Méthodes computationnelles pour améliorer les phases primaires de recherche de nouveaux médicaments

Jeremy Grignard

INSTITUT POLYTECHNIQUE DE PARIS

NNT : 2022IPPAX045

Thèse de doctorat

Méthodes computationnelles pour améliorer les phases primaires de recherche de nouveaux médicaments

Thèse de doctorat de l'Institut Polytechnique de Paris
préparée à l'École polytechnique

École doctorale n°626 : l'École Doctorale de l'Institut de Polytechnique de Paris (ED IP Paris)
Spécialité de doctorat : Informatique

Thèse présentée et soutenue à Palaiseau, le 17 juin 2022, par

M. Jeremy Grignard

Composition du Jury :

M^me^ Elaine Del Nery Ingénieure de Recherche, BIOPHENICS, Institut Curie	Présidente du Jury
M^me^ Laurence Calzone Ingénieure de Recherche, Computational Systems Biology of Cancer, Institut Curie	Rapporteur
M. Olivier Dameron Professeur, Dyliss IRISA, Université de Rennes	Rapporteur
M. Amedeo Napoli Directeur de Recherche émérite, Orpailleur LORIA, CNRS	Examinateur
M. Fréderic Pascal Professeur, Laboratoire des Signaux et Systèmes, CentraleSupélec, CNRS, Université Paris-Saclay	Examinateur
M. Demian Wassermann Chargé de Recherche classe normale, PARIETAL, Inria Saclay	Examinateur
M. Thierry Dorval Directeur du département Data Sciences & Data Management, Institut de Recherches Servier	Directeur de thèse
M. François Fages Directeur de Recherche, LIFEWARE, Inria Saclay	Directeur de thèse

Résumé

Le processus de découverte de nouveaux médicaments est long, coûteux et très risqué. L'objectif de cette thèse de doctorat est d'améliorer la pertinence des phases primaires de recherche pharmaceutique en développant des méthodes computationnelles.

La première contribution porte sur le développement du graphe de connaissances Pegasus afin de capitaliser sur les données pharmaco-biologiques hétérogènes et de provenances multiples du secteur pharmaceutique. Les applications industrielles de Pegasus répondent à des problématiques de projets thérapeutiques et permettent de caractériser des effets hors cibles de perturbateurs, de concevoir une nouvelle expérience, et d'identifier des librairies de criblage focalisées.

La deuxième contribution porte sur le développement d'un algorithme d'identification de composés contrôles positifs et d'un algorithme de normalisation afin d'améliorer la conception et l'analyse d'expériences de criblage phénotypiques à haut contenu. Ces algorithmes permettent de normaliser les signatures phénotypiques obtenues à partir de campagnes de criblage et d'intégrer des similarités phénotypiques informatives dans le graphe de connaissances Pegasus.

La troisième contribution porte sur le développement d'un modèle mathématique du cycle de tyrosination des microtubules qui explique, d'une part, l'inactivité de composés chimiques dans les cellules montrés actifs hors cellule, et d'autre part, suggère la nécessité d'activer deux réactions de ce cycle, en synergie, pour obtenir un effet dans les modèles cellulaires. Ceci illustre l'apport de la modélisation mathématique pour, d'une part, prédire et comprendre la dynamique contre-intuitive de processus biochimiques, qui n'est pas représentable par des graphes de connaissances statiques comme dans Pegasus, et d'autre part, guider la conception de nouvelles expériences de criblage.

Les contributions scientifiques et les applications industrielles de cette thèse sont développées dans le cadre des phases primaires de recherche de nouveaux médicaments et ont vocation à s'étendre aux phases cliniques du processus pharmaceutique.

Remerciements

La vie est un chemin semé de rencontres que nous sommes amenés à faire et d'expériences que nous sommes amenés à vivre. La découverte du monde de la recherche n'aurait pas été possible sans la rencontre avec mon premier directeur de thèse. François, merci de m'avoir accueilli dans ton équipe en tant que stagiaire alors étudiant ingénieur. Je te remercie sincèrement pour ton esprit critique, ta volonté inaltérable de creuser les connaissances en profondeur et pour le nom que tu as donné à ton équipe, Lifeware, le logiciel du vivant, qui manifeste notre volonté de montrer que les cellules peuvent être vues comme des machines et les réactions biochimiques qui s'y opèrent comme des programmes. À nous de continuer d'appréhender les fonctions associées aux codes d'un logiciel hautement complexe et qui nous fait douter.

Je tiens à remercier sincèrement mon second directeur de thèse. Thierry, sans toi ce projet doctoral n'aurait simplement pas été possible. Je te remercie pour l'opportunité que tu m'as donné de découvrir ce domaine extrêmement riche mais très complexe qu'est la recherche de nouveaux médicaments. Je te remercie de m'avoir accompagné sereinement en me laissant une très grande liberté tout en me propulsant, non linéairement, dans des champs de recherche qui m'étaient encore inconnus. Je te remercie pour la confiance profonde que tu m'accordes, pour ce que nous sommes en train de bâtir avec l'équipe, et pour l'opportunité de poursuivre cette aventure après la thèse. Je suis convaincu que nous développons des maillons qui font avancer le processus de recherche de nouveaux médicaments.

J'adresse mes sincères remerciements à Laurence Calzone et Olivier Dameron pour avoir accepté d'être rapporteur de mes travaux et de participer au jury. Je remercie également Elaine Del Nery, Amedeo Napoli, Frédéric Pascal et Demian Wassermann pour être examinateur de cette thèse.

Cette expérience de doctorat, formation par la recherche pour la recherche, aurait été tout autre sans la rencontre avec toutes les personnes avec qui j'ai pu travailler et échanger. Je vous adresse ces petits mots pour vous remercier pour ce que vous m'avez apporté scientifiquement et humainement. Nicolas Boisseau pour ta passion pour les technologies informatiques, tes bonnes pratiques et tes paroles positives. Le travail présenté dans le chapitre 2 section 2.2.3 je te le dois car tu m'as poussé à utiliser un framework pour le développement de Pegasus. Sofia Lotfi pour m'avoir accueilli le premier jour, ta bonne humeur communicative et nos collaborations quotidiennes. Le travail présenté dans le chapitre 2 section 2.3.1 suit naturellement tes travaux. François-Xavier Blaudin de Thé et Clotilde Mannoury La Cour pour nos discussions riches sur plusieurs aspects de la biologie notamment ceux concernant le monde à ARN. Les travaux présentés dans le chapitre 2 section 2.3.2 illustrent notre travail. Arnaud Gohier pour m'avoir fait découvrir le monde chimique. Le travail présenté dans le chapitre 2 section 2.3.3 reflète notre collaboration. Pour cette même section, je tiens à remercier Fany Panayi et Patricia Machado pour votre confiance, vos suggestions et les perspectives de Pegasus pour vos activités. Les méthodes du chapitre 3 ont été améliorées au cours du temps à la suite de nombreux échanges avec Shantanu Singh du Broad Institute, de Arnaud Ogier et Nicolas Wiest-Daesslé de l'entreprise Ksilink. Merci pour cette aventure qui continue avec les phases d'analyses de JUMP-CP. Les parties expérimentales du chapitre 3 sont réalisées par Iffat Sumia Khader, Célia Gautier, Selma Abjabi, Émilie Christ et Anne-Laure Ong. Merci pour vos résultats d'expériences sans lesquelles nous ne pouvons pas faire grand-chose de concret. Philippe Delagrange, je te remercie pour nos points scientifiques qui ne devaient pas durer plus d'une heure. Les articles

scientifiques sur les microtubules, autoroutes cellulaires, nous amenaient dans des contrées toujours plus complexes. Le chapitre 4 illustre notre collaboration. Les parties expérimentales du chapitre 4 sont réalisées par plusieurs collègues : Véronique Lamamy, Céline Legros, Séverine Nicolas, et Eva Vermersh, je vous remercie. Les images acquises et vos activités sont essentielles. Un petit mot pour Eva qui fut ma voisine de bureau et avec qui nous étions séparés par un simple calendrier en carton (avant la crise sanitaire, j'entends). Dans la foulée, je remercie Sylvie Magny, Léa Ragonnet et Aurélie Thomas mes anciennes voisines de bureau. Un petit mot pour Sylvie qui mettait le chauffage en hiver pour avoir un bureau dont la température optimale était appréciable à 7h20. Xavier Bernasconi, une rencontre de Qualité. Xavier Scerri et Jean Damiens pour nos discussions autour de cafés sur des modèles de voitures thermiques en voie d'extinction. Merci également de m'avoir fait visiter les locaux de nos robots de criblage, roboticiens que vous êtes. Une pensée chaleureuse pour Stéphanie Castier. Un petit mot pour Jean-Philippe Stephan pour ta vision, ta disponibilité et la confiance que tu m'accordes. Enfin je remercie tous mes anciens collègues du département du criblage : Fernando, Sandrine, Stéphanie, Laurence, Émilie, Anne, Marie-Élodie, Benjamin, Sylvian, Jayson, Sébastien.

J'ai passé la majorité du temps de ma thèse au sein de l'Institut de Recherches Servier mais je tiens à remercier sincèrement et amicalement tous les membres de l'équipe Lifeware pour vos retours systématiquement constructifs et qui ont profondément amélioré mon sens critique et mon esprit scientifique. Merci à Mathieu Hemery, Aurélien Naldi, Anna Niarakis, Éléa Greugny, Marine Collery, Sahar Aghakhani et Natalia Alves. Un petit mot pour Sylvain Soliman pour tes commentaires bienveillants, tes conseils avisés et tous les mots que tu écrivais sur mon tableau de stagiaire. Un petit mot pour Julien Martinelli et Éléonore Bellot pour nos discussions et nos écoles de recherche sur l'Île de Porquerolles et à Marseille, de doux souvenirs.

Je dédie cette thèse à ma famille. Mon père car tu es un modèle d'excellence et de perfectionnement. Une phrase qui m'a particulièrement marqué et qui me suivra : « La créativité est la base commune entre la musique et l'informatique et la capacité de composer s'acquiert et s'apprécie en passant le temps qu'il faut passer ». D'une certaine façon, la recherche se résume à développer notre curiosité et notre créativité afin de finalement composer. Ma mère car tu es un modèle de bienveillance, d'altruisme et d'accompagnement. Tu aspirais à ce que j'entreprenne des études littéraires, puis tu rêvais de me voir docteur en médecine, fortement influencée par la série Urgence par ailleurs, et j'espère te rendre fière avec ce doctorat en sciences qui s'inscrit dans des aspects profonds d'une science qui m'a toujours intéressée : la biologie. Ma sœur pour ton amour inconditionnel. Clarisse, mon amie, pour m'accompagner chaque jour sur les chemins de la vie. À mes grands-pères, partis beaucoup trop tôt, Gienek et Jean-Louis, et mes grands-mères Francizka et Régine. À ma belle-famille pour nos discussions, nos vacances, nos repas et vos accueils chaleureux. À mes amis pour être des sources d'inspirations bien que nous évoluions tous dans des directions différentes.

Cette thèse est sous convention CIFRE. Je tiens à remercier l'existence d'un tel dispositif et toutes les personnes travaillant dans l'ombre qui ont pris part de façon directes ou indirectes à la faisabilité de ce projet doctoral, notamment, les membres de l'ANRT et de l'École polytechnique - Institut Polytechnique de Paris. Ce dispositif CIFRE et les travaux réalisés dans le cadre de cette thèse ont permis mon recrutement dans l'équipe de Thierry pour une durée indéterminée après la thèse. Ce dispositif CIFRE a également permis de tisser des liens étroits avec l'équipe de François que nous souhaitons pérenniser.

Table des matières

Liste des figures

Liste des tables

Liste des algorithmes, requêtes, équations, fonctions et formules

Liste des abréviations

ADN	Acide DésoxyriboNucléique
ARNm	Acide RiboNucléique Messager
ASO	oligonucléotide antisens (*AntiSens Oligonucleotide*)
BIOCHAM	machine abstraite biochimique (*Biochemical Abstract Machine*)
CDT_N	modèle mathématique du cycle de tyrosination des microtubules paramétré pour les cellules neuronales
CDT_P	modèle mathématique du cycle de tyrosination des microtubules paramétré pour les cellules prolifératives
CRN	réseau de réaction chimique (*Chemical Reaction Network*)
GPE	Graphe à Propriétés Étiquetés
HCS	criblage à haut contenu (*High-Content Screening*)
HTS	criblage à haut débit (*High-Throughput Screening*)
JUMP-CP	consortium international pour le profilage cellulaire (*Joint Undertaking in Morphological Profiling – Cell Painting*)
LDA	analyse discriminante linéaire (*Linear Discriminant Analysis*)
miRNA	micro acide ribonucléique (*Micro RiboNucleic Acid*)
ORF	cadre de lecture des transcrits (*Open Reading Frame*)
OWL	langage d'ontologie web (*Ontology Web Language*)
PCA	analyse en composantes principales (*Principal Component Analysis*)
PROTAC	chimères ciblant la protéolyse (*PROteolysis-TArgeting Chimeras*)
RDF	cadre de description des ressources (*Resource Description Framework*)
RDFS	schéma du cadre de description des ressources (*Resource Description Framework Schema*)
siRNA	petit acide ribonucléique (*Small Interfering RiboNucleic Acid*)
SMILES	*Simplified Molecular Input Line Entry Specification*
TCP	*Tubulin Carboxy Peptidase*
TTL	*Tubulin Tyrosine Ligase*
uORF	cadre de lecture en amont des transcrits (*Upstream Open Reading Frame*)

Chapitre 1

Introduction

« La simplicité est la sophistication suprême. »

Léonard de Vinci

Cette thèse de doctorat s'inscrit dans le cadre d'une « Convention Industrielle de Formation par la Recherche » (CIFRE) entre l'Institut de Recherches Servier et l'équipe-projet Lifeware du centre Inria Saclay de l'Institut National de Recherche en Informatique et en Automatique. L'Institut de Recherches Servier est une entreprise pharmaceutique dont la mission est la découverte de nouveaux médicaments pour traiter des maladies humaines. Le processus de recherche et de développement pharmaceutique, illustré en (Figure 1), est très long, très coûteux et très risqué [1]. En moyenne, il faut plus de dix ans et trois milliards d'euros d'investissement entre le début des phases de recherche à la mise sur le marché d'un nouveau médicament [2].

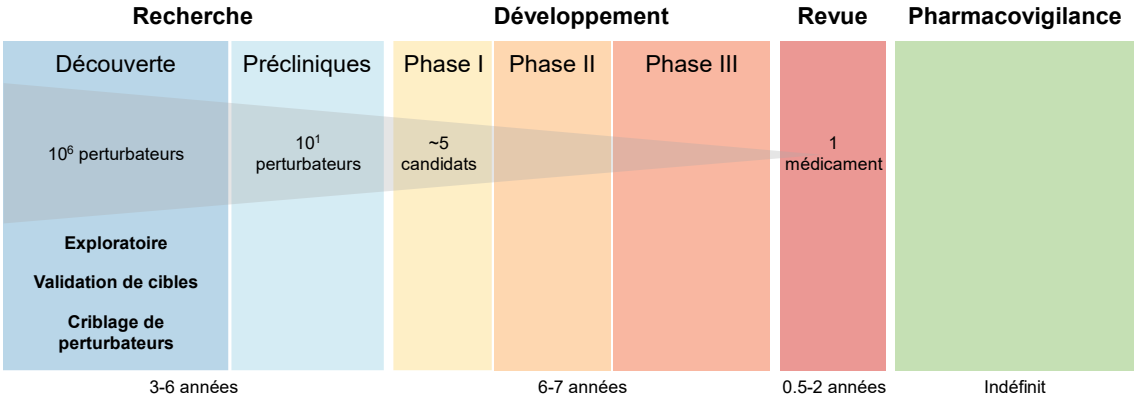

Figure 1. Processus de recherche et de découverte de nouveaux médicaments. Les phases de recherche débutent par la formulation d'une hypothèse de causalité entre la dérégulation d'une cible thérapeutique et une maladie [3]. Une cible thérapeutique est un gène, un transcrit, ou une protéine dont la modulation permet d'obtenir des effets bénéfiques au regard d'une maladie à traiter [4]. Lorsque la cible est identifiée et validée par des méthodes transverses [5], des perturbateurs capables de moduler la cible thérapeutique sont identifiés lors des phases de criblage [6]. Puis, les perturbateurs actifs sont testés sur des modèles animaux lors des phases précliniques. Le développement clinique comprend trois phases. La phase I permet de tester la non-toxicité des candidats-médicaments sur une cohorte de volontaires sains. La phase II consiste à tester le candidat-médicament pour évaluer son efficacité et sa non-toxicité sur des humains présentant la maladie à traiter. La phase III comprend les essais du candidat-médicament afin d'évaluer son efficacité et sa sécurité sur des populations de patients avant que les agences de régulation ne l'approuvent. L'utilisation d'un médicament est surveillée au sein de la population durant l'étape de pharmacovigilance.

Les échecs des projets thérapeutiques en phases cliniques sont critiques car des années de recherche et des investissements considérables ont été réalisés [7]. Trente pour cent des candidats-médicaments qui entrent dans les études de phase II ne progressent pas et plus de cinquante-huit pour cent des candidats-médicaments échouent en phase III [8]. La phase II est critique car c'est à ce stade que l'hypothèse qui lie la cible thérapeutique et le mécanisme de la maladie est réellement mis à l'épreuve. En effet, dans les étapes qui précèdent la phase II, les tests expérimentaux sont réalisés sur des modèles cellulaires et vivants ne présentant pas les caractéristiques réelles de la maladie à traiter.

Les candidats-médicaments n'arrivent pas sur le marché à cause d'un manque d'efficacité ou une toxicité importante [9]. Ces deux indicateurs n'incluent pas d'indications de compréhension de la biologie sous-jacente à la maladie et il est possible qu'un médicament soit approuvé sans que son mécanisme d'action ne soit clairement compris [10]. Malgré les approches de criblage basées sur les cibles thérapeutiques, de nombreux candidats-médicaments n'ont pas d'effets avérés sur ces dernières [10]. Notons en plus, qu'un certain nombre de médicaments sont retirés du marché pour causes d'effets secondaires non identifiés lors des phases cliniques [11]. Les raisons de ces échecs peuvent se résumer, en partie, par une compréhension limitée des processus biochimiques impliqués et une caractérisation incomplète des voies de signalisation engagées par les candidats-médicaments [12].

L'Institut de Recherches Servier transforme depuis plusieurs années son modèle opérationnel de recherche pour pallier les raisons de ces échecs et pour répondre au nouvel objectif stratégique fixé par le Groupe Servier : obtenir une autorisation de mise sur le marché, et ce, tous les trois ans, d'une nouvelle entité ou une combinaison d'entités chimiques ou biologiques. Cet objectif a été défini afin que l'entreprise puisse rester viable et profitable. Le département « Data Sciences & Data Management » a été créé l'année dernière afin de maximiser les chances de succès des candidats-médicaments en déployant des méthodes computationnelles dès les phases primaires de recherche. Les contributions de cette thèse, dont une vue d'ensemble est présentée en (Figure 2), sont alignées avec les activités de mon département et s'intègrent dans les phases exploratoires, de validation de cibles ainsi que de criblage. L'objectif est d'améliorer les phases primaires de recherche de nouveaux médicaments en développant des méthodes computationnelles issues des graphes de connaissances, de la science des données et de la modélisation mathématique, et ce, pour répondre à des problématiques de projets thérapeutiques et exploratoires.

Les données pharmaco-biologiques que nous produisons et exploitons sont hétérogènes et de provenances multiples. Ces données sont silotées et manquent souvent de contextualisation. La conception de graphes de connaissances permet de représenter des données hétérogènes [13]. Ainsi, le premier objectif de la thèse est de développer le graphe de connaissances Pegasus pour capitaliser sur des données actuellement disponibles, et ce, afin de répondre de façon innovante à des problématiques de plusieurs projets thérapeutiques. Nous présentons dans le chapitre 2 le choix du formalisme des graphes à propriétés étiquetés pour ce graphe de connaissances, le modèle de données introduit pour répondre aux problématiques des phases primaires de recherche de nouveaux médicaments, ainsi que l'implémentation et les premières applications industrielles de Pegasus.

Le criblage phénotypique à haut contenu, expérience capturant les phénotypes cellulaires à l'échelle de la cellule unique sous l'effet de traitements, a nourri un regain d'intérêt ces vingt dernières années car ce type d'expérience ne se base pas sur les cibles thérapeutiques [14]. Le développement de nouvelles méthodes computationnelles sont nécessaires pour en tirer tout le potentiel. Ainsi, le deuxième objectif de la thèse est de

développer des algorithmes pour améliorer la conception et l'analyse des expériences de criblage phénotypiques à haut contenu. Nous présentons dans le chapitre 3 un algorithme d'identification de composés contrôles positifs, un algorithme de normalisation de données, et l'intégration de signatures et de similarités phénotypiques informatives au sein du graphe de connaissances Pegasus. Ces résultats sont obtenus à partir d'expériences réalisées en interne et les algorithmes développés seront appliqués sur les données générées dans le cadre du partenariat dans le consortium international Joint Undertaking in Morphological Profiling (JUMP-CP)[1].

Lorsque la dynamique de processus biochimiques, qui n'est pas représentée par des graphes statiques ni par des signatures phénotypiques, est mal comprise, la modélisation mathématique mécaniste permet d'expliquer et de prédire les comportements souvent contre-intuitifs des systèmes biologiques. À cette fin, le Systems Biology Markup Language (SBML) [15], format d'échange standard de modèles, a permis la constitution d'entrepôts de plusieurs milliers de modèles mathématiques de processus biologiques développés manuellement [16]. Le troisième objectif de la thèse s'inscrit dans ces efforts de modélisation afin de développer des modèles mathématiques permettant, d'expliquer les échecs, et guider la conception d'expériences de criblage phénotypiques de composés propriétaires réalisées dans le cadre d'un projet thérapeutique. Nous présentons dans le chapitre 4 deux modèles mathématiques du cycle de tyrosination des microtubules paramétrés d'une part, pour les neurones, et d'autre part, pour les cellules prolifératives, qui expliquent rationnellement les échecs du criblage de composés dans ces modèles cellulaires, identifient de nouvelles cibles thérapeutiques, et permettent de concevoir une nouvelle expérience de criblage par activation de deux réactions en synergie afin d'obtenir un effet.

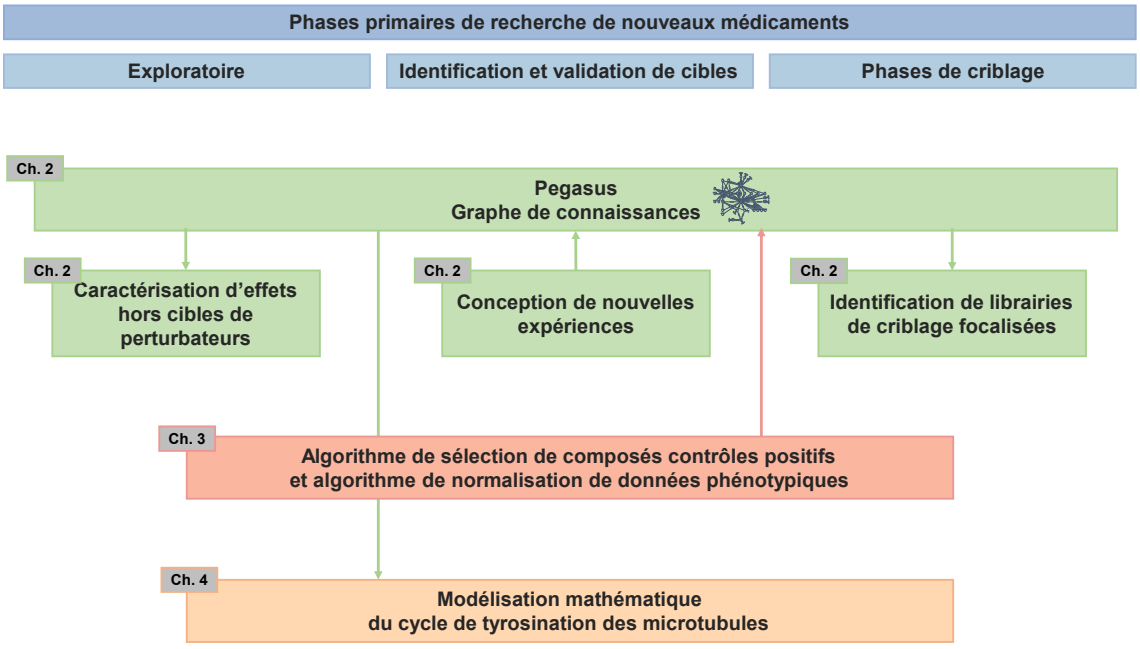

Figure 2. Contributions scientifiques et industrielles intégrées au sein des phases primaires de recherche de nouveaux médicaments. Le chapitre 2 porte sur le développement de Pegasus ainsi que trois applications industrielles. Le chapitre 3 porte sur le développement d'un algorithme d'identification de composés contrôles positifs et d'un algorithme de normalisation de données de criblage à haut contenu. Après normalisation, les signatures phénotypiques obtenues lors de campagnes de criblage sont intégrées

[1] https://jump-cellpainting.broadinstitute.org/

dans Pegasus. Le chapitre 4 porte sur le développement de modèles mathématiques pour expliquer des résultats d'expériences inattendus et identifier de nouvelles cibles. L'utilisation de Pegasus accélère l'identification de perturbateurs pouvant moduler les cibles identifiées par les modèles mathématiques.

Cette thèse de doctorat est une thèse industrielle et mes contributions s'intègrent dans des perspectives pérennes au sein de l'Institut de Recherches Servier. D'une part, l'objectif à long terme des concepts introduits dans le graphe de connaissances Pegasus est de fédérer l'ensemble des résultats expérimentaux et ceux issus de nos algorithmes prédictifs afin d'accélérer les phases primaires de recherche dans leur ensemble. D'autre part, l'analyse de signatures phénotypiques, obtenues dans différentes conditions expérimentales, doit nous permettre d'identifier plusieurs classes de perturbateurs pour moduler des cibles thérapeutiques ou des processus biochimiques. Enfin, la démarche de modélisation mathématique mécaniste vise à s'appliquer aux projets thérapeutiques dès les phases exploratoires afin d'appréhender, le plus tôt possible, la complexité des comportements des systèmes biologiques, souvent contre-intuitifs, en les confrontant aux capacités de prédiction des modèles mathématiques.

Chapitre 2

Pegasus – Graphe de connaissances pour les phases primaires de recherche de nouveaux médicaments

« La science ne renverse pas à mesure ses édifices ;
mais elle y ajoute sans cesse de nouveaux étages et, à
mesure qu'elle s'élève davantage,
elle aperçoit des horizons plus élargis. »

Marcelin Berthelot

Sommaire

2.1 Introduction et motivations

Dans ce chapitre, nous présentons la conception du graphe de connaissances Pegasus pour supporter les phases primaires de recherche de nouveaux médicaments. Pegasus est conçu avec le formalisme des graphes à propriétés étiquetés et est composé de 46.371.784 entités de 66 étiquettes distinctes et de 331.570.883 relations de 14 types distincts.

2.1.1 Contexte – Source de données hétérogènes et de provenances multiples

Les données que nous exploitons sont de nature expérimentale ou prédite par des algorithmes. Les données expérimentales sont issues de domaines scientifiques variés comme les sciences omiques, structurales, cellulaires, chimiques ou phénotypiques, et correspondent à des concepts pharmaco-biologiques hétérogènes. Nous pouvons citer des bases de données traitants de gènes, de transcrits, de protéines [17–21], de perturbateurs chimiques [22–24], de perturbateurs biologiques [25,26], de médicaments [27], de sondes chimiques [28], d'ontologies génomiques et phénotypiques [29,30], de cartes statiques [31–33], d'articles scientifiques [34], de maladies [27], de modèles cellulaires ou animaux [34–36], ou de localisation cellulaire [35]. De plus, nous produisons en interne ou via des consortiums, des données d'activités de perturbateurs modulant des cibles thérapeutiques et des données capturant des phénotypes cellulaires à l'échelle de la cellule unique. Ces données sont issues de campagnes de criblage à haut-débit ou à haut contenu phénotypiques [37,38]. Enfin, à partir de données expérimentales, de la littérature ou de brevets, nous développons des algorithmes prédictifs d'activités et de toxicités de séquences d'acides nucléiques contre des transcrits de cibles thérapeutiques [39].

Nous souhaitons exploiter des sources de données pharmaco-biologiques hétérogènes et de provenances multiples pour répondre aux problématiques des projets thérapeutiques. Ces sources, listées en (Table 1), proviennent du domaine public, de consortiums ou correspondent à des données propriétaires.

Nom	URL	Référence
Bioplex	https://bioplex.hms.harvard.edu/	[31]
Chembl	https://ftp.ebi.ac.uk/pub/databases/chembl/ChEMBLdb/latest/	[22]
ChemicalProbe	https://www.chemicalprobes.org/browse_probes	[28]
Cheng2018	https://www.nature.com/articles/s41467-018-05116-5#Sec17	[32]
CmpoOntology	https://raw.githubusercontent.com/EBISPOT/CMPO/master/cmpo.obo	[29]
Ensembl	http://ftp.ensembl.org/pub/	[17]
Genecode	https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/	[18]
GeneEssentiality	https://www.nature.com/articles/nrg.2017.75	[40]
GeneOntology	http://geneontology.org/docs/download-ontology/	[30]
HGNC	https://www.genenames.org/download/statistics-and-files/	[21]
HumanProteinAtlas	https://www.proteinatlas.org/about/download	[35]
IDR	https://github.com/IDR/idr-metadata	[41]
InternalAsos InternalChemical InternalHTSHCS	Données propriétaires correspondant aux oligonucléotides antisens, aux perturbateurs chimiques, et aux résultats d'expériences de criblage. Les références pointent vers les concepts biologiques et expérimentaux.	[37–39]

MGI	http://www.informatics.jax.org/downloads/reports/index.html	[36]
Mirdb	http://mirdb.org/download.html	[25]
NCBI	https://ftp.ncbi.nlm.nih.gov/	[19]
Opentargets	http://ftp.ebi.ac.uk/pub/databases/opentargets/platform/21.06/output/etl/json/	[23]
Protacdb	http://cadd.zju.edu.cn/protacdb/downloads	[24]
Pubtatorcentral	https://ftp.ncbi.nlm.nih.gov/pub/lu/PubTatorCentral/	[34]
Reactome	https://reactome.org/download-data	[33]
TargetScan	http://www.targetscan.org/cgi-bin/targetscan/data_download.vert80.cgi	[26]
TTD	http://db.idrblab.net/ttd/full-data-download	[27]
Uniprot	https://www.uniprot.org/downloads	[20]
UorfDB	https://www.compgen.uni-muenster.de/tools/uorfdb/downloads.hbi?lang=en	[42]
UorfTool	https://github.com/Biochemistry1-FFM/uORF-Tools	[43]

Table 1. Sources de données pharmaco-biologiques hétérogènes et de provenances multiples en lien avec les problématiques de projets thérapeutiques, et intégrées au sein du graphe de connaissances Pegasus.

2.1.2 Problématique – Comment capitaliser sur des données hétérogènes et de provenances multiples pour améliorer les phases primaires de recherche ?

L'objectif de la conception et de l'implémentation du graphe de connaissances Pegasus est de répondre à des problématiques de projets thérapeutiques. Par exemple, étant donné des cibles thérapeutiques participant dans un processus biologique d'intérêt et dérégulées dans une maladie, comment identifier des perturbateurs chimiques qui ont une activité sur ces cibles, des perturbateurs chimiquement similaires à ces derniers et qui induisent, en plus, des signatures phénotypiques similaires ? Pegasus permet de répondre à ces questions et identifie dans le même temps, les transcrits des cibles modulables par d'autres classes de perturbateurs comme des micro acides ribonucléiques ou des oligonucléotides antisens.

Du fait de leur nature et de leur provenance, les données que nous exploitons sont silotées et manquent de contextualisation. Or, l'ensemble de ces données permettent, avec les connaissances expertes, d'étayer des hypothèses biologiques, de concevoir et d'analyser des expériences, et supportent toutes les étapes des projets thérapeutiques. La mise en relation de données à forte volumétrie est un objectif souhaitable afin de développer des applications innovantes pouvant bénéficier de l'hétérogénéité des données disponibles non encore connectées.

Afin de répondre à cette problématique, nous présentons deux formalismes de représentation des connaissances : les graphes Resource Description Framework (RDF) et les graphes à propriétés étiquetés (GPE). Puis, nous présentons la conception du graphe de connaissances Pegasus qui utilise, étend et introduit de nouveaux concepts au regard des manques de ces formalismes pour nos activités de recherche.

2.1.3 État de l'art – Représentation des connaissances sous forme d'ontologies et de graphes à propriétés étiquetés

Les graphes RDF et les GPE sont deux formalismes de représentation des connaissances qui mettent en relation des données hétérogènes issues du secteur pharmaceutique [13,44]. Ces formalismes permettent d'une part, de représenter et d'organiser, en tout en ou partie, les connaissances de différents domaines, et d'autre part, d'exploiter ces représentations pour répondre à des problématiques spécifiques.

Les graphes RDF représentent les données de façon atomique en les décomposant sous forme de triplets (sujet, prédicat, objet) [45]. Ils sont conçus pour catégoriser, classer et partager différents concepts en essayant de créer un consensus dans la définition des concepts manipulés [45]. Nous présentons en (Figure 3) un modèle de données dans le formalisme RDF[2] qui met en relation des concepts de gènes et de maladies.

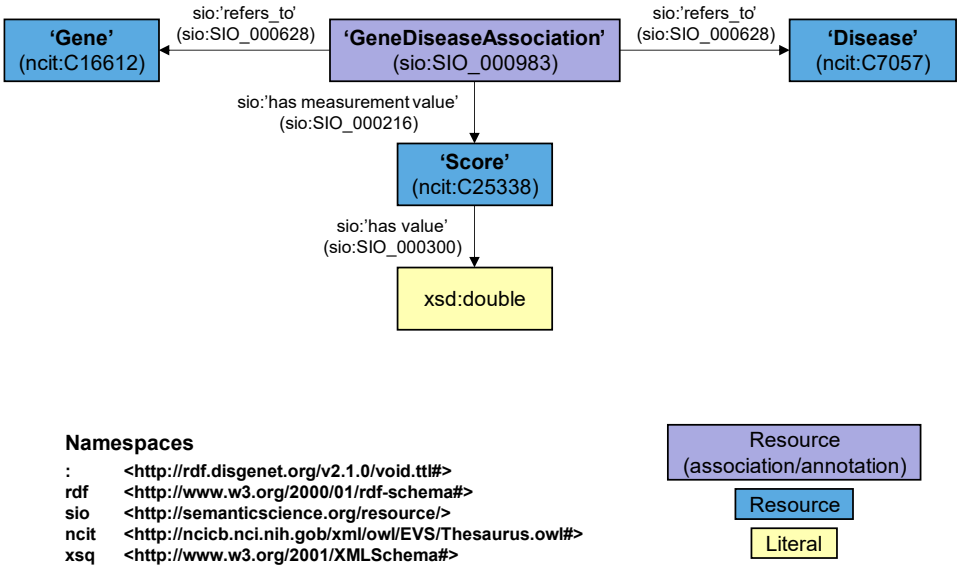

Figure 3. Modèle de données associant des gènes à des maladies dans le formalisme RDF. Ce modèle de données RDF correspond à un sous modèle extrait de DisGeNET [46]. Les rectangles représentent des nœuds RDF (sujets, objets). Les liens entre les nœuds RDF sont des prédicats et décrivent de façon atomique les relations entre des nœuds RDF. Les nœuds et les propriétés sont définis sémantiquement à l'aide d'ontologies standards telle que le thésaurus de l'Institut national du cancer (`ncit`) représentées elles-mêmes par des triplets RDF (sujet, prédicat, objet). Les nœuds `GeneDiseaseAssociation` (`GDA`) sont harmonisés à l'aide de classes Semanticscience Integrated Ontology (`SIO`). Un sujet `GDA` est mis en relation avec les objets (`Gene`, `Disease`, `Score`) par différents prédicats afin de décrire des relations d'association.

Des langages de représentation tels que le Resource Description Framework Schema (RDFS) et le Web Ontology Language (OWL) permettent d'étendre l'expressivité d'un graphe RDF et de concevoir des ontologies [45]. Une ontologie informatique est une spécification formelle des connaissances qui la rend apte à être traitée par des programmes et se base sur un graphe RDF. Cette spécification consiste en une terminologie de termes contrôlés et en des relations sémantiques entre les termes comme des relations de généralité, de spécificité, d'équivalence, ou encore de transitivité [45]. RDFS apporte le concept de schéma et permet de déclarer une ressource RDF comme une classe, définit

[2] Nous changeons de police d'écriture pour décrire les nœuds et les relations issus des graphes RDF et des GPE ainsi que pour l'écriture des requêtes SPARQL et CYPHER.

des hiérarchies de classes et précise les types de propriétés. OWL introduit une sémantique au schéma afin d'inférer de nouvelles connaissances ou de vérifier la cohérence des données. Les concepts de sémantique sont définis par des règles de relations qui peuvent reposer sur la logique de description [47]. Notons que RDFS et OWL sont des langages ontologiques développés initialement dans le cadre du web sémantique et il existe d'autres langages ontologiques basés, par exemple, sur les règles [48].

Plusieurs centaines d'ontologies existent dans le domaine biomédical et utiles pour le secteur pharmaceutique. Nous pouvons citer celles propres aux fonctions des gènes [30], à la description de modalités expérimentales à haut-débit [49], à la description de résultats d'expériences phénotypiques [29], à la description de modèles cellulaires d'études [50], ou à la classification de maladies [51]. Certains consortiums, comme la fondation OBO, centralisent et essayent de rendre interopérables les ontologies du domaine biomédical [44]. D'autres ontologies sont constituées de concepts variés comme les projets Life Sciences Linked Open Data et WikiDataGene [52,53].

Les ontologies biomédicales représentent des concepts en les catégorisant et en les classant, et ce, afin de les partager et pour les requêter. Un des langages de requêtage des graphes RDF est SPARQL [54]. Nous présentons en (Requête 1) une requête SPARQL pour identifier les gènes dérégulés dans la maladie d'Alzheimer à partir du graphe RDF présenté en (Figure 3).

```
SELECT DISTINCT ?gene str(?geneName) as ?name ?score

WHERE {
        ?gda sio:SIO_000628 ?gene, ?disease ;
        ?gda sio:SIO_000216 ?scoreIRI ; # Équivalent à : sio:SIO_000216 ?scoreIRI . (simplification)

        ?gene rdf:type ncit:C16612 ;
        ?gene dcterms:title ?geneName ;

        ?disease rdf:type ncit:C7057 ;
        ?disease dcterms:title "Alzheimer's Disease"@en ;

        ?scoreIRI sio:SIO_000300 ?score ;

        FILTER (?score > 0.4)
}
```

Requête 1. Identification des gènes dérégulés dans la maladie d'Alzheimer avec une requête SPARQL. Cette requête retourne les résultats distincts suivants : l'identifiant du gène (`?gene`), le nom du gène (`?geneName`), le nom de la maladie (`?name`) et le score d'association entre un gène et une maladie (`?score`). Les résultats doivent satisfaire les contraintes définies dans le patron de recherche du graphe inclus dans la clause `WHERE`. Ces contraintes sont des patrons de triplets RDF représentant les contraintes à satisfaire. Dans un patron de triplet, les éléments RDF (sujet, prédicat, objet) sont remplacés par des variables (p. ex. `?gda`). La première contrainte (`?gda sio:SIO_000628 ?gene, ?disease; sio:SIO_000216 ?scoreIRI.`) récupère les objets qui sont de type `Gene`, `Disease` reliés au sujet (`GeneDiseaseAssociation`) par le prédicat (`sio:SIO_000628`) et les objets de type `Score` reliés au sujet (`GeneDiseaseAssociation`) par le prédicat (`sio:SIO_000216`). Plusieurs contraintes de sélection sous forme d'une conjonction de triplets RDF sont présentes dans la clause `WHERE`. Les triplets du graphe respectant l'ensemble de ces contraintes sont retournés. Le mot-clé `FILTER` permet de restreindre les solutions sur l'ensemble du groupe dans lequel les valeurs des objets (qui sont des littéraux) reliés aux sujets `Score` par le prédicat (`sio:SIO_000300`) ont une valeur supérieure à `0.4`.

La principale application des graphes RDF dans le domaine biomédical est le partage d'informations et de données cohérentes. L'implémentation d'un graphe RDF est pertinent lorsque son modèle de données est bien établi, éprouvé par plusieurs

communautés, et que la mise à jour du graphe ne consiste principalement qu'en l'ajout non massive de données. Or, le modèle de données de Pegasus évolue en fonction de nouveaux besoins à couvrir pour les projets thérapeutiques et le formalisme RDF ne permet pas, de façon pratique, une flexibilité de mise à jour du modèle ainsi qu'une intégration efficace et massive des données que nous exploitons. De plus, nous manipulons des concepts pharmaco-biologiques hétérogènes et il est difficile d'identifier une sémantique commune à l'information atomique à décrire, à extraire et à exploiter. Nous ne concevons pas le graphe de connaissances Pegasus avec le formalisme RDF bien que ce dernier soit alimenté de données issues de fichiers d'annotations et de termes provenant d'ontologies (Table 1) [29,30].

Les graphes à propriétés étiquetés sont un autre formalisme de représentation des connaissances. Dans le secteur pharmaceutique, les GPE sont principalement développés avec la technologie Neo4j et nous nous restreignons aux GPE Neo4j. Un GPE permet de concevoir des modèles de données flexibles. En effet, un GPE se base sur quatre concepts simples : entités (nœuds), relations, propriétés et étiquettes. Les entités représentent des objets réels ou abstraits et sont caractérisées par une ou plusieurs étiquettes. Les relations sont typées et décrivent des liens entre des entités sources et des entités destinations. Les entités et les relations sont caractérisées directement au sein de leurs structures respectives. Nous présentons un modèle de données dans le formalisme d'un GPE[3] qui met en relation des concepts de gènes et de maladies en (Figure 4). Ce GPE correspond à une transformation possible du graphe RDF présenté en (Figure 3).

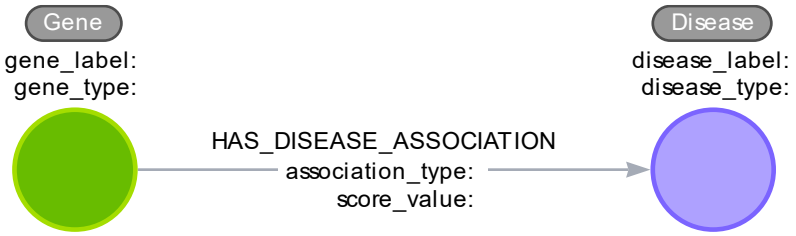

Figure 4. Modèle de données associant des gènes à des maladies dans le formalisme GPE. Les concepts de gènes et de maladies sont représentés par des entités avec les étiquettes `Gene` et `Disease`. Ces entités sont décrites par des propriétés, par exemple, avec un nom de gène (`gene_label`) ou un type de maladie (`disease_type`). L'association entre un gène et une maladie est représentée par la relation typée `HAS_DISEASE_ASSOCIATION`. Cette relation est caractérisée avec des propriétés comme le type d'association (`association_type`) et un score d'association (`score_value`).

Les principales applications des GPE sont l'analyse de graphe, la recherche de chemins en profondeur, l'importation et le stockage massif de données, et servent de support à l'application d'algorithmes d'apprentissage [55]. Les GPE sont conçus pour répondre à des questions spécifiques et nous pouvons citer des GPE utilisés dans le secteur pharmaceutique pour le repositionnement de médicaments [56], l'identification d'effets secondaires liés à la polypharmacologie [57], la facilitation de l'analyse de données cliniques et l'intégration de données omiques [58–60], l'amélioration de la compréhension de maladies [61,62], ou encore l'identification de cibles thérapeutiques impliquées dans des maladies [63].

Le langage de requêtage des GPE Neo4j est CYPHER [64]. Nous présentons une requête CYPHER en (Requête 2) pour identifier les gènes dérégulés dans la maladie d'Alzheimer à partir du GPE présenté en (Figure 4).

