Polymères et Matériaux de Strasbourg sous la direction de Sylvie Bégin-Colin.

La mise en place de cette nouvelle Majeure Chimie&IA est le fruit d'une longue réflexion de la part de groupes de travail spécifiques, alimentée par la consultation d'acteurs industriels partenaires de l'école ou impliqués dans le Conseil de l'ECPM, ainsi que celle d'anciens élèves. S'il est vrai que la presse « grand public » relaye régulièrement les miracles accomplis par les IA, ces dernières constituent également des outils de choix dans les activités de recherche académique mais aussi en R&D dans les entreprises. L'objectif de cette première intervention est de définir le besoin et le positionnement d'un nouveau profil d'ingénieur chimiste, compétent en science des données, recherché par les entreprises du monde de la chimie. Quel est son rôle et les connaissances/compétences qu'il doit posséder pour exercer ses fonctions ? Quels enseignements et sous quelle forme

1. Institut de Science et d'Ingénierie Supramoléculaires, Laboratoire de Chimie Supramoléculaire, basé à l'Université de Strasbourg/CNRS.

doivent-ils lui être dispensés pour assurer une formation adéquate de haut niveau et ainsi répondre aux besoins des entreprises de la chimie ?

1 Pourquoi une nouvelle Majeure à l'ECPM ?

Cela fait quelques années maintenant que l'intérêt de l'IA est apparu dans certains domaines de recherche, comme celui des sciences analytiques, où l'acquisition de grandes quantités de données est relativement aisée. Ainsi, on a vu apparaître dans les titres d'articles scientifiques de plus en plus de termes tels que « *machine learning*[2] », « *deep learning*[3] », ou « réseaux de neurones[4] ». Aujourd'hui, les outils proposés par l'IA, sont présents dans tous les domaines de la recherche en chimie, ce qui représente une petite révolution dans nos laboratoires. À l'ECPM, comme dans d'autres écoles d'ingénieurs chimistes, une cellule de veille scrute les nouvelles évolutions dans l'industrie, les différentes offres d'emploi ou de stage, afin de toujours adapter au mieux notre enseignement aux besoins de l'entreprise et si possible les anticiper. Dans des grands groupes de chimie, comme MICHELIN, SOLVAY ou NOVARTIS, des propositions d'embauche ou de stage ont été publiées mentionnant le besoin de compétences en IA, mais une ambiguïté demeurait car ces profils combinaient des compétences qui semblaient à priori très éloignées : s'agissait-il de propositions à l'attention de chimistes ou de « *data scientists* » ?

Ainsi, au cours de l'année 2017-2018, nous nous sommes donc demandés si le moment n'était pas venu de former des ingénieurs chimistes à l'IA, leur conférant ainsi des connaissances et des compétences solides en sciences des données. Pour répondre à cette question, nous nous sommes entretenus avec de nombreux anciens élèves qui occupent actuellement des postes à responsabilité dans diverses entreprises de la chimie, et nous avons également consulté les membres industriels du Conseil de l'ECPM. C'est précisément à cette période, que des acteurs industriels se sont adressés à la direction de l'école pour leur faire part de leur besoin à venir d'ingénieurs chimistes maîtrisant les nouveaux outils numériques tels que les « IAs ». Le premier d'entre eux est Philippe Robin, qui dirige la société ALYSOPHIL[5], une entreprise spécialisée dans le développement de procédé de production chimique en flux continu contrôlé et optimisé

2. Domaine scientifique qui est considéré comme une sous-catégorie de l'intelligence artificielle, consiste à laisser des algorithmes découvrir des modèles dans les ensembles de données.

3. Technologie principale du *machine learning*, algorithmes capables de mimer les actions du cerveau humain grâce à des réseaux de neurones artificiels.

4. Ensemble organisé de neurones interconnectés permettant la résolution de problèmes complexes.

5. ALYSOPHIL est une société qui développe un nouveau concept de chimie industrielle pour la production de molécules à haute valeur ajoutée en combinant la chimie en flux et l'intelligence artificielle.

par l'IA. Sa venue et ses propos nous ont confortés dans la pertinence de notre réflexion en nous interpellant : « Est-ce que vous êtes conscients qu'aujourd'hui dans l'entreprise, on a certes besoin de chimistes, mais on a besoin de chimistes qui ont des compétences en IA, qui savent ce que c'est et qui savent l'utiliser ? » Il s'agit d'un métier nouveau pour une position nouvelle au sein des entreprises. Ces discussions nous ont poussés à poursuivre et intensifier notre réflexion. Nous avons continué à travailler et à échanger avec d'autres entreprises comme CHEMINTELLIGENCE, MAYFAIR VILLAGE, L'ORÉAL ou MANE.

En début d'année 2019, la célébration des 100 ans de l'ECPM conviant l'ensemble des élèves de toutes les promotions précédentes a rassemblé énormément d'anciens élèves, ce qui a représenté une occasion unique de les interroger à travers un questionnaire leur demandant, entre autres : « Vous semble-t-il important d'intégrer des outils de sciences des données dans la formation de l'ingénieur chimiste à l'ECPM ? ». Bien entendu, nous nous attendions à une réponse majoritairement affirmative, mais nous n'imaginions pas que celle-ci serait à hauteur de 93 % (***Figure 1***) ! Un questionnaire plus complet a ensuite été envoyé à nos anciens élèves, ce qui nous a tout d'abord conforté dans la nécessité d'une formation à l'IA pour nos élèves ingénieurs, mais aussi apporté des premiers éléments d'informations précis sur les réels besoins des entreprises. Le besoin de ce nouveau profil d'ingénieur, déjà en 2019, nous est apparu évident, et nous avons ainsi anticipé que ce besoin allait s'intensifier. Il était donc nécessaire que l'ECPM non seulement forme ses élèves aux outils de l'IA mais réagisse aussi en proposant une nouvelle offre de formation en adéquation avec les attentes industrielles. Ainsi, nous avons entamé une course contre la montre pour pouvoir proposer une formation en chimie et IA adéquate, qui ouvrirait dès la rentrée universitaire 2019-2020, soit quelques mois plus tard.

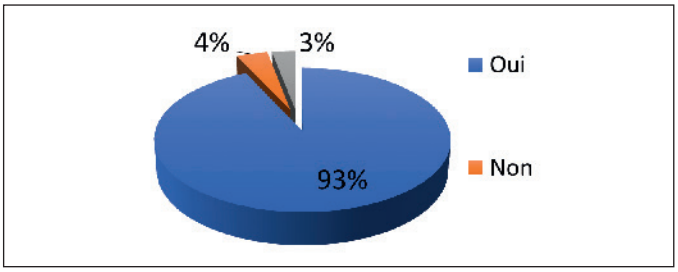


Figure 1

Statistiques des réponses des anciens élèves de l'ECPM à la question : « Est-ce qu'il est important selon vous d'intégrer des outils de sciences des données dans la formation de l'ingénieur chimiste à l'ECPM ? ». Question posée lors de la célébration des 100 ans de l'ECPM, en début d'année 2019 (bleu = OUI ; orange = NON ; gris = sans opinion).

2 Positionnement de l'ingénieur chimiste compétent en science des données au sein des entreprises de la chimie et compétences ciblées

Un aspect important a émergé très rapidement de nos consultations : cet ingénieur compétent en science des données sera avant tout un chimiste et non un *data scientist*. Il

doit être en mesure de comprendre parfaitement la nature des données qu'il manipule et savoir comment les exploiter. L'ingénieur chimiste compétent en science des données doit pouvoir parfaitement s'intégrer dans des équipes de R&D&I et dans celles dédiées à l'exploitation des outils numériques.

Nous avons établi les principales compétences de ce nouveau profil d'ingénieur à l'interface de la chimie et de la science des données :

• **Concevoir** des cibles moléculaires, macromoléculaires ou de formulations originales et des nanomatériaux hybrides nouveaux.

• **Prédire** les propriétés d'une molécule, d'un matériau, d'un polymère, d'une formulation.

• **Optimiser** une formulation de matériaux moléculaires ou inorganiques ou hybrides pour cibler des propriétés.

• **Piloter** des procédés de production pour optimiser les paramètres de synthèse.

3 Que doit-il savoir faire et comment le former ?

3.1. Programme d'enseignement

Nous avons établi la nécessité de dispenser des bases solides en programmation dans différents principaux langages. Les élèves qui intègrent l'ECPM ont déjà quelques notions en langage Python et il est essentiel de consolider et d'approfondir la maîtrise de celui-ci ainsi que de les former à d'autres langages comme le langage R. En plus de ces langages de programmation, les élèves doivent maîtriser des logiciels comme Knime ou Matlab ainsi que l'utilisation de logiciels d'interprétation de données généralement multidimensionnelles comme WEKA ou encore ISIDA. Nous ne visons pas de former les élèves à l'établissement d'algorithmes complexes, mais souhaitons qu'ils soient capables de trouver et d'utiliser des algorithmes déjà existants. En effet, beaucoup d'algorithmes sont accessibles et leur nombre ne fait que croître. Certes, connaître des langages de programmation, des logiciels de visualisation, maîtriser l'utilisation d'algorithmes sont des éléments de formation importants, mais le plus important réside dans la capacité à construire **une base de données fiables** à partir de données expérimentales acquises dans l'entreprise mais aussi à partir de bases de données chimiques existantes (PubChem, SciFinder…). Ce nouvel ingénieur chimiste doit savoir où sont ces bases de données, savoir les extraire des données variées, nettoyer ces données et **vérifier leur fiabilité**.

Ces savoir-faire de base ont tout d'abord été visés en termes de formation. Rappelons encore que nous souhaitons former avant tout des ingénieurs chimistes, pas des statisticiens, ni des informaticiens. L'étape suivante a été d'intégrer cette formation en IA dans l'ensemble des enseignements dispensés à l'ECPM, sans altérer la qualité de la formation en chimie.

3.2. Dispense de l'IA aux élèves ingénieurs de l'ECPM – Création de la nouvelle Majeure Chimie&IA en septembre 2019

De notre travail de prospection en amont est ressorti que l'apprentissage des sciences des données nécessite de la pratique qui s'acquiert principalement au cours de travaux par projets. C'est un apprentissage long, qui ne peut prendre la forme d'un cours de quelques dizaines d'heures dispensé en fin de formation. D'autre part, l'enseignant/encadrant doit être chimiste à la base, afin de pouvoir illustrer au mieux l'intérêt de son enseignement.

Il était donc important de trouver un nombre conséquent d'heures disponibles pour l'enseignement en science des données, étalées dans le temps, dispensées par des enseignants chimistes et compétents dans l'utilisation des outils numériques. Sans porter atteinte à la qualité de la formation d'ingénieur chimiste dispensée à l'ECPM.

Pour remplir ce cahier des charges, nous avons mis à profit une des particularités de l'ECPM, qui est celle d'inclure un grand nombre d'heures de travaux pratiques (TP). En effet, dès leur intégration en première année à l'ECPM et jusqu'à la fin de la 2e année, les élèves suivent en moyenne et de façon alternée, une semaine de cours et une semaine de TP. Nous avons décidé que des élèves appelés à être formés aux sciences des données suivraient uniquement des TP dits « de base » (soit un total de 8 semaines) mais qu'ensuite ces semaines de TP seraient dédiées à l'enseignement de l'IA : un ingénieur chimiste formé par l'IA quand les autres majeures forment des ingénieurs chimistes par l'expérimentation.

Le profil d'enseignants compétents à la fois en chimie et en science des données était aussi rare au moment de la création de cette Majeure que les ingénieurs présentant le profil que nous souhaitions former. Notre étude prospective nous avait permis d'identifier un ensemble d'entreprises, qui étaient prêtes à nous aider dans cette mission de formation en se proposant d'intervenir. Parmi ces entreprises, nous avons pu compter tout d'abord et surtout sur **ALYSOPHIL** dirigée par Philippe Robin, puis sur **CHEMINTELLIGENCE** dirigée par Thomas Galeandro, Alexandre Bouqueau (un Alumni de chez MANNE) et ensuite sur l'entreprise **MAYFAIR VILLAGE** avec Christophe Wilmort, également un de nos Alumni. Ils nous ont tous beaucoup aidés dans la phase de construction des enseignements et interviennent également dans la formation auprès des élèves. Des enseignants du Master Chemoinformatique de la Faculté de chimie de l'Université de Strasbourg, co-dirigé par le Docteur Gilles Marcou, et des enseignants de l'école Télécom Physique Strasbourg interviennent également dans des enseignements très spécifiques de science des données. Par ailleurs, nous bénéficions du soutien d'Alsace Tech (regroupement des écoles

d'ingénieurs, en management et architecture d'Alsace) pour tout ce qui va être acculturation à l'IA pour l'ensemble des élèves de l'école.

La formation d'ingénieur ECPM standard s'étend sur trois années à partir de BAC + 2 (***Figure 2***). Au cours de ces 3 années, les élèves suivent un tronc commun de chimie pour acquérir les bases en chimie moléculaire, sciences analytiques, ingénierie des polymères et « matériaux et nanosciences ». À partir du milieu de la deuxième année, les élèves sont invités à choisir ce qu'on appelle une majeure expérimentale qui peut être chimie moléculaire, ingénierie des polymères, matériaux de fonction et nanosciences et sciences analytiques. Ils sont ainsi divisés en 4 majeures, ce qui fait en moyenne 25 élèves par majeure.

Conscients de la nécessité de commencer la formation spécifique en science des données très tôt dans le cursus ingénieur, en particulier parce que les enseignements par projet requièrent de s'étaler sur plusieurs semaines, nous avons décidé d'ouvrir aux élèves la Majeure Chimie&IA dès la fin du premier semestre de la 1re année (***Figure 2***). À la fin du 1er semestre de la 2e année, les élèves de cette Majeure Chimie&IA vont, comme les autres, être invités à choisir une majeure « expérimentale » en même temps que l'enseignement de science des données : ils suivront tous les mêmes enseignements de chimie (cours, travaux dirigés, projets) que les élèves qui ne suivent pas la Majeure Chimie&IA ; en revanche, ils n'assisteront pas aux travaux pratiques en chimie à partir du 2e semestre de la 1re année. Mais ils effectuent un stage en laboratoire de 6 semaines lors de la 2e année au cours duquel ils peuvent choisir de revenir à l'expérimentation ou combiner les deux : expérimentation et IA. Les élèves de cette Majeure effectuent les mêmes stages et projets que les autres élèves. Ils peuvent choisir d'effectuer un stage de 1re année qui est soit l'interface de l'IA et de la chimie ou un stage purement « chimie ». Pour les stages de 2e année et de 3e année, les élèves doivent choisir une thématique spécifique « IA et chimie ». Ils participent également à un projet élève entreprise en 2e année qui s'étale sur 6 mois et traite d'une problématique industrielle « Chimie et IA ».

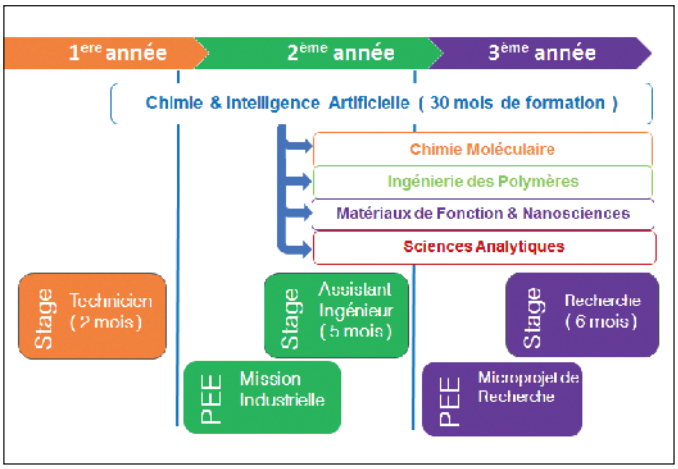


Figure 2

Représentation schématique de la formation d'ingénieur chimiste à l'ECPM, et implémentation de la nouvelle Majeure Chimie&IA.

4 Enseignements dispensés au sein de la formation en Majeure Chimie&IA

4.1. Enseignements dispensés en première année

Voici une brève description des différents enseignements que l'on dispense à l'école. En première année, les élèves suivent à partir du second semestre :

• Un enseignement d'**Introduction à la science des données** où ils vont revoir les bases de la programmation en langage Python, des études statistiques, de la modélisation et introduction aux bases de données générales.

• Un enseignement « **Chemical Databases »** en anglais sur la construction de bases de données chimiques fiables.

• Un cours d'introduction au « **Machine Learning »** et aux différents modèles utilisés, appliqués à la chimie.

• Une introduction aux **Systèmes d'exploitation**, **Linux** et autres.

• Un renforcement en **Mathématiques** avec beaucoup de notions de statistiques et une introduction au langage R.

Les élèves assistent également régulièrement à des conférences industrielles et chaque année, des conférenciers sont invités pour sensibiliser les élèves à l'intérêt de l'IA en chimie.

4.2. Enseignements dispensés en deuxième année

La formation en deuxième année s'intensifie avec de nombreux enseignements :

• Des enseignements de « **Data Mining** » où on poursuit tout ce qui est exploitation de « **Machine Learning** », arbre de décision, réseau de neurones, etc.

• Un approfondissement conséquent de l'enseignement sur les modèles prédictifs et sur les algorithmes d'optimisation.

• Un cours de renforcement en **Python**.

• Un cours dédié au **traitement d'images** et à **Matlab**.

• Un module très important de **modélisation moléculaire**, **DFT**[6] ; il s'agit de branches de la chimie qui génèrent des données très utiles pour le développement d'outils à base d'IA.

• Un cours d'approfondissement en **Langage R**.

• Une **mission industrielle**. Il s'agit d'un projet proposé par un industriel où des élèves de la Majeure Chimie&IA vont répondre à une problématique réelle.

• Le **stage de deuxième année** (4-5 mois) où les élèves vont obligatoirement choisir un stage à l'interface de la chimie et de l'IA.

• 6 semaines de « **TP Projet Recherche IA** » en chimie et IA qui ont lieu dans les laboratoires de l'école. Dans le cadre de ces 6 semaines, l'élève va travailler avec un tuteur expérimental au sein du laboratoire mais il travaillera également avec un tuteur « IA ».

6. *Density Functional Theory* : théorie de la fonctionnelle de la densité, méthode de calcul quantique permettant l'étude de la structure électronique.

4.3. Enseignements dispensés en troisième année

Pendant la troisième année, trois enseignements sont prévus pour le moment, mais d'autres devraient compléter cette première liste très rapidement :

• Approfondissement du langage **Python**.

• Utilisation d'outils de type « **Deep Learning** » en chimie.

• Enseignements spécifiques au **traitement du signal**.

La troisième année à l'école peut se dérouler suivant différents parcours au choix pour les élèves de la Majeure Chimie&IA :

• Cursus standard de l'école et réalisation d'un stage de 6 mois à partir de fin janvier dans le domaine de la chimie et de l'IA.

• 3e année en contrat de professionnalisation dans une entreprise (alternance en entreprise et à l'ECPM).

• 3e année en contrat de professionnalisation « augmenté » dans une entreprise avec « coaching » par l'entreprise Mayfair Village. Ce contrat de professionnalisation « augmenté » permet à l'élève de bénéficier d'une supervision tripartite entre un superviseur de l'ECPM, un tuteur dans l'entreprise d'accueil et un tuteur de l'entreprise Mayfair qui va le guider au sein de l'entreprise sur le projet Chimie et IA qui lui a été confié.

• 3e année en master de Chemoinformatique à l'Université de Strasbourg.

• 3e année dans une université étrangère grâce aux divers partenariats internationaux de l'école avec un programme proche de celui de l'école.

Une année césure entre la 2e année et la 3e année en entreprise ou sur un projet personnel est possible. Depuis 2023, les élèves Chimie et IA ont la possibilité de suivre un programme Erasmus Mundus sous une thématique axée Chemoinformatique. Ces élèves ont une bourse pour suivre des enseignements dans un autre pays et reviennent ensuite suivre leur troisième année à l'école.

5 Bilan – Où sont les élèves et que font-ils dans l'entreprise ?

Que savent faire les élèves formés par la Majeure Chimie&IA et quelle est leur place dans l'entreprise ? Voici un ensemble de quelques titres de stage, qui peut donner une idée des sujets traités par nos élèves :

– Compréhension et prédiction des propriétés à froid des biocarburants par une approche chimiométrique appliquée aux données de CPG[7].

– Parallélisation du traitement d'image de wafers afin d'y accélérer l'exploitation par « Machine Learning ».

– Conception de capteurs optiques assistée par des IAs.

– Structuration d'une base de données pour la conception de produits innovants pour l'impression 3D.

7. Chromatographie en Phase Gazeuse : méthode de séparation des composés chimiques volatils ou semi-volatils d'un mélange.

Les élèves en stage de deuxième, troisième année ou en contrat de professionnalisation ne sont jamais « à la paillasse ». Ils sont derrière un ordinateur et supervisés par quelqu'un qui est, soit un *data scientist*, soit un chimiste qui s'est formé en interne à la science des données. Au sein des entreprises, les élèves de la Majeure Chimie&IA sont généralement au sein des équipes de R&D. Tous les élèves ont déclaré qu'ils ont eu besoin de leurs connaissances en chimie pour réaliser leur stage. Cela conforte le besoin de ce profil d'ingénieurs compétents en chimie et qui savent ce que sont les outils à base d'IA.

Le succès de notre formation tient certainement à la mise en place d'un enseignement avec les conseils d'industriels dans ce domaine et à la pédagogie par projet. En outre, ces élèves sont très « adaptables » : ils savent se former très rapidement et facilement aux outils spécifiques développés dans les entreprises. L'école a aujourd'hui environ 50 élèves dans la Majeure Chimie&IA et la 2e promotion va être diplômée cette année 2023.

Des élèves intègrent spécifiquement notre école pour suivre cette formation. Il y a par ailleurs actuellement un réel engouement des élèves pour les enseignements de science des données ; les élèves semblent parfaitement conscients de l'importance de ces nouveaux outils aujourd'hui et pour demain.

Les élèves de la Majeure Chimie&IA ne rencontrent pas de problèmes pour trouver des stages et il y a actuellement plus d'offres de stage que d'élèves disponibles, ce qui nous incite à augmenter les effectifs de cette majeure. Des grands groupes chimiques comme des start-ups recrutent nos élèves.

Nous sommes dans un processus d'amélioration continue de cette formation et nous sommes très attentifs aux retours des industriels et des élèves après les différents stages ou projets effectués. Notre objectif à moyen terme est de renforcer les partenariats industriels et de monter une Chaire Industrielle pour consolider le programme des études et développer une recherche en partenariat avec l'industrie.

Deuxième partie
Présentation d'une étude incluant l'IA, réalisée par l'équipe de Sylvie Begin-Colin et des élèves de la Majeure Chimie&IA en stage laboratoire :
Iron oxide nanoplates synthesis guided by artificial intelligence to design theranostic iron oxide nanoparticles combining photothermal and magnetothermal therapies[8].

Introduction

L'objectif de cette deuxième partie est de montrer les travaux effectués en six semaines par deux étudiants de 2e année issus de la Majeure Chimie et IA et comment leur collaboration nous a aidés dans nos recherches en synthèse de nanoparticules adaptées à une application visée.

Notre objectif en recherche est de développer des nanoparticules[9] d'oxyde de fer fonctionnalisées pour diagnostiquer et traiter des cancers. Nous avons constaté que des nanoparticules de forme plaquette étaient prometteuses, mais malheureusement le rendement en nanoplaquettes lors de la synthèse était plutôt faible et les paramètres de synthèse très nombreux. L'approche IA nous a permis de trouver très rapidement les paramètres optimaux pour obtenir un rendement élevé en nanoplaquettes.

8. Synthèse de particules nanométriques d'oxyde de fer guidée par intelligence artificielle, dans le but d'avoir des nanoparticules d'oxyde de fer théranostiques, combinant des possibilités de diagnostic par IRM et une thérapie par photothermie et/ou hyperthermie magnétique.

9. Particules dont la taille est comprise entre 1 et 100 nanomètres (millionième de millimètre).

1 Contexte de l'étude

Notre élément de départ, ce sont des nanoparticules d'oxyde de fer fonctionnalisées à partir desquelles nous développons des plateformes théranostiques[10] (***Figure 3***), c'est-à-dire des composés capables de cibler spécifiquement des organes malades, de les imager (diagnostic), de traiter ces organes, de suivre l'effet du traitement par imagerie

10. Néologisme construit à partir des termes thérapie et diagnostic, qui correspond à une nouvelle approche médicale visant à privilégier le développement simultané de ces aspects.

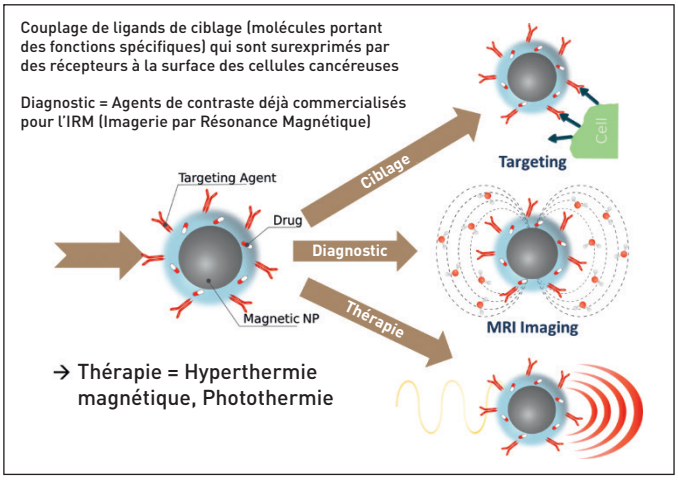


Figure 3

Ingénierie de nanoparticules à base d'oxyde de fer magnétique combinant des propriétés de ciblage, d'imagerie par IRM et de thérapie par hyperthermies.

et d'adapter le traitement suivant cette imagerie pour une meilleure prise en charge de la maladie du patient (nanomédecine personnalisée). Un enjeu est de pouvoir tester différents traitements avec une seule formulation de nanoparticules. À l'heure actuelle, il y a donc un fort engouement pour ces nanoplateformes ou nanobjets théranostiques.

2 L'hyperthermie magnétique

Une des particularités de certaines nanoparticules d'oxyde de fer magnétique est qu'elles peuvent s'échauffer lorsqu'elles sont soumises à un champ magnétique alternatif[11]. Elles permettent de procurer une thérapie par hyperthermie magnétique exploitant le fait que les cellules tumorales sont beaucoup plus sensibles aux élévations de température que les cellules saines. Une société, MagForce, est déjà en phase clinique II[12] et elle teste aussi ce traitement par hyperthermie magnétique sur d'autres types de cancers. Leurs études ont montré que combiner l'hyperthermie magnétique avec la chimiothérapie[13] ou la radiothérapie[14] donnait de très bons résultats. Le problème actuel est qu'il est nécessaire d'injecter dans la tumeur une grande quantité de nanoparticules pour y produire un effet d'hyperthermie magnétique. D'intenses recherches sont menées pour obtenir une nanoparticule ayant un potentiel de chauffe plus important que ce que l'on a actuellement.

Pour résumer le mécanisme actuel expliquant le phénomène d'hyperthermie magnétique : vous avez une nanoparticule portant un moment magnétique, vous la soumettez à un champ magnétique alternatif et la particule s'échauffe localement parce que, soit le moment magnétique tourne dans la particule, soit c'est la particule qui tourne dans ce milieu.

Pour avoir de bons résultats en hyperthermie magnétique, il y a des paramètres extrinsèques qui sont importants : la fréquence[15] et l'amplitude de champ magnétique, la viscosité du milieu, ou encore les conditions cliniques d'utilisation. Mais il y aussi des paramètres intrinsèques liés aux nanoparticules comme leur anisotropie[16] et il a été montré que des nanoparticules qui ont des formes anisotropes sont intéressantes pour augmenter l'anisotropie et donc le pouvoir chauffant. Ainsi, des nanocubes de 19 nm ont montré des capacités de chauffage très élevées par rapport à des nanoparticules de forme sphérique. Mais d'autres formes sont également intéressantes à étudier comme la forme « plaquette ». De plus, les

11. Champ magnétique dont le sens varie périodiquement.
12. Phase d'essai pendant laquelle le médicament est testé sur un échantillon de 50 à 100 patients.
13. Traitement par des substances chimiques.
14. Application thérapeutique des rayonnements ionisants pour détruire les cellules cancéreuses.
15. Nombre de changements de sens du champ magnétique par seconde.
16. Propriétés liées à la cristallinité et à la forme des nanoparticules.

nanoparticules d'oxyde de fer peuvent également s'échauffer sous un faisceau laser et on parle de traitement par photothermie[17].

Il y a donc un grand intérêt à développer une nanoparticule qui chauffe par hyperthermie magnétique et qui chauffe aussi par photothermie. L'effet de la forme des nanoparticules ne semble actuellement pas important pour avoir un effet photothermique mais il est important pour la magnétothermie.

3 Les différentes formes de nanoparticules

3.1. Synthèse des formes sphériques

La synthèse des nanoparticules sphériques est réalisée par la méthode de décomposition thermique. Elle consiste à décomposer thermiquement (2 500-400 °C) un précurseur de fer[18] en présence d'un surfactant[19], l'acide oléique, dans un solvant organique. Cette technique permet d'obtenir des nanoparticules non agrégées et très stables colloïdalement (***Figure 4***).

Pour nos études, nous avons utilisé deux types de précurseur de fer : des stéarates de fer avec des rapports, stéarate/fer, de 2 et 3 (***Figure 5***). Ces deux précurseurs conduisent à des nanoparticules sphériques mais nous avons modifié les conditions de synthèse pour obtenir d'autres formes : des cubes ou des plaquettes.

3.2. Synthèse des nanocubes et des nanoplaquettes

Un moyen pour obtenir des nanoparticules anisotropes par cette méthode de synthèse est de modifier la nature des surfactants : nous utilisons pour les nanosphères, l'acide oléique, mais si l'on remplace une partie de l'acide oléique par de l'oléate de sodium[20] plus chélatants, les oléates vont

17. Production de chaleur à partir de l'énergie lumineuse.
18. Espèce chimique qui, une fois dans le milieu réactionnel, va libérer du fer.
19. Tensioactif.
20. Sel, base conjuguée de l'acide oléique.

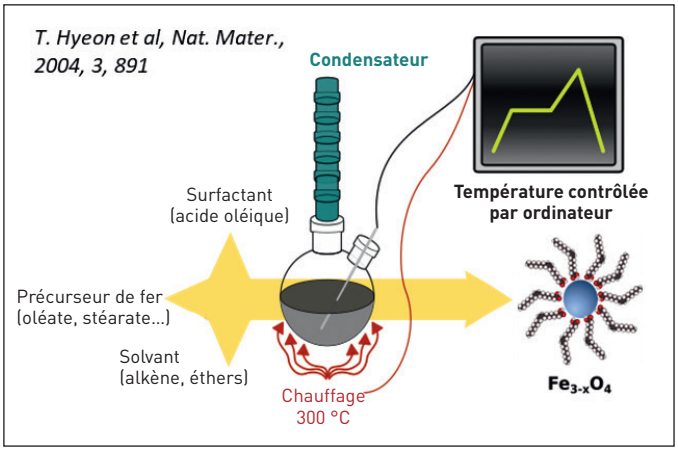


Figure 4

La décomposition thermique : méthode de synthèse des nanoparticules.

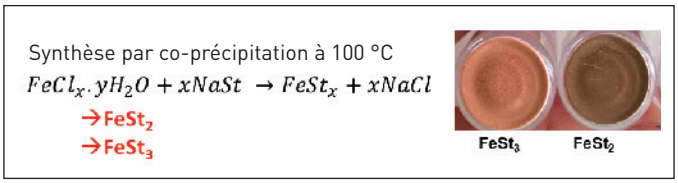


Figure 5

Précurseurs de fer.

s'adsorber sur les faces les plus énergétiques des germes et bloquer la croissance selon ces faces, permettant ainsi de contrôler la croissance suivant certaines faces et d'avoir des formes non sphériques, plus anisotropes.

Cependant, le type de précurseur peut également avoir un effet sur la forme des nanoparticules, tout comme le rapport des surfactants, la quantité de surfactant, la température de réaction, le temps de réaction et la vitesse de chauffage et il y a ainsi de nombreux paramètres à contrôler. Dans un premier temps, nous avons étudié l'influence du précurseur et de la quantité de surfactant. Avec le précurseur de fer, FeSt2 (***Figure 5***), nous avons observé que nous obtenions plutôt des nano plaquettes avec le ratio 80/20 alors que le précurseur, FeSt3 (***Figure 5***), conduit à des cubes mais avec des mélanges de différentes formes pour chaque condition expérimentale.

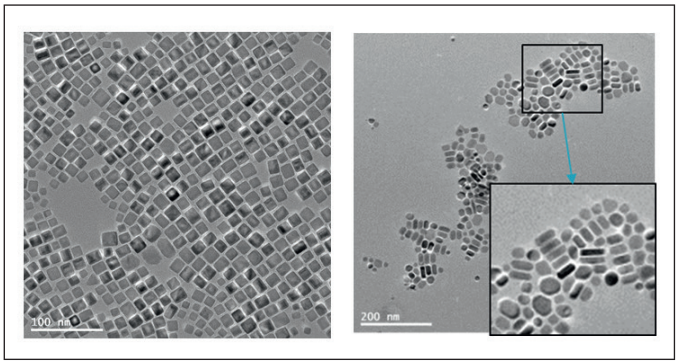


Figure 6

Image par microscopie électronique à transmission de (gauche) nanocubes formés à partir de FeSt2 grâce à une vitesse de montée en température élevée et (droite) de nanoplaquettes (avec un mélange de sphères et cubes) formées à partir de Fe(St)2 grâce à une faible vitesse de chauffage et une étape de nucléation à 180 °C.

Nous avons optimisé les synthèses et réussi à obtenir des nanocubes avec des bords très réguliers avec le précurseur FeSt2 en travaillant avec une vitesse de montée en température élevée (***Figure 6 gauche***). Les nanocubes ont beaucoup été étudiés et la formation de nanoplaquettes était intéressante à tester pour combiner hyperthermie magnétique et photothermie. Nous avons donc recentré notre étude sur le précurseur FeSt2 en faisant varier différents paramètres. Si on diminue la vitesse de chauffage et si on introduit une étape de nucléation[21], des nanoplaquettes étaient obtenues (***Figure 6 droite***). Cependant, quelles que soient les conditions, un mélange de nanoplaquettes avec des cubes, des sphères est obtenu. Le rendement en nanoplaquettes était au maximum de 50 %.

Au vu du nombre très élevé de paramètres à varier pour optimiser la synthèse des nanoplaquettes et surtout le rendement en nanoplaquettes, nous avons fait appel à l'IA : quels sont les paramètres les plus importants pour obtenir un rendement élevé en nanoplaquettes ?

4 Approche IA pour résoudre ce problème

Nous avons décidé d'accueillir dans notre laboratoire des stagiaires de la majeure Chimie et IA, qui ont été supervisés par Thomas Galeandro de ChemIntelligence. Nous avions déjà 35 expériences,

21. Formation de germes qui vont ensuite croître.

qui pouvaient être utilisées comme données. Cependant, le comptage du % de forme par image se fait manuellement et nous avions besoin d'un programme d'analyse d'image pour réaliser des études statistiques automatiques précises. La difficulté est que, sur une image, les plaquettes peuvent être visualisées de profil (visualisation du « côté » / « tranche » de la plaquette) ou de face (forme hexagonale ou triangulaire ou entre les deux) (***Figure 6 droite et Figure 8***). Cette analyse d'images est donc complexe, diverses méthodes sont testées : **OpenCV module** qui est une extension de Python[22] tout d'abord et actuellement nous travaillons sur la méthode **crYOLO**. Ce travail d'analyse d'image est encore en cours.

Ensuite, pour l'optimisation des paramètres de synthèse des nanoplaquettes, différents modèles prédictifs ont été étudiés : *linear regression* (régression linéaire), *decision tree* (arbre de décision), *aleatory forest* (forêt aléatoire)... Au final, c'est le modèle ExtraTree qui s'est avéré le plus efficace (***Figure 7***). Le logiciel de *machine learning* nous a permis de construire des graphiques et nous a demandé de varier divers paramètres, même des paramètres conduisant à des expériences qui ne marchent pas, à des fins de comparaison. Les paramètres les plus importants se situent dans les cases 1 ou –1 (***Figure 7 bas***). Les paramètres les plus importants sont le type de précurseur et la vitesse de montée

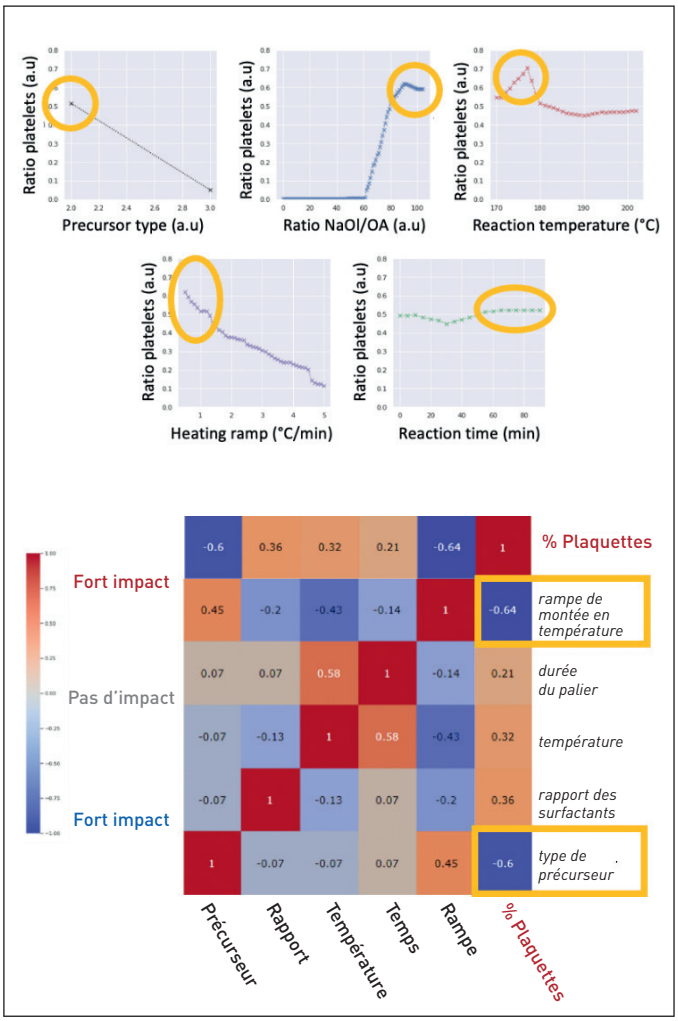


Figure 7

Impact des différents paramètres de synthèse sur la formation des nanoplaquettes (haut) et « ExtraTree Heatmap » pour comprendre l'influence des paramètres (bas).

en température, sachant que le temps, la température et le rapport des surfactants ne sont pas déterminants mais peuvent être optimisés avec bénéfice.

Ainsi, avec environ 10 expériences supplémentaires, nous avons réussi à atteindre un rendement de 80 % de plaquettes

22. Logiciel de codage.

(*Figure 8*). Cette approche permet de mieux comprendre l'effet des paramètres et la corrélation entre eux. Ces nanoplaquettes ont montré des propriétés de chauffage par hyperthermie très intéressantes à faible concentration. Nous poursuivons ce travail d'optimisation du rendement.

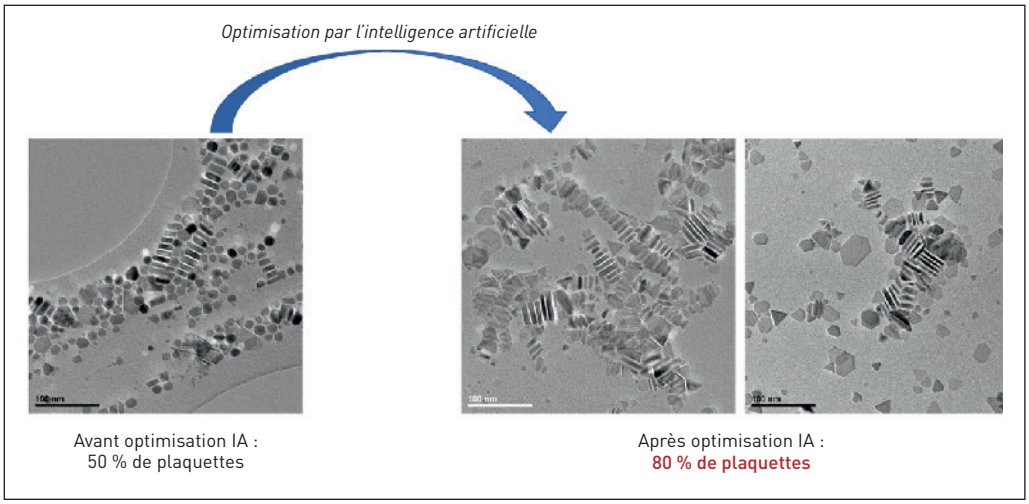


Figure 8

Pourcentage de plaquettes obtenu avant (50 %) et après optimisation (80 %) par l'approche IA.

Conclusion
Les vrais succès de l'IA en chimie sont nés

Au-delà des espoirs théoriques voire journalistiques que tire l'intelligence artificielle dans son sillage, il y a incontestablement des réalisations que l'on peut sans exubérance taxer de « révolutions techniques ». La chimie n'apparaît pas comme la discipline technique la mieux placée pour en profiter au regard de son extrême diversité de composés (les « dix puissances beaucoup » de molécules) et d'arrangements et interactions entre eux et de la complexité des paramètres qui les déterminent… Mais depuis une dizaine d'années, les recherches des chimistes font aussi des miracles et s'apprêtent

à en faire de plus grands ; pas seulement sur la gestion des données, le domaine principal de l'IA, mais sur son utilisation des connaissances propres des sciences de la chimie. Un signal qui ne trompe pas, c'est l'effort maintenant fortement croissant des industriels dans ce domaine pour optimiser/adapter leurs recherches et leurs procédés.

Dans ce chapitre, on voit, très concrètement et sans langue de bois, la description des efforts de l'école d'ingénieurs chimistes strasbourgeoise, l'EPCM, pour former des étudiants que des industriels maintenant demandent instamment. En un petit nombre d'années, cette école s'est fait la réputation locale de centre de ressources pour l'exploitation, et donc indirectement pour la réputation de l'IA en chimie. Et les travaux sur la théranostic (pour la mise au point de médicaments anticancéreux) présentés à la suite nous laissent sans voix : comment le contrôle impossible de la texture détaillée des systèmes moléculaires mixtes, organisés, mal organisés, réputé impossible est-il devenu possible ? Et les conséquences sont là dans la mise au point de nouveaux anticancéreux par des méthodes, à commencer par les exploitations si pertinentes de l'IA, qui nous laissent pantois par leur ambition scientifique et technique.

Très beau chapitre qui sait marier le concret, la pédagogie, le recours aux possibilités de la science et l'attraction pour les applications, le tout dans une atmosphère à la fois académique et industrielle très séduisante.

Sébastien PREYS

L'expérience d'Ondalys dans la formation continue aux outils opérationnels de la chimiométrie et du *machine learning*

Sébastien Preys, Chef de projet Data Science et Machine Learning[1] chez Ondalys.
Monsieur Sébastien Preys est, depuis 2006, docteur en chimiométrie[2] à l'INRA (Institut national de la recherche agronomique) de Montpellier. Sa thèse portait sur l'analyse de données multi-blocs[3] combinant différents jeux de données multivariées, provenant de différentes techniques analytiques. Depuis 15 ans, il travaille chez Ondalys pour fournir des services et des formations.

1. *Machine learning* : apprentissage automatique.
2. La chimiométrie est un outil utilisé afin d'extraire de l'information pertinente et utile à partir de données physicochimiques mesurées ou connues brutes. Il est basé sur la construction, puis l'exploitation d'un modèle de comportement à l'aide d'outils statistiques.
3. La chimiométrie est un outil utilisé afin d'extraire de l'information pertinente et utile à partir de données physicochimiques mesurées ou connues brutes. Il est basé sur la construction, puis l'exploitation d'un modèle de comportement à l'aide d'outils statistiques.

Introduction

Qu'est-ce qu'Ondalys ? C'est une entreprise créée il y a 20 ans (***Figure 1***), au départ, une jeune entreprise innovante et maintenant une petite équipe d'une dizaine de *data scientists*[4] avec de l'expérience, un leader dans la chimiométrie et le *machine learning* en France. Ondalys travaille principalement pour l'industrie de process[5] : pharmaceutique, biotechnologie, chimie, agroalimentaire, cosmétique, etc. Son activité principale est d'accompagner les industriels dans la mise en place des outils de *machine learning* et d'intelligence artificielle.

Ondalys conduit aussi une activité de formation continue, qui fait l'objet du présent chapitre. Enfin, Ondalys distribue également les logiciels d'analyse de données de certains de ses partenaires.

Voici quelques mots-clés que nous aborderons dans ce chapitre : *data mining*[6], calibration spectroscopique, modèle prédictif, combinaison de capteurs, ou monitoring[7], qui implique une supervision de procédés en continu[8] ou

4. *Data scientist* : scientifique des données, exploite les données de l'entreprise.
5. Industrie de process : industrie dans laquelle les matières premières subissent une transformation chimique et/ou physique.
6. *Data mining* : exploration de données, analyse de données depuis différentes perspectives et le fait de transformer ces données en informations utiles, en établissant des relations entre les données ou en repérant des patterns.
7. Monitoring : supervision.
8. Procédé continu : mode de production industriel destiné à fabriquer, construire ou traiter des matériaux sans interruption.

Figure 1

Résumé des informations générales sur l'entreprise Ondalys.

en batch[9], et plans d'expériences[10]. Nous sommes une équipe d'ingénieurs qui a développé un réseau de partenaires de logiciels sur lesquels nous formons des professionnels.

Un projet de *machine learning* ou de data science[11] peut s'aborder au niveau de l'audit chez le client pour voir ce qui a été fait et donner notre avis (***Figure 2***). L'activité principale d'Ondalys réside dans le conseil et l'étude de faisabilité ; le *proof of concept*[12], pour tester les outils de *machine learning* sur des applications précises. Si le projet est satisfaisant, l'idée est d'accompagner le client à développer et valider des modèles et à réaliser l'implémentation logicielle. Comme un modèle possède un cycle de vie, il faut vérifier qu'il ne dérive pas, qu'il continue à bien prédire et cela sous-tend toute une activité de maintenance de modèle. La formation continue, qui est le thème de ce chapitre, représente à peu près 20 % de notre activité, en incluant du coaching après une formation accompagnée.

Ce chapitre revient d'abord sur la sémantique et sur certains concepts de chimiométrie et de *machine learning* qui font partie du vaste champ de l'intelligence

9. Procédé batch : mode de production industriel par lots avec interruption.
10. Plans d'expériences : consiste à sélectionner et ordonner les essais afin d'identifier, à moindres coûts, les effets des paramètres sur la réponse du produit.
11. Data science : science des données.
12. *Proof of concept* : preuve de concept.

