Identification des interactions médicamenteuses délétères et des variations génomiques structurales entraînant des effets secondaires chez l'homme à l'aide d'outils d'apprentissage automatique

Bryan Dafniet

Université Paris Cité

École doctorale Pierre Louis de Santé Publique (ED393) :
Épidémiologie et Sciences de l'Information Biomédicale

Laboratoire de Biologie Fonctionnelle et Adaptative, INSERM UMR 1133,
équipe : *Modélisation Computationnelle des Interactions Protéine-Ligand*

Identification des interactions médicamenteuses délétères et des variations génomiques structurales entraînant des effets secondaires chez l'homme à l'aide d'outils d'apprentissage automatique

Par Bryan DAFNIET

Thèse de Doctorat de Bioinformatique

Dirigée par Pr. Olivier Taboureau

Présentée et soutenue publiquement le 15 Décembre 2022 devant un jury composé de :

Pr. Laurence Payrastre, DR	Université de Toulouse Paul Sabatier	Rapporteure
Pr. Ronan Bureau, PU	Université de Caen	Rapporteur
Pr. Armelle Baeza, PU	Université Paris Cité	Examinatrice
Pr. Pascal Bonnet, PU	Université d'Orléans	Examinateur
Dr. Jean-Christophe Gelly, MCU	Université Paris Cité	Examinateur
Pr. Olivier Taboureau, PU	Université Paris Cité	Directeur de thèse

Résumé

Identification des interactions médicamenteuses délétères et des variations génomiques structurales entraînant des effets secondaires chez l'homme à l'aide d'outils d'apprentissage automatique

Le développement d'un médicament est un processus long et couteux pouvant durer plus de 10 ans et ayant un cout dépassant le milliard d'euros. Si les cibles thérapeutiques (cibles protéiques primaires) que doivent cibler ce médicament sont généralement connues, il a été établi qu'un médicament toucherait de nombreuses protéines supplémentaires (cibles protéiques secondaires) responsables d'effets secondaires. Cet effet peut être bénéfique (*repurposing)* ou délétère (effet indésirable). Ces cibles secondaires sont aujourd'hui mal ou non identifiées et seraient responsables d'environ 200 000 morts par an, engendrant un coût dépassant le milliard d'euros en Europe et aux Etats-Unis. De plus, il a été démontré que les variations génétiques d'un individu seraient susceptibles d'impacter l'efficacité d'un médicament et contribueraient à l'apparition d'effets indésirables associés à un médicament.

Dans ces conditions, l'objectif de ce travail de thèse a été de développer des méthodes computationnelles permettant 1) d'intégrer un ensemble de données pharmacologiques provenant de différentes sources dans une seule base de données et de sélectionner 5100 composés avec leur activité biologique sur l'ensemble du protéome humain et des informations provenant de criblage phénotypique. Les informations composés-cibles seront ensuite reliées, si possible, à des informations provenant de criblages phénotypiques, pour effectuer des calculs d'enrichissement sur le rôle des protéines dans diverses fonctions biologiques et maladies. Ensuite de 2) déterminer les protéines fréquemment associées à un effet indésirable en se basant sur les interactions connues médicament-effet indésirable et médicament-protéine. A l'aide de l'implémentation d'une fonction de score, la contribution d'une protéine dans l'apparition d'un effet indésirable a été caractérisée et quantifiée. L'impact des mutations de type *Copy Number Variations* (CNV) sur l'apparition de tels effets a aussi été étudié. Enfin de 3) développer des modèles prédictifs permettant d'anticiper les effets indésirables d'un médicament potentiellement dus à des variations génétiques. En nous référant à la nomenclature SOC (System Organ Class) et aux variations génétiques de type *Single Nucleotide Polymorphisms* (SNPs) nous pourrons ensuite analyser les cibles thérapeutiques qui seraient les plus à même d'être la cause de ces effets.

Mots-clés : Effets indésirables ; Médicaments ; Science des réseaux ; Bases de données ; Réseaux de neurones profonds ; Mutations ; Phénotypes ; Intégration de données

Abstract

Identification of drug-target interactions and genetic variations leading to adverse drug reactions by using computational approaches.

The development of a drug is an expensive and time-consuming process that could take more than 10 years and cost around a billion euros. Although the drug's target is usually known (primary target), it is also established that a drug interacts with multiple additional proteins (secondary targets) that can be responsible for side effects. These effects can be beneficial (repurposing) or noxious (adverse drug reaction). These secondary targets are not always identified and would cause 200 000 deaths/year, leading to a cost reaching a billion euros in Europe and USA. Moreover, it has been shown that genetic variations might also impact the efficacy of a drug and could contribute to adverse drug reactions.

In these circumstances, the objective of the thesis was to develop computational methods allowing us to 1) integrate pharmacology data from multiple sources into one database and then select 5100 compounds with their biological activity on all of the human proteome with information from phenotypic screening. Compound-Target information was then linked to phenotypic screening information, and enrichment calculations were performed to assess the role of proteins on biological functions and diseases. Then 2) proteins frequently associated with adverse drug reactions have been determined based on known drug-target and adverse drug reaction information. By implementing a scoring function, the possible contribution of a protein in the apparition of adverse drug reactions has been assessed. In addition, the impact of *Copy Number Variations* (CNVs) on the apparition of such reactions has been analyzed. Finally in 3), predictive models have been developed allowing us to suggest if a drug would induce adverse drug reactions, based on the SOC nomenclature (System Organ Class) and *Single Nucleotide Polymorphisms* (SNPs), to overall analyze adverse drug reactions due to genetic variations on these drug targets.

Keywords: Adverse drug reactions ; Drugs ; Network sciences ; Databases creation ; Deep neural networks ; Mutations ; Phenotypes ; Data integration

Remerciements

Tout d'abord je tiens à remercier mon directeur de thèse, Olivier, sans qui ce manuscrit n'aurait jamais vu le jour. Merci de m'avoir fait confiance et de m'avoir guidé tout au long de ces trois années, et de votre réactivité lors de l'écriture de cette thèse malgré des mails envoyés durant les week-end ou lors de vos vacances. Si j'ai pu gravir cette montagne c'est avant tout grâce à vous et à votre bienveillance (et à quelques friandises sucrées), merci.

Je tiens à remercier Jean-Marie Dupret pour m'avoir accueilli au sein de BFA, et Pierre Tuffery pour son accueil au sein de l'équipe.

Merci aux membres du jury Dr. Laurence Payrastre et Pr. Ronan Bureau d'avoir accepté de rapporter ma thèse. Merci au Dr. Jean-Christophe Gelly, Pr. Armelle Baeza et Pr. Pascal Bonnet d'avoir accepté de faire partie de mon jury de thèse et de prêter attention à mes travaux.

Merci à Anne Badel pour une gestion des enseignements me permettant de me concentrer sur la pédagogie sans trop de stress et dans de bonnes conditions.

Je remercie également les personnes avec qui j'ai travaillé chez Servier, Arnaud Gohier, Anaëlle Clary, Thierry Dorval et David Brown pour la reprise de projet et une formation sur de nouveaux outils qui m'ont été bien utiles par la suite.

Merci à Pierre, qui malgré son départ avant moi m'aura bien épaulé pendant mes deux premières années et été un camarade précieux dans ce début de thèse, notamment au babyfoot.

D'ailleurs, merci au babyfoot pour sa présence qui malgré le fait de ralentir la recherche française, lui permet au moins de ne pas devenir folle.

Camarades de babyfoot et de bureau je vous salue. Merci Guillaume pour l'impulsion nécessaire à ma reprise du sport. Merci Natacha, pour les articles publiés ensemble (on a publié un article sur la DrugBank ?!), tes propositions d'aide à chaque micro-blocage, et tes résolutions bien plus efficaces que si je l'avais codé moi-même. Merci à Vanille qui a amélioré mon quotidien dès son arrivée en absorbant toutes les ondes négatives pour elle. Je n'imagine pas le prix que tu dois payer au quotidien pour que je puisse un peu goûter au succès. Je ne sais pas si ce mini-paragraphe te satisfait mais dans tous les cas ne m'en veux pas, pense à ton karma.

Un énorme merci à mes amis, évidemment, de France comme de Belgique que je n'ai pas vu depuis plusieurs mois au moment où j'écris ces mots. Je ne citerai pas de noms, par peur d'en oublier un, mais vous vous reconnaitrez. Ces trois années de thèse auraient été bien fades et angoissantes sans vous, ces soirées, ces bières.
Petite entorse à ce que je viens d'écrire pour Loïck, et nos habituels traquenards au diable qui m'ont permis de souffler de nombreuses fois.

Merci papa pour la relecture et la remise en forme de ce manuscrit, en des temps records !

Impossible de ne pas remercier ma famille pour leur soutien, Xavier pour m'avoir permis de passer une thèse dans plus de 6m^2 à Paris, Natalie pour les peu nombreuses mais Ô combien appréciés déjeuners (et les foies gras). Axel et Lison, pour ces réunions familiales et ces moments nécessaires, que ce soit autour d'un rhum arrangé ou d' « une » bière. Lola et Marius,

pour la légèreté apportée dans mon monde compliqué, et pour me pousser à devenir la meilleure version de moi-même pour être un modèle, je l'espère, inspirant.

Merci infiniment Pragya pour ta patience ces derniers mois où je me suis transformé en ermite, pour ta présence et ton soutien sans faille, pour tous ces plats cuisinés qui ont rendu cette thèse bien moins difficile.

Enfin, un **gigantesque** merci à mes parents et leur confiance aveugle en ma réussite, leurs sacrifices. Si j'ai pu arriver au sommet à la fin, c'est parce qu'en étant assis sur les épaules de si grandes personnes il ne me restait plus beaucoup à gravir pour y parvenir. Cette thèse c'est aussi et surtout la vôtre.

Abréviations

2D/3D	Deux Dimensions / Trois Dimensions
ADME(T)	Absorption-Distribution-Métabolisme-Excretion(-Toxicité)
ADR	Adverse Drug Reaction
AMM	Autorisation de mise sur le marché
ANSM	Agence nationale de sécurité du médicament et des produits de santé
BBBC	Broad Bioimage Benchmark Collection
CNV	Copy number variation
EMA	European Medicines Agency / Agence Européenne du médicament
FP	Fingerprint
KB	Knowledge-Based
MD	Molecular Dynamics
PDB	Protein Data Bank
QSAR	Quantitative Structure-Activity Relationship
SNP	Single Nucleotide Polymorphism
SOC	System Organ Class

Table des matières

Liste des figures

Liste des tableaux

Introduction

Les effets secondaires sont apparus en même temps que les premiers médicaments et s'ils sont aujourd'hui considérés lors des essais cliniques, leur présence reste inévitable.
Lors du processus de développement, les cibles thérapeutiques sont identifiées et les molécules sont synthétisées dans le but de maximiser les interactions avec la protéine cible (cible primaire). Cependant, un médicament est rarement actif sur une seule protéine. En effet il a été montré qu'un médicament interagit avec de multiples protéines supplémentaires [Garcia-Serna R. *et al*. 2010]. Ils peuvent donc impacter la fonction des protéines de façon non intentionnelle (cibles secondaires). De manière générale, ces problèmes sont à l'origine du développement de la polypharmacologie : l'étude de la capacité d'un médicament à affecter plus d'une protéine [Proschak E. *et al* 2018].

Lorsqu'un médicament agit sur plusieurs protéines, deux effets peuvent alors se produire. Pour le premier l'effet secondaire est bénéfique et l'on peut réutiliser cette molécule dans le traitement d'autres pathologies, elle est « réaffectée » (*repurposing*). Pour le second l'effet sera dit « indésirable » (ADR). Il sera délétère et sa gravité entrainera ou non l'arrêt de son développement. Certaines protéines à éviter sont connues et testées *e.g*. le canal hERG au niveau du cœur pouvant entrainer des effets indésirables importants comme l'arrêt cardiaque ou des torsades de pointe si un médicament l'inhibe [Hancox J.C. *et al* 2008]. Les effets indésirables seraient la cause d'environ 200 000 morts/an et auraient un coût dépassant le milliard d'euros en Europe et aux Etats-Unis [Bouvy J.C *et al.* 2015 ; Giardina C. 2018].

Un autre point pouvant aggraver ou être la cause d'un effet indésirable concerne la variabilité génétique humaine. La mutation d'un seul nucléotide (*single nucleotide polymorphism* ou *SNP*) peut affecter l'affinité de liaison d'un médicament pour sa cible, pouvant le rendre inefficace ou entrainer de graves effets indésirables. C'est également le cas pour des mutations entrainant des variations dans le nombre de copies d'un gène (*Copy number variation* ou *CNV*) entrainant des effets directs sur les dosages d'un médicament. Par exemple, le cytochrome CYP2D6 (une enzyme du métabolisme) possède de nombreuses mutations de type CNV, créant des duplications ou délétions du gène codant la protéine. Cela entraine un surdosage de certains médicaments dont de nombreux antidépresseurs qui ont besoin de cette enzyme pour être métabolisés et devenir actifs chez certains patients [Jarvis J.P. *et al.* 2019].

La problématique est multiple. Si des études ciblées mettant en évidence certaines protéines spécifiques ou mutations, responsables d'effets indésirables existent, des études générales sur des méthodes de prédiction d'effets indésirables utilisant des approches statistiques sont plus rares. Des bases de données regroupant des effets indésirables pour des molécules thérapeutiques existent, de même que des liaisons médicament-cible, mais le recoupement de ces informations n'est pas systématique.

Ce manuscrit regroupera les différentes publications effectuées lors de ma thèse pour répondre à cette problématique et se découpera en quatre chapitres.

Le premier chapitre sera consacré à l'état de l'art, de la conception d'un médicament à la recherche d'effets indésirables utilisant des approches *in silico* et les données existantes pouvant être utilisées dans ce but.
Le deuxième chapitre sera consacré à la création d'une base de données intégrant des données composés-protéines-maladies. Cet outil permet d'étudier la polypharmacologie d'un composé chimique de façon rapide, interactive, d'en suggérer autant les effets thérapeutiques potentiels que les risques d'effets indésirables. De plus, à partir de cette base de données, nous avons développé une bibliothèque de 5100 composés chimiques en prenant en compte les *scaffolds*, permettant de représenter l'ensemble du génome « *druggable* » lors de campagne de criblage phénotypique.
Le troisième chapitre présentera une méthode d'identification et de prédiction de protéines causant des effets indésirables en se basant uniquement sur les interactions médicament-cible et médicament-effet indésirable connues. Le rassemblement des données disponibles et leur analyse vont permettre d'étudier la répartition des effets indésirables parmi les cibles et grâce au développement d'une fonction de score de prioriser les cibles les plus à même de causer ces effets. De plus une ouverture sur l'analyse de mutations de type CNVs et leur implication sur certains effets secondaires sera également présentée.
Le dernier chapitre sera concentré uniquement à la mise en place de méthodes de prédiction d'un effet indésirable en se basant sur la présence de mutations de type SNPs grâce à des modèles d'apprentissage profond, ou *deep learning*. Identifier les mutations pouvant être responsables d'effets indésirables lors de la prise d'un médicament conclura donc cette thèse.

Chapitre 1 : État de l'art

1.1 La conception d'un médicament

La conception d'un médicament est un processus extrêmement long et couteux pour une entreprise pharmaceutique. Passant par de nombreuses phases et s'étalant sur 12 ans en moyenne, ce processus possède un coût dépassant le milliard d'euros (figure 1) [Di Masi J.A., *et al* 2016]. Elle se caractérise par 4 axes : 1) La recherche et découverte, 2) Le développement non clinique, 3) Le développement clinique, 4) La pharmacovigilance.

Contenu retiré pour des raisons de droits d'auteurs

1.1.1 Recherche et découverte

Cette toute première phase de développement est là pour répondre à un besoin, une nouvelle pathologie émergente ou une maladie dont les traitements existants ne sont pas assez efficaces. Il convient donc d'identifier les cibles protéiques associées à une pathologie, les caractéristiques de ces dernières et ses mécanismes d'action [Sinha S., *et al.* 2018]. Cette étape va paver la voie du développement d'un médicament et orienter les recherches. On peut notamment penser à l'identification de la protéine Spike permettant au SARS-CoV-2 de pénétrer nos cellules devenant la cible prioritaire lors du développement de vaccins [Tortorici M.A *et al.* 2019]. C'est également lors de cette étape que les traitements existants et les composés disponibles sont étudiés. Plusieurs types de composés peuvent être utilisés, chacun apportant ses avantages et inconvénients. L'origine d'un composé peut être naturelle, ce qui a donné de bons résultats pour

les anti-cancéreux ou les maladies infectieuses. Ce sont des composés qui couvrent un espace chimique plus large que les traditionnelles librairies de petites molécules synthétiques. Néanmoins, trouver un composé d'intérêt peut s'avérer compliqué et des précautions doivent être prises lors de la génération de composés analogues pour ne pas redécouvrir des composés déjà utilisés [Atanasov A.G., *et al.* 2021]. Un autre moyen favorisé par les industries pharmaceutiques se trouve dans la réutilisation de médicaments déjà commercialisés. L'avantage provient de la connaissance déjà acquise pour ce composé, notamment sur sa toxicité, les doses et effets indésirables, et sa disponibilité immédiate. En revanche, il est impossible d'optimiser le composé sans devoir repasser les étapes de développement, ce qui limite la facilité d'utilisation de ces composés [Gil C. *et al.* 2021].

1.1.2 Développement non clinique

Cette étape intervient avant la mise en place de tests sur les humains. Elle est constituée de tests *in silico*, *in vitro et in vivo* permettant de sélectionner des candidats médicaments prometteurs. En général des méthodes de criblages virtuelles (*in silico*) vont être utilisées dans un premier temps permettant à partir de bases de données de millions de composés de tester une dizaine de milliers, voire des centaines de milliers de composés (avec le criblage à haut débit (*High-Throughput Screening* en anglais, HTS) [Aldewachi H., *et al.* 2021]) qui peuvent être testés *in vitro* pour au final n'en garder que quelques-uns. Ils vont être ensuite optimisés pour leur donner différentes propriétés supplémentaires (passage de la barrière intestinale, barrière hémato-encéphalique…) nécessaires à leur efficacité [Subbaiah M.A.M., *et al.* 2021]. Enfin des tests sur animaux peuvent être mis en place pour vérifier sur un organisme complet, idéalement le plus proche possible de l'homme, sa toxicité et ses propriétés d'absorption, distribution, métabolisme et excrétion (ADME) [Honek J. 2017].

1.1.3 Développement clinique

Cette partie de développement clinique va introduire cette fois-ci des patients sur lesquels la ou les molécules médicamenteuses vont être testées. Cette partie est divisée en plusieurs phases :

- Phase I : 20 à 100 volontaires, en général en bonne santé, sont sélectionnés. Il peut arriver que ces volontaires soient déjà atteints de la pathologie lorsque l'on sait que le traitement rend malade des individus en bonne santé (traitements contre le cancer, VIH…). Cette étape a pour but premier de tester la toxicité et la dose maximale tolérée. Un gradient de concentration de dose est donc administré aux participants. C'est également ici que les interactions nourriture-médicament sont étudiées [Fisher J.A. 2014].
- Phase II : Lors de cette phase un panel plus important de 50 à 300 personnes atteints de la pathologie est étudié. Cette phase a deux objectifs. Le premier est de démontrer l'efficacité clinique et l'activité biologique. Le deuxième est la dose à laquelle le médicament a une activité biologique avec le moins d'effets indésirables. Si cette phase peut se dérouler en double aveugle ce n'est pas systématiquement le cas, et

seulement entre 18 et 30% des candidats médicaments passent cette phase [Simon R., *et al.* 1985].

- Phase III : Cette phase est la plus large puisqu'elle peut inclure jusqu'à plusieurs milliers de patients *e.g.* 43 548 personnes pour le vaccin Pfizer contre le SARS-Cov-2 [Polak F.P., *et al.* 2020]. Elle a pour objectif de valider sur un niveau représentatif de la population l'efficacité du traitement ainsi que sa performance comparée aux traitements déjà existant. Elle sert également à révéler les effets indésirables les plus rares notamment à cause de composantes génétiques.

1.1.4 Autorisation de mise sur le marché (AMM)

En Europe plusieurs procédures sont possibles pour obtenir une AMM. Toutes ces procédures commencent par l'envoi d'un dossier contenant toutes les informations disponibles sur le nouveau médicament ; de la composition chimique et sa fabrication à sa posologie et sa notice, en passant par toutes les méthodes et résultats des différentes phases. Par exemple, une procédure est dite « centralisée » lorsque le dossier est envoyé à la commission européenne et accepté après consultation de l'Agence Européenne du médicament (*EMA*) rendant le traitement disponible pour tous les états membres. Une procédure « nationale » est possible lorsque le dossier est envoyé à l'autorité compétente du pays pour validation; comme l'Agence nationale de sécurité du médicament et des produits de santé (ANSM) pour la France [Guerriaud M. 2016].

Plusieurs dérogations existent pour ces procédures dans le cas où une mise sur le marché accélérée est nécessaire pour des raisons de santé publique. On peut notamment citer une mise sur le marché « conditionnelle » d'un an au lieu de cinq, ou « accélérée » avec une procédure d'évaluation plus courte. Lors de la mise sur le marché d'un médicament un ensemble de contre-indications et d'effets indésirables observés durant les essais cliniques sont aussi indiqués.

1.1.5 Gestion post-AMM et pharmacovigilance

Cette phase regroupe la surveillance du traitement sur le long terme pour la découverte d'éventuels effets indésirables apparaissant après la mise sur le marché. De nouvelles études peuvent avoir lieu également pour toucher un public différent exclu initialement des phases cliniques (femmes enceintes, enfants…) ou pour étudier les interactions avec d'autres médicaments en cas de prises conjointes [Suvarna V. 2010].

1.2 Les effets indésirables et polypharmacologie

Un effet indésirable est défini comme tel : « *réaction nocive et non voulue à un médicament, se produisant aux posologies normalement utilisées chez l'homme pour la prophylaxie, le diagnostic ou le traitement d'une maladie ou pour la restauration, la correction ou la modification d'une fonction physiologique* » [World Health organization technical report series No. 498 ; Roulet L., *et al.* 2015].
Il n'est donc pas à confondre avec un effet secondaire qui encadre en plus les effets bénéfiques pouvant survenir à la suite de la prise du médicament. On peut citer le sildénafil, commercialisé sous le nom Viagra, qui était initialement en développement pour traiter des douleurs cardiaques qui a fini par être commercialisé pour traiter les problèmes d'érection [Papapetropoulos A., *et al* 2018].

1.2.1 Quelques classifications d'effets indésirables

Plusieurs méthodes de classifications existent, chacune intégrant un but différent. On peut ainsi retenir les classifications ci-après.

1.2.1.1 La classification selon Kaufman G

On peut citer la classification en différent types allant de A à E proposé par Kaufman G. [Kaufman G. 2016] :

- Type A : Un effet indésirable de type A est en général lié à la dose et peut inclure des exagérations de son effet prévu (un saignement excessif pour un anticoagulant) ou des effets en dehors du cadre prévu mais référencés dans les effets indésirables connus. Ils sont la plupart du temps atténués et facile à détecter mais peuvent être problématiques lorsque la marge entre effet thérapeutique et effet toxique est faible.
- Type B : Cela correspond cette fois ci à des effets non référencés sortant du cadre d'action du médicament. Par exemple un choc anaphylactique due à la pénicilline, dépendant d'un facteur génétique. Ces effets sont moins courants mais plus dangereux que les types A.
- Type C : Ce sont des réactions qui persistent dans le temps.
- Type D : Ces réactions sont retardées après la prise du traitement. Elles apparaissent plusieurs jours ou semaines après la première prise et peuvent être relativement difficiles à détecter.
- Type E : Ce sont des effets apparaissant après la prise du traitement directement liés à son arrêt.

Ce classement prend en compte le type d'effet en se basant sur la prise du traitement et la temporalité des effets observés.

1.2.1.2 La classification selon le Dictionnaire medDRA

Un autre type de classement qui sera utilisé dans nos études par la suite vient du Dictionnaire Médical pour les Activités Réglementées (*medDRA*) et classe cette fois-ci les effets indésirables en fonction du type d'effet qu'il induit [https ://www.meddra.org/]. C'est le classement en *system organ class* (ou SOC). C'est un classement en 27 groupes basé sur l'étiologie ou le lieu d'apparition de l'effet (tableau 1). Par exemple l'effet indésirable « arrêt cardiaque » se situera dans le SOC « *Cardiac disorders* » alors que « douleur vésicale » sera dans « *Renal and urinary disorders* ».

SOC	**SOC_abbrev.**
Blood and lymphatic system disorders	Blood
Cardiac disorders	Card
Congenital, familial and genetic disorders	Cong
Ear and labyrinth disorders	Ear
Endocrine disorders	Endo
Eye disorders	Eye
Gastrointestinal disorders	Gastr
General disorders and administration site conditions	Genrl
Hepatobiliary disorders	Hepat
Immune system disorders	Immun
Infections and infestations	Infec
Injury, poisoning and procedural complications	Inj&P
Investigations	Inv
Metabolism and nutrition disorders	Metab
Musculoskeletal and connective tissue disorders	Musc
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	Neopl
Nervous system disorders	Nerv
Pregnancy, puerperium and perinatal conditions	Preg
Psychiatric disorders	Psych
Renal and urinary disorders	Renal
Reproductive system and breast disorders	Repro
Respiratory, thoracic and mediastinal disorders	Resp
Skin and subcutaneous tissue disorders	Skin
Social circumstances	SocCi
Surgical and medical procedures	Surg
Vascular disorders	Vasc
Product issues	Prod

Tableau 1 : Liste des SOCs et des abréviations associées.

1.2.1.3 La classification selon leur gravité et leur fréquence

Enfin les effets indésirables peuvent être classés par leur gravité en se basant sur une version modifiée de l'étude de Hartwig Siegel [Hartwig SC., *et al.* 1992 ; Geer M.I., *et al.* 2016] ou leur fréquence (tableau 2) [CIOMS 1999].

Sévérité	Effets
Bénin	Réactions mineures, pas de traitement supplémentaire ou d'hospitalisation prolongée.
Modéré	Modification du traitement nécessaire, peut prolonger l'hospitalisation et nécessiter un traitement spécifique.
Sévère	Potentiellement mortel. Nécessite l'arrêt du traitement et la prise en charge de l'effet indésirable.
Mortel	Effet indésirable lié directement ou indirectement à la mort d'un patient.

Fréquence	Probabilité
Très fréquent	> 10%
Fréquent	Entre 1 et 10%
Peu commun	Entre 0,1 et 1%
Rare	Entre 0,01 et 0,1%
Très rare	< 0,01%

Tableau 2: Tableau des différentes sévérités d'effets indésirables et de leurs conséquences (à gauche) et du classement des fréquences d'effets indésirables en fonction de leur probabilité d'apparition (à droite).

Les effets indésirables dus aux médicaments sont donc définis et classifiés, mais les mécanismes entrainant leurs apparitions sont très peu connus. Si la cible principale d'un médicament est connue, on sait qu'un médicament toucherait en moyenne 5 à 7 cibles supplémentaires en moyenne qui seraient mal ou non identifiées causant la majorité des effets indésirables [Vogt I., *et al.* 2010]. De nombreuses approches *in silico* ont donc été développées pour traiter ce problème que ce soit par une identification de cibles, de ligands ou de prédictions d'apparition d'effets indésirables directement.

1.2.2 La polypharmacologie

Cette idée qu'un médicament touche plus de cibles que la cible principale a conditionné la recherche de nouveaux médicaments (*drug discovery* en anglais*)* qui ne s'oriente plus sur un médicament/une cible mais sur un médicament/plusieurs cibles. Cette idée s'appuie sur la découverte des cibles secondaires pouvant mener au développement de médicaments agissant sur de multiples symptômes. L'aspirine est un bon exemple de médicament à effets multiples ayant notamment une activité antipyrétique et analgésique [Reddy A., *et al.* 2013]. L'identification des cibles peut donc amener à une meilleure compréhension des effets indésirables et de la toxicité des composés, permettant ainsi de développer des options de repositionnement (*repurposing)* de médicaments existants. De plus, une telle approche est essentielle dans le traitement de maladies multifactorielles comme les cancers, les maladies du métabolisme ou les maladies mentales. Pour preuve, aucun antipsychotique spécifique à une cible n'a atteint le marché en 60 ans [Peters, J.U. 2013]. Plus récemment la combinaison de plusieurs médicaments (remdesevir, omipalisib et tipifarnib) à des fins de *repurposing* a montré des effets synergétiques anti-SARS-Cov-2, permettant une prise simultanée à des dosages plus

faibles diminuant ainsi les risques de toxicité et l'apparition d'effets indésirables [Jang W.D., *et al*. 2021]. Plusieurs méthodes sont utilisées pour découvrir ces nouvelles cibles ou permettre la synthèse de nouveaux ligands comme la science des réseaux, ou des méthodes de *docking* qui sont des domaines développés par la suite. La polypharmacologie est une approche qui comporte également des risques lorsque les cibles secondaires sont mal identifiées et peut mener au retrait de médicament *a posteriori* étant donné l'augmentation des risques pour la santé, e.g. le retrait par Merck du rofecoxib car il entrainait une augmentation des risques d'arrêt cardiaque et d'accident vasculaire cérébrale [Dieppe P.A., *et al.* 2004]. L'identification des protéines secondaires non souhaitées est donc primordiale au développement de médicaments s'inscrivant dans une optique polypharmacologique.

1.3 La place des approches *in silico*

1.3.1 Les approches basées sur la structure

Ces approches (*structure-based* en anglais) se fondent sur l'étude des protéines à partir de leur reconstitution en 3D. Une base de données nommée *Protein Data Bank* (PDB) sert de plate-forme principale pour étudier ces structures protéiques puisqu'elle recense en 2022 presque 200 000 d'entre elles. Ces structures sont obtenues expérimentalement via trois méthodes principales [Zardecki C., *et al.* 2022] : la cristallographie au rayons X (i), la spectroscopie RMN (ii) et la cryo microscopie électronique (iii).

(i) La cristallographie au rayonx X est la plus utilisée et consiste à étudier le résultat de la diffraction d'un laser sur une protéine purifiée et cristallisée. Le résultat obtenu est ensuite traité pour former une carte de densité d'électron permettant de déduire les coordonnées de chaque atome. Cette méthode est plutôt pensée pour des protéines rigides, ou peu flexibles [Maveyraud L., *et al.* 2020].

(ii) Pour la spectroscopie RMN, la protéine purifiée est placée dans un champ magnétique et les noyaux atomiques sont déterminés à partir des résonnances obtenues en confrontant la protéine à des ondes radios. Cette méthode est limitée à des protéines de petite taille et est davantage conçue pour des protéines flexibles. De plus au lieu d'un cristal la protéine peut être étudiée en solution [Alderson T.R., *et al.* 2020].

(iii) La cryo microscopie électronique s'affranchie du besoin de cristalliser la protéine et est combinée avec une cryogénisation rapide de la protéine ou de la structure sous différents angles. Les électrons vont ainsi « scanner » ces échantillons et, grâce à la construction d'une carte de densité électronique, reconstruire la protéine en 3D) [García-Nafría J., *et al.* 2020].

Une fois ces structures obtenues, de nombreuses approches de modélisation moléculaire et de bioinformatique structurale sont utilisées. L'intérêt de la modélisation moléculaire est l'identification des sites de liaison des composés actifs, de pouvoir caractériser les propriétés de ces poches, les résidus d'importance et la conformation en 3D de ces paramètres. Deux types d'approches sont couramment utilisées, souvent en combinaison : la dynamique et l'amarrage moléculaire (respectivement *molecular dynamics* ou MD, et *docking* en anglais).

1.3.1.1 La dynamique moléculaire

Une fois la structure de la protéine d'intérêt obtenue en 3D, les simulations de MD vont pouvoir simuler le comportement de cette protéine en fonction d'un temps donné, allant de la nanoseconde à la microseconde, dans un environnement reproduisant celui de la cellule dans laquelle elle se trouve. Pour ce faire, le modèle va être soumis à un champ de force permettant à chaque atome d'interagir avec son environnement [Hollington S.A., *et al.* 2018]. Ces simulations vont mener à plusieurs options principales permettant un gain d'informations différent. Deux de ces options les plus couramment utilisées sont les simulations d'une protéine sous forme *apo* (sans ligand), ou *holo* (avec ligand) (figure 2).

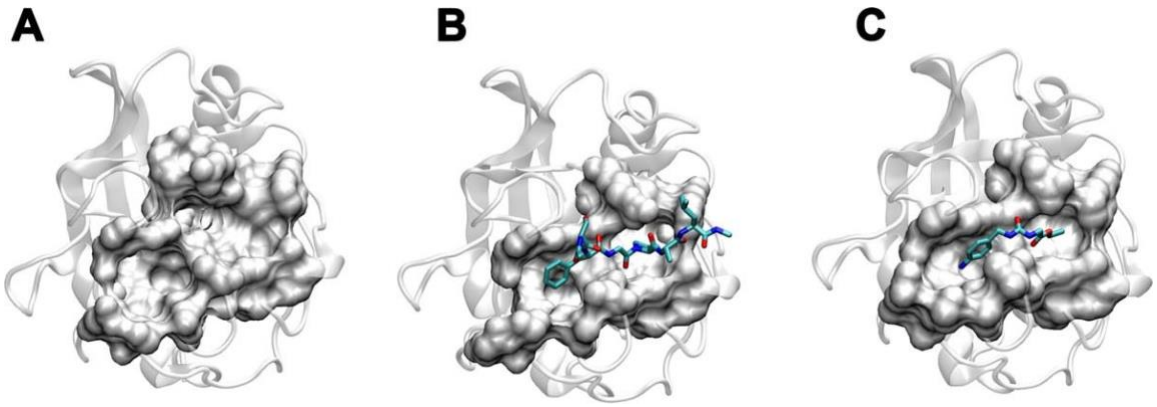


Figure 2 : Exemple de structure de la cyclophiline A A) sous forme apo, B) & C) Sous forme holo avec un ligand, montrant le changement de conformation de la protéine. [Rodriguez-Bussey I.G., et al. 2015]

Les simulations *apo* vont explorer les différentes conformations que peuvent prendre une protéine au cours du temps dans un environnement donné (souvent aqueux et/ou en présence de lipides). Cela va permettre l'identification de poches permettant les interactions avec des médicaments qui n'apparaissent pas lors de l'étude d'une structure statique. [Brown S.P., *et al.* 2006].

Les simulations *holo* vont permettre de caractériser les propriétés du site de liaison et le comportement de la protéine lors de sa liaison au ligand. En effet, certaines protéines lors de leur liaison vont changer de conformation permettant des interactions supplémentaires avec d'autre cofacteurs. Enfin certains sites de liaisons sont profondément enfouis à l'intérieur de la protéine. Pour y accéder le ligand doit traverser un « canal » dont la présence et les interactions permettant l'entrée et la sortie du ligand sont difficilement discernables sur une structure statique [Motta S., *et al*. 2018].

1.3.1.2 Le docking

Le *docking* est une méthode permettant de prédire les interactions entre une protéine et un jeu de ligands en calculant leur force d'interaction grâce à une fonction de score. Contrairement aux approches de DM, le *docking* est effectué le plus souvent sur des protéines statiques, la flexibilité étant gardée pour les ligands seulement. La notion de temps est également retirée. Le principe sera donc de trouver la meilleure « pose » possible d'interaction entre un ligand et sa protéine. Si le site de liaison est bien défini cette recherche peut se faire de manière localisée au niveau de la poche, dans le cas contraire un *docking* sur l'ensemble de la protéine est possible ; on parle alors de *docking* en aveugle. Deux composants principaux vont différencier les différents logiciels de *docking* : l'algorithme d'échantillonnage ou de recherche et la fonction de score [Marjana N., *et al.* 2016].

