Identification of new therapeutic targets for cystic fibrosis using systems biology and chemogenomics approaches

Matthieu Najm

PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Mines Paris-PSL

Recherche de nouvelles cibles thérapeutiques pour la mucoviscidose par des approches de biologie des systèmes et de chémogénomique

Identification of new therapeutic targets for cystic fibrosis using systems biology and chemogenomics approaches

Soutenue par

Matthieu Najm

Le 30 Novembre 2023

Ecole doctorale n° 621

Ingénierie des Systèmes, Matériaux, Mécanique, Énergétique

Spécialité

Bio-informatique

Composition du jury :

Michael, NILGES Professeur, Institut Pasteur	*Président*
Anaïs, BAUDOT Directrice de recherche, Marseille Medical Genetics	*Rapporteure*
Pascal, BARBRY Directeur de recherche, Institut de Pharmacologie Moléculaire et Cellulaire	*Rapporteur*
Benno, SCHWIKOWSKI Directeur de recherche, Institut Pasteur	*Examinateur*
Philippe, REIX Professeur des universités praticien hospitalier, CHU de Lyon HCL	*Examinateur*
Laurence, CALZONE Ingénieure de recherche, Institut Curie	*Directrice de thèse*
Véronique, STOVEN Professeure, Mines Paris - PSL	*Directrice de thèse*

MINES PARIS | PSL

Résumé

La mucoviscidose est la maladie autosomale grave la plus fréquente dans la population caucasienne. Elle est causée par des mutations du gène codant pour la protéine CFTR (Cystic Fibrosis Transmembrane Regulator), qui agit comme un canal de chlorure (Cl^-) à la membrane des cellules épithéliales. La mucoviscidose est principalement délétère pour les poumons, où l'infection chronique et les lésions tissulaires provoquent progressivement une insuffisance respiratoire. Plus de 2000 mutations sont connues pour le gène CFTR, mais 70% des patients sont homozygotes pour la délétion du résidu F508 (F508del). Le traitement de la mucoviscidose est resté longtemps symptomatique, mais des modulateurs pharmacologiques de CFTR sont disponibles depuis peu. Cependant, ils ont un effet limité chez les patients homozygotes F508del et n'arrêtent pas l'évolution de la maladie. De plus, ils restent spécifiques à certaines mutations, et environ 15% des patients ne peuvent pas en bénéficier. Enfin, leurs cibles protéiques, leurs mécanismes d'action et leurs effets secondaires à long terme sont encore inconnus. Par ailleurs, la pathophysiologie globale de la mucoviscidose ne peut être expliquée uniquement par la perte de la fonction du canal chlorure CFTR. Notre hypothèse est que CFTR appartient à un réseau de protéines qui n'ont pas encore été toutes identifiées et dont les fonctions sont perturbées par l'absence de CFTR, participant ainsi à certains des phénotypes cellulaires anormaux qui caractérisent la maladie. En utilisant des approches de biologie des systèmes et des méthodes d'apprentissage automatique chémogénomique, les objectifs du projet sont les suivants : (1) identifier in-silico des cibles thérapeutiques candidates en construisant le réseau des dérégulations moléculaires de la mucoviscidose causées par l'absence de CFTR à l'aide de données transcriptomiques; (2) identifier les cibles protéiques des modulateurs de CFTR afin de déchiffrer leurs mécanismes d'action. À terme, le projet devrait permettre d'identifier de nouvelles stratégies thérapeutiques combinant des médicaments ciblant la restauration de la maturation et de la fonction de CFTR, à des médicaments ciblant le réseau de dérégulations de la maladie. Cette approche systémique pourrait apporter des solutions thérapeutiques aux patients présentant des mutations pour lesquelles il n'existe actuellement aucune thérapie.

Mots clés : mucoviscidose, biologie des systèmes, chemogenomics, machine-learning, cibles thérapeutiques, transcriptomics

Abstract

Cystic Fibrosis (CF) is the most frequent life-limiting autosomal disease in the Caucasian population. It is caused by mutations in the gene coding the Cystic Fibrosis Transmembrane Regulator (CFTR) protein, acting as a chloride (Cl^-) channel at the membrane of epithelial cells. CF is mainly deleterious for the lung where chronic infection and tissue damage progressively cause respiratory insufficiency. More than 2000 mutations are known in the CFTR gene, but 70% of the patients are homozygous for the deletion of residue F508 (F508del). CF treatment remained symptomatic for a long time, but pharmacologic CFTR modulators became recently available. However, they have a limited effect in F508del homozygous patients, and do not stop disease evolution. They remain mutation specific, and around 15% of CF patients cannot benefit. Moreover, their protein targets, mechanisms of action and long-term side effects are still unknown. In addition, CF overall physiopathology cannot be solely explained by the loss of the CFTR chloride channel function. Our hypothesis is that CFTR belongs to a yet not fully deciphered network of proteins, whose functions are disrupted by the absence of CFTR, thus participating in some of the abnormal cellular phenotypes that characterise CF. Using systems biology approaches and machine-learning chemogenomics methods, the aims of the project are to: (1) identify in-silico candidate therapeutic targets by building the network of CF molecular dysregulations caused by the absence of CFTR based on transcriptomic data; (2) identify protein targets of CFTR modulators to decipher their mechanisms of action. At term, the project should help identify new therapeutic strategies combining drugs targeting restoration of CFTR maturation and function, to drugs targeting the network of CF molecular dysregulations. This systemic approach may provide therapeutic solutions for CF patients with mutations for which there is currently no specific therapy.

Keywords : cystic fibrosis, systems biology, chemogenomics, machine-learning, therapeutic targets, transcriptomics

Remerciements

L'épanouissement et la réussite de cette aventure sont les fruits de plusieurs rencontres que j'ai faites avant et durant ces quatre années. Ces quelques mots sont destinés à remercier toutes les personnes qui ont participé de près ou de loin à cet aboutissement.

Je remercie tout d'abord les membres du jury pour avoir lu ce manuscript et assisté à la soutenance: merci aux rapporteurs Anaïs Baudot et Pascal Barbry pour vos remarques et vos réflexions sur le projet, merci à Michael Nilges d'avoir accepté de présider la soutenance et merci à Benno Schwikowski et Philippe Reix pour vos questions, parfois déroutantes mais toujours intéressantes, pendant la soutenance.

Je remercie aussi les associations Vaincre la Mucoviscidose et Blanche pour Vaincre la Mucoviscidose pour avoir financé cette thèse ainsi que La Fondation Dassault Systèmes et MSD Avenir qui ont participé au financement du projet global.

Je remercie chaleureusement la direction de ma thèse et particulièrement Véronique Stoven. Merci de m'avoir donné la chance de faire partie de cette aventure alors que nous n'avions jamais travaillé ensemble. Merci pour la confiance que tu m'as accordée dès les premiers jours, mais aussi pour toutes les discussions passionnantes sur le projet et les connaissances que tu m'as transmises. Je me souviendrai toujours de cette première année de thèse où nous réfléchissions tous les deux au projet, qui était très loin de ce qu'il est aujourd'hui. Merci surtout pour ton exigence et pour ton franc parler. Même si parfois tes remarques franches m'ont remué, je pense que c'est aussi grâce à cela que la qualité du projet et notre relation ont perduré du début jusqu'à la fin de la thèse. Au delà de la science, merci pour tous les jours où nous avons partagé la V319, à refaire le monde avec Gwenn, à "péter des câbles" à tour de rôle et à essayer de garder le calme et le sourire à l'aide de carrés de chocolat. C'était vraiment de très belles journées.

C'est aussi grâce à toi que j'ai fait la connaissance de Laurence et Loredana. Merci à toutes les deux pour votre supervision complémentaire que je qualifie d'exemplaire. Merci pour toutes les connaissances que vous m'avez transmises et surtout de m'avoir transmis le goût de la "sysbio" que je ne connaissais pas avant de commencer la thèse. Merci enfin pour votre enthousiasme, votre gentillesse, et votre soutien durant ces quatre années.

Depuis le début de cette aventure, j'admire la diversité et la complétude de vos connaissances scientifiques, que ce soit en mathématiques, en informatique, en biologie structurale ou cellulaire et même en physique. Cela m'inspire chaque jour à m'approcher de cette diversité de connaissance. Je trouve sincèrement que nous faisions une très belle équipe tous les quatre. J'ai toujours eu l'impression que je pouvais exprimer une

opinion ou des critiques et qu'elles étaient surtout entendues. J'ai beaucoup apprécié cela et je sais que c'est rare lorsque l'on est doctorant. Je sais qu'à court terme nous continuerons de collaborer mais j'espère très sincèrement que nous garderons contact longtemps tous les quatre que ce soit ou pas dans un contexte de science.

J'ai eu de la chance que mon projet de thèse s'inscrive dans une collaboration étroite avec l'équipe d'Isabelle Sermet-Gaudelus à l'INEM. Je remercie Isabelle et son équipe pour cette collaboration, notamment Mairead et Agathe pour votre patience à me transmettre vos connaissances en biologie. Les après-midis passés à réfléchir ensemble, chacun apportant sa propre perspective, sont parmi mes souvenirs préférés de ces quatre années. Je remercie aussi Benoit C. et Alexandre pour notre collaboration sur le projet de l'interactome de CFTR. Travailler sur ce projet avec vous a été comme une bouffée d'air frais au moment ou le projet de thèse avait du mal à avancer.

Enfin, Matthieu, même si nos chemins se sont éloignés plus tôt que prévu, je veux te remercier pour ces deux années de thèse que nous avons partagées. J'ai apprécié travailler avec toi sur ce projet.

De part sa direction, mon projet de thèse s'est déroulé entre les équipes du CBIO des Mines et l'équipe de sysbio de Curie. Ce fut une chance car j'ai pu bénéficier de la complémentarité de ces deux équipes. J'ai surtout eu beaucoup de chance de faire la rencontre de plein de chercheurs novices et chevronnés, avec qui j'ai beaucoup appris et je me suis amusé. Ainsi je remercie tous les membres de ces deux équipes et ceux qui y sont passés pendant ces quatre années.

Je remercie particulièrement Thomas W. pour ta gentillesse et ton incroyable patience envers tous nos problèmes administratifs.

Je remercie aussi Benoit P., pour ton aide généreuse en début de thèse, même lorsqu'il fallait venir aux Mines un dimanche à 10h, et pour ta gentillesse lorsqu'il fallait répondre à toutes mes questions de débutant.

Mention spéciale pour Tristan: outre les nouveaux joujous IA et autres tubercules que tu m'as fait découvrir, c'était surtout un réel bonheur de passer ces 4 années à tes côtés, à essayer les chants polyphoniques, les sessions pomodoro et partager nos hauts et nos bas doctoraux.

Une pensée vient aussi aux doctorants actuels qui ont rendu le quotidien de la thèse agréable et parfois moins durs: aux Mines, Gwenn pour tous les moments de boulot et surtout pour tous les autres moments (nos discussions procrastrinatrices, nos critiques cinéma, Gand etc), Philippe pour ton amitié, Thomas B. pour ta serenité exemplaire et Thomas D. pour ton énergie et ton enthousiasme nécessaire à l'ambiance du CBIO; à Curie, Andrea, Jonathan, Marco et Loïc pour nos quelques tentatives de collaborations scientifiques échouées mais aussi pour tous les bons moments en conférence, à Barcelone et à Budapest, qui ont apporté un peu de folie à ces 4 ans, et enfin Anne-Claire, Daniel, Lucie et Nicolas pour votre énergie et votre enthousiasme. J'espère sincèrement que je vous recroiserai rapidement et que l'on continuera un peu notre aventure scientifique ensemble.

L'enseignement a aussi tres fortement participé à mon épanouissement durant la thèse, je remercie ainsi tous les étudiants que j'ai pu suivre ou encadrer officieusement et en particulier Paul et Leopold que j'ai beaucoup apprécié encadrer. Merci aussi

Chloé pour m'avoir donné l'opportunité de participer à tes cours aux Mines, cours que je relis régulièrement et que je conseille souvent.

Enfin, je remercie les personnes exterieures aux Mines et à Curie qui ont aussi contribué à ce que cette aventure ait existé et soit réussie. Je remercie ainsi Elli et surtout Franck pour m'avoir accepté en stage en 2016 à MSK, où j'ai fait mes premiers pas dans le domaine de la biologie computationnelle, et à Elsa pour tes conseils qui me suivent tout au long de mon éducation de chercheur.

Je tiens à remercier chaleureusement tous mes amis qui m'ont soutenu pendant ces 4 ans. Cela m'a beaucoup touché que vous vous déplaciez et que vous assistiez à la soutenance. Je remercie particulièrement ceux qui ont partagé l'expérience de faire une thèse: le partage de nos conseils, nos expériences, nos colères et nos angoisses a sans doute aidé à faire passer les moments les plus sombres.

Enfin, je remercie toute ma famille en France, au Canada et au Liban. Je sais qu'une partie d'entre vous aurait aimé être présent le jour de la soutenance mais votre soutien à distance s'est fortement ressenti. Je termine par remercier mes parents pour votre confiance, votre soutien, et de toute évidence votre amour.

Contents

List of figures

List of tables

Glossary

ADR Adverse Drug Reaction
AUPR Area Under the Precision-Recall curve
BMI Body Mass Index
CF Cystic Fibrosis
CFTR Cystic Fibrosis Transmembrane Regulator protein
CV Cross Validation
DEG Differentially Expressed Genes
DTI Drug-Target Interaction
ER Endoplasmic Reticulum
FCS Functional Class Scoring
FN False Negative
FP False Positive
FPR False Positive Rate
GO Gene Ontology
GSEA Gene Set Enrichment Analysis
HAEC Human Airway Epithelial Cells
HT High-Throughput
IDR Innate Defense Regulator
KS Kolmogorov Smirnov
MAPK Mitogen-Activated Protein Kinase
ML Machine-Learning
MoA Mechanism of Action
NBD Nucelotide Binding Domain
ORA Over-Representation Algorithm
PB Pathway-Based
PBMC Peripheral Blood Mononuclear Cell
PC1 first Principal Component
PCA Principal Component Analysis
PKN Prior-Knowledge network

PM Plasma Membrane
PN Positive-Negative
PPI Protein-Protein Interaction
PU Positive-Unknown
QSAR Quantitative structure–activity relationship
RF Random Forests
ROC-AUC Area Under the Receiver Operating Characteristic curve
SBGN Systems Biology Graphical Notation
scRNA-seq Single-cell RNA-seq
ssGSEA Single Sample Gene Set Enrichment Analysis
SVD Single Value Decomposition
SVM Support Vector Machine
TF Transcription factor
TMD Trans-Membrane Domain
TN True Negative
TP True Positive
WT Wild Type

Part I

Introduction

Chapter 1

Introduction

Contents

Abstract

Cystic fibrosis (CF) is the most common life-limiting autosomal disease in the Caucasian population. It is caused by mutations in the gene encoding CFTR. However, the overall physiopathology such as uncontrolled pro-inflammatory response, oxidative stress, or impaired epithelial regeneration cannot be easily linked to the loss of the CFTR chloride channel function alone. CF treatment has long been symptomatic, but pharmacological CFTR modulators have recently become available. However, they have a limited effect in F508del homozygous patients, and do not halt disease progression. They remain mutation specific, and about 15% of CF patients do not benefit. In addition, their protein targets, mechanisms of action and long-term side effects are still unknown. Our hypothesis is that CFTR belongs to a yet not fully deciphered network of proteins, whose functions are disrupted in the absence of CFTR, thus contributing to some of the abnormal cellular phenotypes that characterise CF. After addressing the challenges posed by the molecular mechanisms of CF and its therapeutic solutions, I develop in this introduction the specific questions that led to the formulation of our research hypothesis. I then present the methods developed during the project and the contributions of this thesis.

Résumé

La mucoviscidose est la maladie autosomale grave la plus fréquente dans la population caucasienne. Elle est causée par des mutations du gène codant pour la protéine CFTR. Cependant, la physiopathologie globale de la maladie, telle que la réponse pro-inflammatoire incontrôlée, le stress oxydatif ou l'altération de la régénération épithéliale, ne peut être expliquée uniquement par la perte de la fonction du canal chlorure CFTR. Le traitement de la mucoviscidose a longtemps été symptomatique, mais des modulateurs pharmacologiques de CFTR sont disponibles depuis peu. Ils ont cependant un effet limité chez les patients homozygotes F508del et n'arrêtent pas la progression de la maladie. Ils restent spécifiques à certaines mutations et environ 15% des patients ne peuvent pas en bénéficier. En outre, leurs cibles protéiques, leurs mécanismes d'action et leurs effets secondaires à long terme sont encore inconnus. Notre hypothèse est que CFTR appartient à un réseau de protéines qui n'ont pas encore été toutes identifiées et dont les fonctions sont perturbées en l'absence de CFTR, participant ainsi à certains des phénotypes cellulaires anormaux qui caractérisent la maladie. Après avoir abordé les défis posés par les mécanismes moléculaires de la mucoviscidose et ses solutions thérapeutiques, je développe dans cette introduction les questions spécifiques qui ont conduit à la formulation de notre hypothèse de recherche. Je présente ensuite les méthodes développées au cours du projet et les contributions de cette thèse.

1.1 Preface

In this thesis, I addressed the challenges posed by the molecular mechanisms of cystic fibrosis (CF) by applying two areas of computational biology: systems biology approaches and chemogenomics algorithms. This manuscript presents both methodological developments and their practical applications related to CF.

Although a proper introduction to both of these areas is necessary to understand the work of this thesis, I have chosen to keep the primary focus in this introduction on the biological question. For the sake of clarity, systems biology approaches and chemogenomics algorithms are defined separately in two dedicated chapters prior to their application. Indeed, the importance of the biological question has remained central during the research journey. Of course, it had led us to address methodological issues but these emerged while trying to solve the biological problem.

The section 1.2 of the introduction focuses on presenting the disease from a medical and molecular perspective, emphasising its complex and heterogeneous nature. Then I discuss in section 1.3 the current treatment landscape for CF and I develop, in section 1.4, the specific questions that have driven the formulation of our research hypothesis. In section 1.5, I introduce the methods developed during the project which allowed us to address the biological questions raised. Finally, in section 1.6, I highlight the contributions of this work developed in the main body of this manuscript.

1.2 Cystic Fibrosis: a monogenic disease

1.2.1 The prevalence, the symptoms, and the diagnosis

Cystic Fibrosis is a inherited disorder caused by a mutation on the cystic fibrosis transmembrane regulator (*CFTR*) gene. It is considered as a rare disease, that is a disease which affects less than 1 person per 2000 in Europe. Although rare, CF is the most common life limiting autosomal recessive genetic disease in the Caucasian population, affecting approximately one in 3500 birth [Farrell, 2008]. In France, the prevalence varies according to the regions, from 1/2500 in the North-West to 1/10000 in the South-East.

CF was first considered pediatric, but the life expectancy of CF patients has greatly increased over the past decades, in particular thanks to patient care and development of new therapies. It has gone from less than 5 years old in the 1950s to over 40 years old today [Lopes-Pacheco, 2016].

CF affects the cells that express *CFTR*, and in particular cells involved in the production of mucus, sweat and digestive juices. These fluids are thin and slippery but CF makes them abnormally viscous and thick. These secretions plug up tubes and ducts (Figure 1.1) and cause severe damage to the lungs, to the digestive systems and to other organs in the body. In the lung, CF disease is characterised by altered airway surface liquid pH, decreased host defenses at the airway surface, and chronic bacterial infections [Tarran, 2005; Tang, 2016]. These contribute to chronic inflammation of the airways, which gradually causes damage to lung tissue and leads to respiratory insufficiency. CF morbidity and mortality are mostly caused by the chronic and progressive lung dysfunction.