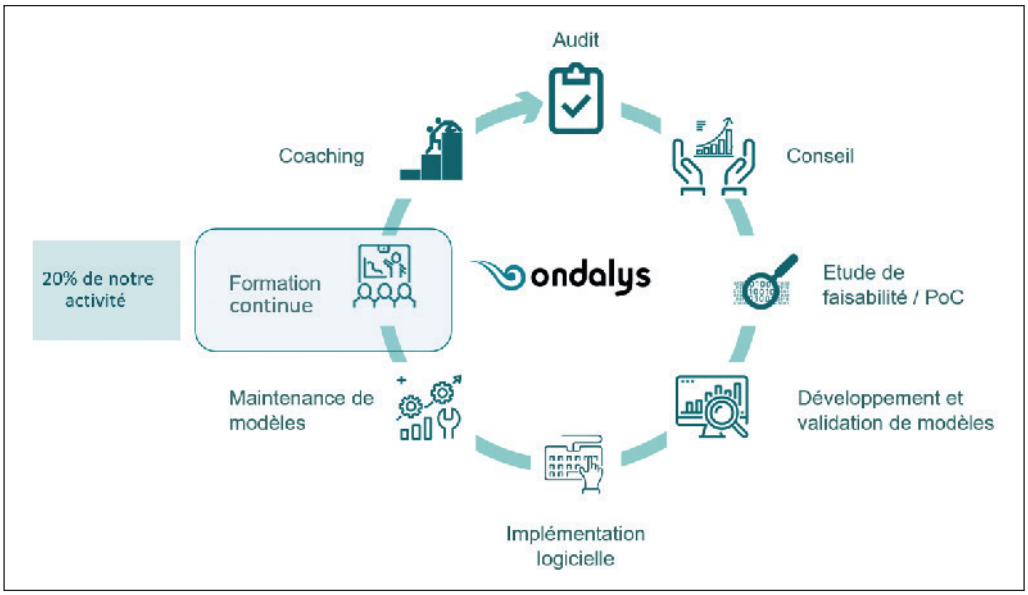


Figure 2

Différentes entrées de l'entreprise dans un projet.

artificielle. Il développe ensuite les activités de formation continue aux outils opérationnels et termine par la présentation de trois applications rendues possibles par ces outils.

1 Chimiométrie, *machine learning* et intelligence artificielle : sémantique et concepts

1.1. Sémantique

Le *deep learning* est une application du *machine learning* utilisant des algorithmes complexes de type réseaux de neurones. La chimiométrie quant à elle utilise des algorithmes linéaires multivariés. Enfin, d'autres outils non linéaires, comme les SVM (*Super Vector Machines*)[13], les arbres de régression, de classification[14], et les forêts aléatoires[15] font également partie du *machine learning*.

On peut signaler un certain nombre de mots-clés, signalés sur la ***Figure 3*** : *Big data*[16], *data science*, *data analytics*[17], MVA (*multivariate analysis*) – en français, analyse de données multivariées – *metabolomics* (analyse du métabolome[18]), etc. Mais retenons

13. *Super Vector Machines* : machines à vecteurs de support, ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression.

14. Arbres de régression, de classification : techniques de groupement des données.

15. Forêt aléatoire : regroupement d'arbres.

16. *Big data* : données massives, ensemble très volumineux de données qu'aucun outil classique de gestion de base de données ou de gestion de l'information ne peut vraiment travailler.

17. *Data analytics* : analyse de données.

18. Métabolome : ensemble des métabolites, des petites molécules telles que les intermédiaires métaboliques, les hormones et autres molécules signal ainsi que les métabolites secondaires, qui peuvent être trouvées dans un échantillon biologique.

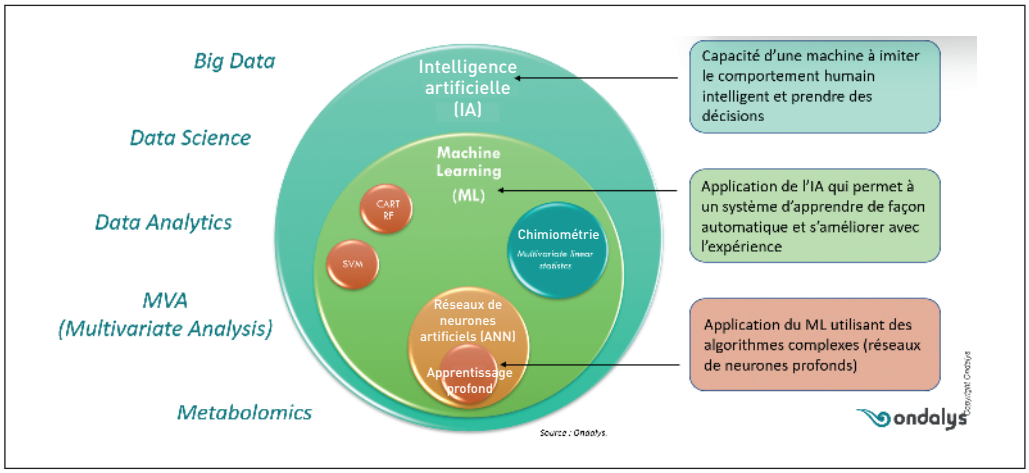


Figure 3

Distinction entre les notions importantes liées à l'intelligence artificielle.

que l'intelligence artificielle est une activité beaucoup plus vaste qui utilise des outils de *machine learning* et prend des décisions autonomes.

1.2. Exemples de concepts mis en œuvre

1.2.1. Signal multivarié

Pour (entre autres) la chimiométrie ou le *machine learning*, on peut utiliser différents types de présentation pour les données : soit la **description univariée**, des paramètres isolés, soit une présentation multidimensionnelle dite en **données multivariées**.

À gauche (*Figure 4*), est présenté un tableau de données consistant en une seule colonne : une variable ou un paramètre est décrit avec des valeurs continues ou qualitatives sur un certain nombre d'échantillons ou sur une population d'individus ; cela correspond à ce qu'on appelle un vecteur. C'est un objet mathématique bien précis, qui va être traité par des statistiques classiques, dites univariées.

Pour traiter simultanément plusieurs variables, mesurées de façon concomitante sur les mêmes échantillons, on établit un tableau (une matrice) de données avec plusieurs colonnes et toujours les mêmes lignes (voir à droite ***Figure 4***). Cette présentation permet de travailler sur les statistiques multivariées, utilisant les méthodes de l'algèbre linéaire et le calcul matriciel.

À quoi ressemblent ces données en chimie, ces signaux multivariés ? Il peut s'agir des paramètres procédés (***Figure 5*** en haut à gauche) qu'ils soient mesurés *at-line*[19] ou *in-line*[20] sur un réacteur chimique. C'est classiquement le pH, la température, la pression, etc.

Mises bout à bout, ces mesures constituent un signal

19. *At-line* : près de la ligne.
20. *In-line* : dans le milieu.

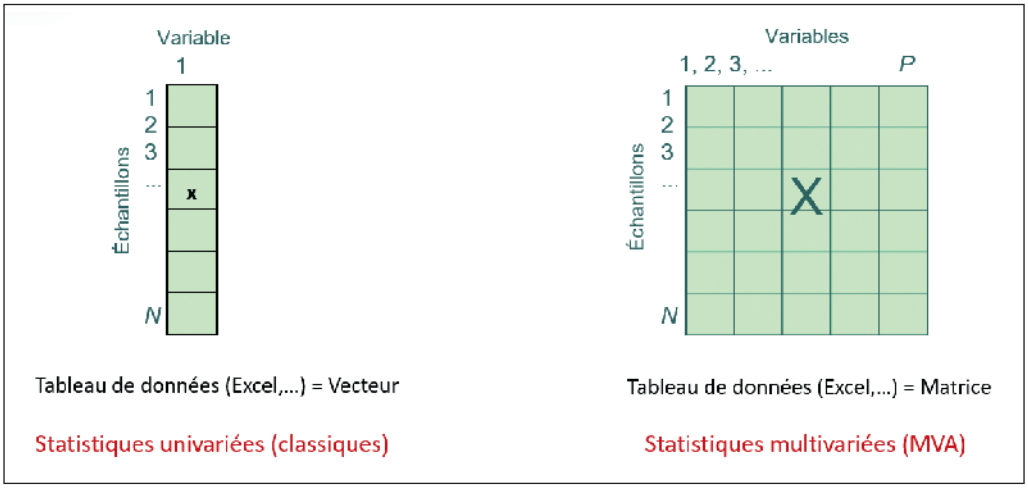


Figure 4

Comparaison entre les données univariées et multivariées.

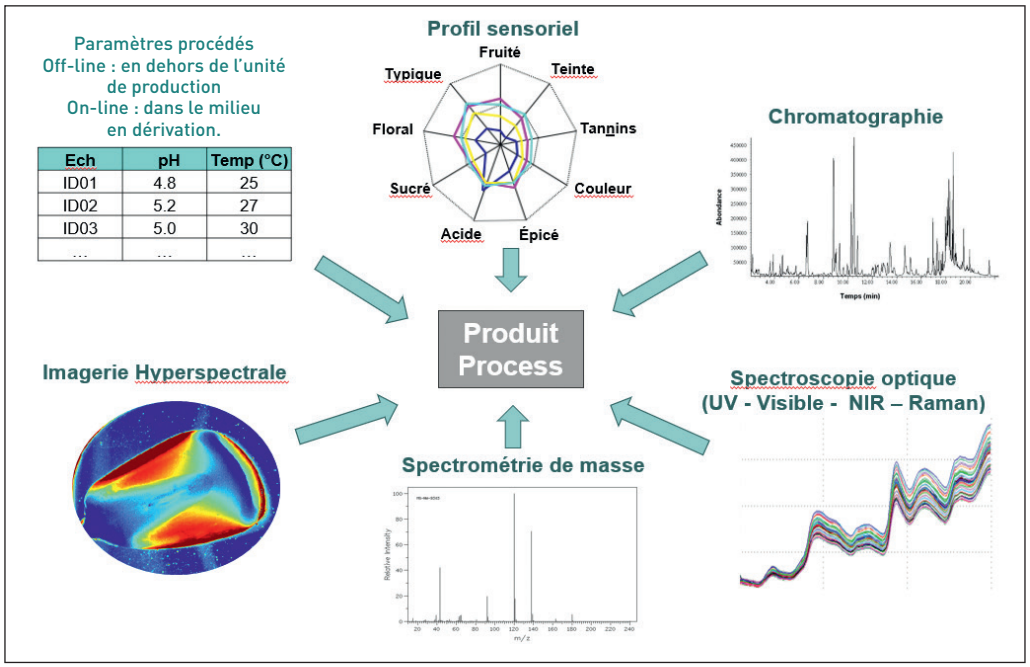


Figure 5

Différents types de données multivariées.

multivarié. En haut au centre, c'est l'exemple d'un profil sensoriel, rare en chimie (quoique peut-être au niveau des odeurs dans le milieu de l'automobile) mais très fréquent en agroalimentaire. À droite, c'est un chromatogramme issu de séparation sur colonne chromatographique. On présente aussi des spectres de masse, des spectres optiques type UV visible ou proche infrarouge, Raman[21] et de l'imagerie hyperspectrale[22]. Avec la spectroscopie optique, on enregistre une mesure d'intensité d'absorbance pour chacune des longueurs d'onde et on peut en extraire un « signal multivarié » (qu'on peut aussi appeler une empreinte ou un *fingerprint*[23]).

Les données de type chromatographie et spectrométrie de masse permettent l'utilisation de la chimiométrie et du *machine learning* en permettant de constituer des **données**

21. Spectroscopie Raman : méthodes non destructives d'observation et de caractérisation de la composition moléculaire et de la structure externe d'un matériau, qui exploite le phénomène physique selon lequel un milieu modifie légèrement la fréquence de la lumière y circulant.

22. Imagerie hyperspectrale : technologie permettant d'obtenir l'image d'une scène dans un grand nombre (généralement plus d'une centaine) de bandes spectrales à la fois étroites et contiguës.

23. *Fingerprint* : empreinte digitale.

métabolomiques[24] ; elles permettent de disposer de signaux synthétiques très utiles pour la comparaison d'échantillons.

1.2.2. Corrélation

Le concept de corrélation est majeur et d'utilisation constante.

Prenons l'exemple du suivi d'un paramètre, par exemple la température sur un procédé physicochimique. On suit la température dans le temps et on la reporte sur une carte de contrôle. L'ensemble de la population de points mesurés est représenté sur l'axe des abscisses (***Figure 6***) pour identifier les points compris entre des limites de contrôle (en l'occurrence fixées ici à plus ou moins 3 écarts-types), pour détecter des dérives du procédé (des non-conformités). Sur ce cas assez simple (axe des abscisses sur la ***Figure 6***), on a deux points, un à température un peu basse et un autre à température un peu haute, qui vont générer des alertes.

Si on suit un deuxième paramètre ou une deuxième variable, par exemple la pression, mesurée au même moment sur le même échantillon, on établit une carte analogue – ici sur l'axe des ordonnées (***Figure 6***) – permettant éventuellement d'observer des points anormaux, en l'occurrence un à pression inférieure et un autre à pression supérieure qui constituent des alarmes.

En considérant simultanément les deux paramètres, on a des informations beaucoup plus puissantes. Par exemple, le premier point situé à gauche (***Figure 6***) paraît conforme en température et en pression si on le considère isolément, mais comme atypique par rapport à la population étudiée car il s'écarte du nuage général. C'est typiquement le cas d'une anormalité **non détectée en univarié, mais détectée en multivarié**, c'est-à-dire, ici, en considérant simultanément la température et la pression. Une situation en sens inverse, un cas détecté non conforme à tort en univarié apparaît aussi sur le diagramme : c'est le cas d'un point détecté conforme sur la carte de contrôle multivariée mais qui déclencherait en univarié une alerte dans l'autre sens.

Ces exemples illustrent l'intérêt de travailler en multidimensionnel, tenant compte de la structure de corrélation entre les variables, et donc de l'utilisation des outils de chimiométrie et de *machine learning*.

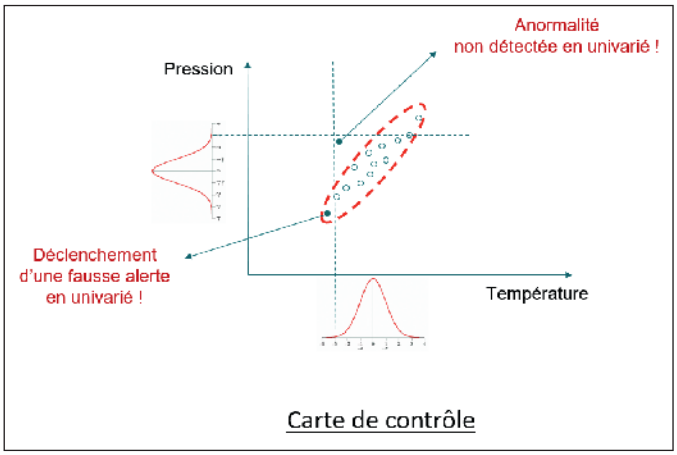


Figure 6

Carte de contrôle de la pression et de la température.

24. Étude des métabolites issus de l'organisme ou provenant de l'environnement.

RECONNAÎTRE AUTOMATIQUEMENT L'IMAGE D'UN CHAT

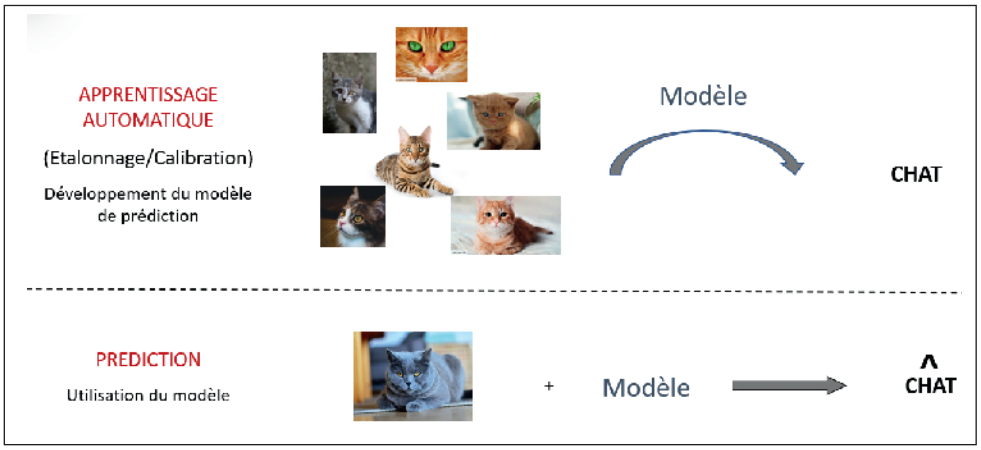


Figure 7

L'automatique pour reconnaître des images de chat.

Nous l'illustrons ici par l'apprentissage d'un algorithme sur des images de chats. Il s'agit de reconnaître ce qu'est un chat, au moyen d'un jeu d'étalonnage, d'entraînement ou de calibration, pour développer un modèle de prédiction (***Figure 7***). Une fois entraîné et optimisé, **le modèle s'appliquera sur l'image d'un nouveau chat : s'il marche bien, il prédira correctement qu'elle correspond bien à un chat.**

1.2.3. Apprentissage automatique

Le concept d'apprentissage automatique du *machine learning* est souvent évoqué (voir Encart : Reconnaître automatiquement l'image d'un chat).

On peut l'aborder en soulignant une correspondance entre données traitées traditionnellement et approches de *machine learning*. Dans les premières, dans le domaine de la chimie, on s'appuie sur des spectres qui se ressemblent tous et construisent une base à partir de différents échantillons. De la même façon, les outils de *machine learning* permettent d'établir une base de calibration spectrale (***Figure 8***) à partir d'un grand nombre de spectres et d'optimiser un modèle de prédiction sur une valeur quantitative ou qualitative. Quantitativement, cela peut être la viscosité, une teneur en molécule, alors que qualitativement, cela peut être une conformité, une qualité.

Le « modèle prédictif » sera appliqué à un nouveau spectre pour prédire le paramètre d'intérêt. C'est l'application typique de développement de calibration spectroscopique utilisant le *machine learning*.

Les deux grandes catégories d'apprentissage automatique sont : les **apprentissages non**

supervisés, utilisés pour le *data mining* ou le *clustering*[25] pour explorer les données de façon multivariée, et les **apprentissages supervisés** pour développer des modèles prédictifs, quantitatifs ou qualitatifs (*Figure 9*).

25. *Data clustering* : partitionnement de données, méthode d'analyse de données.

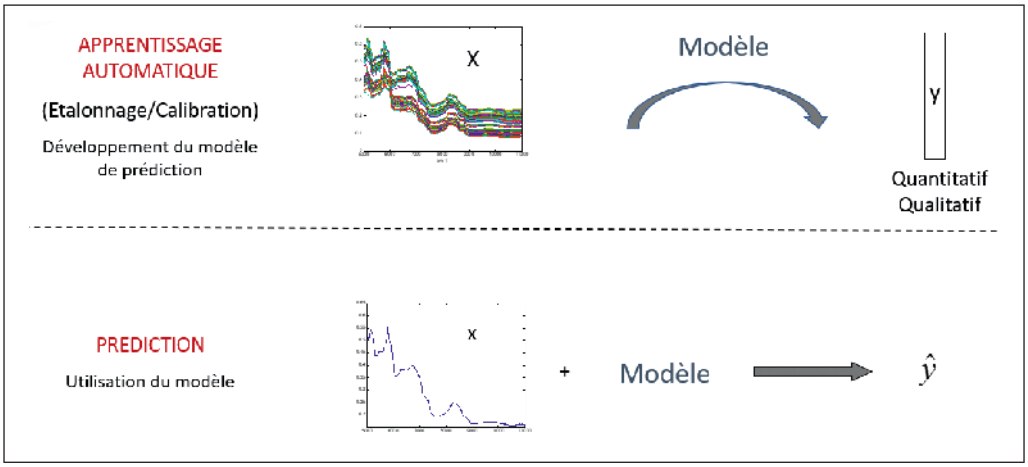


Figure 8

Concept de l'apprentissage automatique pour traiter des spectres.

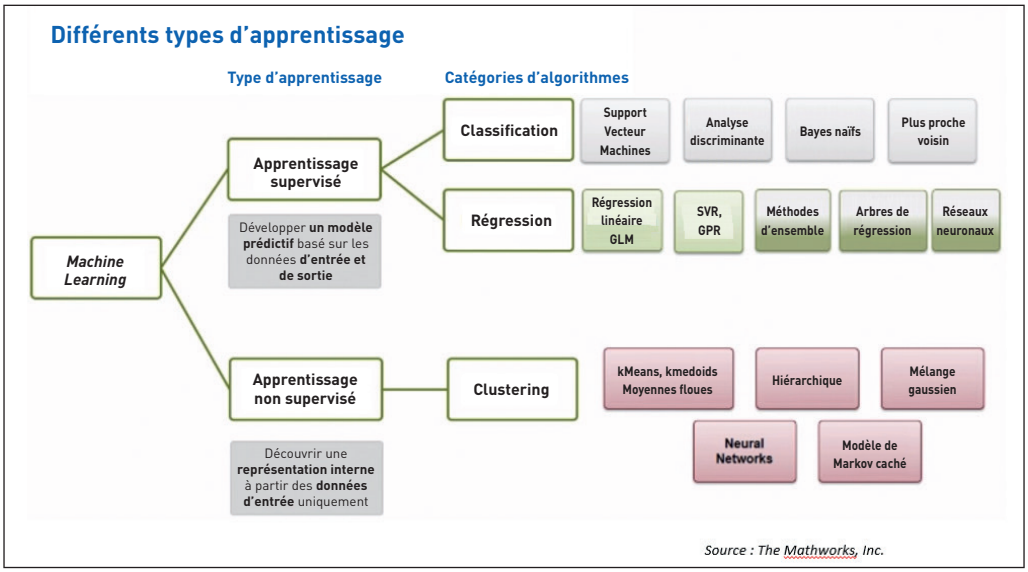


Figure 9

Différents types d'apprentissage automatique.

Souvent on passe beaucoup de temps pour trouver le meilleur algorithme de *machine learning*, mais il faut souligner qu'une étape préalable reste tout à fait indispensable : c'est **le nettoyage et l'exploration des données qui seront utilisées,** processus appelé « conciliation ».

On appelle « conciliation des données » tout ce qui est alignement et synchronisation des données provenant de différents instruments ou capteurs. Le nettoyage des données utilise des statistiques classiques (distribution, quartiles, moyenne, etc.) ; il utilise aussi des statistiques multivariées de la même manière avec, notamment, le choix de la composante principale pour faire ressortir ce qui est outlier[26], tendance, cluster[27], etc.

De nouveau, il y a lieu d'insister sur la tâche de fiabilisation des données (***Figure 10***) à effectuer au préalable et qui est nécessaire pour augmenter la performance et faciliter l'interprétation ultérieure des résultats. On voit là qu'il y a quand même une « valeur ajoutée de l'humain » – on parlera de *human learning*[28] pour illustrer cette notion avec un peu d'humour –, c'est en fait une étape très chronophage dont on ne doit pas se passer avant d'entrer toutes les données dans les algorithmes de *machine learning*.

26. Outlier : individu atypique.
27. Cluster : groupe.

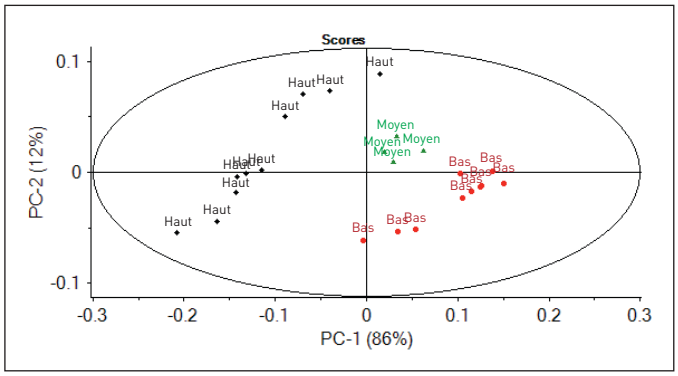


Figure 10

Graphique permettant de juger la fiabilisation de nos données.
Source : The Mathworks, Inc.

2 La formation continue aux outils opérationnels de chimiométrie et *machine learning* chez Ondalys

La partie principale du chapitre a pour but de partager l'expérience développée chez Ondalys dans la formation continue sur ces outils de chimiométrie et de *machine learning*.

2.1. Présentation de la formation

Nous sommes cinq formateurs au sein d'Ondalys. La petite valise au centre (***Figure 11***) représente les compétences multidisciplinaires et c'est un message adressé aux étudiants : quand vous avez les choix à faire dans vos études à un moment donné, pensez aussi multidisciplinarité. Il ne s'agit pas d'avoir seulement les compétences en mathématiques, en statistiques, en modélisation ou en algorithmique ; il faut aussi connaître « les métiers », soit tout ce qui

28. *Human learning* : apprentissage humain.

Figure 11

Équipe de formateurs d'Ondalys.

est procédés, produits, production, laboratoire, R&D, et puis au-delà, les techniques de codage et les langages de programmation. En chimiométrie et *machine learning*, on est à l'interface de plusieurs disciplines entre le métier, la chimie, les compétences en mathématiques, statistiques et informatique.

La ***Figure 12*** présente un exemple de déroulé pédagogique sur la formation en analyse de données spectroscopiques de type proche infrarouge ou Raman sur « Python™ ». Le processus est certifié Qualiopi, et comporte une quantité de milestones[29] et de points à surveiller, notamment positionner les besoins des différents stagiaires en début de formation. La formation mélange la théorie, la méthodologie, les petites astuces dans le traitement de données et la pratique, car il y a beaucoup de pratiques à acquérir sur des logiciels partenaires ou sur des langages de programmation type Python™.

La formation est ponctuée de démonstrations du formateur, ainsi que d'exercices réalisés en autonomie et corrigés par un des stagiaires. L'idée est que le stagiaire sorte de cette formation, qui dure en général deux ou trois jours, en étant capable de traiter les données lui-même sur un logiciel ou avec un langage de programmation. Elle est adressée à des débutants et se conclut par un bilan effectué avec un QCM à remplir sur l'évaluation du contenu, une fiche d'évaluation sur la formation et sur le formateur, puis un tour de table final.

2.2. Formation sur mesure

On propose des programmes de formation sur mesure, notamment en intra-entreprise

29. Milestones : jalons.

ANALYSE DE DONNEES SPECTROSCOPIQUES AVEC PYTHON

Horaires et Durée	Objectifs pédagogiques de la séquence	Contenu de la séquence	Méthodes, moyens pédagogiques et comportement spécifique par séquence	Méthode d'évaluation	Scénario réalisé : préciser les réajustements significatifs
9h00 9h20	Accueil des stagiaires Présentation des stagiaires/Python	- Présentation des stagiaires Présentation des objectifs du programme et de l'organisation Les questions des stagiaires	- Tour de table + paper board		
9h20 10h45	Présenter les principales librairies Python pour le Machine Learning	- Librairies Python et leurs fonctions de base - Distribution Anaconda Notebooks Jupyter	- Powerpoint Démonstration avec un notebook Jupyter	Questionnement direct	
10h45 11h00	Pause				
11h00 12h30	Réaliser une ACP sur des données spectroscopiques avec un notebook Jupyter	Réalisation d'une ACP sur des données réelles	- Présentation des données et de l'objectif de l'exercice sur Powerpoint - Démonstration et utilisation du notebook Jupyter - Réalisation en autonomie par chaque stagiaire sur les données réelles	- Vérification des résultats obtenus suite à la démonstration - Questions/Réponses - Correction de l'exercice par un des stagiaires - Questions/Réponses	
12h30 13h30	Pause déjeuner				
13h30 15h00	Réaliser une régression par modèle PLS sur des données spectroscopiques avec Python dans un notebook Jupyter	Réalisation d'une PLS sur des données réelles	Présentation des données et de l'objectif de l'exercice sur Powerpoint - Réalisation en autonomie par chaque stagiaire sur les données réelles	Questionnement direct - Vérification de l'utilisation du notebook	
15h00 15h15	Pause				
15h15 16h45	Développer des modèles PLS en utilisant plusieurs prétraitements avec Python dans un notebook Jupyter	- Présentation des prétraitements disponibles sous Python - Optimisation de prétraitements sur des données réelles	- Présentation des données et de l'objectif de l'exercice sur Powerpoint - Réalisation en autonomie par chaque stagiaire sur les données réelles	- Correction de l'exercice par un des stagiaires Questionnement direct	
16h45 17h00	Bilan de la journée	Synthèse des points abordés	- Powerpoint - Paper board	Questionnement direct	
17h00 17h30	Finaliser la formation et les acquis	Finalisation des conclusions suite à la formation - Evaluation de la formation	QCM formatif - Fiche d'évaluation individuelle - Tour de table + paper board	Correction du QCM	

Copyright Ondalys + Confidentiel

ondalys

Figure 12

Déroulé pédagogique d'une formation en analyse de données spectroscopiques avec Python™.

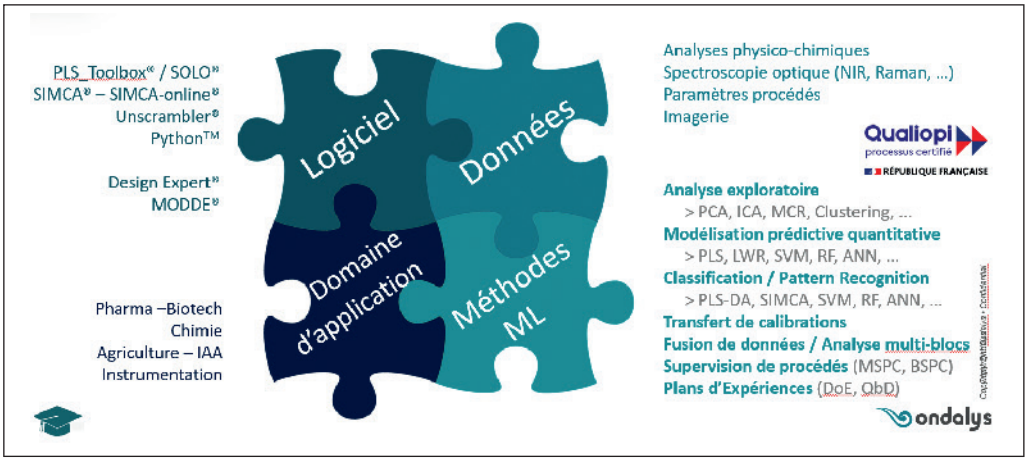


Figure 13

Les quatre composantes de la formation continue.

sur site, qui sont le résultat de quatre composantes (***Figure 13***) :

- le type des données qui seront traitées ultérieurement, que ce soit des analyses physico-chimiques, de la spectroscopie optique, des paramètres procédés, de l'imagerie ;
- le domaine dans lequel le client intervient : en pharma,

biotech, chimie ou autre industrie ;

• les différentes thématiques de *machine learning*, que ce soit l'analyse exploratoire, la modélisation prédictive quantitative, classification, fusion de données, analyses multiblocs, supervision de procédés continus ou en batchs, ou plans d'expériences ;

• les pratiques variées sur des logiciels abordés dans les formations ; le client doit choisir le logiciel sur lequel il veut être formé ; s'il n'a pas d'idées, il est conseillé.

2.3. Logiciels de formations

La ***Figure 14*** montre les différents logiciels sur lesquels sont effectuées les formations. Il y en a deux grandes catégories : les logiciels à interface graphique, qui intéressent plutôt les débutants ; et les langages de programmation, type Matlab ou Python™, qui intéressent des ingénieurs confirmés ou qui veulent en faire leur métier. Dans les logiciels à interface graphique, nous avons quatre partenaires : Eigenvector Research, Sartorius, Aspentech et StatEase, qui proposent chacun différents logiciels ou des suites de logiciels pour adresser le développement de modèles de *machine learning*, leur mise en œuvre et éventuellement le développement de plans d'expériences.

L'offre de formation continue est proposée en intra-entreprise sur site. Elle consiste en formations sur mesure ; alternativement des thématiques fixées avec des dates, des lieux et des logiciels prédéfinis sont proposées. Depuis la période Covid, nous proposons aussi des formations « online ». Le public concerné est constitué d'industriels, mais également de centres de recherche

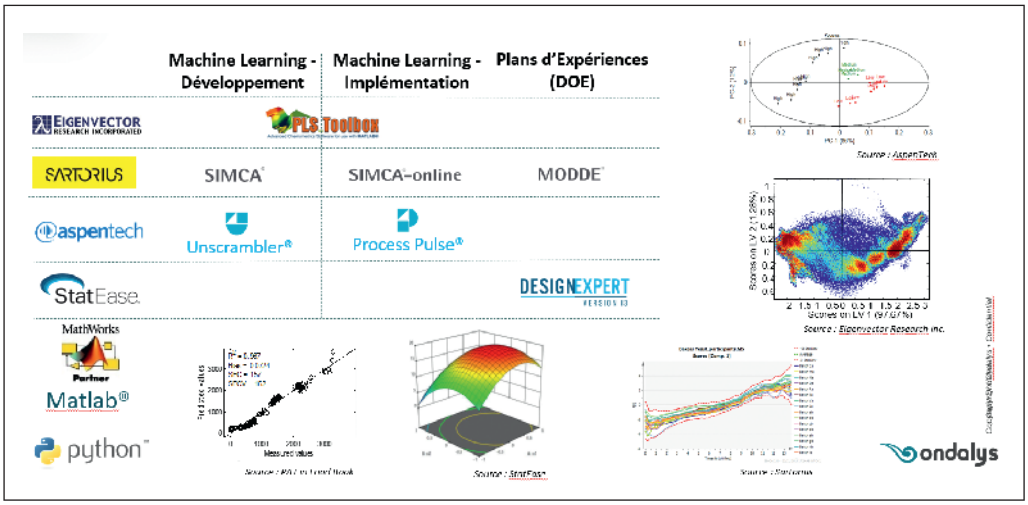


Figure 14

Logiciels partenaires.

publics et de centres techniques ; il est constitué de chercheurs, de techniciens, d'ingénieurs et occasionnellement de managers. Les formations sont données en français ou en anglais.

On propose également après la formation des services associés de type coaching, pour accompagner les stagiaires dans la mise en œuvre de ces outils avec support client, visio ou téléphone, ainsi que des prestations de traitement de données. Le gros de notre activité dans ce domaine est de traiter les données des clients sur des problématiques un peu complexes. On distribue également certains logiciels de nos partenaires.

La société Ondalys est référencée dans différents catalogues de partenaires pour la formation continue, notamment CPE Lyon, EASE Training et MabDesign pour la biotech. Quelques chiffres issus des questionnaires d'évaluation de la formation montrent que les gens sont « satisfaits » pour l'instant de ce que nous proposons depuis plus d'une quinzaine d'années maintenant. Quelques références qui nous ont fait confiance pour la formation continue dans les différents domaines sont listées sur la *Figure 15* : pharma, biotech, chimie, agriculture, agroalimentaire, instrumentation et centre de recherche.

3 Pour quoi faire ? Quelques exemples d'applications

Dans sa dernière partie, le chapitre présente des applications très différentes qui ont été développées avec les outils de *machine learning* et de chimiométrie.

3.1. Calibrations spectroscopiques

La première application traite du **développement de**

Figure 15

Clients d'Ondalys dans les différents domaines.

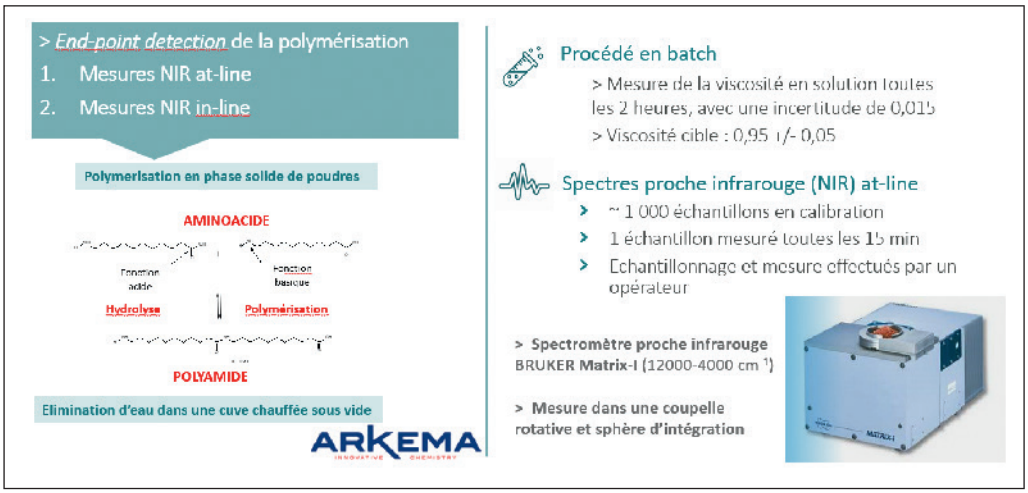


Figure 16

Résumé des besoins d'Arkema dans le modèle at-line.

calibrations spectroscopiques mises au point avec l'industriel Arkema en Normandie (***Figure 16***). Le but était d'avoir une aide au pilotage d'un procédé de polymérisation, notamment en gérant ce qu'on appelle le End-point[30] (fin du procédé) où il faut arrêter le process. Le procédé est en batch[31] classique, ce qui demande une mesure de la viscosité toutes les deux heures. Le besoin était d'avoir une mesure plus rapide, même éventuellement indirecte sur un spectromètre NIR (*Near InfraRed*)[32] en faisant usage dans un premier temps d'un échantillonnage opérateur par prélèvement (modèle at-line).

La base de calibration (d'entraînement) était constituée d'environ mille échantillons pour lesquels on disposait à la fois des spectres NIR et de la viscosité mesurée en laboratoire. Le modèle a été entraîné avec des outils de *machine learning* classiques et a fourni le modèle présenté ***Figure 17***, qui a permis de suivre la cinétique. On voit en rouge les mesures de référence au labo et les petites croix violettes qui correspondent aux prédictions proche infrarouge.

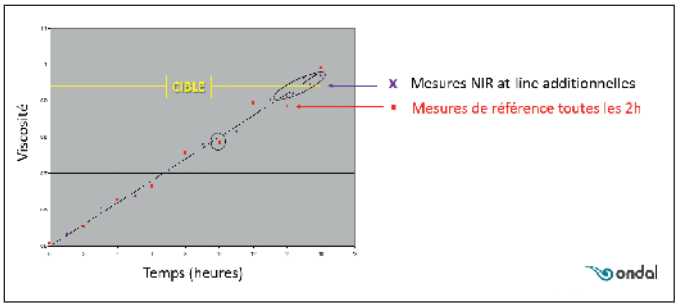


Figure 17

Détection du End-point de la polymérisation avec des mesures proche infrarouge at-line.

30. End-point : fin du procédé.
31. Traitement industriel par lots dans lequel le produit fini est obtenu à la suite d'une série de tâches plutôt qu'en continu.
32. NIR (*Near InfraRed*) : proche infrarouge.

Le modèle fonctionnait bien, avec une faisabilité qui donnait satisfaction ; à ceci près que cela nécessitait toujours un échantillonnage manuel et le fait d'analyser les échantillons en laboratoire engendrait une influence de la température et de l'humidité du milieu environnant.

Pour améliorer la performance, on a regardé si les contrôles pouvaient marcher également sur un analyseur NIR « en ligne » (***Figure 18***). Une deuxième base de données d'étalonnage a été construite en mesurant environ 10 000 échantillons avec, cette fois-ci, un instrument de process, équipé d'une sonde et d'une fibre optique rattachée, permettant une mesure en ligne toutes les deux minutes.

La ***Figure 19*** montre la cinétique de prédiction. On voit la montée en viscosité du produit sur un batch jusqu'à la cible, qui permet de décider du Endpoint. En rose, est présentée la température, puis, en bleu foncé, ce qu'on appelle la distance de Mahalanobis qui est une mesure de la proximité spectrale d'un nouvel échantillon par rapport à la base d'étalonnage et permet de déterminer le domaine de validité du modèle de prédiction ; ici, il est valide au-dessus d'une certaine température. Cette application a très bien marché et ces analyseurs proche-infrarouge ou Raman sont de plus en plus montés sur les installations en chimie, en pharma, en biotech. Nous faisons face à beaucoup de demandes.

3.2. Chromatographie couplée à la spectrométrie de masse

La deuxième application abordée concerne un type de données complètement différent : il s'agit de traiter des données de laboratoire issues d'une chromatographie couplée à

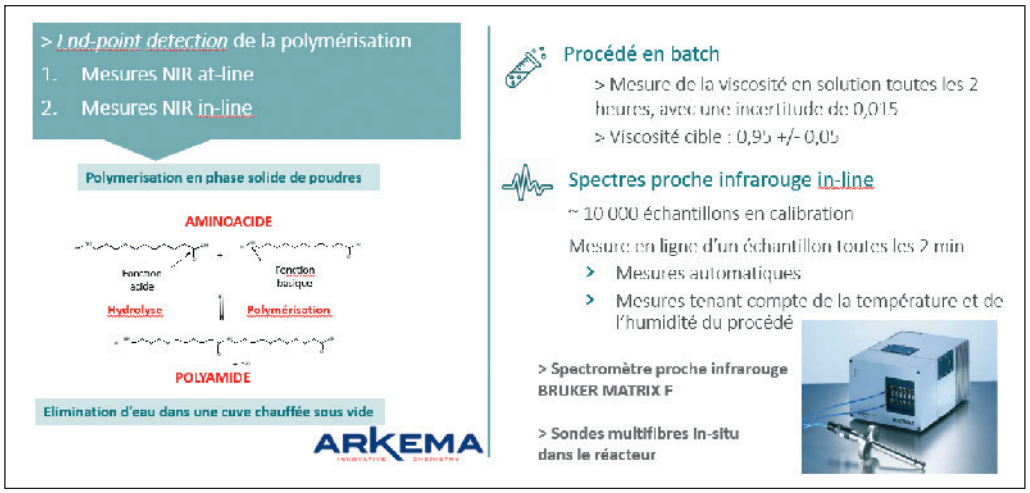


Figure 18

Résumé des besoins d'Arkema dans le modèle in-line.

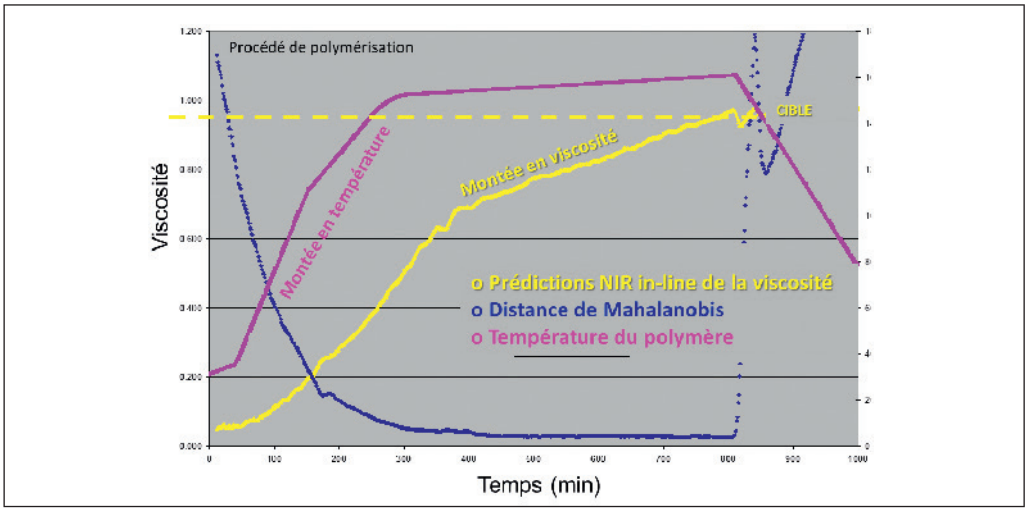


Figure 19

Détection du End-point de la polymérisation avec des mesures proche infrarouge at-line.

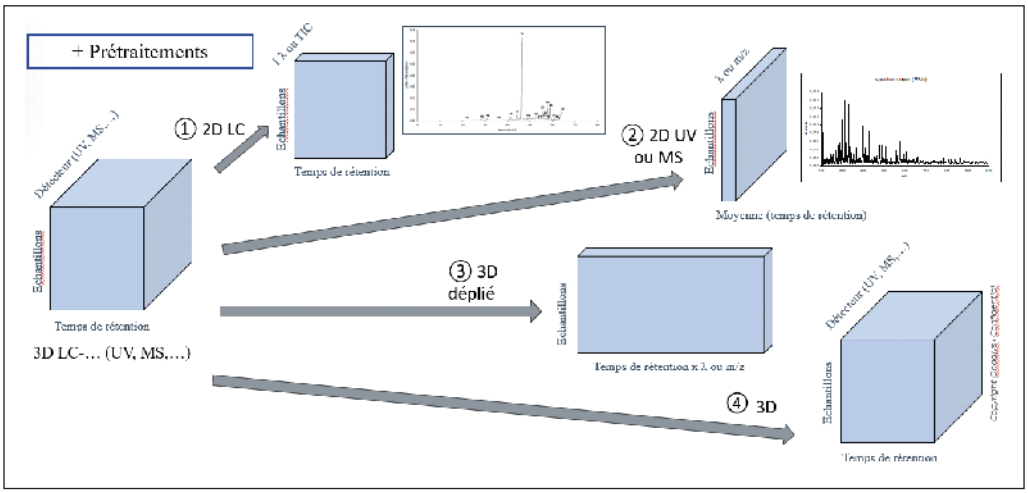


Figure 20

Différentes stratégies pour traiter les données de la chromatographie par machine learning.

une spectrométrie de masse. Cette technique génère une masse de données importantes très complexes présentées ici sur trois dimensions (***Figure 20***) : la dimension échantillon, une dimension chromatographique (évolution du temps de rétention) et une dimension détectrice, dans le cas présent un détecteur masse, qui donne un spectre de masse à chaque temps de rétention du chromatographe. Les mesures fournissent des cubes de données dont il faut

déterminer s'ils peuvent être traités facilement par de la chimiométrie et du *machine learning*. La réponse étant non, cela a conduit à adopter plusieurs stratégies alternatives.

Parmi les stratégies présentées (***Figure 20***), nous avons choisi de travailler sur la dimension chromatographique (voie 1), sur la dimension masse (voie 2), ou sur du déplié (voie 3), mais d'éviter de traiter directement les données 3D (voie 4) très complexes dans le cas présent.

Dans la stratégie 1 (***Figure 20***) – dimension chromatographique – le signal retenu ici correspond à un résumé de

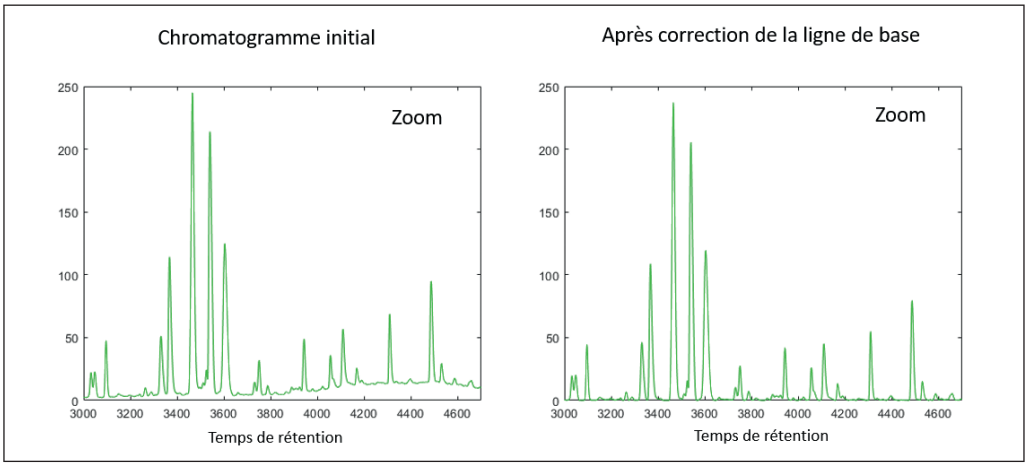


Figure 21

Correction de la ligne de base.