L'algorithme d'échantillonnage va s'occuper de déterminer les interactions à favoriser et la flexibilité. Certains algorithmes peuvent considérer le ligand comme rigide (parfois la protéine est également flexible). L'algorithme de recherche va s'occuper de lier le ligand à la protéine à des endroits favorisés par les calculs d'interactions inhérents à l'algorithme. L'espace sera ensuite échantillonné, en effectuant des micro rotations ou translations de certaines parties du ligand, afin de générer un nombre de poses permettant une exploration la plus exhaustive possible. Une fois ces poses générées, la fonction de score va permettre d'évaluer les différentes conformations et orientations obtenues (figure 3). L'intérêt de cette fonction est double : elle va classer les poses obtenues pour un ligand et classer les ligands eux-mêmes en fonction de leur score. Ce score peut être énergétique ou non mais va évaluer la viabilité des poses et interactions obtenues par l'algorithme de recherche.

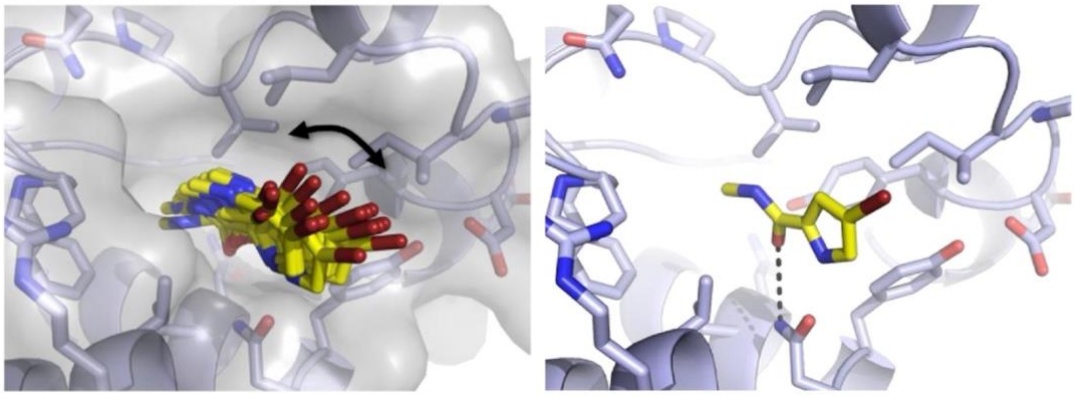


Figure 3: Illustration du positionnement et génération de poses d'un ligand (à gauche) et la meilleure pose déterminée par la fonction de score (à droite). Logiciel SEED. Image modifiée et tirée de Śledź P. et al. 2017

Le principal avantage du *docking* comparé à la MD est sa rapidité. Des milliers de composés peuvent être analysés en quelques minutes contre plusieurs heures voire plusieurs jours selon la complexité d'une MD. Les méthodes de *docking* et de MD sont souvent utilisées ensemble. La première pour filtrer le jeu de données pour ne conserver qu'un groupe de ligands actifs ou juste le plus actif qui sera ensuite analysé par la seconde lors de simulations. Cela permettra de confirmer leur affinité avec la protéine grâce à des analyses énergétiques et de liaisons plus poussées et stabiliser le complexe protéine-ligand formé.

Ces approches de docking et de MD ne vont pas permettre la découverte ou la prédiction d'effets indésirables directement. Néanmoins leur rôle dans la compréhension des mécanismes d'interaction permet le développement de composés spécifiques réduisant en théorie l'apparition d'effets indésirables rendant le développement d'un médicament plus efficace, de manière plus rapide, et moins coûteuse. Ce sont donc des méthodes fondamentales dans les processus de *drug discovery* aujourd'hui.

1.3.2 Les approches basées sur le ligand

Ces approches (*ligand-based* en anglais) sont principalement utilisées lors des phases de développement non cliniques. Elles peuvent être associées avec des approches structurales mais cela n'est pas systématique car ces méthodes peuvent par elles-mêmes chercher à prédire (des activités, de la toxicité, des effets indésirables…) en utilisant différentes informations associées à une petite molécule (données chemo-génomiques, physico-chimiques, structurales…). Des bases de données ont été développées répertoriant certaines de ces informations, notamment la chEMBL [Gaulton A., *et al*. 2017] et PubChem [Kim S., *et al*. 2021].

La ChEMBL est une base de données européenne contenant 2.2 millions de molécules bioactives (figure 4) avec des propriétés médicamenteuses. Les données sont manuellement ajoutées et de nombreuses informations sont présentes notamment sur les différents *assays* effectués, les cibles identifiées, ou encore les types cellulaires et les composés testés étant actifs sur celles-ci.

Contenu retiré pour des raisons de droits d'auteurs

En plus des propriétés obtenues expérimentalement, les outils *in silico* sont également utilisés pour effectuer des prédictions d'interactions pour ces composés avec certaines cibles, ou pour calculer certaines propriétés physico-chimiques importantes dans le développement d'un médicament ; comme les règles de Lipinski ou de Veber [C.A. Lipinski, *et al*. 1997 ; Veber D.F., *et al.* 2002].

PubChem est une base de données américaine regroupant les informations de 869 sources (en constante augmentation) accumulant plus de 111 000 000 de composés (tableau 3).

Data Collection	***Live Count***	***Description***
Compounds	111,507,152	*Unique chemical structures extracted from contributed PubChem Substance records*
Substances	281,974,768	*Information about chemical entities provided by PubChem contributors*
BioAssays	1,466,011	*Biological experiments provided by PubChem contributors*
Bioactivities	295,079,255	*Biological activity data points reported in PubChem BioAssays*
Genes	103,622	*Genes tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents*
Proteins	185,291	*Proteins tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents*
Taxonomy	112,547	*Organisms of proteins/genes tested in PubChem BioAssays and those involved in PubChem Pathways and identified in PubChem Patents*
Pathways	238,908	*Interactions between chemicals, genes, and proteins*
Cell Lines	1,964	*Cell Lines tested in PubChem BioAssays*
Literature	34,357,919	*Scientific publications with links in PubChem*
Patents	42,395,312	*Patents with links in PubChem*
Data Sources	869	*Organizations contributing data to PubChem*

Tableau 3: Résumé des données présentes dans la base de données PubChem (source : https://pubchemdocs.ncbi.nlm.nih.gov/statistics).

Si les données ne sont pas ajoutées manuellement comme pour la chEMBL, son répertoire de composés est beaucoup plus large ; il y ajoute également des informations *in silico* comme les différentes structures en 3D résolues avec ce composé grâce à l'intégration de la PDB ou la structure du composé lui-même.

1.3.2.1 Les modèles *Quantitative Structure-Activity Relationship (QSAR)*

Les modèles QSAR ne sont pas limitées aux approches *ligand-based.* En effet des modèles basés sur la structure des protéines ou des donnée expérimentales peuvent être également développés. En revanche, la majorité des approches QSAR sont implémentées pour du *ligand-based* (prédiction de toxicité, propriétés ADME, activité biologique…), d'où le développement de ces méthodes sur ce chapitre. Elles vont se baser sur les propriétés structurales physico-chimiques des composés, leur empreinte moléculaire (*fingerprint* ou FP en anglais) [Muratov E.N., *et al.* 2020]. Les FPs peuvent aller d'une à trois dimensions allant du nombre d'atomes de carbone aux coordonnées/distances entre deux pharmacophores (figure 5).

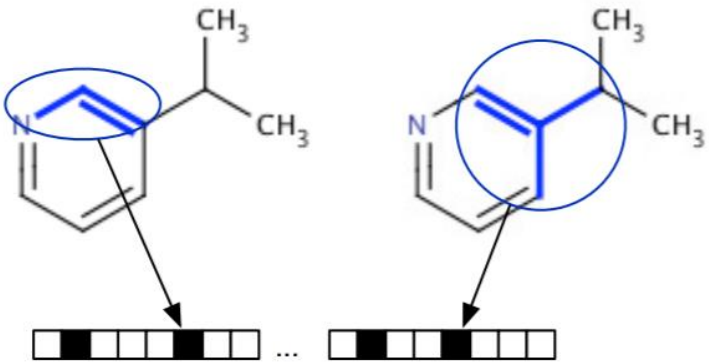


Figure 5 : Transcription de propriétés chimiques en FPs, matrice de 0 et 1, image provenant de la présentation de Gregory Landrum, "fingerprints in the RDKit" RDKit UGM 2012, London (2012).

Cela va permettre de créer des tableaux, reprenant toutes ces propriétés pour chaque composé chimique, pouvant ainsi être analysés. Ces analyses peuvent permettre la prédiction d'interactions protéine-ligand, ou la prédiction d'effets indésirables en se basant sur la similarité de certains composés entre eux grâce à des approches d'apprentissage automatique ou de réseaux de neurones [Dey S., *et al.* 2018]. En plus des FPs d'autres données peuvent être collectées comme les concentrations d'activité, les type d'expériences (*assays* en anglais) effectuées pouvant ajouter des informations utiles pour les analyses.

1.3.2.2 Les pharmacophores

Connaître la structure 2D et 3D de ligands peut conduire à la mise en place de modèles basés sur des pharmacophores (figure 6). Ces modèles s'appuient sur des propriétés moléculaires indiquant les types de groupement et les distances nécessaires entre eux pour qu'un ligand puisse interagir avec une protéine.

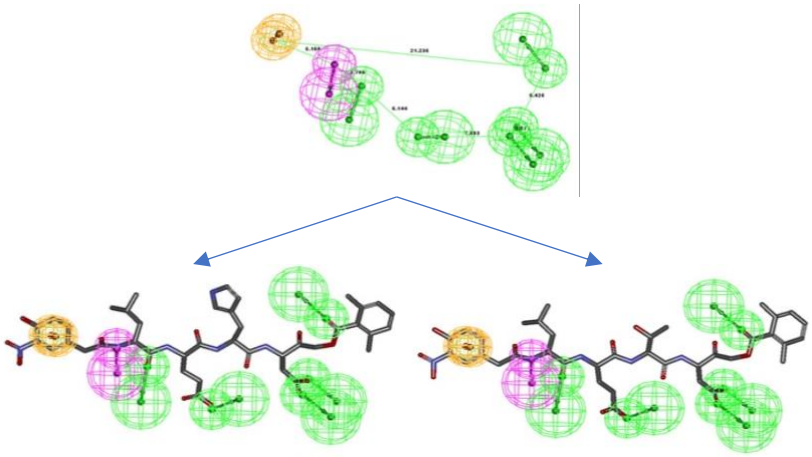


Figure 6: Représentation d'un modèle de pharmacophore en 3D et son application sur deux molécules. En orange les cycles aromatiques, les atomes donneurs d'hydrogène en violet et accepteurs en vert. Image modifiée et tirée de la publication Ahmad K. et al. 2019

Ces informations peuvent être ajoutées dans des modèles QSAR mais l'approche est également intéressante pour constituer un jeu de données structurellement diversifiées ayant les mêmes propriétés clés que le ligand actif de référence.

Les approches basées sur la structure de la protéine et le ligand sont souvent combinées dans des perspectives de découvertes de nouveaux candidats médicaments. Dans l'étude menée par Kumar S. (Kumar S., *et al*. 2020) la combinaison d'approches de *docking* et de MD a permis de mettre en évidence de potentiels candidats médicaments agissant sur la protéase principale du SARS-CoV-2 appelée $3CL^{pro}$. Cette protéine essentielle à la réplication des virus de type coronavirus dans les cellules est une cible thérapeutique intéressante à étudier. Dans un premier temps le *docking* de 13 antiviraux approuvés et agissant sur la protéine pour d'autres coronavirus a été effectué afin de trouver le médicament interagissant le mieux avec celle du SARS-CoV-2. Ensuite les composés analogues à l'indinavir, antiviral ayant obtenu le meilleur score de *docking* sur la $3CL^{pro}$ du SARS-CoV-2, a permis de mettre en évidence 25 composés ayant une meilleure affinité avec la protéine que l'antiviral. Ces composés ont été générés en conservant un pharmacophore considéré important, l'hydroxyethylamine. Un composé respectant les règles de Lipinski a été sélectionné et les MDs ont montré des interactions stables dans la durée avec la $3CL^{pro}$ et similaires à l'indinavir. Un potentiel candidat médicament spécifique à une protéine est donc ainsi mis en évidence pour des études approfondies *in vitro* et *in vivo* par la combinaison d'approches *in silico* et ce à peine trois mois après le début de la mise en place de l'urgence sanitaire en France.

1.3.3 Les approches basées sur la connaissance

Les approches dites *knowledge-based* (KB) en anglais sont des méthodes s'appuyant sur la récupération et l'analyse de données regroupées dans des bases de données. Contrairement aux approches *ligand-based* qui utilisent les propriétés physico-chimiques, les KB se basent sur les interactions connues entre les différents groupes à analyser. De la même manière que les QSARs, des modèles prédictifs peuvent être développés sur ces interactions. De nombreuses bases de données se sont développées regroupant notamment les interactions médicament-cible, médicament-effet indésirable [Mohamed S.K., *et al* 2020 ; Ye Q., *et al.* 2021]. On peut par exemple citer la *drugbank* [Wishart D.S., *et al.* 2018] qui est une base de données regroupant toutes les informations sur les médicaments mis sur le marché, retirés ou en développement, représentant 14 665 composés (version 5.1.9 de la *Drugbank*). Les données provenant de la littérature sont manuellement ajoutées et vérifiées, certaines propriétés physico-chimiques, surtout celles permettant de valider ou non les différentes règles définissant la viabilité d'un candidat-médicament (Lipinski, Veber…), sont prédites. C'est également le cas pour les propriétés ADMET (passage de la membrane hématoencéphalique, l'absorption intestinale…). Cette base de données donne également accès aux interactions médicament-cible (19 198), et aux effets indésirables dus aux médicaments. *Drugcentral* [Ursu O., *et al.* 2017] est une autre base de données qui se concentre principalement sur les médicaments approuvés par la *Food and drug administration (FDA)*. Les données recueillies proviennent de la littérature ou d'autres bases de données comme la *chEMBL*. Cette base de données possède (sur la version du 12 Novembre 2021) des informations sur 4 714 composés actifs, et une liste d'effets indésirables conséquente pour chacun de ces composés.

Les sources permettant de collecter ces interactions étant souvent hétérogènes dans leur construction, les informations recueillies peuvent être formatées et intégrées dans une base de données sous forme de réseaux (figure 7) permettant d'interpréter les différents liens formés ; c'est la science des réseaux.

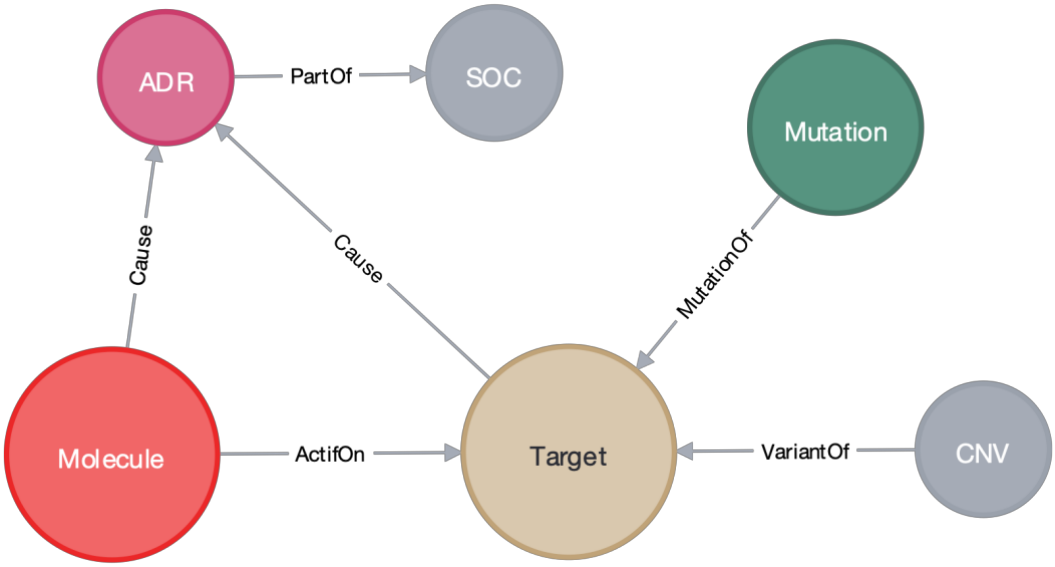


Figure 7: Exemple d'architecture de base de données graphique, les nœuds sont les différentes catégories représentées par des cercles, les relations qui caractérisent ces nœuds sont représentées par des liens, ici orientés.

La science des réseaux est un outil de plus en plus utilisé qui a l'avantage de pouvoir visualiser différentes sources d'information et leur connections sur un seul graphe. Il est basé sur la théorie des graphes et a été remis au goût du jour par Barabási qui exploite ce type d'approche dans de nombreuses études pharmacologiques (notamment le Covid-19) [Barabási A.L., *et al.*, 2010 ; Gysi D.M, *et al.* 2021].

Plusieurs outils ont été développés comme Cytoscape [Shannon P*, et al. 2003*] qui est un logiciel permettant de représenter graphiquement et d'analyser par des approches de graphes les données implémentées. Cytoscape a été développé dans l'idée d'être utilisé en recherche et possède de nombreux *plugins* permettant d'intégrer des données de différentes sources directement via le logiciel et de les traiter comme par exemple stringApp pour la base de données STRING [Doncheva N.T., *et al,* 2019]. On peut également citer le logiciel Neo4J qui est un outil NoSQL (pour *not only SQL)* de haute performance grâce à son système de requête basé sur le langage Cypher. Ce logiciel, utilisé principalement par de grandes entreprises comme la NASA, Allianz ou encore eBay, commence à être utilisé par des groupes pharmaceutiques comme Novartis ou Servier [https://neo4j.com/case-studies/novartis/] dans une optique KB pour associer des données dans une relation composés-gènes-maladies.
Cette science combinant la création de « méta-bases » de données aux approches statistiques telles que l'apprentissage automatique ou les réseaux de neurones permet la création de protocoles de prédictions de maladies ou d'effets indésirables. Dans l'étude menée par Zitnik M., *et al.* (2018) les effets de la polypharmacie sont étudiés pour déterminer les paires de

médicaments liés à la présence d'effets indésirables. En regroupant plusieurs bases de données dont STITCH et SIDER [Szklarczyk D., *et al.* 2015 ; Khun M., *et al.* 2016] avec l'intégration de réseaux développés par d'autres équipes, une méta-base de données de plusieurs millions d'interactions a été générée. Enfin un réseau de neurones apprenant sur les différents liens et nœuds de ce graphique pour prédire l'apparition d'effets indésirables en cas de prise de deux médicaments conjointement a été mis en place.

1.3.4 Les approches phénotypiques

Les approches phénotypiques ont été étudiées dès le début des années 1990 mais furent écartées pour des approches basées sur l'identification des cibles protéiques, plus rapide et moins coûteuses [Haasen D., *et al.* 2017]. Néanmoins avec le développement des outils informatiques les approches phénotypiques redeviennent pertinentes car elles ne nécessitent pas de connaissances *à priori* de la maladie ou des cibles interagissant avec elles. Elles se basent sur des lectures de l'effet d'un composé à un niveau cellulaire, voire au niveau d'un organisme entier, en utilisant des techniques d'imageries [Haasen D., *et al.* 2017]. En combinant ces approches, notamment l'imagerie cellulaire, avec des méthodes d'apprentissage automatique ou d'apprentissage profond (*deep learning* en anglais), ces sorties peuvent être analysées et la perturbation d'un composé à différents niveaux d'un compartiment d'une cellule peut être étudiée et associée à un phénotype (notamment la toxicité d'un composé) [Aulner N., *et al.* 2019]. Une de ces techniques tirant partie de cette combinaison se nomme le *Cell Painting* [Bray M.A., *et al.* 2016]. Cette dernière va relever les caractéristiques morphologiques des cellules à partir d'images et transcrire celles-ci en données quantitatives. Grâce à cela, de nombreuses informations peuvent être tirées des perturbations cellulaires par un composé donné.

Dans le cadre de mes travaux de thèse, j'ai exploité une partie de ces données et outils liés aux sciences des réseaux afin d'explorer et de combiner des approches graphiques et statistiques pour tenter de prédire les effets indésirables causés par les médicaments, leurs possibles origines, tout en prenant en compte la variabilité génétique des individus.

Chapitre 2 : Intégration de données pharmacologiques

Cette partie sera consacrée à l'intégration de données d'interaction protéine-composé provenant de la ChEMBL pour ensuite y associer les données phénotypiques provenant de la *Broad Bioimage Benchmark Collection* (BBBC) [Ljosa V., *et al.* 2012]. Enfin des informations sur les voies de signalisation, ou sur les maladies associées aux protéines provenant de *Gene Ontology* [Mi H., *et al.* 2019], *Disease Ontology* [Schriml L.M, *et al.* 2018] et KEGG *pathway* [Kanehisa M., *et al.* 2000] vont venir compléter cette base de données. Avec la création d'une chimiothèque à partir de ces données, cela permettra un criblage phénotypique plus efficient et d'assister dans les approches basées sur les cibles. L'intérêt sera de déterminer un nombre de composés (5100) touchant le plus grand nombre de cibles thérapeutiques existantes. Si le criblage phénotypique ne se base pas sur la connaissance de cibles particulières, l'utilisation d'une telle librairie peut permettre d'explorer les perturbations cellulaires. La finalité sera de possiblement les lier à une famille, une protéine spécifique, une maladie, nous donnant plus d'information sur les mécanismes d'action des maladies, de la protéine aux effets phénotypiques.

2.1 Les *scaffolds*

Il est estimé à 10^{23} to 10^{60} le nombre de molécules thérapeutiques potentielles, et à 10^{8} le nombre de composés ayant été synthétisés [Lim J., *et al.* 2020]. Des stratégies doivent être utilisées pour explorer l'espace chimique de manière efficace. Pour ce faire un protocole de sélection va être mis en place et reposera sur le découpage de molécules en *scaffolds*. Les *scaffolds* représentent les parties principales des molécules, leur « squelette » (figure 8).

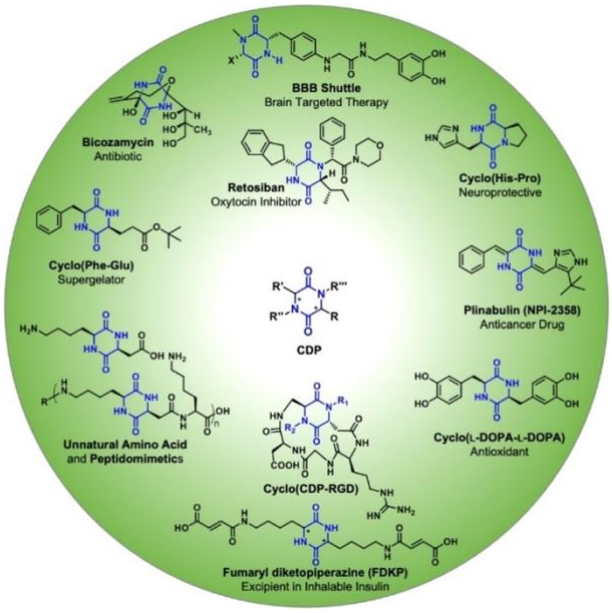


Figure 8: Exemple d'un scaffold (dipeptide cyclique) et de différentes molécules thérapeutiques le possédant [*Balachandra C., et al. 2021].*

Ce découpage permet de garder la structure d'une molécule connue comme étant active sur une cible, pour en garder les caractéristiques principales et d'explorer ou de synthétiser les parties spécifiques à associer à une protéine, par exemple de la même famille que la protéine initiale sur laquelle interagit le composé ayant servi à extraire le *scaffold* [Card G.L., *et al.* 2005]

2.2 L'utilisation d'enrichissement

En utilisant les données de *Cell Painting* obtenues grâce à l'utilisation de la BBBC, les données morphologiques de composés en commun entre la BBBC et la chEMBL peuvent être fusionnées. Pour avoir l'information la plus complète possible sur l'effet de ces composés et connaître l'importance d'une protéine avec laquelle ils interagissent dans un mécanisme biologique ou une maladie, un enrichissement peut être effectué. Des packages existent comme *cluster profiler* [Wu T., *et al.* 2021] permettant grâce à la liste de gènes interagissant avec un composé de calculer leur implication dans les processus biologiques, à partir de bases de données spécifiques, regroupant les voies de signalisation (KEGG *pathway*), les maladies (*Disease ontology*) ou les fonctions biologiques ou moléculaires (*Gene ontology*). Chaque gène ou groupe de gènes agissant sur un même processus va être analysé par rapport au nombre total de gènes sur ce processus afin de déterminer l'importance de ce gène, et par extension de ce composé, sur cette voie biologique en se basant sur une distribution hypergéométrique.

Journal of Cheminformatics

RESEARCH ARTICLE

Open Access

Development of a chemogenomics library for phenotypic screening

Bryan Dafniet[1], Natacha Cerisier[1], Batiste Boezio[1], Anaelle Clary[2], Pierre Ducrot[2], Thierry Dorval[2], Arnaud Gohier[2], David Brown[2], Karine Audouze[3] and Olivier Taboureau[1*]

Abstract

With the development of advanced technologies in cell-based phenotypic screening, phenotypic drug discovery (PDD) strategies have re-emerged as promising approaches in the identification and development of novel and safe drugs. However, phenotypic screening does not rely on knowledge of specific drug targets and needs to be combined with chemical biology approaches to identify therapeutic targets and mechanisms of actions induced by drugs and associated with an observable phenotype. In this study, we developed a system pharmacology network integrating drug-target-pathway-disease relationships as well as morphological profile from an existing high content imaging-based high-throughput phenotypic profiling assay known as "Cell Painting". Furthermore, from this network, a chemogenomic library of 5000 small molecules that represent a large and diverse panel of drug targets involved in diverse biological effects and diseases has been developed. Such a platform and a chemogenomic library could assist in the target identification and mechanism deconvolution of some phenotypic assays. The usefulness of the platform is illustrated through examples.

Keywords: Phenotypic screening, Phenotypic drug discovery, Chemical biology, System pharmacology network, Network pharmacology, Chemogenomics

Introduction

In the past 2 decades, the drug discovery paradigm has shifted from a reductionist vision (one target—one drug) to a more complex systems pharmacology perspective (one drug—several targets) [1]. The reasons are related, notably, to the number of failures of drug candidates in advanced stages of clinical trials due to a lack of efficacy and clinical safety [2]. Furthermore, the traditional expectations that selective ligands act on a single target are now challenged with new drug discovery processes, especially for complex diseases like cancers, neurological disorders and diabetes as they are often caused by multiple molecular abnormalities rather than being the result of a single defect [3–5].

To accelerate drug discovery research in chemogenomic, systematic screening programmes of targeted chemical libraries against a set of protein families have emerged. For example, to discover new drugs to treat cancer, a library consisting of known kinase inhibitors may be screened to identify hit compounds and then start a medicinal chemistry programme. Similar exercises have been performed with GPCR-focused libraries [6] and protein–protein interaction inhibitors [7].

More general chemical libraries were also built up representing collections of selective small pharmacological molecules that can modulate protein's targets across the human proteome and be involved in a phenotype perturbation. With the increased facility for academics to get access to large chemical libraries, chemogenomic, proteochemometric or polypharmacology approaches have started to be developed allowing to mine this vast amount of protein–ligand interactions and to predict

*Correspondence: olivier.taboureau@u-paris.fr
[1] Université de Paris, INSERM U1133, CNRS UMR8251, 75006 Paris, France
Full list of author information is available at the end of the article

a single ligand against a set of heterogeneous targets [8–10]. Associations between drug-target and gene-disease started to be investigated through druggable genome studies [11–13]. Collection and processing of a wide array of genomic, proteomic, chemical and disease-related resource data were also explored using network pharmacology approaches [14, 15]. Network pharmacology combines network sciences and chemical biology allowing the integration of heterogeneous sources of data and the possibility to look over the action of a drug on several protein targets and their related biological regulatory processes in system biology [16]. Multiple studies have reported new insights in drug target clinical outcomes based on the combination of chemogenomics, network analysis and diseases [17–19].

Among chemical libraries considered in chemogenomic studies, many of them have been built by industrial companies like the Pfizer chemogenomic library, the GlaxoSmithKline (GSK) Biologically Diverse Compound Set (BDCS), Prestwick Chemical Library and the Sigma-Aldrich Library of Pharmacologically Active Compounds, but some of them are also available for public screening programmes like the Mechanism Interrogation PlatE (MIPE) library that was developed by the National Center for Advancing Translational Sciences (NCATS). More details about these chemogenomics libraries can be found here [20].

For a few years, there has been a revival of phenotypic screening in drug discovery. However, the chemical libraries discussed previously are not always optimised for such studies. In fact, with the advances in various technologies for cell-based phenotypic screening, including the development of induced pluripotent stem (iPS) cell technologies, gene-editing tools such as CRISPR-Cas and imaging assays technologies, new phenotypic drug discovery studies are reported in the literature [21–25]. Image-based high-content screening (HCS) on 30,000 small molecules has been for example used with a generative adversarial network to propose new small molecule structures that share similar morphological profile [25]. Therefore, as phenotypic drug discovery studies do not rely on knowledge of the molecular target perturbed by a specific drug, the translation of the molecular mechanism of action in the context of a disease-relevant cell system i.e., molecular phenotyping is the next challenge.

In this context, we decided to develop a pharmacology network for phenotypic screening, integrating the ChEMBL database [26], pathways, diseases and a high-content image-based assay for morphological profiling, Cell Painting [27], in a high-performance NoSQL graphics database (Neo4j®). The aim is to identify proteins modulated by chemicals that could be related to some morphological perturbations at the cell level and lead to some phenotypes, diseases and/or adverse outcomes. Furthermore, a chemogenomic library of 5000 small molecules that represents a large panel of drug targets involved in diverse biological effects and diseases was built. Using filtering based on scaffolds, this library encompasses the druggable genome represented within our network pharmacology and that can be of interest for phenotypic screening. The protocol considered in the development of the network pharmacology is discussed further through examples in the next sections.

Materials and methods

Database

ChEMBL

The ChEMBL database (version 22) [28] was used for this analysis. ChEMBL accumulates standardised bioactivity, molecule, target and drug data extracted from multiple sources (including literature). It contained 1,678,393 molecules with bioactivities defined as Ki, IC50, EC50 among others, and 11,224 unique targets for different species.

Kyoto Encyclopedia of Genes and Genomes (KEGG)

The KEGG pathway database (Release 94.1, May 1, 2020, https://www.kegg.jp) is a collection of manually drawn pathway maps representing the known molecular interactions, reactions and relations networks for several pathway categories such as the metabolism, cellular processes, genetic information processes, human diseases, or drug development [29]. The KEGG pathway was integrated into the drug-target library collected from ChEMBL.

Gene ontology (GO)

The Gene ontology (GO) resource (release 2020-05, http://geneontology.org) provides computational models of biological systems from many different organisms, from humans to bacteria, at the molecular level to pathways level. It can provide an annotation to the biological function and process of a protein. It contained more than 44,500 GO terms, 29,211 biological process terms, 11,113 molecular function terms and 4184 cellular component terms for ~1.4 M of annotated gene products and 4593 Annotated species [30].

Human disease ontology (DO)

The DO resource (release 45, v2018-09-10, http://www.disease-ontology.org) provides a human-readable and machine-interpretable classification of biomedical data that are associated with human disease [30]. The DO resource includes 9069 DO identifiers (DOID) disease terms.

Morphological profiling

Morphological profiling data from 20,000 compounds were gathered from the Broad Bioimage Benchmark Collection (BBBC) using the BBBC022 dataset called "Human U2OS cells—compound-profiling Cell Painting experiment" [32] (information: https://data.broadinstitute.org/bbbc/BBBC022/). Basically, U2OS osteosarcoma cells were plated in multiwell plates, perturbed with the treatments to be tested, stained, fixed, and imaged on a high-throughput microscope. Then, an automated image analysis using CellProfiler (http://cellprofiler.org/) identified individual cells and measured morphological features on each of them in the aim to produce a cell profile [33]. In the end, the comparison of the cell profiles treated with different molecules (or experimental perturbations) allowed to suit different objectives such as identifying the phenotypic impact of chemical or genetic perturbations, grouping compounds and/or genes into functional pathways, and identifying signatures of disease [34]. In the BBBC022 dataset, there are 1779 morphological features measuring intensity, size, area shape, texture, entropy, correlation, granularity, angle between neighbours, etc. These parameters concern three "cell objects": the cell, the cytoplasm and the nucleus. For our study, only the relevant information was kept. As each compound has been tested between 1 and 8 times, the average value of each feature for each compound was used. Features with a non-zero standard deviation and not correlated with each other (less than 95%) were kept in each of the three classes. Finally, we have extracted the data matching the compounds extracted from the ChEMBL database.

Methods

Scaffold hunter

We used a software called ScaffoldHunter [35] to cut each molecule into different representative scaffolds and fragments as follow:

(i) Removing all terminal side chains preserving double bonds directly attached to a ring.

(ii) Removing one ring at a time using a set of deterministic rules in a stepwise fashion to keep the most characteristic "core structure" until only one ring is left.

Scaffolds are distributed in different levels based on their relationship distance from the molecule node (Fig. 1).

Neo4J®

The main tool used to create the graph database is Neo4j® (https://neo4j.com/). It allows the integration of large scales of data from numerous sources. Its architecture is composed of nodes that represent a specific object (e.g., molecules, scaffolds, proteins, pathways, diseases…) linked by edges representing a relationship between two nodes (e.g., a scaffold being part of a molecule, a molecule targeting a protein, a target that acts in a pathway, etc.).

R package (cluster profiler, ggplot…)

R package cluster profiler (version 3.14.3) was used to calculate the GO enrichment and KEGG enrichment [36]. The R package DOSE (version 3.12.0) was used to perform the DO enrichment [37]. All the enrichment functions were used with the adjustment method "Bonferroni" and the p-value cutoff set at 0.1.

The R package org.Hs.eg.db [38] (version 3.10.0) was used to translate "EntrezID" [unique gene ID from the Entrez Gene database at the National Center for Biotechnology Information, (http://www.ncbi.nlm.nih.gov/gene)] to SYMBOL (Gene Name) and GO term.

Network pharmacology building

The heterogeneous sources of data were integrated into a network pharmacology database. First, we only selected compounds that have at least information on one bioassay (5,03,000 molecules) and integrated them in two main nodes of our network: "Molecule", containing InchiKey and SMILES information and "CompoundName", containing the chemical name and the database

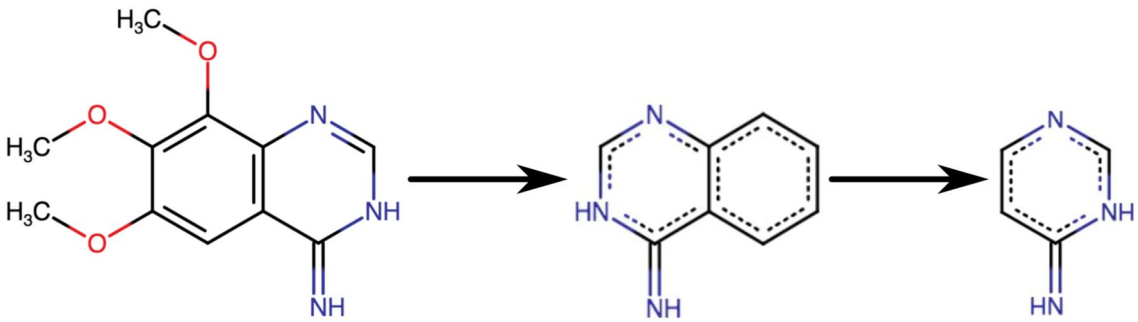


Fig. 1 Illustration of the cutting process of Scaffold Hunter, from the whole molecule to one ring

from which that name was extracted. We added 3 types of nodes related to assays: "Result" which mainly contains the value of the assay (from IC50, Ki…), linked to an "AssayParameter" providing the type of assay (IC50, Ki…) and unit of the value. The type of assay between A (ADME), B (Binding), F (Functional), U (Unassigned) and the confidence score defined by ChEMBL with a scale between 0: uncurated data and 9: direct target assigned were included. We integrated the "Target" node corresponding to protein targeted by the assays and only considered three species: human, rat and mouse. Then, we created a node "UniprotInter" (UI) which contains the generic ChEMBL name without species information and the added UniProt [39] (corresponding to the "Entry_name" in Uniprot).