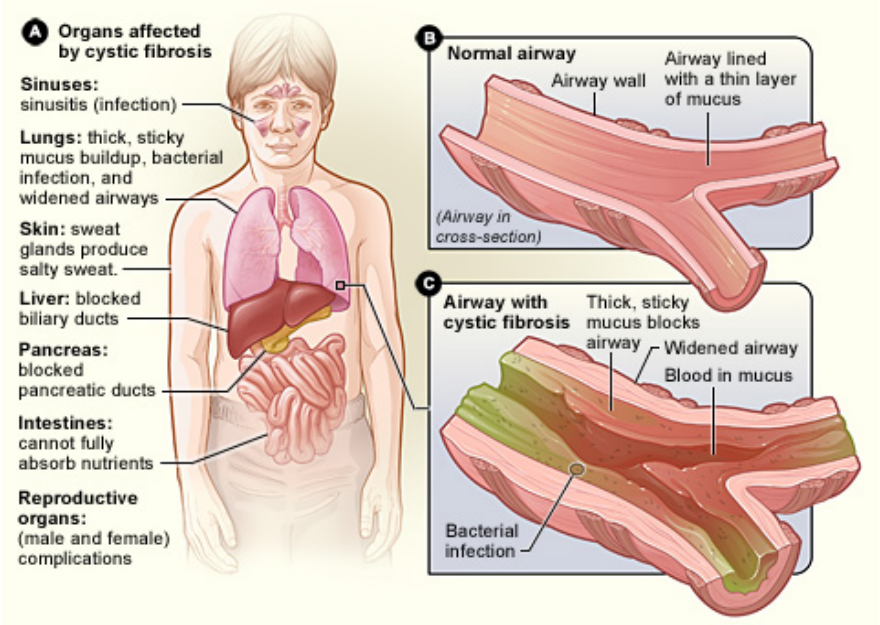


Figure 1.1 – Summary of CF symptoms.
(A) Organs affected by CF, (B) normal airway and (C) airways affected by CF. Illustration from National Heart Lung and Blood Institute (NIH), 12 November 2013. Public Domain.

France has introduced neonatal screening to diagnose CF, which involves measuring the levels of Immunoreactive trypsin (IRT) in the blood of newborns. If the test gives a positive result, further investigation searches for CFTR mutations and a final diagnosis is confirmed by a *sweat test* [Cornet, 2022a]. The so-called *sweat test* consists in measuring the Cl^- ions concentration present in the sweat after sweat stimulation because the sweat of CF patients is highly concentrated in chloride (Cl^-) ions [Di Santagnese, 1953]).

1.2.2 The CFTR protein

The *CFTR* gene codes for the Cystic Fibrosis Transmembrane Regulator (CFTR) protein [Rommens, 1989]. CFTR is a chloride channel situated at the apical membrane of polarised epithelial cells, in particular those lining the airways [Trezise, 1991]. A default in ion transport causes a reduction in liquid hydration of the airway surface. This reduction then prevents efficient mucociliary clearance, which is one of the basic immune defense mechanisms of the respiratory tract [Stoltz, 2015].

CFTR belongs to the ATP Binding Cassette (ABC) superfamily of proteins which carry out substrates inside and outside the cells by hydrolyzing Adenosine TriPhosphate (ATP). Like most ABC proteins, CFTR is composed of two trans-membrane domains (TMD), TMD1 and TMD2, two Nucleotide Binding Domains (NBD), NBD1 and NBD2, and one Regulatory domain R situated between NBD1 and TMD2. The role of the R domain of CFTR is to regulate opening and closing of the chloride chan-

nel. The activation of the channel depends on several elements represented in Figure 1.2. First, the R domain must be phosphorylated by protein kinase A or C (PKA and PKC) [Picciotto, 1992]. As long as the R domain is not phosphorylated, steric effects maintain the channel closed [Bozoky, 2013]. The phosphorylation induces structural changes that displace R from it steric-interfering position and allows dimerisation of the NBDs [Meng, 2017]. ATP binding at the NBDs triggers the opening of the channel, whereas ATP hydrolysis and the subsequent release of ADP result in the closure of the channel [Vergani, 2005]. See Figure 1.2 for a schematic view of CFTR chloride channel opening.

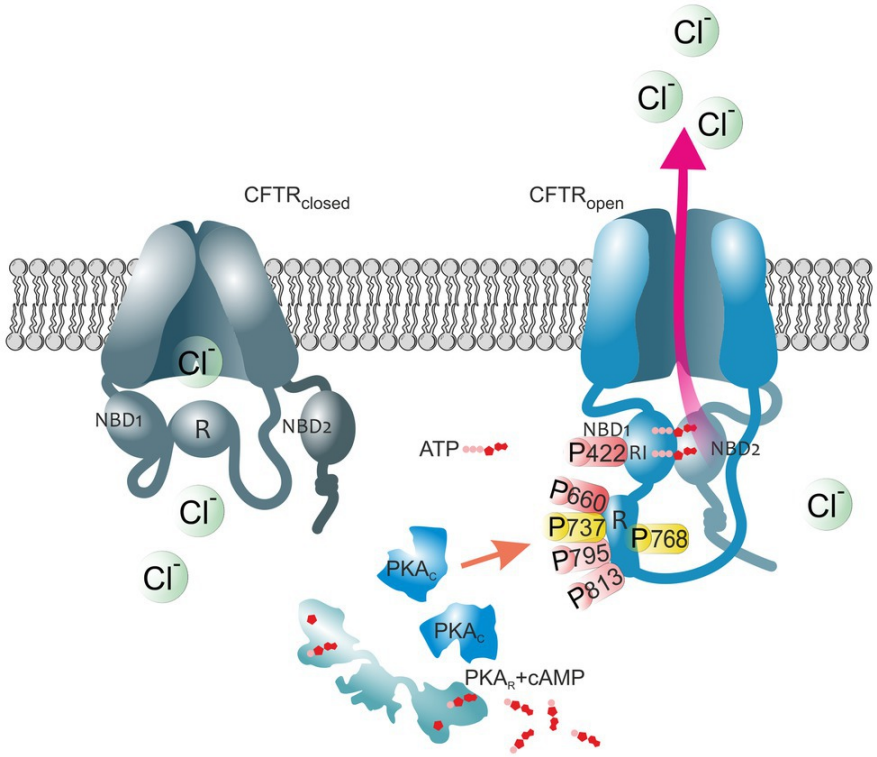


Figure 1.2 – Schematic view of the chloride channel opening.
Illustration from [Della Sala, 2021] used under CC BY 4.0.

The maturation of CFTR involves several steps, including protein synthesis, folding, and trafficking to the plasma membrane (PM), which are collectively referred to as the CFTR proteostatis pathway. CFTR RNA is first synthetized in the nucleus, then it joins the cytosolic ribosomes for the translation and then the endoplasmic reticulum (ER) for post-translational modifications, including folding.

In the ER, CFTR is also subject to the ER Quality Control System (ERQC). ERQC ensures that CFTR is correctly folded and functional, and can migrate to the cell PM [Farinha, 2017]. Conversely, misfolded CFTR is targeted for ER-associated degradation (ERAD) [Jensen, 1995; Ward, 1995]. Once CFTR has passed the ERQC, the protein migrates to the Golgi apparatus to be glycosylated and thus forming the mature protein. The CFTR protein is finally transported to the PM, where it functions as a chloride

channel. Misfolded proteins which are mostly retained in the ER and targeted for ERAD can sometimes bypass the degradation pathway and reach the PM [Sermet-Gaudelus, 2002]. Therefore, the transport of a mutated protein to the PM is possible, it justifies improving the rescue of mutated CFTR as a therapeutic solution [Cornet, 2022a].

1.2.3 *CFTR* mutations

More than 2,000 unique mutations of *CFTR* have been identified. The most common mutation is the deletion of a phenylalanine at position 508 (F508del). This mutation represents about 70% of CF alleles worldwide [Gentzsch, 2018]. The mutations have been grouped into six different classes [Marson, 2016] based on their functional impact on CFTR (See Figure 1.3):

- class I: CFTR nonsense or splicing mutations abrogate CFTR production.
- class II: Many missense mutations impair proper folding of CFTR and lead to its retention in the ER and its degradation by the proteasome. It includes the most common mutation, F508del.
- Some missense and splicing mutations produce CFTR chloride channels that reach the cell surface but are not fully functional due to various defects [Gentzsch, 2018]:
  - class IIII: gating blocking in the closed position even in presence of ATP.
  - class IV: diminished ion conductance.
  - class V: reduced amount of functional CFTR.
  - class VI: decreased membrane residence time at the apical surface.

.

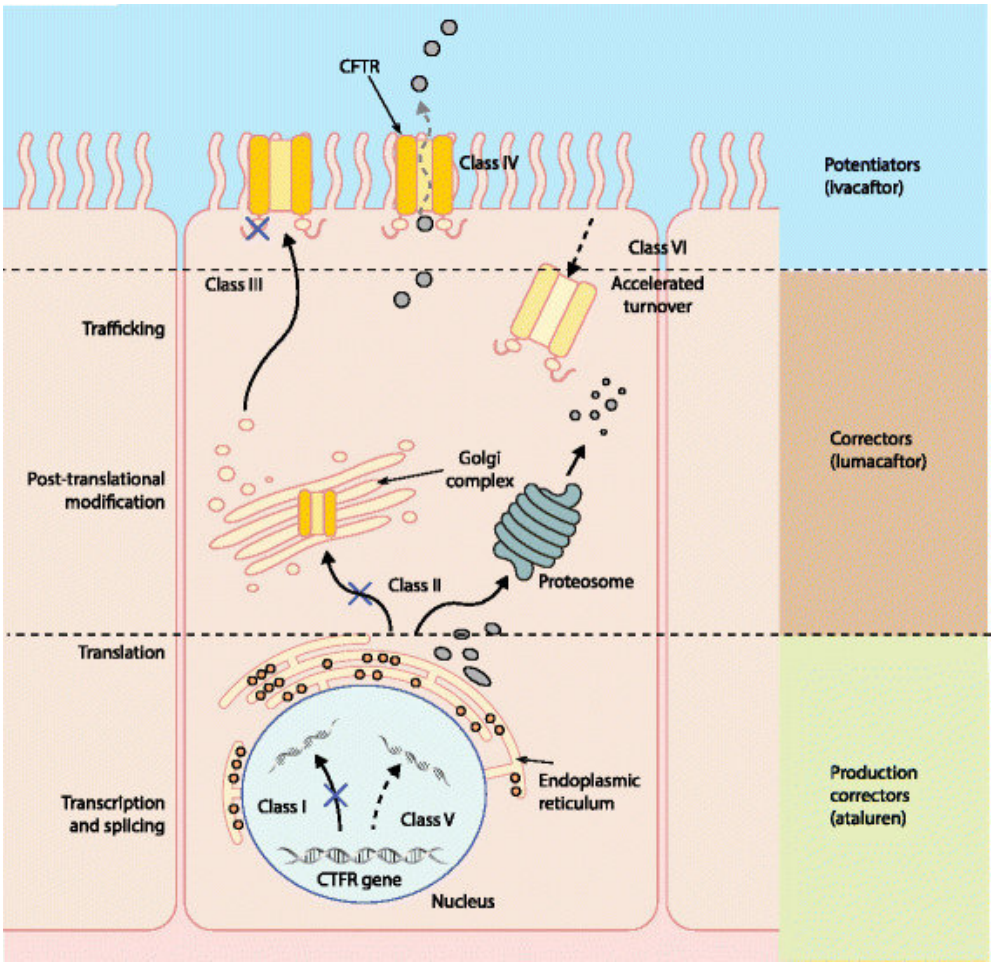


Figure 1.3 – The classes of CFTR mutations and current pharmacologic approaches to restore CFTR function.

Illustration from [Brodlie, 2015] used under CC BY 4.0.

1.3 CF treatments

There is no cure for cystic fibrosis. Treatments for CF have remained mainly symptomatic but some small molecule compounds have recently been developed to improve the processing and activity of mutated CFTR. Although this is not the main focus of this section, we define below the term of *biomarker*, which is closely linked to treatments, before discussing the latter and their effects.

Indeed, the term "biomarker" is derived from "biological marker" and refers to a measurable and reproducible indicator of a patient's medical state [Strimbu, 2010]. Biomarkers are useful to categorize disease severity and thus assess patients heterogeneity, to monitor disease progression and to measure therapeutic efficacy. Biomarkers that show evolution indicating a beneficial effect on patients are particularly valuable and are commonly used to validate treatments.

1.3.1 CF main biomarkers

The primary biological consequence of *CFTR* mutations is a defect in the chloride channel function of the CFTR protein. Therefore, the concentration of Cl^- ions, particularly in the sweat, is a biomarker frequently used to measure the efficacy of a treatment targeting CFTR channel function. Besides, both the American FDA (Food and Drug Administration) and the European Medicines Agency (EMA) recommend measuring the force expiratory volume in 1 second (FEV_1) as a CF biomarker for pulmonary function. FEV_1 is the maximum amount of air a patient can exhale during the first second following maximal inhalation [David, 2023]. FEV_1 decline has been associated with morbidity and mortality among CF patients [Corey, 1997; Liou, 2001].

Other parameters may be important for monitoring disease progression, in particular for assessing the heterogeneity of patient response to treatments. For example, a study by Cornet et al, 2022 showed that the FEV_1 measurement is not sufficient to measure the response to Orkambi of patients aged six to twelve years but that the combination of the FEV_1 measurement and the Body Mass Index (BMI) measurement could be more useful [Cornet, 2022b].

1.3.2 Symptomatic treatments and gene therapy

Historically, treatments have remained symptomatic for a long time, managing all the symptoms. They include antibiotics to fight against chronic infection, or pancreatic enzymes to fill pancreatic deficiencies (reviewed in [Davies, 2007]), and they are still widely used today.

In 1989, the discovery of *CFTR* gene has enabled the development of therapy targeting this gene. However, gene therapy in CF involves challenges such as producing vector in quantities sufficient for treating the entire human lung, developing accurate CF animal models, and identifying airway cell types capable of reversing disease progression once CFTR is expressed [Choi, 2021]. Although there have been some promising results in preclinical studies, no gene therapy drugs for CF have been approved yet, and it is still an area of active research and development.

1.3.3 CFTR modulators

In addition to advances in gene therapy, there have been recent successes in the development of molecules restoring CFTR channel function. Indeed, understanding the structure and function of CFTR has helped to develop and optimize small-molecule compounds designed to restore the processing, maturation and activity of mutant CFTR [Gentzsch, 2018]. These molecules are collectively designated as CFTR *modulators* [Amaral, 2007]. We can categorize them into two classes (See Figure 1.3):

— CFTR *potentiators* increase the activity of mutant CFTR at the cell surface. They have been typically developed for patients with class III mutations, such as the G551D mutation.

— CFTR *correctors* improve defective protein folding and processing to the cell surface, and can be combined with a potentiator.

Since 2012, the Vertex company has developed four of these modulators subsequently approved by the FDA (see [Cornet, 2022a] for a detailed review in French of

these four modulators).

VX-770 (trade name: Ivacaftor) is the only potentiator marketed by Vertex. Clinical trials showed significant decrease of Cl^- concentration in sweat, and improvement in lung function for CF patients with mutations of class III and superior [Ramsey, 2011]. It has been approved for adult patients in 2012 and it has been expanded to patients aged 4 months and older since then.

VX-809 (trade name: Lumacaftor) was the first corrector developed by Vertex. Although results on CFTR maturation and activity were encouraging in primary cultures of patients, no improvement in lung function was reported from clinical trials [Clancy, 2012]. Therefore, for mutations of class II, including F508del, research focused on combotherapies with one or multiple correctors of CFTR folding/processing and one potentiator of CFTR activity at the PM. A combination therapy of VX-770 and VX-809 (trade name: Orkambi) has been developed. A clinical trial on F508del-homozygous patients showed a significant improvement [Wainwright, 2015] whereas conversely a clinical trial on heterozygous patients did not. In 2015, this combotherapy has been approved for 12 years and older F508del-homozygous patients and it has been expanded to 2 years and older patients since 2019 [Konstan, 2017].

VX-661 (trade name: Tezacaftor) is a second-generation corrector developed from VX-809 to stay active longer in the ER, thanks to a better metabolising profile [Donaldson, 2018]. A combination therapy of VX-661 and VX-770 (trade name: Symdeko) has been developed to enhance the beneficial but limited effects of Orkambi. Clinical trials conducted on F508del-homozygous patients did not show better outcomes than those observed with Orkambi. Conversely, clinical trials on heterozygous patients, with one F508del allele and one allele of a class III mutation or superior, showed beneficial effects on both FEV_1 and sweat Cl^- concentration [Donaldson, 2018]. Treatment with Symkevi was approved for these patients.

Finally, VX-445 (trade name: Elexacaftor) is a third-generation corrector. It was developed to act synergistically with other correctors. The triple combination therapy (VX-445/VX-661/VX-770, trade name: Trikafta) enabled significant decrease of sweat Cl^- concentration and significant improvement on FEV_1, at a higher level than all other therapies mentioned above, for patients carrying at least one F508del mutation [Keating, 2018]. Nevertheless the clinical benefit remains heterogeneous, since about 30% of CF patients have less than 5% improvement in FEV_1 [Heijerman, 2019]. In February 2023, clinical benefits were however observed in a subset of CF patients with advanced lung disease and *CFTR* variants not currently approved for CFTR modulators [Burgel, 2023]. This study could expand treatment approval for patients that were not eligible to these therapy until now.

1.3.4 Mechanism of action of CFTR modulators

CFTR modulators were identified by High Throughput phenotypic Screening (HTS) based on improving the chloride channel transport in CF human bronchial cells. Therefore, their molecular mechanisms are not fully understood and their overall mechanism of action (MoA) is partially known.

On the one hand, the maturation of CFTR by VX-770 would be dependent on the state of phosphorylation of the channel, but independent of ATP since it is effective on proteins whose NBD1 site is mutated and in others where the NBD2 site is absent

[Mutyam, 2017]. On the other hand, several studies on the mechanism of action of VX-809 point to different effects: it could stabilize the protein folding by acting on the TMD1 site, on the NBD1 site, or on the TMD2-NBD1 interaction.

Besides, both *in vitro* and *in vivo* studies have suggested potential off-target proteins for CFTR modulators. The improvement in sweat chloride levels during Orkambi treatment did not show a statistically significant association with progression of pulmonary function [Sagel, 2021]. This indicates that the benefit brought by CFTR modulators to pulmonary function would not depend only on improvement of the CFTR channel function. In addition, a study conducted by Rehman [Rehman, 2021] showed a positive correlation between airway inflammation and Trikafata-induced improvement in lung function. This suggests that the potential off-targets are related to inflammatory processes.

Lastly, the response to treatment varies among patients. Several clinical parameters, including FEV_1 expressed in percentage predicted ($ppFEV_1$) and Body Mass Index (BMI) can explain most of the heterogeneity in children's response to Orkambi[Cornet, 2022b]. Specifically, patients with more severe respiratory dysfunction tend to show less improvement.

Understanding the MoA of CFTR modulators is essential to optimize their efficacy, or to define new therapeutic strategies. For instance it has been shown that VX-809 limits the effect of VX-770 [Donaldson, 2018]. A deeper understanding of how each one work could help to tackle this issue.

Among the modulators, Trikafta stands as the most promising treatment with notable outcomes. A deeper understanding of the MoA of elexacaftor seems worthwhile, since its inclusion in the therapy appears to offer the greatest benefits to patients.

1.4 Unravelling CF molecular mechanisms

CF therapeutic efforts mainly focus on targeting the *CFTR* gene or the chloride channel. However, as detailed below, some CF symptoms or CF phenotypes at the cellular level do not seem to depend solely on CFTR function. It is therefore of great interest to understand and unravel the molecular mechanisms of the disease beyond the genetic mutation and the function of the CFTR chloride channel. This understanding is necessary to grasp the variability in patient response and to improve new therapeutic strategies, particularly for patients who are not eligible to CFTR modulators.

1.4.1 Unrelated CF symptoms

The link between the loss of CFTR chloride channel function and the overall CF pathophysiology is not fully understood.

It has been established that CF is characterised by excessive inflammation prior to infection [Bodas, 2010; Nichols, 2015]. This observation cannot be related to CFTR channel function in a direct manner. This is one of the most debated topics related to CF [Khan, 1995; Balough, 1995]: although infection worsen inflammation, the initial causal relationship between infection and inflammation remains to be established. Pig models of CF showed that there is no intrinsic inflammation in non-infected lungs [Stoltz, 2010] although they show blunted early response to pro-inflammatory

environment [Bartlett, 2016]. Conversely, CF ferret models demonstrated that muco-inflammatory processes are present in absence of clinically apparent infections [Rosen, 2018]. Finally, studies on young children bronchioalveolar lavage fluid (BALF) [Esther, 2019] support the idea that inflammation is an intrinsic defect in the immune response, rather than a response to airway infection.