[3] Dans la suite du manuscrit, les relations d'un GPE sont écrites en majuscule.

```
MATCH (g:Gene)-[a:HAS_DISEASE_ASSOCIATION]-(d:Disease)

WHERE d.disease_label = "Alzheimer’s Disease" AND a.score_value > 0.4

RETURN DISTINCT g.gene_label, d.disease_label, a.score_value;
```

Requête 2. Identification des gènes dérégulés dans la maladie d’Alzheimer avec une requête CYPHER. Cette requête retourne les résultats distincts suivants : le nom du gène (`gene_label`), le nom de la maladie (`disease_label`) et le score d’association entre un gène et une maladie (`score_value`). Les résultats retournés doivent satisfaire le motif de recherche définit par le mot-clé `MATCH`. Ce motif de recherche permet de récupérer les entités du graphe possédant l’étiquette `Gene` reliées par la relation de type `HAS_DISEASE_ASSOCIATION` aux entités possédant l’étiquette `Disease`. Les entités et les relations peuvent être ancrées par des variables (p. ex. `g`, `a`, `d`). Plusieurs contraintes de sélection dans le motif de recherche sont décrites par la clause `WHERE`. L’ensemble des chemins du graphe possédant les entités `Gene` reliées à l’entité `Disease` qui possède comme valeur `Alzheimer’s Disease` pour la propriété `disease_label` et pour lesquels la relation `HAS_DISEASE_ASSOCIATION` possède une valeur de score associée à la propriété `score_value` supérieure à `0.4`, sont retournés.

La syntaxe du langage CYPHER (Requête 2) est plus claire que la syntaxe du langage SPARQL (Requête 1). En effet, l’identification des prédicats est réalisée nécessairement par un URI (Uniform Resource Identifier URI) ce qui complexifie à la fois l’écriture et la lecture d’une requête SPARQL (Requête 1).

Il existe des différences entre un graphe RDF et un GPE [65–68]. Un GPE permet de décrire directement les entités et les relations par un ensemble de propriétés. Un graphe RDF ne permet pas de caractériser directement les relations (prédicats) entre un sujet et un objet du fait de la description atomique des données bien qu’une extension de RDF, RDF-Star, permet d’ajouter des descriptions aux prédicat d’un triplet RDF. Dans un GPE, plusieurs relations de même type peuvent exister entre deux mêmes entités. Un graphe RDF ne peut pas mettre en relation directement une même paire de nœuds (sujet, objet) par un même prédicat.

Dans le domaine des représentations des connaissances, nous pouvons citer, en plus des graphes RDF et des GPE, les cartes statiques biologiques. Il existe des initiatives de création de cartes pour différentes maladies [69–71], de réseaux d’interactions [32,72,73] ou encore des voies de signalisation [74,75]. Ces représentations utilisent des concepts communs comme des gènes, des maladies, des médicaments et des processus biologiques. L’objectif de ces cartes est de représenter une partie des connaissances statiquement et nous utilisons certaines données issues de ces cartes au sein de Pegasus (Table 1) [31–33].

2.1.4 Manques des représentations existantes et choix de conception du graphe de connaissance Pegasus

Nous avons choisi de concevoir le graphe de connaissances Pegasus avec le formalisme des GPE pour des raisons de flexibilité de mise à jour du modèle de données, d’efficacité d’importation de données, et pour le requêtage déterministe avec le langage CYPHER.

En effet, certains concepts nécessaires à la réalisation de nos activités de recherche sont mal représentés ou absents des sources de données existantes. Par exemple, il n’existe pas de concepts de signatures phénotypiques ni de similarités phénotypiques dans les représentations existantes. Les signatures phénotypiques représentent les états phénotypiques de cellules obtenus par le criblage de perturbateurs [38]. Afin de capitaliser sur les données issues de campagnes de criblage à haut contenu, nous

modélisons les signatures phénotypiques dans Pegasus sous forme d'entités. Ces entités sont reliées par des relations de similarités phénotypiques. De façon analogue, nous intégrons dans Pegasus des similarités chimiques entre perturbateurs sous forme de relations comme il en existe entre des médicaments dans des graphes hétérogènes [76].

Certains perturbateurs qui nous sont d'intérêt sont absents et nous les modélisons dans Pegasus sous forme d'entités comme nos perturbateurs propriétaires, les sondes chimiques [28], les molécules bi-fonctionnelles chimères ciblant la protéolyse (PROTACs) [77], les oligonucléotides antisens (ASOs) [39,78], ou encore les médicaments, les petits acides ribonucléiques (siRNA), les micros acides ribonucléiques (miRNA) ou les molécules issues de la littérature.

De plus, de multiples bases de données identifient des ressources fonctionnellement identiques comme des gènes, des molécules ou des maladies. Par exemple, les bases de référence Ncbi et Ensembl traitent de concepts de gènes, de transcrits et de protéines [17,19]. Cependant, les instituts qui maintiennent ces bases possèdent leurs propres systèmes d'annotations et les nombres de gènes, de transcrits et de protéines annotés sont différents [79]. Les références croisées entre les ressources pharmaco-biologiques ne se recouvrent pas, ne sont pas toujours mises à jour et ne sont pas caractérisées par une association un-à-un entre différents systèmes. Ainsi, pour ne pas perdre d'information par rapport aux sources de données et pour faciliter l'intégration des données que nous exploitons, nous ne modélisons pas un concept, comme un gène, par une entité unique dans Pegasus mais par un ensemble d'entités. Ces entités sont caractérisées par les données issues de plusieurs sources et sont reliées entre elles par des références croisées.

Les concepts de gènes et de protéines sont souvent mélangés et il est commun qu'une protéine soit référencée par un identifiant de gène. Par exemple, dans les cartes d'interactions protéines-protéines, ce sont majoritairement des identifiants de gènes qui sont présents [32]. Notons également que les protéines sont souvent identifiées par des symboles de gènes menant à des résultats d'expériences faussés même en phases cliniques [80]. Or, un gène est différent d'une protéine et l'emploi d'un identifiant correct devrait être systématique. En effet, un gène peut être transcrit en une multitude de transcrits qui peuvent être traduits en différentes isoformes protéiques [81]. En outre, nous considérons les transcrits comme des cibles thérapeutiques d'intérêt et nous développons des algorithmes prédictifs de séquences d'acides nucléiques modulant les transcrits. Nous modélisons les gènes, les transcrits et les protéines par des entités distinctes dans Pegasus. Un graphe de connaissances récemment publié distingue également ces trois entités [60].

Les concepts existants comme des termes provenant d'ontologies, de maladies, de références bibliographiques, de cartes statiques ou de modèles cellulaires d'études sont modélisés dans Pegasus sous forme d'entités.

Enfin, nous intégrons dans Pegasus un nœud intermédiaire pour relier et annoter contextuellement plusieurs entités entre elles.

2.2 Modèle de données et implémentation du graphe de connaissances Pegasus

2.2.1 Modèle de données

Les ressources pharmaco-biologiques fonctionnellement identiques sont modélisées dans Pegasus par plusieurs entités qui sont reliées entre elles par des références croisées (Figure 5). Les ressources qui ne sont pas référencées par un identifiant unique sont

modélisées avec une étiquette principale et des étiquettes secondaires pour indiquer leurs provenances (Figure 5).

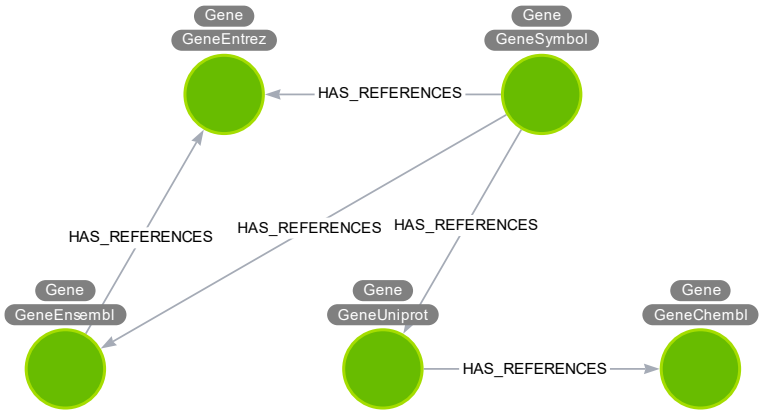

Figure 5. Modélisation de concepts fonctionnellement identiques par plusieurs entités au sein de Pegasus. Des concepts fonctionnellement identiques sont représentés par plusieurs entités et reliées entre elles par des références croisées. Par exemple, un gène est représenté par un ensemble d'entités avec l'étiquette principale `Gene` et des étiquettes secondaires `GeneEntrez`, `GeneSymbol`, `GeneEnsembl`, `GeneUniprot`, `GeneChembl` pour indiquer leurs provenances. Les références croisées sont modélisées par la relation `HAS_REFERENCES`.

Les gènes, transcrits et protéines par trois entités distinctes et les concepts de transcription et de traduction sont modélisés par deux relations dans Pegasus (Figure 6).

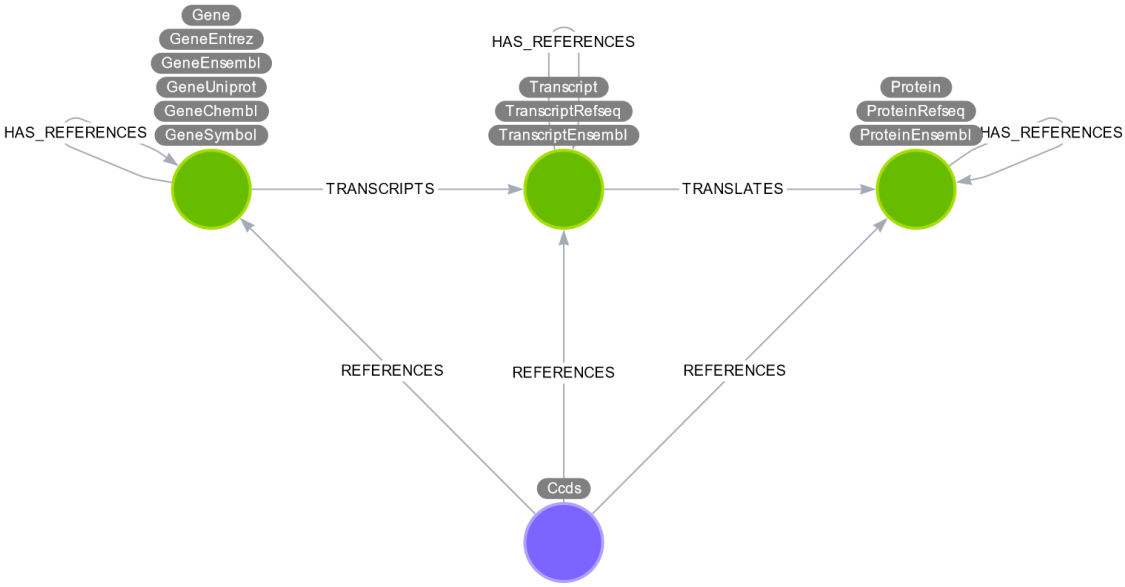

Figure 6. Modélisation des gènes, transcrits et protéines au sein de Pegasus. Les gènes, transcrits et protéines sont modélisés par des entités avec les étiquettes principales : `Gene`, `Transcript`, `Protein` et des étiquettes secondaires (p. ex. `TranscriptRefseq`, `ProteinEnsembl`) pour indiquer leurs provenances. Les transcrits (`Transcript`) et protéines (`Protein`) possèdent des références croisées (`HAS_REFERENCES`). Les concepts de transcription et de traduction sont modélisés respectivement par les relations `TRANSCRITS` et `TRANSLATES`. L'entité `Ccds` est reliée aux entités `Gene`, `Transcript`, `Protein` par la relation `REFERENCES`.

L'entité `Ccds` est introduite dans Pegasus pour identifier une unité fonctionnelle (gène, transcrit, protéine) (Figure 6). L'identifiant de l'entité `Ccds` provient du projet Consensus CDS (CCDS) qui référence de façon unique des unités fonctionnelles (gène,

transcrit, protéine) malgré des identifiants de transcrits et de protéines différents alors que les ressources référencées peuvent être identiques [82].

Dans Pegasus, les différentes classes de perturbateurs sont modélisées par des entités distinctes (Figure 7), les signatures phénotypiques par des entités et les similarités phénotypiques par des relations (Figure 8). Ces deux derniers concepts sont détaillés dans le chapitre suivant.

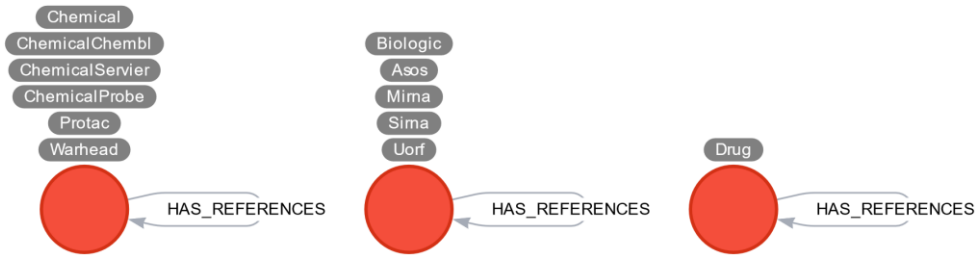

Figure 7. Modélisation de différentes classes de perturbateurs chimiques et biologiques au sein de Pegasus. Les perturbateurs sont modélisés par trois étiquettes principales : `Chemical`, `Biologic` et `Drug`. Ces entités indiquent si le perturbateur est de type chimique, biologique ou si c'est un médicament. Des étiquettes secondaires permettent de préciser la nature de la ressource comme des molécules Servier (`ChemicalServier`), de la littérature (`ChemicalChembl`), des PROTACs (`Protac`) ou des miRNAs (`Mirna`).

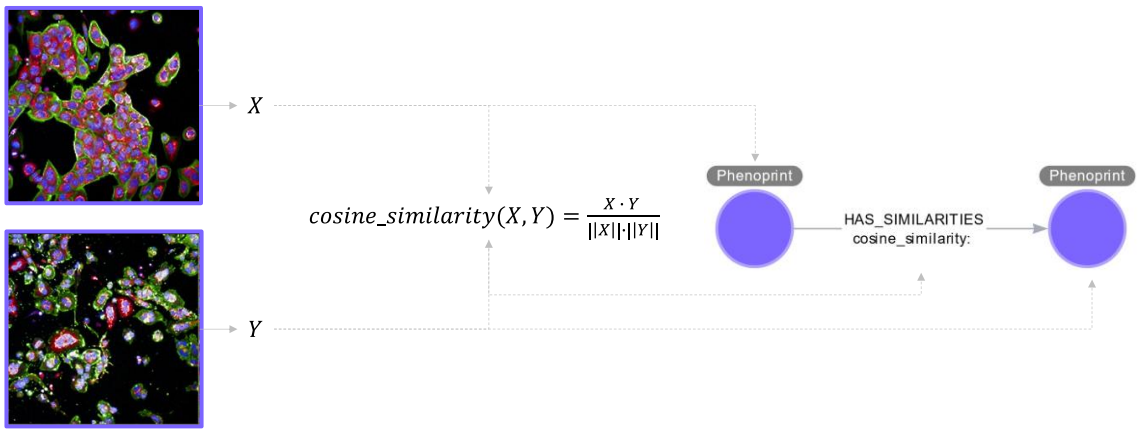

Figure 8. Modélisation des signatures et des similarités phénotypiques au sein de Pegasus. À gauche, des cellules après le criblage de deux perturbateurs. Des étapes d'extraction de descripteurs phénotypiques à l'échelle de la cellule unique, par imagerie, permettent d'extraire des signatures phénotypiques (X, Y). Des mesures de similarités phénotypiques, comme la mesure cosine, peuvent être calculées afin d'identifier des signatures phénotypiques similaires [83]. Les signatures phénotypiques sont intégrées dans Pegasus sous forme d'entités avec l'étiquette `Phenoprint` et sont reliées avec la relation `HAS_SIMILARITIES`. Cette relation possède comme propriétés des mesures de similarités phénotypiques calculées (p. ex. `cosine_similarity`).

De même, les similarités chimiques entre perturbateurs sont modélisées dans Pegasus par des relations (Figure 9).

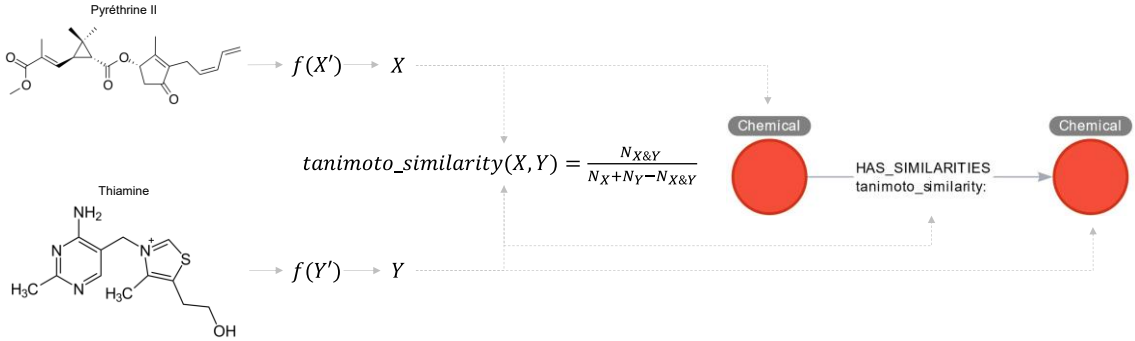

Figure 9. Modélisation des similarités chimiques entre perturbateurs au sein de Pegasus. À gauche, deux molécules chimiques. Chaque molécule possède une représentation Simplified Molecular Input Line Entry Specification (SMILES) [84]. Le SMILES représente une molécule sous forme d'une chaîne de caractères ASCII. Par exemple, la représentation SMILE de la thiamine (Y') est `OCCc1c(C)[n+](=cs1)Cc2cnc(C)nc(N)2)`. À partir des représentations SMILES, une fonction d'encodage f permet de générer des signatures chimiques (X, Y) [85]. Des mesures de similarités chimiques, comme l'indice de Tanimoto, peuvent être utilisées afin d'identifier des molécules chimiques similaires [86]. Des perturbateurs chimiques sont introduits dans Pegasus comme présenté en (Figure 7). En fonction des mesures de similarités chimiques calculées, les perturbateurs de type `Chemical` sont reliés entre eux avec la relation `HAS_SIMILARITIES`. Cette relation possède comme propriétés des mesures de similarités calculées (p. ex. `tanimoto_similarity`).

Les concepts existants comme les termes provenant d'ontologies, de modèles cellulaires, de cartes statiques, de références scientifiques ou de maladies se retrouvent sous forme d'entités (Figure 10). Ces entités représentent des concepts que nous qualifions de haut niveaux au sein de Pegasus.

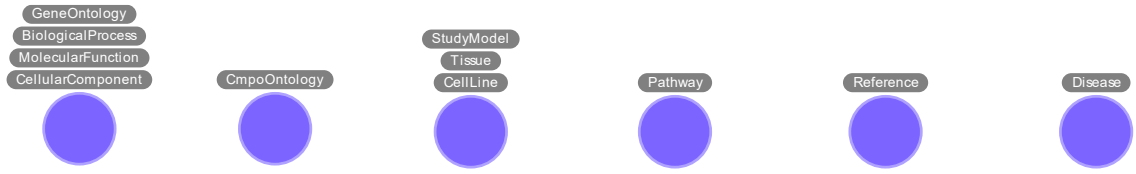

Figure 10. Modélisation d'ontologies, de modèles cellulaires, de cartes statiques, de références scientifiques et de maladies au sein de Pegasus. Les termes ontologiques provenant d'ontologies sont représentés par des entités possédant les étiquettes `GeneOntology`, `BiologicalProcess`, `MolecularFunction`, `CellularComponent` et `CmpoOntology`. Les modèles cellulaires d'études possèdent l'étiquette principale `StudyModel` et des étiquettes secondaires comme `Tissue` ou `CellLine`. Les termes issus des cartes statiques comme des voies de signalisation sont représentées par l'entité avec l'étiquette `Pathway`. Les publications scientifiques sont représentées par l'entité avec l'étiquette `Reference`. Les maladies sont représentées par l'entité avec l'étiquette `Disease`.

Nous présentons en (Figure 11) le modèle de données de Pegasus dans son ensemble rendant compte de tous les concepts introduits notamment du nœud intermédiaire qui permet de relier contextuellement des entités entre elles (Figure 11).

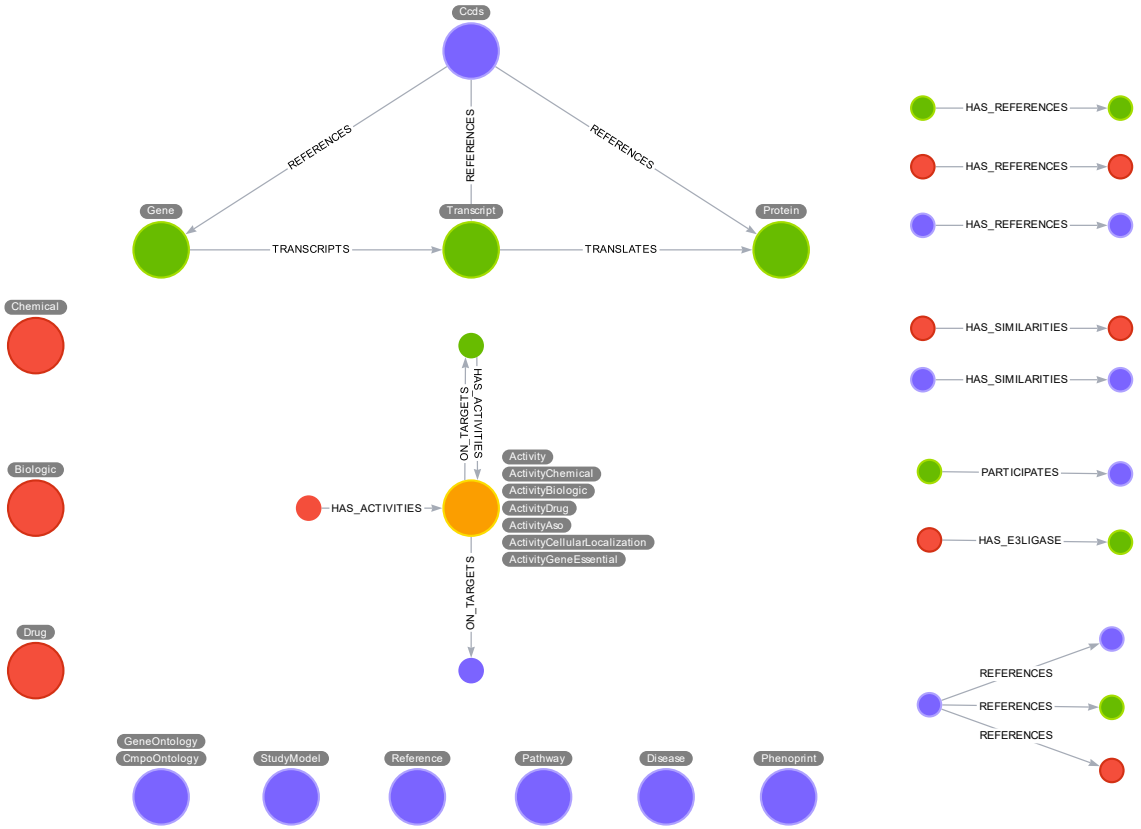

Figure 11. Modèle de données du graphe de connaissances Pegasus pour améliorer les phases primaires de recherche de nouveaux médicaments. Les concepts introduits dans Pegasus sont les entités représentant des objets de la biologie moléculaire (en vert), des objets de type perturbateur (en rouge), des objets de haut niveaux (en violet) et des objets contextuels (en orange). Les entités de la biologie moléculaire possèdent les étiquettes principales `Gene`, `Transcript`, `Protein`. Les perturbateurs possèdent les étiquettes principales `Chemical`, `Biologic` et `Drug`. Les entités de haut niveaux possèdent les étiquettes principales `GeneOntology`, `CmpoOntology`, `Reference`, `Pathway`, `Disease`, et `Phenoprint`. Des références croisées permettent de relier, par les relations `HAS_REFERENCE`, des concepts fonctionnellement identiques mais référencés par des identifiants différents. Les perturbateurs chimiques (`Chemical`) sont reliés entre eux par des relations `HAS_SIMILARITIES` qui contiennent, dans des propriétés, des mesures de similarités chimiques. Les signatures phénotypiques (`Phenoprint`) sont reliées entre elles par des relations `HAS_SIMILARITIES` qui contiennent, dans des propriétés, des mesures de similarités phénotypiques. Les entités contextuelles possèdent l'étiquette principale `Activity` afin de contextualiser des relations entre des entités au sein de Pegasus. Ces entités peuvent avoir des étiquettes secondaires afin de préciser le type d'activité. Par exemple, un `Chemical` peut avoir une activité de type `ActivityChemical`, un `Asos` peut avoir une activité de type `ActivityAso`, ou un `Gene` peut avoir une activité de type `ActivityGeneEssential`. Cette entité contextuelle permet, par exemple, de relier un perturbateur biologique (`miRNA`) qui a une activité (`ActivityBiologic`). Cette entité `ActivityBiologic` est reliée à un modèle cellulaire d'étude (`StudyModel`) ; à une cible thérapeutique qui peut être un (`Transcript`) ; et à une signature phénotypique (`Phenoprint`). Dans le modèle de données de Pegasus, certaines entités sont reliées directement entre elles. Par exemple, les gènes (`Gene`) sont reliés à des processus biologiques (`BiologicalProcess`), des voies de signalisation (`Pathway`) ou des maladies (`Disease`) par la relation `PARTICIPATES`. Les PROTACs (`Protac`) sont reliés à leur E3 ligase (`Gene`) par la relation `HAS_E3LIGASE`. Certaines références scientifiques (`Reference`) sont reliées à des entités biologiques (`Gene`), des perturbateurs (`Chemical`), des maladies (`Disease`) et des modèles cellulaires d'études (`StudyModel`) par la relation `REFERENCES`.

2.2.2 Implémentation et déploiement

La plateforme Pegasus permet de prétraiter les sources de données hétérogènes et générer le graphe de connaissances automatiquement. Cette plateforme, dont une vue d'ensemble est présentée en (Figure 12), est développée avec le framework Python Kedro 0.17.5 qui emprunte des pratiques de l'ingénierie logicielle [87]. Les principales caractéristiques du graphe de connaissances et de la plateforme Pegasus sont résumées en (Table 2).

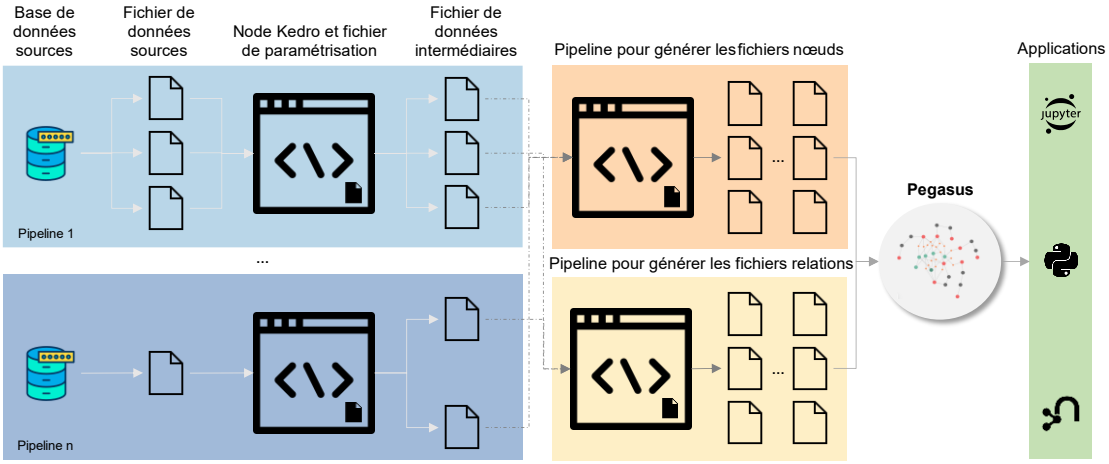

Figure 12. Plateforme Pegasus.

Une `Pipeline` Kedro est développée spécifiquement pour chaque base de données sources. Dans chaque `Pipeline`, nous développons un ou plusieurs `Node` Kedro pour prétraiter les données sources et générer des fichiers de données intermédiaires normalisés selon une charte de nommage. Cette charte de nommage permet, étant donné les noms des fichiers intermédiaires, d'extraire les données correspondant au modèle de donnée de Pegasus définit dans un fichier de paramétrisation. Nous développons une fonction Python spécifique pour chaque `Node` Kedro contenant le code expert pour les traitements. Ces fonctions renomment les champs selon une nomenclature et traitent les données selon des règles expertes. Deux `Pipeline` spécifiques génèrent de façon dynamique des `Node` Kedro pour traiter les données issues des fichiers intermédiaires et pour produire les fichiers nœuds et relations injectables dans un graphe Neo4j. L'exécution d'un script permet de prétraiter les données sources et de générer le graphe de connaissances Pegasus localement sur une machine ou dans un Docker. Nous développons des Jupiter Notebook et des scripts Python et nous utilisons le moteur d'exploration Bloom qui se connectent directement à Pegasus pour réaliser des analyses.

Nombre d'entités et de relations	Type d'entités et de relations	Nombre de Pipeline et de Node Kedro	Temps de génération des fichiers et de Pegasus[4]
46.371.784	66	35	4 heures[5]
331.570.883	14	155	6 minutes[6]

Table 2. Caractéristiques principales du graphe de connaissances et de la plateforme Pegasus.

[4] Machine MSI-GE76 Raider - Windows 10, i7, 2To SSD, 32 Go Ram

[5] `kedro run --pipeline=pipeline_complete_build --parallel`

[6] `neo4j-admin import --nodes "nodes_*.csv" --relationships "relations_*.csv"`

2.2.3 Exemples de requêtes illustrant les concepts introduits

Dans cette section, nous illustrons des concepts[7] introduits sur trois exemples en requêtant le graphe de connaissances Pegasus.

La requête CYPHER (Requête 3) identifie les transcrits et les protéines issus d'un gène formant des unités fonctionnelles avec leurs références croisées. Un exemple de résultat obtenu est présenté en (Table 3) qui illustre la diversité d'identifiants pouvant référencer des ressources biologiques.

```
MATCH p=(g1:Gene)-[:HAS_REFERENCES]-(g2:Gene)-[:TRANSCRIPTS]-(t1:Transcript)-
[:HAS_REFERENCES]-(t2:Transcript)-[:TRANSLATES]-(p1:Protein)-[:HAS_REFERENCES]-
(p2:Protein)-[:REFERENCES]-(c:Ccds)

WHERE g1.geneId = '4137' # L'identifiant Entrez 4137 de la base Ncbi correspond au gène MAPT

RETURN DISTINCT g1.geneId, g2.geneId, t1.transcriptId, t2.transcriptId, p1.proteinId,
p2.proteinId, c.ccdsId;
```

Requête 3. Identification des unités fonctionnelles (gène, transcrits, protéines) étant donné un gène. Cette requête CYPHER identifie les entités `Gene` reliées par la relation `HAS_REFERENCES` à l'entité `Gene` dont la valeur de propriété `geneId` est `4137`, les transcrits (`Transcript`) avec leurs références croisées (`HAS_REFERENCES`) reliés à ces `Gene` par la relation `TRANSCRITS`, les protéines (`Protein`) avec leurs références croisées (`HAS_REFERENCES`) reliées aux `Transcript` par la relation `TRANSLATES`, et les entités `Ccds` reliées à ces `Protein` par la relation `REFERENCES`. Le résultat de cette requête retourne vingt-neuf résultats que nous présentons en (Table 3).

g1.geneId	g2.geneId	t1.transcriptId	t2.transcriptId	p1.proteinId	p2.proteinId	c.ccdsId
4137	ENSG00000186868	ENST00000351559	NM_005910	NP_005901	ENSP00000303214	CCDS11499.1
4137	ENSG00000277956	ENST00000620070	NM_005910	NP_005901	ENSP00000303214	CCDS11499.1
4137	ENSG00000276155	ENST00000621329	NM_005910	NP_005901	ENSP00000303214	CCDS11499.1
4137	ENSG00000186868	ENST00000446361	NM_016834	NP_058518	ENSP00000408975	CCDS11500.1
4137	ENSG00000277956	ENST00000622106	NM_016834	NP_058518	ENSP00000408975	CCDS11500.1
4137	ENSG00000276155	ENST00000626571	NM_016834	NP_058518	ENSP00000408975	CCDS11500.1
4137	ENSG00000186868	ENST00000571987	NM_016835	NP_058519	ENSP00000458742	CCDS11501.1
4137	ENSG00000277956	ENST00000618825	NM_016835	NP_058519	ENSP00000458742	CCDS11501.1
4137	ENSG00000276155	ENST00000627711	NM_016835	NP_058519	ENSP00000458742	CCDS11501.1
4137	ENSG00000186868	ENST00000334239	NM_016841	NP_058525	ENSP00000334886	CCDS11502.1
4137	ENSG00000277956	ENST00000633047	NM_016841	NP_058525	ENSP00000334886	CCDS11502.1
4137	ENSG00000276155	ENST00000628393	NM_016841	NP_058525	ENSP00000334886	CCDS11502.1
4137	ENSG00000277956	ENST00000633047	NM_001377268	NP_001364197	ENSP00000334886	CCDS11502.1
4137	ENSG00000276155	ENST00000628393	NM_001377268	NP_001364197	ENSP00000334886	CCDS11502.1
4137	ENSG00000186868	ENST00000334239	NM_001377268	NP_001364197	ENSP00000334886	CCDS11502.1
4137	ENSG00000186868	ENST00000415613	NM_001123066	NP_001116538	ENSP00000410838	CCDS45715.1
4137	ENSG00000277956	ENST00000618029	NM_001123066	NP_001116538	ENSP00000410838	CCDS45715.1
4137	ENSG00000276155	ENST00000629368	NM_001123066	NP_001116538	ENSP00000410838	CCDS45715.1
4137	ENSG00000186868	ENST00000680542	NM_001123067	NP_001116539	ENSP00000413056	CCDS45716.1
4137	ENSG00000277956	ENST00000612872	NM_001123067	NP_001116539	ENSP00000413056	CCDS45716.1
4137	ENSG00000276155	ENST00000613360	NM_001123067	NP_001116539	ENSP00000413056	CCDS45716.1
4137	ENSG00000186868	ENST00000420682	NM_001123067	NP_001116539	ENSP00000413056	CCDS45716.1
4137	ENSG00000186868	ENST00000680542	NM_001123067	NP_001116539	ENSP00000505258	CCDS45716.1
4137	ENSG00000277956	ENST00000612872	NM_001123067	NP_001116539	ENSP00000505258	CCDS45716.1
4137	ENSG00000276155	ENST00000613360	NM_001123067	NP_001116539	ENSP00000505258	CCDS45716.1
4137	ENSG00000186868	ENST00000420682	NM_001123067	NP_001116539	ENSP00000505258	CCDS45716.1
4137	ENSG00000186868	ENST00000431008	NM_001203252	NP_001190181	ENSP00000389250	CCDS56033.1
4137	ENSG00000277956	ENST00000620818	NM_001203252	NP_001190181	ENSP00000389250	CCDS56033.1
4137	ENSG00000276155	ENST00000620981	NM_001203252	NP_001190181	ENSP00000389250	CCDS56033.1

Table 3. Identification des unité fonctionnelle (gène, transcrits, protéines) du gène MAPT. Dans cet exemple, le gène MAPT (identifié par `4137` depuis la base Ncbi et `ENSG00000276155, ENSG00000277956` et `ENSG00000276155` depuis le base Ensembl) possède vingt-deux transcrits curés issus de la base Ensembl

[7] D'autres concepts de Pegasus et requêtes CYPHER sont présentés dans la section suivante.

(ENST*), huit transcrits curés issus de la base Ncbi (NM_*), et huit protéines curées issues de la base Ensembl et Ncbi (ENSP*, NP_*). Nous observons qu'il n'existe que sept isoformes protéiques issues du gène MAPT (colonne c.ccdsId) malgré un nombre d'identifiant de protéines supérieur.

La requête CYPHER (Requête 4) identifie une classe de perturbateurs, les siRNAs, qui ont des activités dans un modèle cellulaire et décrites par des termes ontologiques. Un exemple de résultat obtenu est présenté en (Figure 13) qui illustre le nœud intermédiaire (Activity) afin d'annoter contextuellement des entités entre elles dans Pegasus. Dans cet exemple (Figure 13), nous présentons l'utilisation de termes ontologiques phénotypiques. En décrivant des résultats d'expériences obtenues dans un modèle cellulaire avec des termes issus d'un vocabulaire contrôlé [29], nous pouvons identifier des perturbateurs qui induisent, par exemple, une augmentation du nombre de cellules en apoptose (CMPO_0000221) ou encore des cellules dont la phase S du cycle cellulaire est arrêtée (CMPO_0000204), et ce, quel que soit le protocole expérimentale qui a été réalisé (modèle cellulaire, instruments de mesures, perturbateurs).

```
MATCH path1=(:Sirna)-[:HAS_ACTIVITIES]-(a:Activity)-[:ON_TARGETS]-(c:Cmpoontology)
WHERE c.cmpoontologyId = 'CMPO_0000274'
WITH path1, a

MATCH path2=(a)-[:ON_TARGETS]-(c:CellLine)
WHERE c.studymodelId = 'S2R+'

RETURN path1, path2;
```

Requête 4. Identification de perturbateurs biologiques dans un modèle cellulaire particulier dont les effets sont caractérisés par des termes ontologiques phénotypiques. Cette requête CYPHER identifie des siRNA (Sirna) qui ont une activité (Activity) sur un modèle cellulaire d'étude (CellLine) nommé S2R+ et qui induisent un phénotype cellulaire décrit par un terme ontologique (Cmpoontology) nommé CMPO_0000274. Le modèle cellulaire S2R+ est une drosophile melanogaster et le terme ontologique CMPO_0000274 correspond à un phénotype observé de cytosquelette d'actine cortical désorganisé. Un résultat de cette requête est présenté en (Figure 13).

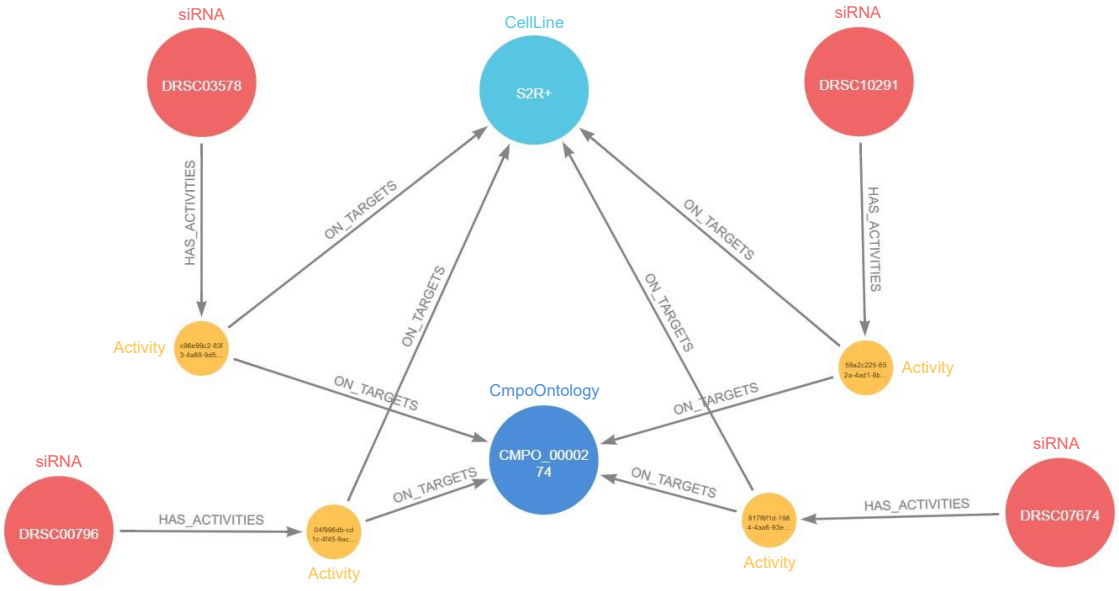

Figure 13. Identification de petits ARN interférents induisant un cytosquelette d'actine cortical désorganisé dans un modèle cellulaire de drosophile. Dans cet exemple, quatre siRNAS (Sirna), criblés dans une campagne de criblage à haut contenu, induisent (Activity) dans le modèle cellulaire S2R+ (CellLine) un phénotype cellulaire décrit par un terme ontologique (CMPO_0000274) correspondant à un

phénotype observé de type cytosquelette d'actine cortical désorganisé. L'entité contextuelle `Activity` est un nœud intermédiaire qui relie différentes entités du graphe. Les résultats présentés dans cette figure ne sont pas exhaustifs par souci de clarté.

La requête CYPHER (Requête 5) identifie, étant donné un processus biologique d'intérêt, les cibles thérapeutiques participant à ce processus biologique et des perturbateurs qui les modulent, des perturbateurs chimiquement similaires à ces derniers, et des perturbateurs modulant les transcrits de ces cibles. Un exemple de résultat obtenu est présenté en (Figure 14) qui illustre une application pour les phases primaires de recherche notamment pour l'identification de cibles et la conception de librairies de criblage focalisées avec différentes classes de perturbateurs.

```
MATCH path1=(g:Gene)-[:HAS_REFERENCES*0..2]-(:Gene)-[:PARTICIPATES]-
(bp:BiologicalProcess)
WHERE bp.geneontology_label = 'autophagosome assembly'
WITH g, path1

MATCH path2=(g)-[:ON_TARGETS]-(:Activity)-[:HAS_ACTIVITIES]-(:Chemicalprobe)
WITH g, path1, path2

MATCH path3=(g)-[:HAS_REFERENCES*0..2]-(:Gene)-[:ON_TARGETS]-(:Activity)-
[:HAS_ACTIVITIES]-(:Protac)
WITH g, path1, path2, path3

MATCH path4=(g)-[:HAS_REFERENCES*0..2]-(:Gene)-[:ON_TARGETS]-(:Activity)-
[:HAS_ACTIVITIES]-(:ChemicalChembl)-[:HAS_SIMILARITIES]-(:ChemicalServier)
WITH g, path1, path2, path3, path4

MATCH path5=(g)-[:HAS_REFERENCES*0..2]-(:Gene)-[:TRANSCRIPTS]-(:Transcript)-
[:HAS_REFERENCES]-(:Transcript)-[:ON_TARGETS]-(:Activity)-[:HAS_ACTIVITIES]-(:Asos)
WITH g, path1, path2, path3, path4, path5

MATCH path6=(g)-[:HAS_REFERENCES*0..2]-(:Gene)-[:TRANSCRIPTS]-(:Transcript)-
[:HAS_REFERENCES]-(:Transcript)-[:ON_TARGETS]-(:Activity)-[:HAS_ACTIVITIES]-(:Mirna)

RETURN path1, path2, path3, path4, path5, path6;
```

Requête 5. Identification de différentes classes de perturbateurs chimiques et biologiques modulant les cibles thérapeutiques impliquées dans un processus biologique. Cette requête CYPHER identifie des cibles thérapeutiques (`Gene`) qui participent (`PARTICIPATES`) à l'assemblage de l'autophagosome (`BiologicalProcess`) et pour lesquelles : des perturbateurs chimiques (`Protac`, `ChemicalProbe`) ont une activité (`Activity`) sur ces cibles (`Gene`) ; des perturbateurs propriétaires (`ChemicalServier`) sont chimiquement similaires (`HAS_SIMILARITIES`) à des perturbateurs chimiques (`ChemicalChembl`) qui ont une activité sur ces cibles (`Gene`) ; et les transcrits (`Transcript`) de ces cibles qui sont modulables par des miRNAs (`miRNA`) et des ASOs (`Asos`). Un résultat de cette requête est présenté en (Figure 14).