The "UniprotInter" nodes were linked to a "ProteinClass" node extracted from the ChEMBL and containing information on the protein class to which a protein belongs. This classification schema has several levels (from 1 to 7) and goes from a specific classification (i.e., Metallo Protease M10A subfamily) to a general one (i.e., Enzyme). An example of this classification schema is illustrated in Additional file (Additional file 1: Figure S1).

For compounds present in the network and for which morphological profile is known, 3 nodes ("CellDesc", "NuclDesc" and "CytoDesc") including major features on these respective compartments (cell, nucleus and cytoplasm) were linked to the compound (CompoundName node).

The KEGG, DO and GO nodes are linked to the targets that are involved in the pathways and diseases respectively. As one target may act in several pathways or diseases, a single pathway and disease node can be linked to several targets.

Compound's selection

For the compounds' selection, only bioactive molecules with level 2 scaffolds and first-level protein classes were considered. It allows removing large series of molecules having too many analogues that can be kept with level 1 scaffolds and limit the association of a large set of molecules to general scaffolds such as benzene. Also, to limit promiscuous compounds, all scaffolds that were linked to more than 6 targets were removed.

As the "Target" information is regrouping 3 species, one target may be represented multiple times with only the species varying (e.g., 5HT1A_HUMAN and 5HT1A_RAT). To remedy this issue, we use the "UniprotInter" (UI) node that does not take species into account, so the information is not redundant.

Then, a binary matrix that annotated the bioactivity profile for each scaffold (in rows) with all the targets (in columns) was created. Scaffolds belonging to an active compound with a bioactivity for a target was noted as 1, 0 otherwise. Based on this matrix, hierarchical clustering was performed to separate the scaffolds into clusters.

We decided to select one scaffold per cluster using the following principle:

- The scaffold with the lowest distance, based on a distance matrix using the dist function in R with the binary method, equivalent to Jaccard/Tanimoto indices.
- If there were scaffolds with the same distance, we selected them based on the number of targets they hit, the highest being prioritised.
- Finally, we chose the scaffold that is linked to the highest number of molecules.

If all of these criteria were not able to filter one scaffold by cluster, we considered the scaffolds to be similar and took one among the ones remaining.

Once all the scaffolds were selected, we extracted all active molecules linked to them and performed a multiobjective Pareto optimisation [40] using Pipeline Pilot to select 5000 molecules that will represent the chemogenomic space present in ChEMBL.

Similarly, to the scaffold selection, the compound selection by Pareto was based on 3 criteria:

- Prioritise molecules with the most targets to maximise the different biological profiles.
- Prioritise molecules to maximise the number of scaffolds selected.
- Prioritise molecules to maximise the average number of times a target is hit.

The Pareto method uses a genetic algorithm to generate the best subsets possible. The considered parameters were several subsets created up to 1000, a subset size of 5000 compounds and 600 iterations. The mutation rate parameter was unchanged.

Results

Network pharmacology development

A representation of the final graph database developed with Neo4J is shown in Fig. 2. Globally, 1,61,468 molecules that have a Ki/IC50 activity below 1 µM, a confidence score of 9 among bioassays of type B and bioactive in mouse, rat and human were integrated into the network. This ensemble of compounds modulates 1975 targets which will be considered for further filtering steps. A direct link between the node "Molecule" and "Target" called actifConf9 was created to facilitate the database manipulation.

From this set of bioactive compounds, 1,13,853 distinct scaffolds were generated and integrated into the network. For the protein classes, ChEMBL has defined

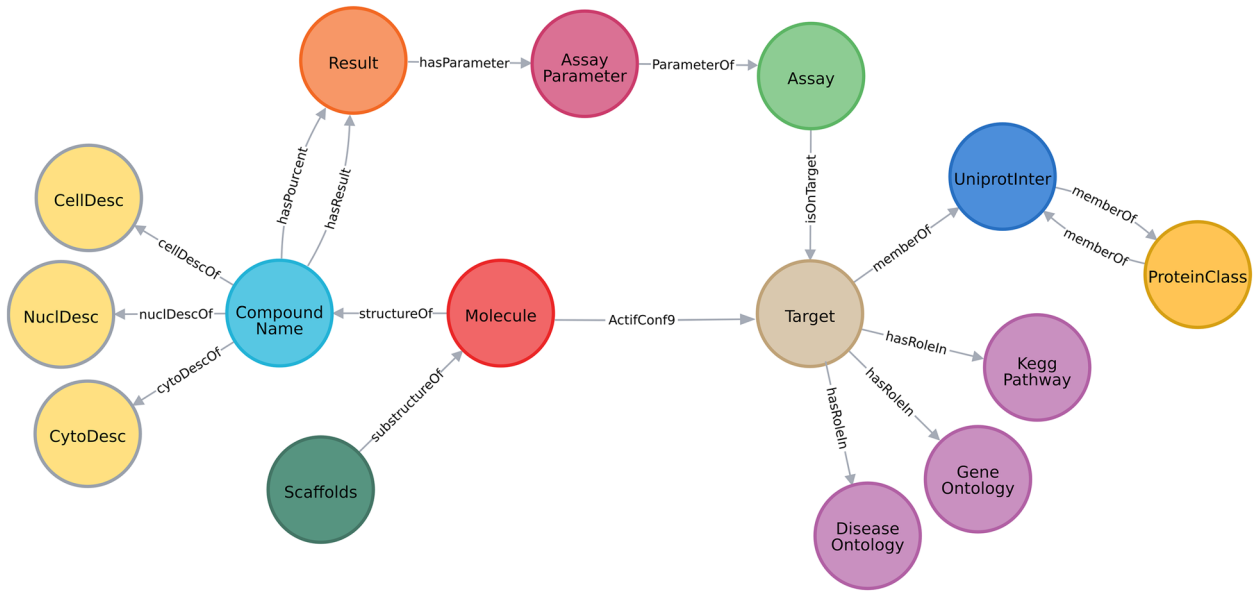
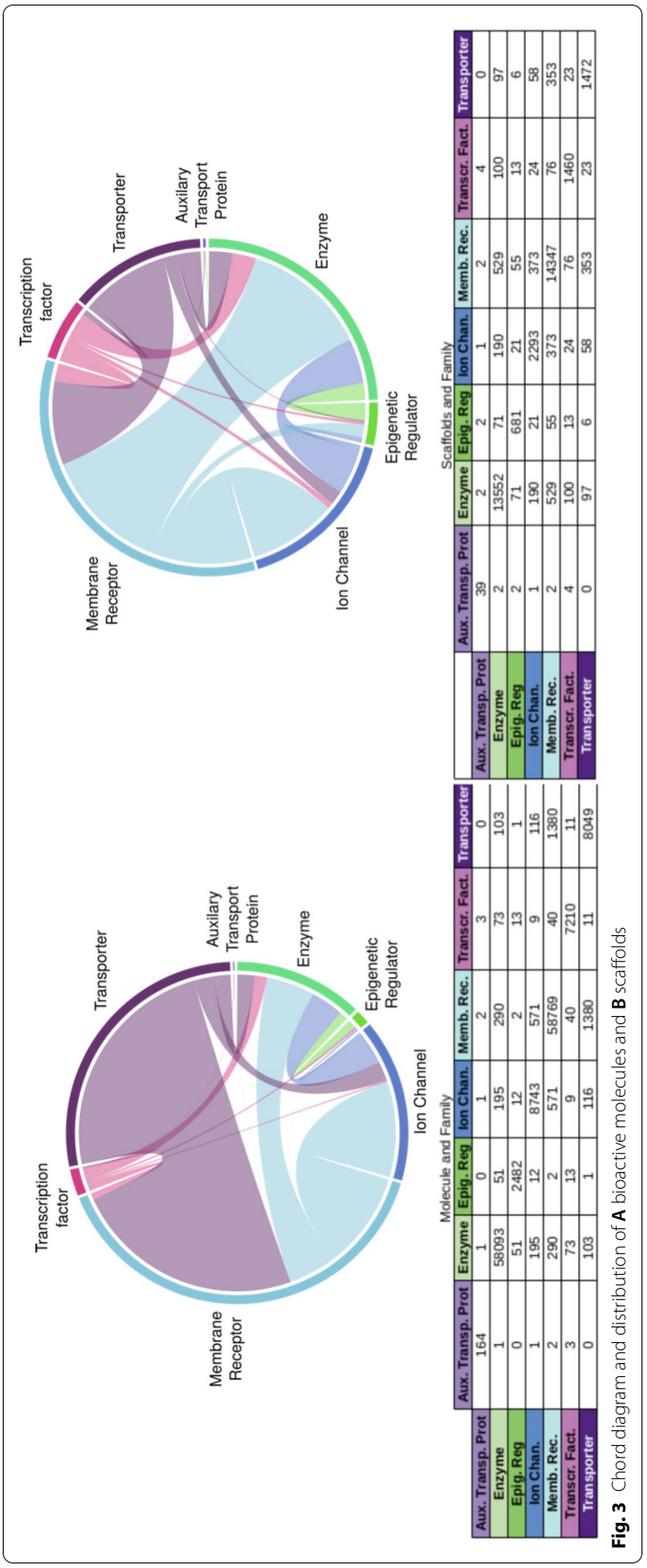


Fig. 2 Overview of the database, coloured by the type of node, the arrows corresponding to the type of relationship between the nodes. The CellDesc, the NuclDesc and the CytoDesc represent features related to the compartment Nucleus, Cytoplasm and Cell observed on the U2OS cell line

1073 distinct classes distributed in 7 main protein families. They represent the main area of drug discovery investigation, notably the membrane receptor [with the G protein-coupled receptors (GPCR)] and the enzyme. Only the protein classes at level 1 protein classes directly connected to a UniprotInter node were considered in this study ending up with 363 protein classes. The distribution of molecules and scaffolds into the 7 families are depicted in Fig. 3. Number of molecules (and scaffolds) that have been reported active on several protein families are also depicted in chord diagrams (Fig. 3). We noticed that among the molecules active on Transporter, many of them (1380 molecules) are also active on membrane receptors. In opposite only few molecules active on Auxiliary Transport Protein have been reported to be active on another protein family.

From the pharmacology network, several information can be obtained such as multiple targets profile associated with a compound and its scaffold. For example, crizotinib, an inhibitor of tyrosine kinase receptors used for the treatment of non-small lung cancer targeted several proteins that belong to different protein classes but all membered of one main protein family, kinase (Additional file 1: Figure S2). Interestingly, looking through the network, the number of molecules sharing the same scaffolds with crizotinib can be collected. This set of molecules could be suggested to have activities on these tyrosine kinase receptors. Similarly, potential new bioactivities not observed in previous studies could be proposed to crizotinib based on scaffold similarity with bioactive molecules.

Furthermore, based on the drug-targets network, it was possible to include pathways and diseases information allowing us to highlight known links between chemicals, proteins, pathways and diseases. In this network, we were able to add 766 GO terms, 301 KEGG pathway terms and 562 diseases ontology terms (DO) and performed enrichment analyses. For each compound linked to at least two proteins that are involved in the same pathway, a p value (adjusted according to the number of genes involved) was computed. It allowed to directly link compounds to pathways and to determine pathways that are statistically enriched in a protein's list. For example, the tozasertib molecule (pan-Aurora kinase inhibitor, anticancer treatment) is linked to 4 proteins: FLT3 (Fms-like tyrosine kinase 3), DDR2 (Discoidin domain receptor tyrosine kinase 2), AURKB (Aurora kinase B) and AURKA (Aurora kinase A) (Fig. 4A) in our network. Two of these targets (FLT3 and DDR2) are involved in the same gene ontology (GO) term "transmembrane receptor protein tyrosine kinase" (GO:0004714). The enrichment for this GO term showed a calculated p value of 2.54e-24, meaning that the tozasertib has a significant influence on the transmembrane receptor protein tyrosine kinase activity. Interestingly, the AURKA and AURKB genes are also involved in kinase activities (histone kinase activity and protein S/T/Ykinase activity) whose activations are necessary for cell division processes in the regulation and



Molecule and Family

	Aux. Transp. Prot	Enzyme	Epig. Reg	Ion Chan.	Memb. Rec.	Transcr. Fact.	Transporter
Aux. Transp. Prot	164	1	0	1	2	3	0
Enzyme	1	58093	51	195	290	73	103
Epig. Reg	0	51	2482	12	2	13	1
Ion Chan.	1	195	12	8743	571	9	116
Memb. Rec.	2	290	2	571	58769	40	1380
Transcr. Fact.	3	73	13	9	40	7210	11
Transporter	0	103	1	116	1380	11	8049

Scaffolds and Family

	Aux. Transp. Prot	Enzyme	Epig. Reg	Ion Chan.	Memb. Rec.	Transcr. Fact.	Transporter
Aux. Transp. Prot	39	2	2	1	2	4	0
Enzyme	2	13552	71	190	529	100	97
Epig. Reg	2	71	681	21	55	13	6
Ion Chan.	1	190	21	2293	373	24	58
Memb. Rec.	2	529	55	373	14347	76	353
Transcr. Fact.	4	100	13	24	76	1460	23
Transporter	0	97	6	58	353	23	1472

Fig. 3 Chord diagram and distribution of **A** bioactive molecules and **B** scaffolds

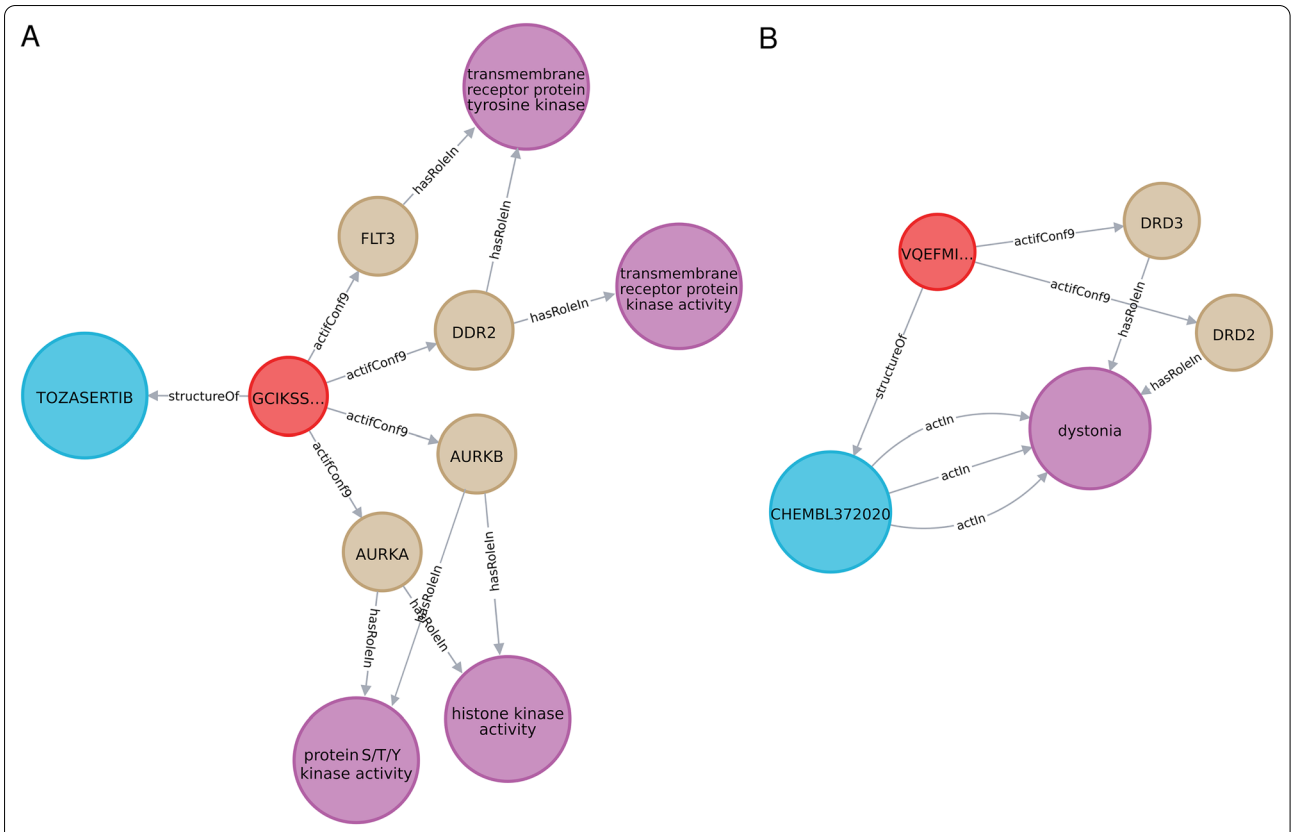


Fig. 4 Example of pathway enrichment for the **A** tozasertib molecule involved in 4 GO and **B** Chembl372020 molecule [(S)-7-Dipropylamino-5,6,7,8 -tetrahydro-indolizine-3- carbonitrile] involved in DO "dystonia"

control of mitosis. All of these proteins play an important role in a wide range of cancers and it explains the interest of tozasertib as an anticancer treatment. As a second example, the molecule Chembl372020 [(S)-7-Dipropylamino-5,6,7,8-tetrahydro-indolizine-3-carbonitrile] is linked to two targets/genes, DRD3 (dopamine receptor D3) and DRD2 (Dopamine receptor D2), both involved in dystonia. The calculated p value enrichment for the DO term (represented in Fig. 4B by the relation arrows "actIn") is 8.42e-05. It means that the molecule is significantly involved in dystonia through DRD3 and DRD2 genes.

Morphological profile integration

Finally, we integrated the morphological profiles for compounds in common between the ChEMBL and the BBBC dataset. We found 2473 compounds common to both datasets. It means that for this set of compounds, proteins are annotated and can be suggested to the morphological perturbations observed in the U20S cell line. Morphological features are included in the network according to the 3 cellular components described in Cell Painting: the nucleus, the cytoplasm and the whole cell itself (respectively named "nucl.", "cyto." and "cell"). This phenotypic information could highlight links between the target compartment and the phenotypic variations associated with the molecule. Among others, features may be a measure of the mean radius of the cytoplasm area shape ("Cytoplasm_AreaShape_MeanRadius"), the location of the centre of the cell according to the X-axis ("Cells_Location_Center_X") or the entropy in the nucleus of the cell ("Nuclei_Texture_Entropy").

A features selection was applied for features concerning the same cellular component. Among the 1779 features, only 767 were kept: 250 for cell, 261 for cyto and 256 for nucl respectively. Overall, a relation between a bioactive molecule on specific proteins and morphological perturbation can be suggested. For example, ciglitazone, a thiazolidonedione with potential interest in ovarian hyperstimulation syndrome or as an anti-hyperglycemic agent is a selective agonist to the nuclear receptor PPARy (Peroxisome proliferator-activated receptor gamma) and shows morphological perturbations for different features i.e., "Cytoplasm_Correlation_Manders_DNA_ER", "Cytoplasm_Correlation_Manders_RNA_ER"" or "Cells_Correlation_Manders_Mito_ER". So, this analysis could suggest

a relation between the activation of PPARγ and the morphological disturbance of some compartments in cells.

Chemogenomics library development

Based on our graph database, we decided to develop a chemogenomic library of 5000 molecules that would cover the chemogenomic space and could be used for phenotypic screening. A workflow of the protocol is shown in Fig. 5.

In the first step, from the set of bioactive molecules, we selected sub-scaffolds at level 2. Such selection allowed to remove too specific scaffolds of a molecule observed at level 1, but still capturing selectivity of molecules associated with some proteins. The main objective is to avoid a general scaffold (i.e., only a ring) that would not be specific enough to discriminate between molecules when trying to select active ones for a target. Then, to limit promiscuity, all scaffolds that were linked to more than 6 targets were removed, (being the beginning of the curves' elbow in Additional file 1: Figure S3), retaining 32,038 scaffolds.

In a second step, we focused on the protein's space. From the 7 main protein's classes defined in ChEMBL, only the first level protein classes were selected and connected to 1221 protein's targets resulting in 363 protein classes. This left us with 32,038 scaffolds linked to 1221 targets belonging to 363 Protein Classes. On average, there were 3.4 molecules/scaffold and 1.5 targets/scaffold.

In our network pharmacology, the 1221 targets correspond to 850 UniprotInter nodes (UI) i.e., proteins having a unique function, independently of the species. From there, the third step consisted of establishing a set of 5000 molecules that cover as much as possible the 850 UI. We decided to select 5000 scaffolds, to have a high diversity of molecules covering the protein's space. To do that we developed a hierarchical clustering that allowed us to select 5000 scaffolds linked to 41,620 molecules and hitting 850 UI. Then, we performed a Pareto multiobjective optimisation which selected the best subsets of 5000 molecules satisfying the criteria defined in the method section.

The Pareto optimisation created multiple "fronts" which correspond to a dataset containing multiple subsets of 5000 molecules. They had the same range of results concerning the criteria and we decided to select the one maximising both the biological profile and the number of scaffolds. As such the 43rd out of 170 subsets from the 1st front matched those criteria and was chosen to represent the 5000 compounds (Additional file 1: Figure S4).

We figured out that by selecting protein classes at level one, some proteins (94 proteins) were not targeted by one of the 5000 compounds or their scaffolds. This is due to the ChEMBL proteins classification schema for

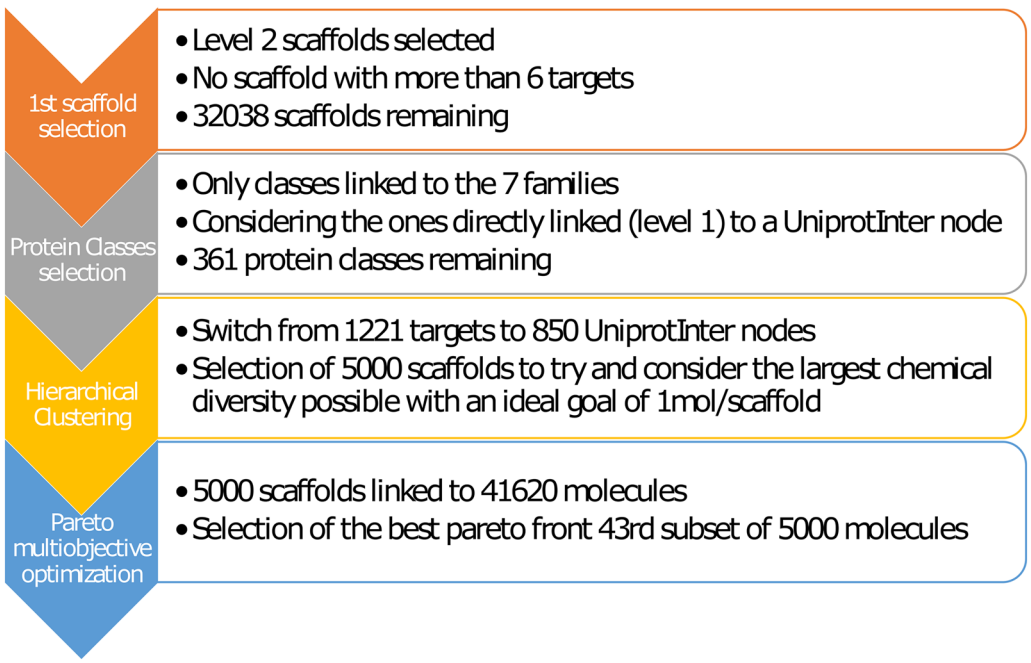


Fig. 5 Workflow of the 5000 molecules selection

which some proteins were not associated with one of the 7 main families. Therefore, in a final step, to cover the maximum of the chemogenomic space, bioactive compounds to missing proteins and capturing most of the molecules through their scaffold were included in the prior set of molecules. Overall, we obtained a library of 5100 molecules with a high diversity of scaffold targeting all bioactive proteins in ChEMBL and that could be used for phenotypic screening. We can observe in this library that on average, there are a little more than two active compounds per protein i.e., with a Ki or IC_{50} lower than 1 μM. Many compounds are active on several proteins (see Additional file 1: Figure S5) which allow associating several scaffolds to a specific protein but also to determine the promiscuity of proteins with scaffolds that could be of interest in the design of drugs acting on multiple targets or disease pathways i.e., polypharmacology.

Interestingly, only a few chemicals from this library also had information on the Cell painting data and around 10% of the compounds in the phenotypic data is also present in ChEMBL. In addition, many of these compounds did not pass the confidence score (score of 9) applied in ChEMBL and a bioactivity threshold (< 1 μM) that allow selecting highly active compounds. It means that only a few chemicals are shared between the two databases. Nevertheless, for these chemicals a relation between their morphological profiles and molecular mechanisms could be proposed.

Discussion

With the aim to relate the modulation of the protein's function by chemicals to some phenotype variations, we created a system pharmacology network, integrating chemical-protein-pathways and phenotypic screening from two different sources, disease ontology and morphological features of cells. The representation of the molecules into scaffold facilitates the recognition of chemotypes i.e., chemical patterns (opioid, benzodiazepine…) associated with specific proteins, the diversity of scaffolds linked to a protein and the diversity of proteins targeted by a series of molecules with a unique scaffold. The incorporation of phenotypic data allows us to go one step further and to assist in the target deconvolution of phenotypic assays. Although high content imaging analysis allow to observe and to measure the morphological disturbance of a cell by a chemical, such technology do not give information about the molecular mechanism that underlies the cell perturbation. The integration of chemical-protein activity from ChEMBL with chemical-morphological profile from Cell Painting, can help to identify proteins that could explain the morphological change of a cell by a chemical and so the potential phenotypic and/or disease impact. The drug-targets-pathways-diseases relationships might help in the investigation of repurposing drugs or a combination of bioactive drugs on two complementary proteins involved in the same pathway. The system pharmacology network is not fully accomplished and phenotypic outcomes could be caused by some targets not yet determined for a compound. Other databases could be integrated. Among them, PubChem [41], ChemProt, DrugCentral [42] databases would be useful to enrich drug-target interactions. Furthermore, with microarray and next-generation sequencing technology, deregulation of genes and pathways caused by a compound in specific conditions (dose, time, cell type, organ, species) like for example in LINCS [43] would be beneficial for obtaining a more comprehensive chemogenomic network. Several initiatives have been developed to identify modes of action of bioactive compounds based on transcriptomics data to suggest new therapeutic indications for a variety of diseases [44, 45]. For example, Iskar et al. combined drug-target information and gene expression profiles after drug treatment to identify the deregulation of new drug-target interactions that could explain the repurposing of drugs or potential side effects associated with them [46]. It is important to notice that the scaffold composition is highly dependent on screening libraries considered and methods used to generate scaffolds [47, 48]. Recently, the implementation of scaffold network has been introduced as a powerful method to navigate and to analyze large screening data sets and could be an alternative to the scaffold selection used in our study [49, 50]. Also, in addition to scaffolds that can help to recognize certain chemotypes, other methods based on activity cliffs could be interesting to integrate as it consists of interpreting a set of structurally similar compounds with a large difference in potency against their target [51].

Overall, our systems pharmacology network captures a large ensemble of drug-target interactions with high confidence and based on a state-of-the-art NOSQL graphics database (Neo4J) facilitating the manipulation of large sets of data in a fast and efficient manner. The integration of biological data such as pathways, diseases and phenotypic screening allows to study the effect of a molecule not only at the molecular level but also in more complex layers of a systems biology and can reveal novel repurposing and synergistic therapeutic opportunities or drug safety issues.

Once the systems pharmacology network was developed, we decided to develop a chemical library limited to 5000 molecules that could be of interest in phenotypic drug discovery campaigns. Several aspects have been considered in the development of the library such as (i) accuracy about drug's bioactivity (ii) diversity of molecular scaffolds (iii) diversity of targets and target family

across the human proteome (iv) diversity of pathways perturbations and diseases associated with chemicals.

Eventually, we obtained a library of 5100 compounds targeting a large ensemble of the proteome i.e. 1234 proteins corresponding to 944 UI (Additional file 2). Compared to GSK and Pfizer libraries which are dominated by kinase, GPCR (Pfizer also includes ion channels), our chemical library is more diverse as it contains transcription factor, enzyme and epigenetic receptors among others. The number of 5000 compounds was chosen based on the fact that it converges to the size of libraries reported by pharmaceutical compagnies (~3000 for Pfizer and ~6000 for GSK libraries respectively)[52]. Our library certainly not covers the complete chemogenomic space but it is more affordable compared to a full HTS, still encompassing a large set of chemical-protein interactions represented in ChEMBL, that is suitable for a hit identification study in early drug discovery program.

The diversity of scaffolds and biological profiles obtained through the Pareto selection give also a much more comprehensive representation of the proteome. Further selections of compounds impacting the genome, and thus other targets, could be performed using other technologies from genomic screening (si/shRNA, CRISPR-Cas9, RNAi, transcriptomics).

Based on this study, we identified 2473 chemical-target interactions from ChEMBL with morphological profiles from Cell Painting. At the scaffold level, common chemotype associating scaffold-proteins and morphological profiling can be suggested. The fact that our chemical library is essentially based on compounds with pharmacological interest will probably have a better merit in deciphering pharmacological mechanisms with disease phenotypic screening. Including some compounds known to generate a broad range of toxic mechanisms would be necessary to predict cellular phenotypic profiles with molecular perturbations.

Conclusion

The developed systems pharmacology network is an interesting tool that can be used in drug recommendation and repurposing. The integration of pathways and phenotypic data allows linking molecular mechanisms to disease pharmacological compounds. Additional data such as high-throughput transcriptomic would be interesting to incorporate in such a network to get insights into the genome-scale perturbation of a compound. Expanding on our previous efforts with a combination of proteome and transcriptome modulations by compounds and linking these data with phenotypic screening would pave the way in phenotypic drug discovery. Furthermore, optimization of a chemical library that would encompass the information coming from these new chemical biology technologies would facilitate the identification of molecular mechanisms to phenotype and the discovery of novel pharmacological entities.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-021-00569-1.

Additional file 1: Figure S1. The "proteinClass" node (in yellow) Serine protease is a level 1 protein class, and the node Protease is a level 2 protein class for the "UniprotInter" node Serine protease hepsin (colored in blue). They are linked by a relationship member of which indicates their belonging to a specific family. **Figure S2.** Example of network representation with crizotinib. 1 molecule, multiple targets hit in multiple protein classes in one main family. **Figure S3.** Repartition of the number of targets for each scaffold, with the repartition curve in red. **Figure S4.** Overview of the 43th pareto front selection between the maximization of the different biological profiles (x axis) and the average number of times a UI is hit (y axis). Each iteration is of a different colour, each point equal 1 out of the 5000 molecules. **Figure S5.** Bar chart of the number of UI targeted by the final selection of molecules.

Additional file 2: Table S1. List of 5100 compounds bioactives on proteins. The compounds are encoded with a ChEMBLID, InChiKey and SMILES code.

Acknowledgements

We would like to thank the doctoral school "Pierre Louis de santé publique" and the pharmaceutical company Servier for their support on this study. This study contributes to IdEx Université de Paris ANR-18-IDEX-0001.

Authors' contributions

Conceived and designed the experiments: OT, PD, AG, TD. Performed the experiments: BD, NC, BB AG, AC. Wrote the manuscript: BD, NC, OT. Review the manuscript: all. All authors read and approved the final manuscript.

Funding

The study has been funded by the doctoral school "Pierre Louis de santé publique", the pharmaceutical company Servier, the Université de Paris and INSERM.

Availability of data and materials

The chemical library, with ChEMBL ID, SMILES, InchiKey and bioactive proteins associated, is available on Additional file. The code to reproduce the work is available on GitHub at this link: bit.ly/3Bs1w3u.

Declarations

Competing interests

The authors declare no competing interests.

Author details

[1]Université de Paris, INSERM U1133, CNRS UMR8251, 75006 Paris, France. [2]Institut de Recherche Servier, 125 Chemin de Ronde, 78290 Croissy-sur-Seine, France. [3]Université de Paris, INSERM UMR S-1124, 75006 Paris, France.