Moreover, other abnormal cellular phenotypes observed in CF do not appear to be related to a defect in Cl^- conductance, such as unbalanced oxidative stress with increased Reactive Oxygen Species (ROS) [Kelly-Aubert, 2011; Jeanson, 2012], impaired epithelial regeneration [Hajj, 2007], proteostatis and autophagy [Bodas, 2019].

1.4.2 Is CFTR just a chloride channel ?

Moreover, recent studies on CF animal models fuel the idea that CFTR is involved in other non-channel functions (see [Hanssens, 2021] for a review). Indeed studies on CFTR -/- knockdown mice [Crites, 2015], CFTR -/- knockdown piglets [Fleurot, 2022] and mutated cell lines in which CFTR is inactivated by the CRISPR/Cas9 technology [Hao, 2020] have reported that the absence of CFTR affects cell signalling and transcriptional regulation, and in particular inflammatory responses.

1.4.3 Patient heterogeneity

Lastly, these symptoms are also found to be very heterogeneous among patients, even among those bearing the same CFTR mutations. The heterogeneity of CF symptoms cannot be explained by *CFTR* mutations, and thus by the defect in the channel function alone. Indeed, several studies showed that CFTR genotypes do not solely correlate with lung function [Wright, 2011; Consortium, 1993].

1.4.4 Main hypothesis of this thesis

The issues raised above suggest the following hypothesis: although CF is a monogenic disease, its symptoms and patients response to treatments are not only caused by the dysfunction of CFTR, but by the dysfunction of a yet unknown network of proteins that functionally interact with CFTR in the cell. This protein network involved in various biological functions may explain how the absence of a functional CFTR leads to perturbations of various biological pathways, leading to an array of CF cellular phenotypes.

Previous research has already studied networks of proteins involved in the processing and maturation of wild-type or mutated CFTR (such as chaperons HSP70 and HSP90) [FARINHA, 2002]. However, our hypothesis goes beyond the processing and proteostasis network of CFTR and relates to the functional protein network to which CFTR belongs once it reaches the membrane, located near other ion channels, membrane receptors and cytoskeleton proteins.

1.5 Project description

The primary focus lies in deciphering the intricate molecular dysregulations of CF caused by the absence of CFTR at the plasma membrane. The hypothesis of this thesis may address the challenges presented in the previous sections. Specifically, it aims at relating the absence of functional CFTR protein to well-known cellular CF phenotypes.

We combined two complementary fields of computational biology: first, we developed systems biology approaches for CF based transcriptomic data, and second, we predicted off-targets for CFTR modulators using machine-learning algorithms.

1.5.1 Systems biology approach to study CF

Systems biology is a modern interdisciplinary field that uses computational methods and mathematical models to unravel the complexity of biological processes. It relies in the premise that biological functions arise from the interaction of multiple cellular components, rather than being attributed to a single one. This paradigm allows for a holistic understanding of the biological processes that goes beyond the isolated study of individual elements. To do this, systems biologists generally represent biological systems as networks, and use tools derived from graph theory and network modelling to analyse them (see chapter 2 for a more detailed introduction of systems biology approaches and notions).

This field is particularly useful for describing complex biological processes. It is therefore valuable for the study of complex diseases, such as cancer. Indeed, cancer is described as a systemic disease involving multiple genetic and environmental factors. It is then necessary to combine knowledge in molecular biology, chemistry, physics, mathematics or/and informatics to understand the complexity of the disease, which makes the systems biology paradigm appropriate for its study.

In this PhD project, we consider CF as a complex disease whose consequences are not limited to the loss of CFTR function. We want to go beyond the CFTR-centric vision that has been the main focus of CF research and study the links between CFTR and the different molecular dysregulations of the CF cell. In this context, it appeared relevant to adopt a systems biology approach to tackle these questions.

Two directions have emerged in systems biology. The first is data-driven: the aim is to collect and analyse large amounts of data to reconstruct a global view of systems in the form of networks and extract biological information from these data. In particular, the high-throughput measurement of transcript or protein data, known as omics data, enables to highlight genes and biological processes that are disrupted in diseases across the whole genome or proteome, without any a priori hypothesis. These data are now being generated and analysed on a massive scale for deriving biological hypotheses. Conversely, the second direction, the model-driven approach, consists of integrating detailed knowledge of the systems' subunits to understand the system as a whole, and translating them into mathematical models. The aim of these models is to understand why certain phenotypes occur in certain biological contexts.

In the context of this project, this second direction is not realistic, as it assumes that all the cellular components and functions relevant for the problem at hand are known. Most of the molecular mechanisms dysregulated in CF have not been extensively stud-

ied, which makes it impossible to build a model exclusively from the literature. For this reason, we have adopted a data-driven approach for this project.

The first step was to retrieve overall CF molecular dysregulations by analysing omics data. We focused on modelling one particular system: the human airway epithelial cells with the F508del mutation. This choice was made because the most severe symptoms of CF are in the lungs, and because this mutation is the most prevalent. We have analysed these omics data adopting a pathway-level approach rather than focusing on individual genes, as interpretation at the gene level is often controversial and not always robust. Finally, we have built a biological network comprising all the molecular dysregulations obtained from a comparative analysis of CF omics data.

1.5.2 Predict CFTR modulators targets

Another original way to understand CF molecular dysregulations is to explore the mechanisms of action of CF treatments and especially CFTR modulators. As presented in section 1.3, these molecules were designed to target CFTR but their MoA have not been yet elucidated. Their overall protein interaction profile in unknown, and therefore, their MoA could involve potential off-targets involved in biological processes that are actually dysregulated in CF. Therefore, we propose to investigate the MoA of these molecules to decipher molecular dysregulations in CF.

Unravelling their MoA means predicting their target profiles, i.e. the panel of proteins with which they directly interact in the cell. We propose to identify these targets using in silico approaches. These methods are an alternative to experimental methods, and have been applied to many problems such as predicting physico-chemical properties or biological information from computational models based on databases built from *in vivo* and *in vitro* tests. Ideally, experimental validation of these predictions are then performed. The goal of this approach is to reduce the number of experiments to be performed to the most probable ones.

Identification of CFTR modulators targets can be formulated as a classification problem in which all (CFTR modulator, protein) couples are predicted as "interacting" or "not interacting". The problem of target identification can thus be tackled in the form of Drug-Target Interaction (DTI) prediction. Several computational methods, such as ligand-based, also known as QSAR, or docking, can be used to solve this problem. We used a supervised machine-learning (ML) chemogenomic algorithms, because unlike the methods cited above, chemogenomics is designed to formulate predictions over the human proteome (see chapter 6 for a brief introduction of chemogenomics basics).

We have developed state-of-the-art ML methods for chemogenomics. These learning algorithms were trained on all drug-protein interactions available in specific DTI databases, and then applied to predict CFTR modulators most probable targets. High scoring predicted targets were finally tested in *in vitro* experiments, in collaboration with the Eurofins company.

1.5.3 Long-term objectives

Beyond providing a deeper understanding of the disease, this project aims at helping to find better therapeutic strategies for CF. First, by looking at the predicted targets of

CF modulators in the biological network, we could improve our understanding of their MoA and propose explanations for the heterogeneity of patient responses. Secondly, the analysis of the network could help us to highlight proteins important for the propagation of the dysregulations and maybe find new therapeutic targets for CF.

For some of these potential targets, inhibitors or drugs might already be available. These drugs could be used in synergy with some CFTR modulators to improve the recovery of patients symptoms. For proteins for which nothing is known for the moment, we could used the chemogenomic algorithm to find available drugs that could target these proteins and use them also in synergy with CFTR modulators.

The position of the project implies that the predicted therapeutic targets, or at least the proteins of interest in the molecular dysregulations, are independent of CFTR. These proteins could therefore be included in therapeutic strategies agnostic to the CFTR mutation and be considered for patients bearing "unrescuable mutations".

Furthermore, CFTR is implicated in other diseases and biological processes than CF: cancer [Zhang, 2013; Xia, 2017; Duan, 2021], COPD [Saint-Criq, 2017], cigarette smoke exposure [Valdivieso, 2018]. Studying the biological network governed by the absence of CFTR might suggest molecular targets beyond the field of CF.

1.6 Contributions

In this introduction, I have highlighted the complexity of the molecular mechanisms of the CF disease. The contributions of this thesis are presented in this manuscript, as separate chapters, each corresponding to an article that is submitted or already published in international scientific journals with reviewing process.

As mentioned above, the fundamentals of systems biology and machine-learning chemogenomics are necessary to understand these articles. These fields are thus introduced in two dedicated chapters: an introduction of the field of computational systems biology is presented in chapter 2 and the chemogenomics basics to predict DTI are presented in chapter 6. I also decided to review the studies on systems biology approaches applied to CF, presented in chapter 4, before introducing our own work. Reading these chapters may be propaedeutic to the comprehension of the articles.

As mentioned above, we started this project by the investigation of systems biology approaches using CF transcriptomic data in part II. The analysis of omics data at the gene level is often controversial and not always robust. There is a need to develop more complex mathematical methods that are more robust statistically and more interpretable. As the biological pathways level is much more interpretable in terms of biological mechanisms, gene-set-based algorithms seem to meet this demand.

The manuscript is organised as follows:
In chapter 3, we present a new and extended version of the rROMA package, an algorithm for fast and accurate computation of the activity of gene sets with coordinated expression. Indeed, identification of gene sets that define biological pathways is an essential step for CF systems biology study. This initial algorithm was developed in 2016 by Martignetti et al. [Martignetti, 2016], and rROMA is widely used in the community of omics data analysis. The improved algorithm enabled the detection of gene sets that

were not detected with the previous version, and provides a clearer classification of the gene sets with a high variance in the data. This methodological work was done in collaboration with Matthieu Cornet, and was supervised by Loredana Martignetti.
In chapter 5, we tackle the main issue of the thesis: how the absence of functional CFTR leads to the overall dysregulations, thereby contributing to some of the deleterious CF phenotypes. We applied pathway-based algorithms to publicly available CF transcriptomic datasets to identify the most frequently dysregulated biological pathways. We adopted a systems biology approach to connect these pathways and thereby defined a CF signalling network. The biological pathways present in the network and their resulting phenotypes were found consistent with today's CF knowledge. The topological analysis of the network highlighted a few proteins that may initiate dysregulations from CFTR into the network, and may explain the observed CF phenotypes. This work was submitted in October 2023.

This initial network model can be refined with other types of omics data such as proteomics data. For instance, we could integrate data from proximity labeling mass spectrometry approaches to provide a more comprehensive picture of CFTR protein partners. In line with this idea, I performed statistical analysis of proximity labeling data of wild type (WT) and mutated CFTR cells, that allows to identify proteins in the proximity of CFTR in the cell (although not necessary in direct interaction). This information could be used in the future to refine our current CF network. This complementary work was published in International Journal of Molecular Sciences in 2022, and is provided in appendix B.

In part III, we investigated the potential off-targets of CFTR modulators using chemogenomics machine-learning algorithms, particularly for elexacaftor, the most recently approved drug. This led us to propose a new algorithm to choose negative examples for training these algorithms, which was published in International Journal of Molecular Sciences in June 2021. Indeed, CFTR modulators were developed to restore the processing of CFTR, but their mechanisms of action are not fully deciphered yet. Identification of potential off-targets would allow a better understanding of the molecular dysregulations of the cell, and propose better therapeutic strategies with drugs designed specifically for these targets. Importantly, it would be very interesting to identify whether these off-targets belong to the proposed CF network, and whether they are common to the list of target candidates proposed based on the systems biology study presented in chapter 5.

This work on target identification is not completed yet, as these approaches need to be further optimized, particularly to improve predictions for interaction with proteins for which very few, or even no ligands are known. However, in collaboration with Philippe Pinel and Gwenn Guichaoua, we showed that the proposed algorithm displays state-of-the-art performances to solve scaffold-hopping problems. This led to a publication in the journal Molecular Informatics in 2023, provided in Appendix D.

Overall, the results obtained in this thesis underline that CF is a complex disease for which a systems biology can contribute to a better understanding while providing suggestions for new experimental work and and therapeutic strategies. However, modelling such a complex and heterogeneous disease in one single model appears somewhat

reductive and simplistic. Therefore, in chapter 8, after summarising the results of this thesis, I discuss evidence of the disease heterogeneity. I propose some perspectives to extend the analyses performed for the F508del mutation to the study of other mutations in CFTR, to investigate the heterogeneity between patients bearing the same mutation, and to investigate the heterogeneity between the different cell types of the airway epithelial cells. Indeed, it is important to address these three topics, keeping in mind the long-term aim of improving therapeutic solutions for CF.

Part II

Systems biology approaches to study CF

Chapter 2

Computational systems biology to study diseases

Contents

Abstract

Systems biology approaches combine computational and mathematical methods with biological data in order to construct biological models, and to decipher the complexity of biological processes. Omics data refers to large-scale datasets that capture biological information at various molecular levels. Therefore, they are commonly used in computational systems biology to understand biological mechanisms that determine phenotypes. This chapter provides an overview of computational systems biology approaches and explain how omics data analysis can be used to understand complex systems.

Résumé

Les approches de la biologie des systèmes combinent des méthodes informatiques et mathématiques avec des données biologiques afin de construire des modèles biologiques et de déchiffrer la complexité des processus biologiques. Les données omiques sont des ensembles de données à grande échelle qui contiennent des informations biologiques à différents niveaux moléculaires. Elles sont donc couramment utilisées en biologie systémique computationnelle pour comprendre les mécanismes biologiques qui déterminent les phénotypes. Ce chapitre donne un aperçu des approches de la biologie systémique computationnelle et explique comment l'analyse des données omiques peut être utilisée pour comprendre des systèmes complexes.

2.1 Introduction to systems biology

Although molecular biology has revealed a multitude of information regarding genome sequences and protein properties, it alone does not provide a complete understanding of biological systems [Kitano, 2002]. Indeed biological functions can rarely be attributed to a single cellular component, but rather to intricate interactions between multiple components. As our understanding of biological systems becomes more complex, and thus less intuitive, we need to adopt a systemic and integrative view that encompasses the components involved in cellular processes. This is the main aim of systems biology approaches, that combine computational and mathematical methods with biological data to construct biological models, and to decipher the complexity of biological processes. These models do not aim to replace biology or explain the biology better than experiments, but rather they can accompany biologists in their understanding and interpretations of their experimental results.

Computational systems biology approaches allow to build two types of models: static models and dynamical models. The first ones aim at describing biological processes in the form of maps or networks from knowledge or/and data mining. These networks can be used as tools to hypothesise protein functions or discover mechanisms in diseases. The second ones aim at reproducing experimental observations by simulating hypotheses dynamically. They allow to predict systems response to external factors (cellular context, micro-environment, genetic alterations, etc.) such as side effects, and suggest or optimize therapeutic solutions.

Static biological modelling comes first, before any dynamical modelling. Therefore, we built a static biological network of the CF dysregulations. It could be late be used to simulate CF cells response to treatments, whether they are CFTR modulators or any inhibition of a newly discovered target. The study of CF dynamical models is beyond the scope of this thesis. Therefore this section will be dedicated to the concepts underlying static biological models, and the dynamical models will not be addressed.

2.1.1 From protein interactions to intracellular biological networks

Biological functions do not usually stem from a single cellular component, but rather from interactions among multiple ones (genes, proteins, small molecules, enzymes, etc.). Indeed, cells process information by physical interactions of molecules. Proteins interact in intra- and inter-cellular signalling, in transcriptional and post-transcriptional regulation, and sometimes in complexes. Databases have been collecting information about protein and gene interactions: STRING DB [Szklarczyk, 2019], the Human Reference Interactome [Luck, 2020], BioGrid [Oughtred, 2021] , IntAct [Toro, 2022]. These databases usually store *undirected interactions*, i.e. they compile protein-protein interactions (PPI), without information on causality or consequence of this interaction.

Several knowledge bases have collected causal interactions by incorporating directionality information, called hereafter "directed interactions" [Csabai, 2022; Lo Surdo, 2022]. A "causal interaction" refers to an upstream component (also called as *source*) exerting a regulatory effect on a downstream component (also called as *target*). These causal interactions serve as fundamental building blocks for understanding the flow of information in biological processes within cells.

The Saez-Rodriguez team at Heidelberg combined more than 100 resources in the OmniPath database [Türei, 2016] covering undirected and directed interactions, that are accessible via a website and packages in R/Bioconcudctor and Python (https://omnipathdb.org/).

A sequence of directed interactions is known as a *cascade of interactions*: a target node (a protein) of one causal interaction becomes the source node of a subsequent causal interaction. When these cascades are associated with biological processes, they are commonly referred to as *biological pathways*. Biologists sketch these biological pathways as graphs or maps to illustrate the biological processes. *Signal transduction pathway*, and *metabolic pathway* are the most common types of biological pathways.

Finally, proteins and genes often belong to multiple biological pathways. It is not even rare to observe an interaction belonging to multiple pathways. For example the activation of AKT by PI3KCA is the core interaction of the PI3K-AKT signalling pathway but is also involved in the regulation of the actin cytoskeleton, according to the KEGG pathway database [Kanehisa, 2012]. When we need to model a complex cell phenotype from a particular disease, or a rare cell type differentiation, multiple biological pathways are often involved. Biologist need to sketch a *biological network*, where biological pathways are intertwined. Such a network can be called *signalling transduction network*, if it consists of signalling transduction pathways, i.e. proteins interact with biochemical reactions or *metabolic network*, if it consists of metabolic pathways.

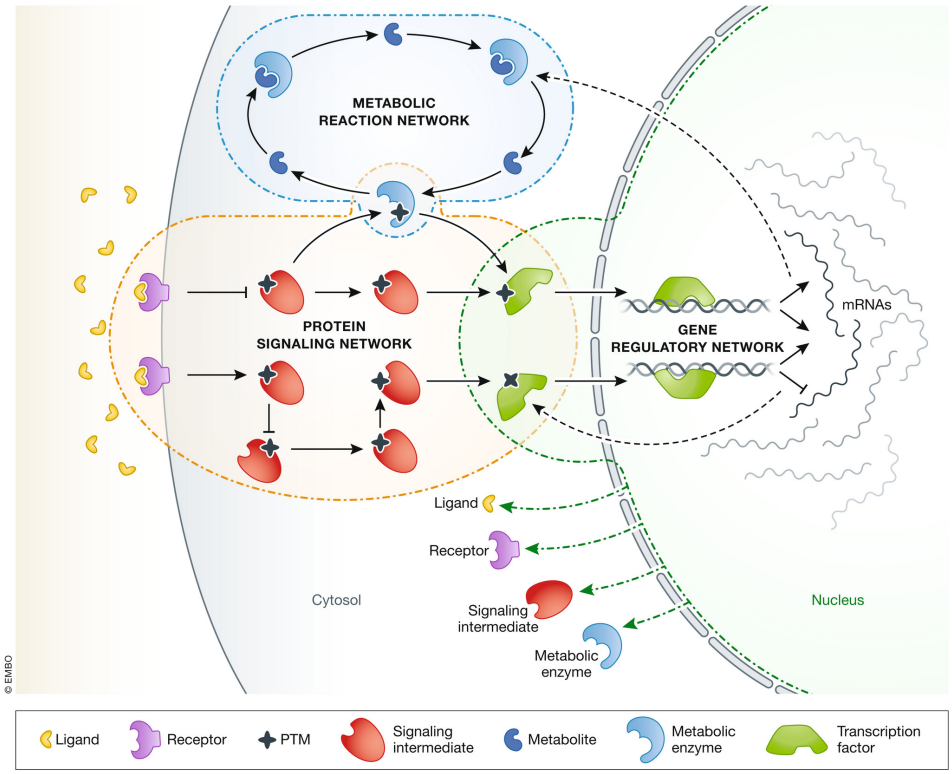


Figure 2.1 – The three major types of intracellular biological networks: signalling networks, metabolic networks, and gene regulatory networks.
Illustration from [Garrido-Rodriguez, 2022] used under CC BY 3.0.

2.1.2 Other biological networks

Biological networks can be used to represent biological pathways, but also to represent any graphs whose nodes and edges carry biological information.

For example, biological networks can represent different types of biological information:

— *Gene regulatory networks*: nodes represent genes and edges denote functional influences between genes.

— *Genetic interaction networks*: nodes correspond to genes and edges to functional relationships between these genes, either through physical interactions, or to account for the synergic effects of combined alterations.