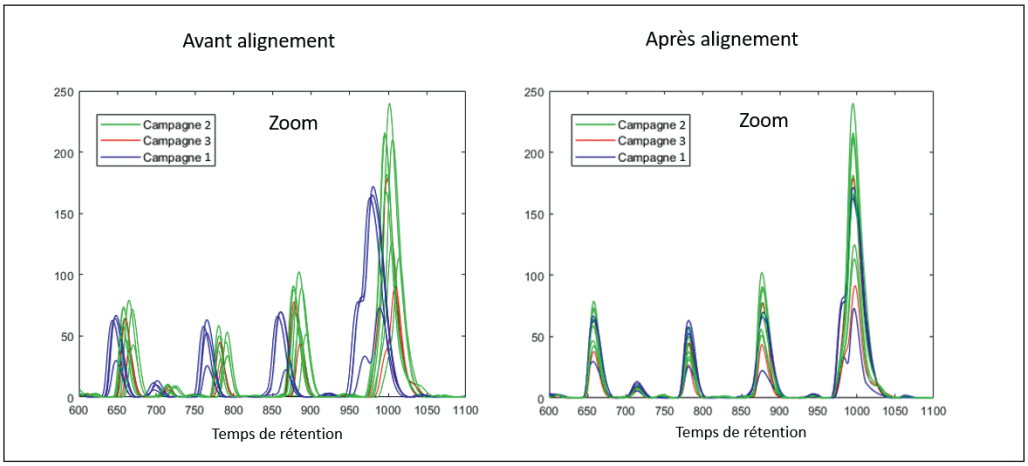


Figure 22

Correction de l'alignement.

l'information de masse TIC (*Total Ion Current*)[33], synthèse des m/z[34] recueillis. On obtient ce genre de signaux représentés sur la ***Figure 21*** qui comportent une forêt de pics.

Un certain nombre de prétraitements sont appliqués pour nettoyer le signal, avant de le soumettre au *machine learning*. Ce peut être par exemple une correction ligne de base (***Figure 21***) ou une correction d'alignement (***Figure 22***).

Il faut aussi aligner les chromatogrammes entre eux (***Figure 23***), parce qu'il y a des « décalages » qui se créent avec la colonne et qu'avant de rentrer ces données dans les algorithmes de chimiométrie et de *machine learning*, il est nécessaire d'aligner tous ces pics.

L'application d'un algorithme, qui s'appelle SIMCA (*Soft Independent Modeling for Class Analogies*)[35], a permis de réaliser la discrimination. Dans ce cas-là, les petits triangles bleus (***Figure 23***) correspondent aux échantillons de calibration, des échantillons conformes qui ont permis d'entraîner le modèle et qui sont bien qualifiés comme conformes. Ce modèle a été testé sur des échantillons conformes en vert, et des échantillons non conformes, en rouge. La validation a été satisfaisante : le cas fonctionne bien.

Pour cette application, nous avons utilisé la dimension chromatographique comme signal global. Cette approche diffère de l'approche classique

33. *Total Ion Current* : courant ionique total.
34. m : masse, z : charge.

35. *Soft Independent Modeling for Class Analogies* : Modélisation douce et indépendante pour les analogies de classes.

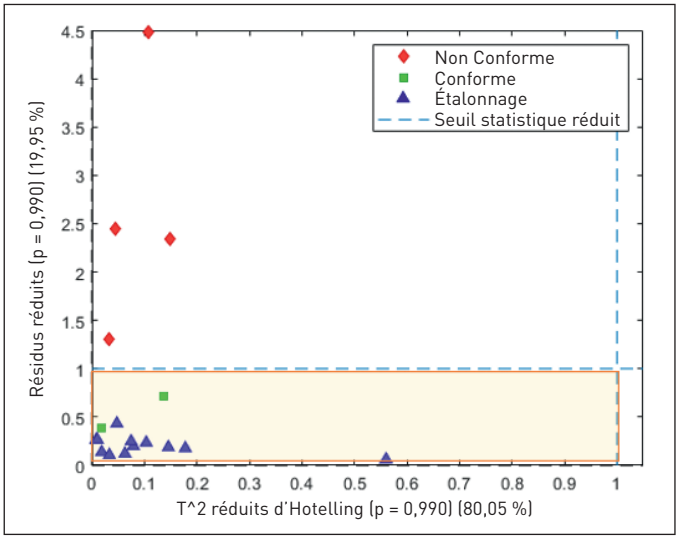


Figure 23

Modélisation de la conformité.

où l'on utilise la masse pour identifier les pics, et où l'on intègre ensuite les aires des pics identifiés pour effectuer de la quantification.

3.3. Modélisation de procédés batch

Le dernier exemple est encore très différent : il concerne la modélisation de procédés batchs. Un certain nombre de variables de procédés (éthanol, température, niveau, temps, pH, etc.) ont été mesurées tout au long de la cinétique, ce qui donne lieu à un cube de données (***Figure 24***). Sur ces mêmes batchs, des paramètres « qualité », caractérisant le produit fini, ont été mesurés ; des conditions initiales sur des variables caractérisant les matières premières entrant dans les batchs ont également été intégrées. Nous disposons alors d'une collection de données hétérogènes, très différentes entre elles et la question est : comment combiner ces données et permettre soit de gagner en compréhension, soit de faire des modèles prédictifs, soit de faire la supervision du procédé en multivarié, ce qui donne le résultat de la ***Figure 25***. Finalement, toutes ces variables sont résumées de façon multivariée par ce qu'on appelle des scores dans une enveloppe statistique, qu'on appelle aussi le « *golden batch* ». L'objectif consiste soit de piloter les procédés en temps réel, soit en différé de faire du troubleshooting pour voir pourquoi un batch de production s'est mal passé. On utilise ainsi une méthode qu'on appelle le « *contribution plot* » : on va cliquer à un endroit où le procédé peut dériver et on va vérifier les variables qui peuvent être responsables de cette dérive (***Figure 26***).

Ces trois exemples montrent bien la diversité des applications qui peuvent être mises en œuvre dans l'industrie de process avec des outils de chimiométrie et de *machine learning*.

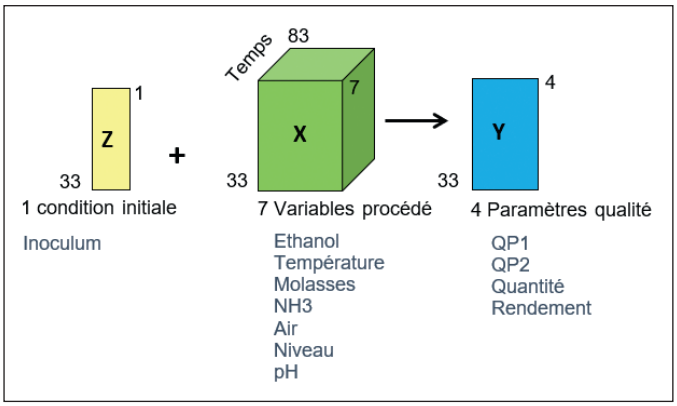


Figure 24

Modélisation d'un procédé batch.

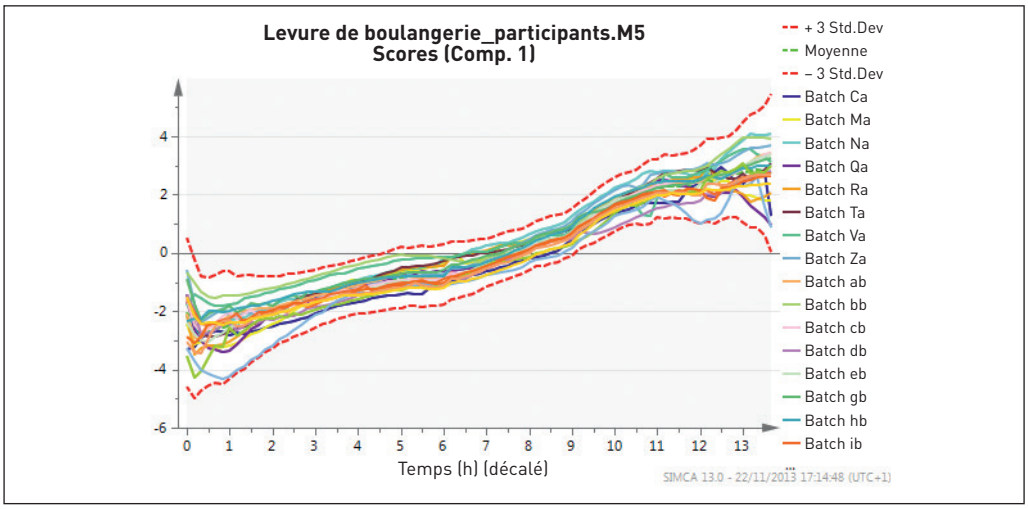


Figure 25

Exemple de process monitoring (supervision de procédé).

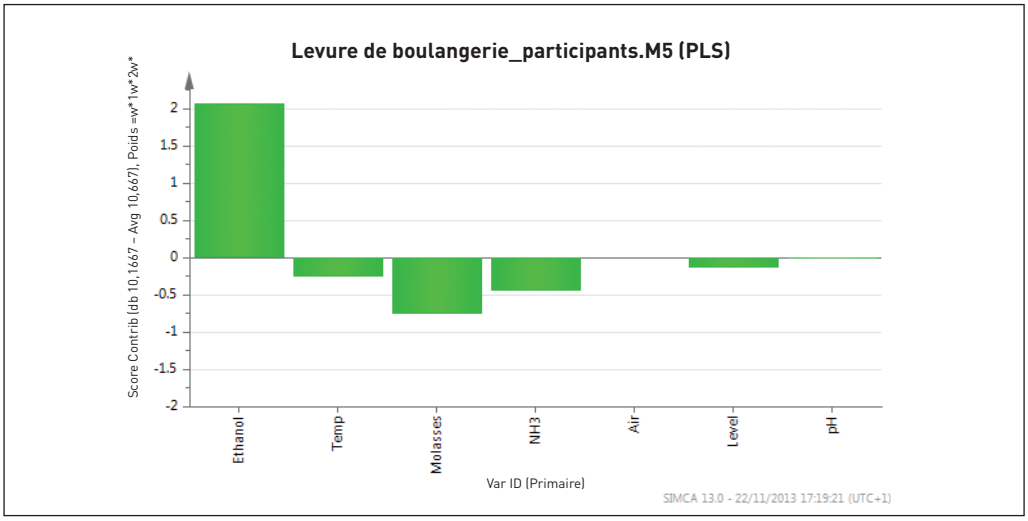


Figure 26

« Contribution plot ».

Conclusion
En chimie comme ailleurs, l'intelligence artificielle n'a pas dit son dernier mot !

Pour terminer, je ne résiste pas à l'envie de vous montrer une petite expérience de la semaine dernière sur des outils d'intelligence artificielle qui sont devenus très à la mode tout d'un coup. C'est l'interface qui s'appelle Dall-E d'OpenAI, c'est un peu un analogue de Chat GPT, mais qui traite les images au lieu de traiter les textes.

Par curiosité, j'ai mis la définition de *machine learning* sur Wikipédia et le logiciel m'a transformé la définition en image. Cela donne la ***Figure 27*** : chacun aura un avis sur cette image ! J'ai fait une deuxième expérience en faisant un résumé en anglais de mon intervention présente sur la formation continue (parce que ça marche mieux pour l'instant en anglais

Figure 27

Image représentant la définition de « Machine Learning » par Dall-E.

qu'en français), et il m'a sorti l'image de la ***Figure 28***, qui est, finalement, assez réaliste malgré les quelques défauts du visage.

L'engouement actuel du grand public pour ces outils d'intelligence artificielle a beau être tout jeune, on le voit déjà, après quelques mois, s'amplifier à toute vitesse. Il s'agit là d'exemples ludiques, mais je crois avoir montré dans ce chapitre que la dynamique est tout aussi vigoureuse pour l'industrie en général et **pour l'industrie chimique en particulier**. Des résultats très nouveaux sont déjà obtenus. Ils sont annonciateurs de performances extraordinaires qu'on n'ose même pas évoquer en craignant de quitter le réalisme… Mais ces techniques d'intelligence artificielle sont parties et nous aurons à travailler pour les suivre et en tirer des bénéfices inimaginables ! Allons-y !

Figure 28

Image représentant la formation continue à la chimiométrie et au « Machine Learning ».

Partie 2

Intelligence artificielle dans la recherche en chimie, notamment dans la recherche de matériaux innovants

Le projet DIADEM : accélérer la découverte de nouveaux matériaux grâce à l'intelligence artificielle

Mario Maglione, Directeur de recherche CNRS ICMCB Bordeaux, Co-pilote du PEPR DIADEM.

Introduction

Le programme DIADEM (***Figure 1***) (un acronyme qui marche aussi bien en anglais qu'en français) signifie en français : Dispositif Intégré pour l'Accélération du Déploiement de Matériaux Emergents. Ce programme copiloté par le CNRS et le CEA dans le cadre du programme France 2030 a été lancé, pour ce qui nous concerne, en 2021, dans le cadre du PIA[1]. L'objectif de ce type de programme est de construire des équipements qui vont devenir différenciants et qui vont provoquer des changements notables, structurants pour la recherche en France. Ce chapitre décrit ce programme en essayant de donner quelques exemples de ce qu'on veut faire pour la découverte accélérée des matériaux.

1. PIA : Programme d'Investissement d'Avenir.

1 La découverte accélérée des matériaux

1.1. Les défis

De nombreux domaines technologiques reposent sur la découverte des matériaux : énergie, transport, santé, transition numérique. Nous connaissons les enjeux dans ces différents domaines et notamment la **notion de temps limité pour réaliser ces transitions**.

La mise en œuvre effective des nouveaux matériaux est d'autant plus retardée (plus d'une décennie d'essais et d'erreurs) que leur **complexité** augmente : c'est le cas par exemple des batteries qui sont des assemblages très complexes de matériaux.

Le troisième défi est qu'en plus, et c'est une nécessité absolue dont tous les chercheurs sont parfaitement convaincus, il faut respecter des contraintes de plus en plus drastiques concernant l'environnement, la maîtrise du cycle de vie des matériaux, la sobriété énergétique et les matériaux critiques.

Notre programme a pour objectif d'utiliser des techniques de l'intelligence artificielle pour relever ces défis.

L'accélération de la découverte des matériaux a été initiée pour une grande part aux États-Unis avec ce que les Américains ont appelé la *Materials Genome Initiative*[2] dans les années 2010. On voit (*Figure 2*) la photo de l'inauguration de cette initiative, centrée sur ce que les Américains appellent les MAPs (*Materials*

2. *Materials Genome Initiative* : Initiative sur le génome des matériaux.

Figure 1

Le projet DIADEM.

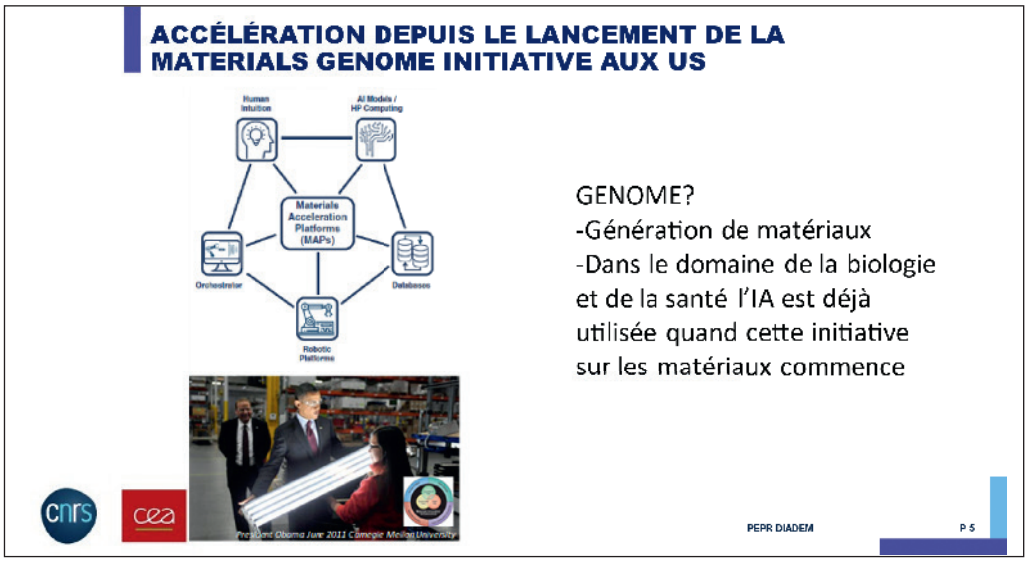


Figure 2

La Materials Genome Initiative.

Acceleration Platform)[3]. Depuis ce lancement, de nombreuses initiatives ont fait jour, et on assiste à des travaux de plus en plus nombreux dans ce domaine.

Des échanges avec des collègues américains, dont le porteur de ce projet *Materials Genome Initiative*, m'ont montré que l'utilisation du mot « genome » n'est pas un hasard. Certes il s'agit de générer de nouveaux matériaux, mais l'acronyme GENOME se réfère aussi à tout ce qui a été fait dans le domaine de la biologie et de la santé sur la base de l'utilisation de l'IA[4]. C'est donc aussi pour dire que dans le domaine des matériaux, avec un certain décalage par rapport aux domaines de la biologie, de la pharmacie et de la santé, on veut utiliser des techniques dérivées de l'intelligence artificielle pour accélérer les découvertes de matériaux.

1.2. Les projets en développement

1.2.1. Les projets européens

Il y a des projets européens en cours (***Figure 3***) comme en Suisse le projet Marvel à l'EPFL[5], porté par Nicola Marzari, qui a commencé en 2013. Il y a des projets d'initiatives à l'échelle européenne pour le partage des données scientifiques : un projet tout à fait similaire à DIADEM a été lancé en Allemagne en 2020 qui s'appelle FAIRMAT et nous avons beaucoup d'échanges

3. *Materials Acceleration Platform* : Plateforme pour l'accélération sur les matériaux.
4. IA : intelligence artificielle.

5. EPFL : École polytechnique fédérale de Lausanne.

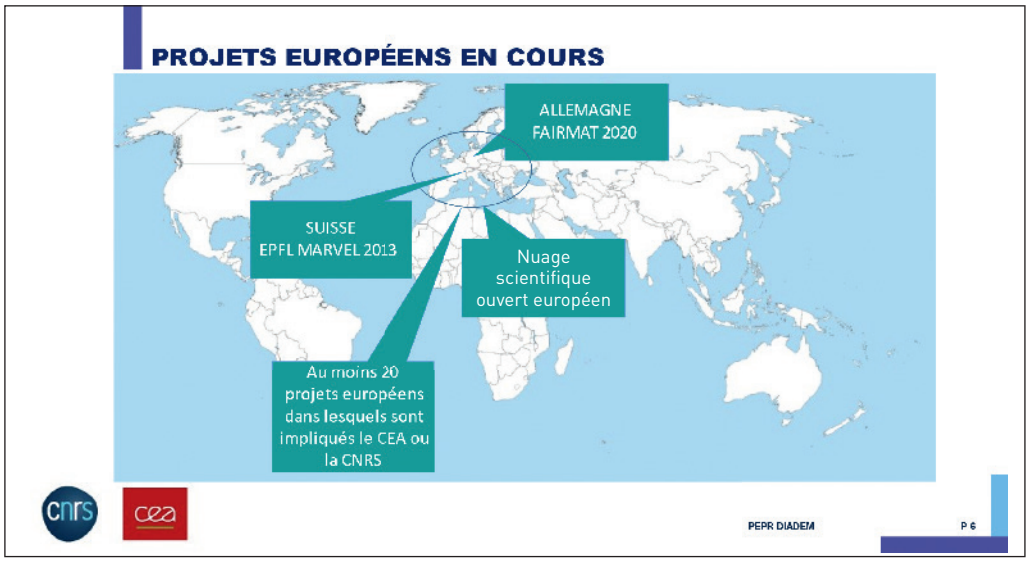


Figure 3

Projets européens en cours.

avec Claudia Draxl qui porte ce projet. Une vingtaine de projets européens, qui pour beaucoup ont à voir avec l'accélération de la découverte des matériaux, dans lesquels le CNRS, le CEA et les universités françaises sont impliqués, ont vu le jour dans la dernière période.

1.2.2. Les projets français

Au niveau national, de nombreuses initiatives préexistaient bien sûr à DIADEM (***Figure 4***). Nous avons de nombreux GDR[6], le dernier en date a été lancé en juillet dernier, il s'appelle IAMAT : Intelligence Artificielle pour les Matériaux. Des initiatives ont été prises par le CEA il y a déjà quelques temps à Saclay, à Grenoble, pour la métallurgie combinatoire. Donc tout un panel de projets et de plateformes de recherches est déjà en cours. DIADEM vient en complément pour essayer de structurer la politique dans ce domaine de recherche et l'un des objectifs est d'essayer de proposer une vision cohérente et une structuration des initiatives en cours.

2 Principes de base du Programme Équipement Prioritaire de Recherche (PEPR) DIADEM

2.1. Lignes directrices et spécificités

Nous avons trois principaux enjeux :

– la découverte accélérée de matériaux pour les grandes transitions : énergétique (batteries filière hydrogène, nucléaire du futur…),

6. GDR : Groupement de recherches.

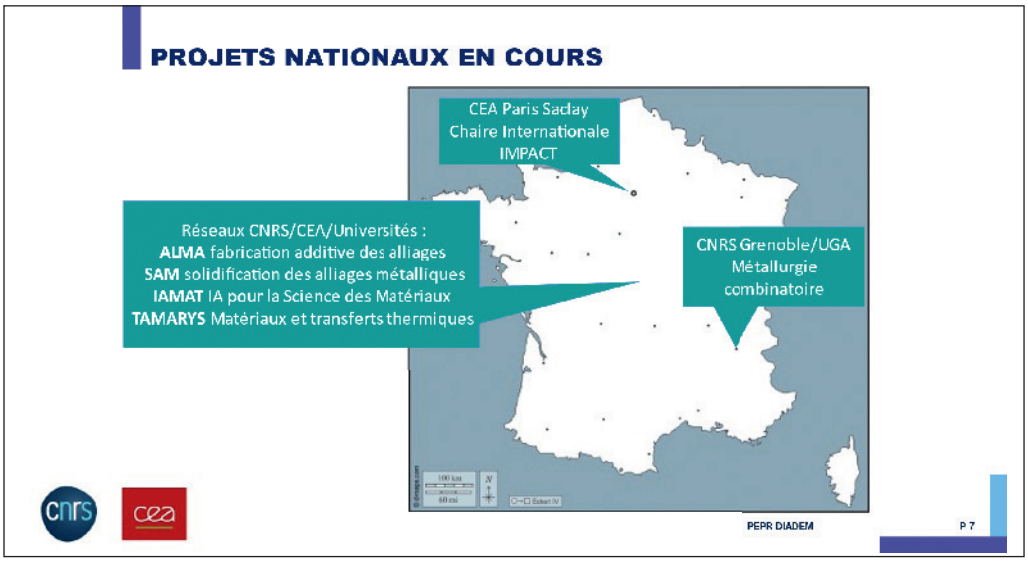


Figure 4

Projets nationaux en cours.
UGA : Université Grenoble-Alpes

environnementale, numérique (électronique…) et sanitaire ;

– la substitution des matières premières critiques et/ou toxiques, qui est un point important de DIADEM, y compris au niveau international ;

– la diversité importante des matériaux et des procédés de mise en œuvre. Donc dans la notion de matériaux, j'insiste sur le fait que la mise en forme, la mise en œuvre, sont des clés pour aboutir à la fonctionnalité. Cela est une spécificité de DIADEM et une complication très importante dans la mise en œuvre de l'intelligence artificielle.

Nos objectifs sont :

– de doter la France d'un réseau de plateformes dédiées à la découverte accélérée des matériaux ;

– ensuite, de mettre à la disposition de la communauté scientifique ces plateformes sous la forme d'appel à projet ;

– et de développer à l'échelle nationale une synergie entre la science des matériaux et l'intelligence artificielle.

2.2. Les différents objets du PEPR DIADEM

1. Les plateformes ont pour objet :

– la conception numérique : à la fois la conception numérique des matériaux et des procédés ;

– la synthèse et la mise en forme à haut débit des matériaux ;

– la caractérisation à haut débit des matériaux ;

– la dernière mais certainement la plus importante pour

notre projet : le lien avec les bases de données et les outils d'intelligence artificielle.

2. Dans un premier temps, nous avons mis en place des démonstrateurs méthodologiques, des projets ciblés, qui ont pour objectif principal la construction des plateformes.

3. Ensuite nous aurons trois vagues d'appels à projets type ANR[7] pour l'utilisation de ces plateformes et leur ouverture à l'international. Ces appels à projets seront au nombre de 30 à 40, pour une aide comprise entre 800 k€ et 1 M€. Ils seront ouverts à toute la communauté scientifique. Nous avons estimé que le nombre de chercheurs et d'ingénieurs dans le domaine des matériaux est actuellement de 4 000 en France.

4. De manière complémentaire, la formation est une clé importante, et donc nous mettrons en place des outils de formation et des Écoles internationales.

2.3. Les plateformes de DIADEM

Les plateformes sont distribuées dans le territoire (***Figure 5***) et sont dédiées à la synthèse, la caractérisation et la modélisation. Il faut noter que les deux synchrotrons[8] localisés sur le territoire français sont impliqués, ce qui est particulièrement important

7. ANR : Agence Nationale de la Recherche.

8. Synchrotron : appareil électromagnétique de grande taille destiné à l'accélération de particules élémentaires.

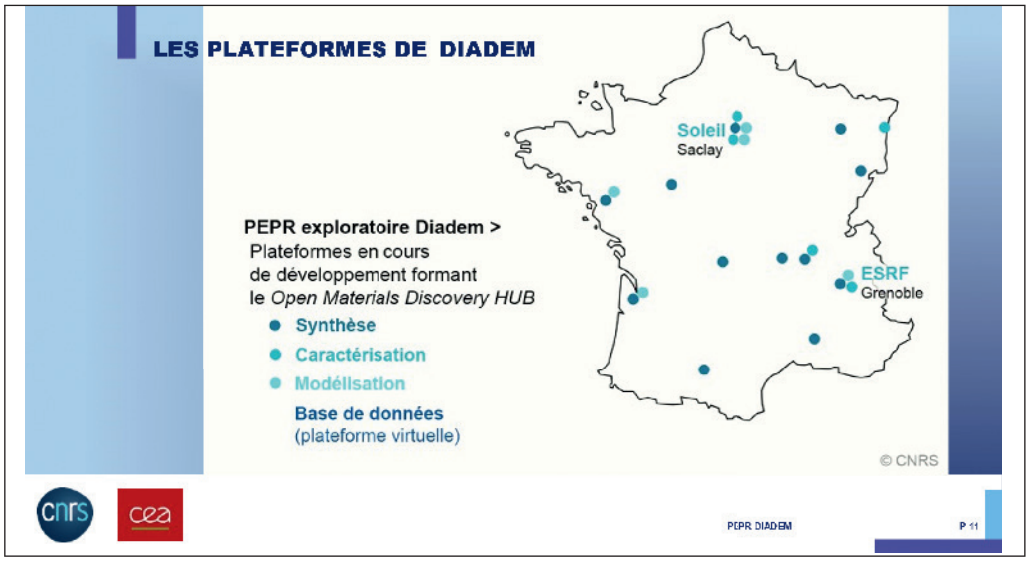


Figure 5

Les plateformes de DIADEM.
Open Materials Discovery HUB : Plateforme ouverte de découverte de matériaux
ESRF (European Synchrotron Radiation Facility) : Équipement européen de radiations synchrotron

quand il s'agit de parler de caractérisation haut débit[9].

3 Construction et validation des plateformes

3.1. Les projets ciblés de démonstration

La construction et la validation des plateformes se fait sur la base de projets ciblés. La *Figure 6* présente le schéma du DIADEM : au centre, au cœur du programme, figurent les plateformes de design[10] numérique et de caractérisation haut débit. Ce qui est en vert représente la mise en œuvre de la science des données, des bases de données et de l'intelligence artificielle qui interagit bien évidemment avec toutes les plateformes.

Les projets que vous avez autour et qui complètent le diadème, sont des projets de synthèse et mise en forme haut débit recouvrant une diversité de matériaux et qui sont en cours de mise en place.

Par exemple, vous avez dans ces différents projets, des projets plutôt autour de la métallurgie comme A Dream (voir le chapitre de Stéphane Gorsse, dans cet ouvrage). Vous avez des projets autour des polymères[11], des matériaux organiques[12], des matériaux poreux[13], par exemple

9. Caractérisation haut débit : méthode de caractérisation avec un nombre important de données.
10. Design : conception.

11. Polymère : grosse molécule formée d'une chaîne d'unité de base : le monomère.
12. Matériaux organiques : matériaux composés principalement de carbone et surtout issus des êtres vivants.
13. Poreux : perméable.

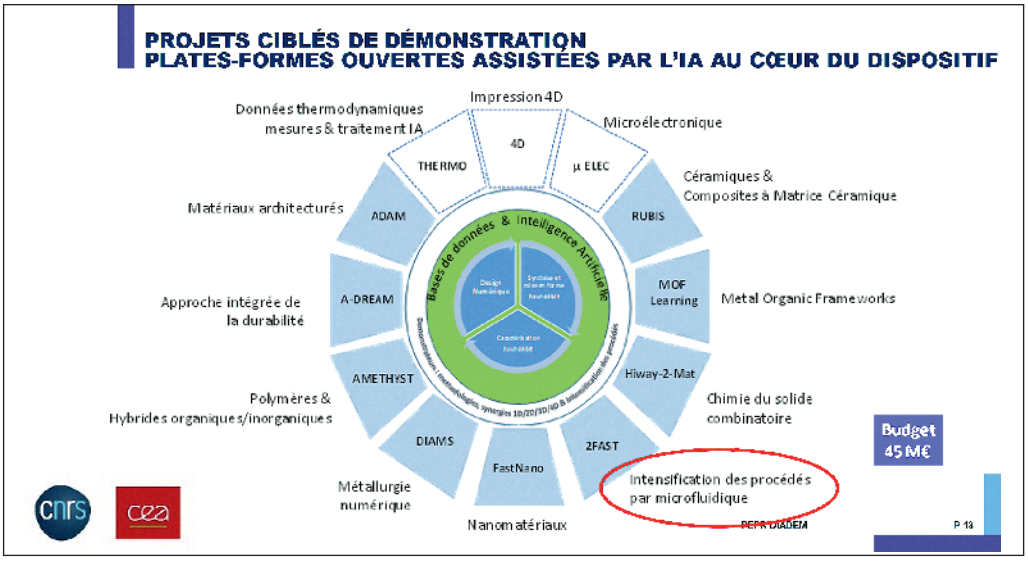


Figure 6

Projets ciblés de démonstration – accent sur l'intensification des procédés par microfluidique.

les MOF[14] (voir le chapitre de François Xavier Coudert qui fait partie de ce projet sur les MOF). Vous avez les projets sur les nanomatériaux[15]. L'ensemble de ces projets, soit au total 17 projets, est déjà mis en place.

Je vais maintenant donner quelques exemples, bien évidemment sans être exhaustif.

3.2. Intensification des procédés par microfluidique : projet 2FAST

3.2.1. Principe

Il s'agit d'un projet qui concerne différents laboratoires, dont le LOF[16] avec la société Solvay. Il s'agit de fabriquer, sur quelques centimètres carrés, des réacteurs dans lesquels vous allez avoir à certains endroits de ce réacteur un mélange et une réactivité entre différents précurseurs chimiques[17] sous forme liquide (***Figure 11***). Ensuite, ces réacteurs microfluidiques[18] génèrent des gouttelettes (ici en bleu), et chacune de ces gouttelettes est un réacteur liquide. Dans chacun de ces réacteurs, on va pouvoir changer de manière très contrôlée et de manière « très haut débit », la synthèse de différents matériaux. Donc la première étape est la synthèse haut débit.

3.2.2. Caractérisation en ligne

Un point très important qui va revenir dans tous les projets de DIADEM, est d'avoir une détection en ligne pendant que la synthèse se réalise. C'est sur le schéma de droite de la ***Figure 7***, la condition pour avoir ce qu'on appelle un réacteur orchestré.

Nous avons différentes spectroscopies et différentes méthodes d'analyse en ligne qui génèrent des données. Dans cet exemple, cela peut être des données optiques, des données de spectroscopie[19], par exemple Raman, des données plus structurales, par exemple de diffusion de rayon X aux petits angles, qui génèrent ces données, et l'objectif est d'utiliser l'intelligence artificielle comme un outil d'orchestration du procédé et d'optimisation de la réactivité en fonction des données qui sont produites en ligne.

3.2.3. Laboratoire autonome

Un objectif important est que cette intelligence artificielle en ligne puisse interagir avec le protocole de la réaction. On a ce que les Américains appellent les « *Materials Acceleration Plateforms* », un laboratoire autonome, en l'occurrence ici, un laboratoire fluidique autonome où la boucle de rétroaction est automatisée. Donc on peut avec toute cette boucle, si on a bien entraîné l'intelligence artificielle et bien mis en

14. MOF (*Metal Organic Frameworks*) : réseaux organométalliques.
15. Nanomatériaux : matériaux à l'échelle du nanomètre (10^{-9} mètres).
16. LOF (*Laboratory of Futur*) : Laboratoire du futur.
17. Précurseurs chimiques : éléments initiaux à partir desquels une réaction chimique peut être réalisée.
18. Microfluidique : techniques de réactions dans des micro canaux (de l'ordre ou inférieur au millimètre).
19. Spectroscopie : étude des spectres électromagnétiques issus de matériaux pour étudier leur émission ou absorption d'énergie suite à une excitation (par de la lumière par exemple).

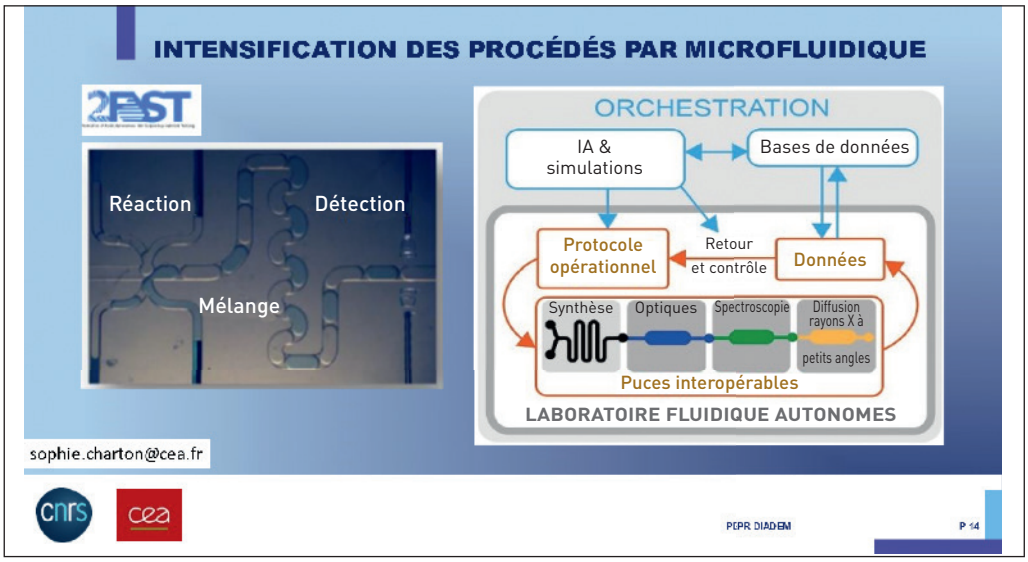


Figure 7

Intensification des procédés par microfluidique.

place les approches d'apprentissage, avoir une rétroaction en direct sur le processus de synthèse. On comprend bien ici l'importance de l'utilisation de l'intelligence artificielle.

Laboratoire autonome, IA, orchestration, base de données, sont tous des mots clés importants, et l'idée des collègues qui portent ce projet est d'orchestrer cet ensemble et de mettre en place les différentes briques pour le construire.

3.3. Un projet de couches minces : Hiway-2-Mat

Le deuxième exemple concerne la chimie du solide combinatoire[20] (***Figure 8***) au sein du projet Hiway-2-Mat. Là on parle de synthèse haut débit combinatoire. Il y a différentes manières de procéder, je vais en décrire une ou deux.

3.3.1. Conception des matériaux

La ***Figure 9*** présente un exemple d'un substrat de quelques centimètres carrés avec trois composés A, B, C. Quand on fait des solutions solides[21] entre ces trois composés, les caractérisations de ces composés peuvent prendre de nombreux mois. Le matériau qui est concerné ici est un matériau qui peut être utilisé pour les piles à combustibles[22] et les électrolyseurs

20. Chimie combinatoire : méthodes de réactions chimiques mettant en place un grand nombre de réactions entre des composés A et B (et parfois C) en faisant varier leurs concentrations.

21. Solution solide : mélange de deux corps purs formant un solide homogène.

22. Pile à combustible : générateur de courant par des réactions d'oxydo-réduction.

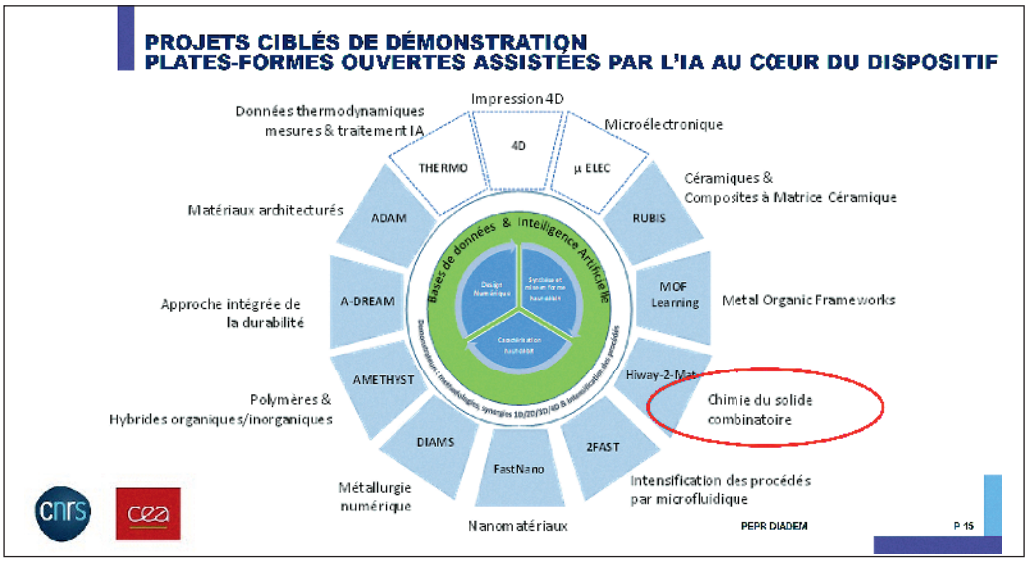


Figure 8

Projets ciblés de démonstration – accent sur la chimie du solide combinatoire.

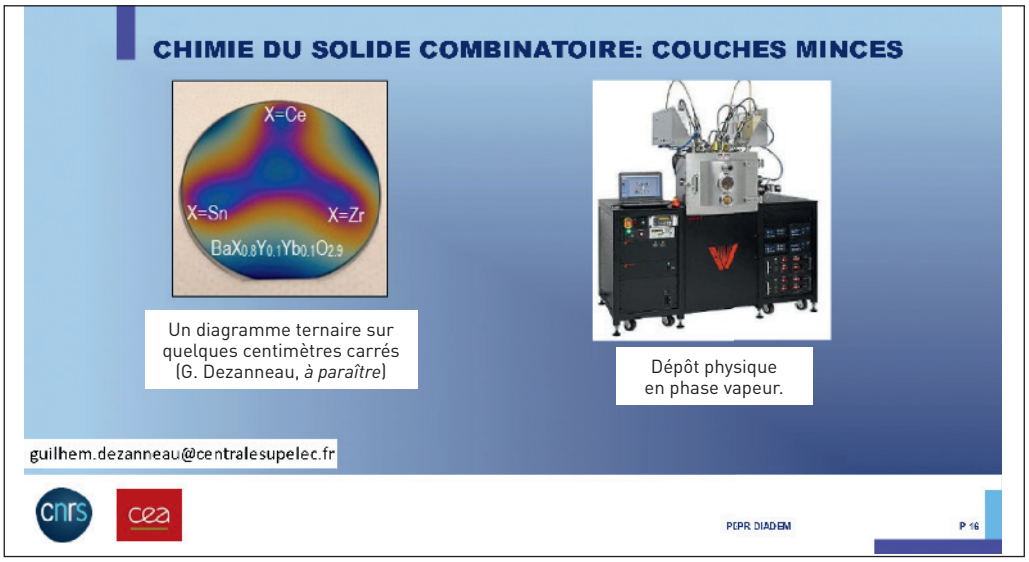


Figure 9

Chimie du solide combinatoire : couches minces.

pour la filière hydrogène, et vous avez ici les trois sommets du triangle du diagramme ternaire pour lequel X représente l'étain, le zirconium et le cérium. Vous avez une continuité de composition chimique sur un même substrat de quelques centimètres carrés.

Cette première étape a été réalisée par une technique dérivée de la PVD « pulvérisation cathodique ». En fait, c'est une technique qui a un gros inconvénient : elle génère des gradients. C'est-à-dire que dans les dépôts sur une grande surface réalisés à partir de cette technique de dépôts, la composition n'est pas fixe en fonction de la position sur le substrat. C'est un gros inconvénient, mais les collègues qui l'ont mis en œuvre ont tiré parti de ces gradients pour avoir un panel de composition continu sur un même substrat. C'est la première étape, il va falloir maintenant analyser les caractéristiques importantes, les descripteurs qui pourront être inclus dans la boucle d'intelligence artificielle.

3.3.2. Analyse des propriétés locales/globales

Si on s'intéresse à la structure des matériaux, quand on balaye la composition, la structure change, mais aussi les propriétés fonctionnelles, en l'occurrence la conductivité ionique. C'est une propriété qui n'est pas locale, donc le challenge[23] c'est d'aller regarder à l'échelle locale sur le substrat, aux différents endroits du substrat, comment varie la conductivité. Donc c'est aussi ce qu'il va falloir mettre en place dans ce projet. Cette machine pour le faire actuellement n'est pas disponible en France, le résultat rapporté ici a été obtenu à l'Université de Barcelone. Le projet dans DIADEM est de monter une machine similaire qui va être déployée au CEA Tech Bordeaux.

23. Challenge : défi.

3.4. Un projet de fabrication de poudre Hiway-2-MaT

3.4.1. Automatisation parallèle

L'objectif dans ce second projet de Hiway-2-MaT (***Figure 10***) est de fabriquer des poudres. Il faut automatiser notre réacteur et notre réaction chimique, les caractérisations structurales, les caractérisations fonctionnelles éventuellement locales et l'analyse des données. On peut automatiser et robotiser ces quatre étapes qui peuvent être faites en parallèle.

3.4.2. Boucle autonome

Mais ensuite il s'agit de combiner toutes ces approches et c'est là qu'on rentre vraiment dans le laboratoire autonome et dans l'utilisation de l'intelligence artificielle pour avoir une boucle qu'on appelle la boucle autonome qui, à partir des données qui sont produites, peut optimiser la préparation, la synthèse des matériaux en l'occurrence. Bien évidemment, si en plus dans cette boucle on arrive à inclure la modélisation, par exemple la modélisation type DFT[24], et si on gère les données de manière reproductible et accessible, on peut peut-être inclure cette boucle d'optimisation par modélisation, et donc

24. DFT : théorie fonctionnelle de la densité, méthode de modélisation.

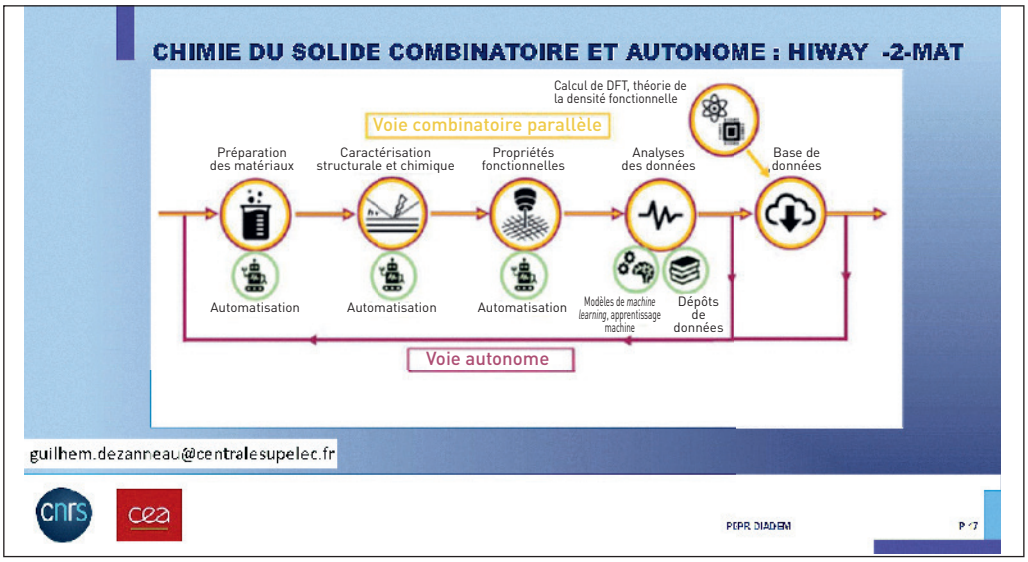


Figure 10

Chimie du solide combinatoire et autonome : Hiway-2-Mat.

on aura un laboratoire autonome expérimental qui en plus sera connecté directement à la modélisation.

Donc il y a beaucoup d'éléments technologiques à mettre en place, et en l'occurrence la robotique va jouer un rôle très important, mais également la science des données. Il faudra que l'ordinateur soit capable de caractériser le composé X, Y que j'ai voulu synthétiser. Je fixe un critère, par exemple la raie de diffraction 100 du composé, et l'ordinateur devra être capable de dire en fonction des paramètres de synthèse si on s'approche ou pas du composé cristallisé qu'on voulait faire.

3.5. Un projet de microélectronique : μ ELEC

Le dernier exemple que je voulais montrer concerne la microélectronique (***Figure 11***).

3.5.1. Conception des substrats

C'est un projet plus récent de DIADEM. Il s'agit aussi d'une technique de dépôt couche mince, donc ici de l'ablation laser combinatoire, où, en utilisant un laser et en pulvérisant des cibles différentes, les collègues ont été capables de réaliser sur un substrat, ici c'est sur quelques centimètres carrés, 600 échantillons. Ce dispositif est disponible depuis plusieurs années au GREMAN à Tours (***Figure 12***). On voit ici ce qu'on appelle les condensateurs[25], les petits rectangles sont des électrodes, et chacun de ces composants peut être testé individuellement du point de vue de sa structure et du point de vue de sa fonctionnalité. Ici il s'agissait

25. Condensateur : composé électronique permettant de stocker des charges.

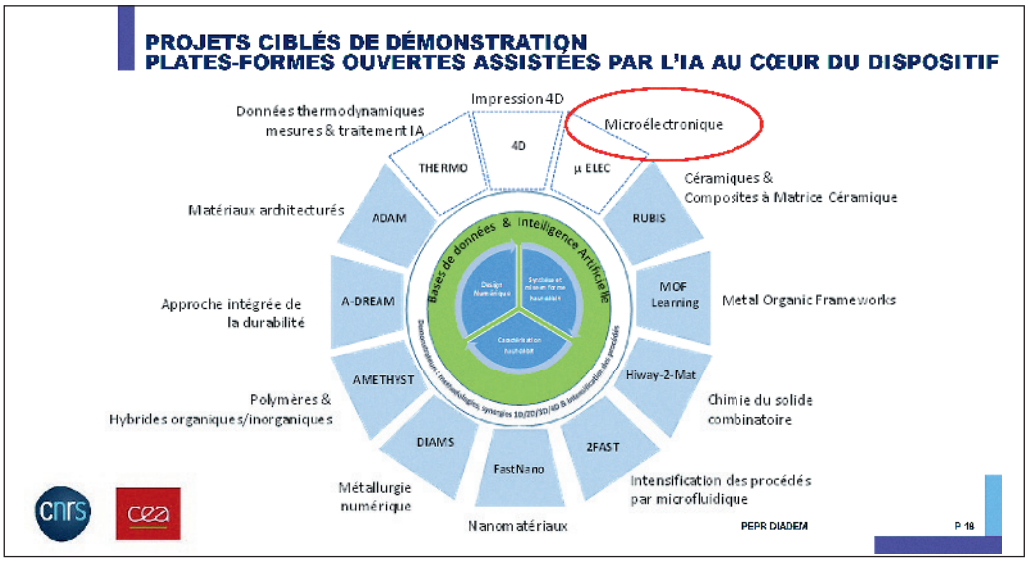


Figure 11

Projets ciblés de démonstration – accent sur la microélectronique.