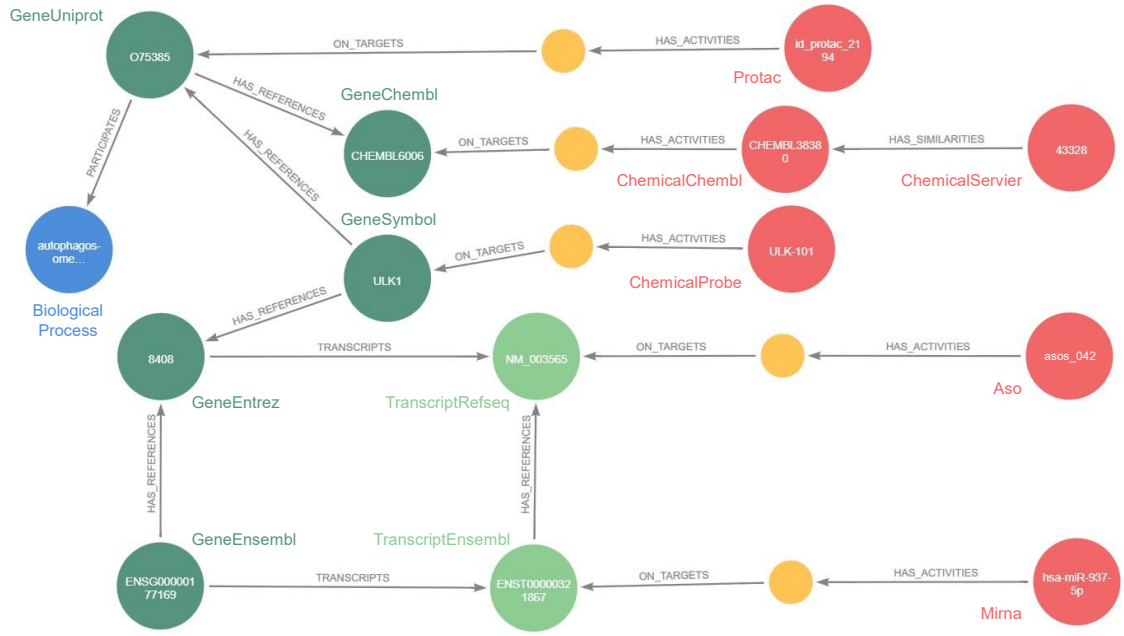

Figure 14. Identification de perturbateurs chimiques et biologiques modulant les cibles thérapeutiques impliquées dans l'assemblage de l'autophagosome. Dans cet exemple, le gène `ULK1` est représenté par plusieurs entités `Gene` (en vert) reliées entre elles par des références croisées (`HAS_REFERENCES`). Ce gène participe (`PARTICIPATES`) à l'assemblage de l'autophagosome (`BiologicalProcess`). Il existe un `Protac` et une `ChemicalProbe` qui ont une `Activity` (en jaune) sur ce gène et un `ChemicalChembl` qui a une activité sur `ULK1` et un `ChemicalServier` chimiquement similaire à ce dernier. Il existe un `Asos` et un `Mirna` qui ont une activité sur les transcrits (`Transcript`) du gène `ULK1`. Les transcrits sont reliés entre eux par des références croisées (`HAS_REFERENCES`). Au sein des différentes entités `Activity` des propriétés sont présentes. Par exemple, des propriétés de l'entité `Activity` entre un `Chemical` et un `Gene` sont : `experimental_parameter`, `experimental_value` et correspondent respectivement au type de paramètre expérimental calculé (p. ex. une concentration inhibitrice médiane) et la valeur associée à ce paramètre. Les résultats présentés dans cette figure ne sont pas exhaustifs par souci de clarté.

2.3 Applications industrielles de Pegasus pour des projets thérapeutiques

Dans les sections précédentes, nous avons introduit et illustré le graphe de connaissances Pegasus. Nous présentons, dans cette section, trois applications industrielles qui répondent à des problématiques de projets thérapeutiques internes et montrent notre utilisation de Pegasus dans le processus de recherche de nouveaux médicaments. Ces applications s'inscrivent dans un cadre scientifique complexe mais nous les abordons simplement. Pour chaque application, nous présentons brièvement la question scientifique et industrielle. Puis, nous présentons, en plus du développement d'algorithmes et d'analyses spécifiques, les requêtes CYPHER permettant de répondre aux problématiques des différents projets.

2.3.1 Caractérisation d'effets hors cibles de perturbateurs

Un projet thérapeutique interne a pour objectif de développer un oligonucléotide antisens (ASO) unique qui sera administré chez un bébé en 2023 pour moduler une cible thérapeutique que nous appellerons cible X. Un ASO est un fragment d'ADN qui peut se

lier à un ARNm [39]. L'objectif de ce projet est d'inactiver la cible X du bébé atteint d'encéphalopathie épileptique, une maladie rare et enfantine [88].

Les ASOs que nous développons ont une taille de 20 paires de bases et sont conçus par des algorithmes bio-informatiques, d'apprentissage profond, d'analyses de conservations entre espèces et d'analyse de brevets. Les ASOs sont une modalité stratégique interne et nous ne pouvons pas présenter nos méthodes de conception. Cependant, nous présentons en (Table 4), à titre indicatif, certaines des caractéristiques calculées pour chaque ASO.

ASOs	Séquences	Homologie	Transcrits hors cibles
Identifiant Séquence Positions	Palindrome, DG37, Tm, Entropie, Nombre de PSN	Souris, Rat, Singe, Brevet	Liens avec gènes essentiels et du développement, Niveau d'expression

Table 4. Caractéristiques calculées pour prédire l'activité et la toxicité d'oligonucléotides antisens pour moduler des cibles thérapeutiques. Chaque ASO possède un identifiant, une séquence d'acides nucléiques et des positions par rapport au transcriptome de référence. Des caractéristiques pour chaque séquence d'ASO sont calculées comme le degré palindromique, l'énergie de liaison (DG37), la température de fusion (Tm), l'entropie de la séquence, ou encore le nombre de polymorphisme nucléotidique. Certaines de ces caractéristiques sont détaillées dans [89]. L'homologie de la séquence d'un ASO est calculée au regard d'une cible dans différentes espèces (souris, rat, singe) et un contrôle est réalisé pour déterminer si la séquence n'est pas brevetée. Avec Pegasus, le nombre de caractéristiques calculées pour chaque ASO est augmenté en déterminant si les transcrits hors cibles d'ASOs sont transcrits à partir de gènes essentiels ou liés au développement. Le niveau d'expression de chaque transcrits hors cible est également quantifié pour les différents tissus humains.

Étant donné leur taille de 20 paires de bases, les ASOs peuvent se fixer et moduler des cibles en dehors de leur cible principale : ce sont des transcrits hors cibles. L'objectif de l'utilisation de Pegasus est de fournir des caractéristiques supplémentaires liées à ces transcrits hors cibles pour supporter la sélection d'ASOs qui seront criblés. En effet, nous avons développé 6.143 ASOs contre la cible X qui peuvent se fixer et donc potentiellement moduler 15.134 transcrits distincts (Figure 15).

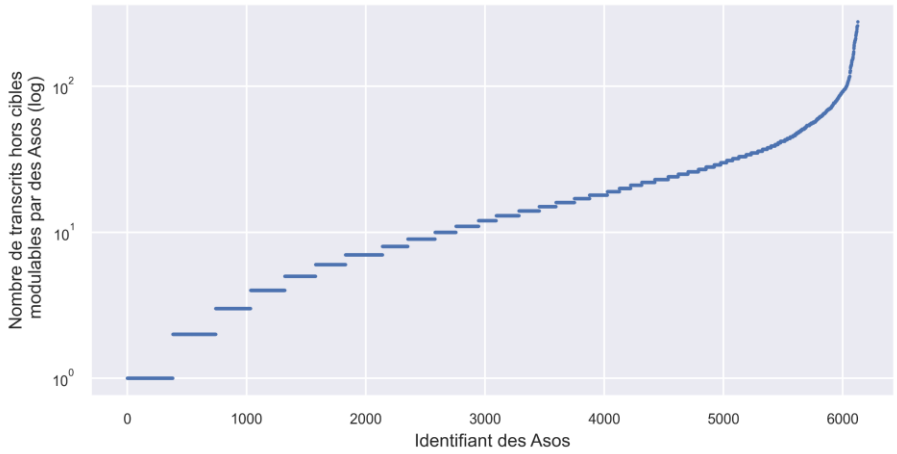

Figure 15. Nombre de transcrits hors cibles modulables par des ASOs développés en interne pour inhiber une cible dérégulée chez un bébé atteint d'encéphalopathie épileptique. En abscisse, les identifiants des 6.143 ASOs développés en interne pour moduler la cible X. En ordonné, le nombre de transcrits hors cibles pour chaque ASO. Les ASOs peuvent se lier à 15.134 transcrits distincts transcrits à partir de 6.186 gènes distincts.

Nous requêtons Pegasus afin de répondre à deux questions : quels sont les transcrits hors cibles qui sont transcrits à partir de gènes essentiels ? Quels sont les transcrits hors cibles qui sont transcrits à partir de gènes participant au développement ? Les concepts de gènes essentiels et liés aux développement, issues de [30,40], sont importants dans le cadre de ce projet puisque nous développons un ASO unique pour le traitement d'un bébé et nous souhaitons minimiser de potentiels effets toxiques. Les deux requêtes CYPHER permettant de répondre à ces deux questions sont présentées en (Requête 6).

```
### Requête 1

MATCH path=(a:Asos)-[:HAS_ACTIVITIES]-(:ActivityAsos)-[:ON_TARGETS]-(t1:Transcript)-
[:HAS_REFERENCES]-(t2:Transcript)-[:TRANSCRIPTS]-(g2:Gene)-[:HAS_REFERENCES]-
(g1:Gene)-[:HAS_ACTIVITY]-(act:ActivityGeneEssential)-[:ON_TARGETS]-(d:Disease)
return path;

### Requête 2

MATCH path=(a:Asos)-[:HAS_ACTIVITIES]-(:ActivityAsos)-[:ON_TARGETS]- (t1:Transcript)
-[:HAS_REFERENCES*0..2]-(t2:Transcript)-[:TRANSCRIPTS]-(g2:Gene)- [:HAS_REFERENCES]-
(g1:Gene)-[:PARTICIPATES]- (go:BiologicalProcess)
WHERE go.geneontology_label contains 'development' AND go.geneontology_evidence IN
['EXP', 'IDA', 'IPI', 'IMP', 'IGI', 'IEP', 'HTP', 'HDA', 'HMP', 'HGI', 'HEP']
RETURN path;
```

Requête 6. Identification des transcrits hors cibles transcrits à partir de gènes essentiels ou liés au développement pour caractériser les effets d'oligonucléotides antisens. La première requête identifie les `Asos` qui ont une activité de type `ActivityAsos` sur des `Transcript` qui sont transcrits à partir de `Gene` qui ont une activité de gènes essentiels (`ActivityGeneEssential`). L'entité `ActivityGeneEssential` est également reliée à une entité avec l'étiquette `Disease`. Un gène en plus d'être essentiel selon la définition de [40] peut être dérégulé dans une maladie. Les propriétés de l'entité `ActivityGeneEssential` sont : `invitro_essential`, `invivo_essential`, `mice_essential`, `number_of_pathogeneic_variants` indiquant respectivement si le gène est essentiel *in vitro*, *in vivo*, dans la souris ainsi que le nombre de variants pathogéniques connus. Par soucis de clarté, nous n'avons pas écrit dans la requête l'ensemble des entités et des relations ainsi que les propriétés retournées qui sont ancrées par les variables (`a`, `g1`, `g2`, `t1`, `t2`, `act`, `d`) et contenues dans la variable `path`. Les propriétés retournées sont : les identifiants des `Asos`, `Gene`, `Transcript`, `Disease` ainsi que toutes les propriétés contenues dans l'entité `ActivityGeneEssential` que nous venons de décrire. La deuxième requête identifie les `Asos` qui ont une activité de type `ActivityAsos` sur des transcrits (`Transcript`) qui sont transcrits à partir de gènes (`Gene`) qui participent (`PARTICIPATES`) dans des processus biologiques (`BiologicalProcess`) dont les noms de ces processus biologiques (`geneontology_label`) contiennent le terme `development`. Le motif de recherche permet de restreindre les résultats aux annotations entre les gènes (`Gene`) et les processus biologiques (`BiologicalProcess`) qui possèdent des preuves expérimentales. Par exemple, le score `HTP` correspond à une preuve expérimentale qui indique que l'annotation entre un gène et un processus biologique est soutenue par des méthodologies à haut débit. Les définitions des scores entre un `Gene` et un `BiologicalProcess` sont présentés en [30]. Par souci de clarté, nous n'avons pas écrit dans la requête l'ensemble des entités et des relations ainsi que les propriétés retournées par la requête qui sont ancrées par les variables (`a`, `g1`, `g2`, `t1`, `t2`, `go`) et contenues dans la variable `path`. Les propriétés retournées sont : les identifiants des `Asos`, `Gene`, `Transcript`, `BiologicalProcess` ainsi que le nom (`geneontology_label`) des processus biologiques.

La première requête (Requête 6) identifie les transcrits hors cibles qui sont transcrits à partir de gènes essentiels et nous illustrons en (Figure 16) un exemple de résultat obtenu.

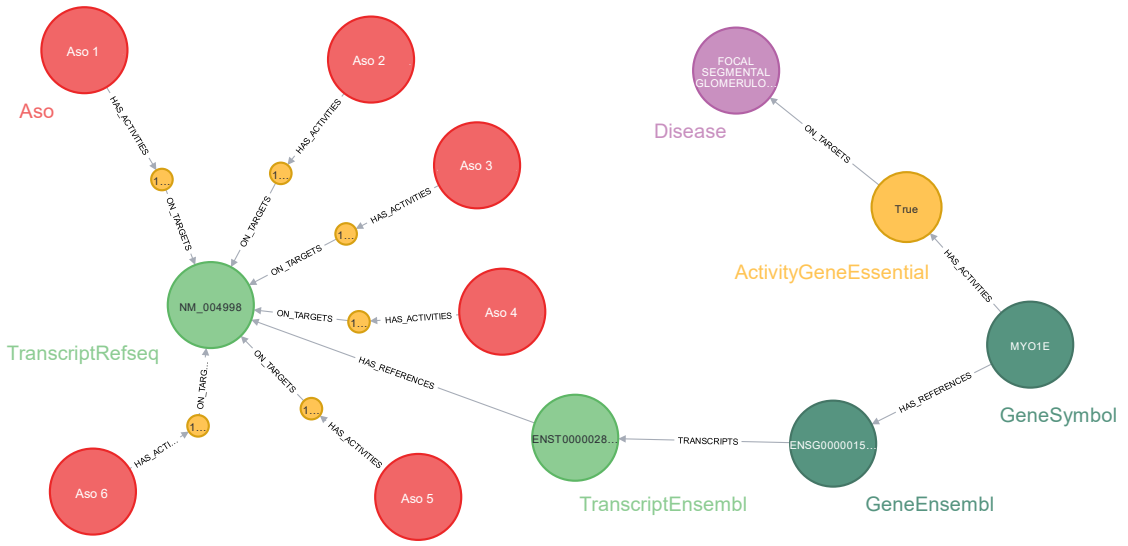

Figure 16. Identification de transcrits hors cibles transcrits à partir de gènes essentiels pour caractériser les effets d'oligonucléotides antisens. Dans cet exemple, six ASOs (`Asos`) sont reliés à un transcrit (`TranscriptRefseq`) par une entité contextuelle (`ActivityAsos`). Le transcrit (`TranscriptRefseq`) est relié par une référence croisée (`HAS_REFERENCE`) à un transcrit (`TranscriptEnsembl`). Ces transcrits sont identiques mais possèdent des identifiants provenant de deux bases de données de référence (Ensembl et Ncbi). Le transcrit est transcrit (`TRANSCRIPTS`) à partir d'un gène (`GeneEnsembl`) qui est relié par une référence croisée à un de ses symboles (`GeneSymbol`). Ce gène possède une activité de type gène essentiel (`ActivityGeneEssential`) et qui est dérégulé, en plus, dans une maladie (`Disease`). Les résultats présentés dans cette figure ne sont pas exhaustifs par souci de clarté.

Parmi les 6.143 ASOs, 2.073 ASOs sont reliés à 1.080 transcrits qui sont transcrits à partir de 372 gènes essentiels. Ce résultat indique qu'approximativement 1/3 des ASOs développés ne sont pas spécifiques de la cible X.

La deuxième requête (Requête 6) identifie les transcrits hors cibles qui sont transcrits à partir de gènes liés au développement. Parmi les 6.143 ASOs, 1.344 ASOs sont reliés à 625 transcrits qui sont transcrits à partir de 400 gènes participants au développement. Ce résultat indique qu'approximativement 1/6 des ASOs ne sont pas spécifiques de la cible X. En comparant les deux listes de gènes obtenues après le requêtage de Pegasus, nous observons qu'il existe 518 ASOs qui ont des activités prédites sur des transcrits hors cibles issus de gènes essentiels et participant au développement.

À partir des caractéristiques calculées par nos méthodes internes (Table 4), nous sélectionnons 784 ASOs parmi les 6.143 ASOs développés. La sélection des 784 ASOs est basée sur les caractéristiques issues des analyses d'homologies et des brevets, ainsi que sur une métrique de score qui avantage les ASOs ayant de bonnes activités prédites et pénalise les ASOs ayant des activités sur des transcrits hors cible. Parmi les 784 ASOs retenus, vingt pour cent des ASOs ont des activités prédites sur des transcrits issus de gènes essentiels ou liés au développement.

Les 784 ASOs sélectionnés sont ensuite criblés dans un modèle cellulaire neuronal et dans les souris. Cette campagne de criblage a permis d'identifier expérimentalement 13 ASOs actifs contre la cible X. L'identification de 13 ASOs actifs contre la cible X à ce stade est de valeur. En effet, nos méthodes couplées à une validation expérimentale ont permis d'identifier 13 candidats précliniques en six mois. Le développement de candidats précliniques prend en moyenne trois à six années dans le processus standard de recherche pharmaceutique. Parmi les 13 ASOs actifs, l'analyse issue de Pegasus montre que 8 ASOs n'ont aucun effet hors cible et parmi ces 8 ASOs, 3 ASOs ont les meilleures activités contre la cible X. Ces 3 ASOs seront criblés en dose-réponse en juin 2022 sur

des cellules neuronales issues de cellules souches pluripotentes induites de la patiente afin de les valider. À l'issue de cette campagne de criblage et d'expériences complémentaires pour optimiser la perméabilité cellulaire, l'accessibilité et la stabilité des 3 ASOs, nous sélectionnerons un ASO unique qui sera administré chez la patiente en mars 2023.

Nous travaillons sur de nouveaux concepts à introduire dans Pegasus pour augmenter la caractérisation des effets hors cibles de perturbateurs. Par exemple, nous intégrons des gènes orthologues de souris transgéniques pour lesquelles les modifications d'allèles entraînent des phénotypes pathologiques [36]. Ce type d'ajout illustre la flexibilité d'évolution du modèle de données de Pegasus, présenté en (Figure 17), et nous permet d'identifier, les gènes orthologues de souris transgéniques dont les modifications induisent des phénotypes pathologiques, aux gènes humains dont les transcrits sont modulés par les ASOs que nous développons. Ceci nous permettrait d'identifier des ASOs avec des effets potentiellement toxiques notamment pour les exclure des expériences réalisées sur des modèles animaux dans les phases précliniques puisque les effets des ASOs sont similaires à l'inhibition des gènes.

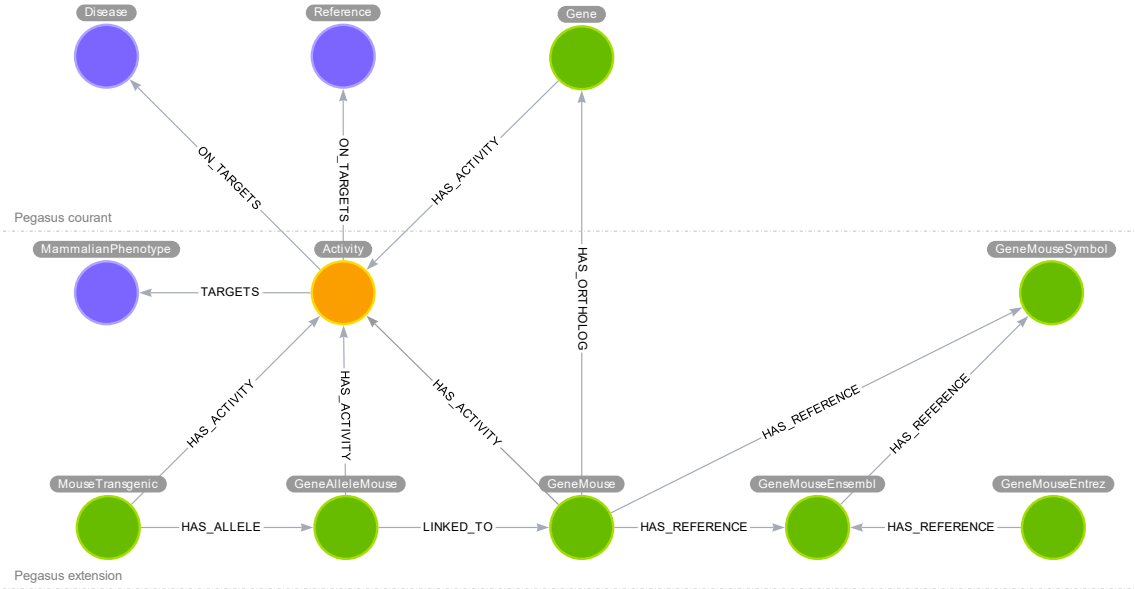

Figure 17. Extension du modèle de données de Pegasus pour introduire des concepts de souris transgéniques et de phénotypes pathologiques. Nous intégrons des concepts de gènes orthologues de souris (`GeneMouse`) qui sont liés à des allèles de ces gènes (`GeneAlleleMouse`), eux-mêmes reliés à des souris transgéniques (`MouseTransgenic`). Une souris transgénique, un de ses gènes ou allèles mutés peuvent avoir des effets (`Activity`) caractérisés par des termes ontologiques (`MammalianPhenotype`) et caractéristiques de certaines maladies (`Disease`). Ces termes ontologiques décrivent des phénotypes pathologiques observés. Les gènes de souris (`GeneMouse`) sont reliés à leurs orthologues humains (`Gene`) présents dans Pegasus par la relation `HAS_ORTHOLOG`.

2.3.2 Conception d'une nouvelle expérience

Nous présentons dans cette section la conception d'une nouvelle expérience couplant les oligonucléotides antisens (ASOs) aux cadres de lecture en amont des transcrits (uORFs). La rationnel thérapeutique est présenté en (Figure 18). Ce travail est réalisé dans le cadre d'un projet thérapeutique interne dont l'objectif est d'augmenter la concentration d'une cible thérapeutique, que nous appellerons cible X, qui est sous-exprimée dans le syndrome amyotrophique latéral [90].

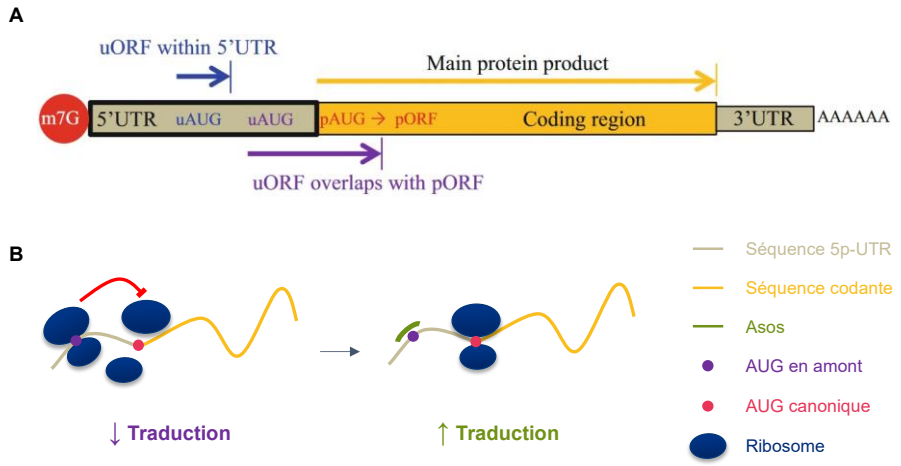

Figure 18. Rationnel thérapeutique des oligonucléotides antisens couplés aux cadres de lecture en amont des transcrits pour augmenter la concentration de protéines sous exprimées dans les maladies. (A) Un cadre de lecture en amont des transcrits (uORF) est un cadre de lecture (ORF) situé dans la partie 5'-UTR d'un ARNm [91]. **(B)** Les uORFs peuvent réguler l'expression des gènes eucaryotes. La traduction de l'uORF inhibe généralement l'expression en aval de l'ORF canonique de l'ARNm [91]. Ceci a pour conséquence une diminution du mécanisme de traduction de la protéine. Ainsi, masquer un uORF par un ASO permettrait d'augmenter la machinerie de traduction et donc d'augmenter la concentration d'une protéine. Ce type de modalité est prometteuse pour traiter des maladies dans lesquelles les protéines sont en plus faible concentration. Notons que soixante-dix pour cent des uORFs sont conservés entre les souris, les rats et les humains ce qui indiquerait un rôle physiologique notable et que deux tiers des oncogènes, qui doivent être étroitement régulés, possèdent des uORFs [92]. La modulation des transcrits offre des perspectives thérapeutiques prometteuses [93–95]. L'image présentée en **(A)** provient de [96].

Nous avons identifié dans la littérature deux sources de données d'uORFs présents dans les transcrits humains : uORFDb et uORFTool [42,43]. La base de données uORFdb a été établie manuellement à partir de la littérature relative aux uORFs répertoriés dans la base de données PubMed [42]. La procédure d'identification des uORFs présentée dans uORFTool se base sur des données expérimentales de profilage du ribosome [43]. Cependant, ces sources de données indiquent seulement si un gène ou un transcrit possède un uORF mais n'indiquent pas sa position spécifique au niveau du transcrit.

Nous avons développé un algorithme trivial avec un double objectif : identifier les positions des uORFs des transcrits humains tout en identifiant les plages de séquences des transcrits pour lesquels des ASOs complémentaires peuvent être développés. Nous présentons en (Algorithme 1) le pseudo-code associé à l'algorithme qui permet d'identifier les uORFs de transcrits et les plages de séquences possibles pour le développement d'ASOs de taille 20 paires de bases.

Entrée : Ensemble S des triplets (s_{5p}, s_{cds}, s). s_{5p} correspond à la séquence 5'-UTR, s_{cds} correspond à la séquence codante du transcrit et s correspond à la concaténation de s_{5p} et s_{cds}.

Sortie : Ensemble $\{(s_{uorf}, s_{asos})\}$ pour chaque transcrit humain où (s_{uorf}, s_{asos}) correspondent respectivement à la séquence d'un uORF et à la plage de séquence possible pour les ASOs.

Algorithme :

Pour (s_{5p}, s_{cds}, s) dans S **faire**
  Si 'ATG' dans s_{5p} **alors**
    Pour idx_{start_uorf} dans idx_{start_codon} **faire**
      Pour idx_{end_uorf} dans idx_{end_codon} **faire**
        $s_{uorf} := s[idx_{start\_uorf} : idx_{end\_uorf}]$
        Si TAILLE(s_{uorf}) modulo 3 égale 0 **alors**
          $s_{asos} := s[idx_{start\_uorf} - 20 : idx_{start\_uorf} + 20]$

          Retourner (s_{uorf}, s_{asos})

Algorithme 1. Algorithme d'identification des cadres de lecture en amont des transcrits humains et d'identification des plages de séquences possibles pour la conception d'oligonucléotides antisens. Le fonctionnement de l'algorithme est le suivant : pour chaque transcrit du transcriptome humain (S) contenant une séquence 5'-UTR (s_{5p}) et codante (s_{cds}), nous vérifions si des codons start (ATG) sont présents dans la séquence (s_{5p}). Si oui, alors nous récupérerons l'ensemble des positions possibles des codons start (idx_{start_codon}) de cette séquence dans laquelle apparait le codon start (idx_{start_uorf}). Puis, nous recherchons la position correspond aux codons stop (idx_{end_uorf}), qui peuvent être TAA, TGA, TAG pour l'ensemble des positions de codons de fin de séquence possibles (idx_{end_codon}). La séquence de l'uORF est comprise entre ces deux positions (idx_{start_uorf}, idx_{end_uorf}). Puis, nous récupérons la plage de séquence possible des ASOs qui doivent être centrés et contenir le codon start ATG de l'uORF ($idx_{start\_uorf} \pm 20$).

L'algorithme identifie 169.312 uORFs distincts pour 46.483 transcrits distincts. Nous retrouvons que quarante-huit pour cent des transcrits humains possèdent au moins un uORF ce qui est cohérent avec ce qui est connu dans la littérature [91]. De plus, l'algorithme retrouve les 1.933 transcrits possédant des uORFs identifiés expérimentalement comme actifs dans uORFTool [43]. Nos hypothèses sur la différence entre le nombre d'uORFs potentiels (169.312) identifiés par notre algorithme et les uORFs actifs (1.933) identifiés dans [43] sont : un certain nombre d'uORFs ne sont pas transcrits du fait de la compaction de l'ADN sous forme d'hétérochromatine, les banques de séquences et le protocole expérimental utilisés dans [43] ne couvrent pas tous les états biologiques cellulaires, et que les uORFs sont régulés par des mécanismes encore inconnus [91].

Nous présentons en (Figure 19) le résultat de l'exécution d'une requête CYPHER qui illustre l'intégration et l'identification dans Pegasus des uORFs obtenus de uORFDb, uORFTool et par notre algorithme étant donné un gène.

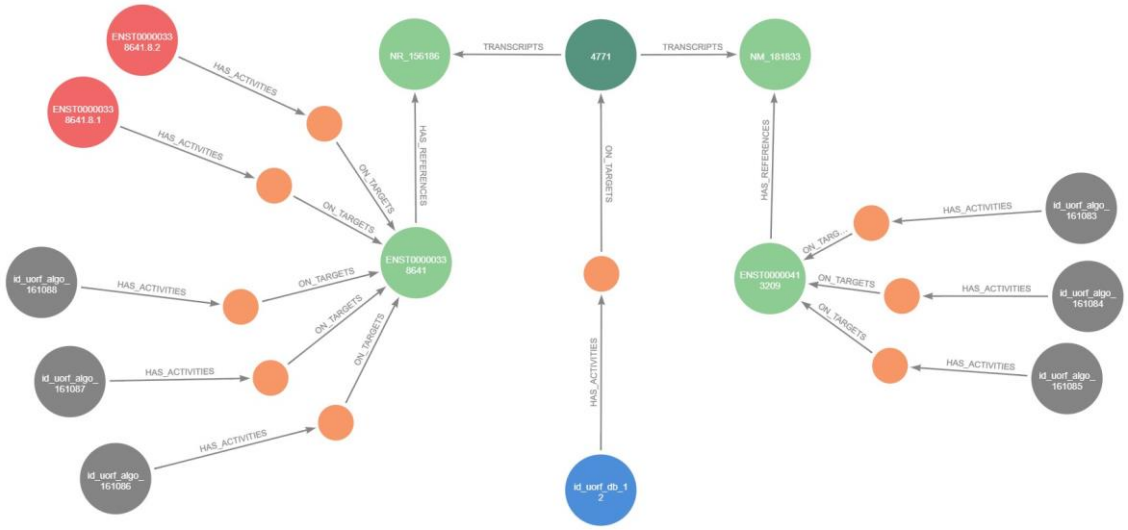

Figure 19. Identification des cadres de lecture en amont des transcrits humains modulables par des oligonucléotides antisens pour augmenter la concentration de protéines sous exprimées dans des maladies. Les uORFs de uORFDb (en bleu) sont reliés aux gènes (vert foncé) par une entité contextuelle (en orange). Les uORFs de uORFTool (en rouge) et de notre algorithme (en gris) sont reliés aux transcrits (vert clair). Dans cet exemple, pour le gène NF2, référencé ici par l'identifiant 4771, il existe un uORF de uORFDb, deux uORFs provenant de uORFTool reliés à un des transcrits du gène NF2 et six uORFs provenant de notre algorithme reliés à deux des transcrits du gène NF2. La requête CYPHER associée à cet exemple est : `MATCH path=(:Uorf)-[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-(:Transcript)-[:HAS_REFERENCES*0..2]-(:Transcript)-[:TRANSCRIPTS]-(g:Gene)-[:HAS_REFERENCES*0..2]-(:Gene)-[:ON_TARGETS]-(:Activity)-[:HAS_ACTIVITIES]-(:Uorf) WHERE g.geneId = '4771' RETURN path;`

Afin d'identifier les uORFs associés aux transcrits de la cible X, nous requêtons Pegasus avec la requête CYPHER présentée en (Requête 7). La cible X ne possède pas d'uORFs provenant d'uORFDb et nous restreignons l'identification des uORFs provenant de uORFTool et de notre algorithme.

```
MATCH path1=(u1:UorfTool)-[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-
(t11:Transcript)-[:HAS_REFERENCES*0..2]-(t12:Transcript)-[:TRANSCRIPTS]-(:Gene)-
[:HAS_REFERENCES*0..2]-(g:Gene)
WHERE g.geneId = 'cible X'

WITH path1, g

MATCH path2=(u2:UorfAlgo)-[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-
(t21:Transcript)-[:HAS_REFERENCES*0..2]-(t22:Transcript)-[:TRANSCRIPTS]-(:Gene)-
[:HAS_REFERENCES*0..2]-(g)

RETURN path1, path2;
```

Requête 7. Identification des cadres de lecture en amont des transcrits d'une cible thérapeutique sous exprimée dans le syndrome amyotrophique latéral. Cette requête CYPHER identifie les uORFs provenant de uORFTool et de notre algorithme (`uORFTool, uORFAlgo`) qui ont une activité (`Activity`) sur des transcrits (`Transcript`) de la cible X. Par soucis de clarté, nous n'avons pas écrit dans la requête l'ensemble des entités et des relations ainsi que les propriétés retournées par la requête qui sont ancrés par les variables (`u1`, `u2`, `t11`, `t12`, `t21`, `t22`) et contenus dans les variables `path1` et `path2`. Les propriétés retournées sont : les identifiants des `uORFTool`, des `uORFAlgo`, et des `Transcript`.

L'exécution de la requête (Requête 7) identifie 11 séquences d'uORFs distinctes pour 7 transcrits codants de la cible X (Figure 20).

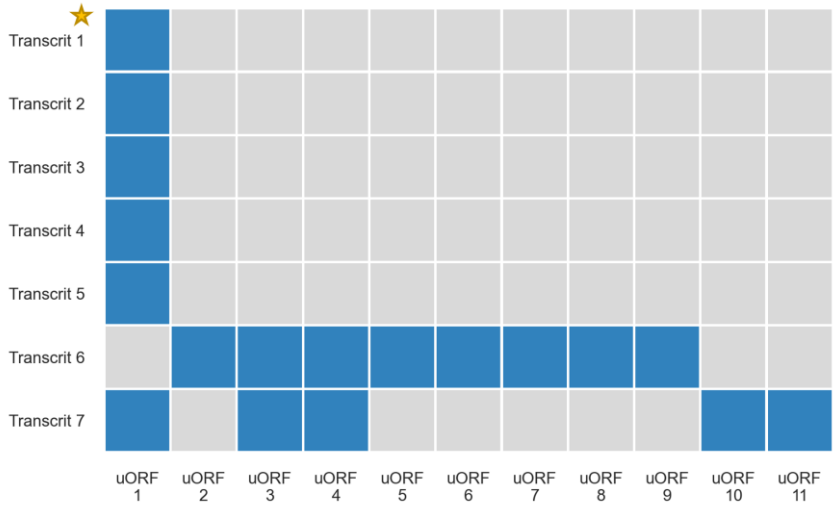

Figure 20. Séquences d'uORFs identifiées pour les transcrits codants d'une cible thérapeutique sous exprimée dans le syndrome amyotrophique latéral. En ordonnée, l'identifiant des transcrits codants de la cible X. En abscisse, l'identifiant des séquences des uORFs identifiés par notre algorithme pour les transcrits de la cible X. Un carré bleu, à l'intersection d'une ligne i et d'une colonne j, indique la présence d'une séquence spécifique d'un uORF j dans un transcrit i. Le transcrit 1 de la cible X correspond au transcrit identifié par uORFTool et notre algorithme. L'uORF 1 est présent dans six transcrits codants de la cible X. Les transcrits 6 et 7 sont traduits dans des formes protéiques tronquées de la cible X ce qui indiquerait pourquoi plusieurs séquences d'uORFs sont présents dans ces transcrits. Un transcrit peut avoir plusieurs cadres de lecture en amont.

Certaines séquences d'uORFs sont spécifiques à cent pour cent des transcrits de la cible X, notamment l'uORF majoritaire. Notre hypothèse sur cette spécificité est que certaines protéines possèdent des systèmes de régulation propres.

Les plages de séquences des ASOs identifiées par l'algorithme (Algorithme 1) sont utilisées ensuite pour extraire les séquences d'ASOs de taille de 20 paires de bases. Nous identifions 151 séquences d'ASOs distinctes complémentaires et contenant le codon ATG des 11 séquences d'uORFs.

Une expérience Enzyme-Linked Immunosorbent Assay (ELISA) est en train d'être mise en place afin de valider cette nouvelle modalité expérimentale qui nous permettra d'identifier les oligonucléotides antisens, parmi les 151 développés, qui augmentent la concentration de la cible X, et ce, en masquant les cadres de lectures en amont des transcrits. Cette application de Pegasus pourra s'étendre à tous les projets thérapeutiques travaillant sur des cibles thérapeutiques sous exprimées dans des maladies.

2.3.3 Identification de librairies focalisées pour le criblage de perturbateurs

Un projet thérapeutique interne a pour objectif de développer un perturbateur chimique pour inhiber l'agrégation anormale d'une cible thérapeutique dans la maladie de Parkinson que nous appellerons cible X. Nous utilisons Pegasus pour identifier des librairies chimiques focalisées pour le repositionnement de perturbateurs propriétaires. La probabilité d'identifier des perturbateurs actifs contre une cible thérapeutique est augmentée en criblant des perturbateurs chimiquement similaires, par exemple, au sens de la mesure de Tanimoto [97].

Nous avons présenté en (Requête 5), une requête permettant d'identifier différentes classes de perturbateurs pour moduler des cibles thérapeutiques participant à un processus biologique d'intérêt. Dans le cadre de ce projet, nous nous restreignons au repositionnement de perturbateurs propriétaires. La requête CYPHER (Requête 8) identifie des perturbateurs propriétaires chimiquement similaires aux perturbateurs de la littérature qui ont une activité sur la cible X et nous présentons la distribution des mesures de similarités chimiques entre les perturbateurs propriétaires et les perturbateurs de la littérature en (Figure 21).

```
MATCH path=(g:Gene)-[:HAS_REFERENCES]-(:Gene)-[:ON_TARGETS]-(act:ActivityChemical)-
[:HAS_ACTIVITIES]-(c2:ChemicalChembl)-[r:HAS_SIMILARITIES]-(c1:ChemicalServier)

WHERE g.geneId = 'cible X' AND r.tanimoto_similarity > 0.6

RETURN c1.chemicalId, r.tanimoto_similarity, c2.chemicalId,
act.experimental_parameter, act.experimental_value;
```

Requête 8. Identification de perturbateurs propriétaires pour réaliser des expériences de criblage focalisées. Cette requête CYPHER identifie les perturbateurs chimiques (`ChemicalChembl`) de la littérature qui ont une activité (`ActivityChemical`) sur la cible X, et les perturbateurs propriétaires (`ChemicalServier`) qui sont chimiquement similaires (`HAS_SIMILARITIES`) et dont l'indice de Tanimoto (`tanimoto_similarity`) est supérieur à `0.6`. La requête retourne les identifiants des perturbateurs chimiques (`ChemicalServier`, `Chemical`), la valeur de la mesure de Tanimoto (`tanimoto_similarity`), ainsi que différentes propriétés de l'entité (`ActivityChemical`) comme le paramètre expérimental qui a été calculé (`experimental_parameter`) ainsi que sa valeur (`experimental_value`).

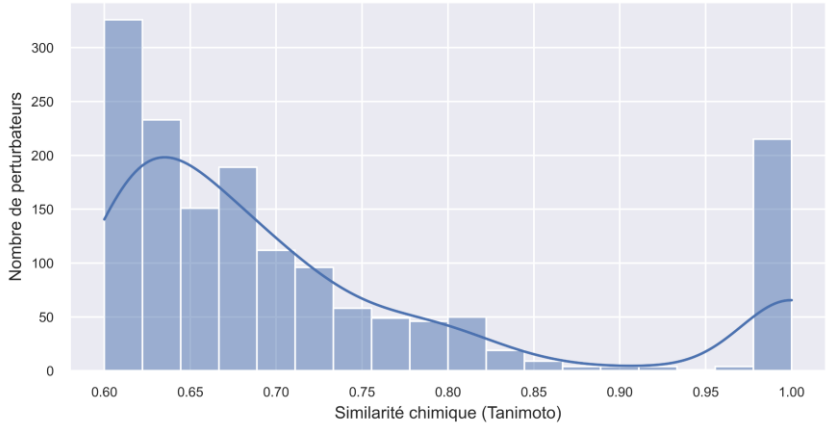

Figure 21. Mesures de similarités chimiques entre les perturbateurs propriétaires et les perturbateurs de la littérature qui ont une activité sur une cible thérapeutique. En ordonnée, le nombre de perturbateurs propriétaires qui ont une mesure de similarité chimique de type Tanimoto supérieur à 0.6 (en abscisses) avec des perturbateurs de la littérature qui ont une activité sur la cible X. Il existe des perturbateurs propriétaires qui ont une valeur pour l'indice de Tanimoto de 1. Notre librairie chimique est composée de perturbateurs qui ont été achetés ou synthétisés à partir de molécules du domaine public. Nous pourrions facilement annoter les perturbateurs de la libraire chimique Servier.

L'exécution de la requête (Requête 8) identifie 1.118 perturbateurs propriétaires. Nous avons également identifié une vingtaine de références scientifiques correspondant aux publications à partir desquelles sont issus les perturbateurs de la littérature qui ont une activité sur la cible X. De façon surprenante, dans ces publications, les auteurs n'ont pas tous identifiés des molécules ayant une activité sur la forme monomérique de la cible X alors que l'identifiant utilisé pour référencer la cible X correspond à sa forme monomérique. Certains perturbateurs ont en fait des activités sur des agrégats de la cible

X. Or, nous sommes intéressés par des perturbateurs qui peuvent inhiber la formation des agrégats de la cible X à partir de sa forme monomérique. Nous avons donc restreint notre recherche en sélectionnant 984 perturbateurs propriétaires chimiquement similaires aux perturbateurs ayant une activité sur la forme monomérique de la cible X, et ce, en nous basant sur les références scientifiques identifiées par des chimistes.

Des identifiants distincts devraient être systématiquement utilisés dans la littérature pour référencer une protéine dans sa forme monomérique, hétérodimérique, sous forme d'agrégat ou en fonction des modifications post-traductionnelles qui peuvent modifier sa structure et ses fonctions [98].

Nous avons réalisé une expérience MicroScale Thermophoresis (MST) afin de valider expérimentalement les 984 perturbateurs propriétaires identifiés par Pegasus. Une expérience MST permet de mesurer la force d'interaction entre deux entités, ici entre un perturbateur propriétaire et la cible X marquée avec un anticorps [99]. Parmi les 984 perturbateurs testés, nous identifions que 153 perturbateurs propriétaires sont actifs, c'est-à-dire, qu'ils se lient à la cible X. Le taux de perturbateurs identifiés comme actifs est généralement très faible dans les phases primaires de nouveaux médicaments, généralement de l'ordre de quelques pourcents, et ce, en utilisant des librairies composées de plusieurs dizaine à centaine de milliers de perturbateurs [100,101]. L'utilisation de Pegasus, dans le cadre de ce projet thérapeutique, nous a permis d'obtenir un taux d'actif de quinze pour cent avec une librairie focalisée composée de moins de mille perturbateurs propriétaires. Des chimistes sélectionnerons 50 perturbateurs parmi les 153 perturbateurs actifs en se basant sur des propriétés structurelles afin de les valider dans une nouvelle expérience MST en dose-réponse.

