Université de Lille – Faculté de Pharmacie
Année Universitaire 2020/2021

THESE
POUR LE DIPLOME D'ETAT
DE DOCTEUR EN PHARMACIE

Soutenue publiquement le 27 mai 2021
Par Mr. QUENTIN VICENTINI

_____

BIG DATA ET INTELLIGENCE ARTIFICIELLE EN DRUG DISCOVERY

_____

Membres du jury :

Président et Directeur de Thèse : Monsieur Philippe Chavatte, Professeur des Universités et Assesseur en charge des Relations Internationales à la Faculté de Pharmacie de Lille

Assesseur(s) : Monsieur Pascal Dao Phan, Professeur Associé à la Faculté de Pharmacie de Lille et Directeur des Opérations Cliniques à Bayer Loos

Monsieur Karim Ali Belarbi, Maître de Conférence à la Faculté de Pharmacie de Lille.

Membre extérieur : Monsieur Simon Cazin, Docteur en Pharmacie, à la Pharmacie Simon Cazin de Wailly-Beaucamp

Faculté de Pharmacie de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX

Tel. : 03.20.96.40.40 - Télécopie : 03.20.96.43.64

http://pharmacie.univ-lille2.fr

Université de Lille

Président :	Jean-Christophe CAMART
Premier Vice-président :	Nicolas POSTEL
Vice-président formation tout au long de la vie :	Christophe MONDOU
Vice-président recherche :	Lionel MONTAGNE
Vice-président relations internationales :	François-Olivier SEYS
Vice-présidente ressources :	Georgette DAL
Directrice Générale des Services :	Marie-Dominique SAVINA

Faculté de Pharmacie

Doyen :	Bertrand DÉCAUDIN
Vice-doyen et assesseur à la recherche :	Patricia MELNYK
Assesseur aux relations internationales :	Philippe CHAVATTE
Assesseur aux relations avec le monde professionnel :	Thomas MORGENROTH
Assesseur à la vie de la faculté :	Claire PINÇON
Assesseur aux études :	Benjamin BERTIN
Responsable des Services :	Cyrille PORTA
Représentant étudiant :	Augustin CLERGIER

Professeurs des Universités - Praticiens Hospitaliers (PU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	ALLORGE	Delphine	Toxicologie et Santé publique	81
M.	BROUSSEAU	Thierry	Biochimie	82
M.	DÉCAUDIN	Bertrand	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	DINE	Thierry	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
Mme	DUPONT-PRADO	Annabelle	Hématologie	82
Mme	GOFFARD	Anne	Bactériologie - Virologie	82
M.	GRESSIER	Bernard	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	ODOU	Pascal	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	POULAIN	Stéphanie	Hématologie	82
M.	SIMON	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	STAELS	Bart	Biologie cellulaire	82

Professeurs des Universités (PU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	ALIOUAT	El Moukhtar	Parasitologie - Biologie animale	87
Mme	AZAROUAL	Nathalie	Biophysique - RMN	85
M.	BLANCHEMAIN	Nicolas	Pharmacotechnie industrielle	85
M.	CARNOY	Christophe	Immunologie	87
M.	CAZIN	Jean-Louis	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	CHAVATTE	Philippe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	COURTECUISSE	Régis	Sciences végétales et fongiques	87
M.	CUNY	Damien	Sciences végétales et fongiques	87

Mme	DELBAERE	Stéphanie	Biophysique - RMN	85
Mme	DEPREZ	Rebecca	Chimie thérapeutique	86
M.	DEPREZ	Benoît	Chimie bioinorganique	85
M.	DUPONT	Frédéric	Sciences végétales et fongiques	87
M.	DURIEZ	Patrick	Physiologie	86
M.	ELATI	Mohamed	Biomathématiques	27
M.	FOLIGNÉ	Benoît	Bactériologie - Virologie	87
Mme	FOULON	Catherine	Chimie analytique	85
M.	GARÇON	Guillaume	Toxicologie et Santé publique	86
M.	GOOSSENS	Jean-François	Chimie analytique	85
M.	HENNEBELLE	Thierry	Pharmacognosie	86
M.	LEBEGUE	Nicolas	Chimie thérapeutique	86
M.	LEMDANI	Mohamed	Biomathématiques	26
Mme	LESTAVEL	Sophie	Biologie cellulaire	87
Mme	LESTRELIN	Réjane	Biologie cellulaire	87
Mme	MELNYK	Patricia	Chimie physique	85
M.	MILLET	Régis	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	MUHR-TAILLEUX	Anne	Biochimie	87
Mme	PERROY	Anne-Catherine	Droit et économie pharmaceutique	86
Mme	ROMOND	Marie-Bénédicte	Bactériologie - Virologie	87
Mme	SAHPAZ	Sevser	Pharmacognosie	86
M.	SERGHERAERT	Éric	Droit et économie pharmaceutique	86
M.	SIEPMANN	Juergen	Pharmacotechnie industrielle	85
Mme	SIEPMANN	Florence	Pharmacotechnie industrielle	85
M.	WILLAND	Nicolas	Chimie organique	86

Maîtres de Conférences - Praticiens Hospitaliers (MCU-PH)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	BLONDIAUX	Nicolas	Bactériologie - Virologie	82
Mme	DEMARET	Julie	Immunologie	82
Mme	GARAT	Anne	Toxicologie et Santé publique	81
Mme	GENAY	Stéphanie	Biopharmacie, Pharmacie galénique et hospitalière	81
M.	LANNOY	Damien	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	ODOU	Marie-Françoise	Bactériologie - Virologie	82

Maîtres de Conférences des Universités (MCU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	AGOURIDAS	Laurence	Chimie thérapeutique	85
Mme	ALIOUAT	Cécile-Marie	Parasitologie - Biologie animale	87
M.	ANTHÉRIEU	Sébastien	Toxicologie et Santé publique	86
Mme	AUMERCIER	Pierrette	Biochimie	87
M.	BANTUBUNGI-BLUM	Kadiombo	Biologie cellulaire	87
Mme	BARTHELEMY	Christine	Biopharmacie, Pharmacie galénique et hospitalière	85
Mme	BEHRA	Josette	Bactériologie - Virologie	87
M.	BELARBI	Karim-Ali	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	BERTHET	Jérôme	Biophysique - RMN	85
M.	BERTIN	Benjamin	Immunologie	87
M.	BORDAGE	Simon	Pharmacognosie	86
M.	BOSC	Damien	Chimie thérapeutique	86
M.	BRIAND	Olivier	Biochimie	87

Mme	CARON-HOUDE	Sandrine	Biologie cellulaire	87
Mme	CARRIÉ	Hélène	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
Mme	CHABÉ	Magali	Parasitologie - Biologie animale	87
Mme	CHARTON	Julie	Chimie organique	86
M.	CHEVALIER	Dany	Toxicologie et Santé publique	86
Mme	DANEL	Cécile	Chimie analytique	85
Mme	DEMANCHE	Christine	Parasitologie - Biologie animale	87
Mme	DEMARQUILLY	Catherine	Biomathématiques	85
M.	DHIFLI	Wajdi	Biomathématiques	27
Mme	DUMONT	Julie	Biologie cellulaire	87
M.	EL BAKALI	Jamal	Chimie thérapeutique	86
M.	FARCE	Amaury	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	FLIPO	Marion	Chimie organique	86
M.	FURMAN	Christophe	Institut de Chimie Pharmaceutique Albert Lespagnol	86
M.	GERVOIS	Philippe	Biochimie	87
Mme	GOOSSENS	Laurence	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	GRAVE	Béatrice	Toxicologie et Santé publique	86
Mme	GROSS	Barbara	Biochimie	87
M.	HAMONIER	Julien	Biomathématiques	26
Mme	HAMOUDI-BEN YELLES	Chérifa-Mounira	Pharmacotechnie industrielle	85
Mme	HANNOTHIAUX	Marie-Hélène	Toxicologie et Santé publique	86
Mme	HELLEBOID	Audrey	Physiologie	86
M.	HERMANN	Emmanuel	Immunologie	87
M.	KAMBIA KPAKPAGA	Nicolas	Pharmacologie, Pharmacocinétique et Pharmacie clinique	86
M.	KARROUT	Younes	Pharmacotechnie industrielle	85
Mme	LALLOYER	Fanny	Biochimie	87

Mme	LECOEUR	Marie	Chimie analytique	85
Mme	LEHMANN	Hélène	Droit et économie pharmaceutique	86
Mme	LELEU	Natascha	Institut de Chimie Pharmaceutique Albert Lespagnol	86
Mme	LIPKA	Emmanuelle	Chimie analytique	85
Mme	LOINGEVILLE	Florence	Biomathématiques	26
Mme	MARTIN	Françoise	Physiologie	86
M.	MOREAU	Pierre-Arthur	Sciences végétales et fongiques	87
M.	MORGENROTH	Thomas	Droit et économie pharmaceutique	86
Mme	MUSCHERT	Susanne	Pharmacotechnie industrielle	85
Mme	NIKASINOVIC	Lydia	Toxicologie et Santé publique	86

Mme	PINÇON	Claire	Biomathématiques	85
M.	PIVA	Frank	Biochimie	85
Mme	PLATEL	Anne	Toxicologie et Santé publique	86
M.	POURCET	Benoît	Biochimie	87
M.	RAVAUX	Pierre	Biomathématiques / Innovations pédagogiques	85
Mme	RAVEZ	Séverine	Chimie thérapeutique	86
Mme	RIVIÈRE	Céline	Pharmacognosie	86
M.	ROUMY	Vincent	Pharmacognosie	86
Mme	SEBTI	Yasmine	Biochimie	87
Mme	SINGER	Elisabeth	Bactériologie - Virologie	87
Mme	STANDAERT	Annie	Parasitologie - Biologie animale	87
M.	TAGZIRT	Madjid	Hématologie	87
M.	VILLEMAGNE	Baptiste	Chimie organique	86
M.	WELTI	Stéphane	Sciences végétales et fongiques	87
M.	YOUS	Saïd	Chimie thérapeutique	86

M.	ZITOUNI	Djamel	Biomathématiques	85

Professeurs certifies

Civ.	Nom	Prénom	Service d'enseignement
Mme	FAUQUANT	Soline	Anglais
M.	HUGES	Dominique	Anglais
M.	OSTYN	Gaël	Anglais

Professeurs Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
M.	DAO PHAN	Haï Pascal	Chimie thérapeutique	86
M.	DHANANI	Alban	Droit et économie pharmaceutique	86

Maîtres de Conférences Associés

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUCCHI	Malgorzata	Biomathématiques	85
M.	DUFOSSEZ	François	Biomathématiques	85
M.	FRIMAT	Bruno	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
M.	GILLOT	François	Droit et économie pharmaceutique	86
M.	MASCAUT	Daniel	Pharmacologie, Pharmacocinétique et Pharmacie clinique	85
M.	MITOUMBA	Fabrice	Biopharmacie, Pharmacie Galénique et Hospitalière	86
M.	PELLETIER	Franck	Droit et économie pharmaceutique	86
M.	ZANETTI	Sébastien	Biomathématiques	85

Assistants Hospitalo-Universitaire (AHU)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	CUVELIER	Élodie	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81
M.	GRZYCH	Guillaume	Biochimie	82
Mme	LENSKI	Marie	Toxicologie et santé publique	81
Mme	HENRY	Héloïse	Biopharmacie, Pharmacie galénique et hospitalière	80
Mme	MASSE	Morgane	Biopharmacie, Pharmacie galénique et hospitalière	81
Mme	VAISSIÉ	Alix	Pharmacologie, Pharmacocinétique et Pharmacie clinique	81

Attachés Temporaires d'Enseignement et de Recherche (ATER)

Civ.	Nom	Prénom	Service d'enseignement	Section CNU
Mme	GEORGE	Fanny	Bactériologie - Virologie / Immunologie	87
Mme	N'GUESSAN	Cécilia	Parasitologie - Biologie animale	87
M.	RUEZ	Richard	Hématologie	87
M.	SAIED	Tarak	Biophysique - RMN	85
M.	SIEROCKI	Pierre	Chimie bioinorganique	85

Enseignant contractuel

Civ.	Nom	Prénom	Service d'enseignement
M.	MARTIN MENA	Anthony	Biopharmacie, Pharmacie galénique et hospitalière

Université de Lille

Faculté de Pharmacie de Lille

Faculté de Pharmacie de Lille

3, rue du Professeur Laguesse - B.P. 83 - 59006 LILLE CEDEX

Tel. : 03.20.96.40.40 - Télécopie : 03.20.96.43.64

http://pharmacie.univ-lille2.fr

L'Université n'entend donner aucune approbation aux opinions émises dans les thèses ; celles-ci sont propres à leurs auteurs.

Remerciements

Je tiens tout d'abord à remercier Monsieur le Professeur Philippe Chavatte, pour sa gentillesse et pour avoir accepté de m'encadrer durant la rédaction de cette thèse.

Je remercie les Docteurs Pascal Dao Phan et Karim Belarbi, pour avoir accepté de juger mon travail, mais également pour le précieux soutien qu'ils m'ont apporté durant ce parcours.

Je remercie également le Docteur Simon Cazin, pour avoir accepté de faire partie de mon jury, et pour avoir été le premier à m'accueillir en tant qu'étudiant pharmacien.

Messieurs, c'est avec beaucoup de plaisir que je termine mon cursus en votre présence.

Cette thèse est dédiée :

A mes parents, pour être des personnes extraordinaires, et pour m'avoir fait confiance tout au long de ces années.

A Claire et Gauthier, ma sœur et mon frère adorés.

A ma grand-mère Monique, pour notre complicité qui m'est chère.

A mes meilleurs amis, Kévin et Thomas, qui ne cesseront jamais de me faire rire.

A Alexandre, mon très cher ami et perpétuel business partner.

A Antoine et Valentin, pour leur amitié qui m'est chère et pour avoir égayé ces six années.

A Paul, pour sa bienveillance et son œil toujours attentif à mon égard.

Et à celles et ceux qui, de près ou de loin, m'ont accompagné lors de ce parcours, merci.

Table des matières

Liste des Abréviations

IA : Intelligence Artificielle
ML : Machine Learning
RN : Réseau Neuronal

I. Introduction

A. Big Data

Si deux citations pouvaient nous donner une idée de ce qu'est la data aujourd'hui, ce seraient celles de Peter Sondergaard et d'Eric Schmidt.

> *"Information is the oil of the 21st century, and analytics is the combustion engine." - Peter Sondergaard (1), Vice-Président et responsable de la recherche à Gartner, Inc.*

> *"There was 5 Exabytes of information created between the dawn of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing.". - Eric Schmidt (2), ancien CEO de Google.*

Même si cette dernière citation n'est pas entièrement exacte, (cela prendrait plutôt une semaine (3)), elles capturent à elles deux l'essence de ce qu'est la data aujourd'hui : un nouveau type de ressource produit en abondance. Elle est générée tellement rapidement et avec une telle complexité, que l'esprit humain à lui seul n'est plus capable de l'interpréter. Un terme a été pour cela inventé : la Big Data.

> *La Big Data est tout ensemble de données, trop volumineuses, trop complexes pour être interprétées et utilisées sans aides technologiques et infrastructurelles.*

De nos jours, la Big Data est utilisée partout et particulièrement par les GAFAM (pour Google, Amazon, Facebook, Apple et Microsoft). Ces géants du web remplissent leurs

serveurs d'une quantité impressionnante d'informations. Il est estimé que Google, Amazon et Facebook regroupent à eux seuls 30 milliards de giga-octets d'information (4). Et pour cause, leurs stratégies commerciales nous poussent à utiliser leurs services et à « parler » de nous. Nous utilisons leurs outils, exprimons nos intérêts pour une marque de vêtements ou un artiste, cherchons *via* leur moteur de recherche, ou encore, utilisons une smartwatch qui traquent nos mouvements et apprennent de nos habitudes. Ces données qui peuvent paraître anodines se révèlent alors de puissantes informations pour généraliser le comportement des utilisateurs, et mieux cibler leur mode de consommation.

Là où notre esprit aurait énormément de mal à trouver des tendances dans ces multi-milliards de points d'information, les GAFAM se servent de ce qu'on appelle le Machine Learning (ML). Ce sont des algorithmes particulièrement puissants et efficaces pour interpréter des amas de données. Nous les retrouvons dans notre quotidien sans que nous ne nous en apercevions : depuis une suggestion de film sur Netflix, jusqu'à nos recommandations d'achat sur Amazon.

Tout comme un enfant apprend par l'exemple de ses parents, les algorithmes de ML apprennent par la data. Plus il y en a, mieux ils fonctionnent. La popularité du ML a notamment été aidée par les capacités de stockage de plus en plus performantes et de moins en moins coûteuses. C'est ce que la conjecture de Moore avait énoncé.

Selon Martin Hilbert (5), la somme de tous les dispositifs de stockages de 2014 représentaient plus de 4.6 zettaoctets, soit $4.6*10^{21}$ mégaoctets d'informations. Nous parlons ici de plus de 1000 milliards de milliards de mégaoctets. En 2020 la capacité moyenne d'un disque dur d'ordinateur portable est de 1 téraoctet, cela représenterait alors plus de 4 milliards et demi de disques durs ! En seulement quelques années, la capacité de stockage du digital a explosé (Figure 1).

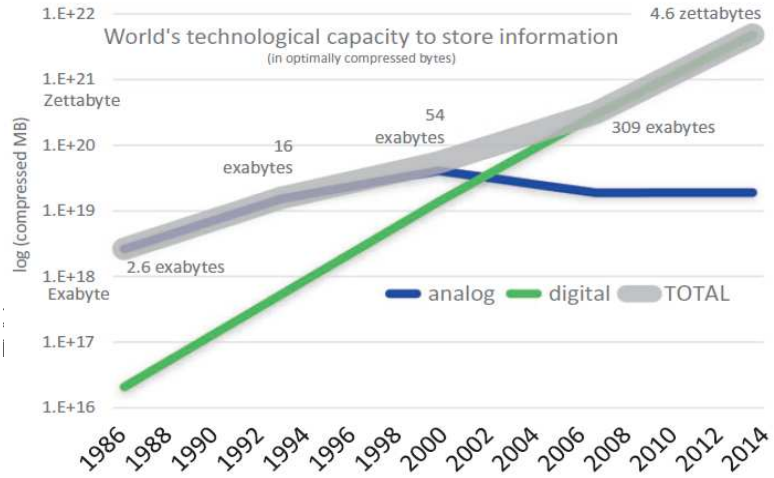

Figure 1 : Echelle logarithme de la capacité totale de stockage des données (en Mo). Figure et estimation réalisée par Martin Hilbert (5) (2017)

B. Les 5 V de la Big Data

Typiquement, on retrouve dans toutes les bases de données type Big Data cinq descripteurs appelés les 5 V. Le volume, la variété, la vélocité, la véracité et la valeur.

i. Le volume

Le volume est certainement le descripteur qui apparaît le plus évident. Il va décrire la quantité d'informations générée et stockée au sein de la banque de données. Il n'y a pas de valeur seuil pour laquelle on commence à parler de Big Data, c'est une mesure relative. Certains auteurs parlent de Big Data lorsque la vitesse de lecture et de traitement des informations n'est plus « raisonnable » et dépasse l'ordre du jour (6).

ii. La variété

La variété, elle, se réfère au(x) format(s) des données disponibles. Les données sont aujourd'hui contenues dans une image, une coordonnée GPS, un fichier vidéo ou encore un texte. Elles ont besoin d'être pré-travaillées, standardisées, avant d'être exploitables. C'est un problème car le traitement et le recoupement des données demandent des choix adaptés à chaque situation, il n'existe aucune méthode universelle.

iii. La vélocité

La vélocité est la vitesse à laquelle une nouvelle information est générée. Certaines données arrivent par « lots » et changent peu fréquemment, c'est le cas par exemple d'un séquençage de génome humain dans un point donné du temps. Entre aujourd'hui et l'année suivante, le génome ne devrait pas changer énormément. Certaines données, à l'inverse, sont générées en temps réel comme celles des marchés financiers par exemple. Procéder à une analyse en temps réel représente un véritable avantage concurrentiel dans ce domaine. Un autre exemple serait celui de Twitter, où la plateforme génère plus de 350 000 tweets par minute (7).

iv. La véracité

La véracité est un des problèmes majeurs de la Big Data. Une expression couramment utilisée dans le domaine est « *garbage-in, garbage-out* ». Face à une telle quantité d'informations, il est difficile d'obtenir des informations de qualité auxquelles on peut attribuer une grande confiance. Facebook a par exemple clôturé plus de 5,4 milliards de faux comptes en 2019 (8), ce qui affaiblit les prédictions effectuées sur ses utilisateurs. Un autre exemple serait celui des brevets déposés et traduits par un logiciel. Souvent, la traduction se révèle erronée et il devient alors difficile d'extraire une information correcte de façon automatisée et à grande échelle. Il est considéré qu'un projet de Machine Learning consiste à 80% de structuration et de nettoyage de données, pour 20% de réelle utilisation de l'algorithme.

v. La valeur

La valeur, le dernier descripteur, est certainement l'élément le plus important lorsqu'un projet de recherche démarre. La valeur de ce que l'on veut trouver doit justifier l'effort et l'investissement fourni dans le processus.

La Big Data représente aujourd'hui beaucoup plus qu'une simple information : c'est un outil stratégique et décisif pour les entreprises, parfois même pour certaines institutions politiques. Les campagnes électorales de Barack Obama ainsi que celle de Donald Trump (9) ont été menées grâce à l'aide de la Big Data. En repérant les électeurs indécis et en leur envoyant un message personnalisé, il devient plus simple de récupérer leur voix.

Pour répondre à ces nouveaux besoins, le métier de Data Scientist a explosé. Grâce aux langages informatiques tels que SQL ou encore Python, le Data Scientist est capable d'exploiter ce nouveau type de ressource.

C. La Big Data en Drug Discovery

La création d'un nouveau médicament, depuis la découverte de sa cible thérapeutique jusqu'à sa commercialisation, génère un très grand nombre de données. A titre d'illustration, les phases de Recherche Fondamentale apportent des informations tridimensionnelles sur la cible, des informations sur les voies de synthèse des composés et sur leurs propriétés physico-chimiques. Les phases précliniques, elles, génèrent des données d'Absorption-Distribution-Métabolisation-Excrétion (ADME) sur modèles animaux, d'imageries, de pharmacocinétique et bien d'autres encore. Tous les jours, de nouvelles données sont publiées dans les journaux scientifiques, de nouveaux brevets sont déposés, et des résultats d'essais cliniques sont partagés. Si nous reprenons un à un les 5 V, nous retrouvons bien les caractéristiques d'une information générée à grande échelle et à grande vitesse, variée et source de potentielles erreurs.

L'intérêt des groupes de recherche biomédicaux pour la Big Data a connu une forte croissance. En réalisant une analyse rapide sur PubMed, le nombre d'articles associés aux termes « big data », « machine learning » ou « deep learning » trouve plus de 16 000 papiers publiés en 2019 (10). Ce domaine est passé en 5 ans d'un sujet de niche à un sujet très à la mode (Figure 2). L'industrie pharmaceutique y porte aussi un grand intérêt, des groupes tels que Bayer ont annoncé des collaborations avec des entreprises spécialisées dans l'intelligence artificielle (11).

