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

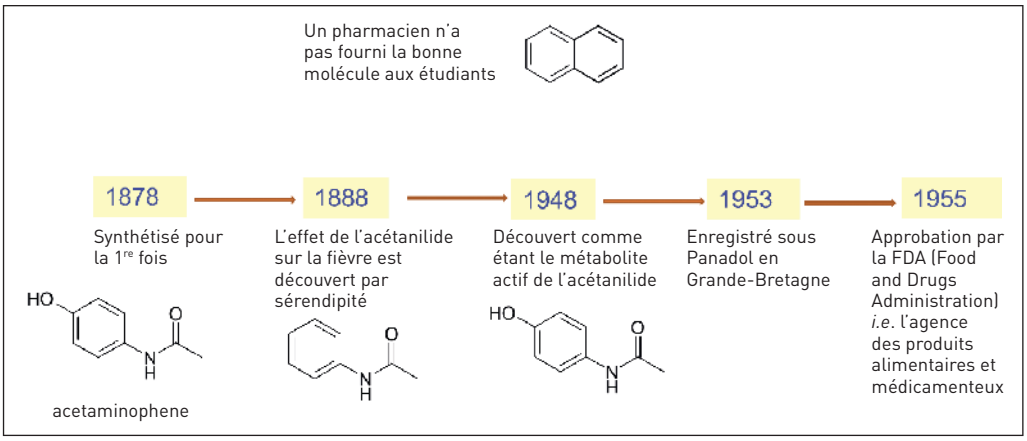


Figure 1

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.

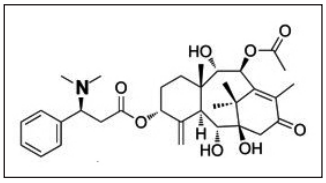


Figure 2

Molécule de taxine, isolée en 1856.

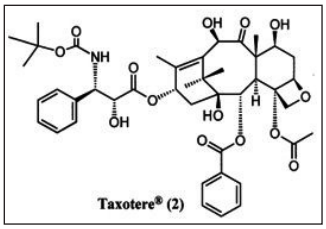


Figure 3

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.