Une campagne de criblage sur des cellules neuronales aurait également dû débuter avec les 984 perturbateurs propriétaires identifiés par Pegasus. Cependant, nous avons perdu les conditions de seeding dans ce modèle cellulaire. Le seeding sert à initier et à quantifier le niveau d'agrégation d'une protéine par un signal de transfert d'énergie par résonance de bioluminescence [102], test indispensable pour observer l'effet de perturbateurs sur l'agrégation de la cible X. Les difficultés rencontrées pour ce protocole expérimental sont liées à la variabilité des lots de protéines utilisées, au protocole d'agrégation ou aux conditions de culture neuronale. Le développement de modèles cellulaires neuronaux pertinents et informatifs sont complexes à mettre en place et manquent, souvent, de robustesse.

En outre, dans le cadre de ce projet thérapeutique, nous investiguons l'approche de type PROTAC afin de diminuer l'agrégation de la cible X. Les PROTACs sont des molécules chimiques spécifiques qui se lient à la fois à une protéine cible et à une E3 ligase. L'E3 ligase recrute des protéines d'ubiquitination qui permet au complexe cible X – PROTAC d'être dégradé par le protéasome [77]. L'utilisation de PROTACs nous permettrait de diminuer le niveau d'agrégation de la cible X en utilisant les processus biochimiques intrinsèques des cellules. Nous avons requêté Pegasus pour identifier les PROTACs qui ont une activité sur la cible X avec une requête CYPHER similaire à celle présentée en (Requête 5). Nous avons identifié six PROTACs qui sont en train d'être synthétisés par des chimistes afin d'être testés dans un modèle cellulaire neuronal en cours de mise au point.

2.4 Conclusion

Nous venons de présenter le graphe de connaissances Pegasus pour répondre à des problématiques de projets thérapeutiques en capitalisant sur des données pharmaco-biologiques hétérogènes et de provenances multiples. Le choix du formalisme des GPE Neo4j se justifie, d'une part, par la flexibilité de la conception d'un modèle de données dans lequel des entités sont mises en relation avec nos connaissances expertes, et d'autre part, par l'efficacité du requêtage, de l'importation et du stockage de données volumineuses.

Nous utilisons, étendons et introduisons des concepts au regard des manques des représentations existantes pour nos activités de recherche. Des ressources fonctionnellement identiques sont représentées par des entités reliées entre elles par des références croisées. Nous distinguons les concepts de gènes, de transcrits et de protéines. Nous intégrons différentes classes de perturbateurs, des similarités phénotypiques et chimiques, ainsi qu'un nœud intermédiaire pour annoter contextuellement des relations entre des entités.

Nous présentons trois applications industrielles pour des projets thérapeutiques afin d'illustrer l'utilisation de Pegasus dans le processus de recherche de nouveaux médicaments. Les concepts introduits dans Pegasus et les applications que nous développons permettent de caractériser des effets hors cibles de perturbateurs, de concevoir une nouvelle expérience, et d'identifier des librairies de criblage focalisées. Nous développons Pegasus pour qu'il soit un outil de support d'aide à la décision afin de réaliser des expériences pour valider nos hypothèses et pour supporter les phases primaires de recherche de nouveaux médicaments dans leur ensemble.

Nous avons été invités à présenter Pegasus lors de workshops et d'évènements internes.

Communications scientifiques

J. Grignard, F. Fages, T. Dorval. Pegasus: A Knowledge Graph To Support Drug Discovery.
Neo4j Health Care & Life Sciences Workshop, 9-10 November, 2021.
Neo4j Graph Data Platform Webinars, 7 April, 2022.

Communications industrielles

J. Grignard, N. Boisseau, S. Lotfi, A. Gohier, A. Clary, F-X. Blaudin de Thé, C. Mannoury la Cour, F. Panayi, P. Machado, H. Tran, F. Fages, T. Dorval. Pegasus: A Knowledge Graph To Support Drug Discovery.
Servier Corporate Strategy & Executive Director, 2021
Symposium Servier – Applications of Artificial Intelligence to New Drug Development, 2020 (360 collaborateurs)
Neurology Immuno Inflammation Research Conference, 2021 (80 collaborateurs)

Un des axes de travail de Pegasus porte sur l'intégration de mesures de similarités phénotypiques obtenues à partir de campagnes de criblage phénotypiques à haut contenu. Nous présentons dans le chapitre suivant deux algorithmes pour améliorer tout d'abord la conception et l'analyse de ces expériences. Puis, nous présentons l'intégration de signatures et de similarités phénotypiques informatives au sein du graphe de connaissances Pegasus.

Chapitre 3

Algorithmes pour améliorer la conception et l'analyse d'expériences de criblage phénotypiques à haut contenu

« Tout est poison, rien n'est poison : c'est la dose qui fait le poison. »

Paracelse

Sommaire

3.1 Introduction et motivations

Dans le chapitre précédent, nous avons présenté le graphe de connaissances Pegasus pour aider à la conception et à l'analyse des expériences réalisées dans les phases primaires de recherche de nouveaux médicaments. Dans ce chapitre, nous présentons un algorithme d'identification de composés contrôles positifs ainsi qu'un algorithme de normalisation pour améliorer la conception et l'analyse d'expériences de criblage à haut contenu. Ces algorithmes permettent d'intégrer dans Pegasus des signatures et de traiter les similarités phénotypiques particulièrement informatives.

3.1.1 Contexte – Expériences de criblage phénotypiques à haut contenu

Les expériences de criblage phénotypiques à haut contenu (HCS) sont réalisées dans les phases primaires de recherche [6]. L'objectif est d'identifier des perturbations chimiques, biologiques ou génétiques qui induisent des phénotypes cellulaires (Figure 22).

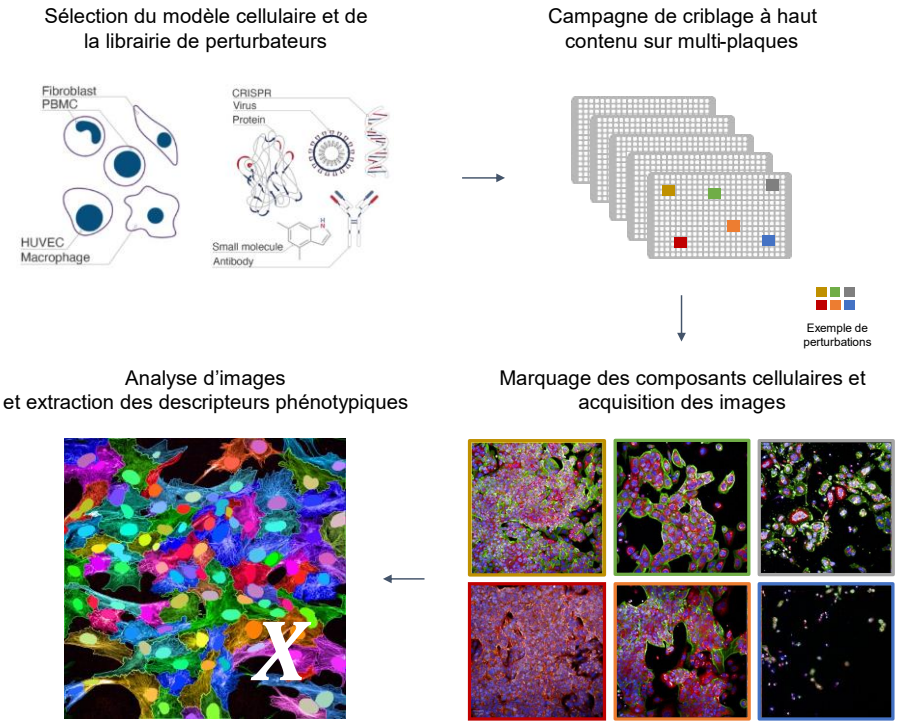

Figure 22. Criblage phénotypique à haut contenu révélant des phénotypes cellulaires hétérogènes à l'échelle de la cellule unique. Un modèle cellulaire est sélectionné compte tenu d'une question biologique à élucider et la faisabilité technique à haut-débit. Une librairie de perturbateurs est identifiée. Le modèle cellulaire est ensemencé dans chacun des puits des plaques d'une campagne de criblage. En moyenne, il y a 2.000 cellules dans chaque puit et 800 plaques seront utilisées pour le criblage de 300.000 perturbateurs. Des logiciels d'analyse couplés aux instruments de microscopie acquièrent et analysent en moyenne 10 images par puit. L'analyse d'image permet d'extraire des descripteurs phénotypiques à l'échelle de la cellule unique formant leurs signatures phénotypiques. Les images en haut à gauche proviennent de [103].

Des protocoles d'imageries permettent de capturer des images représentatives de cellules sous l'effet de traitements [104]. Le HCS est une approche multiparamétrique qui mesure simultanément des centaines à des milliers de descripteurs phénotypiques, représentant les signatures phénotypiques des cellules, après le criblage de perturbateurs.

Une expérience HCS est considérée comme agnostique aux cibles thérapeutiques car des phénotypes cellulaires sont induits sans connaissance, a priori, des cibles thérapeutiques modulées par les perturbateurs criblés [38]. Ce type d'expérience permet d'identifier des perturbateurs actifs et biologiquement pertinents dans les phases primaires de recherche [105], notamment lorsque des algorithmes d'analyses de données et d'apprentissage sont utilisés [106–108], menant au développement de nouveaux médicaments [109]. L'utilisation de mesures de similarités dans des espaces phénotypiques réduits permet de repositionner des perturbateurs pour restaurer des états phénotypiques sains [103] et a permis de mieux comprendre des processus biochimiques essentiels comme la division cellulaire en identifiant de nouveaux gènes impliqués [110].

L'utilisation de modèles cellulaires augmente la pertinence des perturbateurs actifs identifiés par rapport aux méthodes historiques et traditionnelles de criblage à haut-débit (HTS) [111]. En effet, une expérience HTS est réalisée dans un système biochimique dans lequel la cible thérapeutique, extraite de son milieu cellulaire d'origine, est placée seule avec un perturbateur [111]. Dans une expérience HCS, le criblage de perturbateurs est réalisé sur des cellules vivantes, souvent immortalisées, modifiées génétiquement ou pathologiques, permettant de capturer leurs effets dans un contexte biologique [107]. Après avoir identifié un modèle cellulaire (Figure 22), des marqueurs fluorescents sont sélectionnés, comme des anticorps, pour se lier spécifiquement à certaines espèces moléculaires du noyau, du réticulum endoplasmique, des mitochondries, du cytosquelette, de l'appareil de Golgi ou à l'ARN [38]. Le marquage permet de capturer des images représentatives à l'échelle de la cellule unique [112] afin d'observer la modulation de phénotypes cellulaires sous l'effet de perturbations (Figure 23).

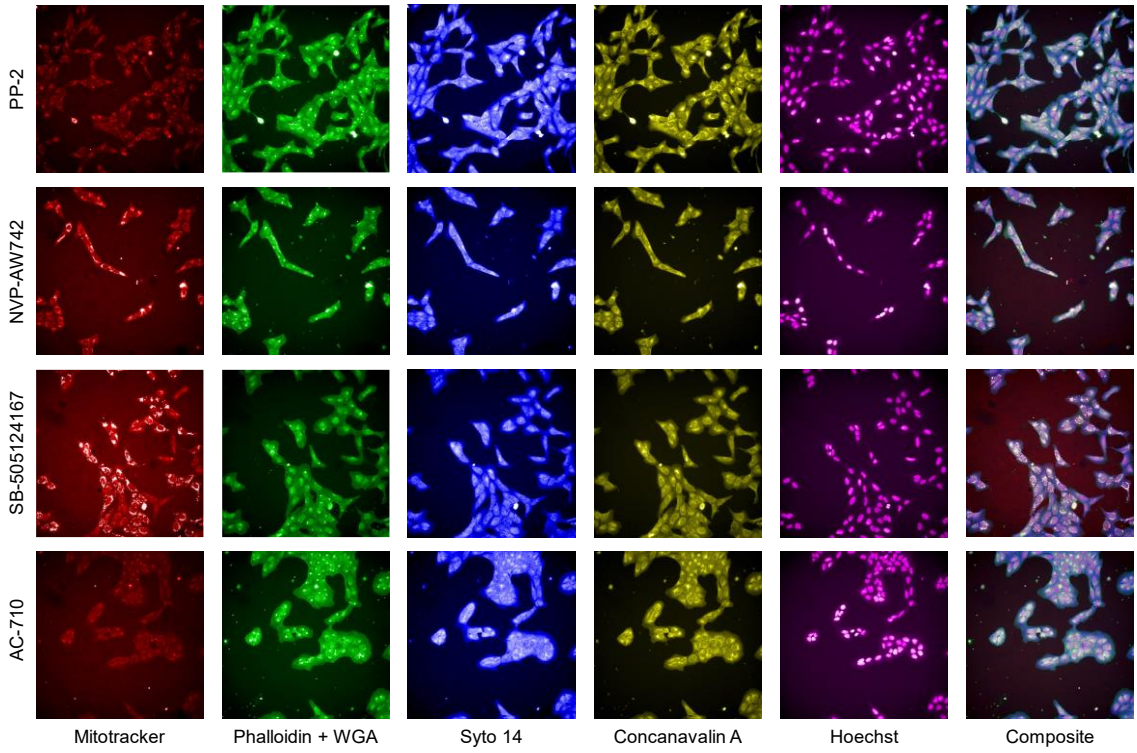

Figure 23. Phénotypes cellulaires hétérogènes obtenus après le criblage de perturbateurs chimiques en modalité HCS. Visualisation de phénotypes cellulaires induits par le criblage de quatre composés chimiques (en ordonnée), observables par l'utilisation de cinq marqueurs fluorescents (en abscisse). Le marqueur MitoTracker permet d'observer les mitochondries. Le marqueur Phalloïdine permet d'observer le cytosquelette F-actine, l'appareil de Golgi et la membrane plasmique. Le Marqueur Syto 14 permet

d'observer l'ARN cytoplasmique et le nucléole. Le marqueur Concanavalin A permet d'observer le réticulum endoplasmique. Le marqueur Hoechst permet d'observer le noyau. L'image composite (à droite) représente la superposition de tous les marqueurs simultanément. Le recouvrement des spectres de fluorescence des marqueurs fluorescents limite le nombre de marqueurs à utiliser et donc le nombre de composants cellulaires observables simultanément [113].

Des logiciels d'analyse d'images extraient des descripteurs phénotypiques fournissant les données brutes pour l'analyse de données [114]. Ces descripteurs incluent l'intensité des marqueurs, la texture d'organelles, la forme de compartiments cellulaires et permet d'obtenir un contexte spatial entre les objets cellulaires du micro-environnement [114]. Les valeurs numériques calculées pour les descripteurs phénotypiques correspondent notamment : aux intensités des marqueurs qui sont calculés pour chaque compartiment segmentés (p. ex. pour les cellules, cytoplasmes ou noyaux) ; aux textures des organelles cellulaires qui sont calculées sur la régularité des intensités des marqueurs (p. ex. pour les mitochondries) ; aux formes des compartiments cellulaires (p. ex. des membranes plasmiques, nucléaires ou l'appareil de Golgi) qui sont calculées sur les limites des compartiments segmentés et comprennent des mesures de tailles et de formes comme le périmètre, la surface et la rondeur ; et au micro-environnement comme des relations spatiales entre les compositions cellulaires et de comptage d'objets (p. ex. de cellules, de micro-noyaux ou d'organelles).

Nous présentons en (Figure 24), des exemples de distributions de descripteurs phénotypiques obtenus après le criblage de huit composés chimiques montrant une hétérogénéité de réponses des signatures phénotypiques.

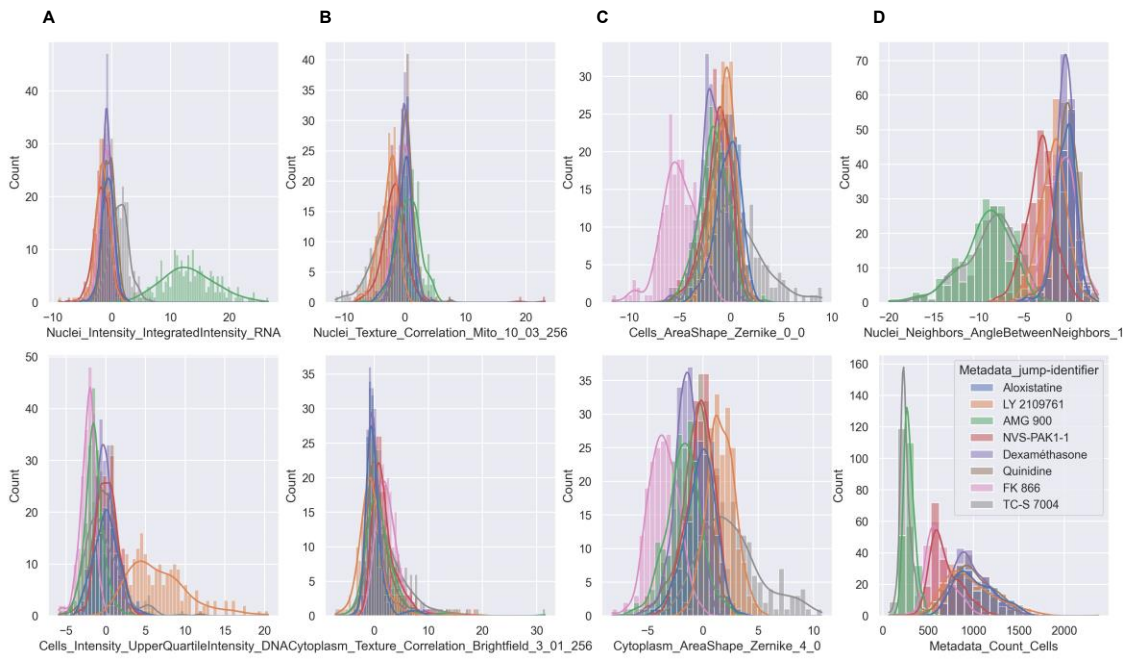

Figure 24. Hétérogénéité des réponses des descripteurs phénotypiques sous l'effet de perturbations chimiques. (A) Exemples de descripteurs liés à l'intensité de marqueurs fluorescents de l'ARN et de l'ADN. **(B)** Exemples de descripteurs liés à la texture de compartiments cellulaires comme le noyau et le cytoplasme. **(C)** Exemples de descripteurs liés à la forme, ici la surface, des cellules ou des cytoplasmes. **(D)** Exemples de descripteurs liés au microenvironnement, notamment le nombre de cellule et des mesures d'angles entre les noyaux des cellules. Chaque couleur de chaque figure représente la distribution des descripteurs phénotypiques sous l'effet de huit composés chimiques.

3.1.2 Problématique – Comment normaliser les données HCS ?

La normalisation des données HCS, espace de haute dimension, est un problème ouvert. L'objectif d'une méthode de normalisation est de rendre, dans un premier temps, comparables et similaires, les distributions de composés identiques dispensés dans les plaques d'une campagne de criblage, comme des composés contrôles, afin d'avoir la capacité d'analyser les données dans leur ensemble. De plus, l'application d'une méthode de normalisation doit permettre d'augmenter la pertinence des perturbateurs identifiés comme actifs.

Notre problématique est la suivante : comment normaliser les données HCS provenant des plaques d'une campagne de criblage phénotypique ?

3.1.3 État de l'art – Normalisation des données de criblage

Les expériences HCS s'opposent aux expériences traditionnelles de criblage HTS bien que ces approches soient complémentaires dans le processus de recherche [6]. Une expérience HTS est réalisée dans un système biochimique dans laquelle une seule variable est mesurée comme l'activité d'une cible thérapeutique [37] alors qu'une expérience HCS permet de mesurer, simultanément, des centaines à des milliers de descripteurs phénotypiques après traitements [38].

Un contrôle positif, composé chimique actif de référence, permet d'évaluer les effets des perturbateurs criblés en fournissant la borne maximale anticipée pour les valeurs que peut prendre une variable. Un contrôle négatif est utilisé comme référence pour démontrer que des perturbateurs n'ont pas d'activité. La dispense de composés contrôles, positifs et négatifs, dans les plaques d'une campagne de criblage, permet d'utiliser des méthodes de normalisation bien définies pour une expérience HTS [115,116]. La présence de contrôles permet d'évaluer également la qualité des données en calculant le degré de séparation d'une variable. Par exemple, le facteur Z', que nous présentons en (Formule 1), est adapté pour quantifier les effets entre deux conditions dans une expérience HTS [117].

$$\text{Facteur Z'} = 1 - \frac{3(\sigma_p + \sigma_n)}{|\mu_p - \mu_n|}$$

Formule 1. Formule du facteur Z' pour quantifier les effets de perturbateurs. Le facteur Z' calcule le degré de séparation entre deux conditions expérimentales [117]. μ_p et σ_p sont la moyenne et l'écart-type de la variable HTS mesurée sous l'effet d'un traitement (p. ex un contrôle positif) et μ_n et σ_n sont respectivement ceux d'un autre traitement (p. ex. un contrôle négatif). La valeur du facteur Z' varie de -∞ à 1. Un facteur Z' tel que Z' >= 0.5 indique que les distributions statistiques d'un variable HTS entre deux conditions expérimentales sont séparées. L'utilisation de ce facteur suppose une distribution gaussienne des valeurs des contrôles, une hypothèse rencontrée pour une variable HTS en raison du théorème central limite [117,118].

La communauté du criblage a adopté, entre autres, le facteur Z′ comme métrique de contrôle des descripteurs phénotypiques HCS [118]. Or, le facteur Z′ n'est pas pertinent pour les données HCS. En effet, il existe un décalage entre l'objectif de ce facteur, développé historiquement pour une expériences HTS et ce que mesure les descripteurs phénotypiques dans une expérience HCS. Nous ne pouvons pas contrôler la qualité des données ni normaliser de façon indépendante chaque descripteur phénotypique pris un à un dans une expérience HCS.

Certaines méthodes de normalisation HCS n'utilisent pas de contrôles positifs pour normaliser les données HCS et se basent uniquement sur la présence d'un contrôle négatif [116]. Cependant, l'utilisation d'un contrôle négatif est problématique lorsqu'une campagne de criblage HCS, composées de plusieurs dizaines à centaines de plaques, est réalisée. En effet, des erreurs peuvent se produire comme la non-dispense de composés et l'utilisation d'un contrôle négatif ne permet pas de distinguer un composé inactif d'un composé non-dispensé.

Il n'existe pas de méthode de normalisation de référence, à notre connaissance, pour normaliser globalement l'entièreté de l'espace phénotypique induit par les descripteurs phénotypiques en utilisant la présence de multiples contrôles positifs et négatifs.

3.1.4 Approche méthodologique – Développement de deux algorithmes pour améliorer la conception et l'analyse d'expériences de criblage phénotypiques

Dans un premier temps, nous développons un algorithme d'identification de composés contrôles positifs. Cette algorithme permet d'automatiser la sélection de quatre contrôles positifs à partir d'une librairie chimique interne constituée de vingt-sept composés de référence. La validation de l'algorithme est semi-quantitative en couplant des algorithmes de réduction de dimensions et une vérification visuelle des phénotypes cellulaires obtenus. Nous montrons que les composés contrôles positifs identifiés par l'algorithme maximisent les effets des descripteurs phénotypiques et induisent des signatures phénotypiques variées.

Les composés contrôles positifs sont dispensés dans chaque plaque d'une campagne de criblage en plus de la présence d'un contrôle négatif.

Dans un deuxième temps, nous développons un algorithme de normalisation de données HCS qui utilise la présence de composés contrôles comme points de référence pour effectuer un recalage global et paramétrique des données dans un espace phénotypique réduit. Nous montrons que l'algorithme de normalisation permet de rendre similaires les signatures phénotypiques issues d'une campagne de criblage HCS interne dans laquelle, 20.000 composés propriétaires, en plus des composés contrôles, ont été dispensés dans soixante plaques.

Après normalisation, nous intégrons les signatures et des similarités phénotypiques informatives au sein du graphe de connaissances Pegasus.

Les contributions de ce chapitre sont également réalisées dans le cadre du consortium international JUMP-CP. L'objectif est de cribler, en modalité HCS [113], 120.000 composés chimiques issus de librairies de partenaires industriels et académiques sur un modèle cellulaire de type U2OS (ostéosarcome), en réplicats, à travers différents sites dans le monde. Une expérience CRISPR/Cas9 sur l'ensemble des gènes du modèle cellulaire sera également réalisée. Nous intégrerons les signatures et des similarités phénotypiques issues de JUMP-CP dans Pegasus.

3.2 Algorithme de sélection de composés contrôles positifs

3.2.1 Contexte – Librairie interne de 27 composés chimiques de référence

Un composé contrôle positif, dispensé dans les plaques d'une campagne de criblage, doit induire un effet caractérisé et reproductible. Or, il n'existe pas de composé unique qui puisse moduler simultanément et de façon caractérisé tous les descripteurs phénotypiques qui peuvent être calculés dans une expérience HCS.

Nous avons développé une librairie chimique interne, la ToolBox, comprenant vingt-sept composés chimiques (Table 5). Ces composés, du fait de leurs mécanismes d'action variés, ont été sélectionnés à partir de la littérature pour maximiser les réponses observables par les descripteurs phénotypiques (Table 5).

Nom	Identifiant	Mécanisme d'action
Cycloheximide	002-070-550	Bloque la biosynthèse des protéines en inhibant la phase d'initiation et d'élongation de la synthèse protéique
Forskoline	006-168-487	Activateur de l'adénylate cyclase permettant son activation sans passer par les récepteurs et protéines G
Acide valproique	003-666-306	Inhibe l'histone déacétylase
Ciprofibrate	003-666-795	Active les processus biochimiques de l'oxydation des acides gras
Doxorubicine	003-933-100	Inhibe la réplication de l'ADN et la topoisomérase II
Méthotrexate	001-779-666	Inhibe l'enzyme dihydrofolate réductase entrainant une inhibition de la synthèse de l'ADN et de l'ARN
Parthenolide	008-268-168	Inhibe les protéines HDAC1 et TCP et module des effets inflammatoires médiées par les voies de signalisation NF-κB
Colchicine	002-507-437	Interfère avec la polymérisation de la tubuline, perturbe la mitose, et empêche la réponse inflammatoire
Paclitaxel	001-742-627	Stimule l'assemblage des dimères de tubuline et stabilise les microtubules en empêchant leur dépolymérisation
Demecolcine	002-507-437	Inhibe la formation du fuseau mitotique
Vinblastine, sulfate	002-518-262	Interfère avec le métabolisme des acides aminés et inhibe l'assemblage du fuseau mitotique
Acide rétinoïque	000-883-857	Active les récepteurs de l'acide rétinoïque entraînant une diminution de la prolifération cellulaire et inhibe la télomérase
Z-Leu-Leu-Leu-al	009-019-420	Inhibiteur des processus biochimiques du protéasome
17-AAG	003-983-836	Inhibe la protéine HSP90 favorisant la dégradation protéasomique des protéines de signalisation oncogènes
Sodium orthovanadate	044-193-358	Inhibiteur compétitif des ATPases, des phosphatases alcalines et acides, et des protéines-phosphotyrosine-phosphatases
Brefeldin A	001-739-555	Inhibe la formation et le transport des vésicules entre le réticulum endoplasmique et l'appareil de Golgi et entraîne la fusion membranaire de ces deux compartiments
3-Isobutyl-1-methylxanthine	001-792-510	Inhibiteur de la phosphodiestérase des nucléotides cycliques augmentant l'AMP et le GMP cyclique
Puromycine, dichlorhydrate	003-939-172	Analogue de l'extrémité 3' de l'aminoacyl-ARNt provoquant sa terminaison prématurée et inhibe la synthèse des protéines

Spiramide	003-983-510	Antagoniste dopaminergique et sérotoninergique
Paracétamol	000-150-777	Inhibe la synthèse et la libération des prostaglandines dans le système nerveux central
Nocodazole	001-889-558	Inhibe la dépolymérisation des microtubules en se fixant sur une arginine de la β-tubuline et bloque le développement des cellules en phase M de la mitose
Chloroquine	001-783-623	Inhibe les processus d'autophagie et anticoronaviral
Actinomycine D	001-739-741	Inhibe la transcription de l'ADN par l'ARN polymérase et provoque des cassures de l'ADN simple brin
Tunicamycin	008-268-178	Inhibe la N-glycosylation nécessaire à la fixation des précurseurs des N-hétérosides sur le dolichol diphosphate
Ulinastatin	046-860-046	Inhibe plusieurs protéases pro-inflammatoires et réduit les niveaux de cytokines inflammatoires
Cytochalasin D	003-929-647	Inhibe l'association et la dissociation des sous-unités de l'extrémité cannelée des filaments d'actine
Thapsigargine	003-959-790	Augmente le taux de calcium dans le lumen du réticulum endoplasmique engendrant l'apoptose

Table 5. Composés chimiques de la ToolBox, une librairie chimique interne, induisant des signatures phénotypiques variées.

Avant de réaliser une campagne de criblage, les composés de la ToolBox sont criblés en dose-réponse, sur une gamme de seize concentrations, de 10^{-7} µM à 10^{1} µM, en intra-réplicats, et en inter-réplicats (Figure 25). Ce protocole expérimentale permet de capturer la variabilité des réponses des descripteurs phénotypiques à plusieurs concentrations tout en augmentant la robustesse des analyses de données.

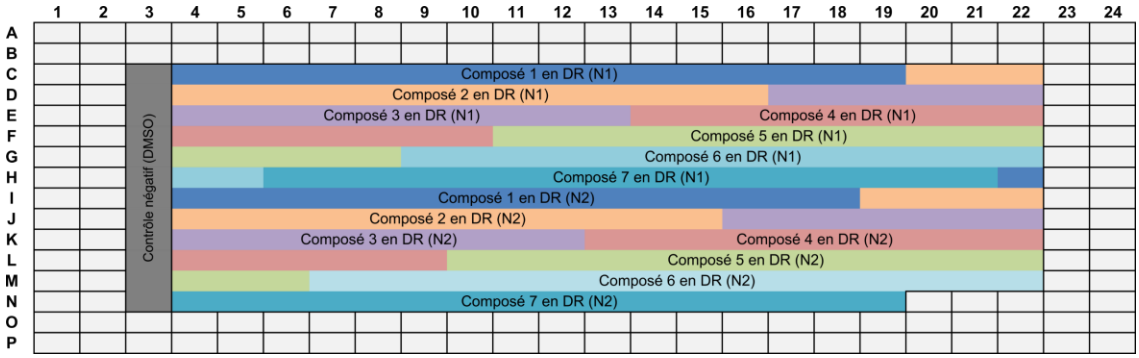

Figure 25. Plan de plaque de criblage des composés de la ToolBox. Les composés chimiques de la ToolBox sont criblés en dose-réponse en intra-réplicats (N1, N2) et en inter-réplicats (un même composé est criblé sur deux plaques distinctes). Ici, le plan de plaque de sept composés chimiques est présenté. Le composé 1 est criblé à des concentrations croissantes entre les puits C4 (dose la plus faible (~10^{-7} µM)) et C19 (dose la plus forte (~10^{1} µM)). Les bords des plaques ne sont pas utilisés pour minimiser les effets de bords couramment observés dans les campagnes de criblage [119]. Le contrôle négatif (DMSO) est criblé à une dose unique (~5 µM) en colonne 3.

Seul quatre composés de la ToolBox peuvent être dispensés dans les plaques d'une campagne de criblage pour assurer un nombre de réplicats suffisant. L'objectif de l'algorithme d'identification de contrôles positifs est de sélectionner une combinaison de quatre composés, à une concentration chacun, parmi les vingt-sept composés de la ToolBox, qui maximisent, ensemble, les réponses des descripteurs phénotypiques calculés.

3.2.2 Étapes de l'algorithme de sélection de contrôles positifs

Nous présentons les différentes étapes de l'algorithme dans le cadre d'une expérience HCS réalisée en interne avant de le résumer en pseudo-code en (Algorithme 2). Les composés de la ToolBox ont été criblés dans un modèle cellulaire de cancer du poumon humain porteur d'une mutation KRAS (H358). L'analyse des images du criblage de la ToolBox extrait cinquante-six descripteurs phénotypiques qui incluent l'intensité, la texture et la forme des cellules marquées avec trois marqueurs fluorescents (BP676/29 : Phalloidine, BP445/45 : Hoescht, BP600/37 : Mitotracker) ainsi que des descripteurs liés à la forme des cellules et leur nombre. Les données HCS calculées à l'échelle de la cellule unique sont agrégées par puits à la suite de l'analyse d'images.

La première étape de l'algorithme consiste à recaler les données HCS des plaques par rapport au contrôle négatif (DMSO) avec la formule présentée en (Formule 2). Nous n'appliquons pas la formule de standardisation classique, présentée en (Formule 4), car les composés sont criblés en dose-réponse. Puis, les valeurs des descripteurs phénotypiques, pour chaque composé et à chaque concentration, sont agrégées par valeurs médianes des quatre réplicats.

$$X^P = X^P - median(X^P_{DMSO})$$

Formule 2. Formule de recalage des données phénotypiques par rapport au contrôle négatif. Les colonnes de la matrice X représentent les descripteurs phénotypiques calculés après l'extraction des caractéristiques plus des méta-informations comme le nom de la plaque P, l'identifiant d'un puit, le nom et la concentration d'un composé. Les lignes de cette matrice correspondent aux valeurs des descripteurs phénotypiques d'un composé qui a été criblé dans un puit de la plaque P a une concentration donnée. La matrice X^P représente l'ensemble des descripteurs d'une plaque P. La matrice X^P_{DMSO} correspond aux descripteurs d'une plaque P pour le composé DMSO. $median(X^P_{DMSO})$ correspond à la valeur médiane de tous les points du contrôle négatif pour tous les descripteurs d'une plaque P. Cette formule permet de recaler les données HCS de tous les composés de chacune des plaques par rapport aux valeurs médianes du contrôle négatif (DMSO) de ces plaques.

Ensuite, les concentrations pour chaque composé, qui seront utilisées dans les étapes suivantes de l'algorithme, sont déterminées. Le calcul d'une concentration n'est pas trivial puisque les descripteurs phénotypiques peuvent varier indépendamment en fonction des concentrations, ne sont pas forcément sigmoïdales et des comportements non linéaires peuvent être observés. La concentration efficace semi-maximale (EC50) est définie comme la concentration nécessaire pour obtenir un effet de cinquante pour cent. À l'EC50, la majorité des descripteurs sont modulés par les composés. Pour des concentrations plus fortes nous observons une mortalité cellulaire importante.

Nous utilisons une procédure d'optimisation des moindres carrés non linéaire[8] pour ajuster les paramètres d'une fonction sigmoïdale aux données du descripteur correspondant au nombre de cellules. Cette étape de l'algorithme permet d'identifier les EC50 respectives des composés et nous illustrons en (Figure 26) des résultats obtenus pour cinq composés chimiques de la ToolBox dans le modèle cellulaire d'étude.

[8] Fonction `scipy.optimize.curve_fit` de la librairie Python SciPy [120]

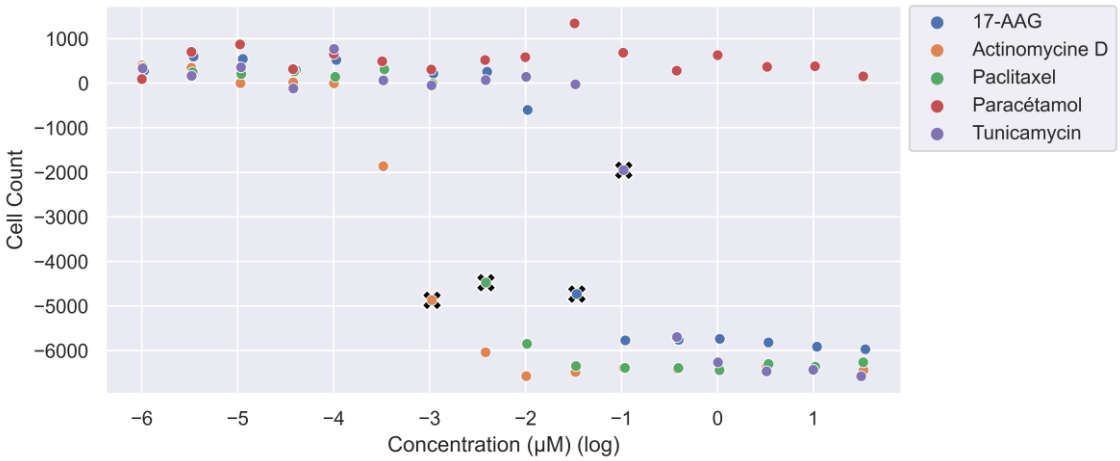

Figure 26. Concentrations efficaces semi-maximales hétérogènes pour cinq composés chimiques de la ToolBox criblés dans un modèle cellulaire. Diagrammes en dose-réponses du descripteur phénotypique correspondant au nombre de cellules après le criblage de composés issus de la ToolBox. La croix noire correspond au point de concentration la plus proche de l'EC50 calculée. Nous observons que quatre EC50 distinctes ont été identifiées pour les quatre composés suivants (Actinomycine D, Paclitaxel, 17-AAG, Tunicamycin). Le composé (Paracétamol) n'induit pas d'effet et aucun EC50 n'a été identifié.

Les composés qui n'induisent pas d'effet sur le nombre de cellules sont écartés car nous cherchons à identifier des composés qui induisent des effets, ils doivent donc être distinct du contrôle négatif. Les neufs composés suivants (3-Isobutyl-1-methylxanthine - Acide rétinoique - Acide valproique, sodique - Chloroquine, Ciprofibrate - Forskoline - Paracétamol - Sodium orthovanadate - Ulinastatin) sont écartés.

Puis, pour les dix-huit composés restants, les valeurs absolues des descripteurs phénotypiques à leurs EC50 sont calculées. En effet, l'objectif de l'algorithme réside dans la sélection de quatre composés qui maximisent les réponses des descripteurs phénotypiques qui peuvent varier positivement ou négativement. Nous illustrons en (Table 6) des exemples de valeurs de descripteurs phénotypiques induites par cinq composés aux EC50 calculées.

Composé	Concentration	d1	d2	d3	d4
17-AAG	-1.46 (10)	4728.75	0.005945	6199.972287	1332.589390
Actinomycine D	-2.97 (7)	4866.75	0.021566	5284.250606	64.275405
Colchicine	-1.98 (9)	2817.00	0.028851	2504.165348	599.645854
Cycloheximide	-0.42 (13)	2499.00	0.000924	1606.705865	353.861908
Cytochalasin D	-0.95 (12)	3029.75	0.027043	4619.898187	336.586241

Table 6. Données phénotypiques en entrée de l'algorithme de sélection de composés contrôles positifs. Exemples de valeurs numériques de quatre descripteurs phénotypiques (d1 : Cell Count, d2 : Cytoplasm Roundness, d3 : Intensity Cytoplasm BP600/37 Mean, d4 : Intensity Nucleus BP445/45 Mean) pour cinq composés à la valeur de l'EC50 calculée sur le descripteur mesurant le nombre de cellule. Les valeurs entre parenthèses pour la colonne concentration représente l'index de la concentration calculée. Les concentrations affichées sont en logarithme base 10.

Nous calculons ensuite la matrice des rangs des valeurs absolues des descripteurs phénotypiques pour tous les composés (Figure 27). Nous calculons une matrice de rang car les valeurs des descripteurs phénotypiques varient sur des plages de valeurs non comparables entre elles (ordre de grandeur allant de 10^{-4} à 10^{10}).

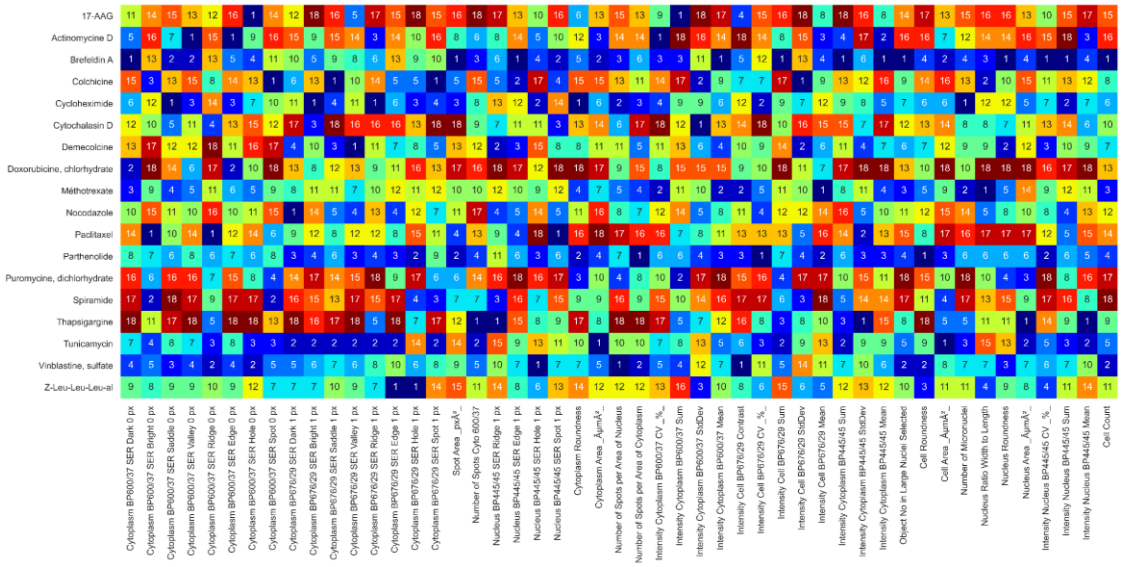

Figure 27. Matrice de rang des valeurs des descripteurs phénotypiques induits par les composés chimiques de la ToolBox. Chaque ligne i représente un composé chimique qui a passé les étapes de prétraitement. Chaque colonne j correspond au score associé d'un descripteur j pour une composé i. La valeur calculée, à l'intersection d'une ligne i et d'une colonne j, représente le rang de la valeur pour le descripteur j pour l'ensemble des composés i. Par exemple, la valeur de rang de 18 à l'intersection de la 15ème ligne (Thapsigargine) et de la 1ère colonne (Cytoplasm BP600/37 SER Dark 0 px) indique que le composé Thapsigargine a induit la 18ème valeur la plus forte pour le descripteur Cytoplasm BP600/37 SER Dark 0 px par rapport aux autres composés. Pour ce descripteur phénotypique, le composé Brefeldin A (3ème ligne) a induit la valeur la plus faible.

À partir de la matrice de rang présentée en (Figure 27), un score est calculé pour chacune des 3.060 combinaisons de quatre composés parmi les dix-huit à une concentration chacun. Ce score correspond à la somme des valeurs maximales de rang pour les descripteurs phénotypiques et est calculé avec la formule présentée en (Formule 3). Cette formule permet de calculer un score qui maximise les réponses des descripteurs phénotypiques issus d'une combinaison de quatre composés à une concentration chacun.

$$S_{c1,c2,c3,c4} = \sum_{i=1}^{d} \max_{c1,c2,c3,c4} \left(M^{d}_{c1,c2,c3,c4}\right)$$

Formule 3. Formule associant un score à une combinaison de quatre composés à une dose chacun qui maximise les effets des descripteurs phénotypiques. La matrice M correspond à la matrice de rang calculée et présentée en (Figure 27). La matrice $M_{c1,c2,c3,c4}$ correspond aux rangs des valeurs des données HCS pour l'ensemble des descripteurs phénotypiques pour une combinaison de quatre composés ($c1$, $c2$, $c3$, $c4$). Pour un descripteur d, $\max_{c1,c2,c3,c4}\left(M^{d}_{c1,c2,c3,c4}\right)$ calcule la valeur maximale pour le descripteur d entre les quatre composés. Le score associé à la combinaison de quatre composés ($S_{c1,c2,c3,c4}$) correspond à la somme des valeurs de rang maximales de chaque descripteur d.

Nous présentons le calcul du score (Formule 3) associé à une combinaison de quatre composés et la distribution des scores pour l'ensemble des combinaisons en (Figure 28).