Received: 14 August 2021 Accepted: 6 November 2021
Published online: 24 November 2021

References

1. Paolini GV, Shapland RH, van Hoorn WP, Mason JS, Hopkins AL (2006) Global mapping of pharmacological space. Nat Biotechnol 24(7):805–815. https://doi.org/10.1038/nbt1228

2. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J (2014) Clinical development success rates for investigational drugs. Nat Biotechnol 32(1):40–51. https://doi.org/10.1038/nbt.2786
3. Hopkins AL (2008) Network pharmacology: the next paradigm in drug discovery. Nat Chem Biol 4(11):682–690. https://doi.org/10.1038/nchembio.118
4. Kola I, Landis J (2004) Can the pharmaceutical industry reduce attrition rates? Nat Rev Drug Discov 3(8):711–715. https://doi.org/10.1038/nrd1470
5. Chaudhari R, Fong LW, Tan Z, Huang B, Zhang S (2020) An up-to-date overview of computational polypharmacology in modern drug discovery. Expert Opin Drug Discov 15(9):1025–1044. https://doi.org/10.1080/17460441.2020.1767063
6. Heilker R, Wolff M, Tautermann CS, Bieler M (2009) G-protein-coupled receptor-focused drug discovery using a target class platform approach. Drug Discov Today 14(5–6):231–240. https://doi.org/10.1016/j.drudis.2008.11.011
7. Bosc N, Muller C, Hoffer L, Lagorce D, Bourg S et al (2020) Fr-PPIChem: an academic compound library dedicated to protein-protein interactions. ACS Chem Biol 15(6):1566–1574. https://doi.org/10.1021/acschembio.0c00179
8. Rognan D (2007) Chemogenomic approaches to rational drug design. Br J Pharmacol 152:38–52. https://doi.org/10.1038/sj.bjp.0707308
9. Keiser M, Setola V, Irwin J et al (2009) Predicting new molecular targets for known drugs. Nature 462:175–181. https://doi.org/10.1038/nature08506
10. Ni E, Kwon E, Young LM, Felsovalyi K, Fuller J (2020) How polypharmacologic is each chemogenomics library? Future Drug Discov 2(1):FDD26. https://doi.org/10.4155/fdd-2019-0032
11. Finan C, Gaulton A, Kruger FA, Lumbers RT, Shah T et al (2017) The druggable genome and support for target identification and validation in drug development. Sci Transl Med. https://doi.org/10.1126/scitranslmed.aag1166
12. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN et al (2018) Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov 7(5):317–332. https://doi.org/10.1038/nrd.2018.14
13. Gaspar H, Hübel C, Breen G (2019) Drug Targetor: a web interface to investigate the human druggome for over 500 phenotypes. Bioinformatics 35(14):2515–2517. https://doi.org/10.1093/bioinformatics/bty982
14. Kringelum J, Kjaerulff SK, Brunak S, Lund O, Oprea TI et al (2016) ChemProt-3.0: a global chemical biology diseases mapping. Database 2016:bav123. https://doi.org/10.1093/database/bav123
15. Zahoranszky-Köhalmi G, Sheils T, Oprea TI (2020) SmartGraph: a network pharmacology investigation platform. J Cheminform 12:5. https://doi.org/10.1186/s13321-020-0409-9
16. Vermeulen R, Schymanski EL, Barabási AL, Miller GW (2020) The exposome and health: where chemistry meets biology. Science 367(6476):392–396. https://doi.org/10.1126/science.aay3164
17. Oprea TI, May EE, Leitão A, Tropsha A (2011) Computational systems chemical biology. Methods Mol Biol 672:459–488. https://doi.org/10.1007/978-1-60761-839-3_18
18. Boezio B, Audouze K, Ducrot P, Taboureau O (2017) Network-based approaches in pharmacology. Mol Inform 36(10):1700048. https://doi.org/10.1002/minf.201700048
19. Dafniet B, Cerisier N, Audouze K, Taboureau O (2020) Drug-target-ADR network and possible implications of structural variants in adverse events. Mol Inform 39(12):2000116. https://doi.org/10.1002/minf.202000116
20. Jones LH, Bunnage ME (2017) Applications of chemogenomic library screening in drug discovery. Nat Rev Drug Discov 16:285–296. https://doi.org/10.1038/nrd.2016.244
21. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M (2017) Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov 16(8):531–543. https://doi.org/10.1038/nrd.2017.111
22. Childers WE, Elokely KM, Abou-Gharbia M (2020) The resurrection of phenotypic drug discovery. ACS Med Chem Lett 11(10):1820–1828. https://doi.org/10.1021/acsmedchemlett.0c00006
23. Lin S, Schorpp K, Rothenaigner I, Hadian K (2020) Image-based high-content screening in drug discovery. Drug Discov Today 25(8):1348–1361. https://doi.org/10.1016/j.drudis.2020.06.001
24. Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE (2021) Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov 20(2):145–159. https://doi.org/10.1038/s41573-020-00117-w
25. Méndez-Lucio O, Baillif B, Clevert DA, Rouquié D, Wichard J (2020) De novo generation of hit-like molecules from gene expression signatures using artificial intelligence. Nat Commun 11:10. https://doi.org/10.1038/s41467-019-13807-w
26. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D940. https://doi.org/10.1093/nar/gky1075
27. Bray MA, Singh S, Han H, Davis CT, Borgeson B et al (2016) Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc 11:1757–1774. https://doi.org/10.1038/nprot.2016.105
28. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J et al (2017) The ChEMBL database in 2017. Nucleic Acids Res 45(D1):D945–D954. https://doi.org/10.1093/nar/gkw1074
29. Kanehisa M, Goto S (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28(1):27–30. https://doi.org/10.1093/nar/28.1.27
30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H et al (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29. https://doi.org/10.1038/75556
31. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M et al (2018) Human disease ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Res 47(D1):D955–D962. https://doi.org/10.1093/nar/gky1032
32. Ljosa V, Sokolnicki KL, Carpenter AE (2012) Annotated high-throughput microscopy image sets for validation. Nat Methods 9(7):637. https://doi.org/10.1038/nmeth.2083
33. Kamentsky L, Jones TR, Fraser A, Bray MA, Logan DJ et al (2011) Improved structure, function, and compatibility for Cell Profiler: modular high-throughput image analysis software. Bioinformatics 27(8):1179–1180. https://doi.org/10.1093/bioinformatics/btr095
34. Bray N, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34:525–527. https://doi.org/10.1038/nbt.3519
35. Schäfer T, Kriege N, Humbeck L, Klein K, Koch O et al (2017) Scaffold Hunter: a comprehensive visual analytics framework for drug discovery. J Cheminform 9:28. https://doi.org/10.1186/s13321-017-0213-3
36. Yu G, Wang L, Han Y, He Q (2012) ClusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16(5):284–287. https://doi.org/10.1089/omi.2011.0118
37. Yu G, Wang L, Yan G, He Q (2015) DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. Bioinformatics 31(4):608–609. https://doi.org/10.1093/bioinformatics/btu684
38. Carlson M (2019) org.Hs.eg.db: genome wide annotation for human. R package version 3.8.2. Springer, Berlin. https://doi.org/10.18129/B9.bioc.org.Hs.eg.db
39. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1):D480–D489. https://doi.org/10.1093/nar/gkaa1100
40. Deb K, Agrawal S, Pratap A, Meyarivan T (2000) A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization. In: Schoenauer M et al (eds) Parallel problem solving from nature PPSN VI. PPSN. Lecture notes in computer science, vol 1917. Springer, Berlin. https://doi.org/10.1007/3-540-45356-3_83
41. Kim S, Chen J, Cheng T, Gindulyte A, He J et al (2019) PubChem 2019 update: improved access to chemical data. Nucleic Acids Res 47(D1):D1102-1109. https://doi.org/10.1093/nar/gky1033
42. Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL et al (2019) DrugCentral 2018: an update. Nucleic Acids Res 47:D963–D970. https://doi.org/10.1093/nar/gky963
43. Stathias V, Koleti A, Vidovic D, Cooper DJ, Jagodnik KM et al (2018) Sustainable data and metadata management at the BD2K-LINCS Data Coordination and Integration Center. Sci Data 5:180117. https://doi.org/10.1038/sdata.2018.117
44. Iwata M, Yamanishi Y (2019) The use of large-scale chemically-induced transcriptome data acquired from LINCS to study small molecules. Methods Mol Biol 1888:189–203. https://doi.org/10.1007/978-1-4939-8891-4_11

45. Lee H, Kim W (2019) Comparison of target features for predicting drug-target interactions by deep neural network based on large-scale drug-induced transcriptome data. Pharmaceutics 11(8):377. https://doi.org/10.3390/pharmaceutics11080377
46. Iskar M, Campillos M, Kuhn M, Jensen LJ, van Noort V (2010) Drug-induced regulation of target expression. PLoS Comput Biol 6(9):e1000925. https://doi.org/10.1371/journal.pcbi.1000925
47. Shelat A, Guy RK (2007) Scaffold composition and biological relevance of screening libraries. Nat Chem Biol 2007(3):442–446. https://doi.org/10.1038/nchembio0807-442
48. Hu Y, Stumpfe D, Bajorath J (2016) Computational exploration of molecular scaffolds in medicinal chemistry. J Med Chem 59:4062–4076. https://doi.org/10.1021/acs.jmedchem.5b01746
49. Kruger F, Stiefl N, Landrum GA (2020) rdScaffoldNetwork: the scaffold network implementation in RDKit. J Chem Inf Model 60:3331–3335. https://doi.org/10.1021/acs.jcim.0c00296
50. Scott OB, Chan WE (2020) ScaffoldGraph: an open-source library for the generation and analysis of molecular scaffold networks and scaffold trees. Bioinformatics 36:3930–3931. https://doi.org/10.1093/bioinformatics/btaa219
51. Hu H, Bajorath J (2020) Simplified activity cliff network representations with high interpretability and immediate access to SAR information. J Comput Aided Mol Des 34:943–952. https://doi.org/10.1007/s10822-020-00319-9
52. Jones L, Bunnage M (2017) Applications of chemogenomic library screening in drug discovery. Nat Rev Drug Discov 16:285–296. https://doi.org/10.1038/nrd.2016.244

Publisher's Note

SUPPLEMENTARY DATA

Figure S1: The "proteinClass" node (in yellow) Serine protease is a level 1 protein class, and the node Protease is a level 2 protein class for the "UniprotInter" node Serine protease hepsin (colored in blue). They are linked by a relationship *memberOf* which indicates their belonging to a specific family.

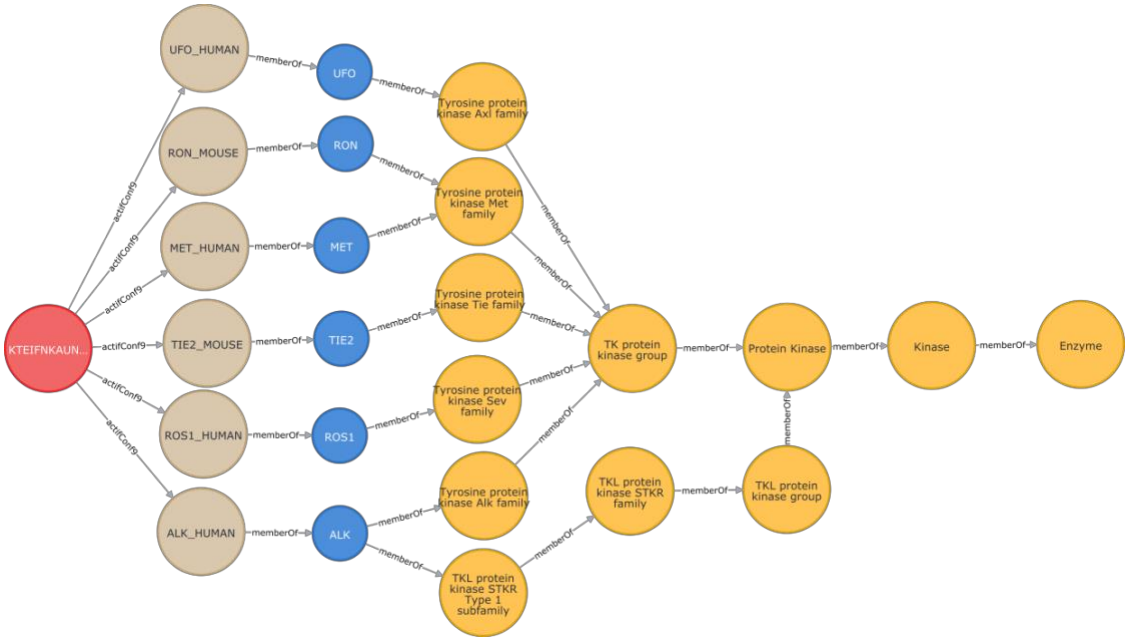


Figure S2: Example of network representation with crizotinib. 1 molecule, multiple targets hit in multiple protein classes in one main family.

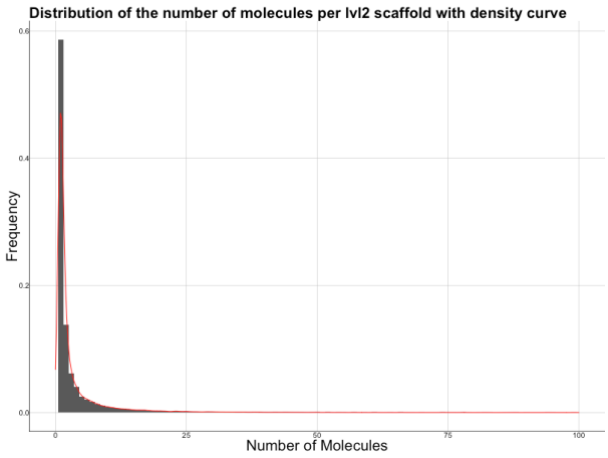


Figure S3: Repartition of the number of targets for each scaffold, with the repartition curve in red.

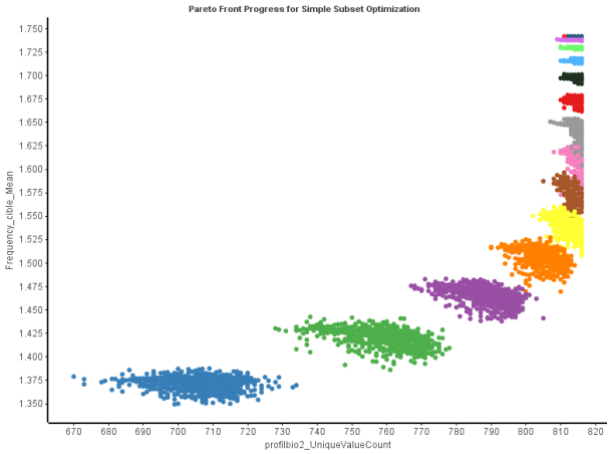


Figure S4: Overview of the 43th pareto front selection between the maximization of the different biological profiles (x axis) and the average number of times a UI is hit (y axis). Each iteration is of a different colour, each point equal 1 out of the 5000 molecules.

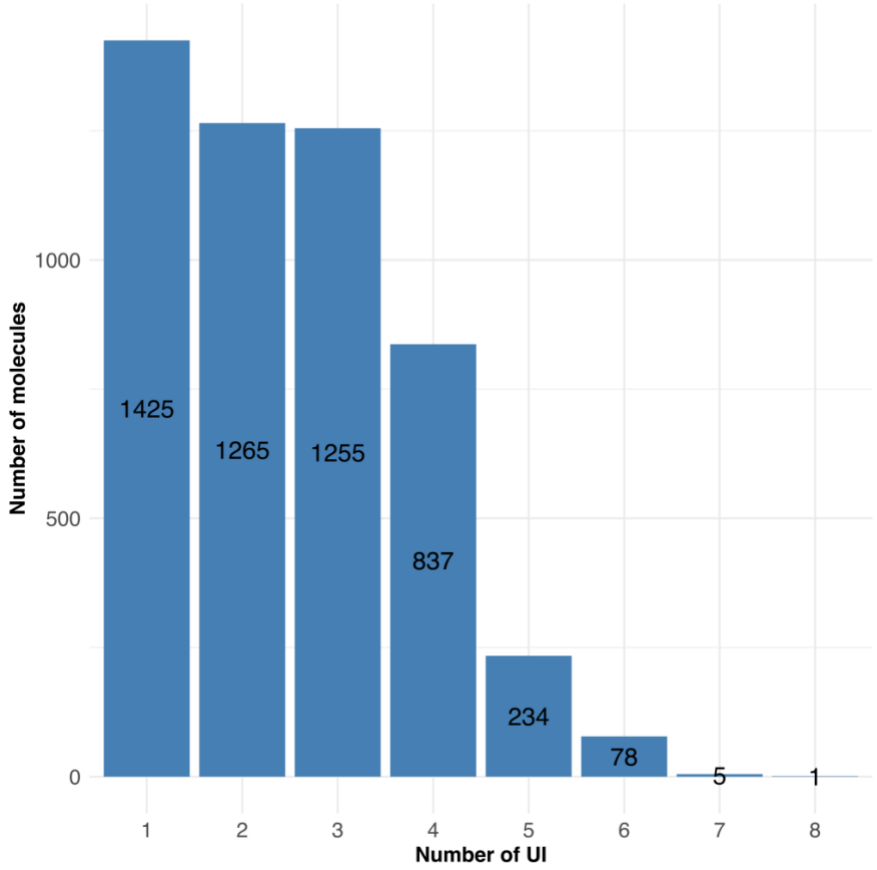


Figure S5: Bar chart of the number of UI targeted by the final selection of molecules

chEMBLID	InChIKey	SMILES	Target
CHEMBL3235493	GTMRUYCIJSNXGB-BJUDXGSMSA-N	[11CH3]N1CC2CN(C3=CC=C(C4=CC=CC=C4)N=N3)CC2C1	ACHA7_RAT
CHEMBL520970	IGIFNSQJPOGUFY-BJUDXGSMSA-N	[11CH3]OC1=CC=C2CN(CCCC3=CNC4=CC=C(I)C=C34)CCC2=C1	SC6A4_HUMAN;SC6A2_HUMAN;5HT1A_RAT
CHEMBL3598052	ZWHNWMJLCJLVSY-GKOSEXJESA-N	[2H]C([2H])([2H])N(C(=O)C1=C(Cl)C=CC=C1Cl)C1=CC=C(C2=CC(N=C(C)O)=NN2C(C)C)C=C1N1CC2CC2C1	NR1I2_HUMAN;RORB_HUMAN;RORG_MOUSE;RORG_HUMAN
CHEMBL3598048	JGWLJYUWJAGUIP-GKOSEXJESA-N	[2H]C([2H])([2H])N(C(=O)C1=C(F)C=CC=C1Cl)C1=CC=C(C2=CC(N=C(C)O)=NN2C(C)C)C=C1N1CCCC1	NR1I2_HUMAN;RORB_HUMAN;RORG_MOUSE;RORG_HUMAN
CHEMBL3598049	CVPCMJOHHJOBAS-GKOSEXJESA-N	[2H]C([2H])([2H])N(C(=O)C1=C(F)C=CC=C1Cl)C1=CC=C(C2=CC(N=C(C)O)=NN2C(C)C)C=C1N1CCCCC1	RORB_HUMAN;RORG_HUMAN;RORG_MOUSE
CHEMBL3598075	UZBUSCTWIRRONO-HPRDVNIFSA-N	[2H]C([2H])([2H])N(C(=O)C1=C(F)C=CC=C1Cl)C1=CC=C(C2=CC(N=C(C)O)=NN2C(C)C2CC2)C=C1N1CC2CC2C1	NR1I2_HUMAN;RORG_HUMAN
CHEMBL3598070	HLRDMFKUBGUXIF-BMSJAHLVSA-N	[2H]C([2H])([2H])N(C(=O)C1=C(F)C=CC=C1Cl)C1=CC=C(C2=CC(N=C(C)O)=NN2C2CCCC2)C=C1N1CC2CC2C1	NR1I2_HUMAN;RORG_HUMAN;RORG_MOUSE
CHEMBL3084930	KSYXCGBEMVAEDX-GPOLMCQNSA-N	[N-]=[N+]=NC1=CC=C(CCCN2[C@H]3CC[C@@H]2C[C@H](OC(C2=CC=C(F)C=C2)C2=CC=C(F)C=C2)C3)C=C1I	SC6A3_RAT
CHEMBL1163419	LQZBDVDATBCNNN-UHEGPQQHSA-N	[NH-]C1=NC=NC2=C1N=CN2[C@@H]1O[C@H](COP(=O)(O)OP(=O)(O)OP(=O)(O)O)[C@H]2OC3(O[C@H]21)C([N+](=O)[O-])=CC(=[N+]([O-])O)C=C3[N+](=O)[O-]	P2RX4_HUMAN
CHEMBL74283	CETMNKCDAHCKRC-UHFFFAOYSA-N	BrC1=C(NC2=NCCN2)C=CC2=C1NCCN2	ADA2B_HUMAN;ADA2C_HUMAN;ADA2A_HUMAN
CHEMBL2325091	WRQKYTMPGPMLPI-UHFFFAOYSA-N	BrC1=CC=C(C2=NN(C3=CC=CC=C3)C(C3=CC=CC4=CC=CC=C43)C2)C=C1	EGFR_HUMAN
CHEMBL3622097	MLPGCAWHLVVATJ-UHFFFAOYSA-N	BrC1=CC=C(NCCN2CCN(CCC3=CNC4=CC=CC=C34)CC2)C=C1	DRD3_HUMAN;DRD4_HUMAN;DRD1_HUMAN;DRD2_HUMAN

CHEMBL177293	SNWLVVNOSVKKRA-UXBLZVDNSA-N	BrC1=CC=C2C(=C1)CC/C2=C\C1=CN=CN1	C11B2_HUMAN;C11B1_HUMAN;CP19A_HUMAN
CHEMBL2164560	YWSNJWSAQDJCBJ-UHFFFAOYSA-N	BrC1=CC=CC(C2=NC(C3=CC=CC=N3)=NO2)=C1	GRM5_HUMAN
CHEMBL486511	HYLBWYXHCHUAGL-UFIFRZAQSA-N	BrC1=CN=CC(N2C[C@@H]3CNC[C@@H]3C2)=C1	ACHA7_RAT;ACHA4_RAT
CHEMBL327626	GKWMWOXKMZWVQU-NBYYMMLRSA-N	BrC1=NOC(CO/N=C2\CN3CCC2C3)=C1	ACM1_HUMAN
CHEMBL574719	PBNNHHJLMBMBNF-UHFFFAOYSA-N	C(#CC1=CC=C(CN2CCC(CC3=CC=CC=C3)CC2)C=C1)CCN1CCCCC1	HRH3_HUMAN
CHEMBL355656	HAQJDPAHYIINDW-YUUWEXFGSA-N	C[C@](CC1=CNC2=CC=CC=C12)(N=C(O)OC1C2CC3CC(C2)CC1C3)C(O)=N[C@@H](CN=C(O)CCC(=O)O)CC1=CC=CC=C1	CCKAR_RAT;GASR_MOUSE
CHEMBL3353541	MPMKMQHJHDHPBE-RUZDIDTESA-N	C[C@]1(C(=O)N(CCCC(=O)O)CC2=CC=CC(Cl)=C2)CCN1C(=O)C1=CSC2=CC=CC=C21	FFAR2_HUMAN
CHEMBL3359926	QASDPHCNHAXDKX-IUODEOHRSA-N	C[C@]1(C#N)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL2347215	DVMUZHLUMHPCGZ-QGZVFWFLSA-N	C[C@]1(C2=CC(NC(=O)C3=CC=C(C#N)C=N3)=CC=C2F)NC(=N)OCC1(F)F	BACE2_HUMAN;BACE1_HUMAN
CHEMBL3127490	HZWLXTGQNYLLDT-LWQYYNMXSA-N	C[C@]1(C2=CC=CC=C2)C[C@@H](C2=CC(N=C(O)C3CC3)=CC(C3=CC=C(C(=N)O)C=C3C(=O)O)=C2)NC2=CC=C(C(=N)N)C=C21	FA7_HUMAN;FA10_HUMAN;FA11_HUMAN
CHEMBL2323961	POSSBUJNCSVZHA-DQEYMECFSA-N	C[C@]1(CN2C=NC3=CC=C(C#N)C=C32)CCC[C@@]2(CN(CC3=CC=CC=C3F)C(=O)O2)C1	TRPV4_HUMAN
CHEMBL3359924	GATQRZORHMQKNV-IUODEOHRSA-N	C[C@]1(CO)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL3648214	OTNFBGWJFGXRRO-UFHPHHKVSA-N	C[C@]1(COC2=CC=C(Cl)C=N2)CN(C(=O)C2CCN(C3=CC=C(Cl)C=N3)CC2)C[C@@H]1C1=CC=C(Cl)C=C1	NK3R_HUMAN
CHEMBL3359927	UPBRYBJZTWHUQK-RISCZKNCSA-N	C[C@]1(F)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL166444	CCCIJQPRIXGQOE-XWSJACJDSA-N	C[C@]1(O)CC[C@H]2[C@@H]3CCC4=CC(=O)CCC4=C3C=C[C@@]21C	ANDR_RAT
CHEMBL235871	IHBSVQHHJXJVFU-IZHYOYPJSA-N	C[C@]1(O)CC[C@H]2[C@H]3[C@H](CC[C@@]21C)[C@@]1(C)CCC(=O)C=C1C[C@H]3CCCCC1=CC=CC(OCCCC(=O)O)=C1	ANDR_HUMAN
CHEMBL1255771	DLWKVBPPVYZEML-ZIAGYGMSSA-N	C[C@]12C[C@@]1(C1=CN=CN1)CC1=CC=CC=C12	ADA2B_HUMAN;ADA2C_HUMAN;ADA1B_HUMAN;ADA2A_HUMAN
CHEMBL2172650	FCFPLTRQOYJAKG-RGQDEETRSA-N	C[C@]12CC[C@@H]3[C@H]4CCC(C(=O)O)=CC4=CC[C@H]3[C@@H]1CCC2=O	S5A2_HUMAN
CHEMBL2105738	PAFKTGFSEFKSQG-PAASFTFBSA-N	C[C@]12CC[C@H]3[C@@H](CC=C4C[C@@H](O)CC[C@@]43C)[C@@H]1CC=C2N1C=NC2=CC=CC=C21	CP17A_HUMAN;ANDR_HUMAN
CHEMBL371376	DMJGJEOQFVLVQU-PUHATCMVSA-N	C[C@]12CC3=C(C=C1CCC[C@@H]2[C@@H](O)C1=COC=C1)N(C1=CC=C(F)C=C1)N=C3	GCR_HUMAN;GCR_MOUSE
CHEMBL363179	GSBXDPQKEHNRQR-NPAAKHOSSA-N	C[C@]12CC3=C(C=C1CCC[C@@H]2[C@@H](O)C1=CSC2=CC=CC=C21)N(C1=CC=C(F)C=C1)N=C3	GCR_MOUSE;GCR_HUMAN
CHEMBL186948	MBOPYKZQVGYOJO-CLYCCHKQSA-N	C[C@]12CC3=C(C=C1CCC[C@@H]2C(O)C1=CC=C2C=CCC=CC2=C1)N(C1=CC=C(F)C=C1)N=C3	GCR_RAT
CHEMBL255521	OBJXJRIGGFJBSZ-HSZRJFAPSA-N	C[C@]12CC3=C(C=C1CCN(S(=O)(=O)C1=CC=C(F)C=C1)C2)N(C1=CC=C(F)C=C1)N=C3	GCR_HUMAN
CHEMBL419934	KSFRRPAHIDJCHM-XEYOUOGVSA-N	C[C@]12CCC3C(CC=C4CC(O)CC[C@@]43C)C1CCC2C1=NC=CN1	CP17A_HUMAN
CHEMBL78003	YIWNIULQOQYDHI-CEJNIXRPSA-N	C[C@]12CCC3C(CCC4=CC(=O)CC[C@@]43C)C1CC=C2N1C=CN=N1	CP17A_HUMAN;CP17A_RAT
CHEMBL575448	LQVXSNNAFNGRAH-QHCPKHFHSA-N	C[C@@]1(C(O)=NC2=CC=C(F)N=C2)CCCN1C1=NC(=NC2=NNC(C3CC3)=C2)C2=CC=CN2N1	IGF1R_MOUSE;AKT1_MOUSE
CHEMBL3352892	QASDPHCNHAXDKX-DOMZBBRYSA-N	C[C@@]1(C#N)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN

CHEMBL3359925	GATQRZORHMQKNV-DOMZBBRYSA-N	C[C@@]1(CO)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL3359928	UPBRYBJZTWHUQK-SMDDNHRTSA-N	C[C@@]1(F)CCC[C@@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL3415630	JMBODOHYRWZQEW-GOSISDBHSA-N	C[C@@]1(NC2=CC=C(C(=O)NO)C=C2)CCN(C2=CC=C(Cl)C=C2)C1=O	HDAC6_HUMAN;HDAC8_HUMAN
CHEMBL3359923	QJPWFJUJHXVLEI-BXUZGUMPSA-N	C[C@@]1(O)CCC[C@H]1NC1=C(C(=N)O)C=NN2C=CC=C12	JAK2_HUMAN;TYK2_HUMAN;JAK1_HUMAN;JAK3_HUMAN
CHEMBL3317811	MYZPAEFNFJXMAN-VCQYNLKMSA-N	C[C@@]12CSC(=N1)C1=CSC(=N1)CN=C(O)C[C@@H](/C=C/CCS)OC(=O)[C@H](CC1=CC=C(O)C=C1)N=C2O	HDAC1_HUMAN;HDAC3_HUMAN;HDAC2_HUMAN
CHEMBL3317812	BTCGBXANODJFOB-CIDTVBPUSA-N	C[C@@]12CSC(=N1)C1=CSC(=N1)CN=C(O)C[C@@H](/C=C/CCS)OC(=O)[C@H](CC1=CN=CN1)N=C2O	HDAC1_HUMAN;HDAC3_HUMAN;HDAC2_HUMAN
CHEMBL1683452	JAPLGNWCNXCEJT-XDHUDOTRSA-N	C[C@@H](C[C@@](C)(CS(=O)(=O)N1CCC(CCC2=CC(F)=CC=C2Cl)CC1)N(O)C=O)C1=NC=C(F)C=N1	MMP13_HUMAN;MMP14_HUMAN;MMP2_HUMAN
CHEMBL1683449	KFGJNQLPIZMSFB-XDHUDOTRSA-N	C[C@@H](C[C@@](C)(CS(=O)(=O)N1CCC(CCC2=CC=C(C(F)(F)F)C=C2Cl)CC1)N(O)C=O)C1=NC=C(F)C=N1	MMP13_HUMAN;MMP14_HUMAN;MMP2_HUMAN
CHEMBL206501	JHIFTPQOVZRODM-ZYJMRSDMSA-N	C[C@@H](C1=CC=C(C2=CN(C)C(=O)C=C2)C=C1)[C@H](N)C(=O)N1CCC(F)(F)C1	DPP4_HUMAN;DPP2_HUMAN
CHEMBL2431822	SQYNPZYBHGTSOY-ZDUSSCGKSA-N	C[C@@H](C1=CC=C2N=CC=CC2=C1)C1=NN=C2C=CC(C3=CC=C(F)C(F)=C3)=NN21	PDE3B_HUMAN;PDE1C_HUMAN;PDE1A_HUMAN;PDE10_HUMAN;MET_HUMAN
CHEMBL2431826	QRWHTFXNHZLJAL-HNNXBMFYSA-N	C[C@@H](C1=CC=C2N=CC=CC2=C1)C1=NN=C2C=CC(C3=CC=CC(C#N)=C3)=NN21	PDE3B_HUMAN;PDE1A_HUMAN;PDE11_HUMAN;PDE10_HUMAN;MET_HUMAN

Table S1: 50 first lines of the supplementary table titled “List of 5100 compounds bioactives on proteins. The compounds are encoded with a ChEMBLID, InChiKey and SMILES code.”. The entire file is available at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8611952/

2.3 Conclusion

Dans cet article, nous avons relié différentes sources d'information biologique permettant une caractérisation des perturbations majeures d'un ensemble de composés chimique au niveau moléculaires et cellulaires et de leur possible implication dans le développement d'un ou plusieurs phénotypes. De plus la création d'une librairie de 5100 composés permet de couvrir un espace chemo-génomique important permettant de définir un grand nombre d'interactions composé-protéine rapidement analysable. Enfin, ce travail fournit une base pour des approches d'identification de potentiels candidats médicaments et de leurs effets tant phénotypiques qu'au niveau des voies de signalisation.
Lors de l'application de l'optimisation pareto les profils biologiques ont pu être optimisés afin d'apporter un maximum d'information structurale sur les interactions composé-protéine. Néanmoins malgré ce paramètre le plus gros obstacle reste le nombre moyen de protéines interagissant avec un composé qui reste dépendant des *assays* et est relativement faible. L'intégration de nouvelles bases de données comme PubChem qui contient, entre autres, des données de criblage à haut débit peut être inclus pour apporter plus d'information. Des données transcriptomiques peuvent également être utilisées afin d'analyser des dérégulations de gènes causés par un composé. Enfin, concernant la similarité structurale des composés chimiques, nous nous sommes focalisés sur les 'charpentes' (*scaffolds*) des molécules. D'autres paramètres pourraient être pris en compte comme les propriétés physicochimiques ou les pharmacophores d'une molécule chimique comme proposé dans l'article de Métivier J.P., *et al*. 2020.

Concernant l'aspect technique de l'analyse, le code et les fichiers ont été déposés sur GitHub. Cette démarche s'inscrit dans les principes FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) proposé par l'Europe et qui donne accès à toutes les informations permettant de reproduire, modifier et/ou optimiser les résultats que nous avons obtenus. La seule composante aléatoire dans la création de notre librairie est le pareto qui utilise une composante d'algorithme génétique. Néanmoins cette composante a été « seedée » rendant les résultats reproductibles.

Chapitre 3 : Analyse des effets indésirables et prédiction des protéines responsables

Après avoir exploré l'effet de composés au niveau protéomique, phénotypique et leurs possibles implications sur les processus biologiques ou les maladies, l'étape suivante est de relier les protéines à des effets indésirables. Pour cela une deuxième base de données va être créée à partir de la *drugbank* et de *drugcentral*. L'intérêt de cette dernière va être l'inclusion de données provenant de la *Food and Drug Administration* (FDA) *Adverse Event Reporting System* (FAERS) qui encode 8501 effets indésirables reportés pour les médicaments avec leur identifiant *drugbank*, rendant possible le lien entre ces bases de données.

3.1 Analyses de graphe

Une fois le réseau construit la première approche a été d'analyser le graphe en utilisant des approches classiques comme la centralité, mesure permettant de déterminer à quel point un nœud est connecté dans le réseau, et la *betweenness* permettant d'analyser l'importance d'un nœud de par le nombre de chemins les plus courts passant par celui-ci [Boezio B., *et al.* 2017] Cette dernière mesure est intéressante et peut servir par exemple à déterminer les protéines les plus influentes menant à un effet indésirable spécifique, par exemple la nausée, de par le nombre de composés ayant un nombre de liens minimums pour atteindre cet effet passant par une même protéine. Ces analyses vont permettre un premier niveau de lecture du graphe et d'en extraire les premières informations.

3.2 Fonction de score

La création de cette fonction va permettre de directement lier un effet indésirable à une liste de protéines. Lors de cette étude, des liens ont pu être fait grâce à des données expérimentales pour les relations médicament-effets indésirables et médicament-cibles. Une hypothèse a donc été faite en considérant que si un médicament possède des effets indésirables et interagit avec une cible, cette cible est liée à ces effets indésirables. Compte-tenu du nombre d'effets indésirables pour chaque médicament les cibles ne sont probablement pas responsables de chacun de ces effets. Pour analyser et préciser cette hypothèse une fonction de score a été créée. Les effets indésirables vont être représentés par la somme des coefficients des protéines reliées ; chaque coefficient sera calculé comme étant la somme, pour chaque médicament interagissant avec la protéine, de la fraction ayant pour dénominateur le nombre total de protéine interagissant avec ce médicament. Cela permettra de mettre un poids plus important aux protéines touchées par des médicaments qui seraient spécifiques à celles-ci, ou avec peu de cibles car ce sont elles qui auraient le plus de chance de causer cet effet indésirable.

3.3 *Copy Number Variation*

Une fois l'analyse des effets indésirables faite, le début de l'analyse de la variabilité génétique va conclure cet article avec l'analyse de possibles relations entre les effets indésirables et les CNVs. Ces mutations incluent deux types de modifications : la délétion, qui implique la suppression de séquences répétées du génome, et la duplication qui à l'inverse augmente le nombre de certaines répétitions de séquences génomiques [Shlien A., *et al.* 2009]. Lors de la traduction en protéine cela peut mener à une variation du nombre de protéines disponibles. Cela peut avoir un impact sur le dosage des médicaments si la protéine traduite fait partie de la voie de métabolisation de ceux-ci, pouvant mener à des effets indésirables car la dose prescrite ne correspond pas à ce type de profil génétique [Hauser A.S., *et al.* 2017 ; Hollox E.J., *et al.* 2021].

Full Paper

www.molinf.com

molecular informatics

DOI: 10.1002/minf.202000116

Drug-target-ADR Network and Possible Implications of Structural Variants in Adverse Events

Bryan Dafniet,[a] Natacha Cerisier,[a] Karine Audouze,[b] and Olivier Taboureau*[a]

Abstract: Adverse drug reactions (ADRs) are of major concern in drug safety. However, due to the biological complexity of human systems, understanding the underlying mechanisms involved in development of ADRs remains a challenging task. Here, we applied network sciences to analyze a tripartite network between 1000 drugs, 1407 targets, and 6164 ADRs. It allowed us to suggest drug targets susceptible to be associated to ADRs and organs, based on the system organ class (SOC). Furthermore, a score was developed to determine the contribution of a set of proteins to ADRs. Finally, we identified proteins that might increase the susceptibility of genes to ADRs, on the basis of knowledge about genomic structural variation in genes encoding proteins targeted by drugs. Such analysis should pave the way to individualize drug therapy and precision medicine.

Keywords: Network sciences · drug safety · pharmacology · adverse drug reactions · precision medicine

1 Introduction

The occurrence of adverse drug reactions (ADRs) is an important concern for the health of patients as well as for the healthcare sector as it costs several billion dollars every year. ADRs account for 5% to 7% of all hospitalized individuals and represent the fifth most common cause of death in hospitals.[1-3] ADR is defined as a noxious and unintended response to drug therapy at a normal dose. Several factors, including polypharmacy, age, type of prescribed medicines, and genomic variations, might influence its occurrence.[4] For example, drug-drug interactions (DDIs) from combined medication have been reported to account for 30% of all ADRs.[5] In addition, genetic factors and structural variations may predispose a person to some ADRs. It has been reported that pharmacogenomics accounts for about 80% of the variability in drug efficacy and safety.[6] Therefore, identifying the underlying mechanisms of these ADRs is necessary to limit their severity and mortality and to improve drug safety.