— *Co-expression networks*: is an undirected network, where nodes correspond to genes, and edges account for significant co-expression relationship between them.

— *Protein-protein interaction networks*: is an undirected network, where nodes represent proteins and edges physical interactions between proteins.

— *Disease networks*: is an undirected network, where nodes account for diseases or disease genes, and edges for an association between two genes, or between a disease and a gene.

2.1.3 Network biology applied to complex diseases

Alberto-Lazlo Barabasi have inspired research on biological networks and their analysis, also known as *network biology*, demonstrating their potential for knowledge extraction and identification of new candidate targets genes [Barabasi, 2004]. Indeed network biology approaches have been applied since then to study complex diseases, involving multiple genetic and environmental factors, such as cancer [Barillot, 2012]. Signalling networks have been particularly studied to decipher cancer cellular dysregulations, knowing that cancer cells hijack preexisting molecular processes observed in normal tissues to achieve their phenotypes, for instance cell proliferation processes.

Today, systems biology focuses on using these networks to understand disease complex mechanisms and optimize treatments. One of the great challenges in network biology is currently to support decision in what we called *personalized medicine* or *precision medicine*. The latter is defined by the National Cancer Institute as *a form of medicine that uses information about a person's genes, proteins, and environment to prevent, diagnose, and treat disease*. Genetic backgrounds are now commonly used in clinical trials, but mathematical models on molecular profiles could also be included to guide the choice of treatments.

Precision medicine have been applied so far to complex diseases where patient heterogeneity is evident. However, patient heterogeneity have also been observed in genetic diseases where same causal genes can lead to different phenotypes, which also motivates our interest for systems biology and network biology to study the monogenic disease CF.

There are two ways to build biological networks. The first strategy is to build the network from the literature, gathering a list of proteins or biological pathways known to

be involved in the disease, and their interactions, according to published experimental results. The second type of models is to infer this list of proteins (or pathways) from data related to the disease, in order to avoid potential bias in the literature, or in the interpretation of published experimental results. Data-driven approaches are usually based on omics studies, because the number of available omics studies is increasing rapidly. In the next section, we present the types of omics experiments that can be used in data-driven systems biology approaches.

2.2 Omics data for systems biology

Omics data refers to large-scale datasets that capture biological information at various molecular levels: DNA, RNA, proteins or small molecules like metabolites. Monitoring tens of thousands of cellular components provides a global vision of the biological systems. Therefore, they are commonly used in computational systems biology to understand biological mechanisms that determine phenotypes.

In this section we will show how omics data analysis can be used to understand complex systems.

2.2.1 Omics data

Omics techniques

Omics techniques refer generally to high throughput sequencing techniques or gene/protein profiling techniques that generate omics data. They can be grouped into five categories:

— *Genomics* studies the genome of organisms, i.e. the complete DNA sequences comprising genes, regulatory elements and non-coding regions. Genomics data are analysed to study genetic variations, heredity and evolutionary relationships.

— *Transcriptomics* studies RNA molecules transcribed from the genome, also known as *the transcriptome*. Types and abundances of RNA molecules are monitored to identify which genes are expressed, and how they are regulated. Transcriptomic data are analysed to examine gene expression patterns and infer biological activities.

— *Proteomics* studies proteins detected in specific conditions, such as specific cell type, tissue or organism. Proteomic data, generated by liquid chromatography coupled to mass spectrometry (LC-MS), are analysed to provide insights about protein functions, post-translational modifications and proteins interactions in the cell. *Phospho-proteomics* is an increasingly studied branch of proteomics, focusing on the phosphorylation status of proteins.

— *Metabolomics* focuses on small molecules, known as metabolites, that can be viewed as specific fingerprints of cellular biochemistry. It quantifies a wide range of low-molecular weight metabolites, such as amino acids, sugars, fatty acids, lipids and steroids.

— *Epigenomics* studies the so-called epigenetic modifications, such as DNA methylation or histone modification, which affect gene expression without altering the DNA sequence. Epigenomic data are the most recent types of omics data, and state as an emerging field in the recent years.

With the recent technological advances in the different omics fields, the generated data can be available at the level of individual cells. The study of biological systems at the single-cell level is referred to as *single-cell omics*, as opposed to *bulk omics*. This emerging field offers new biological insights, such as development, cellular heterogeneity and gene expression dynamics, that could not be studied with bulk data. Additionally, omics data can be overlayed onto tissue images using *spatial omics* technologies. These approaches provide information on the spatial distribution of cellular populations within the tissue of origin, their proximity to one another and with other tissue components.

Furthermore, it is becoming increasingly clear that addressing research questions with a single type of omics data is incomplete. Scientists are now combining different omics types through integrative approaches, commonly known as *multi-omics* approaches [Hasin, 2017]. *Single-cell multi-omics* data are even starting to be published and *spatial multi-omics* will surely soon follow.

2.2.2 Omics data analysis

Treating omics data is complex and requires a computational workflow including quality-control, pre-processing, analysing and statistical steps. Recently machine-learning techniques are even replacing the statistical methods, due to the growing size of data generated by high-throughput technologies. One of the main challenges in analysing omics data is to analyse the outputs of the statistical methods into a interpretable, useful and last but not least, relevant mechanistic insights [Yamada, 2021].

In this manuscript, we do not deal with the quality control and pre-processing steps, but we are directly interested in the analysis of the already preprocessed omics data. We focus also on proteomics data and transcriptomics data, as they are the natural choice for describing protein activity [Szalai, 2020]. In this section, we use gene expression data as example data to discuss the approaches, although the practical guidelines presented below are equally applicable to any type of data generated by omics technologies.

Gene-level analysis

The vast majority of the biological and clinical applications compare samples from different conditions, such as disease and healthy (control) samples. The first approach consists in comparing the conditions at the individual level: statistical tests are done at the single gene or protein level. The output yields in a list of differentially expressed genes (DEG) and the interpretation lies in the identification of biomarkers (gene or protein indicating a particular disease state) or driver genes causally linked to a biological process [Barillot, 2012]. This level of analysis often fails to provide meaningful biological understanding. Indeed, in various systemic diseases, the disruption of a signalling pathway can arise from different genes within the same pathway, and these gene alterations may vary from one patient to another.

Pathway-level analysis

In systemic diseases such as cancer, it has become apparent that the molecular profiles of patient samples are more similar at the pathway level than at the individual

gene level [Wang, 2010]. This observation led to the development of pathway-based (PB) approaches in the analysis of omics data, to capture biological information undetectable by the analysis at the individual genes level. By PB approaches, we mean computational methods that combine omics data and pathway knowledge from public databases, in order to identify pathways altered in one condition compared to the other. There are generally two outputs of these methods depending on the kind: it can be the activities of the pathways for each sample of the dataset, or if the method does not enable the analysis at the sample level, the output is a list of differentially expressed pathways (DEPs). These approaches have two advantages: on the one hand, they make it possible to reduce the dimension of the problem to the number of DEPs (rather than the number of DEGs), and therefore the complexity of the system; on the other hand, they have lead to more biologically interpretable results than a list of DEGs.

It is important to distinguish pathway 'mapping'-based methods and 'footprint'-based methods [Szalai, 2020]. These latter do not infer pathway activity from the omics measurement of the pathway members but from those of the genes regulated by the pathway of interest. Both kinds of methods are used on gene expression data and produced significant results. We will refer to both indiscriminately as PB approaches. The aim of this part is not to do a review of all the PB methods but give an idea of their evolution and the computational foundations.

The first generation of approaches are grouped under the name of *over-representation analysis* (ORA). They all remain on the following principle: statistically evaluate the fraction of genes belonging to a pathway that is found among the set of genes showing changes in expression. All the methods follow the following steps: first, compute a gene-level statistics using omics measurements, and then define a list of over-expressed or under-expressed genes according to a given threshold on the statistic measure. Next, for each pathway tested, genes belonging to the pathway are counted inside the list of DEG. Finally, over-representation of the pathway is tested in the list of DEGs thanks to a statistical test. ORA is the approach most widely used by the biologists to identify DEPs, but this method has significant limitations: it only considers the differentially expressed genes (DEGs), and does not take the other genes into account; it only considers the number of genes belonging to the tested pathways that belong to the DEGs, but does not use their level in expression change (e.g. the fold change); finally ORA considers genes as being independent, and does not take into account the interactions between these genes [Khatri, 2012].

Systems biologists developed the second generation of PB methods grouped under the term of *Functional Class Scoring* (FCS) methods to avoid the use of the arbitrary statistical thresholds. A pathway-level statistics is computed from the gene-level statistic using all available molecular measurements of the omics data (e.g. sum, mean of gene-level statistic). The last step is to assess the significance of the pathway-level statistic. The most used FCS method is Gene Set Enrichment Analysis (GSEA) which defines a metric based on the ranks of differential expression of the genes belonging to the tested pathway [Subramanian, 2005]. FCS methods have a main limitation: many of them use changes in gene expression to rank genes, but do not use the fold change. Moreover, they do not address the third limitation of ORA related to potential gene interactions [Khatri, 2012].

The third generation of PB methods overcome this limitation by including knowl-

edge of gene interactions in the computation of the pathway-level statistics. The other steps are identical to those used by the FCS methods. Nevertheless, these third generation methods are still less widely used than the second generation PB methods, although various methods have been developed in recent years (for instance DEGgraph [Jacob, 2012], CLIPPER [Martini, 2013] or the DEAP methods [Haynes, 2013]).

Towards building biological network

Despite their limitations, PB approaches are effective in detecting disrupted biological processes, and are therefore often used to build network representing biological dysregulations. For example, DEGs can be connected together into one biological network based on interactions databases, or DEPs can be merged thanks to their overlapping elements.

In fact, these methods analyse gene sets/pathways independently but this paradigm can be widely discussed. Indeed, many genes/proteins belong to several pathways, and the latter are strongly intertwined. As a result, if one pathway is impacted, other pathways may also be significantly dysregulated because they share common genes [Szalai, 2020].

In the last five years, many computational methods have been developed to infer directly biological networks to tackle this issue. They generally combine a high throughput omics datasets and one large biological network, called *prior-knowledge network* (PKN) or several layers of PKNs. This field is very recent and still less established than ORA approaches and PB methods. One of the first review of these methods was published in 2022 [Garrido-Rodriguez, 2022]. The review classified the methods based on different characteristics such as the omics data properties, the PKN properties and the computational methods used to infer the subnetworks. The review highlighted the fact that these methods do not use the same mathematical formalisms or the same vocabulary, which makes difficult to compare them. It is, however, a very promising area of research that will improve our understanding of biological systems.

Prior knowledge collections

All the presented methods are highly dependent on a reference database that gathers prior biological knowledge such as PPIs or pathways.

Gene set collections

Several public databases store and update the gene-set knowledge as gene signatures: The Gene Ontology (GO) Resource [Ashburner, 2000; Garcia-Alonso, 2019], the Hallmark gene sets of the Molecular Signatures Database (MSigDB) [Liberzon, 2015], the Pathway Interaction Databse (PID) [Schaefer, 2009], or DoRothEA [Garcia-Alonso, 2019]. The gene sets can correspond to genes sharing the same functional annotations or regulatory motifs, genes belonging to the same pathway or genes forming a group of frequently co-expressed genes. Besides, these databases were built for various purposes: for example GO database focus on molecular functions or cellular components, MSigDB Hallmarks focus on cancer biological processes, and DoRothEA on transcription factors (TF) and their transcriptional targets.

Many studies combine ORA and FCS methods with these gene set databases. The term *gene set* is used here instead of *pathway* because the links (interactions) connecting the genes belonging to the same set are not known, at least for the majority of these collections. These gene sets generally result from studies exploring gene expression data or from various biological knowledge. The number of available gene sets is increasing due to the increasing amounts of quantitative data produced by high throughput techniques, providing researchers with a wider range of biological processes.

However, not all sets (or signatures) in these collections are equally informative: a lot of signatures are redundant, and the number of gene sets representing the same biological process is not balanced [Cantini, 2017]. Besides, in these collections, gene or protein members of these sets have different cellular functions and gene/protein interactions are not mentioned (See figure 2.2 for the representation of the notion of *pathway*, *gene set* and *tf regulons*). This leads to a lack of interpretability, and sometimes prevents from comprehensive biological analysis.

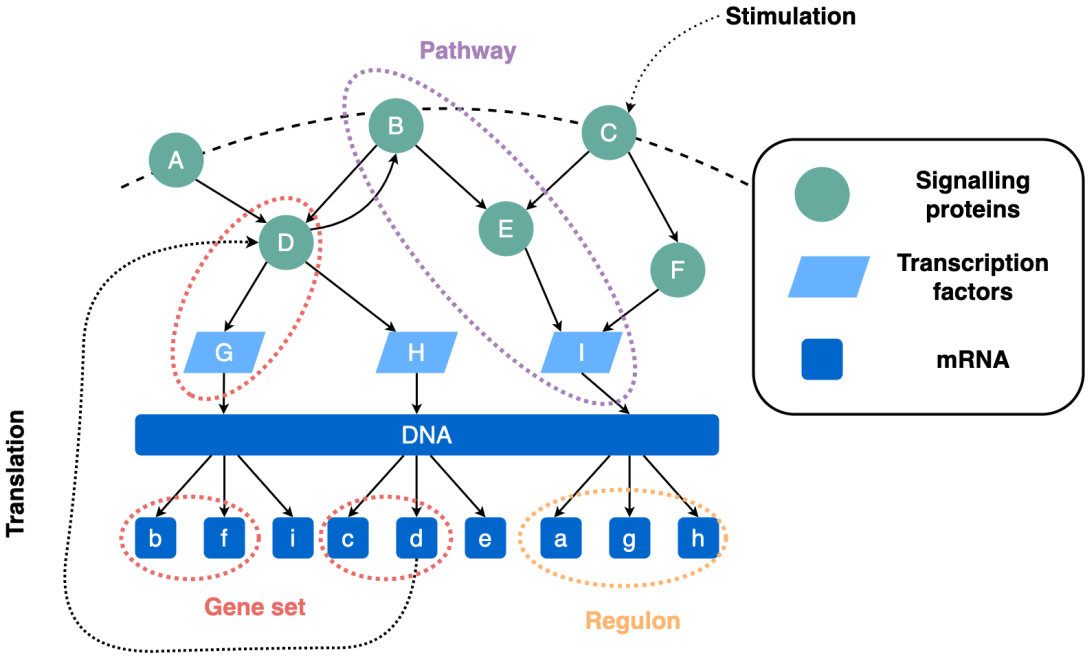


Figure 2.2 – Concept of pathways, gene sets and TF regulons.
Illustration adapted from [Szalai, 2020] used under CC BY 4.0.

Pathways collections

KEGG [Kanehisa, 2012], Reactome [Gillespie, 2022] and WikiPathways [Martens, 2021] are well-known repositories that store biological pathways as graphs. These databases are more informative than gene sets databases because they also store interactions between the components. Although each database has its own representation of interaction information, an important community effort has been done to standardise the way these graphs are drawn (Systems Biology Graphical Notation, also known as SBGN [Novère, 2009]). SBGN defines three languages for three types of diagram: *Process Description* corresponding to biochemical reactions, *Entity Relationship* which shows all the relationships (interactions, regulations etc...) in which a given entity is in-

volved, and finally *Activity Flow* which describes influences of one entity (gene, protein etc...) on another one.

These representations still present some limitations. First, they may present bias towards specific proteins and diseases [Garrido-Rodriguez, 2022] that have been more extensively studied than others. It is unclear whether this is a bias in knowledge, or correspond to intrinsic characteristics of these proteins or diseases. Second, they are used to infer biological properties in various contexts, whereas they gather information that was collected from biological experiments carried out in very specific conditions (tissue, cell type). Condition- or cell-type- specific pathways knowledge networks are not yet available and are much needed. In addition to these two major limitations, these databases are still under construction and lack important information. For instance, they only indicate which genes/proteins are activated or inhibited but give no information on the transcripts that are concerned. They are also incomplete regarding pseudo-genes, i.e. non-coding genes which influence the mechanisms. These are important limitations for their use to model biological processes, which reflects the fact that most biological processes are far from been understood, and the scientific community still needs to put in considerable effort to correctly represent them.

Chapter 3

Representation and quantification Of Module Activity from omics data with rROMA

Contents

Abstract

The efficiency of analyzing high-throughput data in systems biology has been demonstrated in numerous studies, where molecular data, such as transcriptomics and proteomics, offers great opportunities for understanding the complexity of biological processes. One important aspect of data analysis in systems biology is the shift from a reductionist approach that focuses on individual components to a more integrative perspective that considers the system as a whole, where the emphasis shifts from differential expression of individual genes to determining the activity of gene sets. In this context, identification of gene sets with coordinated expression and involved in the same biological pathway or function in the system under study play a key role for biological interpretation of omics data. Here, we present the rROMA software package for fast and accurate computation of the activity of gene sets with coordinated expression. The rROMA package incorporates significant improvements with respect to the initial version of the algorithm, along with the implementation of several functions for statistical analysis and visualizing results. These improvements greatly expand the package's capabilities and offer valuable tools for data analysis and interpretation. It is an open-source package available on github at: www.github.com/sysbio-curie/rRoma. Based on publicly available transcriptomic datasets, we applied rROMA to cystic fibrosis, highlighting biological mechanisms potentially involved in the establishment and progression of the disease and the associated genes. The results notably identified a significant mechanism relevant to cystic fibrosis, raised awareness of a possible bias related to the medium used for cell culture, and uncovered an intriguing gene that warrants further investigation.

Résumé

L'analyse de données à haut débit en biologie des systèmes a été démontrée comme efficace dans de nombreuses études. En effet les données moléculaires, telles que les données transcriptomiques et protéomiques, offrent de grandes possibilités pour comprendre la complexité des processus biologiques. Un aspect important de l'analyse de ces types de données est le passage d'une approche réductionniste, qui se concentre sur les composants individuels, à une perspective plus intégrative considérant le système dans son ensemble. L'accent n'est plus sur l'expression différentielles des gènes individuels mais sur la mesure de l'activité d'ensemble de gènes. Nous présentons ici le package R rROMA pour le calcul rapide et précis de l'activité des ensembles de gènes dont l'expression est coordonnée. L'outil rROMA intègre des améliorations significatives dans l'algorithme de calcul, ainsi que l'implémentation de plusieurs fonctions d'analyse statistique et de visualisation des résultats. Ces ajouts élargissent considérablement les capacités du logiciel et offrent des outils précieux pour l'analyse et l'interprétation des données. Le package est libre de droit, disponible sur github à l'adresse suivant : www.github.com/sysbio-curie/rRoma. Nous avons appliqué rROMA à des données transcriptomiques publiques sur la mucoviscidose et avons mis en évidence les mécanismes biologiques potentiellement impliqués dans l'établissement et la progression de la maladie ainsi que les gènes impliqués. Les résultats ont notamment permis d'identifier un mécanisme important lié à la mucoviscidose, d'attirer l'attention sur un éventuel biais lié au milieu de culture cellulaire, et de découvrir un gène intriguant qui mérite d'être étudié plus en détail.

3.1 Preface

Research projects in CF have traditionally focused on improving CFTR processing to the PM or its Cl^- channel function. Nevertheless many other dysfunctional biological functions have been identified in specific studies (compiled in the review of Ideozu et al. [Ideozu, 2019]). They mainly belong to the signal transduction system, including those related to the inflammatory system or to the immune system. By analysing the molecular dysregulations at the pathway level, we can gain a better understanding of the molecular basis of some of the CF disease phenotypes.

To address this issue, we have adopted a data-driven approach rather than relying solely on a literature review of the pathways dysregulated in CF. We believe that this methodology mitigates potential biases present in the scientific literature.

As outlined in chapter 2.2.2, omics technologies provide a comprehensive view of protein and transcript levels across the entire genome (or proteome). They are therefore interesting for inferring molecular dysregulations without focusing on a specific dysregulations or part of the cell. Hence, they allow a global view of the biological systems by giving information of thousands of components in the cell. While gene-level analysis is often not sufficient to provide comprehensive biological interpretation, pathway-level analysis is a more informative approach. We therefore investigated the CF omics data quantitatively at the pathway level.

As part of my PhD, I worked on a pathway-based algorithm, called ROMA (Representation and quantification Of Module Activity). ROMA was designed to quantify the activity of gene sets characterized by coordinated gene expression. Pathways activities are measured by computing the largest amount of one-dimensional variance across samples explained by the genes in the gene set.