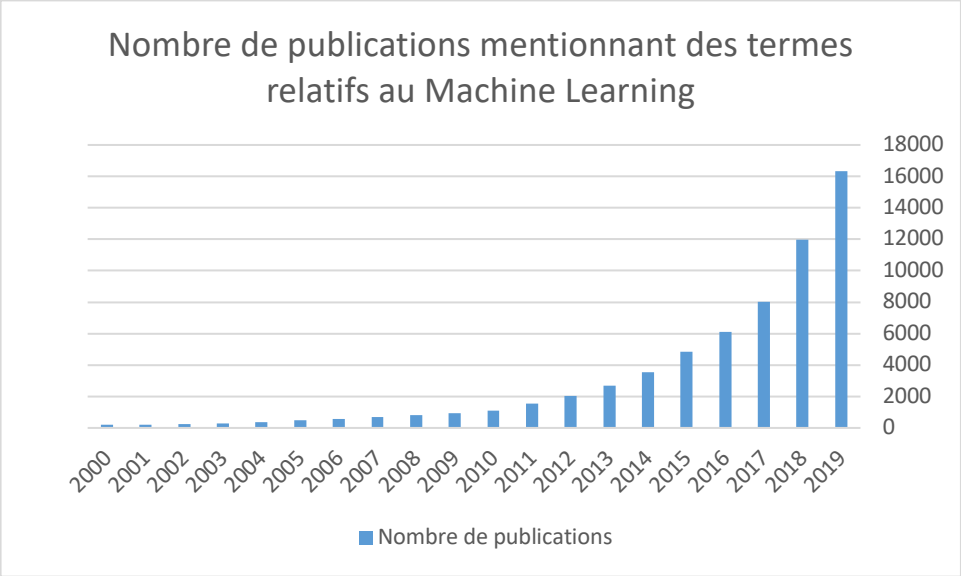

Figure 2 : Nombre de publications référencées par PubMed des années 2000 à 2019, mentionnant l'intelligence artificielle et termes relatifs

Les chercheurs peuvent aujourd'hui se former à ce tout nouveau domaine, gratuitement et depuis chez eux. Il existe un très grand nombre de formations gratuites et ouvertes à toutes et à tous. La Finlande (12) a par exemple mis à disposition un cours rédigé par l'Université d'Helsinki afin d'aider la population à appréhender ces nouvelles technologies. Leur objectif est de démystifier l'intelligence artificielle et de promouvoir l'accès à la programmation. C'est également le cas des Massive Open Online Courses (MOOCs), où des universités comme Harvard et MIT offrent des formations à la bio-informatique (13).

D. Les ressources disponibles pour la Drug Discovery

Chaque jour des articles scientifiques sont publiés, des brevets déposés, des annonces de presse effectuées. Et dans chacune de ces ressources se trouvent des points d'information.

Cependant, il est impossible pour un chercheur d'extraire ces informations une à une et de les agencer manuellement. C'est pour cela que des initiatives institutionnelles se sont mises en place pour faciliter l'accès et l'exploitation de données biomédicales.

C’est par exemple le cas de l’European Bioinformatics Institut (EMBL-EBI) et de son homologue américain, le National Centre for Biotechnology Information.

A titre d’illustration, l’EMBL-EBI possédait en 2018 une capacité de stockage de 273 petaoctets (14). Cela semble faible comparé aux capacités de Google qui sont estimées entre 10 et 15 exaoctets, mais la Big Data est toujours relative à son domaine d’étude.

En Sciences Biomédicales, on retrouve des données dans de très nombreux champs d’exploration. La Protein Data Bank contient près de 160 000 structures biologiques et aide les chimistes médicaux et les bio-informaticiens à concevoir de façon rationnelle de nouvelles molécules. La plateforme ChEMBL de l’EMBL-EBI concentre des données de binding, de bioactivités et de propriétés ADME de composés. Un autre exemple est celui du projet américain « Cancer Genome Atlas » qui se concentre sur la caractérisation moléculaire de plus de 20 000 cancers primaires.

Ces données, prises individuellement, ne peuvent répondre qu’à un pan restreint de questions. Si le Cancer Genome Atlas peut par exemple aider un chercheur à trouver une cible thérapeutique potentielle, il est inutile à la prédiction des propriétés ADME d’une molécule. Un potentiel encore plus grand existe lorsque différentes sources sont intégrées les unes aux autres. Une association entre la PDB, ChEMBL et le CGA pourrait par exemple relier la surexpression d’une protéine observée dans un cancer particulier, et grâce à la structure tridimensionnelle de la protéine et la connaissance de ses ligands, de nouveaux composés pourraient être générés très rapidement (Figure 3).

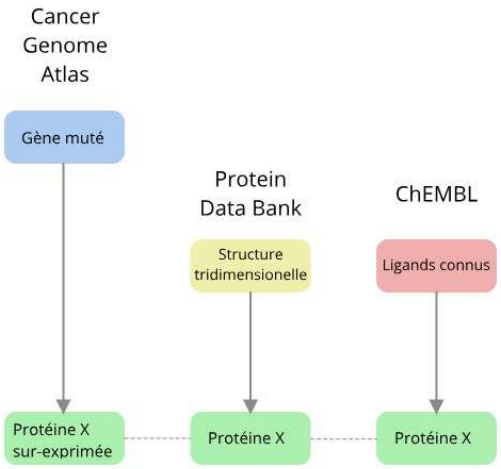

Figure 3 : Exemple d'intégration de différentes données. A elle seule, une banque ne peut répondre qu'à certaines questions. Combinées, elles peuvent soutenir un projet de drug discovery. Figure adaptée de Brown et al. 2018 (15)

En réalité, l'intégration de données n'est pas aussi simple, mais là aussi certaines initiatives existent. On peut noter Open Targets, qui est issu d'une collaboration privée-publique lancée par l'EMBL-BL et de nombreux industriels comme Sanofi, GSK ou encore Celgene (16). La plateforme combine différents types d'informations, afin de relier une pathologie à de potentielles cibles thérapeutiques.

Une liste non exhaustive des différentes bases de données est disponible en index.

E. Les attentes de l'Industrie Pharmaceutique ?

L'industrie Pharmaceutique fait face à de nombreux défis : les programmes de Recherche et Développement sont toujours plus longs et plus risqués. Le coût de développement d'un médicament a doublé entre 2003 et 2013, pour être aujourd'hui estimé à 2,6 milliards de dollars, d'après le Tufts Center for the Study of Drug Development(17). Si les coûts d'investissement augmentent, les taux de réussite, eux, diminuent de plus en plus. De la phase 1 jusqu'à l'autorisation de mise sur le marché, le succès d'une molécule n'est que de 12% environ. Aujourd'hui, les 10 plus grosses entreprises pharmaceutiques investissent près de 80 milliards de dollars par année pour un retour sur investissement de deux cents par dollar investi (18). C'est alors une industrie à 1 trillion de dollars qui est en difficulté : 9 composés sur 10 qui entrent en phase d'essai clinique sont retirés pour mauvaise efficacité, toxicité etc. L'Intelligence Artificielle peut-elle alors soutenir ces projets ?

Le domaine des petites molécules thérapeutiques a été enrichi par les campagnes de screening à haut débit réalisées par l'industrie pharmaceutique. Malgré le nombre impressionnant de composés testés (une entreprise possède en moyenne entre 1 à 4 millions de composés avec, pour chacun, les résultats de tests biologiques correspondants) une campagne n'offre que 0 à 0,01% de succès car le choix des hits validés par le test de la campagne est une étape critique. Chaque pourcentage de gagné, même infime, se révèle très avantageux.

L'utilisation de l'Intelligence Artificielle est implantable à tous les nouveaux d'un programme de recherche. Par exemple, les méthodes de prédiction conventionnelles de composés inhibiteurs des cytochromes P450 sont peu performantes dans un tiers des cas (18). Cela conduit à de lourdes conséquences lorsque la molécule a déjà atteint le stade d'essais cliniques. Bristol-Myers-Squibb a développé pour cela un modèle de ML ayant une précision de plus de 95% sur la prédiction des composés inhibiteurs des cytochromes.

Finalement, ce que les laboratoires attendent d'un tel outil aujourd'hui, ce n'est pas de générer une molécule prête à aller sur le marché, mais plutôt un algorithme capable de mettre une fin prématurée à des composés qui se révèleraient toxiques ou inefficaces lors d'étapes plus critiques, comme celles des essais cliniques. Lorsqu'une molécule entre dans une phase plus mûre, beaucoup de temps et de moyens financiers ont déjà été investis, un échec peut donc être très lourd de conséquences.

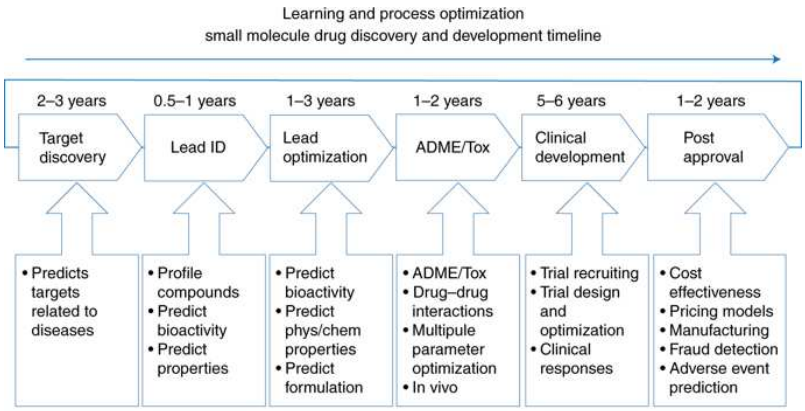

Figure 4 : Implémentation de l'IA à toutes les étapes d'un projet de Drug Discovery. Pour l'industrie, chaque année de gagnée est considérablement avantageux une fois le produit sur le marché. Figure extraite de Ekins et al. 2019 (19)

II. Le Machine Learning

Le Machine Learning (ML), encore appelé Apprentissage Automatique, est l'outil clé pour traiter la Big Data. Il ne se passe pas une journée sans que les médias parlent « d'intelligence artificielle », de « deep learning » en inter-changeant les termes et sans explications claires. Ce chapitre a pour but de démystifier ce qu'est le ML, et d'offrir au lecteur une compréhension plus précise de ce qu'est cette technologie.

L'Intelligence Artificielle se définit par :

> *« The study of how to produce machines that have some of the qualities that the human mind has, such as the ability to understand language, recognize pictures, solve problems, and learn » (20)*

Le Machine Learning est une classe de l'IA qui apprend ces qualités grâce à des données. Le Deep Learning est une classe du Machine Learning, qui lui-même est une classe de l'Intelligence Artificielle (21) (Figure 5).

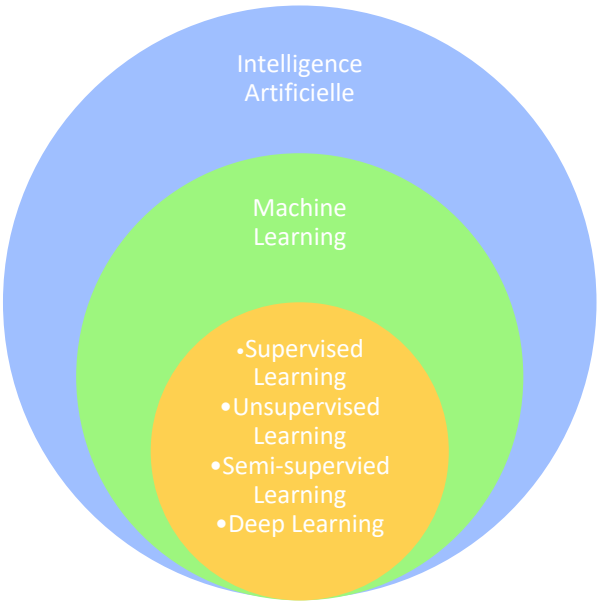

Figure 5 : Intelligence Artificielle et ses différents sous-groupes

C’est Arthur Samuel (Figure 6) qui en 1959 est le premier à utiliser et à définir le terme de Machine Learning comme :

« A field of study that gives computers the ability to learn without being explicitly programmed. »

Figure 6 : Arthur Samuel, « père » de l’intelligence artificielle. Photographie de l’Université de Stanford (22)

Arthur Samuel est un pionner du ML. Il était chercheur chez IBM et enseignant à l’Université de Standford. Il a développé un programme capable de jouer au jeu de dames et de s’améliorer de parties en parties. Son programme a notamment été capable de battre le 4ème meilleur joueur de dames des Etats-Unis (22). Aujourd’hui, on retrouve des algorithmes encore plus puissants, notamment AlphaGo conçu par DeepMind (une filiale de Google). Le jeu de Go est connu pour être un jeu complexe et possédant bien plus de possibilités que le jeu de dames ou même encore celui d’échecs. Le logiciel s’est amélioré en jouant à la fois des parties contre d’autres robots, d’autres joueurs « humains » mais essentiellement contre lui-même. Cette capacité à jouer contre lui-même lui a permis de réaliser un nombre gigantesque de parties, et de faire face à énormément de scénarios. En 2017, AlphaGo a battu le champion du monde Ke Jie (23).

Tout comme un enfant apprend par des exemples, les algorithmes de ML apprennent par des « sets d’apprentissage ». Traditionnellement, pour programmer l’addition de deux nombres ou le décollage d’une fusée, il était nécessaire d’écrire un algorithme long et compliqué. Avec le ML cela est différent : les algorithmes définissent eux-mêmes leurs règles grâce aux données. Plus ils en ont, plus ils peuvent s’affiner et devenir performants.

Le ML est omniprésent dans notre quotidien, sans même que nous en soyons forcément conscient. Lorsque par exemple Spotify nous suggère un nouveau morceau en fonction de ce que l'on a écouté, ou encore lorsque notre boîte mail classe nos SPAM. D'autres exemples plus impressionnants, comme la reconnaissance vocale, les moteurs de traduction ou encore la conduite automobile automatique de Tesla passent aussi par le ML.

F. Généralités sur le Machine Learning

Un algorithme de ML, quel qu'il soit, apprend grâce à un set de données. Grâce à cette phase d'apprentissage, le modèle va générer une fonction de prédiction. C'est elle qui permettra d'estimer la valeur de nouveaux points.

Dans un set de données, chaque point est appelé observation et est décrit au moyen de deux types de variables : les variables prédictives, ce sont les variables qui aideront l'algorithme à émettre une prédiction et les variables cibles, celles que l'on souhaite prédire. A titre d'illustration, si l'on veut prédire le prix d'une maison (variable cible), nous nous aiderons de sa superficie, sa localisation etc… (variables prédictives).

Lorsque l'on impose au modèle une forme particulière pour effectuer une prédiction, on parle de modèle paramétrique (on peut penser à l'équation y = ax + b, où l'algorithme doit ajuster les paramètres a et b). A l'inverse, quand aucune forme ne lui est imposée, on parle de modèle non paramétrique.

vi. Gradient Descent

Afin de formuler une prédiction optimale et se rapprocher au mieux de la réalité, le ML passe par la technique de l'algorithme du gradient (en anglais : Gradient Descent). En d'autres termes, le modèle va évaluer l'écart entre la prédiction effectuée et la valeur cible connue, puis va s'adapter afin de rendre cette différence-là plus faible possible.

Afin de trouver cet optimum, l'algorithme va itérer et se modifier étape par étape jusqu'à ce qu'il trouve ce point. En langage mathématique il suffit de tracer la tangente de ce point et de regarder si elle « monte », ou si elle « descend ». Le processus continue jusqu'à ce que le modèle trouve une droite ayant un coefficient nul ou le plus

faible possible. Dans un cas simple (Figure 7), il n'existe qu'un seul minimum et le modèle pourra le trouver rapidement.

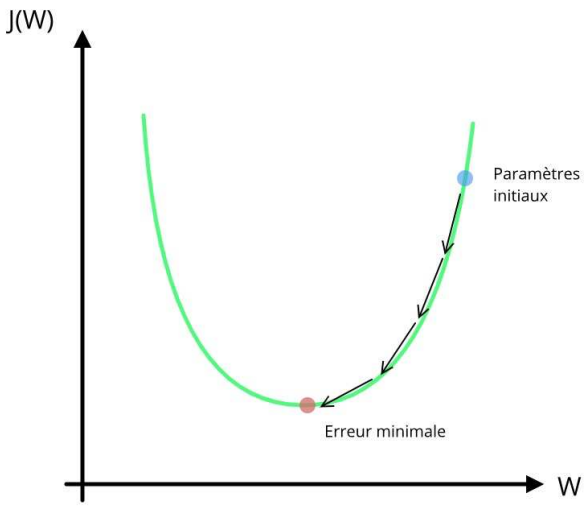

Figure 7 : Erreur en fonction des paramètres. Chaque flèche représente la taille des « pas » de l'itération.

Toutefois dans le cadre de la Big Data ce sont parfois plusieurs centaines voire milliers de paramètres qui entrent en jeu. De cette façon, plusieurs points locaux minimums peuvent exister, et l'algorithme considèrera qu'un minimum local est la meilleure solution possible. Une façon de contourner ce problème est de moduler la vitesse d'itération de l'algorithme (Figure 8).

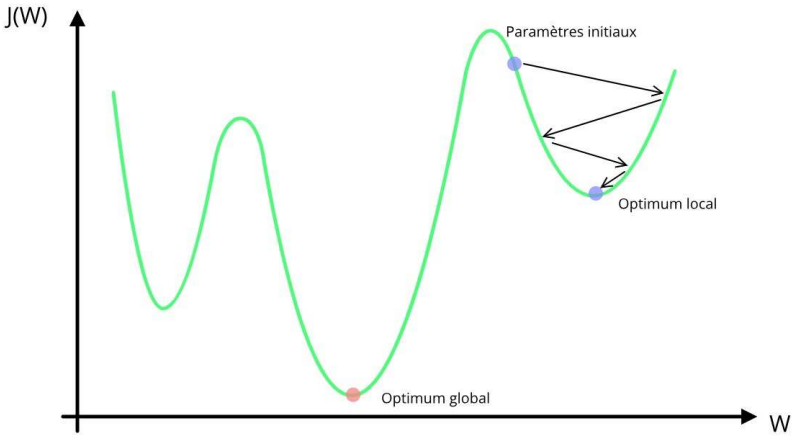

Figure 8 : Lorsque le nombre de paramètres augmente, des optimums locaux peuvent être interprétés à tort comme l'optimum global. De ce fait les paramètres du modèle ne seront pas correctement ajustés.

vii. Performance et sur-apprentissage

La performance d'un algorithme de ML est basée sur sa capacité à généraliser les données et à prédire de nouveaux points. Un algorithme peu performant n'arrivera pas à généraliser, et va substituer l'apprentissage par de la mémorisation. Il va reproduire avec une trop grande précision les résultats connus (Figure 9). On appelle ce phénomène « l'overfitting » ou « sur-ajustement ».

C'est un problème notamment avec la Big Data car le « bruit » que contiennent ces données est important, et l'algorithme n'est parfois pas capable de généraliser par lui-même. Afin d'éviter ce genre de problème, une parade existe. On divise un set de données en deux : 80% des données, le « training set », et les 20% restant, « le test set » seront testés sur l'algorithme. C'est la validation croisée. A l'issue de cette étape une erreur moyenne sera estimée entre les deux sets, et des corrections pourront être apportées.

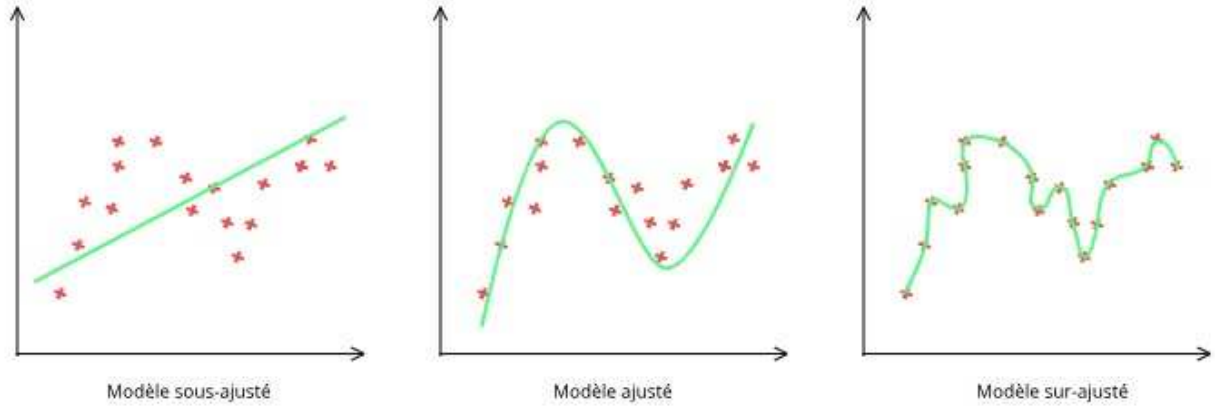

Figure 9 : Un modèle performant doit être capable de généraliser une tendance et non reproduire des données qu'il a déjà croisées.

viii. L'apprentissage supervisé

L'apprentissage supervisé est certainement la technique de Machine Learning la plus utilisée à ce jour. L'algorithme est entrainé par un set de données dites « étiquetées ». Ce sont des observations caractérisées par des variables prédictives pour lesquelles on connait la variable cible. Grâce à cet apprentissage, l'algorithme va pouvoir généraliser l'association entre les variables prédictives et la variable cible.

L’exemple classique est celui d’un set de photographies d’animaux, utilisé pour entraîner un modèle à reconnaître chaque animal. L’algorithme analyse chaque image une à une, étiquetée comme « chat », « chien », ou « girafe » etc... C’est la phase d’apprentissage : l’algorithme détermine le modèle à partir des données qu’il reçoit. La seconde phase est celle de test, l’algorithme doit prédire l’étiquette d’une nouvelle donnée.

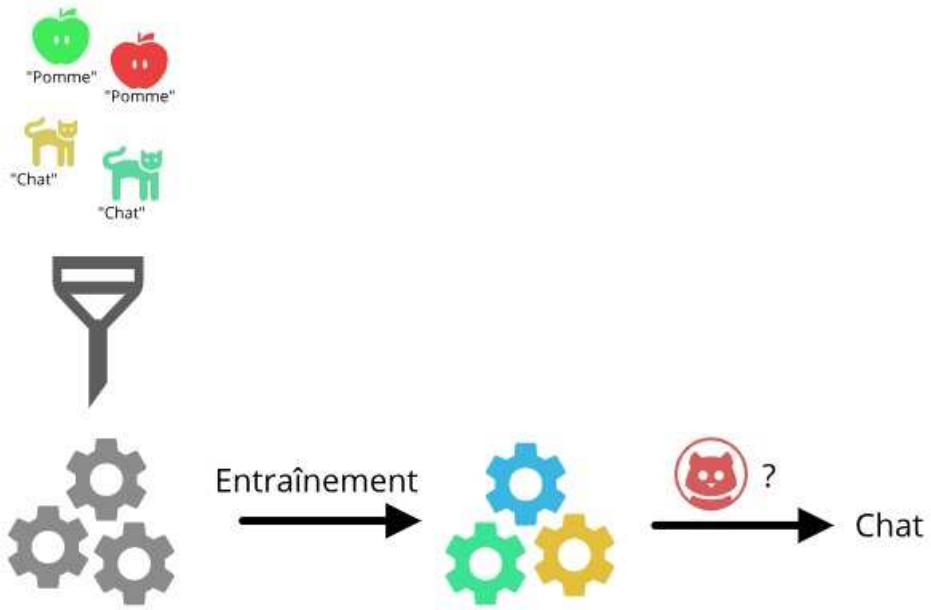

Figure 10 : Schéma du fonctionnement d'un algorithme de ML supervisé. Chaque donnée utilisée dans la phase d'entraînement est dite étiquetée, on connait sa valeur cible.

L’apprentissage supervisé est très utile afin de créer des modèles de régression ou de classification.

Exemples d’algorithmes utilisés en apprentissage supervisé

Modèle de régression linéaire

Le modèle de régression linéaire fait partie des algorithmes les plus simples à comprendre et à utiliser. Lors de la phase d’apprentissage, le modèle va tenter de prédire une tendance de la forme $y = ax + b$, on est ici dans un modèle paramétrique.