As a large number of drugs interact with more than a single target, perturbated protein interaction network system-wide approaches may be more suitable to capture the effects of drugs on the human body.[7–8] A variety of methods linking ADRs to drug actions have been proposed. One common approach is to correlate the chemical structure of a drug compound with a particular set of ADRs.[9–11] However, chemically unrelated structures might share ADRs, targeting similar off-targets or pathways. To overcome this limitation, methods based on target profiling similarity and side effect similarity have been investigated.[12–13] Campillos et al.[14] proposed a method based on side effect similarity to associate drug pairs with common protein targets, whereas Fliri et al.[15] adopted a systems biology approach, showing that drugs with similar bioactivity profiles tend to cause similar side effects. In another study, Lounkine et al. developed an enrichment score that associates targets with ADRs based on the likelihood of the target-ADR pairs co-occurring as compared to random associations.[16] Garcia-Serna and Mestres[17] assigned a strength score between drug and secondary effects (SE) depending on the reporting frequency among the five SE sources used in their study, where "1" denoted presence of SE in all sources, and "0.2" denoted presence in only one source.

More recently, systems pharmacology approaches, combining network sciences and chemical biology, have been developed to predict and understand ADRs. Network sciences allow the integration of heterogeneous data sources and the quantification of their interactions.[18–19] Several studies have reported new insights on ADRs based on network representation and analysis. A bipartite graph and supervised machine learning were developed to predict new drug-protein pairs by combining chemical space (chemical structure similarity), genomic space (amino acid similarity), and pharmacological

[a] *B. Dafniet, N. Cerisier, O. Taboureau*
Université de Paris, INSERM U1133, CNRS UMR 8251, 75006 Paris, France
E-mail: Olivier.taboureau@u-paris.fr
[b] *K. Audouze*
Université de Paris, T3S, INSERM UMR S-1124, 75006 Paris, France
phone: +331 57278279
fax: +33157278372

Supporting information for this article is available on the WWW under https://doi.org/10.1002/minf.202000116

effect.[20] Based on the topological properties of protein-protein interactions, network biology methods were applied to identify proteins involved in specific ADRs.[21–22] Chen et al. performed an ADR-protein network and identified 41 network modules related to specific ADRs.[23] Oprea et al. included tissue information on a drug-target-SE network and reported that a drug is more likely to cause SE in the organ/tissue where it is more likely to accumulate.[24] Recently, a combined deep learning and biomedical tripartite network approach to predict drug-ADR associations was reported.[25]

In the present study, we developed a network biology model that complements the ones previously mentioned to identify and prioritize drug targets involved in specific ADRs as well as in more general terms, based on the system organ class (SOC) implemented in the Medical Dictionary for Regulatory Activities (MedDRA).[26] In addition, we included genomic structural variation (SV) information in the models to determine drug target associations contributing to the highest ADR susceptibility in individuals.

2 Material and Methods

2.1 Data

2.1.1 DrugBank

To build the network model, we used the DrugBank database v5.1.5.[27] It is a free online database with a wide range of information on drugs, notably drug-target relationships. Considering all drugs, and only human proteins, we collected 11,355 small molecules targeting 3510 proteins, reaching 24,579 drug-target interactions.

2.1.2 DrugCentral

DrugCentral is a drug information resource, which includes, among others, Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS).[28] The likelihood ratio test (LRT) for the safety signal detection method proposed by Huang *et al.* was used.[29] Any drug-ADR with a likelihood ratio (llr) superior to the likelihood ratio threshold was conserved in our analysis, giving us 92,794 associations with 8501 unique ADRs. Moreover, on the basis of the Medical Dictionary for Regulatory Activities (MedDRA), ADRs can be categorized using the System Organ Class (SOC) classification, which is the highest level in MedDRA hierarchy.[26] Therefore, all the compiled ADRs were categorized through the 27 SOCs, representing them at the level of organs and body systems, including other special categories (e.g., social factors, surgery, poison, and injury). The SOC abbreviations are described in Table S1 in the Supporting Information.

2.1.3 Database of Genomic Variants

For information on the genomic structural variation observed in the population, we used the Database of Genomic Variants (DGV).[30] DGV provides high-quality structural variations (SVs), defined as a region of DNA elements approximately 1 kb and larger and can include inversions and balanced translocations or genomic imbalances (insertions and deletions), commonly referred to as copy number variants (CNVs). The content of DGV represents SV identified in healthy control samples from large published cohorts and integrated by the DGV team. This database contained 8 million entries in 2019. We worked with the latest release available from the GRCh37(hg19) assembly of supporting variants section (http://dgv.tcag.ca/dgv/docs/GRCh37_hg19_supportingvariants_2020-02-25.txt). We extracted SVs with variant subtypes, including "deletion", "duplication", "loss" and "gain". SVs without frequency and gene information were removed. This leads to 83541 SVs with frequencies. For clarity, we combined deletion and loss under the term "loss" and duplication and gain to the term "gain".

2.2 Tools

2.2.1 Network Development & Representation

All compiled data are represented as a graph. A graph is a mathematical model consisting of a set of nodes defined by properties that are linked by relationships (edges). We used the Neo4j tool (www.neo4j.com), which is a high-performance NOSQL graphics database using Cypher-based query commands. Data collected from the DrugBank, DrugCentral, and DGV databases were integrated together into a network. First, the drug-target reported in DrugBank was considered. The SV information from the DGV was then connected to the targets, forming a drug-target-SV network. In parallel, the drug-ADR-SOC network was built from DrugCentral data. Finally, we merged both networks. As we have no information about specific target-ADRs, putative links were added to the drug-ADR associations, based on the assumption that if a drug has an ADR, then the protein targets may be causing it.

We deleted the nodes that either did not have any link, or did not have a link considering all integrated collected data (for instance, if a drug is connected to a target, but had no ADR linked to it, we would remove this specific molecule node). This step allowed to refine and reduce the compiled data, and to keep only a fully linked graph (Figure 1A). An example is also represented in Figure 1B with the drug bivaluridin

2.2.2 Networks Analysis

Several parameters were considered to analyze the three networks (Drug-ADR, Drug-Target and, Drug-Target-ADR). Among them, we investigated the following:

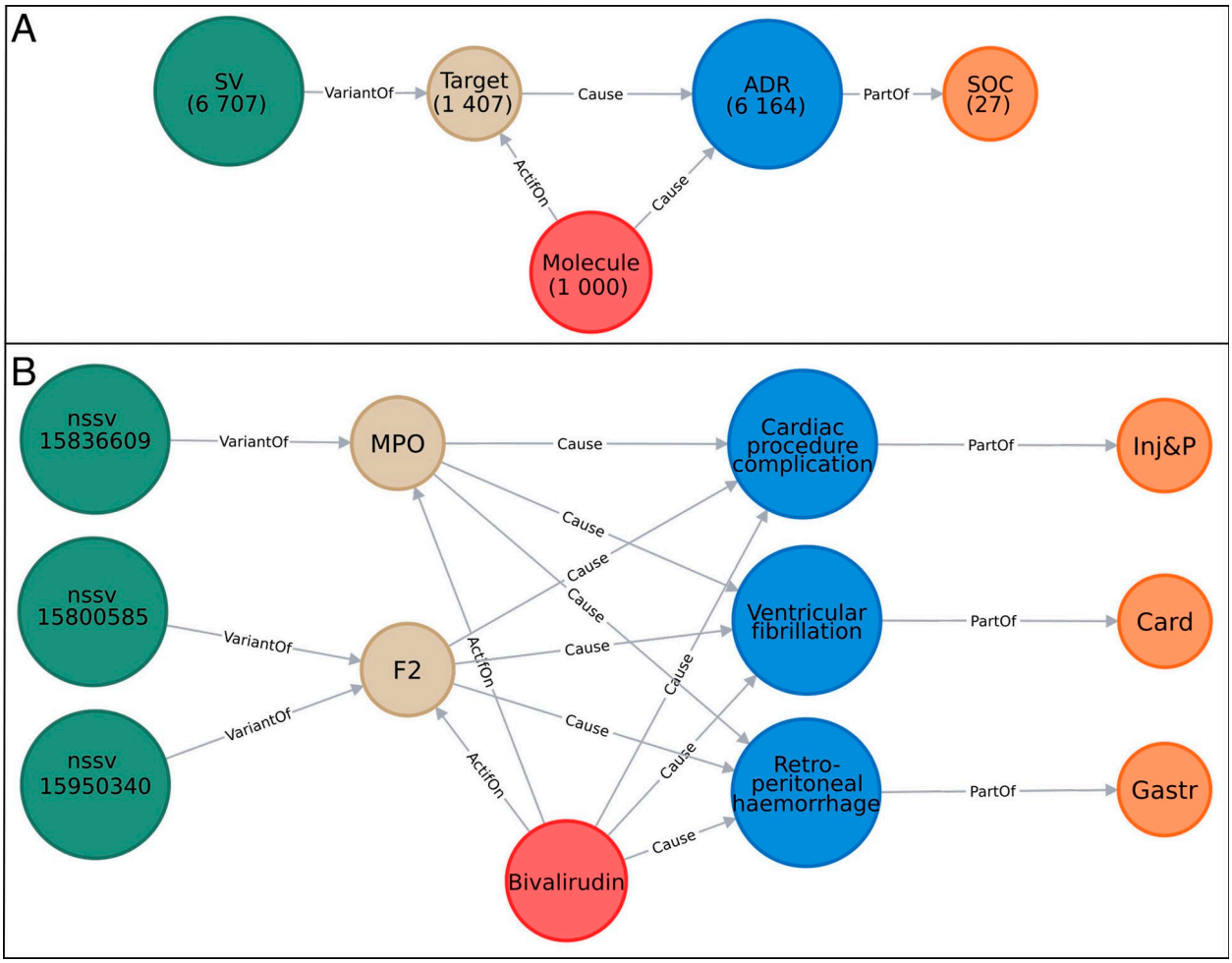


Figure 1. A) Overview of the integrative developed network. The nodes are represented by circles and colored according to their labels. The numbers of nodes in the entire network are written below their labels. B) Example of a part of the network. Here, 3 nodes "SV" are represented in dark green and named after their variant accession number, 2 nodes "Target" in beige named after their gene name, 1 node "Molecule" in red named after the molecule name, 3 nodes "ADR" in blue named after their ADR term, and 3 nodes "SOC" in orange named after their SOC abbreviation. The edges (in gray) show the relations between different nodes. This network was created using the Neo4j tool.

- *Network density* measures the fraction of edges in the network, compared to the theoretical maximum number of edges connecting each node.
- *Network centrality* is a measure of how central its most central node is in relation to all the other nodes.
- The *degree* of a node corresponds to the number of edges connected to a node.
- *Network heterogeneity* measures the variance of the connectivity distribution in a network.
- *Betweenness centrality* is the frequency with which the shortest paths between any pair of nodes pass through that node. Betweenness centrality estimates the global influence of a node in controlling the interaction between a pair of nodes passing through this node in the network. A node with high betweenness centrality is usually a node concentrated in a dense subnetwork with many connections to other nodes. Such a node has a larger impact on the information flow in the global network.[31] Betweenness centrality is usually associated with *closeness centrality*, measuring the importance of a node in a subnetwork. It assumes that the closer a node is to other nodes, the more likely it is to be included in the shortest paths.
- *Radiality* is a node centrality index. If the radiality is high, it means that, with respect to its diameter, the node is generally closer to the other nodes, whereas low radiality means that the node is peripheral.
- The *topological coefficient* is a relative measure of the extent to which a node shares its neighbors with other nodes. Nodes that have one or no neighbors are assigned a topological coefficient of zero.
- The *connectivity* of a node is the number of its neighbors. The *neighborhood connectivity* of a node n is defined as the average connectivity of all neighbors of n.

For, the drug-target-ADR network, the drug-ADR linkages were removed, to conserve only the Drug-Target-ADR-SOC connections in the network indices calculation.

All analyses were performed using R[32] and the Network Analysis plugin from Cytoscape (v.3.6.1) (www.cytoscape.org). Some figures required special R packages such as "circlize"[33] to produce the chord diagram, and "lattice"[34] and "ggplot2",[35] which produce various types of plots.

3 Results and Discussion

3.1 Drug-ADR and Drug-SOC Networks

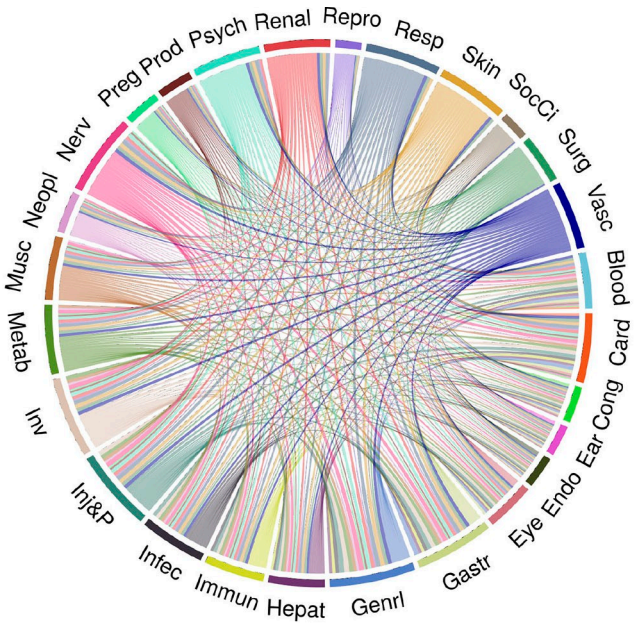
Based on the 1000 drugs from the Drugbank, we collected 6164 ADRs from DrugCentral. To analyze the multiple ADR-drug annotations, an undirected network-based model was developed. Although such a large model is difficult to visualize, network analysis allows the identification of some interesting features and modules related to the topology of the graphs. The obtained drug-ADR network was sparse, with a total of 7164 nodes (drugs and ADRs combined). It has a low network density (0.003), a network heterogeneity of 2.98 (tendency to contain hubs), an average number of 24 neighbors, and a network centrality of 0.21 (close to 0 when all nodes have the same centrality and close to 1 when one actor has the maximal centrality). 'Nausea', 'Vomiting' and 'Drug ineffective' are the ADRs with the highest closeness centrality (around 0.48), and radiality (around 0.86) i.e. they have the highest number of drugs connected (407, 402 and 390 respectively). Methotrexate (used in acute lymphoblastic leukemia, breast cancer, rheumatoid arthritis), alendronic acid (indicated for the treatment of osteoporosis), and prednisone (an anti-inflammatory or immunosuppressive drug derived from cortisone) show the highest number of links to ADRs (1561, 1422, and 1377, respectively). Interestingly, 155 drugs are associated with only one ADR (293 drugs have less than five ADRs), and 1626 ADRs are only related to one drug (3718 ADRs are connected to less than five drugs). However, a single ADR for a drug does not mean that this ADR appears only for this drug. For example, prucalopride, a drug indicated for the treatment of chronic idiopathic constipation, shows 'Vomiting' as unique ADR, even if 'Vomiting' is connected to 402 drugs. Similarly,"Agoraphobia" was the only ADR seen with the antidepressant paroxetine. However, this drug is annotated to 449 ADRs. An interesting feature is the topological coefficient (TC), which measures how much a node shares its neighbors with other nodes. For example, the ADR 'Mental disability' is indicated with two drugs and has a TC of 1, that is, the drugs aminophylline and theophylline are annotated for the same 55 ADRs (Neighborhood Connectivity score). Drugs sharing the same set of ADRs can be analogs – such as droxidopa and norepinephrine for the 'Tracheal atresia' ADR, mycophenolate mofetil and mycophenolic acid for 'Wound infection pseudomonas' ADR – but also structurally different, although indicated for the same treatment (fluorouracil and capecitabine for 'Tumor perforation' or atovaquone and proguanil for 'Plasmodium falciparum infection').

Based on the ADR-drug interaction pairs, it is possible to create an ADR-ADR network that reaches close to 5 million unique interactions between 6164 ADR terms. This network allows the identification of ADRs interrelated to other ADRs by large sets of drugs. This is the case, for example, with the ADRs 'Nausea' and 'Vomiting' sharing 360 drugs and 'Dizziness' and 'Nausea' sharing 288 drugs. Some ADRs are also related to common blood tests to evaluate liver problems such as increased alanine aminotransferase and increased aspartate aminotransferase, which share 219 drugs. Therefore, although many drugs have common general ADRs, drugs may also be related to specific ADRs. In the second step, each ADR has been annotated to one of the 27 SOCs defined in MedDra. Using this classification, we developed an SOC-SOC network to visualize the 27 SOC interactions by drugs (Figure 2). The larger the SOC, the more the drugs depict ADRs for this SOC. Similarly, the larger the edge between the two SOCs, the less the drugs have specific ADRs. We observed a fully connected network between the 27 SOCs. Globally, the 'General disorders and administration site conditions', the 'Injury, poisoning and procedural complications', and the 'Nervous System disorders' are the most targeted SOCs by drugs with 684, 653, and 616 drugs, respectively. 'Reproductive system and breast disorders', 'Endocrine disorders', and 'Ear and labyrinth disorders' are less impacted with 144, 156 and 170 drugs, respectively (Figure 2). The majority of the SOCs are highly connected to the 'General disorders and administration site conditions'. In addition, many compounds with'Nervous System disorders' also present 'Pregnancy, puerperium, and perinatal condition' (182 drugs). Similarly, 'Cardiac disorders' and 'Vascular disorders' (361 drugs) are linked as well as 'Gastrointestinal disorders' with 'Injury, poisoning, and procedural complications' (458 drugs). We can observe that drugs annotated to 'Infections and infestations' are also linked to 'Respiratory, thoracic, and mediastinal disorders', 'Gastrointestinal disorders', and 'Renal and urinary disorders', which might give some input about the comorbidities seen with the Covid-19.[36]

3.2 Drug-Target-ADR and Drug-Target-SOC Networks

After analyzing both drug-ADR and drug-SOC interactions, we included available drug target information from the DrugBank database in the previously developed network and performed a tripartite network analysis to assess drug targets that are potentially associated with an ADR or an SOC. The principle is the more drugs share common ADRs and proteins, the more the proteins are associated to these ADRs. By integrating the drug target information into the drug-ADR model, we built a network of 515959 interactions between drugs, targets, ADRs, and SOCs. The full list of the network scores obtained for each node is available in the Supporting Information (Table S2). The

A



B

SOC	Blood	Card	Cong	Ear	Endo	Eye	Gastr	Genrl	Hepat	Immun	Infec	Inj&P	Inv	Metab	Musc	Neopl	Nerv	Preg	Prod	Psych	Renal	Repro	Resp	Skin	SocCi	Surg	Vasc
Blood	361	243	144	126	123	180	310	331	256	226	289	295	311	277	240	196	296	127	115	233	268	100	292	272	85	187	255
Card		464	163	134	140	230	375	424	250	294	296	414	396	335	292	184	409	165	172	354	325	117	399	330	128	219	361
Cong			205	92	97	139	179	188	151	153	158	191	182	171	165	107	178	123	96	160	156	81	171	166	79	136	158
Ear				156	82	119	152	155	128	137	140	148	148	141	142	92	154	83	83	141	137	71	140	146	71	117	135
Endo					170	116	153	163	131	123	141	158	161	142	145	106	157	89	90	141	138	80	149	148	76	124	146
Eye						287	253	271	187	211	221	262	260	228	220	136	261	128	130	226	217	102	249	239	102	174	223
Gastr							562	520	304	323	368	458	476	403	354	220	477	173	181	379	365	127	428	404	132	251	373
Genrl								684	317	356	413	546	527	426	367	237	543	186	200	430	400	136	481	439	144	261	410
Hepat									335	238	266	288	305	276	249	175	293	147	117	248	272	95	282	284	97	187	253
Immun										407	271	342	330	279	266	150	327	162	152	290	273	102	318	314	110	199	283
Infec											428	362	377	323	289	216	357	150	149	284	307	109	339	331	111	230	297
Inj&P												653	469	392	355	214	507	205	198	433	372	132	453	397	144	253	395
Inv													583	407	350	223	489	180	186	386	375	128	437	410	138	259	392
Metab														458	306	208	397	161	160	327	333	114	353	347	120	226	337
Musc															381	183	353	153	157	307	282	117	316	318	124	219	301
Neopl																243	206	88	87	169	191	94	208	194	73	154	194
Nerv																	616	182	200	430	362	133	452	404	142	250	396
Preg																		212	95	164	153	68	175	173	76	141	150
Prod																			208	173	150	69	178	167	83	123	169
Psych																				487	308	115	381	340	136	214	327
Renal																					427	112	358	331	111	205	326
Repro																						144	118	121	61	97	119
Resp																							531	362	130	236	379
Skin																								478	125	229	330
SocCi																									144	107	125
Surg																										274	218
Vasc																											441

Figure 2. A) System Organ Class (SOC) network. Each colored segment (node) corresponds to one SOC and is named according to its SOC abbreviation name (Supporting information). The segment size depicts the number of drugs associated with their corresponding SOC. Each edge represents a drug that causes an ADR linked to their corresponding SOCs. The chord diagram was developed using the R package circlize. B) SOC association counting table. Each value represents the number of drugs common to both SOCs.

drug-target-ADR network is still sparse with a network density of 0.014 between the 8571 nodes and a similar network heterogeneity (2.43). The average number of neighbors (120) as well as the network centrality (0.66) is higher, meaning that some nodes (proteins) are more central than others (drugs and ADRs). Proteins targeted by drugs are essentially cytochromes and transporters, except for the serum albumin protein (ALB) (Table S3 in Supporting Information). As primary targets are drug-metabolizing enzymes (DMEs), these results are not surprising.[37–38] If we remove the cytochrome enzymes, proteins such as ALB, SLC22 A6, PTGS1, UGT2B7, UGT1 A9, and PTGS2 are highly related to 'Acute kidney injury' (Figure 3). Some of these proteins have been associated with this adverse effect in studies.[39–41] 'Somnolence' is highly related to ADRA1 A, but so far, no clear relationship has been reported in the literature. Finally, 'Toxicity to various agents' is highly associated with the human H1 receptor (HRH1). Although antihistamine drugs are associated with many ADRs, a recent study has reported that the polymorphism of HRH1 may be related to the severity of ADR, notably sedation.[42]

In the tripartite network, 84 drugs were annotated to a unique protein (449 drugs with less than five proteins). Through the drug-ADR relationships, it is possible to assume that some proteins have an impact on specific ADRs. For example, the 4-hydroxyphenylpyruvate dioxygenase protein (HPD) is the unique target of the drug nitisone in DrugBank and is linked to the ADR 'Amino acid level increased', 'Liver transplant' and 'Hepatocellular carcinoma'. The squalene

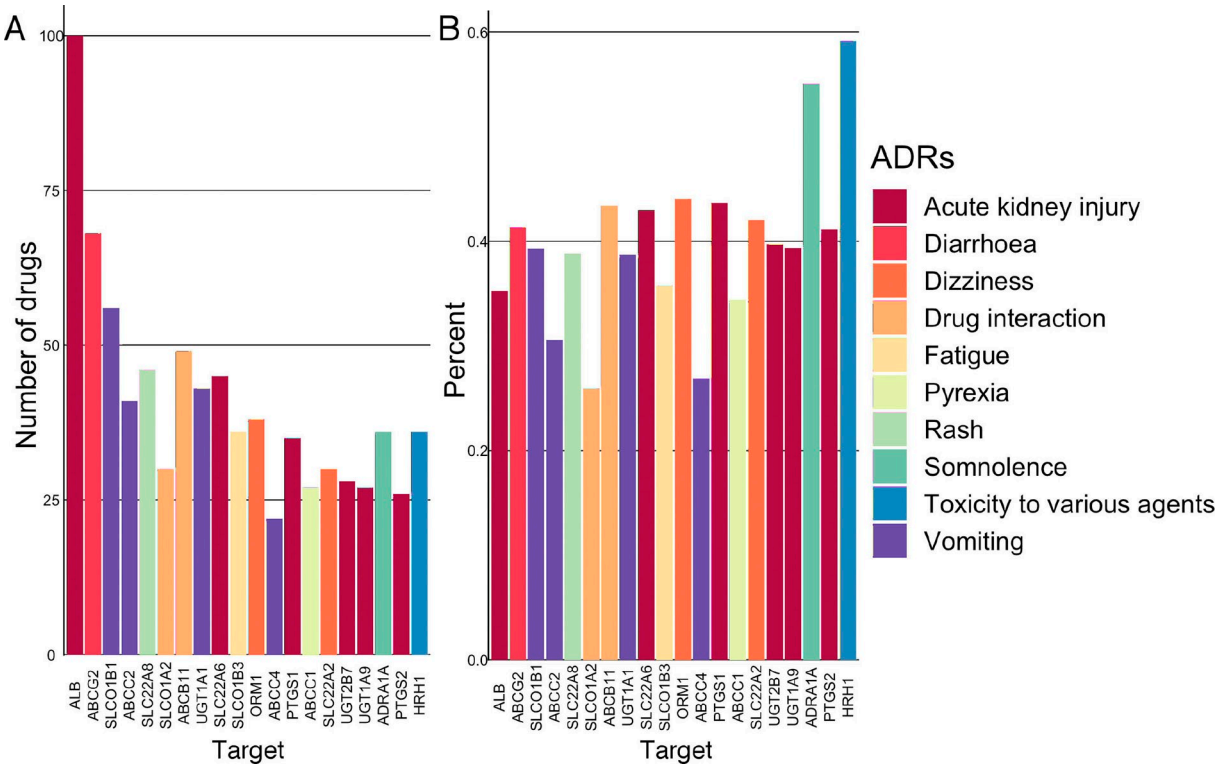


Figure 3. A) The 20 proteins associated with the highest number of drugs for an ADR. B) The proportion of these 20 proteins associated with an ADR among the other proteins linked to this ADR.

monooxygenase protein (SQLE) is linked to the drug naftifine and the ADR 'Blister', and the Oxytocin Receptor (OXTR) connect the drug carbetocin and the ADR 'Acute kidney injury'. Interestingly, 166 proteins (11% of the drug targets) connect one drug to only one ADR. Similarly, we can estimate whether these proteins are central in the linkage between drugs and ADRs. For example, the protein GNRHR2 is linked to the ADR 'Ovarian hyperstimulation syndrome' and the drug nafarelin (Figure 4A). GNRHR2 has a low Neighborhood Connectivity score (16.5), and its topological coefficient (TC) is one of the lowest (0.51). This means that GNRHR2 is relatively specific to this ADR and has a high possibility of being involved with this ADR. Protein DDAH2 has the highest Neighborhood Connectivity score (424), but a higher radiality (0.67) and a lower TC (0.50). This is due to DDAH2 being linked to citrulline that is annotated to 14 targets and the ADR "drug ineffective", which is linked to 834 proteins. This protein does not seem to be a central partner in this ADR.

In contrast, there is a set of 23 proteins (P5CR2, P3H2, PPIC, EPRS, P3H1, PYCR1, PPIA, P4HA1, L3HYPDH, P3H3, PPIG, PYCR2, SLC6 A14, PYCRL, PROSC, PPIF, SLC6 A7, PPIB, P4HA2, PARS2, PPIH, SLC16 A10, PRODH) that are linked to the nutraceutical Proline and the ADR 'Fetal growth restriction'.

With a TC of 0.64, these nodes are more specific to this ADR. The proteins with the highest connections are essentially cytochromes and transporters. CYP3A4 has the highest closeness centrality (0.66) and radiality (0.91). This is the most central node. However, its TC is the lowest (0.06), meaning that there is no clear relationship between a set of drugs and a set of ADRs with which CYP3 A is involved.

Finally, 16 ADRs are associated with only one protein. For example, 'Niemann-Pick disease' is linked only with the UGCG protein, and the drug miglustat (Figure 4B); the ADR 'Mixed dementia' is linked with the cytochrome CYP1A2, and the drug bendamustine; and the 'Implant site rash' is associated to the GNRHR receptor, and the drug histrelin. However, these proteins can be related to other ADRs, as many drugs influence these proteins. The Neighborhood Connectivity score can inform us if the nodes (drugs and proteins) linked to an ADR are highly connected to other ADRs. For 'Niemann-Pick disease', the UGCG protein is connected to 18 other ADRs through the drug miglustat (see Figure 4B where only 13 of its ADRs are shown), whereas for 'Mixed dementia', CYP1A2 is linked to 3547 ADRs through 131 drugs. On average, 18.5 proteins are predicted to have an effect on more than one ADR, confirming the possible role of some proteins in many ADRs. To estimate the contribution of each protein to an ADR, we developed an equation (eq. 1) combining the proportion of the proteins targeted by a drug and adding the ensemble of proportions for all drugs involved in the same ADR.[Eq. 1]

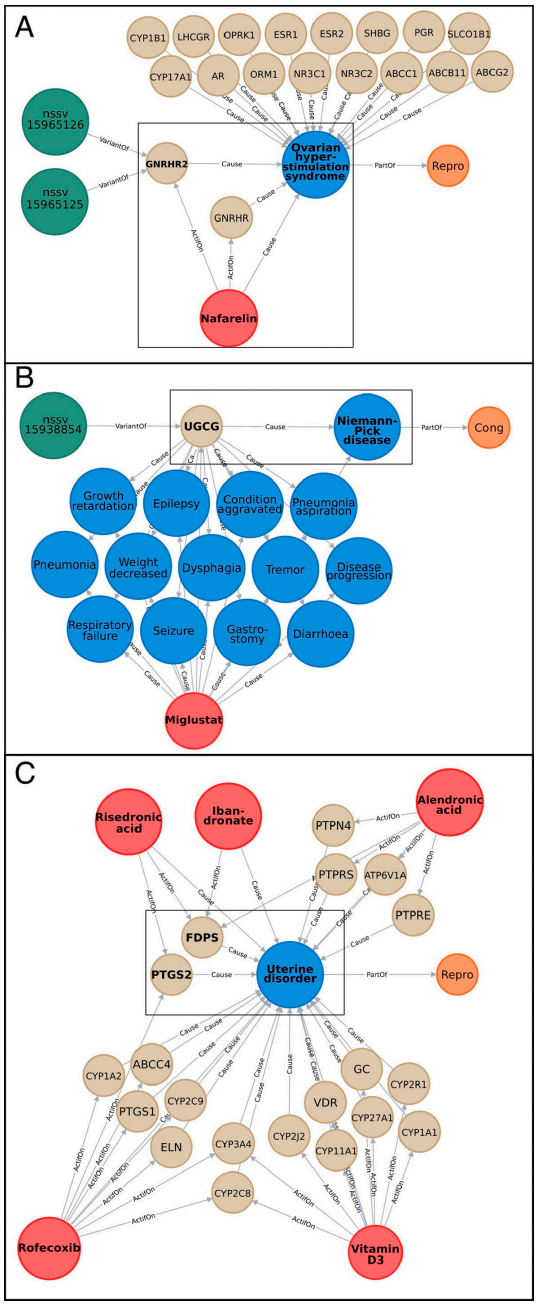


Figure 4. Zoomed-in parts of the network as example. "SV" nodes are represented in dark green and named after their variant accession number, "Target" nodes in beige named after their gene name, "Molecule" nodes in red named after their name, "ADR" nodes in blue named after their name, and "SOC" nodes in orange named after their SOC abbreviation. The edges (in gray) show the relations between different nodes. Nodes cited in the text are framed in black and written in bold. A) Example of subnetwork involving the target "GNRHR2" and the ADR "Ovarian hyper-stimulation syndrome". B) Example of subnetwork involving the target "UGCG" and the ADR "Niemann-Pick disease". Only 13/18 ADRs linked to the miglustat molecule are represented. C) Example of subnetwork involving the ADR "Uterine disorder", its 5 molecules and 20 targets.

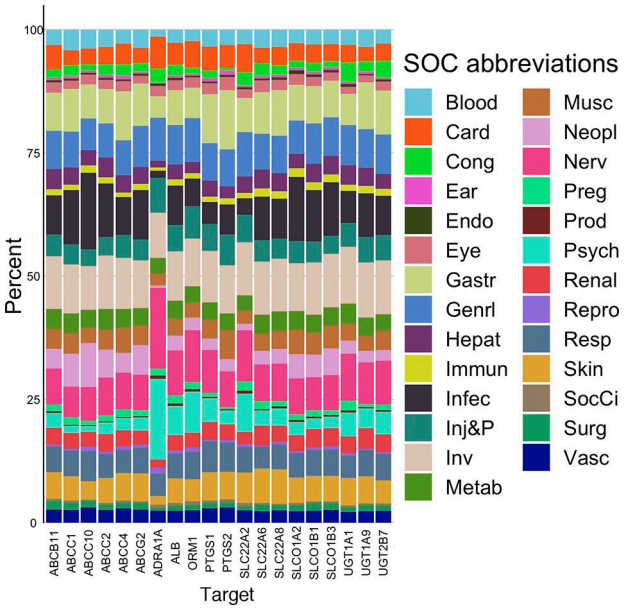


Figure 5. Proportion of drug-SOC associated with the 20 proteins having the most drug-ADR associations. Each color corresponds to an SOC and the size of the color bar the proportion of drug-ADR associations for the protein to a specific SOC.

$$ADRx = \left(\sum^{h} \frac{1}{TD_g} Dg\right) Protein_1 + \left(\sum^{j} \frac{1}{TD_i} Di\right) Protein_2 + \ldots + \left(\sum^{l} \frac{1}{TD_k} Dk\right) Protein_n$$

With

$D_{g,i,k}$: Drug known to cause ADR_x, That also interact with the protein associated with ADR_x.

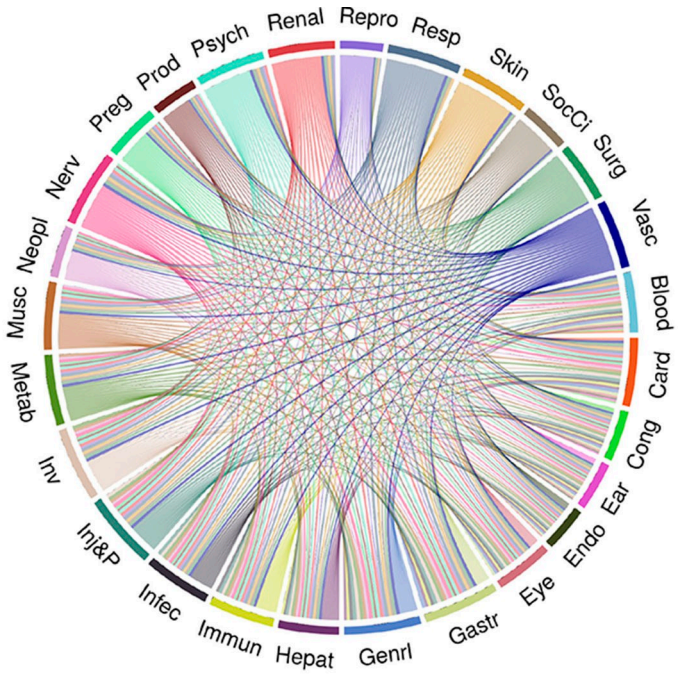
$T_{D\ g,i,k}$: Total number of proteins interacting with $D_{g,i,k}$.

$Protein_{1\ldots n}$: Protein associated with ADR_x.

Such an equation allows the assessment of the contribution of each protein to an ADR. The information is available in the Supporting Information (Table S4).