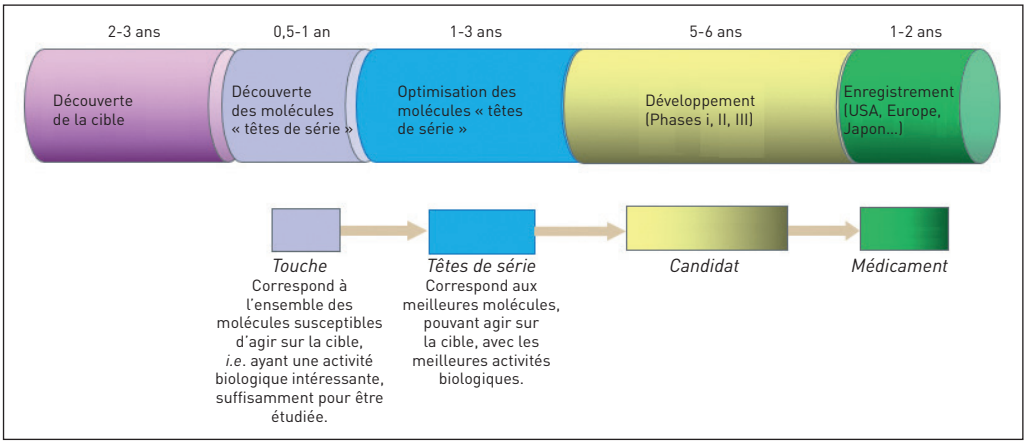


Figure 4

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

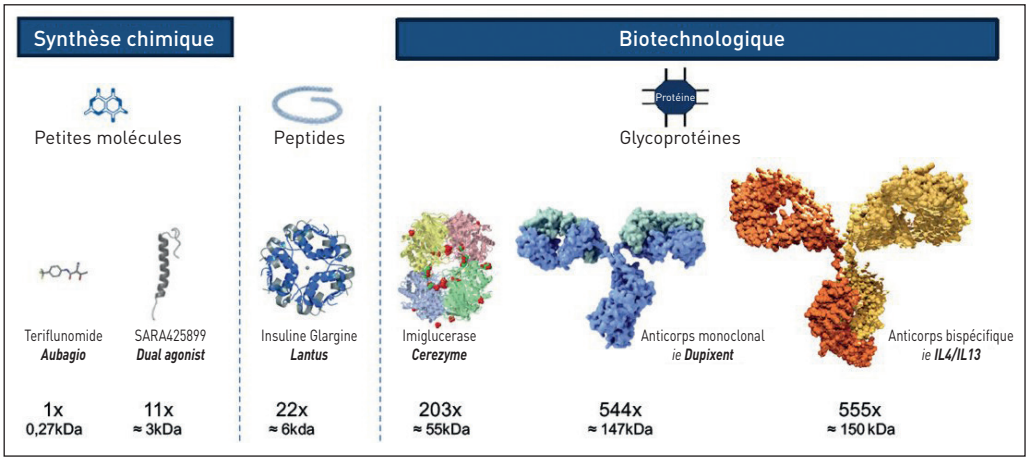


Figure 5

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.

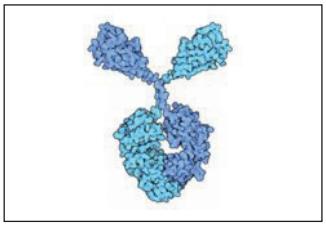


Figure 6

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.

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.

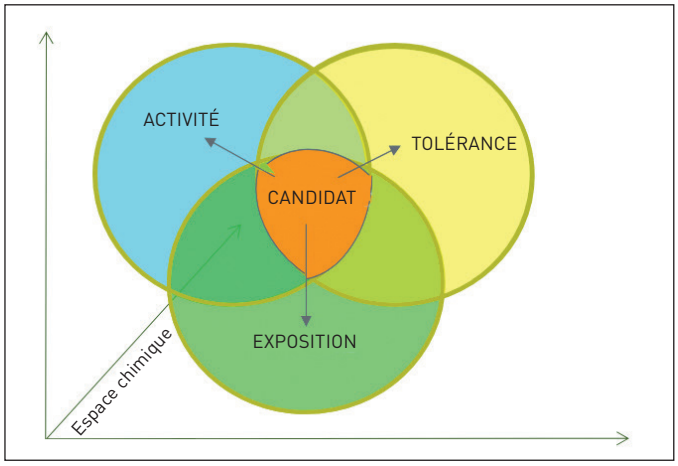


Figure 7

Diagramme correspondant au compromis pour obtenir le médicament désiré.

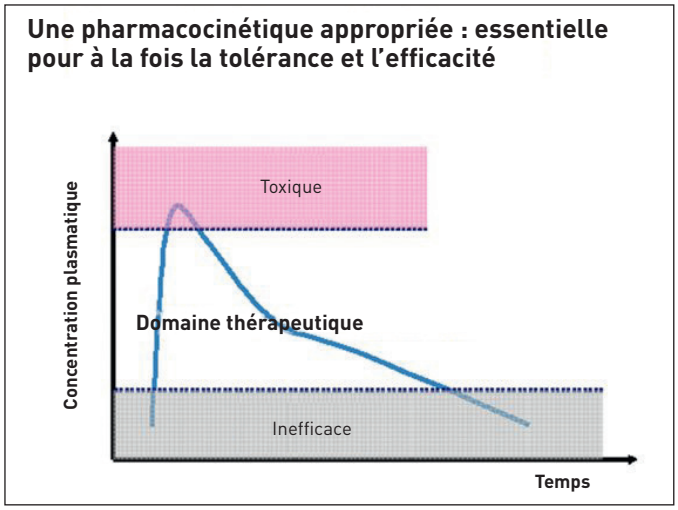


Figure 8

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

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.