En Machine Learning, les termes sont différents et la fonction prédictive, annotée $h\theta(x)$, est aussi appelée « hypothèse ».

$$h\theta(x) = \theta_0 + \theta_1 x_1 + \cdots + \theta_n x_n$$

Avec $x1 \ldots xn$ étant les différentes variables prédictives.

Le modèle se complique lorsque la relation entre les paramètres n'est pas linéaire, on parle alors par exemple de régression polynomiale.

La méthode des moindres carrés (voir équation ci-dessous) permet de définir l'erreur de l'algorithme par rapport à ce qu'on attend en réalité. Elle est également appelée « Cost Function » en Anglais. Elle est définie comme la somme des carrés des écarts entre les valeurs prédites et les valeurs observées. L'erreur élevée au carré permet notamment d'éviter que certaines valeurs ne s'annulent (Figure 11).

Avec $h\theta(x_i)$ étant la prédiction proposée par l'algorithme, et y_i le résultat attendu en réalité.

$$\sum_{i=1}^{n} (\, h\theta(x_i) - y_i)^2$$

Plus cette somme se rapproche de 0 et plus le modèle proposé par l'algorithme collera à la réalité.

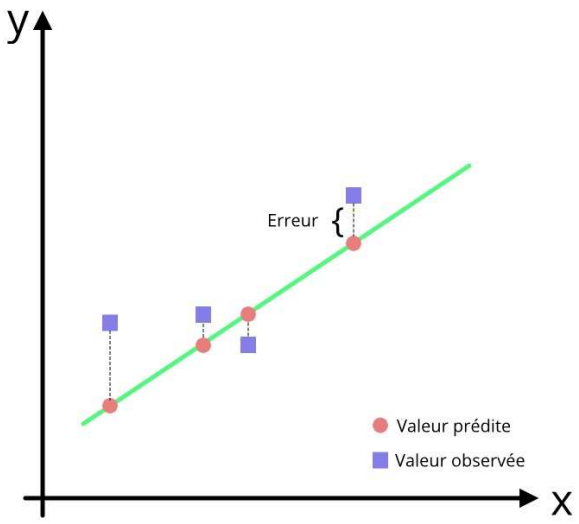

Figure 11 : Exemple typique d'une droite de régression linéaire, avec pour modèle f(x) = ax + b

Les k plus proches voisins ou K Nearest Neighbors (KNN)

L'algorithme des k plus proches voisins est un algorithme de classification supervisé et non paramétrique. Il consiste à supposer qu'une nouvelle valeur observée est similaire à celle de ses voisins. Dans ce modèle, il existe une notion de distance entre les points, on le qualifie de modèle géométrique (Figure 12).

Cet algorithme est toutefois très sensible au bruit. Lorsque le nombre de descripteurs est grand, les calculs utilisent beaucoup plus de ressources. Il est souvent nécessaire de passer par une technique de réduction dimensionnelle.

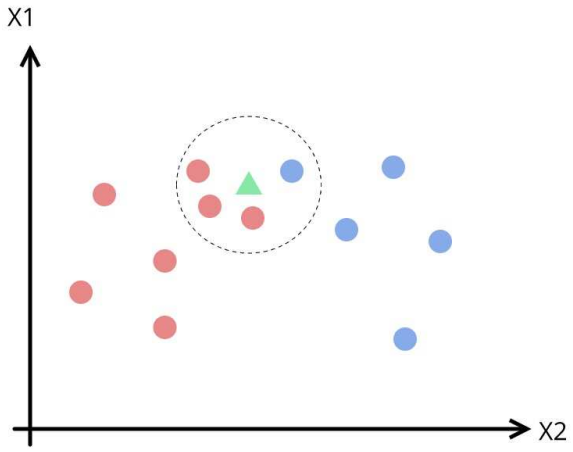

Figure 12 : Modèle des plus proches voisins, en regardant ici les 4 plus proches. Le nouvel élément (triangle vert) sera classé comme appartenant à la classe des points rouges de par sa proximité aux autres points rouges.

La classification naïve bayésienne

La classification naïve bayésienne est un outil probabilistique. Elle considère que toutes les variables d'une donnée sont indépendantes les unes des autres. Grâce à cela, le modèle repérera des tendances dans différents clusters. Il calculera la probabilité qu'une nouvelle donnée appartienne à telle ou telle classe, basée sur ses variables prédictives. L'algorithme prend une décision par maximum à posteriori.

A titre d'illustration, si un nouvel élément E a 85% de chance d'appartenir à la classe X et 63% de chance d'appartenir à la classe Y, l'algorithme décidera que E appartient à X.

Le terme naïf est employé car on assume que les descripteurs sont indépendants les uns des autres, et qu'ils ont tous la même importance. En pratique cela se révèle faux, mais l'algorithme donne cependant de très bons résultats. Le modèle va se baser sur ses connaissances antérieures, c'est-à-dire sur ce qu'il a analysé lors de sa phase d'entraînement. L'algorithme a analysé plusieurs classes et a pu extraire des tendances dans chacune d'entre elles, du type : la classe Z est décrite en majorité par des variables descriptives x et y. Lorsque l'on veut analyser une nouvelle donnée, qui n'a alors pas de classe, l'algorithme va réaliser l'opération inverse : connaissant ses variables descriptives H et U, quelle est la probabilité que cette donnée appartienne à la classe Z ?

Les machines à vecteurs de support ou SVM

Aussi appelés séparateurs à vaste marge, ce sont de puissants algorithmes de classification binaire, non linéaire. Les SVM ont pour but de construire une ligne ou une marge de largeur maximale afin de séparer deux groupes d'observation (Figure 13).

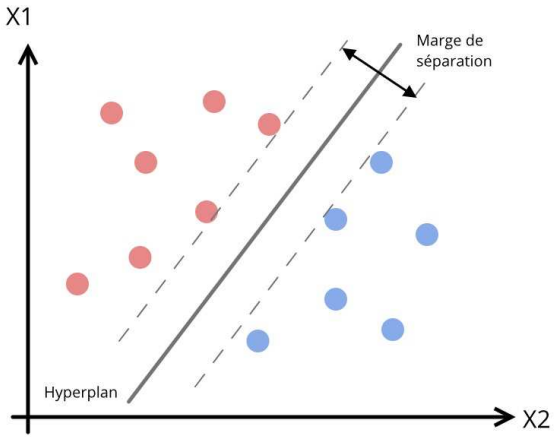

Figure 13 : Représentation d'un SVM. Le modèle va chercher à séparer au maximum deux classes distinctes grâce à un hyperplan.

Il arrive que dans certains cas un modèle linéaire ne soit pas adapté au set de données. L'algorithme n'arrive pas à formuler d'hyperplan. Il est dans ce cas nécessaire de décrire un nouvel espace de représentation des données. Pour cela, l'algorithme va utiliser les variables du set de données différemment, les transformer

mathématiquement, jusqu'à ce qu'un plan linéaire soit exploitable. Sur la Figure 14 les deux classes de données ne sont pas séparables linéairement. Après transformation et élévation des variables au carré, il est possible de définir un hyperplan.

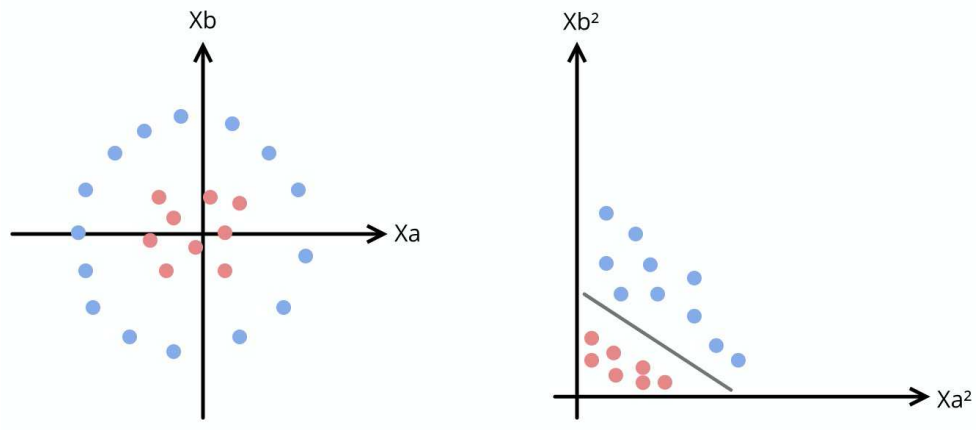

Figure 14 : A gauche, un cercle semble plus pertinent pour séparer les données. A droite, après élévation des descripteurs au carré, un nouvel hyperplan apparaît. Les données sont linéairement séparables.

Les forêts aléatoires

Aussi appelé « random forest », cet outil se base sur des arbres de décision. Il est à ce jour un des meilleurs algorithmes utilisés en classification.

L'arbre de décision est un concept très intuitif dans lequel l'algorithme va répondre à une question concernant les descripteurs d'une observation. A chaque question (aussi appelée nœud), la branche de l'arbre se divisera et la classe de la variable sera déterminée de façon de plus en plus précise (Figure 15).

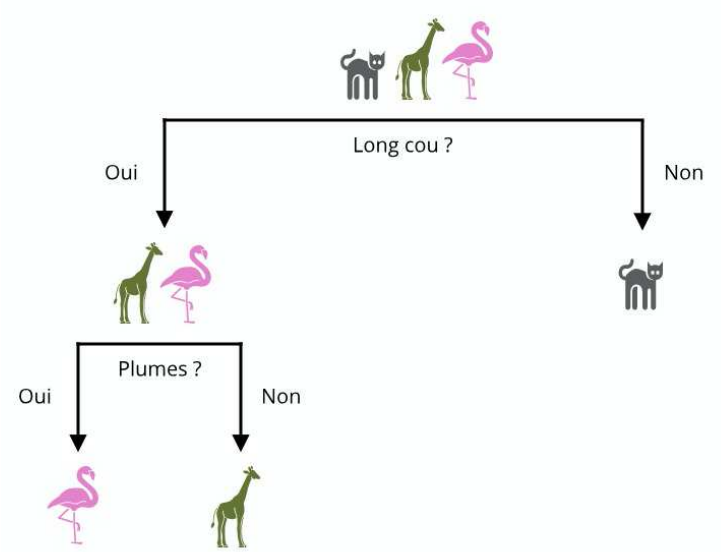

Figure 15 : Illustration d'un arbre de décision. En passant chaque nœud, une donnée pourra être classée de façon de plus en plus précise.

Les forêts aléatoires se basent sur trois principes :

- Dans un set de données de N observations, chaque donnée est décrite par p variables descriptives. B arbres de décisions de taille N seront créés et entrainés. La particularité est que le set d'entraînement pourra utiliser plusieurs fois une même observation (si le set de données est [1 ; 2 ; 3 ; 4], l'arbre peut utiliser [1 ;1 ;2 ;3] pour s'entraîner). C'est la technique du « bootstrap ».
- Sur les différentes variables descriptives disponibles, un nombre m inférieur à p, sera choisi au hasard et utilisé pour effectuer la meilleure segmentation possible.
- L'algorithme va combiner les différents arbres de décision, et la classe qui obtient le plus haut score sera sélectionnée.

Cet algorithme est particulièrement efficace car il combine plusieurs arbres qui sont non corrélés entre eux. Ainsi, si l'on retrouve une même réponse dans plusieurs arbres, la signification de la classification est également plus élevée (Figure 16).

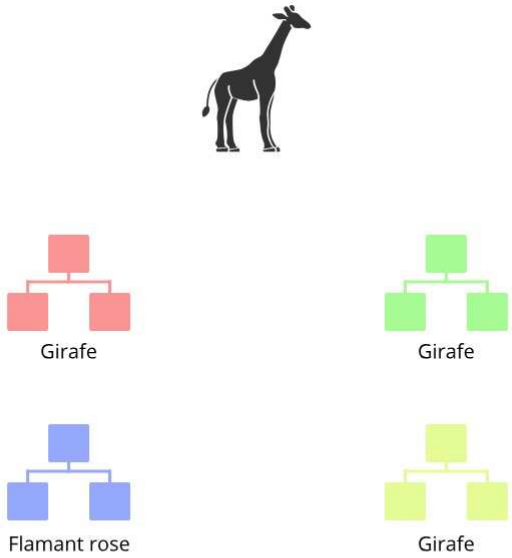

Figure 16 : Illustration d'un modèle de forêts aléatoires. 3 forêts prédisent la valeur "Girafe" contre une forêt qui prédit la valeur "Flamant Rose"

ix. L'apprentissage non-supervisé

L'apprentissage non-supervisé ne présuppose aucune donnée étiquetée. On peut parler de données brutes. Les algorithmes devront d'eux-mêmes regrouper les éléments qui se ressemblent en différents clusters. Il est utilisé afin d'identifier des relations cachées dans les données, ou encore afin de comprendre leur structure intrinsèque et de les regrouper de manière significative (Figure 17).

Posséder un set de données étiquetées n'est pas toujours possible car cela est très laborieux. C'est notamment le travail de certaines entreprises d'étiqueter tout un ensemble de données manuellement. Google propose ce type de service, où l'on peut faire étiqueter 1000 images pour environ 30 dollars (24). Nous étiquetons aussi certaines données sans le savoir, lorsqu'un site web nous demande, par exemple, de remplir un captcha avant de pouvoir poursuivre notre navigation (25).

Les utilisations de ML non-supervisé sont différentes de l'apprentissage supervisé. On les retrouve principalement dans les modèles de clustering et de généralisation.

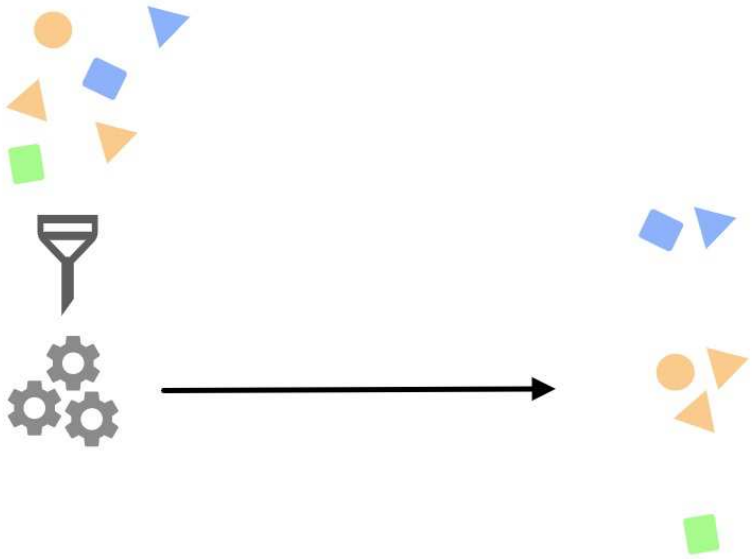

Figure 17 : Illustration d'un modèle de ML non supervisé. Les données sont brutes et l'algorithme va devoir trouver des tendances dans ces données. Ici par couleur.

Exemples d'algorithmes

L'algorithme des k-moyennes

L'algorithme des k-moyennes est l'algorithme le plus utilisé en apprentissage non-supervisé. Il est capable d'analyser la structure des données et de générer différents clusters.

Chaque observation étant décrite par un nombre p de descripteurs, l'algorithme va créer un nombre prédéfini k de clusters homogènes. Tout comme l'algorithme des k plus proches voisins, il va s'aider de l'espace en positionnant chaque observation dans un référentiel à p dimensions.

L'algorithme va alors placer aléatoirement k points intermédiaires dans l'espace et itérer de la façon suivante :

1) Il va d'abord relier chaque observation au point k le plus proche afin de définir un cluster intermédiaire.

2) Il va ensuite calculer le centre de gravité de ce cluster et le définir comme nouveau point central.

L'itération continuera jusqu'à ce que la somme des carrés de la distance des observations au point central k soit minimale (Figure 18).

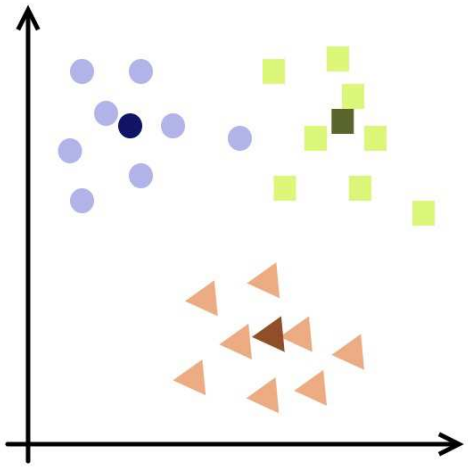

Figure 18 : Illustration du KNN. Ici k = 3, on veut différencier nos observations en 3 classes différentes. Les centres de gravités sont de couleur foncée.

x. L'apprentissage par renforcement ou semi-supervisé

L'apprentissage semi-supervisé se situe à mi-chemin entre l'apprentissage supervisé et le non-supervisé.

Dans ce cas, les données étiquetées sont remplacées par une évaluation donnée à l'algorithme, appelée agent. En fonction de ses actions sur un environnement, l'agent recevra une récompense ou une punition. Cette évaluation ne dit pas à l'algorithme si sa décision était optimale ou non, mais l'algorithme utilisera ce système de feedback afin d'améliorer son habilité à prendre des décisions et de maximiser les récompenses (Figure 19).

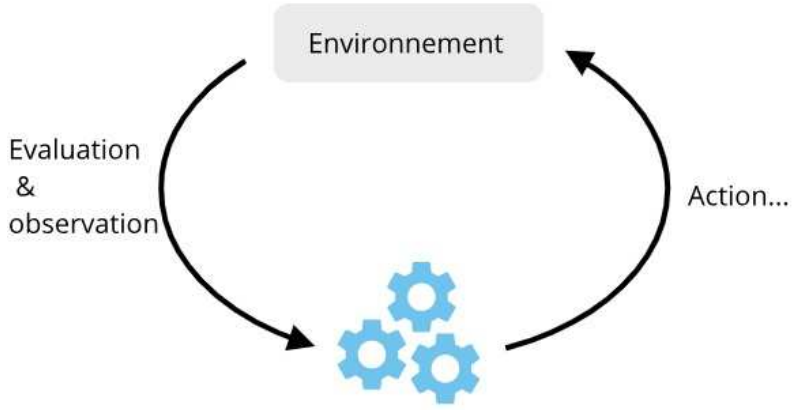

Figure 19 : Illustration de l'apprentissage renforcé. L'algorithme apprend selon un système de punition-récompense.

xi. L'apprentissage profond ou Deep Learning

Le Deep Learning (DL) est un sous-ensemble du Machine Learning qui est retrouvé dans de nombreuses applications. C'est par exemple le cas de la reconnaissance d'images (particulièrement de la reconnaissance faciale utilisée par les réseaux sociaux), de la reconnaissance vocale (utilisée par Alexa d'Amazon ou par Google Home) ou encore dans le Natural Language Processing (NLP). Ce dernier est une application très intéressante car l'algorithme va être capable d'extraire le sens d'un ensemble de mots donnés. Le DL est particulièrement efficace lorsque le contexte d'apprentissage et les relations entre les données sont complexes et non linéaires.

Le DL s'est construit sur une catégorie d'algorithmes appelés les « réseaux de neurones ». Ces derniers ont été inspirés par la même logique retrouvée dans le cerveau humain : chaque unité de processing ou chaque neurone est relié l'un à l'autre afin de générer un ensemble capable de résoudre des opérations complexes. McCulloch et Pitts ont été les premiers à proposer un modèle informatique de neurone en 1943 (26).

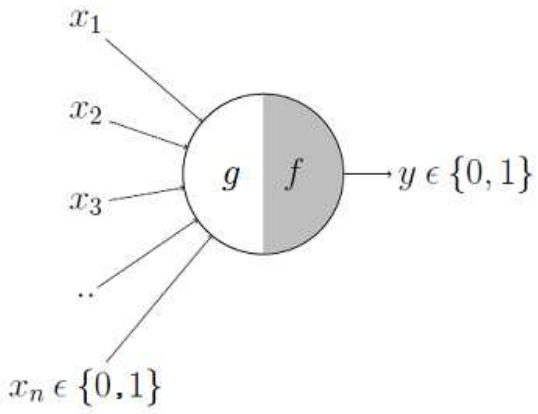

Figure 20 : Le premier modèle de neurone informatique.

Figure 20 : Tout comme un neurone reçoit des informations par ses dendrites, le corps du neurone informatique reçoit différents signaux (ou inputs) x_n. Dans ce modèle, les valeurs ne peuvent être que booléennes, c'est-à-dire vraies ou fausses. Le corps du neurone est par la suite divisé en deux parties, g et f. g est une fonction qui agrégera tous les différents inputs de la sorte : $g(x) = \sum_{i=1}^{n} x_i$.

L'autre partie du neurone, f, va prendre une décision selon que $g(x)$ soit supérieur ou non à une valeur seuil, appelée θ. f prendra une valeur y, elle aussi booléenne, vraie ou fausse {0,1}. Soit :

$$y = f\big(g(x)\big) = 1 \; if \; g(x) \geq \theta$$
$$= 0 \; if \; g(x) < \theta$$

Prenons un exemple afin d'illustrer ce premier concept. Imaginons que l'on conçoive un neurone de façon à ce qu'il nous aide à prendre la décision de partir en promenade ($y = 1$) ou non ($y = 0$).

Pour cela je base mon avis sur 3 inputs, x_1 : Il fait beau dehors, x_2 : Un ami m'accompagne et x_3 : Je suis de bonne humeur.

Au moment où je souhaite prendre une décision, il pleut (x_1 = 0), je suis accompagné d'un ami (x_2 = 1) et de bonne humeur (x_3= 1). Ainsi $g(x) = 2$. La valeur seuil θ indique en elle-même ma volonté naturelle à aller me promener, si elle était par exemple nulle ($\theta = 0$), qu'importe la météo, si une personne m'accompagne ou non et si je suis triste, j'irai me promener.

Dans notre cas $\theta = 2$, comme $g(x) = 2 \geq \theta$ainsi $y = 1$, j'irai me promener.

Il existe des signaux excitateurs et inhibiteurs. Ces derniers ont un effet maximum sur la valeur de y. Par exemple si je n'aime pas la pluie, qu'importe la valeur que prendront les deux autres inputs, je n'irai pas me promener s'il pleut dehors. Les signaux excitateurs sont ceux qui, une fois combinés entre eux, vont potentiellement activer le neurone.

Le perceptron

Un autre modèle, plus puissant, sera créé par Minsky et Papert, appelé le perceptron. Ce modèle va apporter deux innovations importantes. La première, les inputs pourront prendre des valeurs autres que booléennes, comprises entre 0 et 1. La seconde est que chaque input sera associé à un « poids »de valeur *w* (de l'anglais *weight*). Soit la fonction u définie par la somme des produits des inputs x_i et des poids w_i. Ainsi :

$$u_i = \sum_{i=1}^{n} w_i * x_i$$

Le poids symbolise la force synaptique que pourrait avoir un vrai neurone. Il y a également une différence importante, la valeur seuil θ ne sera plus choisie arbitrairement mais sera un nouvel input x_0 = 1 ayant pour poids $w_0 = -\theta$. La particularité de cette manipulation est que, maintenant, l'algorithme pourra déterminer par lui-même la valeur de ce seuil. Ainsi la convention établit :

$$u_i = \sum_{i=1}^{n} w_i * x_i - \theta$$

Il sera par la suite amélioré vers la fin des années 1960 par un algorithme de rétro propagation. Ce dernier correspond à la recherche des paramètres qui minimisent l'écart entre les prédictions et les valeurs observées, cela va notamment permettre à l'algorithme d'attribuer par lui-même des poids à ses différents signaux.