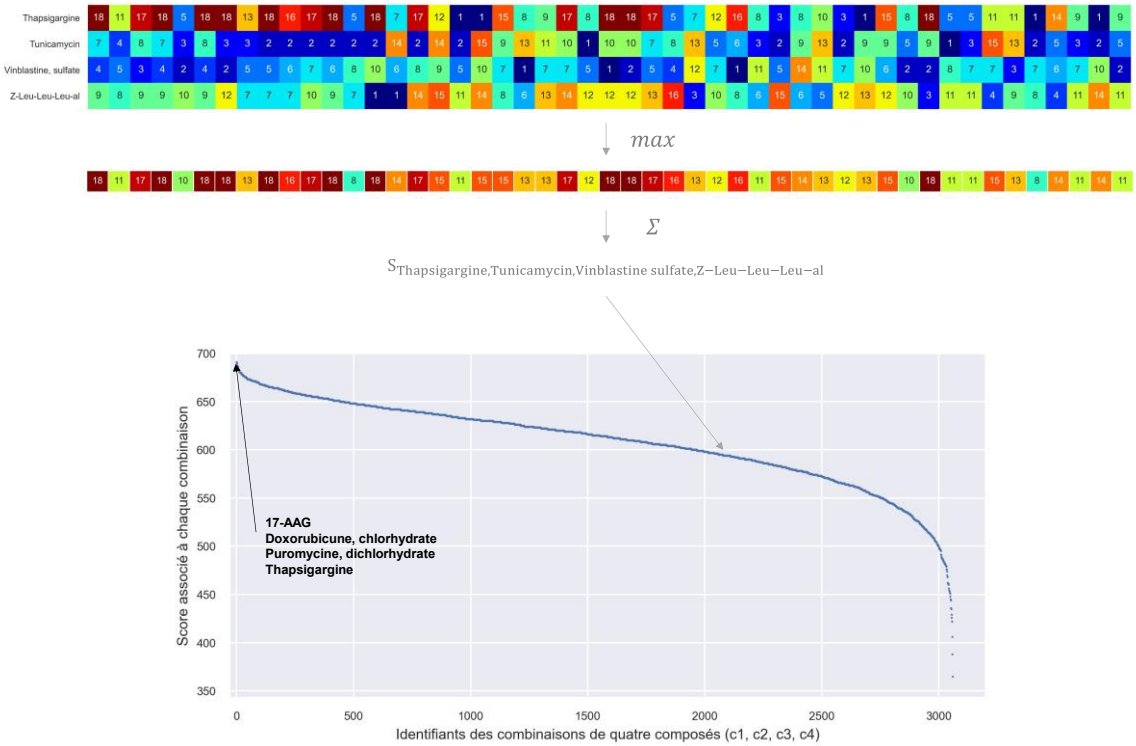

Figure 28. Scores des combinaisons de quatre composés maximisant les réponses des descripteurs phénotypiques. Nous illustrons le calcul d'un score associé à une combinaison des quatre composés (Thapsigargine, Tunicamycin, Vinblastine, sulfate, Z-Leu-Leu-Leu-al). La première étape consiste à prendre la valeur des rangs maximale des descripteurs pour les quatre composés de la combinaison. La deuxième étape consiste à prendre la somme des valeurs maximales pour tous les descripteurs. Le score associé à une combinaison correspond à la somme des valeurs maximales pour chaque composé. Sur la figure en bas, la distribution des scores associés pour chaque combinaison de quatre composés. En abscisses les identifiants des combinaisons de quatre composés. En ordonnée, le score associé à ces combinaisons calculé avec la formule présentée en (Formule 3).

L'algorithme d'identification de contrôles positifs identifie les quatre composés de la combinaison avec la valeur de score maximale calculée à partir de la formule (Formule 3). Dans cette expérience, à l'issue de l'exécution de l'algorithme, les composés de la combinaison (17-AAG, Doxorubicine, chlorhydrate, Puromycine, diclorhydrate, Thapsigargine) aux concentrations en logarithme base 10 respectives (-1.62, -1.46, -0.60, -2.37) sont identifiés comme ceux maximisant le score calculé avec la (Formule 3). Ces composés maximisent les réponses des descripteurs phénotypiques calculés et nous présentons en (Algorithme 2) le pseudo-code associé à l'algorithme de sélection de composés contrôle positifs.

Entrée : Ensemble des données HCS $\boldsymbol{X}_{m,c}^{p,d}$ issues du criblage de la ToolBox où p, d, m, c représentent respectivement les indices d'une plaque, d'un descripteur phénotypique, d'un composé chimique et d'une concentration. Les valeurs numériques N_p, N_d, N_m représentent respectivement le nombre de plaques, de descripteurs phénotypiques et de composés chimiques.

Sortie : Combinaison de quatre composés distincts m à des doses uniques c : $\{(m_1, c_1), (m_2, c_2), (m_3, c_3), (m_4, c_4)\}$

Algorithme :

Étape 1 : Recaler les valeurs des descripteurs phénotypiques de chacune des plaques.

Pour p = 1, …, N_p **faire**
  Pour d = 1, …, N_d **faire**
    $\boldsymbol{X}_{:,:}^{p,d} = \boldsymbol{X}_{:,:}^{p,d} - median(\boldsymbol{X}_{DMSO,:}^{p,d})$

Étape 2 : Agréger les valeurs des descripteurs phénotypiques par réplicats pour chaque composé. $\widetilde{\boldsymbol{X}}_{m,c}^{d}$ *correspond aux données après le recalage et l'agrégation des descripteurs phénotypiques.*

Pour m = 1, …, N_m **faire**
  Pour d = 1, …, N_d **faire**
    Pour c = 1, …, N_c **faire**
      $\widetilde{\boldsymbol{X}}_{m,c}^{d} = median(\boldsymbol{X}_{m,c}^{:,d})$

Étape 3 : Calculer la concentration efficace semi-maximale (EC50) sur le descripteur phénotypique mesurant le nombre de cellules pour chaque composé. $\widehat{\boldsymbol{X}}_{m}^{d}$ *correspond aux valeurs des descripteurs phénotypiques pour chaque composé à leurs concentrations EC50.*

Pour m = 1, …, N_m **faire**
  $EC50_m = CURVE\_FITTING(\widetilde{\boldsymbol{X}}_{m,:}^{cell\_count})$
  Pour d = 1, …, N_d **faire**
    $\widehat{\boldsymbol{X}}_{m}^{d} = \widetilde{\boldsymbol{X}}_{m,EC50_m}^{d}$

Étape 4 : Calculer les scores de chaque combinaison de quatre composés à une dose chacun à partir de la matrice des rangs des valeurs des descripteurs phénotypiques. $\boldsymbol{M}_s^d$ *correspond à la matrice des rangs des valeurs des descripteurs phénotypiques présentée en (Figure 27).*

Pour s = $[\{(m_1, EC50_{m1}), (m_2, EC50_{m2}), (m_3, EC50_{m3}), (m_4, EC50_{m4})\}, \ldots, \{(m_i, EC50_{mi}), (m_j, EC50_{mj}), (m_k, EC50_{mk}), (m_l, EC50_{ml})\}]$ **faire**

$$S_s = \sum_{i=1}^{d} \underset{s}{max}(\boldsymbol{M}_s^d)$$

Étape 5 : Retourner la combinaison de quatre composés à leurs EC50 qui maximise le score calculé à l'étape précédente.

Retourner $\{(m_1, c_1), (m_2, c_2), (m_3, c_3), (m_4, c_4)\} := \underset{m1,m2,m3,m4}{\max} S_{m1,m2,m3,m4}$

Algorithme 2. Algorithme de sélection de composés contrôles positifs pour identifier une combinaison de quatre composés à une concentration chacun qui maximisent les réponses des descripteurs phénotypiques.

3.2.3 Validation de l'algorithme

La validation de l'algorithme de sélection de contrôles positifs est semi-quantitative. Nous couplons des algorithmes de réduction de dimension et une vérification experte à partir des images. L'utilisation d'algorithmes de réduction de dimensions permet d'observer la répartition des composés de la ToolBox dans un espace phénotypique réduit. La vérification experte des images nous permet de valider que les composés identifiés par l'algorithme induisent des phénotypes variés. Nous présentons en (Figure 29), l'espace engendré par l'application d'une analyse discriminante linéaire[9] (LDA) sur les données de la ToolBox et des images représentatives des cellules après le criblage des composés chimiques. Ces figures montrent que les quatre composés identifiés par l'algorithme en plus du contrôle négatif, à quatre concentrations distinctes, induisent bien des signatures phénotypiques variées.

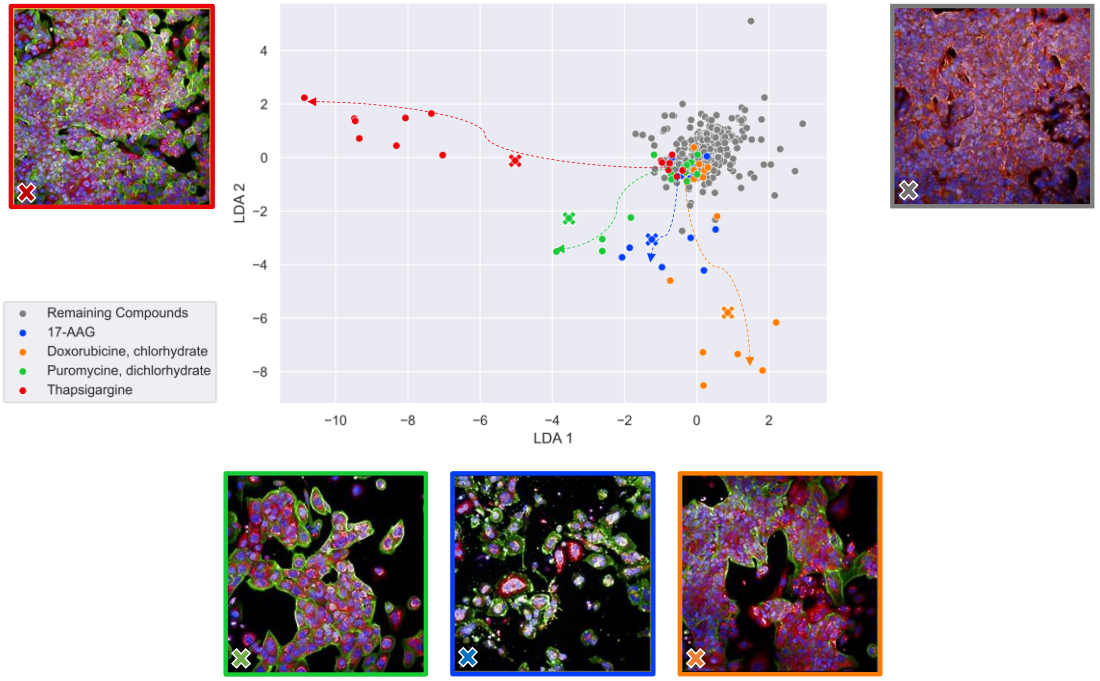

Figure 29. Espace phénotypique réduit montrant l'hétérogénéité des réponses des composés contrôles positifs identifiés par l'algorithme de sélection dans un modèle cellulaire. Analyse discriminante linéaire (LDA) appliquée sur l'ensemble des données HCS après recalage pour observer la séparation des contrôles (classes) identifiés par l'algorithme par rapport aux autres composés. Les classes de la LDA correspondent aux contrôles positifs (classes 1, 2, 3, 4), au contrôle négatif (classe 5) et aux autres composés de la ToolBox (classe 6). L'espace réduit engendré par la LDA montre la répartition des composés de la ToolBox. Le nuage de points (en gris) centré en zéro correspond au contrôle négatif (qui n'est pas affiché par souci de clarté), aux composés à leurs premières concentrations et aux composés inactifs dont les signatures phénotypiques sont identiques à celles du contrôle négatif. Les composés contrôles identifiés par l'algorithme séparent l'ensemble des classes. Pour les quatre contrôles positifs, les symboles croix correspondent à leurs EC50. Cinq images représentatives issues du criblage sont présentées. Nous observons que les composés contrôles positifs induisent des phénotypes variés et distincts du contrôle négatif (image aux bords gris correspond au nuage de point en gris). Nous montrons également l'évolution des signatures phénotypiques par des flèches en pointillées depuis la dose la plus faible (qui part du nuage du DMSO), en passant par la dose sélectionnée pour les contrôles positifs (points avec des croix) jusqu'à la dose la plus forte qui induit une mortalité cellulaire importante. En fonction des concentrations, des composés chimiques n'induisent pas les mêmes signatures phénotypiques.

[9] Fonction `analysis.LinearDiscriminantAnalysis` de la librairie scikit-learn [121].

3.3 Algorithme de normalisation de données phénotypiques

3.3.1 Contexte – Criblage de composés contrôles dans les plaques d'une campagne de criblage

Nous développons un algorithme de normalisation de données HCS qui utilise des contrôles positifs et un contrôle négatif dispensés dans les plaques d'une campagne de criblage. Nous présentons les étapes de l'algorithme sur une expérience HCS interne constituée de soixante plaques dans laquelle 20.000 composés propriétaires ont été dispensés. Le modèle cellulaire et les descripteurs phénotypiques calculés sont identiques à ceux présentés dans la section précédente. Le contrôle négatif est le DSMO et les quatre contrôles positifs sont : Doxorubicine, Thapsigargine, Cytochalasin D et le 17-AAG. Ces contrôles positifs et leurs concentrations ont été choisis historiquement manuellement et sont dispensés huit fois dans chaque plaque. Les composés propriétaires sont criblés à une concentration unique de 5 µM.

La première étape de l'algorithme consiste à standardiser les données HCS de chaque plaque par rapport au contrôle négatif avec la formule présentée en (Formule 4).

$$X^P = \frac{X^P - median(X^P_{DMSO})}{std(X^P_{DMSO})}$$

Formule 4. Formule de standardisation des données phénotypiques par rapport au contrôle négatif. La matrice X^P représente l'ensemble des descripteurs phénotypiques d'une plaque P. Cette formule permet de standardiser les données HCS en centrant et réduisant les données HCS d'une plaque P par rapport aux valeurs médianes et à l'écart-type du contrôle négatif (DMSO) d'une plaque P.

Nous observons en (Figure 30) une variabilité des réponses des descripteurs phénotypiques pour les composés contrôles dispensés dans les plaques de la campagne de criblage sans méthode de normalisation. L'objectif de l'algorithme de normalisation est de rendre similaire les distributions des contrôles entre les plaques afin d'avoir la capacité d'analyser les données dans leur ensemble.

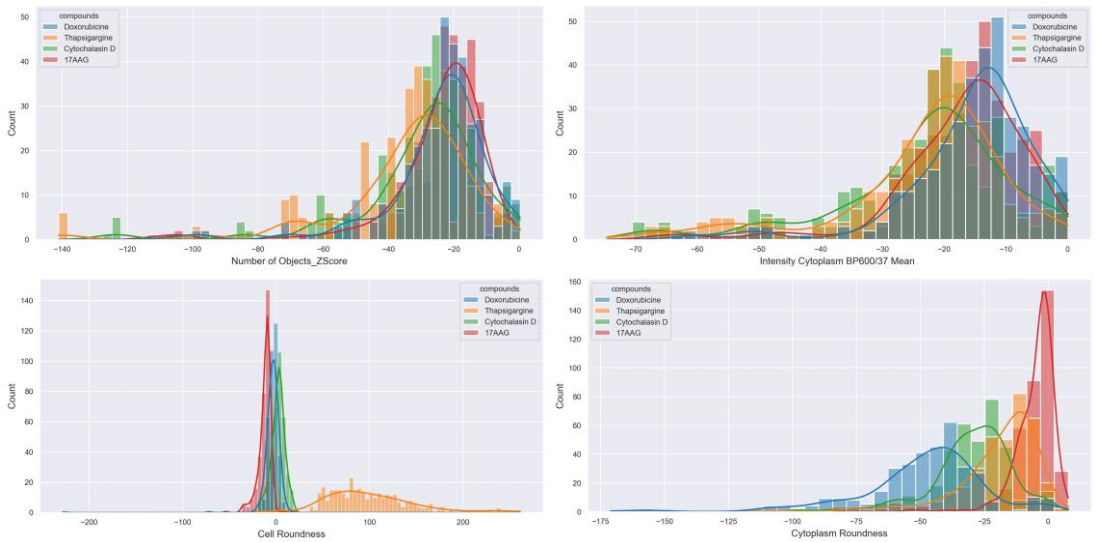

Figure 30. Hétérogénéité des réponses de descripteurs phénotypiques dans une campagne de criblage sous l'effet de composés contrôles positifs sans normalisation. Chaque figure représente les distributions

des valeurs de quatre descripteurs phénotypiques (un par figure) pour les quatre composés contrôles positifs issus des soixante plaques de la campagne de criblage. De fortes variabilités des réponse des descripteurs phénotypiques sont observées, et ce, pour un même composé contrôle positif.

3.3.2 Étapes de l'algorithme de normalisation

Les étapes de l'algorithme de normalisation sont tout d'abord illustrées en (Figure 31) avant d'être décrites. L'algorithme est résumé en pseudo-encode en (Algorithme 3).

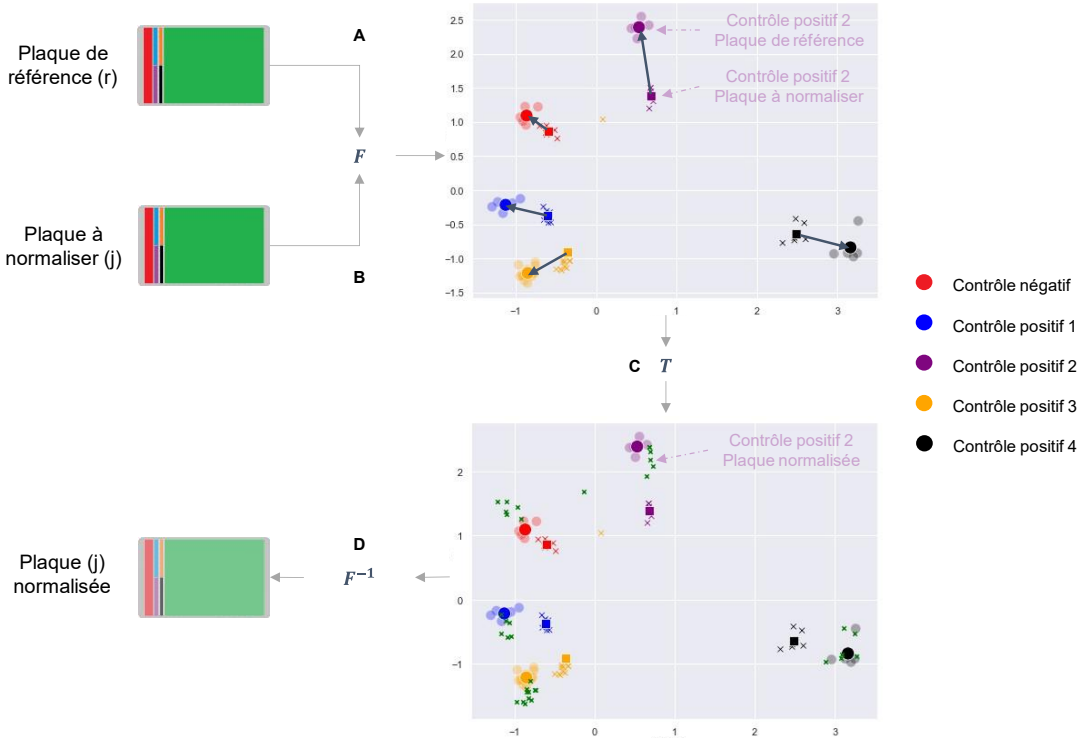

Figure 31. Algorithme de normalisation des données phénotypiques obtenues à partir d'une campagne de criblage à haut contenu.

La première étape de l'algorithme de normalisation consiste à optimiser les paramètres d'une transformation F de l'espace des descripteurs phénotypiques à un espace réduit. Les paramètres de F sont optimisés en utilisant uniquement les données HCS correspondant aux contrôles dispensés dans une plaque de référence (Figure 31A). La transformation F obtenue à partir de la plaque de référence doit minimiser la perte d'information par rapport aux données de l'espace original. Nous appliquons une analyse par composante principale[10] (PCA) dont la variance expliquée par deux axes principaux est supérieure à quatre-vingt-dix pour cent. Puis, la transformation F est appliquée aux contrôles de la plaque de référence (Figure 31A) et de la plaque à normaliser (Figure 31B).

La deuxième étape de l'algorithme consiste à optimiser les paramètres d'une transformation globale paramétrique géométrique T dans l'espace engendré par F. L'optimisation des paramètres de la transformation T considère tous les points de données simultanément contrôles-à-contrôles, de la plaque de référence et de la plaque à

[10] Fonction `decomposition.PCA` de la librairie scikit-learn [121].

normaliser (Figure 31C). La transformation T est de type rigide[11] et les paramètres géométriques de rotation et de translation sont optimisés.

Après optimisation des paramètres de la transformation T, toutes les données HCS de la plaque à normaliser, correspondant aux composés criblés et aux contrôles, sont projetées dans l'espace induit par F (Figure 31B) et sont recalées par T (Figure 31C). Ensuite, l'inverse de la transformation F est appliquée afin de revenir dans l'espace phénotypique de dimension le nombre de descripteurs phénotypiques (Figure 31D). La transformation F est unique et les paramètres des transformations T sont optimisées pour chaque plaque à normaliser.

Nous présentons en (Algorithme 3) le pseudo-code associé à l'algorithme de normalisation.

Entrée : Ensemble des données HCS $\boldsymbol{X}_m^{p,d}$ issues d'une campagne de criblage où p, d, m représentent respectivement les indices d'une plaque, d'un descripteur phénotypique et d'un composé chimique.

Sortie : Ensemble des données HCS $\widetilde{\boldsymbol{X}}_m^{p,d}$ issues d'une campagne de criblage après l'application de la méthode de normalisation.

Algorithme :

Étape 1 : Standardiser les valeurs des descripteurs phénotypiques de chacune des plaques par rapport au DMSO.

Pour $p = 1, \ldots, N_p$ **faire**
  Pour $d = 1, \ldots, N_d$ **faire**
    $\boldsymbol{X}_{:}^{p,d} = (\boldsymbol{X}_{:}^{p,d} - median(\boldsymbol{X}_{DMSO}^{p,d}) \,/\, std(\boldsymbol{X}_{DMSO}^{p,d})$

Étape 2 : Optimiser les paramètres d'une transformation F inversible de l'espace des descripteurs phénotypiques à un espace réduit.

$F_\theta := F(\boldsymbol{X}_{controls}^{plate\_reference,:})$

Étape 3 : Optimiser les paramètres d'une transformation globale T géométrique et paramétrique de type rigide à partir des contrôles de la plaque de référence et des plaques à normaliser. Puis appliquer F et T sur l'ensemble des données HCS des plaques à normaliser avant d'appliquer F^{-1}.

Pour $p = 1, \ldots, N_p - 1$ **faire**
  $T_\vartheta = T(F_\theta(\boldsymbol{X}_{controls}^{plate\_reference,:}), F_\theta(\boldsymbol{X}_{controls}^{p,:}))$
  $\widetilde{\boldsymbol{X}}_{:}^{p,:} = F^{-1}(T_\vartheta(F_\theta(\boldsymbol{X}_{:}^{p,:}))$

  Retourner $\widetilde{\boldsymbol{X}}_{:}^{p,:}$

Algorithme 3. Algorithme de normalisation de données à haut contenu issues des plaques d'une campagne de criblage.

L'idée de la deuxième étape de l'algorithme de normalisation de données HCS, correspondant à un recalage global paramétrique de type rigide, est similaire aux méthodes de recalage d'images obtenues par Imagerie par Résonance Magnétique (IRM) [123].

[11] Fonction `transform.estimate_transform` de la librairie scikit-image [122].

3.3.3 Validation de l'algorithme

Dans un premier temps, nous inspectons visuellement les distributions entre les descripteurs phénotypiques issues des plaques de la campagne de criblage avant et après normalisation (Figure 32).

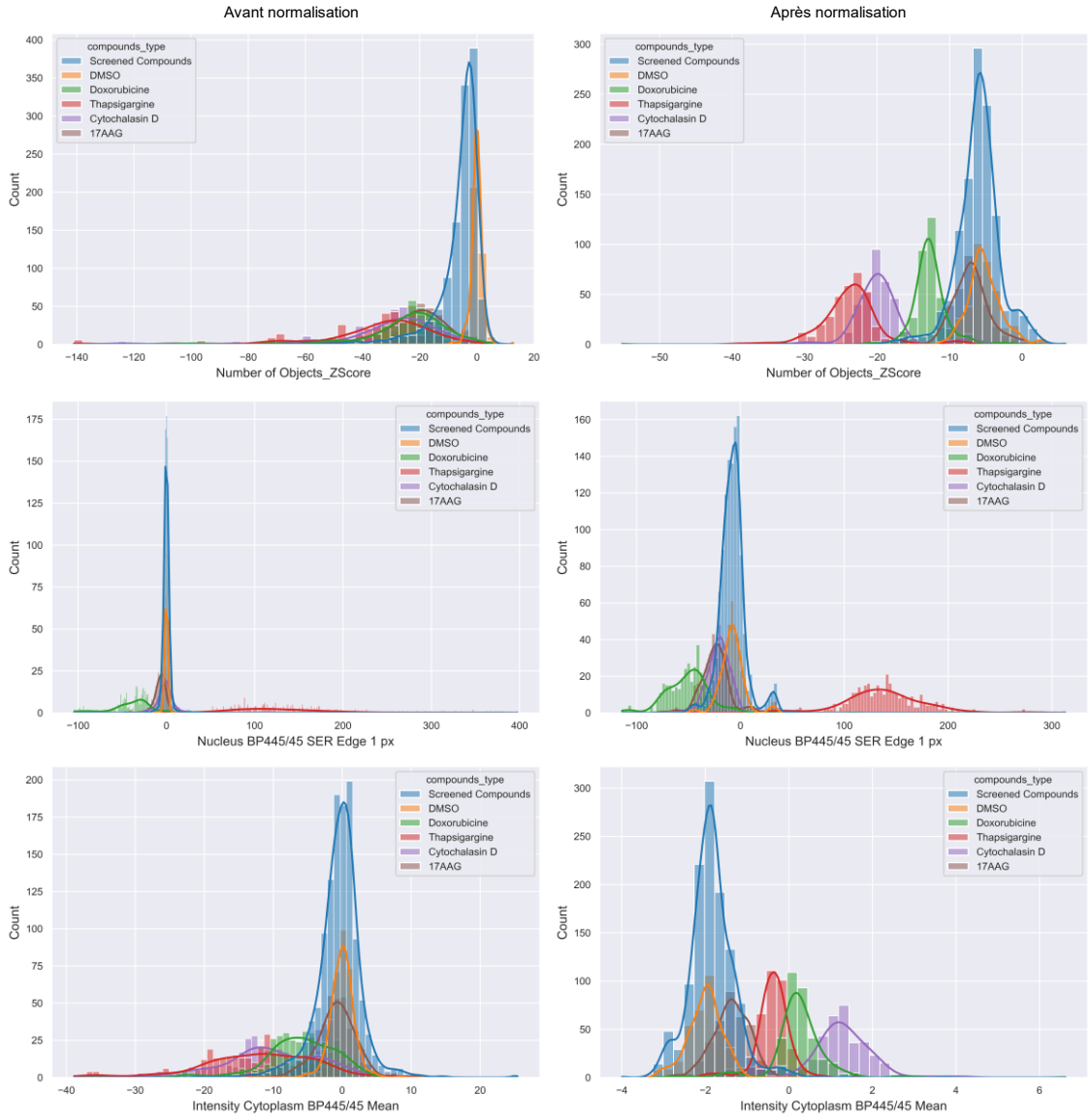

Figure 32. Application de l'algorithme de normalisation rendant les signatures phénotypiques similaires entre les plaques d'une campagne de criblage. Chaque distribution représente les valeurs de descripteurs phénotypiques pour les contrôles et un sous ensemble de composés criblés sélectionnés de façon aléatoire à partir des plaques de la campagne de criblage. À gauche, les distributions de trois descripteurs phénotypiques avant normalisation. À droite, les distributions des mêmes descripteurs phénotypiques après l'application de l'algorithme de normalisation montrant que l'algorithme permet de rendre les distributions des contrôles similaires entre les plaques.

Nous observons que l'algorithme de normalisation permet de rendre les distributions des descripteurs phénotypiques des contrôles plus similaires entre eux, et ce, entre toutes les plaques de la campagne de criblage (Figure 32). De plus, nous observons que les distributions des descripteurs phénotypiques des composés criblés, autres que les

contrôles, ne semblent pas être impactés par la normalisation puisque leurs distributions sont similaires à celles du contrôle négatif (Figure 32). En effet, le nombre de composés identifiés comme actifs dans une campagne de criblage est de l'ordre d'un pour cent et sont répartis aléatoirement dans toutes les plaques. La majorité des composés criblés n'ont pas d'effet et les signatures phénotypiques obtenues sont similaires à celles du contrôle négatif.

Dans un deuxième temps, afin de valider quantitativement l'algorithme de normalisation, nous comparons les valeurs des mesures de similarités cosine pour les composés contrôles entre les plaques de la campagne de criblage. La mesure de similarité de type cosine, présentée en (Formule 5), est utilisée dans le domaine du criblage pour mesurer les similarités phénotypiques induitent par des composés [83]. Cette mesure calcule la similarité de deux vecteurs à dimensions n-descripteurs phénotypiques.

$$cosine\_similarity(X, Y) = \frac{X \cdot Y}{||X|| \cdot ||Y||}$$

Formule 5. Mesure cosine pour calculer des similarités entre des signatures phénotypiques. Les vecteurs X et Y représentent deux signatures phénotypiques obtenues après le criblage de deux composés. Cette mesure de similarité cosine calcule la similarité de deux vecteurs (X, Y) à n dimensions en déterminant le cosinus de leur angle et est obtenue en prenant le produit scalaire des vecteurs X et Y divisé par le produit de leurs normes. Les valeurs de cette mesure sont comprises entre -1 et 1. La valeur de -1 indique des vecteurs opposés, la valeur de 0 indique des vecteurs indépendants et la valeur de 1 indique des vecteurs colinéaires de coefficients positifs. Les valeurs intermédiaires, entre -1 et 1, permettent d'évaluer le degré de similarité des signatures phénotypiques.

Des composés identiques dispensés entre les plaques doivent normalement induirent des signatures phénotypiques similaires. Or ce n'est pas ce que nous observons avant normalisation pour les composés contrôles dispensés dans les soixantes plaques (Figure 33).

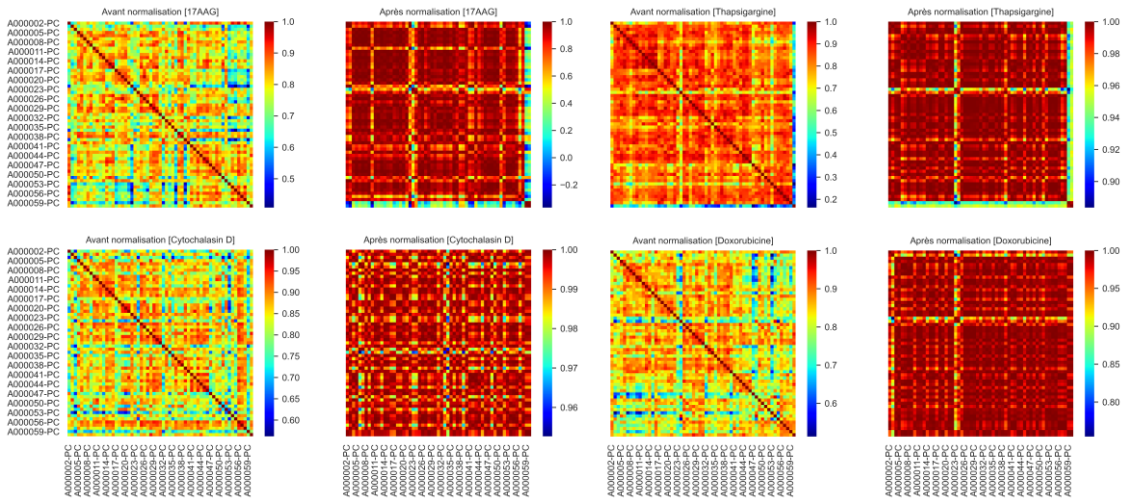

Figure 33. Matrices de similarités phénotypiques des composés contrôles positifs des plaques d'une campagne de criblage à haut contenu avant et après normalisation. L'intersection d'une ligne et d'une colonne représente la valeur de la mesure de similarité cosine calculée entre les signatures phénotypiques des composés contrôles avant et après normalisation. Les lignes et colonnes de chaque matrice représentent une plaque de la campagne de criblage.

Nous observons que l'algorithme de normalisation permet de rendre les signatures phénotypiques des contrôles similaires entre elles, et ce, entre les plaques de la campagne

de criblage (Figure 33). Notons que l'algorithme de normalisation permet d'identifier des plaques problématiques à partir des différents composés contrôles qui semblent mal normalisés. Par exemple, pour les deux dernières plaques (Figure 33), le composé 17-AAG est mal normalisé car les signatures phénotypiques obtenues après normalisation sont moins similaires qu'avant normalisation. Nous pouvons utiliser ce résultat pour identifier si des plaques ont subit des problèmes techniques durant la campagne pour les évaluer plus précisément.

3.4 Intégration des signatures et des similarités phénotypiques dans Pegasus

Nous avons brièvement introduit les concepts de signatures et de similarités phénotypiques dans Pegasus dans le chapitre 2 (Figure 8). Les données associées à ces concepts sont intégrées dans le graphe de connaissances après normalisation.

Les signatures phénotypiques sont représentées sous forme d'entités avec l'étiquette `Phenoprint`. Nous ne modélisons pas dans Pegasus tous les descripteurs phénotypiques : l'entité `Phenoprint` est une entité abstraite correspondant à la signature phénotypique induite par un perturbateur, dans un puit d'une plaque, à une concentration. Nous relions les signatures phénotypiques (`Phenoprint`) avec des relations de type `HAS_SIMILARITIES` qui possèdent des valeurs de propriétés en fonction des mesures de similarités calculées. Ces mesures sont issues de la littérature [83] comme la mesure cosine que nous avons présenté en (Formule 5).

Nous présentons une application envisagée de l'utilisation des similarités phénotypiques en les mettant en perspectives avec d'autres concepts introduits dans Pegasus (Figure 34) afin de repositionner des perturbateurs en utilisant des similarités phénotypiques et chimiques (Requête 9).

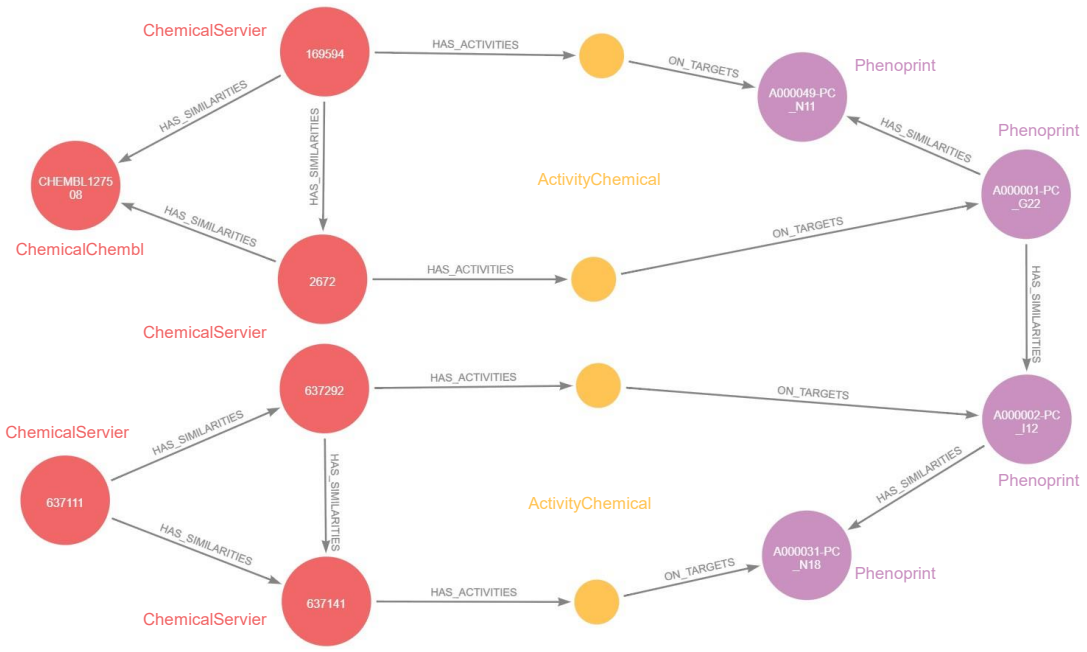

Figure 34. Repositionnement de perturbateurs par l'utilisation de signatures phénotypiques et de similarités phénotypiques et chimiques. Dans cet exemple, des perturbateurs propriétaires (`ChemicalServier`), dont les identifiants sont `2672` et `637292`, induisent (`ActivityChemical`) deux signatures phénotypiques (`Phenoprint`) similaires identifiées par `A000002-PC_G22` et `A000001-PC_I12`.

Ces `Phenoprint` sont reliées par la relation `HAS_SIMILARITIES`. Le perturbateur chimique (`ChemicalServier`) est chimiquement similaire, avec un indice de Tanimoto (`tanimoto_similarity`) supérieur à `0.6`, à deux perturbateurs chimiques identifiés par les identifiants `CHEMBL127508` (`ChemicalChembl`) et `169594` (`ChemicalServier`). Les résultats présentés dans cette figure sont obtenus par l'exécution de la requête (Requête 9) et ne sont pas exhaustifs par souci de clarté.

```
MATCH path=(c1:Chemical)-[rc:HAS_SIMILARITIES]-(c2:ChemicalServier)-
[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-(:Phenoprint)-[rp:HAS_SIMILARITIES]-
(:Phenoprint)-[:ON_TARGETS]-(:Activity)-[:HAS_ACTIVITIES]-(c1)-[:HAS_SIMILARITIES]-
(c2)

WHERE rc.tanimoto_similarity > 0.6 AND rp.cosine_similarity > 0.8

return path;
```

Requête 9. Repositionnement de perturbateurs par des similarités phénotypiques et chimiques. Cette requête CYPHER identifie des perturbateurs chimiques (`Chemical`) qui sont chimiquement similaires (`HAS_SIMILARITIES`) à des perturbateurs propriétaires (`ChemicalServier`) avec une valeur de similarité de Tanimoto (`tanimoto_similarity`) supérieur à `0.6`, et les perturbateurs (`ChemicalServier`) qui induisent (`Activity`) des signatures phénotypiques (`Phenoprint`) similaires, reliés entre elles par la relation `HAS_SIMILARITIES` et dont la valeur de propriété (`cosine_similarity`) est supérieure à `0.8`.

L'objectif à long terme de l'utilisation de signatures phénotypiques au sein de Pegasus est d'avoir la capacité d'identifier différentes classes de perturbateurs pour moduler des cibles thérapeutiques et déconvoluer leurs mécanismes d'actions par les activités annotées présentes dans le graphe. Cependant, l'utilisation de différents protocoles expérimentaux qui ne calculent pas les mêmes descripteurs phénotypiques rendent impossible la comparaison de signatures phénotypiques entre elles. Les liens manquants de la (Figure 35) correspondent aux relations de similarités phénotypiques que nous souhaitons introduire afin de relier des signatures phénotypiques obtenues par différents protocoles expérimentaux et par différentes classes de perturbateurs. L'utilisation des similarités phénotypiques illustre également une façon de relier des entités biologiques, chimiques et phénotypiques de Pegasus entre elles (Figure 35).

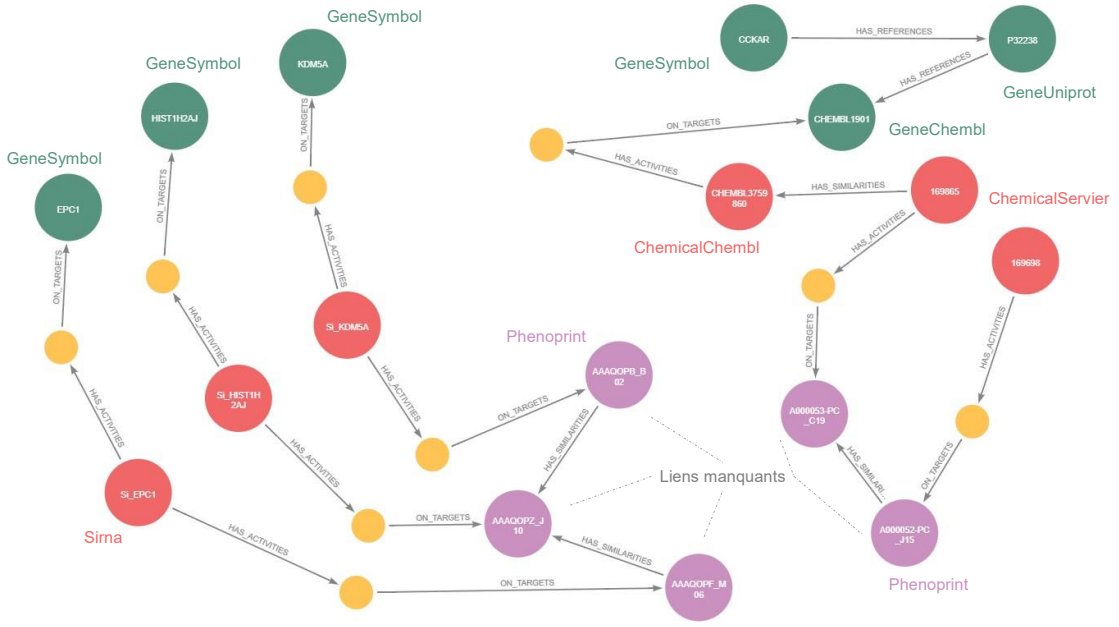

Figure 35. Les signatures phénotypiques comme liens manquant pour relier les concepts biologiques, chimiques et phénotypiques introduits dans le graphe de connaissances Pegasus. Dans cet exemple, à

gauche, des siRNAs (`Sirna`) qui inhibent de façon spécifique des gènes (`GeneSymbol`) et qui induisent des signatures phénotypiques (`Phenoprint`) similaires (`HAS_SIMILARITIES`). À droite, des perturbateurs chimiques (`ChemicalServier`) qui induisent des signatures phénotypiques (`Phenoprint`) similaires. Un des perturbateurs chimique (`ChemicalServier`) est chimiquement similaire à un perturbateur (`ChemicalChembl`) de la littérature qui a une activité sur une cible (`Gene`).

3.5 Conclusion

Nous venons de présenter un algorithme d'identification de composés contrôles positifs et un algorithme de normalisation globale de données HCS. L'algorithme d'identification permet d'identifier de façon automatique, avant la réalisation d'une campagne de criblage, quatre composés contrôles positifs à une concentration chacun qui, pris ensemble, maximisent les réponses des descripteurs phénotypiques calculés. Les composés contrôles positifs sont dispensés dans chaque plaque d'une campagne de criblage en plus d'un contrôle négatif. L'algorithme de normalisation effectue un recalage global des données HCS à partir de composés contrôles utilisés comme point de référence dans un espace phénotypique réduit dans lequel une transformation paramétrique géométrique de type rigide est appliquée. Cet algorithme permet de normaliser des données phénotypiques et d'intégrer des signatures et des similarités phénotypiques informatives dans le graphe de connaissances Pegasus.

Les contributions de ce chapitre feront l'objet de la rédaction de deux articles scientifiques. Le premier article reprendra les deux algorithmes présentés et a fait l'objet d'une présentation lors d'un workshop. Le deuxième article sera rédigé avec les partenaires du consortium JUMP-CP qui accompagnera la mise à disposition des données produites dans le domaine public en 2023.