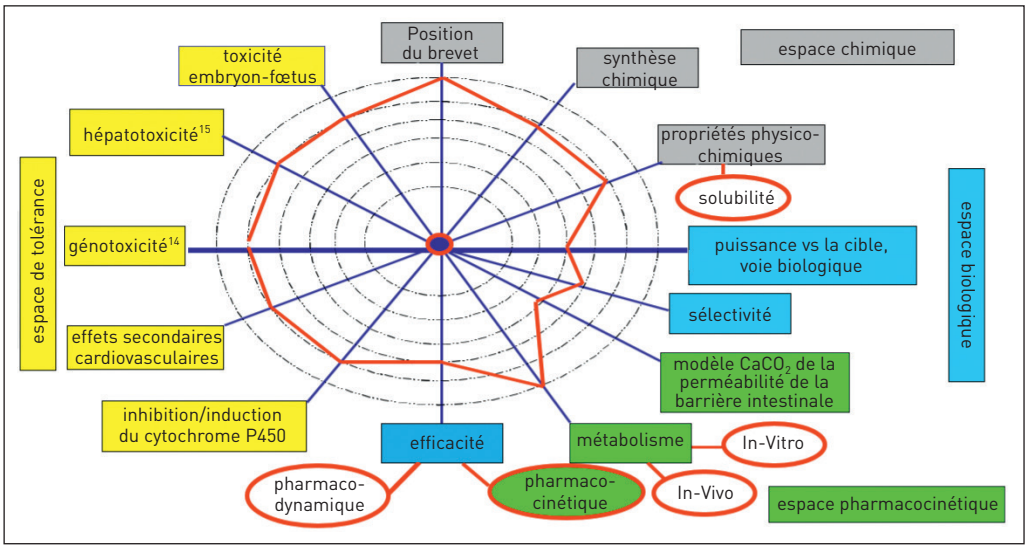


Figure 9

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.

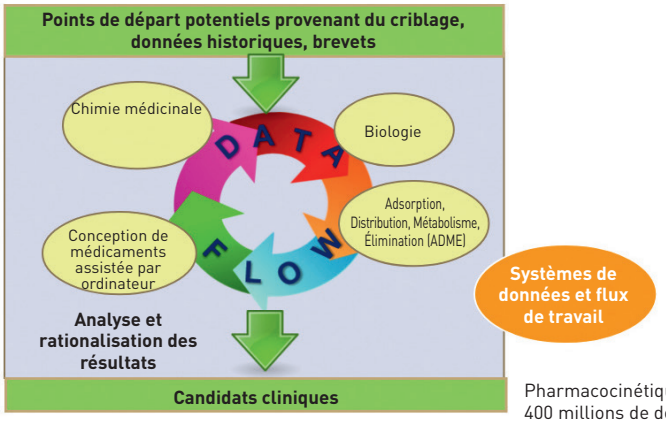
Notre méthodologie type suit le cycle DMTA : *Design, Make, Test and Analysis* (***Figure 10***). Le temps que l'on mettra entre le « Hit » et le candidat suit le nombre de cycles qu'on réalisera ; on a donc intérêt à ce que ce cycle d'optimisation soit le plus efficace possible.

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.



Pharmacocinétique (PK) & tolérance
400 millions de données internes contre 265 millions de données bio-actives en PubChem

Figure 10

Cycle DMTA.