La somme de ces n signaux multipliés par leur poids et du biais b forme une fonction linéaire ui. L'algorithme va alors calculer le signal de sortie par une fonction non-linéaire de $f(ui)$, c'est la fonction d'activation. Cette dernière fonction est généralement une fonction sigmoïde : c'est le modèle le plus utilisé à ce jour.

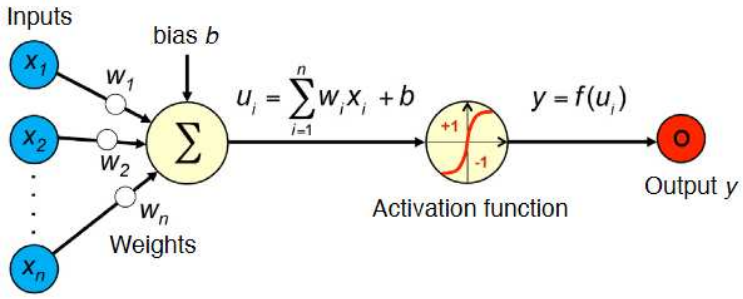

Figure 21 : Extrait de Lavecchia 2019 (27). Le modèle de perceptron va calculer un signal de sortie, selon toutes les entrées qu'il reçoit.

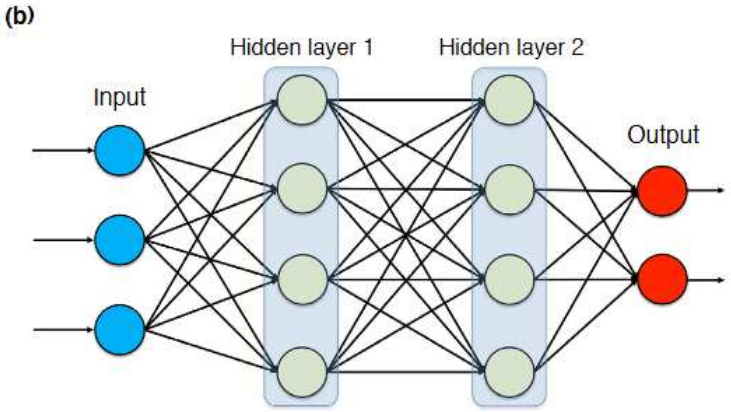

Figure 22 : Extrait de Lavecchia 2019 (27). Un modèle de réseau neuronal à 2 couches. Chaque neurone d'une couche n va avoir un « impact » la couche n+1.

Si l'on connecte maintenant de nouveaux neurones entre eux, nous obtenons un réseau de neurones. Selon les connexions choisies, le nombre de couches de neurones successives, il est possible de créer différentes architectures. Dans le Réseau neuronal (RN), il y a une couche d'entrée, un nombre x de couches cachées, et une couche de sortie. Quand le nombre de couches cachées est de plus de deux ou trois, le réseau est alors qualifié de « profond ». C'est pour cela qu'on parle de *Deep Learning* ou *d'apprentissage profond*. Chaque neurone étant connecté l'un à l'autre,

l'information sera alors traitée de façon de plus en plus abstraite au fil du réseau. Le désavantage de cette technique est que, l'information devenant de plus en plus abstraite, il devient de plus en plus difficile de la comprendre. Les RN sont de très puissants modèles prédictifs, mais de faibles modèles explicatifs. On parle de « black box ».

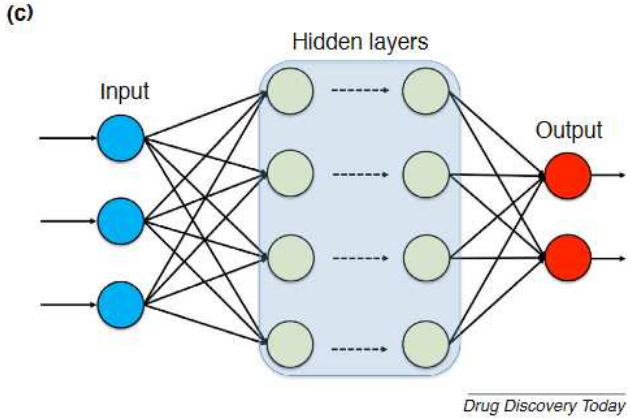

Figure 23 : Extrait de Lavecchia 2019 (27). Lorsque le nombre de couches devient important, l'information traitée devient de plus en plus abstraite. On parle alors d'apprentissage profond.

Le CNN : Convolutional neural networks

Le CNN est en partie responsable de l'amélioration de la vision informatique. Cette dernière est la capacité qu'a un ordinateur à reconnaître une image, à repérer certains éléments d'une vidéo, etc... C'est un type de réseau neuronal (presque littéralement, puisqu'il a été inspiré du cortex visuel des félins) particulièrement efficace pour cette tâche mais également dans le NLP. L'ordinateur va être capable de déchiffrer, tout comme nous, ce qu'on lui montre.

Dans notre cortex visuel, l'information captée par notre rétine suit le même modèle : notre première couche de neurones capte les contrastes, la seconde détecte les courbes, la suivante intègre la géométrie et ainsi de suite. Grâce à cela, l'information traitée est encodée de façon de plus en plus abstraite. C'est ce que tentent de reproduire les CNN.

Un CNN simple consiste en une succession de couches : une couche de convolution, une de correction, une autre de pooling, et enfin une couche dite « entièrement

connectée » qui donnera le résultat de classification qui correspond le mieux à l'image analysée.

La couche de convolution est la plus importante du réseau : c'est elle qui va extraire les motifs de l'image. Pour cela, elle va utiliser des filtres qui viendront échantillonner l'image à la recherche de lignes, de bords, de coins.

La couche de pooling permet de réduire le nombre de paramètres générés. Elle sert notamment à réduire la taille des données en sélectionnant par exemple le pixel ayant la plus grande activation.

Enfin, la couche entièrement connectée va analyser tous les motifs trouvés précédemment et prédire la classe de l'image analysée.

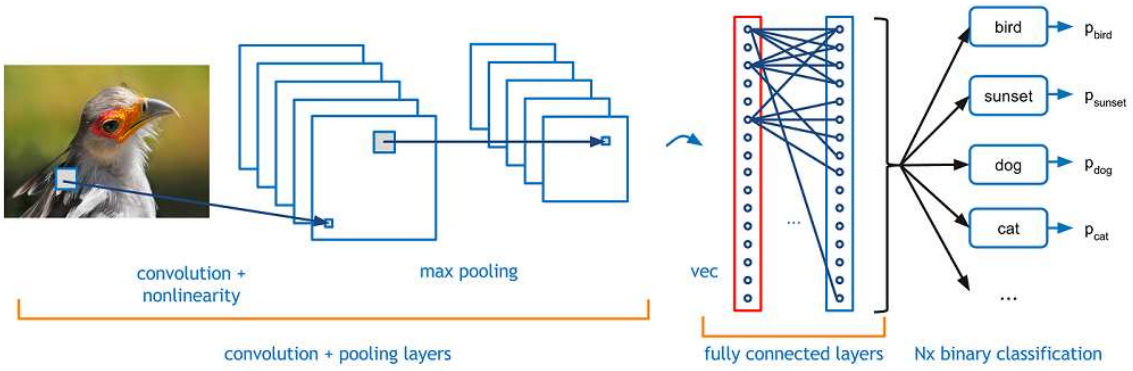

Figure 24 : Architecture d'un CNN. La partie de gauche se charge de repérer et de condenser les formes, les couleurs. L'information sera ensuite travaillée par un réseau de neurone, comme nous les avons déjà décrits. Figure extraite de TowardsDataScience.com (28)

Le réseau neuronal récurant

Le réseau neuronal récurant (RNN) est un algorithme capable de conserver une « mémoire » du passé. Lorsque l'on veut par exemple traduire une longue séquence de mots, l'algorithme peut perdre le sens de la phrase au fur et à mesure que l'on avance dans sa séquence. Quand l'on veut par exemple traduire une longue phrase, l'algorithme peut par exemple oublier à qui l'adjectif se réfère. En ayant cette mémoire, l'intégralité ou en tout cas une grande partie de la signification sera conservée, et une traduction plus fidèle sera proposée.

Le modèle du réseau neuronal récurrent possède une boucle de rétroaction. Ainsi, la sortie du signal dépend non seulement de ce que le neurone reçoit, mais également de ce qu'il possède en mémoire interne (Figure 25).

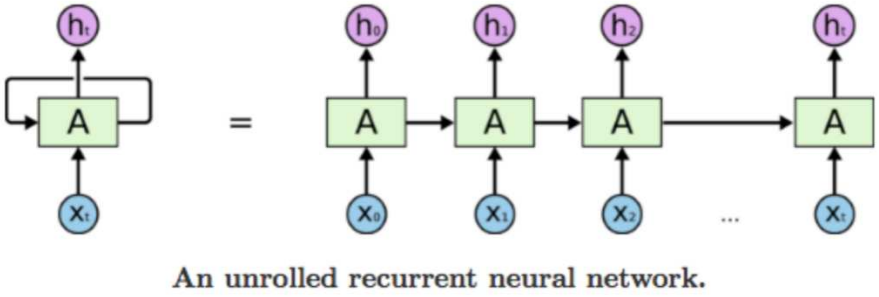

Figure 25 : Architecture du RNN. Chaque neurone possède une mémoire interne, témoin du passé. Figure extraite de colah.github.io (29)

Le RNN est couramment utilisé en tant que modèle génératif pour des données ayant une nature séquentielle, comme le NLP ou encore la musique. Il a reçu un intérêt croissant pour la génération de molécules *de novo* (30).

Les Réseaux antagonistes génératifs ou Generative Adversarial Network

Les Réseaux antagonistes génératifs (GAN) sont des algorithmes bien particuliers du monde du Machine Learning. Ils sont en effet capables de générer de nouvelles données en s'inspirant du set d'entraînement. Les GAN ont par exemple été exploités pour générer des œuvres d'art (31), des visages de personnes qui n'existeront jamais (Figure 26) (32) ou encore des morceaux de musique (33).De cette façon, un morceau a été entièrement généré dans le style de Sinatra, par l'Open AI jukebox (34).

Figure 26 : Tous ces visages n'existent pas et ont été entièrement généré par un GAN. Le site https://thispersondoesnotexist.com/ génère à chaque page un visage qui n'existe pas.

Dans cet algorithme, deux réseaux sont placés en compétition. Le premier est un générateur, il va produire une nouvelle donnée. Le second est le discriminateur, il va tenter de classifier la donnée comme réelle (i.e. appartenant au set d'apprentissage) ou comme générée par le premier réseau (Figure 27).

L'objectif du générateur sera de maximiser l'erreur de classification du discriminateur, c'est-à-dire de le rendre incapable de pouvoir différencier une donnée vraie d'une donnée générée. A l'inverse, le discriminateur va tenter de réduire cette erreur et informer le générateur de ses performances afin de l'entraîner. Cette boucle de rétro-propagation durera ainsi de suite, jusqu'à ce le modèle soit satisfaisant.

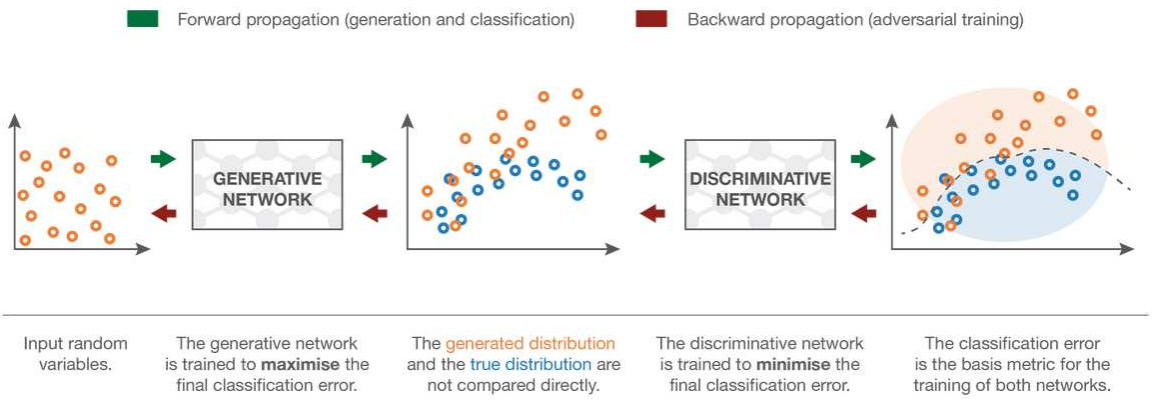

Figure 27 : Principe du GAN. Deux modèles sont en compétition l'un avec l'autre. Le génératif va essayer de duper le réseau discriminateur, jusqu'à ce que celui-ci ne soit plus capable de faire la différence entre une vraie donnée et une donnée générée. Figure extraite de towardsdatascience.com (35)

Ayant maintenant ces quelques notions en tête, nous allons pouvoir nous pencher de plus près sur ce que les sciences biomédicales ont développé.

III. Explorer la Chimie

Les bases de données regroupant des structures chimiques, des activité(s) biologique(s) etc, constituent une immense ressource pour les chimistes médicinaux. En effet, les programmes de Drug Discovery peuvent construire des modèles, allant de la prédiction de paramètres physico-chimiques, à la génération de molécules *de novo.* Cela conduira à une sélection priorisée des structures actives, et à l'accélération des projets de recherche.

Avant les années 2000, les banques de données étaient souvent constituées par les éditeurs de revues scientifiques, et étaient ensuite commercialisées. Aujourd'hui, de nombreux projets en accès libre sont disponibles à tous les chercheurs. Une banque de données comme ChemSpider, créée par la Royal Society of Chemistry, regroupe plus de 67 millions de structures, propriétés et informations associées. PubChem possède plus de 274 millions de données d'activités biologiques, réparties sur 109 millions de composés différents. La RCSB Protein DataBank (ou PDB) contient un grand nombre de structures cristallographiques et se révèle très utile lors de projets de structure-based design.

En Janvier 2020, un cap important a été franchi : le premier composé entièrement conçu par IA va être testé en clinique. C'est l'entreprise de biotechnologie Exscientia (36),(37) qui est à l'origine de cette avancée. La pathologie ciblée est le trouble obsessionnel compulsif. Le projet a mis cinq fois moins de temps à arriver à ce stade comparé à un projet classique. Le ML a énormément de potentiel pour tous les autres projets à venir.

G. Prédire une voie de synthèse

Lorsqu'un chimiste imagine une molécule, il passe par ce qu'on appelle la rétrosynthèse. Le composé cible est progressivement découpé en précurseurs plus simples, jusqu'à ce qu'ils soient disponibles commercialement.

Connaître la faisabilité chimique d'une molécule est important : cela permet tout d'abord au chimiste de choisir une voie de synthèse plus rapidement, d'estimer si une

molécule sortie d'une campagne de criblage virtuel est synthétisable, de prédire la faisabilité d'analogues lors de la phase de hit-to-lead, ou encore d'estimer la faisabilité de la synthèse industrielle lors de la phase de scale-up.

Historiquement, les outils de prédictions étaient programmés *via* des lignes de code qui suivaient des règles brutes, imposées par les chimistes. Aujourd'hui, elles passent par des méthodes statistiques qui classifient les types de réactions, et qui identifient les centres réactifs des molécules. L'avantage des algorithmes de ML est qu'ils vont généraliser ce qu'ils apprennent des données, les règles ne devront plus répondre avec exactitude à ce que le code établi à la main exigeait.

Les aides de synthèse informatique peuvent être divisées en deux groupes : celles qui passent par la rétrosynthèse : comment faire cette molécule ? Le deuxième : si je fais réagir A et B ensemble que se passe-t-il ? Et à quelle condition ?

Les données peuvent être extraites de la littérature, des brevets, mais également des carnets de laboratoires électroniques. Elles peuvent être également générées lors de campagnes de screening purement chimiques, lorsque les chercheurs testent aléatoirement différentes combinaisons dans l'espoir de trouver de nouvelles réactions.

Marwin Segler *et al.* (38) ont entraîné un réseau de neurones profonds combiné à une recherche arborescente Monte Carlo sur plus de 12 millions de réactions associées à des petites molécules. Le modèle est capable de trouver une voie de synthèse pour deux fois plus de composés qu'un logiciel codé manuellement, et est également 30 fois plus rapide. Leur algorithme aidera les chimistes à gagner du temps, et également à permettre des synthèses totalement robotisées. Le projet est même allé plus loin en testant leur algorithme sur un panel de chimistes en double aveugle. L'exercice consistait à proposer deux voies de synthèse, une reportée dans la littérature et l'autre proposée par l'algorithme. Les chimistes devaient ensuite choisir celle qui leur paraissait la plus probable. Étonnamment, ce sont les synthèses de l'algorithme qui arrivent en tête (57% vs 43% pour la littérature).

Cependant le modèle n'est pas encore capable de prédire la synthèse d'une molécule naturelle, beaucoup plus complexe en termes de chiralité et d'architecture. L'équipe estime toutefois que cela n'est pas impossible, un algorithme plus puissant et plus entraîné devrait être capable de résoudre ce type de problème.

xii. Une plateforme totalement automatisée pour la chimie en flux

Malgré les avancées en termes d'automatisation dans les laboratoires de chimie, l'exploration de nouvelles voies de synthèse s'effectue encore majoritairement à la main.

Dans le cas, par exemple, de la synthèse d'un composé complexe, il n'est pas rare que les chimistes passent du temps à tester différents catalyseurs pour optimiser une étape, et finissent par complètement abandonner la réaction en question car elle ne fonctionne pas dans la stratégie globale de synthèse. Afin de réduire cette consommation de ressources type « temps et travail », Coley *et al.* (39) du MIT ont conçu une plateforme robotisée gérant des techniques de chimie en flux continu. En s'entraînant via la base de données Reaxys, l'algorithme est capable de planifier des voies complètes de synthèse, tout en gérant le processus propre de la réaction (aller chercher tel produit à tel endroit, le laisser réagir autant de temps, etc.…) Le robot gère également les paramètres de stœchiométrie, qui sont particuliers à la chimie en flux. Le robot a ainsi achevé une étape importante en synthétisant 15 composés drug-like.

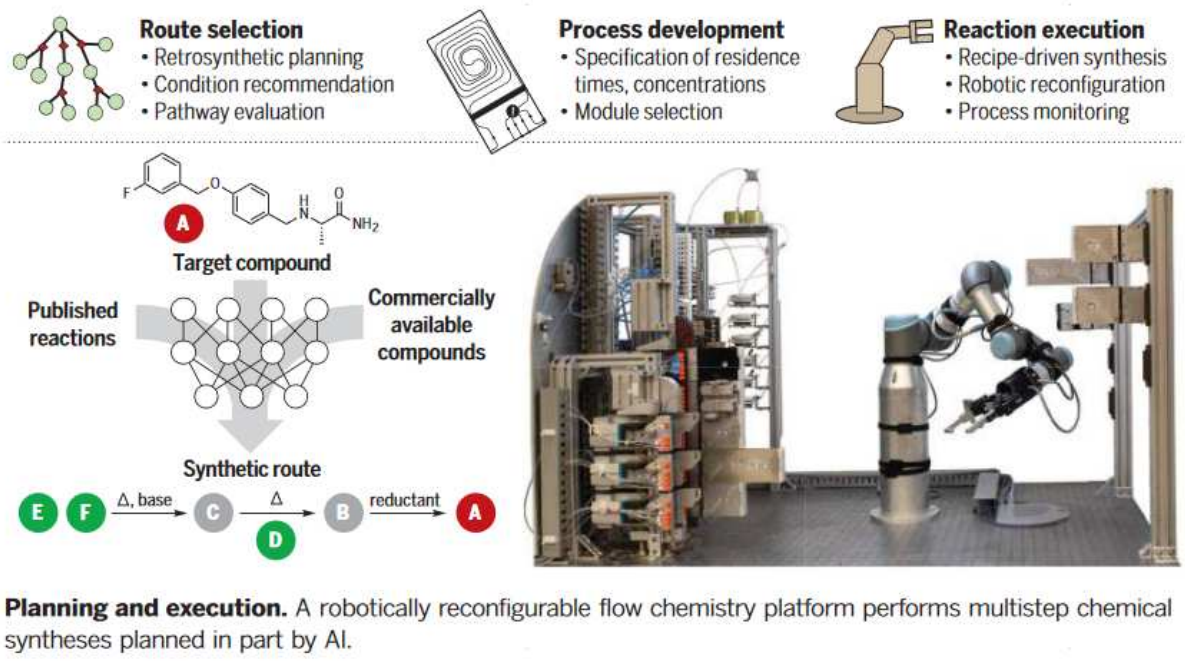

Figure 28 : Extrait de Coley et al. 2019 (40). Le robot (représenté sur la droite) planifie de façon presque autonome la réaction chimique, et apprend de ces propres expériences.

Dans le même esprit d'exploration de nouvelles réactions chimiques, Caramelli *et al.*(41) ont développé de leur côté un robot capable de choisir aléatoirement deux substrats et d'analyser le résultat grâce à la spectroscopie infrarouge et à la résonance magnétique nucléaire. Grâce aux résultats, le robot s'entraîne de façon autonome et prédit le résultat de futures réactions.

H. Prédire les propriétés physico-chimiques d'un composé

Les premiers travaux étudiant les relations quantitatives entre structures chimiques et activités variées, sont attribués à Corwin Hansch. En anglais on parle de Quantitative Structure–Activity Relationship (QSAR). Dans les années 60, ses recherches ont permis de mettre en évidence la corrélation entre l'activité de composés phytosanitaires, et leur hydrophobicité. Depuis, de nombreux outils se sont développés et de nombreuses propriétés physico-chimiques peuvent être estimées *in silico*.

Par exemple, le point de fusion d'un composé est une propriété particulièrement intéressante. Ce dernier est corrélé à sa solubilité aqueuse. Ainsi, lors de l'optimisation du composé sur ses paramètres ADME, estimer la solubilité permet d'estimer la biodisponibilité du futur médicament.

En analysant les points de fusions de plus de 275 000 composés extraits de OCHEM, Enamine et de brevets, Withnall *et al.* (42) ont ainsi pu comparer l'influence de simples descripteurs moléculaires sur le point de fusion des composés. L'objectif était d'analyser deux molécules similaires, qui ne diffèrent que par un changement de structure mineur. L'avantage de cette méthode est que l'algorithme va se passer de la structure entière, et ne se concentrer que les différences observées. Ainsi, même les industriels peuvent publier leurs résultats, sans révéler le « squelette » de molécules sensibles.

Comme un chimiste aurait pu l'estimer, le modèle a trouvé de façon totalement naïve que le point de fusion des composés est influencé par le nombre de liaisons hydrogène intermoléculaires, ou encore par l'interconversion de certaines fonctions chimiques par d'autres (Figure 29).

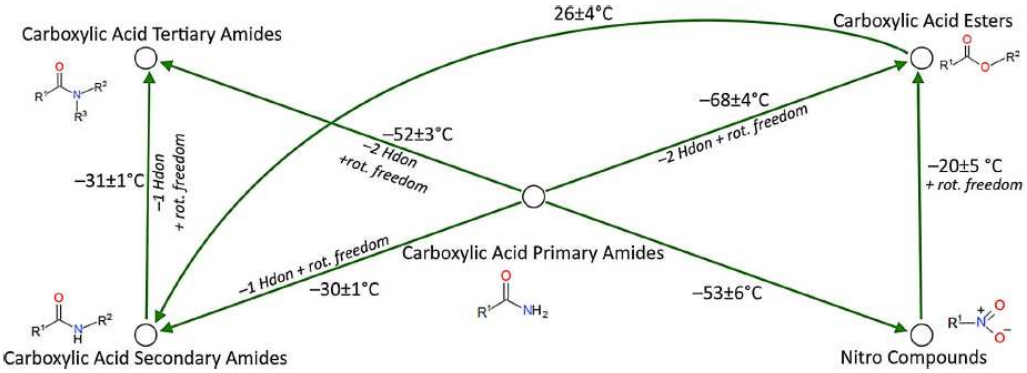

Figure 29 : Extrait de Withnall et al. 2018 (42). De façon totalement naïve et en comparant des molécules proches, l'algorithme a pu par exemple trouver que de passer d'un ester à un amide secondaire, augmentait le point de fusion du composé de 26°C en moyenne.