The research paper presented in this thesis introduces the R package *rROMA*, an implementation of the algorithm in R programming language. It includes significant improvements in the computational algorithm, along with several functions for statistical analysis and graphical visualisation of the results. In particular, the new version of the algorithm allows the detection of shifted sets of the genes.

The algorithm was first developed in Java by Loredana Martignetti in 2016 [Martignetti, 2016], and then translated into R by Luca Albergante in 2018. I implemented the functionalities and improvements in the new version, in collaboration with Matthieu Cornet and under the supervision of Loredana Martignetti. Loredana and I gave a tutorial on rROMA, internal to Curie, in June 2022 and I presented the new version of the algorithm in a contributed talk in the "Logical modelling for quantitative data" session of the workshop *Statistical Methods for Post Genomic Data (SMPGD)* in Ghent, Belgium in February, 2023.

This work was made in collaboration with Matthieu Cornet, Luca Albergante, Andrei Zinovyev, Isabelle Sermet-Gaudelus, Véronique Stoven, Laurence Calzone and Loredana Martignetti. It was submitted to the scientific journal *npj Systems Biology and Application* in May 2023 and is currently under review. In the following section, the article is transcribed as currently under review, and followed by a broader discussion in the context of the PhD project.

3.2 Representation and quantification Of Module Activity from omics data with rROMA

3.2.1 Introduction

The use of high-throughput molecular techniques, such as transcriptomics and proteomics, is becoming easier with the improvement of data acquisition tools, leading to a drastic decrease in the costs associated with such analyses. This allows for precise measurement of the molecular profiles of biological systems at several levels. However, the amount of data produced during such experiments is very important, and requires the use of dedicated software and algorithms to analyze them. Moreover, the ability to interpret the data in terms of biological processes becomes a crucial issue. Dedicated analyses are needed to synthesize and transform the data into comprehensive biological information [Hawkins, 2010].

A commonly employed approach in genomics involves comparing measurements at the individual gene or protein level, to identify distinctive markers indicative of specific disease states (biomarkers), or genes that play a causal role in the studied disease [Zinovyev, 2012].

Nonetheless, in numerous systemic diseases, the disruption of a signalling pathway can arise from distinct genes within that pathway, and these gene alterations may vary from one patient to another. For example, in cancer It has become apparent in recent years that the same pathways are affected by defects in different genes and that the molecular profiles of patient samples are more similar at the pathway level than at the individual gene level [Wang, 2010].

Therefore, quantification of gene set activity from transcriptomic or proteomic measurements is now widely used to transform gene-level data into associated sets of genes representing biological processes [Levine, 2006; Ramos-Rodriguez, 2012; Borisov, 2014]. By employing gene set-based approaches in the analysis of omics data, it becomes possible to capture valuable biological insights that would otherwise remain undetectable when solely focusing on individual genes. In this study, we developed an algorithm, implemented as an R package called rROMA, which was designed to quantify the activity of sets of genes defined by their participation in a common functional role. These gene sets can thus correspond to genes with the same functional activities, genes regulated by the same motifs, genes belonging to the same signalling pathway, or genes forming a group of frequently co-expressed genes. The underlying hypothesis of rROMA is to assess the activity of a gene set by determining the maximum amount of one-dimensional variance, which is represented by the first principal component (PC1) derived from the genes within the set. This quantity is considered to be proportional to the influence of a single latent factor on the gene expression within the gene set, and reflects the variability of this factor's activity across the studied samples. This setting corresponds to the uni-factor linear model of gene expression regulation [Schreiber, 2007].

Naive quantification of the activity of a gene set often consists in calculating the mean or median of the expression of the gene set in each sample. Alternatively, it may rely on a single marker gene that represents the overall activity of the gene set. In contrast, rROMA differs from these approaches by its ability to effectively model

scenarii where individual genes within the set do not contribute equally to its activity. This is particularly relevant when some genes have a more significant impact than others, such as transcription factors downstream of signalling pathways. Furthermore, rROMA is well-suited to cases where some genes exhibit a correlation of opposite sign with respect to the overall activity of the gene set. In such situations, the first component of rROMA can capture this effect, whereas a simple averaging approach would not provide effective results.

Other more complex gene set quantification methods have already been proposed to compute the activity of a gene set, by calculating the first principal component of the expression matrix restricted to the genes in the gene set [Tomfohr, 2005]. In the study by Bild et al. [Bild, 2006], a similar strategy was exploited to define activity of several cancer-related pathways on a large collection of human cancer transcriptomes. In another study by Fan et al. [Fan, 2016], the authors suggested the notion of an overdispersed pathway in the context of single-cell transcriptomic analysis. Other methods have been developed to estimate the activity scores of gene sets in individual samples, such as the extension to a single sample of GSEA (ssGSEA) [Barbie, 2009] or OncoFinder [Borisov, 2014]. Our algorithm expands the repertoire of existing methods by introducing unique functionalities that are increasingly relevant in various contexts of systems biology. One distinguishing feature of the rROMA algorithm is that it computes a p-value that denotes the significance of the gene set's activity. This reflects the probability of obtaining the observed activity for a specific gene set by chance. The rROMA algorithm uses a random gene set procedure to generate a null distribution for the L1 amount of variance explained by the PC1, and calculates the p-value by comparing the observed L1 to the null distribution. Usually, a p-value threshold of 0.05 is employed to determine the significance of the gene set's activity.

In addition, the algorithm estimates the statistical significance of the distribution of samples along the first component for a gene set in two ways: it distinguishes between shifted and over-dispersed gene sets. A shifted set of genes corresponds to the situation where the median expression of all the genes in the gene set is significantly different from the one of all the genes studied, i.e. that the gene set shows a particularly high expression in at least one sample (see Figure 3.1 A). Over-dispersion of a gene set corresponds to the situation where the amount of variance explained by PC1 calculated for only the genes in that gene set is significantly greater than the variance of a randomly selected set of genes of the same size. Thus, overdispersion means greater variability in a set of genes among the considered samples (see Figure 3.1 B). The fact that rROMA distinguishes these two situations is particularly useful because, in many cases, the activity of a gene set does not correspond to overdispersion of the module in the global gene expression space, but to a shift of the genes in a particular direction. Therefore, analysis of shifted gene sets can highlight findings that would not be identified with overdispersion analysis alone.

Importantly, the algorithm provides the activity level of the gene set for each individual sample, and does not require a predefined labeled classification of the samples into various conditions or groups. These activity values can be subsequently compared, bringing to light the heterogeneity present in the dataset in relation with the analyzed gene sets. This may be useful to define groups of samples or patients, when such stratification is unknown.

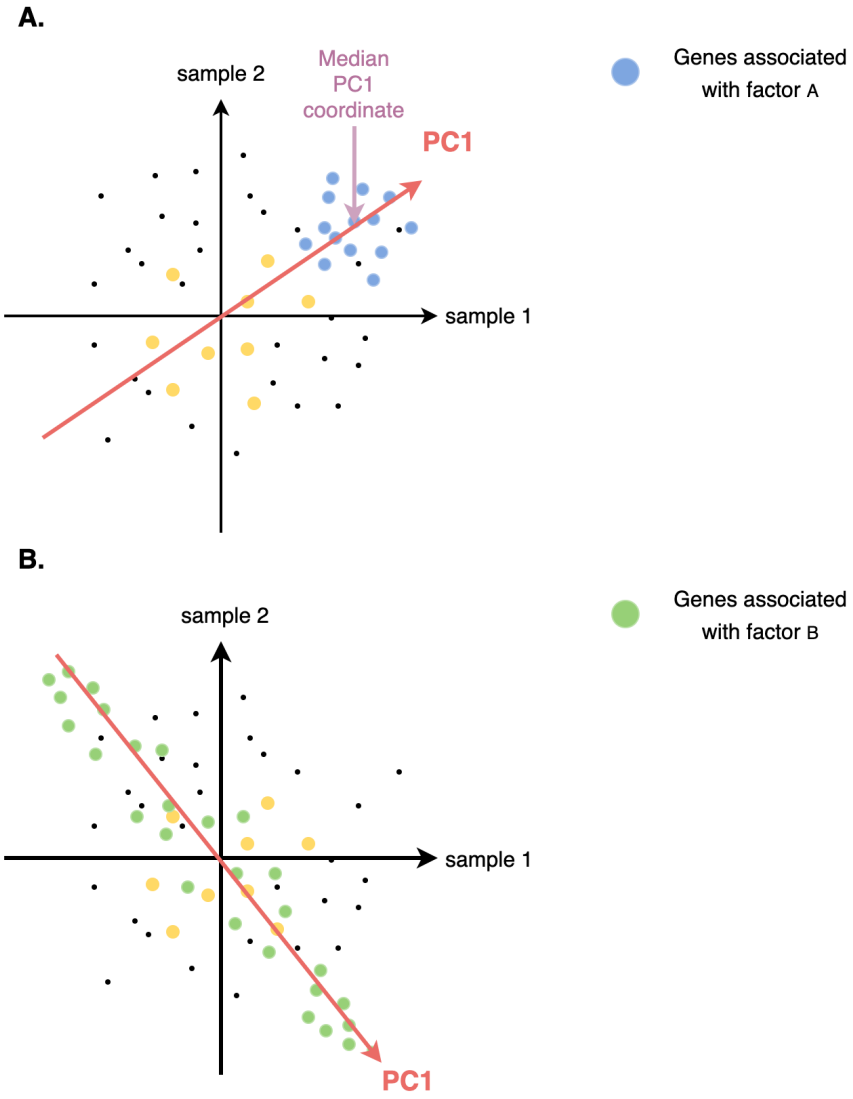


Figure 3.1 – Representation of gene sets in the case of two samples.
Each dot represents one gene, its horizontal (resp. vertical) value corresponding to its expression in sample 1 (resp. sample 2). Genes associated with latent Factor A are plotted in blue, and the corresponding PC1 direction is plotted in red (A). This example corresponds to a shifted pathway, as assessed by a median of gene projections onto PC1 direction far from the origin of the distribution. Genes associated with latent Factor B are plotted in green (B) and the corresponding PC1 direction is plotted in red. This example corresponds to an overdispersed pathway, as the PC1 is well aligned with the dots' distribution. Genes in yellow are neither overdispersed nor shifted, as PC1 explains a relatively small fraction of variance (not represented on the figure) and the median of projections onto PC1 is close to the origin for this group of genes.

The rROMA algorithm is specifically designed to handle situations where genes within a gene set do not equally contribute to its activity. The unifactor model underlying rROMA presumes that an unobserved factor (i.e., a latent factor) acts on the gene expression variables observed in the gene set, and that this action is characterized by the calculated weights. The weights indicate the strength and direction of effect, and can have opposite signs, as in the case in which a transcription factor has an activating action on some genes of the set and a repressor action on others. In particular, identification of the genes displaying the stronger contribution to a gene set's activity allows data analysis in the context of network modeling. Indeed, the output of rROMA, in particular the top weighted genes of significantly active gene sets, can be interpreted as nodes comprising genes and proteins of importance in the system under investigation, which can be used to construct mathematical models.

The package rROMA is an evolution of the algorithm ROMA, a program originally developed in Java [Martignetti, 2016]. The new rROMA package incorporates significant improvements in the calculation algorithm, along with several functions for statistical analysis and graphical visualization of results. These additions greatly expand the package's capabilities and offer valuable tools for data analysis and interpretation. It is an open-source package available on github at: www.github.com/sysbio-curie/rROMA.

As an example of its interest to study complex diseases, we applied rROMA to publicly available transcriptomic datasets in the context of cystic fibrosis (CF). CF is a genetic disease caused by mutations in a single gene, the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) gene, coding for the CFTR protein that functions as a chloride channel in various epithelial cells, including airways epithelial cells. Absence of a functional CFTR protein causes further functional dysregulations in various biological pathways, leading to various deleterious phenotypes observed in CF patients that cannot be related to the chloride channel function in a direct manner. Analysis of dysregulations at the pathways level may provide a better understanding of the molecular basis of some of the CF disease phenotypes.

Therefore, we applied rROMA to investigate pathway activities in airway epithelial cells from CF patients and from healthy donors [Saint-Criq, 2020]. The analysis shows that rROMA can identify biological pathways associated with diseases from transcriptomic data, allowing both a clearer interpretation of high-throughput data from a biological point of view, and the interpretation of molecular changes in a functional way. Results highlighted a relevant mechanism in the context of CF, a potential bias due to cell culture, and an interesting gene that could be studied further. The analysis workflow is schematized in Figure 3.2. A detailed vignette to reproduce all the analysis presented here is also available on the gitHub containing the source code of the algorithm.

3.2.2 Methods

First Principal Component and the Simplest Uni-Factor Linear Model of Gene Expression Regulation

The main idea of rROMA is based on the simplest uni-factor linear model of gene regulation in which it is assumed that the expression of a gene G in sample S is pro-

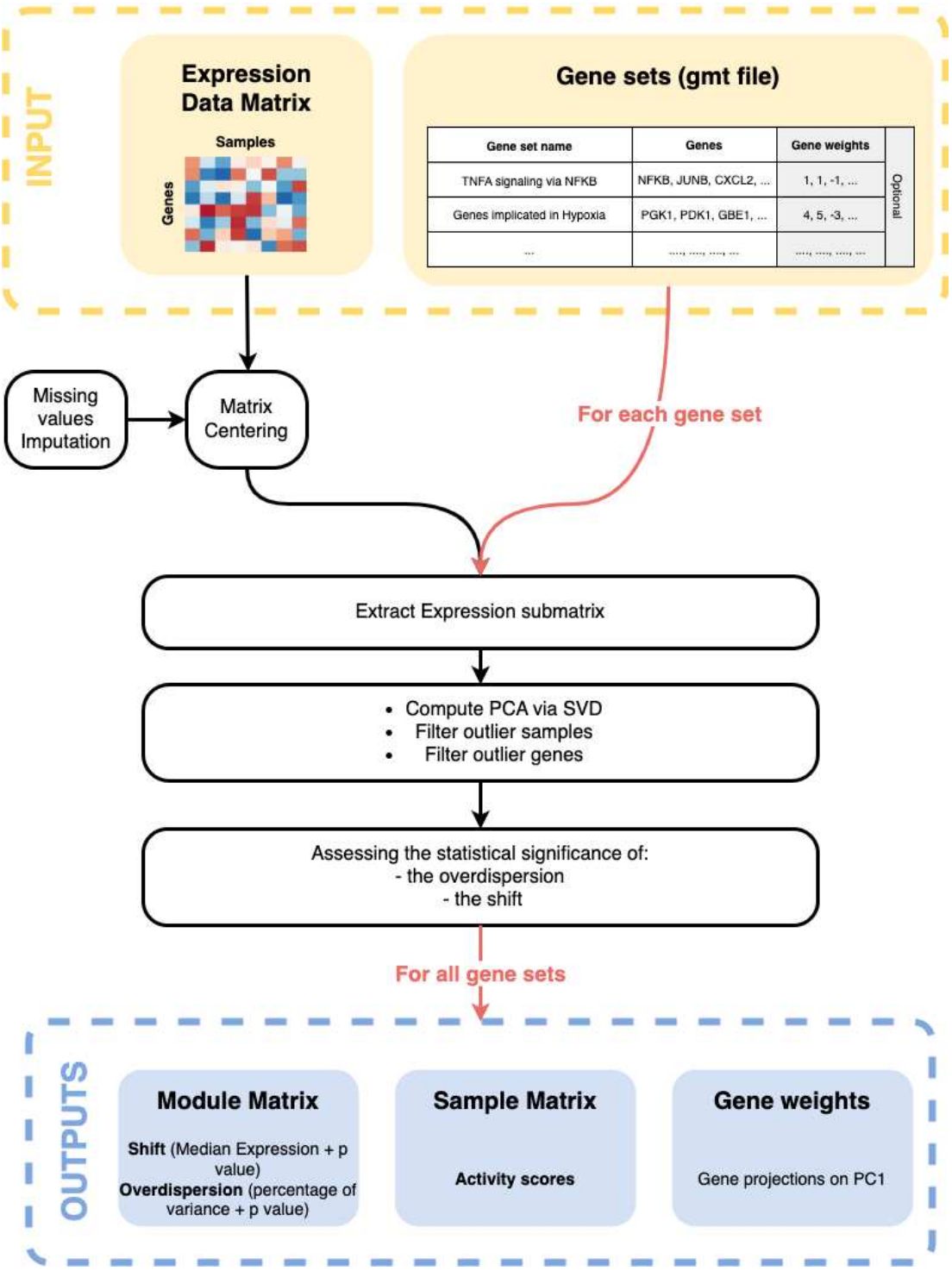


Figure 3.2 – Schematic diagram illustrating the workflow of the rROMA algorithm.

portional to the activity of one latent biological factor F (which can be a transcription factor or any other endogenous or exogenous factor affecting gene expression) in sample S with positive or negative (response) coefficient. Within this model, the expression of a gene G in a sample S is proportional to the activity of a factor F, so that we can write:

$$Expression(gene\,G, sample\,S) \approx \alpha_G^F\, Activity_S^F$$

where α_G^F is the response coefficient of the gene G to the factor F and $Activity_S^F$ is the activity of factor F in the sample S. These two values can be easily determined by considering the first component of the principal component analysis (PCA) of the genes of the considered gene set in the space of the samples. In this case, the vector containing the activities of the different samples corresponds to the first eigenvector, i.e., the first column of the weight matrix. The vector containing the response coefficients of the different genes in the considered gene set corresponds to the projection of the genes onto the first component.

As many other applications of PCA, rROMA uses singular value decomposition to speed up the calculation. However, to be valid, the use of this method requires centering the data beforehand. We define a dataset made up of n individuals, described by p numerical variables, which we represent by a matrix X of $n \times p$ size. In this matrix, the i th column corresponds to the vector x_i of observations of the variable i. The PCA consists in determining the eigenvalues and the eigenvectors of the covariance matrix C of X. This can be written as:

$$C = \frac{(X - X_{mean})\,(X - X_{mean})^\top}{(n-1)}$$

Where X_{mean} contains the mean values of x_i for each column. When data are centred, this equation can be simplified as $C = \frac{XX^\top}{(n-1)}$. Therefore, we have a symmetric matrix that can be diagonalized. If we denote C as the covariance matrix of X, there exists a diagonal matrix L and a matrix W such that $C = WLW^\top$. Following the single value decomposition (SVD) theorem, any matrix X can be written as the product of a diagonal matrix S and two matrices U and V such that: $X = USV^\top$. By analogy, the eigenvectors of the X matrix are obtained from the V matrix, and its eigenvalues are proportional to the diagonal of the S matrix. The computation of the SVD is therefore equivalent to computing the covariance matrix of X. The former is also much faster, and there are algorithms that allow it to accelerate even more by focusing only on the first columns of V.

rROMA uses the R irlba package [Baglama, 2005], which allows to focus on the first principal component, by computing only the first column of V and the first value of S. However, if the data are not centred, the simplification is no longer valid, and we must now consider the fact that the mean is not zero. The rROMA algorithm therefore starts by centering the data of the global matrix.

Pre-processing of data for rROMA analysis

The input format for gene or protein expression for rROMA is a tab-delimited text file with columns corresponding to biological samples, and rows corresponding to

genes or proteins. The first row is assumed to contain the sample identifiers while the first column is assumed to contain the non-redundant gene or protein names. If the data table contains missing values, they can be imputed using an approximation of the data matrix with missing values by a full matrix of lower rank. To do this, the user must specify the rank of the approximate full matrix that he wants to use. Then, the principal components are calculated up to the specified rank, using an algorithm capable of working with missing data [Gorban, 2010]. This PCA decomposition is used to construct the full approximate matrix of lower rank, from which the missing values of the original data are imputed. In the rest of the algorithm, the full imputed matrix is used.

Orientation of the PC1

In standard PCA, all components are calculated with an undefined sign of orientation: there is mirror symmetry, which makes it difficult to determine whether a given set of genes is over- or under-activated. In rROMA, several methods exist to solve this ambiguity. If knowledge exists about the role of a gene in a given collection, we recommend using it by associating a sign with the effect of the gene in the collection: negative for an inhibitor, and positive for an activator, for example. rROMA then uses the information about these signs to choose the orientation that maximizes the number of genes associated with a positive sign whose projection in PC1 is positive, and the number of genes associated with a negative sign whose projection in PC1 is negative. Some a priori fixed weights can be associated to each gene in the gene set file (Figure 3.2) to fix their contribution.