Productions et communications scientifiques

Amgen, AstraZeneca, Bayer, Biogen, Broad Institute of MIT and Harvard, Eisai, Janssen Pharmaceutica NV, Ksilink, Merck KGaA, Darmstadt, PerkinElmer, Pfizer, Servier, Takeda. Joint Undertaking In Morphological Profiling (JUMP-CP) Consortium. *En préparation, Nature Methods, 2023.*

J. Grignard, S. Adjabi, E. Christ, C. Gauthier, A-L. Ong, S. Lotfi, F. Fages, T. Dorval. An End-To-End Pipeline To Normalize High-Content Screening Data. *En préparation pour le journal SLAS Discovery*.

J. Grignard, F. Fages, T. Dorval. An End-To-End Pipeline To Normalize And Maximize The Phenotypic Information From High-Content Data For Drug Screening Applications. *Short talk - CytoData 2020 - 5th Annual CytoData Society Meeting, October 21-22, 2020.*

Les expériences de criblage phénotypiques à haut contenu permettent de capturer des phénotypes cellulaires sous l'effet de différentes classes de perturbateurs. Cependant, les données phénotypiques ne permettent pas de comprendre, comme les représentations statiques présentées dans le chapitre 2, la dynamique de processus biochimiques modulés par le criblage de perturbateurs.

Afin d'illustrer l'apport de la modélisation mathématique au regard de ce manque, nous présentons dans le chapitre suivant un modèle mathématique mécaniste du cycle de tyrosination des microtubules qui est développé pour expliquer l'inactivité de composés propriétaires criblés en modalité HCS dans des modèles cellulaires prolifératifs et neuronaux alors que ces composés ont été identifiés comme actifs dans un système biochimique en modalité HTS.

Chapitre 4

Modèle mathématique mécaniste du cycle de tyrosination des microtubules

« La maladie d'Alzheimer enlève ce que l'éducation a mis dans la personne et fait remonter le cœur en surface. »

Christian Bobin

Sommaire

4.1 Introduction et motivations

Dans les chapitres précédents, nous avons présenté le graphe de connaissances Pegasus et deux algorithmes pour améliorer la conception et l'analyse d'expériences phénotypiques. Lorsque la dynamique de processus biochimiques est mal comprise car non capturable par des cartes statiques ni par des signatures phénotypiques, le développement de modèles mathématiques permet de prédire et d'expliquer des comportements de systèmes biologiques contre-intuitifs.

Nous présentons dans ce chapitre un modèle mathématique du cycle de tyrosination des microtubules paramétré d'une part, pour les neurones, et d'autre part, pour les cellules prolifératives. Ces modèles montrent que la structure en chaîne du cycle de tyrosination ne permet pas d'augmenter le statut de tyrosination dans les cellules, par le recours à un seul activateur de la chaîne, mais prédisent l'obtention d'un effet par l'inhibition d'une réaction ou l'activation de deux réactions en synergie.

4.1.1 Contexte – Cycle de tyrosination des microtubules dérégulé dans la maladie d'Alzheimer

La maladie d'Alzheimer est un trouble neurodégénératif [124]. Les médicaments disponibles sont insuffisants car ils atténuent certains symptômes mais n'adressent pas les causes de la maladie [125]. Le lien entre les maladies neurodégénératives et la dérégulation des modifications post-traductionnelles du cytosquelette est clair [126]. Les microtubules et leurs régulateurs sont étudiés comme des cibles thérapeutiques [127,128].

Nous nous intéressons aux réactions de tyrosination et de détyrosination, deux modifications post-traductionnelles de la tubuline et des microtubules. Ces réactions forment le cycle de tyrosination des microtubules que nous illustrons en (Figure 36).

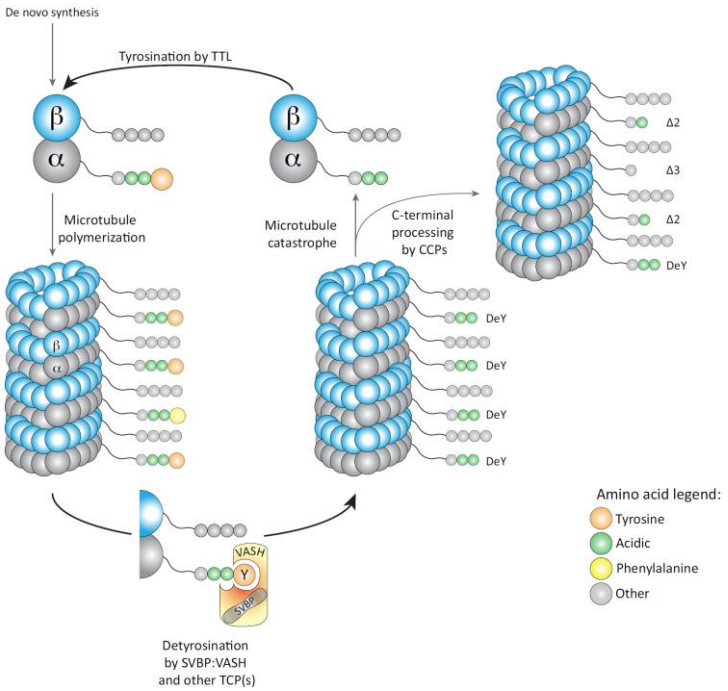

Figure 36. Le cycle de tyrosination des microtubules dérégulé dans les maladies neurodégénératives. Les hétérodimères α/β-tubuline tyrosinés polymérisent pour former un microtubule. Les enzymes Tubuline Carboxy Peptidase (TCP) tels que VASH en complexe avec SVBP détyrosinent les microtubules [129–

132]. Les enzymes CarboxyPeptidase Cytosolique (CCP) peuvent transformer l'extrémité C-term de la tubuline en D2- ou D3-tubuline par d'autres modifications post-traductionnelles. Après la dépolymérisation de la tubuline incorporée dans les microtubules, les hétérodimères α/β-tubuline détyrosinés peuvent être retyrosinés par l'enzyme Tubuline Tyrosine Ligase (TTL) [133,134]. Cette figure provient de [129].

Le cycle de tyrosination des microtubules est initié par la détyrosination des hétérodimères α/β-tubuline tyrosinés et incorporés dans les microtubules. La détyrosination est catalysée par l'enzyme Tubuline Carboxy Peptidase (TCP) comme les vasohibines (VASH1/VASH2) avec la protéine chaperonne Small Vasohibin Binding Protein (SVBP) [129–132]. Après la dépolymérisation des microtubules, l'hétérodimère α/β-tubuline détyrosiné soluble peut être retyrosiné par l'enzyme Tubuline Tyrosine Ligase (TTL) [133,134].

Le statut de tyrosination de la tubuline et des microtubules est essentiel pour la plasticité cérébrale [135], la régénération des axones [136], leur capacité de trouver correctement leur chemin [137], la régulation du transport de complexes protéiques [138], le recrutement de protéines et d'organelles [139] et le câblage correct des neurones [140].

Nous observons une diminution de la quantité des espèces tyrosinées et de l'enzyme TTL dans les cerveaux des personnes atteintes de la maladie d'Alzheimer [141,142]. Les microtubules tyrosinés et les tubulines tyrosinées sont en plus faible concentration. La dérégulation du cycle de tyrosination corrèle avec la perte notable des principales structures neuronales responsables des processus d'apprentissage et de mémorisation [143,144]. La perte du statut de tyrosination est une cause de la maladie d'Alzheimer dans ses stades précoces [145].

4.1.2 Problématique – Comment expliquer les échecs d'expériences de criblage phénotypiques ?

À ce jour, aucun activateur de l'enzyme TTL n'a été publié. Des expériences historiques de criblage HTS ont été réalisées en interne afin d'identifier des activateurs de l'enzyme TTL dans un système biochimique dans lequel seul étaient présents : le composé criblé, l'enzyme TTL et le peptide de tubuline dans sa partie C-terminal. Dans ces expériences, plusieurs composés propriétaires ont augmenté le statut de tyrosination de la tubuline en C-terminale (Figure 37). Ces composés augmentent l'activité de l'enzyme TTL dans un système biochimique.

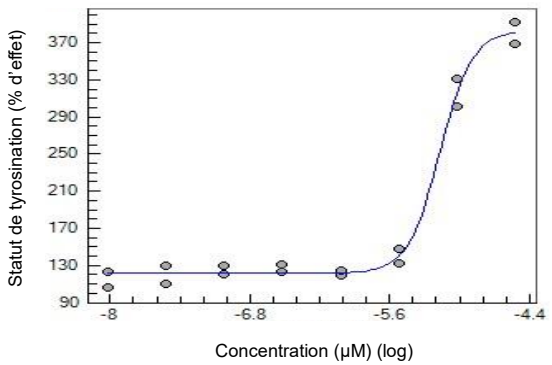

Figure 37. Augmentation du statut de tyrosination en activant l'enzyme TTL par un composé chimique propriétaire dans un système biochimique en modalité HTS. Diagramme en dose-réponse montrant une augmentation du statut de tyrosination (pourcentage d'effet) du peptide de tubuline en C-

terminal à la suite de l'activation de l'enzyme TTL par le criblage d'un composé propriétaire. L'axe des abscisses correspond aux concentrations du composé criblé en échelle logarithmique.

Ces composés actifs ont ensuite été criblés en modalité HCS dans plusieurs modèles cellulaires prolifératifs, notamment dans des fibroblastes embryonnaires de souris (MEF), des cellules épithéliales de rétine humaine immortalisés (hTERT), et dans des neurones obtenues à partir de cellules souches pluripotentes induites (CNS.4U). Ces composés chimiques n'ont pas augmenté le statut de tyrosination dans ces modèles cellulaires (Figure 38).

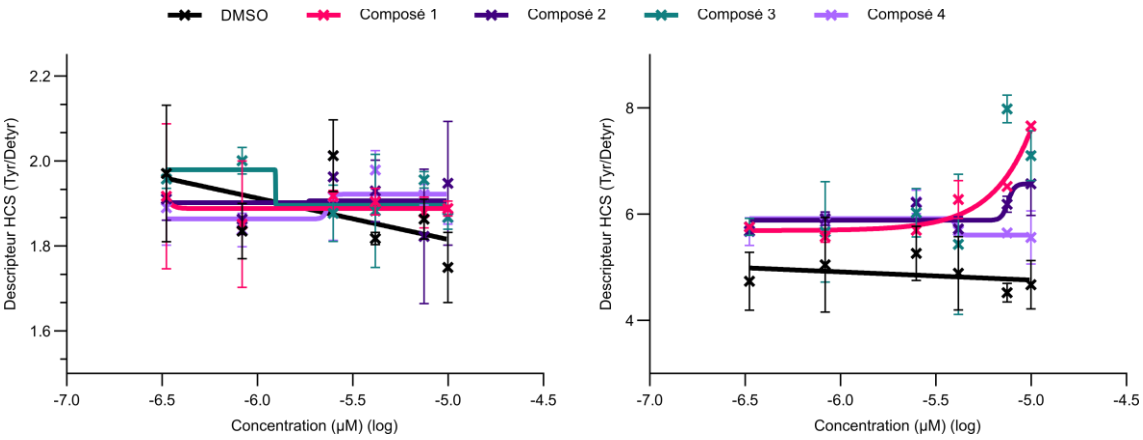

Figure 38. Échecs de l'augmentation du statut de tyrosination dans les cellules prolifératives et neuronales par des composés chimiques propriétaires en modalité HCS. Diagrammes en dose-réponse ne montrant pas d'augmentation significative du statut de tyrosination représenté ici par le descripteur phénotypique `Tyr/Detyr` à la suite du criblage de quatre composés propriétaires. L'axe des abscisses correspond aux concentrations des composés criblés en échelle logarithmique. Le composé DMSO correspond au contrôle négatif de l'expérience. La figure de gauche correspond au criblage de composés propriétaires réalisés dans des cellules neuronales de type CNS.4U. La figure de droite correspond au criblage de ces mêmes composés dans des cellules prolifératives de type MEF. Pour les concentrations de -5 µM (log), les morphologies cellulaires sont altérées, les cytoplasmes sont réduits et les cellules semblent fortement stressées rendant les résultats obtenus à cette concentration non fiables.

Cette différence d'activité des composés entre les expériences biochimiques (Figure 37) et cellulaires (Figure 38) est la principale motivation de notre approche de modélisation mathématique mécaniste. Le processus de modélisation que nous entreprenons a pour objectif de répondre à trois questions :

- expliquer l'inactivité des activateurs de l'enzyme TTL dans les modèles cellulaires prolifératifs et neuronaux
- identifier d'autres cibles thérapeutiques afin d'augmenter le statut de tyrosination
- proposer de nouvelles stratégies de criblage pour repositionner les activateurs de l'enzyme TTL dont le coût de développement a été important

4.1.3 État de l'art – Manque d'un modèle mathématique du cycle de tyrosination des microtubules

À ce jour et à notre connaissance, il n'existe pas de modèle mathématique mécaniste du cycle de tyrosination des microtubules bien que des travaux antérieurs aient été réalisés

sur la modélisation de propriétés des microtubules telles que le trafic, la dynamique des microtubules, l'instabilité dynamique ou encore l'interaction de protéines régulatrices spécifiques avec les microtubules [146–150].

4.1.4 Approche méthodologique – Développement de modèles mathématiques mécanistes

Du fait du manque d'un modèle mathématique du cycle de tyrosination des microtubules, nous entreprenons une approche de modélisation mécaniste. Nous combinons la modélisation mathématique avec des données de la littérature et obtenues par expériences HCS pour développer un modèle mathématique du cycle de tyrosination et de la dynamique des microtubules. La partie de notre modèle concernant la polymérisation des microtubules est basée sur le mode linéaire de la dynamique des microtubules décrit dans [150].

Nous présentons deux modèles mathématiques paramétrés respectivement pour les neurones (CDT_N) et pour les cellules prolifératives (CDT_P).

Ces modèles expliquent à posteriori l'échec inattendu de la confirmation dans ces modèles cellulaires de l'activité de composés précédemment identifiés comme des activateurs de l'enzyme TTL dans un système biochimique. En effet, la tubuline tyrosinée est le produit d'une chaîne de deux réactions dans le cycle : la dépolymérisation du microtubule détyrosiné suivie de la tyrosination de la tubuline détyrosinée. Les niveaux d'espèces tyrosinées à l'équilibre sont donc limités par les deux taux de réaction. L'activation de la réaction de tyrosination seule n'est pas efficace.

En outre, les analyses de sensibilité et les simulations numériques montrent que la diminution du paramètre cinétique de la réaction de détyrosination peut augmenter les concentrations des espèces tyrosinées. En criblant le parthénolide en dose-réponse, un inhibiteur de référence de l'enzyme TCP [127], nous confirmons la prédiction du modèle mathématique dans les cellules prolifératives de type MEF.

Enfin, afin de concevoir de nouvelles expériences de criblage dans les modèles cellulaires neuronaux, le modèle mathématique identifie une combinaison consistant à augmenter, en synergie, la réaction de tyrosination et la dépolymérisation du microtubule détyrosiné, afin d'augmenter du statut de tyrosination.

Dans la suite de ce chapitre, nous présentons notre processus de modélisation mécaniste.

4.2 Modélisation mathématique mécaniste

4.2.1 Structure du modèle mathématique du cycle de tyrosination des microtubules

Dans un premier temps, nous concevons un réseau de réactions chimiques (CRN) axé sur les principales espèces moléculaires et les principaux paramètres régissant le cycle de tyrosination (Figure 39) en se basant sur la vue standard du cycle de tyrosination des microtubules (Figure 36).

La dynamique et les fonctions des microtubules sont modulées par des protéines régulatrices appartenant à des voies de signalisation croisées et sous le contrôle de signaux extracellulaires et intracellulaires [151–154]. Les protéines régulatrices des microtubules peuvent agir sur les microtubules pour favoriser leur polymérisation, leur

dépolymérisation, leur stabilisation et permettent le recrutement et le transport de complexes protéiques [155]. Nous n'introduisons pas les protéines régulatrices des microtubules en raison de leurs nombres élevés et de leurs cinétiques inconnues [155,156].

Nous optons ici pour une approche réductionniste dans laquelle nous considérons, de façon experte, les espèces et les réactions minimales mais essentielles à modéliser pour reproduire des comportements expérimentaux. Le diagramme d'influence du CRN du modèle mathématique du cycle de tyrosination, présenté en (Figure 39A), montre les différentes espèces et réactions modélisées. Le modèle de réaction associé au CRN est développé avec le logiciel Biochemical Abstract Machine (BIOCHAM) et est présenté en (Figure 39B). BIOCHAM est un environnement de modélisation pour la biologie des systèmes et la biologie synthétique [157,158].

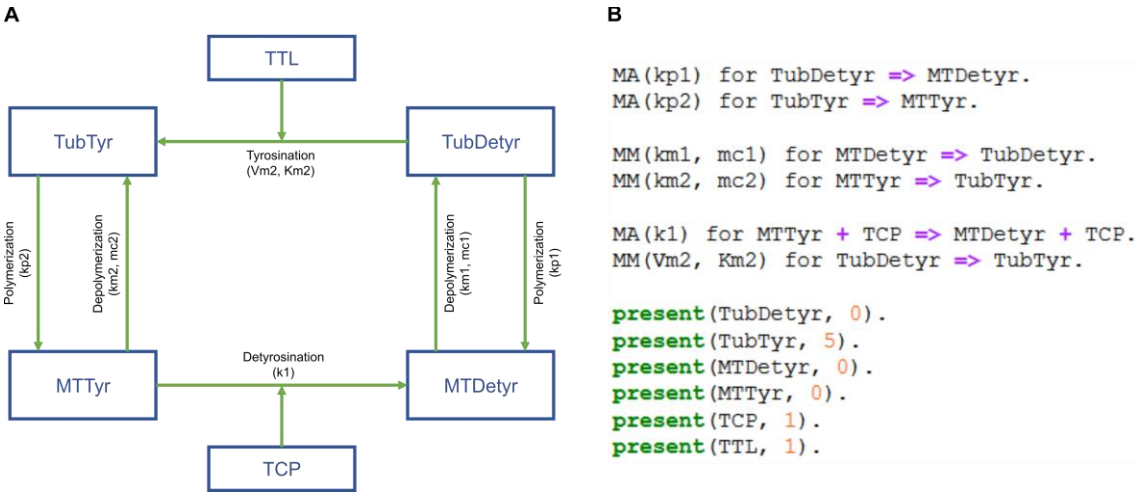

Figure 39. Diagramme d'influence et réseau de réactions chimiques en syntaxe BIOCHAM du modèle mathématique du cycle de tyrosination des microtubules. (A) Diagramme d'influence du cycle de tyrosination des microtubules. **(B)** Réseau de réactions chimiques du cycle de tyrosination des microtubules en syntaxe BIOCHAM dans lequel est indiqué des cinétiques de loi d'action de masse (MA), des cinétiques de Michaelis-Menten (MM) et les concentrations initiales des espèces du modèle.

Le modèle mathématique générique du cycle de tyrosination (Figure 39) est composé de 6 réactions : 2 réactions de polymérisation de la tubuline détyrosinée et tyrosinée, 2 réactions de dépolymérisation du microtubule détyrosiné et tyrosiné et les réactions de détyrosination et de tyrosination catalysées par les enzymes TCP et TTL respectivement.

La réaction de tyrosination est modélisée avec une cinétique Michaelienne basée sur la caractérisation enzymatique de l'enzyme TTL réalisée dans le cerveau bovin [159] (Figure 39B). Lorsque le microtubule est détyrosiné, des protéines déstabilisatrices, recrutées sur le microtubule tyrosiné, sont libérées des microtubules [160,161]. Nous avons choisi d'associer une cinétique de Michaelis-Menten aux réactions de dépolymérisation pour considérer ce phénomène saturant sans inclure de nouvelles espèces dans le modèle mathématique (Figure 39B). Les facteurs limitant des réactions de dépolymérisation du microtubule tyrosiné et détyrosiné sont respectivement le microtubule tyrosiné et détyrosiné (Figure 39B). À notre connaissance, il n'existe pas de caractérisation enzymatique de l'enzyme TCP publiée. La réaction de détyrosination est donnée ici avec une cinétique de loi d'action de masse (Figure 39B).

Un tel ensemble de réactions donné par des fonctions de taux peut être interprété dans BIOCHAM par une chaîne de Markov à temps continu (sémantique stochastique) ou par des équations différentielles ordinaires (EDO) [157]. Pour le modèle du cycle de

tyrosination, impliquant un nombre relativement élevé de molécules, nous considérons la sémantique différentielle de BIOCHAM afin de dériver, à partir du CRN présenté en (Figure 39), le système d'équations différentielles ordinaires présenté en (Équation 1).

$$\begin{cases} \dfrac{dTubTyr}{dt} = \dfrac{km2 \cdot MTTyr}{mc2 + MTTyr} - kp2 \cdot TubTyr + \dfrac{Vm2 \cdot TubDetyr}{Km2 + TubDetyr} \\ \dfrac{dTubDetyr}{dt} = \dfrac{km1 \cdot MTDetyr}{mc1 + MTDetyr} - kp1 \cdot TubDetyr - \dfrac{Vm2 \cdot TubDetyr}{Km2 + TubDetyr} \\ \dfrac{dMTDetyr}{dt} = kp1 \cdot TubDetyr - \dfrac{km1 \cdot MTDetyr}{mc1 + MTDetyr} + k1 \cdot MTTyr \cdot TCP \\ \dfrac{dMTTyr}{dt} = kp2 \cdot TubTyr - \dfrac{km2 \cdot MTTyr}{mc2 + MTTyr} - k1 \cdot MTTyr \cdot TCP \end{cases}$$

Équation 1. Système d'équations différentielles ordinaires du modèle mathématique paramétrable et générique du cycle de tyrosination des microtubules.

Le modèle mathématique paramétrable et générique du cycle de tyrosination donne lieu à deux modèles mathématiques paramétrés : l'un pour les cellules neuronales (CDT_N) et l'autre pour les cellules prolifératives (CDT_P). En effet, nous observons une différence dans la dynamique des microtubules et dans le statut de tyrosination entre ces deux modèles cellulaires. Dans les cellules prolifératives, les microtubules sont globalement dynamiques et tyrosinés, tandis que dans les neurones, les microtubules sont globalement stables et détyrosinés [127,162]. Nous modélisons ces différences en changeant les valeurs des deux paramètres cinétiques (V_{m2}, k_{m1}) (Table 7).

Nous présentons les paramétrisations respectives des modèles mathématiques CDT_N et CDT_P dans les deux sections suivantes.

Taux de réactions	**Paramètres**	**Unités**	**CDT_N**	**CDT_P**
Polymérisation du microtubule détyrosiné	k_{p1}	$\mu M^{-1}.min^{-1}$	0.975	0.975
Polymérisation du microtubule tyrosiné	k_{p2}	$\mu M^{-1}.min^{-1}$	0.975	0.975
Dépolymérisation du microtubule détyrosiné	k_{m1}	min^{-1}	**0.478**	**11.74**
	mc_1	µM	2.75	2.75
Dépolymérisation du microtubule tyrosiné	k_{m2}	min^{-1}	4.78	4.78
	mc_2	µM	0.48	0.48
Détyrosination	k_1	$\mu M^{-2}.min^{-1}$	1	1
Tyrosination	V_{m2}	min^{-1}	**0.2**	**7.70**
	K_{m2}	µM	1.9	1.9

Table 7. Valeurs des paramètres cinétiques des modèles mathématiques paramétrés pour les neurones (CDT_N) et pour les cellules prolifératives (CDT_P).

Les concentrations initiales des espèces des modèles CDT_N et CDT_P sont fixées par rapport aux données de la littérature (Table 8). Dans la Figure 2E de [131], les auteurs observent l'évolution temporelle des espèces du cycle de tyrosination pour quatre points de temps (0, 2, 5, 10 minutes). Les espèces tyrosinées à savoir la tubuline et les microtubules tyrosinés sont seules présentes au début de leur expérience. Dans notre modèle, la concentration initiale de la tubuline tyrosinée est fixée à 5 µM, conformément aux concentrations de tubuline indiquées dans [131] et de façon cohérente avec d'autres modèles cellulaires [163,164]. Les concentrations initiales des autres espèces du cycle

sont fixées à 0 µM car nous souhaitons observer la polymérisation du microtubule tyrosiné. De plus, nous contrôlons que les concentrations initiales ne changent pas l'état d'équilibre du système. En effet, nous vérifions avec BIOCHAM les conditions nécessaires à l'existence de plusieurs états stables non dégénérés dans la dynamique différentielle du modèle de réaction en contrôlant l'existence de circuits positifs. Ce contrôle est réalisé dans le graphe d'influence étiqueté associé à la structure du réseau de réaction [165]. La commande BIOCHAM `check_multistability` appliquée à notre modèle de réaction renvoie l'absence de circuits positifs. Ce résultat prouve l'absence de multiples états stables non dégénérés dans le système EDO et assure l'unicité de l'état stable atteint à partir de différentes conditions initiales.

Concentrations initiales	Unités	CDT_N	CDT_P
Tubuline détyrosinée	µM	0	0
Tubuline tyrosinée	µM	5	5
Microtubule détyrosiné	µM	0	0
Microtubule tyrosiné	µM	0	0
Tubulin Tyrosine Ligase (TTL)	µM	1	1
Tubulin CarboxyPeptidase (TCP)	µM	1	1

Table 8. Concentrations initiales des espèces moléculaires des modèles mathématiques paramétrés pour les neurones (CDT_N) et pour les cellules prolifératives (CDT_P).

4.2.2 Paramétrisation du modèle CDT_N avec des valeurs cinétiques issues de la littérature

Le modèle CDT_N est paramétré avec des données issues de la littérature, d'hypothèses et d'une procédure de recherche de paramètres. Les paramètres cinétiques de polymérisation de la tubuline détyrosinée et tyrosinée (k_{p1}, k_{p2}), le paramètre cinétique de la dépolymérisation du microtubule tyrosiné (k_{m2}) et les paramètres cinétiques de la réaction de tyrosination (V_{m2}, K_{m2}) proviennent de la littérature [147,159].

Nous supposons ici que le paramètre cinétique de la dépolymérisation du microtubule détyrosiné (k_{m1}) est dix fois plus petite que celle du microtubule tyrosiné (k_{m2}). Cette hypothèse provient d'abord de références indiquant que les microtubules détyrosinés sont plus stables que les microtubules tyrosinés [127,162,166]. De plus, il est établi que la demi-vie des microtubules tyrosinés est de l'ordre de quelques minutes alors que la demi-vie des microtubules détyrosinés est de l'ordre de plusieurs heures [167–169]. Cette différence d'un ordre de grandeur entre les demi-vies des microtubules détyrosinés et tyrosinés se reflète dans notre modèle par le choix d'une constante de vitesse de dépolymérisation pour le microtubule détyrosiné (k_{m1}) dix fois plus petite que pour le microtubule tyrosiné (k_{m2}).

Les paramètres cinétiques mc_1et mc_2 correspondent aux constantes de Michaelis-Menten pour les réactions de dépolymérisation du microtubule détyrosiné et tyrosiné, et en l'absence de données expérimentales directes, leurs valeurs sont inférées à partir d'une procédure de recherche de paramètres pour obtenir la demi-vie connue des espèces tyrosinées de l'ordre de 5 minutes [167–169]. Cette contrainte est exprimée dans

BIOCHAM par une formule logique linéaire temporelle du premier ordre avec contraintes linéaires sur les réels (FO-LTL(Rlin)) [158] que nous présentons en (Formule 6).

```
F(Time == 5 /\ Tyr = factor1)
```

Formule 6. Formule logique linéaire temporelle du premier ordre avec contraintes linéaires sur les réels (FO-LTL(Rlin)) spécifiant un comportement expérimental à reproduire obtenu à partir de la littérature.

La formule FO-LTL(Rlin) (Formule 6), utilisée dans la procédure de recherche de paramètres cinétiques présentée en (Fonction 1), stipule que finalement (`F`), à un moment autour de `5` unités de temps, la concentration des espèces tyrosinées (`Tyr`) a une certaine valeur assignée à la variable `factor1`. Cette formule FO-LTL(Rlin) est évaluée sur la trace de simulation du modèle CDT_N dont les valeurs de (mc_1, mc_2) sont fixées à 1 par défaut et retourne un degré de satisfaction continu dans l'intervalle [0,1]. Le degré de satisfaction indique à quel point la formule est loin de la satisfaction entre faux (0) et vrai (1). Le degré de satisfaction de la spécification formelle est utilisé comme fonction objective pour guider la recherche pendant l'optimisation des paramètres (Fonction 1). Le logiciel BIOCHAM utilise la stratégie d'évolution par adaptation de la matrice de covariance (CMA-ES), un algorithme d'optimisation non linéaire continu [170], pour inférer les ensembles de paramètres satisfaisant des contraintes FO-LTL(Rlin) [157,158].

```
search_parameters(
      F(Time == 5 /\ Tyr = factor1),
      [0 <= mc1 <= 10, 0 <= mc2 <= 10],
      [factor1 -> 2.5]
).
```

Fonction 1. Fonction de recherche de paramètres cinétiques BIOCHAM pour inférer les valeurs des constantes de Michaelis-Menten (mc_1, mc_2) des réactions de dépolymérisation des microtubules détyrosinés et tyrosinés. Dans notre étude, nous avons optimisé les valeurs des paramètres cinétiques (mc_1, mc_2) en utilisant ce schéma de commande BIOCHAM. Les valeurs inférées sont respectivement (2.75, 0.48) et le meilleur degré de satisfaction obtenu est de 1 indiquant que la contrainte est pleinement satisfaite.

Nous n'avons pas à notre connaissance de valeur pour le paramètre cinétique de la réaction de détyrosination (k_1). Nous fixons sa valeur à 1 $\mu M^{-2}.min^{-1}$ sachant que l'enzyme TCP devrait agir lentement sur les microtubules [171].

Avec ces valeurs de paramètres cinétiques, l'intégration numérique du système EDO (Équation 1) associée au modèle mathématique CDT_N paramétré pour les neurones permet d'observer l'évolution temporelle des espèces moléculaires du cycle de tyrosination (Figure 40). La simulation numérique du modèle CDT_N est cohérente avec l'évolution temporelle des espèces moléculaires obtenue expérimentalement dans les neurones (Figure 40). Le modèle CDT_N capture la dynamique des espèces du cycle de tyrosination des microtubules dans les cellules neuronales.

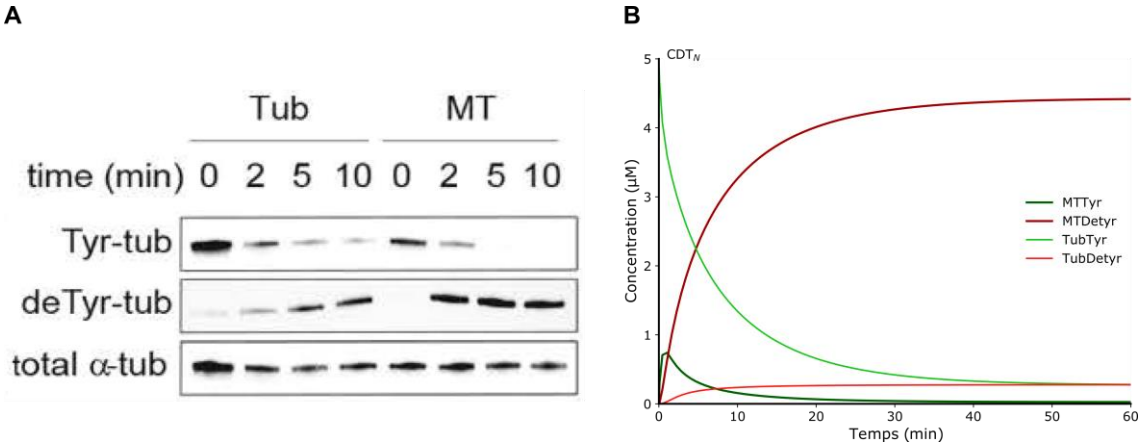

Figure 40. Comparaison des évolutions temporelles expérimentales et numériques montrant que le modèle mathématique paramétré pour les neurones (CDT_N) capture la dynamique des espèces du cycle de tyrosination des microtubules. (A) Évolution temporelle des espèces du cycle de tyrosination dans les neurones obtenue expérimentalement par immunoblot. Cette image a été extraite de la Figure 2E de [131]. **(B)** Simulation numérique de l'évolution temporelle des espèces du cycle de tyrosination dans le modèle CDT_N. En comparant **(A)** et **(B)**, nous observons que le microtubule détyrosiné est la principale espèce moléculaire à l'état d'équilibre tandis que la tubuline détyrosinée augmente légèrement au cours du temps et les demi-vies des microtubules tyrosinés et détyrosinés sont de l'ordre de quelques minutes et heures respectivement [167–169].

4.2.3 Paramétrisation du modèle CDT_P en ajustant le modèle CDT_N à des données expérimentales d'imagerie à haut contenu

Le modèle CDT_P est paramétré avec des données issues d'expériences HCS internes et d'une procédure de recherche de paramètres qui modifie deux valeurs de paramètres cinétiques (V_{m2}, k_{m1}) à partir du modèle CDT_N.

Nous avons réalisé des expériences HCS pour quantifier le statut de tyrosination dans les cellules prolifératives de type hTERT et MEF. Nous observons que le rapport de fluorescence entre les espèces tyrosinées et les espèces détyrosinées est de l'ordre de dix dans ces modèles cellulaires (Figure 41). Le statut de tyrosination à l'état d'équilibre entre les cellules prolifératives est donc différent de celui observé dans les cellules neuronales (Figure 40). En effet, dans le modèle CDT_N, à l'état d'équilibre, les espèces détyrosinées sont majoritaires (Figure 40) alors que dans les cellules prolifératives les espèces tyrosinées sont majoritaires (Figure 41).

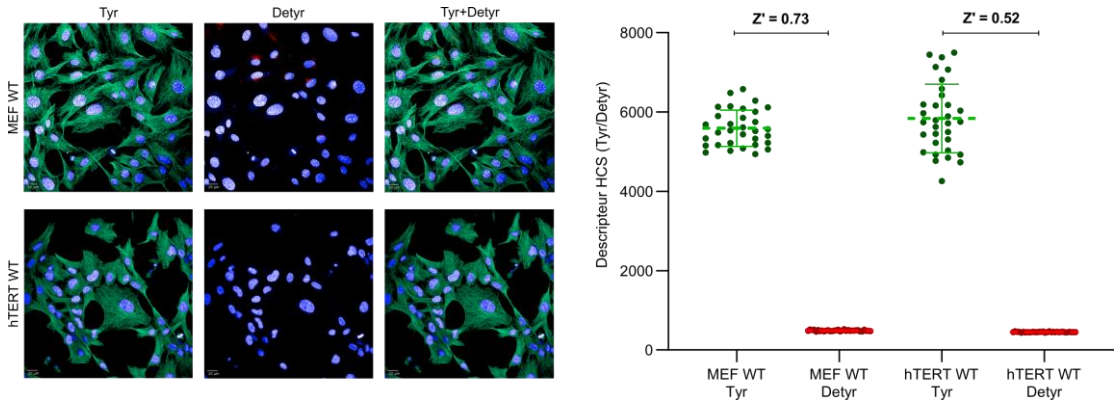

Figure 41. Quantification du statut de tyrosination en modalité HCS dans les cellules prolifératives montrant que les espèces du cycle de tyrosination sont majoritairement tyrosinées. À gauche, des

images représentatives de l'immunomarquage des espèces tyrosinées (Tyr) en vert, des espèces détyrosinées (Detyr) en rouge, dans des cellules prolifératives de type MEF et hTERT. Barres d'échelle : 20 µm. À droite, quantification du statut de tyrosination en modalité HCS. Les espèces du cycle de tyrosination des microtubules sont principalement tyrosinée (facteur Z' > 0.5). En HCS, le facteur Z' est classiquement utilisé pour évaluer le degré de séparation d'une variable entre deux conditions [172] et la formule est présentée en (Formule 1) dans le chapitre précédent.

Afin de paramétrer le modèle CDT_P, notre objectif est donc de trouver les changements minimaux à partir du modèle CDT_N qui font que le système se stabilise avec un ratio de dix et ce au plus tard à cinq minutes. En effet, nous observons un ratio entre les espèces tyrosinées et détyrosinées de dix dans les cellules prolifératives (Figure 41). De plus, nous savons que la demi-vie des microtubules tyrosinés est d'environ cinq minutes [167–169].

Nous formalisons ces observations expérimentales sous forme de contraintes en logique temporelle quantitative pour rechercher les valeurs des paramètres qui permettent de reproduire le comportement observé dans les cellules prolifératives par optimisation continue. Ces contraintes sont exprimées dans BIOCHAM par la formule FO-LTL(Rlin) que nous présentons en (Formule 7).

```
F(Temps == 5 /\ Tyr = ratio1 * Detyr /\ F(Temps == 20 /\ Tyr = ratio2 * Detyr))
```

Formule 7. Formule logique linéaire temporelle du premier ordre avec contraintes linéaires sur les réels (FO-LTL(Rlin)) spécifiant un comportement expérimental à reproduire obtenu par des données d'imagerie à haut contenu.

La formule FO-LTL(Rlin) (Formule 7), utilisée dans la procédure de recherche de paramètres cinétiques présentée en (Fonction 2), stipule que finalement (`F`), à un moment autour de `5` unités de temps, le rapport `Tyr` sur `Detyr` a une certaine valeur assignée à la variable `ratio1`, et plus tard (`F`) à un moment autour de `20` unité de temps, le rapport `Tyr` sur `Detyr` a la valeur `ratio2`. Cette formule FO-LTL(Rlin) est évaluée sur la trace de simulation du modèle CDT_N présentée en (Figure 40B) et retourne un degré de satisfaction continu dans l'intervalle [0,1].

```
search_parameters(
      F(Time == 5 /\ Tyr = factor1 * Detyr /\ F(Time == 20 /\ Tyr = factor2 * Detyr)),
      [b11 <= p1 <= b12, b21 <= p2 <= b22], # [b11 <= p <= b12]
      [factor1 -> 10, factor2 -> 10]
).
```

Fonction 2. Fonction de recherche de paramètres cinétiques BIOCHAM pour satisfaire la contrainte FO-LTL(Rlin) afin d'augmenter le statut de tyrosination. Dans notre étude, nous avons optimisé les valeurs des paramètres cinétiques en utilisant ce schéma de commande BIOCHAM. Ici, deux paramètres cinétiques sont optimisés (p_1, p_2) entre des valeurs spécifiées par les bornes (b_{11}, b_{12}, b_{21}, b_{22}) Pour optimiser un seul paramètre p, il suffit de remplacer la ligne par ce qui est présenté en commentaire (#).

Les paramètres cinétiques de polymérisation (k_{p1}, k_{p2}) sont supposés être les mêmes entre les modèles cellulaires étudiés et ne sont pas optimisées. En effet, le statut de tyrosination n'a aucun effet sur la capacité de polymérisation de la tubuline et les vitesses de polymérisation de la tubuline tyrosinée et détyrosinée sont similaires [173,174].

Dans un premier temps, nous réalisons des cycles d'optimisation de notre contrainte FO-LTL(Rlin) en recherchant un seul paramètre cinétique à la fois en utilisant la fonction de recherche de paramètres cinétiques présentée en (Fonction 2). Aucun des cycles d'optimisation n'a pu reproduire le comportement expérimental observé lors de la modification d'un paramètre cinétique unique (Figure 42A). Il s'agit d'une indication

forte, bien qu'il ne s'agisse pas d'une preuve formelle, que le comportement observé ne peut pas être obtenu en modifiant un seul paramètre cinétique.

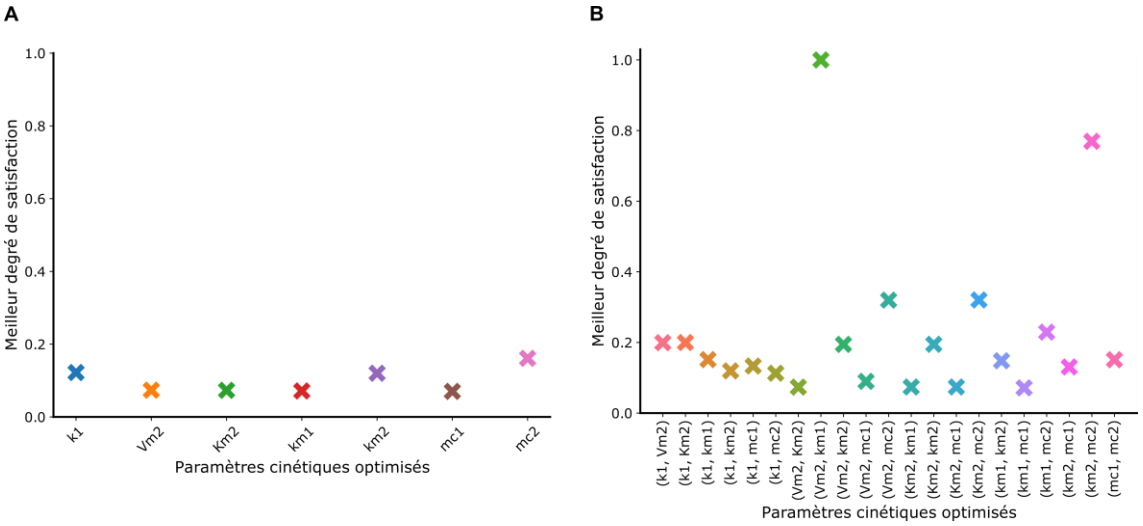

Figure 42. Résultats de la procédure de recherches de paramètres cinétiques pour paramétrer le modèle CDT$_P$ montrant que seul le couple (V_{m2}, k_{m1}) permet de satisfaire la formule logique. (A) Meilleur degré de satisfaction obtenu par la procédure de recherche de paramètres en faisant varier un seul paramètre cinétique montrant l'incapacité de reproduire le comportement observé. La commande BIOCHAM utilisée est présentée en (Fonction 2) avec un seul paramètre cinétique p à optimiser. **(B)** Meilleur degré de satisfaction obtenu par la procédure de recherche de paramètres en faisant varier des couples de deux paramètres cinétiques simultanément montrant une satisfaction parfaite de la spécification avec un seul couple de paramètres : (V_{m2}, k_{m1}). La commande BIOCHAM permettant d'obtenir ces résultats est présentée en (Fonction 2).

Dans un deuxième temps, nous effectuons des optimisations sur toutes les paires de paramètres cinétiques (Figure 42B). Il est remarquable que la procédure d'optimisation ait réussi à satisfaire la spécification temporelle pour une seule paire de paramètres cinétiques (V_{m2}, k_{m1}). La modulation de la réaction de tyrosination (V_{m2}) en synergie avec la modulation de la réaction de dépolymérisation du microtubule détyrosiné (k_{m1}) semble donc suffisante pour reproduire l'augmentation du statut de tyrosination observée expérimentalement dans les cellules prolifératives, et ce, avant cinq minutes. Bien qu'il ne soit pas particulièrement intuitif et donc particulièrement instructif, ce résultat peut être comparé à certaines connaissances connues de la littérature. En effet, dans les cellules prolifératives, l'enzyme TTL agit rapidement sur la tubuline, ce qui se traduit ici par une modification de V_{m2}, et les microtubules sont dynamiques, ce qui se traduit ici par une modification de k_{m1} [161].

Il est intéressant de noter que la paire de paramètres cinétiques (k_{m2}, mc_2), bien que ne satisfaisant pas la spécification temporelle, a pu néanmoins atteindre un degré de satisfaction de 0.76 (Figure 42B). Dans ce jeu de paramètres optimisés, le système semble toujours plus lent à se stabiliser (vingt minutes) et les valeurs inférées ne semblent pas biologiquement réalistes puisque le paramètre mc_2 de (k_{m2}, mc_2) a une valeur faible en 10^{-8} ajoutée à une concentration de MTTyr jusqu'à 10^{-2} rendant la réaction indépendante du réactant.

Les changements sur le couple de paramètres (V_{m2}, k_{m1}) peuvent satisfaire pleinement la spécification (Formule 7), cependant avec de nombreuses solutions. Afin de visualiser l'ensemble de ces solutions, c'est-à-dire, le paysage du degré de satisfaction de notre spécification, nous balayons les valeurs des paramètres du couple (V_{m2}, k_{m1})

dans des intervalles raisonnables (Figure 43A). Le paysage indique que la formule est satisfaite pour un ensemble infini de solutions pour ce couple de paramètres (Figure 43A).

