Les apports du machine learning dans la synthèse de molécules médicamenteuses

Hubert Gorny

UFR DE PHARMACIE
Université Clermont Auvergne

UNIVERSITÉ CLERMONT AUVERGNE
UFR DE PHARMACIE

Année : 2020 N°

THÈSE D'EXERCICE
pour le
DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE

Présentée et soutenue publiquement
le 13 Novembre 2020
par
Hubert GORNY

Les apports du machine learning dans la synthèse de molécules médicamenteuses

Directeur de thèse : Mr Olivier CHAVIGNON

Jury

Président : Mr Olivier CHAVIGNON	Professeur des universités, UFR Pharmacie de Clermont-Ferrand (63)
Membres : Mme Magali VIVIER	Maître de conférences, Docteur en pharmacie UFR Pharmacie de Clermont-Ferrand (63)
Mme Sophie LEVESQUE	Docteur en pharmacie, Centre Jean Perrin, Clermont-Ferrand (63)
Mme Marion TEMPIER	Docteur en pharmacie, Centre Jean Perrin, Clermont-Ferrand (63)

UFR DE PHARMACIE
Université Clermont Auvergne

UNIVERSITÉ CLERMONT AUVERGNE
UFR DE PHARMACIE

Année : 2020 N°

THÈSE D'EXERCICE
pour le
DIPLÔME D'ÉTAT DE DOCTEUR EN PHARMACIE

Présentée et soutenue publiquement
le 13 Novembre 2020
par
Hubert GORNY

Les apports du machine learning dans la synthèse de molécules médicamenteuses

Directeur de thèse : Mr Olivier CHAVIGNON

Jury

Président : Mr Olivier CHAVIGNON	Professeur des universités, UFR Pharmacie de Clermont-Ferrand (63)
Membres : Mme Magali VIVIER	Maître de conférences, Docteur en pharmacie UFR Pharmacie de Clermont-Ferrand (63)
Mme Sophie LEVESQUE	Docteur en pharmacie, Centre Jean Perrin, Clermont-Ferrand (63)
Mme Marion TEMPIER	Docteur en pharmacie, Centre Jean Perrin, Clermont-Ferrand (63)

Remerciements

Je tiens tout d’abord à remercier Mr Olivier CHAVIGNON pour avoir accepté d’encadrer mon travail et de présider mon jury et ce malgré son emploi du temps pris par la gestion de la première année des études de santé très compliquée aux vues des circonstances actuelles. Je voulais aussi vous remercier car vous êtes indubitablement à la base de mon intérêt pour la chimie organique que m’a mené à réaliser une thèse de recherche dans ce domaine.

Je remercie ensuite Mme Magali VIVIER, pour avoir accepté de faire partie de mon jury. Je vous remercie aussi pour avoir suivi mon parcours et m’avoir permis d’arriver où j’en suis, en encadrant mon tout premier stage de recherche et pour m’avoir suivi lors de ma formation d’ingénieur chimiste.

J’adresse évidemment mes remerciements à Mme Sophie LEVESQUE et Mme Marion TEMPIER pour faire partie de mon jury mais aussi pour avoir fait de mon stage hospitalier une superbe expérience professionnelle et personnelle. Vous avez fait des mois passés à vos côtés à la radiopharmacie de Jean Perrin des moments parmi les plus agréables de ma scolarité.

Merci à mes parents qui m’ont depuis toujours donné l’amour de la science et des études sans lesquelles je n’aurais pu réaliser mon parcours. C’est grâce à vous que j’ai pu me focaliser sur mes études sans jamais avoir à m’inquiéter de quoi que ce soit d’autre et aussi sans aucun doute grâce à vous pour de nombreuses autres raisons que j’en suis là aujourd’hui.
Merci aussi à mon frère toujours à l’écoute en cas de problèmes ou autres questionnements auxquels je peux être confronté.

Je remercie Julie, tout d’abord pour les nombreuses relectures évitant ainsi les excès de fautes d’orthographes et d’erreurs de frappe. Mais surtout pour me supporter depuis aussi longtemps, pour m’avoir canalisé et rendu plus sérieux et plus globalement pour l’ensemble de ton œuvre durant ces sept années.

Je remercie enfin mes amis, que vous soyez amis d’enfance ou rencontrés lors des études. Sachez que sans vous je n’aurais pas autant apprécié toutes ces années. Je ne donnerai pas de noms car j’ai tendance à en oublier mais je n’oublierai pas tous les moments de joie et de rire que vous m’avez offerts et que j’espère vous continuerez à m’offrir.

Table des matières

Liste des figures et des tableaux

Tableaux :

Figures :

Liste des abréviations

AMM : Autorisation de mise sur le marché

CAA : Cartes Auto-Adaptatives

FDA : Food and Drug Administration

GAINS : Genome-associated Interaction Networks

GANs : Generative Adversarial Networks

GENTRL : Generative Tensorial Reinforcement Learning

GWAS : Genome-Wide Association Study

HIPAA : Health Insurance Portability and Accountability Act

IA : Intelligence Artificielle

NDD : Network-driven Drug Discovery

NLP : Natural Language Processing

PA : Principe Actif

QSAR : Quantitative Structure Activity Relationship = relations structures activité

R&D : Recherche et Développement

RMN : Résonance Magnétique Nucléaire

RNN : Recurrent Neural Network

Introduction

La création d'un médicament est un processus long et fastidieux. A partir de la découverte de la cible thérapeutique et/ou de la molécule d'intérêt il peut se passer des dizaines d'années et des centaines de milliers d'euros peuvent être dépensés avant d'arriver au médicament.

La découverte de médicaments se faisait historiquement par l'observation ou la sérendipité. Au fur et à mesure des années, la recherche est devenue de plus en plus structurée. La création de molécules thérapeutiques se faisant par amélioration de composés existants, extraction de molécules naturelles et tests sur animaux en vue d'observer des effets d'intérêt.

L'informatique a ensuite joué un rôle important dans le processus de découverte de molécules d'intérêt. Les stratégies de recherches ont ainsi utilisé ce type d'outils afin de réaliser des criblages à haut débit de molécules intéressantes. Les systèmes informatiques ont aussi permis d'avoir une meilleure vision des cibles biologiques affinant ainsi la recherche sur celles-ci notamment grâce à des techniques de docking ou de visualisation 3D.

Le processus de R&D (recherche et développement) s'est donc modernisé et est devenu plus précis. Cela laisse penser qu'il est plus facile de découvrir de nouveaux médicaments de nos jours. Or, ce n'est pas la cas, bien au contraire. La création de nouveaux médicaments est aujourd'hui moins rentable que jamais. Cela est dû à une augmentation des coûts et une diminution des revenus. L'une des principales causes est la difficulté à trouver de nouvelles molécules d'intérêt ce qui a amené à un constat amer qui porte le nom de « loi de Eroom » qui sera expliqué ultérieurement.

Nous expliquerons ainsi dans ce document ce qui a conduit à cet état de la recherche médicamenteuse mais aussi ce qui pourrait la relancer. Depuis quelques années, une technologie particulière est devenue à la mode. Ou plutôt est redevenue à la mode. Il s'agit du machine learning (apprentissage automatique) qui permet d'utiliser des IA (intelligences artificielles) afin de réaliser des tâches jusqu'alors accomplies par l'homme.

Dans un monde où de plus en plus de données sont générées, il devient impossible pour un cerveau humain de suivre le rythme, même sur un sujet précis. Il devient donc important de résoudre ce problème et l'utilisation du machine learning semble être une solution. De plus, le

traitement de données n’est pas la seule option qu’offre le machine learning. Comme cela sera décrit dans ce manuscrit, l’utilisation d’IA peut prendre un grand nombre de formes dans la création de nouveaux médicaments « le drug design ».

Un focus sera fait sur les techniques et sur les entreprises d’IA impliquées dans le drug design. Dans cette catégorie, les techniques des entreprises spécialisées dans le « de novo drug design », c’est-à-dire la création de molécules à partir de zéro, seront développées. Nous verrons notamment l’utilisation d’un type particulier de système de machine learning, les modèles génératifs spécialisés dans la génération de structures moléculaires.

Il sera donc décrit dans ce texte un bref historique de la création de remède suivi d’un état de l’art de la recherche avec une description des techniques *in silico* modernes. Viendra ensuite une description de la loi de Eroom relative au cul-de-sac dans lequel se trouve actuellement l’industrie pharmaceutique. Enfin, nous verrons des moyens semblant futuristes (mais actuels) qui peuvent être la solution pour se sortir de cette loi de Eroom ainsi que des idées permettant l’intégration de ce type de technologies dans le monde de la recherche actuelle.

1. Les méthodes de découverte de nouveaux médicaments

1.1. Généralités

La recherche de traitements ne date pas de notre époque moderne. La notion de remède existe depuis des millénaires. Il est possible de citer les tablettes sumériennes de Nippur où l'on retrouve la première pharmacopée qui date de -2100 avant J-C. L'art de soigner était principalement réservé aux religieux et résultait de démarches basées sur l'observation (du comportement des animaux ou des effets de certaines plantes lors de l'ingestion etc.). Les premiers traitements étaient majoritairement d'origine végétale mais avec tout de même certains minéraux ou parties extraites d'animaux.(1)

Il est possible de citer ensuite Hippocrate, le père de la médecine en Grèce en –460 avant J-C qui a lui établit une pharmacopée, le « Corpus Hippocraticum » basé sur l'expérience. C'est-à-dire que les remèdes étaient identifiés en fonction de symptômes observés. Une vraie logique était instaurée derrière chaque traitement. On y retrouve aussi la description des premières formes pharmaceutiques comme entre autres des infusions et des onguents, mais aussi des suppositoires et des pilules.(2)

Il est ensuite important d'évoquer Galien, le père de la pharmacie galénique à Rome vers 130 après J-C. Il fonde ses principes sur la théorie des humeurs très utilisée par Hippocrate et il établit qu'il faut soigner les tempéraments (obtenus lors d'un déséquilibre des 4 humeurs) par leurs opposés. C'est l'apparition de l'allopathie «c'est par les contraires que l'on soigne les contraires» qui domine toujours aujourd'hui la façon de soigner les maladies.(3)

Vient ensuite la contribution de la médecine Arabe avec Avicenne en 980, auteur du Canon de la médecine dont une partie est consacrée aux préparations pharmaceutiques. On y retrouve la création de nouveaux remèdes avec des formes pharmaceutiques originales préparées avec des récipients adaptés, de nouveaux types d'instruments (alambics) et selon des mesures et des pesées précises.(4)

Au milieux du seizième siècle, on retrouve Paracelse, médecin suisse à qui l'on doit l'analyse pertinente « tout est poison rien n'est poison c'est la dose qui fait le poison ». Il est le premier à avoir recherché un médicament spécifique pour chaque maladie. Il est aussi à l'origine de la notion de principe actif (PA) et fonde la chimie pharmaceutique qui sont tous deux au centre de la

recherche actuelle. Il a cependant commis quelques erreurs comme renier la théorie des contraires au profit de la théorie des semblables ou théorie des signatures se basant sur des observations telles que « les tiges de bambous peuvent permettre un redressement de la colonne car sont un enchaînement de nœuds ».(5) (6)

La recherche de nouveaux traitements a ensuite été majoritairement basée sur la théorie des contraires tout en suivant malgré-tous les idées de Paracelse sur les PA et la chimie pharmaceutique. C'est-à-dire que l'on s'est mis à chercher des principes actifs ayant un effet inverse à celui présent chez le malade. Cela notamment à l'aide de la chimie médicinale qui a pu évoluer grâce à l'émergence de nouvelles technologies et de nouvelles connaissances au cours des siècles.

Un principe actif peut être découvert selon différents processus. Cela peut être par hasard à partir de données empiriques comme pour la trinitrine découverte par l'observation de céphalées des ouvriers travaillant à la fabrication de la dynamite ou par sérendipité comme la pénicilline. Mais l'approche actuelle est plus rationnelle. Elle se base sur des observations et des données existantes qu'elle va exploiter pour pouvoir créer un nouveau médicament. Par exemple commencer par trouver un composé ayant une action sur une cellule ou sur un organisme, puis améliorer cet effet et réduire la toxicité pour en faire un médicament et enfin associer une cible biologique à cet effet. C'est ce que l'on appelle la « forward pharmacology » (pharmacologie prospective) ou pharmacologie traditionnelle qui est en opposition avec une méthode plus récente, la « reverse pharmacology » (pharmacologie inversée). Dans cette dernière approche, l'idée est d'identifier une cible, caractériser sa structure et son rôle afin de pouvoir créer des médicaments la prenant pour cible.(7)

Pour ces deux méthodes, il faut d'abord identifier un premier composé intéressant que l'on appelle le « Hit ». Cela peut se faire entre autre par observation d'un effet d'une molécule connue, par criblage d'une bibliothèque de composés sur des cellules ou des organismes ou bien par criblage virtuel sur la structure de la cible lorsque celle-ci est connue. Mais le Hit ne va pas directement donner un médicament. Lors des tests préliminaires, il existe de nombreux Hits présentant une activité. Mais celle-ci n'est pas forcément suffisante et le composé n'a pas forcément les caractéristiques d'un bon médicament. Un tri est donc fait pour sélectionner les plus intéressants. Les structures similaires sont rassemblées et les plus aptes à devenir des médicaments (en fonction de la brevetabilité, la synthétisabilité, l'efficacité etc.) sont gardées. C'est l'étape du « Hit to Lead ».(8)

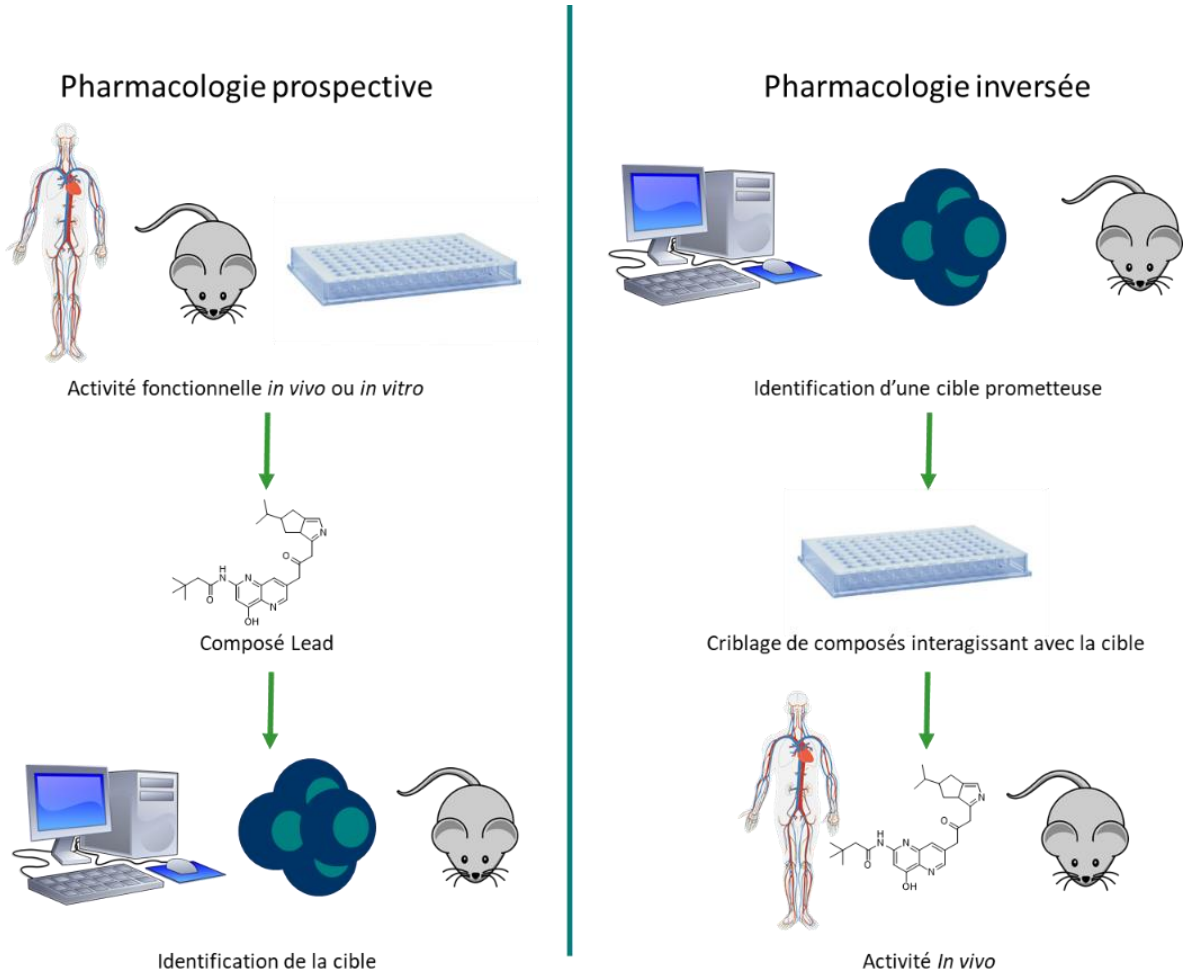

Figure 1: schéma explicatif comparant la pharmacologie prospective "traditionnelle" à gauche et la pharmacologie inverse plus moderne à droite

Ensuite, il faut s'occuper des Leads (des composés pouvant être amenés au stade médicament). Il va falloir réussir à améliorer leur activité et leur biodisponibilité tout en réduisant leur toxicité. C'est la phase de « Lead optimization ». Les Leads subissent de nombreuses modifications qui sans changer toute la structure pour ne pas perdre l'effet désiré vont permettre de jouer sur tel ou tel aspect de la molécule pour en faire un composé à tester in vitro puis in vivo. Ce travail est un véritable casse-tête pour le chimiste médicinal car maximiser un caractère va en dégrader un autre. Cette étape concentre 50% des coûts de la recherche représentant elle-même 40% des dépenses de la R&D. (9) Les améliorations sont le plus souvent le résultat de ce que l'on appelle des études de relation structure activité. C'est-à-dire que l'on va étudier l'effet de telle ou telle modification de la structure du Lead sur tel ou tel effet du composé jusqu'à arriver au composé « optimal » qui va subir les tests nécessaire en vue de devenir un médicament.(10)

Toutes ces étapes ont fait l'objet d'amélioration lors de ces dernières décennies. L'une des avancées majeure a été l'utilisation d'ordinateurs permettant de faciliter et d'accélérer le processus grâce à de nouvelles méthodes. C'est ce que l'on appelle les techniques de recherche « *in silico* ».

1.2.Les méthodes de recherches *in silico* (11)

Depuis les années 1990, nous avons assisté à un changement dans les stratégies de recherche de nouveaux médicaments. La norme est passée d'un criblage *in vivo* assez lent suivi d'une chimie d'optimisation à un screening haut débit de grandes bibliothèques de composés sur des cibles définies. La première méthode est lente et chère tandis que la seconde est rapide et économique. Cette transition s'est notamment faite grâce à l'apparition de nouvelles techniques inhérentes à l'utilisation d'ordinateurs dans le système de recherche.

La création de médicament à l'aide d'ordinateurs s'appelle le « *in silico* drug design ». Cela regroupe un grand nombre de techniques qui peuvent s'étendre sur toutes les phases de la R&D, des tests préliminaires aux tests cliniques. Parmi elles, on en retrouve un grand nombre spécialisé dans la recherche exploratoire, qui nous intéresse ici. Il est possible de citer entre autres (11):

- **La modélisation d'homologie :** Cela permet de générer la structure 3D d'une cible protéique à partir de sa séquence d'acide aminé en la comparant à des protéines déjà existantes. Il est ainsi possible de visualiser une cible biologique même si sa structure 3D n'a pas été élucidée par RMN (résonance magnétique nucléaire) ou par cristallographie, ce qui est très utile dans les phases préliminaires de la recherche pour la caractérisation de la cible.

- **Criblage virtuel à haut débit :** cette technique permet d'évaluer la capacité de nombreux composés d'une banque de données (ZINC database par exemple) à lier un site spécifique d'une cible, une protéine le plus souvent. Cette méthode est très utile lorsque l'on a la structure de la cible afin d'obtenir rapidement et facilement une grande quantité de Hits.

- **La cartographie 3D de pharmacophore :** permet d'analyser différents composés agissant sur une cible et d'évaluer les pharmacophores. C'est-à-dire les similarités structurales permettant d'avoir une action sur la cible. Cette étape est importante dans la sélection et l'optimisation du lead.

- **Le docking moléculaire :** permet de visualiser les orientations favorables de deux molécules lorsqu'elles sont liées ensemble. Cela permet de visualiser par exemple comment un médicament se lie à sa cible. Il est possible de classer les ligands selon leurs affinités (utile pour le Hit to Lead) et d'optimiser un ligand pour augmenter son intéraction avec la cible (Lead optimisation).

- **L'analyse conformationnelle :** est une technique où l'on analyse les configurations énergétiques minimales de certaines molécules à l'aide de diverses techniques de calcul pour comparer le site d'un récepteur moléculaire d'une autre molécule afin de calculer sa conformation 3D la plus satisfaisante. Cela permet par exemple de savoir comment se comporte le site de liaison d'une protéine cible pour ensuite réaliser du docking dessus par exemple.

- **La relation quantitative structure à activité** : (QSAR = quantitative structure activity relationship) : ces techniques sont utilisées pour décrire le lien entre la structure d'un composé et ses propriétés. Au début, ces propriétés comme l'hydrophobicité, l'encombrement stérique, les effets électroniques etc, étaient obtenues de façon empirique mais sont maintenant déterminées par informatique. Ce type de données est fondamental lors du lead optimisation.

Un focus a été fait sur ces technologies car elles peuvent être très adaptées à l'utilisation de machine learning. Il existe cependant beaucoup d'autres méthodes comme la simulation de dynamique moléculaire, ou l'utilisation des puces à ADN pour ne citer qu'elles.

Les techniques *in silico* sont d'une grande aide dans la création de nouveaux médicaments, elles permettent à la fois d'identifier et de visualiser les cibles thérapeutiques, de tester de grands nombres de composés et enfin d'améliorer et d'évaluer ces composés pour créer le meilleur médicament le plus rapidement possible. Or, comme nous allons le voir, ces avancées ne sont pas suffisantes et à moins d'un changement important dans les plus brefs délais, l'investissement dans la R&D ne sera plus un investissement rentable .

2. L’état de la recherche de molécules médicamenteuses

2.1. Les chiffres

Lorsque l’on parle de recherche et développement, deux problématiques sont récurrentes, le temps et les coûts. Si l’informatique et l’apparition de techniques de recherche *in silico* ont permis une amélioration du processus de recherche, elles n’ont pas pu répondre à ces problématiques.

Comme le montre le schéma suivant, la création d’un nouveau médicament est un parcours long (environ douze ans entre l’innovation et l’administration au malade) et fastidieux car très peu de réussite. On considère en effet que dix milles molécules intéressantes sont criblées en cinq ans pour à terme en avoir cent qui seront brevetées puis testées en clinique. De ces cent molécules, dix seulement réussiront les tests et pourront être testées sur l’homme en clinique. A la fin de ces tests cliniques soit dix ans de recherche, un seul composé , sur dix mille de départ pourra devenir un médicament.

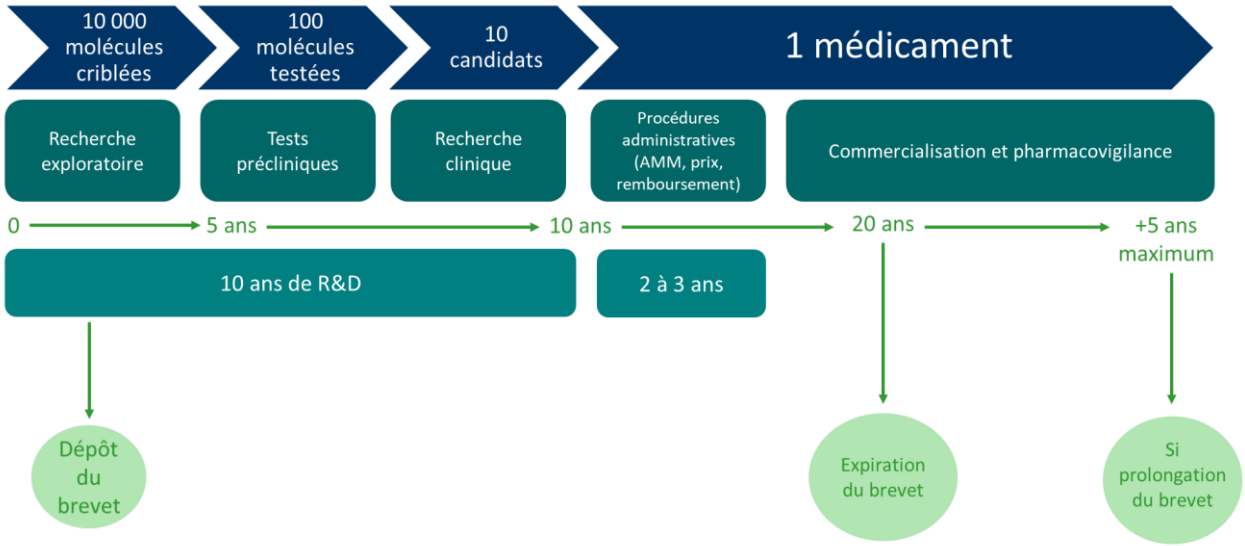

Figure 2: frise décrivant les étapes de création d'un nouveau médicament, le temps associé et le nombre de molécules présentes à chaque étape (12)

Une étude de 2012 a estimé que la mise au point d’un médicament coûtait environ de neuf cents millions à un milliard et demi de dollar.(12) En 2014, une autre étude, du Tufts Center for the study of Drug Development, a évalué que le coût pour développer un nouveau médicament jusqu’à la mise sur le marché était de 2.6 milliards de dollar(13). L’une des causes de cette augmentation des coûts de R&D est la croissance des frais de développement qui est estimée à plus de 10% par an et représente plus des deux tiers des coûts de R&D.(12)

Le développement de médicament est donc très coûteux et est par conséquent financé par les grosses entreprises qui ont plus de capacités à allouer des budgets conséquents à la recherche et au développement. Les entreprises du médicament dépensent en moyenne 9.8% de leur chiffre d'affaire dans la R&D.(12) Une analyse a été faite par le magazine Forbes(14) sur les 227 médicaments mis sur le marché par des big pharma ces dix dernières années. La conclusion est que en divisant les coûts de R&D par le nombre de molécules développées, pour les entreprises ayant sorti plus de trois nouvelles molécules depuis 2002, la somme médiane par molécule est de 3.149 milliards d'euros. Pour celles en ayant lancé plus de quatre, la médiane monte à 3.975 milliards d'euros. Pour celle ne lançant qu'un seul médicament, la médiane reste élevée à 351 millions de dollars alors que tout est centré sur une molécules et que des partenariats sont passés avec de plus grandes entreprises pour en assumer le coût.(15) La R&D est donc un des secteurs majeur en terme d'investissement pour les big pharma et est en constante augmentation. Cela est mis en évidence par les sommes déboursées mais aussi par l'augmentation du nombre d'employés dans ce secteur au cours des trois dernières décennies.

Année	Effectif de la R&D	Dont chercheurs ingénieurs de R&D
1980	6 998	1 901
1985	9 614	3 019
1990	11 175	3 916
1995	17 748	6 056
2000	18 227	6 792
2005	22 555	9 814
2011	20 263	9 498
2012	20 048	9 774
2013	20 054	9 938
2014	18 768	9 136
2015	19 109	9 635

Figure 3: évolution du nombre d'employés dans le secteur de la R&D chez les big pharma de 1980 à 2015

L'innovation thérapeutique est donc chère, longue et incertaine. Cela en fait un pari financier important qui peut s'avérer perdant pour les grandes entreprises. Il faut en effet mobiliser de grosses sommes pour un médicament qui ne verra possiblement jamais le jour, ou n'aura pas le succès attendu. Peu d'entre eux génèrent assez d'argent pour compenser leur coût de R&D. Il faut donc compter sur ceux se vendant bien et ne coûtant plus rien à l'entreprise considérés comme « vaches à lait ». Cependant, ces composés rapportent tant qu'ils sont protégés par un brevet.

Celui-ci dure vingt ans avec une prolongation possible de cinq ans à l'aide d'un certificat complémentaire de protection. Contrairement à ce que l'on peut penser, le médicament n'est pas protégé pendant vingt ans car le brevet est couramment déposé lorsque la molécule est identifiée, hors celle-ci subit les tests précliniques puis cliniques ce qui prend une dizaine d'années. Après avoir obtenu son autorisation de mise sur le marché (AMM) le médicament ne bénéficie d'une protection que pour dix ans.

De plus, un médicament novateur met entre deux et trois ans à rejoindre sa population cible thérapeutique,(12) En effet, il faut par exemple que les médecins prennent connaissance de son existence et changent leurs habitudes de prescription. Evidemment, ce délai est beaucoup moins long pour les génériques car le princeps a déjà eu le temps de faire sa place.

Un médicament arrivant sur le marché doit être un succès pour être rentable, mais encore faut-il qu'il obtienne son AMM. Un composé ayant atteint un stade avancé a déjà coûté des sommes conséquentes et s'il ne parvient pas au stade final, les pertes n'en sont que plus importantes. C'est ce qui est arrivé par exemple au rimonabant (Acomplia) qui devait être le nouveau blockbuster de Sanofi-Aventis (devenu Sanofi) pour lutter contre l'obésité. Ce composé disponible à travers 32 marchés dans le monde et prescrit à cinquante mille personnes en France devait rapporter un chiffre d'affaire de plus d'un milliard de dollars. Le retrait du marché de cette pilule miracle à cause de ses effets secondaires psychiatriques a signé le départ du PDG de l'époque de Sanofi Gérard le Fur. (15) On peut aussi citer dans la même idée un composé de Biogen contre l'Alzheimer en développement depuis des années, ayant coûté des milliards de dollars qui n'a pas passé les tests cliniques de phase trois. L'entreprise a donc perdu dix-huit milliards de dollars soit 30% de sa valeur au marché en un seul jour.(16)

Les coûts de R&D, les pertes sèches induites par les échecs et le profit ne durant qu'une dizaine d'année font que les retours sur investissement sont en constante diminution depuis une dizaine d'années. Comme le montre le schéma suivant, ceux-ci sont passés de 10.1% en 2010 à 1.9% en 2019. Certaines prédictions prévoient même que ces retours sur investissement tombent à zéro en 2020. (13)

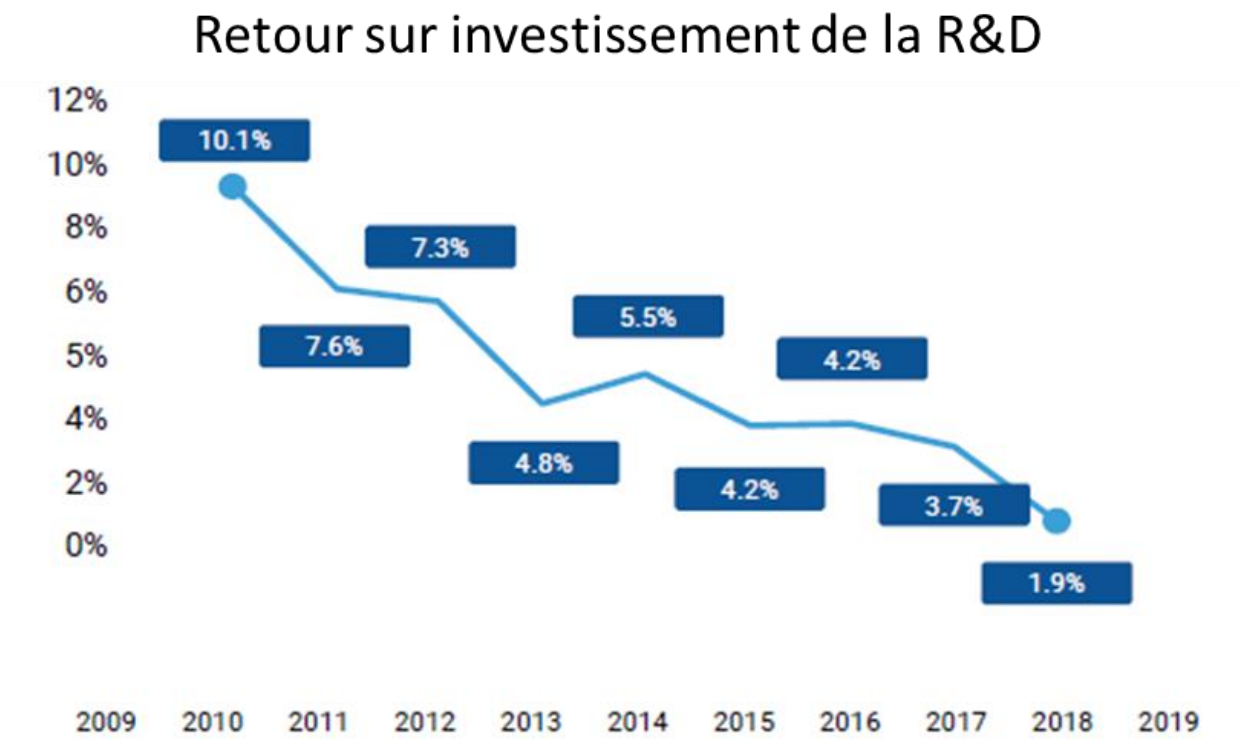

Figure 4: courbes montrant la diminution des retours sur investissement de la R&D de 2010 à 2019(13)

La recherche de médicament est donc un risque pris par les entreprises mais qui leur est nécessaire pour rester compétitives. Cependant, le constat est que ce risque est plus important d'année en année car les études sont plus chères et plus longues et les échecs sont de plus en plus néfastes. Cet état de fait est une réalité de plus en plus étudiée avec une loi qui lui est dédiée, la loi de Eroom.