Although less efficient, other methods of orienting the first component exist in the case where there is no a priori knowledge about the genes in the reference gene sets. The most efficient method consists in considering only the genes associated with the most extreme weights according to the first component (according to a percentage defined by a modifiable hyperparameter), and then summing these weights. If the result is negative, then the orientation of the PC1 is reversed. The principal behind this method is to orient the first component according to the most contributing genes for the gene set studied.

Filtering of outlier samples

Measurements may have been performed incorrectly in some samples, and keeping them may lead to erroneous results from rROMA. By default in the algorithm, no sample filtering is performed, as the matrix used as input is assumed to contain only correct samples. However, sample checking can be activated by a hyperparameter, and two filtering steps are then performed. First, the algorithm ensures that a similar number of genes is detected in all samples, and the allowed difference threshold is defined manually. Then, the samples are projected in the gene space, and rROMA ensures that no sample is too far from the others. The number of PCs used to perform the filtering and the allowed difference threshold are defined by two hyperparameters. It is also possible to perform only one of the two filtering steps.

Filtering of outlier genes

The calculation of PC1 can be affected by the presence of an outlier gene in the dataset. This outlier could indeed artificially affect the PC1. In order to increase the robustness of the PC1 calculation, we use in rROMA the "leave-one-out" cross-validation approach [Hastie, 2009]. This method works as follows: first, for each gene in the considered gene set, we calculate the percentage of variance L1 explained by PC1 when the gene is removed from the dataset. Each gene is then associated with a L1 value. In a second step, the distribution of these L1 values is centred and reduced to obtain z-scores. In a third step, all genes whose associated z-score above a threshold value set in the algorithm are considered as outliers and removed from the analysis. The idea behind this method is to identify genes that have too much impact on the PC1 on their own: if the percentage of variance explained by PC1 increases significantly in the absence of a single gene, this means that this gene does not follow the alignment of all the others. It is then considered as an outlier gene.

When a gene is considered an outlier for a given gene set, it is only removed from that gene set. This default behavior can be modified by a hyperparameter, so that genes are completely removed from all analyses, as soon as they are considered outliers in at least one gene set. However, in some cases, these genes may still be of interest for analysis. Additional analysis steps for these genes are therefore available in rROMA.

In particular, the greater the number of gene sets analyzed containing a given gene, the more likely it is that the gene will be considered as an outlier in at least one gene set, and therefore be removed from the analysis. In order to avoid such abusive withdrawals, a Fisher test is performed. This consists of comparing the average proportion of collections in which the genes in the analysis are considered as outliers to this same proportion for a particular gene. If the proportion of outliers is close to the average proportion for all the genes, then it is no longer considered an outlier. Conversely, if the proportion of aberrations is significantly higher (threshold determined by a modifiable hyperparameter), then it is still considered an outlier. Instead, a gene present only in a small number of gene sets may be important for the understanding of these gene sets. In such cases, it is not desirable to remove it, even if it is considered an outlier. Thus, if a gene is present in less than a defined number of gene sets (set by a modifiable hyperparameter), then it is not considered an outlier, regardless of the results of the “leave-one-out” approach for it.

Finally, it is possible that a significant proportion of genes are considered outliers for a given gene set. However, removing too many genes can totally distort the detectable behavior for a collection. A final filter therefore exists to limit the maximum proportion of genes that can be considered as aberrant for a given gene set and removed from the analysis (the proportion is set by a modifiable hyperparameter). In this way, only genes with the most extreme leave-one-out variations are effectively considered outliers for this gene set.

Interpretation of results

The core of the rROMA algorithm starts once these pre-processing steps, involving imputing missing values, removing outlier samples, centering the data at the global

level, and then removing outlier genes from each gene set have been performed. Each gene set is then analyzed separately. First, the algorithm computes the first principal component of the genes in the gene set in the sample space. Two measures of interest are then considered: the percentage of variance explained by the first component alone, and the average expression value of the genes projected onto this first component.

The same measures are then performed for a hundred randomly generated gene sets of the sam size, which will constitute the reference null distribution. As explained in the previous section, the use of SVD is only justified if the data are centred. If this condition is not met, the median projection on the first component will not be zero. For this reason, the average value of the projection of the genes in the gene set under study onto PC1 is compared to the values obtained for the random collections of genes that constitute the null distribution and are therefore assumed to be centred. These values can be positive or negative, but since we are interested in the deviation from the center, the absolute values are considered. If less than 5% (value defined by a tunable hyperparameter) of the values obtained in the null distribution are lower than the one obtained for the gene set under study, the latter is said to be shifted (*ppv Median Exp* <0.05). This means that the average expression of the genes in the gene set is different from the average expression of all genes for at least one sample. In such a case, additional analyses are needed to determine precisely the origin of this shift. If the collection is not shifted, then the centred data assumption is considered valid. We are then interested in the percentage of variance explained by the first component. If less than 5% (value defined by a modifiable hyperparameter) of the values obtained in the null distribution are lower than the one obtained for the gene set under study, then the latter is said to be overdispersed ($ppv\,L1 < 0.05$).

Analysis of a shifted gene set

For a shifted gene set, the measurement of sample activity is particularly important. It determines which samples are responsible for the shift of the gene set. If the samples were already separated into several conditions prior to the analysis, it is possible to verify that this separation into groups is indeed responsible for the shift, by checking that the activity scores are significantly different between the conditions. Conversely, if the conditions are not known a priori, it is possible to determine new groups by performing a hierarchical clustering analysis on the gene sets that are shifted, and thus determine potential new groups of interest from the analysis.

Analysis of an over-dispersed gene set

The analyses mentioned above for the case of shifted gene sets are also valid for *over-dispersed* gene sets. But in the case of the latter, the analysis of gene weights also becomes interesting. The genes associated with the highest weights are the driving force in the activity scores of the gene sets. They summarize the information of the gene sets, which can be particularly useful, especially for interpretations using systems biology approaches.

The analysis of the sign of the genes' weights is also particularly interesting. For example, gene sets containing both activators and inhibitors can be highlighted by rROMA by being overdispersed, and the associated genes highlighted. Such gene sets

would not be detected as overdispersed by methods based on the average expression of genes in the samples.

Optimization of the calculation of null distributions

In practice, it is often necessary to test many different gene sets available in large reference databases such as KEGG [Kanehisa, 2012] or MSigDB [Liberzon, 2015]. Estimating the null distribution for each set of genes can lead to very time consuming calculations. rROMA does not compute the significance scores of overdispersion and shift for all gene sets, but approximates them on a predefined grid of values, depending on the size of the considered gene set. Indeed, since these two values are dependent on the size of the gene set, it is not possible to use the same null distribution for all gene sets. In order to rapidly estimate the importance of the over-dispersion and shifting scores, rROMA constructs null distributions for a representative list of gene set sizes. These are selected to be uniformly distributed in the logarithmic scale between the minimum and maximum size of the reference database. For a given gene set, the null distribution that is closest in size in the log scale is then chosen.

3.2.3 Case study

rROMA identifies active signalling pathways in CF tissues compared to healthy donors

We applied rROMA to investigate the activity of pathways in airway epithelial cells from CF patients and healthy donors. More precisely, we compared the transcriptomes of primary cultures of airway epithelial cells from patients (N=6) with those of healthy controls (N=6), based on RNAseq data publicly available in the NCBI's GEO database, under the accession ID GSE176121 [Rehman, 2021]. rROMA was run by specifying the pathway database to use and the expression matrix to analyze, as shown in the accompanying vignette.

Here, the Molecular Signature Database MsigDB hallmark gene set collection [Liberzon, 2015] was used, a gene set collection of 50 gene sets specifically curated to represent core biological processes and pathways that are commonly dysregulated in cancer. However, to provide a more complete view of the biological processes involved in a study, rROMA can be applied with different reference databases.

The results of rROMA highlight pathways that are provided in the *ModuleMatrix* output. Pathways with a *ppv Median Exp* lower than a given threshold were deemed as shifted, while those with a *ppv L1* lower than this threshold were overdispersed. The *Plot.Genesets.Samples* function allows for the visualization of activity scores for significantly shifted and overdispersed pathways across samples, in the form of a heatmap representation (see Figure 3.3). rROMA identified two shifted pathways, APICAL SURFACE which is found to have higher activity scores in healthy controls than in CF patients, and FATTY ACID METABOLISM which has higher scores in CF patients than in controls, and an overdispersed pathway, COAGULATION, with higher activity scores in healthy controls than in CF patients.

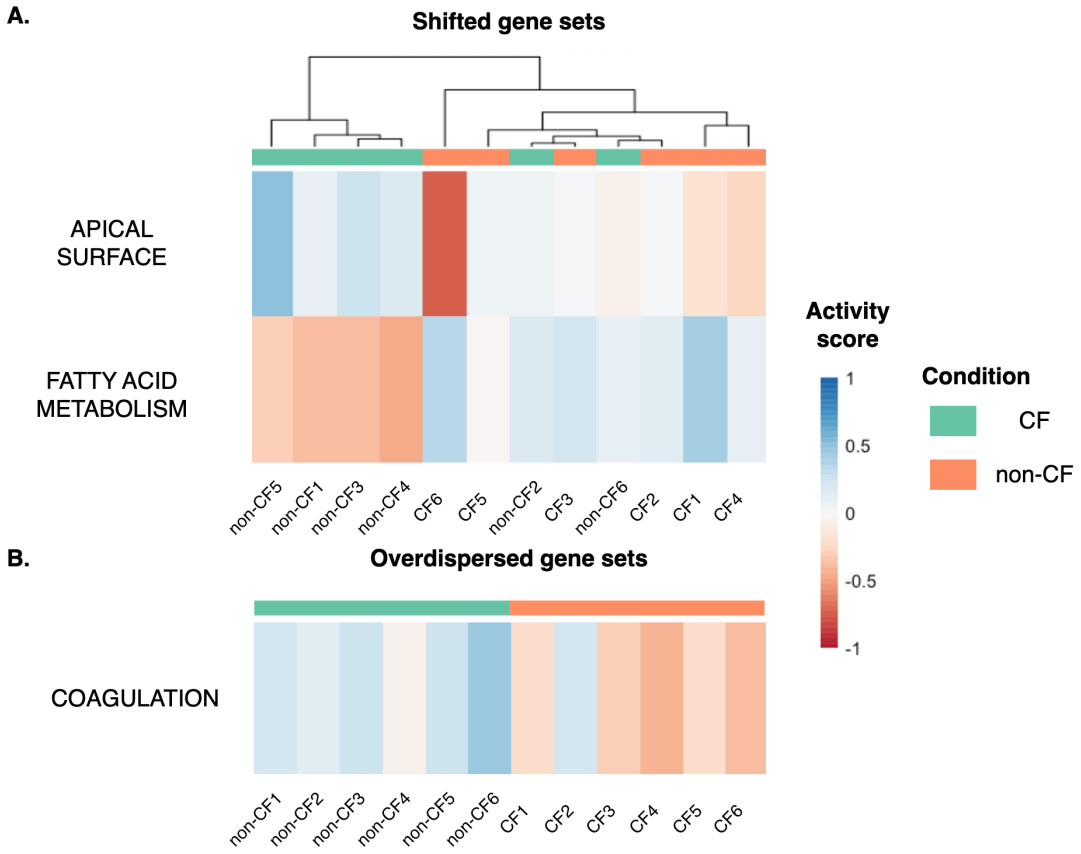


Figure 3.3 – Heatmap of activity scores for gene sets identified as significantly shifted (A) or significantly overdispersed (B) in GSE176121 dataset.
Samples are in columns, gene sets are in rows. Horizontal sidebar color encodes true class labels.

When sample groups have been pre-defined, as the two CF or control groups in our case, these groups can be compared based on the activity scores of the gene sets observed in the samples belonging to the two groups. Boxplot of the activity scores based on predefined groups can help differential analysis. In our study, shifted and overdispersed pathways behaved significantly differently in CF patients versus healthy controls, as shown in Figure 3.3. Alternatively, when the groups are not predefined, analyzing the shifted and overdispersed pathways can reveal clusters of samples exhibiting similar pathway activity.

The analysis of top contributing genes in each pathway also provides crucial information. The weights assigned to each gene in the PC1 vector allow to identify the key genes that most contribute to variations in pathway activities as those with the higher weights. For each pathway, gene weights are provided by the *PlotGeneWeight* function. For example, plotting the gene weights for the COAGULATION pathway (see Figure 3.4) highlighted the GSN gene, encoding the protein *Gelsolin*, as the highest contributor to the activity score. Notably, *Gelsolin* has been previously reported to play a role for CFTR activation [Vasconcellos, 1994; Cantiello, 1996], and to promote

mucus fluidification in CF [Bucki, 2015].

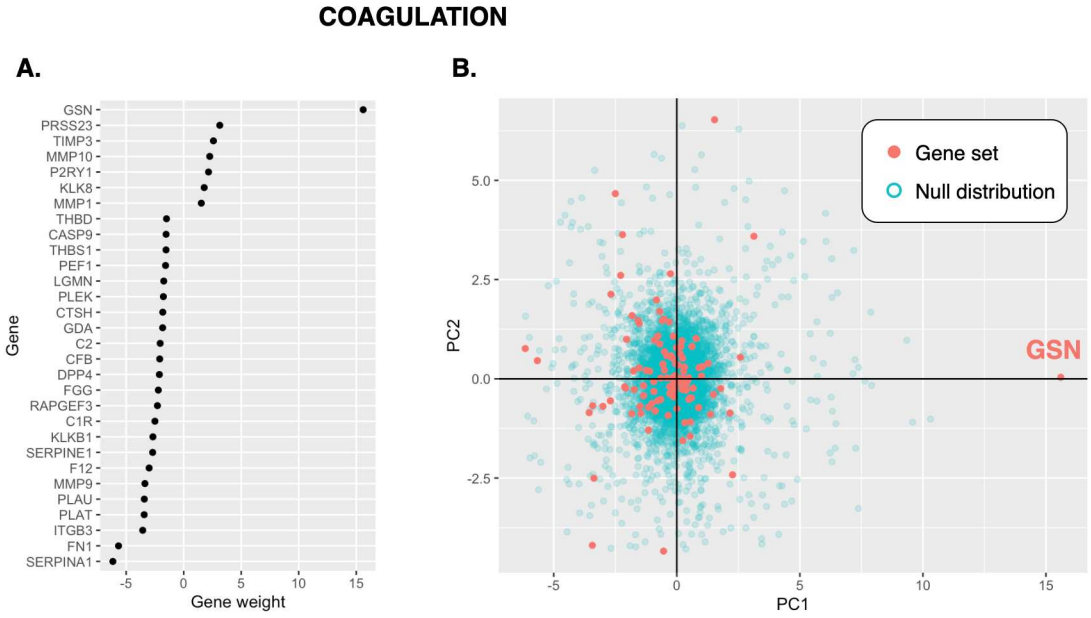


Figure 3.4 – Plots illustrating the contribution of genes to the COAGULATION gene set activity score.
The weights in panel A indicate the gene projections on PC1, limited to the genes that have the greatest contribution to the observed variation in the COAGULATION gene set. In panel B the genes of the COAGULATION gene set are represented in the PCA space. Red dots are genes from the gene set, blue dots show randomly selected genes used to generate a null distribution.

Finally, many hyperparameters can be specified and changed to modify rROMA speed, precision, or behavior regarding outlier detection. Details about all available hyperparameters are described in the vignette accompanying the software. The computational time required to run the algorithm typically depends on the number of considered pathways and their relative sizes. It also depends on whether parallelization is enabled. In the present work, the algorithm ran in about 3 minutes and 15 seconds on a MacBook Pro equipped with a 2,6 GHz Intel Core i7 6 cores processor. A single 60 genes pathway took roughly 5 seconds to be analyzed. Parallelization was not used, but this would have increased the speed of the analysis.

rROMA estimates cell type abundances from bulk transcriptomic data

In addition, our case study also demonstrates the ability of rROMA to investigate cell type abundance based on bulk transctiptomic data. Saint-Criq et al. investigated the impact of two differentiation media on primary cultures of CF and non-CF airway epithelial cells, as determined by transcriptomic data [Saint-Criq, 2020]. They built a gene signature for each cell type using the 50 most expressed markers derived from a single-cell RNA sequencing (scRNAseq) dataset [Plasschaert, 2018]. They observed

a significant overexpression of genes belonging to the signature of the secretory cell subtype in cultures grown in one of the media (referred to as UNC), compared to the other medium (referred to as SC). Conversely, gene markers of the ciliated subtype were overexpressed in primary cultures grown using the SC medium compared to UNC medium.

We applied rROMA to the Saint-Criq RNA seq dataset to estimate cell type abundance in CF and non-CF samples. More precisely, the Plasschaert signature for each cell type was used and gene reference gene sets, and the activity scores of these gene sets across the samples were represented in the form of a heat map. Samples were found to be clustered according to the differentiation medium in which they were grown, and our results confirmed the higher abundance of the ciliated cell subtype in UNC medium and higher abundance of the secretory cell subtype in SC medium (Figure 3.5A). We repeated the rROMA analysis using an alternative signature [Okuda, 2021]. In contrast with the Plasschaert signature, the Okuda signature includes the most differentially expressed genes in each cell type, encompassing both overexpressed and underexpressed genes. As illustrated in figure 3.5B, rROMA's analysis using a reference gene set corresponding to this alternative signature consistently revealed the same relative abundances of secretory and ciliated subtypes between UNC and SC growing media, as observed with the initial signature. Thus rROMA allows us to clearly highlight differences in cell-type abundances, facilitating the use of gene signatures that contain both upregulated and downregulated genes and thus potentially more accurate.

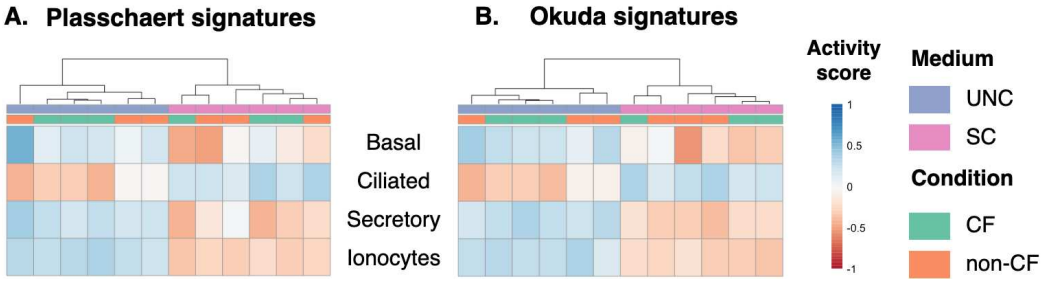


Figure 3.5 – Heatmap of rROMA scores obtained for Plasschaert (A) and Okuda (B) signatures of cell types in the Saint-Criq RNA seq dataset.
Samples are in columns, gene sets corresponding to cell types are in rows. Horizontal sidebar color encodes true class labels.

3.2.4 Discussion

Quantifying the activity of biologically related gene sets is a commonly employed approach to extract valuable biological insights from high-throughput data. The use of gene sets as aggregated variables from molecular data enables the capture of biological information that may not be detectable when solely focusing on individual genes. To address this challenge, we introduced the rROMA algorithm. Based on a gene expression data matrix, this algorithm implements a linear model of gene regulation and efficiently and reliably quantifies the activity of gene sets by computing the first principal component (PC1), while also evaluating the statistical significance of this

approximation.

We applied rROMA to CF transcriptomic datasets, highlighting some biological mechanisms potentially involved in the initiation or progression of the disease, and their associated genes. In our study, out of the 50 hallmark pathways tested, 3 were significantly active: FATTY ACID METABOLISM, APICAL SURFACE, and COAGULATION. The FATTY ACID METABOLISM pathway has significantly different activity scores between CF patients and healthy donors. This pathway has been extensively studied in CF, and essential fatty acid deficiency is a well known CF phenotype (for a review, see Strandvik [Strandvik, 2010]). The APICAL SURFACE can be related to another well known hallmark of CF, i.e. a perturbation of airway surface secretory mucus content. Finally, the COAGULATION pathway, the only overdispersed pathway in our study, seems to be highlighted due to one specific gene with a very high associated weight, that is by far the most contributing gene to the activity score of this pathway: *Gelsolin* (GSN). *Gelsolin* has been reported as playing a role for CFTR activation [Vasconcellos, 1994; Cantiello, 1996], which suggests that the role of this gene in CF disease may be interesting to study in more detail. The goal of this use case was not to undertake a detailed systems biology approach of CF, which is beyond the scope of the present paper. In particular, it would require us to include additional transcriptional dataset to take more samples into account and increase the statistical power of our analyses, and to test several reference databases of gene sets.