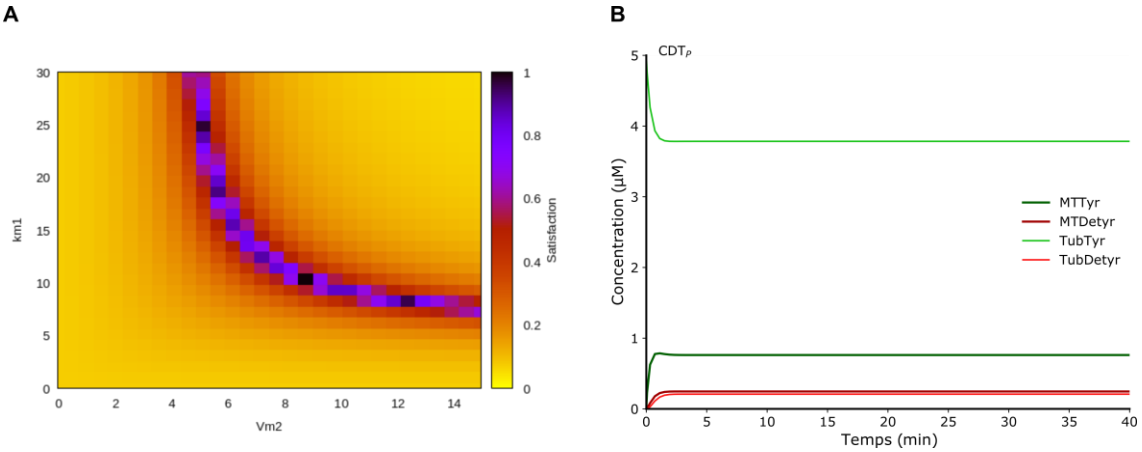

Figure 43. Paramétrisation du modèle mathématique pour les cellules prolifératives (CDTP) par modification minimale des deux paramètres cinétiques (V_{m2}, k_{m1}). (A) Le paysage montre une infinité de solution pour le couple (V_{m2}, k_{m1}) qui permet de satisfaire la contrainte. La commande BIOCHAM pour obtenir ce paysage est `scan_parameters(F(Time == 5 /\ Tyr = factor1 * Detyr /\ F(Time == 20 /\ Tyr = factor2 * Detyr)), (0 <= Vm2 <= 15), (0 <= km1 <= 30), [factor1 -> 10, factor2 -> 10], resolution: 30)`. Afin de paramétrer le modèle CDTP, nous prenons les valeurs correspondant à la distance minimale par rapport aux valeurs du modèle CDTN en exécutant la fonction présentée en (Fonction 3). **(B)** Simulation numérique du modèle CDTP après paramétrisation de ce dernier en ajustant le modèle CDTN à nos données expérimentales couplé à la procédure de recherche de paramètres cinétiques. La simulation du modèle CDTP montre que les espèces tyrosinées se stabilisent avant cinq minutes et à une concentration correspondant à un facteur dix par rapport aux espèces détyrosinées, et ce, après une modification minime des paramètres cinétiques (V_{m2}, k_{m1}) par rapport au modèle CDTN.

Nous avons effectué une nouvelle optimisation des paramètres (V_{m2}, k_{m1}) afin d'inférer les valeurs des paramètres cinétiques minimisant la différence avec les valeurs originales présentes dans le modèle CDTN avec la fonction de recherche présentée en (Fonction 3).

```
search_parameters(
      F(Time == 5 /\ Vm2 = VarVm2 /\ km1 = Varkm1 /\ Tyr = factor1 * Detyr /\ F(Time
== 20 /\ Tyr = factor2 * Detyr)),
      [0 <= Vm2 <= 15, 0 <= km1 <= 30],
      [VarVm2 -> 0.2, Varkm1 -> 0.478, factor1 -> 10, factor2 -> 10]
).
```

Fonction 3. Fonction de recherche de paramètres cinétiques BIOCHAM pour satisfaire la contrainte FO-LTL(Rlin) pour deux paramètres cinétiques avec une modification minimale de leurs valeurs.

Ce nouveau cycle d'optimisation infère les nouvelles valeurs de V_{m2} et de k_{m1} et qui sont choisies pour paramétrer le modèle CDTP résultant en une augmentation d'un facteur 24.56 pour k_{m1} et d'un facteur 38.54 pour V_{m2} (Table 8). Cette modification minime par rapport au modèle CDTN suffit à reproduire par simulation, l'observation expérimentale selon laquelle les espèces moléculaires tyrosinées se stabilisent autour de cinq minutes et à une concentration supérieure d'un facteur dix à celle des espèces détyrosinées dans les cellules prolifératives.

4.2.4 Explication mécaniste des échecs de campagnes de criblage phénotypiques

Comme indiqué dans la section introductive de ce chapitre, il n'existe aucun activateur de l'enzyme TTL publié dans la littérature. Néanmoins, des expériences HTS dans un système biochimique nous ont permis d'identifier des composés chimiques propriétaires qui augmentent le statut de tyrosination en activant l'enzyme TTL (Figure 37). Ces composés actifs ont été criblés dans des modèles cellulaires prolifératifs (MEF) et neuronaux (CNS.4U) mais n'ont pas augmenté le statut de tyrosination (Figure 38).

Nous utilisons les modèles mathématiques paramétrés CDT_P et CDT_N pour comprendre l'inactivité des activateurs de l'enzyme TTL dans les cellules prolifératives et neuronales. Tout d'abord, nous réalisons une analyse de sensibilité pour déterminer la sensibilité du statut de tyrosination, représenté dans nos modèles par le rapport des espèces tyrosinées sur les espèces détyrosinées (TyrDetyr), pour le paramètre cinétique de la réaction de tyrosination (V_{m2}) en utilisant BIOCHAM. Les indices de sensibilité des modèles CDT_P et CDT_N indiquent une tolérance de cinq cents pour cent pour le paramètre (V_{m2}) avant que le statut de tyrosination ne s'écarte de sa valeur d'équilibre de quatre-vingts et quinze pour cent, respectivement (Figure 44).

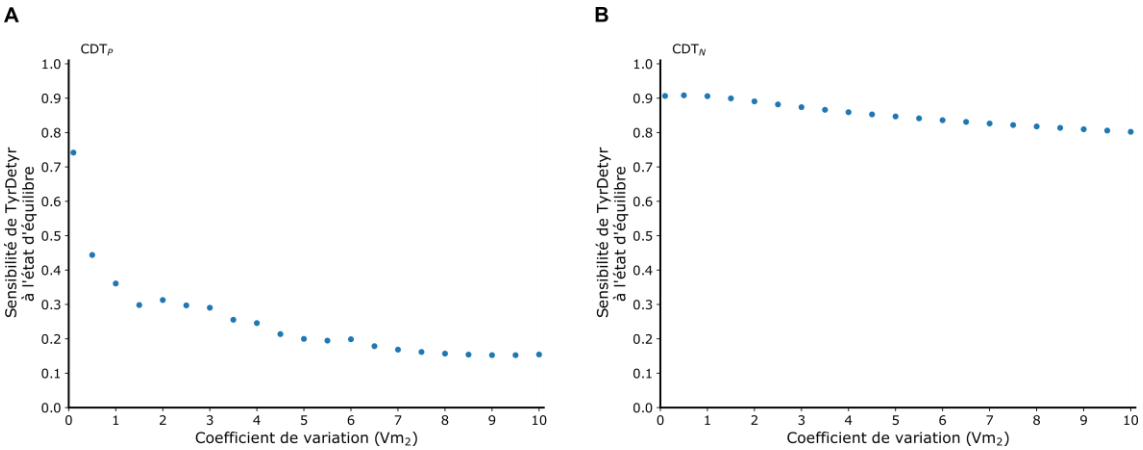

Figure 44. Analyses de sensibilités de la valeur de TyrDetyr à l'état d'équilibre obtenue pour différents coefficients de variation du paramètre cinétique V_{m2} dans les modèles mathématiques CDT_P et CDT_N. (A) Analyse de sensibilité dans le modèle CDT_P indiquant une tolérance de cinq cents pour cent pour le paramètre V_{m2} avant que `TyrDetyr` ne s'écarte de son état d'équilibre de quatre-vingts pour cent. La commande BIOCHAM utilisée est : `sensitivity(F(G(TyrDetyr = x)), [Vm2], [x -> 10], robustness_coeff_var : c)`, où `c` est la valeur du coefficient de robustesse. **(B)** Analyse de sensibilité similaire dans le modèle CDT_N indiquant une tolérance de cinq cents pour cent pour le paramètre V_{m2} avant que `TyrDetyr` ne s'écarte de son état d'équilibre de quinze pour cent. La commande BIOCHAM utilisée est : `sensitivity(F(G(TyrDetyr = x)), [Vm2], [x -> 0.065386], robustness_coeff_var : c)` où `c` est la valeur du coefficient de robustesse.

Les indices de sensibilité sont calculés en estimant le degré de satisfaction moyen, c'est-à-dire la robustesse [175] d'une spécification temporelle en faisant varier un ou des paramètres indépendamment [158].

Les analyses de sensibilité indiquent que l'augmentation du statut de tyrosination nécessiterait une modulation élevée du paramètre cinétique (V_{m2}) dans les cellules neuronales, ce qui pourrait ne pas être réalisable sur le plan pharmacologique, et une modulation modérée dans les cellules prolifératives. Cependant, nous n'avons pas observé expérimentalement une augmentation significative du statut de tyrosination avec

les activateurs de l'enzyme TTL. Par conséquent, nous avons simulé l'ajout d'un activateur de l'enzyme TTL qui pourrait augmenter l'activité de tyrosination d'un facteur dix dans les modèles CDT_P et CDT_N en augmentant V_{m2} après que les systèmes se sont stabilisés (Figure 45). Dans les deux modèles, les espèces tyrosinées ont légèrement augmentées sans dépasser les espèces détyrosinées (Figure 45). L'augmentation du paramètre cinétique de la réaction de tyrosination n'est donc pas suffisante pour déclencher une augmentation significative du statut de tyrosination dans nos modèles mathématiques.

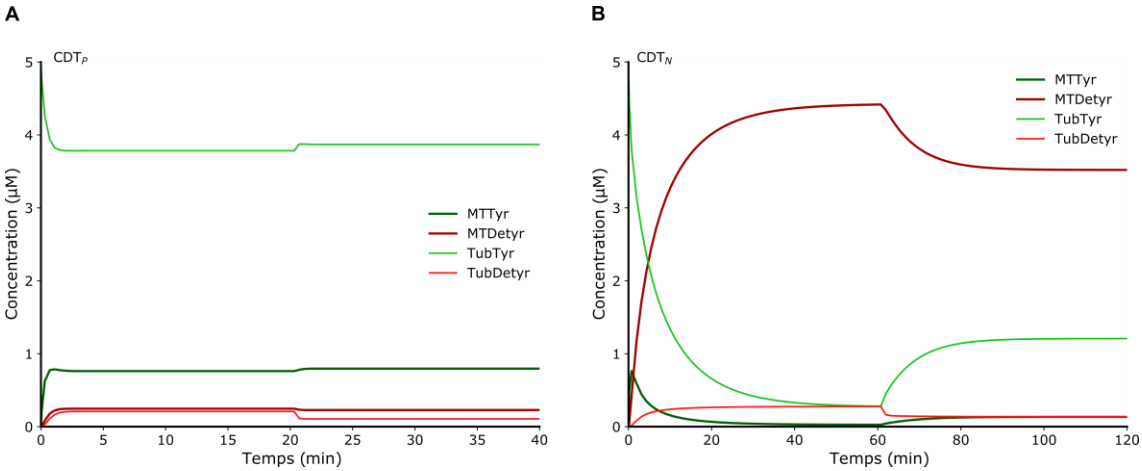

Figure 45. Prédiction de l'augmentation de l'activité de l'enzyme TTL dans les modèles mathématiques prolifératifs et neuronaux montrant l'incapacité d'augmenter le statut de tyrosination significativement. (A) Simulation numérique perturbée dans le modèle CDT_P. La constante de vitesse de tyrosination V_{m2} est augmentée à vingt unités de temps (min) par un facteur dix. La simulation numérique montre que le statut de tyrosination n'augmente pas. **(B)** Simulation numérique perturbée dans le modèle CDT_N. La constante de vitesse de tyrosination V_{m2} est augmentée au temps soixante (min) par un facteur dix. La simulation numérique montre que la concentration des espèces tyrosinées augmente légèrement mais n'est pas supérieure aux espèces détyrosinées à l'état d'équilibre.

Ces observations s'expliquent par le fait que la tubuline tyrosinée est le produit d'une chaîne de deux réactions dans le cycle de tyrosination : la dépolymérisation du microtubule détyrosiné suivie de la tyrosination de la tubuline (Figure 39). Le niveau des espèces tyrosinées à l'équilibre est donc limité par les deux vitesses de réaction. L'activation de la réaction de tyrosination seule n'est pas efficace. Dans le modèle CDT_P, le niveau de tubuline détyrosinée reste très faible, et dans le modèle CDT_N, le niveau du microtubule détyrosiné reste prédominant.

Nos modèles mathématiques expliquent les échecs expérimentaux inattendus en fournissant une explication mécaniste de l'incapacité d'augmenter directement le statut de tyrosination de manière pharmacologique en activant uniquement l'enzyme TTL dans les modèles cellulaires. Les modèles mathématiques rationalisent ainsi l'inactivité des composés étudiés dans les cellules prolifératives et neuronales. La modélisation mathématique permet de soutenir les étapes de validation de cibles thérapeutiques.

4.2.5 Prédiction de l'effet de l'inhibition de la réaction de détyrosination validée expérimentalement

Nous avons utilisé le modèle CDTP afin de déterminer la sensibilité du statut de tyrosination pour d'autres paramètres cinétiques notamment celui lié à la réaction de détyrosination (k_1).

Les indices de sensibilité du modèle CDTP indiquent une tolérance de cent pour cent pour le paramètre cinétique k_1 avant que le statut de tyrosination ne s'écarte de son état d'équilibre de soixante-dix pour cent (Figure 46A). L'analyse de sensibilité indique que le statut de tyrosination peut être modulé par une forte modulation de la réaction de détyrosination (k_1). Nous avons simulé l'ajout d'un inhibiteur qui pourrait diminuer l'activité de détyrosination en imitant une expérience de dose-réponse dans les cellules prolifératives (Figure 46B).

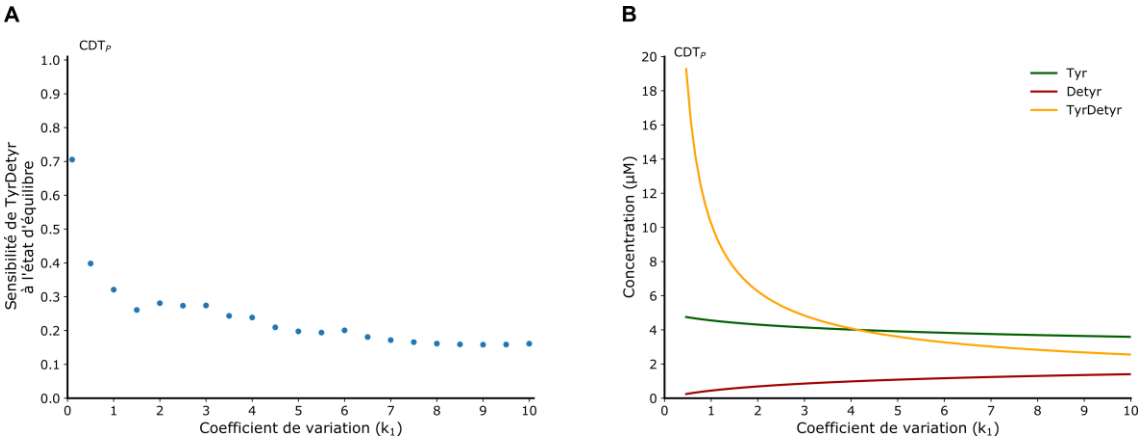

Figure 46. Prédiction de l'inhibition de l'activité de l'enzyme TCP en dose-réponse montrant une augmentation du statut de tyrosination dans les cellules prolifératives. (A) Analyse de sensibilité de la valeur d'équilibre de `TyrDetyr` obtenue pour différents coefficients de variation du paramètre cinétique k_1 dans le modèle de calcul CDTP, indiquant que la valeur d'équilibre de `TyrDetyr` est sensible pour une forte variation de k_1. La commande BIOCHAM utilisée est : `sensitivity(F(G(TyrDetyr = x)), [k1], [x -> 10], robustness_coeff_var : c).` où `c` représente la valeur du coefficient de robustesse. **(B)** Diagramme en dose-réponses du modèle CDTP en faisant varier le paramètre cinétique k_1. Les commandes BIOCHAM utilisées sont : `change_parameter_to_variable(k1)` et `dose_response(k1, 0, 10, time:100, show: TyrDetyr)`. Le résultat de l'exécution de ces commandes BIOCHAM dessine un diagramme dose-réponse par variation linéaire de la concentration initiale (la dose) de l'objet d'entrée, ici k_1, et trace l'objet de sortie (la réponse), ici les espèces moléculaires : `Tyr`, `Detyr` et `TyrDetyr`. Ce diagramme dose-réponse numérique montre une augmentation lisse du statut de tyrosination avec une diminution de k_1.

Dans le modèle CDTP, le rapport entre les espèces tyrosinées et détyrosinées augmente avec la diminution de k_1 (Figure 46). La diminution du paramètre cinétique lié à la détyrosination (k_1) est donc censée augmenter le statut de tyrosination.

Afin de valider la prédiction du modèle mathématique, nous avons réalisé une expérience HCS sur des cellules prolifératives de type MEF en criblant l'inhibiteur de référence de l'enzyme TCP, le parthénolide [127]. Le criblage du parthénolide augmente le statut de tyrosination (Figure 47). Les expériences réalisées (Figure 47) confirment la prédiction du modèle CDTP (Figure 46) sur l'effet de l'inhibition de la réaction de détyrosination sur le statut de tyrosination dans les cellules prolifératives.

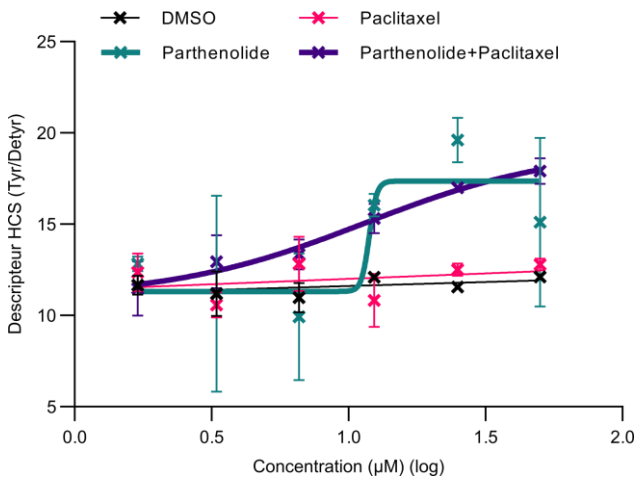

Figure 47. Validation expérimentale de la prédiction du modèle mathématique en inhibant l'enzyme TCP par le parthénolide en dose-réponse montrant une augmentation du statut de tyrosination. Digrammes dose-réponses du descripteur HCS Tyr/Detyr sous l'effet du parthenolide dans des cellules prolifératives du type MEF. Le statut de tyrosination augmente en inhibant la réaction de détyrosination avec le parthenolide, composé de référence pour inhiber l'enzyme TCP. Les concentrations de parthénolide sont indiquées sur l'axe des abscisses en échelle logarithmique. La concentration de paclitaxel est fixée à 5 µM. Nous observons que les barres d'erreur du parthénolide et du DMSO se chevauchent à des concentrations supérieures à 1.5 µM (log). À cette concentration de parthénolide, les morphologies cellulaires sont altérées, les cytoplasmes sont réduits et les cellules semblent fortement stressées.

Nous observons qu'une diminution de k_1 dans le modèle CDT_P (Figure 46B) induit une augmentation du statut de tyrosination mais pas selon une réponse sigmoïdale comme observé expérimentalement (Figure 47). Nous pouvons émettre l'hypothèse, à partir de ces observations et de nos précédentes hypothèses de modélisation, que le parthénolide a un effet sigmoïdal sur l'enzyme TCP alors que l'action de TCP, portée dans le modèle mathématique par le paramètre cinétique k_1, a une action lisse sur le microtubule tyrosiné. Le modèle mathématique ne rend pas compte de l'effet saturant d'un inhibiteur de l'enzyme TCP sur le ratio entre les espèces tyrosinées et détyrosinées. Nous devrions modifier la loi de la réaction de détyrosination en incluant un effet saturant, par exemple, par l'inclusion d'une nouvelle espèce dans le modèle pour représenter un composé chimique agissant sur l'enzyme TCP.

Nos modèles prédisent et capturent l'effet de l'inhibition de l'enzyme TCP [127]. Ces prédictions sont validées par des expériences d'imagerie à haut contenu et des données de la littérature. De plus, les modèles fournissent une explication mécaniste de la capacité d'augmenter directement le statut de tyrosination en inhibant uniquement l'enzyme TCP. En effet, ces résultats s'expliquent par le fait que le microtubule tyrosiné est un réactant direct de la réaction de détyrosination (Figure 39). Le niveau de microtubule tyrosiné peut donc être augmenté en diminuant l'activité de la réaction de sa transformation en microtubule détyrosiné. L'inhibition de la réaction de détyrosination seule est donc efficace pour augmenter le statut de tyrosination dans les cellules prolifératives. En plus d'expliquer des résultats d'expériences inattendues, les modèles mathématiques développés permettent de diriger sur le choix de nouvelles cibles thérapeutiques.

4.2.6 Conception d'une nouvelle expérience de criblage avec une combinaison

Dans la perspective de concevoir de nouvelles expériences de criblage dans des modèles cellulaires neuronaux et afin de repositionner des activateurs de l'enzyme TTL, nous étudions une stratégie de combinaison en utilisant le modèle CDT_N. En effet, le coût en développement des activateurs de TTL est important. Ainsi, des méthodes permettant de repositionner ces molécules sont utiles.

Comme indiqué précédemment, la tubuline tyrosinée est le produit d'une chaîne de deux réactions dans le cycle de tyrosination : la dépolymérisation du microtubule détyrosiné (k_{m1}) suivie de la tyrosination de la tubuline détyrosinée (V_{m2}). L'activation d'une seule des deux réactions n'est pas efficace. En effet, nous avons précédemment montré que l'augmentation seul de l'activité de TTL (V_{m2}) ne permet pas d'augmenter le statut de tyrosination. Nous simulons, ici, l'ajout d'un activateur de la réaction de dépolymérisation du microtubule détyrosiné (k_{m1}) après que le système a atteint son état d'équilibre (Figure 48).

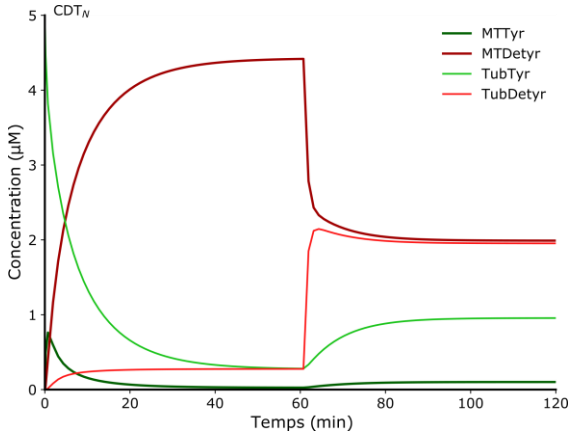

Figure 48. Prédiction de l'effet de l'augmentation de la vitesse de réaction de dépolymérisation du microtubule détyrosiné ne permettant pas d'augmenter le statut de tyrosination significativement. Simulation numérique perturbée dans le modèle CDT_N. La constante de vitesse de dépolymérisation du microtubule détyrosiné k_{m1} est augmentée d'un facteur dix à soixante unités de temps (min). Les espèces tyrosinées augmentent mais ne sont pas supérieures aux espèces détyrosinées à l'état d'équilibre.

Cette simulation numérique (Figure 48) suggère que lorsque le système atteint son état d'équilibre, l'augmentation de la dépolymérisation du microtubule détyrosiné (k_{m1}) seul ne permet pas une augmentation du statut de tyrosination. Cependant, nous observons une augmentation importante de tubuline détyrosinée ce qui suggère qu'un pool de tubuline devient disponible pour réintégrer le cycle de tyrosination (Figure 48). Par conséquent, nous simulons une combinaison d'effets en augmentant les paramètres cinétiques de tyrosination (V_{m2}) et de dépolymérisation du microtubule détyrosiné (k_{m1}) en synergie dans le modèle CDT_N (Figure 49). Le niveau des espèces tyrosinées devient rapidement plus important que celui des espèces détyrosinées (Figure 49).

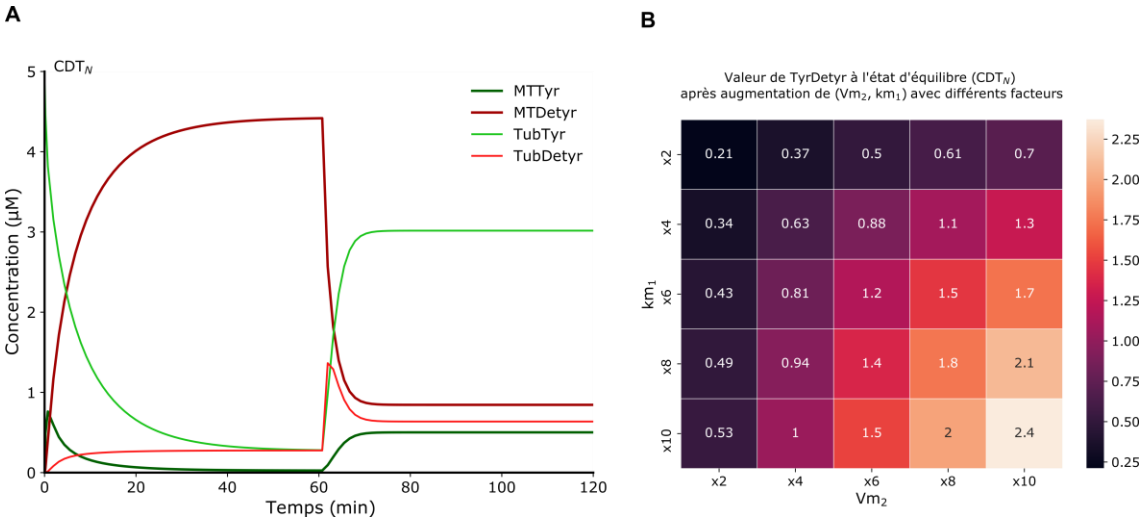

Figure 49. Prédictions des effets des augmentations en synergie des vitesses des réactions de tyrosination et de dépolymérisation du microtubule détyrosiné permettant d'augmenter le statut de tyrosination dans les neurones. (A) Simulation numérique perturbée dans le modèle CDT_N. La constante de vitesse de dépolymérisation du microtubule détyrosiné k_{m1} et la constante de vitesse de tyrosination V_{m2} sont augmentées à soixante unités de temps (min) par un facteur dix. La simulation numérique montre que le niveau des espèces tyrosinées devient rapidement plus important que celui des espèces détyrosinées. **(B)** Prédiction du statut de tyrosination à l'état d'équilibre en augmentant, en synergie, les paramètres (V_{m2}, k_{m1}) par différents facteurs montrant une augmentation du statut de tyrosination. Les valeurs représentées correspondent aux valeurs du ratio (`TyrDetyr`) entre les espèces tyrosinées et détyrosinées à l'état d'équilibre.

L'augmentation en synergie des réactions de dépolymérisation du microtubule détyrosiné et de la réaction de tyrosination devrait donc déclencher une augmentation significative des concentrations des espèces tyrosinées.

D'un point de vue biologique, l'augmentation des facteurs dépolymérisant devrait accroître la dépolymérisation du microtubule détyrosiné et permettre à la tubuline d'être directement disponible pour réintégrer le cycle de tyrosination, alors que dans le même temps, l'augmentation de l'activité de tyrosination devrait accroître la probabilité que la tubuline soit tyrosinée. En outre, nous pouvons simuler diverses augmentations des paramètres liés à la tyrosination et à la dépolymérisation du microtubule détyrosiné (Figure 49B). Il est intéressant de noter que pour différentes augmentations des paramètres (V_{m2}, k_{m1}), une augmentation du statut de tyrosination est observée (Figure 49B).

Les modèles mathématiques indiquent que l'augmentation de la réaction de tyrosination en synergie avec l'augmentation de la dépolymérisation du microtubule détyrosiné apparait comme une stratégie pertinente afin d'augmenter le statut de tyrosination dans les neurones. Cette prédiction serait intéressante à valider dans une campagne de criblage pour repositionner les activateurs de l'enzyme TTL avec des composés chimiques pouvant activer la réaction de dépolymérisation du microtubule.

4.3 Identification de cibles thérapeutiques et de perturbateurs par Pegasus

Le graphe de connaissances Pegasus permet d'identifier des perturbateurs pouvant moduler des cibles thérapeutiques associées aux paramètres cinétiques identifiés comme pertinents par nos modèles mathématiques.

Nos modèles mathématiques du cycle de tyrosination des microtubules montrent qu'en inhibant l'enzyme TCP avec le parthenolide nous pouvons augmenter le statut de tyrosination. Or, le parthénolide est une molécule brevetée. Ainsi, nous pouvons requêter Pegasus, à titre illustratif, pour identifier des perturbateurs propriétaires similaires au parthénolide pour de futures validations expérimentales si nous souhaitons augmenter le statut de tyrosination en inhibant TCP (Requête 10).

De même, nous pouvons identifier des perturbateurs propriétaires qui induisent des signatures phénotypiques similaires à celles induites par le parthenolide dans des expériences de criblages phénotypiques sous l'hypothèse que des composés qui induisent des signatures phénotypiques similaires agissent sur les mêmes voies de signalisation (Requête 10).

Enfin, nos modèles mathématiques prédisent qu'en augmentant la tyrosination et la dépolymérisation des microtubules, nous pouvons augmenter le statut de tyrosination. Nous pouvons requêter Pegasus pour identifier des perturbateurs qui modulent des cibles thérapeutiques qui participent à la régulation positive de la dépolymérisation des microtubules (Requête 10) afin de les tester en synergie, dans de futures campagnes de criblage, avec les activateurs de l'enzyme TTL.

```
##### Requête 1

MATCH path=(c1:Chemical)-[sim:HAS_SIMILARITIES]-(c2:ChemicalServier)
WHERE c1.chemicalId = 'CHEMBL540445' and sim.tanimoto_similarity > 0.8
RETURN path;

##### Requête 2

MATCH path=(:ChemicalServier)-[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-
(:Phenoprint)-[sim:HAS_SIMILARITIES]-(:Phenoprint)-[:ON_TARGETS]-(:Activity)-
[:HAS_ACTIVITIES]-(c:Chemical)
WHERE c.chemicalId = 'CHEMBL540445' and sim.cosine_similarity > 0.95
RETURN path;

##### Requête 3

MATCH path=(:Chemical)-[:HAS_ACTIVITIES]-(:Activity)-[:ON_TARGETS]-(:Gene)-
[:HAS_REFERENCES*0..2]-(:Gene)-[:PARTICIPATES]-(go:BiologicalProcess)
WHERE go.label = 'positive regulation of microtubule depolymerization'
RETURN path;
```

Requête 10. Identification de nouvelles cibles thérapeutiques et de perturbateurs associés aux paramètres cinétiques des modèles mathématiques. La première requête CYPHER identifie les perturbateurs propriétaires chimiquement similaires au parthénolide avec un indice de Tanimoto supérieur à `0.8`. La deuxième requête CYPHER identifie les perturbateurs propriétaires qui induisent des similarités phénotypiques similaires à celles induites par le parthénolide avec une mesure de similarité de type cosine supérieures à `0.95`. La troisième requête CYPHER identifie des perturbateurs chimiques qui ont des activités sur des cibles participant à la régulation positive de la dépolymérisation des microtubules.

4.4 Conclusion

Nous venons de présenter un modèle mathématique mécaniste, le premier à notre connaissance, du cycle de tyrosination des microtubules couplant des données d'imagerie et de la littérature. Le modèle mathématique générique est paramétré d'une part, pour les neurones et d'autre part, pour les cellules prolifératives, et permet de comprendre la dynamique du cycle de tyrosination sans introduire toutes les protéines régulatrices modulant les réactions de ce cycle.

Ces modèles mathématiques expliquent les échecs inattendus de la confirmation de l'activité de composés identifiés comme des activateurs de l'enzyme TTL dans ces modèles cellulaires et leur système biochimique complexe. La tubuline tyrosinée est le produit d'une chaîne de deux réactions, la dépolymérisation du microtubule détyrosiné suivie de la tyrosination de la tubuline détyrosinée. Le statut de tyrosination à l'équilibre est donc limité par les deux taux de réaction et l'activation de la réaction de tyrosination seule n'est pas efficace. Les analyses de sensibilité et les simulations numériques prédisent que la diminution du paramètre cinétique de la réaction de détyrosination permet d'augmenter le statut de tyrosination. Nous avons validé cette prédiction dans les cellules prolifératives de type MEF en criblant le composé parthénolide en dose-réponse.

Enfin, les modèles mathématiques prédisent une combinaison consistant à augmenter, en synergie, la vitesse de réaction de tyrosination et de dépolymérisation du microtubule détyrosiné. L'enzyme TTL pourrait être activée directement, bien qu'aucun composé n'ait encore été approuvé, ou indirectement, via l'inhibition de ses inhibiteurs, puisque son activité est diminuée par phosphorylation [134]. Une autre approche consisterait à moduler, en synergie, les voies de signalisation impliquant des facteurs de dépolymérisation telles que les voies PKC, BDNF/TrkB, JNK, Stathmin [176–180]. Il est intéressant de noter que ces voies de signalisation sont dérégulées dans les maladies neurodégénératives [181–184].

Le modèle mathématique paramétrique développé vise à être étendu au-delà des modèles cellulaires que nous avons étudiés et devrait impacter les recherches relatives à d'autres modifications post-traductionnelles des microtubules et leurs dérégulations dans d'autres maladies comme le cancer ou les cardiomyopathies [185–187]. Le travail présenté dans ce chapitre est basé sur l'article de journal et les communications suivantes.

Productions et communications scientifiques et industrielles

J. Grignard, V. Lamamy, E. Vermersch, P. Delagrange, J-P. Stephan, T. Dorval, F. Fages. Mathematical Modeling of The Microtubule Detyrosination/Tyrosination Cycle For Cell-Based Drug Screening Design. *PLOS Computational Biology, 2022.* [188]

J. Grignard, F. Fages, T. Dorval. Mathematical Modeling Of The Microtubule Tyrosination Cycle For Cell-Based Drug Screening Design.
Séminaire GT-BIOSS, 2022
Servier Research Executive Committee, 2020
Neurology Immuno Inflammation Research Conference, 2021 (60 collaborateurs)

La modélisation mathématique mécaniste permet de comprendre la dynamique de processus biochimiques, dont les comportements parfois contre-intuitifs, ne peuvent pas être analysés avec des représentations statiques ni des signatures phénotypiques. Les modèles mathématiques permettent en outre de soutenir les étapes d'identification et de validations de cibles et de conception rationnelle d'expériences de criblage.

Chapitre 5

Conclusion et perspectives

« Le futur appartient à ceux qui croient à la beauté de leurs rêves. »

Eleanor Roosevelt

La première contribution de la thèse porte sur le développement du graphe de connaissances Pegasus avec le formalisme des graphes à propriétés étiquetés. Les concepts introduits permettent de capitaliser sur des données pharmaco-biologiques hétérogènes et de provenances multiples. Nous intégrons des signatures phénotypiques sous forme d'entités reliées entre elles par des relations de similarités. De façon analogue, nous intégrons des similarités chimiques entre perturbateurs. Une ressource non identifiable de façon unique est modélisée par plusieurs entités reliées par des références croisées. Nous distinguons les gènes, transcrits et protéines et nous intégrons différentes classes de perturbateurs. Un nœud intermédiaire permet de relier contextuellement plusieurs entités.

Le modèle de données de Pegasus introduit permet de répondre à des problématiques de projets thérapeutiques. La première application s'inscrit dans le cadre du développement d'un oligonucléotide antisens unique qui sera administré chez un bébé atteint d'encéphalopathie épileptique. Pegasus permet de caractériser les effets hors cibles d'oligonucléotides antisens et supporte la sélection des perturbateurs à cribler et à optimiser pour les phases précliniques. La deuxième application permet de concevoir une nouvelle expérience couplant les oligonucléotides antisens aux cadres de lecture en amont des transcrits. La future validation expérimentale permettra de déterminer si les perturbateurs identifiés par Pegasus sont actifs contre une cible thérapeutique dérégulée dans le syndrome amyotrophique latéral. Cette application de Pegasus pourra s'étendre à tous les projets thérapeutiques travaillant sur des cibles sous exprimées dans des maladies. La troisième application permet de concevoir des librairies de criblage focalisées. Nous avons illustré comment Pegasus permet d'identifier différentes classes de perturbateurs afin de moduler directement ou indirectement des cibles thérapeutiques participant dans des processus biochimiques.

Le graphe de connaissances Pegasus est un outil de support d'aide à la décision et les perspectives portent notamment sur l'intégration de nouvelles sources de données

issues des consortiums européens MELLODY et EUbOPEN [12] et sur l'utilisation d'algorithmes d'apprentissage automatiques et profonds [76,189].

La deuxième contribution de la thèse porte sur le développement de deux algorithmes pour améliorer la conception et l'analyse d'expériences phénotypiques à haut contenu. Le premier algorithme permet, à partir de composés chimiques caractérisés, d'identifier des composés contrôles positifs à une concentration chacun qui maximisent, ensemble, les réponses des descripteurs phénotypiques. L'algorithme de normalisation, après avoir réduit les dimensions de l'espace phénotypique d'entrée, effectue un recalage global des données dans l'espace réduit, à partir de composés contrôles utilisés comme point de référence. Cet algorithme normalise les données phénotypiques et nous permet de calculer des similarités informatives que nous intégrons dans Pegasus.

Le première perspective consiste à valider l'algorithme de normalisation sur les données issues de JUMP-CP. Les données générées représenteront les états phénotypiques de cellules traitées par 120.000 composés chimiques en plus de perturbations génétiques. Une expérience CRISPR/Cas9 sur l'ensemble des gènes du modèle cellulaire sera également réalisée. Si l'algorithme de normalisation ne passe pas à l'échelle car nous calculerons plus de six-mille descripteurs phénotypiques qui proviendront de plusieurs milliers de plaques, nous investiguerons des méthodes de sélection de caractéristiques et d'autres méthodes de réduction des dimensions d'entrées [190]. Lorsque les données de JUMP-CP seront normalisées nous les intégrerons au sein de Pegasus.

L'objectif à long terme de l'utilisation des signatures phénotypiques, obtenues dans différentes conditions expérimentales, est d'avoir la capacité d'identifier de multiples modalités de criblage pour moduler des cibles thérapeutiques ou des processus biochimiques. Cependant, une question ouverte concerne la normalisation de données HCS provenant de différents protocoles expérimentaux notamment lorsque les descripteurs phénotypiques calculés sont différents, rendant impossible leurs comparaisons. Nous travaillons sur une approche couplant des réseaux de neurones profonds avec des termes ontologiques. Nous investiguons tout d'abord cette méthode à partir de données omiques. Les nœuds de la couche d'entrée, les gènes, sont reliés, de façon experte, aux nœuds des couches profondes représentant des termes ontologiques comme des processus biologiques [191]. L'introduction de mesure d'interprétabilité permet d'identifier les nœuds importants du réseau [192]. Nous avons testé cette méthode sur des données d'expression omiques à l'échelle de la cellule unique et les résultats sont encourageants. Nous identifions des processus biologiques et des gènes importants, au sens de la mesure d'interprétabilité [192], et ces derniers sont dérégulés dans des maladies dont les données d'entraînement sont représentatives. Nous étendrons ensuite cette méthode pour l'analyse de données HCS. Les nœuds de la couche d'entrée correspondraient à des descripteurs phénotypiques reliés, de façon experte, aux nœuds des couches profondes représentant des termes ontologiques phénotypiques [29]. Nous envisageons d'annoter automatiquement des signatures phénotypiques avec des termes ontologiques pour comparer ensuite, les réseaux de neurones optimisés, et ce, afin de déduire des similarités phénotypiques de haut niveaux.

La troisième contribution de la thèse porte sur le développement d'un modèle mathématique mécaniste du cycle de tyrosination des microtubules paramétrés d'une part, pour les neurones, et d'autre part, pour les cellules prolifératives. Les modèles expliquent l'inactivité de composés propriétaires dans ces modèles cellulaires alors qu'ils ont

[12] https://www.melloddy.eu/ et https://www.eubopen.org/

précédemment été identifiés comme actifs dans un système biochimique. La tubuline tyrosinée étant le produit d'une chaîne de deux réactions dans le cycle, la dépolymérisation du microtubule détyrosiné suivie de la tyrosination de la tubuline détyrosinée, le statut de tyrosination à l'équilibre est donc limité par les vitesses de ces deux réactions. L'activation de l'enzyme TTL seule n'est pas efficace. De plus, les modèles prédisent que l'inhibition de la réaction de détyrosination permet d'augmenter le statut de tyrosination dans les modèles cellulaires. Nous avons validé expérimentalement cette prédiction en criblant le parthénolide, un inhibiteur de l'enzyme TCP, dans les cellules prolifératives. Enfin, les modèles mathématiques permettent de concevoir une nouvelle expérience de criblage consistant à augmenter, en synergie, la vitesse de réaction de la tyrosination et de la dépolymérisation du microtubule détyrosiné, afin d'augmenter le statut de tyrosination dans les cellules neuronales.

Ce travail de modélisation illustre la complémentarité et l'apport de la modélisation mathématique pour l'identification et la validation de cibles thérapeutiques ainsi que pour concevoir rationnellement des expériences de criblage notamment lorsque la dynamique de processus biochimiques est mal comprise et que leurs représentations statiques ne suffit pas à en rendre compte.

La perspective de la modélisation mathématique experte est de déployer son utilisation pour les projets thérapeutiques dès les phases exploratoires afin d'expliciter nos a priori sur les systèmes biologiques dont les comportements sont parfois contre-intuitifs, et les confronter aux capacités de prédictions des modèles mathématiques. Nous travaillons actuellement sur un modèle mathématique pour modéliser les réactions de complexation ternaire des PROTACs avec des cibles thérapeutiques [193] ainsi que sur un modèle mathématique pour comprendre les réactions de liaisons concurrentes d'anticorps bispécifiques propriétaires à leurs récepteurs membranaires [194].

Une autre perspective industrielle porte sur le développement d'algorithmes d'apprentissage automatique de modèles à partir de données. Ces algorithmes permettraient d'accélérer la mise à disposition de modèles mathématiques pour les projets thérapeutiques car le processus de modélisation expert peut s'avérer long et complexe. Nous avons développé initialement un algorithme d'inférence de règles d'influences à partir de données temporelles [195,196], que nous n'avons pas présenté pas dans ce manuscrit, mais qui s'inscrit pleinement dans cette perspective.

Productions scientifiques

J. Martinelli, J. Grignard, S. Soliman, F. Fages. A Statistical Unsupervised Learning Algorithm For Inferring Reaction Networks From Time Series Data. *ICML - Workshop on Computational Biology, 2019*. [195]

J. Martinelli, J. Grignard, S. Soliman, F. Fages. On Inferring Reactions From Data Time Series By A Statistical Learning Greedy Heuristics. *International Conference on Computational Methods in Systems Biology, 352-355, 2019*. [196]

Les contributions scientifiques et les applications industrielles présentées dans cette thèse améliorent les phases primaires de recherche de nouveaux médicaments. Les méthodes computationnelles développées et leurs champs d'applications, s'inscrivent pleinement dans notre effort global afin de répondre au nouvel objectif stratégique fixé par le Groupe Servier, à savoir, obtenir une autorisation de mise sur le marché d'un nouveau médicament tous les trois ans, et ont vocation à s'étendre aux phases cliniques du processus pharmaceutique.