3. La loi de Eroom (Eroom's law)

Pour expliquer la loi de Eroom, il faut tout d'abord comprendre son origine. Le terme Eroom vient de la loi de Moore inventée au milieu des années 60 par le co-fondateur de Intel, Gordon Moore. Il y explique que grâce aux progrès technologiques, le nombre de transistors par circuit intégré double environ tous les deux ans et même si l'avenir de cette prédiction est moins certain, elle s'est tout de même avérée juste pendant près d'un demi-siècle.(17) Cette loi décrit donc comment les progrès techniques permettent d'améliorer les capacités notamment informatiques tout en en diminuant les coûts et les rendant plus abordables.

La loi de Eroom décrit l'effet inverse (comme le symbolise le fait que ce soit le même nom écrit à l'envers) dans le monde de la recherche pharmaceutique. Cette loi est plus récente car inventée en 2012. Elle explique que le prix du développement d'un médicament double environ tous les neufs ans et est rapidement devenu le symbole du déclin de la recherche pharmaceutique.(17) Le schéma suivant met en parallèle ces deux lois permettant de visualiser le contraste entre le domaine du médicament qui s'essouffle dû à la complexité et la compréhension actuelle limitée des systèmes biologiques opposé au monde informatique en plein essor car plus simple ou du moins mieux compris.

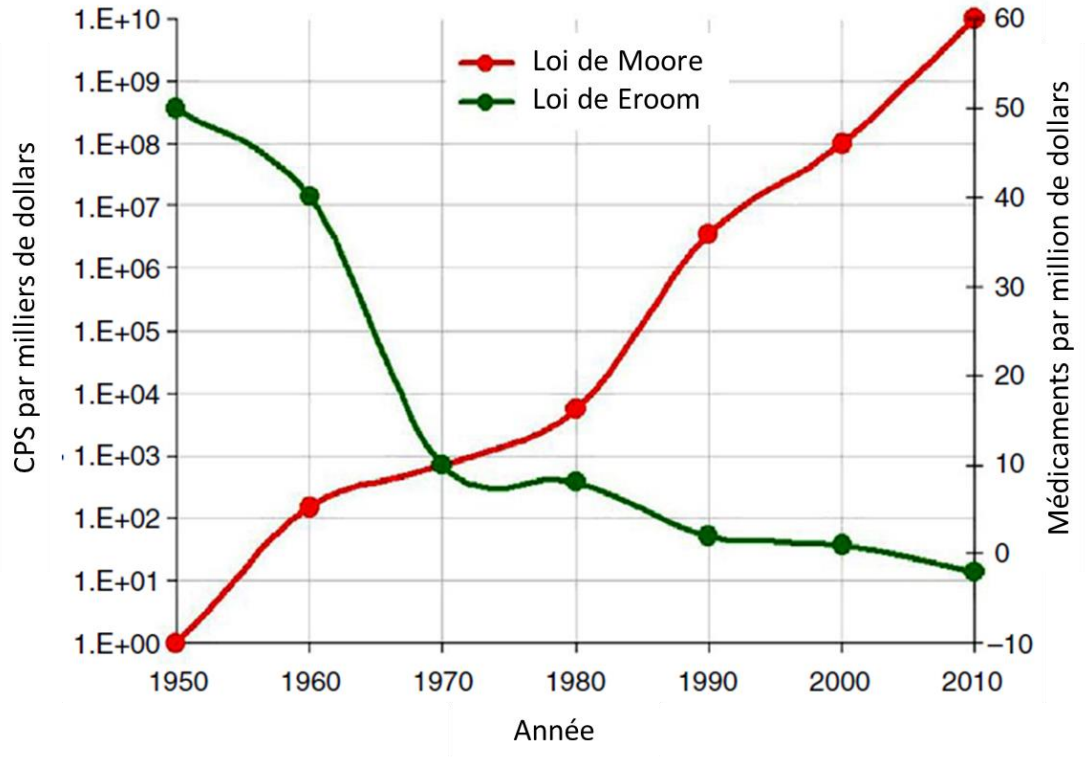

Figure 5: comparaison de la loi de Moore en rouge (CPS = caractère par seconde correspond à un débit de transfert assimilable à la puissance informatique) à la loi de Eroom en vert.(18)

Une étude approfondie du déclin de la recherche pharmaceutique menant à l'établissement de loi de Eroom a été menée et publiée dans nature en 2012 dans la publication de référence : «Diagnosing the decline in pharmaceutical R&D efficiency»(19). Ce papier sert de base aux paragraphes suivants décrivant les causes et conséquences de la loi Eroom.

3.1.Les causes de cet état de fait :

La loi de Eroom explique que des contraintes inhérentes à la recherche ont diminué les bienfaits apportés par les progrès technologiques et scientifiques depuis les 60 dernières années. Elle explique aussi que les améliorations étaient en réalité moins impactantes que ce que l'on pensait.

Certaines personnes pensent que la loi de Eroom peut être contrecarrée grâce à différents changements. Parmi ces changements, on retrouve la mise en place de techniques managériales appliquées à la recherche, une réorganisation de la R&D en petites unités focalisées et en grandes structures avec de forts potentiels financiers, l'externalisation dans les pays à bas coûts et enfin de faires des scientifiques des entrepreneurs. Ce n'est pas le point de vue partagé par les auteurs du «Diagnosing the decline in pharmaceutical R&D efficiency» (19) pour qui la plupart de ces solutions ne s'attaquent pas au cœur du problème. Pour eux, il existe quatre causes primaires à la loi de Eroom qui sont en anglais dans le texte :

- The « better than the Beatles" problem
- The "cautious regulator" problem
- The "throw money at it" tendency
- The "Basic research – Brute force" bias

Ils évoquent aussi l'impact d'un cinquième facteur, le « low hanging fruit » problem moins important que les quatre autres. Ces causes ont des noms très évocateurs qui dépeignent parfaitement des problèmes actuels de la recherche pharmaceutique.

3.1.1. The « better than the Beatles » problem

Ce nom est très imagé et se traduit en français par « comment faire mieux que les Beatles ? En effet, une analogie est faite entre la recherche pharmaceutique et le monde de la musique. Imaginons que toutes les chansons des Beatles soient en libre-service et que les gens ne se lassent jamais de les écouter. Dans ce cas, pour sortir un nouvel album de rock et avoir du succès, il faudrait que celui-ci soit à chaque fois meilleur que les Beatles.

C’est exactement ce qu’il se passe lorsqu’un nouveau médicament est créé. Les Blockbuster d’hier sont les génériques d’aujourd’hui et les gens (tout comme dans la comparaison précédente) ne se lassent pas de ces médicaments, un médecin ne s’arrêtera pas de prescrire de la metformine à un diabétique parce qu’il s’en est lassé. Chaque nouveau traitement doit donc être meilleur que ceux existant ayant la même indication. Sans cela ils ne se vendront pas et ne prendront pas de part de marché aux génériques qui sont de fait moins cher et plus faciles à produire.

Le catalogue des médicaments approuvés est forcé d’être en constante amélioration. Cela complexifie le développement, l’apport de la preuve de l’amélioration de service médical rendu, l’adoption et le remboursement des nouveaux traitements. Un bon exemple est une classe d’antiacide (les bloqueurs d’acidité compétitifs du potassium) qui auraient pu être des blockbusters il y a 15 ans mais qui n’ont pas pu se faire une place à cause des deux classes d’antiacides déjà existantes (les antagonistes des récepteurs H2 de l’histamine et les inhibiteurs de pompe à proton.).

Une autre cause de la loi de Eroom est assez similaire, les « low hanging fruits » se traduisant littéralement par « les fruit des branches basses » qui décrit les cibles faciles. L’idée est que tous les médicaments « faciles à trouver » l’ont été. Cela reste différent du « better than the Beatles problem » qui évoque plutôt le fait que les médicaments restant ont moins de valeur à cause de ceux existants. Pour les auteurs du «Diagnosing the decline in pharmaceutical R&D efficiency» (19), le problème des « low hanging fruits » est moins important car il reste énormément de cibles non exploitées. De plus, il semble que les médicaments tirent de plus en plus leurs bénéfices de leurs actions sur différentes cibles. Si cela s’avère juste, s’inquiéter des cibles faciles reviendrait à s’inquiéter du nombre limité de notes pour faire de la musique. Cette cause reste tout de même un sujet de discussion permettant d’expliquer la loi de Eroom.

3.1.2. The « cautious regulator » problem

Cette cause, traduite littéralement par problème du régulateur prudent, vient du fait que le seuil de tolérance de risque est constamment abaissé. Les médicaments arrivants sur le marché doivent montrer que leur toxicité et les risques qu’ils induisent sont bien maîtrisés. Cela a toujours été le cas mais le seuil de tolérance est constamment abaissé à la suite de scandales sanitaires ou bien de nouvelles découvertes. Cela afin d’être toujours plus sûr. Mais ce seuil ne remonte évidemment jamais. Cela augmente le temps et donc le coût des tests de toxicité et donc de la R&D globale.

Le test d'Ames qui est un vieux test permettant de tester la mutagénicité élimine un grand nombre de candidats en n'améliorant que modérément la sécurité des médicaments. Mais il semble inconcevable pour une institution de santé de dire que l'on va se passer de ce test car en cas de problème, les répercussions seront grandes.

Les preuves d'efficacité et d'innocuité à apporter par les entreprises sont donc de plus en plus conséquentes, le chef scientifique de Novo Nordisk a déclaré que si elles avaient été imprimées, les données sur leurs deux traitements à base d'insulines sorties en 2011 auraient été plus hautes que l'empire state building.

Il y existe cependant un aspect économique pour certaines entreprises à ce que ce seuil de tolérance reste bas. En effet, s'il est coûteux de mettre sur le marché un médicament car les tests de toxicité sont longs et complexes, alors les plus petites entreprises auront bien plus de mal à amener leur produit sur le marché. Les grandes entreprises auraient donc un intérêt à garder ce seuil bas pour éliminer la concurrence, tout en faisant bonne figure face au grand public, même si cela conduit à une augmentation de la R&D.

Ce régulateur est plus tolérant pour les aires thérapeutiques où peu de traitements existent. Mais dit d'une autre façon, cela veut dire que lorsqu'un traitement est existant, la régulation est renforcée et il est plus compliqué de faire sa place sur le marché (ce qui rappelle le « better than the Beatles problem »).

Si les années 1950 / 1960 apparaissent aujourd'hui comme l'âge d'or de l'innovation, cela était probablement dû à un laxisme impossible aujourd'hui.

3.1.3. The « throw money at it » tendency

The « throw money at it » tendency qui se traduit par la tendance à « noyer les problèmes sous l'argent » vient du fait que jusqu'à récemment, il y a eu toujours plus de dépenses dans la R&D. Ce qui est logique car le retour sur investissement était intéressant. De plus, le fait qu'un médicament qui est le deuxième ou troisième sur le marché vaille moins que le premier incite à dépenser plus dans la R&D pour être le premier à commercialiser un traitement.

Cependant, à l'heure actuelle, l'idée est de combattre cette tendance. Beaucoup d'investisseurs pensent qu'il est possible de diminuer le budget de la R&D sans pour autant

diminuer la production. Ceci est possiblement vrai et serait peut-être le moyen de contrecarrer la loi de Eroom, mais à court terme seulement.

L'incompréhension des facteurs influençant les retours sur investissement de la R&D a conduit à une dépense irraisonnée lorsque celle-ci était rentable et cela pourrait mener à une réduction des coûts, elle aussi irraisonnée. Cela pourrait induire une économie en un premier temps, suivie d'une perte de capacité de recherche induisant à terme une baisse de productivité car moins de médicament novateurs menant au final à un empirement de l'état actuel.

3.1.4. The « basic research – brute force » bias

Le "basic research – brute force" bias ou en français le biais de la force brute de la recherche fondamentale est défini par les auteurs de «Diagnosing the decline in pharmaceutical R&D efficiency» (19) comme : « la tendance à surestimer la capacité des progrès de la recherche fondamentale (notamment en biologie moléculaire) et de la puissance brute des techniques de criblage, à augmenter la probabilité d'avoir une molécule sûre et efficace lors des essais cliniques ».

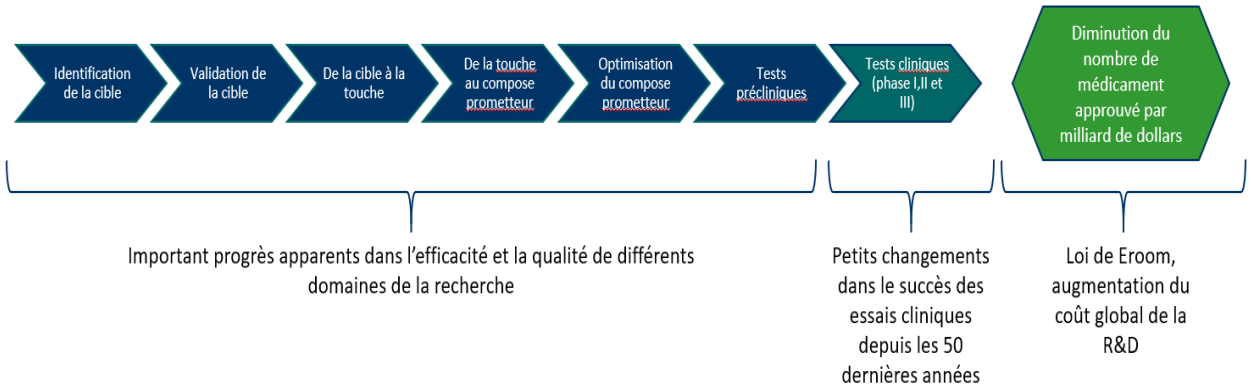

Figure 6: schéma décrivant le processus de développement de médicament comme une série de filtres excluant des molécules à chaque étape. De grandes améliorations ont été apportées aux étapes précédant les tests cliniques. Cela n'ayant malgré tout pas empêché l'apparition de la loi de Eroom.

Comme le montre la figure précédente, au cour des dernières années les étapes préliminaires du développement (de l'identification de la cible aux tests précliniques) se sont grandement améliorées. Ces améliorations auraient dû permettre de n'avoir en test clinique que des molécules ayant un effet sur la cible visée ou du moins d'avoir un meilleur taux de succès et moins de molécules toxiques. Il en résulterait un meilleur rapport sur investissement de la R&D.

Or, comme on le voit grâce à la loi de Eroom, ce n'est pas le cas. De plus, les plus grosses pertes sont dues à des projets échouant aux tests cliniques. Il est donc évident que l'amélioration des techniques n'a pas eu l'effet attendu. En effet, la probabilité qu'une molécule réussisse les tests cliniques est restée la même lors des cinquante dernières années.

Quelle est la cause de ce problème. Les techniques développées n'étaient pas les bonnes ? Parmi les critiques qui ont été faites sur la R&D, les auteurs mettent en exergue deux potentielles raisons. Le premier problème est que la R&D des industries pharmaceutiques est basée sur l'idée qu'une bonne molécule est un ligand de haute affinité pour une cible biologique donnée. Or les cibles biologiques font partie d'un système bien plus complexe ce qui fait que l'activité sur une cible pourra avoir des effets inattendus ou au contraire aucun effet car normalisée par d'autres mécanismes.

Il y a quelques dizaines d'année, il y avait plus de tests à plus grande échelle sur des animaux qui permettaient peut être d'avoir une meilleure vision d'ensemble que les tests moléculaires de screening actuels. Entre 1999 et 2008, la plupart des molécules approuvées ont été découverte majoritairement par essais phénotypiques plutôt que par essai basé sur la cible. Tout cela alors que les techniques de screening étaient majoritairement basées sur la cible. Il se peut donc que l'approche basée sur la cible soit efficace pour valider des hypothèses sur la thérapeutique mais ne soit pas performante pour la découverte de molécules pouvant surpasser le problème « better than the Beatles».

Le second problème avancé par les auteurs est que la technique de recherche moderne consistant à des filtrages en série (comme expliqué dans la figure 6) avec au début un criblage haut débit d'une bibliothèque prédéfinie de composés qui n'est pas adapté à l'exploration de l'espace chimique. Il se pourrait que la technique « ancienne » (avant les années 90) basée sur la chimie médicinale ciblée avec des essais en parallèle soit plus adaptée à trouver la bonne molécule dans l'‘immensité que représente l'espace chimique.

Nous venons de voir les deux causes qui expliquent le « basic research – brute force bias » mais comment se fait-il que l'on en soit arrivé là, pourquoi la recherche est partie dans ce qui semble aujourd'hui être une mauvaise direction ? Les auteurs de l'article donnent trois raisons.

- Au début des années 80 alors que le « cautious regulator problem », le « better than the Beatles problem » et le « low hanging fruit problem » étaient déjà des réalités ou du moins

des craintes, les industriels pressentaient que les modèles animaux pour des maladies peu traitées allaient manquer, ce qui a permis l'essor du criblage haut débit.

- La génétique et la biologie moléculaire ont été considérées comme les meilleurs moyens pour comprendre les systèmes biologiques et pour ensuite intervenir dessus. Cela a défavorisé l'approche empirique qui semblait plus désordonnée.
- Enfin, il y a une inclination des commerciaux, des investisseurs etc. à penser que l'ancien modèle était trop hasardeux. Il semblait trop risqué de laisser des chimistes et des biologistes se lancer dans de long tests pouvant échouer alors que l'automatisation des processus qui a marché dans tant d'autres domaines était accessible à la R&D.

Ce biais a mené à une vision erronée de la recherche qui l'a fait passer pour beaucoup plus précise et rationnelle qu'elle ne l'est réellement. Il est plus vendeur et crédible d'avoir une molécule X qui agit sur la cible A induisant l'effet B grâce au mécanisme C le tout en expliquant pourquoi elle a cet effet et aucun autre. Malheureusement la science ne marche pas comme cela, l'exemple peut être donné par l'anticancéreux iniparib supposé agir contre le cancer du sein car il inhibe la poly(ADP-ribose)polymérase 1. Arrivé en phase III, il s'est avéré qu'il n'avait pas d'effet contre le cancer du sein ni même sur la polymérase.(20)

Pour terminer, il serait possible d'améliorer le ratio de molécules passant les tests cliniques si l'on savait pour quelles raisons les molécules refusées ont échoué. Cela permettrait de trouver des points communs lors des échecs afin de pouvoir repenser le système des études préliminaires du processus de R&D. Les entreprises sont donc très intéressées actuellement par des études sur les raisons des échecs rencontrés lors du développement de médicaments et par des techniques permettant de les éviter. Parmi ces techniques, l'utilisation d'IA afin de mieux comprendre les systèmes biologiques semble être une voie prometteuse, ce qui explique l'apparition récente de nombreuses entreprises dans ce domaine.

3.2.Les symptômes secondaires :

Les quatre causes primaires de la loi de Eroom citées précédemment ont donné naissance à des symptômes secondaires qui ont pour conséquence d'augmenter les coûts de la recherche. L'augmentation se concentre majoritairement sur les essais cliniques qui ne sont pas l'objet de ce manuscrit mais pour lesquels on retrouve de plus en plus d'entreprises d'IA. Il est donc intéressant d'évoquer le sujet ici. Scannell and al. évoquent quatre symptômes dans leur publication.

- **“The narrow clinical search problem”** ou le problème de la recherche clinique correspond au changement qui s’est opéré dans la recherche clinique. Le modèle est passé de techniques de screening sur animaux permettant de chercher largement de nouvelles cibles thérapeutiques pour des composés actifs à une méthode de recherche plus (trop) précise où l’on veut une molécule agissant sur une cible déjà décidée. Ce problème est la conséquence de «cautious regulator problem » et du « basic research-brute force bias ». Le premier car si un test entre en phase clinique et démontre un effet qui n’est pas celui attendu, il échoue, ce qui peut mener à une perte de composés actifs et une moins grande probabilité de découvrir un composé par sérendipité. Le second car on a une plus grande confiance en nos modèles *in vitro* et *in vivo*, ce qui a poussé les chercheurs à s’appuyer dessus, possiblement à l’excès .

- **“The big clinical trial problem”** ou le problème des gros essais cliniques. En effet, au cours des années, le nombre de personnes enrôlées dans les essais cliniques a fortement augmenté. Cela permet d’avoir plus de certitudes quant à la sécurité du produit, d’avoir une meilleure puissance marketing mais cela rend aussi les tests plus chers et plus compliqués à mettre en œuvre.

- **“The multiple clinical trial problem”** ou problème des essais cliniques multiples. A cause du « better than the Beatles problem », l’exercice médical est devenu bien plus compliqué. Les médecin on en effet le choix aujourd’hui entre beaucoup de traitements différents pour une même maladie. Par rapport au temps où le choix du traitement du diabète de type 2 se faisait entre insuline, régime et sport, le médecin a maintenant le choix entre une dizaine de classes thérapeutiques différentes. Le « régulateur prudent » n’était pas prêt à faire face à cette évolution qui a fragmenté la population de malades obligeant les entreprises à réaliser des tests cliniques plus précis. L’exemple peut être donné par le premier analogue de l’insuline à action prolongé autorisé par la FDA (Food and Drug Administration) en 1999 à la suite de trois essais cliniques contre douze essais qui ont dû être passés par le dernier analogue approuvé en 2011.

- **“The long cycle time problem”** ou le problème des longs temps de cycle. En 2005-2006, les tests cliniques duraient environ 9 ans. Dans les années cinquante, la régulation était moins sévère et les tests cliniques étaient bien plus rapide. La preuve en est l’imipramine qui en six ans avait passé les tests précliniques et trois tests cliniques différents.

Ces quatre symptômes montrent que l'on a des tests qui sont aujourd'hui plus importants, plus nombreux, plus longs et plus restrictif que précédemment. Si ces quatre symptômes sont accès sur les tests cliniques, ils reflètent malgré tout bien l'état général de la R&D dans ce contexte de la loi de Eroom. Dans un domaine où créer semble de plus en plus compliqué, les institutions de recherches semblent avoir troqué leur capacité d'innovation contre une meilleure sécurité d'utilisation de leurs médicaments. A cause de cela, les entreprises mettent entre dix et quinze ans à amener une molécule sur le marché pour un rentabilité qui n'est pas toujours au rendez-vous.

Il me semble cependant important de mesurer les propos que l'on peut lire dans la publication «Diagnosing the decline in pharmaceutical R&D efficiency» qui fait une analyse de la loi de Eroom avec une approche très scientifique de la chose. A mon sens, il n'est pas assez évoqué le fait que rendre les tests plus drastiques permet d'éviter des effets possiblement dévastateurs sur la population et que diminuer au maximum les tests sur les animaux, même si cela coûte à la recherche est une bonne chose pour l'aspect éthique de la recherche médicale.

Il faut donc trouver un moyen de lutter contre la loi de Eroom tout en respectant les avancées permettant une plus grande sécurité du patient et un meilleure éthique de la recherche. Il semblerait qu'une des solutions que le futur (qui est paradoxalement déjà là) ait à nous apporter soit l'utilisation d'IA. En effet, le machine learning est actuellement vu comme un bon moyen de réduire le temps et les coûts de la R&D sans avoir à sacrifier la sécurité du patient et sans augmenter les tests sur les animaux tout en partant d'un pool de molécules plus important. Le problème de la loi de Eroom est si important et le possible apport du machine learning est si intéressant que cela a poussé en 2019 le GAO (le US Government Accountability Office, l'organisme d'audit, d'évaluation et d'investigation du congrès des Etats-Unis) à réaliser un rapport sur le sujet intitulé «Artificial Intelligence in Health Care Benefits and Challenges of Machine Learning in Drug Development(21). On peut retrouver dans ce rapport le schéma présent en figure 7 où l'on voit bien l'amélioration que peut apporter le machine learning, plus de composés de départ pour arriver plus rapidement à un seul composé qui pourra devenir un traitement approuvé.

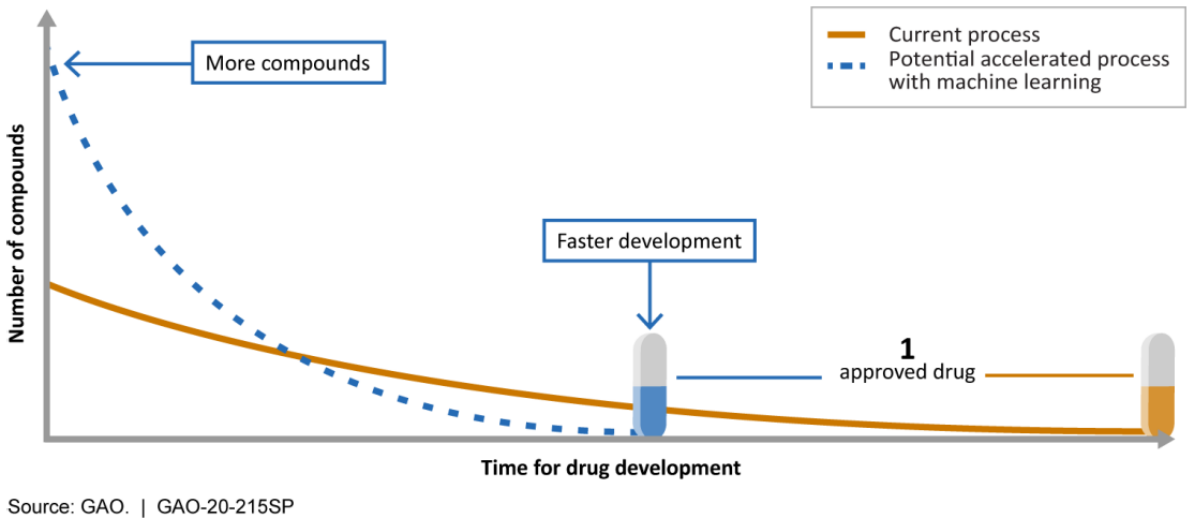

Figure 7: schématisation par le GAO de l'amélioration que peut apporter l'utilisation du machine learning dans le développement de nouveaux médicaments, plus de molécules pour un médicament plus rapidement(21)

4. L’apport des intelligences artificielles

4.1. Définitions

Il sera fait dans la suite de ce manuscrit une liste et une description non exhaustive des domaines de la recherche médicale où le machine learning est utilisé en vue diminuer les coûts et le temps global de la R&D.

Mais tout d’abord, afin de comprendre les bénéfices pouvant être apportés par l’utilisation de l’intelligence artificielle dans la création de molécules médicamenteuses, il est nécessaire de connaître certains concepts. Pour cela, voici un historique agrémenté que quelques définitions et explications de ces termes avec des exemples simples permettant d’aborder sereinement les notions qui seront utilisées par la suite dans ce manuscrit.

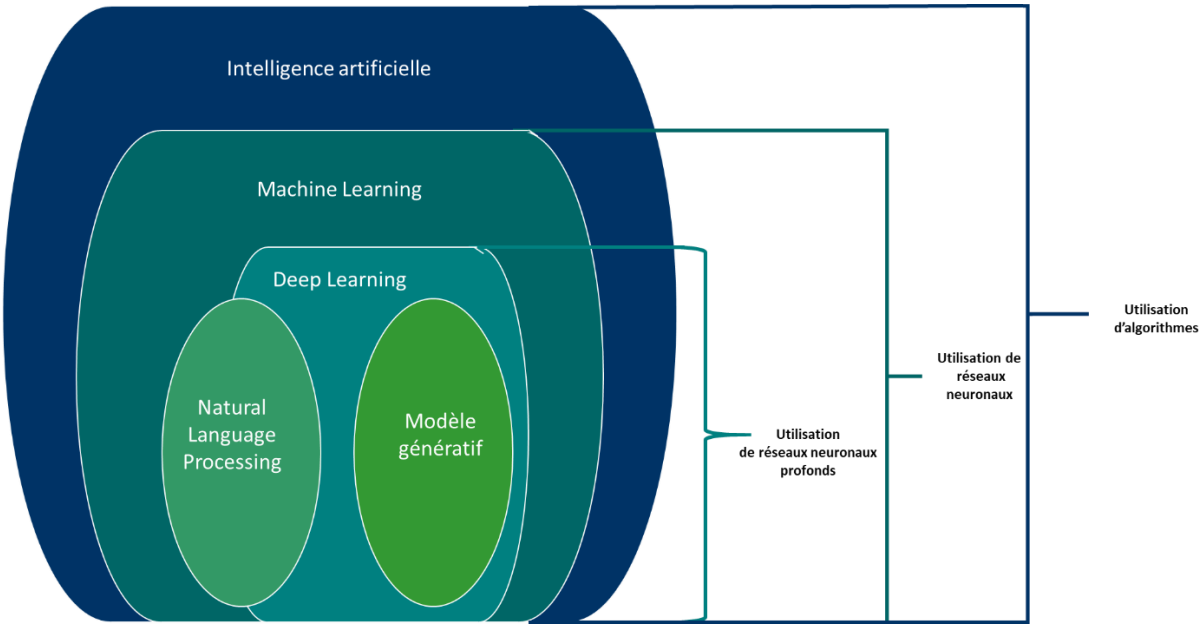

Figure 8: schéma explicatif hiérarchisant de façon simplifiée les liens entre IA, Machine learning, Deep learning et NLP

Algorithme : Un algorithme est un «ensemble de règles opératoires dont l'application permet de résoudre un problème énoncé au moyen d'un nombre fini d'opérations. Un algorithme peut être traduit, grâce à un langage de programmation, en un programme exécutable par un ordinateur. »(22) Il a toujours plus ou moins existé des algorithmes, l'exemple le plus simple et concret est une recette de cuisine. Les ingrédients sont les données de départ, le plat est le résultat et la préparation est divisée en plusieurs étapes qui sont les « nombres finis d'opérations ». Comme sous-entendu dans la définition, la principale utilisation des algorithmes se fait grâce à l'informatique. Cela permet entre autres l'utilisation d'un grand nombre de données, la réalisation de calculs complexes, la combinaison de plusieurs algorithmes etc en vue de répondre à des problèmes plus ambitieux que «comment réussir une omelette». Ils sont utilisés par exemple pour simuler l'évolution de la propagation de la grippe en hiver ou conseiller un acheteur en ligne en fonction de ses commandes précédentes pour ne citer que ça.(23)

L'intelligence artificielle (IA) correspond à tout programme ou système informatique qui fait quelque chose que nous considérons comme intelligent chez l'homme. Les IA extraient des concepts et des relations à partir de données. Elles sont aussi capables d'apprendre indépendamment grâce à des modèles de données (donc des données structurées), augmentant ainsi ce que les humains peuvent faire.(24) C'est un concept qui a été évoqué pour la première fois par Alan Turing dans l'article « Computing Machinery and Intelligence »(25) en octobre 1950. Lors de ce test, un humain interagit à l'aveugle avec une autre personne puis avec une machine programmée à répondre. Si le sujet n'est pas capable de différencier la machine de l'homme, la

machine réussit le test. Ce test sera plus tard jugé inadéquate mais sera le point de départ d'un domaine qui sera appelé intelligence artificielle par John McCarthy (cofondateur du MIT, Massachusetts Institute of Technology, USA) en 1956 et défini par Marvin Lee Minsky (second cofondateur du MIT) comme la "Construction de programmes informatiques qui s'adonnent à des tâches qui sont, pour l'instant, accomplies de façon plus satisfaisantes par des êtres humains car elles demandent des processus mentaux de haut niveau tels que l'apprentissage perceptuel, l'organisation de la mémoire et le raisonnement critique."(26)

L'IA a comme principe de mimer l'intelligence « biologique ». Le développement de ce domaine s'est fait en ce sens. Après leur publication « *What the frog's eye tells the frog's brain* »(27) où sont présentés leurs travaux sur les réseaux neuronaux. Les neurologues McCulloch et Walters Pitts créent le « neurone formel » en 1943.(28) Il s'agit d'une copie simplifiée d'un neurone biologique conçu comme un automate permettant de traiter des données venant de multiples entrées (dendrite) grâce à des règles précises pour donner un résultat en sortie (axone).