I. Prédire l'interaction d'un composé avec sa cible

Les projets de structure-based design (conception des molécules selon les données structurelles de la cible) reposent sur un score de liaison entre la molécule et sa cible. La première étape de ces travaux consiste à identifier les sites de liaison potentiels de la cible étudiée. Classiquement, les chercheurs se servent de quatre méthodes différentes (géométriques, évolutionnaires, énergétiques et statistiques), basées sur les connaissances préliminaires d'autres sites de liaison.

Les CNN sont connus pour être de puissants algorithmes dans la vision informatique. Jiménez *et al.*(43) ont ainsi repositionné ce type de réseaux de neurones, afin d'analyser plus de 7600 cibles protéiques et prédire leurs potentiels sites de liaisons. A l'inverse des algorithmes classiques qui se basent sur des méthodes géométriques pour chercher des poches, des cavités, le modèle ici décrit s'aide de la vision informatique pour estimer où le ligand peut se lier. De cette façon, leur algorithme DeepSite s'est entraîné uniquement par des exemples et sans règles imposées. L'outil s'est révélé plus performant que deux de ses concurrents (fPocket et Concavity). Les chercheurs peuvent accéder librement à l'algorithme sur http://www.playmolecule.org/ . Un fichier .PDB suffit au logiciel pour effectuer une prédiction.

Les méthodes de structure-based design se basent également sur des fonctions qui prédisent l'affinité d'un ligand pour une protéine. En utilisant une nouvelle fois un CNN, Ragoza *et al.*(44) ont pu créer un algorithme capable d'estimer si une molécule peut se lier ou non à la cible, et si le ligand est correctement positionné. Le modèle a été

entraîné sur deux sets d'interactions ligand-protéine en 3D : CSAR-NRC HiQ pour les poses, et DUD-E pour la partie ligand/non-ligand. Le CNN montre de meilleures performances comparé à un logiciel de référence comme Autodock Vina.

J. Explorer l'espace chimique d'une cible

Un espace chimique se définit par un ensemble de molécules possédant un ou plusieurs descripteurs moléculaires communs (45). Par exemple, l'espace chimique GDB-17 se définit par toutes les molécules synthétisables contenant au plus 17 atomes. Il est estimé qu'il regroupe 166.4 milliards de structures uniques. Un espace chimique peut aussi être centré sur une cible spécifique, telle qu'une enzyme. L'espace chimique des molécules drug-like (c'est-à-dire susceptibles de devenir un médicament) peut être toutes les molécules respectant la règle des 5 de Lipinski par exemple.

xiii. L'espace chimique d'une cible : l'exemple des bromodomaines

Il existe dans l'organisme humain 41 protéines possédant un bromodomaine. Les bromodomaines sont des modules capables de lire les modifications post-translationnelles de l'ADN ou d'autres protéines, et notamment l'acétylation des lysines. Les protéines contenant un ou plusieurs bromodomaines sont impliquées dans la régulation de la transcription, et jouent un rôle important dans le processus de cancérisation. Viser ces cibles serait donc une voie thérapeutique potentielle. BRD4 fait partie de ces protéines et constitue une cible intéressante.

Afin de cartographier l'espace chimique des molécules actives sur BRD4, Casciuc *et al.* (46) ont créé des modèles prédictifs grâce aux banques de données Reaxys et ChEMBL (en sélectionnant des molécules actives et non actives connues). Ils ont pu ensuite cribler 2 millions de composés de l'entreprise Enamine, et 29 composés ont été confirmés pour avoir une activité biologique. Les chercheurs se sont basés sur deux algorithmes, le SVM et le Generative Topographic Mapping (GTM).

Le GTM est un outil idéal pour les Big Data. Il permet de visualiser des données possédant de grandes dimensions (i.e. possédant plusieurs variables descriptives) en 2D. Il n'a pas de limite et peu, en théorie, analyser une quantité infinie d'informations. Ainsi, un point projeté sur la carte correspond à un type de structure, et son activité lui est associée (Figure 30).

Dans ce travail, les auteurs réaffirment l'utilité des banques de données publiques, et les considèrent comme des outils clés pour les projets de Drug Discovery.

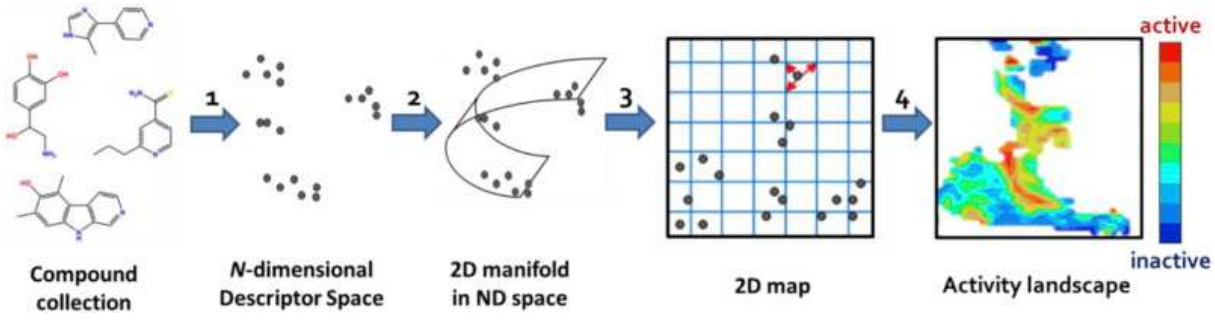

Figure 30 : Extrait de Casciuc et al. 2019 (46). Principe du GMT : technique de réduction de dimension d'un composé à multiples variables. La carte générée permet par exemple de visualiser les composés actifs ou inactifs d'une cible, là où leurs descripteurs étaient beaucoup trop nombreux.

K. Drug Design *de novo* et la data augmentation

Le principe du drug design *de novo* est de créer de nouvelles molécules, ayant une ou plusieurs propriétés désirées. Cela peut-être une activité sur une cible précise, tout comme posséder une solubilité particulière dans un solvant X.

En utilisant un réseau neuronal profond (RNN) ainsi qu'un algorithme d'apprentissage renforcé, Olivecrona *et al.* (47) ont pu construire un modèle capable de proposer de nouveaux ligands pour le récepteur à la dopamine de type 2. L'intérêt de combiner l'apprentissage renforcé au RNN est d'aider le modèle à éviter « d'oublier » ce qu'il a appris. De plus, en codant les structures chimiques en SMILES on retrouve une donnée séquentielle particulièrement adaptée aux RNN, qui pourront ensuite générer de nouvelles molécules.

Plus de 95% des structures générées ont ensuite été prédites comme actives grâce à un modèle de SVM.

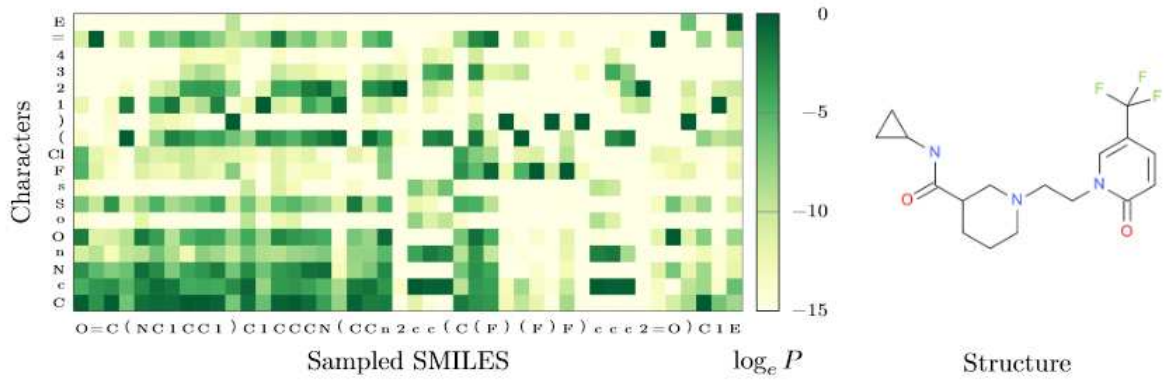

Figure 31 : Extrait de Olivecrona et al. 2017 (47). En abscisse la structure SMILE du composé, en ordonnée la probabilité de changer l'activité du composé. Plus la case est verte, moins

On peut analyser la façon dont le RNN procède pour générer de nouvelles molécules grâce à la Figure 31. Une structure chimique est traduite en séquence SMILES (axe des abscisses), et l'on regarde la probabilité que l'élément à telle position puisse être changé par un autre (axe des ordonnées). Plus la case est verte, plus les chances sont élevées de choisir cet autre élément. La molécule générée de cette façon est représentée à droite.

La data augmentation est le fait d'ajouter virtuellement du bruit à nos données afin de donner plus d'information et rendre le modèle plus robuste à l'overfitting. Afin d'illustrer ce concept, imaginons une photographie d'arbre. Si elle est tournée à 90°, le modèle doit toujours être capable de la reconnaître. De cette même façon, une molécule peut être représentée de différentes façons grâce à sa séquence SMILES, en fonction de là où l'on commence (Figure 32) et augmenter la représentation des données du modèle.

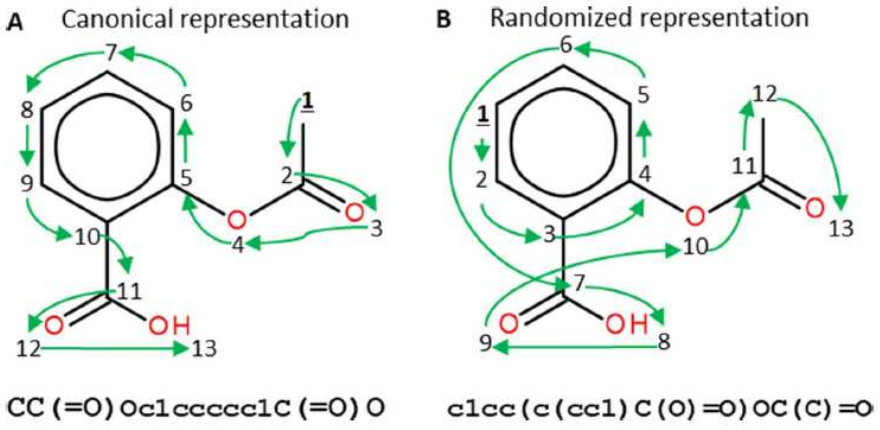

Figure 32 : Extrait de David et al. 2019 (48). Représentation différentes représentations SMILES pour un même composé. Cela permet à l'algorithme d'extraire une information plus abstraite, et de mieux généraliser ce qu'il a appris.

Ainsi, un algorithme entraîné par des séquences SMILES aléatoires génère un espace chimique plus restreint et uniforme : l'algorithme a mieux appris à généraliser et à repérer ce qui est essentiel à la structure du composé.

L. Explorer le biological fingerprint des composés

Les méthodes traditionnelles de Chémoinformatique se basent sur l'utilisation de descripteurs moléculaires afin d'étudier les activités biologiques de différents composés. Une méthode de référence est celle des « structural fingerprints », utilisée pour rechercher des composés de structure similaire à une molécule connue. Si un composé X possède une certaine activité biologique, un composé Y ressemblant à X a des chances de l'être également.

A l'inverse, le biological fingerprint se passe de données structurelles. Il ne se base que sur la similarité des activités biologiques que peut avoir un composé. Si un composé possède des activités sur des cibles A,B et C, un composé pour lequel seules ses activités sur A et B sont connues, a des chances d'être actif sur C également (Figure 33).

Afin de générer ce type de « carte d'identité », de nombreux efforts sont mis en place par les industriels. Novartis notamment, a collecté les résultats de plus de 1,5 million de composés, sur 135 tests cellulaires et biochimiques (48). La combinaison des fingerprints structurelles et biologiques est intéressante lorsque les chercheurs

explorent des possibilités de scaffold hopping, et veulent élargir le champ des hits identifiés lors d'une campagne de screening à haut débit (49).

Une campagne de screening génère toutefois plus de données « d'inactivités » que « d'activités », et il est difficile de générer le biological fingerprint d'un composé qui n'a jamais été testé. Afin de combler ce manque d'information et d'améliorer les prédictions d'activités des composés, Laufkötter *et al.* (50) ont construit un fingerprint « hybride », regroupant les données biologiques et chimiques. Même si leurs travaux ont démontré que les données de biological fingerprint contribuaient en grande partie à la performance du modèle, le structural fingerprint apportait un soutien important, là où les données d'activités biologiques étaient manquantes.

Il existe de nombreuses autres façons de construire l'identité biologique d'un composé. Par exemple en observant le profil d'expression génomique, lorsqu'il est perturbé par la présence d'une petite molécule. Plutôt que d'étudier la réponse d'une seule cible, le génome donne une vue d'ensemble de ce qu'il se passe dans la cellule. L'initiative « The connectivity Map » du MIT (51) permet d'analyser les profils transcriptionnels de plus de 5 lignées cellulaires en réponse à un ensemble de petites molécules approuvées par la Food and Drug Administration.

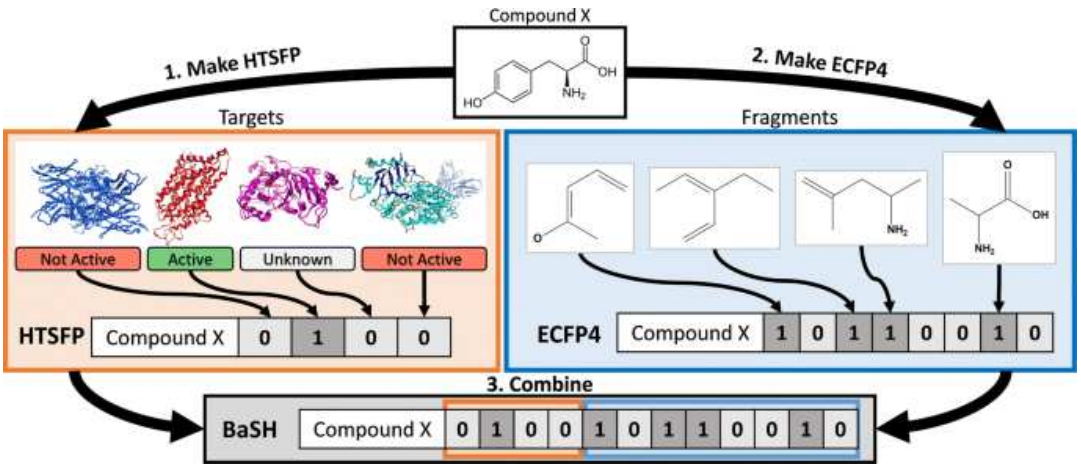

Figure 33 : Extrait de Laufkötter et al. 2019 (50). Une campagne d'HTS ne permet pas d'obtenir des informations sur toutes les cibles moléculaires, mais une approche combinant données structurelles et activités permet de combler ce manque d'information.

M. Insilico Medicine

Insilico Medecine est une entreprise américaine fondée par Alex Zhavoronkov. Le groupe est spécialisé dans l'usage du ML et de la Big Data dans des programmes de Drug Discovery.

En Septembre 2019, le groupe a réalisé un tour de force en générant une molécule biologiquement active en moins de deux mois. Pour cela, ils ont développé un réseau neuronal génératif appelé le GENTLR (pour Generative Tensorial Reinforcement Learning) (52).

Le Discoidin Domain Receptor 1 (DDR1) est un récepteur pro-inflammatoire à tyrosine kinase, activé par le collagène. Des programmes de recherche antérieurs ont pu relier son implication dans la fibrose et jusqu'à ce jour, 8 familles de composés ont démontré une activité bénéfique chez le rongeur lorsque DDR1 était inhibé. Cela a permis à l'équipe de tester leur concept tout en ayant suffisamment de données à utiliser. Il est important de noter à nouveau que tout IA ne peut pas prédire l'imprédictible, sans données préalables la force des prédictions est faible.

Le GENTLR permet de générer des composés *de novo* et a été entraîné sur 6 sets de données différents (voir Tableau 1). Le modèle a été désigné afin de prioriser les composés en fonction de trois paramètres : leur faisabilité chimique, leur efficacité sur cible et l'innovation apportée par la structure (comparée aux molécules déjà brevetées par les industriels pharmaceutiques ou les groupes académiques).

Tableau 1 : GENTLR a été entrainé sur 6 sets de données différents.

Base de données	Rôle
Set « Clean leads » de la banque de données ZINC	Améliorer la qualité « drug-like » d'un composé
Inhibiteurs connus de DDR1	Améliorer l'activité sur DDR1
Inhibiteurs communs de kinases	Améliorer la sélectivité sur DDR1
Molécules actives sur des cibles « non-kinases »	Améliorer la sélectivité et réduire les effets « off-target »

Données de brevets des entreprises pharmaceutiques	Diminuer la « ressemblance » avec d'autres composés déjà brevetés
Structure 3D des inhibiteurs de DDR1	Prédire la conformation et le mode de liaison à la cible

Dans ce modèle, la molécule la plus efficace (Figure 34) possède une IC_{50} de 10 nM sur DDR1 et une bonne sélectivité sur le spectre des kinases.

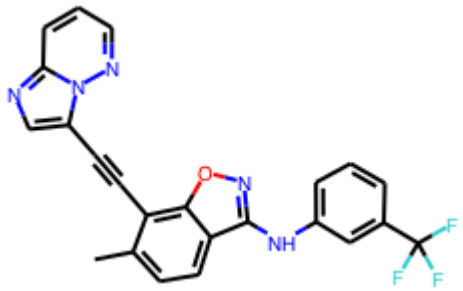

Figure 34 : Structure du composé généré par l'algorithme de InSilico. Il est actif à 10 nM sur DDR1.

L'auteur le rappelle lui-même, la molécule bien qu'active et biodisponible chez le rongeur, n'est pas encore un médicament et requiert davantage d'optimisation.

La communauté scientifique n'est toutefois pas entièrement convaincue, plaidant que la molécule n'est pas assez innovante. Un composé très similaire, le pogatinib a été approuvé par la FDA en 2012 et possède une activité de 9 nM sur DDR1 (Figure 35) (53). Pour certains, le remplacement isostérique de l'amide par un isoxazole n'est pas suffisant. L'entreprise s'est défendue en répondant que le but n'était pas de générer un composé entièrement nouveau mais de prouver le concept de leur modèle. Leurs efforts sont maintenus afin de convaincre davantage la communauté scientifique, qui reste encore sceptique (54).

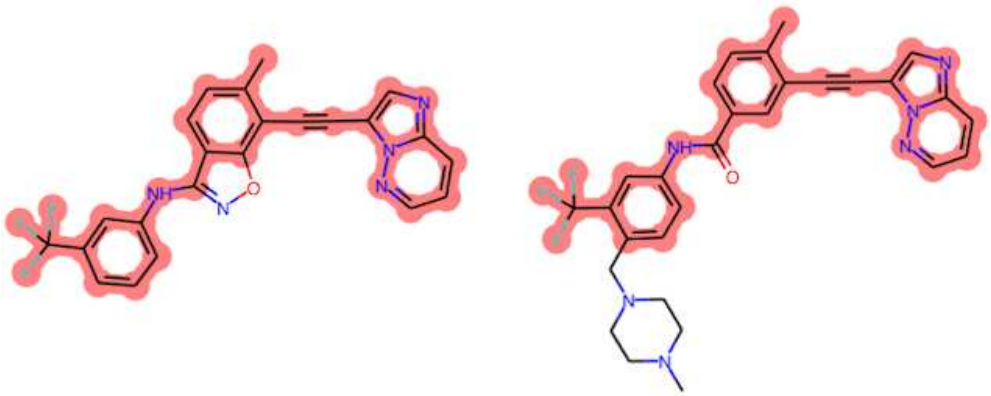

Figure 35 : L'activité du composé de gauche, généré par l'algorithme de InSilico n'est pas surprenante, compte tenu du composé de droite, ponatinib déjà commercialisé.

N. Prédiction des paramètres ADME

En 2012, la compagnie pharmaceutique Merck a organisé une compétition de prédiction de propriété ADMET qui a contribué à l'enthousiasme global des chercheurs pour les algorithmes de Deep Learning. En effet, comparés aux forêts aléatoires classiquement utilisées, les réseaux de neurones profonds étaient beaucoup plus performants (55).

L'optimisation des paramètres ADMET d'une molécule fait partie intégrale du développement d'un médicament. De nombreuses propriétés, telles que la pharmacocinétique, peuvent être dérivées d'essais cellulaires et biologiques de façon précoce. Par exemple, une molécule métabolisée trop rapidement sera plus vite écartée du projet de recherche. Ces données sont d'importants outils décisionnels, et le développement de méthodes *in silico* encourage la réduction de ressources animales (en accord avec les 3 R : Réduction, Raffinement, Remplacement (56)) et le gain de temps.

xiv. Prédire l'absorption d'un composé

Des modèles *in vitro* tels que Caco-2 (cellules cancéreuses du colon) ou PAMPA (Parallel Artificial Membrane Permeability Assay) sont couramment utilisés afin

d'estimer l'absorption d'un composé. Si les résultats sont mauvais, il est possible d'extrapoler que l'absorption du médicament par voie orale a peu de chance d'être efficace.

Même si ces modèles sont populaires et performants, le nombre de molécules à tester est toujours croissant, et cela représente un coût en temps et en argent. Afin de réduire l'utilisation de ces deux ressources, des modèles prédictifs *in silico* sont développés pour trier plus rapidement les molécules.

Shin *et al.*(57) par exemple, ont développé un réseau neuronal profond en extrayant des données de la littérature, ainsi que de ChemSpider. Leur algorithme permet de classer les molécules Caco-2 perméables ou non perméables, en étant plus performant que d'autres algorithmes classiques.

xv. Prédire la distribution d'un composé

Lorsque nous prenons un médicament, de nombreux facteurs influencent la distribution du composé. Si la molécule est lipophile, elle peut avoir un tropisme pour le foie. Ce peut-être la barrière hémato-encéphalique (BHE) qui empêche le médicament d'accéder au système nerveux central, ou encore les protéines plasmatiques qui se lient au composé et le « capturent ».

Les médicaments à visée du système nerveux central (SNC) sont particulièrement complexes à développer. Non pas parce que les composés manquent d'activité sur les cibles, mais parce qu'ils ont du mal à les atteindre. En effet, le SNC est extrêmement bien protégé par ce qu'on appelle la BHE. C'est une protection semi-perméable qui limite le passage de molécules, souvent toxiques. A cause de cette propriété, de nombreuses molécules échouent en clinique.

Miao *et al.*(58) ont récemment développé un réseau neuronal profond, entraîné sur les propriétés physico-chimiques de molécules, mais également sur leurs propriétés cliniques. Les modèles précédents ne se basaient que sur les caractéristiques physico-chimiques, ce qui était suffisant pour prédire un mode de diffusion passif, mais pas assez pour estimer une diffusion *via* des mécanismes plus complexes. Le glucose par exemple, bien qu'étant une molécule très simple, ne repose pas sur un mode de

diffusion passif. En implémentant des paramètres décrits en clinique (indications, effets secondaires), une information plus importante peut-être extraite par l'algorithme, et estimer si une molécule peut traverser ou non la BHE.