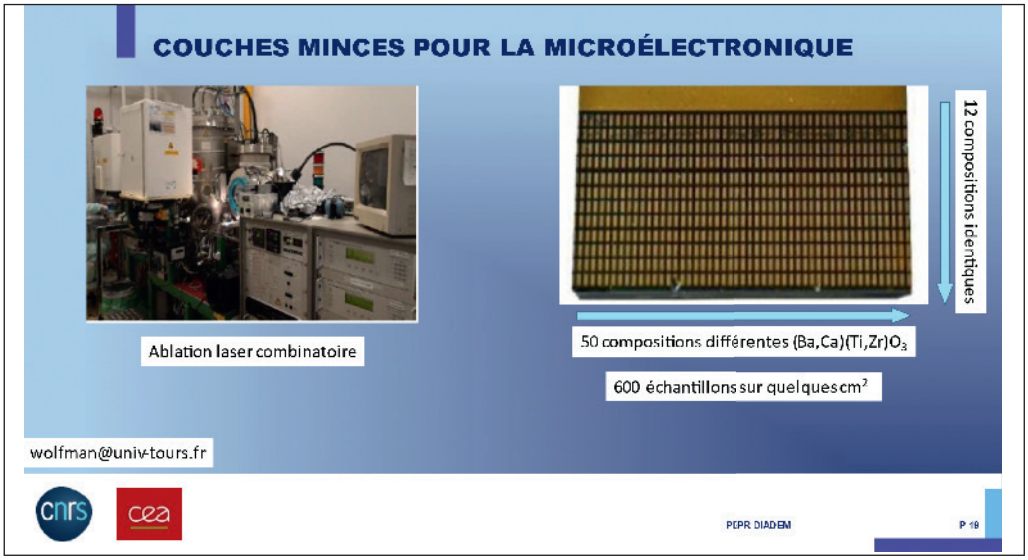


Figure 12

Couches minces pour la microélectronique.

d'une pérovskite contenant du baryum, du calcium, du titane et du zirconium. Le long de l'axe horizontal, c'est la composition en zirconium qui change de manière systématique, donc vous avez 50 compositions le long de l'axe horizontal, et le long de l'axe vertical vous avez une douzaine de compositions qui sont toutes les mêmes, ce qui autorise, en plus, de pouvoir faire de la statistique. Voyons les résultats obtenus.

3.5.2. Résultats des caractérisations

On peut faire une caractérisation structurale et fonctionnelle. Ici nous avons le diagramme ternaire (*Figure 13*), $BaZrO_3$, $BaTiO_3$, $CaTiO_3$, et vous pouvez balayer des lignes dans ce diagramme ternaire qui se traduisent par ces différents échantillons. Vous avez des nappes de fonctionnalité dans cet espace qui est complexe, vous voyez la composition qui varie ici, vous pouvez ajouter un paramètre – ici en l'occurrence c'était le champ électrique – et vous avez votre fonctionnalité qui change de manière systématique en fonction de la composition et d'autres paramètres.

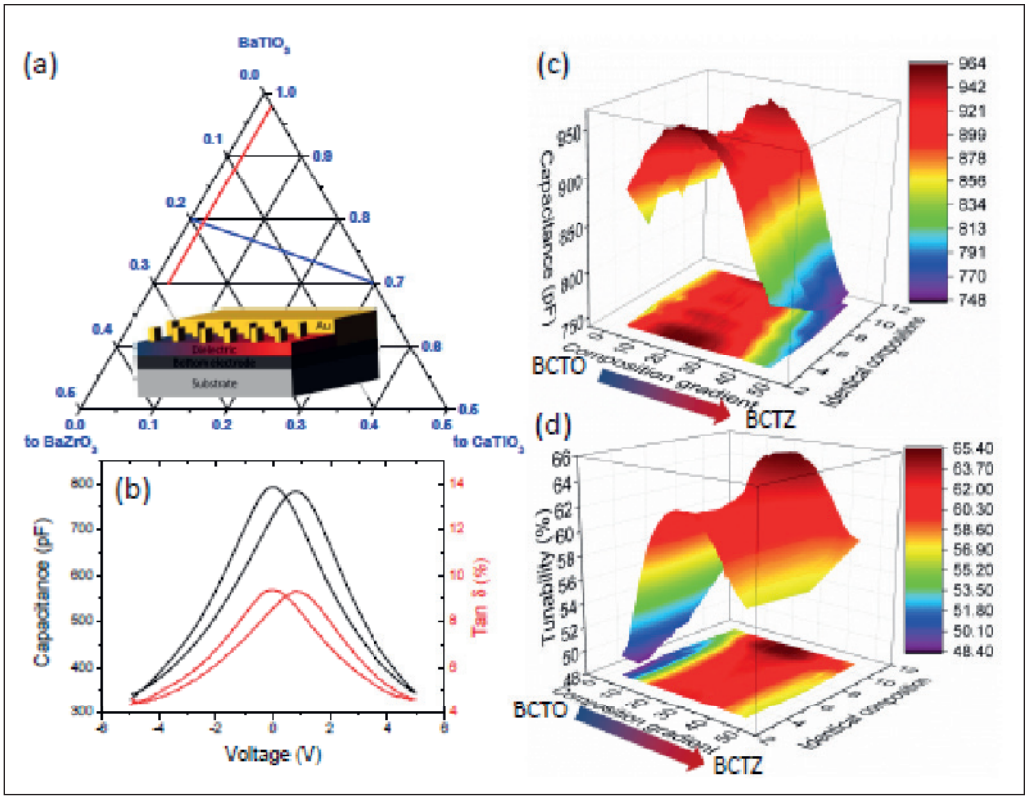


Figure 13

Résultats pour la microélectronique.

Jusqu'à présent le traitement de ces données se faisait à la main, à partir de tableaux Excel qui contenaient un nombre très conséquent de données, et donc l'objectif du projet c'est d'automatiser complètement le travail de traitement de ces données pour optimiser, se retrouver près de la courbe maximale qui concerne ici un maximum de propriétés piézoélectriques de ces matériaux. Le travail long et ingrat qui avait été fait pour optimiser ces matériaux peut être fait de manière totalement automatisée.

4 Déroulement du programme : formations et AAP

4.1. Formations

Le déroulement du programme commence par la formation (*Figure 14*) des 4 000 personnes impliquées au niveau national, pas seulement dans DIADEM, mais des 4 000 chercheurs et ingénieurs en France qui sont focalisés sur les thématiques d'accélération de la découverte des matériaux. Donc la formation sera importante, mais pas que dans la formation initiale, il s'agira de mettre en place une formation continue, et de mettre en place un vrai changement de la culture scientifique des collègues impliqués dans la science des matériaux.

4.2. Budget

Des initiatives seront prises ; on a doté ce programme de trois millions d'euros sur huit ans pour développer des formations initiales et continues dans le domaine de l'accélération de la découverte des matériaux.

Figure 14

Développement d'outils de formation.

4.3. Universités partenaires

Des universités sont partenaires (*Figure 15*) du CNRS et du CEA qui sont les porteurs du projet. Il y a sept universités partenaires dont trois sur la région parisienne : Paris Saclay, Sorbonne Université et Institut Polytechnique de Paris, ainsi que les Université de Lorraine, de Lyon, Université Grenoble-Alpes et l'Université de Bordeaux. Sont aussi impliquées les deux grosses infrastructures de recherche que sont les deux synchrotrons ESRF et SOLEIL. Les plateformes ne sont pas toutes localisées sur les sites universitaires, mais dans les endroits les plus pertinents.

4.4. Déroulement du projet

Ces 17 projets ciblés du programme DIADEM (*Figure 16*) vont se dérouler entre maintenant et 2025-2026. Les plateformes qui sont en cours de mise en place vont se développer pendant toute la durée du projet. Par exemple, la plateforme numérique va monter en régime entre maintenant et la fin du projet en 2029-2030.

À partir de 2024, les plateformes mises en place seront ouvertes à toute la communauté scientifique sous forme d'appels à projets dédiés dont la thématique principale sera l'accélération de la découverte de matériaux grâce aux techniques d'intelligence artificielle, et tout au long du programme nous y associerons un programme de formation.

L'objectif est qu'en 2030, les plateformes soient en place et soient le plus possible utilisables par toute la communauté académique mais aussi industrielle.

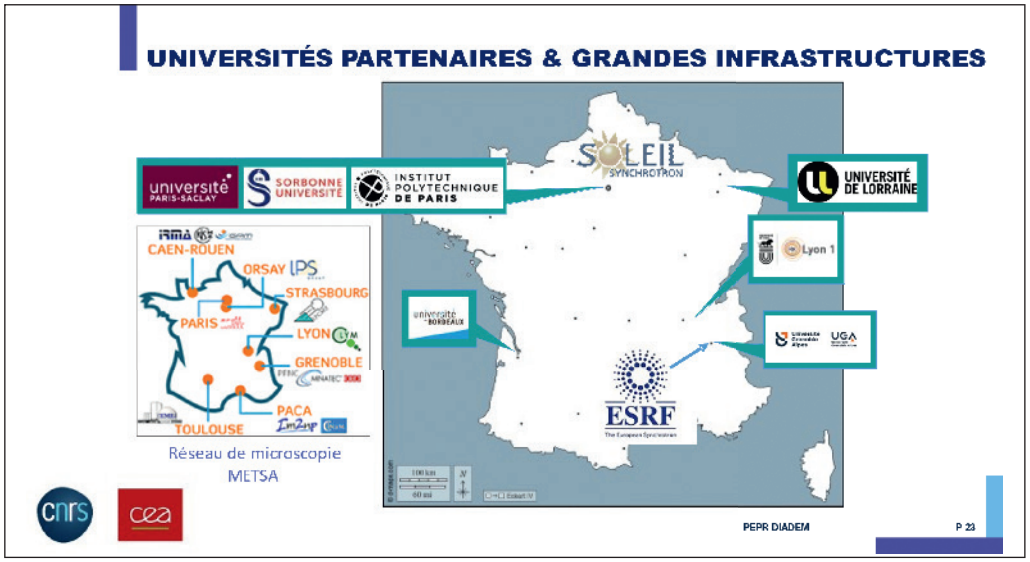


Figure 15

Universités partenaires et grandes infrastructures.

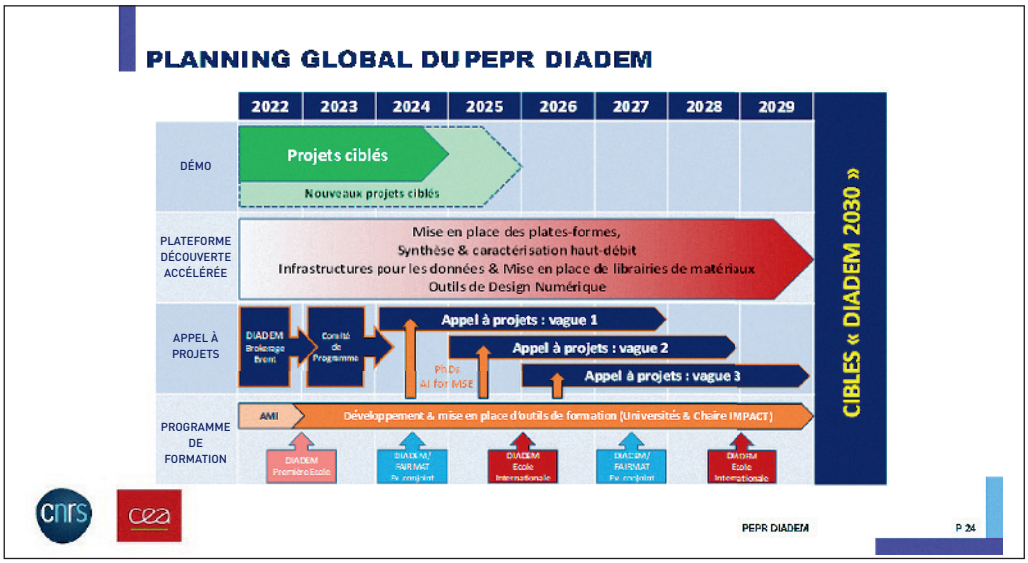


Figure 16

Planning global du PEPR DIADEM.

Conclusion

La ***Figure 17*** montre la photo des 14 projets ciblés qui ont déjà démarré. L'objectif est de découvrir de nouveaux matériaux dans différents domaines, de manière complètement diversifiée grâce à l'intelligence artificielle. Nous revendiquons la diversité des projets car nous voulons montrer qu'approcher une grande diversité de matériaux, c'est le moyen de montrer que l'intelligence artificielle contribue effectivement à la découverte accélérée des matériaux.

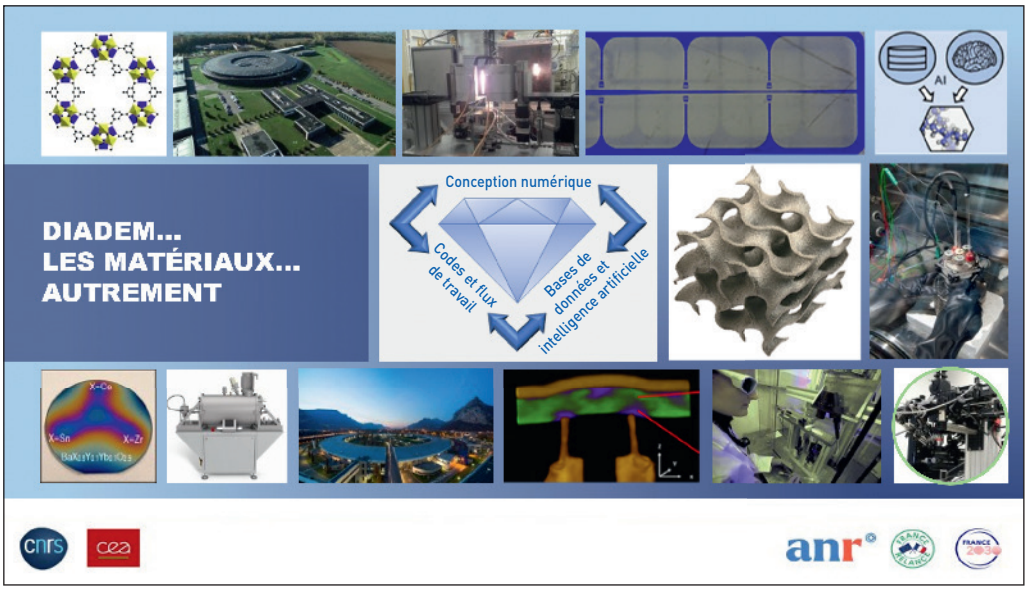


Figure 17

Conclusion : DIADEM... Les matériaux autrement...

Informatique des matériaux : comment combiner puissance des calculs *ab initio*[1] à haut débit et l'intelligence artificielle ?

Gian-Marco RIGNANESE

Gian-Marco Rignanese, Institute of Condensed Matter and Nanosciences (Université catholique de Louvain), School of materials sciences and engineering Northwestern polytechnical University Xian (Chine).

Gian-Marco Rignanese est professeur à l'Université catholique de Louvain (Belgique) et directeur de recherche au F.R.S.-FNRS, l'équivalent du CNRS en Belgique. Son domaine de recherche est la simulation ab initio et l'apprentissage automatique pour la prédiction des propriétés des matériaux. Il a notamment contribué au développement de divers logiciels libres permettant ce type de simulations ainsi que le calcul ab initio à haut débit grâce aux supercalculateurs[2].

1. *Ab initio* : depuis les premiers principes, c'est-à-dire les lois de la mécanique quantique et l'électromagnétisme sans recours à des données expérimentales.
2. Ordinateur capable de réaliser un nombre très élevé d'opérations et de calculs par unité de temps.

Introduction

Tout au long de l'Histoire, les méthodes et concepts ont évolué en science des matériaux. Prenons l'exemple du feu. Il a été découvert un peu par hasard il y a des centaines de milliers d'années. À partir de l'Antiquité, il fut considéré (avec l'eau, la terre, l'éther, et l'air) comme l'un des 5 éléments formant la matière qui constitue l'Univers (***Figure 1***).

Avec l'évolution des sciences vers une compréhension de plus en plus fine des choses (notamment grâce à la mécanique quantique), il s'est avéré que cette vision était complètement erronée. En fait, la chaleur générée par une combustion amène les électrons des molécules (ou des atomes), qui composent l'air ambiant, dans un état d'énergie élevée, dit *état excité*. Et de la lumière est émise lorsque ces électrons redescendent à un niveau d'énergie inférieur, dit *état fondamental* (***Figure 2***).

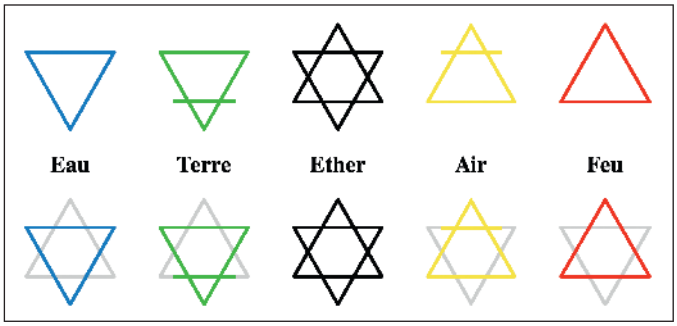


Figure 1

Les 5 éléments.

1 Premier paradigme : la science empirique

Cet exemple du feu montre que l'empirisme a existé depuis le début de l'humanité, dès l'âge de Pierre. La roue n'a d'ailleurs pas été découverte du premier coup, il y a eu de nombreux essais et erreurs (***Figure 3***). Cette approche est encore aujourd'hui le cœur de la recherche expérimentale empirique.

Prenons l'exemple plus récent de **la lampe à incandescence**,

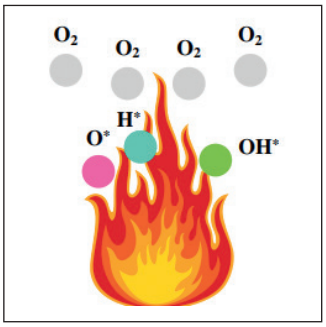


Figure 2

Le feu – les électrons des molécules (ou des atomes) de l'air ambiant excités par la chaleur retournent à leur état fondamental en émettant de l'énergie lumineuse.

Figure 3

L'invention progressive de la roue.

même si elle n'existe plus de nos jours. Lors de sa mise au point en 1879, Edison a testé comme filament près de 3 000 matériaux différents (y compris du papier, du carbone, de l'écorce d'arbre et même un poil de barbe de son assistant) (*Figure 4*). Après ces nombreux essais et erreurs, il a retenu une fibre de coton carbonisée. Ce n'est que quelques années plus tard que le tungstène a fait son apparition comme filament. Il fut pourtant le matériau longtemps utilisé dans nos lampes à incandescence. Cet exemple montre une des limites des essais empiriques. En effet, Edison n'a pu tester que les matériaux qu'il avait sous la main.

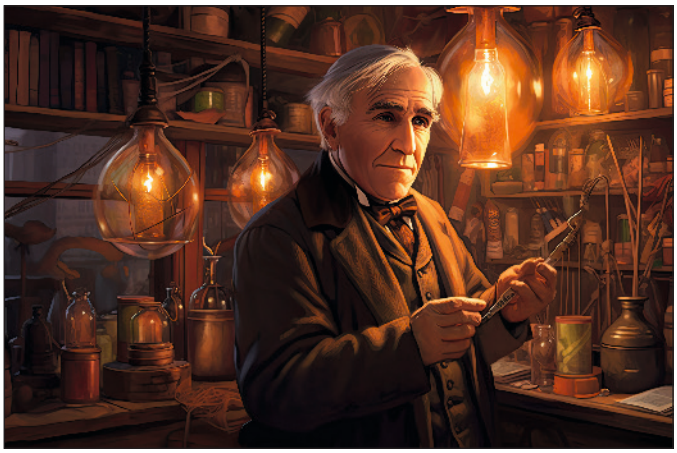


Figure 4

Lampe avec un fil à incandescence de Thomas Edison.

Source Adobe Stock

2 Deuxième paradigme : la science théorique

L'approche expérimentale est restée celle privilégiée pendant longtemps. La **science empirique** est le premier paradigme qui a été utilisé en sciences (*Figure 5*). Vers la fin du XVIe siècle, des lois physiques et chimiques ont petit à petit été établies pour mieux décrire les phénomènes se déroulant aux différentes échelles.

La **science théorique** est ainsi devenue le deuxième paradigme utilisé pour approcher les sciences en général, et donc pour ce qui nous intéresse la connaissance des matériaux.

En sciences des matériaux, il s'agit d'établir des lois scientifiques de cause à effet entre le procédé utilisé, la structure qui en résulte, les propriétés qui en dérivent, et finalement la performance qui en découle dans une application particulière, comme schématisé sur la *Figure 6*.

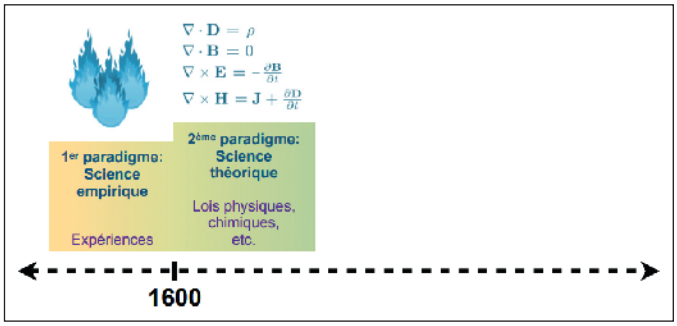


Figure 5

Frise chronologique des paradigmes – 1600.

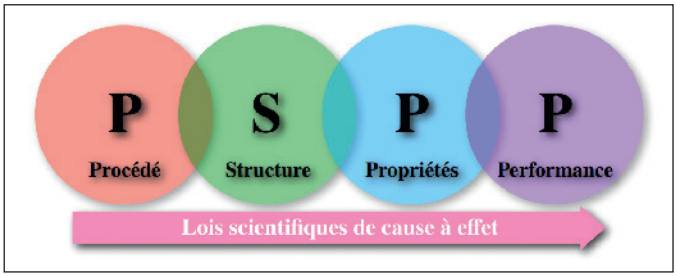


Figure 6

Quadriptyque des lois scientifiques de cause à effet en science des matériaux.

Ces lois scientifiques impliquent des phénomènes et des échelles de longueur/temps différentes. Prenons comme exemple (***Figure 7***) la comparaison des échelles de longueur dans deux cas : la capture de la lumière par une forêt ou par des modules photovoltaïques. Les systèmes de départ sont à l'échelle du mètre. Descendons en dimension dans la connaissance des deux systèmes : à l'échelle du centimètre (10^{-2} m), les feuilles sont à comparer aux cellules photovoltaïques, puis

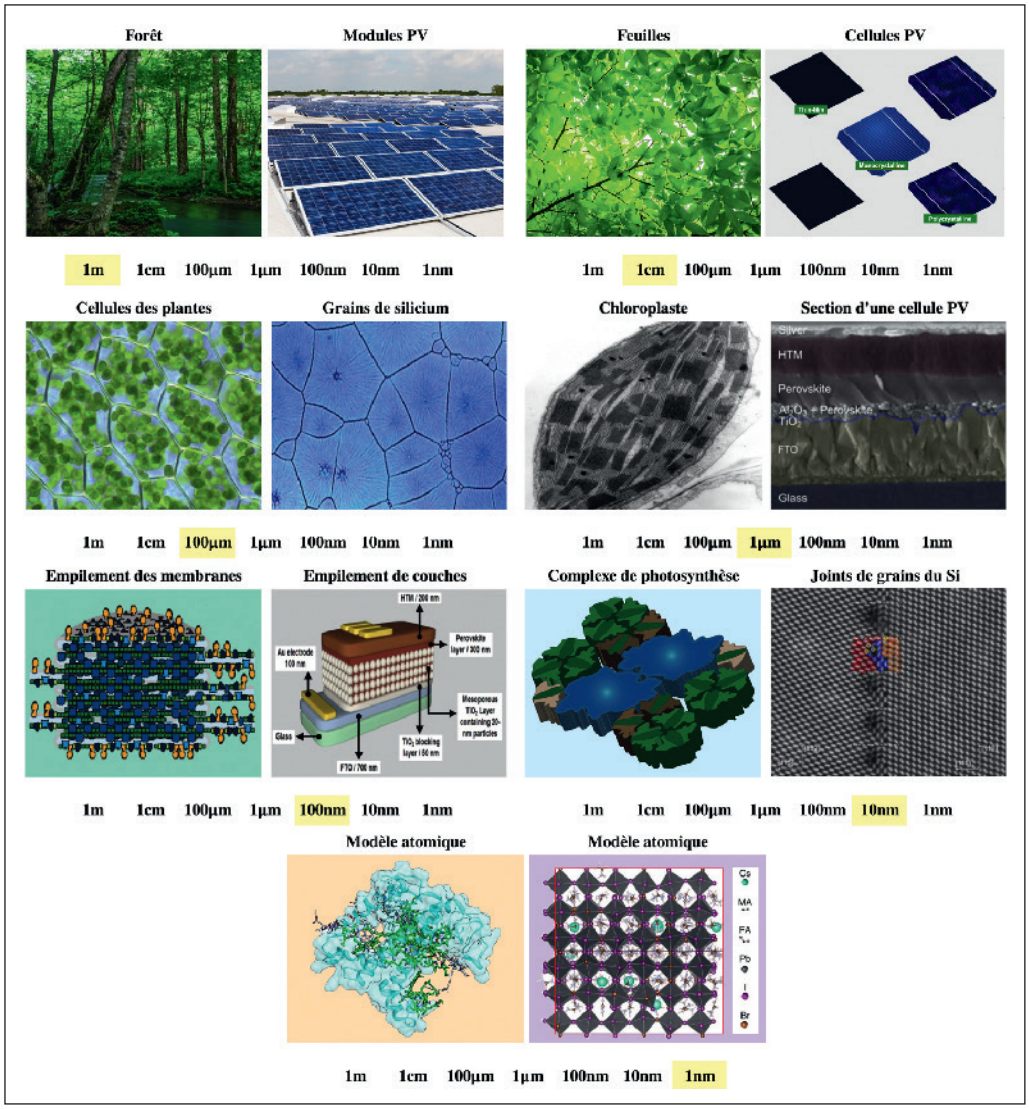


Figure 7

Illustration des différentes échelles de longueur importantes pour la capture de lumière par une forêt d'une part et par un module photovoltaïque d'autre part.

les cellules des plantes aux grains de silicium, à l'échelle du micromètre (10^{-6} m), le chloroplaste[3] peut être comparé à une section de la cellule photovoltaïque, puis les empilements membranaires avec les empilements de couches dans le matériau de la cellule photovoltaïque, et enfin le complexe de la photosynthèse avec les joints de grain du silicium[4]. À l'échelle du nanomètre (10^{-9} m), il y a dans les deux cas des atomes.

Il n'est pas nécessaire de descendre à des échelles plus petites pour essayer de comprendre les propriétés des matériaux. Par contre, selon l'échelle à laquelle on se place, il est important de comprendre que les phénomènes qui se déroulent sont différents, et il en est de même pour les échelles de temps. À titre d'exemple, certains phénomènes peuvent se dérouler sur une durée de l'ordre de la femtoseconde, soit un milliardième de millionième de seconde.

3 Troisième paradigme : la science computationnelle

Le développement des ordinateurs puis des supercalculateurs a révolutionné le monde. En 1945, un des premiers ordinateurs, l'ENIAC, pouvait exécuter de l'ordre de 500 opérations en virgule flottante[5] (***Figure 8***). Il coûtait l'équivalent actuel de 6 millions d'euros. Aujourd'hui, avec une somme environ quinze fois plus grande, vous avez non pas un ordinateur, mais un centre de calcul qui exécute 10^{18} (un milliard de milliards) opérations par seconde. Le gain de puissance a donc été exponentiel.

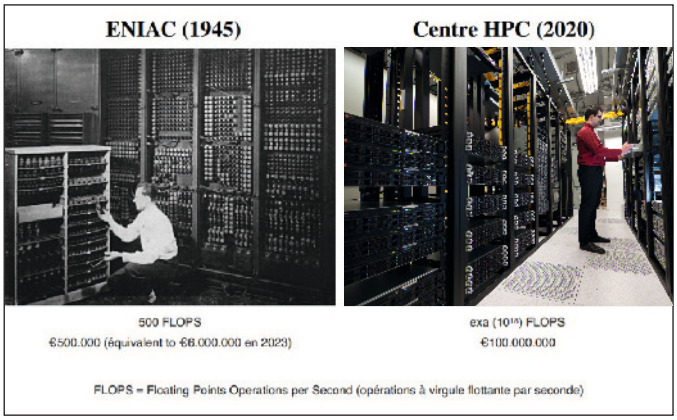


Figure 8

De l'ordinateur au supercalculateur et au centre de calculs.
Sources : Photo Gauche https://fr.wikipedia.org/wiki/ENIAC#/media/Fichier:ENIAC-changing_a_tube.jpg, Photo Droite Adobe Stock

3. Fait partie de la cellule végétale et assure entre autres la photosynthèse.
4. Interface entre deux cristaux qui ont des orientations différentes.

5. La rapidité de calcul d'un ordinateur est mesurée par le nombre d'opérations en virgule flottante qui peuvent être effectuées par seconde (en anglais : *floating-point operations per second* ou FLOPS). Le terme de virgule flottante se réfère au fait que les calculs (additions ou multiplications) peuvent être effectués tant pour de très grands que pour de très petits nombres en les représentant par une mantisse et un exposant (en notation binaire, la mantisse est le nombre initial sans virgule et l'exposant indique la position de la virgule). Les opérations en virgule flottante prennent nettement plus de temps de calcul que des opérations sur les nombres entiers.

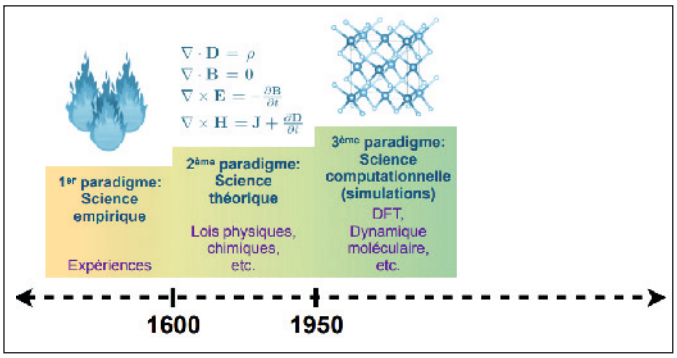


Figure 9

Frise chronologique des paradigmes – 1950.

En science des matériaux, les supercalculateurs ont permis de numériser les lois physiques et donc de réaliser des expériences virtuelles. Il est ainsi devenu possible de faire des prédictions, ce qui a induit un nouveau changement de paradigme : le passage à **la science computationnelle** (*Figure 9*).

Différentes échelles de longueur/temps peuvent être simulées à partir de différentes méthodes selon la précision requise (*Figure 10*). Le **calcul *ab initio*** est la technique la plus précise (en bas à gauche sur le schéma). Il nécessite de prendre en compte explicitement le comportement des atomes et de leurs électrons par le biais des lois de la mécanique quantique et de l'électromagnétisme, sans avoir recours à des données expérimentales. Cette technique permet de prédire toute une série de propriétés des matériaux avec un très bon accord avec celles mesurées expérimentalement. Toutefois, cette précision va de pair avec un temps de calcul nécessaire fort important même sur des supercalculateurs. Dès lors, les systèmes qui peuvent être ainsi simulés sont typiquement constitués de quelques centaines d'atomes, soit sur une échelle allant de l'Angström (1Å = 10^{-10} m) au nanomètre (10^{-9} m), et ce, sur une échelle de temps également très faible, celle des femtosecondes (10^{-15} s).

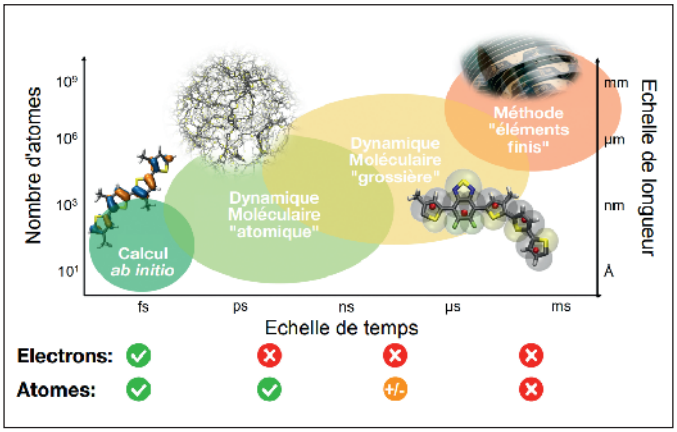


Figure 10

Précision des différentes méthodes de calcul en fonction des échelles longueurs/temps des systèmes et phénomènes à simuler.

Pour passer à des échelles de longueur et de temps supérieures, il n'est plus possible de considérer les électrons explicitement : il faut traiter les atomes dans leur ensemble. La technique de simulation de choix est la **dynamique moléculaire « atomique »**. En acceptant de perdre en précision suite à l'approximation des interactions interatomiques par des potentiels empiriques ou au mieux semi-empiriques[6], il est possible de gagner en termes d'échelle de longueur (pour atteindre quasiment le micromètre, 10^{-6} m) et de temps (pour simuler jusqu'à quelques nanosecondes, 10^{-9} s). En sacrifiant encore la fidélité du modèle en ne traitant plus que les interactions entre groupes d'atomes par le biais de la **dynamique moléculaire « grossière »**, les échelles sont encore augmentées pour atteindre des longueurs de quelques centaines de micromètres et des durées proches de la milliseconde. Finalement, afin d'atteindre des échelles de longueur et de temps encore plus élevées, il devient nécessaire de considérer les matériaux comme des milieux continus (sans même traiter les atomes de façon explicite) et d'avoir recours à la **méthode des « éléments finis »**. Elle permet de simuler des matériaux à des dimensions qu'on peut voir à l'œil nu, mais elle requiert d'introduire divers paramètres dans les modèles dont la valeur est obtenue empiriquement, ce qui réduit encore la précision.

6. Les potentiels empiriques ont une forme analytique et les paramètres qui interviennent sont ajustés de façon à ce que les résultats soient en bon accord avec les expériences disponibles. Dans ces conditions, il est quasiment impossible de se faire une idée de la qualité des prédictions. Les potentiels semi-empiriques ont également une forme analytique mais les paramètres sont ajustés à la fois sur les expériences disponibles et sur d'autres calculs plus précis. La qualité des prédictions peut être testée par comparaison avec quelques nouveaux calculs plus précis.

3.1. Utilisation du calcul *ab initio* pour la conception de matériaux

La conception de matériaux vise à trouver ceux dont les propriétés respectent un ensemble de critères liés à une application particulière. Par exemple, le matériau qui constitue les écrans des téléphones portables doit être à la fois transparent à la lumière (afin que nous puissions voir les informations qui apparaissent dessus) et conducteur (pour permettre la transmission de signaux électriques au contact de nos doigts). Comme nous l'avons vu pour le cas de la lampe à incandescence, l'approche par essai et erreur peut être longue et coûteuse. Aujourd'hui, le calcul *ab initio* a atteint une maturité telle qu'il permet d'accélérer fortement ce processus de sélection grâce au **criblage à haut débit**.

En effet, ayant vu le jour dans les années 1980, le calcul *ab initio* s'est fortement développé dans les années 2000. Alors qu'au début, une thèse de recherche basée sur cette approche permettait d'étudier quelques propriétés pour un seul matériau, de nos jours, grâce à l'augmentation de la puissance des supercalculateurs, à la stabilité des

programmes informatiques et à une automatisation des calculs, il est possible de prédire de nombreuses propriétés pour plusieurs milliers voire centaines de milliers de matériaux. Typiquement, le criblage à haut débit agit comme un entonnoir (*Figure 11*). La propriété requise la moins coûteuse en temps calcul (propriété 1) est déterminée pour tous les matériaux envisagés (de l'ordre de 10^3 à 10^5). Ne sont retenus à l'étape suivante que ceux pour lesquels la propriété 1 rencontre le critère de sélection, ce qui réduit le nombre de matériaux qui seront considérés pour le calcul de la propriété 2. Au fur et à mesure que de nouvelles propriétés sont calculées, les critères de sélection se font de plus en plus stricts et le nombre de matériaux diminue considérablement (ce qui permet au passage de faire des calculs plus complexes et donc plus coûteux).

Ce tri se termine généralement avec de l'ordre de dix à une centaine de matériaux qui remplissent tous les critères et qui peuvent donc être proposés aux expérimentateurs pour qu'ils soient validés. Le dernier mot revient à l'approche expérimentale qui reste donc toujours très importante. Les changements de paradigme ne font pas disparaître les anciens. Mais si le travail de simulation a été bien fait, cette technique fait gagner beaucoup de temps aux expérimentateurs qui font ensuite la synthèse et caractérisent les matériaux triés.

Le haut débit (le fait de pouvoir traiter beaucoup de matériaux en un temps restreint) est fondamental dans cette démarche. C'est lui qui permet d'accélérer la conception. Il est atteint grâce à l'automatisation des calculs.

Afin d'illustrer cette approche, prenons l'exemple de la conception de matériaux absorbeurs des cellules

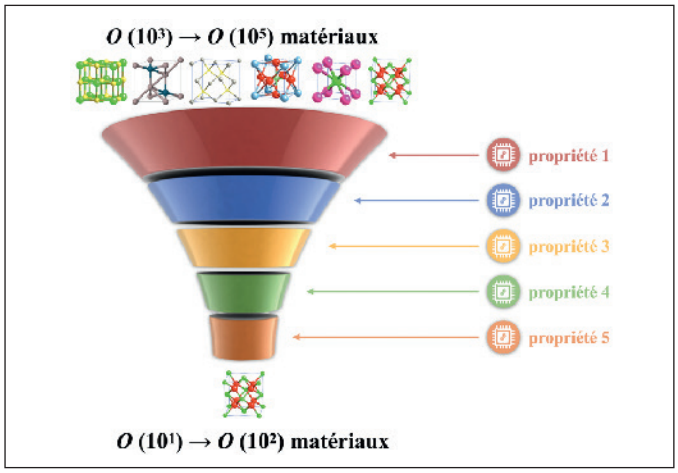


Figure 11

Principe de l'entonnoir pour le criblage à haut débit des matériaux. Partant de 10^3 à 10^5 matériaux, il permet d'en identifier 10 à 100 dont les propriétés rencontrent les critères requis.

photovoltaïques. Ces matériaux constituent la couche qui capture la lumière du soleil et la convertit en électricité. Pour que cela soit possible, ces matériaux doivent avoir des propriétés qui respectent toute une série de critères. Ainsi, au niveau de leurs propriétés électroniques, la bande interdite[7] doit être suffisamment petite pour permettre de capturer des photons[8]. Mais ils doivent aussi être stables thermiquement, être de bons conducteurs, et idéalement être peu sensibles à la présence de défauts qui limitent la production électrique. Ces différentes propriétés seront donc calculées les unes après les autres et permettront un filtrage tout au long de l'entonnoir de criblage (***Figure 12***). Dans l'étude en question [D. Dahliah *et al.*, *Energy Environ. Sci.* 14, 5057 (2021)], il s'agissait de cribler près de 8 000 matériaux à base de cuivre (un élément assez abondant sur terre). En bout de course, il n'en restait plus que 6. Il arrive qu'une même propriété soit calculée plusieurs fois au cours du

7. La mécanique quantique a permis de montrer que dans un atome isolé, les électrons ne peuvent posséder que des énergies de valeurs discrètes et bien définies, par contraste au continuum d'énergie dans le cas d'un électron parfaitement libre (non lié à un atome). Dans un solide, la situation est intermédiaire : l'énergie d'un électron peut avoir n'importe quelle valeur à l'intérieur de certains intervalles, les bandes d'énergies permises. Ces intervalles sont séparés des bandes d'énergie interdites (ou, plus simplement, bandes interdites) dans lesquelles on ne trouve aucun niveau d'énergie accessible aux électrons.

8. En mécanique quantique, un photon est un paquet d'énergie élémentaire associé à une onde électromagnétique (par exemple, une onde de lumière).

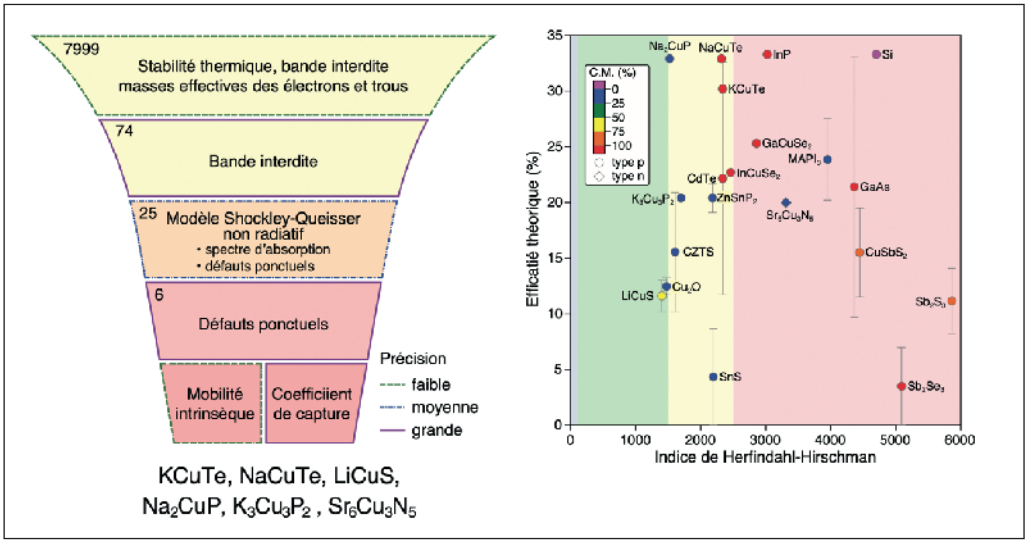


Figure 12

Comment choisir un matériau absorbeur solaire pour le photovoltaïque ?

processus avec une précision différente (voir couleur des bords de l'entonnoir dans la *Figure 12*) : le premier calcul de faible précision (effectué pour beaucoup de matériaux) permet de faire un tri grossier tandis que le calcul de grande précision (pour nettement moins de matériaux) conduit à un tri plus fin.

Suite à ce criblage, il se peut qu'un matériau ait été écarté à tort (les calculs ne sont pas toujours fiables à 100 %). Cependant, les matériaux qui sortent de l'entonnoir sont ceux pour lesquels les certitudes sont les plus fortes. Outre les critères sur les propriétés des matériaux, d'autres contraintes peuvent être prises en compte. Ainsi, dans l'étude ci-dessus, des indicateurs liés au développement durable avaient également été pris en compte. Ces matériaux sont-ils disponibles à différents endroits du globe ? Sont-ils accessibles pour tout le monde ? Y a-t-il des difficultés à les extraire ? Y en a-t-il suffisamment sur la croûte terrestre ? Dans la partie droite de la figure, l'efficacité théorique calculée pour la conversion de la lumière en électricité est représentée pour les différents matériaux en fonction de l'indice de Herfindahl-Hirschman qui mesure la concentration du marché, c'est-à-dire du nombre d'entreprises qui produisent le matériau considéré. De surcroît, les différents matériaux ont été colorés en fonction de leur

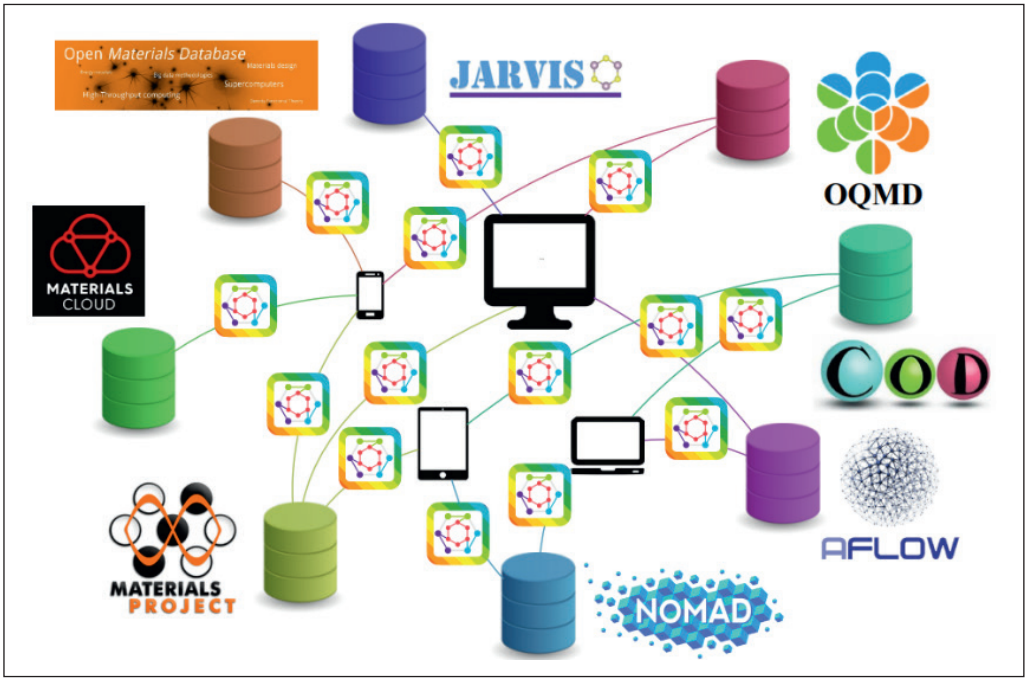


Figure 13

Interrogation des bases de données de matériaux au travers du protocole OPTIMADE.

« compagnonnage » (C.M. en %) qui indique la mesure dans laquelle les éléments qui le constituent ont été obtenus en grande partie ou entièrement en tant que sous-produit de l'extraction d'autres éléments à partir de minerais géologiques. Un matériau de faible compagnonnage est constitué d'éléments qui sont principalement exploités pour eux-mêmes et, à l'inverse, un matériau de haut compagnonnage est constitué d'éléments qui sont obtenus comme les sous-produits de l'exploitation d'autres éléments (par exemple, plus de 90 % des approvisionnements en sélénium et en tellure sont associés au cuivre). Tous les indicateurs disponibles peuvent être pris en compte lors du criblage afin de proposer à l'expérimentation les matériaux les plus pertinents.

3.2. Utilisation des bases de données disponibles en ligne : le consortium OPTIMADE

Tous les calculs effectués au cours du criblage ne sont pas perdus même si un matériau n'est pas retenu au bout du compte. En effet, les résultats sont stockés dans des bases de données qui peuvent être mises à disposition en ligne. Citons par exemple AFLOW, COD, JARVIS, the Materials Cloud, the Materials Project, NOMAD, ou OQMD (*Figure 13*). Chacune de ces bases de données ayant ses spécificités, il est intéressant de pouvoir les combiner. C'est ce que le consortium OPTIMADE s'est attaché à faire en définissant un protocole unique pour interroger toutes ces bases de données. Cela contribue à rendre les données plus facilement « trouvables » (Findable), « accessibles » (Accessible), « interopérables » (Interoperable), « réutilisables » (Reusable) : c'est le concept des « FAIR-data » qui a pris de l'ampleur dernièrement. En effet, pouvoir obtenir le maximum de données est très important pour les modèles de *machine learning*[9].