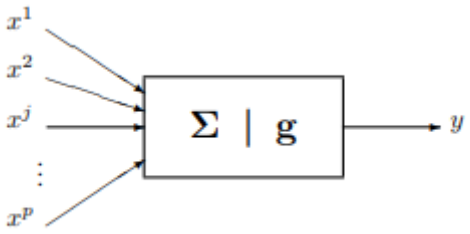

Figure 9: représentation schématique d'un neurone formel avec les x correspondant aux données en entrée, y le résultat en sortie et au milieux sa fonction traitant les données(29)

Un réseau de neurones est une association de neurones formels. L'agencement des connexions et l'efficacité de transmission des résultats d'un neurone à l'autre sont variables. Cela correspond à des « poids synaptiques » pouvant être contrôlés par des règles d'apprentissage mathématiques.(28) Ces poids constituent la mémoire ou connaissance répartie du réseau,(29) ils permettent par exemple d'accorder plus d'importance à une donnée par rapport aux autres. Par conséquent, «apprendre» pour un réseau neuronal, correspond à faire varier ces poids (ou coefficient de pondération) afin de trouver la combinaison optimale pour traiter les données présentées.

Malheureusement, à cette époque-là, il n'y avait pas de techniques pour adapter ces coefficients de pondération. Il faut en effet attendre 1949 et les travaux du physiologiste Donald Hebb dans

son livre « the organization of behavior » qui donne la règle suivante : « Faisons l'hypothèse qu'une activité persistante et répétée d'une activité avec réverbération (ou trace) tend à induire un changement cellulaire persistant qui augmente sa stabilité. Quand un axone d'une cellule A est assez proche pour exciter une cellule B de manière répétée et persistante, une croissance ou des changements métaboliques prennent place dans l'une ou les deux cellules ce qui entraîne une augmentation de l'efficacité de A comme cellule stimulant B. » Cette règle de Hebb est maintenant utilisée comme un standard dans la majorité des réseaux, même les plus sophistiqués pour modérer les relations entre les neurones le composant.(30)

L'évolution des réseaux neuronaux a donné ce que Arthur Samuel, développeur chez IBM mettant au point une machine pour jouer aux échecs appelle pour la première fois en 1952 le machine learning.

Le Machine Learning ou apprentissage automatique en français est l'utilisation d'algorithmes informatiques qui apprennent à partir de données structurées et non structurées, identifient les modèles cachés, font des classifications et prédisent les résultats futurs.(24) Les systèmes de machine learning utilisent des algorithmes et des réseaux neuronaux pour aider les ordinateurs à améliorer progressivement leurs capacités. Pour cela, l'algorithme créé automatiquement un modèle mathématique en utilisant des données d'entrainement pour ensuite prendre des décisions sans que celui-ci soit spécifiquement programmé.(30) On considère que le plus vieil algorithme de machine learning est le Perceptron inventé en 1957 par Frank Rosenblatt en associant les travaux de Hebb et Samuel. Il pouvait effectuer des tâches de reconnaissance de modèles complexes afin de pouvoir séparer deux classes de données. C'est cet algorithme qui sera plus tard utilisé pour reconnaitre des objets sur des images. Considéré comme le premier neuro-ordinateur fonctionnel, le perceptron a cependant rencontré certains problèmes. Ainsi, le domaine de l'IA et du machine learning ne s'est pas arrêté, mais a fait face dans les année 70 à ce qui a été appelé « l'AI winter ». Il faudra attendre les années 1990 pour que les investisseurs fassent confiance de nouveau à ce type de système. La résurgence s'est notamment faite grâce à l'arrivée de l'internet et de l'explosion de la quantité de données digitales disponibles.(31) Cependant, dans le domaine de la santé à la suite de l'AI winter, des travaux dans ce domaine ont continué mais une succession d'échecs a amené à un second AI winter entre la fin des années 80 à la fin des années 2000.

Le machine learning se définit comme la capacité d'un ordinateur à apprendre et à s'améliorer. Et si ici les réseaux neuronaux sont beaucoup évoqués, il existe différentes méthodes pour arriver à ce résultat comme les réseaux bayésiens, les random forest, et le deep learning qui utilisent

chacune des hypothèses et des outils mathématiques différents pour gérer le traitement des données et l'apprentissage au sein de l'algorithme. (21)

Le Deep Learning ou apprentissage profond en français est une sous-catégorie du machine learning qui utilise une structure logique similaire à celle du cerveau appelée " réseaux " neuronaux pour reconnaître et discriminer des modèles tels que la parole, l'image et la vidéo.(24) Jusque-là, le concept est semblable à celui des réseaux neuronaux du machine learning. Le terme Deep (profond) vient de la superposition de couches de neurones artificiels, bien plus importante que dans le machine learning, chacune ayant une fonction bien définie. Chaque couche reçoit et interprète les informations venant de la couche précédente permettant ainsi l'obtention de résultats souvent plus précis et d'avoir un système capable d'apprendre par lui-même et cette fois-ci, sans intervention humaine. Cette capacité à empiler plus de couches de neurones est notamment due à une augmentation des capacité informatiques.(32)

Les applications du machine et du deep learning sont nombreuses. Parmi elles, il est possible de citer notamment le natural language processing et les modèles génératifs qui se répandent de plus en plus dans le milieu de la santé.

Le natural language processing (NLP) ou programmation neuro-linguistique (PNL) est l'application du machine learning à l'analyse et à la synthèse du langage naturel et de la parole.(24) Cela permet aux ordinateurs de comprendre et organiser les langages humains. Ce type de technologie permet d'analyser des documents écrits par l'homme pour en tirer des données. Un exemple d'application est l'évaluation de la progression d'un cancer et sa réponse à la thérapie par l'analyse de rapports de radiologie. (21)

Les modèles génératifs ou Generative Adversarial Networks (GANs) sont « issus de l'apprentissage automatique et permettent à la fois de générer de nouveaux exemples à partir des données d'entraînement et d'évaluer la probabilité qu'un nouvel exemple provienne ou ait été généré à partir des données d'entraînement ».(33) Ce concept a vu le jour en juin 2014 dans la publication « Generative Adversarial Networks »(34) par Ian Goodfellow. Un modèle génératif peut être assimilé à un système de Deep learning monté à l'envers car permettant une fois la phase d'apprentissage réalisée de créer de nouvelles données artificielles.

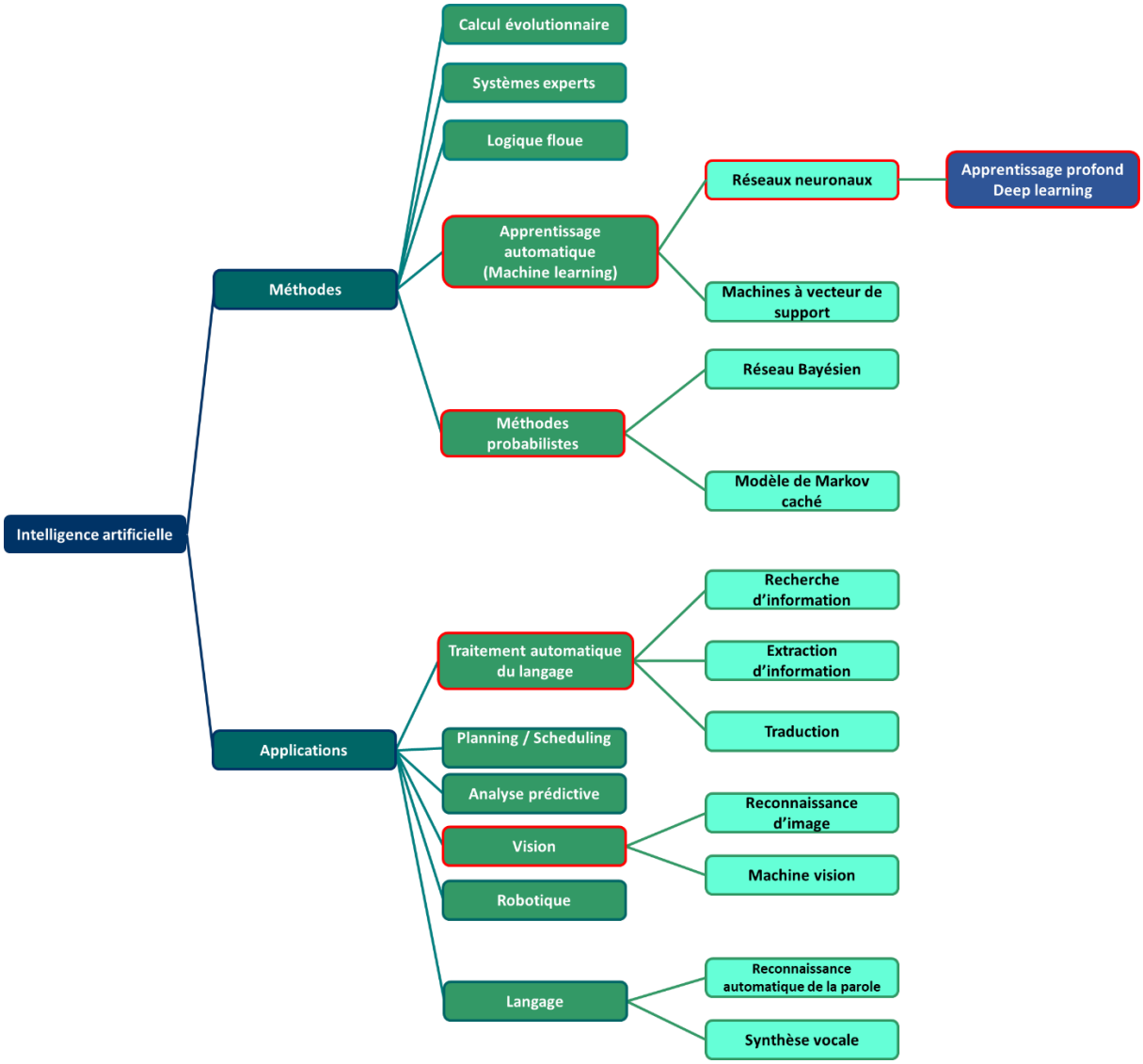

Figure 10: schéma résumant les différentes techniques et méthodes résultant de l'utilisation des IA. Le focus se fait (cadre rouge) sur les techniques les plus évoquées dans ce manuscrit, notamment l'utilisation de réseaux neuronaux qui est loin d'être le seul moyen d'utiliser des IA.

Les concepts inhérents au machine learning, au deep learning ou au modèle génératif sont complexes à saisir à partir de définitions. Le meilleur moyen de s'approprier ces notions est d'utiliser des exemples simples et concrets. En voici quelques-uns utilisant des modèles de réseaux neuronaux car c'est ce type de méthodes qui sera le plus souvent évoqué dans ce manuscrit.

Imaginons que l'on cherche à avoir un logiciel capable d'identifier un chien sur une image, dans le cadre de la création d'un système d'audiodescription par exemple. Il va falloir entrainer les algorithmes du logiciel à reconnaitre ce qui fait qu'un chien est un chien, et non autre chose. La phase d'apprentissage consiste à montrer un grand nombre d'images au logiciel en lui indiquant si c'est un chien ou non.

Pour un logiciel de **machine learning**, il va falloir qu'un programmeur soit très explicite lorsqu'il indiquera à l'ordinateur ce qui doit être recherché dans l'image pour analyser l'animal. C'est l'Homme qui doit indiquer par la création d'un algorithme d'extraction de données qu'il faut qu'il y ait par exemple quatre pattes, des dents, des poils etc. Ce processus long et fastidieux est appelé « l'extraction de caractéristiques » et influence très fortement les capacités du logiciel. Le taux de réussite de l'ordinateur dépend en effet entièrement de la capacité du programmeur à définir les caractéristiques d'un chien et ce qui le différencie d'un autre animal ayant des caractéristiques similaires comme un chat.(35)

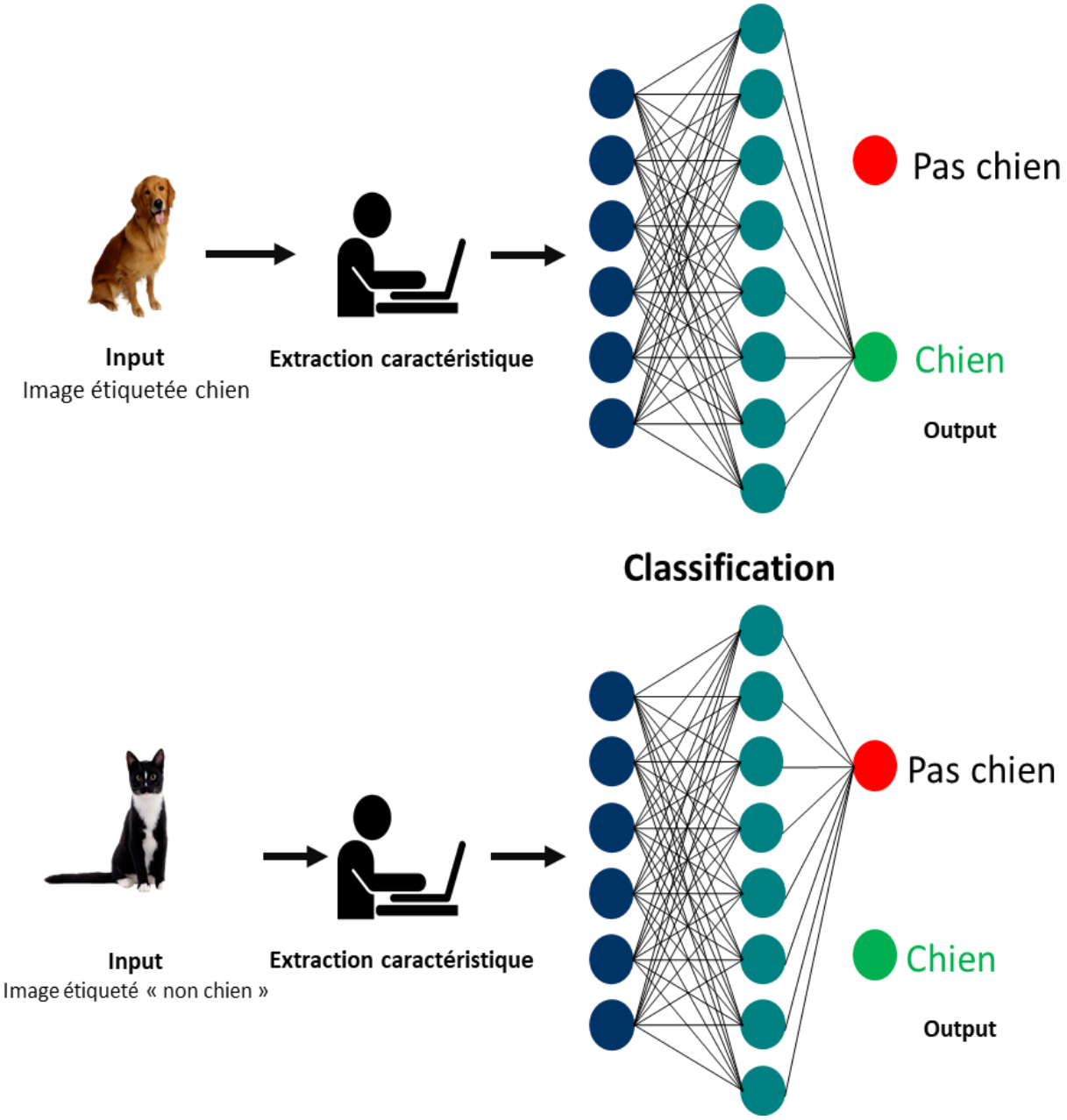

Figure 11: schéma explicatif de la phase d'apprentissage d'un logiciel de machine learning

L'avantage du **deep learning**, c'est que le logiciel va lui-même établir les caractéristiques à avoir pour affirmer que l'image représente un chien et non un chat. Cela permet au logiciel d'être plus rapide et plus précis. Pour la phase d'apprentissage, il suffit d'utiliser des images étiquetées chien / pas chien et le logiciel à force de recevoir des données va comprendre que pour que l'animal soit « chien »,il faut quatre pattes, des dents pointues deux oreilles etc. Pour être efficace et précis, un algorithme de deep learning nécessite un très grand nombre de données lors de l'apprentissage. Cela est aujourd'hui possible dans le monde du « Big Data » car il est facile de trouver une immense quantité d'images décrites et étiquetées. Cela peut venir de professionnels comme ImageNet (36), mais aussi de particuliers sur les réseaux sociaux notamment, où fourmillent des photos balisées par des lieux, des activités etc. Il faut aussi une bien plus grande puissance de calcul car contrairement au machine learning où l'on retrouve deux à trois couches de neurones, on en retrouve bien plus dans le deep learning. Cet aspect a longtemps été un frein au développement du deep learning.

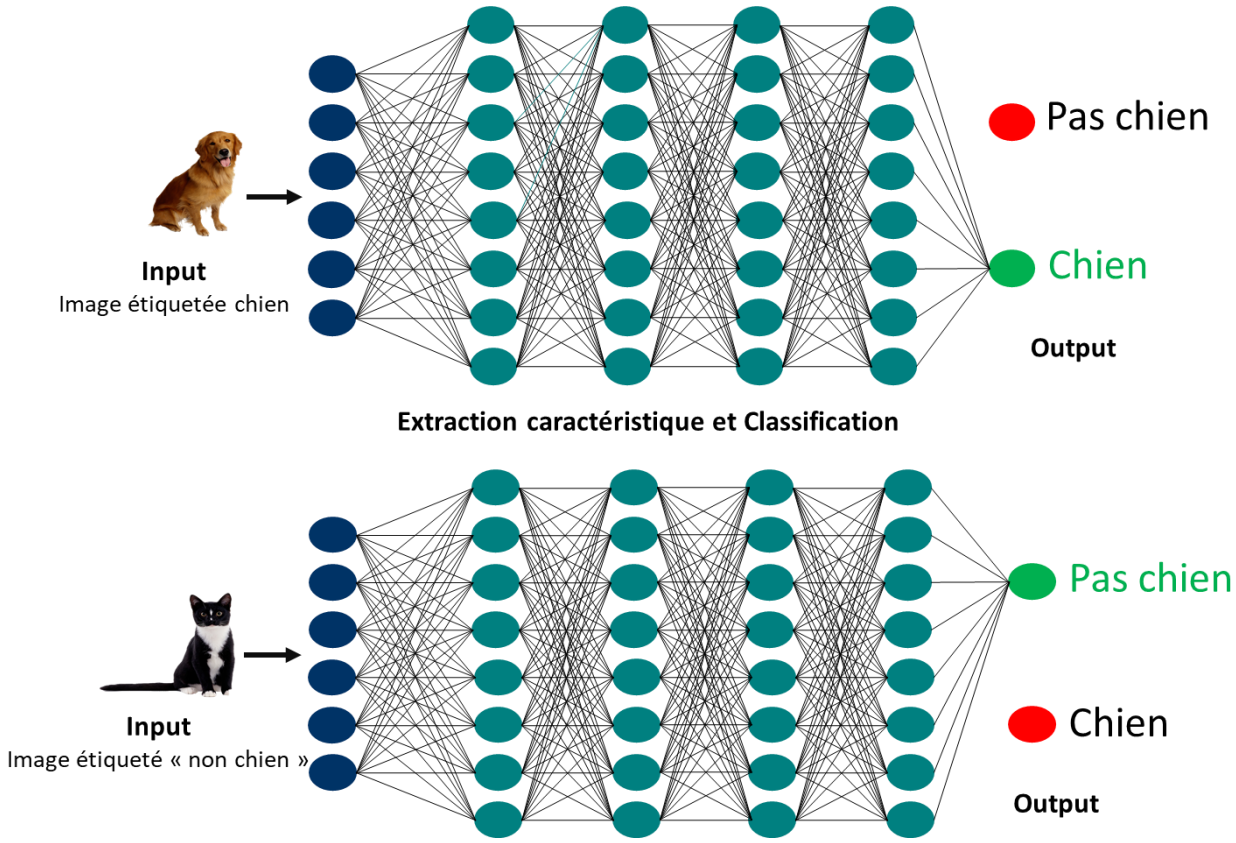

Figure 12: schéma explicatif de la phase d'apprentissage d'un logiciel de deep learning

Un **modèle génératif** aura un autre intérêt. C'est un logiciel de deep learning qui aura subi une phase d'apprentissage. Par exemple pour pouvoir reconnaitre un chien, il aura assimilé les caractéristiques qui font qu'un chien est un chien et sera ensuite utilisé dans « l'autre sens ». Au

lieu de lui montrer une image pour déterminer la présence ou non d'un chien, on lui demande une nouvelle image de chien. Celle-ci sera tout à fait nouvelle et artificielle car ne correspondant à aucune autre image qui lui a été soumise lors de sa phase d'apprentissage mais élaborée à partir des caractéristiques qu'il a identifiées dans ces images.

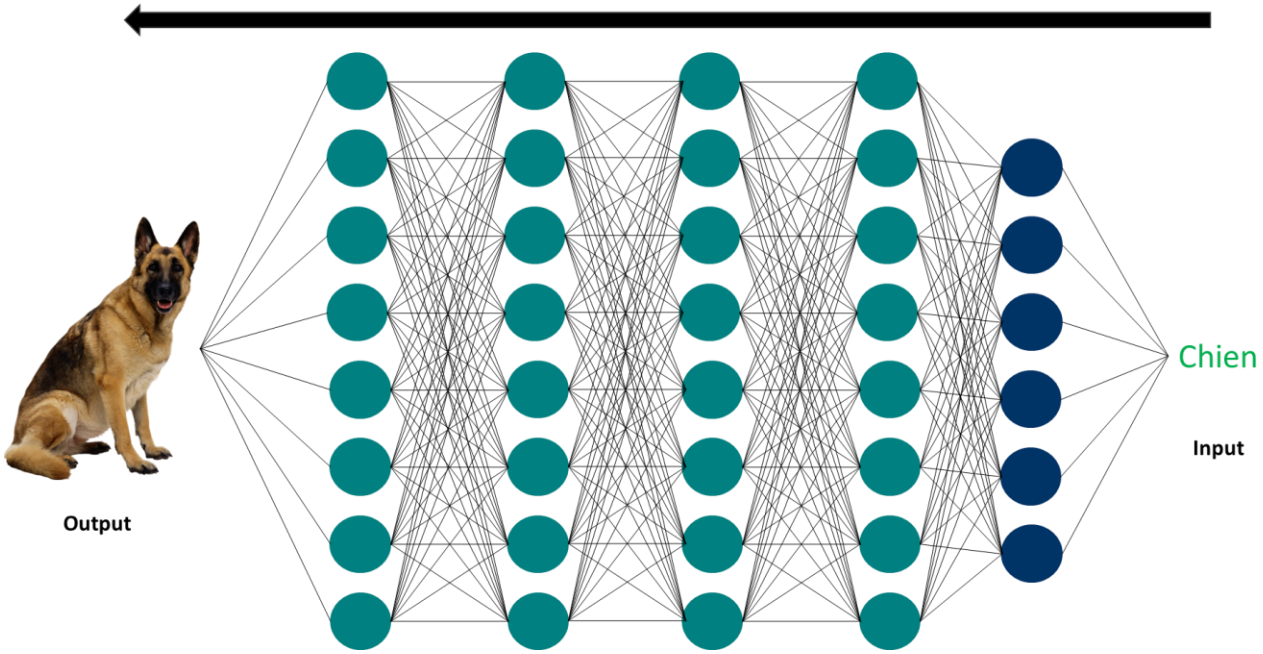

Figure 13: schéma explicatif de l'utilisation d'un modèle génératif

La relation entre un logiciel permettant de reconnaitre un chien et le monde de la santé peut paraître un peu flou à ce stade. Mais si un ordinateur est capable de reconnaitre une image de chien, il peut reconnaitre autre chose, cette idée a été utilisée par exemple pour reconnaitre des cellules cancéreuses parmi des cellules saines et ainsi faciliter le diagnostic. Ou bien un logiciel reconnaissant des points communs retrouvés dans des sites d'actions de récepteurs biologiques pourrait évaluer la capacité d'un médicament à agir sur une cible. Ce type de technologie peut aussi être utilisé pour créer de nouvelles molécules. En effet, s'il est possible d'entraîner un logiciel à dessiner des chiens à partir de zéro, pourquoi ne serait-il pas possible d'en entrainer un à dessiner des molécules ?

Ce sont ce type d'idées parmi d'autres qui ont permis au machine et au deep learning de trouver leur place dans le domaine pharmaceutique. Leur utilisation se fait de nombreuses façons différentes. Ce document fait un focus sur les technologies permettant la création de médicament et plus précisément sur la synthèse de novo de molécules thérapeutiques.

4.2.Les entreprises d’IA dans la découverte médicamenteuse

L’histoire de l’IA commence dans les années cinquante mais l’une des avancées ayant permis à ce domaine de se propager est l’exploit de l’ordinateur d’IBM Deep Blue qui a battu le champion d’échec Garry Kasparov en 1997. IBM Watson a d’ailleurs développé une filiale spécialisée santé et découverte de nouveaux médicaments utilisant les IA et a réalisé un partenariat avec Pfizer en 2016. Un autre fait d’arme qui a permis aux IA d’être prises au sérieux est la première démonstration de voiture autonome par Tesla en 2015. Depuis, les collaborations entre entreprises d’IA et les Big pharmas sont devenues fréquentes et certaines entreprises pharmaceutiques ont même développé des pôles spécialisés dans l’IA comme GSK en 2017. (37)

Comme expliqué précédemment, la loi de Eroom est la preuve que la recherche de nouveaux médicaments est dans une impasse. En effet, même s’il n’existe des traitements que pour un tiers des vingt à trente milles maladies existantes, la FDA n’a approuvé au total que 1578 médicaments qui sont dirigés contre seulement 819 cibles biologiques alors que l’on dénombre entre vingt et vingt-cinq milles gènes chez l’homme.(38) Malgré l’apparente quantité de cibles potentielles à explorer, l’apparition de nouveaux médicaments est toujours plus lente et plus chère.

Lors des cinquante dernières années, la découverte de médicaments s’est grandement basée sur du criblage à haut débit permettant la découverte de petites molécules qui ont souvent la particularité d’être actives mais peu spécifiques à l’inverse des produits biologiques. En 2018, seulement 17 des 59 médicaments approuvés par la FDA étaient des produits biologiques. (38) La recherche de petites molécules médicamenteuses reste donc encore majoritaire. La période de recherche aboutissant à la découverte d’une molécule peut être divisée en quatre étapes :

- Le premier criblage et l’identification d’une touche
- Le passage de la touche au composé prometteur (lead)
- L’optimisation de ce lead
- Les tests précliniques.

Cette période dure en moyenne cinq à six ans et équivaut à environ un tiers du coût de développement d’un médicament.(38) Après cela, le composé doit passer les tests cliniques où il prouve son efficacité et son innocuité. En moyenne, de dix milles molécules criblées, dix seulement arrivent aux tests cliniques.

Après cette étape, les chances de succès sont inférieures à 10%.(38) Il y a donc un manque de précision dans la prédiction de l'effet des composés. Cela est entre autres lié au fait que les connaissances des structures 3D des cibles, des affinités de liaisons ou des propriétés cinétiques ne sont pas assez connues. Il y a un manque de ressources alors que depuis les années quatre-vingt-dix, le nombre de données enregistrées a explosé, conséquence de l'apparition de l'ère du « Big Data ». Cela est d'autant plus vrai dans le monde de la santé avec les omiques (données obtenues à partir de nouvelles techniques comme la génomique, métabonomique etc …). Le problème est qu'il y a justement trop de données, bien plus que ce qu'un cerveau humain peut intégrer en une vie. L'utilisation de l'IA est selon de nombreux spécialistes du milieu le moyen de répondre à ce type de problématique, notamment avec l'émergence du deep learning. (37)

La question est : Que peut apporter l'IA mais aussi pourquoi est-ce que cela émerge maintenant ? Tout d'abord, maintenant car les deux principales conditions sont réunies . Il y a assez de données pour entraîner les IA et les puissances de calculs informatiques sont devenues assez performantes. De plus, cela est adaptable au monde de la santé car les IA peuvent intégrer des données mesurées comme des constantes d'affinités etc. Elles peuvent aussi reconnaitre et prédire des formes, ce qui se transpose parfaitement à la réalisation de structures 3D et à la visualisation des interactions. L'intégration des IA dans les méthodes de recherche s'est faite pour développer de meilleurs moyens d'évaluation ou pour mettre au point de nouveaux biomarqueurs. Mais aussi pour identifier de nouvelles cibles biologiques, ou même mieux comprendre des voies métaboliques dans leur entièreté. On retrouve aussi des entreprises facilitant la synthèse de composé de novo et aussi de nombreuses technologies pour trouver de nouvelles indications à des molécules existantes.(37)

Les IA dans le domaine de la recherche de médicaments permettent donc de :

- **Réduire le temps de découverte et améliorer l'adaptabilité du procédé de recherche.** Cela pourrait accélérer le temps de la découverte aux tests cliniques d'un facteur 15. (38)
- **Améliorer la précision de prédiction de l'efficacité et de la non-toxicité des molécules**. Andrew Witty (ancien PDG de GlaxoSmithKLine) a déclaré : « *If you stop failing so often, you massively reduce cost of drug development* ».(39) En effet, comme dit précédemment, une molécule sur dix passe les tests cliniques car les autres ne sont soit pas efficaces soit toxiques. En comprenant mieux les systèmes biologiques, il est plus facile de se concentrer sur des molécules réellement intéressantes. Une amélioration de 10% permettrait d'économiser des milliards de dollars. (38)

- **Améliorer la diversification des Pipelines**. Trouver de nouvelles applications aux molécules permettrait de s'affranchir des tests de phase I et des tests de toxicologie. Cela aiderait les entreprises à se repositionner sur de nouveaux créneaux plus facilement.

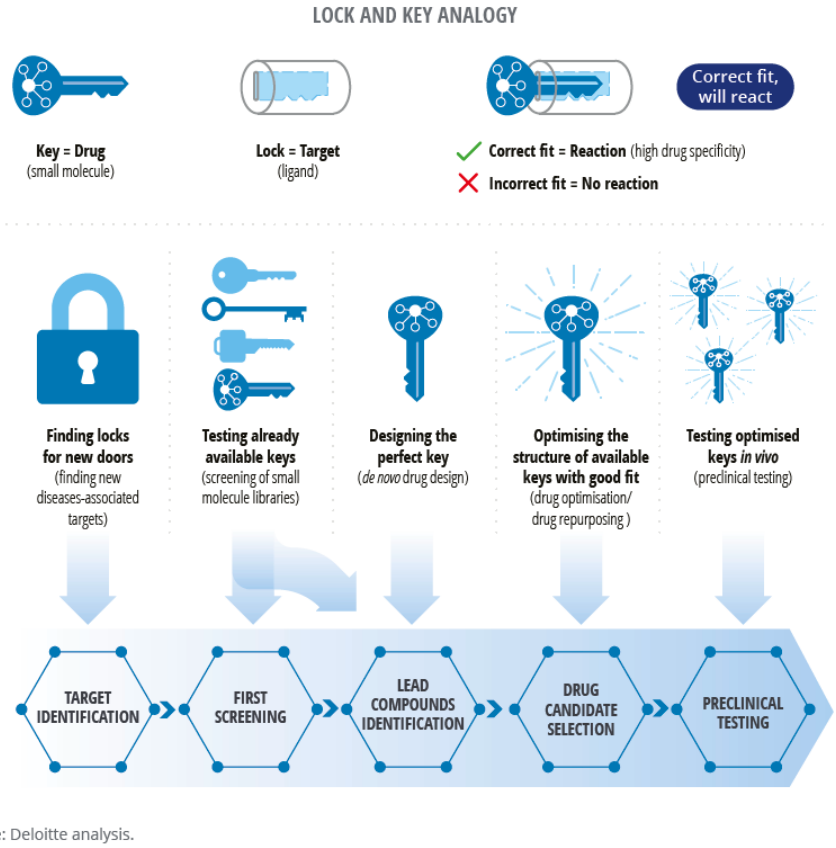

Figure 14: figure représentant l'analogie de la clé (molécule) et de la serrure (cible biologique) avec les cinq étapes où l'IA peut être utilisée = identification de la cible, premier criblage, identification d'un composé prometteur, création d'une nouvelle molécule, sélection du candidat et tests précliniques. (38)

L'utilisation des IA s'est donc majoritairement focalisée sur l'amélioration de la création de petites molécules selon l'idée qu'il faut trouver la bonne clé pour la bonne serrure, et que cela passe par cinq grandes étapes (voir figure précédente). (38) L'utilisation des IA permet ainsi de faire la chasse à la bonne molécule. En effet, il existe environ 10^{60} molécules médicamenteuses dans l'espace chimique, soit plus que le nombre d'étoiles dans l'espace. Avec les techniques actuelles, il est évident que les limites sont atteintes. L'utilisation du machine learning a donc pour but de s'affranchir de ces limites notamment en traitant le plus de données possibles pour une création plus précise de nouveau médicaments.