Comparé aux algorithmes utilisés précédemment, comme le SVM ou encore le KNN, le modèle est plus performant. Toutefois, il n'est pas encore possible de prédire *comment* la molécule diffusera au travers de la BHE.

xvi. Prédire la métabolisation d'un composé

Nous avons dans notre organisme plus de 30 isoformes de cytochromes P450 identifiés à ce jour. Ces enzymes métabolisent plus de 90% des médicaments et ont une grande influence sur la « vie » d'un composé dans notre organisme.

Beaucoup de médicaments inhibent ces cytochromes, ce qui cause de nombreuses interactions médicamenteuses. Le ritonavir est par exemple un inhibiteur connu des CYP P450 utilisé dans le traitement du VIH. Il « retarde » la métabolisation d'autres composés thérapeutiques, et permet de prolonger leurs effets bénéfiques. L'autre versant est qu'un médicament ne pouvant pas être éliminé va être présent en trop grande quantité dans le corps et produire des effets indésirables, néfastes pour la patient.

La détection de ces effets inhibiteurs peut être réalisée expérimentalement. Il est possible de les évaluer sur des cytochromes humains recombinants, *via* des tests de substrat ou encore *via* des sondes fluorescentes. Ces tests sont toutefois chers et prennent beaucoup de temps, même si des techniques de cocktails de CYP P450 ont été développées. Il est intéressant de développer un modèle capable de prédire si un composé est inhibiteur ou non, afin de prioriser les composés lors des phases de tests animaux.

Wu *et al.* (59) ont ainsi développé un modèle de classification de composés, en « inhibiteur de CYP P450 » ou « non inhibiteur de CYP P450 » sur les 5 isoformes majeurs. Leur algorithme XGBoost s'est révélé plus performant que les algorithmes de Deep Learning précédemment employés.

xvii. Prédire la toxicité d'un composé

Le profil toxicologique d'un médicament est considéré comme le paramètre le plus important lors de son optimisation. C'est également l'étape préclinique la plus critique. De nombreux médicaments ont été retirés du marché à cause de leur toxicité hépatique ou encore à cause de leur activité sur le canal hERG.

hERG est canal à potassium voltage dépendant, particulièrement connu pour avoir un rôle dans la cardiotoxicité de certains composés. Lorsque le canal est bloqué par un ligand, des effets néfastes du type prolongement du QT et des torsades de pointe sont observés. Actuellement, le screening des bloqueurs de hERG passe par des techniques de patch-clamp, particulièrement laborieuses et longues. Afin de répondre à cette problématique, Kim *et al.* (60) ont développé un modèle non seulement capable de prédire si un composé est antagoniste du canal, mais également de repérer les sous-structures chimiques responsables de cette activité (Figure 36).

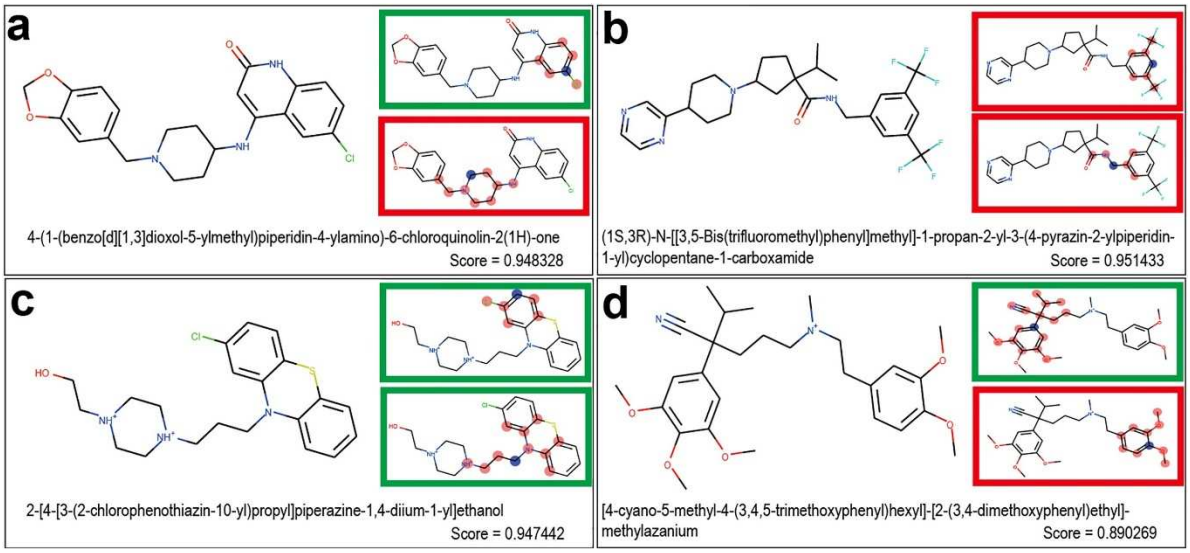

Figure 36 : Extrait de Kim et al. 2020 (60). Les encarts en vert sont les structures chimiques connues pour être ligand de hERG. En rouge, les structures non connues à ce jour, suspectées par l'algorithme de l'être également.

O. Le repositionnement d'un médicament

Le repositionnement d'un médicament est défini comme la découverte de nouvelles indications d'un composé ancien. C'est par exemple le cas du sildénafil qui était originellement utilisé dans le dysfonctionnement érectile, et qui est maintenant prescrit chez le patient atteint d'hypertension pulmonaire artérielle. Le repositionnement est très prisé car il permet aux laboratoires de donner un nouveau souffle à une molécule, ayant déjà été approuvée par les autorités de santé. Comme de nombreuses données ont déjà été générées (de toxicité, de pharmacocinétique etc.), cela est très avantageux pour le laboratoire qui n'a pas à investir une seconde fois.

Pour cela il est possible d'extraire des données de santé et de rechercher des effets indésirables mais thérapeutiques intéressants. Le challenge est important car une molécule qui a un profil d'action périphérique peut-être difficilement repositionnée dans un contexte où l'on veut agir sur le système nerveux central. Il a été par exemple prouvé par Khatri *et al.* (61) que l'utilisation de l'atorvastatine était favorable chez les patients greffés d'un rein, grâce à une étude rétrospective des journaux médicaux électroniques des patients inclus en essais cliniques. C'est également le cas de la metformine qui a montré un effet anti-cancéreux par Xu *et al.* (62) ou encore un effet protégeant des morbidités liées à la vieillesse.

En Février 2020, une équipe du MIT dirigée par James Collin (63) a identifié l'halicine (Figure 37), originellement étudiée dans le cadre du diabète, comme étant un puissant antibiotique. Le composé est actif sur de nombreuses souches bactériennes, dont des souches tuberculeuses et des souches jusqu'alors considérées comme résistantes. Cette découverte a été permise grâce à des algorithmes de Deep Learning, qui ont analysé plus de 2300 molécules antibactériennes, connues pour leur activité sur *E. coli* (comprenant des antibiotiques mais également des produits naturels).

L'algorithme a appris à prédire l'activité d'une molécule sans aucune information préalable, ce qui a lui a permis de découvrir des structures actives jusque-là inédites. Le modèle a ensuite été testé sur plus de 6000 molécules du Drug Repurposing Hub, avec pour contrainte de choisir des molécules différentes des antibiotiques classiques. 100 candidats ont été sélectionnés et l'halicine a été trouvée. Elle possède notamment des activités contre des souches de *Clostridioides difficile* (anciennement connue

comme *Clostridium difficile*) ou encore *Acinetobacter baumannii,* souches responsables d'infection nosocomiales pour lesquelles des composés thérapeutiques sont urgemment nécessaires.

Le mécanisme de l'halicine n'est pas totalement connu, mais il semble que le composé perturbe le flux de proton au travers de la membrane bactérienne (64). Il est efficace aussi bien *in vitro* qu'*in vivo* chez la souris. L'étape suivante a été de tester le modèle sur la banque de données ZINC15, sur plus de 107 millions de structures, et où 8 composés ont montré une activité antibactérienne.

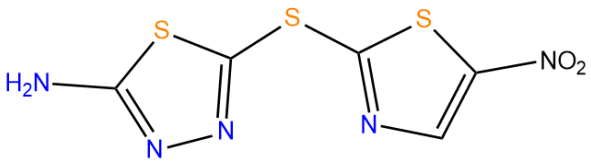

Figure 37 : l'halicine, une petite molécule initialement utilisée contre le diabète.

Du côté privé, Cyclica (65), une entreprise canadienne, a identifié une protéine clé dans la sclérodermie systémique. Une molécule qui existait déjà dans la prise en charge VIH a été repositionnée et testée dans cette nouvelle indication (66). Le groupe effectue de nombreux partenariats, notamment avec les géants Bayer et Merck.

IV. Explorer la Biologie et la Santé

L'avènement des méthodes dites « –omics » a permis une récolte importante de données biologiques. Ce champ d'exploration biologique s'attarde sur des domaines tels que la génomique, le protéomique ou encore le transcriptomique. Cela permet aux scientifiques de quantifier et d'interpréter ce qu'il se passe dans un organisme, une cellule, à différents niveaux d'observations.

Les technologies de séquençages actuelles, appelées les « NGS » pour Next Generation Sequencing, permettent d'obtenir très rapidement des données sur un génome complet, et à moindre coût (Figure 38). Ce sont par exemple des entreprises comme Oxford Nanopores ou encore Illumina qui ont développé ces outils et qui participent à la croissance toujours plus grande des bases de données génétiques (67,68).

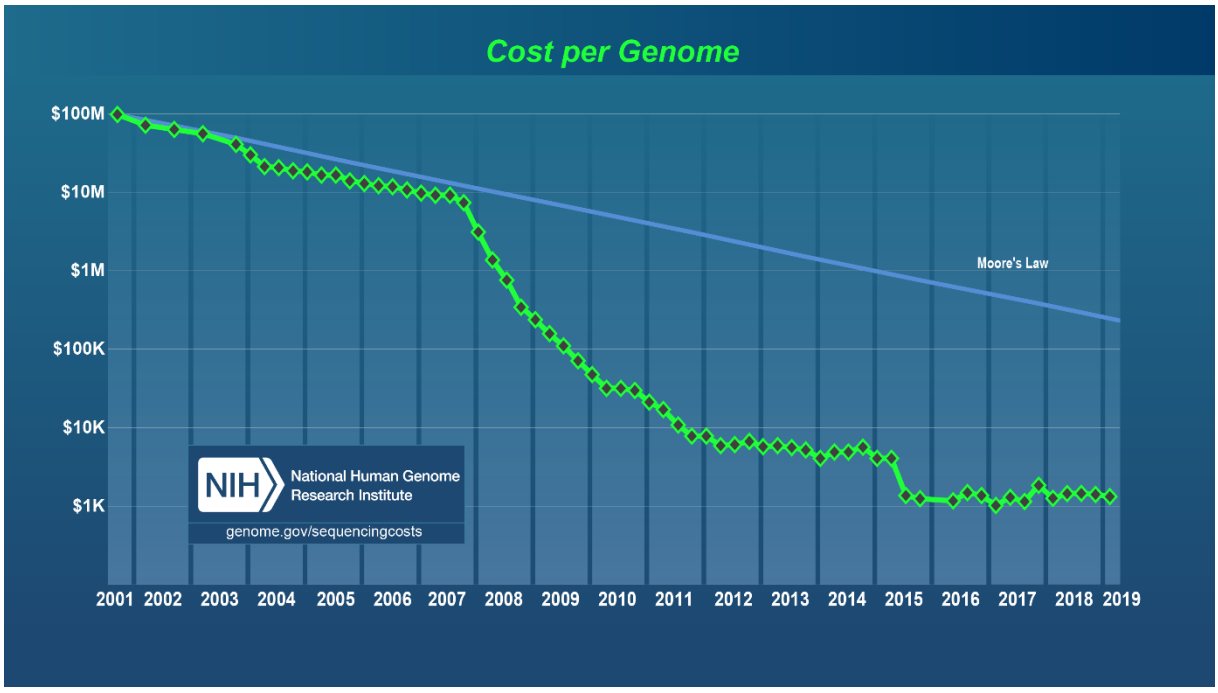

Figure 38 : le prix du séquençage complet d'un génome a diminué drastiquement au fil des années. En 2020, un test coûte environ $1000. Source : genome.gov (69)

Le génome humain a été complètement séquencé en 2003 mais il n'est pas encore entièrement compris. A titre d'exemple, nous ne comprenons toujours pas à quoi servent les grandes régions non-codantes du génome.

On arrive aujourd'hui à caractériser et à dresser le profil moléculaire de patients sains comme celui des patients malades à différentes échelles. Toutefois, les données étant extrêmement riches, le ML est nécessaire à la construction des modèles. Ces technologies permettent de se rapprocher toujours plus d'une médecine ultra-personnalisée. Demain, un médicament sera destiné à un patient unique de par son profil moléculaire. C'est par exemple le cas de Mila, une petite fille atteinte de la maladie de Batten, qui a reçu un oligonucléotide anti-sens conçu uniquement pour elle (70).

P. Découvrir et valider des nouvelles cibles thérapeutiques

Lorsque l'on veut traiter une pathologie, la première étape consiste à identifier une cible. Il faut ensuite émettre l'hypothèse que lorsque celle-ci est altérée, cela va moduler ou causer la pathologie. La deuxième étape, appelée validation de la cible, consiste à utiliser des modèles *ex vivo* ou *in vivo* afin de confirmer la causalité entre la cible et la pathologie.

Les études GWAS (pour Genome Wide Association Studies) consistent à rechercher des petites variations dans le génome de certains patients, afin de découvrir des gènes potentiellement responsables de maladies (Figure 39). Ces petites mutations, appelées Single Nucleotide Polymorphism (ou SNP), ne consistent qu'en une seule base mutée et sont difficiles à déceler car il faut analyser et comparer un grand nombre de patients. Grâce à une analyse statistique puissante, le SNP sera par exemple retrouvé à une fréquence plus élevée dans un groupe de patients malades.

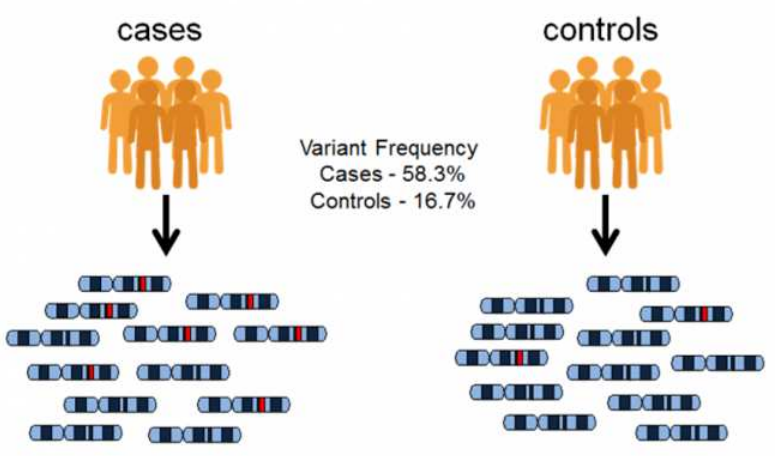

Figure 39 : Principe d'une étude type GWAS. Une mutation sera retrouvée à une fréquence plus élevée chez les malades. Figure extraite de ebi.ac.uk (71)

En se basant sur cette méthode, Finan *et al.* (72) ont développé un modèle aidant à l'identification et à la validation de cibles potentielles. Ils ont également élaboré un outil de repositionnement de molécules déjà autorisées. Le principe est le suivant : les SNP affectant un gène associé à l'expression ou à l'activité d'une protéine sont équivalents à une intervention médicamenteuse sur cette même cible. A titre d'illustration, une mutation affectant le gène *HMGCR,* impliqué dans la synthèse de précurseurs du cholestérol, produit chez les patients le même phénotype qu'un patient sain traité par statine (73).

L'équipe a pu ainsi connecter les loci identifiés lors des GWAS, au génome druggable (potentiellement cibles de médicaments) et aux molécules déjà existantes. De cette façon, ils ont pu découvrir plus de 1400 gènes druggables impliqués dans des pathologies, et 240 cibles pour lesquelles des composés existent déjà.

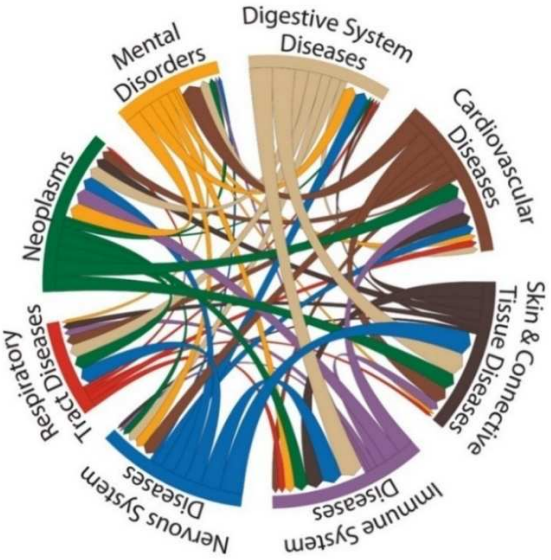

Figure 40 : Repositionnement potentiel de composés déjà indiqués dans une aire thérapeutique. Plus la flèche qui relie deux aires est épaisse, plus il y a de composés potentiellement repositionnables dans cette classe. Figure extraite de Finan et al. 2017 (72)

En effectuant ces études de GWAS, l'équipe a également pu découvrir que certaines classes de médicaments pourraient éventuellement être repositionnées dans d'autres pathologies, là où le même gène est dysfonctionnel (Figure 40).

Une autre étude de Jeon *et al.* (74) a décrit la construction d'un modèle de SVM, capable de classifier des protéines en cibles de médicaments ou non, pour certains cancers. Ils ont utilisé divers sets de données et ont analysé l'expression d'ARN messagers, le nombre de copies d'ADN, l'occurrence des mutations et la topologie des interactions protéines-protéines. En tout, 122 cibles communes ont été repérées dont 69 cibles déjà connues, preuve que le modèle fonctionne.

xviii. 23andMe et l'utilisation de nos données

L'entreprise 23andMe développe des tests ADN disponibles en vente libre. Le produit se veut ludique, et propose à ses utilisateurs de retrouver leurs traits ancestraux ou encore d'estimer leur sensibilité à certaines pathologies. En se basant sur les données de leurs utilisateurs (plus de 10 millions de tests), les chercheurs ont pu découvrir une cible impliquée dans de nombreuses maladies inflammatoires (75). La cible n'est pas

révélée à ce jour, mais un partenariat avec la société espagnole Almirall a été établi, dans le but de développer de nouveaux principes actifs.

Même si plus de 80% des usagers ont accepté les conditions légales du test (qui consistent à autoriser la recherche sur leurs données génétiques), peu connaissent la valeur réelle de leur génome. Les données patientes sont compliquées à obtenir depuis la loi RPGD, surtout pour les entreprises à visée commerciale. De cette façon, l'entreprise a su créer sa propre base de données, tout en restant lucrative. Les utilisateurs ne déboursent pas moins de $100 pour un test, mais à qui reviendraient les profits tirés de cette nouvelle cible ? Il est légitime de questionner l'éthique de l'entreprise.

Q. Prédire la structure tertiaire d'une protéine

De nombreuses pathologies sont fortement liées au dysfonctionnement d'une protéine. Ainsi, comprendre comment la protéine se replie et se positionne dans l'espace aide à comprendre comment la pathologie se développe. C'est ce qui arrive par exemple pour la mucoviscidose (76) ou encore pour la maladie d'Alzheimer (77). Connaître cette structure permet également de développer de nouveaux ligands, et de potentiellement soulager la pathologie.

Nous avons aujourd'hui la structure 1D de nombreuses protéines : la séquence d'acides-aminés qui les compose est connue, mais ce n'est pas pour autant que le problème de la structure en 3D soit résolu (Figure 41). Il est principalement dû au fait qu'une protéine classique possède un nombre astronomique de configurations possibles (estimées entre 3^{300} et 10^{143}), et qu'il prendrait plus de temps pour toutes les énumérer que le temps écoulé entre la naissance de l'Univers à aujourd'hui. C'est ce qu'on appelle le paradoxe de Levinthal (78), énoncé en 1969.

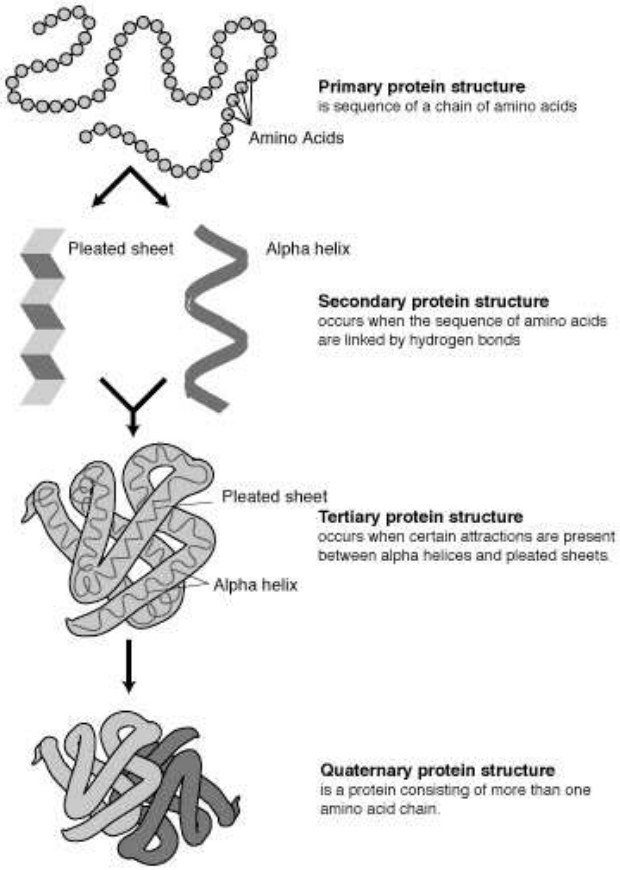

Figure 41 : Les différentes dimensions que peut prendre la séquence d'une protéine. Figure extraite de genome.gov (79)

Les techniques expérimentales actuelles passent par la résonance magnétique nucléaire, la cryo-électro microscopie ou encore la cristallographie par diffraction des rayons X. Ces méthodes requièrent toutefois un nombre important d'essais pour un faible taux de succès et ce, à un prix élevé. Le coût peut aller jusqu'aux millions de dollars par structure (80). C'est pour cela que la prédiction de ces structures en 3D par différents algorithmes est particulièrement intéressante.

En pratique, la protéine est décomposée en paquets d'acides aminés, et l'algorithme va tenter de reproduire les interactions idéales entre acides aminés, souvent *via* des liaisons hydrogènes mais des ponts disulfures sont également possibles.

Une autre approche consiste à récupérer les données de structures 3D de protéines connues, et d'identifier celles qui ont des structures similaires. Cela permet de construire un modèle pour les protéines inconnues, on appelle cela « la modélisation de protéines par enfilage » ou encore le « fold recognition » (81).

Les prédictions actuelles ne sont pas encore parfaites, mais les algorithmes d'apprentissage profond ont montré des résultats prometteurs et la Protein Data Bank est la source idéale pour entraîner les différents modèles.

En 2015, Spencer *et al.* ont par exemple utilisé un Deep Belief Network, afin de prédire la structure secondaire de protéines avec une précision de 80,7%. Un Deep Belief Network est un type de réseau de neurones qui ne connecte plus les neurones directement entre eux, mais les couches entièrement entre elles. Toutefois, même si les résultats des structures secondaires sont bons, il est difficile d'accéder aux structures tertiaires et encore plus aux structures quaternaires.