4 Quatrième paradigme : la science des données

Cette multiplication de bases de données de propriétés des matériaux a rendu possible l'utilisation de l'intelligence artificielle, de l'apprentissage automatique et du minage de données. Cette évolution a conduit au quatrième changement de paradigme apparu au début des années 2000 : la science des données (*Figure 14*).

9. Processus d'apprentissage automatique par l'intelligence artificielle.

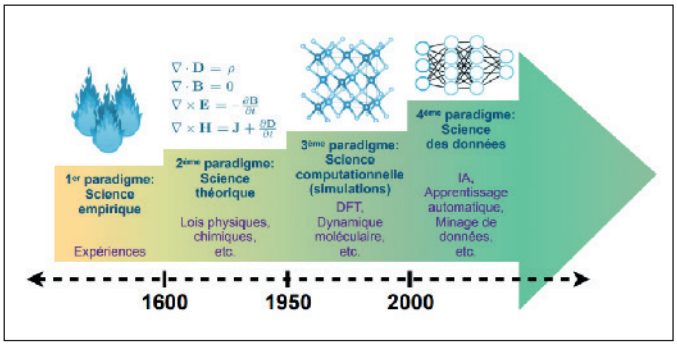


Figure 14

Frise chronologique des paradigmes – 2000 et plus.

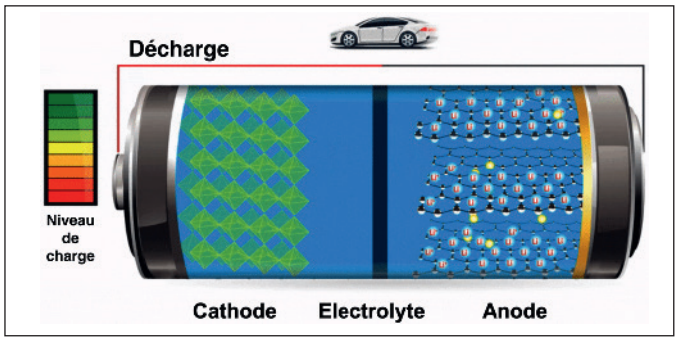


Figure 15

Schéma d'une batterie.

4.1. Intelligence artificielle prédictive

En science des matériaux, les techniques de régression de l'apprentissage automatique permettent ainsi d'établir des modèles pour les relations entre le procédé, la structure, les propriétés et les performances (***Figure 6***). Il est ainsi possible de faire des prédictions en une fraction de seconde, ce qui constitue une accélération considérable par rapport aux calculs *ab initio* présentés précédemment.

L'apprentissage automatique facilite également la connexion entre les différentes échelles de longueurs et de temps (***Figure 10***). Par exemple, pour la dynamique moléculaire atomique, des potentiels décrivant les interactions entre les atomes peuvent être « appris » sur la base de calculs *ab initio* et en utilisant de l'apprentissage automatique. De même, pour passer à l'échelle supérieure, les paramètres nécessaires peuvent être appris au départ d'une série de simulations de dynamique moléculaire. L'avantage principal est d'avoir à l'échelle supérieure des calculs de précision quasi similaire à celle de l'échelle inférieure, mais avec une vitesse nettement supérieure.

Une des utilisations récentes de cette approche a porté sur l'amélioration des matériaux pour les électrolytes des batteries Li-ion. Celles-ci consistent en une cathode[10], un électrolyte[11] et une anode[12] (***Figure 15***). Quand la batterie se charge, les ions Li+ quittent le matériau de la cathode et vont se stocker dans le matériau de l'anode. Quand elle se décharge, c'est-à-dire quand elle produit le courant électrique (par exemple pour alimenter une voiture), les ions Li+ font le mouvement inverse, migrant de l'anode à la cathode. Les électrons suivent le mouvement opposé aux ions Li+ tant à la charge qu'à la décharge. Le mouvement de migration des ions Li+ se fait au travers de l'électrolyte.

Dans les batteries actuelles, l'électrolyte est un liquide. Celui-ci est malheureusement inflammable, ce qui a conduit à divers problèmes avec les batteries (***Figure 16***).

Pour éviter de tels problèmes, qui ont immobilisé des avions Boeing au sol pendant de nombreux mois en entraînant de grosses pertes économiques, une des solutions proposées est de remplacer l'électrolyte liquide par un électrolyte solide qui n'a plus ce problème d'inflammabilité (***Figure 17***).

10. La cathode est l'électrode positive et le siège des réactions de réduction.
11. L'électrolyte constitue la jonction entre les 2 électrodes : elle doit permettre de conduire le courant électrique et les ions.
12. L'anode est l'électrode négative et le siège des réactions d'oxydation.

Cependant, la mobilité des ions Li+ dans les solides est nettement moindre que dans les liquides, ce qui affecte l'efficacité des batteries à électrolyte solide.

La recherche actuelle vise donc à trouver des électrolytes solides avec des mobilités des ions Li+ similaires à celles dans les électrolytes liquides. L'intelligence artificielle a permis d'accélérer la dynamique moléculaire atomique tout en gardant la précision des calculs *ab initio*. La durée des simulations a donc pu être fortement allongée tout en réduisant le calcul, ce qui a permis de travailler à des températures plus réalistes dans les simulations. Les simulations ont donc fourni des prédictions nettement plus proches des résultats expérimentaux. De nouveaux électrolytes solides ont ainsi vu le jour.

Figure 16

Batteries qui prennent feu : un danger venant de l'électrolyte.
Sources : Photo Gauche : https://renew.org.au/renew-magazine/sustainable-tech/lithium-battery-fires-and-safety/ « Image : Jim Heaphy, California (CC BY-SA 3.0) », Photo Droite Adobe Stock

4.2. Intelligence artificielle générative

L'intelligence artificielle peut aussi être utilisée pour faire de la conception de matériaux dite « inversée », à savoir d'inverser le sens des relations de la ***Figure 17***. Il faudrait être capable de déterminer quel est le procédé à mettre en œuvre pour obtenir une propriété ? Ceci pourrait être rendu possible grâce à l'intelligence artificielle générative. Ce type d'approche vise à apprendre la structure des données d'entrée : il s'agit de trouver un espace de dimension la plus réduite possible qui permette néanmoins la description la plus complète des données. Tous les points de cet **espace latent** permettent de générer

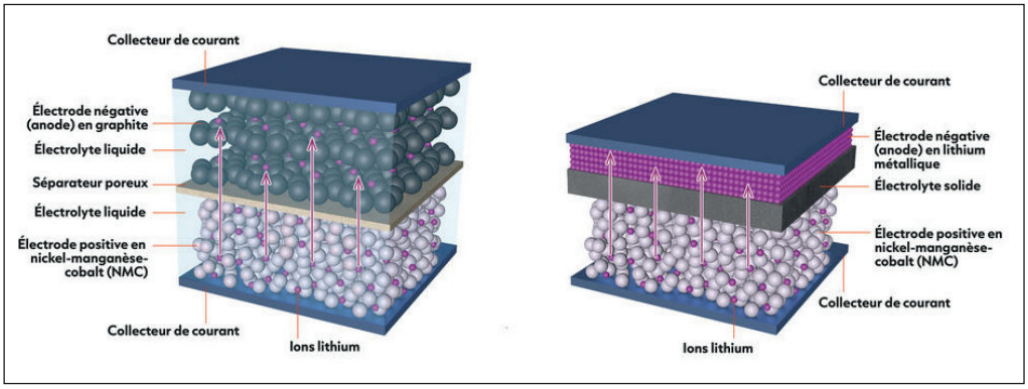


Figure 17

Le remplacement de l'électrolyte liquide par un électrolyte solide.

un nouveau contenu similaire aux données d'apprentissage mais avec un certain degré de nouveauté (plutôt que de simplement classer ou prédire les données).

Cette approche a connu un essor considérable récemment dans de nombreux autres domaines. Le plus notoire est évidemment ChatGPT. Mais, l'IA générative permet aussi de produire des visages de personnes qui n'existent pas. Au départ d'un grand nombre d'images de visages, la structure de celles-ci est encodée dans l'espace latent. Dans l'exemple de la ***Figure 18***, il est représenté en deux dimensions pour rendre les choses plus compréhensibles (mais en pratique le nombre de dimensions est plus élevé). En considérant les images représentant des femmes et des hommes, il est possible d'identifier un axe « féminité-masculinité ». En se déplaçant sur cet axe, au départ d'une image de visage (existant ou non), il est possible de générer de nouvelles images de visage dont la masculinité est augmentée. En pratique (vu que le nombre de dimensions est plus élevé), on peut aussi changer divers autres attributs tels que la longueur des cheveux, ou l'âge de la personne... Il existe divers sites web qui permettent maintenant de faire ce genre de manipulation.

Des recherches actuelles visent précisément à faire la même chose en science des matériaux. Il faut trouver l'espace latent pour les matériaux et identifier les dimensions, les directions qui permettront de changer les propriétés des matériaux de façon générative. Par exemple, le matériau devrait être « plus dur », « plus bleu », « plus léger »...

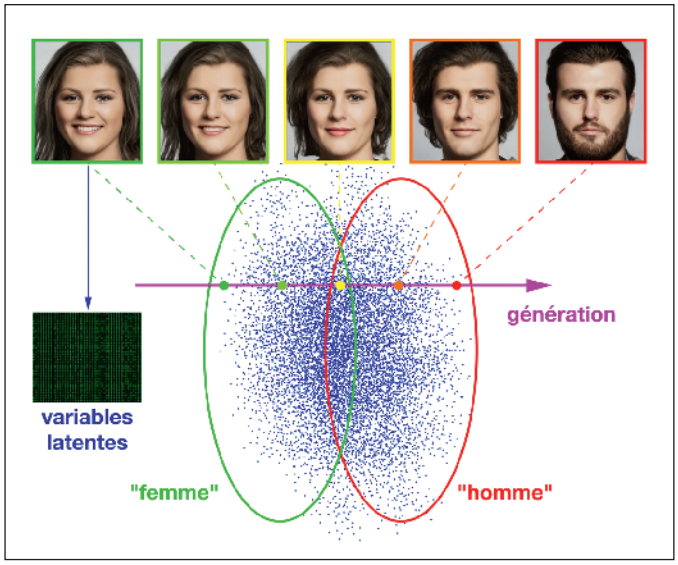


Figure 18

« Cette personne n'existe pas » – Génération de visages qui n'existent pas.

Conclusion

Les méthodes et concepts ont évolué tout au long de l'histoire de la science des matériaux, entraînant ainsi des changements de paradigme. Au départ de la **science empirique**, des lois physiques et chimiques ont peu à peu été mises au point pour expliquer les phénomènes qui se déroulent à différentes échelles de longueur et de temps. Ce fut l'avènement de la **science théorique**. Le développement des ordinateurs a permis de numériser ces équations et donc de faire des expériences virtuelles, donnant ainsi naissance à la **science computationnelle**. Avec l'augmentation de la puissance des ordinateurs, ce type de simulations a permis de mettre en œuvre le criblage à haut débit des matériaux et de produire ainsi de nombreuses données. Leur combinaison à l'intelligence artificielle ouvre la porte à un tout nouveau paradigme : la **science des données**. Cet article a montré comment l'informatique des matériaux (simulation et intelligence artificielle) a permis (et permettra encore plus dans le futur) la découverte de nouveaux composés avec des propriétés spécifiques. Il convient cependant de souligner que les autres paradigmes (***Figure 14***) restent et resteront toujours d'une grande nécessité.

Intelligence artificielle et nouvelles approches méthodologiques pour la maîtrise des risques industriels

Guillaume FAYET

Guillaume Fayet, docteur de l'Université Paris VI, Responsable d'études et de recherche à l'Ineris.

Guillaume Fayet a soutenu à l'Université Paris VI en 2010, une thèse sur « Le développement de méthodes prédictives des propriétés d'explosivité de substances chimiques » financée par l'Ineris. Il intègre alors l'Institut en tant que Responsable d'études et de recherche sur la sécurité des substances et procédés. Il pilote actuellement un Axe de Recherche sur la « Sécurité des substances, des matériaux énergétiques et des réactions chimiques » à la Direction Incendie Dispersion Explosion de l'Ineris.
Ses recherches concernent en particulier le développement et l'utilisation de méthodes prédictives, par chimie computationnelle, pour étudier et gérer les risques associés aux substances et procédés industriels. Ces travaux portent notamment sur la prédiction des dangers physiques (inflammabilité, explosivité) des substances pures et des mélanges, ainsi que sur l'étude théorique des réactions chimiques dangereuses (telles que les incompatibilités chimiques).

Introduction

Les activités et produits industriels sont sujets à des risques qu'il convient de maîtriser afin d'éliminer ou au moins limiter leur impact sur l'Homme et l'environnement. L'évaluation des risques industriels repose sur des approches expérimentales et de modélisation qui visent entre autres à caractériser les dangers (éco)-toxicologiques et physiques des substances chimiques, évaluer les risques associés aux réactions chimiques dangereuses ou encore estimer les conséquences des phénomènes dangereux associés.

L'expert public pour la maîtrise des risques industriels et environnementaux

L'Ineris est un institut public sous la tutelle du ministère en charge de l'environnement. Sa mission est de contribuer à la prévention des risques que peuvent induire les activités économiques pour l'Homme, l'environnement et les biens. L'Institut intervient sur un champ d'activité large (***Figure 1***) incluant les risques d'incendie et d'explosion à l'origine des accidents majeurs, la toxicologie et l'écotoxicologie des substances, l'impact des rejets industriels sur les milieux, la qualité de l'air ou encore la sécurité des sols et des cavités souterraines.

Naturellement, les thèmes des recherches de l'Ineris s'intéressent aux enjeux sociétaux actuels et à la sécurité des nouvelles technologies, que ce soit dans un contexte de transition vers de nouvelles énergies, de l'économie circulaire ou en ce qui concerne l'exploitation post-industrielle des mines et des énergies fossiles.

Les activités de l'Ineris sont de trois ordres. En premier lieu, l'Ineris apporte un appui technique aux pouvoirs publics en développant et validant des outils et des méthodes, et en leur apportant son expertise technique sur les risques industriels dans la mise en place des politiques publiques ou via des expertises réglementaires. L'Ineris accompagne également les industries dans l'évaluation des risques auxquels elles sont confrontées et vers des solutions de maîtrise des risques, ce qui permet ainsi à l'Institut d'être en contact direct avec le monde industriel. Enfin, cette expertise s'appuie sur

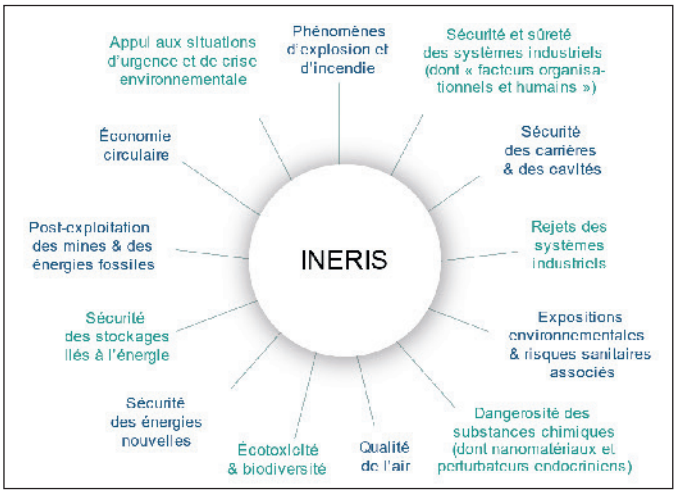


Figure 1

Champs d'activité de l'Ineris[1].

1. https://www.ineris.fr/fr/ineris/institut-bref/ineris-expert-public-maitrise-risques-technologiques

des activités de recherche de pointe, sur des programmes de recherche propres, via des projets de recherche collaboratifs nationaux et européens ou dans le cadre de projets de recherche partenariale.

Une expertise basée sur l'approche expérimentale, la modélisation et la connaissance du monde industriel

L'Ineris met en œuvre des moyens expérimentaux et de modélisation de pointe. L'Institut dispose en effet d'un grand nombre d'installations expérimentales (***Figure 2***) depuis l'échelle laboratoire, pour étudier par exemple les dangers des nanomatériaux ou pour caractériser les paramètres clés des réactions chimiques dangereuses, jusqu'à l'échelle réelle. L'Ineris dispose notamment pour cela d'une plateforme pyrotechnique pour réaliser des essais sur des matières explosives ou des réactions chimiques dangereuses à plus grande échelle, d'une plateforme incendie pour tester le comportement de produits pris dans des incendies ou encore de mésocosmes, des rivières artificielles représentant les écosystèmes naturels, dans lesquels sont étudiés les impacts que peuvent avoir des polluants sur les milieux naturels.

Les laboratoires de l'Ineris s'adaptent en permanence pour répondre aux enjeux réclamés par le développement des nouvelles technologies. Ainsi, une plateforme, dénommée STEEVE, a été mise en place pour l'évaluation de la sécurité des batteries.

L'Ineris dispose également de nombreux moyens de modélisation, là encore à différentes

Figure 2

Exemples d'installations expérimentales à l'Ineris.

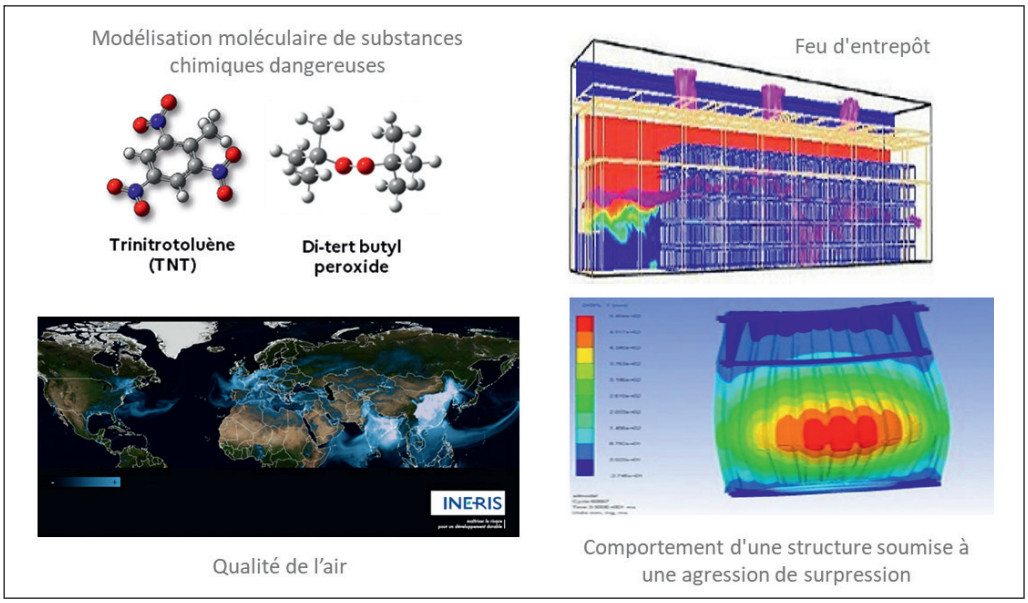


Figure 3

Exemples de modélisations réalisées à l'Ineris.

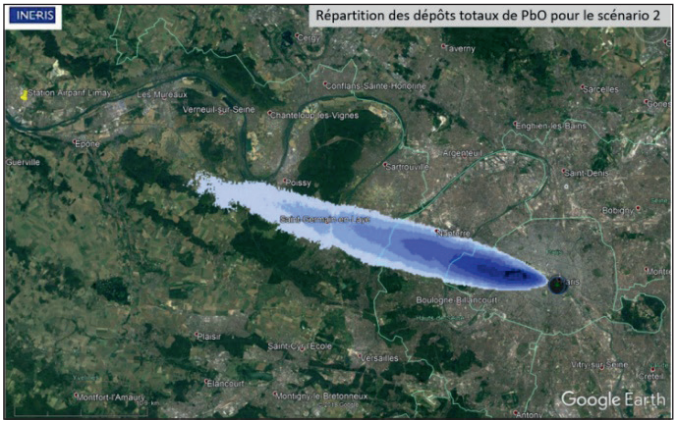


Figure 4

Modélisation de la dispersion des particules de plomb du panache de l'incendie de Notre-Dame (2019)[2].

échelles (***Figure 3***). Des approches à l'échelle moléculaire permettent de prédire les dangers des substances et de comprendre leur réactivité. À plus grande échelle, l'Ineris modélise par exemple des feux d'entrepôt, l'impact de phénomènes dangereux sur des structures et met en œuvre des modèles jusqu'à une échelle « géographique » pour l'évaluation de la qualité de l'air. L'Ineris intervient notamment dans un contexte post-accidentel pour modéliser des dispersions de fumées ou de particules, comme la dispersion des particules de plomb dans le panache de l'incendie de Notre-Dame en 2019 (***Figure 4***).

Apport des nouvelles approches méthodologiques et de l'intelligence artificielle

Si l'approche expérimentale représente une pièce fondamentale de l'expertise de l'Ineris, depuis la caractérisation

2. https://www.ineris.fr/fr/ineris/actualites/incendie-dame-rapport-ineris-est-paru

en laboratoire jusqu'aux essais à grande échelle, les progrès scientifiques et techniques ont donc élargi au cours du temps le panel d'outils disponibles par des outils de modélisations de plus en plus poussés. Aujourd'hui, de nouvelles approches méthodologiques (notamment celles basées sur l'intelligence artificielle) permettent d'accéder à des informations plus nombreuses, plus rapidement et parfois plus complètes.

Il s'agit soit de compléter les moyens existants (pour exploiter de manière plus précise, plus étendue et efficace les données produites par les essais et modélisations), soit de disposer de méthodes alternatives pour gagner du temps ou optimiser des plans d'expérience. Ces outils s'intègrent dans une stratégie de digitalisation et de capitalisation des données disponibles à la fois à l'Ineris et à l'extérieur afin que ce retour d'expérience contribue à une meilleure maîtrise des risques.

Au niveau des activités expérimentales, au-delà de la gestion et de l'exploitation des données, des métrologies innovantes permettent le traitement d'informations plus nombreuses et en temps réel pendant les essais. Ces outils ouvrent la voie à des analyses et des traitements plus automatisés en faisant gagner du temps aux opérateurs et aux experts, leur permettant de se consacrer pleinement aux enjeux de recherche qui y sont associés.

Ces données permettent également de développer de nouveaux types d'outils basés en construisant des modèles prédictifs et des méta-modèles de substitution ou compléments des modèles existants qui nécessitent dans certains cas des données et des temps de calculs importants.

Ces approches numériques nouvelles permettent d'accompagner *in fine* le développement de procédés et de substances plus sûres. Elles peuvent par exemple aider à la formulation et à la substitution en prenant en compte les risques industriels au plus tôt dans les démarches de développement des substances chimiques.

Quelques exemples d'application à l'Ineris de ces nouvelles approches méthodologiques sont ici présentés.

1 Prédire les dangers des substances

1.1. Le règlement REACH et les méthodes prédictives

Des recherches ont été initiées à l'Ineris il y a environ 15 ans pour le développement de méthodes prédictives des dangers physiques des substances, dans le contexte de la mise au point du règlement européen REACH[3].

3. Le règlement européen REACH (Enregistrement, Évaluation et Autorisation de Produits Chimiques en français), adopté en 2007, vise à mieux protéger la santé humaine et l'environnement contre les risques liés aux substances chimiques. Toute substance mise sur le marché ou transitant dans l'Union européenne à plus d'une tonne par an doit y être enregistrée. Le règlement prévoit des mesures d'interdiction ou de restriction pour les substances les plus préoccupantes.

Cette nouvelle réglementation concerne l'autorisation de mise sur le marché des substances sur le territoire européen et a impliqué une quantité de travaux très importante, pour l'enregistrement d'un grand nombre de substances dans un calendrier défini et serré.

Or, les essais de caractérisation des dangers des substances peuvent dans certains cas être très coûteux et la disponibilité des laboratoires d'essai peut poser problème. Par ailleurs, la dangerosité des propriétés et substances testées peut imposer des contraintes importantes (essais sur des explosifs ou des substances toxiques). La caractérisation de la toxicologie et de l'écotoxicologie pose également des problèmes d'éthique avec une volonté de réduire le recours aux essais sur animaux, une recommandation toujours très actuelle pouvant entraîner le besoin d'utiliser des méthodes alternatives en substitution des essais actuels, qu'ils soient par des calculs numériques comme illustré par la suite, ou par des essais *in vitro*[4].

Ainsi le règlement REACH encourage le recours au partage et à la capitalisation des données ainsi qu'à l'utilisation des méthodes alternatives, telles que les modèles QSAR[5] ou QSPR[6], qui reposent sur des relations structures chimiques-propriétés en utilisant des approches de *machine learning*[7]. Très connues pour la toxicologie et en général pour des produits purs, l'Ineris les développe et les utilise également sur les dangers physiques et dans le cas de mélanges.

L'approche QSAR/QSPR repose sur un principe de similarité selon lequel des molécules avec des structures chimiques similaires présenteront probablement des propriétés et des activités biologiques similaires. Tout l'enjeu de cette approche est donc de trouver des descripteurs pertinents pour caractériser la structure moléculaire qui soient corrélés avec la propriété ciblée (***Figure 5***).

Les données disponibles pour chercher ces corrélations sont la base du développement des modèles QSAR/QSPR. Ils sont entraînés sur un jeu de données expérimentales qui doivent être suffisamment nombreuses, fiables et obtenues dans des conditions homogènes.

Pour représenter les structures moléculaires, différents types de descripteurs sont utilisés. Certains sont simples à calculer et à utiliser comme des nombres

4. L'approche *In vitro* consiste à réaliser des tests sur des modèles cellulaires en laboratoire. Ces tests sont utilisés comme alternatives aux essais sur animaux.

5. QSAR pour *Quantitative Structure-Activity Relationships* : Relations Quantitatives Structure-Activité.

6. QSPR pour *Quantitative Structure-Proprety Relationships* : Relations Quantitatives Structure-Propriété.

7. Méthode d'apprentissage automatisée, forme d'intelligence artificielle, consistant à utiliser des approches statistiques pour entraîner un modèle à partir d'une base de données d'apprentissage.

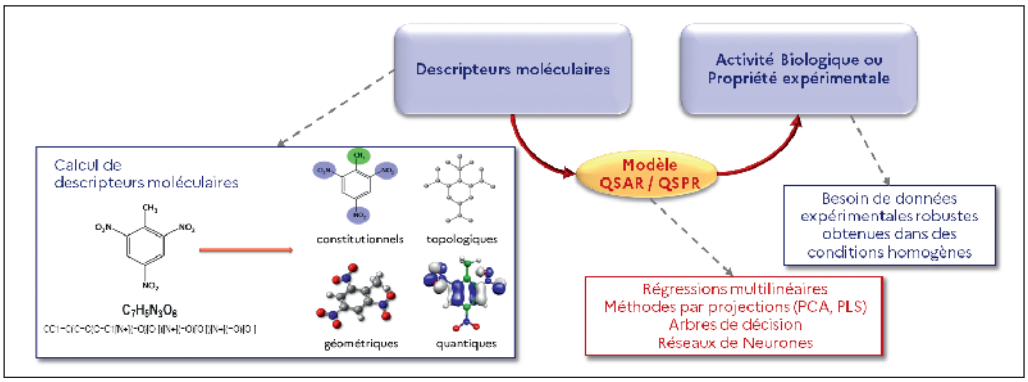


Figure 5

Approche QSAR/QSPR.

d'atomes. D'autres sont plus complexes pour caractériser des propriétés électroniques ou de réactivité (à partir de calculs de chimie quantique). Ces derniers seront plus compliqués à utiliser mais peuvent apporter une compréhension physico-chimique plus complète.

Le travail consiste ensuite à analyser les données et à établir des corrélations au sein d'un modèle mathématique reliant les descripteurs moléculaires les plus pertinents à la propriété que l'on cherche à prédire. Pour cela, différentes méthodes de *machine learning* sont utilisées, depuis les régressions linéaires simples jusqu'à des réseaux de neurones artificiels plus complexes et plus adaptés lorsque des quantités de données plus importantes sont disponibles. Les méthodes pourront être « supervisées » ou « non supervisées » pour obtenir des « données quantitatives » ou bien des « classifications » selon les besoins et les systèmes étudiés.

1.2. Cas des dangers physiques des matières auto-réactives

Un premier exemple de modèle QSPR développé à l'Ineris concerne les matières auto-réactives. Une matière auto-réactive est une substance thermiquement instable, susceptible de générer une décomposition exothermique[8], potentiellement forte et rapide. Une des particularités de ces substances est qu'elles n'ont pas forcément besoin d'un apport de l'oxygène de l'air pour se décomposer. On retrouve dans cette classe de substances dangereuses différentes familles de composés chimiques (***Figure 6***).

La caractérisation des dangers physiques de ces substances

8. Exothermique : qui dégage de la chaleur.

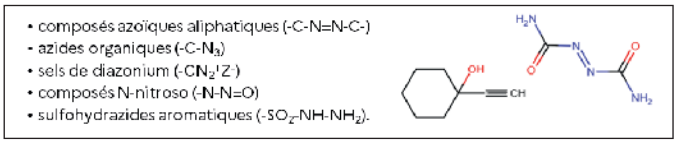


Figure 6

Groupes chimiques typiques rencontrés dans des matières auto-réactives.

est importante, car elles sont susceptibles de se décomposer de manière explosive. Les effets sont alors très importants. Elles brûlent rapidement. Prises dans un feu, elles peuvent amplifier l'incendie. Elles peuvent aussi présenter une sensibilité aux chocs ou à la friction, ce qui peut générer des accidents. Dans certains cas, elles sont susceptibles de réagir dangereusement au contact d'autres substances, on parle d'incompatibilité chimique.

Du fait de ces propriétés, elles ont un classement particulier dans le cadre des réglementations, par exemple pour le transport de marchandises dangereuses ou dans les réglementations liées à l'étiquetage et à l'emballage des produits. Des contraintes particulières et un étiquetage particulier leur sont alors appliqués.

9. G. Fayet *et al.*, "First QSPR models to predict the thermal stability of potential self-reactive substances", *Process Safety and Environmental Protection* 163 (2022) 191-199.

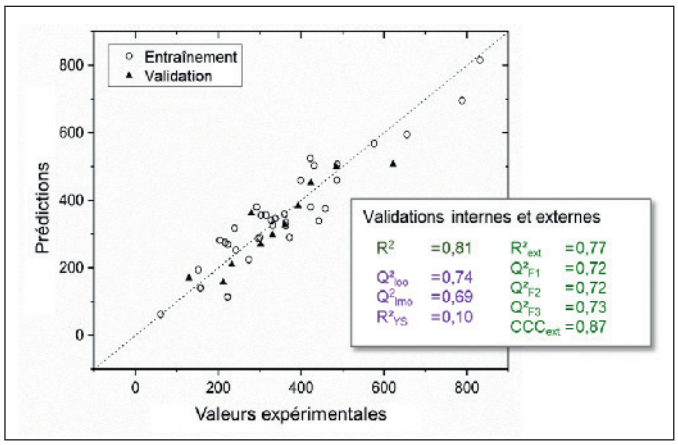


Figure 7

Modèle développé pour la chaleur de décomposition des matières auto-réactives[9].

Jusqu'à présent, ces substances sont caractérisées exclusivement sur la base d'essais expérimentaux, car aucun modèle prédictif n'existait. L'Ineris a donc engagé des travaux sur ce sujet dans le cadre d'un projet européen (en collaboration notamment avec le BAM[10] en Allemagne) pour combler ce manque de modèles et pouvoir anticiper le caractère auto-réactif des substances chimiques.

Dans le cadre de ce travail, l'Ineris a développé un modèle pour évaluer la chaleur de décomposition de substances susceptibles de présenter un caractère auto-réactif[11] (***Figure 7***). Cette propriété est utilisée comme critère de présélection pour le classement en tant que matière auto-réactive. Il permet de décider s'il est nécessaire ou non d'engager toute une campagne d'essai sur le caractère auto-réactif de la substance.

Il s'agit du premier modèle développé pour cette famille de substances. Il a été développé à partir d'une base de données générée spécifiquement pour cette étude à partir d'essais sur une série d'échantillons fournis par la société Bayer[12]. Ils ont tous été réalisés par le BAM selon des protocoles identiques pour s'assurer de conditions expérimentales uniformes.

10. BAM pour *Bundesanstalt für Materialforschung und prüfung* : Institut fédéral pour la recherche et les essais des matériaux (Allemagne).
11. Chaleur de décomposition : la température à laquelle la substance se décompose chimiquement.
12. Société pharmaceutique et agrochimique allemande.

Dans ce type d'étude, l'enjeu clé du travail est de sélectionner les paramètres à retenir pour le modèle final. Ceci a été réalisé à l'aide d'un « algorithme génétique » qui a permis de choisir les descripteurs les plus pertinents.

Une contrainte particulière pour l'Ineris réside dans la volonté que les modèles développés puissent être utilisés dans un cadre réglementaire. Ils doivent donc répondre à des principes de validation mis en place par l'OCDE[13]. Sont en particulier demandées : une définition claire de la propriété visée, une définition du domaine d'applicabilité du modèle, des validations statistiques approfondies et si possible une interprétabilité du modèle.

Néanmoins, le fait qu'un modèle satisfasse à ces principes ne suffit pas à valider toute prédiction issue de ce modèle. L'OCDE propose également des recommandations pour la validation des prédictions issues de ces modèles afin de vérifier qu'elles soient suffisamment fiables et permettent une prise de décision dans le contexte réglementaire visé.

13. OCDE : Organisation de coopération et de développement économiques. L'Ineris participe aux groupes de travail de l'OCDE définissant les principes de validation des modèles et prédictions QSAR pour un usage réglementaire.

1.3. Extension de la modélisation au cas des mélanges – Exemple du point d'éclair

En pratique, il est très courant qu'une substance chimique ne soit pas un composé pur mais un mélange dont la composition est déterminée pour obtenir des fonctions particulières. Cette situation nécessite de modifier la méthodologie en développant des approches permettant de prendre en compte les spécificités des mélanges. En effet, le principe d'un modèle QSPR est de corréler la variation de structure de la molécule à ses propriétés. Pour un mélange, plusieurs composés sont présents et différents facteurs influençant la propriété sont à prendre en compte comme leurs concentrations respectives et les interactions qu'ils peuvent avoir entre eux.

Une telle adaptation a été réalisée pour la prédiction du point d'éclair, la propriété qui sert de base au classement des liquides inflammables. Il s'agit de la température à partir de laquelle le liquide s'enflamme à l'approche d'une flamme. Plus la température du point éclair est basse, plus le liquide est inflammable (*Figure 8*).

L'adaptation de l'approche a consisté à définir des

Figure 8

Point d'éclair et liquides inflammables.

descripteurs de mélanges, calculés à partir des descripteurs moléculaires et de la fraction molaire[14] des différents constituants du mélange sur lesquels la méthodologie QSPR est ensuite appliquée, de la même manière que pour les substances pures.

Ce travail a été un succès puisque, comme le montre la ***Figure 9***, le modèle obtenu fournit de bonnes prédictions, assez proche des performances attendues pour des modèles de produits purs. Même si d'autres approches (thermodynamiques) peuvent être utilisées pour estimer le point d'éclair de liquides inflammables, cette modélisation fait maintenant partie de la boîte à outils à disposition de l'Ineris.

14. Grandeur quantifiant la proportion de quantité de matière d'un constituant dans un mélange. La fraction molaire est comprise entre 0 et 1 (1 pour un composé pur, 0 pour un composé absent).

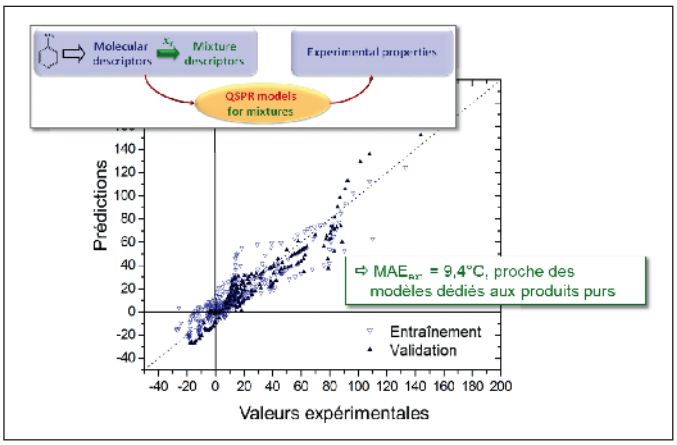


Figure 9

Modèle développé pour le point d'éclair de mélanges liquides inflammables[15].

1.4. Utilisation des modèles QSPR

1.4.1. Sécurité des substances et aide à la formulation/substitution

Les modèles QSPR sont intéressants d'un point de vue réglementaire pour combler des données manquantes. Au-delà des dangers physiques, l'Ineris s'intéresse de la même manière aux modèles QSAR pour la toxicologie et l'écotoxicologie. De telles prédictions permettent également d'accéder à des données utiles dans des études de sécurité des procédés. Elles permettent en particulier de faire varier des paramètres comme la concentration de tel ou tel composé et d'évaluer ainsi l'influence de la composition d'un mélange sur la sécurité d'une installation industrielle.

La modélisation QSPR est très précieuse dans le cadre d'études R&D[16] pour le développement de substances nouvelles ou alternatives à des substances existantes. Elle permet de prendre en compte leurs dangers le plus en amont possible dans les phases de recherche. Cela peut éviter d'importants efforts de recherche, synthèses et caractérisations sur des molécules à priori intéressantes pour leurs propriétés fonctionnelles, mais qui présentent finalement

15. G. Fayet & P. Rotureau, "New QSPR Models to Predict the Flammability of Binary Liquid Mixtures", *Molecular Informatics* 38 (2019) 1800122.
16. R&D : Recherche et Développement.

des dangers induisant des contraintes de sécurité trop importantes.

Les modèles QSPR sont donc utilisés non seulement pour prédire les propriétés de substances particulières, mais aussi à des fins de criblage au sein de listes de molécules. À partir d'une base de données importante de substances, des modèles QSPR peuvent être utilisés dans une analyse multicritère incluant les propriétés dangereuses et la toxicité (***Figure 10***). À l'aide de ces prédictions, il est ainsi possible de sélectionner les candidats les plus pertinents pour les étapes ultérieures de synthèse et de caractérisations expérimentales.

Il est également possible d'aller plus loin en proposant de nouvelles molécules par *Design in Silico*[17]. Le principe est alors de se baser sur les modèles et les paramètres qu'ils génèrent pour construire des bases de données de nouvelles molécules (virtuelles) sur lesquelles est réalisé le criblage (***Figure 11***).

Une telle approche permet la prise en compte précoce de la sécurité et sera particulièrement intéressante dans la recherche de solutions durables et plus sûres dans des démarches de type « *safe and sustainable by design*[18] ».

1.4.2. Application au cas des tensioactifs dérivés de sucre

Des travaux visant ce type d'application ont été menés dans le cadre de projets financés par l'ITE[19] Pivert, à savoir Amphipred et Amphifoam dont l'objectif était de développer des tensioactifs biosourcés, afin de les substituer aux produits pétro-sourcés actuellement sur le marché.

Ce travail a été réalisé en collaboration avec différents

17. *Design in Silico* : Conception par ordinateur.

18. *Safe and Sustainable by Design* : sûr et durable par conception.

19. ITE : Institut de Transition Energétique.

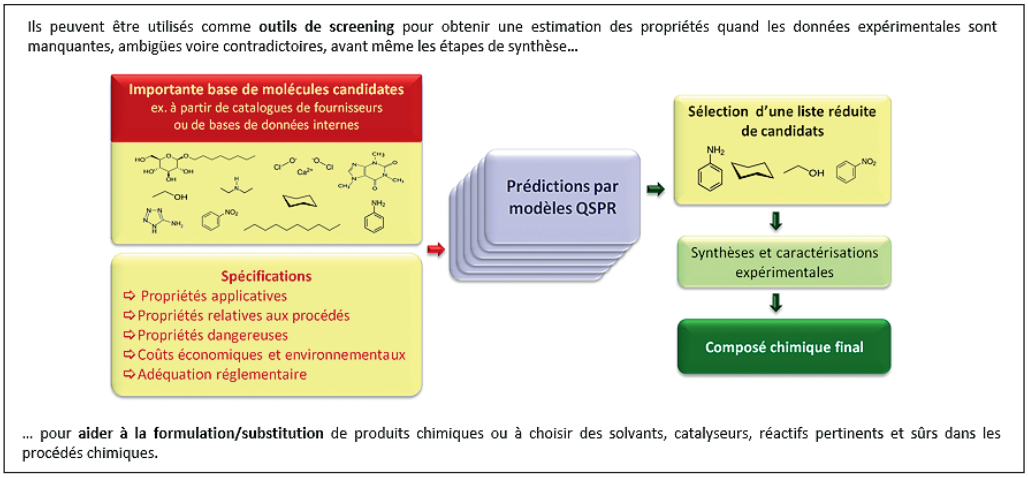


Figure 10

Utilisation des modèles QSPR comme outils de screening.

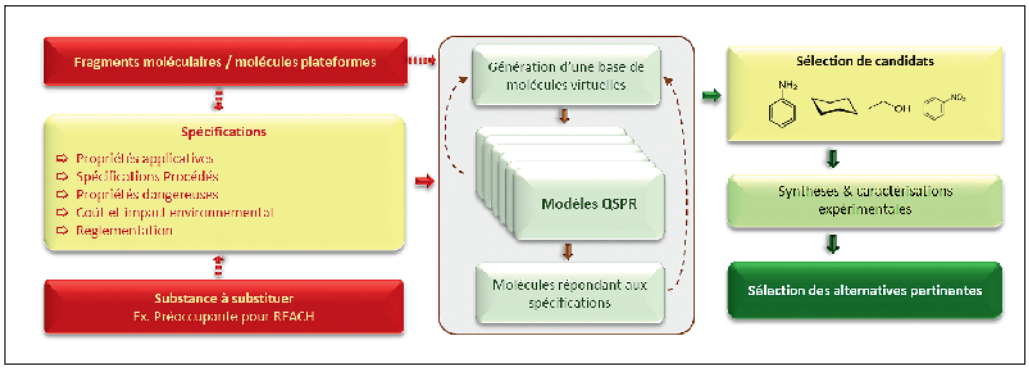


Figure 11

Modèle du Design in silico.

partenaires dont l'Université de Technologie de Compiègne et l'Université Picardie Jules Vernes à Amiens. Une des caractéristiques de ces substances est qu'elles sont constituées d'une tête polaire[20] (hydrophile) et d'une chaîne alkyle[21] (hydrophobe) qui leur donnent des propriétés spécifiques en solution (formation de mousses). Cette structure moléculaire particulière a encouragé l'utilisation de descripteurs de fragments pour le développement des modèles QSPR permettant de prédire les propriétés de ces substances mais aussi pour créer de nouveaux tensioactifs. Les propriétés visées étaient la concentration micellaire critique[22] (la CMC), la tension de surface[23] à la CMC, l'efficience (pC20) et le point de Krafft (TK)[24].

De nouveaux modèles spécifiques aux tensioactifs dérivés de sucre ont tout d'abord été développés (***Figure 12***), car les modèles existants pour les propriétés d'intérêt étaient dédiés aux molécules pétrosourcées et ne montraient pas les mêmes niveaux de performance pour les tensioactifs dérivés de sucre.

Une démarche de *Design in Silico* a ensuite été développée pour obtenir de bons candidats pour de futures synthèses et caractérisations (***Figure 13***).

La première étape consiste à définir un cahier des charges des propriétés désirées pour ces substances. Étant donné qu'il s'agissait d'une démarche de substitution, il a

20. Polarité : non homogénéité de la répartition des charges au sein de tout ou partie d'une molécule.
21. Succession d'atomes de carbone, chacun lié à deux hydrogènes.
22. Concentration en surfactants au-delà de laquelle il leur est favorable de se regrouper sous forme de structures supramoléculaires nommées micelles pour minimiser leur surface de contact avec l'eau.
23. Grandeur caractérisant la résistance d'un fluide à augmenter sa surface de contact avec un autre fluide.
24. Température minimale pour qu'un tensioactif forme des micelles. Cette propriété permet de vérifier si un tensioactif est utilisable à température ambiante.

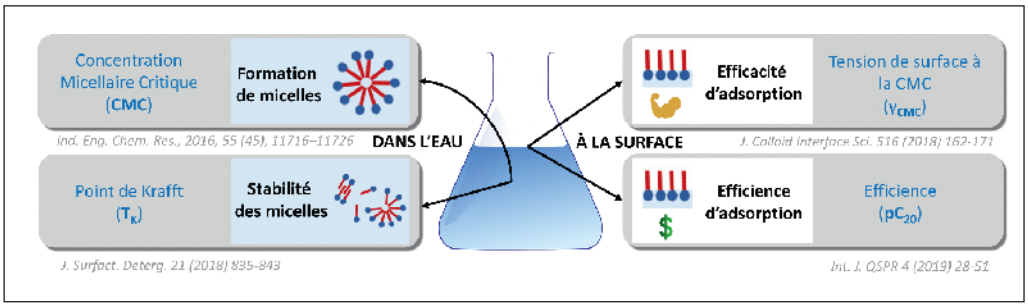


Figure 12

Modèles développés pour les tensioactifs dérivés de sucre.

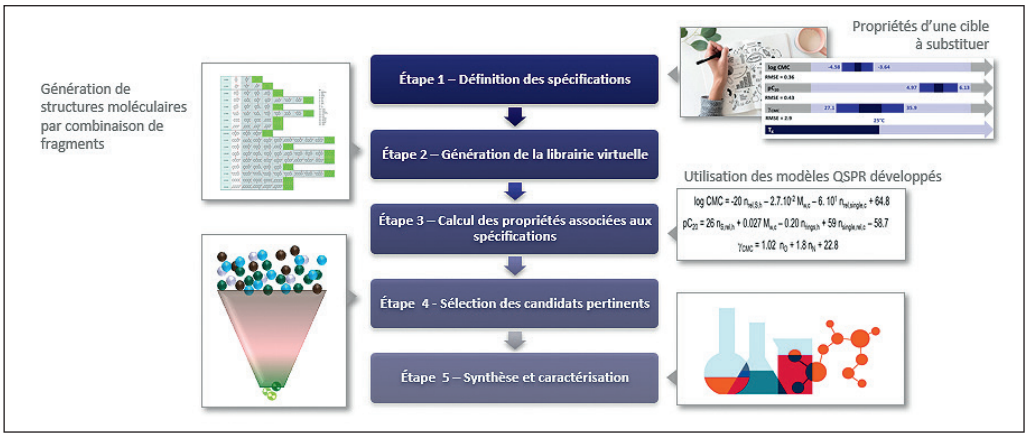


Figure 13

Méthode de Design in Silico.

été basé sur une sélection de cibles à substituer et sur leurs propriétés, l'objectif étant de rechercher des molécules qui pouvaient avoir des propriétés proches.

Les fragments moléculaires modélisés pour le développement des modèles ont été exploités pour générer une base de données de molécules virtuelles par combinaison des fragments hydrophiles et hydrophobes. De cette manière, plus de 2 500 molécules ont déjà pu être générées à partir de fragments représentés dans les molécules étudiées pour le développement des modèles QSPR et cette base pourrait être facilement étendue en ajoutant plus de fragments.

Un criblage au sein de cette base de tensioactifs virtuels a ensuite été réalisé pour identifier les candidats les plus pertinents, qui pourraient être considérés dans une étape suivante de synthèse et de caractérisation.

Le nombre de candidats proposés était bien supérieur à

ce qui a pu être trouvé dans la base de données AmphInnov qui correspondait au recensement de surfactants dérivés de sucre le plus exhaustif connu (*Figure 14*). Les candidats recensés répondaient à chaque critère du cahier des charges. Par ailleurs, les prédictions issues des modèles étaient en bon accord avec les valeurs expérimentales disponibles, validant la pertinence de l'approche proposée.