For example, the ADR 'Prostate cancer stage IV', which is linked to one drug (leuprolide) that targets two proteins (GNRHR and CYP3A4), will have a contribution of 0.5 to GNRHR and 0.5 to CYP3A4. The ADR 'Uterine disorder' is linked to five drugs that are linked to 20 proteins (Figure 4C). Integrating proteins targeted by these drugs, the farnesyl pyrophosphate synthase (FDPS) obtains a score of 1.7, whereas the second one (PTGS2) obtains 0.625, and the other proteins less than 0.23. Therefore, these two proteins could contribute the most to this ADR. Of course, for some general ADR like 'Nausea' for which 944 proteins are linked, the contribution of each protein might be questionable, but for

A



B

SOC	Blood	Card	Cong	Ear	Endo	Eye	Gastr	Genrl	Hepat	Immun	Infec	Inj&P	Inv	Metab	Musc	Neopl	Nerv	Preg	Prod	Psych	Renal	Repro	Resp	Skin	SocCi	Surg	Vasc
Blood	747	654	544	480	470	600	719	742	651	617	688	713	711	691	659	587	715	521	435	661	668	452	706	665	445	589	690
Card	914	552	490	478	643	872	878	666	752	737	877	862	820	742	550	884	613	505	808	751	451	872	784	554	638	805
Cong	583	430	412	525	565	578	534	531	548	579	563	547	550	456	580	485	420	546	549	403	562	544	414	496	563
Ear	566	409	481	564	566	481	532	550	536	565	547	556	421	566	442	399	519	513	399	523	519	406	535	507
Endo	494	440	493	492	480	446	481	490	494	479	480	421	490	416	373	473	484	393	486	479	382	465	488
Eye	710	700	697	604	613	621	685	681	652	656	504	702	513	447	659	632	442	696	641	486	579	662
Gastr	1073	1041	725	855	836	995	984	925	857	598	1034	668	532	898	828	481	983	856	574	722	860
Genrl	1142	723	858	863	1042	1009	924	859	614	1055	671	537	913	847	481	982	884	579	723	876
Hepat	731	621	642	703	710	690	679	532	712	533	443	686	671	447	711	656	466	576	686
Immun	883	716	840	811	772	743	513	844	643	496	781	687	451	804	738	524	648	717
Infec	865	823	838	787	753	599	844	563	482	722	742	454	784	760	487	690	764
Inj&P	1197	956	881	828	592	1011	721	542	898	828	472	968	864	579	694	860
Inv	1034	913	829	603	989	629	526	856	811	470	923	853	577	726	856
Metab	944	784	589	917	606	485	799	759	455	857	793	549	682	806
Musc	864	561	858	583	500	774	740	461	805	740	520	700	763
Neopl	619	602	441	390	547	567	417	586	578	389	533	596
Nerv	1105	661	541	908	835	480	984	871	579	721	880
Preg	723	456	643	575	407	667	608	467	529	572
Prod	555	512	502	360	530	490	392	448	503
Psych	927	761	464	894	801	571	644	795
Renal	855	453	828	748	483	638	779
Repro	482	477	468	349	440	466
Resp	1009	843	572	676	851
Skin	891	540	658	801
SocCi	579	474	534
Surg	727	658
Vasc	886

Figure 6. Chord diagram of the SOC-SOC interaction network based on the target SOC information.

many ADRs for which we have fewer proteins, this approach can be used to prioritize proteins causing ADRs.

Finally, we developed a drug-target-SOC network. At the SOC level, 177 proteins are linked specifically to one SOC. Many proteins do not seem to be specific to one SOC. Many proteins targeted by drugs are linked to 'Metab', 'Infec', 'Gastr', and 'Inv' in majority (Figure 5 and Table S5 in Supporting Information).

Through the SOC-SOC network, we can observe that many proteins are involved in two SOCs. The 'Nervous System disorders' share more than 1000 targets with the 'General disorders and administration site conditions', the 'Injury, poisoning and procedural complications', and the 'Gastrointestinal system'. The 'Respiratory system' is also highly connected with other systems. The 'Reproduction systems' and the 'Social circumstances' are SOCs that are the least connected to other systems with 400 to 500 proteins involved in the two SOCs (Figure 6).

3.3 Structural Variations (SVs) on Drug Targets Associated with ADRs

Matching the 1407 proteins with the data from the Database of Genomic Variants (DGV), we identified 1117 drug targets having SVs; 794 SVs were defined as 'gain' (replication of the protein) and 957 SVs defined as 'loss' (deletion of the protein) (Table S6 in Supporting Information). Figure 7 shows the drug targets with the highest frequency of deletion, replication, and SV associated with the highest number of drugs.

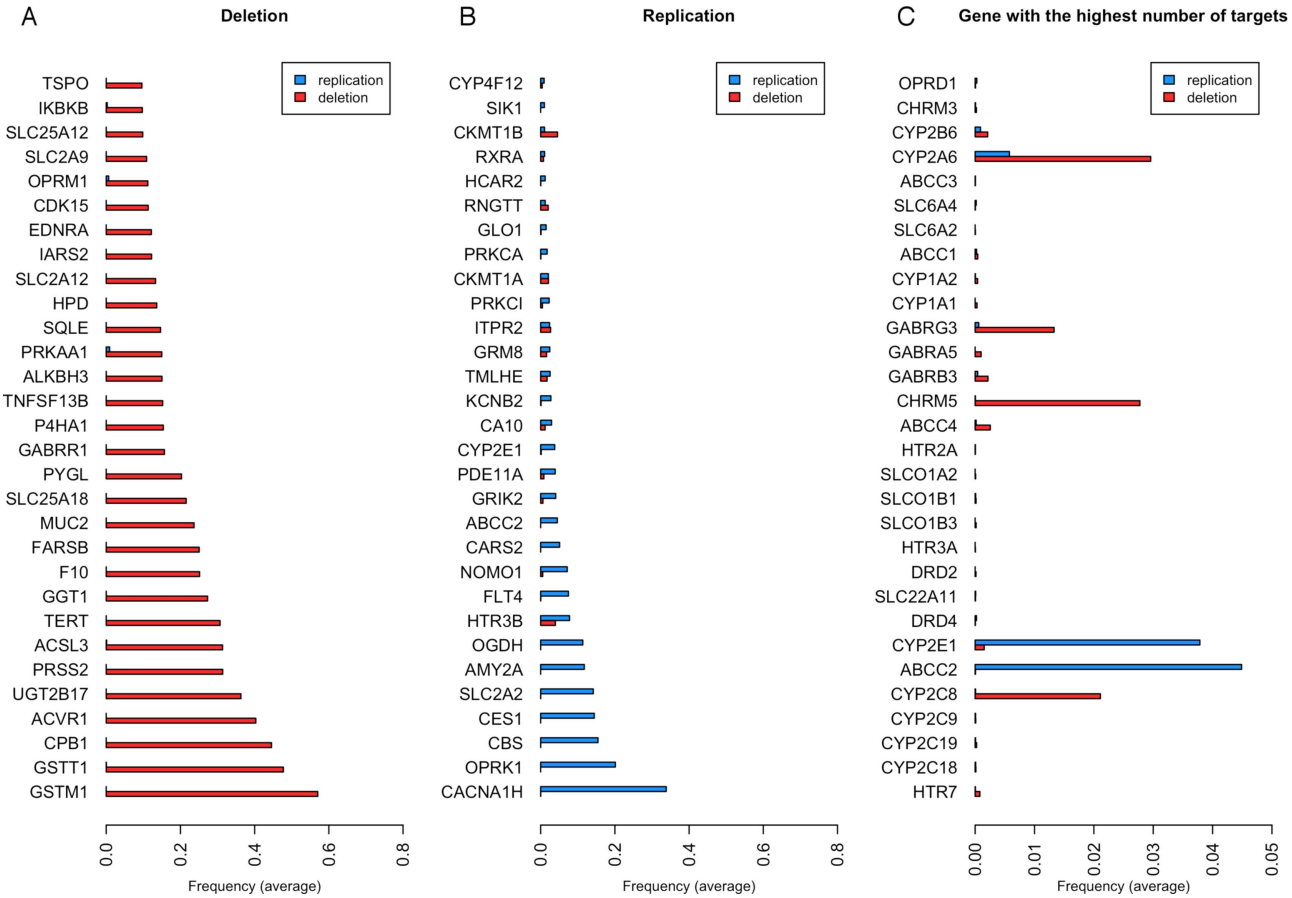


Figure 7. Frequency of structural variations (SVs) according to the variant subtype (deletion or duplication) for A) the 30 genes with the highest frequency of deletion, B) the 30 genes with the highest frequency of replication, and C) the 30 genes with the highest number of drugs.

One of the proteins that shows the highest frequency of replication is opioid receptor kappa (OPRK1). The polymorphism of this protein has recently been reported to affect pain relief from opioids.[43] Studies on cancer and postoperative patients show that carriers of the homozygote G allele have higher pain scores and require higher morphine doses, thereby indicating reduced signaling efficacy and a possibly declined receptor expression.[44] A total of 52 drugs target this protein, and some of them show as ADR 'Drug abuse', 'Drug tolerance', 'Drug hypersensitivity', 'Drug ineffective', 'Drug intolerance', or 'Drug withdrawal syndrome' and so might be related to this SV. CES1 also has a high frequency of replication. CES1 is involved in the catalysis of the hydrolytic biotransformation of a variety of compounds containing an ester, amide, or carbamate function to their respective free acids and alcohols.[45] Numerous drugs, including the psychostimulant methylphenidate (MPH) used in the treatment of attention-deficit hyperactivity disorder (ADHD), angiotensin-converting enzyme inhibitors (quinapril, imidapril, temocapril, and cilazapril), anti-cancer agents (CPT-11), and narcotics and analgesics (cocaine and meperidine) are all hCES1 substrates, and the ADRs associated with these drugs could involve in some individuals an SV of this protein.[46] On the other hand, glutathione S-transferase, GSTM1, and GSTT1 genes are more frequently subject to deletion. Studies have reported the susceptibility of these enzymes to clinical toxicities, notably gastrointestinal toxicity.[47] This is a typical ADR that we see, for example, with acetaminophen (paracetamol). The opioid receptor mu (OPRM1) is also among the most frequently deleted genes. Therefore, opioid receptors seem to have frequent SV (deleted or replicated) that might explain ADRs related to drug inefficacy. In general, we retrieved drug-metabolizing enzymes (notably cytochromes) with the most frequent SV. As these proteins are also highly targeted by drugs, we can assume that the susceptibility to ADRs caused by an SV on these targets is not negligible. For example, the antidepressant fluoxetine targets CYP2C9, CYP2C19, CYP1A2, CYP2B6, CYP3A4, CYP3A5, ORM1, SLC6A4, HTR2 C, and KCNH2, which have some of the highest frequencies among SV. We could assume that patients having an SV on one of these proteins and taking this treatment have a greater susceptibility to one of the ADRs linked to this drug. Interestingly, at the top of the list, we can observe the serotonin, dopamine, and noradrenaline neurotransmitter

transporters. As they are the primary targets of many antidepressants, it could be one of the explanations for the inefficacy or ADR associated with them.[48]

4 Conclusions

The present study explored the reported ADRs of drugs and the potential drug-target associations causing these ADRs, using network sciences. Although it is possible to assume the role of some proteins for some specific ADRs, many biological targets are related to several ADRs. Grouping ADRs into SOCs allowed the observation of organs more affected by ADRs, and whether the drug-specific ADRs were more localized in an SOC or spread over several organs. We have to be aware that some terms such as 'Drug ineffective', 'Device ineffective' and 'Therapeutic product ineffective' are listed as ADRs in the reporting system provided by the FDA, although these terms mean that there is no reaction occurring by the administration of a drug in a normal dose. This is a known situation. As Wysowsky et al. observed, the most frequently reported adverse event was 'Drug ineffective'.[49]

The findings of ADRs are essentially based on clinical trials or spontaneous reports but are rarely related to genomic variations. The inclusion of structural variations in the drug-target data is an interesting avenue as the high polymorphism of some genes might contribute to increase the susceptibility to an ADR. In our study, we considered structural variations of more than 1 kb, which cover a complete loss or a large gain of a gene that is important for the functionality of a protein. There are more local genetic variations such as single-nucleotide polymorphisms (SNPs) which can also be related to the occurrence of ADRs, but the impact of such local variations in a protein target is more difficult to assess. Pharmacogenomic studies investigating the role of genetic variations in drug response and results for some drugs have been reported and can be accessed.[50–51]

Some challenges remain in utilizing the full potential of network sciences to decipher the mechanisms behind drug-ADR associations and network pharmacology.[52] The coverage of the drug-target associations is not fully accomplished, and ADRs might be caused by some targets not yet determined for a drug. In addition, our network does not consider the binding affinity value between a drug and a target. Besides, drugs might directly impact the expression of genes. With the opportunity to access omics data, notably transcriptomic data, toxicogenomic studies would allow an analysis of the deregulation of genes and pathways in a specific cell type, tissue, or organ, in the presence of a drug. The integration of such information will be beneficial for obtaining a more comprehensive pharmacological profile of drugs.

Conflict of Interest

None declared.

Acknowledgements

We would like to thank the doctoral school "Pierre Louis de santé publique" and the pharmaceutical company Servier for their support on this study. This study contributes to IdEx Université de Paris ANR-18-IDEX-0001.

References

[1] J. Lazarou, B. H. Pomeranz, P. N. Corey, *JAMA.* **1998**, *279*, 1200–1205.
[2] C. Giardina, P. M. Cutroneo, E. Mocciaro, G. T. Russo, G. Mandraffino, G. Basile, F. Rapisarda, R. Ferrara, E. Spina, V. Arcoraci, *Front. Pharmacol.* **2018**, *9*, 350.
[3] C. Kongkaew, P. R. Noyce, D. M. Ashcroft, *Ann. Pharmacother.* **2008**, *42*, 1017–1025.
[4] T. K. Patel, P. B. Patel, *Curr. Drug Saf.* **2016**, *11*, 128–136.
[5] S. V. Iyer, R. Harpaz, P. LePendu, A. Bauer-Mehren, N. H. Shah, *J. Am. Med. Inform. Assoc.* **2013**, *21*(2), 353–62.
[6] R. Cacabelos, N. Cacabelos, J. C. Carril, *Expert Rev. Clin. Pharmacol.* **2019**, *12*, 407–442.
[7] A. L. Hopkins, *Nat. Chem. Biol.* **2008**, *4*, 682–690.
[8] K. Audouze, A. S. Juncker, F. J. Roque, K. Krysiak-Baltyn, N. Weinhold, O. Taboureau, T. S. Jensen, S. Brunak, *PLoS Comput. Biol.* **2010**, *6*: e1000788.
[9] J. Scheiber, J. L. Jenkins, S. C. K. Sukuru, A. Bender, D. Mikhailov, M. Milik, K. Azzaoui, S. Whitebread, J. Hamon, L. Urban, et al., *J. Med. Chem.* **2009**, *52*, 3103–3107.
[10] N. Atias, R. Sharan, *J. Comput. Biol.* **2011**, 18, 207–218.
[11] T–B. Ho, L. Le, D. T. Thai, S. Taewijit, *Curr. Pharm. Des.* **2016**, *22*, 3498–3526.
[12] M. R. Boland, A. Jacunski, T. Lorberbaum, J. Romano, R. Moskovitch, N. Tatonetti, *Wiley Interdiscip Rev. Syst. Biol. Med.* **2016**, *8*, 104–22.
[13] H. Zhou, M. Gao, J. Skolnick, *Sci. Rep.* **2015**, *5*, 11090.
[14] M. Campillos, M. M. Kuhn, A. C. Gavin, L. J. Jensen, P. Bork, *Science* **2008**, *321*, 263–266.
[15] A. F. Fliri, W. T. Loging, R. A. Volkmann, *ChemMedChem* **2007**, 2, 1774–1782.
[16] E. Lounkine, M. J. Keiser, M. S. Whitebread, D. Mikhailov, J. Hamon, J. L. Jenkins, P. Lavan, P. E. Weber, A. K. Doak, S. Côté, et al., *Nature* **2012**, 486, 361–367.
[17] R. Garcia-Serna, J. Mestres, *Expert Opin. Drug Metab. Toxicol.* **2010**, 6, 1253–1263.
[18] A. L. Barabási, *Network Science* (Cambridge Univ. Press, **2016**).
[19] B. Boezio, K. Audouze, P. Ducrot, O. Taboureau, *Mol. Inf.* **2017**, *36*, 1700048.
[20] Y. Yamanishi, M. Kotera, M. Kanehisa, S. Goto, *Bioinformatics.* **2010**, *26*, 246–54.
[21] Y. Jiang, Y. Li, Q. Kuang, L. Ye, Y. Wu, L. Yang, M. Li, *Anal. Methods* **2014**, *6*, 2692–2698.
[22] Y. Hwang, M. Oh, G. Jang, T. Lee, C. Park, J. Ahn, Y. Yoon, *Mol. BioSyst.* **2017**, *13*, 1788–1796.
[23] X. Chen, X. Liu, X. Jia, F. Tan, R. Yang, S. Chen, L. Liu, Y. Wang, Y. Chen, *Sci. Rep.* **2013**, *3*, 1744.
[24] T. I. Oprea, S. K. Nielsen, O. Ursu, J. J. Yang, O. Taboureau, S. L. Mathias, Irene Kouskoumvekaki, L. A. Sklar, C. G. Bologa, *Mol. Inf.* **2011**, *30*, 100–111.
[25] R. Xue, J. Liai, X. Shao, K. Han, J. Long, L. Shao, N. Ai, X. Fan, *Chem. Res. Toxicol.* **2020**, *33*, 202–210.

[26] Brown E. G., Wood L., Wood S. The Medical Dictionary for Regulatory Activities (MedDRA), *Drug Saf.***1999**. *20*, 109–117.
[27] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, et al., *Nucleic Acids Res.* **2018**, *46*, D1074-D1082.
[28] O. Ursu, O. J. Holmes, C. G. Bologa, J. J. Yang, S. L. Mathias, V. Stathhias, D. T. Nguyen, S. Schürer, T. Oprea, *Nucleic Acids Res.* **2019**, *47*, D963-D970.
[29] L. Huang, J. Zalkikar, R. C. Tiwari, *J. Am. Stat. Assoc.* **2011**, *106*, 1230–1241.
[30] J. R. MacDonald, R. Ziman, R. K. Yuen, L. Feuk, S. W. Scherer, *Nucleic Acids Res.* **2013**, *42*, D986–92.
[31] J. M. Harrold, M. Ramanathan, D. E. Mager, *Clin. Ther.* **2013**, *94*, 651–658.
[32] R Core Team. **2017**. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[33] Z. Gu, *Bioinformatics.* **2014**, *30*, 2811–2.
[34] D. Sarkar. Lattice: Multivariate data visualization with R. Springer, New York. **2008**. ISBN 978-0-387-75968–5.
[35] H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. **2016**. ISBN 978-3-319-24277–4.
[36] S. Richardson, J. S. Hirsch, M. Narasimhan, T. McGinn, K. W. Davidson, et al., *JAMA.* **2020**, doi: 10.1001/jama.2020.6775.
[37] F. P. Guengerich, J. S. MacDonald, *Chem. Res. Toxicol.* **2007**, *20*, 344–369.
[38] P. B. Danielson, *Curr. Drug Metab.* **2002**. *3*, 561–97.
[39] V. E. Blanco, C. V. Hernandorena, P. Scibona, W. Belloso, C. G. Musso, *Healthcare.* **2019**, *7*,10.
[40] J. Faria, S. Ahmed, K. G. F. Gerritsen, S. M. Mihaila, R. Masereeuw, *Arch. Toxicol.* **2019**, *93*, 2297–3418.
[41] H.F Cheng, R. C. Harris, *Curr. Pharm. Des.* **2005**. *11*, 1795–804.
[42] J. Li, W. Chen, C. Peng, W. Zhu, Z. Liu, W. Zhang, J. Su, J. Li, X. Chen, *Pharmacogenomics J.* **2020**, *20*, 87–93.
[43] J. W.D Ho, M. R. Wallace, R. Staud, R. B. Fillingim, *Pharmacogenomics J.* **2019**, doi: 10.1038/s41397-019-0131-z.
[44] L. M. Nielsen, A. E. Olesen, R. Branford, L. L. Christrup, H. Sato, *Pain Pract*. **2015**. *15*(6), 580–594.
[45] M. R. Redinbo, P. M. Potter, *Drug Discovery Today* **2005**, *10*, 313–25.
[46] G. S. Nzabonimpa, H. B. Rasmussen, S. Brunak, O. Taboureau, *Drug Metab. Pers. Ther.* **2016**, 31, 97–106.
[47] M. Abbas, V. S. Kushwaha, K. Srivastava, S. T. Raza, M. Banerjee, *Br. J. Biomed. Sci.* **2018**, *75*, 169–174.
[48] C. A. Bousman, M. Forbes, M. Jayaram, H. Eyre, C. F. Reynolds, M. Berk, M. Hopwood, C. Ng, *BMC Psychiatry.* **2017**, *17*, 60.
[49] D. K. Wysowski, L. Swartz, *Arch. Intern. Med.* **2005**, *165*, 1363–9.
[50] J. M. Barbarino, M. Whirl-Carrillo, R. B. Altman, T. E. Klein, Wiley *Interdiscip Rev. Syst. Biol. Med.* **2018**, *10*, e1417.
[51] J. A. Luzum, R. E. Pakys, A. R. Elsey, C. E. Haidar, J. F. Peterson, M. Whirl-Carrillo, S. K. Handelman, K. Palmer, J. M. Pulley, et al., *Clin. Pharmacol. Ther.* **2017**, *102*, 502–510.
[52] I. Vogt, J. Mestres, *Mol. Inf.* **2019**, *38*, e1900032.

Received: May 12, 2020
Accepted: July 28, 2020
Published online on August 28, 2020

SUPPLEMENTARY DATA

Table S1: SOC abbreviations

SOC	abbrev
Blood and lymphatic system disorders	Blood
Cardiac disorders	Card
Congenital, familial and genetic disorders	Cong
Ear and labyrinth disorders	Ear
Endocrine disorders	Endo
Eye disorders	Eye
Gastrointestinal disorders	Gastr
General disorders and administration site conditions	Genrl
Hepatobiliary disorders	Hepat
Immune system disorders	Immun
Infections and infestations	Infec
Injury, poisoning and procedural complications	Inj&P
Investigations	Inv
Metabolism and nutrition disorders	Metab
Musculoskeletal and connective tissue disorders	Musc
Neoplasms benign, malignant and unspecified (incl cysts and polyps)	Neopl
Nervous system disorders	Nerv
Pregnancy, puerperium and perinatal conditions	Preg
Product issues	Prod
Psychiatric disorders	Psych
Renal and urinary disorders	Renal
Reproductive system and breast disorders	Repro
Respiratory, thoracic and mediastinal disorders	Resp
Skin and subcutaneous tissue disorders	Skin
Social circumstances	SocCi
Surgical and medical procedures	Surg
Vascular disorders	Vasc

Table S2: Network analysis - Adverse Drug Event

50 first lines of the supplementary table 2. The full data is available at DOI: 10.1002/minf.202000116.

Name	Degree	Closeness Centrality	Betweenness Centrality	Neighborhood Connectivity	Radiality	TopologicalCoefficient
"Allergic hepatitis"	1	0.259	0.0	144.0	0.523	0.0
"Allergy to chemicals"	1	0.258	0.0	127.0	0.522	0.0
"Bladder instillation procedure"	1	0.274	0.0	957.0	0.560	0.0
"Blood ketone body"	1	0.258	0.0	130.0	0.522	0.0
"Chemical burn"	1	0.258	0.0	127.0	0.522	0.0
"Conjunctival erosion"	1	0.265	0.0	467.0	0.539	0.0
"Implant site rash"	1	0.257	0.0	110.0	0.520	0.0
"Mixed dementia"	1	0.334	0.0	3708.0	0.668	0.0
"Multiple cardiac defects"	1	0.289	0.0	1742.0	0.591	0.0
"Niemann-Pick disease"	1	0.255	0.0	19.0	0.514	0.0
"Scleral thinning"	1	0.265	0.0	467.0	0.539	0.0
"Scleromalacia"	1	0.265	0.0	467.0	0.539	0.0
"Sensitisation"	1	0.258	0.0	127.0	0.522	0.0
"Skin test negative"	1	0.40	0.0	5824.0	0.750	0.0
"Urine ketone body"	1	0.258	0.0	130.0	0.522	0.0
"Urticaria chronic"	1	0.289	0.0	1742.0	0.591	0.0
"Agraphia"	2	0.260	0.0	208.0	0.525	1.0
"Anterior chamber fibrin"	2	0.296	0.0	1350.0	0.604	0.678
"Anterior chamber inflammation"	2	0.296	0.0	1350.0	0.604	0.678
"Application site pustules"	2	0.258	0.0	119.0	0.520	1.0
"Atrophy of globe"	2	0.290	0.0	1183.5	0.593	0.664
"Autonomic dysreflexia"	2	0.264	0.0	305.0	0.536	0.962
"Blood cholinesterase decreased"	2	0.271	0.0	537.0	0.553	0.688
"Blood product transfusion dependent"	2	0.268	0.0	361.5	0.545	0.574
"Cerebral cyst"	2	0.276	0.0	720.5	0.564	0.674
"Choroidal dystrophy"	2	0.290	0.0	1183.5	0.593	0.664
"Coagulation time abnormal"	2	0.268	1.0E-8	310.0	0.546	0.580
"Congenital cyst"	2	0.276	0.0	720.5	0.564	0.674
"Coronary artery perforation"	2	0.268	1.0E-8	310.0	0.546	0.580
"CSF culture positive"	2	0.264	0.0	305.0	0.536	0.962
"CSF pressure decreased"	2	0.264	0.0	305.0	0.536	0.962
"Device battery issue"	2	0.264	0.0	305.0	0.536	0.962
"Device electrical finding"	2	0.264	0.0	305.0	0.536	0.962

"Device extrusion"	2	0.264	0.0	305.0	0.536	0.962
"Device inversion"	2	0.264	0.0	305.0	0.536	0.962
"Device kink"	2	0.264	0.0	305.0	0.536	0.962
"Device material issue"	2	0.264	0.0	305.0	0.536	0.962
"Device physical property issue"	2	0.264	0.0	305.0	0.536	0.962
"Device related thrombosis"	2	0.263	0.0	233.0	0.533	0.602
"Duodenal atresia"	2	0.276	0.0	720.5	0.564	0.674
"Epiphysiolysis"	2	0.401	9.2E-7	2967.0	0.751	0.508
"Extraocular muscle disorder"	2	0.290	0.0	1183.5	0.593	0.664
"Eye excision"	2	0.290	0.0	1183.5	0.593	0.664
"Eyelid rash"	2	0.290	0.0	1183.5	0.593	0.664
"Forced vital capacity decreased"	2	0.334	5.0E-8	1885.5	0.668	0.508
"General symptom"	2	0.260	0.0	208.0	0.525	1.0
"Head banging"	2	0.338	1.0E-8	2130.5	0.674	0.549
"Hepatic infarction"	2	0.291	0.0	1667.0	0.595	0.941
"Hereditary angioedema"	2	0.259	1.0E-8	83.0	0.524	0.532

Table S3: The 100 proteins the most targeted by drugs.

Rang	Gene	NbrDrug	Rang	Gene	NbrDrug	Rang	Gene	NbrDrug	Rang	Gene	NbrDrug	Rang	Gene	NbrDrug
1	CYP3A4	486	21	DRD2	62	41	GABRA3	43	61	GABRB3	36	82	HTR1B	27
2	CYP2D6	224	22	PTGS1	62	42	HTR2C	43	62	GABRG1	36	83	HTR3A	27
3	CYP2C9	218	23	ADRA1B	61	43	SLC6A4	43	63	ADRB2	35	84	SLC22A7	27
4	ALB	189	24	SLCO1B3	61	44	GABRA2	42	64	GABRB2	35	85	SLC6A3	27
5	CYP2C19	178	25	ABCC2	59	45	SLCO2B1	42	65	GABRD	35	86	CYP3A43	26
6	CYP1A2	161	26	ADRA2A	56	46	GABRA5	41	66	GABRE	35	87	DRD4	26
7	CYP3A5	154	27	CYP2A6	56	47	SLC15A1	41	67	GABRG3	35	88	SLC22A11	26
8	CYP2C8	151	28	HRH1	55	48	SLC22A1	40	68	GABRP	34	89	CYP1B1	25
9	CYP2B6	112	29	HTR2A	55	49	ADRA2C	39	69	GABRQ	34	90	OPRD1	25
10	ABCG2	99	30	CHRM1	53	50	DRD1	39	70	CHRM5	32	91	CES1	24
11	SLCO1B1	99	31	SLCO1A2	51	51	GABRG2	39	71	DRD3	32	92	DRD5	24
12	CYP3A7	94	32	CHRM3	50	52	ABCC1	38	72	OPRK1	32	93	HTR1D	24

13	SLC22A6	85	33	PTGS2	50	53	ABCC4	38	73	OPRM1	32	94	HTR2B	23
14	CYP2E1	82	34	SLC22A2	50	54	GABRA4	38	74	SLC22A5	32	95	MAOA	23
15	ABCB11	74	35	CHRM2	49	55	GABRA6	38	75	KCNH2	31	96	NR3C1	22
16	SLC22A8	73	36	SLC6A2	49	56	UGT1A9	38	76	UGT1A3	31	97	SCN5A	22
17	ADRA1A	72	37	HTR1A	48	57	UGT2B7	38	77	BCHE	29	98	SLC16A10	22
18	ORM1	71	38	ADRA1D	46	58	CHRM4	37	78	SLC15A2	29	99	ABCC3	21
19	UGT1A1	67	39	GABRA1	46	59	ADRB1	36	79	HTR7	28	100	CYP2C18	21
20	CYP1A1	63	40	ADRA2B	43	60	GABRB1	36	80	NR1I2	28	100	HTR6	20

Table S4: Proteins potentially involved in the cause of ADR

10 first lines of the table S4. The entire file is available at DOI: 10.1002/minf.202000116.

ADR	Score of proteins involved in the ADR
17 ketosteroids urine abnormal	0.077xABCG2 + 0.077xCYP2C18 + 0.077xCYP2C19 + 0.077xCYP2C8 + 0.077xCYP2C9 + 0.077xCYP2D6 + 0.077xCYP3A4 + 0.077xCYP3A5 + 0.077xCYP3A7 + 0.077xOPRD1 + 0.077xOPRK1 + 0.077xOPRL1 + 0.077xOPRM1
5-hydroxyindolacetic acid increased	0.2xCYP3A4 + 0.2xMPO + 0.2xSSTR1 + 0.2xSSTR2 + 0.2xSSTR5
5q minus syndrome	0.346xCYP3A4 + 0.283xCYP3A5 + 0.25xCDH5 + 0.25xCRBN + 0.25xPTGS2 + 0.25xTNFSF11 + 0.25xCYP2B6 + 0.216xCYP2A6 + 0.216xCYP2C19 + 0.216xCYP2C8 + 0.2xDCK + 0.2xPOLA1 + 0.2xRRM1 + 0.2xSLC28A3 + 0.2xSLC29A1 + 0.178xCYP2C9 + 0.172xCYP3A7 + 0.168xABCC2 + 0.168xABCG2 + 0.168xALB + 0.15xNR1I2 + 0.138xABCB11 + 0.135xSLCO1A2 + 0.129xABCC1 + 0.129xABCC10 + 0.129xABCC3 + 0.111xCYP2C18 + 0.105xCYP3A43 + 0.105xHSD11B1 + 0.105xNR3C1 + 0.1xRALBP1 + 0.1xCYP1B1 + 0.096xSLCO1B1 + 0.096xSLCO1B3 + 0.072xNOS2 + 0.068xSLC22A8 + 0.067xSLC22A3 + 0.067xTUBA4A + 0.067xTUBB + 0.067xSERPINA6 + 0.038xANXA1 + 0.038xCYP11B1 + 0.038xCYP17A1 + 0.038xCYP1A1 + 0.038xCYP2E1 + 0.038xCYP4A11 + 0.038xHSD11B2 + 0.038xNR0B1 + 0.033xABCB8 + 0.033xABCC6 + 0.033xAKR1A1 + 0.033xAKR1C3 + 0.033xCBR1 + 0.033xCBR3 + 0.033xCYP2D6 + 0.033xNDUFS2 + 0.033xNDUFS3 + 0.033xNDUFS7 + 0.033xNOLC1 + 0.033xNOS1 + 0.033xNOS3 + 0.033xNQO1 + 0.033xPOR + 0.033xSLC22A16 +

	0.033xTOP2A + 0.033xXDH + 0.029xABCC11 + 0.029xABCC4 + 0.029xAOX1 + 0.029xATIC + 0.029xDHFR + 0.029xFOLR1 + 0.029xFOLR2 + 0.029xFPGS + 0.029xGGH + 0.029xMTHFR + 0.029xPGD + 0.029xSLC15A1 + 0.029xSLC16A1 + 0.029xSLC19A1 + 0.029xSLC22A11 + 0.029xSLC22A6 + 0.029xSLC22A7 + 0.029xSLC36A1 + 0.029xSLC46A1 + 0.029xSLCO1C1 + 0.029xSLCO3A1 + 0.029xSLCO4C1 + 0.029xTYMS
Abdominal abscess	0.678xCYP3A4 + 0.604xALB + 0.449xCYP3A5 + 0.319xNR3C1 + 0.296xSLCO1A2 + 0.287xSLCO1B1 + 0.267xSERPINA6 + 0.266xABCG2 + 0.258xUGT1A1 + 0.258xUGT1A9 + 0.253xCYP3A7 + 0.222xCYP2C8 + 0.216xABCC2 + 0.195xIMPDH1 + 0.195xIMPDH2 + 0.195xUGT1A6 + 0.195xUGT1A7 + 0.195xUGT2B7 + 0.182xCYP2B6 + 0.175xCYP1A2 + 0.169xCYP2A6 + 0.169xCYP2C9 + 0.151xABCC4 + 0.143xABCA5 + 0.143xFKBP1A + 0.143xORM1 + 0.125xGSTA1 + 0.125xGSTA2 + 0.125xGSTM1 + 0.125xHPRT1 + 0.125xRAC1 + 0.125xTPMT + 0.125xXDH + 0.119xCYP1B1 + 0.119xCYP2C19 + 0.119xCYP3A43 + 0.119xHSD11B1 + 0.115xCES1 + 0.092xABCC1 + 0.082xSLCO1B3 + 0.079xABCC3 + 0.079xMTHFR + 0.079xSLC22A7 + 0.079xTYMS + 0.071xCSF1R + 0.071xFLT1 + 0.071xFLT3 + 0.071xFLT4 + 0.071xKDR + 0.071xKIT + 0.071xPDGFRA + 0.071xPDGFRB + 0.062xBCHE + 0.062xSLC22A3 + 0.062xTOP1 + 0.062xTOP1MT + 0.053xCES2 + 0.053xPTS + 0.053xUGT1A10 + 0.053xUGT1A8 + 0.053xAKR1C1 + 0.053xAKR1C2 + 0.053xAKR1C3 + 0.053xAKR1C4 + 0.053xANXA1 + 0.053xHSD11B2 + 0.05xABCC5 + 0.05xDPYD + 0.05xPPAT + 0.05xSERPINA7 + 0.05xSLC29A1 + 0.05xTYMP + 0.05xUMPS + 0.05xUPP1 + 0.05xUPP2 + 0.029xABCC10 + 0.029xABCC11 + 0.029xAOX1 + 0.029xATIC + 0.029xDHFR + 0.029xFOLR1 + 0.029xFOLR2 + 0.029xFPGS + 0.029xGGH + 0.029xPGD + 0.029xSLC15A1 + 0.029xSLC16A1 + 0.029xSLC19A1 + 0.029xSLC22A11 + 0.029xSLC22A6 + 0.029xSLC22A8 + 0.029xSLC36A1 + 0.029xSLC46A1 + 0.029xSLCO1C1 + 0.029xSLCO3A1 + 0.029xSLCO4C1
Abdominal adhesions	0.2xATP6V1A + 0.2xFDPS + 0.2xPTPN4 + 0.2xPTPRE + 0.2xPTPRS
Abdominal cavity drainage	0.333xP2RY12 + 0.333xPTGIR + 0.333xPTGIS + 0.19xABCG2 + 0.19xCYP3A4 + 0.19xCYP3A5 + 0.19xCYP3A7 + 0.19xSLCO1B1 + 0.19xSLCO1B3 + 0.143xCYP3A43 + 0.048xABCB11 + 0.048xABCC10 + 0.048xABCC2 + 0.048xABCC3 + 0.048xCAMLG + 0.048xCYP2C19 + 0.048xCYP2C9 + 0.048xCYP2D6 + 0.048xPPIA + 0.048xPPIF + 0.048xPPP3R2 + 0.048xSLC10A1 + 0.048xSLC10A2 + 0.048xSLC22A6 + 0.048xSLCO1A2
Abdominal compartment syndrome	0.2xCYP2C19 + 0.2xCYP2C9 + 0.152xCHRM2 + 0.152xCYP2D6 + 0.152xHRH1 + 0.125xPTGS1 + 0.125xSLC22A2 + 0.125xSLC22A5 + 0.104xCYP3A4 +

	0.077xSLC22A8 + 0.077xSLCO1A2 + 0.075xCYP2A6 + 0.075xCYP2C8 + 0.048xAKR1D1 + 0.048xANXA1 + 0.048xCYP11B1 + 0.048xCYP11B2 + 0.048xCYP1B1 + 0.048xCYP2B6 + 0.048xCYP3A5 + 0.048xCYP3A7 + 0.048xHSD11B1 + 0.048xHSD11B2 + 0.048xNR3C1 + 0.048xSERPINA6 + 0.048xSHBG + 0.048xSRD5A2 + 0.029xABCC1 + 0.029xABCC10 + 0.029xABCC11 + 0.029xABCC2 + 0.029xABCC3 + 0.029xABCC4 + 0.029xABCG2 + 0.029xALB + 0.029xAOX1 + 0.029xATIC + 0.029xDHFR + 0.029xFOLR1 + 0.029xFOLR2 + 0.029xFPGS + 0.029xGGH + 0.029xMTHFR + 0.029xPGD + 0.029xSLC15A1 + 0.029xSLC16A1 + 0.029xSLC19A1 + 0.029xSLC22A11 + 0.029xSLC22A6 + 0.029xSLC22A7 + 0.029xSLC36A1 + 0.029xSLC46A1 + 0.029xSLCO1B1 + 0.029xSLCO1B3 + 0.029xSLCO1C1 + 0.029xSLCO3A1 + 0.029xSLCO4C1 + 0.029xTYMS + 0.027xADRA1A + 0.027xADRA1B + 0.027xADRA2A + 0.027xADRA2B + 0.027xADRA2C + 0.027xCALY + 0.027xCHRM1 + 0.027xCHRM3 + 0.027xCHRM4 + 0.027xCHRM5 + 0.027xCYP1A1 + 0.027xCYP1A2 + 0.027xDRD1 + 0.027xDRD2 + 0.027xDRD3 + 0.027xDRD4 + 0.027xFMO3 + 0.027xGSTP1 + 0.027xHRH4 + 0.027xHTR1A + 0.027xHTR1B + 0.027xHTR1D + 0.027xHTR1E + 0.027xHTR2A + 0.027xHTR2C + 0.027xHTR3A + 0.027xHTR6 + 0.027xHTR7 + 0.027xUGT1A4

Table S5: Table counting drugs and targets associated with drugs, based on the 27 System Organ Class (SOC).