However, overall, this case study illustrates that rROMA is able to identify disease-associated pathway dysregulations from transcriptomic data, allowing a more comprehensive and functional interpretation of the data. It is also a versatile tool that can shed light on various biological questions such as highlight the key genes driving these dysregulations, identify clusters of samples, study samples' cell-type composition, or other cellular changes in a broader biological perspective.

3.3 A broader discussion related to the PhD project

In this paper, we presented *rROMA*, an algorithm for the quantification of pathway activity from bulk omics data. *rROMA* computes sample-wise gene set scores and does not require a predefined label classification of samples into groups (e.g. "disease" or "control" labels). Therefore it enables the clustering of samples with a greater biogical interpretability than at the gene level. Other methods exist, but *rROMA* is unique because it combines several functionalities that other methods do not offer, or at least not all in once:

— the detection of both overdispersed and shifted pathways.

— the estimation of the statistical significance of the distribution of sample activities.

— the ability to address scenarii where genes within a gene set do not equally contribute to its activity, i.e. where certain genes hold more significance than others in defining the activity of the module, or where specific genes are expected to negatively correlate with the activity of the module.

— and finally, the detection of outlier genes in the dataset.

A comparison of the most used linear sample-wise methods of pathway quantification

is presented in Table 3.1. All these functionnalities make *rROMA* very well suited for broader applications and exploratory analyses with heterogeneous samples.

Table 3.1 – Feature-wise comparison of rROMA to existing tools.

Method	**z-score**	**GSVA**	**ssGSEA**	**PLAGE**	**rROMA**
Type	Aggregate expression	Ranking based	Ranking based	SVD	SVD
Gene-set score	X	X	X	X	**X**
Overdispersed gene-sets				X	**X**
Shifted gene-sets	X	X	X		**X**
Stat significance					**X**
Outlier detection					**X**
A priori gene weights					**X**
Code availability	R	R	Java, R	R	R
Reference	18989396	23323831	19847166	16156896	NA

In the context of the PhD project, the aim was to build a biological network of CF dysregulations.

rROMA can be useful to build biological networks. Except in the case of network-based approaches are used, a list of genes or proteins is initially needed to build a biological network. There are generally two ways to gather this list of genes: retrieve them from a review of the literature, or extract them from data. *rROMA* can be particularly appropriate for the latter. The algorithm is specifically designed for cases where the genes in a gene set do not contribute equally to its activity. It computes the weights of each gene in the gene set, which indicates the strength and the effect of the "unseen" factor on each gene. Therefore, the genes associated with the highest weights are the driving force in the activity scores of the gene sets.

It would be then very interesting to connect these "heavy" genes into one network, as they are the most characteristic ones in the system under study. For instance, following the pipeline proposed in figure 3.6, we could connect these genes by retrieving all direct interactions between them, or interactions involving one or several intermediates from PPI databases.

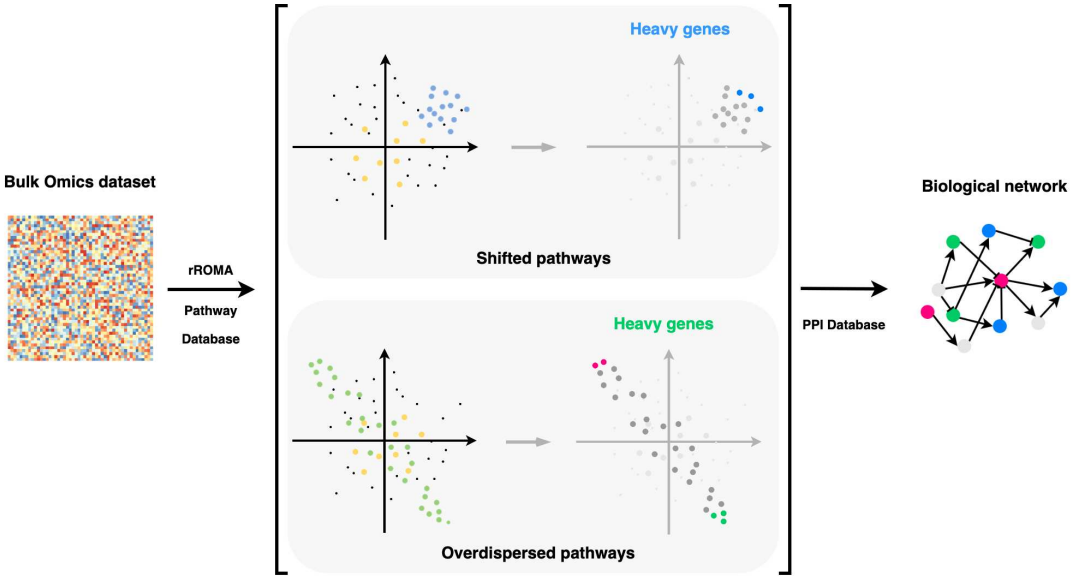


Figure 3.6 – Pipeline using rROMA to build biological networks.

We applied *rROMA* to the CF public omics data in order to retrieve the "heavy" genes of CF dysregulations. The challenge then was to ensure that the altered pathways were related to CF.

Indeed, the algorithm enables to identify altered pathways among all the samples, regardless of the samples conditions. This means that a pathway can be considered as shifted even if the average expression of the genes is different from the average expression of all genes for just one sample. Thus, when applying *rROMA* to compare "disease" and "control" samples, some pathways can be found significantly altered in some samples, but their alteration may not be linked to the disease, but to other factors such as the experimental conditions, sex, age etc.

In order to identify only the altered pathways related to CF, we proposed to perform a Kolmogorov Smirnov (KS) test on the pathways activities computed by *rROMA*, between the CF and the control samples. Indeed, the KS test enables to determine if the distribution of pathway activities in the CF samples differ significantly from the control ones. The resulting p value of the statistical test corresponds to the probability that the pathways activities in the disease samples follow the same distribution as the pathways activities of the control samples.

Unfortunately, we faced some statistical problems. The number of samples per CF dataset was very low: it varied from 5 to 10 (See Chapter 5 for more details about data selection). Therefore, for each dataset, the statistical results obtained from the KS test were not precise enough. The correction of the p values for multiple testing led to no pathways significantly altered for many datasets, making it difficult to compare the results between the different datasets.

This limitation is due more to the size of the CF datasets than to the methods themselves. It would have occurred with any sample-wise method applied to these datasets. Moreover, in our case, the labels were known. Therefore, we opted for Gene Set Enrichment Analysis (GSEA) [Subramanian, 2005], one of the the most commonly

used pathway-based methods. This method evaluates enrichment between two conditions, and requires a single statistical test, conversely to *rROMA* which requires two in this context: the test to identify shifted and/or overdispersed pathways among all the samples, and the KS test to compare the two conditions.

In fact, the greater the number of samples they are in the input expression matrix of rROMA, the better the algorithm will detect dysregulated pathways between the conditions. Indeed, as mentioned in the introduction of this chapter, the computation of rROMA pathway activities is based on the variance across the samples explained by the genes in the gene set. The greater the number of samples in the dataset, the finest the variance and the pathway activities. Overall, this suggests that rROMA is particularly interesting for analysis of large datasets. Such large datasets are now common in diseases like cancer, but not in rare diseases like CF.

rROMA could be applied on larger CF cohorts, for example to help decipher CF patients' hetereogeneity at the biological pathway level. For example, one could investigate CF patients with different mutations. Similar pattern of coordinated expression may be found between group of samples with mutations of different classes (See chapter 1 for more details about mutation classes). Such results could highlight similar molecular dysregulations even among patients bearing mutations of different classes, and indicating the use of similar therapeutic solutions for these patients. This could be very helpful for patients with unrescuable mutations that are not eligible to the CFTR modulators.

A second application would be to investigate the hererogeneity of CF patients bearing the same mutation. Indeed, symptoms of CF patients with the same mutation can be heterogeneous [Cornet, 2022b]. Applying rROMA to a large dataset of CF patients with the same mutation could identify subgroups of patients with similar molecular dysregulations. Then, it could help to understand why some patients have more severe symptoms than others, or are more responsive than others to the current treatments, and find better treatments to the less responsive patients.

Chapter 4

State of the art in systems biology approaches to study CF

Contents

Abstract

In this chapter, I provide an overview of the the different omics studies available in CF. Transcriptomics techniques have been the most common omics techniques that have been used in CF research. Additionnally, a significant portion of CF-related proteomic studies have been focusing on the partners of CFTR in the cell (the so-called CFTR interactome). Secondly, I review the studies that address molecular mechanisms in CF through a systems biology approach. These studies have mainly focused on the molecular mechanisms centered around CFTR, offering a valuable resource for understanding CF.

Résumé

Dans ce chapitre, je donne un aperçu des différentes études omiques disponibles sur la mucoviscidose. Les techniques transcriptomiques ont été les techniques omiques les plus couramment utilisées. En outre, une part importante des études protéomiques liées à la mucoviscidose s'est concentrée sur les partenaires de CFTR dans la cellule (ce que l'on appelle l'interactome de CFTR). Dans un second temps, je passe en revue les études qui abordent les mécanismes moléculaires de la mucoviscidose par le biais d'une approche de biologie des systèmes. Ces études se sont principalement concentrées sur les mécanismes moléculaires centrés sur CFTR, offrant ainsi une ressource précieuse pour la compréhension de la maladie.

4.1 CF omics data

Studying monogenic diseases may appear easier than complex systemic diseases, because it would mainly rely on studying the protein resulting from the mutated gene, and focus on restoring its expression or function. However, proteins interact with one another and with other cellular components, so that absence of a given functional protein can lead to broader cell dysregulations, involving the dysfunction of its partners. Systems biology approaches allow to provide a global vision of the biological processes that characterize the disease, and thus facilitate the study of its complexity. These approaches are therefore also useful for studying monogenic diseases such as CF.

One of the most important questions in a systems biology approach is what kind of data are available, and how they can be used to build the model. High throughput technologies are a good valuable to infer molecular dysregulations, and enable a global vision of the biological system by giving information of thousands of components in the cell (See chapter 2 for an introduction of systems biology approaches and omics techniques). Since the last decade, the use of such technologies has grown for what concerns the research on systemic diseases, but also on monogenic diseases, and CF is no exception.

This chapter will present the state of the art of systems biology approaches to CF. In a first part, I will first present the different omics studies available in CF, because they fuel systems biology approaches, and I will focus on how they helped to unravel some CF molecular dysregulations. Then, I will present existing initiatives in the scientific literature that employ systems biology to investigate CF.

4.1.1 Transcriptomic studies

Bulk transcriptomic studies

Transcriptomics techniques are the most common omics techniques that have been used in CF research. More than 30 transcriptomics studies have been conducted so far to answer questions related to CF. All the CF human transcriptome profiling studies have been reviewed in 2019 [Ideozu, 2019]. An up-to-date list of studies is presented in Table 4.1, including new studies which became available since then together with some studies which had been overlooked.

Table 4.1 – Human transcriptome profiling studies in CF

Author	Year	Focus	Methods	Tissue/Cell	Mutation	Reference
Virella-Lowell	2004	Transcriptional changes induced by CFTR gene correction, Interleukin-10 and P. aeruginosa	Microarray	Epithelial	F508del heterozyguous	[Virella-Lowell, 2004]
Zabner	2005	CF vs. non-CF samples	Microarray	Epithelial	F508del homozygous	[Zabner, 2005]
Worgall	2005	Transcriptional changes induced by CF Pseudomonas aeruginosa and Burkholderia cepacia	Microarray	Alveolar macrophages	not mentioned	[Worgall, 2005]
Wright	2006	Mild vs. severe CF lung disease	Microarray	Epithelial	F508del homozygous	[Wright, 2006]
Verhaeghe	2007	CF vs. non-CF samples	Microarray	Epithelial	F508del homozygous	[Verhaeghe, 2007]
Ribeiro	2009	Transcriptional changes induced by Azithromycin	Microarray	Epithelial		[Ribeiro, 2009]
Ogilvie	2011	CF vs. non-CF samples	Microarray	Epithelial	F508del homozygous	[Ogilvie, 2011]
Hampton	2012	Transcriptional changes induced by CF P. aeruginosa	Microarray	Epithelial	F508del homozygous	[Hampton, 2010]
Levy	2012	Transcriptional changes induced by plasma of CF and non-CF	Microarray	PBMCs	F508del homozygous	[Levy, 2012]
Clarke	2013	CF vs. healthy controls	Microarray	Epithelial	F508del homozygous	[Clarke, 2013]
Mayer	2013	Transcriptional changes induced by innate defense regulator 1018	Microarray	Epithelial, PBMCs	varied	[Mayer, 2012]
Stanke	2013	Transcriptional changes influenced by a CF modifier gene-EHF	Microarray	Epithelial	F508del homozygous	[Stanke, 2014]
Chesne	2014	Transcriptional changes influenced by CF and other lung diseases	Microarray	Blood, PBMCs	not mentioned	[Chesné, 2014]
Voisin	2014	Transcriptional changes induced by oxidative stress	Microarray	Epithelial	F508del homozygous	[Voisin, 2014]
McKiernan	2014	CF vs. non-CF samples	Long non-coding Microarray	Epithelial	varied	[McKiernan, 2014]
Balloy	2015	Transcriptional changes influenced by P. aeruginosa in CF and non-CF samples	RNA-Seq	Epithelial	F508del homozygous	[Balloy, 2015]
O'Neal	2015	Transcriptional changes influenced by CF	Microarray	Lymphoblasts	F508del homozygous	[ONeal, 2015]
Kormann	2017	Mild vs. severe lung phenotype	RNA-Seq	Leukocytes	F508del homozygous	[Kormann, 2017]
Zeitlin	2017	Transcriptional changes induced by digitoxin	Microarray	Epithelial	varied	[Zeitlin, 2017]
Polineni	2018	Transcriptional changes influenced by genomic variation	RNA-Seq	Epithelial	varied	[Polineni, 2018]
Jiang	2018	CF before vs. after treatment for exacerbation	RNA-Seq	Blood, neutrophils	varied	[Jiang, 2019]
Levy	2018	Transcriptional changes induced by plasma of CF and its phenotypes	Microarray	PBMCs	varied	[Levy, 2018]
Bardin	2018	CF vs. non-CF samples	small RNA-Seq	Epithelial	F508del homozygous	[Bardin, 2018]
Zoso	2019	CF and non-CF samples before and after mechanical wounding, exposed or not to flagellin	RNA-Seq	Epithelial	F508del homozygous	[Zoso, 2019]
Kamei	2019	CF vs. non-CF samples	Long non-coding Microarray	Epithelial	F508del homozygous	[Kamei, 2019]
Sun	2019	Transcriptomic response to ivacaftor	RNA-Seq	PBMCs	G551D heterozyguous	[Sun, 2019]
Kopp	2019	Transcriptomic response to lumacaftor/ivacaftor	RNA-Seq	Whole-blood	F508del homozygous	[Kopp, 2020]
Bampi	2020	CF vs. non-CF samples	RNA-Seq	Epithelial	varied	[Bampi, 2020]
Ling	2020	Transcriptional changes induced by rhinovirus infection in CF and non-CF samples	RNA-Seq	Epithelial	F508del homozygous	[Ling, 2020]
Saint-Criq	2020	Choice of differentiation media in transcriptional characteristics in CF and non-CF samples	RNA-Seq	Epithelial	F508del homozygous	[Saint-Criq, 2020]
Rehman	2021	Transcriptional changes induced by TNF-α and IL-17 in CF and non-CF samples	RNA-Seq	Epithelial	varied	[Rehman, 2021]

Microarrays have been the most used transcriptome profiling approach until 2015, and since then, RNAseq technologies are the platform of choice because of sensitivity and because of the greatly reduced cost of this technology.

Most of the studies have been done on models of human airway epithelial cells (HAEC). This can be justified by the fact that the most severe symptoms of CF are in the lungs, and as we previously mentioned that epithelial cells are affected by CFTR dysfunction. Lung biopsies are the ideal sample source for these molecular characterization studies but they cannot be obtained from children with CF easily or without considerable risk [Levy, 2018]. Bushing techniques are less invasive and are preferred for CF transcriptomic studies on HAEC primary cultures.

Recently, blood cells have emerged as targets for transcriptome profiling, and especially pheripheral blood mononuclear cell (PBMC). These are round-nucleus blood cells (e.g., lymphocytes, monocytes, or natural killer cells) in the circulatory system that are present at sites of CF airway injury [Saavedra, 2008]. In particular, circulating leukocyte RNA transcripts are systemic markers of inflammation, and are thus currently investigated as assessment for pulmonary treatment response in CF [Levy, 2018].

Most studies focus on the F508del mutation and consider patients homozygous for this mutation. Very few transcriptomic studies have considered rare mutations, genotypes for which CFTR modulators are not currently available.

Finally, regarding cell culture models, transcriptomic studies were done on immortalized cell lines (mostly CFBE41o^-) and bronchial or nasal primary cultures. Compared to primary cell cultures, cell lines are difficult to compare to the actual HAEC of patients. Conversely, primary cell cultures derived from patients better reflect the CF physiopathology, and, specifically, the heterogeneity among patients.

These studies have explored a wide range of questions related to CF, mainly the identification of specific genes that play a role in CF and its various phenotypes, as well as the CF cells response to external stimuli at the transcriptional level (mostly infection by Pseudomonas aeruginosa or response to treatments). The highlighted genes are generally considered as candidates for further studies. Interestingly, the first studies concluded that the level of CFTR mRNA was not significantly different in cells homozygous for the F508del mutation compared to cells from healthy patients.

More recently, with the approval of the CFTR modulators for the treatment of CF patients with particular mutations, studies have evaluated global gene expression before and after these treatments in order to develop gene expression signatures for the prediction of treatment response [Sun, 2019; Kopp, 2020]. All these studies were done on blood samples. To our knowledge, no study has yet focused on the effect of modulators on the transcriptomic profile of airway epithelial cells from CF patients.

Omics data are usually publicly available in the NCBI's GEO database. However, some of the studies presented in Table 4.1 are not shared, and this is particularly the case for the study including 124 CF samples [Polineni, 2018]. This is currently the largest cohort of gene expression data of CF patients, but unfortunately, it is not accessible, in a context where sample sizes of hundreds or thousands of individuals for transcriptomic studies are very difficult to obtain for rare disease and in particular for CF.

Single-cell transcriptomic studies

Single-cell RNA-seq (scRNA-seq) techniques have been developed over the last 10 years [Stark, 2019], and studies are gradually emerging for each disease. A first study of single-cell gene expression on murine tracheal epithelial and primary human airway cells was published in 2018 [Montoro, 2018]. They showed for the first time that CFTR is predominantly expressed in a newly discovered cell type, the so-called *pulmonary ionocytes*. The same year, another study investigated single-cell profiling of human bronchial epithelial cells and mouse tracheal epithelial cells [Plasschaert, 2018]. Again, they reported the identification of a novel, rare cell type called *pulmonary ionocytes* as well as the the fact that this cell type is the major source of CFTR activity in the conducting airway epithelium. Since then, three other scRNA-Seq studies have been conducted to chart the cellular landscape of healthy upper and lower airways [Ruiz García, 2019; Vieira Braga, 2019; Deprez, 2020].

In 2020, a study conducted by Okuda [Okuda, 2021] performed scRNA sequencing to identify cell types that contribute to CFTR expression and function, within the proximal-distal axis of the normal human lung. ScRNA-seq data analysis identified secretory cells as dominating CFTR expression in human airway superficial epithelial although ionocytes expressed the highest CFTR levels but were rare. The expression in ciliated cells was infrequent and low. In conclusion, they suggest that CFTR therapies should act on secretory cells. Finally, in May 2021, the first and only to date scRNA-seq study on CF cells was published from a multi-institute consortium led by the team of Gomperts in UCLA [Carraro, 2021]. Proximal airway of CF donors undergoing transplantation for end-stage lung disease were compared with that of healthy lung donors. This study confirmed that secretory cells, as well as basal cells, account for the vast majority of CFTR expression in the proximal airway epithelium.