2 Faciliter et améliorer l'identification de contaminants dans l'environnement

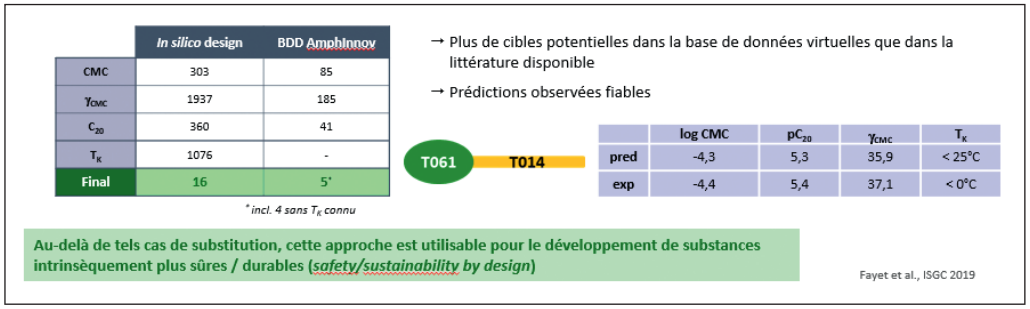
2.1. Identifier les substances contaminant l'environnement

La présence de contaminants dans l'environnement, notamment dans les milieux aquatiques, peut avoir un impact néfaste et interroge l'Ineris. Pour identifier les substances potentiellement polluantes, des échantillons sont prélevés puis analysés au laboratoire à l'aide d'instruments de pointe. Malgré les technologies disponibles très avancées, l'identification de contaminants peut s'avérer longue et difficile.

Une étude a donc été lancée avec pour objectif d'aider à l'identification des contaminants dans le cadre d'analyses LC-HRMS (chromatographie liquide[25] couplée à la spectrométrie de masse à haute résolution[26]), en utilisant les nouvelles techniques issues de l'intelligence artificielle. La *Figure 15* présente la procédure expérimentale de base, qui conduit à l'identification d'un très grand nombre de molécules, plusieurs milliers de molécules, via environ 20 000 spectres générés pour un échantillon.

L'immense quantité de données générées par chaque analyse est traitée par des logiciels constructeurs ou libres afin d'aider l'opérateur à identifier des substances présentes dans les échantillons. Concrètement, les données

25. Méthode de séparation de composés basée sur leur temps d'écoulement dans une colonne.
26. Méthode d'analyse d'un mélange de composés basée sur leurs différences de masse.



	In silico design	BDD AmphInnov
CMC	303	85
γ_{CMC}	1937	185
C_{20}	360	41
T_K	1076	-
Final	16	5*

* *incl. 4 sans T_K connu*

→ Plus de cibles potentielles dans la base de données virtuelles que dans la littérature disponible

→ Prédictions observées fiables

T061 T014

	log CMC	pC_{20}	γ_{CMC}	T_K
pred	-4,3	5,3	35,9	< 25°C
exp	-4,4	5,4	37,1	< 0°C

Au-delà de tels cas de substitution, cette approche est utilisable pour le développement de substances intrinsèquement plus sûres / durables (*safety/sustainability by design*)

Fayet et al., ISGC 2019

Figure 14

Comparaison du nombre de tensioactifs proposés par l'approche de Design in Silico avec ceux trouvés dans la littérature scientifique disponible (Base de données AmphInnov).

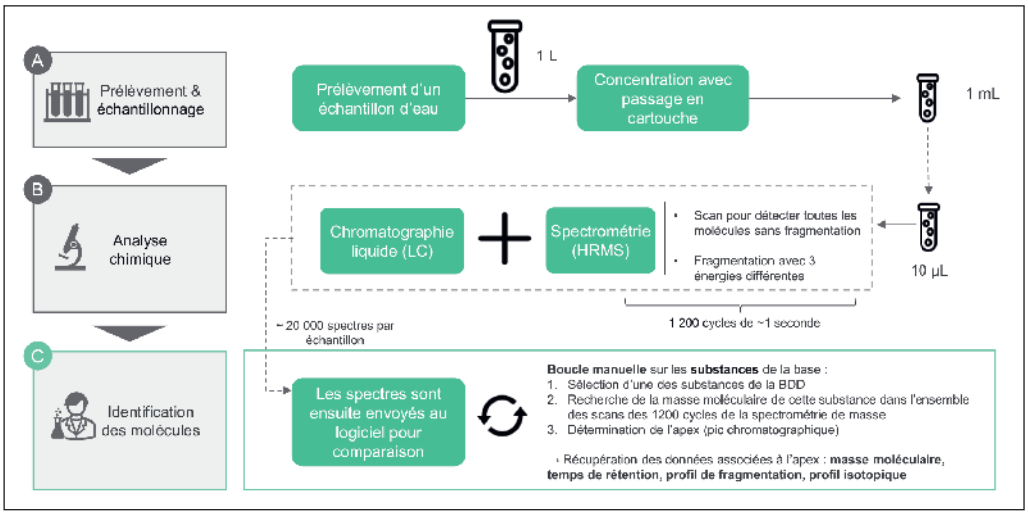


Figure 15

Schéma du processus analytique pour l'identification de contaminants par LC-HRMS.

générées sont comparées à des bases de données, et le « matching » (la correspondance) entre les données générées et celles des bases de données permet d'identifier des substances. L'opérateur doit tout de même analyser les données générées et traitées par le logiciel pour valider ou non les identifications. Cette procédure est longue, fastidieuse et peut s'avérer peu reproductible d'un opérateur à un autre.

D'autre part, les bases de données utilisées pour identifier des correspondances, fournies par les constructeurs ou créées par les laboratoires utilisateurs, sont peu fournies ; ceci limite les possibilités d'identification de substances dans l'échantillon. Les méthodes d'automatisation permises par l'IA couplées à de plus larges bases de données disponibles en libre accès éviteraient ces inconvénients et limitations.

Un autre intérêt serait le gain de temps qu'elles offriraient : l'opérateur doit actuellement vérifier manuellement chaque concordance, substance par substance pour chaque échantillon. L'objectif de ce travail était également de chercher à automatiser et accélérer ce travail de vérification.

2.2. Intégrer l'intelligence artificielle dans le processus métier

L'automatisation des identifications grâce à l'IA a été réalisée dans le cadre d'un « projet empreinte environnementale » financé par un Appel à Manifestation d'Intérêt (AMI IA) en 2020, et une aide dans le cadre des projets de France Relance[27] 2021. L'enjeu était de créer et d'utiliser des

27. Aide publique au financement de projets à visée écologique ou sociale.

features[28] particuliers à partir des spectres de masse pour construire un modèle capable de reproduire le travail de vérification de l'opérateur, a minima pour une partie des identifications, afin de lui permettre de se concentrer sur les cas les plus litigieux. L'utilisation de l'IA de type *machine learning* a permis d'élaborer un modèle capable d'identifier de façon automatique en moyenne la moitié des substances dans un échantillon.

De plus, l'inclusion des connaissances propres à l'Ineris ainsi que des données librement disponibles dans des bases de données européennes (MassBank) a permis de multiplier par 4 la taille de la base de données originale (la base constructeur contenait 1 200 molécules).

2.3. Plus-values observées et perspectives

Les gains obtenus avec l'approche IA sont importants. L'outil constructeur (***Figure 16***) est basé sur un matching automatique focalisé sur la masse moléculaire et la masse de certains fragments, l'opérateur doit ensuite vérifier la correspondance des intensités de fragment, du temps de rétention et du profil isotopique pour valider ou non l'identification. L'outil développé avec l'intelligence artificielle peut vérifier, plus rapidement, les correspondances entre échantillon et base de données au regard de ces cinq paramètres. Le taux d'analyse est accru et la systématisation de l'analyse faite par l'outil informatique permet également d'augmenter la reproductibilité et la fiabilité des résultats obtenus. Par ailleurs, l'augmentation de la taille de la base de données a permis de multiplier le nombre de substances identifiées par 2.

En pratique, pour l'analyse d'environ 50 échantillons (***Figure 17***), le temps de travail qui était réalisé jusqu'à présent avec l'outil constructeur était de 39 jours pour 500 identifications. Avec l'outil IA, le temps de travail de l'opérateur est divisé par plus de deux et

28. *Features* : caractéristiques.

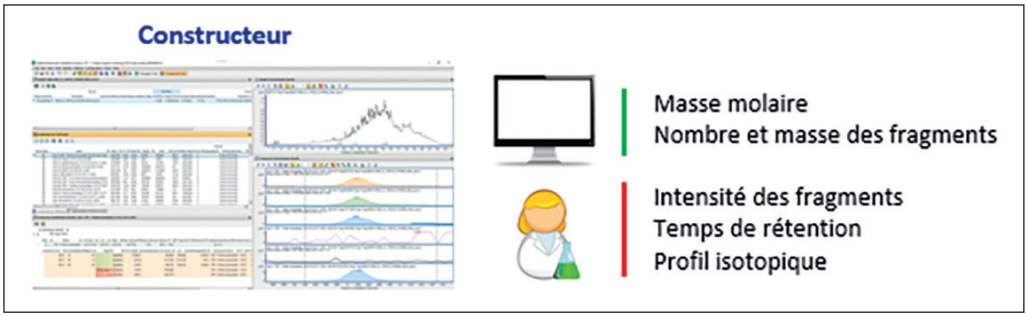


Figure 16

Données traitées par le logiciel constructeur (en vert) et par l'opérateur (en rouge) pour conduire à une identification.

un nombre légèrement plus élevé d'identifications est obtenu, en utilisant la même base de données. L'utilisation de la base de données étendue permet d'identifier deux fois plus de substances en 50 jours environ.

L'outil IA est plus fiable parce que plus reproductible, mais des perfectionnements supplémentaires pour l'automatisation seront encore possibles dans l'avenir. Cependant, il faut une nouvelle fois rappeler que la présence d'un opérateur est nécessaire pour les cas les plus litigieux.

Sur la thématique, d'autres travaux sont encore à conduire avec l'aide de l'IA, comme la détermination des sources de polluants (par exemple, relier les rejets aux sources de formation de particules).

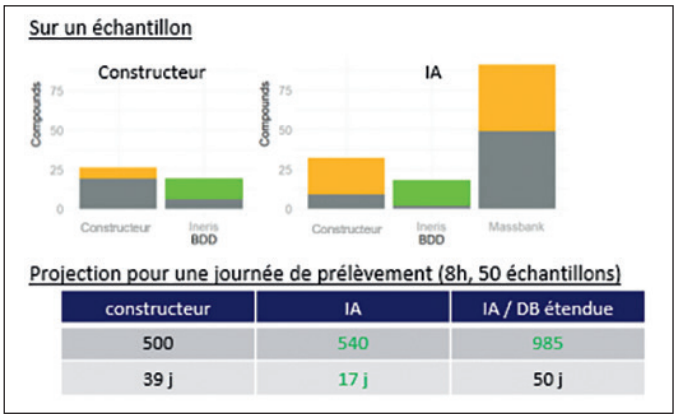


Figure 17

Comparaisons des performances entre un outil constructeur et l'outil par l'intelligence artificielle.

3 D'autres applications de l'intelligence artificielle

Deux autres applications moins en lien avec les substances chimiques et leurs dangers sont citées pour illustrer d'autres utilisations de l'IA à l'Ineris.

3.1. Surveillance microsismique

La première concerne la surveillance microsismique, classiquement utilisée pour surveiller l'exploitation industrielle du sous-sol : mines, réservoirs, stockages géologiques, géothermie profonde entre autres. Dans certains cas, les anciennes exploitations souterraines font l'objet d'une activité microsismique résiduelle de long terme durant leur phase de post-exploitation, le temps nécessaire au sous-sol de retrouver un nouvel équilibre hydromécanique.

La surveillance microsismique est habituellement réalisée via l'enregistrement et l'analyse de signaux provenant de stations de surface ou en forages, réparties autour de la zone d'intérêt à surveiller. Dans le cas de la surveillance d'anciennes mines, les signaux enregistrés peuvent bien évidemment être générés par l'endommagement et/ou la rupture d'une partie des ouvrages souterrains, mais également avoir pour origine des activités anthropiques de surface (travaux, bruits parasites divers, tirs) ou une origine naturelle (séismes naturels, orages et surtensions) (***Figure 21***). Les signaux, issus de ces différents phénomènes, présentent des similarités et sont usuellement analysés dans un délai court par des opérateurs spécialistes, pour

en déterminer l'origine (phase de qualification), les traiter (phase de traitement du signal et de calculs à la source) et ainsi aider l'expert à définir s'ils correspondent au déclenchement d'un évènement redouté. L'automatisation de la phase de qualification de données en quasi-temps réel et de manière fiable est donc un enjeu important.

Les méthodes d'apprentissage (*machine learning*) associées à l'intelligence artificielle sont bien adaptées pour réaliser de la classification automatique d'informations. Plusieurs types de réseaux neuronaux (CNN, Inception et LSTM) ont pu être testés pour évaluer leur capacité à prédire, sans intervention humaine, 24 h/24, l'origine des signaux microsismiques issus de travaux souterrains. Le cas d'application de l'ancien bassin minier de Gardanne, objet d'une microsismicité résiduelle de long terme, a été utilisé car tout à fait représentatif de la surveillance d'opérations industrielles sismogéniques en milieu souterrain.

La ***Figure 18*** montre que jusqu'à plus de 98 % des signaux sont correctement classés pour ce site, ouvrant la perspective vers d'autres applications. Mais l'outil ne remplacera pas l'expert qui intervient pour confirmer les prédictions de l'IA et évaluer les conséquences possibles des signaux précurseurs d'endommagement potentiel en surface.

3.2. Méta-modèles – Explosions en milieu confiné

L'intelligence artificielle peut également être utilisée pour développer des méta-modèles, c'est-à-dire des modèles prédictifs, basés sur des approches d'apprentissage automatiques (*machine learning*), en substitution des modèles phénoménologiques existants (parfois trop lourds ou nécessitant la connaissance

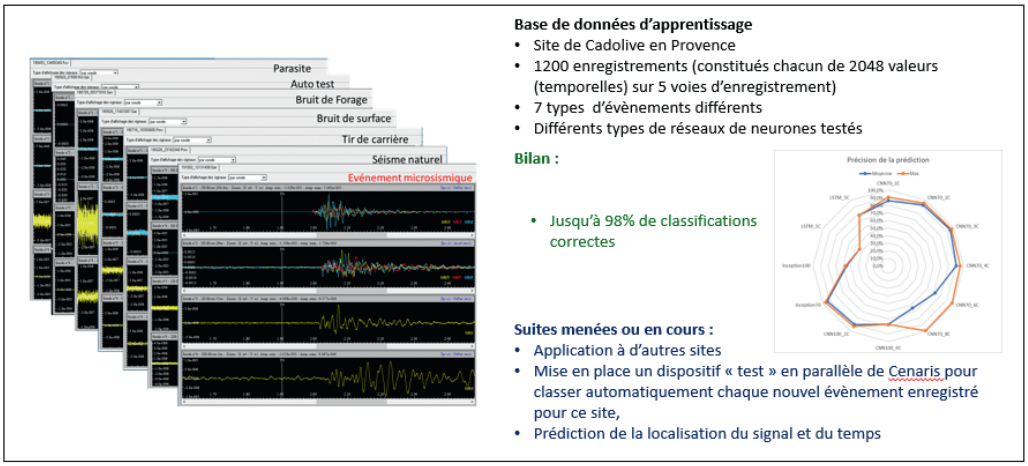


Figure 18

Résultats obtenus pour la surveillance microsismique.

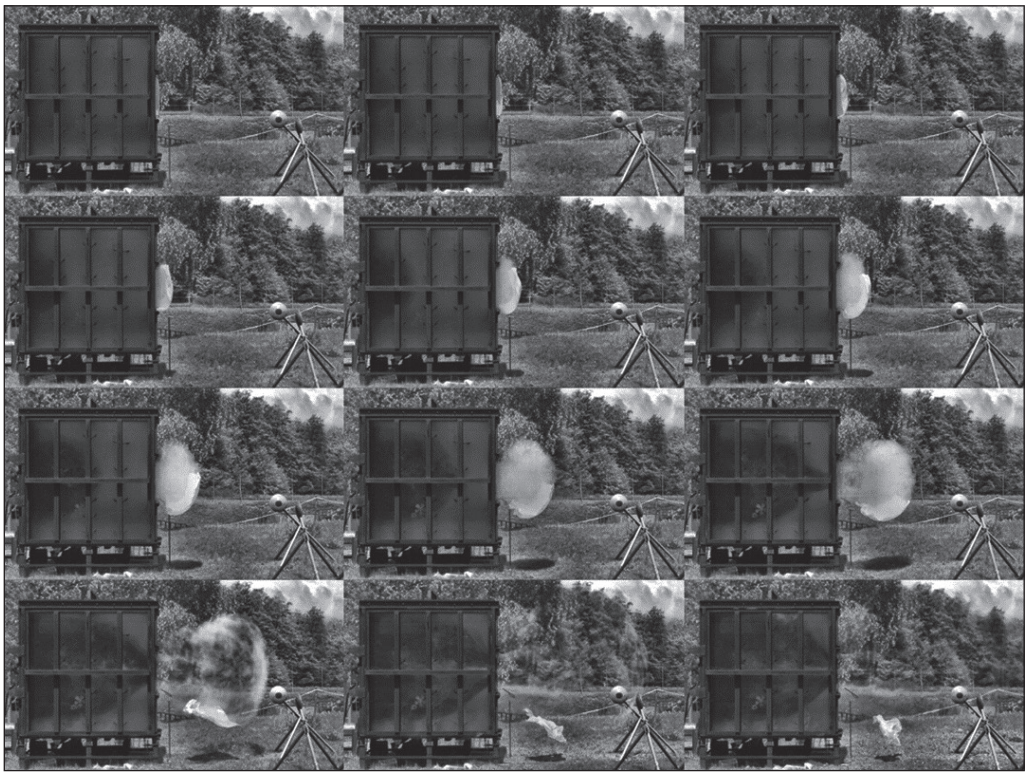


Figure 19

Exemple d'essais d'explosion de mélange hydrogène-air.

de nombreux paramètres). Ce principe est appliqué ici à l'étude des explosions en milieu confiné.

La ***Figure 19*** présente des images acquises avec une caméra rapide lors d'une explosion de gaz dans une enceinte équipée d'un évent d'explosion[29]. On y observe l'explosion d'un mélange air-hydrogène, ainsi que l'éjection de gaz frais puis brûlés, à l'extérieur, par le biais de l'ouverture.

Lors de ce type d'expériences, les mesures effectuées consistent en des signaux de pression, comme illustré dans la ***Figure 20***. Pour modéliser ce

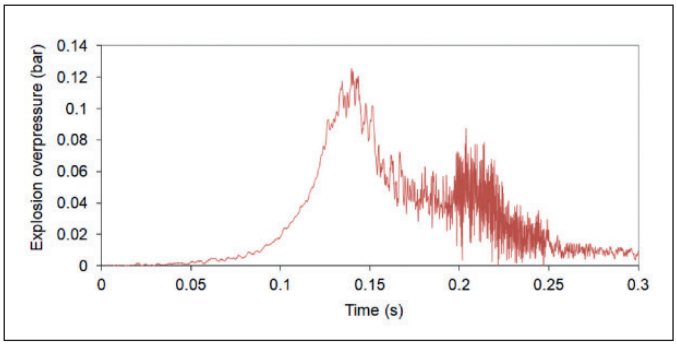


Figure 20

Exemple de signal de pression mesuré.

29. Évent d'explosion : il s'agit d'une sorte de porte calibrée pour s'ouvrir à une pression choisie, qui vise à s'ouvrir en cas d'explosion pour évacuer les gaz et limiter la surpression dans l'enceinte.

type de phénomène, il existe des modèles empiriques[30] ou phénoménologiques, qui reposent sur des modèles de physique connus et assez bien maîtrisés, mais dépendent de nombreux paramètres, dont certains ne sont pas aisément accessibles. L'idée sous-jacente est de développer, grâce à l'intelligence artificielle, des métamodèles substitutifs plus facilement utilisables, plus robustes lorsqu'il manque une partie des données ou dans le cadre de criblages préliminaires.

Des premiers travaux ont été engagés en ce sens en exploitant des données historiques recensées dans la littérature, englobant environ 270 essais réalisés au cours des 50 dernières années pour différentes configurations d'explosions et différents types de gaz. La *Figure 21* schématise la structure du réseau de neurones retenu qui utilise en entrée des variables aisément accessibles décrivant à chaque fois une configuration d'un essai d'explosion, telles que la nature du gaz (hydrogène, méthane, propane, éthylène), sa concentration dans l'air, le volume de l'enceinte, la taille de l'évent. Il a été choisi de ne pas introduire dans ces données des paramètres plus difficilement accessibles tels que la vitesse de flamme.

Malgré l'absence d'une optimisation complète de la structure du réseau, des premiers résultats prometteurs ont été obtenus. La *Figure 22* compare un signal simulé par le métamodèle avec un signal expérimental issu de mesures réelles, démontrant une bonne adéquation entre les prédictions du modèle et l'expérience. La comparaison des prédictions obtenues avec ce réseau de neurones avec les modèles empiriques et/ou phénoménologiques existants, montre une bonne cohérence et met en avant une grande performance de cet outil, globalement plus précis que tous les autres.

Néanmoins, les cas testés restent proches des données d'entraînement et il est encore difficile de se prononcer sur les capacités réelles de généralisation de ce modèle. Également, le non-respect de certaines tendances physiques, telles que l'écart entre pression atteinte dans l'enceinte et pression d'ouverture de l'évent qui devrait toujours être positif, a été mis en évidence.

Ces travaux sont donc poursuivis afin d'optimiser le réseau de neurones sur lequel est basé le modèle, mettre à jour

30. Basés sur l'expérience.

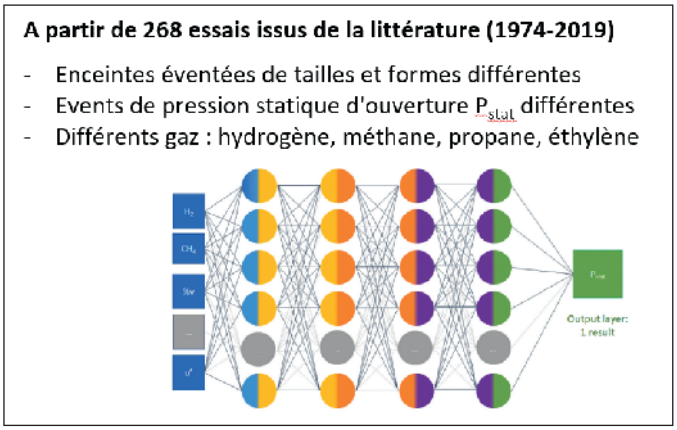


Figure 21

Structure du réseau de neurones du métamodèle développé.

la base de données (qui avec l'attrait actuel pour l'hydrogène s'est considérablement étoffée) et l'intégration de contraintes physiques.

Au-delà de la mise en place de ce métamodèle, des recherches sont également en cours pour exploiter l'intelligence artificielle afin d'analyser les séquences d'images issues de vidéos d'essais de l'Ineris. En effet, ils mettent en évidence des phénomènes explosifs, caractérisés par des durées et des effets externes souvent extrêmes et difficiles à mesurer. L'imagerie rapide se révèle être une méthode particulièrement intéressante pour étudier ces phénomènes, car elle permet d'enregistrer localement ou globalement les explosions ou leurs conséquences de manière non intrusive avec une cadence d'acquisition et une durée d'enregistrement correspondant, dans une certaine mesure, aux temps caractéristiques des explosions. Les travaux menés visent par exemple à mesurer les vitesses caractéristiques des éléments projetés ou des nuages à partir de ces enregistrements vidéo. Un autre enjeu concerne le traitement des images enregistrées dont la qualité est très variable (en fonction notamment de la nature de l'essai réalisé et des conditions extérieures).

31. Y. Grégoire, J. Daubech, C. Proust, E. Leprette, "Vented gas explosion overpressure calculation based on a multi-layered neural network", *Journal of Loss Prevention in the Process Industries* 74 (2022) 104641.

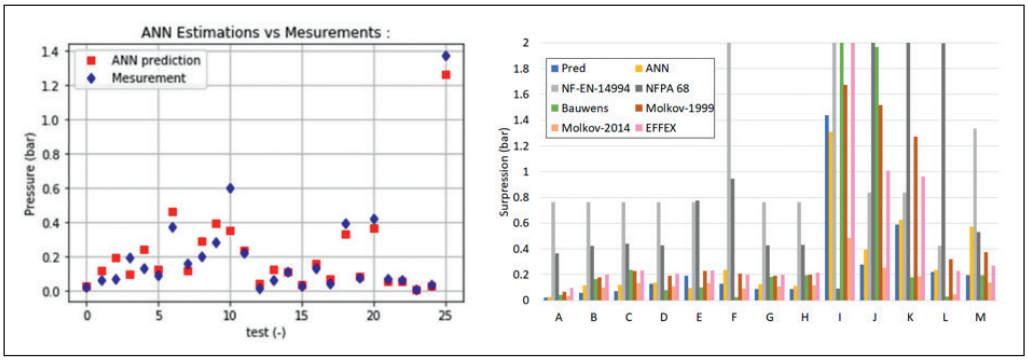


Figure 22

Performances du métamodèle en comparaison des données de pressions (pressure) mesurées lors des essais (à gauche) et des modèles phénoménologiques existants[31].

Conclusion

L'intelligence artificielle et les nouvelles approches méthodologiques donnent accès à des outils et des méthodes aux potentiels indéniables dans les différents secteurs d'activité de l'Ineris, et trouvent peu à peu leur place en complément des outils expérimentaux et de modélisation plus classiques.

Ces développements sont conduits en parallèle d'une stratégie de digitalisation des process et des plateformes d'essais de l'Ineris, et ne se limitent pas au développement de modèles prédictifs. Il s'agit également de disposer d'outils et de méthodes de travail améliorés, plus rapides et plus efficaces, dans nos travaux expérimentaux et de modélisation permettant de se focaliser sur les aspects les plus importants et complexes de l'évaluation des dangers et des risques.

Les démarches actuelles vont permettre d'optimiser les campagnes expérimentales, pour ne réaliser que les essais les plus intéressants, notamment lorsqu'ils nécessitent une logistique importante et en prenant en compte la dangerosité des essais.

Ces démarches visent également à capitaliser au mieux toute l'expérience disponible de manière étendue, à partir des données d'essais, mais aussi en exploitant le retour d'expérience de l'accidentologie (pour identifier les signaux faibles et des risques émergents par exemple).

Au-delà de l'exploitation de ces approches pour les besoins de l'Ineris (tel que présenté ici), l'Ineris s'intéresse aux enjeux de sécurité posés par ces nouvelles technologies et notamment la question de la certification des systèmes et barrières de sécurité utilisant l'intelligence artificielle ou celle de cybersécurité qu'ils pourraient poser.

Remerciements

Guillaume Fayet remercie Azziz Assoumani, Yann Gregoire et Jean-Bernard Kazmierczak pour leurs contributions respectives à cette présentation.

Partie 3

Intelligence artificielle et industrie

Julien ROMESTANT

Intelligence artificielle et parfumerie cosmétique : nouvelles expériences client et réduction du *time to market*

Julien Romestant est Directeur de l'intelligence économique de Cosmetic Valley, Chartres.

Introduction : la Cosmetic Valley

La cosmétique française a une longue histoire qui remonte à Versailles et aux parfums de la cour du roi Louis XIV, et même bien avant cela. Ces productions artisanales sont devenues une véritable industrie sous l'Empire avec entre autres la création de la société Guerlain en 1828. Aujourd'hui, la Cosmetics Valley c'est le cœur battant de l'industrie mondiale dans le domaine de la parfumerie cosmétique française.

La France est le premier exportateur mondial de parfums et cosmétiques, loin devant les États-Unis et l'Allemagne. Alors qu'en 2022 notre commerce extérieur a été déficitaire de 163 milliards en cosmétique, nous exportons 19,2 milliards d'euros de produits cosmétiques avec un excédent commercial de

15 milliards d'euros. C'est donc un secteur qui fonctionne bien et qui concentre l'intégralité de la chaîne de valeur et l'intégralité de toutes les composantes industrielles qui sont localisées sur le territoire national (*Figure 1*). Cela représente aujourd'hui 45 milliards de chiffres d'affaires, 3 200 entreprises, et également une recherche dédiée, pluridisciplinaire, dont la chimie fait intégralement partie.

1 Le marché des cosmétiques évolue vers la personnalisation

Selon une étude Euromonitor, 49 % des consommateurs veulent des produits et des services spécifiquement personnalisés. 77 % des consommateurs sont prêts à payer plus pour des produits personnalisés. Une expérience personnalisée est donc plus recherchée et entraîne des intentions d'achats plus importantes.

En termes de comportements d'achat, la génération X et des milléniaux sont les plus enclins à rechercher des cosmétiques personnalisés, suivis de près par la génération Z et les boomers. Longtemps, la cosmétique a été développée pour des peaux caucasiennes. Désormais, le relais de croissance du marché se trouve dans l'inclusivité et le développement d'une offre pour les peaux de toutes carnations. On voit sur le graphique que la demande est supérieure pour l'Asie pacifique, l'Amérique latine, et la zone Moyen-Orient - Afrique (*Figure 2*).

Figure 1

La Cosmetic Valley : schéma des métiers de la filière. © Cosmetic Valley.

2 La collecte des données en parfumerie cosmétique

Aujourd'hui, les données en parfumerie cosmétique se structurent en quatre grandes catégories.

1. Les données environnementales qui influent sur la peau et qui sont appelées l'**exposome**[1]. C'est-à-dire, « Est-ce que je vis dans un environnement qui est plutôt froid et sec, plutôt chaud avec un fort taux d'humidité, avec une exposition aux UV importante ? », « Est-ce que je suis plutôt à la campagne exposé aux pollens ou plutôt en ville entouré de pollution ? ». Tous ces facteurs vont avoir un effet

1. Ensemble des expositions extérieures auxquelles est confronté un être humain au cours de sa vie.

sur la peau comme une occlusion de la peau par la pollution et une irritation de la peau par les UV.

2. Le mode de vie : « Est-ce que j'ai un niveau de sommeil suffisant ? », « Est-ce que mon alimentation comporte 5 fruits et légumes par jour ? », « Est-ce que je suis fumeur et exposé aux radicaux libres par le fait de fumer du tabac ? » Le mode de vie influe sur notre peau.

3. La biologie et la génétique : la nature de la peau diffère selon les individus, qui peuvent avoir une peau plutôt grasse, plutôt sèche ou plutôt mixte. Les marqueurs de vieillissement sont différents en fonction du type génétique : une peau asiatique va avoir des marqueurs de vieillissement qui vont être plutôt des taches de vieillissement alors qu'un Caucasien va plutôt avoir des rides. Entre en compte également la topographie du visage sur laquelle nous reviendrons un peu plus tard.

4. Les émotions et les neurosciences parce qu'en fait la cosmétique est une industrie du bien-être. On se parfume, on se maquille, on fait sa toilette pour des interactions sociales entre êtres humains. Le but c'est de se sentir bien dans sa peau et d'aller vers les autres et la composante émotionnelle est très importante. Dans le domaine émotionnel, on regardera tout ce qui va concerner la prosodie[2], c'est à dire l'intonation de la voix mais aussi l'expression faciale. C'est typiquement le type d'émotion qu'on peut arriver à analyser par la reconnaissance faciale et par l'intelligence artificielle.

2. Ensemble des modulations de la voix, de l'expression verbale permettant de nuancer les émotions de l'interlocuteur.

2.1. Les capteurs de l'exposome

Prenons un exemple de capture des données de l'**exposome** qui a été développé par La Roche Posay, du groupe l'Oréal dans une application appelée My Skin UV Track[3] (*Figure 3*), créée en 2018.

3. Mon programme UV pour la peau.

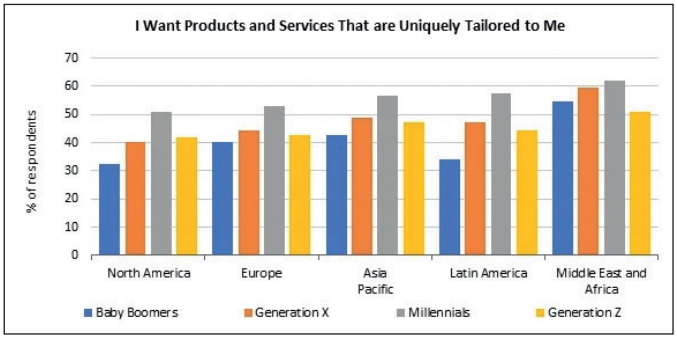


Figure 2

Personnalisation des produits cosmétiques. © Euromonitor.

Figure 3

Exposome : exemples de capteurs de données. © La Roche Posay.

La base est un capteur UV connecté à un smartphone avec un petit clip qui peut s'accrocher à votre maillot de bain et qui va mesurer votre exposition aux rayons solaires. L'avantage est que ce système est capable de dire « attention, ça fait deux heures que vous êtes exposé, si vous n'appliquez pas à nouveau la crème, vous allez avoir un coup de soleil ». Peut-être que si vous êtes sous les tropiques, il faudra appliquer de la crème un peu plus souvent que si vous êtes sur une plage en Méditerranée.

Nous utilisons tous une plateforme météo sur notre smartphone, cela est permis par la startup israélienne BreezoMeter, qui analyse les données environnementales (qualité de l'air, pollen et particules liées aux incendies). L'Oréal et BreezoMeter combineront leur expertise à la fois dans la science du vieillissement et dans celle de l'environnement dans **le but de développer une plateforme d'exposome unique axée sur la beauté.** Cette plateforme permettra de découvrir de nouvelles données sur la façon dont l'environnement affecte le vieillissement de la peau, et de fournir ainsi de nouveaux services aux consommateurs du monde entier, capables d'accompagner leurs besoins en matière de soin de la peau avec des routines de soin personnalisées et des conseils sur leur mode de vie.

2.2. L'analyse de la peau

Grâce aux smartphones, on prend des milliers de selfies. Ces selfies permettent également d'arriver à faire une analyse de la peau. Les caméras des smartphones sont tellement puissantes qu'on peut voir la vascularité[4] de la peau et ainsi faire des diagnostics de la peau juste par une analyse d'image réalisée par la startup Modiface (rachetée par L'Oréal). On peut voir les ridules et les pores dilatés et même mesurer leur dilatation exacte tellement les caméras sont de haute résolution. Grâce à un simple selfie, il est donc possible d'obtenir une cartographie de votre actualité cutanée : par exemple, on saura dire que vous n'avez pas assez bu car vous avez des rides un peu plus creusées. Tout cela peut être réalisé grâce à des algorithmes du *deep learning*[5] et du *machine learning*[6] qui ont été entraînés grâce à 6 000 images cliniques issues d'évaluation de recherche et testés sur plus de 4 500 selfies de femmes issues de populations asiatique, caucasienne et afro-américaine. Il sera ainsi possible de véritablement capitaliser ces données grâce à l'intelligence artificielle pour avoir des atlas de la peau et du vieillissement cutané. Grâce à cette technologie, Vichy a développé le diagnostic Skinconsult AI (***Figure 4***).

4. Inflammation des vaisseaux sanguins pouvant se traduire par des rougeurs au niveau de la peau.
5. Apprentissage approfondi. Pour une intelligence artificielle, il s'agit de reproduire le plus fidèlement possible les actions humaines grâce à des algorithmes.
6. Apprentissage d'une machine par sa capacité à simuler et commenter les résultats d'une expérience grâce à des bases de données.

Figure 4

Analyse de la peau avec l'intelligence artificielle. © Vichy

2.3. Les données concernant les molécules chimiques entrant dans les formulations

Du point de vue des ingrédients et de la formulation, on va jouer sur l'identification et le croisement de données : on va croiser les avis des consommateurs sur les différents segments de beauté. Aujourd'hui on dispose de millions d'informations sur les réseaux sociaux. On peut également analyser les comportements de consommateurs par rapport à des molécules. On va relier ces tendances aux connaissances déjà intégrées dans les portefeuilles d'actifs et, grâce à une discussion client, on va pouvoir dire assez précisément quel est l'éventail de molécules qu'on va pouvoir proposer pour répondre à ses besoins.

Cela se traduit, par exemple chez BASF, par des solutions basées sur l'intelligence artificielle qui permettent d'identifier le mix d'émollients optimal selon plusieurs paramètres ou alors sur des remplacements de molécules. On comprendra assez facilement comment on pourra remplacer des huiles minérales ou des diméthicones[7] dans la formule d'un produit, puisque les consommateurs commencent à s'éloigner des silicones. Cela permettra de reformuler plus facilement, de réduire le temps de mise sur le marché et d'éviter les essais/erreurs qu'on avait autrefois. Un large éventail pourra ainsi être présélectionné.

3 Applications de l'IA en parfumerie cosmétique

3.1. Assistance à la formulation de parfums

Les goûts concernant les parfums sont liés aux découvertes des odeurs et saveurs dans la petite enfance, c'est l'effet « madeleine de Proust ». Ainsi, un Brésilien aura une base culturelle qui sera complètement différente de celle d'un Français ou d'un Chinois.

7. Polymère à base de silicone, généralement utilisé comme agent protecteur de la peau.

En termes d'application de l'intelligence artificielle dans les parfums, il y a quelques années les précurseurs ont été IBM et Symrise pour la société brésilienne O'Boticario (***Figure 5***). Ils ont créé un système d'IA capable d'accumuler les connaissances relatives aux formules, aux ingrédients, à l'historique des succès et aux tendances. Et ce, afin de **fournir aux parfumeurs des assistants robotisés intelligents**. Les maîtres-parfumeurs auraient ainsi le loisir d'employer leur temps à l'amélioration de fragrances plutôt qu'à la recherche de nouvelles combinaisons.

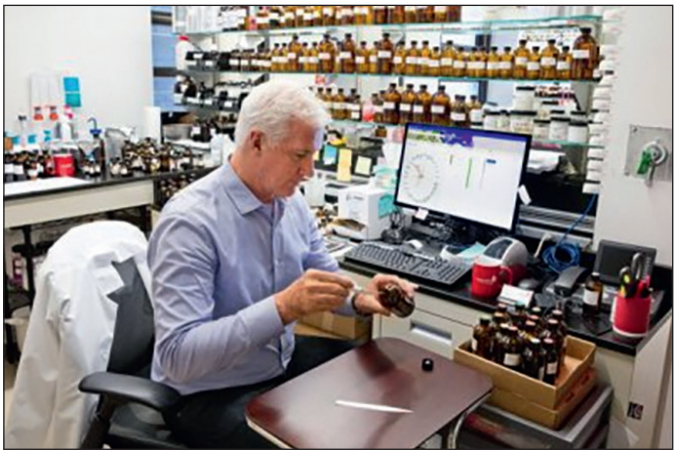


Figure 5

David Apel, parfumeur senior chez Symrise, a collaboré avec l'IA Philyra d'IBM pour créer deux parfums pour O'Boticario. © AFP/Relaxnews.

Givaudan, une société de création de parfums qui travaille pour les marques, est allé encore plus loin récemment avec EVE et Digipulse™. Cette solution combine plusieurs critères : le profil des consommateurs cibles, la région prévue pour le lancement, les exigences de conformité réglementaire, le positionnement du produit, les revendications, l'alignement avec les tendances du marché, le type de formule, la fourchette de prix de vente au consommateur. Tous ces éléments vont permettre un gain de temps considérable pour la mise sur le marché.

3.2. Intelligence artificielle, odeurs et émotions

Le dernier parfum « Phantom » développé pour la marque Paco Rabanne a été réalisé avec quatre parfumeurs et avec l'aide d'une IA (***Figure 6***). Le parfumeur Loc Dong a eu une idée folle : une overdose d'une molécule vintage l'acétate de styrallyle[8]. L'IA a recommandé d'utiliser 10 fois la dose des parfums modernes, l'overdose optimale pour un effet feel good. Cette overdose n'aurait pu être obtenue par la méthode

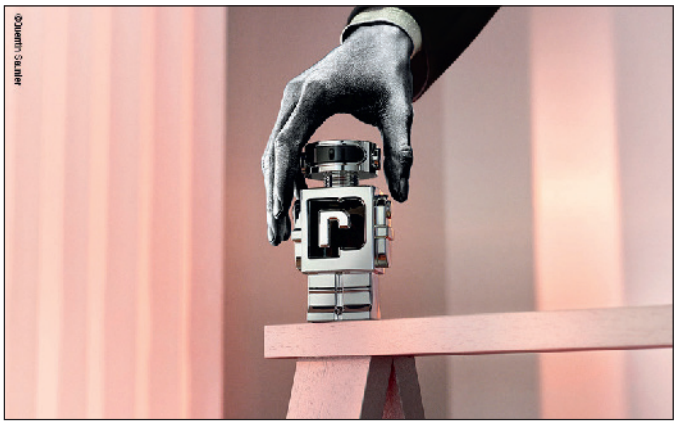


Figure 6

Parfum Phantom de Paco Rabanne développé grâce à une IA. © Quentin Saunier.

8. Molécule artificielle parfois utilisée pour apporter des notes végétales aux produits.

essai-erreur et on aurait pu mettre des années avant d'oser aller jusqu'à 10 fois la dose utilisée auparavant. Dans ce cas-là, l'intelligence artificielle a permis vraiment d'optimiser un développement et de réduire les essais-erreurs qu'on aurait pu avoir avec une formulation traditionnelle.

Passons à l'analyse des émotions. Les parfums Bvlgari-Tygar ont réalisé une expérience multisensorielle au travers de la vue, des sons, des odeurs et du toucher à partir de la mesure des ondes cérébrales, des micromouvements et du rythme cardiaque d'une personne. On lui met un casque, on lui donne un carton pulvérisé du parfum, et on le fait entrer dans une cabine sensorielle où on lui déclenche des stimuli visuels et musicaux choisis pour évoquer l'esprit du parfum. On essaye donc d'immerger le consommateur dans un environnement qui va retraduire le parfum.

Aujourd'hui, des start-up travaillent avec Givaudan pour permettre de traduire des odeurs en couleurs. Ce sont des processus assez obscurs mais qui sont déjà possibles. On a ainsi une nouvelle expérience client qui intègre la composante émotionnelle et on ressort avec une œuvre d'art transformée en NFT[9].

3.3. Les couleurs et le maquillage

Dans le maquillage, l'analyse d'image permet maintenant de détecter la teinte de la peau à partir de selfies, ou encore de faire des essais virtuels par la réalité augmentée. On pourra voir, en se regardant en selfie, l'effet d'une teinte de vernis à ongles sur ses ongles. L'expérience client est très améliorée puisque plutôt que se démaquiller et se remaquiller, il sera ainsi possible d'essayer très facilement des teintes de rouges à lèvres ou des teintes de fonds de teint. Les hommes ne sont pas en reste puisqu'on peut également essayer des teintes pour la barbe. Ainsi, le Chanel LIPSCANNER (***Figure 7***) permet de savoir dans le catalogue des rouges à lèvres Chanel, lequel est le mieux assorti à votre sac à main.

3.4. La formulation

Le Hylab[10] est un bon exemple de l'utilisation de l'IA pour formuler des soins cosmétiques personnalisés. C'est un genre de machine type « Nespresso »

9. *Non fungible token* (objet non tangible). Les NFT sont un type de monnaie électronique, associée à des objets purement numériques.

10. Gadget électronique permettant de créer des produits cosmétiques chez soi.

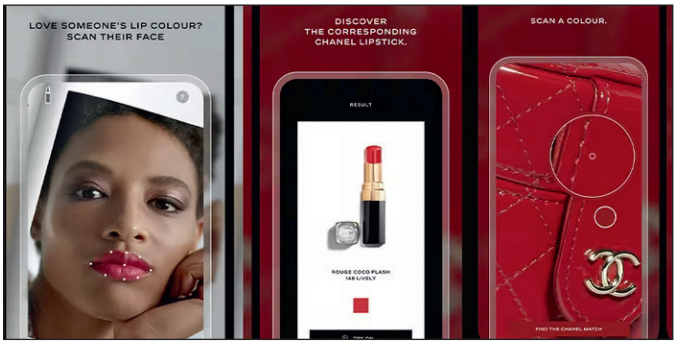


Figure 7

Couleurs, amélioration de l'expérience client. © Chanel LIPSCANNER.

qui va permettre d'avoir une base et des actifs sous forme de capsule pour formuler en fonction de vos besoins et de l'analyse des données quotidiennes. Une application smartphone vous guide pour vous proposer une formule et vous donner un protocole de formulation.

Grace à l'intelligence artificielle, on peut formuler en fonction du besoin et du lieu où l'on se trouve. Car le besoin d'un Français qui aujourd'hui est à Paris dans une atmosphère polluée à moins 1 degré n'est pas le même que celui d'un Chinois vivant dans un climat humide, ou de celui de quelqu'un qui est à Dubaï avec un climat désertique et sec. Cela permet même, pour quelqu'un qui voyage, d'adapter sa crème de soin en fonction du lieu dans lequel il se trouve pour que celle-ci soit en adéquation avec ses besoins.

YSL Rouge sur Mesure par Perso (***Figure 8***), un nouveau dispositif de beauté connectée de L'Oréal. Ici, l'intelligence artificielle et l'Internet des objets permettent de créer, sur mesure, des milliers de teintes personnalisées. L'application permet de tester la couleur souhaitée grâce à la réalité augmentée, et ensuite de faire une formulation sur mesure en fonction d'une photo de vous, de placer des épingles sur les zones clés de votre tenue, de votre peau, de vos cheveux ou de votre maquillage. La technologie intelligente générera des recommandations de teintes basées sur les règles d'harmonie des couleurs pour créer une teinte assortie ou non. L'appareil créera le rouge à lèvres parfait pour compléter votre tenue.

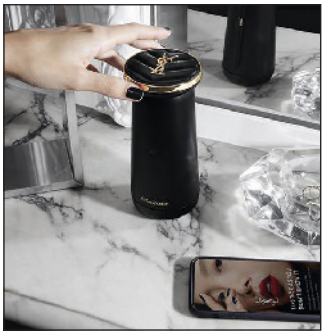


Figure 8

Formulation, exemples d'analyse par l'intelligence artificielle. © YSL Beauté (L'Oréal).

3.5. Les cheveux

Le domaine capillaire n'est pas en reste avec Schwarzkopf SalonLab™ (***Figure 9***), un appareil qui permet au coiffeur de scanner l'intérieur de la structure du cheveu grâce à la technologie du proche infrarouge[11]. Alimenté par l'application SalonLab, qui guide à la fois le coiffeur et le client dans le processus de consultation et d'analyse, le SalonLab Customizer fabrique des produits de soins capillaires sur place, pour chaque client, en fonction des données recueillies.

Figure 9

Schwarzkopf SalonLab™. © Henkel.

3.6. L'IA en amont de l'impression 3D

Quand on applique un masque de beauté sur la peau, parfois la technologie n'est pas forcément adaptée : vous ne

11. Analyse des émissions de rayonnement infrarouge par les différentes espèces chimiques.

tombez pas en face des yeux, ou vous avez le menton un peu plus long.

L'application Neutrogena MaskiD, qui se base sur la technologie FaceID[12], est un système de reconnaissance faciale qui permet avec un smartphone de reconnaître les visages, et notamment les reliefs des visages, permettant de réaliser la topographie du visage pour faire des masques sur mesure.

Il va ainsi être possible de personnaliser les zones du masque et donc de personnaliser la délivrance des actifs. Par exemple, si chez vous une ride est plus prononcée, il va être possible d'en tenir compte. Si vous avez une peau mixte avec une zone T[13] plus grasse, on va mettre moins d'actifs gras sur cette zone là et plutôt rajouter des actifs matifiants[14]. On peut ainsi créer une personnalisation en fonction de la zone du visage. Cela va de plus en plus loin : on va pouvoir ensuite imprimer le masque de manière personnalisée, avec des actifs ciblés selon les zones. Tout cela est permis par l'analyse et par l'intelligence artificielle en amont.