SOC	Number of drugs	Number of targets	SOC	Number of drugs	Number of targets
Blood	361	747	Musc	381	864
Card	464	914	Neopl	243	619
Cong	205	582	Nerv	616	1105
Ear	156	565	Preg	212	722
Endo	170	494	Prod	208	555
Eye	287	709	Psych	487	927
Gastr	562	1073	Renal	427	855
Genrl	684	1142	Repro	144	482
Hepat	335	731	Resp	531	1009
Immun	407	882	Skin	478	891
Infec	428	865	SocCi	144	579
Inj&P	653	1197	Surg	274	727
Inv	583	1034	Vasc	441	886
Metab	458	944			

Table_S6: Genes encoding for drug's targets and having an SV. The values correspond to the frequency of the SV (deletion or replication) observed in healthy individuals.
First 50 lines of the table S6 The rest of the file is available at DOI: 10.1002/minf.202000116.

Gene	deletion	replication
AKR1C1	4,60E-05	0,00012
AKR1C2	0,013	0,00015
AKR1C3	0,0003	7,67E-05
AKR1C4	0,0305	0,00012
PRKCQ	6,13E-05	NA
CACNB2	0,0562	0,006500833
MSRB2	6,90E-05	NA
THNSL1	4,60E-05	NA
GAD2	6,90E-05	NA
ALOX5	7,40E-05	4,60E-05
MAPK8	6,13E-05	6,13E-05
HK1	0,003	0,0002
RET	4,60E-05	NA
P4HA1	0,1539	NA
ADK	0,0005	9,20E-05
KCNMA1	0,0092	7,67E-05
MAT1A	0,0002	0,0004
GRID1	0,0005	2,00E-04
HTR7	0,0008	4,60E-05
PDE6C	0,0003	NA
CYP2C18	0,0001	9,20E-05
CYP2C19	0,0003	7,67E-05
CYP2C9	0,0001	7,67E-05
CYP2C8	0,0211	6,90E-05
ALDH18A1	9,25E-05	NA
GOT1	5,83E-05	4,60E-05
ABCC2	5,75E-05	0,045
CHUK	4,70E-05	NA
NOLC1	5,91E-05	NA
NT5C2	4,60E-05	4,60E-05
GSTO2	6,17E-05	NA
MGMT	0,00012	0,00023
CHRNA10	0,00005	0,00005
RRM1	0,00420	0,00016
HBB	0,00015	0,00005
CALY	0,00014	0,00132
CYP2E1	0,00152	0,03786
KCNQ1	0,00941	0,00031

CARS	0,0002	NA
ABCC8	0,0064	6,13E-05
KCNC1	4,60E-05	NA
TPH1	0,0008	4,60E-05
PTPRE	0,0176	4,60E-05
SLC1A2	7,91E-05	0,000107667
MUC2	0,2368	4,60E-05
PDE3B	0,0003	8,05E-05
CYP2R1	4,60E-05	9,20E-05
FGFR2	6,44E-05	0,0012
ACADSB	8,13E-05	NA

3.4 Conclusion

La création d'une fonction de score permettant de mettre un poids sur chaque protéine associée à un effet indésirable reporté dans *drugcentral* est une implémentation intéressante ouvrant des pistes sur des mécanismes d'actions et des protéines plus à risque à engendrer un effet indésirable. À ma connaissance, aucune analyse systématique de ce genre pour chaque effet indésirable n'a été effectuée. Néanmoins, ces interactions se basant sur des essais cliniques de nombreuses interactions ne sont pas connues, et les effets indésirables de chaque médicament sont parfois trop nombreux pour identifier de manière effective une protéine spécifique qui en serait responsable (e.g : fièvre, nausée…). De plus, notre hypothèse initiale reliant les ADRs aux protéines en fonction des médicaments interagissant avec elles ajoute beaucoup de bruit de fond puisque certaines protéines sont ciblées pour leur effet thérapeutique. Les interactions multiples (un effet indésirable causé par plusieurs interactions) n'ont également pas été étudiées et notre analyse propose cet aspect combinatoire qui à mon sens est plus proche de la réalité. L'utilisation de Neo4J rend l'ajout de nouvelles données transcriptomiques ou toxicogénomiques intéressant et rapide et pourrait apporter de l'information sur l'effet de ces médicaments à un niveau cellulaire ou génomique.
Enfin, l'étude des CNVs permet de mettre en lien certains effets toxiques avec une augmentation ou délétion de certains gènes limitant ou favorisant la réplication de protéines. Associée avec des études sur le polymorphisme génétique, cela permettrait d'avoir une étude approfondie d'effets indésirable en fonction de mutations.
La programmation nécessaire à de telles analyses est totalement reproductible en se basant sur les mêmes données car aucune composante aléatoire n'est présente.

Chapitre 4 : Rôle du polymorphisme génétique dans les effets indésirables

4.1 Le Polymorphisme génétique

Un autre paramètre pouvant expliquer la présence d'effets indésirables est la variabilité génétique de l'être humain. Le sujet a commencé à être exploré dans le chapitre précédent et va se poursuivre avec cette dernière étude sur un autre type de mutation très fréquente chez l'humain, les SNPs (fréquence >1%) [Schork N.J., *et al.* 2000]. Comme leur nom l'indique il s'agit de la mutation d'une seule base nucléotidique pouvant causer plusieurs types d'effets (figure 9).

Contenu retiré pour des raisons de droits d'auteurs

Le premier est une mutation ne causant aucun changement de résidu lors de la traduction, c'est une mutation dite "synonyme". Le code génétique étant redondant, plusieurs combinaisons de triplets peuvent mener à un même résidu (e.g UUA, UUG sont tous deux traduits en leucine).

Le second arrive lorsque la mutation a un effet lors de la traduction, c'est une mutation non-synonyme. Deux résultats sont possibles : une mutation non-sens et une mutation faux-sens. La première entraine l'apparition d'un codon stop prématuré, tronquant la protéine et la rendant le plus souvent non fonctionnelle. Le deuxième type de mutation non synonyme se traduit par un changement de résidu lors de la traduction, C'est sur ce type de mutation que se portera l'étude. Les effets de ces mutations peuvent être multiples allant d'un effet bénin voire nul, à de graves effets indésirables [Ansari M., *et al.* 2007 ; Ozeki T., *et al.* 2010].

4. 2 Réseau neuronal profond

Les réseaux de neurones sont des sous-catégories de l'apprentissage automatique, utilisant énormément d'approches différentes. L'une d'entre elles, les réseaux de neurones profonds, est beaucoup utilisée en se basant sur la reconnaissance d'images pour prédire la toxicité ou les intéractions de potentiels candidats [Ciregan D., *et al.* 2012; Ibragimov B., *et al.* 2018]. Elle peut également être utilisée en combinant des approches de QSARs, de calculs de propriétés et

de descripteurs pour prédire des interactions entre protéines [Zhang L., *et al.* 2019]. Les résultats obtenus sont souvent très bons malgré un risque de sur-apprentissage dépendant des données, d'où le développement de ces approches avec le développement des outils informatiques toujours plus puissants. Le réseau de neurones profond possède une structure se composant de trois types de couches (figure 10).

Contenu retiré pour des raisons de droits d'auteurs

Une couche d'entrée, permettant d'introduire les données, comporte une ou plusieurs couches cachées qui vont récupérer les données avec un poids ajouté pour calculer une valeur temporaire, soumise à une fonction d'activation. Cette fonction va s'activer, ou non, à partir d'un certain seuil et seuls les nœuds activés vont se propager sur les couches cachées suivantes. Enfin la couche de sortie reçoit la prédiction finale et la compare avec les résultats du jeu d'entraînement. C'est un processus de propagation avant (*forward propagation* en anglais*)*. Si la prédiction est mauvaise le réseau va apprendre de ses erreurs et corriger ses différents poids en fonction de la marge d'erreur qu'il avait obtenue et réapprendre le modèle : c'est de la rétropropagation (*backward propagation* en anglais). Ces mécanismes permettent au réseau d'optimiser lui-même ses paramètres en pouvant résoudre des problèmes non linéaires, d'où son intérêt et les résultats encourageants souvent obtenus. La mise en place de réseaux de neurones similaires avec des variations sur le nombre de couches cachées va donc être mis en place pour échantillonner les modèles permettant le mieux de prédire l'effet d'un composé sur un SOC.

Prediction of adverse drug reactions due to gene polymorphisms using deep neural networks.

Bryan Dafniet[1], Olivier Taboureau[1*]
[1] Université de Paris, INSERM U1133, CNRS UMR 8251, 75006, Paris, France
[*] Corresponding author: olivier.taboureau@u-paris.fr

Abstract

The development of drugs is a long and costly process, often limited by the toxicity and adverse drug reactions (ADRs) caused by drug candidates. Even on the market, some drugs can cause strong ADRs that can vary depending on an individual polymorphism. The development of Genome-wide association studies (GWAS) allowed the discovery of genetic variants of interest that may cause these effects. In this study, the objective was to investigate a deep learning approach with the aim to predict genetic variations potentially related to ADRs. We used single nucleotide polymorphisms (SNPs) information from dbSNP to create a network based on ADR-drug-target-mutations and extracted matrixes of interaction to build deep Neural Networks (DNN) models. These DNNs predicted the association of a compound with a potential adverse effect category based on the MedDRA System Organ Classes (SOCs) with an average balanced accuracy of 0.61 having only information about mutations as variables. Including molecular fingerprints representing structural features of the drugs didn't improve the performance of the models. To our knowledge, this is the first model that exploits DNN to predict ADR-drug-target-mutations. Although some improvements are suggested, these models can be of interest to analyze multiple compounds over all of the genes and polymorphisms information accessible and thus pave the way in precision medicine.

Introduction

Adverse drug reactions (ADRs) are of major concern in drug safety as it accounts for 5% to 7% of all hospitalized individual and represents the fifth most common cause of death in hospitals[1-3]. Although several factors including age, gender, and polypharmacy contribute to the occurrence of ADR, it has been reported that the genetic polymorphism of an individual can affect the pharmacokinetics or pharmacodynamics of a drug and induce ADRs[4]. Pharmacogenomics account for ≈80% of drug efficacy and safety variations and about 60% of patients are exposed to potential ADRs[5].

Genome-wide association studies (GWAS) provide an approach for the discovery of possible mechanisms and pathways underlying human characteristics, diseases, and chemical responses[6]. Furthermore, it has been shown that pharmacogenomic studies can identify genetic variants that lead to variations in chemical activity and can cause adverse drug reactions and death by overdose. This is, for example, the case of the genetic variants in CYP2C19, which help the choice of antiplatelet drugs and their dose[7] or CYP2D6 which is involved in the metabolism of many antidepressants and the modulation of the drug efficacy[8].

For two decades, pharmacogenomics analysis has been collected in the Pharmacogenomic Knowledgebase (PharmGKB), an integrated knowledge resource that integrated pharmacogenomics studies from scientific literature, drug labels, and clinical guidelines and allows to understand how genetic variations contribute to the modulation in drug response[9]. Globally, the database contains genetic variation annotation for 1700 genes associated with 784 drugs. However, despite the growing effort on pharmacogenomics, associations between genetic variants and drug responses remain a huge challenge in predisposition to ADRs. Furthermore, pharmacogenomics study is relatively expensive and the development of computational tools that could assess the risks of ADRs caused by genetic variation would be of interest in both clinical medicine and precision medicine. Existing tools like PolyPhen-2[10] and SIFT[11] started to be developed to predict the functional effects of non-synonymous SNPs in the human genome. Schärfe *et al.* (2017)[12] studied the functional genetic variability in drug-related genes, using exomes from more than 60,000 individuals and over a thousand FDA-approved drugs. Virtual Pharmacists connected SNP data with mined information from several databases and showed the interaction between drugs and SNPs[13].

Here, our objective was to use deep learning approaches to predict if a compound could induce ADR due to genetic variations. To perform such analysis, drug targets, drug-ADR, ADR-Systems/Organs, and Single Nucleotide Variants (SNVs) on these targets were collected, cleaned, and concatenated in a matrix. Then, deep learning models were developed to evaluate for each system/organ, the SNVs that are more involved in ADRs and to predict for any drug candidates the potential risk of associated ADRs due to SNVs. The results obtained from our analysis are presented and discussed below.

Materials & Methods

Data integration

Drug-protein-ADR

DrugCentral is a drug information resource, including the Food and Drug Administration (FDA) and Adverse Event Reporting System (FAERS). It contains information on interactions for 2068 molecules with bioactivities on 1910 genes and ADR information. Also, the ADRs are classified by 27 systems organs codes based on the Systems Organs Classification (SOC). This information will be considered in our analysis We included information from DrugBank[14] v5.1.5 which contains 11 355 molecules with bioactivities on 3510 proteins to complete our drug-protein-ADR repository.
Overall, by merging the information from both databases, and by considering only nodes fully linked (ADR-Drug-Target), we obtained 1090 drugs (with their chemical structure represented in SMILES), 1999 proteins, and 6180 ADRs.

dbSNP

To collect the mutations detected on human proteins from previous human genomic studies, we used dbSNP[12] which contains more than a million references on mutations for the Homo Sapiens. Only Single Nucleotide Polymorphism (SNP) and Single Nucleotide Variation (SNV) with a frequency of mutation equal to or superior to 0.01 were used. We integrated all the SNV and SNP mutations related to the 1999 drug targets concatenated previously to obtain at the end 5050 mutations susceptible to having an impact on the drug-ADR relationships.

Neo4J®

The use of Neo4J®(https://neo4j.com/), a high-performance NoSQL database, allowed us to integrate data from numerous sources and to create our database from which we extracted the interactions information we will use in this article. The database was already created[16] and we only added information from DrugCentral[17] and dbSNP.

Data preparation

The data frames and general handling of data were done using R v4.1.2 and Python v3.9.12.
We took mutations with a frequency higher than 0.01 and molecules only related to proteins having a mutation, and having an inchikey for the final data leaving us with 1029 molecules (in a matrix row) for 2893 mutations (in a matrix column) to study. Every drug is annotated with a SOC if it is connected to an adverse drug reaction happening in its category, with the possibility to be associated with several SOCs.
For each drug-protein, when an interaction was reported, a value of 1 was integrated into the matrix for the drugs interacting with that protein's mutations, otherwise, a value of 0 was put (defined as Y variables) (table 1).

		Mutations			
		Mut 1	Mut 2	…	Mut Y
Compounds	Cpds 1	0	1	…	0
	Cpds 2	1	0	…	1
	…	…	…	…	1
	Cpds X	0	0	…	0

Table1: Example of a matrix template integrating the presence or absence of a drug-target/mutation interaction

For the machine learning models, the dataset was split 4/5 for the training set and 1/5 for the test set. As datasets were imbalanced, weights were dynamically added with the *Sklearn*[18] package for each SOC to prevent heavy bias for the model to predict a drug having no effect on that SOC (prediction 0) and still reach a good accuracy despite not being able to predict an effect (prediction 1).

Neural network model

Multiple sequential neural network models have been set up by using the *tensorflow*[19] package v2.7.0 on python to predict the possibilities of a drug being associated with adverse drug reactions (ADRs) and organs, based on the system organ class (SOC) while taking mutations (single nucleotide polymorphism and single nucleotide variants; SNVs & SNPs) into account. Three different architectures have been implemented as follow. The first used the SNPs information without chemical descriptors to see if knowledge-based only would be sufficient to predict a SOC. The second mixed SNPs with RDKit 1D and 2D descriptors amounting to 72 new variables. The third only contained RDKit descriptors. All the neural networks were created using a grid search on batch sizes and epochs, as well as hidden layers and dropout combinations. Multiple architectures were tested ranging from 5 hidden layers with 256 nodes, reducing the number of nodes by half in each following layer, to one with 16. One or two

dropouts of 0.2 were added and always in the layers directly following the first hidden layer. For example, if a NN started at 256 nodes and had every subsequent layer halving their number of nodes compared to the previous layer (256, 128, 64…), and if two dropouts are added there would be one between the 256 nodes and 128 nodes layers, and one between 128 nodes and 64 nodes layers. Following the different layout combinations, 12 NN were created covering all combinations of layers and dropouts possible (table 2).

Hidden Layers nodes	16	32	64	128	256
Possible dropouts	0	0-1	0-1-2	0-1-2	0-1-2

Table 2: Possible NN layout combinations with every number of dropouts

For every model, a grid search has been performed by varying the batch size and the number of epochs to find the combination giving the best accuracy (table 3).

Batch_size	5	10	20	32	64
Epochs	50	100	150	200	

Table 3: Grid search parameters for neural networks

The batch size here was the number of samples that were forwarded in the neural network at once, specifically how many compounds will go at the same time for predictions. With a batch size of 5, the model was trained 5 and adjusted 5 compounds at a time until reaching the last one. An epoch is how many times all of the compounds will go through that process. As the weights are set randomly initially and modified while training the higher the number of epochs the higher the combination of weights explored, possibly leading to better predictions. The downside is the increase in computational power and time required.
To select the best NN for further analysis we extracted balanced accuracy for each SOC's model to estimate if there were any significant differences between the groups.

Mutations analysis

Once the NN is selected the top mutations can be analyzed to show possible important mutations leading a compound to cause adverse effects related to a SOC. We selected every true positive compound for every SOC and calculated the number of occurrences a mutation had to calculate the frequency for each of them. It gave us an idea of which mutation was prioritized for the NN.
This step concluded the study by introducing a prediction method for the impact of a drug on a SOC (figure 1). This method can be refined to predict one level lower, on adverse drug reactions themselves.

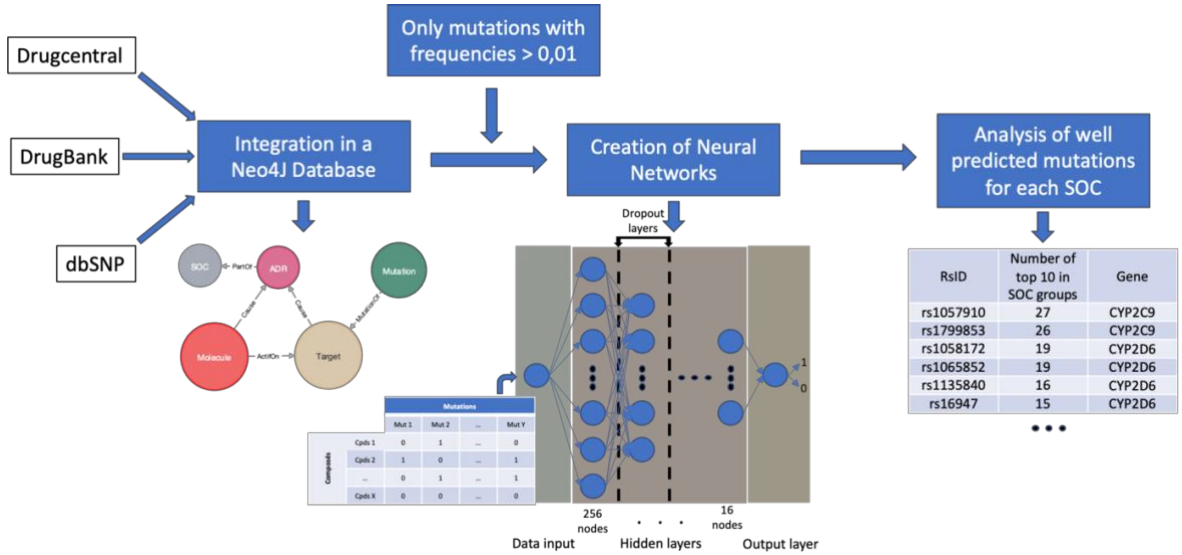


Figure 1: Summary of the protocol used in the article from data collection to mutation analysis

Results & Discussion

Three matrixes were created for the three different methods and if a molecule was binding to a protein associated with one or more mutations it was written as 1s in the table for these mutations, 0 otherwise (figure 2).

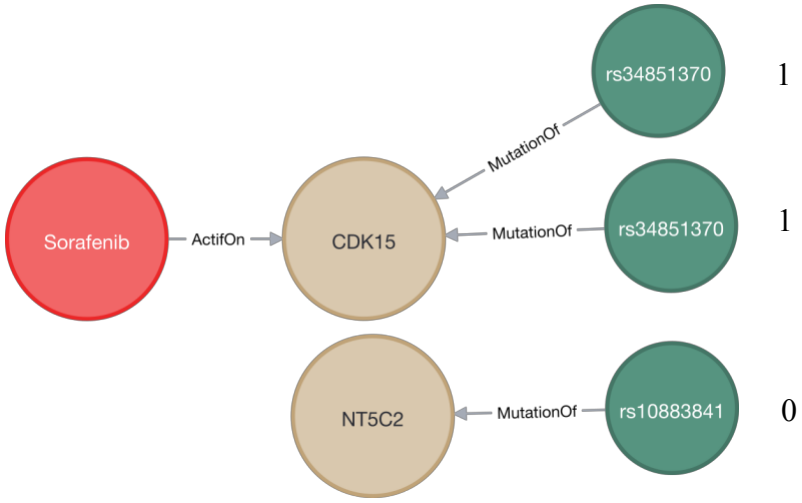


Figure 2: Explanation of how the matrix was built, molecules in red, target in brown, mutations in green, 1 if the mutation is linked to a target bound by a molecule, 0 otherwise.

Splitting the dataset gave us 823 compounds on the training set and 206 for the test set. Class weight was set on “balanced” and the grid search was performed with the parameters given in table 1 and 5-fold cross-validation.

Neural Network

When using the matrix with only information on mutations, the average balanced accuracy (BA) was 0.71 for the training sets and 0.61 for the test sets when averaging all NN combinations and all the SOCs. When the descriptors were added to the mutations making a mixed model the averages for the training and test sets were the same. When only the descriptors were taken into account, with the same filters as previously described, the Neural Network failed to predict most of the SOC, barely averaging at 0.55 for the training sets and 0.54 for the test sets barely predicting better than random and in numerous NN models failing to do so. Therefore, this network will not be considered in the next analysis.

Permutation test

Since normality and variance equality were not verified and since the sample size was small, the t-test was not an option hence the permutation test was chosen to evaluate the robustness of the models. Given the results for the "No descriptor" and the mixed models, we took one model in each category having the best predicting results. For the models without descriptors, the best model predicting the highest number of SOCs with the best results was the one starting at only 16 nodes (table 4).

SOC	BA_train	Sens_train	Spe_train	BA_test	Sens_test	Spe_test
Blood	0.75	0.77	0.72	0.73	0.75	0.71
Card	0.72	0.81	0.64	0.59	0.69	0.50
Cong	0.73	0.80	0.67	0.59	0.62	0.57
Ear	0.73	0.86	0.59	0.60	0.69	0.51
Endo	0.69	0.89	0.50	0.63	0.81	0.46
Eye	0.69	0.80	0.57	0.56	0.63	0.49
Gastr	0.72	0.77	0.67	0.66	0.71	0.60
Genrl	0.74	0.74	0.73	0.62	0.59	0.64
Hepat	0.74	0.84	0.64	0.66	0.72	0.59
Immun	0.69	0.84	0.54	0.63	0.73	0.52
Infec	0.67	0.67	0.68	0.63	0.62	0.63
Inj&P	0.73	0.78	0.68	0.55	0.66	0.44
Inv	0.69	0.86	0.51	0.59	0.80	0.38
Metab	0.77	0.73	0.82	0.61	0.54	0.68
Musc	0.73	0.80	0.67	0.56	0.60	0.53
Neopl	0.75	0.86	0.63	0.74	0.82	0.66
Nerv	0.73	0.75	0.72	0.62	0.71	0.53
Preg	0.69	0.93	0.46	0.55	0.70	0.40
Prod	0.62	0.79	0.45	0.49	0.53	0.45
Psych	0.70	0.80	0.60	0.66	0.71	0.62
Renal	0.71	0.82	0.60	0.64	0.76	0.52

Repro	0.72	0.87	0.56	0.56	0.67	0.45
Resp	0.72	0.77	0.67	0.61	0.68	0.55
Skin	0.73	0.82	0.63	0.62	0.67	0.57
SocCi	0.72	0.92	0.53	0.57	0.66	0.48
Surg	0.73	0.84	0.62	0.57	0.64	0.51
Vasc	0.74	0.81	0.68	0.66	0.72	0.60

Table 4: Detailed results of the NN with one layer at 16 nodes and no dropout for every SOC on the training and test set. Sensitivity and specificity are included as well.

For the mixed models, the neural network was the one without dropout and starting at 256 nodes (Suppl. 1). It led to no significant mean differences for SOCs when we compared the NN with mutations and the one mixed with descriptors.

Selecting one NN model is complex for results may greatly differ depending on the number of layers, nodes and dropouts selected. Overall, the simplest one, with one layer of 16 nodes and no dropout, has the maximum balanced accuracy in 7 SOCs and one of the highest global BA among the NN models (Suppl. 2). Among the SOCs, the prediction accuracies vary from the highest for "Blood" or "Neopl" at 0.73 and 0.74 respectively to the lowest "Prod" which is the only one predicted below 0.5 BA at 0.49 (table 5).

An interesting detail to note for these predictions is that sensitivity seems to be better predicted than specificity. One hypothesis could be that while we know if a compound binds to a target, the 0s in the matrix also come from a lack of information on drug-target interactions. Interactions that might happen with specific targets causing adverse effects might have not been tested, bringing here a limit, or at least an axis of development for this method.

Mutations analysis

We end up with mutation tables giving us insights on which ones are the most represented for every SOC. For example, in the "Cardiac Disorders" SOC, 11 mutations have the same frequency of 0.4 among the well-predicted compounds and we can mostly see mutations affecting cytochromes of the 2C and 2D families (Suppl. 3).

To see if that happens in every SOC, we extracted every top 10 mutations and counted the number of times they appeared in the top 10 among every SOC (table 5).

RsID	Number of top 10 in SOC groups	Gene
rs1057910	27	CYP2C9
rs1799853	26	CYP2C9
rs1058172	19	CYP2D6
rs1065852	19	CYP2D6
rs1135840	16	CYP2D6
rs16947	15	CYP2D6
rs28371703	15	CYP2D6
rs28371704	15	CYP2D6
rs10509681	12	CYP2C8

rs1058930	12	CYP2C8
rs28371706	12	CYP2D6
rs11572076	11	CYP2C8
rs11572080	11	CYP2C8
rs11572103	10	CYP2C8
rs2275622	10	CYP2C8
rs3915951	10	CYP2D6
rs17878459	8	CYP2C19
rs11045819	5	SLCO1B1
rs3758581	5	CYP2C19
rs2306283	3	SLCO1B1
rs2231137	2	ABCG2
rs2231142	2	ABCG2
rs10841795	1	SLCO1A2
rs11568563	1	SLCO1A2
rs34671512	1	SLCO1B1
rs4149056	1	SLCO1B1
rs769258	1	CYP2D6

Table 5: Number of occurrences for mutations present in the top 10 of a SOC group among the 27 existing.

Cytochromes are the main drug targets, logically, a mutation would impact drug assimilations leading to adverse effects. We can also note the presence of SLCO and ABC superfamilies in the top occurrences which is expected since they are transporters.

Even if analyzing the mutations for these genes as they can cause issues in drug metabolization we tried to remove genes from the CYP, SLCO, UGT and ABCG superfamilies to prevent metabolism and transporter enzymes that are bound to appear and explore genes specific to one SOC. It gave us more insights when we look at specific SOCs such as “Card”, short for “Cardiac disorders”, where mutations on the gene KCNH2 become predominant with two mutations at 15 occurrences among the well-predicted compounds. It has been shown that mutations in this gene, including two in our list of mutations linked to well-predicted compounds: rs36210421 and rs1805123, can be responsible for variations of QT intervals that may lead to more serious adverse effects[20-21]. Multiple drugs interacting with KCNH2, like fluoxetine which has been well predicted in our SOC group, are linked to this adverse drug reaction when taken either alone or in combination with other drugs[22].

On another SOC “Gastr”, short for “Gastrointestinal disorders”, the gene PDE3A and its polymorphism can cause nausea, especially one mutation we could find on our variable list: rs12305038[23]. We can link this mutation to the well-predicted compounds known to cause nausea and to interact with that target such as sildenafil and dipyridamole[24-25].

The objective of this study was to investigate the performance of Artificial Intelligence in precision medicine. More precisely, predictive models have been developed considering SNPs' impact on adverse drug reactions (ADR). Grouping the ADRs into SOCs, we obtained, on average, a BA of 0.71 for the training set and 0.61 in the test set. Developing a unique deep learning model, including each drug-SNP-SOC in one matrix was attempted with no good

results (BA < 0.5). One reason is that many drugs are related to many ADRs that belong to several SOCs, so the algorithm might not be able to discriminate between SNPs involved in a SOC or not.

There are a lot of improvements that could be investigated further in the optimization of such models. First, in this work, all the reported missense SNVs were considered. However, it is known that many SNPs do not modify the function of the proteins and/or the drug effect or have a small impact on the residue properties (for example Gly to Ala) which do not impact the drug activity. So, reducing the SNP to only those that are known to contribute to individual drug response variability could improve the model performance. Such information is available in PharmGKB although the data is quite sparse. Some tools like SIFT or PolyPhen 2 have been developed to predict if a mutation is damaging or not in the function of a protein. Such information could be included in our model. Also, proteins are involved in protein-protein interactions (PPI), so mutations in a hot spot of a PPI might influence the drug's response. Including SNPs from proteins that are involved in a PPI with another drug target could be beneficial.

Finally, with the progress in structural biology and notably the development of Alphafold that allows the generation of 3D structures of proteins[26], many drug targets can be modelled. The environment around the location of the SNP in a protein can be visualized and analyzed to determine the impact of an SNP on the binding of a drug. For example, it has been shown that the mutation G143E in the carboxylesterase CES1 had an impact on the efficacy of methylphenidate, a drug used for ADHD[27]. Using docking approaches, it has been observed that this mutation led to a weaker interaction between CES1 and the drug and might be an explanation for the weak response to methylphenidate[28].

Overall, there are many ways to optimize the prediction of the SNP's impact on the individual drug response variability and with the boost of technologies in sequencing, we believe that our study will pave the way to determine the risk of ADR led by genetic variations which is a key issue in pharmacogenomics and precision medicine.

Conclusion

With advances in artificial intelligence, high-throughput SNP detection, functional annotation, and fast and cost-effective sequencing technologies, it becomes possible to determine the individual drug response variability and thus the efficacy and ADRs associated with a drug. Although the high polymorphism of drug metabolism enzymes has an important contribution to this variability, a mutation in a drug target directly involved in drug activity can also influence drug efficacy and safety. Knowing the genome of individual computational approaches and Deep Neural Networks can provide insight into individual responses to a drug and ADR risks. It can be also used to predict the possible ADRs for new drugs due to genetic variations.

It could be interesting to take advantage of the available resources and techniques in structural biology to carry out further structure-based studies on drug targets with SNP to give insight into the molecular mechanisms perturbed by this variation. It would improve individual therapeutic outcomes, hence providing clues to personalized drug treatments.