Afin de tester leurs algorithmes, les bioinformaticiens sont invités au Critical Assessment of protein Structure Prediction (CASP). Le but de cette compétition est simple : prédire la structure de trois protéines de difficulté croissante, grâce à leur seule séquence primaire. La véritable structure de la protéine est connue, et l'équipe qui s'en rapproche le plus a gagné.

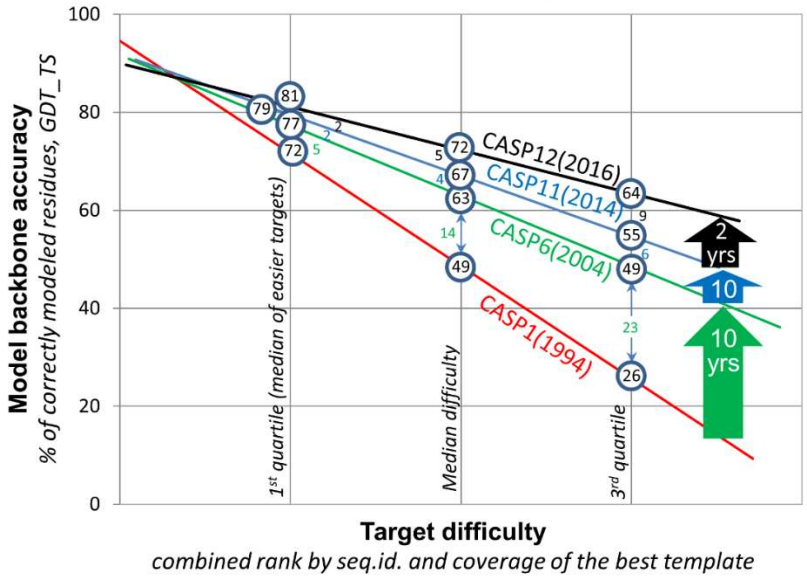

Figure 42 : Pourcentages de prédictions correctes en fonction des années du concours de CASP. En seulement deux ans la cible la plus difficile à modéliser a vu une nette amélioration. Figure extraite de predictioncenter.org (82)

La Figure 42 analyse les performances des équipes au fil des années. On peut voir que pour la troisième cible (la plus difficile) il y a eu peu d'amélioration entre 2004 et 2014, mais avec qu'avec le renouveau des réseaux de neurones (et notamment de

DeepMind), la cible a été prédite avec une précision de 9% supérieure en seulement deux ans (contre 6% en 10 ans lors des années précédentes).

xix. DeepMind et AlphaFold

En 2018, c'est DeepMind (une filiale de Google) qui a remporté la compétition du CASP, grâce à leur algorithme AlphaFold (83).

L'algorithme se repose sur deux méthodes, basées sur des réseaux de neurones profonds. La première consiste à prédire la distance entre les paires d'acides aminés. La deuxième à prédire l'angle entre les liaisons chimiques de ces acides aminés.

Ils ont également entraîné un GNN afin d'inventer de nouveaux fragments d'acides aminés, et d'améliorer le score de la structure prédite.

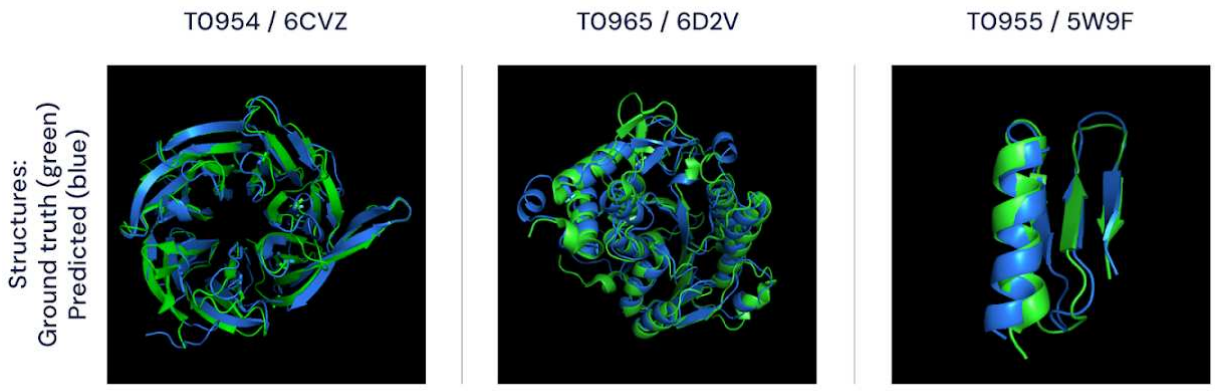

Figure 43 : Exemples de la précision du modèle DeepFold, en vert le "vrai" modèle et en bleu le modèle généré par l'algorithme de DeepMind. Source DeepMind.com (83)

Cet exemple est très important : il démontre que les GAFAM sortent de leur fenêtre marketing et investissent dans la santé et la recherche fondamentale. Google s'est imposé en seulement quelques années grâce à ses moyens colossaux.

Récemment, DeepMind est allé encore plus loin avec le nouvel algorithme « AlphaFold 2 ». En 2020, le laboratoire a résolu la structure d'une protéine à près de 90% de

similarité (Figure 44), score considéré comme similaire à la protéine sous forme cristalline (84). L'algorithme pourra soutenir les projets de structure-based design, mais la structure quaternaire, elle, est encore un challenge à résoudre.

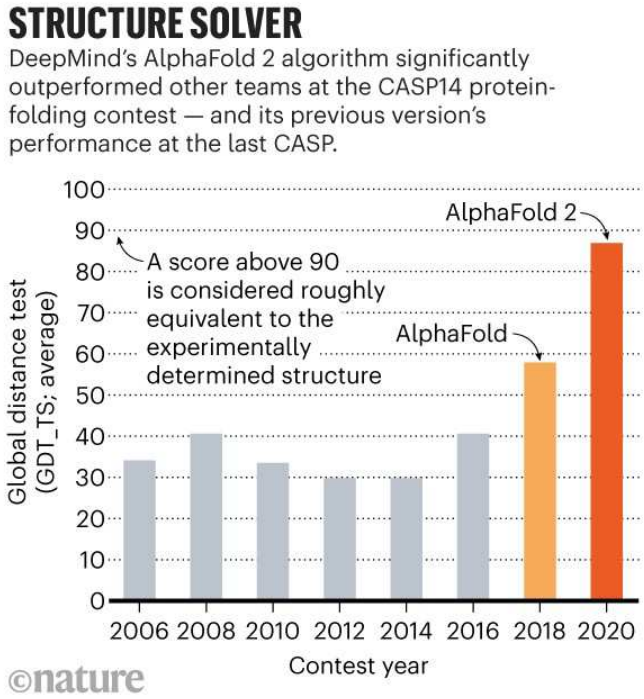

Figure 44: Extrait de Calaway 2020 (84). Performance de AlphaFold 2 durant le CASP14. L'algorithme approche une telle similarité avec la structure cristallographique, que de nombreux scientifiques parlent de "Révolution".

R. L'analyse d'images

L'analyse d'images est particulièrement utile pour les biologistes. L'utilisation de sondes fluorescentes, de marqueurs ou encore et simplement d'un organisme entier (depuis la cellule jusqu'à l'animal entier) génèrent une information très riche et souvent très bruyante. L'habilité des algorithmes de Machine Learning (particulièrement les CNN) à exploiter ce genre de données a été prouvée à tous les stades de développement préclinique.

De plus en plus d'images de hautes qualités sont disponibles, permettant un meilleur entraînement des algorithmes. Une banque de données comme The Cell Image Library regroupe plusieurs centaines de clichés biomédicaux (85). L'utilisation d'automates pour classifier les images n'est pourtant pas récente. Toutefois, les CNN

se sont révélés bien supérieurs car plus aptes à généraliser et à extraire un niveau d'information beaucoup plus abstrait. Les CNN sont appliqués partout, depuis la super résolution (permettant de mieux « voir » l'image) jusqu'à la classification de cellules en fonction de leur stade de division.

Une campagne de HCS (pour High Content Screening) contient une information beaucoup plus riche qu'une campagne de HTS, là où un seul endpoint est analysé. Le High-throughput imaging est régulièrement utilisé pour découvrir des composés modifiant la morphologie cellulaire ou encore la localisation subcellulaire de certaines protéines.

En se basant sur des images entières de cellules cancéreuses du sein, Kraus *et al.* (86) ont ainsi pu classifier des petites molécules selon 12 catégories différentes en fonction de leur mode d'action.

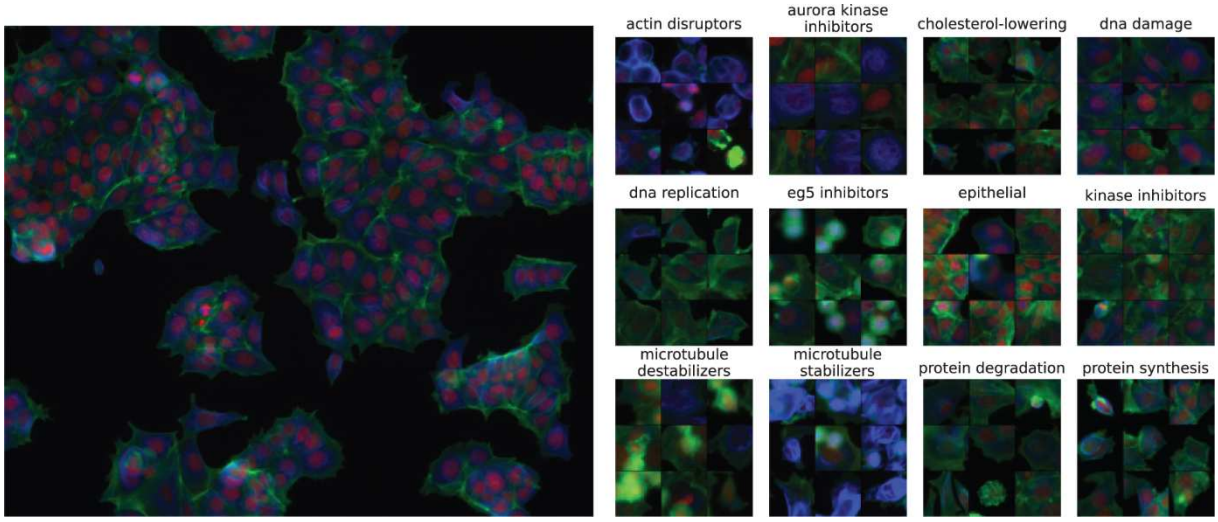

Figure 45 : Extrait de Kraus et al. 2016 (86). Le modèle a pu classer des images en fonction du mode d'action des petites molécules.

Google a également développé un algorithme capable de reconnaître des échantillons de tissus de cancer de la prostate avec plus de précision que ne l'a fait un groupe d'histopathologistes entraînés (87). A nouveau, les GAFAM se montrent de plus en plus agressifs dans le secteur de la Santé.

Les CNN peuvent être également utilisés afin de segmenter les cellules en compartiments sub-cellulaires(88) ou compter le nombre de colonies bactériennes sur agarose (89). Ils sont également très efficaces dans la pratique clinique courante. En effet, ils seront bientôt implémentés dans les hôpitaux et cabinets afin d'aider les

cliniciens à interpréter des IRM, des tomographies, des radios par rayons-X. C'est le cas de l'interprétation des ECG (90).

S. Biomarqueurs et médecine de précision

Un biomarqueur est une caractéristique biologique mesurable. Il est utilisé comme indicateur dans le diagnostic, la réponse de l'organisme à un médicament ou encore comme outil de pronostic (91). On peut par exemple noter le biomarqueur ALK dans le cancer du poumon à non petites cellules, qui une fois détecté chez le patient, conduit à la prescription d'un traitement particulier (92) (i.e crizotinib).

Découvrir ces biomarqueurs le plus tôt possible est très avantageux : les phases cliniques sont longues et coûteuses, et stratifier les patients permet d'augmenter les chances de réussites lors de l'essai. Li *et al.* (93) ont par exemple utilisé des données *in vitro* pour définir la sensibilité de certaines lignées cellulaires cancéreuses à deux traitements différents (i .e l'erlotinib et le sorafenib). L'étude a ensuite adapté le modèle afin de classer des patients selon leurs caractéristiques, et leur prescrire le traitement le plus adapté.

Un autre exemple vient du UC Davis MIND Institute, où des chercheurs californiens ont pu mettre en relation la présence de certains autoanticorps chez la femme enceinte, et le risque d'autisme chez le nouveau-né. Ramirez-Celis *et al.*(94) ont ainsi étudié la réactivité du plasma maternel à 8 différentes protéines, retrouvées en abondance dans le cerveau du fœtus. En comparant le plasma de mères d'enfants autistes et de mères témoins, le modèle de ML a pu identifier des motifs récurrents, et prédire avec une grande précision si le fœtus est à risque d'autisme. Les chercheurs veulent poursuivre leurs efforts, et développer des thérapies capables de neutraliser ces autoanticorps.

Toutefois, de nombreuses données cliniques ne sont pas disponibles, notamment car les laboratoires pharmaceutiques ne partagent pas entièrement leur base de données interne, pour des raisons stratégiques. Toutes les open-data en essais cliniques apparaissent alors comme une solution innovante pour permettre la recherche de nouveaux biomarqueurs chez les patients. Cela permettrait notamment de « sauver »

les médicaments ayant échoué lors des phases cliniques et de les repositionner chez une population répondeuse.

T. Deep Longevity et la prédiction des âges biologiques

Les médicaments nous font vivre plus longtemps mais vieillir est inévitable. L'âge est considéré comme le principal responsable du dérèglement de notre organisme, et de l'installation de différentes pathologies chroniques. La question du « vieillir bien » prend de plus en plus d'importance dans l'industrie, il est estimé que la longévité est un domaine à plusieurs trillions de dollars (95).

Des outils sont actuellement en cours de développement afin de repérer des facteurs qui pourraient améliorer notre bonne santé sur le long terme. La metformine par exemple, a déjà été identifiée comme bénéfique à un meilleur vieillissement (96). Il est toutefois difficile de savoir quelle intervention a un réel impact sur notre vie, sachant que les données s'étalent et se diluent sur plusieurs décennies. Encore une fois, l'apprentissage profond est l'outil idéal pour ce genre de situation, où les données sont à variables multiples et bruyantes.

Nous avons tous un âge chronologique, légal, mais nous ne vieillissons pas tous de la même façon. Il arrive que l'âge biologique prenne de l'avance sur l'âge chronologique, et cette observation est corrélée à des taux de mortalité plus élevés. Zhavoronkov *et al.* (97), de la société Deep Longevity, ont développé un réseau neuronal profond afin d'analyser les âges biologiques de différentes populations, afin de les comparer à leur âge légal. L'étude a par exemple permis de mettre en évidence que la population de l'Europe de l'Est est « plus vieille » qu'elle ne l'est, et qu'à l'inverse, la population coréenne apparaît « plus jeune » comparé à son âge chronologique. Le réseau de neurones s'est entraîné sur des millions d'échantillons sanguins, en se basant sur 20 biomarqueurs sanguins et le nombre de cellules comptées.

L'utilisation de ce type d'algorithme pourrait permettre aux médecins de proposer à leurs patients des interventions personnalisées, conduisant à une vie plus longue et en meilleure santé.

U. Le minage des réseaux sociaux

Les réseaux sociaux sont riches en information type santé : il ne se passe pas une journée sans que quelqu'un ne parle de ses problèmes de santé, de sa pathologie chronique sur un forum, ou évoque sa qualité de vie sous un traitement particulier. Grâce à cela, des actions de pharmacovigilance peuvent être mises en place, en récoltant des informations utiles notamment au niveau du management du patient et plus particulièrement lors de pathologies à longue durée. Ce sont également ce qu'on appelle les données de vraie vie (*i.e* phase IV) qui émanent de ces partages.

En retour, ces données peuvent être analysées afin de déceler tout effet secondaire causé par le médicament lui-même, ou encore par des phénomènes d'interactions médicamenteuses. Le mésusage est également détectable, notamment pour les composés utilisés à des fins récréatives par exemple.

Toutefois la plupart des informations, bien qu'utiles au corps scientifique, sont bloquées par les réseaux sociaux eux-mêmes. Bien que visibles par tous, elles permettraient de relever des effets indésirables peu fréquents.

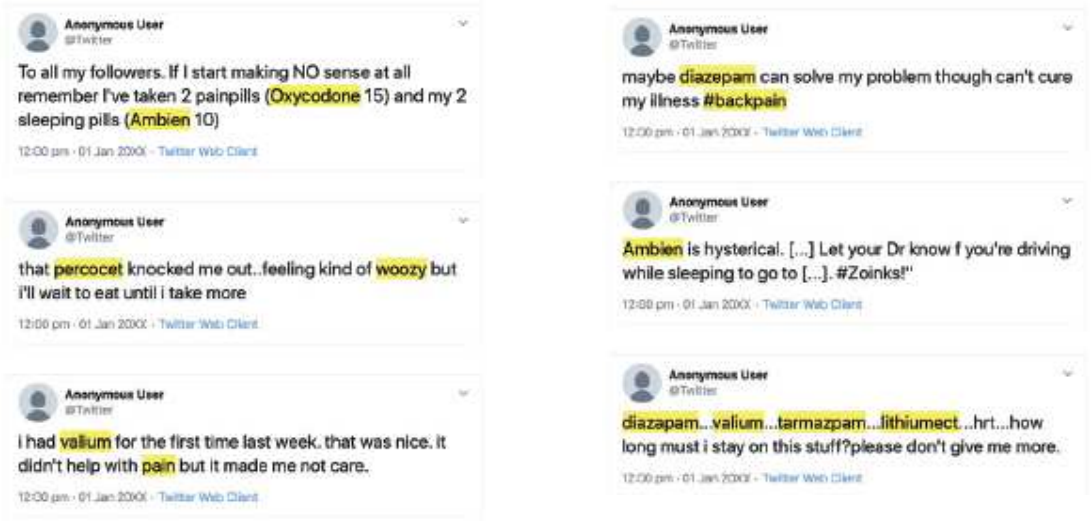

Figure 46 : Exemple de messages retrouvés sur Twitter. Tweets mentionnant des médicaments connus pour leur mésusage. Figure extraite de Correia et al. 2020 (98)

Dans un autre exemple, des mathématiciens de l'Université du Vermont ont analysé depuis 2008 le bonheur de la population grâce aux tweets et à leur outil appelé « *l'hedonometer* »(99) (littéralement mesureur de bonheur). L'année 2020 a été particulièrement difficile, et cela a été quantifié par les contenus partagés (Figure 47). Chaque jour, l'algorithme prélève 10% des tweets et analyse les mots employés. L'échelle employée pour mesurer la « positivité » ou la « négativité » des mots est arbitraire, car établie sur un panel de 50 personnes. Toutefois, cela reste un indicateur pertinent pour mesurer le bien-être général de la population.

Le score actuel est disponible sur le site https://hedonometer.org/timeseries/en_all/

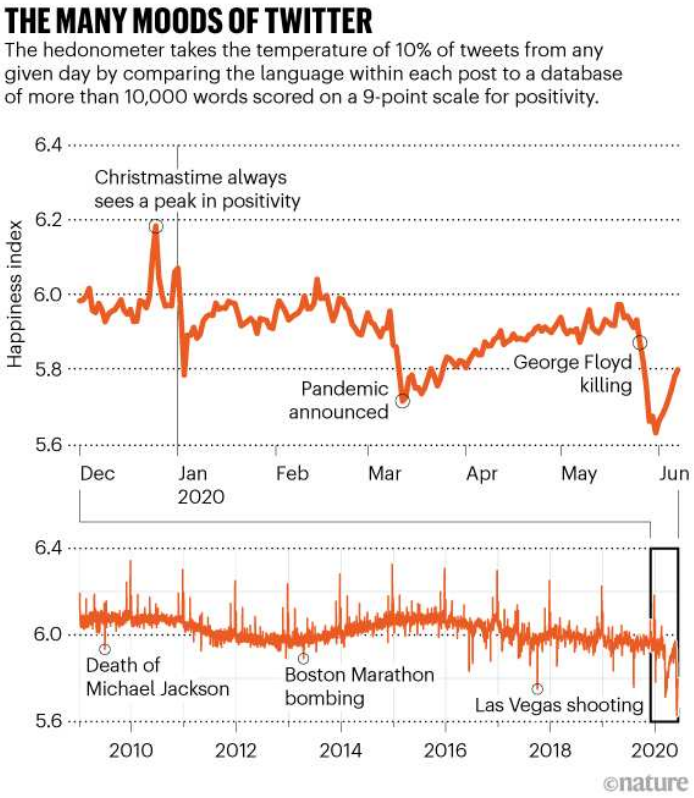

Figure 47 : Score du bonheur, 2020 est une année particulièrement difficile pour la population mondiale. Figure extraite de Viglione et al. 2020 (99)

V. Conclusion et perspectives

Face à l'immensité des données disponibles, l'Homme n'est plus capable de les interpréter à lui seul. La nécessité du ML n'est plus à prouver aujourd'hui, et son importance ne cessera de grandir dans les années à venir. Que ce soit en Chimie, en Biologie ou en Clinique, plusieurs idées d'algorithmes et de modèles ont déjà été exploitées avec succès. Il n'est pas à exclure que dans la prochaine décennie, les modèles *in vitro* voire même *in vivo* soient totalement remplacés par des modèles mathématiques *in silico*, construits à l'aide de l'intelligence artificielle. De la même façon, les essais cliniques seront potentiellement eux-aussi totalement virtuels (100).

Sachant que la performance des modèles repose sur la qualité des données collectées, il reste cependant beaucoup de travail pour la construction de bases harmonisées et exploitables. Dans l'industrie pharmaceutique, le ML est déjà d'actualité et les géants effectuent de nombreux partenariats avec des experts du domaine.

Les promesses du ML peuvent paraître encore loin de notre quotidien mais il ne faut pas oublier que l'évolution des technologies est exponentielle. CRISPR est par exemple arrivé en clinique en 2019, alors que la technologie n'a été découverte et maîtrisée que dans les années 2010 (101).

Beaucoup d'outils et de formations sont aujourd'hui accessibles de façon totalement libre, demain ne serions-nous pas tous des « data scientists » ? Tout comme il n'est pas nécessaire de savoir comment un ordinateur fonctionne pour l'utiliser, il en va de même pour les modèles de ML. De nombreux efforts sont fournis pour délivrer de puissants outils prêts à l'emploi qui pourront aider les scientifiques dans toutes les facettes de leurs projets de recherche.

Certaines voitures sont déjà capables de conduire par elles-mêmes, l'assistant virtuel Alexa de comprendre les nuances de nos phrases. Certes les modèles ne sont pas encore parfaits, mais ils apprennent toujours plus et toujours plus vite. Notre vie quotidienne est déjà impactée par le ML, ce n'est maintenant qu'une question de temps avant que notre Santé ne la soit aussi.

VI. Index

Nom	Description	Nombre de points
Open-Target	Identification de cibles Plateforme intégrative, recense les associations cible-pathologies.	6,3 M données d'association liées à 27 K cibles sur 13 K pathologies
Bio2RDF	Plateforme intégrative Première base de données qui a tenté d'intégrer les différentes données biomédicales grâce à la sémantique	11 milliards de données biomédicales reliées entre elles.
HMDD	Identification de cibles Banque de données regroupant des associations miRNA-pathologies extraites de PubMed	35000 associations entre miRNA et pathologies extraits de plus de 19 000 publications
DisGeNET	Identification de cibles Regroupe les associations entre gène/variant et pathologies	Contient plus de 1 million d'associations entre gène et pathologies.
DrugBank	Pharmacologie et repositionnement Banque de données sur les médicaments ainsi que leurs cibles.	Contient les données de plus de 14000 médicaments

ChEMBL	Chimie et Essais Biologiques Contient des données de binding et activités ADME	Contient des informations sur près de 2 millions de composés, avec des données d'activités (16 millions) sur plus de 13000 cibles.
PDB	Données structurales Contient les structures tridimensionnelles de protéines, séquences nucléiques et de structures plus complexes.	Recense plus de 17000 structures tridimensionnelles.
Tox21 dataset	Données toxicoloqiques Banque de données contenant les informations de 12 tests relatifs à la toxicité humaine	Plus de 7800 composés avec leurs activités et toxicités associées.
ChemSpider	Données chimiques Structures chimiques	Contient plus de 100 millions de différentes structures chimiques
PROMISCOUS	Repositionnement Contient les informations de médicaments et leurs effets secondaires.	Interactions et effets secondaires de 25000 médicaments
ArrayExpress	Identification de cibles et repositionnement Banque de données de génomique fonctionnelle. Mesure des ARN messagers	Données sur plus de 74000 expériences.