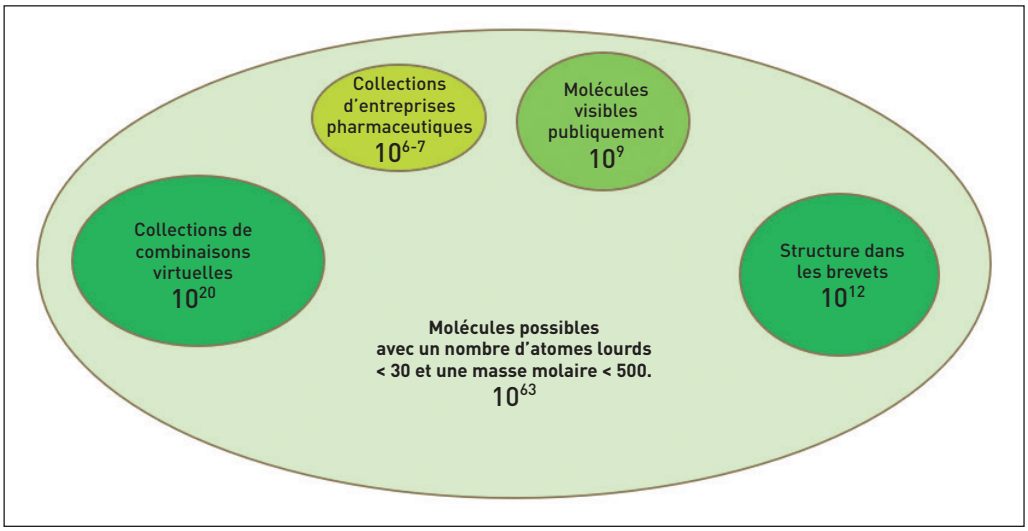


Figure 11

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.

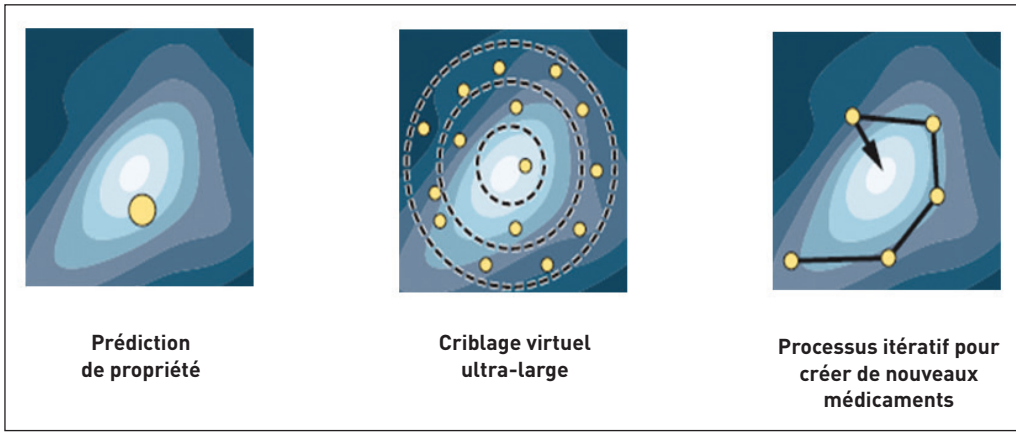


Figure 12

Trois types d'approche dans l'intelligence artificielle pour la recherche pharmaceutique.

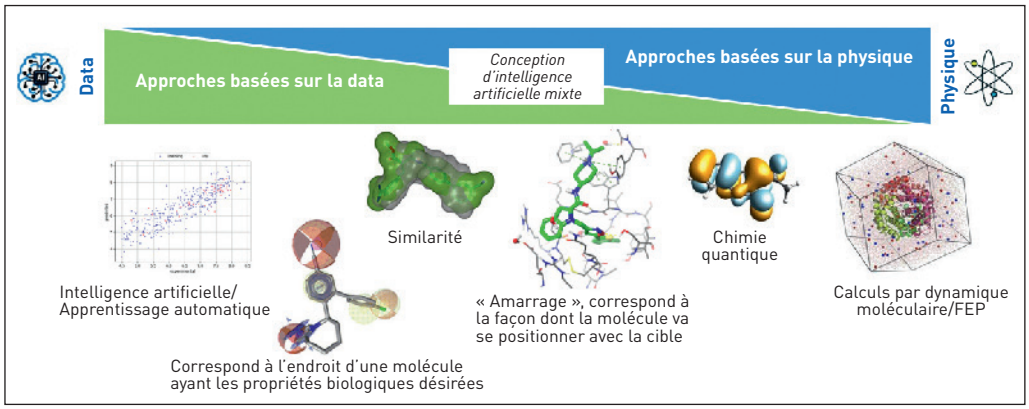


Figure 13

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.