L'utilité des IA a mis du temps à être intégrée mais aujourd'hui comme le montre les figures 15 et 16 une multitude d'entreprises a émergée dans ce domaine . Cela démontre bien que l'intérêt scientifique et économique est important.

Figure 15: schéma de la distribution des entreprises d'AI intervenant dans le procédé de la création de médicaments(13), le focus de cette thèse sera fait sur la partie «drug design »

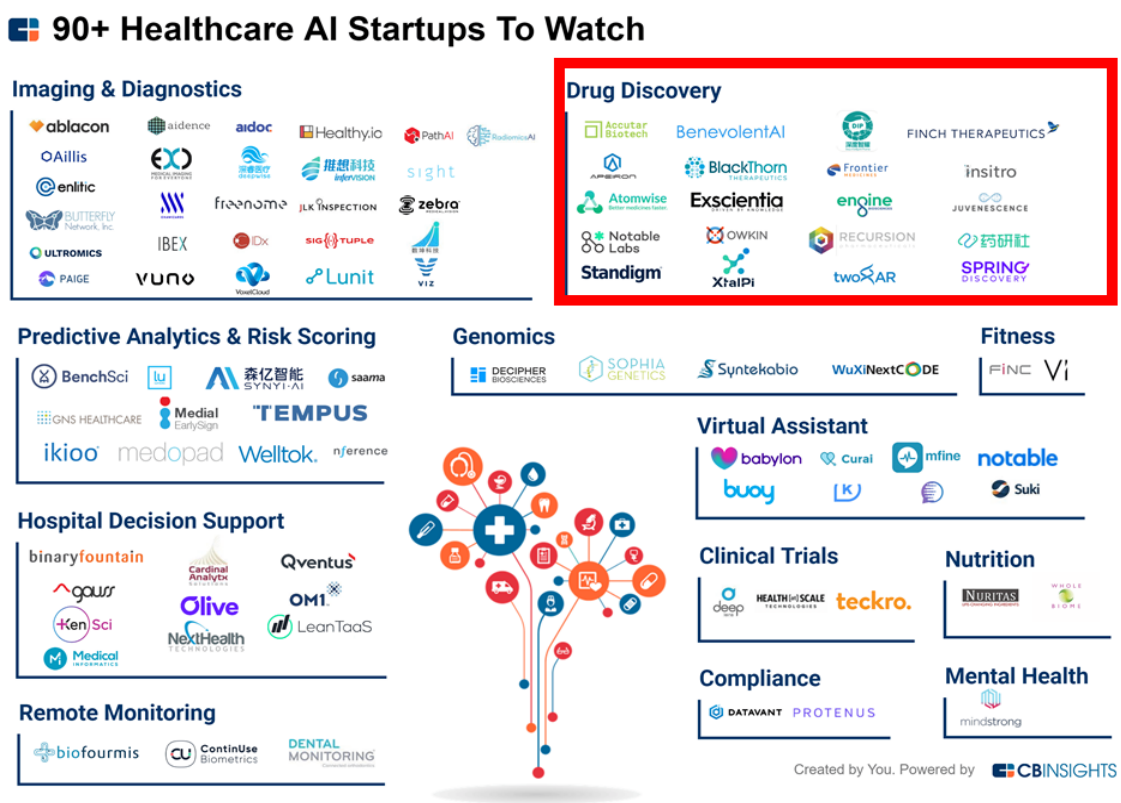

Figure 16: Second schéma représentant la distribution des entreprises d'IA intervenant dans le système de santé(24), focus de cette thèse sur la catégorie drug discovery

Comme le montrent les deux figures précédentes, il existe de nombreuses entreprises d'IA dans le monde de la santé. Les domaines possibles peuvent aller de l'amélioration des essais cliniques (qui comme vu lors de l'explication de la loi de Eroom en ont bien besoin) à la favorisation de la compliance du patient en passant par une amélioration de l'engagement du professionnel de santé. Ces aspects ne seront pas traités dans ce manuscrit mais sont évoqués en autre dans la thèse « l'intelligence artificielle : nouveau levier de croissance pour les industries pharmaceutiques ».(28)

Il est possible de classer les entreprises spécialisées dans la R&D en trois catégories importantes :

- **Drug discovery** : Identification de nouvelles cibles , sélection de composés connus pour de nouvelles application et création de novo.

- **Recherche préclinique :** Augmentation du nombre et possible remplacement de tests précliniques afin de prédire la toxicité et l'efficacité des médicaments avant de les amener à l'homme.

- **Essais cliniques :** Amélioration de la conception des essais cliniques notamment au niveau du recrutement et de répartition des patient pour obtenir une évaluation optimale à l'étape où la plupart des médicaments échouent.

Il existe dans les autres aires d'activité des entreprises d'IA très importantes qui ne seront pas vues ici telles que cyto reason, recursion pharmaceutical, biovista, owkin , Berkely lights. Ces entreprises réalisent des partenariats avec de grandes entreprises pharmaceutiques afin de faciliter le développement de médicaments mais sans pour autant être spécialisées « Drug Design ».

Dans ce manuscrit, une liste non exhaustive d'entreprises d'IA spécialisées dans le drug discovery sera réalisée. Les entreprises présentes dans cette liste ont été sélectionnées selon leur champ d'activité, et de leur importance selon certains articles (40) (41) et selon le rapport Q3 2019 du DKA (Deep Knowledge Analytics) (13). Ce dernier fait des classements des entreprises d'IA dans le monde de la pharmacie que ce soit selon leur nombre de partenariats réalisés avec les grandes entreprises pharmaceutiques, leur distribution géographique, leur fonctionnement etc.

- Numerate  (42)

Numerate est une entreprise fondée en 2007 dont le siège social est situé à San Francisco (Californie).(43) Numerate a créé un centre d'algorithme permettant de faciliter la prise de décision en pré-clinique de l'identification de la touche à la nomination d'un candidat pour les tests cliniques.

L'entreprise utilise une IA capable d'utiliser tout type de données même incomplètes afin d'amorcer l'optimisation d'un lead ou l'indentification d'un candidat et ceux en quelques mois au lieu de quelques années. Leur système n'est pas contraint d'utiliser des données exhaustives. Cela a permis à Numerate de construire un catalogue de 6000 mécanismes d'action pour plus de 2500 protéines cibles accélérant ainsi le choix de la cible biologique à viser. L'entreprise a aussi développé des modèles prédictifs pour l'étude des propriétés ADME plus performants que ceux déjà existant afin de faciliter la création de nouveaux médicaments. (44)

La plateforme de Numerate a été testée et améliorée pendant plus de 10 ans à travers 25 projets où les procédés traditionnels n'étaient pas applicables ou bien avaient échoué. Cela a permis à l'entreprise de se confronter avec succès à des problématiques industrielles notamment avec des collaborations avec des entreprises pharmaceutiques ou de biotechnologie.(44)

- Twoxar  (45)

Twoxar est une société de découverte de médicaments assistée par intelligence artificielle fondée en 2014 et dont le siège social se trouve à Mountain view (San José, Californie).(46)

Cette entreprise utilise une plateforme dirigée par l'IA afin d'identifier plus rapidement des candidats-médicaments, réduire les risques grâce à leurs algorithmes entrainés sur de nombreuses données biomédicales comme l'expression génique, les réseaux d'interactions protéiques et les dossiers cliniques. Cette technologie a permis de passer de vingt-cinq milles candidats potentiels contre le cancer du foie à dix dont le plus prometteur est entré en test clinique.

Sachant que le seul traitement pour ce type de cancer a pris cinq ans à être développé et qu'ici, il n'a fallu que 4 mois(41), on voit l'impact que peut apporter ce type de technologie.

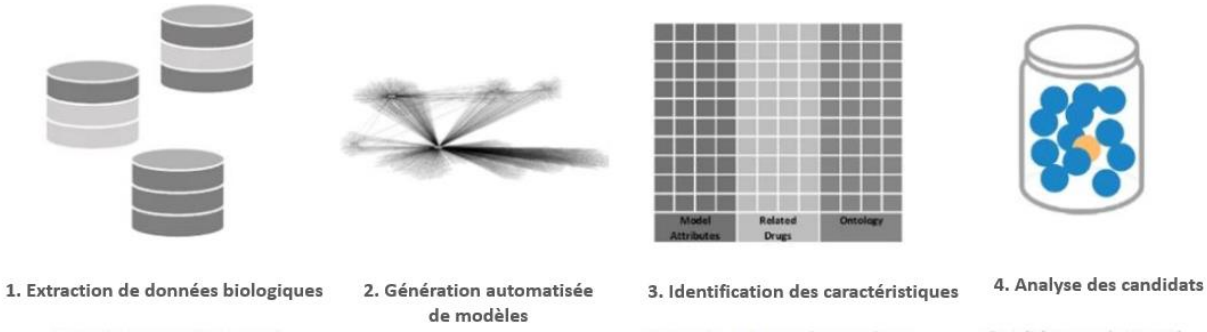

Figure 17: schéma représentant les étapes permettant au système de TwoXar d'accélérer le processus de drug discovery traduit à partir du schéma utilisé dans l'article (40)

- NuMedii

NuMedii est une société de biotechnologie spécialisée dans l'utilisation d'IA fondée en 2008 avec son siège social à San Mateo (Californie).(47)

Cette société exploite depuis 2010 les IA afin d'analyser le « Big data » en vue de découvrir rapidement les liens entre les médicaments et les maladies qu'ils traitent. Leur système a été construit à la suite de l'extraction d'un grand nombre de données sur des centaines de maladies et des milliers de composés. En effet, leur modèle couvrirait 5 milliards de faits biologiques et des trillions (milliards de milliards) de points de données (mesures diverses). Cela leur permet d'aller plus loin que l'approche par étude de la cible et facilite par la même occasion l'approche polypharmacologique.(48)

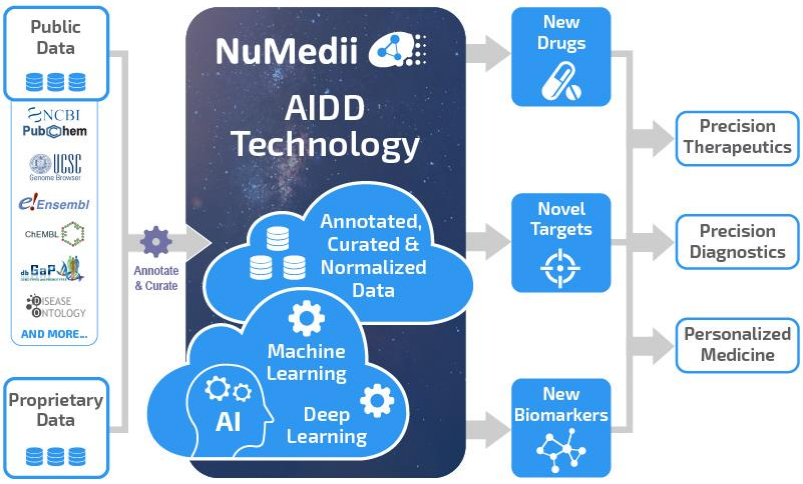

Figure 18: schéma représentant l'utilisation de l'IA chez NuMedii afin de traiter un grand nombre de données pour accélérer le processus de développement d'un médicament

L’AIDD (Artificial Intelligence for Drug Discovery) de NuMedii permet d’accélérer le processus de découverte de nouveaux médicaments mais aussi d’aboutir à une médecine plus efficace. Il est en effet considéré que 90% des médicaments les plus vendus ne traitent que 30 à 50% des patients chez qui ils sont prescrits. Leur technologie permet donc d’éviter ce type d’écart en permettant une meilleur compréhension des mécanismes mis en jeu dans les maladies.(48)

Cette société est déjà bien implantée dans le monde de la pharmacie car elle annonçait déjà sa troisième collaboration dans des projets de découvertes pharmaceutiques en 2016 (49). Aucune annonce sur des résultats n’a cependant été dévoilée à ma connaissance.

- Berg 

(50)

Berg est une entreprise de biotechnologie spécialisée dans l’IA fondée en 2006 dont le siège social se trouve à Framingham (Massachussetts).(51)

Berg utilise sa plateforme d’IA pour cartographier les maladies afin d’améliorer les traitements. Au lieu de supposer le mécanisme d’une maladie et se focaliser sur quelques composés agissant sur ces cibles, l’idée est d’analyser différents fluides et cellules de patients ainsi que les données cliniques. C’est leur plateforme bAlcis qui s’occupe de cette analyse pour créer un réseau de cause à effet, identifier les responsables de la maladie pouvant être des biomarqueurs ou de nouvelles cibles thérapeutiques. Cette technologie qu’ils appellent « interrogative biology » permet entre autres de cibler la maladie mais aussi de trouver des modèles prédictifs pour analyser et trier les patients et leur fournir le traitement le plus adapté.(52)

Les domaines ciblés par Berg sont l’oncologie, la neurologie et les maladies rares. L’entreprise possède un candidat le BPM31510 (ubidecarenone + Gemcitabine), un anticancéreux réactivant l’apoptose actuellement en phase clinique. Un autre de leurs composés est en phase clinique, le BPM31543 (calcitrol) contre l’alopécie chimioinduite. Leur technologie sert aussi à aider le diagnostic, grâce notamment à deux marqueurs biologiques (l’un pour le cancer de la prostate et l’un pour parkinson) qui sont actuellement en validation clinique.(53)

- E-therapeutics (54)

E-therapeutics est une entreprise basée à Oxford (Royaume-Uni) utilisant le machine learning pour la découverte de médicament.(55)

La technologie de l'entreprise est un modèle NDD (Network-driven Drug Discovery) permettant non pas une approche classique de haut en bas de la maladie mais de bas en haut. C'est-à-dire qu'au lieu de trouver une bonne voie pour pouvoir traiter la maladie, le système va partir de la maladie pour analyser toutes les voies mises en jeu. A partir du moment où la maladie est complètement modélisée *in silico* grâce aux banques de données, au travail de l'IA et aux mesures réalisées sur des tissus humains, le système analytique trouve les points faibles du mécanisme pathologique (là où les actions auront le plus de bénéfices). Ensuite, les composés de leur chimiothèque sont évalués contre ces points faibles afin de réduire le nombre de produits à tester.(56)

Au-delà de leur plateforme NDD, e-therapeutics a mis au point le GAINS (Genome-associated Interaction Networks). Ce système permet d'utiliser des données GWAS (Genome-Wide Association Study) mettant en parallèle les variations du génome et les maladies. Le chemin entre ces GWAS et la maladie ne sont pas simples mais grâce à leur GAINS, e-therapeutics a permis la découverte de mécanismes centraux à la maladie de Parkinson qui peuvent être utilisés par leur plateforme NDD ou pour la recherche classique.(57) L'entreprise travaille sur de multiples voies métaboliques dont deux pour lesquelles les recherches en sont à l'étape de l'optimisation du lead, il s'agit du métabolisme du tryptophane et de l'inhibition de la voie Hedgehog (impliquée entre autre dans l'apparition de carcinomes).(58)

- Verge (59)

Verge Genomics est une société de biotechnologie spécialisée dans l'utilisation d'IA fondée en 2015 avec son siège social à San Francisco Californie.(60)

Verge Genomics utilise le machine learning pour rendre plus efficace et accélérer la découverte et le développement de médicaments ciblés contre les maladies neurodégénératives (Alzheimer, Parkinson et la sclérose latérale amyotrophique). Pour cela, le système utilise une

database obtenue à partir d'échantillons de cerveaux de patients décédés de ce type de maladies. L'algorithme crible ensuite des milliers de gènes pour trouver ceux modifiés lors de la maladie. Par exemple pour la sclérose, vingt-deux nouvelles cibles thérapeutiques ont été identifiées. Grâce à ce système, Verge Genomics a permis le développement de deux possibles candidats médicamenteux.(61)

- Deep genomics 

(62)

Deep genomics est une entreprise utilisant les IA dans le domaine de la thérapie génique fondée en 2014 dont le siège social se trouve à Toronto.(63)

Cette entreprise est spécialisée dans la prédiction des altérations phénotypiques induisant des maladies génétiques et dans la création d'oligonucléotides thérapeutiques visant ces altérations. Deep genomics possède une plateforme d'IA contenant un grand nombre de données biologiques d'activité, de toxicité de tous les composés identifiés à l'aide de leur plateforme. Cela leur permet d'explorer l'espace chimique des oligonucléotides thérapeutiques qui comporte plusieurs dizaines de milliards de composés.(62) La conception intelligente de médicament que permet leur technologie consiste à prévoir les résultats dès le départ. L'un de leurs faits d'armes est leur projet Saturn où soixante-neuf milliards d'oligonucléotides ont été testés *in silico* sur un million de cibles ayant généré une liste de mille composés. Ces oligonucléotides sont testés expérimentalement afin de vérifier leur capacité à manipuler la biologie de la cellule de la façon prévue.(64)

- Atomwise 

(65)

Atomwise est une entreprise spécialisée dans le « drug discovery » et le « machine learning » fondée en 2012 et basée à San Francisco (Californie). Elle est la première entreprise utilisant le deep learning pour la découverte de molécules basée sur la structure des cibles biologiques.(66)

La technologie d'Atomwise utilise les réseaux neuronaux convolutifs, les mêmes qui permettent la reconnaissance faciale, ou qui sont utilisés dans les voitures autonomes. Leur système extrait des millions de résultats de mesures d'affinités sur des milliers de structures protéiques afin de prévoir les modes de liaison de petites molécules sur les protéines. Cela permettant d'aider le chimiste dans la recherche de touches, dans l'optimisation des Lead et dans la prédiction de la toxicité de ses composés.

Les systèmes neuronaux convolutifs sont réputés pour obtenir les meilleurs performances dans la reconnaissance d'image en hiérarchisant des données locales simples (de simples traits) jusqu'à arriver à l'image complète. Cette idée a été appliquée par Atomwise à la chimie pour permettre de prédire la bioactivité de petites molécules par la création de leur algorithme AtomNet.(67) Leur système apprend les données tri-dimensionnelles des interactions médicament-cible, identifie les conditions à remplir et sélectionne les touches correspondantes. Ces touches remplissent en plus d'autres qualités à respecter comme la capacité à passer la BHE. Cette technologie permet de gagner du temps de recherche notamment en s'affranchissant de longues phases de criblage à haut débit qui peuvent s'avérer très coûteuses. L'un des faits d'arme d'Atomwise est l'identification rapide d'un traitement possible contre le virus Ebola en partenariat avec l'Université de Toronto.(68)

- Cyclica CYCLICA (69)

Cyclica est une entreprise de biotechnologie, utilisant les IA pour améliorer le système de découverte de médicament, fondée en 2013 dont le siège social est à Toronto (Canada).(70)

Cette entreprise possède deux systèmes d'IA brevetés :

- **Ligand Express**, lancé en 2018 qui est une plateforme de criblage de molécules. Les molécules médicamenteuses sont confrontées à des cibles biologiques afin de déterminer des profils polypharmacologiques. A l'aide d'une approche basée sur le deep learning, le système identifie les protéines cibles en fonction de leur structure et détermine les effets des molécules dessus. Cette technique peut être utilisée pour identifier des effets indésirables et prédire la toxicité. Plusieurs entreprises pharmaceutiques ont utilisé ce système pour cribler leurs composés, permettant par exemple d'identifier deux voies métaboliques pouvant révolutionner le traitement d'Ebola.

- **Ligand Design**, lancé en mai 2019 qui permet la production de médicaments lead, optimisés contre plusieurs cibles pharmacologiques. La plateforme permet la création de molécules par exploration de l'espace chimique tout en respectant des critères définis au préalable. Cette plateforme a été utilisée en partenariat avec Tieös Pharmaceuticals pour créer un anticancéreux ayant une action sur plusieurs cibles tout en présentant de bonnes propriétés pharmacologiques.(24)

- Benevolent AI BenevolentAI (71)

Benevolent AI est une entreprise fondée en 2013 située à Londres.(72) Cette entreprise a intégré l'IA à chaque étape du développement de médicament, de la découverte d'une touche jusqu'au développement clinique.

Figure 19: Benevolent AI agit à chaque étape du drug design(71)

Le processus commence par l'analyse de la bonne cible biologique à l'aide de modèles d'IA identifiant des protéines ou des gènes qui ne sont pas exprimés de la même façon dans une cellule saine que dans une cellule malade. Les cibles obtenues sont triées par une autre IA selon différents critères comme entre autres la capacité à être ciblée par un médicament et la sécurité qu'il y a à les modifier. Cela permet aux scientifiques de se focaliser sur les plus prometteuses pour passer à la création de molécules.(73)

A l'aide de leur système Evochem, Benevolent AI génère de nouvelles molécules « drug-like » synthétisables avec des propriétés définies. Leur système est basé sur l'optimisation multiparamétrique et est doté d'un système de classement permettant de favoriser les molécules générées répondant aux critères voulus. Le meilleur composé est ensuite sélectionné, synthétisé et testé. Les résultats des tests sont utilisés pour améliorer le système. Cette technique permet d'accélérer le processus de sélection de candidats en passant de quatre ans et demi à quatorze mois.(74)

Benevolent AI souhaite aussi accélérer les tests cliniques à l'aide de leurs IA. En effet, comme expliqué dans le paragraphe de NuMedii, pour les maladies traditionnelles, 30 à 50% des médicaments se vendant le plus ne fonctionnent pas chez les patients chez lesquels ils sont prescrits. Leur système a pour rôle d'identifier des groupes de patients en fonction de la signature moléculaire de leur maladie afin d'identifier les plus à même de bien réagir au traitement. Cette approche permet aussi d'élucider des mécanismes d'action de médicaments existants, d'améliorer les diagnostiques et mieux cibler les traitements.(75)

Benevolent AI est donc l'une des entreprises d'IA leader dans le domaine pharmaceutique dans la création de nouveaux médicaments grâce à ses systèmes permettant l'accélération de chaque étape du développement. L'approche de cette entreprise est globale contrairement à d'autres entreprises se positionnant sur une étape précise du développement d'un médicament.

Cette liste non exhaustive est la preuve que l'utilisation d'IA peut se faire à toutes les étapes du développement d'un médicament. Il est possible de classifier ces entreprises en trois grands domaines de recherche :

- **L'analyse de données biologiques** : Numerate, Twoxar, NuMedii, Berg et e-therapeutics.

- **L'analyse du génome** : Deep Genomics et Verge genomics

- **La modélisation de cibles :** chez Atomwise et Cyclica

Il y a aussi des entreprises souhaitant agissant sur plusieurs étapes (comme Cyclica), voire sur toutes les phases de R&D comme Benevolent AI.

Certaines technologies peuvent être comparées à des « améliorations » de techniques connues telles que le docking ou l'analyse QSAR. Et il existe aussi un autre domaine qui va être traité plus en détail dans la partie suivante. Il s'agit de la création de nouvelles molécules à partir de zéro notamment grâce à l'utilisation de modèles génératifs. Cela correspond usuellement au travail du chimiste médicinal qui à partir de données biologiques, de molécules existantes, de structure 3D de cibles biologiques etc va créer des molécules novatrices pouvant à terme donner un médicament. Ces entreprises ne sont pas parmi les plus nombreuses mais comme le montre la figure suivante sont très sollicitées par les « big pharma ». Parmi elles, un focus est réalisé sur les trois plus importantes qui sont Exscientia, *In silico* Medicine et Iktos.

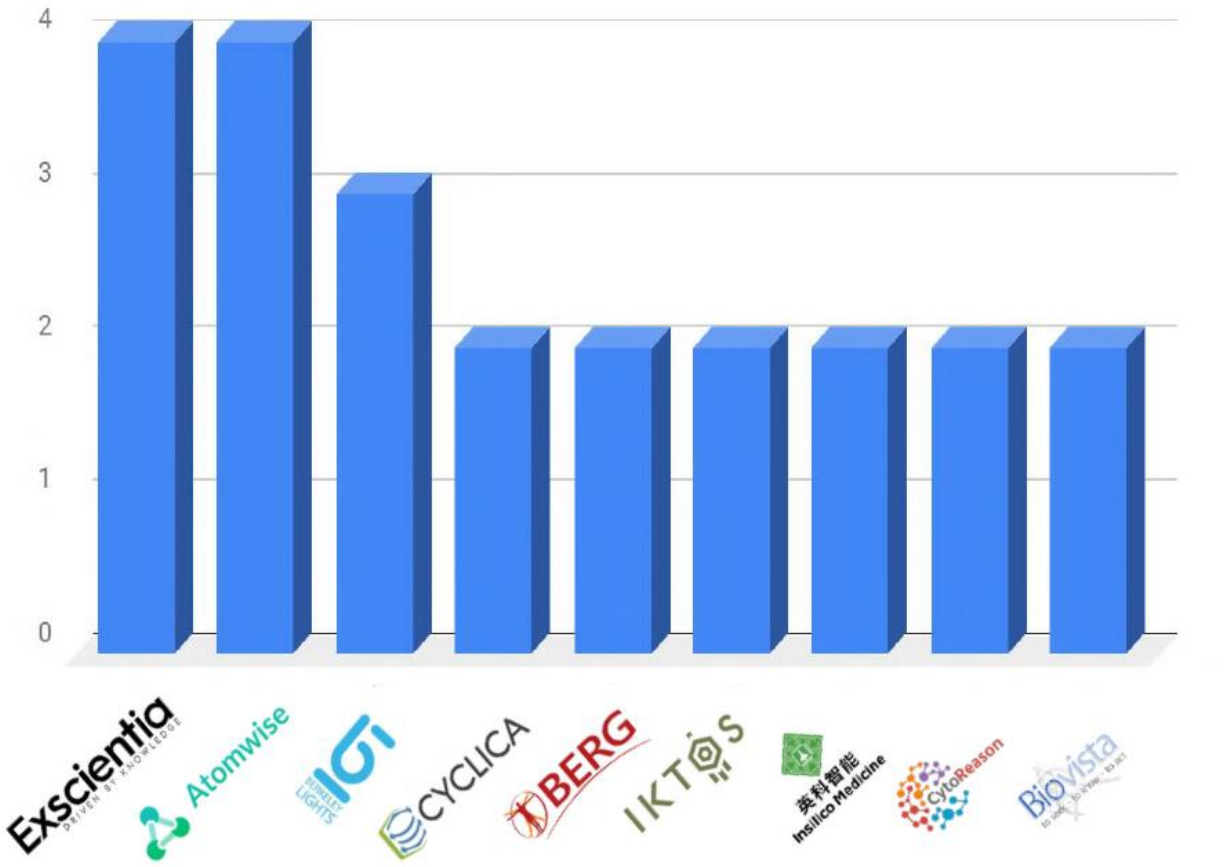

Figure 20 : Classement des entreprises d'IA par rapport à leur nombre de deals avec des compagnies pharmaceutiques. Trois parmi elles sont spécialisées dans la création de novo de molécules, Exscientia, Iktos et In silico Medicine(4)

5. Focus sur l’utilisation dans la modélisation de nouvelles molécules

5.1. Introduction

Ce chapitre fait un focus sur les entreprises d’IA utilisant le machine learning pour la création de novo de molécules médicamenteuse. Les trois compagnies évoquée (Exscientia, Insilico Medicine et Iktos) utilisent des technologies plus ou moins similaires. Elles ont le point commun d’utiliser des systèmes de récompense permettant de créer des structures innovantes répondant à des problématiques définies à l’avance. Parmi ces problématiques, ont peur retrouver : avoir une forte activité, une forte spécificité, un caractère novateur important, une faible toxicité …

Si le focus de ce manuscrit se fait sur l’exposition de technologie d’entreprises d’IA, de grandes avancées ont été apporté par des équipes académiques. C’est grâce à ce type de travaux que l’utilisation d’IA dans l’industrie pharmaceutique est devenue tangible et que le développement de ce type de compagnies est devenu possible. Il est donc important d’expliquer certains travaux académiques et de citer certaines équipes (même si la liste est loin d’être exhaustive).

Initialement les méthodes utilisées étaient basées sur le structure based drug design. C’est-à-dire que la création de ligand se fait à partir de l’étude d’un site de liaison dans la cible. La création du ligand se faisant notamment en fonction des contraintes stériques et électroniques. L’un des problème de cette méthode est que de nombreux composés obtenus s’avéraient ne pas être synthétisables ou bien n’étaient pas des molécules médicamenteuses. (76)

C’est entre autres pour cela que le ligand-based de novo drug design basé sur l’étude de ligands déjà connus s’est répandu. Pour cela, il faut utiliser un catalogue de réactions possibles ainsi que des briques élémentaires servant à la création des nouveaux composés accessibles. Cette technique a été notamment développée par Hartefeller et al qui ont créé DOGS, un logiciel permettant l’association de ces « briques ».(77) Une autre méthode développée par Besnard et al consiste à faire subir des transformations communes de la chimie médicinale à une structure pour concevoir des analogues.(78) Comme évoqué plus tard, ces travaux ont notamment conduit à la création d’Exscientia.

L’utilisation de ces méthodes est cependant limitée par leur manque de flexibilité ne permettant pas une bonne exploration de l’espace chimique. Ce problème peut être évité par l’utilisation d’une troisième méthode, la QSAR inversée. Au lieu de générer de nombreux

composés pour les classer selon leur similarité avec un composé type, la QSAR inversée cherche un moyen explicite de passer des propriétés à un descriptif de la molécules dans l'espace chimique. Par exemple, trouver le moyen qu'à partir d'une activité voulue trouver une structure moléculaire qui y réponde. L'une des difficultés de cette méthode est la représentation moléculaire à utiliser pour être pertinent. Celle s'étant le plus développée étant la représentation SMILE (inventée en 1980).

Le langage SMILE est utilisé notamment par Gomez-Bombarelli et al dans leur proposition de modèle génératif dans «Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules » (79) publiée en 2016. Ce papier présente la création d'un modèle génératif permettant l'exploration de l'espace chimique et est cité dans de nombreux travaux utilisant le machine learning. Il y est décrit une méthode d'encodage de la molécule à partir de son code SMILE suivit de l'exploration de l'espace chimique en vue de trouver une molécule répondant aux critères définis. Celle-ci sera enfin décodées pour être retraduite en SMILE afin d'être utilisables.

Ce langage a aussi été utilisé par Segler et al (80) dans leur démonstration d'entrainement d'un Recurrent Neural Network (RNN). Les RNN sont des types particuliers de réseaux neuronaux où les résultats peuvent être réutilisés comme entrées et dont l'intérêt dans la génération de molécules est grandissant. Dans ce travail, leur modèle à entre autres réussi à générer 28% de 1240 composés actif sur Plasmodium falciparum sans y avoir été confronté. Cela a apporté la preuve qu'il pouvait générer des molécules cohérentes avec une activité désirée.