Bibliographie

1. DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. Journal of Health Economics. 2016;47: 20–33. doi:10.1016/j.jhealeco.2016.01.012

2. Wouters OJ, McKee M, Luyten J. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. JAMA. 2020;323: 844. doi:10.1001/jama.2020.1166

3. Hughes JP, Rees S, Kalindjian SB, Philpott KL. Principles of early drug discovery. Br J Pharmacol. 2011;162: 1239–1249. doi:10.1111/j.1476-5381.2010.01127.x

4. Santos R, Ursu O, Gaulton A, Bento AP, Donadi RS, Bologa CG, et al. A comprehensive map of molecular drug targets. Nat Rev Drug Discov. 2017;16: 19–34. doi:10.1038/nrd.2016.230

5. Blake RA. Target Validation in Drug Discovery. High Content Screening. New Jersey: Humana Press; 2006. pp. 367–378. doi:10.1385/1-59745-217-3:367

6. Dorval T, Chanrion B, Cattin M-E, Stephan JP. Filling the drug discovery gap: is high-content screening the missing link? Current Opinion in Pharmacology. 2018;42: 40–45. doi:10.1016/j.coph.2018.07.002

7. Alteri E, Guizzaro L. Be open about drug failures to speed up research. Nature. 2018;563: 317–319. doi:10.1038/d41586-018-07352-7

8. Van Norman GA. Phase II Trials in Drug Development and Adaptive Trial Design. JACC: Basic to Translational Science. 2019;4: 428–437. doi:10.1016/j.jacbts.2019.02.005

9. Harrison RK. Phase II and phase III failures: 2013–2015. Nat Rev Drug Discov. 2016;15: 817–818. doi:10.1038/nrd.2016.184

10. Lin A, Giuliano CJ, Palladino A, John KM, Abramowicz C, Yuan ML, et al. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. Sci Transl Med. 2019;11. doi:10.1126/scitranslmed.aaw8412

11. Siramshetty VB, Nickel J, Omieczynski C, Gohlke B-O, Drwal MN, Preissner R. Withdrawn - a resource for withdrawn and discontinued drugs. Nucleic Acids Res. 2016;44: D1080–D1086. doi:10.1093/nar/gkv1192

12. Parasrampuria DA, Benet LZ, Sharma A. Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations. AAPS J. 2018;20: 46. doi:10.1208/s12248-018-0204-y

13. Bonner S, Barrett IP, Ye C, Swiers R, Engkvist O, Bender A, et al. A Review of Biomedical Datasets Relating to Drug Discovery: A Knowledge Graph Perspective. 2021. Available: http://arxiv.org/abs/2102.10062

14. Zock J. Applications of High Content Screening in Life Science Research. CCHTS. 2009;12: 870–876. doi:10.2174/138620709789383277

15. Hucka M, Bergmann FT, Chaouiya C, Dräger A, Hoops S, Keating SM, et al. The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core Release 2. Journal of Integrative Bioinformatics. 2019;16. doi:10.1515/jib-2019-0021

16. Li C, Donizelli M, Rodriguez N, Dharuri H, Endler L, Chelliah V, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. BMC Syst Biol. 2010;4: 92. doi:10.1186/1752-0509-4-92

17. Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic Acids Research. 2021;49: D884–D891. doi:10.1093/nar/gkaa942

18. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. 2012;22: 1760–1774. doi:10.1101/gr.135350.111

19. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2016;44: D7–D19. doi:10.1093/nar/gkv1290

20. The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Agivetova R, et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 2021;49: D480–D489. doi:10.1093/nar/gkaa1100

21. Tweedie S, Braschi B, Gray K, Jones TEM, Seal RL, Yates B, et al. Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research. 2021;49: D939–D946. doi:10.1093/nar/gkaa980

22. Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Research. 2012;40: D1100–D1107. doi:10.1093/nar/gkr777

23. Koscielny G, An P, Carvalho-Silva D, Cham JA, Fumis L, Gasparyan R, et al. Open Targets: a platform for therapeutic target identification and validation. Nucleic Acids Res. 2017;45: D985–D994. doi:10.1093/nar/gkw1055

24. Weng G, Shen C, Cao D, Gao J, Dong X, He Q, et al. PROTAC-DB: an online database of PROTACs. Nucleic Acids Research. 2021;49: D1381–D1387. doi:10.1093/nar/gkaa807

25. Wang X. miRDB: A microRNA target prediction and functional annotation database with a wiki interface. RNA. 2008;14: 1012–1017. doi:10.1261/rna.965408

26. Agarwal V, Bell GW, Nam J-W, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. eLife. 2015;4: e05005. doi:10.7554/eLife.05005

27. Chen X. TTD: Therapeutic Target Database. Nucleic Acids Research. 2002;30: 412–415. doi:10.1093/nar/30.1.412

28. Arrowsmith CH, Audia JE, Austin C, Baell J, Bennett J, Blagg J, et al. The promise and peril of chemical probes. Nat Chem Biol. 2015;11: 536–541. doi:10.1038/nchembio.1867

29. Jupp S, Malone J, Burdett T, Heriche J-K, Williams E, Ellenberg J, et al. The cellular microscopy phenotype ontology. J Biomed Semant. 2016;7: 28. doi:10.1186/s13326-016-0074-0

30. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene Ontology: tool for the unification of biology. Nat Genet. 2000;25: 25–29. doi:10.1038/75556

31. Huttlin EL, Bruckner RJ, Navarrete-Perea J, Cannon JR, Baltier K, Gebreab F, et al. Dual proteome-scale networks reveal cell-specific remodeling of the human interactome. Cell. 2021;184: 3022-3040.e28. doi:10.1016/j.cell.2021.04.011

32. Cheng F, Desai RJ, Handy DE, Wang R, Schneeweiss S, Barabási A-L, et al. Network-based approach to prediction and population-based validation of in silico drug repurposing. Nat Commun. 2018;9: 2691. doi:10.1038/s41467-018-05116-5

33. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Research. 2019; gkz1031. doi:10.1093/nar/gkz1031

34. Wei C-H, Kao H-Y, Lu Z. PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research. 2013;41: W518–W522. doi:10.1093/nar/gkt441

35. Uhlen M, Oksvold P, Fagerberg L, Lundberg E, Jonasson K, Forsberg M, et al. Towards a knowledge-based Human Protein Atlas. Nat Biotechnol. 2010;28: 1248–1250. doi:10.1038/nbt1210-1248

36. Eppig JT. Mouse Genome Informatics (MGI) Resource: Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse. ILAR Journal. 2017;58: 17–41. doi:10.1093/ilar/ilx013

37. Szymański P, Markowicz M, Mikiciuk-Olasik E. Adaptation of High-Throughput Screening in Drug Discovery - Toxicological Screening Tests. IJMS. 2011;13: 427–452. doi:10.3390/ijms13010427

38. Lin S, Schorpp K, Rothenaigner I, Hadian K. Image-based high-content screening in drug discovery. Drug Discovery Today. 2020;25: 1348–1361. doi:10.1016/j.drudis.2020.06.001

39. Rinaldi C, Wood MJA. Antisense oligonucleotides: the next frontier for treatment of neurological disorders. Nat Rev Neurol. 2018;14: 9–21. doi:10.1038/nrneurol.2017.148

40. Bartha I, di Iulio J, Venter JC, Telenti A. Human gene essentiality. Nat Rev Genet. 2018;19: 51–62. doi:10.1038/nrg.2017.75

41. Williams E, Moore J, Li SW, Rustici G, Tarkowska A, Chessel A, et al. Image Data Resource: a bioimage data integration and publication platform. Nat Methods. 2017;14: 775–781. doi:10.1038/nmeth.4326

42. Wethmar K, Barbosa-Silva A, Andrade-Navarro MA, Leutz A. uORFdb - a comprehensive literature database on eukaryotic uORF biology. Nucl Acids Res. 2014;42: D60–D67. doi:10.1093/nar/gkt952

43. Scholz A, Eggenhofer F, Gelhausen R, Grüning B, Zarnack K, Brüne B, et al. uORF-Tools - Workflow for the determination of translation-regulatory upstream open reading frames. Jan E, editor. PLoS ONE. 2019;14: e0222459. doi:10.1371/journal.pone.0222459

44. Jackson R, Matentzoglu N, Overton JA, Vita R, Balhoff JP, Buttigieg PL, et al. OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies. Database. 2021;2021: baab069. doi:10.1093/database/baab069

45. Antoniou G, Van Harmelen F. A semantic Web primer. 2nd ed. Cambridge, Mass: MIT Press; 2008.

46. Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, et al. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database. 2015;2015: bav028–bav028. doi:10.1093/database/bav028

47. Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF, editors. The Description Logic Handbook: Theory, implementation, and applications. 2nd ed. Cambridge University Press; 2007. doi:10.1017/CBO9780511711787

48. Kalibatiene D, Vasilecas O. Survey on Ontology Languages. In: Grabis J, Kirikova M, editors. Perspectives in Business Informatics Research. Berlin, Heidelberg: Springer Berlin Heidelberg; 2011. pp. 124–141. doi:10.1007/978-3-642-24511-4_10

49. Visser U, Abeyruwan S, Vempati U, Smith RP, Lemmon V, Schürer SC. BioAssay Ontology (BAO): a semantic description of bioassays and high-throughput screening results. BMC Bioinformatics. 2011;12: 257. doi:10.1186/1471-2105-12-257

50. Gremse M, Chang A, Schomburg I, Grote A, Scheer M, Ebeling C, et al. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. Nucleic Acids Research. 2011;39: D507–D513. doi:10.1093/nar/gkq968

51. Schriml LM, Mitraka E, Munro J, Tauber B, Schor M, Nickle L, et al. Human Disease Ontology 2018 update: classification, content and workflow expansion. Nucleic Acids Research. 2019;47: D955–D962. doi:10.1093/nar/gky1032

52. Kamdar MR, Musen MA. An empirical meta-analysis of the life sciences linked open data on the web. Sci Data. 2021;8: 24. doi:10.1038/s41597-021-00797-y

53. Burgstaller-Muehlbacher S, Waagmeester A, Mitraka E, Turner J, Putman T, Leong J, et al. Wikidata as a semantic framework for the Gene Wiki initiative. Database. 2016;2016: baw015. doi:10.1093/database/baw015

54. Pérez J, Arenas M, Gutierrez C. Semantics and complexity of SPARQL. ACM Trans Database Syst. 2009;34: 1–45. doi:10.1145/1567274.1567278

55. Geleta D, Nikolov A, Edwards G, Gogleva A, Jackson R, Jansson E, et al. Biological Insights Knowledge Graph: an integrated knowledge graph to support drug development. Systems Biology; 2021 Nov. doi:10.1101/2021.10.28.466262

56. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. eLife. 2017;6: e26726. doi:10.7554/eLife.26726

57. Breit A, Ott S, Agibetov A, Samwald M. OpenBioLink: a benchmarking framework for large-scale biomedical link prediction. Lu Z, editor. Bioinformatics. 2020;36: 4097–4098. doi:10.1093/bioinformatics/btaa274

58. Santos A, Colaço AR, Nielsen AB, Niu L, Geyer PE, Coscia F, et al. Clinical Knowledge Graph Integrates Proteomics Data into Clinical Decision-Making. Bioinformatics; 2020 May. doi:10.1101/2020.05.09.084897

59. Pareja-Tobes P, Tobes R, Manrique M, Pareja E, Pareja-Tobes E. Bio4j: a high-performance cloud-enabled graph-based data platform. Bioinformatics; 2015 Mar. doi:10.1101/016758

60. Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, et al. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol. 2022. doi:10.1038/s41587-021-01145-6

61. Zhang R, Hristovski D, Schutte D, Kastrin A, Fiszman M, Kilicoglu H. Drug repurposing for COVID-19 via knowledge graph completion. Journal of Biomedical Informatics. 2021;115: 103696. doi:10.1016/j.jbi.2021.103696

62. CovidGraph - a COVID-19 Knowledge Graph - HealthECCO. Available: https://healthecco.org/covidgraph/

63. Lysenko A, Roznovăţ IA, Saqi M, Mazein A, Rawlings CJ, Auffray C. Representing and querying disease networks using graph databases. BioData Mining. 2016;9: 23. doi:10.1186/s13040-016-0102-8

64. Francis N, Green A, Guagliardo P, Libkin L, Lindaaker T, Marsault V, et al. Cypher: An Evolving Query Language for Property Graphs. Proceedings of the

2018 International Conference on Management of Data. Houston TX USA: ACM; 2018. pp. 1433–1445. doi:10.1145/3183713.3190657

65. Donkers AJA, Yang D, Baken N. Linked Data for Smart Homes: Comparing RDF and Labeled Property Graphs. CEUR Workshop Proceedings. 2636: 23–36.

66. Alocci D, Mariethoz J, Horlacher O, Bolleman JT, Campbell MP, Lisacek F. Property Graph vs RDF Triple Store: A Comparison on Glycan Substructure Search. Helmer-Citterich M, editor. PLoS ONE. 2015;10: e0144578. doi:10.1371/journal.pone.0144578

67. Margitus M, Tauer G, Sudit M. RDF Versus Attributed Graphs: The War for the Best Graph Representation. International Conference on Information Fusion (Fusion). 2015;18: 200–206.

68. RDF Triple Stores vs. Labeled Property Graphs: What's the Difference. Neo4J blog. 201718.

69. Mazein A, Ostaszewski M, Kuperstein I, Watterson S, Le Novère N, Lefaudeux D, et al. Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. npj Syst Biol Appl. 2018;4: 21. doi:10.1038/s41540-018-0059-y

70. Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, et al. AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease. BMC Syst Biol. 2012;6: 52. doi:10.1186/1752-0509-6-52

71. Ostaszewski M, Mazein A, Gillespie ME, Kuperstein I, Niarakis A, Hermjakob H, et al. COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. Sci Data. 2020;7: 136. doi:10.1038/s41597-020-0477-8

72. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Research. 2009;37: D412–D416. doi:10.1093/nar/gkn760

73. Stark C. BioGRID: a general repository for interaction datasets. Nucleic Acids Research. 2006;34: D535–D539. doi:10.1093/nar/gkj109

74. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000;28: 27–30.

75. Croft D, Mundo AF, Haw R, Milacic M, Weiser J, Wu G, et al. The Reactome pathway knowledgebase. Nucleic acids research. 2013;42: D472–D477.

76. Zeng X, Zhu S, Lu W, Liu Z, Huang J, Zhou Y, et al. Target identification among known drugs by deep learning from heterogeneous networks. Chem Sci. 2020;11: 1775–1797. doi:10.1039/C9SC04336E

77. Sun X, Gao H, Yang Y, He M, Wu Y, Song Y, et al. PROTACs: great opportunities for academia and industry. Sig Transduct Target Ther. 2019;4: 64. doi:10.1038/s41392-019-0101-6

78. Roberts TC, Langer R, Wood MJA. Advances in oligonucleotide drug delivery. Nat Rev Drug Discov. 2020;19: 673–694. doi:10.1038/s41573-020-0075-7

79. Zhao S, Zhang B. A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification. BMC Genomics. 2015;16: 97. doi:10.1186/s12864-015-1308-8

80. Braschi B, Seal RL, Tweedie S, Jones TEM, Bruford EA. The risks of using unapproved gene symbols. The American Journal of Human Genetics. 2021;108: 1813–1816. doi:10.1016/j.ajhg.2021.09.004

81. Corcos L, Solier S. Épissage alternatif, pathologie et thérapeutique moléculaire. Med Sci (Paris). 2005;21: 253–260. doi:10.1051/medsci/2005213253

82. Pruitt KD, Harrow J, Harte RA, Wallin C, Diekhans M, Maglott DR, et al. The consensus coding sequence (CCDS) project: Identifying a common protein-coding gene set for the human and mouse genomes. Genome Res. 2009;19: 1316–1323. doi:10.1101/gr.080531.108

83. Reisen F, Zhang X, Gabriel D, Selzer P. Benchmarking of Multivariate Similarity Measures for High-Content Screening Fingerprints in Phenotypic Drug Discovery. J Biomol Screen. 2013;18: 1284–1297. doi:10.1177/1087057113501390

84. O'Boyle NM. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. J Cheminform. 2012;4: 22. doi:10.1186/1758-2946-4-22

85. Pattanaik L, Coley CW. Molecular Representation: Going Long on Fingerprints. Chem. 2020;6: 1204–1207. doi:10.1016/j.chempr.2020.05.002

86. Bajusz D, Rácz A, Héberger K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform. 2015;7: 20. doi:10.1186/s13321-015-0069-3

87. Bălan L, Kiyohito Kunii (Kiyo), Deriabin D, Hoang L, Ivaniuk A, Yetunde Dada, et al. quantumblacklabs/kedro: 0.17.0. Zenodo; 2020. doi:10.5281/ZENODO.4336685

88. Radaelli G, de Souza Santos F, Borelli WV, Pisani L, Nunes ML, Scorza FA, et al. Causes of mortality in early infantile epileptic encephalopathy: A systematic review. Epilepsy & Behavior. 2018;85: 32–36. doi:10.1016/j.yebeh.2018.05.015

89. Karaki S, Paris C, Rocchi P. Antisense Oligonucleotides, A Novel Developing Targeting Therapy. In: Sharad S, Kapur S, editors. Antisense Therapy. IntechOpen; 2019. doi:10.5772/intechopen.82105

90. Batra G, Jain M, Singh R, Sharma A, Singh A, Prakash A, et al. Novel therapeutic targets for amyotrophic lateral sclerosis. Indian J Pharmacol. 2019;51: 418. doi:10.4103/ijp.IJP_823_19

91. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans.

Proceedings of the National Academy of Sciences. 2009;106: 7507–7512. doi:10.1073/pnas.0810916106

92. Barbosa C, Peixeiro I, Romão L. Gene Expression Regulation by Upstream Open Reading Frames and Human Disease. Fisher EMC, editor. PLoS Genet. 2013;9: e1003529. doi:10.1371/journal.pgen.1003529

93. Winkle M, El-Daly SM, Fabbri M, Calin GA. Noncoding RNA therapeutics - challenges and potential solutions. Nat Rev Drug Discov. 2021;20: 629–651. doi:10.1038/s41573-021-00219-z

94. Levin AA. Treating Disease at the RNA Level with Oligonucleotides. N Engl J Med. 2019;380: 57–70. doi:10.1056/NEJMra1705346

95. Lieberman J. Tapping the RNA world for therapeutics. Nat Struct Mol Biol. 2018;25: 357–364. doi:10.1038/s41594-018-0054-4

96. Zhao X, Voutila J, Ghobrial S, Habib NA, Reebye V. Treatment of Liver Cancer by C/EBPA saRNA. In: Li L-C, editor. RNA Activation. Singapore: Springer Singapore; 2017. pp. 189–194. doi:10.1007/978-981-10-4310-9_13

97. Kogej T, Blomberg N, Greasley PJ, Mundt S, Vainio MJ, Schamberger J, et al. Big pharma screening collections: more of the same or unique libraries? The AstraZeneca–Bayer Pharma AG case. Drug Discovery Today. 2013;18: 1014–1024. doi:10.1016/j.drudis.2012.10.011

98. Jensen ON. Interpreting the protein language using proteomics. Nat Rev Mol Cell Biol. 2006;7: 391–403. doi:10.1038/nrm1939

99. Jerabek-Willemsen M, André T, Wanner R, Roth HM, Duhr S, Baaske P, et al. MicroScale Thermophoresis: Interaction analysis and beyond. Journal of Molecular Structure. 2014;1077: 101–113. doi:10.1016/j.molstruc.2014.03.009

100. Rainard JM, Pandarakalam GC, McElroy SP. Using Microscale Thermophoresis to Characterize Hits from High-Throughput Screening: A European Lead Factory Perspective. SLAS Discovery. 2018;23: 225–241. doi:10.1177/2472555217744728

101. Shoichet BK. Virtual screening of chemical libraries. Nature. 2004;432: 862–865. doi:10.1038/nature03197

102. Woo J, Hong J, Dinesh-Kumar SP. Bioluminescence Resonance Energy Transfer (BRET)-Based Synthetic Sensor Platform for Drug Discovery. Current Protocols in Protein Science. 2017;88. doi:10.1002/cpps.30

103. Cuccarese MF, Earnshaw BA, Heiser K, Fogelson B, Davis CT, McLean PF, et al. Functional immune mapping with deep-learning enabled phenomics applied to immunomodulatory and COVID-19 drug discovery. Immunology; 2020 Aug. doi:10.1101/2020.08.02.233064

104. Haney SA, editor. High Content Screening: Science, Techniques and Applications. 1st ed. Wiley; 2008. doi:10.1002/9780470229866

105. Moffat JG, Vincent F, Lee JA, Eder J, Prunotto M. Opportunities and challenges in phenotypic drug discovery: an industry perspective. Nat Rev Drug Discov. 2017;16: 531–543. doi:10.1038/nrd.2017.111

106. Kraus OZ. Automating High Content Screening with Deep Learning. PhD Thesis.

107. Chandrasekaran SN, Ceulemans H, Boyd JD, Carpenter AE. Image-based profiling for drug discovery: due for a machine-learning upgrade? Nat Rev Drug Discov. 2021;20: 145–159. doi:10.1038/s41573-020-00117-w

108. Scheeder C, Heigwer F, Boutros M. Machine learning and image-based profiling in drug discovery. Current Opinion in Systems Biology. 2018;10: 43–52. doi:10.1016/j.coisb.2018.05.004

109. Swinney DC. Chapter 1. Phenotypic Drug Discovery: History, Evolution, Future. In: Isherwood B, Augustin A, editors. Drug Discovery. Cambridge: Royal Society of Chemistry; 2020. pp. 1–19. doi:10.1039/9781839160721-00001

110. Neumann B, Walter T, Hériché J-K, Bulkescher J, Erfle H, Conrad C, et al. Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature. 2010;464: 721–727. doi:10.1038/nature08869

111. Entzeroth M, Flotow H, Condron P. Overview of High-Throughput Screening. Current Protocols in Pharmacology. 2009;44. doi:10.1002/0471141755.ph0904s44

112. Pepperkok R, Ellenberg J. High-throughput fluorescence microscopy for systems biology. Nat Rev Mol Cell Biol. 2006;7: 690–696. doi:10.1038/nrm1979

113. Bray M-A, Singh S, Han H, Davis CT, Borgeson B, Hartland C, et al. Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. Nat Protoc. 2016;11: 1757–1774. doi:10.1038/nprot.2016.105

114. Caicedo JC, Cooper S, Heigwer F, Warchal S, Qiu P, Molnar C, et al. Data-analysis strategies for image-based cell profiling. Nat Methods. 2017;14: 849–863. doi:10.1038/nmeth.4397

115. Mpindi J-P, Swapnil P, Dmitrii B, Jani S, Saeed K, Wennerberg K, et al. Impact of normalization methods on high-throughput screening data with high hit rates and drug testing with dose–response data. Bioinformatics. 2015; btv455. doi:10.1093/bioinformatics/btv455

116. Bray M-A, Carpenter A. Advanced Assay Development Guidelines for Image-Based High Content Screening and Analysis. In: Markossian S, Grossman A, Brimacombe K, Arkin M, Auld D, Austin CP, et al., editors. Assay Guidance Manual. Eli Lilly & Company and the National Center for Advancing Translational Sciences; 2004. Available: http://www.ncbi.nlm.nih.gov/books/NBK126174/

117. Zhang J-H. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. Journal of Biomolecular Screening. 1999;4: 67–73. doi:10.1177/108705719900400206

118. Singh S, Carpenter AE, Genovesio A. Increasing the Content of High-Content Screening: An Overview. J Biomol Screen. 2014;19: 640–650. doi:10.1177/1087057114528537

119. Caraus I, Alsuwailem AA, Nadon R, Makarenkov V. Detecting and overcoming systematic bias in high-throughput screening technologies: a comprehensive review of practical issues and methodological solutions. Briefings in Bioinformatics. 2015;16: 974–986. doi:10.1093/bib/bbv004

120. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

121. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Bertrand Thirion, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12: 2825–2830.

122. Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2: e453. doi:10.7717/peerj.453

123. Friston KarlJ, Ashburner J, Frith CD, Poline J-B, Heather JD, Frackowiak RSJ. Spatial registration and normalization of images. Hum Brain Mapp. 1995;3: 165–189. doi:10.1002/hbm.460030303

124. Cipriani G, Dolciotti C, Picchi L, Bonuccelli U. Alzheimer and his disease: a brief history. Neurological Sciences. 2011;32: 275–279.

125. Briggs R, Kennelly SP, O'Neill D. Drug treatments in Alzheimer's disease. Clinical medicine. 2016;16: 247–253.

126. Eira J, Silva CS, Sousa MM, Liz MA. The cytoskeleton as a novel therapeutic target for old neurodegenerative disorders. Progress in Neurobiology. 2016;141: 61–82.

127. Fonrose X, Ausseil F, Soleilhac E, Masson V, David B, Pouny I, et al. Parthenolide inhibits tubulin carboxypeptidase activity. Cancer research. 2007;67: 3371–3378.

128. Gobrecht P, Andreadaki A, Diekmann H, Heskamp A, Leibinger M, Fischer D. Promotion of functional nerve regeneration by inhibition of microtubule detyrosination. Journal of Neuroscience. 2016;36: 3890–3902.

129. Nieuwenhuis J, Brummelkamp TR. The Tubulin Detyrosination Cycle: Function and Enzymes. Trends in Cell Biology. 2019;29: 80–92. doi:10.1016/j.tcb.2018.08.003

130. Janke C, Chloë Bulinski J. Post-translational regulation of the microtubule cytoskeleton: mechanisms and functions. Nat Rev Mol Cell Biol. 2011;12: 773–786. doi:10.1038/nrm3227

131. Aillaud C, Bosc C, Peris L, Bosson A, Heemeryck P, Van Dijk J, et al. Vasohibins/SVBP are tubulin carboxypeptidases (TCPs) that regulate neuron differentiation. Science. 2017;358: 1448–1453.

132. Nieuwenhuis J, Adamopoulos A, Bleijerveld OB, Mazouzi A, Stickel E, Celie P, et al. Vasohibins encode tubulin detyrosinating activity. Science. 2017;358: 1453–1456.

133. Beltramo D, Arce C, Barra H. Tubulin, but not microtubules, is the substrate for tubulin: tyrosine ligase in mature avian erythrocytes. Journal of Biological Chemistry. 1987;262: 15673–15677.

134. Prota AE, Magiera MM, Kuijpers M, Bargsten K, Frey D, Wieser M, et al. Structural basis of tubulin tyrosination by tubulin tyrosine ligase. J Cell Biol. 2013;200: 259–270.

135. Jaworski J, Kapitein LC, Gouveia SM, Dortland BR, Wulf PS, Grigoriev I, et al. Dynamic microtubules regulate dendritic spine morphology and synaptic plasticity. Neuron. 2009;61: 85–100.

136. Song W, Cho Y, Watt D, Cavalli V. Tubulin-tyrosine ligase (TTL)-mediated increase in tyrosinated α-tubulin in injured axons is required for retrograde injury signaling and axon regeneration. Journal of Biological Chemistry. 2015;290: 14765–14775.

137. Kahn OI, Baas PW. Microtubules and growth cones: motors drive the turn. Trends in neurosciences. 2016;39: 433–440.

138. Nirschl JJ, Magiera MM, Lazarus JE, Janke C, Holzbaur EL. α-Tubulin tyrosination and CLIP-170 phosphorylation regulate the initiation of dynein-driven transport in neurons. Cell reports. 2016;14: 2637–2652.

139. Bieling P, Kandels-Lewis S, Telley IA, van Dijk J, Janke C, Surrey T. CLIP-170 tracks growing microtubule ends by dynamically recognizing composite EB1/tubulin-binding sites. J Cell Biol. 2008;183: 1223–1233.

140. Marcos S, Moreau J, Backer S, Job D, Andrieux A, Bloch-Gallego E. Tubulin Tyrosination Is Required for the Proper Organization and Pathfinding of the Growth Cone. Hendricks M, editor. PLoS ONE. 2009;4: e5405. doi:10.1371/journal.pone.0005405

141. Kato C, Miyazaki K, Nakagawa A, Ohira M, Nakamura Y, Ozaki T, et al. Low expression of human tubulin tyrosine ligase and suppressed tubulin tyrosination/detyrosination cycle are associated with impaired neuronal differentiation in neuroblastomas with poor prognosis. International journal of cancer. 2004;112: 365–375.

142. Zhang F, Su B, Wang C, Siedlak SL, Mondragon-Rodriguez S, Lee H, et al. Posttranslational modifications of α-tubulin in alzheimer disease. Transl Neurodegener. 2015;4: 9–17. doi:10.1186/s40035-015-0030-4

143. Dorostkar MM, Zou C, Blazquez-Llorca L, Herms J. Analyzing dendritic spine pathology in Alzheimer's disease: problems and opportunities. Acta neuropathologica. 2015;130: 1–19.

144. Penzes P, Cahill ME, Jones KA, VanLeeuwen J-E, Woolfrey KM. Dendritic spine pathology in neuropsychiatric disorders. Nature neuroscience. 2011;14: 285.

145. Peris L, Qu X, Soleilhac J-M, Parato J, Lanté F, Kumar A, et al. Impaired α-tubulin re-tyrosination leads to synaptic dysfunction and is a feature of Alzheimer's disease. Neuroscience; 2021 May. doi:10.1101/2021.05.17.443847

146. Dunn S, Morrison EE, Liverpool TB, Molina-París C, Cross RA, Alonso MC, et al. Differential trafficking of Kif5c on tyrosinated and detyrosinated microtubules in live cells. Journal of cell science. 2008;121: 1085–1095.

147. Gardner MK, Charlebois BD, Jánosi IM, Howard J, Hunt AJ, Odde DJ. Rapid microtubule self-assembly kinetics. Cell. 2011;146: 582–592.

148. Iniguez A, Allard J. Spatial pattern formation in microtubule post-translational modifications and the tight localization of motor-driven cargo. Journal of mathematical biology. 2017;74: 1059–1080.

149. Hervy J. Modeling the dynamical interaction Tau Proteins - microtubules. Thèse, Université Grenoble Alpes. 2018. Available: https://tel.archives-ouvertes.fr/tel-02053825

150. Dogterom M, Leibler S. Physical aspects of the growth and regulation of microtubule structures. Phys Rev Lett. 1993;70: 1347–1350. doi:10.1103/PhysRevLett.70.1347

151. Kalil K, Dent EW. Branch management: mechanisms of axon branching in the developing vertebrate CNS. Nat Rev Neurosci. 2014;15: 7–18. doi:10.1038/nrn3650

152. Takano T, Funahashi Y, Kaibuchi K. Neuronal Polarity: Positive and Negative Feedback Signals. Front Cell Dev Biol. 2019;7: 69–78. doi:10.3389/fcell.2019.00069

153. Takano T, Wu M, Nakamuta S, Naoki H, Ishizawa N, Namba T, et al. Discovery of long-range inhibitory signaling to ensure single axon formation. Nature communications. 2017;8: 33.

154. Schelski M, Bradke F. Neuronal polarization: From spatiotemporal signaling to cytoskeletal dynamics. Molecular and Cellular Neuroscience. 2017;84: 11–28.

155. Poulain FE, Sobel A. The microtubule network and neuronal morphogenesis: Dynamic and coordinated orchestration through multiple players. Molecular and Cellular Neuroscience. 2010;43: 15–32.

156. Arimura N, Kaibuchi K. Neuronal polarity: from extracellular signals to intracellular mechanisms. Nat Rev Neurosci. 2007;8: 194–205. doi:10.1038/nrn2056

157. Calzone L, Fages F, Soliman S. BIOCHAM: an environment for modeling biological systems and formalizing experimental knowledge. Bioinformatics. 2006;22: 1805–1807. doi:10.1093/bioinformatics/btl172

158. Rizk A, Batt G, Fages F, Soliman S. Continuous valuations of temporal logic specifications with applications to parameter optimization and robustness measures. Theoretical Computer Science. 2011;412: 2827–2839. doi:10.1016/j.tcs.2010.05.008

159. Deans NL, Allison RD, Purich DL. Steady-state kinetic mechanism of bovine brain tubulin: tyrosine ligase. Biochemical Journal. 1992;286: 243–251. doi:10.1042/bj2860243

160. Homma N, Takei Y, Tanaka Y, Nakata T, Terada S, Kikkawa M, et al. Kinesin superfamily protein 2A (KIF2A) functions in suppression of collateral branch extension. Cell. 2003;114: 229–239.

161. Peris L, Wagenbach M, Lafanechère L, Brocard J, Moore AT, Kozielski F, et al. Motor-dependent microtubule disassembly driven by tubulin tyrosination. The Journal of cell biology. 2009;185: 1159–1166.

162. Webster DR, Gundersen GG, Bulinski JC, Borisy GG. Differential turnover of tyrosinated and detyrosinated microtubules. Proceedings of the National Academy of Sciences. 1987;84: 9040–9044.

163. Hiller G, Weber K. Radioimmunoassay for tubulin: a quantitative comparison of the tubulin content of different established tissue culture cells and tissues. Cell. 1978;14: 795–804.

164. Gard DL, Kirschner MW. Microtubule assembly in cytoplasmic extracts of Xenopus oocytes and eggs. The Journal of Cell Biology. 1987;105: 2191–2201. doi:10.1083/jcb.105.5.2191

165. Baudier A, Fages F, Soliman S. Graphical requirements for multistationarity in reaction networks and their verification in BioModels. Journal of Theoretical Biology. 2018;459: 79–89. doi:10.1016/j.jtbi.2018.09.024

166. Song Y, Brady ST. Post-translational modifications of tubulin: pathways to functional diversity of microtubules. Trends in cell biology. 2015;25: 125–136.

167. Schulze E, Kirschner M. Dynamic and stable populations of microtubules in cells. The Journal of cell biology. 1987;104: 277–288.

168. Bulinski JC, Gundersen GG. Stabilization and post-translational modification of microtubules during cellular morphogenesis. Bioessays. 1991;13: 285–293.

169. Kreitzer G, Liao G, Gundersen GG. Detyrosination of tubulin regulates the interaction of intermediate filaments with microtubules in vivo via a kinesin-dependent mechanism. Molecular Biology of the Cell. 1999;10: 1105–1118.

170. Hansen N, Ostermeier A. Completely Derandomized Self-Adaptation in Evolution Strategies. Evolutionary Computation. 2001;9: 159–195. doi:10.1162/106365601750190398

171. Wehland J, Weber K. Turnover of the carboxy-terminal tyrosine of alpha-tubulin and means of reaching elevated levels of detyrosination in living cells. Journal of Cell Science. 1987;88: 185–203.

172. Zhang J-H. A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. Journal of Biomolecular Screening. 1999;4: 67–73. doi:10.1177/108705719900400206

173. Raybin D, Flavin M. Modification of tubulin by tyrosylation in cells and extracts and its effect on assembly in vitro. The Journal of cell biology. 1977;73: 492–504.

174. Chen J, Kholina E, Szyk A, Fedorov VA, Kovalenko I, Gudimchuk N, et al. α-tubulin tail modifications regulate microtubule stability through selective effector recruitment, not changes in intrinsic polymer dynamics. Developmental Cell. 2021;56: 2016-2028.e4. doi:10.1016/j.devcel.2021.05.005

175. Kitano H. Towards a theory of biological robustness. Mol Syst Biol. 2007;3: 137. doi:10.1038/msb4100179

176. Idriss HT. Phosphorylation of tubulin tyrosine ligase: A Potential Mechanism for Regulation of α-Tubulin Tyrosination. Cell motility and the cytoskeleton. 2000;46: 1–5.

177. Jeanneteau F, Deinhardt K, Miyoshi G, Bennett AM, Chao MV. The MAP kinase phosphatase MKP-1 regulates BDNF-induced axon branching. Nat Neurosci. 2010;13: 1373–1379. doi:10.1038/nn.2655

178. Westerlund N, Zdrojewska J, Padzik A, Komulainen E, Björkblom B, Rannikko E, et al. Phosphorylation of SCG10/stathmin-2 determines multipolar stage exit and neuronal migration rate. Nat Neurosci. 2011;14: 305–313. doi:10.1038/nn.2755

179. Uchida S, Martel G, Pavlowsky A, Takizawa S, Hevi C, Watanabe Y, et al. Learning-induced and stathmin-dependent changes in microtubule stability are critical for memory and disrupted in ageing. Nature communications. 2014;5: 4389.

180. Uchida S, Shumyatsky GP. Deceivingly dynamic: learning-dependent changes in stathmin and microtubules. Neurobiology of learning and memory. 2015;124: 52–61.

181. Gonzalez-Billault C, Jimenez-Mateos EM, Caceres A, Diaz-Nido J, Wandosell F, Avila J. Microtubule-associated protein 1B function during normal development, regeneration, and pathological conditions in the nervous system. Journal of neurobiology. 2004;58: 48–59.

182. Kawauchi T, Chihama K, Nishimura YV, Nabeshima Y, Hoshino M. MAP1B phosphorylation is differentially regulated by Cdk5/p35, Cdk5/p25, and JNK. Biochemical and biophysical research communications. 2005;331: 50–55.

183. Tarrade A, Fassier C, Courageot S, Charvin D, Vitte J, Peris L, et al. A mutation of spastin is responsible for swellings and impairment of transport in a region of axon characterized by changes in microtubule composition. Human Molecular Genetics. 2006;15: 3544–3558. doi:10.1093/hmg/ddl431

184. Chauvin S, Sobel A. Neuronal stathmins: a family of phosphoproteins cooperating for neuronal development, plasticity and regeneration. Progress in neurobiology. 2015;126: 1–18.

185. Lafanechère L, Courtay-Cahen C, Kawakami T, Jacrot M, Rüdiger M, Wehland J, et al. Suppression of tubulin tyrosine ligase during tumor growth. Journal of Cell Science. 1998;111: 171–181.

186. Caporizzo MA, Chen CY, Prosser BL. Cardiac microtubules in health and heart disease. Exp Biol Med (Maywood). 2019;244: 1255–1272. doi:10.1177/1535370219868960

187. Magiera MM, Singh P, Gadadhar S, Janke C. Tubulin Posttranslational Modifications and Emerging Links to Human Disease. Cell. 2018;173: 1323–1327.

188. Grignard J, Lamamy V, Vermersch E, Delagrange P, Stephan J-P, Dorval T, et al. Mathematical modeling of the microtubule detyrosination/tyrosination cycle for cell-based drug screening design. PLoS Comput Biol. 2022;18: e1010236. doi:10.1371/journal.pcbi.1010236

189. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, et al. Graph neural networks: A review of methods and applications. AI Open. 2020;1: 57–81. doi:10.1016/j.aiopen.2021.01.001

190. Siegismund D, Fassler M, Heyse S, Steigele S. Benchmarking feature selection methods for compressing image information in high-content screening. SLAS Technology. 2022;27: 85–93. doi:10.1016/j.slast.2021.10.015

191. Peng J, Wang X, Shang X. Combining gene ontology with deep neural networks to enhance the clustering of single cell RNA-Seq data. BMC Bioinformatics. 2019;20: 284. doi:10.1186/s12859-019-2769-6

192. de Sá CR. Variance-Based Feature Importance in Neural Networks. In: Kralj Novak P, Šmuc T, Džeroski S, editors. Discovery Science. Cham: Springer International Publishing; 2019. pp. 306–315. doi:10.1007/978-3-030-33778-0_24

193. Han B. A suite of mathematical solutions to describe ternary complex formation and their application to targeted protein degradation by heterobifunctional ligands. Journal of Biological Chemistry. 2020;295: 15280–15291. doi:10.1074/jbc.RA120.014715

194. Rhoden JJ, Dyas GL, Wroblewski VJ. A Modeling and Experimental Investigation of the Effects of Antigen Density, Binding Affinity, and Antigen Expression Ratio on Bispecific Antibody Binding to Cell Surface Targets. Journal of Biological Chemistry. 2016;291: 11337–11347. doi:10.1074/jbc.M116.714287

195. Martinelli J, Grignard J, Soliman S, Fages F. A Statistical Unsupervised Learning Algorithm for Inferring Reaction Networks from Time Series Data. ICML 2019 - Workshop on Computational Biology. Long Beach, CA, United States; 2019. Available: https://hal.inria.fr/hal-02163862

196. Martinelli J, Grignard J, Soliman S, Fages F. On Inferring Reactions from Data Time Series by a Statistical Learning Greedy Heuristics. Computational Methods in Systems Biology. 2019. pp. 352–355. doi:10.1007/978-3-030-31304-3_25

ECOLE DOCTORALE

Titre : Méthodes computationnelles pour améliorer les phases primaires de recherche de nouveaux médicaments

Mots clés : graphe de connaissances, science des données, criblage à haut contenu, modélisation mathématique, découverte de nouveaux médicaments

Résumé : Le processus de découverte de nouveaux médicaments est long, coûteux et très risqué. L'objectif de cette thèse de doctorat est d'améliorer la pertinence des phases primaires de recherche pharmaceutique en développant des méthodes computationnelles. La première contribution porte sur le développement du graphe de connaissances Pegasus afin de capitaliser sur les données pharmaco-biologiques hétérogènes et de provenances multiples du secteur pharmaceutique. Les applications industrielles de Pegasus répondent à des problématiques de projets thérapeutiques et permettent de caractériser des effets hors cibles de perturbateurs, de concevoir une nouvelle expérience, et d'identifier des librairies de criblage focalisées. La deuxième contribution porte sur le développement d'un algorithme d'identification de composés contrôles positifs et d'un algorithme de normalisation afin d'améliorer la conception et l'analyse d'expériences de criblage phénotypiques à haut contenu. Ces algorithmes permettent de normaliser les signatures phénotypiques obtenues à partir de campagnes de criblage et d'intégrer des similarités phénotypiques informatives dans le graphe de connaissances Pegasus. La troisième contribution porte sur le développement d'un modèle mathématique du cycle de tyrosination des microtubules qui explique, d'une part, l'inactivité de composés chimiques dans les cellules montrés actifs hors cellule, et d'autre part, suggère la nécessité d'activer deux réactions de ce cycle, en synergie, pour obtenir un effet dans les modèles cellulaires. Ceci illustre l'apport de la modélisation mathématique pour, d'une part, prédire et comprendre la dynamique contre-intuitive de processus biochimiques qui n'est pas représentable par des graphes de connaissances statiques comme dans Pegasus, et d'autre part, guider la conception de nouvelles expériences de criblage. Les contributions scientifiques et les applications industrielles de cette thèse sont développées dans le cadre des phases primaires de recherche de nouveaux médicaments et ont vocation à s'étendre aux phases cliniques du processus pharmaceutique.

Title: Computational methods to improve the early drug discovery

Keywords: knowledge graph, data science, high-content screening, computational modeling, drug discovery

Abstract: The drug discovery process is long, costly and very risky. The objective of this doctoral thesis is to improve the relevance of the primary phases of drug discovery by developing computational methods. The first contribution concerns the development of the Pegasus knowledge graph in order to capitalize on the heterogeneous and multi-sourced pharmaco-biological data of the pharmaceutical sector. The industrial applications of Pegasus address the problems of therapeutic projects and allow to characterize off-target effects of perturbators, to design a new experiment, and to identify focused screening libraries. The second contribution concerns the development of an algorithm for the identification of positive control compounds and a normalization algorithm to improve the design and analysis of high content phenotypic screening experiments. These algorithms allow the normalization of phenotypic signatures obtained from screening campaigns and the integration of informative phenotypic similarities in the Pegasus knowledge graph. The third contribution concerns the development of a mathematical model of the microtubule tyrosination cycle which explains, on the one hand, the inactivity of chemical compounds shown to be active outside the cell, and on the other hand, suggests the need to activate two reactions of this cycle, in synergy, to obtain an effect in cellular models. This illustrates the contribution of mathematical modeling to, on the one hand, predict and understand the counter-intuitive dynamics of biochemical processes that are not representable by static knowledge graphs as in Pegasus, and on the other hand, guide the design of new screening experiments. The scientific contributions and industrial applications of this thesis are developed in the context of the primary phases of drug discovery and are intended to extend to the clinical phases of the pharmaceutical process.

Institut Polytechnique de Paris
91120 Palaiseau, France