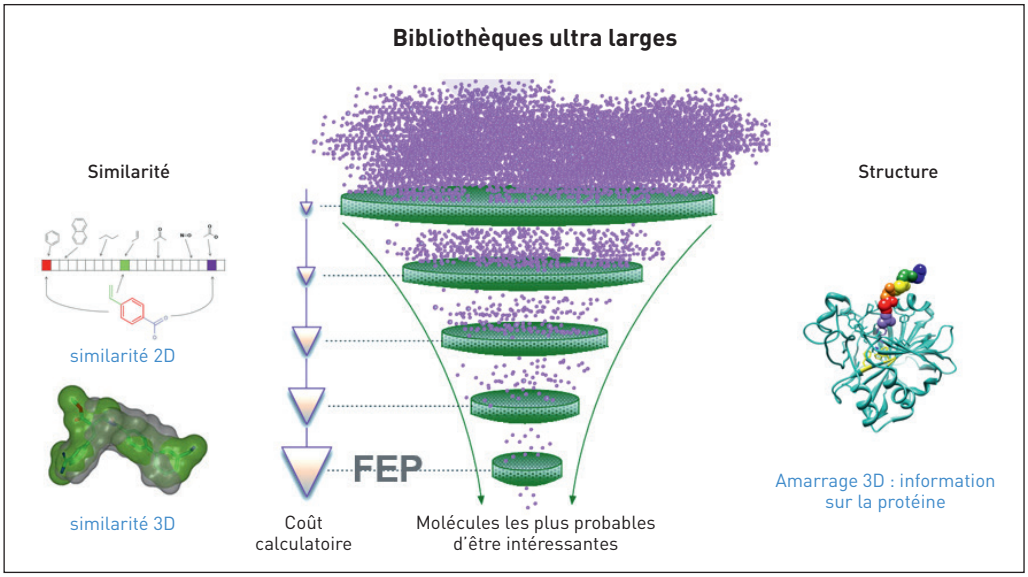


Figure 14

Le screening des molécules par l'intelligence artificielle.