Il est apparu ensuite le problème de la diversité des molécules générées. L'une des réponse à cette problématique est l'utilisation d'algorithme de renforcements. Jacques et al par exemple ont utilisé ce type de technologie pour améliorer la production de leur RNN.(81) Leur système de renforcement fonctionne en récompensant les molécules les plus médicamenteuses et les plus diversifiées. Cela permettrait ensuite d'augmenter la possibilité de générer des composés de ce type. Il faut faire attention au type de récompense que l'on utilise. Cette équipe s'est par exemple retrouvé à générer beaucoup de molécules « simples »pour répondre aux critères.(76)

Il existe bien évidemment de nombreuses autres équipes académique et de nombreuses autres thématiques. Il est impossible de toutes les traiter ici, mais les quelques sujets abordés dans ce chapitre étaient cependant importants à évoquer car au cœur des technologies des entreprises citées, preuve que le travail académique a son rôle à jouer.

5.2.Exscientia

5.2.1. Présentation de la société

Figure 21: Logo société Exscientia (82)

Exscientia est une entreprise fondée en 2012 par Andrew Hopkins. Son siège social se trouve à Oxford au Royaume-Uni. Ses bureaux sont situés à Oxford et Dundee (Ecosse).(83) Cette entreprise est spécialisée dans le de novo drug design par utilisation du machine learning. Exscientia peut être considérée comme dérivée de l'université de Dundee où Andrew Hopkins était le chef de la section informatique médicale. A cette époque, même si le deep learning commençait à se développer, il n'y avait pas assez de données pour construire des modèles de machine learning. Cependant son équipe a tout de même travaillé à l'élaboration de bases de données nécessaires à ce type de techniques. De plus, au bout d'une année, Exscientia avait publié dans Nature un papier détaillant une technique d'approche pour la création de nouvelles molécules.(78) Le travail de Hopkins ne s'est pas arrêté là car il a jugé qu'il pouvait appliquer cette méthode au milieux commercial.

Exscientia a trois domaines de recherche majeurs :(68)

- **Découverte de molécules visant une cible thérapeutique unique:** Cela en identifiant la probabilité que la dite cible se lie à une petite molécule.
- **Découverte de composés bispécifiques**: L'idée est la même que de trouver une molécule active sur une cible. La différence ici est que le médicament doit être actif simultanément sur deux cibles pour potentialiser son effet.
- **La conception phénotypique de médicaments:** Un système extrait automatiquement les indicateurs clés de performance d'une molécule. C'est-à-dire qu'il analyse les meilleures preuves de l'efficacité sur la cible donnée. Il les utilise ensuite pour orienter et optimiser les itérations de créations de nouvelles molécules pour arriver rapidement à un composé répondant aux critères voulus.

Dans ce paragraphe, le système de génération de nouveaux composés d'Exscientia est décrit grâce à la publication «Automated design of ligands to polypharmacoligical profiles » publiée dans nature(78). Dans cette publication, est expliqué comment l'entreprise utilise son IA pour générer des molécules bispécifiques. En effet, la particularité d'Exscientia n'est pas seulement l'utilisation d'IA pour accélérer la découverte médicamenteuse, mais aussi permettre la découverte de composés plus complexes. Car si trouver une molécule active sur une cible peut s'apparenter à trouver une aiguille dans une botte de foin, alors trouver une molécule bispécifique correspond à trouver l'aiguille dans la ferme tout entière.(84)

L'idée n'est donc pas de créer des molécules complexes pour seulement légitimer la technologie mais bien par utilité. Il existe en effet de nombreuses maladies qui doivent être traitées par la modification de différentes voies métaboliques, notamment pour les maladies du système nerveux central. Alors, pourquoi ne pas créer de molécules médicamenteuses agissant sur deux cibles pouvant se potentialiser ? Mais comme dit précédemment, cela est encore plus compliqué que trouver un médicament « classique », d'où l'intérêt d'utiliser le machine learning.

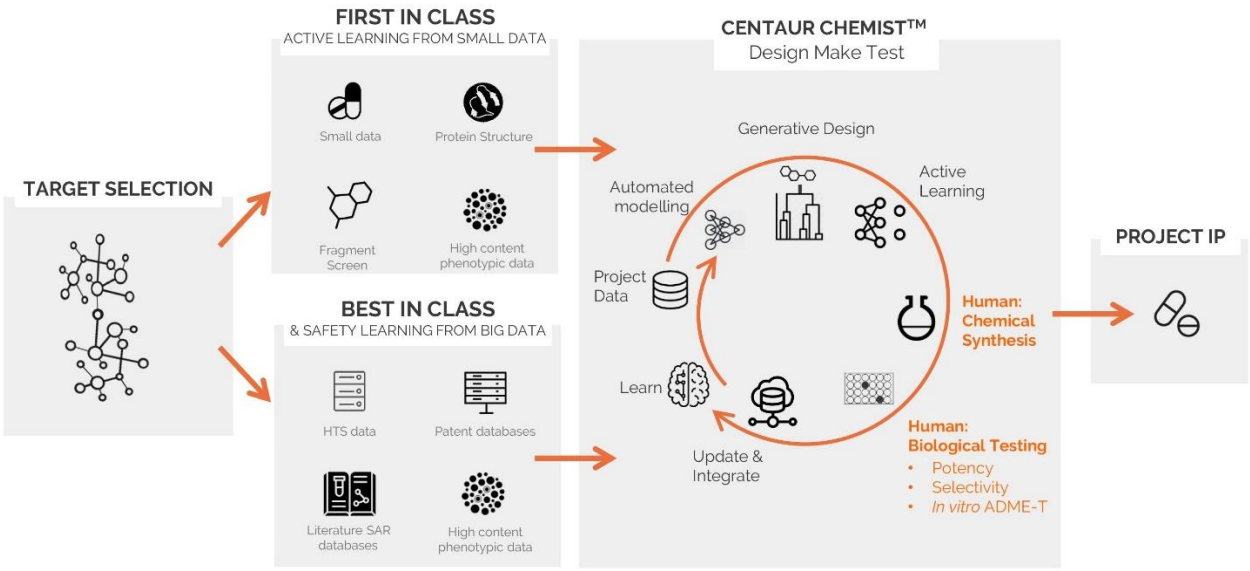

Figure 22: schéma descriptif de la technologie Exscientia, commençant par la sélection de la cible, les données "first in class" obtenues des expériences, les données "best in class" obtenues par Data mining et l'utilisation de leur modèle génératif Centaur Chemist afin d'obtenir un Hit (85)

Le système d'Exscientia est une combinaison de systèmes d'acquisition de données, de machine learning et de modèle génératif. Ils forment ensemble une plateforme d'IA permettant le passage de la touche au candidat.

5.2.2. Présentation de la technologie(78)

La méthode de génération de nouvelles molécules d'Exscientia est basée sur une déclaration de James W. Black (médecin pharmacologue écossais, prix Nobel de médecine 1988, ayant inventé le propanolol et activement participé à la synthèse de la cimétidine).(86) Pour lui, le meilleur moyen de découvrir un nouveau médicament est de commencer par un ancien. A partir de cette idée, Exscientia a mis au point un système d'intelligences artificielles permettant la création de molécules bispécifiques novatrices en se basant sur la structure de médicaments connus.

Le système part d'une molécules connue, avec une efficacité prouvée pour une cible définie. Cette molécule subit ensuite un grand nombre de changements dans sa structure afin de répondre à un second critère. Ce critère peut être d'avoir une activité sur une cible supplémentaire, différente de celle de la molécule de départ tout en conservant cette activité première. Le système mime la méthode de création utilisée par les chimistes pharmaceutiques grâce une approche basée sur les connaissances accumulées par ceux-ci. En effet, grâce au « mining » de la littérature de chimie médicinale, le système va savoir quelles modifications structurales types sont usuellement utilisées par l'homme. Celles-ci peuvent être le remplacement d'un cycle par un équivalent, l'allongement ou rétrécissement d'une chaine carbonée, le remplacement d'une fonction par une autre etc. Il modifie donc la structure comme le ferait un chimiste mais à une vitesse plus importante et sans avoir à synthétiser les composés obtenus pour avoir une idée de son activité. En effet, après chaque itération, les structures générées sont évaluées selon des modèles dits « de statistique bayésienne » servant à déterminer leur efficacité. Les statistiques bayésienne permettent d'évaluer des probabilités sur de petits échantillon, très utiles dans l'exploration de données.(87) Chaque molécule créée subit une autre itération de changement de structure et ainsi de suite jusqu'à ce qu'un composé satisfasse les critères demandés.

La publication "Automated design of ligands to polypharmacoligical profiles"(78) parue dans Nature en 2012 décrit l'utilisation de ce système à partir du Donépézil, un inhibiteur de l'acétylcholine estérase utilisé pour améliorer les fonctions cognitives dans la maladie d'Alzheimer. Des modèles de statistique bayésienne entrainés pour 784 cibles de molécules grâce à la base de données ChEMBL ont prédit que le Donépézil avait vraisemblablement une activité sur les récepteurs D4 à la dopamine et possiblement sur les récepteurs D2 à la dopamine. Cela s'est avéré être vrai une fois le Donépézil testé. L'idée a donc été de faire évoluer sa structure afin d'améliorer son activité sur D2 et de permettre sa pénétration de la barrière hématoencéphalique.

Dans cette approche, les objectifs à remplir sont définis à l'avance et exprimés dans un espace à deux dimensions. Le profil désiré est appelé «le point de réalisation idéal » et est celui qui est au maximum des deux objectifs. Chaque dimension est définie par un score bayésien pour l'activité prévue combiné à un score décrivant les propriétés ADME permettant le passage de la BHE.

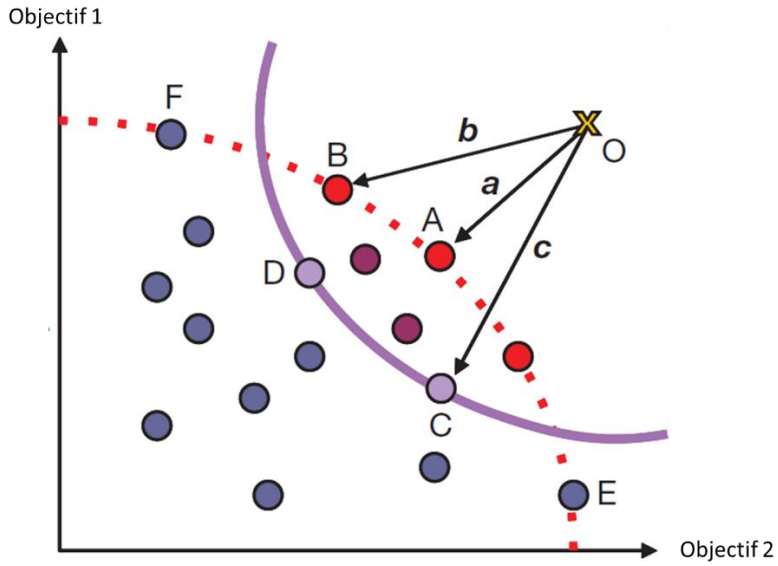

Figure 23: schéma du système d'amélioration de deux objectifs simultanément en vue d'obtenir le point de réalisation idéal (O). Le composé A étant le plus viable car le vecteur « a » le séparant de la croix est le plus petit de tous. La ligne en pointillé rouge est l'optimum de Pareto(88) où l'on ne peut plus améliorer un objectif sans diminuer l'autre.

Des séries de transformations sont appliquées au Donépézil. Les prédictions d'activité sont ensuite calculées pour chaque composé grâce aux modèles bayésiens. Chaque structure ainsi évaluée est ensuite classée et positionnée dans le plan multidimensionnel afin d'être comparée au « point de réalisation idéal ». Les composés sont ensuite filtrés selon leur nouveauté (par rapport à ceux déjà existant), le respect de la règle de Lipinski et leur faisabilité chimique. Les 10 000 meilleurs plus 500 autres composés sont utilisés pour le cycle suivant. Ces étapes sont répétées jusqu'à l'obtention d'une structure très proche du point idéal ou bien lorsqu'il n'y a plus d'amélioration obtenue après chaque itération.

Comme le montre la figure suivante, une focalisation s'est faite sur une série d'isoindoles permettant d'aboutir à la création de 8 composés dont celui appelé « composé 3 » dans la publication. Cette molécule s'est avérée présenter une forte affinité pour le récepteur D2. L'utilisation de l'IA a donc permis d'obtenir un composé bifonctionnel avec une activité d'agoniste

inverse de D2 et une activité agoniste sur D4 tout en étant capable de passer la BHE.

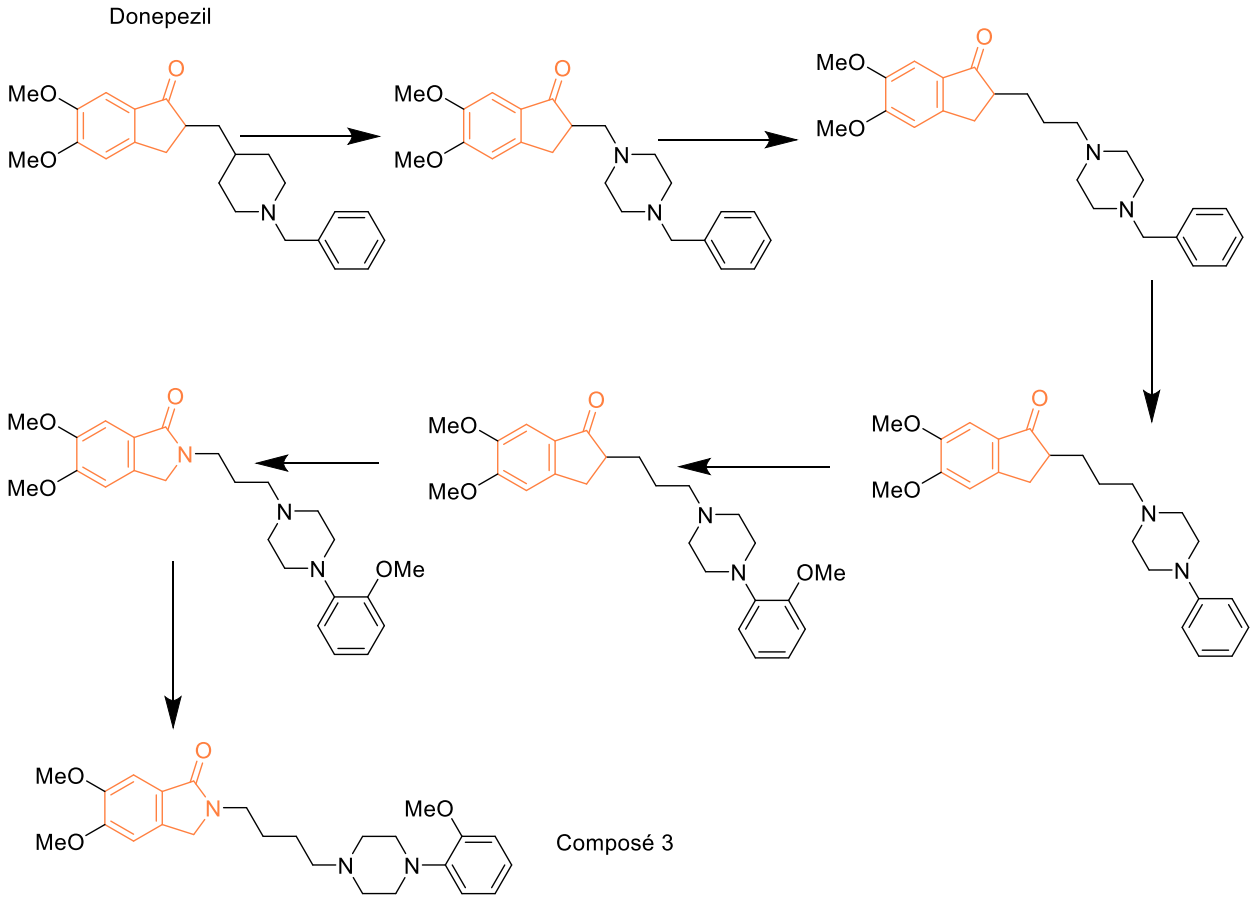

Figure 24: schéma décrivant les évolutions du Donépézil avec pour objectif d'augmenter l'activité sur D2 ayant permis l'obtention du composé 3

La création de nouvelles structures ne s'arrête pas là. Le système est ensuite utilisé pour faire évoluer les 8 isoindoles afin de diminuer leur activité sur les récepteurs α1 adrénergiques tout en conservant la capacité à traverser la BHE. Cette évolution a mené à la formation de benzolactames. Les composés ayant les prédictions d'activité sur α1 les plus basses se sont bien avérés être ceux présentant le moins d'affinité pour ce récepteur un fois testés. Pour les autres composés, on retrouve aussi une confirmation des prédictions par les tests d'activité. Le bon type de structure a été favorisé alors que les benzolactames n'étaient pas présents dans la base de données ChEMBL utilisée pour mettre au point le modèle bayésien. Le modèle est donc capable de générer et favoriser de nouveaux composés.

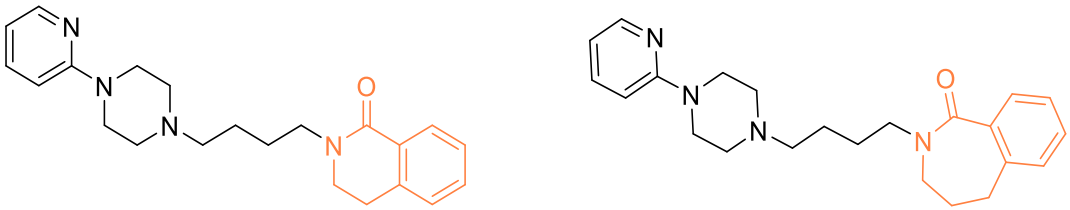

Figure 25: composés obtenus ayant la plus petite affinité pour D2 avec cycles benzolactames en orange

Le système a ensuite été utilisé pour améliorer la sélectivité tout en partant de nouveau du Donépézil. Cela passe par l'amélioration de l'activité sur le récepteur D4, de la pénétration de la BHE et de la sélectivité pour D4. Après six générations, deux composés sont retenus. Celui avec le meilleur rang était inactif, tandis que l'autre appelé « composé 13 » s'est avéré être le plus actif de tous les composés. Ce composé, en plus d'être actif sur D4 présente une faible affinité pour cinq autres RCPG testés et présente une bonne pénétration de la BHE. Le composé 13 a été testé sur souris, sa grande sélectivité a été démontrée même si elle n'est pas absolue.

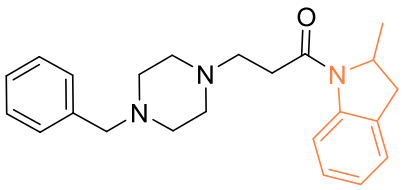

Figure 26: Composé 13 obtenu avec cycle 2,3-dihydro-indol-1-yl en orange

Le composé 13 a ensuite été utilisé comme point de départ pour être évolué en des composés encore plus sélectifs pour D4, avec une meilleure pénétration de la BHE et avec un autre chémotype. Cela a donné des composés de type Morpholino. Un nouveau type de molécule a été obtenu alors que la plupart des composés agissant sur le récepteur D4 sont des pipéridines et des pipérazines 1,4 disubstituées (1,4-DAPS), structures couramment retrouvées chez les ligands des récepteurs aux amines biogènes. Vingt-quatre de ces composés ont été synthétisés et testés. Il ressort de ces tests une très grande sélectivité de ces composés pour le récepteur D4 et peu d'affinité pour les autres récepteurs testés. Deux composés sont ressortis comme étant des têtes de série remplissant tous les objectifs (forte affinité pour D4, excellente sélectivité, pénétration de la BHE et caractère novateur).

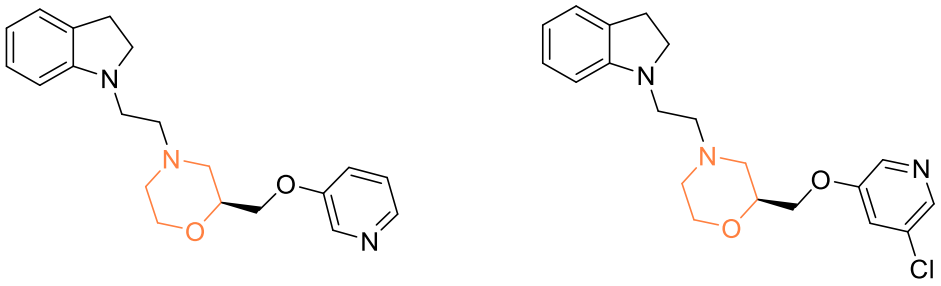

Figure 27: Composés tête de série obtenus avec cycle morpholino en orange

Exscientia a ainsi réussi à prouver l'utilité de son modèle pour générer de nouvelles têtes de série plus efficaces que celles connues pour une cible donnée. L'entreprise a de plus démontré sa capacité à orienter la transformation du composé de départ afin de remplir les objectifs voulus. Elle l'a fait notamment en créant un composé actif sur deux cibles biologiques différentes, la neurologie étant de plus un domaine où cette méthode est particulièrement intéressante. Cette technique est applicable pour tout type de molécule de départ et pour toutes les cibles médicamenteuses. Il y a cependant des contraintes comme l'obligation d'avoir assez de données sur les relations structure / activité afin de créer des modèle bayésiens utiles.

5.2.3. Le futur de l'entreprise

Le CEO Hopkins souhaite que Exscientia devienne une entreprise pouvant identifier les cibles, créer les molécules et les tester en clinique. Pour le moment, Exscientia a perfectionné sa technologie pour les étapes de drug-design, laissant la synthèse et les tests aux chimistes, mais Hopkins compte à terme automatiser ces étapes.(89)

Exscientia est l'entreprise d'IA ayant le plus de collaboration avec des entreprises pharmaceutiques.(13) Parmi elles, une collaboration avec Sanofi qui a donné le jour à un composé bispécifique ciblant deux voies de signalisation de l'inflammation et du développement de la fibrose entrant en tests cliniques. On retrouve aussi une collaboration à 33 millions de livres avec GSK en 2017 qui a permis la création d'un traitement contre la bronchopneumopathie chronique obstructive. Actuellement, la société possède un pipeline d'environ 20 composés, venant de collaborations ou de projets personnels dont un entrant en phase I.(90) Il est intéressant de noter que 5 ont été délivrés en 14 mois contre les 5 ans habituels. Avec notamment une économie de 80% réalisés sur les coûts de découverte de composés et 30% sur le total du développement du médicament. Les améliorations apportées par cette méthode sont conséquentes, mais Exscientia ne compte pas s'arrêter là et prévoit de doubler son pipeline d'ici 2020 – 2021.(68)

5.3.Insilico Medicine

5.3.1. Présentation de la société

Figure 28: Logo société Insilico Medicine(91)

Insilico Medicine est une entreprise de biotechnologie spécialisée dans l'intelligence artificielle dont le siège social actuel est basé à Hong-Kong. Elle a été fondée en 2014 à Rockville (Maryland) par Alex Zhavoronkov, spécialiste en biotechnologie. Cette société a pour ambition d'augmenter l'espérance de vie humaine. Cela passe par un changement des étapes de recherche et de développement des médicaments grâce à la découverte de biomarqueurs, l'amélioration des techniques de drug design, la médecine digitale et des recherches sur le vieillissement. Insilico Medicine a axé ses recherches sur les maladies d'Alzheimer et de Parkinson, la sarcopénie, le cancer, la fibrose, les maladies métaboliques et dermatologiques et enfin la sénescence, c'est-à-dire des maladies dominantes chez les populations vieillissantes. Cette entreprise qui était la première à utiliser l'association de modèles génératifs et de renforcement est considérée comme la référence des entreprises d'intelligence artificielle pour la découverte de médicament. Elle a en effet été placée par NVIDIA en 2017 dans le top 5 des entreprises d'IA pour son potentiel d'impact social, en 2018 par CB insights comme l'une des 100 meilleures entreprises d'IA au monde et enfin, a reçus le « Frost & Sullivan 2018 North American Artificial Intelligence for Aging Research and Drug Development Award ».(92) (93)

En 2015, un évènement a bouleversé le monde du deep learning. Cet évènement est la création de l'IA Alphago par l'équipe Deepmind de Google capable de battre des champions de jeu de go. C'est en 2017 que l'IA bat le champion du monde du titre en trois manches.(94) Un évènement similaire appelé le «pharma's alphago moment» a eu lieu en 2019 lorsque la publication «Deep learning enables rapid identification of potent DDR1 Kinase inhibitors»(95) voit le jour dans Nature Biotechnology. Il y est décrit le nouveau système développé par Insilico Medicine permettant la création et la validation de nouveaux médicaments potentiels en 45 jours. Cette publication sera discutée lors du chapitre suivant décrivant la technologie utilisée par Insilico Medicine.

L’entreprise développe des modèles génératifs et de renforcement depuis 2015 et a publié plus de 330 papiers scientifiques cités plus de 2300 fois. Malgré cela, aucune molécule imaginée par une IA n’a jusque-là était mise sur le marché. Mais ce travail réalisé par Insilico Medicine, est ce qui se rapproche le plus de l’utilisation industrielle des IA. Il montre entre autres que les molécules obtenues sont actives *in vitro* et *in vivo*. La méthode mise au point pourrait réduire le taux d’échecs précliniques de 99% et accélérer le temps nécessaire pour passer de l’étape de recherche et développement au traitement réel.(93)

5.3.2. Présentation de la technologie(95)

Comme évoqué précédemment, en 2019, Insilico Medicine a créé le « pharma’s alphago moment » en publiant dans Nature Biotechnology le papier suivant «Deep learning enables rapid identification of potent DDR1 Kinase inhibitors »(95). Dans cette publication est décrite l’utilisation de leur nouvel outil utilisant l’IA appelé le Generative Tensorial Reinforcement Learning (GENTRL) leur ayant permis de concevoir six nouveaux inhibiteurs de DDR1 (Protéine kinase impliquée en autre dans le mécanisme de la fibrose). La conception de ces six composés s’est faite en 21 jours, ce qui une avancée majeure dans ce domaine.

Avant d’en arriver à ce succès, Insilico Medicine a développé de nombreux GANs et système de renforcement pour générer de nouveaux principes actifs théoriques. Comme expliqué précédemment, les GANs permettent de générer des images avec des propriétés spécifiques. Dans le modèle présenté ici, deux systèmes de deep learning fonctionnent en étroite relation. L’un génère les nouvelles molécules répondant à certains critères précédemment définis et le second, le discriminateur évalue les composés générés en attribuant des récompenses. Ces récompenses dépendent de la capacité du composé généré à remplir les critères imposés.

Il faut bien comprendre que comme le montre la figure 29, il n’y a pas une seule méthode pour générer les molécules et pour les évaluer. Différentes techniques ont pu être testées et publiées par Insilico Medicine (96) (97) mais c’est bien le modèle GENTRL qui leur a permis leur exploit.

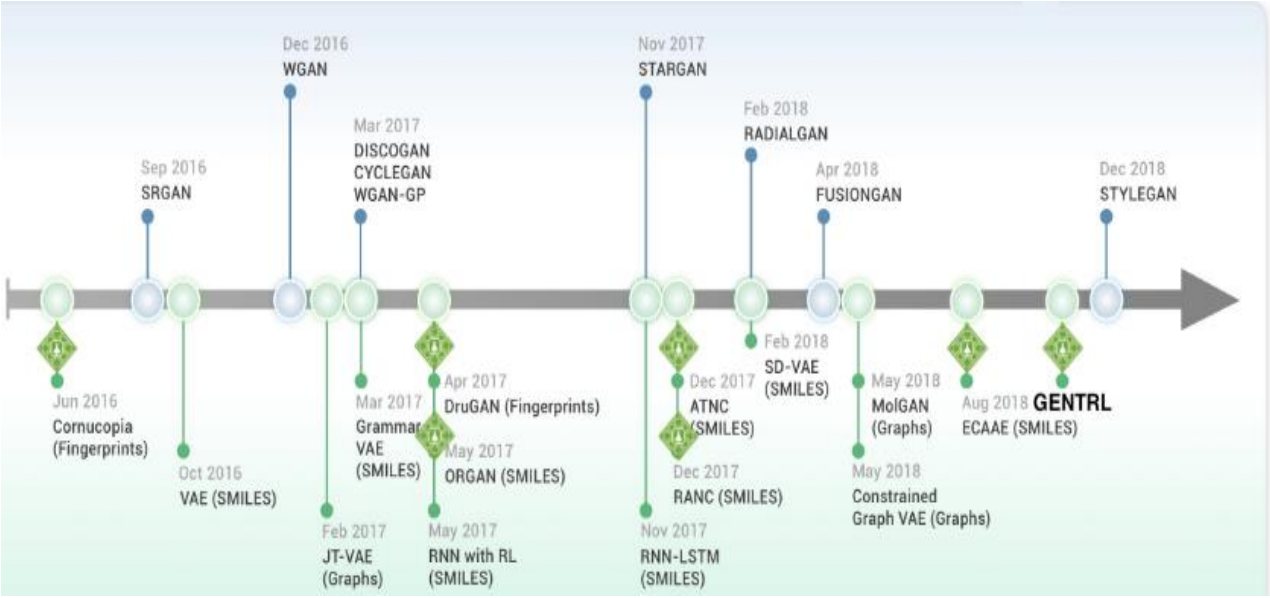

Figure 29: Timeline exposant les différents modèles de GANs généraux en bleu et ceux servant à la création de médicament en vert. Focus de ceux développés par Insilico Medicine avec un détail des moyens de descriptions des molécules utilisés(93)

La création du GENTRL a commencé par l'apprentissage de l'espace chimique. Cela s'est fait à partir de graphes moléculaires ayant permis d'arriver à un espace à 50 dimensions. Un auto-encodeur permet de répartir les molécules dans l'espace chimique en fonction de critères définis. Cette phase est indiquée dans l'espace « learning the chemical space » dans la figure suivante.

La seconde étape consiste à explorer cet espace chimique afin de découvrir de nouveaux composés. Les composés générés sont évalués afin de garder les plus intéressants et orienter le générateur à l'aide d'un système de récompense. Pour ces récompenses, le GENTRL utilise des cartes auto-adaptatives CAA (self-organizing maps = SOM).

- **La CAA de tendance** (the trending SOM): Récompense la nouveauté des composés selon leurs dates d'apparition dans des brevets.
- **La CAA des kinases communes** (the general kinase SOM): Différencie les composés inhibiteurs de kinases des autres classes médicamenteuses.
- **La CAA des kinases spécifiques** (the specific kinase SOM): Isole les inhibiteurs de DDR1 des autres inhibiteurs de kinases.