CMap	Repositionnement Plateforme regroupant les profils d'expression génétique de différentes lignées cellulaires après incubation avec petites molécules.	1,5 million de profils d'expression génétique de 5000 composés type petites molécules
TWOSIDES	Repositionnement Recense l'utilisation off-label de médicaments.	Contient 868221 associations entre 59220 paires de médicaments et 1301 effets secondaires.
ZINC	Criblage virtuel Banque de données qui contient la structure 3D de composés susceptibles d'être synthétisés.	ZINC contient les structures plus de 230 millions de composés commerciaux prêts à être cribler.
OpenTox	Données toxicologiques Contient différentes données toxicologiques et modèles QSAR pertinents pour la toxicité chronique, carcinogénique et génotoxique.	Non communiqué.
PubChem	Données chimiques et activités biologiques associées	109 millions de composés 247 millions d'activités
DUD-E	Données de docking, utiles lors de l'entraînement de modèles.	Plus de 22000 composés actifs et leurs activités associées sur 102 cibles.

VII. Bibliographie

1. Gartner Symposium/ITxpo in October, 2011 in Orlando, Florida.

2. Google's 2010 Atmosphere convention.

3. Was Eric Schmidt Wrong About the Historical Scale of the Internet? [Internet]. ReadWrite. 2011 [cité 18 sept 2019]. Disponible sur: https://readwrite.com/2011/02/07/are-we-really-creating-as-much/

4. Lemberger P, Batty M, Morel M. Big Data et Machine Learning - Manuel du data scientist Ed. 2 [Internet]. Dunod; 2016 [cité 18 sept 2019]. Disponible sur: http://univ.scholarvox.com/catalog/book/docid/88836214?searchterm=big%20data

5. Hilbert M. Information Quantity. In: Schintler LA, McNeely CL, éditeurs. Encyclopedia of Big Data [Internet]. Cham: Springer International Publishing; 2017 [cité 18 sept 2019]. p. 1-4. Disponible sur: http://link.springer.com/10.1007/978-3-319-32001-4_511-1

6. Lemberger P, Batty M, Morel M. Big Data et Machine Learning : Les concepts et les outils de la data science Ed. 3 [Internet]. Dunod; 2019 [cité 9 janv 2020]. Disponible sur: http://univ.scholarvox.com/catalog/book/docid/88875311?searchterm=big%20data

7. The Number of tweets per day in 2019 [Internet]. David Sayce. 2019 [cité 9 janv 2020]. Disponible sur: https://www.dsayce.com/social-media/tweets-day/

8. Business BF and AG CNN. Facebook has shut down 5.4 billion fake accounts this year [Internet]. CNN. [cité 9 janv 2020]. Disponible sur: https://www.cnn.com/2019/11/13/tech/facebook-fake-accounts/index.html

9. Donald Trump election campaign: Big Data in political campaigns [Internet]. Bernard Marr. [cité 14 avr 2020]. Disponible sur: https://www.bernardmarr.com/default.asp?contentID=711

10. pubmeddev. Home - PubMed - NCBI [Internet]. [cité 14 avr 2020]. Disponible sur: https://www.ncbi.nlm.nih.gov/pubmed/

11. Communications BA. Bayer and Exscientia collaborate to leverage the potential of artificial intelligence in cardiovascular and oncology drug discovery [Internet]. Bayer and Exscientia collaborate to leverage the potential of artificial intelligence in cardiovascular and oncology drug discovery. [cité 14 avr 2020]. Disponible sur: http://media.bayer.com/baynews/baynews.nsf/id/Bayer-Exscientia-collaborate-leverage-potential-artificial-intelligence-cardiovascular-oncology

12. A free online introduction to artificial intelligence for non-experts [Internet]. English. [cité 9 févr 2020]. Disponible sur: http://www.elementsofai.com/

13. edX | Free Online Courses by Harvard, MIT, & more [Internet]. edX. [cité 18 janv 2021]. Disponible sur: https://www.edx.org/

14. EMBL-EBI_Scientific_Report-2018.pdf [Internet]. [cité 5 mai 2020]. Disponible sur: https://www.ebi.ac.uk/about/digital-bookshelf/publications/EMBL-EBI_Scientific_Report-2018.pdf

15. Brown N, Cambruzzi J, Cox PJ, Davies M, Dunbar J, Plumbley D, et al. Chapter Five - Big Data in Drug Discovery. In: Witty DR, Cox B, éditeurs. Progress in Medicinal Chemistry [Internet]. Elsevier; 2018 [cité 24 avr 2019]. p. 277-356. Disponible sur: http://www.sciencedirect.com/science/article/pii/S0079646817300243

16. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 4 janv 2017;45(D1):D985-94.

17. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ. mai 2016;47:20-33.

18. Freedman DH. Hunting for New Drugs with AI. Nature. 18 déc 2019;576(7787):S49-53.

19. Ekins S, Puhl AC, Zorn KM, Lane TR, Russo DP, Klein JJ, et al. Exploiting machine learning for end-to-end drug discovery and development. Nat Mater. 2019;18(5):435-41.

20. ARTIFICIAL INTELLIGENCE | meaning in the Cambridge English Dictionary [Internet]. [cité 14 avr 2020]. Disponible sur: http://dictionary.cambridge.org/dictionary/english/artificial-intelligence

21. Azencott C-A. Introduction au Machine Learning [Internet]. Dunod; 2019 [cité 15 avr 2020]. Disponible sur: http://univ.scholarvox.com/catalog/book/docid/88875339?searchterm=Data%20scientist

22. Professor Arthur Samuel | Stanford Computer Science [Internet]. [cité 30 nov 2019]. Disponible sur: https://cs.stanford.edu/memoriam/professor-arthur-samuel

23. AlphaGo Zero: Starting from scratch [Internet]. Deepmind. [cité 8 oct 2019]. Disponible sur: /blog/article/alphago-zero-starting-scratch

24. Tarifs | Service d'ajout d'étiquettes aux données [Internet]. Google Cloud. [cité 11 avr 2020]. Disponible sur: https://cloud.google.com/ai-platform/data-labeling/pricing?hl=fr

25. Chaudhary R. You are a Senior Data Labelling Expert at Google. Read Why? [Internet]. Medium. 2018 [cité 5 févr 2021]. Disponible sur: https://medium.com/datadriveninvestor/you-are-a-senior-data-labelling-expert-at-google-read-why-134175d03d1a

26. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys. 1 déc 1943;5(4):115-33.

27. Lavecchia A. Deep learning in drug discovery: opportunities, challenges and future prospects. Drug Discov Today. oct 2019;24(10):2017-32.

28. dshahid380. Convolutional Neural Network [Internet]. Medium. 2019 [cité 5 juin 2021]. Disponible sur: https://towardsdatascience.com/covolutional-neural-network-cb0883dd6529

29. Understanding LSTM Networks -- colah's blog [Internet]. [cité 5 juin 2021]. Disponible sur: https://colah.github.io/posts/2015-08-Understanding-LSTMs/

30. Segler MHS, Kogej T, Tyrchan C, Waller MP. Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks. ArXiv170101329 Phys Stat [Internet]. 5 janv 2017 [cité 13 avr 2020]; Disponible sur: http://arxiv.org/abs/1701.01329

31. Vincent J. How three French students used borrowed code to put the first AI portrait in Christie's [Internet]. The Verge. 2018 [cité 3 mars 2020]. Disponible sur: https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans

32. This Person Does Not Exist [Internet]. [cité 3 mars 2020]. Disponible sur: https://thispersondoesnotexist.com/

33. MuseGAN [Internet]. MuseGAN. [cité 3 mars 2020]. Disponible sur: https://salu133445.github.io/musegan/

34. Classic Pop, in the style of Frank Sinatra - OpenAI Jukebox [Internet]. [cité 4 mai 2020]. Disponible sur: https://soundcloud.com/openai_audio/jukebox-265820820

35. Rocca J. Understanding Generative Adversarial Networks (GANs) [Internet]. Medium. 2021 [cité 5 juin 2021]. Disponible sur: https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29

36. Wakefield J. AI-created drug to be used on humans for first time. BBC News [Internet]. 30 janv 2020 [cité 7 mai 2020]; Disponible sur: https://www.bbc.com/news/technology-51315462

37. "Bigger than DNA" — how AI is transforming the pharma industry [Internet]. Sifted. [cité 25 févr 2020]. Disponible sur: https://sifted.eu/articles/ai-transforming-pharma/

38. Segler MHS, Preuss M, Waller MP. Learning to Plan Chemical Syntheses. Nature. mars 2018;555(7698):604-10.

39. Coley CW, Green WH, Jensen KF. Machine Learning in Computer-Aided Synthesis Planning. Acc Chem Res. 15 mai 2018;51(5):1281-9.

40. Coley CW, Thomas DA, Lummiss JAM, Jaworski JN, Breen CP, Schultz V, et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. Science [Internet]. 9 août 2019 [cité 3 juin 2021];365(6453). Disponible sur: http://science.sciencemag.org/content/365/6453/eaax1566

41. Caramelli D, Salley D, Henson A, Camarasa GA, Sharabi S, Keenan G, et al. Networking chemical robots for reaction multitasking. Nat Commun. 24 août 2018;9(1):1-10.

42. Withnall M, Chen H, Tetko IV. Matched Molecular Pair Analysis on Large Melting Point Datasets: A Big Data Perspective. ChemMedChem. 20 2018;13(6):599-606.

43. Jiménez J, Doerr S, Martínez-Rosell G, Rose AS, De Fabritiis G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. Bioinformatics. 1 oct 2017;33(19):3036-42.

44. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein–Ligand Scoring with Convolutional Neural Networks. J Chem Inf Model. 24 avr 2017;57(4):942-57.

45. Reymond J-L, Deursen R van, Blum LC, Ruddigkeit L. Chemical space as a source for new drugs. MedChemComm. 1 juill 2010;1(1):30-8.

46. Casciuc I, Horvath D, Gryniukova A, Tolmachova KA, Vasylchenko OV, Borysko P, et al. Pros and cons of virtual screening based on public "Big Data": In silico mining for new bromodomain inhibitors. Eur J Med Chem. 1 mars 2019;165:258-72.

47. Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular De Novo Design through Deep Reinforcement Learning. ArXiv170407555 Cs [Internet]. 29 août 2017 [cité 13 avr 2020]; Disponible sur: http://arxiv.org/abs/1704.07555

48. David L, Arús-Pous J, Karlsson J, Engkvist O, Bjerrum EJ, Kogej T, et al. Applications of Deep-Learning in Exploiting Large-Scale and Heterogeneous Compound Data in Industrial Pharmaceutical Research. Front Pharmacol [Internet]. 2019 [cité 21 janv 2020];10. Disponible sur: lauf

49. Riniker S, Wang Y, Jenkins JL, Landrum GA. Using Information from Historical High-Throughput Screens to Predict Active Compounds. J Chem Inf Model. 28 juill 2014;54(7):1880-91.

50. Laufkötter O, Sturm N, Bajorath J, Chen H, Engkvist O. Combining structural and bioactivity-based fingerprints improves prediction performance and scaffold hopping capability. J Cheminformatics. 8 août 2019;11(1):54.

51. [clue.io] [Internet]. [cité 13 avr 2020]. Disponible sur: https://clue.io/cmap

52. 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.

53. Dissecting the Hype With Cheminformatics [Internet]. Dissecting the Hype With Cheminformatics. [cité 2 mai 2020]. Disponible sur: http://practicalcheminformatics.blogspot.com/2019/09/dissecting-hype-with-cheminformatics.html

54. February DL 4, 2020. Arguing on AI Drug Discovery [Internet]. In the Pipeline. 2020 [cité 2 mai 2020]. Disponible sur: http://blogs.sciencemag.org/pipeline/archives/2020/02/04/arguing-on-ai-drug-discovery

55. Ma J, Sheridan RP, Liaw A, Dahl GE, Svetnik V. Deep Neural Nets as a Method for Quantitative Structure–Activity Relationships. J Chem Inf Model. 23 févr 2015;55(2):263-74.

56. The 3Rs | NC3Rs [Internet]. [cité 6 févr 2021]. Disponible sur: https://nc3rs.org.uk/the-3rs

57. Shin M, Jang D, Nam H, Lee KH, Lee D. Predicting the Absorption Potential of Chemical Compounds Through a Deep Learning Approach. IEEE/ACM Trans Comput Biol Bioinform. avr 2018;15(2):432-40.

58. Miao R, Xia L-Y, Chen H-H, Huang H-H, Liang Y. Improved Classification of Blood-Brain-Barrier Drugs Using Deep Learning. Sci Rep. 19 juin 2019;9(1):8802.

59. Wu Z, Lei T, Shen C, Wang Z, Cao D, Hou T. ADMET Evaluation in Drug Discovery. 19. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches. J Chem Inf Model. 25 nov 2019;59(11):4587-601.

60. Kim H, Nam H. hERG-Att: Self-attention-based deep neural network for predicting hERG blockers. Comput Biol Chem. 1 août 2020;87:107286.

61. Khatri P, Roedder S, Kimura N, De Vusser K, Morgan AA, Gong Y, et al. A common rejection module (CRM) for acute rejection across multiple organs identifies novel therapeutics for organ transplantation. J Exp Med. 21 oct 2013;210(11):2205-21.

62. Xu H, Aldrich MC, Chen Q, Liu H, Peterson NB, Dai Q, et al. Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality. J Am Med Inform Assoc. 1 janv 2015;22(1):179-91.

63. Stokes JM, Yang K, Swanson K, Jin W, Cubillos-Ruiz A, Donghia NM, et al. A Deep Learning Approach to Antibiotic Discovery. Cell. 20 févr 2020;180(4):688-702.e13.

64. Marchant J. Powerful antibiotics discovered using AI. Nature [Internet]. 20 févr 2020 [cité 7 mai 2020]; Disponible sur: http://www.nature.com/articles/d41586-020-00018-3

65. Cyclica [Internet]. Cyclica. [cité 11 avr 2020]. Disponible sur: https://cyclicarx.com

66. Sanchez CG, Molinski SV, Gongora R, Sosulski M, Fuselier T, MacKinnon SS, et al. The Antiretroviral Agent Nelfinavir Mesylate. Arthritis Rheumatol. 2018;70(1):115-26.

67. Learn more [Internet]. Oxford Nanopore Technologies. Z [cité 9 nov 2019]. Disponible sur: http://nanoporetech.com/learn-more

68. Next-Generation Sequencing for Beginners | NGS basics for researchers [Internet]. [cité 9 nov 2019]. Disponible sur: https://www.illumina.com/science/technology/next-generation-sequencing/beginners.html

69. The Cost of Sequencing a Human Genome [Internet]. [cité 5 juin 2021]. Disponible sur: https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

70. The cost of getting personal. Nat Med. déc 2019;25(12):1797-1797.

71. EMBL-EBI. What are genome wide association studies (GWAS)? | GWAS Catalog [Internet]. [cité 5 juin 2021]. Disponible sur: https://www.ebi.ac.uk/training/online/courses/gwas-catalogue-exploring-snp-trait-associations/what-is-gwas-catalog/what-are-genome-wide-association-studies-gwas/

72. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T, Engmann J, et al. The druggable genome and support for target identification and validation in drug development. Sci

Transl Med [Internet]. 29 mars 2017 [cité 10 nov 2019];9(383). Disponible sur: http://stm.sciencemag.org/content/9/383/eaag1166

73. Swerdlow DI, Preiss D, Kuchenbaecker KB, Holmes MV, Engmann JEL, Shah T, et al. HMG-coenzyme A reductase inhibition, type 2 diabetes, and bodyweight: evidence from genetic analysis and randomised trials. Lancet Lond Engl. 24 janv 2015;385(9965):351-61.

74. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. Genome Med [Internet]. 30 juill 2014 [cité 17 mai 2020];6(7). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4143549/

75. Hamzelou J. 23andMe has sold the rights to develop a drug based on its users' DNA [Internet]. New Scientist. [cité 25 févr 2020]. Disponible sur: https://www.newscientist.com/article/2229828-23andme-has-sold-the-rights-to-develop-a-drug-based-on-its-users-dna/

76. Charvis M, Alismail A. Exploring the Mechanism behind G551D and the Effects of Ivacaftor. 2016;7.

77. Alzheimer's disease - Causes [Internet]. nhs.uk. 2018 [cité 1 févr 2021]. Disponible sur: https://www.nhs.uk/conditions/alzheimers-disease/causes/

78. Zwanzig R, Szabo A, Bagchi B. Levinthal's paradox. Proc Natl Acad Sci U S A. 1 janv 1992;89(1):20-2.

79. Protein [Internet]. Genome.gov. [cité 5 juin 2021]. Disponible sur: https://www.genome.gov/genetics-glossary/Protein

80. The cost and value of three-dimensional protein structure – Drug Discovery World (DDW) [Internet]. [cité 1 févr 2021]. Disponible sur: https://www.ddw-online.com/the-cost-and-value-of-three-dimensional-protein-structure-1114-200308/

81. Jo T, Hou J, Eickholt J, Cheng J. Improving Protein Fold Recognition by Deep Learning Networks. Sci Rep. 4 déc 2015;5(1):1-11.

82. Home - Prediction Center [Internet]. [cité 5 juin 2021]. Disponible sur: https://predictioncenter.org/

83. AlphaFold: Using AI for scientific discovery [Internet]. Deepmind. [cité 4 mai 2020]. Disponible sur: /blog/article/AlphaFold-Using-AI-for-scientific-discovery

84. Callaway E. 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. Nature. 30 nov 2020;588(7837):203-4.

85. The Cell Image Library [Internet]. [cité 4 mai 2020]. Disponible sur: http://www.cellimagelibrary.org/home

86. Kraus OZ, Ba JL, Frey BJ. Classifying and segmenting microscopy images with deep multiple instance learning. Bioinformatics. 15 juin 2016;32(12):i52-9.

87. Nagpal K, Foote D, Liu Y, Po-Hsuan, Chen, Wulczyn E, et al. Development and Validation of a Deep Learning Algorithm for Improving Gleason Scoring of Prostate Cancer. ArXiv181106497 Cs [Internet]. 15 nov 2018 [cité 25 mai 2020]; Disponible sur: http://arxiv.org/abs/1811.06497

88. Cireşan DC, Giusti A, Gambardella LM, Schmidhuber J. Deep neural networks segment neuronal membranes in electron microscopy images. In 2012. p. 2843-51.

89. Ferrari A, Lombardi S, Signoroni A. Bacterial colony counting with Convolutional Neural Networks in Digital Microbiology Imaging. Pattern Recognit. 1 janv 2017;61:629-40.

90. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. Nat Med. janv 2019;25(1):70-4.

91. Biomarker [Internet]. European Medicines Agency. [cité 2 févr 2021]. Disponible sur: https://www.ema.europa.eu/en/glossary/biomarker

92. Vincent MD, Kuruvilla MS, Leighl NB, Kamel–Reid S. Biomarkers that currently affect clinical practice: EGFR, ALK, MET, KRAS. Curr Oncol. juin 2012;19(Suppl 1):S33-44.

93. Li B, Shin H, Gulbekyan G, Pustovalova O, Nikolsky Y, Hope A, et al. Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib. PLoS ONE [Internet]. 24 juin 2015 [cité 2 févr 2021];10(6). Disponible sur: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4480971/

94. Ramirez-Celis A, Becker M, Nuño M, Schauer J, Aghaeepour N, Van de Water J. Risk assessment analysis for maternal autoantibody-related autism (MAR-ASD): a subtype of autism. Mol Psychiatry. 22 janv 2021;1-10.

95. Colangelo M. AI Will Drive The Multi-Trillion Dollar Longevity Economy [Internet]. Forbes. [cité 3 févr 2021]. Disponible sur: https://www.forbes.com/sites/cognitiveworld/2019/12/07/ai-will-drive-the-multi-trillion-dollar-longevity-economy/

96. Brutsaert E. Metformin in Longevity Study (MILES). [Internet]. clinicaltrials.gov; 2018 mai [cité 2 févr 2021]. Report No.: NCT02432287. Disponible sur: https://clinicaltrials.gov/ct2/show/NCT02432287

97. Zhavoronkov A, Mamoshina P. Deep Aging Clocks: The Emergence of AI-Based Biomarkers of Aging and Longevity. Trends Pharmacol Sci. 1 août 2019;40(8):546-9.

98. Correia RB, Wood IB, Bollen J, Rocha LM. Mining social media data for biomedical signals and health-related behavior. ArXiv200110285 Cs Q-Bio [Internet]. 28 janv 2020 [cité 25 avr 2020]; Disponible sur: http://arxiv.org/abs/2001.10285

99. Viglione G. Has Twitter just had its saddest fortnight ever? Nature [Internet]. 15 juin 2020 [cité 28 juin 2020]; Disponible sur: http://www.nature.com/articles/d41586-020-01818-3

100. http://www.atcg-partners.com, 2020 RT for atcg-partners. Les essais cliniques in silico : le numérique et la modélisation au service du médicament [Internet]. Eurobiomed - Biocluster Méditerranée. [cité 25 mai 2020]. Disponible sur: https://www.eurobiomed.org/actualite/detail/News/les-essais-cliniques-in-silico-le-numerique-et-la-modelisation-au-service-du-medicament/

101. Single Ascending Dose Study in Participants With LCA10 - Full Text View - ClinicalTrials.gov [Internet]. [cité 25 mai 2020]. Disponible sur: https://clinicaltrials.gov/ct2/show/NCT03872479

Université de Lille

FACULTE DE PHARMACIE DE LILLE

DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE

Année Universitaire 2020/2021

Nom : VICENTINI

Prénom : QUENTIN

Titre de la thèse : Big Data et Intelligence Artificielle en Drug Discovery

Mots-clés : Big Data - Machine Learning - AI – Drug Discovery

---

Résumé : Tous les jours, de nouveaux articles scientifiques sont publiés, de nouvelles informations biomédicales sont partagées, de nouveaux brevets sont déposés. Face à l'immensité de ces données, notre esprit n'est plus capable à lui seul de les interpréter. Afin de s'aider, les chercheurs ont développé des algorithmes de Machine Learning. Depuis les étapes de synthèses jusqu'à la découverte de biomarqueurs, le Machine Learning est déployé à toute les échelles et soutient les projets de Drug Discovery.

<u>Membres du jury</u> :

Président et Directeur de Thèse : Monsieur Philippe Chavatte, Professeur des Universités et Assesseur en charge des Relations Internationales à la Faculté de Pharmacie de Lille

Assesseur(s) : Monsieur Pascal Dao Phan, Professeur Associé à la Faculté de Pharmacie de Lille et Directeur des Opérations Cliniques à Bayer Loos

Monsieur Karim Ali Belarbi, Maître de conférence à Faculté de Pharmacie de Lille.

Membre extérieur : Monsieur Simon Cazin, Docteur en Pharmacie, à la Pharmacie Simon Cazin de Wailly-Beaucamp