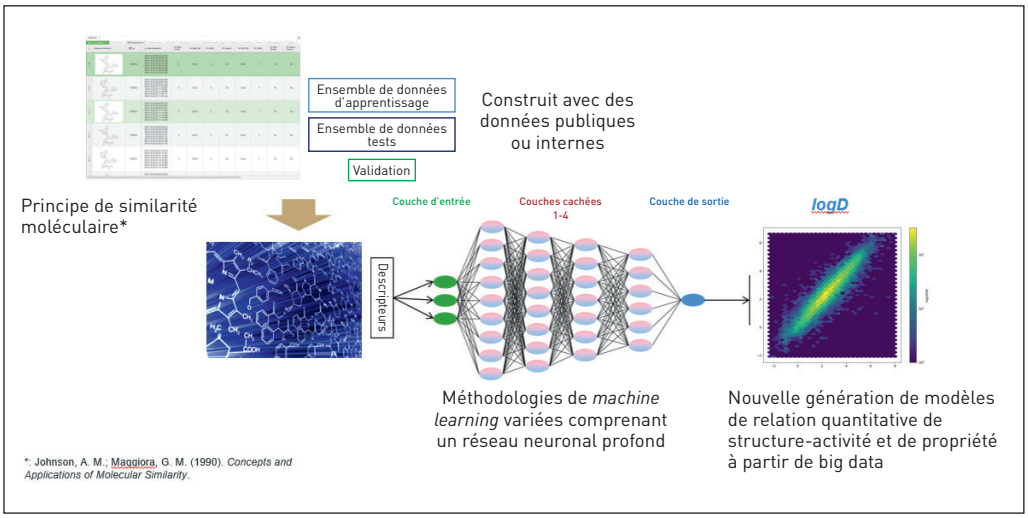


Figure 15

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),

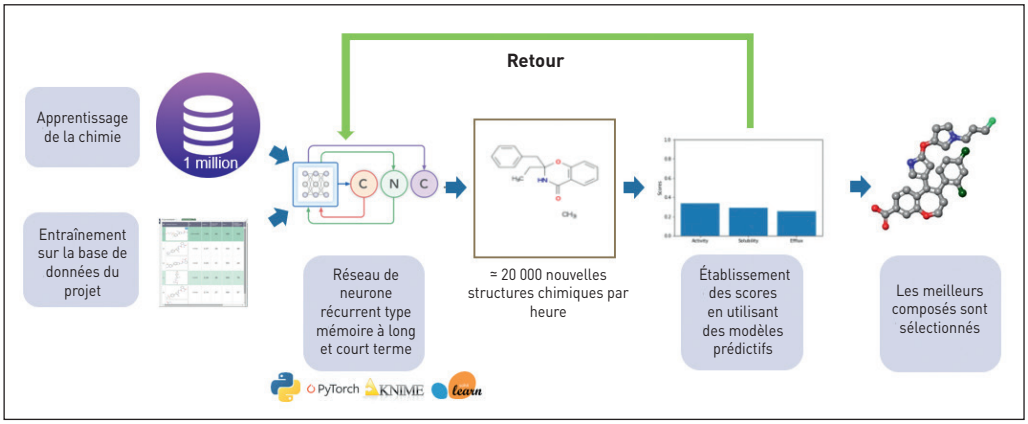


Figure 16

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.

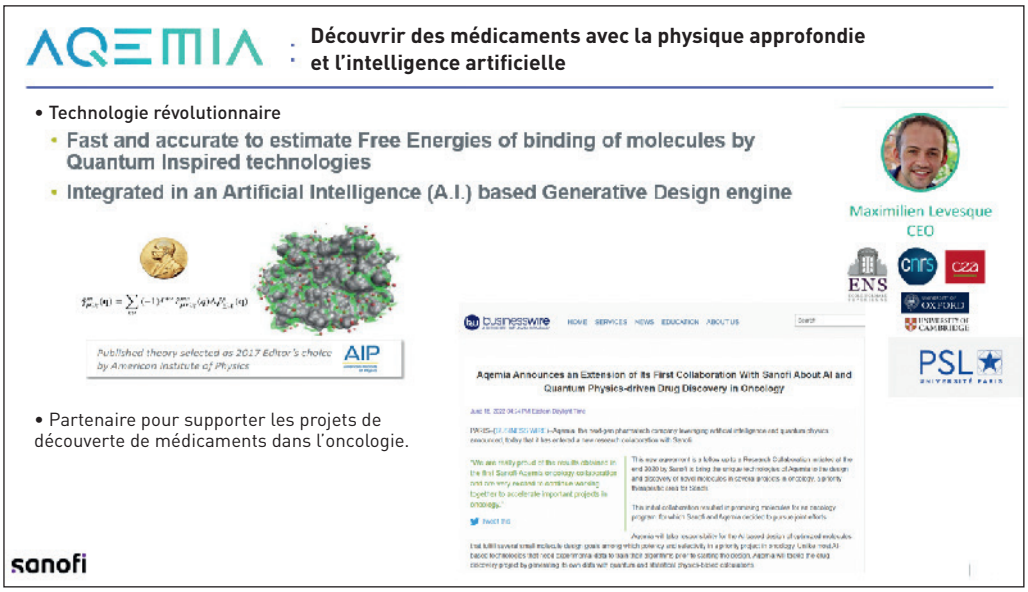


Figure 17

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.