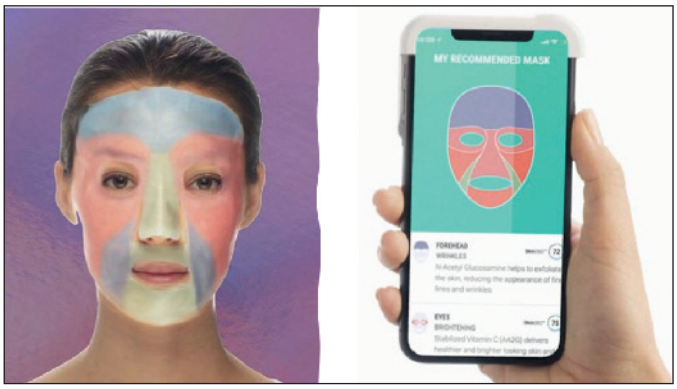


Figure 10

L'analyse topograhique par l'IA couplée à l'impression 3D permet de réaliser des masques de beauté à délivrance d'actifs personnalisée.
© Neutrogena MaskiD.

3.7. L'analyse de l'efficacité des soins cosmétiques personnalisés

L'analyse de l'efficacité de ces nouvelles techniques de soins cosmétiques personnalisés est en train de se mettre en place. La ***Figure 11*** montre l'exemple d'une utilisatrice devant un miroir connecté avec des capteurs d'analyse. Ces capteurs vont analyser leurs résultats par IA et l'outil de soins personnalisé va nous permettre de suivre si les rides ont vraiment diminué. Il est donc possible d'adapter la routine et de faire du *machine learning* afin de continuer à progresser. Énormément de données peuvent être capitalisées sur les différents consommateurs et sur l'efficacité des produits, ce qui va nous permettre d'arriver par *machine learning* à les améliorer encore plus.

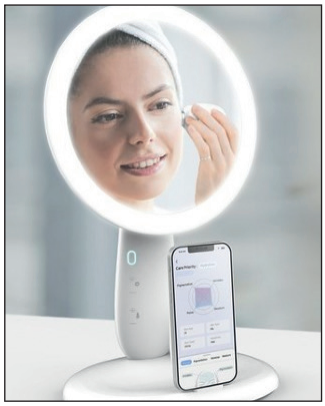


Figure 11

Monitoring, suivi et amélioration des résultats.
© Myskin Recovery Platform.

12. Système de déverrouillage de smartphone en utilisant la reconnaissance faciale.
13. Zone médiane du visage.
14. Produit permettant de gommer l'excès de sébum.

Conclusion

La collecte des données est facilitée par les capteurs miniaturisés comme les smartphones, et permet d'avancer sur ces sujets. L'exploitation des données par l'IA apporte une aide à la décision, une réduction du temps de mise sur le marché et une personnalisation des produits. On commence aujourd'hui à explorer le *machine learning* avec un croisement des données en masse, un monitoring des résultats et de leur efficacité dans le temps pour avoir des produits réellement adaptés aux besoins.

Transition énergétique et technologies numériques : comment la donnée est utilisée pour la stratégie multi-énergies de TotalEnergies

D'après la conférence de **Michel LUTZ**

Michel Lutz est depuis 2016 Chief Data Officer[1] pour la compagnie TotalEnergies. Son activité consiste ainsi à savoir quoi et comment faire pour mieux utiliser les données au niveau stratégique pour la compagnie. Il est aussi Head of Data au sein de TotalEnergies Digital Factory[2], filiale détenue à 100 % par TotalEnergies, et qui développe des solutions numériques pour la Compagnie. À ce titre, il pilote une équipe d'une quarantaine de data scientists, data engineers et autres experts en matière de data science, recherche opérationnelle, MLOps, IA générative, etc.

1. Responsable des données.
2. Usine digitale.

Introduction : la stratégie multi-énergies de TotalEnergies

TotalEnergies considère l'énergie comme l'un des enjeux majeurs de la société d'aujourd'hui et de celle de demain. À cet égard, la compagnie engage ce qu'elle considère être une importante transformation. Celle-ci repose sur deux grands axes pour répondre aux attentes des populations en termes de transition énergétique (*Figure 1*).

D'une part, nous travaillons sur la réduction et la compensation des émissions de dioxyde de carbone liées aux activités industrielles de la compagnie. Cela concerne la production et la distribution d'hydrocarbures, mais aussi toutes les nouvelles énergies. D'autre part, TotalEnergies accompagne des recherches et des projets dans le domaine des nouvelles énergies : l'électricité, le renouvelable, l'éolien, le solaire, l'hydrogène et la biomasse. La chimie joue un grand rôle dans tous ces domaines dans lesquels existent énormément de données. L'utilisation conjointe des données et des technologies digitales est un levier important pour accompagner cette transformation.

La variété des données à notre disposition pour justement accompagner ces transformations de la compagnie est représentée sur la *Figure 2*.

Notre patrimoine dispose d'une grande variété de données avec des enjeux techniques très différents. **Notre cœur historique sur les données de sous-sol** représente des volumes énormes : nous disposons de 50 pétaoctets[3] d'archives sismiques pouvant être traitées avec notre super calculateur. Ces données représentent des grands enjeux pour la connaissance du sous-sol, et peuvent par

3. Unité de mesure de la mémoire, correspondant à un million de gigaoctets.

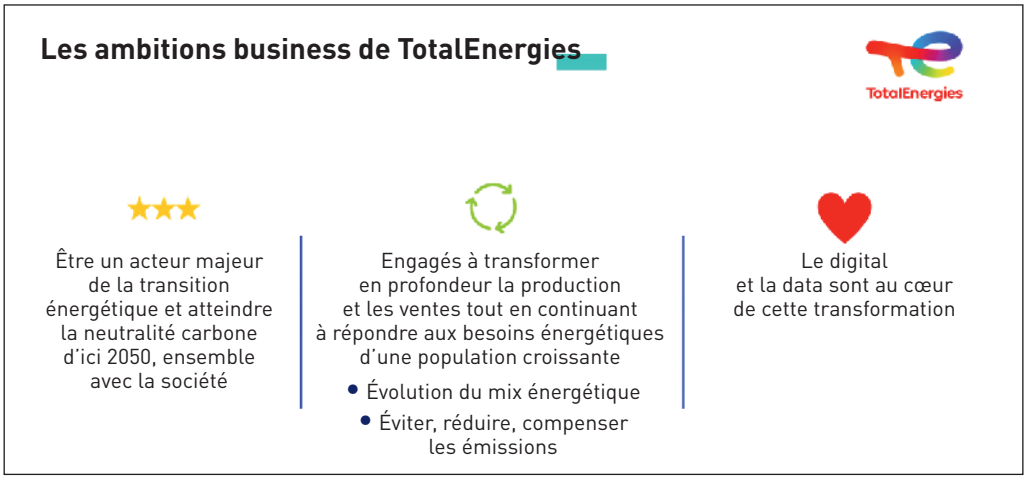


Figure 1

TotalEnergies et la transition énergétique.

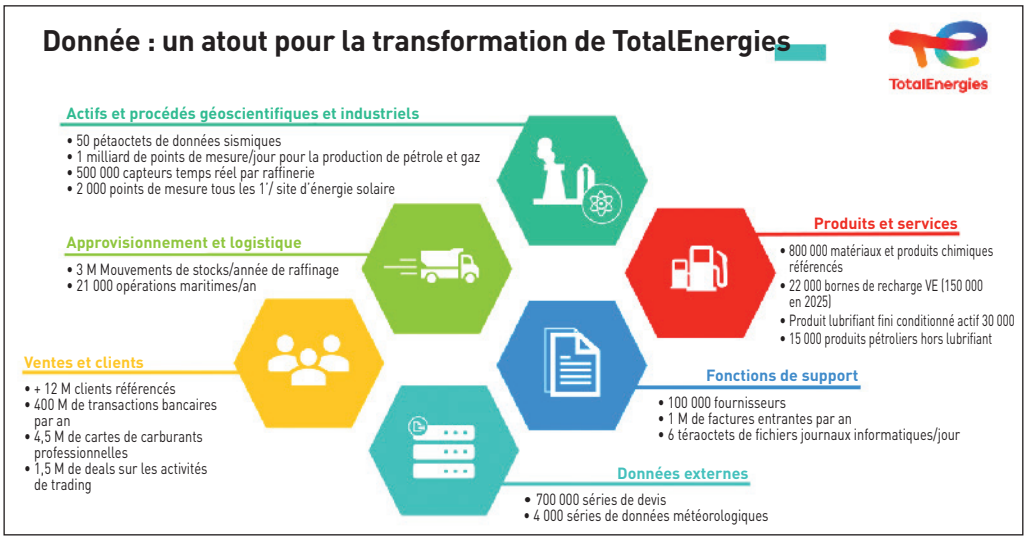


Figure 2

Patrimoine de données de TotalEnergies.

exemple être utilisées pour les projets de capture et de stockage de CO_2.

En ce qui concerne **notre activité industrielle**, nous avons des historiques de données importants sur les activités **de raffinage et de production de pétrole**. Nous travaillons aussi de plus en plus autour de ce qui concerne la production d'énergies renouvelables, notamment le solaire et l'éolien. Ce sont des unités industrielles instrumentées avec des dizaines de millions de capteurs et déployées dans le monde entier qui nous remontent de l'information sur le fonctionnement de nos usines de production d'énergies. C'est précieux, pour optimiser au mieux le fonctionnement de nos usines et diminuer les émissions de dioxyde de carbone.

Nous disposons aussi de données très variées parmi nos 12 millions de clients référencés dans les bases, par exemple pour tout ce qui est logistique client. C'est très important pour optimiser nos mécaniques de distribution d'énergie. Ce domaine est en pleine évolution avec le déploiement de bornes de rechargement.

1 L'utilisation des données et l'écosystème scientifique de la société

Les équipes travaillent de différentes façons sur l'utilisation de ces données. L'écosystème scientifique de la compagnie est représenté sur la ***Figure 3***.

1.1. La recherche et le développement

Les équipes de R&D travaillent en amont. Ces équipes recherchent les solutions pour

les défis actuels et préparent les problématiques de demain. Certaines équipes maîtrisent et utilisent beaucoup l'intelligence artificielle et le *machine learning* dans les procédés chimiques, c'est le cas pour la génération de matériaux. Ce sont des domaines dans lesquels nous travaillons en partenariat avec des laboratoires universitaires.

1.2. Les équipes spécialisées

Nous avons des équipes spécialisées qui utilisent les nouvelles technologies ou les technologies « numériques » sur des problématiques métiers précises.

Historiquement, nous utilisons depuis longtemps la simulation numérique pour la modélisation du sous-sol et des réservoirs. Cependant, nous utilisons désormais de plus en plus les méthodes de *machine learning* d'analyses de données pour enrichir les points de vue et les façons de traiter les problèmes.

Nous avons aussi des équipes historiques pour tout ce qui concerne la modélisation et l'optimisation des processus de raffinage.

Il y a enfin des équipes très spécialisées sur des sujets comme : la chimiométrie, la maîtrise statistique des processus dans les usines, la production d'énergies renouvelables, etc.

1.3. L'équipe *Machine Learning* (ML)

L'équipe data de la Digital Factory, que gère Michel Lutz, est spécialisée en *Learning engineering*[4]. Dans ce domaine, une compétence spécifique

4. Ingénierie de l'apprentissage machine.

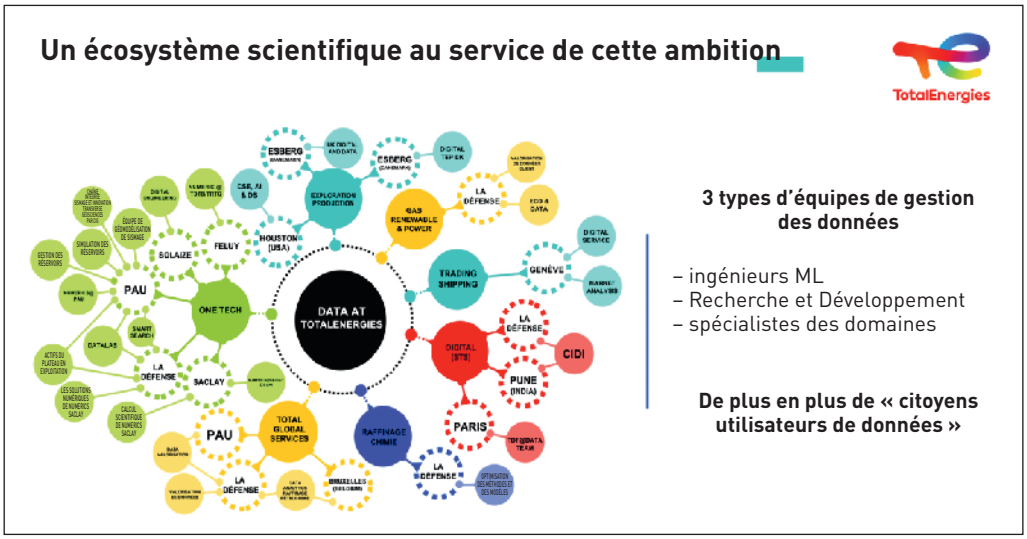


Figure 3

Un écosystème scientifique chez TotalEnergies.

est la mise en production des modèles dans des applications mises dans les mains de l'utilisateur. Nous devons élaborer des applications qui embarquent l'intelligence artificielle avec des interfaces d'utilisation développées avec des solutions logicielles robustes.

Par ailleurs, l'usage des données se démocratise de plus en plus : on appelle souvent cela *Citizen data* ou *Citizen data scientist*[5].

Dans une société comme TotalEnergies, beaucoup de personnes, dans les métiers et dans les usines, ont de bonnes compétences scientifiques et savent manipuler les données. L'enjeu est donc de les rendre plus efficaces, de leur faciliter l'accès à la donnée pour qu'ils soient plus autonomes dans leur capacité à mieux décider au quotidien avec les bons outils et les bonnes données.

5. Donnée accessible. Ces professionnels utilisent régulièrement l'intelligence artificielle dans leur travail, sans pour autant l'avoir mise au point ni en en étant expert.

1.4. L'exemple du déploiement des panneaux solaires (*Figure 4*) : le programme SOLEAD

C'est l'équipe R&D qui à la base a travaillé sur la recherche d'algorithme de traitement d'images pour identifier à partir d'images satellites caractéristiques des zones potentiellement intéressantes pour développer les installations des panneaux solaires en fonction de différents facteurs dont le potentiel solaire de la zone.

Figure 4

L'installation de panneaux solaires : un exemple de coopération pour la recherche d'une zone potentiellement intéressante.

L'équipe (ML) travaille actuellement avec eux sur le résultat de cette recherche pour l'intégrer dans une application capable de délivrer l'information en temps réel, qui sera mise à la disposition des personnes de TotalEnergies qui développent l'activité solaire.

Nous développons en plus les algorithmes pour aider à la prise des décisions. Il existe tout un ensemble de données sur le contexte économique d'une zone géographique qui peuvent être intégrées, à partir desquelles on modélise les préférences des décideurs pour savoir si c'est intéressant ou non de mettre en place des panneaux solaires.

Cet ensemble permet de faire un produit qui sera ensuite utilisé de façon opérationnelle au quotidien pour la stratégie du déploiement solaire de TotalEnergies.

2 La Digital Factory Data Team (l'usine numérique)

L'équipe Data de la Digital Factory rassemble une quarantaine de personnes expertes en science des données appliquées et en ingénierie du *machine learning* pour trouver des solutions opérationnelles déployables et utiles pour les gens au quotidien.

Les compétences scientifiques doivent être pluridisciplinaires : modèles statistiques, *machine learning* (apprentissage automatique), recherche opérationnelle, compétences en génie logiciel et gestion du cycle de vie des modèles.

Il faut avoir une approche multi méthodes pour trouver de bonnes solutions mais ensuite, pour que ces solutions soient utiles, la partie développement de logiciels est importante.

Une dimension importante de notre travail est donc le développement de logiciel pour trouver des solutions répondant aux besoins des utilisateurs. Je citerai deux exemples de produits que nous avons réalisés.

Le premier est un logiciel de vision artificielle qui permet de reconnaître des gravures sur des tuyaux utilisés en industrie. Ces gravures étaient très difficiles à lire, donc nous avons créé un algorithme de vision artificielle que nous avons déployé sur téléphone portable afin de pouvoir scanner les tuyaux avec le téléphone portable. Après avoir eu la référence du tuyau, on peut ainsi l'utiliser pour optimiser les mouvements de déplacements de stocks.

L'autre exemple a été créé pour Saft, une filiale de TotalEnergies qui fabrique des batteries industrielles. Nous avons élaboré un algorithme qui fait de la prédiction de risques de non-qualité en sortie de chaîne de fabrication. Cet algorithme est basé sur toutes les données de capteurs de l'usine, et nous l'avons ensuite déployé sur un multimètre. Quand le technicien qualité fait une mesure avec le multimètre sur une batterie, cela appelle l'algorithme qui fait une prédiction de risque de non-qualité qui s'affiche dans l'usine sur un écran, et tous les techniciens qualité peuvent ainsi optimiser leurs processus d'inspection en fonction de

cette aide à la décision fournie par l'algorithme.

Ce type de travail nécessite une expertise assez spécifique en développement logiciel pour être capable d'amener les solutions jusqu'au bout dans les mains des utilisateurs.

Le deuxième axe de notre métier, au-delà de la partie technique, est de faire de l'intelligence artificielle pour et avec les utilisateurs : MLOps[6].

Très souvent quand on fait un algorithme, il a au début un certain niveau de performance. Ensuite nous le mettons en production dans un logiciel, mais il va peut-être se tromper ou être confronté à des situations qu'il n'a jamais vues et pour lesquelles il pourra ne pas prendre une bonne décision. Nous devons construire des interfaces qui permettent justement de récupérer une boucle de rétroaction avec l'utilisateur. Cela signifie que si l'algorithme n'a pas le comportement attendu, on aura un retour utilisateur pour avoir un entraînement continu, ce qui nécessite de mettre en place des mécaniques logicielles de réapprentissage assez particulières. C'est un objectif au cœur de ce qu'on appelle le MLOps.

Le troisième axe est un peu plus méthodologique mais néanmoins très important. Nous sommes très engagés dans ce qu'on appelle l'intelligence artificielle de confiance pour ne pas mettre en place des systèmes non maîtrisés qui pourraient prendre de façon autonome des mauvaises décisions, afin que nos utilisateurs aient confiance dans ce que nous faisons.

Pour réaliser cela, tout un cadre, qui s'inspire des bonnes pratiques qui existent sur le marché, a été mis en place et nous avons aussi des projets de régulation en cours pour s'assurer qu'on fait bien les choses et aussi donner un cadre de confiance entre les utilisateurs.

6. MLOps (*Machine Learning Operations*). Le MLOps peut se définir comme un ensemble de pratiques combinant le *machine learning*, le DevOps et le *Data Engineering*, qui vise à déployer et à maintenir les systèmes ML en production de manière fiable et efficace.

Conclusion
Perspectives stratégiques

En entreprise, les enjeux de la transformation data ne sont pas que de la science et pas que de la data science. Si on veut que ce soit bien fait, à l'échelle, et que ce soit impactant, il y a tout un ensemble de facilitateurs à mettre en place, et le rôle du Chief Data Officer se résume en 3 axes (***Figure 5***).

Technologies

Historiquement, les données proviennent des serveurs installés dans les usines et les systèmes opérationnels sont des données difficilement accessibles. Il y a donc toute une transformation technologique à mettre en place pour accéder plus facilement à ces données, les rendre plus facilement disponibles, pour enfin être capable de les utiliser avec les technologies qui permettent de faire de l'intelligence artificielle et du *machine learning*. Ce sont des

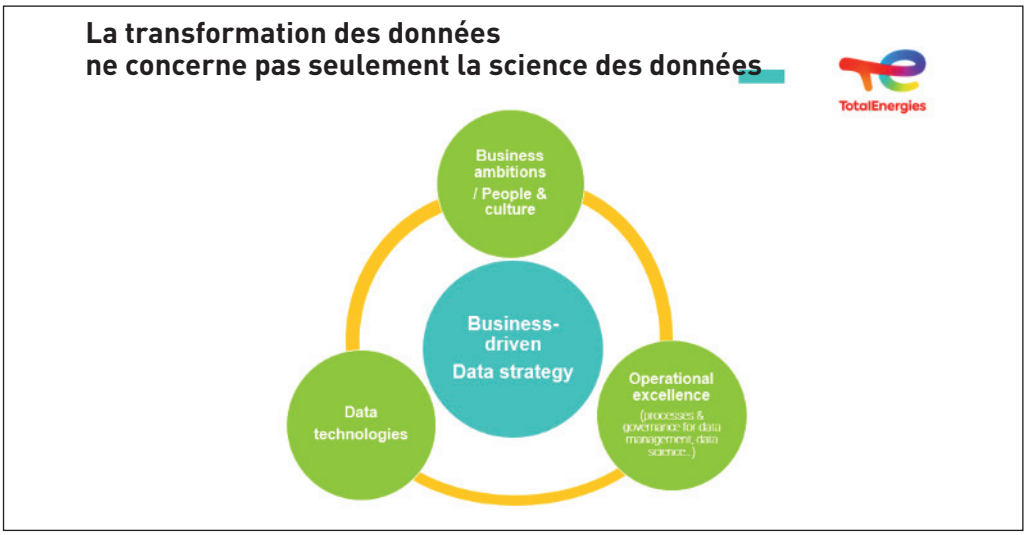


Figure 5

Différents axes de la stratégie data

grands chantiers menés avec la direction informatique de la compagnie.

Data management

Ensuite, il y a l'excellence opérationnelle à avoir, rien qu'au niveau de la donnée. On peut faire de la super data science mais si la donnée est de mauvaise qualité, ou n'est pas maîtrisée, on prendra nécessairement des mauvaises décisions quelle que soit la qualité des scientifiques qui font des modèles. Donc ce sont des sujets qu'il faut travailler avec les métiers pour mettre en place des processus, des outils, des bonnes pratiques pour garder la maîtrise de la qualité de nos données. C'est vrai pour la data science, mais c'est utile de façon générale pour la prise de décision au quotidien : pas de bonnes données, pas de bonnes décisions. C'est un vrai enjeu de toujours progresser sur ce point.

Vision métier et transformation

Le troisième axe, je l'ai appelé vision métier et transformation. L'objectif est de donner à tous les membres de la compagnie la volonté d'aller toujours plus loin dans l'utilisation de la donnée.

Il faut donner l'envie et la capacité à tout le monde, d'être plus autonome pour utiliser la donnée pour mieux décider.

Il faut aussi transformer les façons de faire avec les données. Dans une société comme TotalEnergies, beaucoup de gens n'ont pas attendu les *data scientists* et les informaticiens pour manipuler les données, pour modéliser des problèmes. Cependant les gens qui viennent du *machine learning* apportent un regard un peu différent sur la modélisation d'un problème.

Il faut donc faire se rencontrer les cultures, il faut savoir comment faire cohabiter des gens

qui vont partir avec juste de la donnée et un peu moins de connaissances avec ceux qui modélisent plutôt un problème avec simulation numérique ou avec une connaissance vraiment phénoménologique. Il faut arriver à trouver les ponts pour qu'ils travaillent ensemble et, finalement, qu'ils regardent différemment les problèmes pour mieux décider ensemble. C'est un des enjeux importants et passionnants du développement de l'IA.

Laurent SCHIO

De la sérendipité à l'intelligence artificielle en recherche pharmaceutique

Laurent Schio est Responsable France de la plateforme de recherche Integrated Drug Discovery chez Sanofi.

Introduction

L'origine de la sérendipité

Le mot *serendipity* (sérendipité en français) vient d'un vieux conte, « Les 3 princes de Serendip », une localisation qui doit correspondre au Sri Lanka d'aujourd'hui.

Ces 3 princes, qui avaient refusé de succéder à leur père, sont partis en voyage et ont analysé les traces d'un chameau. Par leurs observations, ils ont pu voir et deviner que le chameau était borgne, qu'il lui manquait une dent, qu'il portait une femme enceinte, qu'il boitait, etc. Au point qu'ils ont été condamnés pour le vol du chameau, puisqu'il avait disparu et donc emprisonnés, condamnés à mort, et puis finalement graciés, récompensés parce que le chameau avait été retrouvé. C'est le genre d'ascenseur émotionnel qu'on vit en *drug discovery*[1][2], et l'observation reste une des facultés les plus importantes et de qualité dans notre métier.

Notion de problème complexe et compliqué

Le processus de la recherche dans le domaine du *drug discovery* est empreint de

1. IDD : *Integrated Drug Discovery*.
2. La découverte de médicaments.

sérendipité, découvertes faites « par hasard ». Le *drug discovery* est un problème complexe, et on doit faire la différence entre un problème compliqué et un problème complexe.

Un problème compliqué, c'est par exemple d'envoyer une fusée sur la Lune. Pour le résoudre, il faut, mais « il suffit », d'avoir les bonnes équations ; on les résout, on a les bons investissements, on a les bonnes personnes, parce qu'on a vu avec la mission Apollo 13 que c'est parfois un peu compliqué de revenir sur Terre. Un problème complexe, c'est par exemple d'élever un enfant : ce qui marche dans une famille ne marche pas dans une autre, ce qui marche avec un enfant ne marche pas avec un autre, ce qui marche un jour ne marche pas le lendemain…

Exemples de médicaments découverts par sérendipité

Le *drug discovery* est un problème complexe. C'est la raison pour laquelle son histoire est jonchée de découvertes dites faites « par hasard », mais c'est plutôt par sérendipité qu'il faudrait dire. Je vais donner deux exemples rapidement.

Le doliprane

Le premier est celui du paracétamol, ou doliprane (***Figure 1***). Cette molécule, l'acétaminophène, a été synthétisée pour la première fois en 1878. Dix ans plus tard, un professeur a demandé à ses étudiants d'aller chercher du naphtalène, et ils sont revenus avec de l'acétanilide. Par erreur, le pharmacien a fourni une mauvaise molécule. C'est comme ça que, par hasard, les propriétés antifièvre de l'acétanilide ont été découvertes. Il a fallu 60 ans de plus pour découvrir qu'en fait les effets qu'on observait

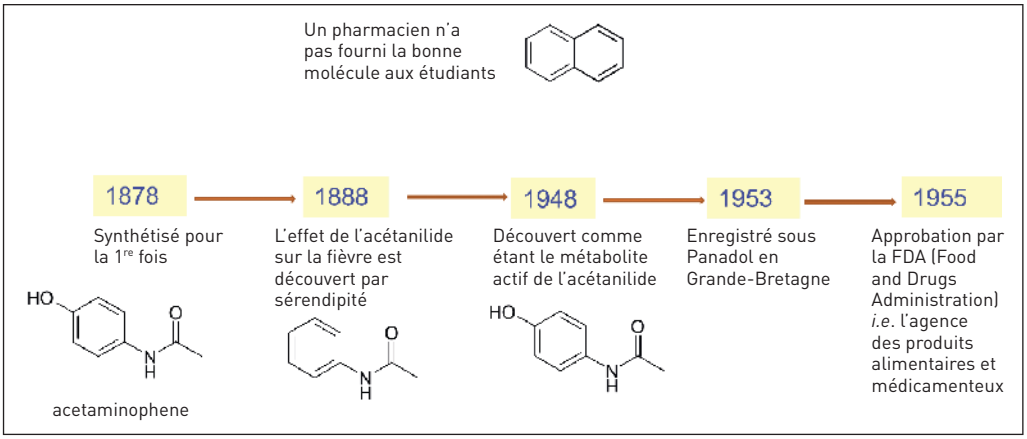


Figure 1

L'histoire du paracétamol.

in vivo[3] n'étaient pas liés à la molécule, mais à un métabolite hydroxylé formé *in vivo*.

Il a donc fallu attendre environ 80 ans entre la première synthèse et l'enregistrement de la molécule sous le nom de paracétamol, pour qu'elle devienne un médicament. Évidemment ce genre de process[4] n'est plus acceptable aujourd'hui, on ne peut pas se permettre d'attendre 80 ans pour fournir un nouveau traitement à des patient.es.

Le Taxotère

Depuis deux millénaires, il est connu que les extraits d'arbres d'if sont très toxiques, et les Gaulois déjà badigeonnaient leurs lances d'extraits d'if pour les rendre plus fatales. Par ailleurs, on retrouvait régulièrement, par exemple, des chevaux morts autour des cimetières où des ifs étaient plantés pour leur ombre, et les chevaux s'en nourrissaient et mourraient.

On a découvert que la toxicité de l'if était liée à une molécule, qui s'appelle la taxine (***Figure 2***) ; par chimie, on a pu maîtriser cette toxicité et en tirer une nouvelle molécule qui va devenir le Taxotère (***Figure 3***), plus spécifique pour les cellules cancéreuses. Ce produit a été approuvé vers les années 1995 pour le traitement du cancer du sein.

Ces deux histoires sont vraies, et leurs processus de génération du médicament sont beaucoup trop longs, trop aléatoires pour qu'aujourd'hui on puisse soutenir un portefeuille de projets de recherche dans des groupes pharmaceutiques.

1 Le processus actuel de découverte des médicaments

1.1. L'approche *« Bed-to-Bench-to-Bed »*

Le nouveau paradigme[5] en vigueur aujourd'hui s'appelle « *from bed-to-bench-to-bed*[6] ». Il consiste à analyser les maladies observées chez des patients, par exemple une patiente atteinte du cancer du sein. On analyse biologiquement les tumeurs, on détermine quels sont les mécanismes qui sont déficients et qui ont créé cette tumeur, par exemple des mutations oncogènes, ou des surexpressions de certaines protéines[7]. À partir de cette compréhension de la maladie, on développe une procédure pour contrer cette déficience en ayant recours aux sciences chimiques ou biologiques pour obtenir, on l'espère, un résultat positif comme la guérison du cancer du sein.

3. Au sein du vivant, de l'organisme, du corps.
4. Procédé, façon de faire les choses.

5. Modèle, représentation.
6. Du lit au laboratoire, jusqu'au lit d'hôpital.
7. Production élevée de protéines.

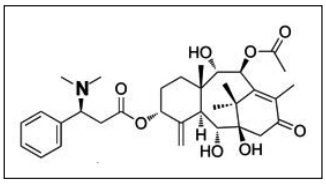


Figure 2

Molécule de taxine, isolée en 1856.

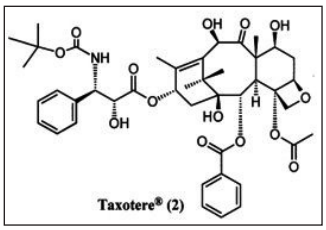


Figure 3

Molécule de Taxotère, approuvée en 1996 pour le traitement du cancer du sein.

Avec ces nouvelles façons de travailler qui datent quand même de 10 à 20 ans et qui sont basées sur l'« analyse translationnelle[8] », on est aujourd'hui capable de mettre un médicament sur le marché, en 10-15 ans à partir de la sélection de la cible thérapeutique. En général, ces techniques demandent de synthétiser entre 5 000 et 10 000 molécules par projet ; cela entraîne un délai d'environ 4 ans de recherche avant de délivrer une molécule pour un essai clinique[9] (***Figure 4***).

Ce délai n'est pas celui du doliprane, mais il est tout de même trop lent : beaucoup d'investissement, en général 2 à 3 milliards, pour mettre une molécule sur le marché. La Recherche n'est pas l'étape la plus coûteuse du processus...

8. Concept qui correspond aux efforts à fournir pour produire des applications concrètes à partir de connaissances fondamentales.
9. A pour but d'évaluer la tolérance et l'efficacité du composé.

1.2. Les médicaments possibles

Après l'analyse de la maladie et la compréhension des mécanismes biologiques (déficiences biologiques) responsables, on peut identifier différents types de modalités d'intervention. On dispose d'un panel de types de modalités thérapeutiques envisageables qui sont classés en 2 catégories (***Figure 5***) : les modalités « synthétiques » (faits par la Chimie), et les biologiques, qui sont produits par les cellules. Les synthétiques comprennent les petites molécules et les peptides[10], les biologiques les protéines, les anticorps monoclonaux[11], bi-spécifiques[12] ou tri-spécifiques. De nouveaux médicaments ont été lancés

10. Molécule composée de plusieurs acides aminés ; un acide aminé est composé d'un groupe amine et acide carboxylique.
11. Synthétisés en laboratoire à partir d'un seul gène.
12. Peuvent reconnaître deux récepteurs (antigènes) à la fois.

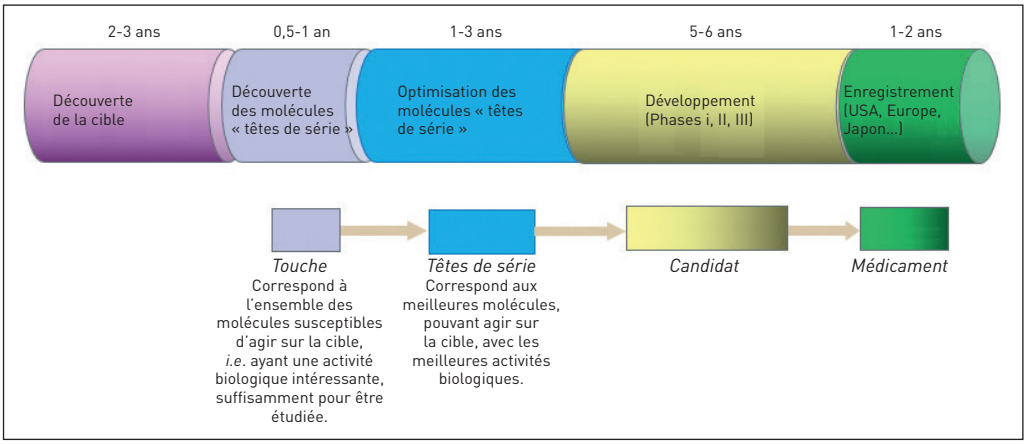


Figure 4

Chaîne de valeur pour la découverte et le développement de médicaments.

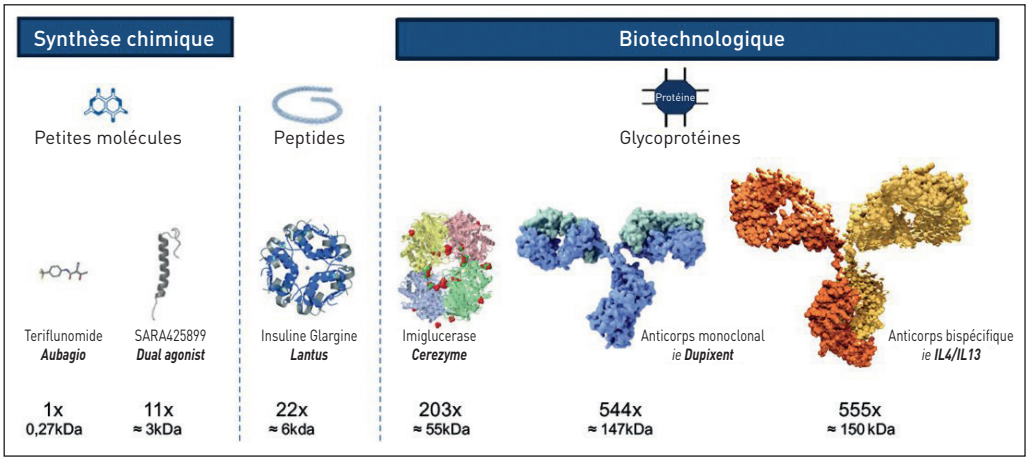


Figure 5

Panel de ressources pour produire un médicament.

pour le cancer avec des anticorps qui ainsi s'attachent à la fois aux cellules cancéreuses et attirent des cellules du système immunitaire pour éliminer les cellules cancéreuses.

Toutes sortes d'essais sont entrepris, par exemple pour voir l'influence de la taille des composants moléculaires sur l'activité ou les propriétés pharmaco-cinétiques. Ainsi, chez Sanofi, on travaille aussi sur des constituants des systèmes immunitaires des chameaux basés sur des « *nanobodies* » (***Figure 6***), beaucoup plus petits que les anticorps humains.

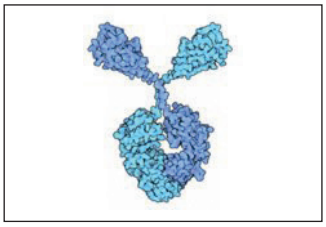


Figure 6

Structure du « nanobody ».

Ces travaux construisent tout un panel de modalités thérapeutiques nouvelles pour soigner différents types de maladies.

2 Comment optimiser les propriétés en drug discovery

2.1. Prendre en compte le principe de marge thérapeutique...

On peut maintenant revenir sur la remarque initiale selon laquelle le *drug discovery* est « complexe », intrinsèquement, quelles que soient les modalités qu'on utilise, quelles que soient les maladies qu'on adresse.

En fait, il s'agit de chercher un compromis entre la puissance du médicament, sa tolérance et puis la possibilité de sa distribution chez le patient, par voie orale, voie sous-cutanée, etc. Ces propriétés s'analysent selon des vecteurs

(nous n'allons pas approfondir ici l'aspect technique), mais les possibilités d'amélioration ont parfois des directions opposées (***Figure 7***). Il faut dégager les meilleurs compromis *in fine* pour respecter l'ensemble, arbitrages entre bonnes expositions, bonnes activités, absence de toxicité.

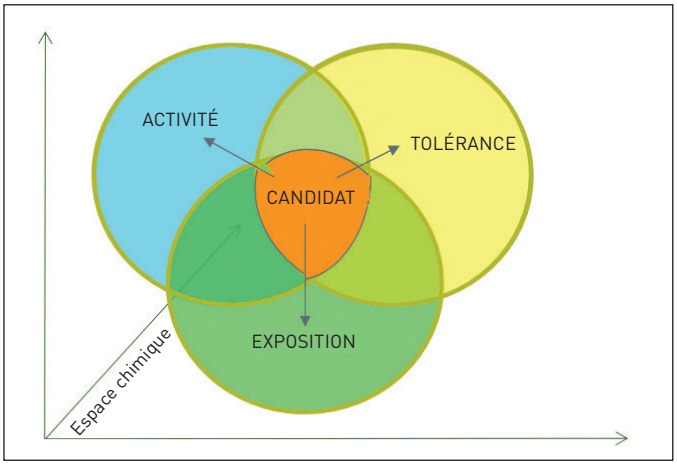


Figure 7

Diagramme correspondant au compromis pour obtenir le médicament désiré.

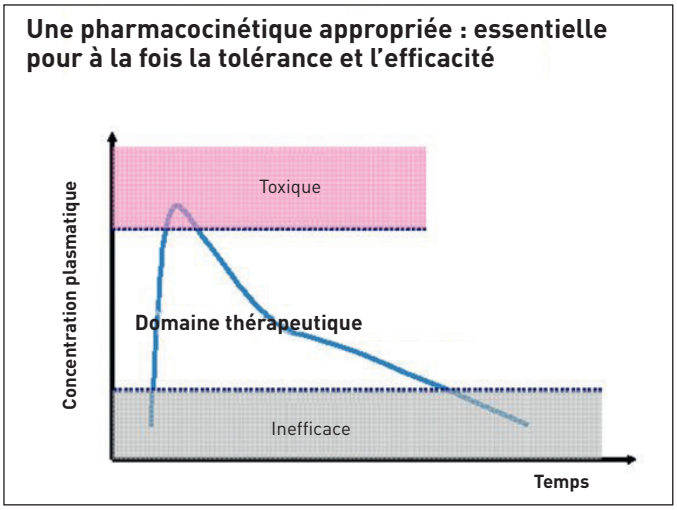


Figure 8

Graphe montrant l'importance de la dose pour l'efficacité d'un médicament.

Une grande vérité reste incontournable : pour tous les médicaments, le facteur clef de l'arbitrage, c'est la question de dose (***Figure 8***). En dessous d'un certain niveau il n'y a pas d'activité, au-dessus il y a activité. Si on monte trop haut la dose (exposition), on induit des effets secondaires indésirables, parfois létaux, et qui sont propres à chaque personne.

2.2. Et maîtriser chaque propriété de drugabilité

Dans ce processus, la *drug discovery* est conduite selon certains paramètres (***Figure 9***) qui sont liés à l'activité, la sélectivité, l'efficacité *in vivo* (en bleu). De surcroît, on doit aussi prendre en compte des paramètres de « drugabilité » (le conditionnement du médicament à la prise par les patients, par exemple absorption orale, exposition, concentration, en vert).

On essaye d'éviter des problèmes en suivant, en mesurant ou en calculant au cours du processus des propriétés cardiovasculaires potentielles, des affinités sur des récepteurs qui peuvent induire des effets cardiovasculaires importants, ou des interactions drug-drug parce que les molécules peuvent interagir avec les cytochromes P450[13], etc. (en jaune). Les diagrammes qui représentent toutes ces études portent le

13. Enzymes qui ont pour fonction de métaboliser des substances dans notre corps.

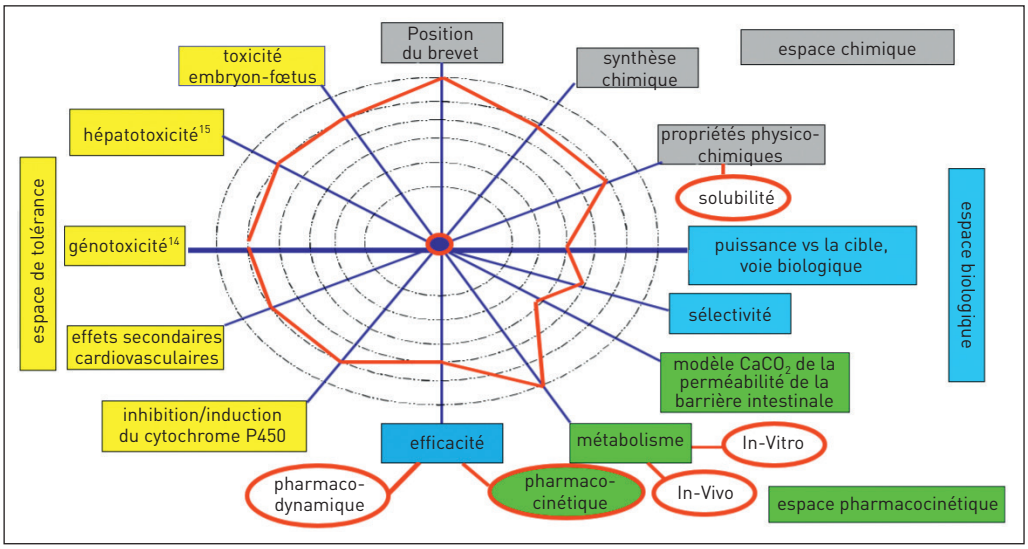


Figure 9

Spider diagramme représentant les propriétés d'un médicament à prendre en compte.

nom de « spider diagramme » ; ils illustrent bien que chaque propriété peut aller dans une direction différente pour un critère ou pour un autre, éventuellement même dans des directions opposées.

Pourtant, le cahier des charges impose de respecter l'ensemble de toutes les propriétés. Typiquement, aujourd'hui, un programme suit en recherche une quinzaine de paramètres au-delà de l'activité. De plus, l'impératif de la brevetabilité s'impose, ce qui rend le processus encore plus compliqué (l'espace chimique ou biologique pour manœuvrer est limité par la compétition).

Pour faire sentir le caractère complexe de la logique du domaine du *drug discovery*, il peut être utile de recourir à l'analogie et la différence avec les énigmes posées par le Rubik's cube, car on ne le manipule pas par une construction bloc par bloc analogue à la manipulation d'un Lego. Avec un Rubik's cube, quand on a résolu une face, il arrive que pour régler la suivante, il faille rechanger la dernière. Mais pour le *drug discovery*, il n'y a pas de solution *in fine* : on ne peut pas trouver sur Internet comment résoudre la dernière face, et il est possible qu'il n'y ait pas de solution pour régler toutes les faces en même temps.

2.3. Utiliser les données existantes pour trouver le candidat au développement

Le travail avec les données (les data) est au centre des techniques nouvelles du *drug*

14. Peut entraîner des dommages à l'ADN.
15. Toxique pour le foie.

discovery. Pour construire la base, on a besoin d'intégrer toutes les données qui existent en interne mais aussi d'inclure les données extérieures disponibles.

Notre méthodologie type suit le cycle DMTA : *Design, Make, Test and Analysis* (***Figure 10***). Le temps que l'on mettra entre le « Hit » et le candidat suit le nombre de cycles qu'on réalisera ; on a donc intérêt à ce que ce cycle d'optimisation soit le plus efficace possible.

Chez Sanofi, on dispose d'à peu près 400 millions de données internes, correspondant à des résultats positifs et négatifs. On les combine avec des données publiques, ce qui aboutit à peu près à 1 milliard de données utilisables pour supporter chaque projet.

À ce stade, il est indispensable de se retourner sur l'analyse dimensionnelle de grands nombres car elle cache un potentiel que nous avions occulté. La notion même d'intelligence artificielle est construite à partir de cela. Prenons des analogies, en commençant avec l'espace sidéral.

L'espace (en fait, le nombre) des molécules « drugables[16] », c'est théoriquement 10^{63} composés (***Figure 11***). Alors que le nombre d'étoiles dans l'univers c'est 10^{24}. D'un autre côté, le nombre de molécules qu'une compagnie pharmaceutique actuelle a synthétisées et stockées s'évalue entre 1 million et 10 millions, et le nombre de molécules qui ont été décrites publiquement (c'est-à-dire publiées dans « *chemical abstract*[17] »), c'est 1 milliard. Pour le nombre de molécules qui ont été décrites plus ou moins précisément dans l'ensemble des brevets

16. Molécules susceptibles d'être un médicament.
17. Molécules possédant un numéro CAS c'est-à-dire enregistrées publiquement.

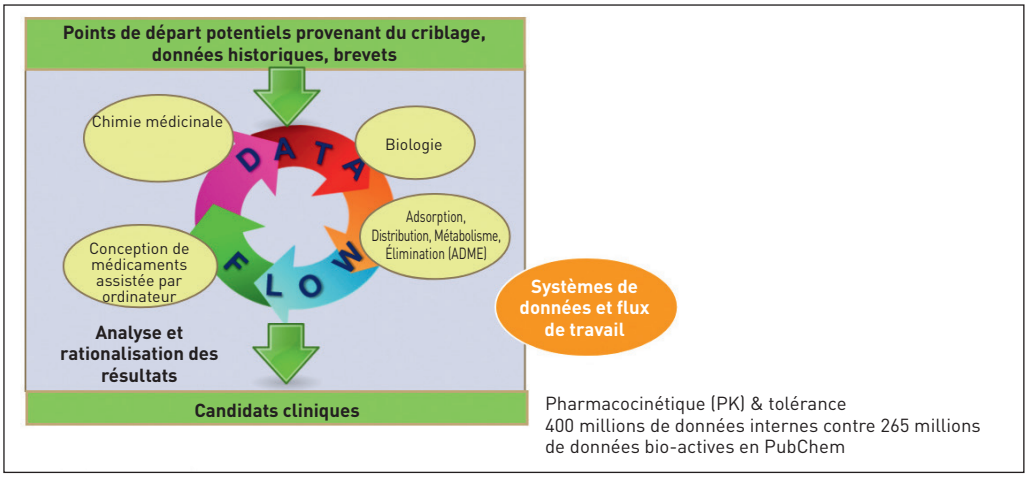


Figure 10

Cycle DMTA.