Le modèle a été construit à partir de 6 bases de données prétraitées afin d'éviter les résultats aberrants et les composés trop similaires:

1- Une grande quantité de molécules venant de la base de données ZINC.
2- Des inhibiteurs de DDR1 connus.
3- Des inhibiteurs communs de kinases (ensemble positif).
4- Des molécules agissant sur d'autres cibles que des kinases (ensemble négatif).
5- Des molécules biologiquement actives brevetées par des entreprises pharmaceutiques.
6- Des structures 3D d'inhibiteurs de DDR1 :

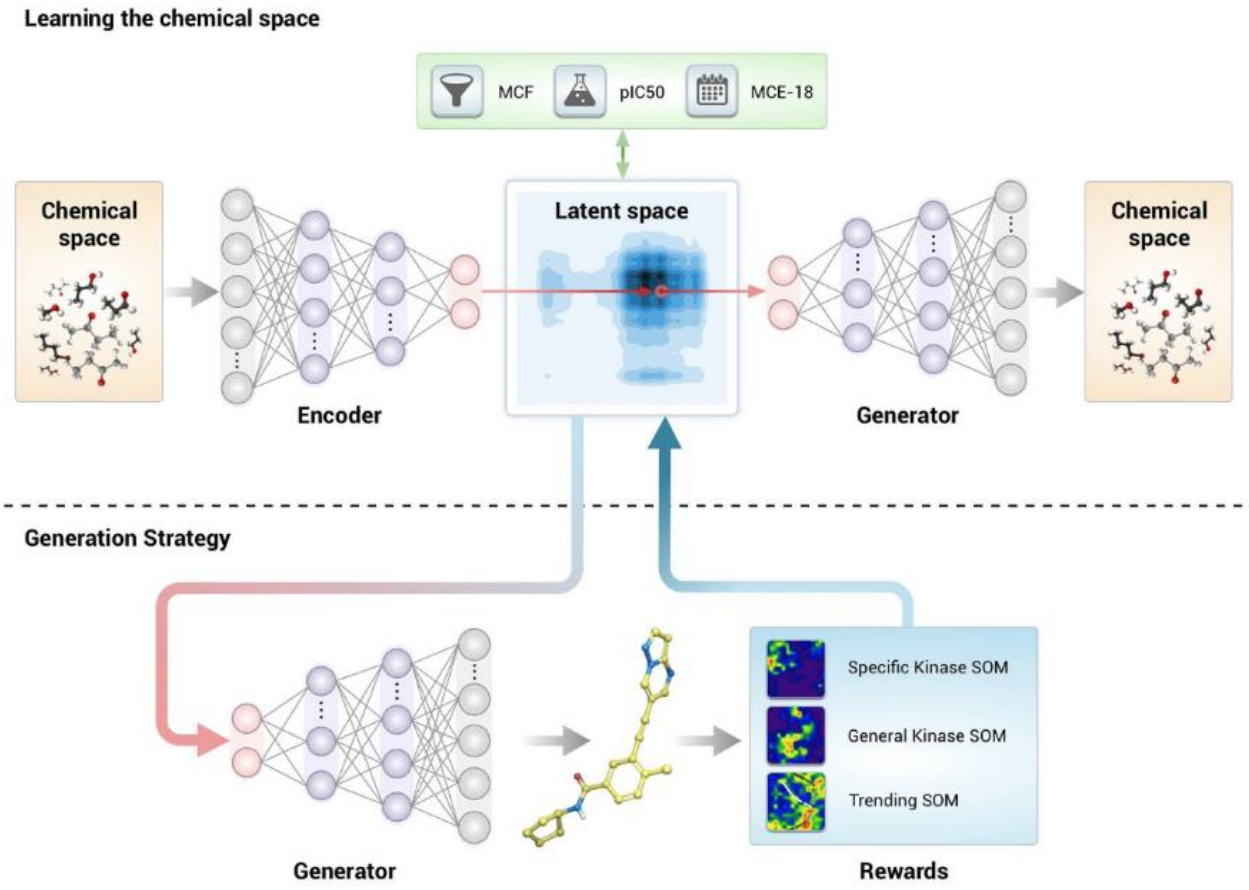

Figure 30: représentation graphique de l'utilisation du modèle génératif de Insilico Medicine permettant la conception de molécules avec des caractéristiques spécifiques(93)

L'entrainement du GENTR a commencé sur les bases de données 1 (ZINC) la plus générale et a continué sur les bases 2 et 3. Ceci est la phase de pré-entrainement. Le renforcement fait ensuite son apparition grâce au système de récompenses. Ce sont finalement 30 000 structures qui sont obtenues et traitées afin d'éliminer celles contenant des anomalies structurales ou des groupes réactifs. Les structures ont ensuite été évaluées par les CAA 2 et 3 (kinases communes et spécifiques) et grâce à des modèles de pharmacophore obtenus à partir de structures cristallographiques de composés complexés à DDR1. A l'aide de descripteurs moléculaires et de calculs de déviation, la distribution de ces molécules dans l'espace chimique a été diminuée (pour

information, la technique est dite « Sammon mapping »). Mais il y a encore à ce stade bien trop de molécules. Quarante d'entre elles occupant la globalité de l'espace chimique ont donc été choisies au hasard. Trente-neuf n'étaient pas brevetées. Parmi celles-ci, six ont été choisies par rapport à leur faisabilité chimique.

Comme le montre la figure suivante, ces 6 candidats principaux ont été obtenus au bout de 23 jours. Au jour 35 ils étaient synthétisés. Parmi ceux-ci, les composés 1 et 2 ont démontré de fortes activités lors des test *in vitro* (IC50 = 10 et 21 nM), les composés 3 et 4 avaient une activité moyenne (IC50 = 1 µM et 278nM) et enfin les composés 5 et 6 étaient inactifs. Les composés 1 et 2, en plus d'être actifs s'avèrent être sélectifs de DDR1 par rapport à DDR2.

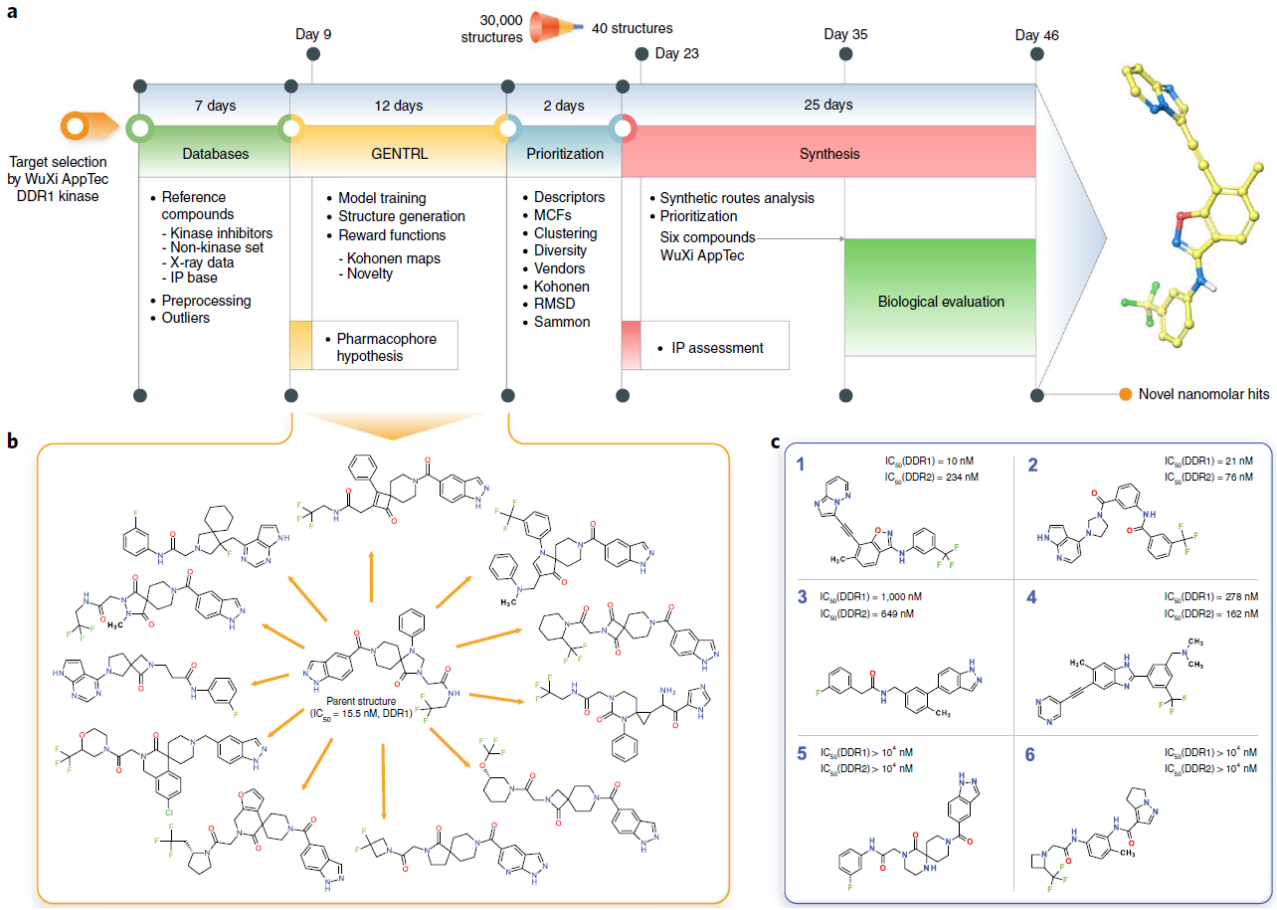

Figure 31: description du passage de la sélection de la cible à l'obtention d'une molécule hit. a, Le flux de travail général et le calendrier pour la conception des principaux candidats utilisant GENTRL .b, Exemples représentatifs de structures générées par rapport à l'inhibiteur de la DDR1 kinase parent. c, Composés générés avec la plus forte activité d'inhibition contre la DDR1 kinase humaine.(95)

L’activité des composés 1 et 2 a ensuite été établie par des tests cellulaires. Ils avaient notamment des valeurs d’IC50 très encourageantes.

Ces composés ont ensuite passé avec succès les tests de stabilité et s’avéraient avoir de bonnes propriétés physicochimiques en remplissant notamment la règle de Lipinski.

Les composés ont enfin été testés sur souris et ont présenté de bonnes propriétés pharmacocinétiques.

Comme le montre la figure 31, l’utilisation du GENTRL a permis la création, la synthèse et la validation expérimentale de deux molécules actives sur un site d’intérêt en moins de deux mois contre en moyenne trois ans pour les méthodes traditionnelles. Ces molécules ont de plus l’avantage d’être innovantes et non brevetées. Enfin, ces deux composés ont été obtenus après n’avoir synthétisé et testé que 6 molécules.

Les gains économique et de temps sont donc indéniables. Il faut cependant retenir que les molécules obtenues s’avèrent être de possibles candidats médicamenteux mais n’ont pas été optimisées. Il est probable que la sélectivité, la spécificité ainsi que d’autres propriétés pharmacologiques puissent être améliorées. Cependant, l’activité en tests pré-cliniques et cliniques n’a pas encore été évaluée et comme vu précédemment, c’est un passage couperet dans le développement de médicament.

L’amélioration des techniques de recherche que permet Insilico Medicine est indéniable, mais comme pour toute innovation, il faut garder du recul. Cette technique, bien qu’elle permette de gagner du temps, ne répond qu’à une partie des problématiques de la création d’un médicament. En effet, la cible biologique doit être définie et il faut des connaissances sur cette cible pour établir les modèles de récompense etc. Enfin comme pour tout médicament, ce ne sont que les tests cliniques qui affirmeront ou non l’intérêt des molécules obtenues de cette façon.

5.3.1. Le futur de l’entreprise

Le CEO, Alex Zhavoronkov a déclaré qu’il considérait que la plateforme d’IA de Insilico Medicine s’était améliorée au point de pouvoir automatiser tout le parcours de développement d’un composé. Il pense pouvoir créer et produire un médicament commercialisable de bout en bout en 24 mois (ou moins).

Comme démontré dans la partie précédente, Insilico Medicine s'est pour le moment focalisée sur le développement préclinique des médicaments, par l'identification des cibles et le développement de molécules thérapeutiques. Mais la société travaille actuellement sur des modèles de prédiction de résultats d'essais cliniques. C'est pour cela que Zhavoronkov cherche à s'associer avec une grande société pharmaceutique dans le cadre d'un défi type XPRIZE où Insilico Medicine serai dans une course contre la montre pour créer le premier médicament entièrement développé par intelligence artificielle. De plus, Zhavoronkov déclare que même si aucun candidat n'en sort, ce type de découverte verra malgré tout le jour dans 4 à 5 ans.(98)

Pour finir, on peut dire que grâce à ses résultats, Insilico Medicine est une entreprise qui attire l'attention. Elle collabore en effet avec plus de 150 partenaires qu'ils soient académiques ou industriels. En 2019 cette entreprise a réussi à soulever plus de 51.3 millions de dollars de financement, preuve que beaucoup d'institutions de santé croient en l'apport représenté par l'utilisation de l'IA dans le processus de recherche de médicaments.(68)

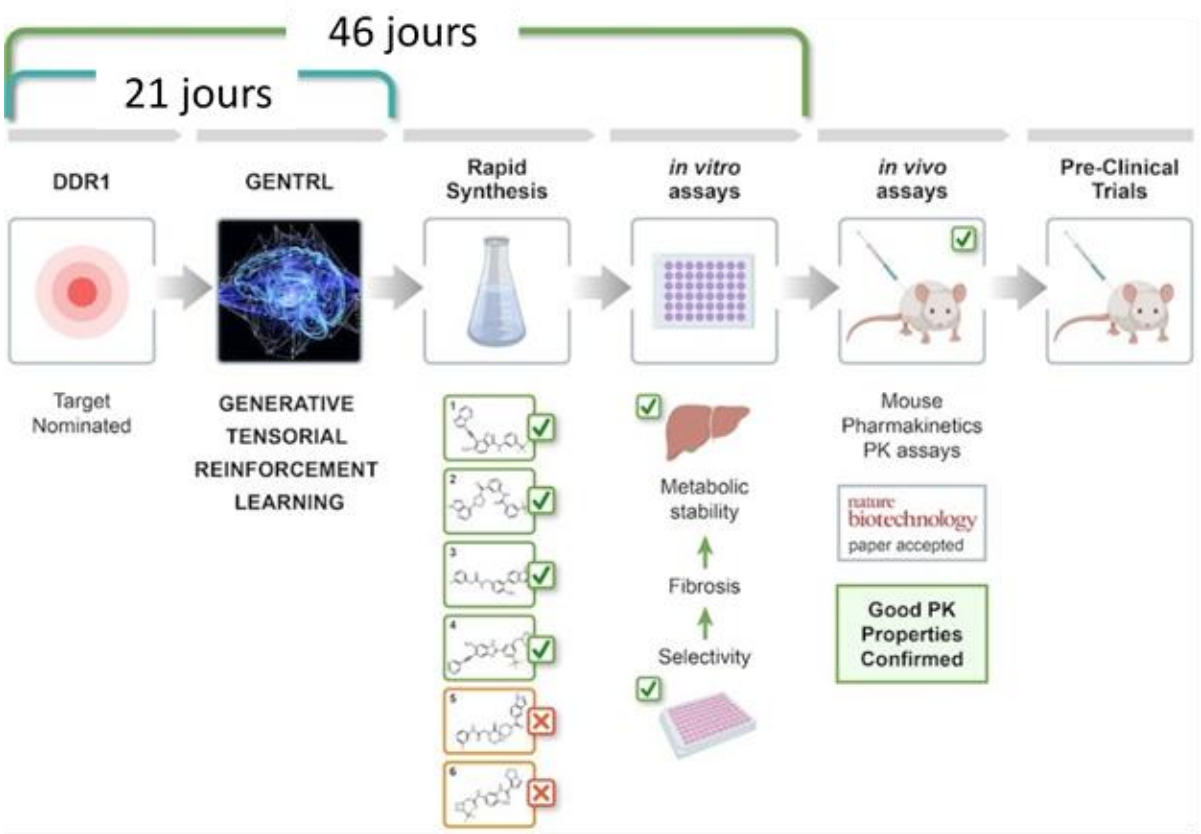

Figure 32: passage du Hit au lead en moins de deux mois à l'aide du GENTRL contre 2 à 3 ans par les méthodes traditionnelles(93)

5.4.Iktos

Figure 33: Logo société Iktos(99)

5.4.1. Présentation de la société

IKTOS est une entreprise française spécialisée dans l'intelligence artificielle appliquée à la conception de nouveaux médicaments. Elle a été fondée en 2016 par trois personnes venant d'horizons variés mais aux profils complémentaires. Yann Gaston-Mathé scientifique spécialisé dans la R&D pharmaceutique, Nicolas Do-Huu spécialiste en deep learning titulaire d'un PhD en intelligence artificielle et Quentin Perron, PhD en chimie organique et spécialiste en chimie informatique.(100)

Comme expliqué précédemment, l'innovation en chimie médicinale s'est complexifiée ces dernières années. Il est notamment long et couteux de développer une nouvelle molécule sans pour autant être satisfait du résultat final. Une des causes de ce problème est la difficulté d'optimiser une molécule d'intérêt. En effet, cette étape peut être assimilée à la résolution d'un Rubik's cube. La première face complétée correspondant à l'activité de la molécule. C'est cependant loin d'être le seul critère à remplir, il faut réussir à compléter les autres faces. En effet la molécule doit être biodisponible, non toxique, spécifique de sa cible etc. Or, quand la molécule est modifiée en vue d'améliorer un des critères, les autres peuvent en pâtir, exactement comme lorsqu'une face d'un Rubik's cube est défaite lors de la réalisation d'une autre.(101)

Pour permettre l'optimisation multicritère, IKTOS a développé un modèle génératif particulier. Celui-ci est capable de créer des molécules virtuelles d'intérêt répondant à des critères prédéfinis et ce en quelques heures. S'il est possible de créer des modèles génératifs capables de produire des images de chats, ou des visages humains totalement artificiels(102), alors pourquoi pas des molécules médicamenteuses. C'est cette idée qui est à l'origine de ce projet.(103)

5.4.2. Présentation de la technologie

Lors de la création de cette entreprise, différentes équipes académiques avaient déjà travaillé et publié sur la création de modèle génératif de nouvelles molécules. Les sujets abordés étaient divers, du langage à utiliser (fingerprints ou langage SMILE) à l'utilisation d'Autoencoders, avec chaque équipe ayant son algorithme. Cependant, les molécules créées n'optimisaient généralement pas plus de cinq paramètres et les modèles génératifs n'étaient pas confrontés à de vrais projets industriels.

La technologie d'IKTOS repose sur l'utilisation de deux algorithmes. Le premier est un modèle génératif entrainé sur plus de quatre-vingt-six millions de molécules tirées de la base de données publique ChemBL. Le deuxième est un système de renforcement qui a pour rôle d'évaluer les molécules générées selon les critères à respecter pour le projet. Les molécules analysées se voient attribuer un score en fonction de leur capacité à répondre aux attentes. Elles sont ensuite réintroduites dans le modèle génératif avec un poids qui a une influence proportionnelle au score obtenu. Les molécules remplissant les critères et ayant donc un bon score vont beaucoup influencer le modèle génératif par rapport aux autres. Cela permettant la création de molécules situées dans la même partie de l'espace chimique remplissant de mieux en mieux les critères choisis après chaque itération. Les déviations sont corrigées par l'attribution d'un mauvais score aux molécules ne remplissant pas les critères définis. Le modèle génératif de ce système est assimilable à un opérateur qui fait ce que le chef d'orchestre lui dicte. Ce chef d'orchestre est le programme de renforcement qui lui, reçoit une claque ou un bon point en fonction de ce qu'il arrive à produire, ce qui le pousse à faire attention à ce qui est généré.(103)

Ce modèle a ensuite été testé. Pour cela, une base de données de 596 molécules a été utilisée. L'activité de ces composés structurellement proches a été évaluée sur deux cibles biologiques (mTor et Pi3k). Sur un graphique avec en ordonné l'activité sur mTor et en abscisse l'activité sur Pi3k, les molécules se répartissent selon une forme assimilable à un rectangle. Trois base de données ont été créées en enlevant à chaque fois un des angles de ce rectangle (sauf celui en bas à gauche n'étant pas utilisé car correspondant aux molécules peu actives sur les deux cibles). L'algorithme est ensuite testé sur sa capacité à générer des molécules similaires à celles enlevée (correspondant aux angles) en fonction des prédicteurs utilisés. Ce test fut une réussite, l'algorithme étant capable de générer des composés actifs sur une des deux cibles et sur les deux cibles à la fois. Ce test reste cependant est encore loin d'une utilisation réelle sur un projet industriel (104)

Le premier projet industriel sur lequel IKTOS a communiqué est la collaboration avec les laboratoires Servier dont le résultat a été dévoilé en septembre 2018.(105) Cette alliance fait suite à 10 années de recherches des laboratoires Servier au cours desquelles 881 molécules ont été synthétisées en vue de répondre à 11 critères.(106)

- L'activité sur la cible
- Six tests d'activité sur d'autres cibles afin d'évaluer la sélectivité des molécules.
- Deux critères de perméabilité.
- Un critère de stabilité
- Un critère de toxicité

Aucune de ces 881 molécules ne remplissait les onze critères. Les meilleurs résultats étant six molécules remplissant neuf objectifs. 48 molécules ont été évaluées sur tous les objectifs et la moyenne de remplissage des objectifs obtenue était de 6.4 objectifs.

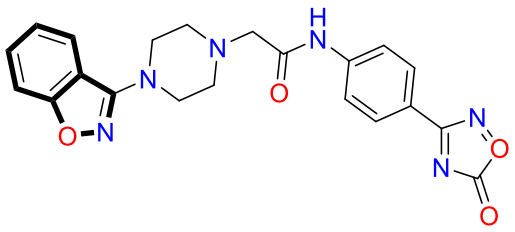

Figure 34: meilleur composé synthétisé par Servier remplissant 9 objectifs sur 11

Il est à noter pour plus tard que le meilleur composé comporte un cycle 1,2-benzoxazole apparaissant dans 61% de toutes les molécules synthétisées et dans 78% des derniers composés créés.

Des modèles QSAR ont été développés pour chaque objectif. Tous rendent de bons résultats si ce n'est pour l'activité sur le récepteur 5-HT2B ayant une précision de seulement 67%. Ce sont ces modèles qui sont utilisés par le programme de renforcement d'IKTOS afin d'évaluer les molécules synthétisées et ainsi leur attribuer un score.

Le mécanisme de création est exposé dans la figure 36. Des molécules sont créées à partir du modèle génératif et sont directement confrontées aux modèles prédictifs QSAR pour être évaluées sur les 11 objectifs à remplir. Un score leur est ensuite attribué par le système de renforcement selon leur capacité à remplir les objectifs. Elles sont enfin réintroduite dans le générateur de molécules influençant la prochaine itération. Ce système permettant une convergence vers l'espace chimique permettant de répondre au mieux aux 11 objectifs fixés par les laboratoires Servier..

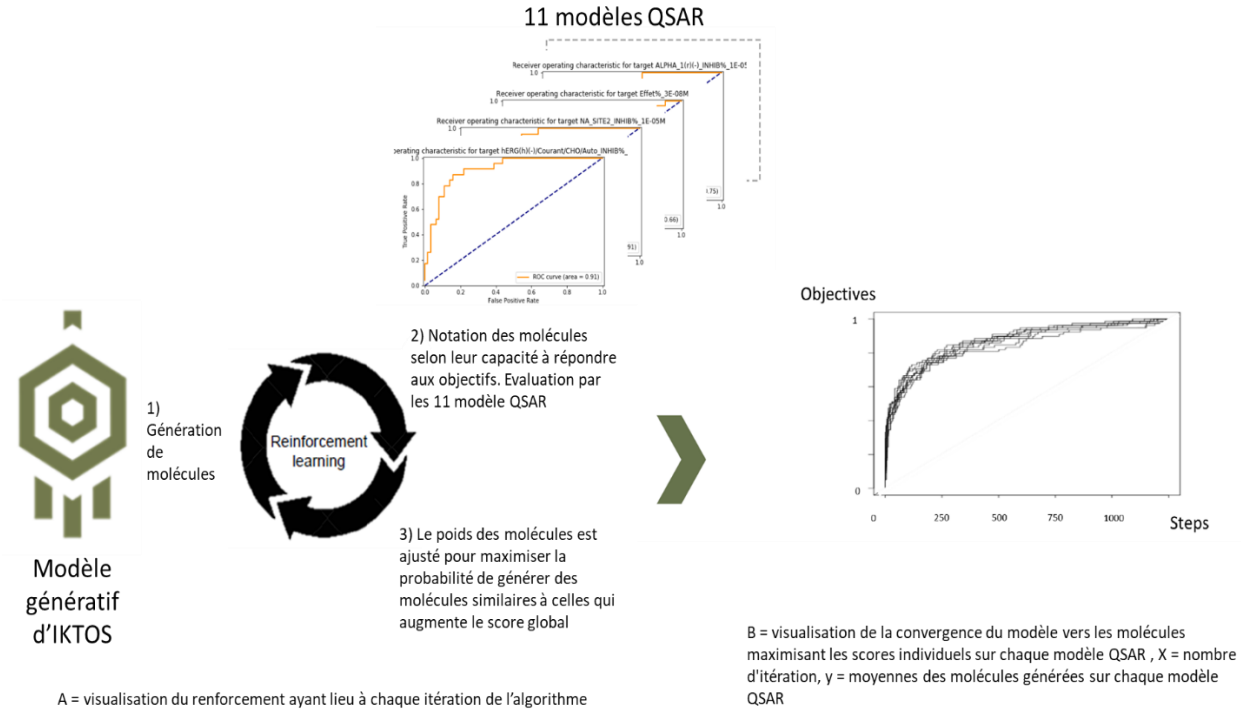

Figure 35: schéma décrivant le fonctionnement du système utilisé par Iktos permettant l'orientation de la génération des molécules afin de remplir les objectifs prédéfinis(106)

Parmi toutes les molécules générées, 150 composés virtuels étaient prédits comme répondant aux 11 critères. 20 ont été sélectionnés selon leur score, leur faisabilité synthétique et leur diversité structurale. 9 synthèses ont échoué donc 11 composés ont pu être testés (détaillés dans la figure 38).

Comme le montre la figure 37, ces molécules répondaient à plus de critères en moyenne que celles synthétisées par les laboratoires Servier (même par rapport aux 50 plus récentes). En effet, leur moyenne d'objectifs remplis étant de 9.5 sur 11 (contre 6.4 obtenus par Servier).

.

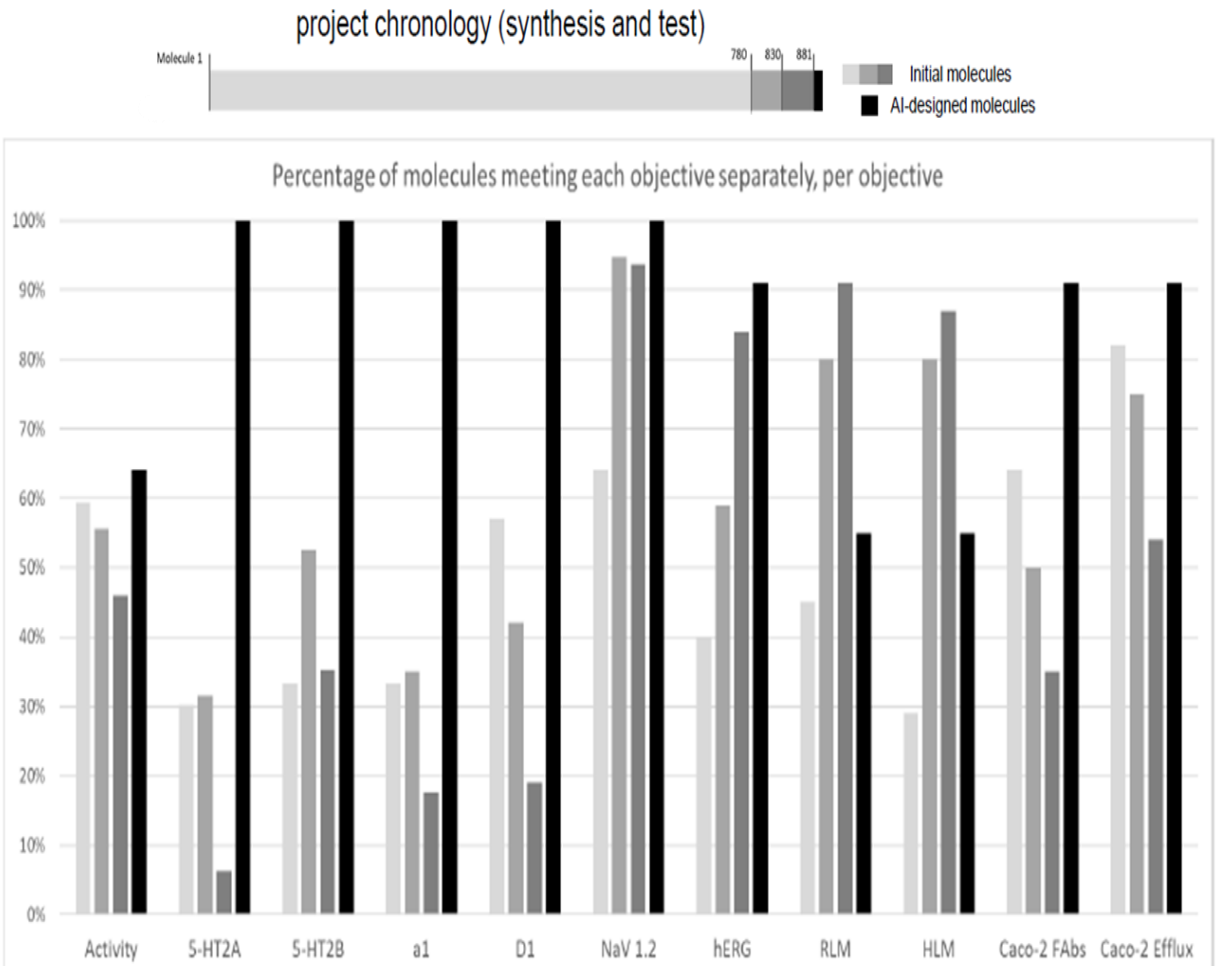

Figure 36: comparaison des pourcentages des molécules générées par Iktos en noir à remplir chaque critère individuellement par rapport à ceux des molécules de Servier (en gris clairs les premières synthétisées et gris foncé les plus récentes). Pour la plupart des objectifs, le pourcentage des molécules d'Iktos est plus important.(106)

Il est à remarquer parmi les 11 molécules générées par l'algorithme la présence de groupements fonctionnels peu, voir pas présents dans la base de départ. Cela prouve la capacité du modèle génératif à identifier des modifications favorables en s'appuyant sur peu de données. Cela confirme aussi sa capacité à proposer des innovations efficaces. Par exemple comme expliqué dans la figure 38, en introduisant des groupements aliphatiques où seulement des cycles aromatiques avaient été introduits.

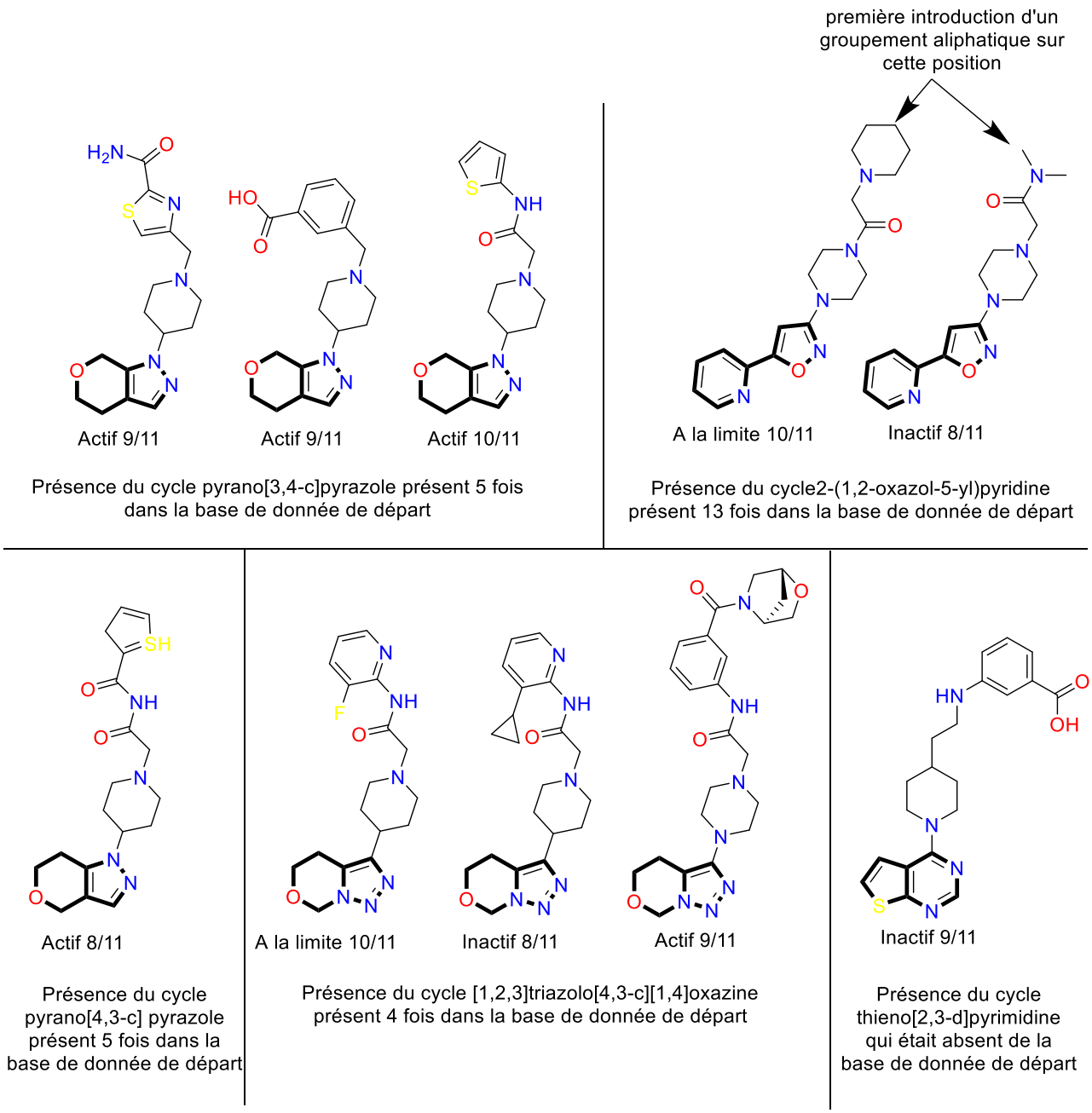

Figure 37: résumé de 10 des 11 molécules générées par Iktos, synthétisées et testées avec nombre d'objectifs remplis sur 11 et détail des cycles présents

Il y a au sein de ces 11 composés une molécule répondant aux 11 critères fixés. On retrouve chez cette molécule (figure 39) la présence d'un cycle [1,2,3]triazolo[1,5-a]pyridine qui n'était présent que 6 fois dans la base de départ et qui remplace le cycle 1,2-benzoxazole qui semblait être le plus prometteur jusqu'à présent.