References

1. Lazarou J., Pomeranz B.H., Corey P.N. Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. 1998 *JAMA* 279: 1200-1205. DOI:10.1001/jama.279.15.1200
2. Giardina C., Cutroneo P.M., Mocciaro E., *et al.* Adverse Drug Reactions in Hospitalized Patients: Results of the FORWARD (Facilitation of Reporting in Hospital Ward) Study. 2018 *Frontiers in Pharmacology* 11(9): 350 DOI: 10.3389/fphar.2018.00350
3. Kongkaew C., Noyce P.R., Ashcroft D.M. Hospital Admissions Associated with Adverse Drug Reactions: A Systematic Review of Prospective Observational Studies. 2008 *The Annals of Pharmacotherapy* 42: 1017-1025 DOI:10.1345/aph.1L037
4. Wake D.T., Ilbawi N., Dunnenberger H.M., *et al.* Pharmacogenomics: Prescribing Precisely. 2019 *Medical Clinics of North America* 103(6):977-990 DOI: 10.1016/j.mcna.2019.07.002.
5. Cacabelos R., Cacabelos N. Carril J.C., The role of pharmacogenomics in adverse drug reactions. 2019 *Expert Review of Clinical Pharmacology* 12(5): 407-442 DOI: 10.1080/17512433.2019.1597706.
6. Giacomini K.M., Yee S.W., Mushiroda T., *et al.* Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. 2017 *Nature Review Drug Discovery* 16(1):1 DOI: 10.1038/nrd.2016.234.
7. Scott S.A., Sangkuhl K., Gardner E.E., *et al.* Clinical Pharmacogenetics Implementation Consortium. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450-2C19 (CYP2C19) genotype and clopidogrel therapy. 2011 *Clinical Pharmacology and Therapeutics* 90(2):328-332 DOI: 10.1038/clpt.2011.132.
8. Thuerauf N., Lunkenheimer J. The impact of the CYP2D6-polymorphism on dose recommendations for current antidepressants. 2006 *European Archives of Psychiatry and Clinical Neuroscience* 256, 287–293 DOI: 10.1007/s00406-006-0663-5
9. Gong L., Whirl-Carrillo M., Klein, T.E. PharmGKB, an integrated resource of pharmacogenomic knowledge. 2021 *Current Protocols* 1: e226 DOI: 10.1002/cpz1.226
10. Adzhubei I.A., Schmidt S., Peshkin L., *et al.* A method and server for predicting damaging missense mutations. 2010 *Nature Methods* 7(4): 248-249 DOI: 10.1038/nmeth0410-248
11. Kumar P., Henikoff S., Ng P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. 2009 *Nature Protocols*. 4(7): 1073-1081 DOI: 10.1038/nprot.2009.86
12. Schärfe C.P.I., Tremmel R., Schwab M., *et al.* Genetic variation in human drug-related genes. 2017 *Genome Medicine* 9: 117 DOI: 10.1186/s13073-017-0502-5
13. Cheng R., Leung R.K.K., Chen Y., *et al.* Virtual Pharmacist: A Platform for Pharmacogenomics. 2015 *PLoS One* 10(10): p. e0141105 DOI: 10.1371/journal.pone.0141105
14. Wishart D.S., Feunang Y.D., Guo A.C., *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. 2018 *Nucleic Acids Research* 46(D1):D1074-D1082 DOI: 10.1093/nar/gkx1037
15. Sherry S.T., Ward M.H., Kholodov M., *et al.* dbSNP: the NCBI database of genetic variation. 2001 *Nucleic Acid Research* 29: 308-311 DOI: doi: 10.1093/nar/gkx1037
16. Dafniet B., Cerisier N., Audouze K., *et al.* Drug-target-ADR network and possible implications of structural variants in adverse events. 2020 *Molecular Informatics* 39: e2000116
17. Ursu O., Holmes J., Knockel J., *et al.* DrugCentral: online drug compendium. 2017 *Nucleic Acids Research* 45(D1): D932–D939 DOI: 10.1093/nar/gkw993
18. Pedragosa F., Varoquaux G., Gramfort A., *et al.* Scikit-learn: Machine Learning in Python. 2011 *Journal of Machine Learning Research* 12: 2825-2830
19. Abadi M., Agarwal A., Barham P., *et al.* TensorFlow: Large-scale machine learning on heterogeneous systems. 2015 Software available from tensorflow.org.
20. Engelbrechtsen L., Mahendran Y., Jonsson A., *et al.* Common variants in the hERG (KCNH2) voltage-gated potassium channel are associated with altered fasting and glucose-stimulated plasma incretin and glucagon responses. 2018 *BMC Genetics* 19: 15. DOI: 10.1186/s12863-018-0602-2

21. Pfeufer A., Jalilzadeh S., Perz S., *et al.* Common variants in myocardial ion channel genes modify the QT interval in the general population: results from the KORA study. 2005 *Circulation Research* 96: 693-701 DOI: 10.1161/01.RES.0000161077.53751.e6
22. Wei A., Peng J., Gu Z., *et al.* QTc prolongation and torsades de pointes due to a coadministration of fluoxetine and amiodarone in a patient with implantable cardioverter–defibrillator. 2017 *Medicine* 96(49): e9071 doi: 10.1097/MD.0000000000009071
23. Colombo F., Pintarelli G., Galvan A., *et al.* Identification of genetic polymorphisms modulating nausea and vomiting in two series of opioid-treated cancer patients. 2020 *Scientific Reports* 10: 542 DOI: 10.1038/s41598-019-57358-y
24. Salonia A., Maga T., Colombo R., *et al.* A Prospective Study Comparing Paroxetine Alone Versus Paroxetine Plus Sildenafil in Patients With Premature Ejaculation. 2002 *Journal of Urology* 168(6):2486-2489 DOI: 10.1016/S0022-5347(05)64174-2
25. Jolma P., Ollikainen J., Uurto I. Oral Dipyridamole-Associated Circulatory Collapse. 2018 *Journal of Stroke and Cerebrovascular Diseases* 27(12): 3460-3462 DOI: 10.1016/j.jstrokecerebrovasdis.2018.08.009
26. Jumper J., Evans R., Pritzel A., *et al.* Highly accurate protein structure prediction with AlphaFold. 2021 *Nature* 596, 583–589 DOI: 10.1038/s41586-021-03819-2
27. Zhu H.J., Patrick K.S., Yuan H.J., *et al.* Two CES1 gene mutations lead to dysfunctional carboxylesterase 1 activity in man: clinical significance and molecular basis. 2008 *American Journal of Human Genetics* 82:1241–1248 DOI: 10.1016/j.ajhg.2008.04.015
28. Nzabonimpa G.S., Rasmussen H.B., Brunak S., *et al.* INDICES Consortium. Investigating the impact of missense mutations in hCES1 by in silico structure-based approaches. 2016 *Drug Metabolism and Personal Therapy* 31(2) DOI: 10.1515/dmpt-2015-0034

Supplementary data

NN parameters	Number of max BA
NN_16_0_Dropout	7
NN_64_0_ Dropout	4
NN_64_1_ Dropout	4
NN_128_1_ Dropout	2
NN_128_2_ Dropout	2
NN_256_1_ Dropout	2
NN_32_0_ Dropout	2
NN_32_2_ Dropout	2
NN_128_0_ Dropout	1
NN_256_0_ Dropout	1
NN_64_2_ Dropout	1

Supplementary 1: NN parameters and the number of max predictions for each

SOC	BA_train	Sens_train	Spe_train	BA_test	Sens_test	Spe_test
Blood	0.76	0.75	0.77	0.70	0.67	0.74
Card	0.72	0.63	0.81	0.59	0.49	0.68
Cong	0.71	0.81	0.62	0.64	0.68	0.59
Ear	0.74	0.76	0.72	0.68	0.69	0.66
Endo	0.64	0.73	0.54	0.64	0.76	0.52
Eye	0.72	0.71	0.72	0.65	0.62	0.68
Gastr	0.73	0.84	0.61	0.66	0.80	0.51
Genrl	0.70	0.56	0.84	0.62	0.47	0.78
Hepat	0.75	0.84	0.66	0.65	0.68	0.61
Immun	0.70	0.85	0.56	0.61	0.76	0.46
Infec	0.75	0.78	0.71	0.62	0.64	0.60
Inj&P	0.70	0.47	0.93	0.58	0.33	0.82
Inv	0.73	0.67	0.78	0.69	0.65	0.73
Metab	0.72	0.60	0.84	0.61	0.45	0.77
Musc	0.68	0.82	0.54	0.60	0.72	0.48
Neopl	0.73	0.93	0.52	0.66	0.82	0.50
Nerv	0.71	0.63	0.78	0.63	0.59	0.67
Preg	0.68	0.82	0.55	0.57	0.70	0.44
Prod	0.62	1.00	0.24	0.54	0.92	0.15
Psych	0.72	0.62	0.82	0.64	0.53	0.75
Renal	0.73	0.84	0.61	0.65	0.72	0.59
Repro	0.65	0.98	0.31	0.57	0.93	0.21
Resp	0.69	0.84	0.54	0.61	0.81	0.41
Skin	0.73	0.87	0.58	0.63	0.75	0.52

SocCi	0.75	0.80	0.69	0.63	0.56	0.70
Surg	0.61	0.95	0.27	0.53	0.88	0.18
Vasc	0.72	0.89	0.56	0.62	0.79	0.45

Supplementary 2: Detailed results of the NN with one layer at 256 nodes and no dropout for every SOC on the training and test set. Sensitivity and specificity are included as well.

rsID	Count	Freq	Gene
rs1057910	26	0.4	CYP2C9
rs1058172	26	0.4	CYP2D6
rs1065852	26	0.4	CYP2D6
rs1135840	26	0.4	CYP2D6
rs16947	26	0.4	CYP2D6
rs1799853	26	0.4	CYP2C9
rs28371703	26	0.4	CYP2D6
rs28371704	26	0.4	CYP2D6
rs28371706	26	0.4	CYP2D6
rs3915951	26	0.4	CYP2D6
rs769258	26	0.4	CYP2D6

Supplementary 3: Number of occurrences for a mutation based on the well-predicted compounds having an effect.

The code used for setting up the neural networks is available on github here : http://bit.ly/3hdOja7

4.3 Conclusion

Lors de cette étude la création de réseaux de neurones profonds pour prédire la présence d'ADRs pour un médicament, à l'aide de leur association aux catégories SOC a été implémenté. Une prédiction si large sur un aussi grand éventail de composés n'a, à ma connaissance, jamais été réalisé. Si les résultats de BA sont en moyenne autour de 0.61, la découverte de nouvelles interactions réduisant le biais favorisant les 0 lorsqu'il n'y a aucune interaction connue avec un médicament devrait permettre d'améliorer ces résultats. Ce protocole mis en place est donc extrêmement intéressant pour commencer une analyse et orienter les recherches sur le type d'effet qu'un composé pourrait avoir, et après affinement pourrait même prédire directement les effets indésirables eux-mêmes. Si l'introduction de données physico-chimiques et structurales des composés chimiques n'a pas amélioré les résultats, des informations de transcriptomique pourraient être intéressantes pour ajouter l'expression des gènes en tant que variable pour prédire un effet indésirable. Il est aussi important de noter que l'ensemble des mutations récoltées proviennent de patients qui, probablement, n'ont pas été traités avec l'ensemble des médicaments étudiés et donc des études en pharmacogénomique permettraient d'affiner nos modèles de prédiction.

De nombreuses composantes aléatoires sont présentes sur la programmation, au niveau de la génération des jeux de données d'apprentissage et de test, et au niveau des réseaux de neurones avec les *batch,* néanmoins dans notre étude elles sont « *seedées* » permettant la reproductibilité des résultats.

Conclusion et perspectives

Le développement d'un médicament est un processus long et coûteux entraînant beaucoup d'échecs avant de réussir à mettre un médicament sur le marché. Des milliers de candidats sont rejetés lors des différentes phases pour leur toxicité et les effets indésirables qu'ils provoquent. Pouvoir caractériser les protéines et l'influence de la variabilité génétique sur ces effets permettrait, en plus de l'aspect financier, de proposer une médecine plus sûre et personnalisée. De nombreuses approches *in silico* s'attachent à identifier ces cibles par des axes qui bien que différents, peuvent se combiner. De nombreuses études de ce type se font sur des référentiels spécifiques, comme des familles de protéines, des maladies précises avec les traitements existants comme base. Les approches générales uniquement basées sur les interactions répertoriées, sorte de « méta-études », sont moins courantes et ont pour avantage de ne pas avoir besoin de beaucoup d'informations sur les données recueillies, pas de cristallisation de structure par exemple. Néanmoins ces travaux de thèse ont pu me permettre d'explorer plusieurs approches à chaque niveau : ligand, protéine, génomique, pour tenter d'apporter le plus d'informations et de réponses possibles au sujet des effets indésirables et de l'identification de cibles.

Lors de ma première étude une librairie de composés touchant la majorité des protéines d'intérêt recensées par la chEMBL a été générée, pouvant permettre un screening plus rapide des cibles, et la mise en place de criblage phénotypique moins contraignante pouvant pallier au principal désavantage de cette méthode, qui est son faible débit de criblage. Avec le développement technologique contribuant à réduire le temps mis pour générer les images cellulaires et le nombre restreint de composés dans la librairie pour chaque cible, ce problème peut être évité. De plus, un enrichissement sur les cibles impliquées dans différents processus biologiques et maladies a été effectué, permettant de lier des informations phénotypiques et protéiques intéressantes pour une compréhension approfondie des mécanismes d'action.

Lors de la deuxième étude, l'intérêt a été d'identifier précisément les protéines les plus à même de causer des effets indésirables grâce à la création d'une fonction de score. Un poids va être mis sur chaque protéine étant liée à ces effets, basé sur le nombre total de protéines que touchent les médicaments interagissant avec celles-ci. Cela a permis de diminuer le biais inhérent à l'hypothèse reliant une protéine aux effets secondaires des médicaments qui les ciblent, créant des protéines liées à plusieurs dizaines d'effets indésirables. Le poids sera plus important pour une protéine ciblée par un médicament qui est spécifique et n'interagit qu'avec un nombre limité de protéines causant un effet indésirable. À l'inverse une protéine touchée par plusieurs médicaments qui eux-mêmes interagissent avec de nombreuses protéines aura un poids plus faible, rendant plus difficile l'association directe d'un effet indésirable à une protéine spécifique. Une partie sur le rôle des CNVs a été étudiée pouvant donner de nouvelles pistes intéressantes sur les dosages médicamenteux menant à des effets indésirables lorsque des récepteurs voient leur expression modifiée par des CNVs.

Enfin la dernière partie a permis de continuer l'exploration de l'effet des mutations sur les effets indésirables en étudiant cette fois-ci le type de mutation le plus fréquent, les SNPs. La création d'un modèle permettant de prédire si un composé aura un effet sur une catégorie de SOC uniquement, basé sur ses interactions avec des gènes possédant des SNPs. a été développé. Cela a permis par la suite l'analyse des protéines d'intérêt les plus souvent liées à des effets

indésirables ciblant des catégories spécifiques ; comme le foie, le cœur, le système endocrinien… Cette première étape peut paver la voie de la mise en place de nouveaux axes de recherche sur les liens SNPs-effets indésirables, en prédisant par exemple chaque SOC en même temps, en pouvant prédire quels effets parmi les plus de 8000 présents vont avoir lieu suite à un traitement spécifique en ayant accès au génome de la personne. Cela pourra par la suite permettre un suivi personnalisé et des traitements aux effets indésirables minimes.

Si ces recherches sont intéressantes et apportent des informations pertinentes des limites sont bien présentes ; des axes d'amélioration et/ou des voies d'exploration plus spécifiques existent. Plusieurs problèmes récurrents sur plusieurs études existent, notamment le manque d'information sur les cibles de composés. Les tests d'activité sont effectués sur les cibles d'intérêt et quelques-unes connues comme causant de graves effets indésirables (comme le canal hERG). Les informations d'interaction sont donc réduites et bien souvent un composé est limité par les *assays* effectués menant à une ou deux interactions au maximum. Cette limitation va se ressentir lors de la sélection de composés puisque les profils biologiques vont être peu variés et finalement, les *scaffolds* vont mener à la sélection de composés interagissant avec une seule protéine la plupart du temps, limitant la diversité et le gain d'information. Enfin, sur la dernière étude lors de la mise en place du réseau de neurones dans la matrice les composés interagissant avec des protéines et représentés par des « 1 » sous le SNPs correspondant à cette protéine sont bien expérimentalement vérifiés. En revanche, les « 0 » n'indiquent pas uniquement une non-interaction, mais également le manque d'information pour cette paire médicament-cible. Cela va impacter les performances du réseau en ajoutant du bruit dans le jeu de données pouvant empêcher une prédiction plus précise.

Les approches basées sur la connaissance ne vont faire que se renforcer avec le développement de nouveaux outils accélérant les informations obtenues et vont profiter des études spécifique étudiant des petits jeux de données. Les mécanismes d'action ne vont pas être identifiés par ces approches mais les prédictions et le nombre de relations ne vont faire qu'augmenter en qualité avec l'ajout d'informations, augmentant la robustesse des premiers et la puissance des approches graphiques grâce au second.

Les axes de développement suite à cette thèse sont nombreux de par la diversité des méthodes employées. Rajouter les informations d'interactions protéine-protéine serait intéressant pour rendre le graphique plus connecté et rendre les approches d'analyses de graphe plus pertinentes notamment au niveau de la *betweenness*. Avec les nouvelles techniques de transcriptomique à haut débit, permettant d'analyser la dérégulation de gènes de plusieurs milliers de composés de manière systématique, ces informations pourraient être ajoutées en plus des données protéomiques. Le modèle de prédiction peut également être perfectionné pour les SNPs par l'ajout de paramètres permettant au modèle d'apprendre de manière plus efficiente en incluant des systèmes de récompenses pour les bonnes prédictions. Par la suite l'ajout d'approches de *docking* ou de MD pour observer l'impact des mutations sur l'interaction médicament-protéine peut être intéressant à explorer si la mutation intervient sur un site actif ou à l'entrée d'une poche. Enfin, idéalement, ce modèle servirait de base pour le développement du prochain permettant de prédire les effets indésirables eux-mêmes.

L’étude des effets indésirables, de leur cause et de l’identification de cibles est un axe majeur et passionnant du processus de développement d’un médicament. Durant ces trois années je pense avoir pu contribuer à des travaux permettant le développement de cet axe de recherche, ou au moins d’en poser des bases, menant à une amélioration des traitements et de leurs effets.

Liste des références

Admera Health, http://www.admerahealth.com

Agence Nationale de sécurité du médicament, https://ansm.sante.fr/page/autorisation-de-mise-sur-le-marche-pour-les-medicaments

Ahmad K., Balaramnavar V.M., Chaturvedi N., *et al*. Targeting Caspase 8: Using Structural and Ligand-Based Approaches to Identify Potential Leads for the Treatment of Multi-Neurodegenerative Diseases. 2019 *Molecules* 24(9):1827 DOI: 10.3390/molecules24091827

Alderson T.R., Kay L.E. Unveiling invisible protein states with NMR spectroscopy. 2020 *Structural Biology* 60: 39-49 DOI: 10.1016/j.sbi.2019.10.008

Aldewachi H., Al-Zidan R.N., Conner M.T., *et al.* High-Throughput Screening Platforms in the Discovery of Novel Drugs for Neurodegenerative Diseases. 2021 *Bioengineering* 8(2): 30 DOI: 10.3390/bioengineering8020030

Ansari M., Lugthart S., Evans W.E. pharmacogenomics of acute leukemia. 2007 *Medical Sciences* 23: 961–967 DOI: 10.1051/medsci/20072311961

Aulner N., Danckaert A., Ihm J., *et al.* Next-Generation Phenotypic Screening in Early Drug Discovery for Infectious Diseases. 2019 *Trends in Parasitol*ogy 35(7): 559-570 DOI: 10.1016/j.pt.2019.05.004

Atanasov A.G., Zotchev S.B., Dirsch V.M. Natural products in drug discovery: advances and opportunities. 2021 *Nature Reviews Drug Discovery* 20: 200–216 DOI: 10.1038/s41573-020-00114-z

Balachandra C., Padhi D., Govindaraju T. Cyclic Dipeptide: A Privileged Molecular Scaffold to Derive Structural Diversity and Functional Utility. 2021 *ChemMedChem* 16(17): 2558-2587 DOI: 10.1002/cmdc.202100149

Barabási A.L., Gulbahce N., Loscalzo J. Network medicine: a network-based approach to human disease. 2010 *Nature Reviews Genetics* 12: 56–68 DOI: 10.1038/nrg2918

Boezio B., Audouze K., Ducrot P., *et al.* Network-based Approaches in Pharmacology. 2017 *Molecular informatics* 36(10): 1700048 DOI: 10.1002/minf.201700048

Bouvy J.C., De Bruin M.L., Koopmanschap M.A. Epidemiology of Adverse Drug Reactions in Europe: A Review of Recent Observational Studies. 2015 *Drug Safety* 38: 437–453 DOI: 10.1007/s40264-015-0281-0

Bray M.A., Singh S., Han H., *et al*. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols* 2016 11(9):1757-1774. DOI: 10.1038/nprot.2016.105

Brown S.P., Hajduk P.J. Effects of Conformational Dynamics on Predicted Protein Druggability. 2006 *ChemMedChem* 1(1): 70-72 DOI: 10.1002/cmdc.200500013

Card G.L., Blasdel L., England B.P., *et al.* A family of phosphodiesterase inhibitors discovered by cocrystallography and scaffold-based drug design. 2005 *Nature Biotechnology* 23: 201–207 DOI: 10.1038/nbt1059

CIOMS - Council for International Organizations of Medical Sciences, CIOMS III - Core Clinical Safety Information 1999

Ciregan D., Meier U., Schmidhuber J. Multi-column deep neural networks for image classification. 2012 IEEE Conference on Computer Vision and Pattern Recognition 3642-3649 DOI: 10.1109/CVPR.2012.6248110

Dieppe P.A., Ebrahim S., Martin R.M., *et al.* Lessons from the withdrawal of rofecoxib. *British Medical Journal* 2004 329: 867 DOI: 10.1136/bmj.329.7471.867

DiMasi J.A., Grabowski H.G., Hansen R.W. 2016 Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *Journal of Health Economics*, 47: 20-33 DOI: 10.1016/j.jhealeco.2016.01.012.

Dey S., Luo H., Fokoue A., *et al.* Predicting adverse drug reactions through interpretable deep learning framework. 2018 *BMC Bioinformatics* 19(21): 476 DOI: 10.1186/s12859-018-2544-0

Doncheva N.T., Morris J.H., Gorodkin J., *et al.* Cytoscape StringApp: Network Analysis and Visualization of Proteomics Data. 2019 *Journal of Proteome Research* 18(2): 623-632. DOI: 10.1021/acs.jproteome.8b00702.

Etalab, https://www.etalab.gouv.fr/

Fisher J.A. Feeding and Bleeding: The Institutional Banalization of Risk to Healthy Volunteers in Phase I Pharmaceutical Clinical Trials. 2014 *Science, Technology, & Human Values* DOI: 10.1177/0162243914554838

García-Nafría J., Tate C.G. Cryo-Electron Microscopy: Moving Beyond X-Ray Crystal Structures for Drug Receptors and Drug Development. 2020 *Annual Review of Pharmacology and Toxicology* 60(1): 51-71 DOI: 10.1146/annurev-pharmtox-010919-023545

Ricard Garcia-S., Mestres J. Anticipating drug side effects by comparative pharmacology, 2010 *Expert Opinion on Drug Metabolism & Toxicology* 6(10): 1253-1263, DOI: 10.1517/17425255.2010.509343

Gaulton A., Hersey A., Nowotka M., *et al.* The ChEMBL database in 2017. 2017 *Nucleic Acids Research* 45(D1): D945-D954 DOI: 10.1093/nar/gkw1074

Geer M.I., Koul P.A., Tanki S.A., *et al.* Frequency, types, severity, preventability and costs of Adverse Drug Reactions at a tertiary care hospital. 2016 *Journal of Pharmacological and Toxicological Methods* 81: 323-334 DOI: 10.1016/j.vascn.2016.04.011

Giardina C., Cutroneo P.M., Mocciaro E., *et al.* Adverse Drug Reactions in Hospitalized Patients: Results of the FORWARD (Facilitation of Reporting in Hospital Ward) Study. 2018 *Frontiers in pharmacology* 9: 350 DOI: 10.3389/fphar.2018.00350

Gil C., Martinez A. Is drug repurposing really the future of drug discovery or is new innovation truly the way forward? 2021 *Expert Opinion on Drug Discovery* 16(8): 829-831 DOI: 10.1080/17460441.2021.1912733

Guerriaud M. Droit Pharmaceutique. 2016 *Elsevier Masson*

Gysi D.M., do Valle Í., Zitnik M., *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. 2021 *Proceedings of the National Academy of Sciences of the United States of America* 118(19): e2025581118

Haasen D., Schopfer U., Antczak C., *et al.* How Phenotypic Screening Influenced Drug Discovery: Lessons from Five Years of Practice. 2017 *ASSAY and Drug Development Technologies* 15(6): 239-246 DOI: 10.1089/adt.2017.796

Hancox J.C., McPate M.J., El Harchi A., *et al.* The hERG potassium channel and hERG screening for drug-induced torsades de pointes. 2008 *Pharmacology & Therapeutics* 119(2): 118-132 DOI: 10.1016/j.pharmthera.2008.05.009

Hartwig S.C., Siegel J., Schneider P.J. Preventability and severity assessment in reporting adverse drug reactions. 1992 *American Journal of Hospital Pharmacy* 49(9): 2229-2232

Hauser A.S., Chavali S., Masuho I., *et al.* Pharmacogenomics of GPCR Drug Targets. 2017 *Cell* 172(1-2): 41-54.e19. DOI: 10.1016/j.cell.2017.11.033

Hollingsworth S.A., Dror R.O. Molecular Dynamics Simulation for All. 2018 *Neuron* 99(6): 1129-1143 DOI: 10.1016/j.neuron.2018.08.011

Hollox E.J., Zuccherato L.W., Tucci S. Genome structural variation in human evolution. 2021 *Trends in Genetics* 38(1): 45-58

Honek J. Preclinical research in drug development. 2017 *Medical Writing* 26(4):5-8

Hurgobin B., Edwards D. SNP Discovery Using a Pangenome: Has the Single Reference Approach Become Obsolete? 2017 *Biology* 6(1): 21 DOI: 10.3390/biology6010021

Ibragimov B., Toesca D., Chang D., *et al.* Development of deep neural network for individualized hepatobiliary toxicity prediction after liver SBRT. 2018 *Medical Physics* 45(10): 4763-4774 DOI: 10.1002/mp.13122

International drug monitoring: the role of national centres. Report of a WHO meeting. 1972 *World Health organization technical report series* 498: 1-25

Jang W.D., Jeon S., Kim S., *et al.* Drugs repurposed for COVID-19 by virtual screening of 6,218 drugs and cell-based assay. 2021 *The Proceedings of the National Academy of Sciences* 118(30): e2024302118 DOI: 10.1073/pnas.2024302118

Jarvis J.P., Peter A.P., Shaman J.A., *et al.* Consequences of CYP2D6 Copy-Number Variation for Pharmacogenomics in Psychiatry. 2019 *Frontiers in Psychiatry* 10: 432 DOI:10.3389/fpsyt.2019.00432

Kanehisa M., Goto S. KEGG: kyoto encyclopedia of genes and genomes. 2000 *Nucleic Acids Research* 28(1): 27-30 DOI: 10.1093/nar/28.1.27

Kaufman G. Adverse drug reactions: classification, susceptibility and reporting. 2016 *Nursing Standard*, 30(50): 53–63 DOI: 10.7748/ns.2016.e10214

Kuhn M., Letunic I., Jensen L.J., *et al.* The SIDER database of drugs and side effects. 2016 *Nucleic Acids Research* 44(D1): D1075-10799 DOI: 10.1093/nar/gkv1075

Kim S., Chen J., Cheng T., *et al*. PubChem in 2021: new data content and improved web interfaces. 2021 *Nucleic Acids Research* 49(D1): D1388-D1395 DOI: 10.1093/nar/gkaa971

Kumar S., Sharma P.P., Shankar U., *et al.* Discovery of New Hydroxyethylamine Analogs against 3CLpro Protein Target of SARS-CoV-2: Molecular Docking, Molecular Dynamics Simulation, and Structure–Activity Relationship Studies. 2020 *Journal of Chemical Information and Modeling* 60(12): 5754–5770 DOI: 10.1021/acs.jcim.0c00326

Landrum G. Fingerprints in the RDKit. 2012 *RDKit UGM 2012*

Lim J., Hwang S.Y., Moon S., *et al.* Scaffold-based molecular design with a graph generative model. 2020 *Chemical Science* 11(4): 1153-1164 DOI: 10.1039/c9sc04503a

Lipinski C.A., Lombardo F., Dominy B.W., *et al.* Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. 1997 *Advanced Drug Delivery Review* 23(1-3): 3-26 DOI: 10.1016/s0169-409x(00)00129-0

Ljosa V. Sokolnicki K.L., Carpenter A.E. Annotated high-throughput microscopy image sets for validation. 2012 *Nature Methods* 9: 637

Métivier J.P, Cuissart B., Bureau R., *et al.* The Pharmacophore Network: A Computational Method for Exploring Structure-Activity Relationships from a Large Chemical Data Set. 2018 *Journal of Medicinal Chemistry* 61(8): 3551–3564 DOI: 10.1021/acs.jmedchem.7b01890

Novič M., Tibaut T., Anderluh M., *et al.* The Comparison of Docking Search Algorithms and Scoring Functions: An Overview and Case Studies. 2016 *Methods and Algorithms for Molecular Docking-Based Drug Design and Discovery* DOI: 10.4018/978-1-5225-0115-2.ch004.

Maveyraud L., Mourey L. Protein X-ray Crystallography and Drug Discovery. 2020 *Molecules* 25(5): 1030 DOI: 10.3390/molecules25051030

Mestres J., Gregori-Puigjané E., Valverde S., *et al.* Data completeness—the Achilles heel of drug-target networks. *Nature Biotechnology* 2008 26: 983–984 DOI: 10.1038/nbt0908-983

Mi H., Muruganujan A., Ebert D., *et al.* PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. 2019 *Nucleic Acids Research* 47(D1): D419-D426 DOI: 10.1093/nar/gky1038

Mohamed S.K., Nováček V., Nounu A. Discovering protein drug targets using knowledge graph embeddings. 2020 *Bioinformatics* 36(2): 603–610 DOI: 10.1093/bioinformatics/btz600

Motta S., Callea L., Giani Tagliabue S. Exploring the PXR ligand binding mechanism with advanced Molecular Dynamics methods. 2018 *Scientific Reports* 8(1): 16207 DOI: 10.1038/s41598-018-34373-z

Muratov E.N., Bajorath J., Sheridan R.P. QSAR without borders. 2020 *Chemical Society Reviews* 49(11): 3525-3564. DOI: 10.1039/d0cs00098a

Ozeki T., Mushiroda T., Yowang A., *et al.* Genome-wide association study identifies HLA-A*3101 allele as a genetic risk factor for carbamazepine-induced cutaneous adverse drug reactions in Japanese population. 2011 *Human Molecular Genetics* 20(5): 1034–1041 DOI: 10.1093/hmg/ddq537

Papapetropoulos A., Szabo C. Inventing new therapies without reinventing the wheel: the power of drug repurposing. 2018 *British Journal of Pharmacology* 175(2): 165-167 DOI: 10.1111/bph.14081

Peters, J.U. Polypharmacology – Foe or Friend? 2013 *Journal of Medicinal Chemistry* 56(22): 8955–8971 DOI:10.1021/jm400856t

Polak F.P., Thomas S.J., Kitchin N., *et al.* Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. 2020 *The New England Journal of Medicine* 383(27):2603-2615 DOI: 10.1056/NEJMoa2034577

Proschak E., Stark H., Merk D. Polypharmacology by Design: A Medicinal Chemist's Perspective on Multitargeting Compounds. 2018 *Journal of Medicinal Chemistry* 62(2):420-444 DOI: 10.1021/acs.jmedchem.8b00760

Reddy A.S., Zhang S. Polypharmacology: drug discovery for the future. 2013 *Expert Review of Clinical Pharmacology* 6(1): 41–47 DOI:10.1586/ecp.12.74

Rodriguez-Bussey I.G., Doshi U., Hamelberg D. Enhanced molecular dynamics sampling of drug target conformations. 2015 *Biopolymers* 105(1): 35-42 DOI: 10.1002/bip.22740

Roulet L., Ballereau F., Lapeyre-Mestre M., *et al.* Iatrogénie médicamenteuse : contribution à l'uniformisation de la terminologie en langue française pour la pratique de soins et la recherche clinique. 2015 *Thérapie* 70(3): 283–292 DOI: 10.2515/therapie/2014215

Schork N.J., Fallin D., Lanchbury J.S. Single nucleotide polymorphisms and the future of genetic epidemiology. 2000 58(4): 250–264 DOI: 10.1034/j.1399-0004.2000.580402.x

Schriml L.M., Mitraka E., Munro J., *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. 2018 Nucleic acids research, 47(D1): D955–D962. DOI:10.1093/nar/gky1032

Shannon P., Markiel A., Ozier O., *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. 2003 *Genome Research* 13(11): 2498-2504 DOI: 10.1101/gr.1239303

Shlien A., Malkin D. Copy number variations and cancer. 2009 *Genome Medicine* 1(6): 62 DOI: 10.1186/gm62

Simon R., Wittes R.E, Ellenberg S.S. Randomized phase II clinical trials. 1985 *Cancer Treatment Reports* 69(12):1375-1381

Sinha S., Vohora D. Chapter 2 - Drug Discovery and Development: An Overview. 2018 *Pharmaceutical Medicine and Translational Clinical Research* 19-32 DOI: 10.1016/B978-0-12-802103-3.00002-X

Śledź P., Caflisch A. Protein structure-based drug design: from docking to molecular dynamics. 2018 *Current Opinion in Structural Biology* 48:93-102 DOI: 10.1016/j.sbi.2017.10.010

Subbaiah M.A.M., Meanwell N.A. Bioisosteres of the Phenyl Ring: Recent Strategic Applications in Lead Optimization and Drug Design. 2021 *Journal of Medicinal Chemistry* 64(19): 14046–14128 DOI: 10.1021/acs.jmedchem.1c01215

Suvarna V. Phase IV of Drug Development. 2010 *Perspectives in Clinical Research* 1(2):57-60

Szklarczyk D., Franceschini A., Wyder S., *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. 2015 *Nucleic Acids Research* 43: D447-452 DOI: 10.1093/nar/gku1003

The European Patients' Academy, https://toolbox.eupati.eu/

Tortorici M.A. Veesler D. Structural insights into coronavirus entry. 2019 *Advances in Virus Research* 105: 93-116 DOI: 10.1016/bs.aivir.2019.08.002

Ursu O., Holmes J., Knockel J., *et al*. DrugCentral: online drug compendium. 2017 *Nucleic Acids Research* 45(D1): D932–D939 DOI: 10.1093/nar/gkw993

Vogt I., Mestres J. Drug-Target Networks. 2010 *Molecular informatics* 29(1-2), 10–14 DOI:10.1002/minf.200900069

Veber D.F., Johnson S.R., Cheng H.Y., *et al.* Molecular properties that influence the oral bioavailability of drug candidates. 2002 *Journal of Medicinal Chemistry* 45(12): 2615-2623 DOI: 10.1021/jm020017n

WhatisDNA, https://whatisdna.net/

Wishart D.S., Feunang Y.D., Guo A.C., *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. 2018 *Nucleic Acids Research* 46(D1): D1074-D1082. DOI: 10.1093/nar/gkx1037.

Wu T., Hu E., Xu S., *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. 2021 *The Innovation* 2(3):100141 DOI: 10.1016/j.xinn.2021.100141

Ye Q., Hsieh C.Y., Yang Z., *et al.* A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. 2021 *Nature Communications* 12(1): 6775 DOI: 10.1038/s41467-021-27137-3

Zardecki C., Dutta S., Goodsell D.S., *et al.* PDB-101: Educational resources supporting molecular explorations through biology and medicine. 2022 *Protein Science* 31(1): 129-140 DOI: 10.1002/pro.4200

Zhang L., Yu G., Xia D., *et al.* Protein–protein interactions prediction based on ensemble deep neural networks. 2019 *Neurocomputing* 324: 10-19 DOI: 10.1016/j.neucom.2018.02.097

Zitnik M., Agrawal M., Leskovec J. 2018 Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34(13): i457-i466. DOI: 10.1093/bioinformatics/bty294

Université Paris Cité

LISTE DES ÉLÉMENTS SOUS DROITS

Liste de **tous les éléments retirés** de la version complète de la thèse
faute d'en détenir les droits

Document à intégrer dans la version partielle de la thèse

Illustrations, figures, images...

Légende de l'image	*N° de l'image*	*Page(s) dans la thèse*
Figure 2: Représentation des différentes étapes de développement d'un médicament (https://toolbox.eupati.eu/).	1	12
Figure 4 : Résumé du contenu de la ChEMBL v30 (source https://www.ebi.ac.uk/ChEMBL/).	4	21
Figure 9: Résumé des variations génétiques SNPs et CNVs et des potentiels effets suite à la prescription d'un traitement. (Images tirées de Hurgobin B., et al. 2017, http://www.admerahealth.com/ et https://whatisdna.net/).	9	69
Figure 10: Schéma des différentes parties d'un réseau neuronal profond. Image tirée de https://www.etalab.gouv.fr/	10	70