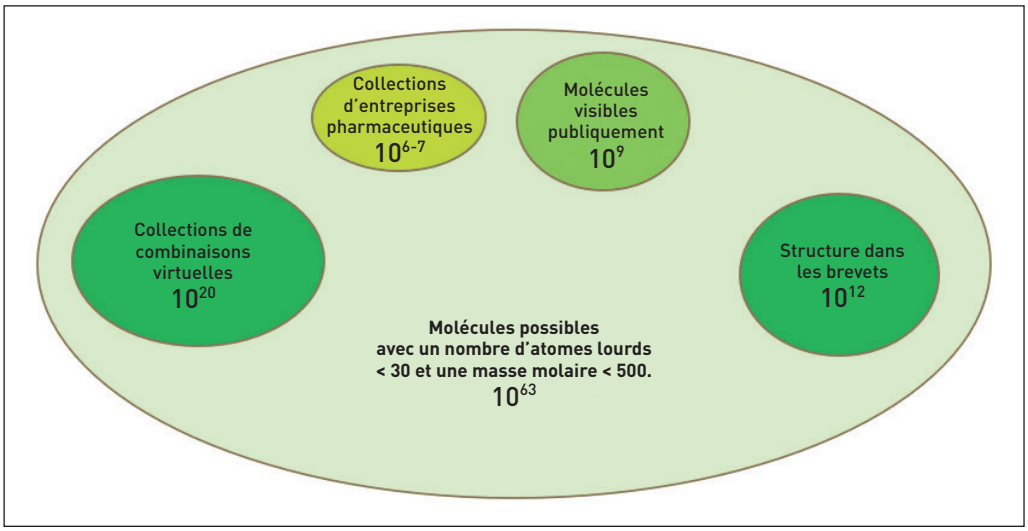


Figure 11

L'espace chimique des médicaments.

qui ont été déposés, on est à 1 000 milliards, ce qui est très loin de10^{63}.

Mais dans le monde des molécules et des bases de données correspondantes, la nouveauté c'est que maintenant il est proposé sur catalogues **des molécules virtuelles**. Ces molécules n'ont pas été faites, mais on sait que la chimie pourrait en faire la synthèse. Pour le faire vraiment, encore faudrait-il que les molécules s'avèrent positives dans un test *in silico*[18]. Aujourd'hui, les plus grandes bases de données virtuelles de molécules que l'on connaisse arrivent à 10^{20}, donc pratiquement à la dimension d'un espace de l'univers, mais c'est encore bien loin des 10^{63}. Cependant, on monte en gamme, en termes de nombre de molécules et de nombre de calculs envisageables : c'est cela la réalité d'aujourd'hui en termes de dimension des nombres.

3 L'intelligence artificielle au cœur de la recherche pharmaceutique

Depuis peu de temps, l'intelligence artificielle (AI), dans son sens le plus large, prend place au niveau du *drug discovery*. On distingue 3 types d'approches (***Figure 12***) : celles qui permettent de réaliser des prédictions de propriétés, celles qui permettent de faire du *screening* dans des espaces très larges (on envoie des sondes *in silico* dans ces espaces très larges de molécules virtuelles). Dans la troisième approche, on utilise des algorithmes de génération de molécules qui permettent de parcourir un chemin (le plus court possible) vers la cible.

18. Correspond à un test réalisé par ordinateur, par simulation

3.1. La classification des approches AI

Avant de détailler ces trois exemples, distinguons parmi nos différents modèles et nos différents outils (***Figure 13***), ceux qui sont basés sur les datas, comme le *machine learning*[19] », et ceux qui sont basés sur la physique, comme les calculs de dynamique et de *Free Energy Perturbation* (FEP). L'intérêt de l'intelligence artificielle est de mixer les deux, de l'analyse de data, jusqu'au calcul de dynamique[20] et vice versa. On ne traitera pas ici

19. Capacité de la machine à apprendre au cours du temps.

20. Simulation du comportement d'une structure en prenant en compte la notion de temps.

Prédiction de propriété

Criblage virtuel ultra-large

Processus itératif pour créer de nouveaux médicaments

Figure 12

Trois types d'approche dans l'intelligence artificielle pour la recherche pharmaceutique.

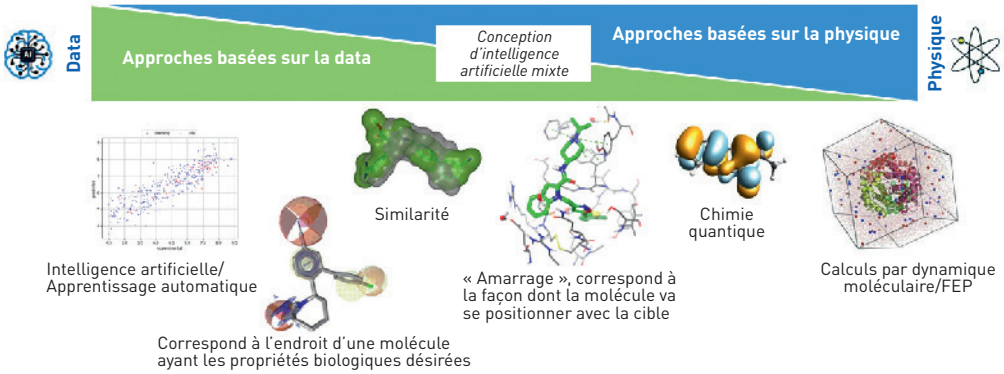


Figure 13

Mix des 2 approches : par la data et par la physique.

de DFT[21] mais de FEP[22] c'est-à-dire des calculs précis d'interactions d'une molécule sur son récepteur. Une étape suivante est de mixer ces résultats dans des modèles qui permettront de prédire des activités potentielles.

3.2. Les différentes applications

3.2.1 Le screening virtuel

Aujourd'hui, on sait faire virtuellement le « screening[23] » des grandes collections de molécules. Ensuite par l'emploi de filtres (***Figure 14***) de plus en plus précis, on réduit la liste sélectionnée pour converger vers un espace de molécules qui soit raisonnable à synthétiser. On n'est en effet pas capables de traiter de nombreuses molécules rapidement et précisément. On utilise la FEP seulement sur un ensemble réduit de molécules priorisées par d'autres méthodologies de calcul plus « grossières.

3.2.2. La prédiction de propriétés

Pour aller plus loin, on entraîne un set de data[24] internes et des data publiques qu'on peut utiliser, pour ensuite, par des modèles de *deep learning*[25], définir des descripteurs moléculaires (***Figure 15***). Ces descripteurs vont prédire des propriétés ou générer des corrélations

21. *Density Functional Theory* (théorie de la densité fonctionnelle) : permet de calculer l'énergie d'un système.
22. FEP : *Free Energy Perturbation*.
23. Screening : dépistage.

24. Datas : données.
25. *Deep learning* : apprentissage en profondeur.

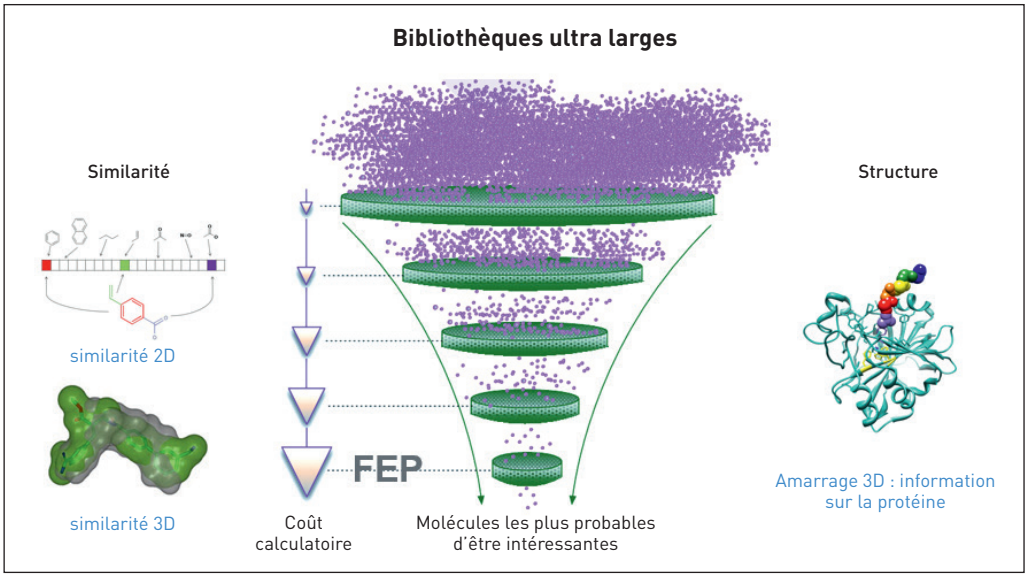


Figure 14

Le screening des molécules par l'intelligence artificielle.

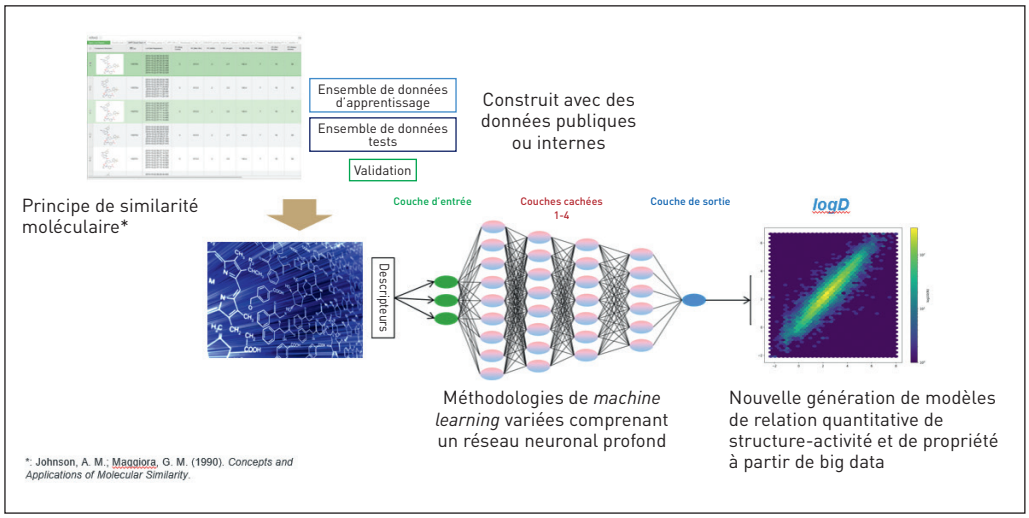


Figure 15

Prédiction de propriété par l'intelligence artificielle.

linéaires[26], sur une propriété, ici le logD[27], dans un grand espace de molécules.

3.2.3. L'exemple phare de DeepBlue *et l'idée de l'apprentissage*

Les deux premières approches citées plus haut sont, d'un certain sens, des optimisations, comme la « Micheline » qui est devenue un TGV, avec le machinateur. Avec l'intelligence artificielle, on fait en effet mieux et plus vite. La vraie différence est que l'on est capable de donner de nouvelles capacités à l'ordinateur. De nouveau, une analogie triviale : l'exemple des échecs avec Garry Kasparov. En 1996, l'ordinateur *DeepBlue* a été battu par Kasparov, mais un an après, Kasparov était battu par la nouvelle génération de *Deepblue* qui s'appelle *DeeperBlue* ; l'ordinateur avait appris les coups et a déstabilisé Kasparov. Il avait appris comment mieux jouer que le meilleur joueur du monde de l'époque.

3.3. L'intelligence artificielle et les modèles génératifs

Aujourd'hui en chimie, on sait apprendre à l'ordinateur à générer des molécules en lui indiquant ce qui est positif et ce qui est négatif sur la base des propriétés qui sont importantes pour le projet. Il est capable de réaliser des cycles d'apprentissage (***Figure 16***) sur la base de modèles prédictifs.

Ainsi, des algorithmes génératifs peuvent produire 20 000 structures chimiques nouvelles par heure ; on est

26. Exprime la notion de liaison entre 2 paramètres.
27. Le logD est la valeur du logP à un pH donné pour un composé d'un certain pKa ; il donne la mesure de la solubilité différentielle entre un solvant organique et l'eau.

alors loin de l'intuition « je fais cette molécule-là parce que je le sens bien ». Cette intuition d'ailleurs ne produit pas beaucoup de molécules *in fine* ; elle donne peut-être la bonne molécule occasionnellement, mais n'enrichit pas beaucoup d'idées.

Grâce aux modèles développés, on peut donc trier de nombreuses molécules chimiques proposées pour ne synthétiser ensuite que celles qui ont les meilleurs scores.

3.3.1. Collaboration avec Aqemia

Le *drug discovery*, tel qu'il vient d'être décrit, n'est pas fait chez Sanofi en interne uniquement, même si des algorithmes sont développés par des experts maison. Pour l'ensemble du travail, Sanofi s'appuie très largement sur des collaborations. On décrit particulièrement la collaboration avec la **société Aqemia** (***Figure 17***). C'est une startup qui a été créée par Maximilien Levesque, un ancien professeur de l'ENS.

À l'origine d'Aqemia, est le fait que Maximilien Levesque a « craqué une équation » qui permet aujourd'hui à ses calculs d'affinité d'être 10 000 fois plus rapides que ceux donnés par la meilleure des méthodes *in silico*. 10 000 fois plus rapides, ça permet d'explorer des espaces autrement inaccessibles, en particulier pour les calculs de FEP, très coûteux en temps.

3.3.2. Collaboration avec Exscientia

Pour terminer un dernier exemple de collaboration. Il s'agit de la **société Exscientia** qui est probablement celle qui est la plus avancée dans l'utilisation de l'intelligence artificielle en *drug discovery* (***Figure 18***). Début 2023, elle annonçait avoir généré une nouvelle molécule qui rentrait en essai clinique après onze mois de recherche, et après avoir synthétisé seulement 150 molécules. Au lieu de faire 4 000 molécules en quatre ans de recherche (en moyenne),

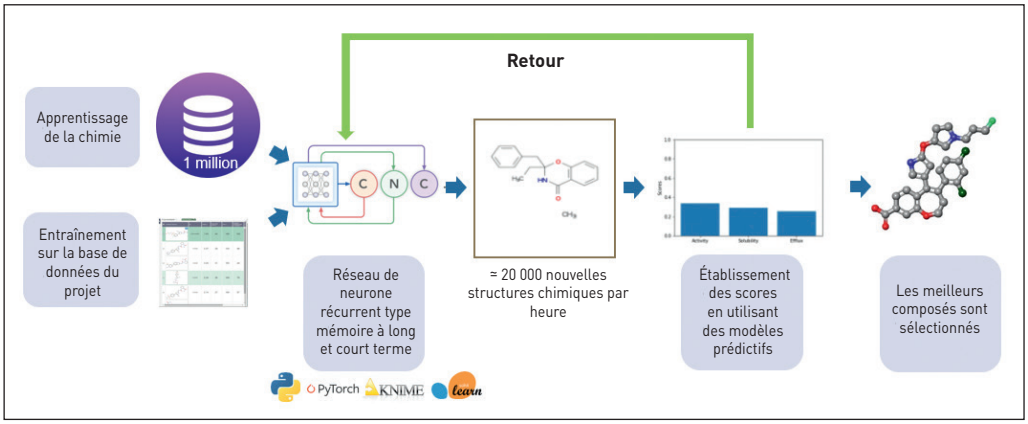


Figure 16

Le cycle d'apprentissage pour l'obtention des meilleures molécules.

l'avenir est peut-être à faire 100 à 200 molécules en une année avant d'aller en essai clinique, grâce à tous les outils de l'IA mentionnés dans ce chapitre.

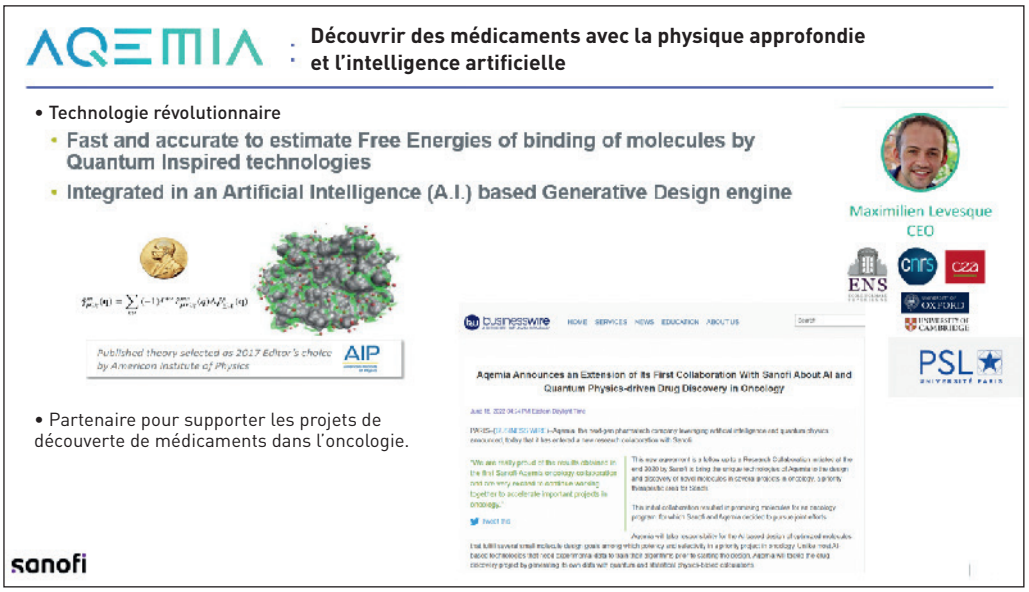


Figure 17

Aqemia.

Figure 18

Exscientia, collaboration avec Sanofi.

Conclusion
L'IA en chimie. Des perspectives stupéfiantes mais des risques apocalyptiques

Ce n'est pas un scoop, mais une réalité stupéfiante à connaître et à expliquer, comme il vient d'être fait dans ce chapitre avec l'exemple du *drug discovery* : **le développement de l'IA constitue pour la chimie une véritable explosion ; elle ouvre une nouvelle ère.**

On dit aujourd'hui que la connaissance de la chimie va doubler toutes les douze heures. À toute question dans le domaine, on peut dire « je répondrai demain parce que je serai deux fois plus intelligent ». Dans les années 2010, la connaissance doublait tous les ans, et jusqu'à 1900 tous les 100 ans.

C'est la raison pour laquelle l'intelligence artificielle est un incontournable, mais l'intelligence artificielle d'aujourd'hui doit aussi évoluer. La question des « biais » est posée : comment l'influence (voulue ou non voulue) des opérateurs humains peut-elle être contrôlée ? les orientations catastrophiques évitées ? Ces risques éternels et permanents deviennent redoutables devant des outils à puissance infinie comme l'IA.

Intelligence artificielle pour la science et l'industrie

Cédric VILLANI, *Université Lyon I et Institut des Hautes Études Scientifiques (Bures-sur-Yvette)*

Cédric Villani a reçu en 2010 la médaille Fields, qui est l'équivalent du prix Nobel pour les mathématiques, discipline où le prix Nobel est inexistant. Professeur, chercheur, passionné de vulgarisation scientifique, homme politique, il a été chargé en 2017 par le Premier ministre Édouard Philippe d'une mission parlementaire sur l'intelligence artificielle. En 2018, il a initié la conférence « AI for humanity », lors de laquelle il a présenté son rapport sur l'IA sous le titre « Donner un sens à l'IA ». En dehors de son activité universitaire, à Lyon et en Essonne, il est engagé dans de nombreux projets associatifs, entrepreneuriaux, publics, en particulier auprès d'une association qui s'appelle SISTEMIC dont le but est de promouvoir les carrières scientifiques et technologiques des jeunes filles.

Introduction : peut-on définir l'intelligence artificielle ?

L'intelligence artificielle dont il est question aujourd'hui n'a rien à voir avec ce qu'on appelait l'intelligence artificielle au milieu des années 1980. À l'époque, adolescent passionné de sciences, je découvrais l'intelligence artificielle dans les ouvrages de vulgarisation de Douglas Hofstadter[1]. C'était le moment où avait été traduite en français la biographie d'Andrew Hodges[2] sur Alan Turing et

1. Douglas Hofstadter : universitaire américain (né le 15 février 1945), travaillant dans les domaines des sciences cognitives, de l'informatique, de l'histoire et de la philosophie des sciences, la philosophie, la littérature comparée et la psychologie.
2. Mathématicien et écrivain britannique, auteur de *Alan Turing : l'Énigme de l'Intelligence*.

son rôle incroyable dans la Seconde Guerre mondiale. Et l'intelligence artificielle était à l'époque un domaine scientifique dans lequel les mots essentiels étaient « comprendre » et « fonctionnement du cerveau humain ». C'était en lien avec les sciences cognitives et avec l'idée de la compréhension et du savoir. La question « comment fonctionnons-nous ? et celle de la « conscience » étaient aussi beaucoup mises en avant. L'image très présente dans la culture populaire, encore présente aujourd'hui, était celle de HAL, le robot de *2001 l'Odyssée de l'espace* qui devient conscient.

L'intelligence artificielle aujourd'hui n'est pas du tout cela. Ce sont des algorithmes qui accomplissent des tâches déterminées, sans l'ambition de les comprendre. Ce sont des méthodes de calcul, et d'exploration dans des espaces de très grande dimension qui permettent d'accélérer une tâche fastidieuse. Cela ne ressemble pas à ce qu'on appelle d'habitude « comprendre ». Et la première chose à dire et à répéter à chaque fois sur l'intelligence artificielle, c'est qu'en l'état actuel des choses, ce n'est pas de l'intelligence. Du moins pas au sens où nous l'entendons en tant qu'êtres humains.

Ne cherchez pas une définition précise. Dans une rencontre toute récente organisée au laboratoire de Google à Paris, réunissant une vingtaine d'experts et de gens intéressés par le sujet de l'intelligence artificielle en général, le responsable en charge chez Google a commencé la soirée en disant : « Si on demandait à chaque personne autour de la table de donner sa définition de l'intelligence artificielle, il y aurait sans doute autant de définitions que de personnes présentes. »

Moi-même, quand j'étais responsable du rapport sur l'intelligence artificielle, en bon mathématicien, je prévoyais de commencer par un chapitre Définition ! Mais après un certain nombre d'auditions et de réflexions sur le sujet, j'ai vite compris qu'il ne fallait pas passer par cette étape**. Chercher la définition de l'intelligence artificielle serait un leurre : c'est un sujet qui se définit par les usages.**

1 L'impact de l'intelligence artificielle dans les sciences

1.1. En mathématiques, un impact du numérique plutôt que de l'intelligence artificielle

Parmi les jeunes scientifiques de ma promotion, tous et toutes ont vu leur carrière scientifique impactée massivement par la numérisation en général ; mais une partie seulement a vu l'intelligence artificielle débouler dans leur sujet. Certains en informatique bien sûr, mais aussi en physique, y ont trouvé leur chemin. Chez les mathématiciens, on peut dire que certes, certains d'entre nous se sont mis à faire de l'intelligence artificielle, au sens de développer des algorithmes, mais aucun n'a vu sa pratique de mathématicien – cherchant,

ciselant, classant, partageant des démonstrations – impactée de façon importante par le sujet.

Certes, nous avons eu une révolution numérique en mathématiques, mais cela a été quoi ?

Il y a eu les articles numérisés qui ont changé la façon d'échanger des données, les plateformes répertoriant les articles, les grandes bases de données, dont celle qui est tenue par *l'American Mathematical Society*, et les grands répositoires comme arXiv[3] qui ont servi à se partager l'information. Tout cela a eu un rôle majeur. Et avec le numérique, mais aussi les courriels, le volume a augmenté massivement !

Et puis il y a eu TeX, l'unique, l'incomparable ! TeX, le logiciel qui sert à tous les mathématiciens et mathématiciennes pour écrire et mettre en forme leurs articles mathématiques, soit directement, soit indirectement, avec une typographie qui est contrôlée. Cela nous a permis, à partir des années 1990, de regagner un domaine qui avait été complètement perdu, celui du contrôle de la clarté de l'écriture mathématique dans les publications : le contrôle des formules, le contrôle des beaux signes « intégrale » et des innombrables autres symboles qui nous sont chers, le contrôle des bons espacements, des parenthèses, tout ce qui fait qu'on pouvait composer nos ouvrages nous-mêmes sans être à la merci d'intermédiaires.

3. arXiv : archive ouverte d'articles dans des domaines scientifiques ou économiques.

Il y a encore eu, évidemment, l'essor des calculs massifs, des simulations numériques, qui ont joué un rôle majeur dans la résolution quantitative des équations différentielles ou variationnelles, dans la mise en œuvre du travail des ingénieurs, dans l'intuition des experts en physique mathématique et dans d'autres domaines mathématiques. Et dans certains cas des preuves assistées par ordinateur, depuis Appel-Haken (le théorème des quatre couleurs, revisité et corrigé récemment) jusqu'à la preuve de la conjecture de Kepler par Hales et ses collaborateurs.

Voilà certes des évolutions notables, mais tout cela n'est pas pour autant de l'intelligence artificielle au sens habituel, et l'impact de cette dernière sur la recherche mathématique est resté très limité.

1.2. L'impact de l'intelligence artificielle dans les autres sciences

Dans d'autres domaines en revanche, en particulier la biologie, la médecine, la chimie, il est devenu impossible d'ignorer l'impact de l'intelligence artificielle. Cette différence avec les mathématiques est avant tout une différence de nature de l'exploration et du savoir.

Certes, toutes les sciences sont en interaction, et je suis résolument favorable à ce que l'on aborde la culture scientifique comme un tout où mathématique, physique, biologie, chimie, et même les sciences humaines sont en interaction

permanente, formant un grand ensemble connecté. Mais dans sa pratique, il y a des singularités de la discipline mathématique. Par exemple, le profil de carrière : on a coutume de dire que les mathématiciens arrivent à maturité en moyenne plus jeunes que les autres scientifiques ; et d'ailleurs, c'est cohérent avec le fait que la médaille Fields soit strictement limitée aux personnes de moins de 40 ans. Un des facteurs derrière cette relative jeunesse, c'est qu'en mathématique les concepts comptent bien plus que les catalogues, et l'expérience joue un rôle moins important. Alors que dans d'autres disciplines (chimie, biologie), il est impossible de travailler sans catalogues, car Il y a beaucoup de propriétés que vous ne pouvez pas deviner, que vous ne pouvez pas inventer. En clair, en mathématique on est beaucoup moins dépendant de catalogues d'expériences.

Il y a d'autres différences sociologiques. La mathématique est une discipline moins hiérarchisée que la chimie ou la biologie ; c'est assez naturel en fait, car les disciplines qui réalisent de grandes expériences ont aussi besoin d'une organisation plus dirigiste… Il y a certainement d'autres facteurs, mais j'en reste là !

Et en tout cas, retenons que les disciplines scientifiques sont différentes dans leurs méthodes de travail et d'acquisition du savoir. Et cela se ressent en permanence. Dans la gouvernance d'une université, on observe en général que, lors des discussions, et quels que soient les thèmes, trois blocs se forment au sein des sciences « exactes » : le premier, les mathématiciens et informaticiens ; le deuxième, celui des physiciens ; le troisième, celui des chimistes, biologistes, géologues.

Ces disciplines différentes ont aussi un rapport différent à la notion d'apprentissage automatique. Les disciplines, dont la pratique repose beaucoup plus sur des concepts et sur une structuration du savoir comme en mathématiques ou en physique théorique, ont été beaucoup moins affectées par l'IA, et, en particulier, par l'IA statistique, qui est aujourd'hui la plus répandue. Cette approche statistique est d'ailleurs très prédominante (mais pas pour autant exclusive) dans les usages de biologie ou médecine.

Prenons un exemple de jeu mathématique comme celui du jeu de go, jeu défini par des règles mathématiques simples. On peut très bien imaginer un apprentissage par l'intelligence artificielle qui, à partir des règles, puisse deviner les propriétés. C'est ce que fait AlphaGo. Après AlphaGo qui utilisait les parties déjà jouées, il y a eu AlphaGo Zero qui ne partait de rien pour lancer l'apprentissage.

En matière de biologie, de chimie, ou de santé, on ne sait pas retrouver ou réinventer les règles à partir de zéro. On peut dire que c'est juste l'état de l'art, mais qu'un jour on y parviendra… On peut quand même attendre longtemps ! Imaginez qu'il faille retrouver toutes les propriétés chimiques à partir de l'équation de Schrödinger[4],

4. Équation fondamentale en physique quantique. En effet, sa résolution permet de connaître à tout instant la position d'un électron.

ou les mécanismes biologiques à partir des réactions de base en biologie. C'est plus dur en chimie qu'en physique, plus dur en biologie qu'en chimie, plus dur en médecine qu'en biologie. Plus le nombre de paramètres croît, plus la diversité et la personnalisation comptent, plus c'est difficile à aborder ou à deviner. Grand nombre de paramètres, grande variabilité, méconnaissance des mécanismes fondamentaux : tout cela est propice à l'IA !

Et la physique ? On sait que la mathématique et la physique ont partie liée dès 1600, quand commence la révolution scientifique. Encore aujourd'hui, on peut argumenter que la moitié des concepts qu'on utilise en mathématiques ont leur origine quelque part en physique.

On ne peut absolument pas en dire autant de l'interface mathématiques/biologie. Cela a été un échec presque constant des mathématiciens que d'essayer de comprendre le caractère extrêmement hétérogène, et varié, des propriétés des molécules, des cellules, et de la biologie en général. C'est une raison pour laquelle la mathématique, a eu tant de mal à se faire sa place en biologie. Mais pour ces mêmes raisons, l'IA, et plus généralement l'algorithmique, sont en train de changer cette donne. L'excellent ouvrage de Ian Stewart sur les avancées mathématiques en biologie dresse un bon panorama.

Avant que l'on parle tant d'intelligence artificielle, d'autres méthodes de calcul jouaient ce rôle. Il n'y a pas si longtemps, c'était la grande mode des méthodes MCMC[5], les méthodes de Monte Carlo Markov Chain, pour résoudre toutes sortes de problèmes de grande dimensionnalité. Encore auparavant ont été utilisées les méthodes d'algorithmes génétiques ou la foisonnante variété de techniques utilisées en construction d'arbre phylogénétique[6], par exemple par Joe Felsenstein.

Et maintenant, c'est l'IA qui est devenu un domaine incontournable dans les disciplines comme la médecine, la biologie, et maintenant la chimie !

2 La place de l'homme face à l'emploi de l'intelligence artificielle

La démarche scientifique n'est certainement pas de faire du calcul. Si vous dites à un scientifique qu'il est un calculateur, il n'appréciera pas, un mathématicien pas plus que les autres. La science, c'est l'organisation en concepts. Mais si on laisse à la machine et à l'algorithme le soin de faire des calculs, et à l'humain qui l'utilise, le soin de tenter d'en extraire les concepts, pourquoi pas ?

Cela signifie qu'aujourd'hui, l'intelligence artificielle n'est pas une révolution méthodologique de **faire de la science, mais plutôt une révolution dans la manière de l'appréhender. Elle**

5. Équation fondamentale en physique quantique. En effet, sa résolution permet de connaître à tout instant la position d'un électron.
6. Figure utilisée en biologie. Les différentes espèces vivantes y sont classées sous forme d'arbre, dans lequel chaque nœud représente une caractéristique commune aux espèces rattachées (par exemple squelette osseux ou ailes).

est notamment utile lorsque les simulations se font en utilisant de nombreuses bases de données, ou encore quand les calculs sont trop fastidieux pour être faits à la main.

Aucun nouveau domaine ou théorème mathématique n'a encore été découvert par l'intelligence artificielle. Et même, aucun domaine scientifique n'a été inauguré par l'IA. Certains contempteurs considèrent brutalement que l'IA n'est pas de la science, mais juste une démarche empirique, une régression par rapport aux grands principes de la science. Mais si le rôle de **l'intelligence artificielle est** justement d'être en soutien à l'analyse et à la démarche méthodologique et conceptuelle, cela veut dire que c'**est un outil qui peut soutenir la science et lui permettre d'explorer de nouvelles pistes.**

Au niveau humain, on peut faire la part des outils d'intelligence artificielle qui sont là pour vous aider à vous développer, à vous rendre plus intelligents, à aiguiser votre curiosité, et ceux qui sont là pour vous remplacer et qui au contraire vous abêtissent : ceux qui écrivent les phrases à votre place ; ceux qui vous suggèrent automatiquement où vous voudriez partir en vacances, au lieu de chercher par vous-mêmes, ceux qui vous emmènent automatiquement du point A au point B sans que vous ayez le moindre effort musculaire à faire et ainsi de suite.

Le philosophe Ivan Illich[7], célèbre entre autres pour son étude des rapports entre humains et technologie, avait l'habitude de distinguer les outils conviviaux, ceux qui viennent en soutien pour vous développer en tant qu'humains avec votre intelligence, votre corps, votre bien-être, et ceux qui, au contraire, parce qu'ils sont trop perfectionnés, trop exigeants, ou trop addictifs, viennent vous remplacer, vous empêcher de vous développer, ou vous mettre en situation de dépendance, et finalement vous amoindrir. Pour l'instant, on peut dire que l'intelligence artificielle dans les sciences a été un outil convivial qui a laissé la primauté à l'humain, en ce sens qu'elle a été là pour aider des personnes à continuer à développer les concepts et à continuer à faire de la science selon les grands principes que l'on connaît.

3 Les acteurs de l'intelligence artificielle, son impact politique

À la question : « Qui est concerné par l'intelligence artificielle ? », j'ai répondu : « Certaines sciences plus que d'autres, au niveau de l'industrie certaines plus que d'autres aussi, et au niveau des technologies certaines plus que d'autres. »

Mais au-delà de la sphère scientifique et technologique, la grande évolution de l'intelligence artificielle sur les 30 dernières années, c'est qu'aujourd'hui cela concerne et intéresse la société. Il y a 30 ans, aucun gouvernement n'avait quoi que ce soit à faire de l'intelligence artificielle. C'était un domaine qui était

7. Ivan Illich (1926-2002) : philosophe, référence importante en écologie politique et critique de la société industrielle.

réservé à des experts, et même à une poignée d'experts, c'était beaucoup moins interdisciplinaire qu'actuellement. Aujourd'hui, cela concerne tout le monde, en particulier les entreprises, les clients, et même dans la mesure où cela a été soutenu et tiré par les usages et par les applications, dans bien des cas cela concerne au moins autant les usagers que les développeurs.

Dans la mesure où cela s'applique d'autant plus qu'il y a beaucoup de paramètres et que c'est imprédictible, évidemment quand on arrive dans la sphère des applications et des questions humaines c'est là que l'impact est le plus important. On ne voit guère à quoi ChatGPT pourra, dans un avenir proche, être utile sur des sujets scientifiques. En revanche, ChatGPT peut impacter la façon dont on communique, les corrections des copies, de cours de sciences politiques ou de qui que ce soit, dans la mesure surtout où cela s'est invité sur le domaine de la culture générale. ChatGPT peut vous aider à préparer une plaidoirie, un argumentaire, un dossier de demande de subvention, un discours généraliste, un article de journal… Les impacts potentiels en termes de communication, en termes de messages, et en termes de débat public sont considérables. L'impact dans la relation au travail est aussi potentiellement considérable. ChatGPT n'est pas programmé pour dire le vrai mais pour dire le plausible. Cela a ses limitations, ses impacts, ses avantages et ses dangers. C'est bien sûr un objet hautement politique, non seulement par son potentiel sur la société, mais aussi parce que la politique se nourrit de paroles et de discours !

Pour toutes ces raisons, on pourrait se dire que la catégorie qui devrait de prime abord, parmi les citoyens, être la plus branchée sur l'intelligence artificielle et se tenir au courant, est ce qu'on appelle les politiques. Ce n'est pas ce qui est observé. Les députés, de la gauche ou de la droite classiques, n'ont pas bien conscience des enjeux de l'IA et ne suivent pas la chose de près comme j'ai pu le constater lors de mon rapport.

Il ne faut pas croire que l'IA nous amènera spontanément vers un monde meilleur. « L'intelligence artificielle est le plus fabuleux concentrateur d'inégalités », m'a dit un jour le ministre d'un État asiatique. Mais justement, on peut s'en emparer pour conjurer cette prédisposition. L'intelligence artificielle pourrait être un sujet de redistributions des richesses pour la gauche classique. Cela pourrait aussi être un enjeu de la droite classique dans un contexte où on parle énormément des questions de souveraineté. Mais en fait, dans la pratique, ce ne fut pas le cas lors de la rédaction du rapport sur l'intelligence artificielle. L'intérêt fut plus grand, mais néanmoins limité, dans le camp de la majorité présidentielle au sens large, du fait qu'il y a eu un renouvellement de la classe politique sur cette case de l'échiquier en 2017 et du fait qu'il y ait eu un appel à la société civile. C'est seulement maintenant que la classe politique commence à s'y mettre.

Je suis convaincu que si Marx était de ce monde, il aurait travaillé sur l'intelligence artificielle et y aurait vu un enjeu énorme, de la même façon qu'il avait travaillé à l'époque sur le calcul différentiel et tenté d'appliquer les méthodes de calcul différentiel[8] aux questions politiques. D'ailleurs, je tiens à le rappeler dans cette belle maison de la Chimie, Marx se passionnait aussi pour les enjeux liés à la chimie, lisait les chimistes de son époque et en particulier Von Liebig, convaincu que c'était une des clés majeures de la transformation politique.

À l'époque du rapport, le thème de l'IA en chimie n'était pas vraiment développé. Pourquoi, 5 ans plus tard, est-il si fort, comme on peut le voir dans cet ouvrage ? Il n'y a pas eu, dans cet intervalle, de révolution dans les concepts majeurs. En revanche, il y a, sans aucun doute, un accroissement de la puissance de l'IA. Des nouvelles techniques ont été découvertes. Les techniques, depuis plus de dix ans, tirent le sujet, les théoriciens et les experts courent après les expérimentateurs. Par exemple pour ChatGPT, ce sont les techniques dites des *transformers*[9]. Et il y a un peu plus de dix ans, le grand choc ce fut la réalisation du pouvoir des réseaux de neurones, alors que les experts les plus reconnus en théorie algorithmique pensaient que c'était une voie sans issue. Et les quelques-uns qui passaient pour des originaux avec leur marotte, Yann Le Cun, Yoshua Bengio, Goeffrey Hinton[10] et quelques autres, sont devenus des héros. Quelques années plus tard, ils ont reçu les plus hautes distinctions en matière d'informatique.

Cela a été pareil en IA pour les développements des grands modèles et les ChatGPT. Saut quantitatif et qualitatif ! En août 2022, Bertrand Braunschweig, expert très respecté en IA, présentait dans une conférence publique quelques failles qui mettaient les meilleurs agents conversationnels en erreur. Quelques mois plus tard, arrive ChatGPT en version publique et il ne faisait plus ces erreurs. L'humilité est de mise !

Dans la partie de cet ouvrage concernant la mise en place de l'enseignement de l'intelligence artificielle dans le domaine de la chimie, il est dit avec raison que c'est aux chimistes de se lancer avec des rudiments de programmation dans les expérimentations, assistés et aidés par l'intelligence artificielle parce que l'opération inverse est tellement plus difficile. On peut apprendre à un chimiste à programmer et à lancer un réseau de neurones ; on peut même apprendre à un élève très motivé de terminale à programmer son petit réseau de neurones, alors qu'il sera plus difficile de lui apprendre un cours de chimie organique,

8. Domaine mathématique s'intéressant aux variations infinitésimales de fonctions.

9. Les transformers sont un type de réseau de neurones développé en 2017, particulièrement adapté dans le traitement automatique du langage.

10. Yann Le Cun, Yoshua Bengio et Goeffrey Hinton ont reçu le prix Turing en 2019 pour leur travail sur le *machine learning*.

et on ne lui apprendra jamais une résolution d'équation de Navier-Stockes[11], ou tout ce genre de choses qui sont bien plus installées dans la science et qui demandent un savoir-métier bien plus solide.

Quelles évolutions à venir pour la discipline ? Il faut d'une part rester très solide sur les fondamentaux de la discipline et ne pas croire qu'on n'aura plus à les enseigner ; mais il faut utiliser autant que possible et à chaque fois que c'est possible les outils d'exploration et familiariser les élèves dès maintenant au fait que, dans la démarche expérimentale, il y a de nouveaux outils qui changent la donne et qu'on ne peut pas ignorer. Il faut donc travailler le plus possible, en mode interdisciplinaire, aussi bien dans les formations que dans la recherche. Il faut que les experts parlent aux experts, échangent leurs résultats, échangent les données, discutent pour savoir ce qui peut se faire, et que les équipes travaillent en partenariat pendant un certain nombre d'années.

Quels blocages sont à craindre ?

Ce qui est le plus dur à traiter systématiquement avec l'intelligence artificielle, ce ne sont pas les questions scientifiques et technologiques, ce sont la plupart du temps les questions politiques et les questions sociétales.

En 2018, la mobilité autonome était un sujet majeur dans l'industrie. Aujourd'hui on en parle beaucoup moins, simplement parce qu'on s'est rendu compte que c'est beaucoup plus dur que ce que l'on croyait. Les ambitions d'innovation en matière d'automatisme et de mobilité sont bien en-dessous de ce qui était espéré il y a encore quelques années. On s'est rendu compte que, sur le plan technique, c'est beaucoup plus difficile que ce qu'on croyait de faire une voiture autonome. Mais dans bien d'autres dossiers, les blocages ont systématiquement été du côté politique, en particulier pour le dossier santé sur lequel j'ai beaucoup travaillé, car le rapport était à l'origine de la naissance de la plateforme nationale de données de santé pour alimenter justement la recherche par l'intelligence artificielle. Les blocages portaient par exemple sur des questions de débat de souveraineté publique, sur des questions de consentement et diffusion des informations, ou sur des questions des rapports de pouvoirs entre la CNIL[12] et les équipes de scientifiques qui opèrent. De sorte que, cinq ans après le rapport, on n'a toujours pas une plateforme de données de santé qui soit opérationnelle, donc on est très loin du niveau qu'on pouvait espérer.

Ce blocage au niveau politique est extrêmement important de façon générale. Quand vous êtes amenés à construire des plateformes de données, vous avez trois défis à résoudre.

Le premier défi est scientifique et technique : les questions de format réseau,

11. Équation fondamentale de la mécanique des fluides, permettant de décrire leur mouvement dans l'espace et le temps.

12. CNIL : Commission nationale de l'informatique et des libertés.

et d'interopérabilité, par exemple. C'est difficile mais on le résout.

Le deuxième défi est d'ordre éthique, légal, et réglementaire. Il porte sur toutes les questions liées à ce qu'on a le droit de faire ou pas. Il est de difficulté à peu près égale et on peut le résoudre.

Le troisième défi concerne le partage des pouvoirs. Qui a le droit de tenir la plateforme ? Qui en a l'accès ? Qui peut la modifier ? Ce point est beaucoup plus difficile à traiter et est même parfois indémêlable. C'est le cas de certains programmes de plateforme de données de santé, qui dépendent de plusieurs autorités.

Finalement, on a beau croire qu'on part d'un dossier technique, à la fin on arrive toujours sur des dossiers qui sont humains, des dossiers de confiance et avec un rôle du politique qui devient d'autant plus important.

Est-ce que la même chose peut se reproduire en chimie ? À coup sûr les obstacles seront moins importants qu'en médecine, domaine hanté par les peurs. Des questions de concentration de pouvoir, de secret industriel, de pollution peuvent bien évidemment se produire, mais on peut parier que les blocages politiques et humains seront moins importants qu'en médecine ou biologie.

Conclusion
L'impact de l'Intelligence artificielle sur la société

Les exemples traités dans les différents chapitres de cet ouvrage montrent l'impact de l'intelligence artificielle sur la chimie, avec des cas dans lesquels les bénéfices à venir pour la société sont incontestables. L'un des exemples utilisés comme illustration est l'importance des nouveaux polymères pour avoir des pales des éoliennes, à la fois résistantes et robustes. Et on se dit bravo la chimie des polymères, c'est exceptionnel. C'est vrai ! Mais attention ! On s'est placé ici sous l'angle de la création de nouveaux polymères utiles… Sauf que les polymères, à travers les plastiques d'emballage, sont aussi un des problèmes d'environnement majeurs dans le monde et des rapports sérieux

sur ce domaine, comme celui fait à l'OPECST[13], concluent que la seule vraie solution est de réduire drastiquement le volume de production des plastiques. On se retrouve alors face à une équation politique insoluble. Ou plutôt, une équation qui ne se résout pas avec une solution technique mais avec un besoin de volonté ! Et le besoin de prendre à bras le corps les conséquences dans les habitudes, les usages, l'économie.

Le remède est on ne peut plus « low-tech » : juste arrêter la production de ces plastiques d'emballage, en tout cas la restreindre drastiquement, c'est la cause numéro 1 de pollution par les plastiques dans les grands continents et l'océan. Chimiquement et scientifiquement trivial, mais politiquement très difficile, cela entraînerait, entre autres, des reconversions liées aux fermetures d'usines. Il faudra bien le faire, c'est ma conviction d'homme politique ! C'est indispensable pour notre santé, pour notre planète, pour nos enfants… Mais cela demandera un vrai courage ! Et si l'on se concentre sur les solutions de haute technologie – avec de l'intelligence artificielle, de nouveaux matériaux, etc. – pour faire oublier qu'il y a aussi des remèdes de basse technologie à appliquer, c'est du « technoblanchiment », un adversaire redoutable au progrès !

Prenez un autre sujet comme l'agriculture : pour améliorer le diagnostic et la mise en œuvre dans ce domaine, on peut utiliser des outils, pas forcément à base d'IA mais en général numériques. Il y a une communauté française active sur ce sujet qui s'appelle « La ferme digitale ». Ils nous disent, exemples à l'appui, que dans les bons cas un agriculteur peut gagner grâce aux outils numériques jusqu'à 5 000 €

13. OPECST : Office Parlementaire d'Evaluation des Choix Scientifiques et Technologiques.

par an et économiser 100 h de travail sur cette même période. Certes ce n'est pas une révolution, sur une année de labeur d'agriculteur mal payé et sans vacances, c'est réellement non négligeable. Il n'empêche que ce n'est pas cela qui changera complètement la donne pour rendre le métier attractif. Et plus généralement, cela ne va pas résoudre les principaux problèmes de l'agriculture, le premier étant l'usage massif des pesticides, herbicides et fongicides et en premier lieu des néonicotinoïdes, les plus ravageurs et destructeurs de biodiversité que le monde ait jamais connus. Ce n'est pas non plus le numérique ou l'intelligence artificielle qui dispenseront de regarder le fait que personne ne sait faire une transition écologique à production de viande constante. Il ne faut pas croire que l'intelligence artificielle ou quoi que ce soit de numérique nous dispensera de regarder ce problème en face : sans évolution de la consommation, en tous cas dans nos assiettes, il n'y aura juste pas de transition écologique. Il n'y aura pas non plus de transition écologique si on ne recrute pas massivement des agriculteurs et pour cela il nous faut une revalorisation du métier. Sortie des pesticides, sortie des énergies fossiles, amélioration de notre alimentation, revalorisation du métier d'agriculteur. Sur tous ces problèmes, on n'est pas sur des questions scientifiques dures, aucune solution technologique, aucun progrès de l'IA ne semble prêt à nous aider, on est sur des questions de société.

C'est bien plus général que le sujet de la chimie, et c'est à bien garder en tête : les techniques mathématiques et l'intelligence artificielle viennent s'inviter dans beaucoup de sujets, mais il faut être très attentif à ce que cela ne distraie pas aussi bien le scientifique que le politique du fait qu'il y a des enjeux et des défis comportementaux énormes, très *low-tech*. Et, si

au niveau des investissements et des directions de recherche on se fourvoie en allant regarder du côté qui avance, plutôt que du côté qui est difficile à résoudre, on est exactement dans la situation du fou qui va chercher ses clefs sous le lampadaire alors qu'il sait bien qu'il les oubliées là-bas dans l'ombre. Mais il n'a juste pas envie d'aller dans l'ombre parce que cela le dérangerait… Tâchons de ne pas être ce fou là.