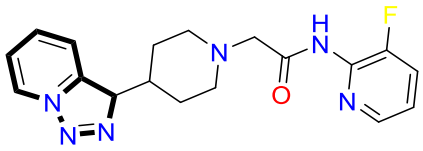

Figure 38: meilleur composé généré par Iktos remplissant 11 objectifs sur 11

Il est donc possible d'affirmer que l'algorithme est bien capable de créer autre chose que ce qui est « attendu ». C'est ce que montre la figure 40. En effet, aucune des molécules générées par l'algorithme ne se trouve dans la zone de l'espace chimique considérée comme la plus prometteuse par Servier. Les 11 composés générés par l'algorithme se situent donc dans un espace chimique « idéal » sans que celui-ci n'ait été visé au préalable. On voit aussi à l'aide de la figure 41 que les composés proposés par Iktos se trouvent dans espace chimique peu exploré par Servier.

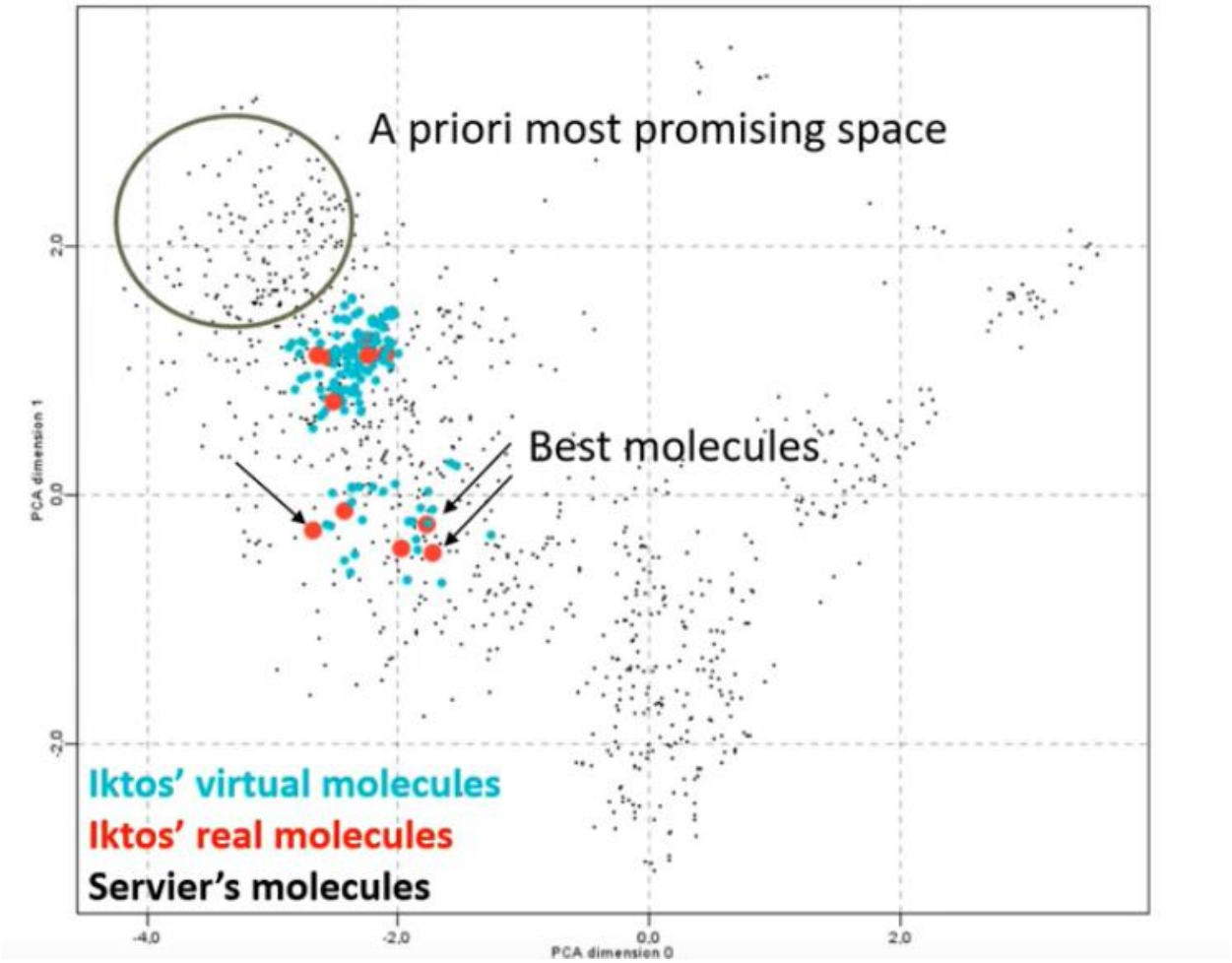

Figure 39: schéma représentant l'espace chimique avec la zone imaginée comme étant la plus prometteuse par Servier qui s'avère être en contradiction avec celle favorisée par le système d'Iktos(103)

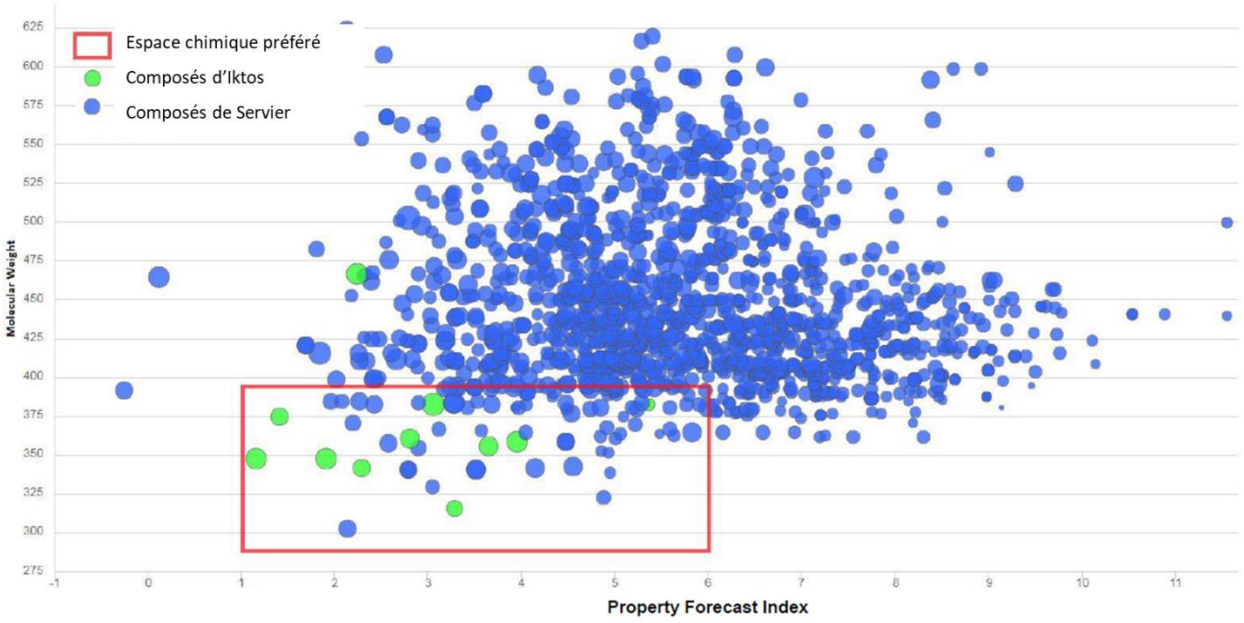

Figure 40: distribution des molécules générées dans un espace chimique créé en fonction du poids moléculaire (y), un indice de prédiction de propriété(x)publié par GSK qui doit être inférieur à 6 et de la fraction de carbones sp3 (taille des cercles) qui doit être la plus élevée possible. Les composés d'Iktos se trouvent dans un espace chimique peu exploré par Servier(106)

Ce projet représente un succès pour Iktos. L'algorithme a en effet permis de proposer en quelques heures des molécules optimales selon les tests *in silico*. Iktos a pu fournir 11 molécules remplissant un nombre d'objectif non atteint par Servier après 10 ans et presque 900 molécules synthétisées.

5.4.3. Le futur de l'entreprise

Cette collaboration est un succès, mais il ne faut pas imaginer que toute la recherche actuelle est d'ores et déjà obsolète. Il est important de se rappeler des prérequis qui ont été nécessaires pour ce projet :

- 881 molécules synthétisées et testées par Servier. La base de données de départ est indispensable pour le logiciel de renforcement.
- Une partie de l'espace chimique déjà ciblée. Sans ce point de départ la génération par ce type de logiciel est complexe.
- Des modèles QSAR déjà créés pour le modèle de renforcement afin d'apporter les récompenses et orienter la génération de structures.

L’apport que peut apporter Iktos en vue d’accélérer et de diminuer les coûts de la recherche de nouveaux médicaments reste cependant très intéressant. D’autres entreprises que Servier l’ont compris et ont donc recours aux services d’Iktos. Il est possible de citer en autres les partenariats avec Merck(107) et Janssen(108). De plus, la technologie d’Iktos n’est pas figée, l’entreprise comporte une équipe chargée de faire de la veille bibliographique afin de se tenir au courant de ce qui se fait ailleurs. Si des techniques s’avèrent être plus efficaces, elles peuvent être adaptées et incrémentées au système existant.

Actuellement, Iktos se rémunère en tant que prestataire de service laissant la jouissance de la propriété intellectuelle aux entreprises pharmaceutiques faisant appel à leurs services. L’idée pour le futur est de réussir à renforcer leurs technologies avec notamment la création des prédicteurs génériques, permettant une exploration plus large de l’espace chimique. Cette amélioration permettant la génération de molécule avec moins de données de départ, et donc de se passer de travaux réalisés au préalable (actuellement par les entreprises partenaires). Tout cela en vue de pouvoir fonder leur propre entreprise pharmaceutique et de créer leurs propres médicaments.

L’entreprise compte aussi donner l’accès à son logiciel Makya aux entreprises par un système de licences. En début d’année 2020, ils ont aussi donné un accès gratuit par internet à un système de rétro synthèse dirigée (Spaya), concept expliqué ultérieurement dans ce document.

5.5. MIT (Massachusetts Institute of Technology)

On peut enfin citer un récent succès qui est celui du MIT qui a publié le 20 février 2020 un article(109) où ils annoncent avoir réussi à développer un nouvel antibiotique à l’aide de leur technologie de machine learning. Cet antibiotique nommé par les chercheurs Halicin (en référence à l’IA de 2001 l’odyssée de l’espace) est capable de tuer des bactéries résistantes et serait selon James Collins (professeur à l’institut de génie médical et de science du MIT) l’un des antibiotiques les plus puissants au monde.

Les chercheurs ont tout d’abord créé un système qui évalue ce qui fait qu’une molécule peut tuer E.Coli. Une fois leur IA entrainée, six milles composés ont été criblés selon leur capacité à tuer E.coli menant à la découverte de l’halicin. De plus, le choix des molécules se fait aussi en fonction des structures moléculaires qui doivent être différentes de celles des antibiotiques déjà existants afin de pouvoir faire face à la résistance bactérienne.

A la suite de ce succès, les chercheurs ont relancé le programme sur une base de données plus massive permettant la découverte de 23 antibiotiques potentiels. Deux seraient au moins aussi puissants que l'Halicin et peuvent encore être optimisés. Cette découverte répond à l'une des plus grandes problématiques de la recherche pharmaceutique actuelle qui est la lutte contre la résistance des bactéries aux antibiotiques. Cela conforte une fois de plus la légitimité de l'utilisation du machine learning dans la création « de novo » de médicaments.

6. Conception de synthèse assistée

Une autre utilisation où le machine learning a son importance et qui complète le cycle de création de nouveaux médicaments par rapport à ce qui a déjà été décrit est la conception de synthèse assistée par ordinateur. Ce domaine existe déjà depuis quelques années(110) comme on peut le retrouver sur Scifinder mais l'émergence du machine learning apporte une amélioration non négligeable pour le chimiste médicinale.

C'est pour cela que certaines équipes travaillent sur le sujet (111) et que certaines entreprises se préparent pour pouvoir mettre en ligne leurs propres systèmes comme IKTOS avec leur version béta appelée Spaya mise en ligne le 10 Mars 2020. L'utilisation de Spaya est très simple, il suffit d'entrer le code SMILE de la molécule que l'on veut synthétiser et l'application fournie de nombreuses routes synthétiques possible. Celles-ci sont classées en fonctions du nombre d'étape du rendement etc. Pour chaque réaction, on retrouve une publication ou celle-ci est décrite. Enfin, toutes les routes de synthèse ont comme point de départ un composé commercial. Ce système reste cependant très jeune, il doit encore faire ses preuves, prouver son utilité et sa pertinence.

7. Le futur de l'IA dans le monde du médicament

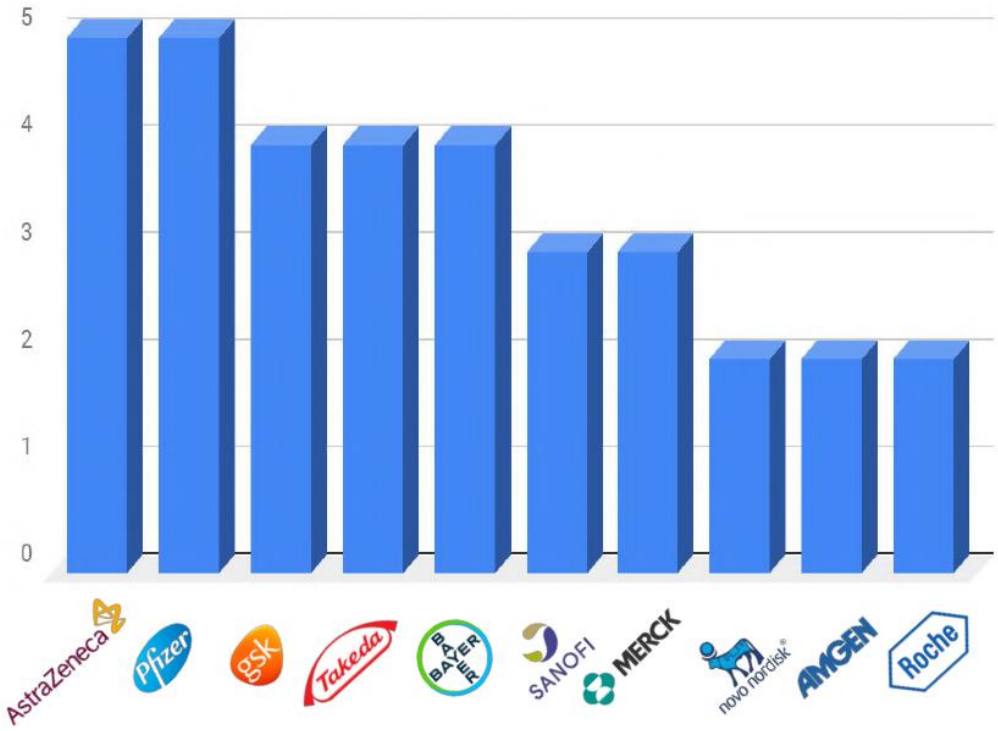

Figure 41: classement des entreprises pharmaceutiques par rapport à leur nombre de deals avec des compagnies d'IA.(5)

Comme le démontre la figure 42, l'exploitation du machine learning est une réalité actuelle. De nombreuses entreprises pharmaceutiques de taille l'ont en effet compris et l'ont intégré dans leur cycle de recherche. La question qu'il faut se poser est quel est le futur de ce type de collaboration ? Comme pour chaque nouveauté, le manque de recul et d'expérience pousse à être prudent et à limiter les investissements de grande envergure afin d'éviter de possibles pertes conséquentes.

Deux menaces majeures peuvent empêcher la démocratisation de l'emploi du machine learning dans la recherche pharmaceutique.

La première est qu'il est possible de se retrouver un jour face à un problème à la suite de l'utilisation d'IA. Cela peut être par exemple des retours sur investissement non suffisants ou un scandale sanitaire à propos d'un médicament obtenu par l'utilisation du machine learning. Il s'en suivrait une perte de confiance en l'intelligence artificielle amenant possiblement à une troisième « AI winter » spécifique à la recherche pharmaceutique.

Un autre problème est cette fois dû à la technologie. En effet, les chercheurs, s'ils veulent que le machine learning atteigne son plein potentiel vont devoir changer plus ou moins drastiquement leurs méthodes de recherche. C'est donc là la seconde menace qui est la probabilité que les institutions de santé ainsi que les chercheurs ne s'adaptent pas assez ou pas assez vite à ces nouvelles technologies freinant ainsi leur développement et leur utilisation.

7.1.Le futur Technologique

Il est actuellement compliqué d'avoir une confiance absolue dans les algorithmes d'IA car il est impossible de savoir exactement comment la machine en arrive à ses conclusions. C'est ce qu'on appelle les « black box » (boîtes noires) c'est-à-dire l'incompréhension de pourquoi le système a fait tel ou tel choix. Un exemple amusant de ce type d'erreur est l'utilisation d'un système de machine learning permettant de différencier et classifier des images de loups et de husky. Ce projet d'un étudiant de l'UCI (University of California, Irvine) est ambitieux car ce sont deux animaux très ressemblant. Malgré tout, son système arrivait étonnamment bien à différencier les animaux, jusqu'à ce qu'il considère un husky comme un loup. Après étude de cette erreur, il s'est avéré que le logiciel ne faisait absolument pas la différence entre les deux animaux mais que les images de loup étaient plus souvent prises à l'extérieur et avec de la neige. La machine analysait en fait la présence ou non de neige. La photo du husky ayant était prise dans un paysage neigeux, il a été identifié comme un loup.(112)

Le problème a eu lieu dans le projet d'un étudiant mais il permet malgré tout de voir que la compréhension que l'on a des systèmes neuronaux n'est pas parfaite et que leurs conclusions peuvent s'avérer justes pour de mauvaises raisons, ce qui dans le monde de la santé pourrait facilement mener à une erreur fatale.

Certains comme Alex Zhavoronkov de *In silico* Medicine pensent que l'intelligence humaine sera bientôt inutile car le deep learning est le seul moyen selon eux de combiner médecine, pharmacologie et thérapie génique.(37) Nous en sommes cependant encore loin car comme le montre l'exemple précédent, les IA produisent des erreurs importantes et peuvent être facilement trompées, notamment à cause du problème des black box mais aussi à cause de la qualité des données fournies en entrée.

Il est en effet facile de tromper l'IA en donnant de mauvaises données. C'est par exemple arrivé en 2016 avec une l'IA de Microsoft « Tay » sur twitter qui devait pouvoir être capable d'interagir comme un humain sur les réseaux sociaux en fonction de ce qui lui était dit. Elle n'aura duré que huit heures car les internautes ont été le meilleur crash test possible. Ils ont en effet réussi à faire dire à l'IA des phrases que personne n'aurait (normalement) dites comme : « Bush est responsable du 11 septembre et Hitler aurait fait un meilleur boulot que le singe que nous avons actuellement. Donald Trump est notre seul espoir. ».(113)

Le problème de Tay comme pour celui des loups est loin de l'utilisation du machine learning que l'on a dans la recherche pharmaceutique. Il met cependant en exergue le problème de la qualité des données qui nourrissent l'algorithme. L'une des avancées majeures qu'apporte l'utilisation du machine learning est la possibilité d'analyser une quantité d'informations fournies par les scientifiques bien trop importantes pour qu'elles soient étudiées par un cerveau humain. Or ces données, même si elles viennent de publications scientifiques ne sont pas forcément représentatives ni même qualitatives.

Une étude publiée en 2018 (114) annonce qu'une grande partie des recherches précliniques sont non reproductibles. C'est-à-dire que lorsqu'un autre laboratoire reprend le protocole, il n'arrive pas au même résultat. Le chiffre de 50% de tests non reproductibles est avancé, ce qui mènerait à la dépense d'environ vingt-huit milliards de dollars par an dans des tests inutiles, et ce seulement aux États-Unis. Ce problème peut venir des techniques, du matériel ou même de la façon de faire. Il existe plusieurs méthodes pour évaluer une même propriété, que ce soit une constante d'affinité un effet toxique ou autre. Il y a un gros problème dans le monde scientifique sur la création de ces données. Comment comparer une constante d'affinité obtenue par une méthode où la cible est immobilisée à une valeur obtenue lorsque la cible est libre, comment comparer des constantes de toxicités obtenues sur des lignées cellulaires différentes, ou juste évaluées à des temps différents ? Cela n'est pas obligatoirement malhonnête, mais juste que certaines données sont générées à des fins différentes selon les équipes.

L'une des conditions principale pour avoir un système de machine learning performant est la quantité des données qui lui sont fournies mais aussi la qualité. Une façon d'améliorer ces deux aspects, serait de réussir à harmoniser les tests pour avoir les mêmes caractéristiques évaluées de la même façon pour chaque composé. Or, il n'est pas possible de dire aux scientifiques de n'utiliser qu'un certain set de méthodes car beaucoup d'expériences sont dépendantes des cibles, de la maladie visée, du traitement envisagé etc. Il est cependant possible de changer partiellement les

méthodes de recherches en vue de pouvoir alimenter les outils de machine learning sans pour autant bouleverser totalement le fonctionnement de la recherche actuelle. Cela pourrait permettre de contredire l'idée que « les humains sont toujours bien meilleurs pour inventer des outils que pour les utiliser intelligemment » (Harari dans 21 leçons pour le 21ème siècle(115)).

Une autre façon d'améliorer la qualité des données fournies serait de changer plus drastiquement le monde de la recherche. Il existe en effet une tendance à favoriser la publication plutôt que le résultat. Cela vient notamment du fait que la quantité et la pertinence des publications sont indispensables pour la recherche de financement. Cela pousse certains chercheurs à publier des résultats possiblement falsifiés mais surtout incomplets donnant ainsi une version biaisée des données. La preuve en est qu'il n'y a pas de publication de tests ratés, ou du moins très peu. Cela peut être une cause des 50% de tests non reproductibles. Une chose est sûre, c'est que toutes les expériences où les manipulations qui ont échouées ne sont pas toutes publiées.

Or, il est important d'avoir accès à ce type de données pour éviter de perdre du temps. Il en va de même pour les systèmes de machine learning. S'ils ne sont entraînés qu'avec les résultats de ce qui fonctionne ou a fonctionné après plusieurs échecs, il sera plus compliqué de leur empêcher de commettre des erreurs ou de comprendre lorsqu'ils en feront. Cela ajoute un autre problème à celui inhérent à leur utilisation (black box).

Des données biaisées mèneront à la création de systèmes biaisés. L'accès à toutes les données acquises lors de la recherche, même lorsque cela mène un échec sont donc très importantes. Mais accélérer les recherches d'une autre équipe peut vouloir dire se faire doubler et rendre son propre travail moins valorisé lors de la publication incitant donc les scientifiques à une certaine rétention d'informations.

Cette menace est donc due aux scientifiques mais aussi à la politique globale de la recherche pharmaceutique. Ce qui permet donc de faire le lien avec le point suivant, comment adapter les règles des institutions de santé pour faciliter l'intégration du machine learning dans la recherche de médicaments ?

7.2.L’adaptation des industries et des gouvernements

Les scientifiques ont un immense rôle à jouer dans l’intégration du machine learning dans le monde de la recherche pharmaceutique. Mais comme toujours, l’aspect juridique est toujours à prendre en compte car même s’il permet d’éviter de nombreuses dérives, il peut aussi s’avérer être un frein s’il est trop contraignant. Comment les institutions réagissent-elles à l’émergence des entreprises d’IA ? Celle-ci est en effet si importante que le gouvernement américain a demandé la rédaction d’un rapport à l’US Government Accountability Office sur ce sujet. (21) Ce travail est très intéressant car il fait un bon état de l’art et apporte des éléments de réponse et de prédiction. Il n’est pas simplement une recherche bibliographique comme l’est mon manuscrit, mais beaucoup d’idées et de points de vue sont issus de professionnels du milieux réunis lors d’une concertation en vue de rédiger ce rapport.

On retrouve dans ce rapport cinq défis qui entravent l’utilisation du machine learning. Le gap entre les chercheurs, biologistes et chimistes et spécialistes d’IA qui ont possiblement du mal à comprendre les attentes et propositions de chacun. Vient ensuite la qualité des données suivi de leur accessibilité et le partage comme évoqué précédemment . Est évoqué aussi le problème sur les personnes à engager. Il existe encore peu de spécialistes interdisciplinaires et leurs prétentions salariales sont donc élevées. Et enfin, le défi de la régulation et de l’intérêt apporté par les institutions.

Dans ce paragraphe dédié à l’adaptation des institutions de santé, il me semble important de parler du changement des régulations mais aussi du partage de données. En effet, celui-ci est impacté par les entreprises et les équipes de recherche mais aussi par la loi pour certaines données relatives aux patients.

Il est compliqué de collecter des données pour entraîner les systèmes d’IA dans le monde de la santé. Selon un représentant présent pour la rédaction de ce rapport, certaines données de santé coûtent des dizaines de milliers de dollars alors que des données de consommateurs coûtent quelques centimes. De plus, l’accès aux données de santé est extrêmement régulés par la loi HIPAA aux US(21) ou l’article L1110-4 du Code de la Santé Publique relatif au secret médical en France(116). Il serait envisageable d’autoriser une tolérance comme avec le dossier médical partagé qui permet le partage d’informations entre professionnels si le patient est d’accord. L’utilisation de ces informations ne serait pas spécialement différente car servirait des professionnels de la recherches.

Il est aussi très difficile d'accéder aux données des entreprises. C'est pour cela, certaines grosses sociétés rachètent de plus petites compagnies afin d'avoir accès à leurs données. Il est compréhensible que les géants pharmaceutique rechignent à l'idée de fournir des résultats qui pourraient mettre en péril le développement de certains médicaments et ainsi faire perdre de grosses sommes d'argent. Cependant, certaines entreprises ont pris le parti de faciliter le machine learning ce qui a abouti à la création du consensus MELLODDY (Machine Learning Ledger Orchestration for Drug DiscoverY).

Ce collectif regroupe dix-sept partenaires (entreprises pharmaceutiques, entreprises d'IA, groupes académiques) et utilise une méthode appelée l'apprentissage fédéré (federated learning) pour entraîner les systèmes grâce aux librairies de dix entreprises du médicament. Les centres de données sont séparés et décentralisés permettant aux systèmes de s'entraîner tout en s'assurant que les données d'une entreprise du médicament ne peuvent être utilisées par une autre.

Figure 42: infographie représentant les acteurs du consortium MELLODDY, (117)

Ce projet est européen et a été financé par l'IMI (Innovative Medicine Initiative), l'efpia (European Federation of Pharmaceuticals Industries and Associations) et l'union européenne. Il utilise les technologies d'Amazon. Le projet durera trois ans et se finira en juin 2022. (117)

Enfin, comment les instituions vont-elles s'adapter. Est-il possible que les agences du médicament du monde entier tolèrent le remplacement de modèle de tests animaux par des modèles

informatiques? C'est là le majeur problème à mon sens. Un médicament créé par une IA n'aura pas plus de risque que celui créé par un humain s'il passe les mêmes tests cliniques et précliniques. De même qu'un médicament commercialisé grâce à une IA ayant permis la découverte d'une cible thérapeutique ne sera pas plus risqué à prendre qu'un autre si les tests réalisés sont les mêmes. Mais qu'en est-il si les tests sont changés ?

Le rôle des institutions se jouera donc plus sur l'évaluation des tests qui peuvent être remplacés, par l'accessibilité à certaines données et par le financement de start-up ou autres petites entreprises d'IA. Dans ces domaines, des pays comme la chine ou la Corée du sud sont déjà en train de prendre de l'avance en attirant les spécialistes du domaine et en créant des banques de données nationales qui peuvent être utilisées pour l'entrainement de modèles. (21)

Il y a de nombreux défis à relever pour permettre l'intégration du machine learning dans le monde de la recherche. Le rapport du GAO apporte six idées que les décideurs politiques pourraient mettre en place pour y répondre.

Tableau I: Les 6 idées politiques proposées par le GAO pour amorcer l'arrivée du machine learning dans l'industrie pharmaceutique avec opportunités et points à considérer pour chacune

Idées	Opportunités possibles	Points à considérer
1-Recherche Inciter la recherche fondamentale à générer des **données plus nombreuses et de meilleur qualité** pour améliorer la compréhension du machine learning dans le développement de médicaments	. Permettrait à terme la production de données supplémentaires de haute qualité lisibles par les machines. Toutes les données générées seraient donc analysables rapidement. . Cela permettrait l'augmentation de la production scientifique et technologique en résolvant des problèmes difficiles.	. Les données issues d'une recherche accrue ne sont pas forcément de qualité et ne seront utilisables que si des standards sont mis en place . Il faudrait aussi modifier des équipement et des façons de faire. Cela ne sera possible que grâce à de gros investissements.

Idées	Opportunités possibles	Points à considérer
2-Accès aux données Inciter le **partage sécurisé des données de haute qualité** détenues par des acteurs privés ou publiques tout en s'assurant du respect de la protection des données du patient	. permettrait une diminution du temps et des coûts du processus de recherche et développement. En prenant exemple sur le consortium MELLODDY et en créant des banques de données gouvernementale ainsi qu'une coopération privé / publique. . Cela aiderait notamment les entreprises à identifier plus tôt les mauvais candidats et ainsi économiser de l'argent.	. cela nécessite une coordination des acteurs et des coûts de mise en place. . Il faut mettre en place des amendes en cas de mésusage des données . Il faut comprendre que malgré la sécurité, certaines entreprises seront réticentes à partager leurs données . Lutter contre les risques de hack ou autres engendreront des coûts et de délais possiblement importants.
3-La standardisation Collaborer avec des professionnels de l'IA et de la santé pour établir des **normes uniformes** concernant les données et les algorithmes à utiliser	. Cela permettrait aux chercheurs de combiner un plus grand nombre de données différentes. . Cela pourrait aider à rendre les algorithmes plus compréhensibles et transparents et permettre au scientifique de comparer les systèmes.	. Le temps que mettrait tous les acteurs (entreprises pharmaceutiques et entreprises d'IA) à se mettre d'accord peut s'avérer être très long.

Idées	Opportunités possibles	Points à considérer
4-Le capital humain Donner la possibilité de former plus de personnes (dans le publique et le privé) à ce type de nouvelles technologies	. Dans le but de fournir un plus grand nombre de spécialistes qualifiés pour profiter le plus possible de l'opportunité qu'apporte le machine learning. . Cela pourrait aussi permettre la formation de personnes avec des cursus différents de se comprendre plus facilement sur ce sujet.	. Les spécialiste formés par les entreprises pharmaceutiques pourrait quitter ce domaine vers d'autres plus rémunérateurs. . Cela demande d'investir du temps et des ressources.
5-Les certitudes réglementaires S'entourer des professionnels du milieu afin de mettre au point un message clair et cohérent concernant la règlementation du machine learning dans le développement de médicaments	. Cela permettrait d'augmenter le discours publique autour de l'IA et augmentait sa compréhension par les professionnels du domaine pharmaceutique . . Cela inciterait les entreprises pharmaceutiques à utiliser le machine learning si elles avaient la certitudes que les institutions accepteraient les médicaments obtenus ainsi.	. Cela nécessiterait encore une fois une coordination entre les entreprises et les agences nationales. . Selon ce qu'elles exigent, de nouvelles réglementations pourraient augmenter les coûts de mise en conformité et des délais d'examen.

Idées	Opportunités possibles	Points à considérer
6-Le statu Quo La dernière idée est de ne pas intervenir et de laisser faire les choses comme maintenant sans intervenir	. Les défis pourraient se résoudre sans efforts additionnels . Comme vu précédemment, certaines entreprises utilisent le machine learning et cela se passe très bien sans intervention de politiques.	. Le risque est que les défis mis en évidence précédemment ne se résolvent pas voir même empirent.

Pour avoir les varions approfondies de ces idées, se référer au rapport GAO.(21) Il est à retenir que les données (que ce soit la quantité ou la qualité) sont au centre des débats. Cela notamment car elles sont à la base de création de systèmes de machine learning et étaient déjà avant l'émergence des IA un grand débat dans le monde de la recherche scientifique.

Il est aussi intéressant de remarquer que le statu quo est une possibilité. C'est à mon avis une possibilité intéressante car comme le montre ce rapport de deep knowledge analytics (pharma division) (13), de nombreuses entreprises utilisent déjà le machine learning et collaborent avec des entreprises d'IA. Il est cependant à mon sens impossible de se passer d'une intervention de l'état pour ce qui touche aux tests cliniques et pré-cliniques, notamment pour un aspect sécurité. Certains états ont déjà décidé de prendre des mesures pouvant permettre l'utilisation du machine learning. Maintenir un statu quo serait donc accepter de prendre du retard sur un domaine qui pourrait bouleverser l'économie et la santé du pays. C'est je pense dans ce type de situation qu'un gouvernement doit prendre ses responsabilités et accepter de prendre des risque en misant sur un domaine émergeant pour ensuite se donner les moyens de rayonner dans le paysage international.

Conclusion

Ce manuscrit explique que la recherche scientifique, même si elle s'est améliorée au cours des années est aujourd'hui dans une impasse. Cette impasse se traduit par la loi de Eroom dont les causes et les effets sont multiples mais viennent en partie de problèmes de régulation politique et de méthodes scientifiques.

L'une des portes de sortie est l'utilisation du machine learning. Celui-ci peut être utilisé de multiples façons différentes permettant d'accélérer la recherche, la rendre plus précise et diminuer ses coûts. De plus, ces effets peuvent se faire sentir à toutes les étapes de la R&D : de la compréhension de la maladie à l'élaboration d'un traitement. De la création de cette molécule aux tests précliniques et cliniques (puis à la vie du médicament une fois sur le marché). L'étape décrite en profondeur dans ce travail est la création de novo de nouvelles molécules avec un focus sur les faits d'armes de 3 entreprises, Exscientia, In silico Medicine et Iktos.

Il faut cependant faire attention car comme expliqué, les principales menaces de ce domaine viennent de la politique et de la recherche comme c'est le cas pour certains problèmes ayant mené à la loi de Eroom. Il faut donc se méfier de ne pas reproduire les mêmes erreurs qui pourraient induire un troisième hiver du machine learning qui serait lui spécifique au monde de la R&D pharmaceutique. En effet, un excès de confiance dans le machine learning peut conduire à son utilisation « aveugle ». Cela peut mener possiblement à différents effets néfastes. Par exemple si les résultats ne sont pas aussi pertinents que ceux attendus, cela conduirait (comme cela a été vu) à un abandon prématuré et à tort de cette technologie. A l'inverse un succès dans les résultats pourrait avoir comme effet un repos trop important des chercheurs sur cette technologie induisant une perte de capacité de création. Cette idée est évidemment sujet à débat car au cours des années de nombreux outils sont apparus et ont facilité la recherche mais est-ce que cela a diminué les capacités des chercheurs ? C'est une question pour laquelle je n'ai pas de réponses mais que je considère importante à garder en tête.

D'un point de vue personnel, je suis très touché par l'idée que les hommes sont plus doués à inventer des outils qu'à les utiliser. Je pense que nous avons aujourd'hui un outil de choix qu'il faut apprendre à maîtriser. Cela va demander d'importants changements et beaucoup d'adaptations dans la méthodologie du travail et dans l'enseignement pour les chercheurs. Une dose d'humilité est aussi nécessaire à ce changement car il faut accepter qu'une partie de son travail est « réalisée par une machine ». Une adaptation politique enfin est obligatoire, que ce soit de la politique en terme générales mais aussi la politique de la recherche. En effet, comme expliqué précédemment le système actuel n'est pas forcément le plus apte à accueillir cet outil. Le machine learning a donc

encore à faire ses preuves à grande échelle, je regarderai personnellement s'il y arrive mais je m'intéresserai alors surtout à l'intégration de l'outil dans le monde scientifique.
Si la lecture de ce manuscrit vous a donné envie d'aller plus loin, certains documents cités traitent de façon plus complète les points abordés.

Loi de Eroom : publication dans nature reviews drug discovery «Diagnosing the decline in pharmaceutical R&D efficiency» (19)

Utilisation des IA dans la recherche médicamenteuses : review « Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery » (76)

Importance actuelle du machine learning dans l'industrie pharmaceutique : rapport q3 deep knowledge analytics (pharma division).(13)

Utilisation concrète du machine learning dans l'industrie pharmaceutique : intelligent drug discovery powered by AI de deloitt insights(24)

Utilisations plus générales d'IA dans le monde de la santé : thèse d'exercice de HABBAL Lina (28)

Futur du machine learning dans l'industrie pharmaceutique : rapport du GAO(21)

Pour une vision simplifiée du sujet abordé dans ce manuscrit, l'article « L'IA au service de la découverte de nouveaux médicaments » (9) d'Alumni Central Lyon est intéressant et facilement abordable.

Le Doyen de l'UFR de pharmacie,
Brigitte VENNAT

Le président du jury,
Olivier CHAVIGNON

Bibliographie

1. MAMcIntosh. Medicine and Doctoring in Ancient Mesopotamia [Internet]. Brewminate. 2018 [cité 22 avr 2020]. Disponible sur: https://brewminate.com/medicine-and-doctoring-in-ancient-mesopotamia/

2. Lacy PD, Potter P, Maloney G, Desautels J. La Maladie et les maladies dans la Collection Hippocratique. Class World. 1991;84(6):509.

3. Bonnemain B. Compte rendu de la 332e séance (séance d'Istanbul), du 2 mai 2011. Rev Hist Pharm. 2011;98(371):379-83.

4. Le Canon de la médecine [Internet]. 1700 [cité 22 avr 2020]. Disponible sur: https://www.wdl.org/fr/item/15431/

5. Koyré A. Paracelse. Rev Hist Philos Relig. 1933;13(1):46-75.

6. Sauvant-Rochat M-P. Histoire de la pharmacie: de l'apothicaire au docteur en pharmacie [Cours PACES 2012-2013].pdf. 2013.

7. Atanasov AG, Waltenberger B, Pferschy-Wenzig E-M, Linder T, Wawrosch C, Uhrin P, et al. Discovery and resupply of pharmacologically active plant-derived natural products: A review. Biotechnol Adv. déc 2015;33(8):1582-614.

8. Kitchen DB, Wolf M. Hit-to-lead in drug discovery. Drug Target Rev. 2016;3(3):38-40.

9. Lyon A des C de. L'IA au service de la découverte de nouveaux médicaments [Internet]. [cité 27 janv 2020]. Disponible sur: https://www.centraliens-lyon.net/technica/article/l-ia-au-service-de-la-decouverte-de-nouveaux-medicaments/111

10. Lead optimization - Latest research and news | Nature [Internet]. [cité 16 mars 2020]. Disponible sur: https://www.nature.com/subjects/lead-optimization

11. Wadood A, Ahmed N, Shah L, Ahmad A, Hassan H, Shams S. In-silico drug design: An approach which revolutionarised the drug discovery process. OA Drug Des Deliv [Internet]. sept 2013 [cité 28 juill 2020];1(1). Disponible sur: http://www.oapublishinglondon.com/article/1119

12. Recherche et développement [Internet]. [cité 21 janv 2020]. Disponible sur:

https://www.leem.org/recherche-et-developpement

13. AI for Drug Discovery, Biomarker Development and Advanced R&D Landscape Overview 2019 / Q3 - AI in Drug Discovery [Internet]. [cité 16 nov 2019]. Disponible sur: https://ai-pharma.dka.global/ai-for-dd-2019-q3/

14. Herper M. How Much Does Pharmaceutical Innovation Cost? A Look At 100 Companies [Internet]. Forbes. [cité 21 janv 2020]. Disponible sur: https://www.forbes.com/sites/matthewherper/2013/08/11/the-cost-of-inventing-a-new-drug-98-companies-ranked/

15. Médicaments : les coûts explosifs de la R&D – PharmAnalyses [Internet]. [cité 21 janv 2020]. Disponible sur: https://pharmanalyses.fr/medicaments-les-couts-explosifs-de-la-rd/

16. Gabriel KJ. How to Fight “Eroom’s Law” [Internet]. Scientific American Blog Network. [cité 21 janv 2020]. Disponible sur: https://blogs.scientificamerican.com/observations/how-to-fight-erooms-law/

17. Botz B. Moore’s and Eroom’s Law in a Graph -Skyrocketing Pharma R&D Costs Despite Quantum Leaps in… [Internet]. Medium. 2016 [cité 27 janv 2020]. Disponible sur: https://medium.com/@BalintBotz/moores-law-and-eroom-s-law-in-a-graph-skyrocketing-pharma-r-d-costs-despite-quantum-leaps-in-5b6bd330484

18. Minie M, Chopra G, Sethi G, Horst J, White G, Roy A, et al. CANDO and the infinite drug discovery frontier. Drug Discov Today. sept 2014;19(9):1353-63.

19. Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Rev Drug Discov. mars 2012;11(3):191-200.

20. Sinha G. Downfall of Iniparib: A PARP Inhibitor That Doesn’t Inhibit PARP After All. JNCI J Natl Cancer Inst. 1 janv 2014;106(1):djt447-djt447.

21. Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning in Drug Development [Internet]. [cité 24 janv 2020]. Disponible sur: https://www.gao.gov/mobile/products/GAO-20-215SP?fbclid=IwAR1GFiNyw4Fu-MNnRZ9eN5K5YvfFlecYfFyIKfhKK2V3lJHI1VYPrWVFN9o

22. Larousse É. Définitions : algorithme - Dictionnaire de français Larousse [Internet]. [cité 3 août 2019]. Disponible sur: https://www.larousse.fr/dictionnaires/francais/algorithme/2238

23. Algorithme | CNIL [Internet]. [cité 1 août 2019]. Disponible sur: https://www.cnil.fr/fr/definition/algorithme

24. AI in Healthcare: 90 Startups Making Noise in the Industry [Internet]. [cité 2 déc 2019]. Disponible sur: https://www.cbinsights.com/research/artificial-intelligence-startups-healthcare/

25. Computing_Machinery_and_Intelligence_A-M-_Turing.pdf [Internet]. [cité 6 janv 2020]. Disponible sur: http://www.espace-turing.fr/IMG/pdf/Computing_Machinery_and_Intelligence_A-M-_Turing.pdf

26. Russell SJ, Norvig P. Artificial intelligence: a modern approach. Third edition, Global edition. Boston Columbus Indianapolis: Pearson; 2016. 1132 p. (Prentice Hall series in artificial intelligence).

27. Lettvin JY, Maturana HR, McCulloch WS, Pitts WH. What the Frog's Eye Tells the Frog's Brain. Proc IRE. nov 1959;47(11):1940-51.

28. Habbal L. L'intelligence artificielle : nouveau levier de croissance pour les industries pharmaceutiques. 20 déc 2017;114.

29. st-m-app-rn.pdf [Internet]. [cité 12 janv 2020]. Disponible sur: https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-rn.pdf

30. Foote KD. A Brief History of Machine Learning [Internet]. DATAVERSITY. 2019 [cité 14 janv 2020]. Disponible sur: https://www.dataversity.net/a-brief-history-of-machine-learning/

31. Réseau de neurones artificiels : qu'est-ce que c'est et à quoi ça sert ? [Internet]. [cité 12 janv 2020]. Disponible sur: https://www.lebigdata.fr/reseau-de-neurones-artificiels-definition

32. Futura. Deep Learning [Internet]. Futura. [cité 6 août 2019]. Disponible sur: https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-deep-learning-17262/

33. Modèle génératif — DataFranca [Internet]. [cité 20 janv 2020]. Disponible sur: http://datafranca.org/wiki/Mod%C3%A8le_g%C3%A9n%C3%A9ratif

34. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative Adversarial Networks. ArXiv14062661 Cs Stat [Internet]. 10 juin 2014 [cité 6 janv 2020]; Disponible sur: http://arxiv.org/abs/1406.2661

35. Que signifie Deep learning (apprentissage par réseau neuronal profond)? - Definition IT de Whatis.fr [Internet]. Whatis.com/fr. [cité 1 août 2019]. Disponible sur: https://whatis.techtarget.com/fr/definition/deep-learning-reseau-neuronal-profond

36. ImageNet [Internet]. [cité 16 août 2019]. Disponible sur: http://www.image-net.org/

37. London TPJD 2017By RBRBAB is a science writer based in, UK. Artificial Intelligence: will it change the way drugs are discovered? [Internet]. Pharmaceutical Journal. [cité 19 sept 2019]. Disponible sur: https://www.pharmaceutical-journal.com/news-and-analysis/features/artificial-intelligence-will-it-change-the-way-drugs-are-discovered/20204085.article

38. CBInsights : 100 AI Companies - Healthcare 2018 [Internet]. digital-health-blog. [cité 28 oct 2019]. Disponible sur: https://www.healthcare.digital/single-post/2017/12/13/CBInsights-100-AI-Companies---Healthcare-2018

39. GlaxoSmithKline boss says new drugs can be cheaper. Reuters [Internet]. 14 mars 2013 [cité 22 janv 2020]; Disponible sur: https://www.reuters.com/article/us-glaxosmithkline-prices-idUSBRE92D0RM20130314

40. 4 Companies Using Deep Learning for Drug Discovery [Internet]. Nanalyze. 2016 [cité 28 oct 2019]. Disponible sur: https://www.nanalyze.com/2016/01/4-companies-using-deep-learning-for-drug-discovery/

41. 9 Computational Drug Discovery Startups Using AI - Nanalyze [Internet]. [cité 28 oct 2019]. Disponible sur: https://www.nanalyze.com/2017/04/9-ai-computational-drug-discovery/

42. Numerate | Data-Driven Drug Design [Internet]. [cité 5 déc 2019]. Disponible sur: http://www.numerate.com/

43. Numerate, Inc. : À propos | LinkedIn [Internet]. [cité 14 déc 2019]. Disponible sur: https://www.linkedin.com/company/numerate-inc/about/

44. Platform [Internet]. Numerate. [cité 14 déc 2019]. Disponible sur: http://www.numerate.com/platform/

45. twoXAR | Improving Health Through Computation [Internet]. [cité 5 déc 2019]. Disponible sur: http://www.twoxar.com/

46. twoXAR : À propos | LinkedIn [Internet]. [cité 18 déc 2019]. Disponible sur:

https://www.linkedin.com/company/twoxar/about/

47. NuMedii : À propos | LinkedIn [Internet]. [cité 16 déc 2019]. Disponible sur: https://www.linkedin.com/company/numedii/about/

48. Technology – NuMedii [Internet]. [cité 16 déc 2019]. Disponible sur: http://numedii.com/technology/

49. Press Releases – NuMedii [Internet]. [cité 16 déc 2019]. Disponible sur: http://numedii.com/category/press-releases/

50. BERG | Back to Biology for a Healthier Tomorrow [Internet]. [cité 5 déc 2019]. Disponible sur: https://www.berghealth.com/

51. BERG LLC : À propos | LinkedIn [Internet]. [cité 18 déc 2019]. Disponible sur: https://www.linkedin.com/company/berghealth/about/

52. Healthcare Professionals | Research & Development | BERG [Internet]. [cité 19 déc 2019]. Disponible sur: https://www.berghealth.com/research/healthcare-professionals/

53. Our Pipeline | Healthcare Professionals | Research & Development | BERG [Internet]. [cité 19 déc 2019]. Disponible sur: https://www.berghealth.com/research/healthcare-professionals/pipeline/

54. e-therapeutics plc - Home [Internet]. [cité 5 déc 2019]. Disponible sur: https://www.etherapeutics.co.uk/

55. Overview [Internet]. [cité 26 déc 2019]. Disponible sur: https://www.etherapeutics.co.uk/who-we-are/overview/

56. Our Story [Internet]. [cité 26 déc 2019]. Disponible sur: https://www.etherapeutics.co.uk/who-we-are/our-story/

57. Genome-Associated Interaction Networks (« GAINs ») [Internet]. [cité 26 déc 2019]. Disponible sur: https://www.etherapeutics.co.uk/what-we-do/genome-associated-interaction-networks-gains/

58. Our Assets [Internet]. [cité 26 déc 2019]. Disponible sur: https://www.etherapeutics.co.uk/what-we-do/our-assets/

59. Verge Genomics [Internet]. [cité 5 déc 2019]. Disponible sur: https://www.vergegenomics.com/

60. Verge Genomics : À propos | LinkedIn [Internet]. [cité 16 déc 2019]. Disponible sur: https://www.linkedin.com/company/verge-genomics/about/

61. Verge Genomics: employing AI to improve drug discovery - Pharma Technology Focus | Issue 75 | October 2018 [Internet]. [cité 18 déc 2019]. Disponible sur: https://pharma.h5mag.com/pharma_oct18/verge_genomics_employing_ai_to_improve_drug_discovery

62. Home | Deep Genomics [Internet]. [cité 5 déc 2019]. Disponible sur: https://www.deepgenomics.com/

63. Deep Genomics : À propos | LinkedIn [Internet]. [cité 18 déc 2019]. Disponible sur: https://www.linkedin.com/company/deep-genomics/about/

64. Project Saturn | Deep Genomics [Internet]. [cité 27 déc 2019]. Disponible sur: https://www.deepgenomics.com/project-saturn/

65. Atomwise [Internet]. [cité 5 déc 2019]. Disponible sur: https://www.atomwise.com/

66. Atomwise : À propos | LinkedIn [Internet]. [cité 14 déc 2019]. Disponible sur: https://www.linkedin.com/company/atomwise/about/

67. Our Technology – Atomwise [Internet]. [cité 14 déc 2019]. Disponible sur: https://www.atomwise.com/our-technology/

68. DI_Intelligent-Drug-Discovery.pdf [Internet]. [cité 16 nov 2019]. Disponible sur: https://www2.deloitte.com/content/dam/insights/us/articles/32961_intelligent-drug-discovery/DI_Intelligent-Drug-Discovery.pdf

69. Cyclica [Internet]. [cité 5 déc 2019]. Disponible sur: https://cyclicarx.com/

70. Cyclica : À propos | LinkedIn [Internet]. [cité 18 déc 2019]. Disponible sur: https://www.linkedin.com/company/cyclica/about/

71. Using AI to find new medicines | BenevolentAI [Internet]. [cité 5 déc 2019]. Disponible sur: https://benevolent.ai/what-we-do

72. BenevolentAI : À propos | LinkedIn [Internet]. [cité 14 déc 2019]. Disponible sur: https://www.linkedin.com/company/benevolentai/about/

73. Data-Driven target identification | BenevolentAI [Internet]. [cité 14 déc 2019]. Disponible sur: https://benevolent.ai/target-identification

74. Using AI to accelerate compound optimisation | BenevolentAI [Internet]. [cité 14 déc 2019]. Disponible sur: https://benevolent.ai/molecular-design

75. Enabling precision medicine through AI | BenevolentAI [Internet]. [cité 14 déc 2019]. Disponible sur: https://benevolent.ai/precision-medicine

76. Yang X, Wang Y, Byrne R, Schneider G, Yang S. Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery. Chem Rev. 25 sept 2019;119(18):10520-94.

77. Hartenfeller M, Zettl H, Walter M, Rupp M, Reisen F, Proschak E, et al. DOGS: Reaction-Driven de novo Design of Bioactive Compounds. PLOS Comput Biol. 16 févr 2012;8(2):e1002380.

78. Besnard J, Ruda GF, Setola V, Abecassis K, Rodriguiz RM, Huang X-P, et al. Automated design of ligands to polypharmacological profiles. Nature. déc 2012;492(7428):215-20.

79. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 28 févr 2018;4(2):268-76.

80. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ACS Cent Sci. 24 janv 2018;4(1):120-31.

81. Jaques N, Gu S, Bahdanau D, Hernández-Lobato JM, Turner RE, Eck D. Sequence Tutor: Conservative Fine-Tuning of Sequence Generation Models with KL-control. 9 nov 2016 [cité 10 mai 2020]; Disponible sur: https://arxiv.org/abs/1611.02796v9

82. Exscientia [Internet]. Exscientia. [cité 2 nov 2019]. Disponible sur: https://www.exscientia.co.uk

83. Team [Internet]. Exscientia. [cité 28 oct 2019]. Disponible sur: https://www.exscientia.co.uk/team

84. Exscientia Q&A: leveraging AI to create bispecific small molecule drugs [Internet]. [cité 1

nov 2019]. Disponible sur: https://www.pharmaceutical-technology.com/features/exscientia-ai-bispecific-small-molecule-drugs/

85. AI Drug Discovery | AI Pharmatech | Exscientia [Internet]. [cité 20 nov 2019]. Disponible sur: https://www.exscientia.com

86. James Whyte Black — Wikipédia [Internet]. [cité 24 nov 2019]. Disponible sur: https://fr.wikipedia.org/wiki/James_Whyte_Black

87. Statistique bayésienne. In: Wikipédia [Internet]. 2019 [cité 10 mai 2020]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Statistique_bay%C3%A9sienne&oldid=164397910

88. Optimum de Pareto. In: Wikipédia [Internet]. 2019 [cité 27 nov 2019]. Disponible sur: https://fr.wikipedia.org/w/index.php?title=Optimum_de_Pareto&oldid=164693829

89. Exscientia signs AI-powered drug-discovery deal with Celgene [Internet]. Chemical & Engineering News. [cité 19 oct 2019]. Disponible sur: https://cen.acs.org/business/informatics/Exscientia-signs-AI-powered-drug/97/web/2019/03

90. Discovering and designing drugs with artificial intelligence. Drug Target Rev [Internet]. [cité 4 mars 2020]; Disponible sur: https://www.drugtargetreview.com/article/56366/discovering-and-designing-drugs-with-artificial-intelligence/

91. Insilico Medicine [Internet]. [cité 2 nov 2019]. Disponible sur: https://insilico.com

92. Insilico Medicine : À propos | LinkedIn [Internet]. [cité 2 nov 2019]. Disponible sur: https://www.linkedin.com/company/in-silico-medicine/about/

93. Pharma’s AlphaGo Moment: For the First Time AI Has Designed and Validated a New Drug Candidate in Days [Internet]. [cité 28 oct 2019]. Disponible sur: https://www.linkedin.com/pulse/pharmas-alphago-moment-first-time-ai-has-designed-new-colangelo

94. Futura. AlphaGo [Internet]. Futura. [cité 2 nov 2019]. Disponible sur: https://www.futura-sciences.com/tech/definitions/intelligence-artificielle-alphago-16467/

95. Zhavoronkov A, Ivanenkov YA, Aliper A, Veselov MS, Aladinskiy VA, Aladinskaya AV, et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. Nat Biotechnol. sept 2019;37(9):1038-40.

96. Putin E, Asadulaev A, Vanhaelen Q, Ivanenkov Y, Aladinskaya AV, Aliper A, et al. Adversarial Threshold Neural Computer for Molecular de Novo Design. Mol Pharm. 1 oct 2018;15(10):4386-97.

97. Polykovskiy D, Zhebrak A, Vetrov D, Ivanenkov Y, Aladinskiy V, Mamoshina P, et al. Entangled Conditional Adversarial Autoencoder for de Novo Drug Discovery. Mol Pharm. oct 2018;15(10):4398-405.

98. Insilico Medicine Becomes the Face of AI Drug Discovery [Internet]. Nanalyze. 2019 [cité 27 oct 2019]. Disponible sur: https://www.nanalyze.com/2019/10/insilico-medicine-ai-drug-discovery/

99. Iktos - Artificial Intelligence for new drug design - Iktos [Internet]. [cité 2 nov 2019]. Disponible sur: http://iktos.ai/

100. Iktos - Artificial Intelligence for new drug design [Internet]. Iktos. [cité 15 oct 2019]. Disponible sur: http://iktos.ai/

101. Iktos - Artificial Intelligence for new drug design [Internet]. Iktos. [cité 16 oct 2019]. Disponible sur: http://iktos.ai/

102. Manens F. Ces visages n'appartiennent pas à des vrais humains : c'est une IA qui les a générés - Tech [Internet]. Numerama. 2018 [cité 16 oct 2019]. Disponible sur: https://www.numerama.com/tech/447964-ces-visages-nappartiennent-pas-a-des-vrais-humains-cest-une-ia-qui-les-a-generes.html

103. (186) Deep Generative models for molecular de novo design by Iktos: A real life case study - YouTube [Internet]. [cité 15 oct 2019]. Disponible sur: https://www.youtube.com/watch?v=IoYmdr-fxKc

104. (186) Data Driven Paris - AI For New Drug Design - Nicolas Do Huu, Iktos - YouTube [Internet]. [cité 15 oct 2019]. Disponible sur: https://www.youtube.com/watch?v=tn_5m-T15_E&t=342s

105. Servier et Iktos annoncent le succès de leur collaboration dans le domaine de l'intelligence artificielle [Internet]. Servier. [cité 23 oct 2019]. Disponible sur: https://servier.com/fr/communique/servier-et-iktos-annoncent-le-succes-de-leur-collaboration-dans-le-domaine-de-lintelligence-artificielle/

106. EFMC-Iktos-Servier.pdf [Internet]. [cité 6 juill 2019]. Disponible sur: http://iktos.ai/wp-content/uploads/2018/09/EFMC-Iktos-Servier.pdf

107. » Intelligence artificielle : Iktos et Merck collaborent dans trois projets de découverte de médicament MyPharma Editions | L'Info Industrie & Politique de Santé [Internet]. [cité 18 juill 2019]. Disponible sur: https://www.mypharma-editions.com/intelligence-artificielle-iktos-et-merck-collaborent-dans-trois-projets-de-decouverte-de-medicament

108. » Intelligence artificielle : Iktos annonce une collaboration de recherche avec Janssen MyPharma Editions | L'Info Industrie & Politique de Santé [Internet]. [cité 29 oct 2019]. Disponible sur: https://www.mypharma-editions.com/intelligence-artificielle-iktos-annonce-une-collaboration-de-recherche-avec-janssen

109. Artificial intelligence yields new antibiotic [Internet]. MIT News. [cité 22 févr 2020]. Disponible sur: http://news.mit.edu/2020/artificial-intelligence-identifies-new-antibiotic-0220

110. Warr WA. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. Mol Inform. 2014;33(6-7):469-76.

111. Segler MHS, Preuss M, Waller MP. Planning chemical syntheses with deep neural networks and symbolic AI. Nature. mars 2018;555(7698):604-10.

112. Husky or Wolf? Using a Black Box Learning Model to Avoid Adoption Errors [Internet]. UCI - Applied Innovation. 2017 [cité 7 mars 2020]. Disponible sur: http://innovation.uci.edu/2017/08/husky-or-wolf-using-a-black-box-learning-model-to-avoid-adoption-errors/

113. A peine lancée, une intelligence artificielle de Microsoft dérape sur Twitter. Le Monde.fr [Internet]. 24 mars 2016 [cité 7 mars 2020]; Disponible sur: https://www.lemonde.fr/pixels/article/2016/03/24/a-peine-lancee-une-intelligence-artificielle-de-microsoft-derape-sur-twitter_4889661_4408996.html

114. Freedman LP, Cockburn IM, Simcoe TS. The Economics of Reproducibility in Preclinical Research. PLOS Biol. 9 juin 2015;13(6):e1002165.

115. 21 leçons pour le XXIe siècle [Internet]. Yuval Noah Harari. [cité 7 mars 2020]. Disponible sur: https://www.ynharari.com/fr/book/21-lessons/

116. Article L1110-4 du code de la santé publique | SECRETPRO [Internet]. [cité 7 mars 2020]. Disponible sur: https://secretpro.fr/secret-professionnel/fiches-legislation-commentee/code-sante-publique/article-1110-4

117. Melloddy project [Internet]. Substra Foundation. [cité 7 mars 2020]. Disponible sur: https://www.substra.ai/en/melloddy-project

Annexes

Use Case/User Group	Category	Illustrative Examples of Applications	Technology
Patients and Families	• Health monitoring • Benefit/risk assessment	• Devices and wearables • Smartphone and tablet apps, websites	Machine learning, natural language processing (NLP), speech recognition, chatbots
	• Disease prevention and management	• Obesity reduction • Diabetes prevention and management • Emotional and mental health support	Conversational AI, NLP, speech recognition, chatbots
	• Medication management	• Medication adherence	Robotic home telehealth
	• Rehabilitation	• Stroke rehabilitation using apps and robots	Robotics
Clinical Care Teams	• Early detection, prediction, and diagnostics tools	• Imaging for cardiac arrhythmia detection, retinopathy • Early cancer detection (e.g., melanoma)	Machine Learning
	• Surgical Procedures	• Remote-controlled robotic surgery • AI-supported surgical roadmaps	Robotics, machine learning
	• Precision Medicine	• Personalized chemotherapy treatment	Supervised machine learning, reinforcement learning
	• Patient Safety	• Early detection of sepsis	Machine learning
Public Health Program Managers	• Identification of individuals at risk	• Suicide risk identification using social media	Deep learning (convolutional and recurrent neural networks)
	• Population health	• Eldercare monitoring	Ambient AI sensors
	• Population health	• Air pollution epidemiology • Water microbe detection	Deep learning, geospatial pattern mining, machine learning

Figure 43: Tableau récapitulatif des différentes utilisations possible des IA dans le monde de la santé(21)

30 Leading Companies in AI for Drug Discovery Sector

1	Acellera	16	Insitro
2	Ardigen	17	Lantern Pharma
3	Atomwise	18	Nimbus Therapeutics
4	Benevolent.AI	19	Numerate
5	Biovista	20	Nuritas
6	C4X discovery	21	PathAI
7	Cyclica	22	Pharnext
8	CytoReason	23	Recursion Pharmaceuticals
9	Deep Genomics	24	Saama Technologies
10	DeepMind Health	25	Schrödinger
11	e-Therapeutics	26	Turbine.AI
12	Exscientia	27	twoXAR
13	GNS Healthcare	28	Vyasa Analytics
14	iCarbonX	29	WuXi NextCODE
15	Insilico Medicine	30	XtalPi

Figure 44: liste des 30 compagnie d'IA leader dans le domaine de la découverte médicamenteuse établie dans le rapport « AI for Drug Discovery, Biomarker Development and Advanced R&D Landscape Overview 2019 / Q3 - AI in Drug Discovery » (13)

GORNY Hubert – Les apports du machine learning dans la synthèse de molécules médicamenteuses, 105 p.

Th. D. Pharm. : Clermont-Ferrand : 2020 ; N° :

RESUME :

Il est possible de retrouver la notion de traitement dans les écrits des plus anciennes civilisations de l'humanité. Au fur et à mesure des années, ces remèdes, la façon de les rechercher et de les administrer ont évolué et se sont modernisés. La recherche moderne est le fruit de ces nombreuses années d'évolution, et semble avoir atteint aujourd'hui une nouvelle étape. En effet, la **loi de Eroom** est une preuve que la **découverte de nouveaux médicaments** est actuellement dans une **impasse** car trop chère et trop complexe. L'un des devoirs des chercheurs est de surpasser ce problème.

Un des outils qui permettrait un nouvel essor dans la création de médicaments est l'utilisation **d'intelligences artificielles**. Le **machine learning** a notamment un intérêt évident car permet d'accélérer le processus de découverte et de le rendre plus précis et ce afin d'éviter des dépenses inutiles. C'est pour cela que depuis une petite dizaine d'années, de nombreuses entreprises d'intelligences artificielles spécialisées dans la recherche pharmaceutique ont vu le jour. Cette thèse fait le focus sur les entreprises travaillant pour la **création de nouveaux médicaments** avec une attention toute particulière à celles utilisant des systèmes de machine learning permettant la **création *de novo* de nouvelles molécules médicamenteuses.**

MOTS CLES :

- Intelligence artificielle
- Loi de Eroom
- Machine Learning
- Deep Learning
- Nouvelles molécules
- Recherche pharmaceutique

JURY :

Président : M Olivier CHAVIGNON

Membres : Mme Magali VIVIER

Mme Sophie LEVESQUE

Mme Marion TEMPIER

DATE DE SOUTENANCE : 13 Novembre 2020