When we started the project, no scRNA-seq data on CF was published or available. Due to time constraints, the analysis of these data has not been considered in this thesis, but it is clear that it is necessary for a complete analysis of the complex system of CF HAEC. Further studies of dysregulations in each major cell type (basal, secretory and ciliated) would allow the search for cell type-specific therapies, as suggested in [Okuda, 2021], where I propose some avenues for such analyses in chapter 8.

4.1.2 Proteomics studies

Proteomic analysis which allows the detection of the presence of thousands of proteins in the cell, is another promising approach to obtain a global picture of the cell components and identify potential cell dysregulations.

CFTR-interactome profiling studies

Proteomic approaches have been first investigated to decipher the partners of CFTR in the cell (the so-called *CFTR interactome*). These *interactome profiling* studies consist in CFTR immunoprecipitation coupled to mass spectometry. Just like CF transcriptomic studies, the majority of these studies focused on the most common mutation in CF, F508del, and compared the interactome of WT-CFTR to that of F508del-CFTR

[Wang, 2006; Pankow, 2015; Canato, 2018; Matos, 2018]. Only one study considered another mutation, the G551D one [Teng, 2012]. All these studies highlighted protein interactions potentially lost (or gained) when CFTR is mutated, and then studied the link between one or several of them and CF phenotypes.

The studies showed that the interactome of WT-CFTR and of F508del-CFTR are very similar, with more than 80% of the proteins interacting with both WT- and F508del-CFTR [Pankow, 2015]. They showed that CFTR interacts with chaperone and co-chaperone proteins whose function is to assist other proteins in their maturation [Wang, 2006]. They also highlighted a group of proteins involved in the degradation of misfolded proteins, and thus interacting with F508del-CFTR [Pankow, 2015].

All these studies focused on the CFTR life cycle, from its folding in the ER [Canato, 2018] to its stabilization at the PM [Matos, 2018] and through its processing in the cytoplasm. It should be interesting to compare WT and mutated CFTR interactomes and their impact on the signalling of the cell. To date, only one study used high throughput proteomics for this purpose [Reilly, 2017]. Their results linked CFTR defect to deficient autophagy and the mTOR (mammalian target of Rapamycin) signalling pathway. Finally, proteomic approaches have also been applied to investigate the interactome of CFTR when the cell is treated: for instance with lumacaftor (VX-809) [Matos, 2018] or when the CFTR processing to the PM is rescued with incubation of low-temperature of 26-30°C [Pankow, 2015].

Note that yeast two-hybrid screens have also been used to identify CFTR partners. Conversely to proteomic approaches, they do not comprise mass spectometry. Both techniques are improving although they have their own limits: yeast-two-hybrid screens cannot detect proteins located at the membrane, whereas immunoprecipitation requires cell lysis which may modify the interactome [Lim, 2022].

Whole-cell proteomic studies

Proteomic studies carried out on the whole cell are still much less used than transcriptomic studies as it is easier, cheaper and quicker to read the whole transcriptome than the whole proteome. To our knowledge, three whole-cell proteomic approaches have been performed on CF lung epithelial cell lines [Pollard, 2006; Ciavardelli, 2013; Puglia, 2018] and four on primary culture of CF airway epithelial cells [Jeanson, 2014; Rauniyar, 2014; Braccia, 2019; Veltman, 2021]. The observation is the same as for the CFTR-interactome profiling studies, and the analyses compare WT-CFTR and F508del-CFTR proteomes, the rescue of CFTR processing being the major issue considered in these studies. To date and to our knowledge, no phosphoproteomic studies have been conducted in the study of CF systems, whether on cell lines or on primary cultures.

The main limitation of proteomic studies is that a few thousand of proteins are detected, when transcriptomic studies detect up to more than 20,000 RNAs.

4.2 Systems biology approaches for CF in the literature

4.2.1 Pathway-based approaches

In-depth pathway-based analyses have never been the main purpose of CF omics studies. They are usually conducted in studies whose purposes are to present transcriptomic data in specific biological contexts, and the analysis at the biological pathway level is limited. The possible mechanisms causing the dysregulated genes or pathways are rarely detailed, and their links with CFTR loss of function even less.

Most CF transcriptomic studies conclude with the statistical analysis at the gene level, with or without an ORA applied to the lists of DEG (see details in chapter 2). In these studies, the pathways found enriched in DEGs belong generally to three categories: CFTR proteostatis pathway, signal transduction and immune response/inflammatory pathways (see [Ideozu, 2019] for a review of defective pathways found from transcriptomic data). The inflammatory response is then rarely tackled as a main defect in CF but very often associated with infection. It would be interesting to study these data with FCS methods, using the whole gene expression data instead of ORA approaches, because FCS methods do not rely on an arbitrary thresholds (e.g., expression fold change), conversely to ORA approaches.

To date, the only study specifically dedicated to the systemic analysis of CF transcriptomic data with the intention of mechanistic understanding is the one conducted by Hodos and colleagues in 2020 [Hodos, 2020]. They performed a meta-analysis of transcriptomic data from both original microarray experiments and public sources. They studied four categories of experiments: one that compared CF vs. non-CF expression in human tissues, and three types of *in vitro* rescue strategies: low-temperature rescue, RNAi-based rescue, and chemical rescue via C18, an analog of the CFTR corrector lumacaftor (VX-809). Systematic comparison of these datasets yielded a core signature of the CF disease phenotype and two core signatures associated with F508del-CFTR rescue. Additionally, 60 gene sets associated with CF or CFTR were compiled from 34 publications leading to the CFG (CFTR functional Genomics) Library, i.e. genes with a functional effect on CFTR. Each core signature was then analyzed by GSEA and also compared to the CFG genes.

This integrative analysis suggested altered activity of SGK1 and EGR1 in CF cells, and also pointed at potential downstream effects on CFTR. The transcriptomic signatures suggested that C18 and the other rescue interventions act via distinct mechanisms. Even if this study is the most integrative and in-depth on the study of CF transcriptomic data, no network approach has been undertaken.

4.2.2 Network-based approaches

Network-based approaches for CF are quite diverse, although few in number. As mentioned in chapter 2, biological networks are multiple and it is important to choose the appropriate one that suits the most the question under study.

CF biological networks have been mainly created directly from data, and especially from DEGs. Two types of networks have been preferred so far: co-expression networks or PPI networks (see chapter 2 for their definition).

PPI networks have been built from DEGs between CF and control cell lines induced by viruses such as bacterial virulence factor [Mayer, 2012] or rhinovirus [Ling, 2020]. Both studies used the InnateDB which store undirected interactions to link the DEGs [Breuer, 2013]. These studies identified modules in the PPI networks before applying ORA approaches to the proteins of the module. Besides, two other PPI networks were built respectively from a study on rectal epithelium and from a study on bronchial epithelial with the STRING Database [Szklarczyk, 2019; Faria Poloni, 2021]. The results suggested that the F508del-CFTR promoted tissue-specific pathways in CF patients.

A study by Strub et al investigated strategies of *in vitro* rescue of CFTR with gene co-expression networks [Strub, 2021]. Networks were built based on DE of genes compared to baseline condition. The analysis of the networks enabled to identify several pathways differentially activated and several genes, including CHRUC1, GZF1 and RPL15, whose knockdown partially restored CFTR function. In the same spirit, a study by Pineau et al focused on the construction of co-expression network from blood samples data to identify genes and pathways that modulate the associated comorbidities [Pineau, 2020].

The use of networks in these cases does not differ much from the pathway-based approaches presented in the previous section, as it does not lead to the proposal of a mechanistic hypothesis to explain pathway dysregulations.

Network biology has also been applied to investigate CFTR interactors. To this end, these approaches can sometimes go further in suggesting hypotheses on the links between CFTR and molecular dysregulations.

A study by Loureiro et al was centred on some interactors of CFTR known to be involved in the stabilization process of CFTR (called *stability factors* hereafter): namely EPAC, NHERF1, EZRIN and SYK [Loureiro, 2019]. After defining 5 PPI networks corresponding to the PPI networks for each CFTR interactor, they mapped these individual networks on a full human PPI network, and explored how strongly the identified PPI networks were connected within the complete PPI network (number of overlapping proteins, number of direct interactions between set members and number of common direct neighbours between these sets). They identified shared neighbours for the CFTR interactor sets, and built a subnetwork including these specific neighbours, CFTR itself and the stability factors. A total of 194 proteins were identified by statistical analysis (bridge score) as candidates for experimental validation of CFTR PM stability modulation.

Topological analysis was also used in a study by Mayer et al, to understand the transcriptional changes induced by an Innate Defense Regulator (IDR) [Mayer, 2012]. They particularly analysed the links between CFTR interactors and TLR5 signalling pathways, influenced by the IDR. They mapped the genes in these two gene sets on human PPI network. The CFTR and TLR5 networks separated into two distinct entities without any evident connections. When genes differentially expressed following treatment with the IDR were merged into the network, CFTR and TLR5 networks became connected via PRKAA1 (AMPK), HSPB1 (Hsp27), and AKT1. The functional significance of these newly discovered interconnections was validated by wet labs experiments, which showed that the treatment did not reduce inflammatory responses in CF cells pre-treated with an AMPK activator or an AKT inhibitor. To date, and to our

knowledge, this study is the only one in which the combination of quantitative analysis of transcriptomic data and network approaches suggested mechanistic hypotheses to link CFTR to CF phenotypes (in this case, inflammation via TLR signalling).

Finally, two very recent systems biology initiatives have led to the construction of knowledge maps for CFTR interactomes. The first one from the Disease Map community [Mazein, 2018] integrated information from the literature and highlighted the complexity of the mechanisms around CFTR at different locations in the cell in WT vs mutated conditions [Pereira, 2021]. The so-called "CyFi-MAP" was implemented following the SBGN Activity Flow [Novère, 2009]. The second initiative called "CFTR Lifecycle Map" [Vinhoven, 2021] is composed of two maps, a core map manually curated from small-scale experiments in human cells, and a coarse map including direct and indirect interactors identified in high-throughput (HT) efforts (including [Wang, 2006; Pankow, 2015; Matos, 2018]). These maps are also written in the SBGN format, adhering to the Process Description language. The interactors retrieved from the literature and integrated to the core map were also mapped on a human PPI network. Basic network analysis was done on the obtained PPI subnetwork leading to the identification of hubs (proteins with highest degrees) but no mechanistic interpretation has followed. The same methodology (mapping CFTR interactors to a human PPI network) was also undertaken by [Sahrawat, 2013].

This review of existing studies highlights the importance and the need for mechanistic interpretation of the existing data. All these studies have mainly focused on the molecular mechanisms centred around CFTR. We propose here to go a step further and use the information extracted from studies on the CFTR interactome studies to enrich the mechanistic knowledge behind CF dysregulations.

Chapter 5

From CFTR to a CF signalling network: A systems biology approach to study Cystic Fibrosis

Contents

Abstract

Cystic Fibrosis (CF) is a monogenic disease caused by mutations in the gene coding the Cystic Fibrosis Transmembrane Regulator (CFTR) protein, but its overall physiopathology cannot be solely explained by the loss of the CFTR chloride channel function. Indeed, CFTR belongs to a yet not fully deciphered network of proteins participating in various signalling pathways. We propose a systems biology approach to study how the absence of the CFTR protein at the membrane leads to perturbation of these pathways, resulting in a panel of deleterious CF cellular phenotypes. Based on publicly available transcriptomic datasets, we built and analyzed a CF network that recapitulates signalling dysregulations. The CF network topology and its resulting phenotype was found to be consistent with CF pathology. Analysis of the network topology highlighted a few proteins that may initiate the propagation of dysregulations, those that trigger CF cellular phenotypes, and suggested several candidate therapeutic targets. Although our research is focused on CF, the global approach proposed in the present paper could also be followed to study other rare monogenic diseases.

Résumé

La mucoviscidose est une maladie monogénique causée par des mutations du gène codant la protéine CFTR (Cystic Fibrosis Transmembrane Regulator), mais sa physiopathologie globale ne peut pas être expliquée uniquement par la perte de la fonction du canal chlorure CFTR. En effet, CFTR appartient à un réseau de protéines qui n'a pas encore été entièrement déchiffré et qui participe à diverses voies de signalisation. Nous proposons une approche de biologie des systèmes pour étudier comment l'absence de la protéine CFTR à la membrane conduit à une perturbation de ces voies, résultant en un panel de phénotypes cellulaires délétères de la mucoviscidose. Sur la base d'ensembles de données transcriptomiques accessibles au public, nous avons construit et analysé un réseau qui récapitule les dérégulations de la signalisation. La topologie du réseau et le phénotype qui en résulte se sont avérés cohérents avec la pathologie de la mucoviscidose. L'analyse de la topologie du réseau a mis en évidence quelques protéines susceptibles d'initier la propagation des dérégulations qui déclenchent les phénotypes cellulaires de la maladie, et a suggéré plusieurs cibles thérapeutiques candidates. Bien que notre recherche soit axée sur la mucoviscidose, l'approche globale proposée dans cet article pourrait être également suivie pour étudier d'autres maladies rares monogéniques.

5.1 Preface

The previous chapter presented studies that address molecular mechanisms in CF through the lens of systems biology. Although they constitute a rich resource for CF, providing detailed knowledge of CFTR interactome, they have mainly focused on the molecular mechanisms centred around CFTR. This thesis does not tackle the same issue: we want to provide a global understanding of how the absence of functional CFTR leads to the overall dysregulations. Therefore, we are not only interested in the life cycle of CFTR, but in all the possible phenotypes observed in CF cells and their links with CFTR.

The first idea was to build a network, in which all CF molecular dysregulations are merged together into a single network. An essential step in the construction of biological networks is the formulation of the biological question to which they respond. The translation of the question into a network requires choices about what to include and what not to include. The two main questions are what type of networks we want to build and how we are going to build it.

5.1.1 What type of biological networks ?

Our goal is to provide a mechanistic understanding of some CF phenotypes, i.e. understand the causal link of the absence of CFTR on the cellular phenotypes of the disease. The dysregulations can then be modelled as cascades of directed interactions between proteins, from CFTR as the starting point to the effector proteins of the CF phenotypes as the end point. Each interaction to add should correspond to physical interactions with a biological function. Various types of biological networks are encountered in systems biology studies, such as *genetic interaction networks*, *gene regulatory networks*, *co-expression networks*, *protein-protein interaction (PPI) networks*. However, we chose to build a signalling network, because our goal was to relate CFTR to overall signalling pathways dysregulations in CF [Ideozu, 2019].

5.1.2 How to build the CF network?

From a practical point of view, there are two different ways to build biological networks: the first one is based on results from the scientific literature, and the second one is based on the direct analysis of HT sequencing data obtained from patients cells or model CF cell-lines (See chapter 2). In fact, building a signalling network recapitulating CF dysregulations exclusively from prior knowledge revealed itself to be a challenge. Very little information is available in signalling knowledge databases with respect to CF, and particularly to CFTR. Indeed, prior studies showed that signalling mechanisms have been much more studied in specific contexts, such as cancer [Magalhães, 2022], compared to others, leading to a small number of proteins more extensively studied [Kustatscher, 2022]. This bias obviously does not favor CF, considered as a rare disease.

Conversely, adopting data-driven approaches in building models allowed us to overcome over-abstraction and over-generalisation when building the CF network.

What data to use ?

(Phospho-)proteomic data are the natural choice for describing pathway activity [Szalai, 2020]. However, transcriptomics data are frequently used in PB methods due to their much higher abundance, and their ability to detect transcripts at the genome scale, whereas proteomic data detect a few thousands of proteins. Therefore, transcriptomic data allow the detection of a larger spectrum of dysregulations than most other types of omics modalities. Concerning CF, transcriptomics data have also been most widely generated and analysed than other modalities (see chapter 4). Therefore, we focused exclusively on this modality, although the resulting network could be refined in the future based on other types of omics data.

We used publicly available bulk transcriptomic data on the most prevalent mutation, F508del, for which the most data and prior knowledge are available. Including studies on other mutations, although interesting in itself, would not have allowed us to build a consistent model. We only retained samples from human airway epithelial cells (HAEC). Indeed, the most severe symptoms of CF are in the lungs (see section 1.2.1) and epithelial cells are affected by CFTR dysfunction, and discarded studies on tissues. Including these data in our analysis would require to investigate tissue-specific pathways, which was beyond the scope of the present thesis.

We did not discard datasets generated from nasal samples. We acknowledge that the transcriptomic datasets might be different from the bronchial ones, but sampling patients from nasal epithelial is less invasive which makes them the model of choice for numerous studies. Besides, a previous meta-analysis study of CF transcriptomic data showed an agreement between dysregulated genes in nasal and bronchial cells, suggesting nasal epithelium as a good surrogate for the CF airway [Clarke, 2013]. Not including them would have further reduced the number of studies to be analysed.

Finally, CF transcriptomics datasets are very heterogeneous in terms of tissue samples, cell culture and technology. Neither statistical methods based on samples variance, such as rROMA (presented in chapter 3), nor integrative methods, which combine all datasets, are suitable. We thus needed to find alternative approaches to take into account data arising from different technologies.

In the research paper presented in this chapter, we performed a meta-analysis of CF transcriptomic datasets at the level of biological pathways to retrieve CF molecular dysregulations. We proposed to take advantage of the knowledge gathered on the CFTR interactome over the past ten years to link CFTR to the dysregulated signalling pathways, and gathered these pathways into a single signalling network. Finally, the topological analysis of the network highlighted a few proteins that may initiate or propagate dysregulations from CFTR into the network, and explain the observed CF phenotypes.

This work was made in collaboration with Loredana Martignetti, Matthieu Cornet, Mairead Kelly-Aubert, Isabelle Sermet-Gaudelus, Laurence Calzone and Véronique Stoven. It was submitted to biorXiv in October 2023, and has been submitted to a journal for peer review. In the following section, the article is transcribed as submitted.

The project was also presented as a talk at the *Conférence des Jeunes Chercheurs sur la mucoviscidose* in February 2020, and preliminary results were presented as a talk

in a panel session at the *European Young Investigators Meeting in Cystic Fibrosis* in Paris, France, in March 2022, and as a poster at the *European Conference on Computational Biology (ECCB)* in Barcelona, Spain in September 2022. Finally, the network construction and its topological analysis were presented as a poster in the *International signalling Workshop (ISW)* in Visegrad, Hungary in July 2023, where I won an award for best poster in the "Disease Modelling" panel.

5.2 From CFTR to a CF signalling network: A systems biology approach to study CF

5.2.1 Introduction

Cystic fibrosis (CF) is the most common life-limiting autosomal disease in the Caucasian population, affecting about 162.000 patients worldwide, of which 105.000 are diagnosed [Guo, 2022]. It is caused by mutations in the *CFTR* gene encoding for the cystic fibrosis transmembrane conductance regulator (CFTR) protein, a chloride ion channel expressed at the apical membrane of polarized epithelial cells [Seibert, 1997]. More than 2000 mutations in *CFTR* have been reported, but the deletion of the F508 amino-acid (F508del) is present in 70% of the mutated alleles in the Caucasian population, and most of the mutations lead to compromised transepithelial anion conductance [Veit, 2016]. Various organs are affected in CF, but the most severe symptoms are in the lungs, where the defective chloride transport leads to the dehydration of surface mucus, chronic bacterial infection, and inflammation, causing lung tissue damage and ultimately, respiratory insufficiency.

However, CF symptoms not only result from the loss of CFTR-mediated anion conductance, but also from perturbations of other CFTR-dependent biological functions [Hanssens, 2021]. Indeed, CFTR belongs to a protein-protein interactions (PPI) network [Pereira, 2021; Farinha, 2021], and the absence of CFTR may perturb its direct or indirect interactors, and propagate dysregulations towards various biological pathways in which these interactors play a role. In agreement with this idea, studies on textitCFTR -/- knockout mice [Crites, 2015], *CFTR* -/- knockout piglets [Fleurot, 2022], and cell lines in which CFTR is inactivated by the CRISPR/Cas9 technology [Hao, 2020] have reported that the absence of CFTR affects cell signalling and transcriptional regulation. These dysregulations may explain various and apparently unrelated cellular phenotypes, including uncontrolled pro-inflammatory response [Jacquot, 2008], unbalanced oxidative stress with increased reactive oxygen species [Jeanson, 2012], impaired epithelial regeneration [Conese, 2021], or perturbation of cell junctions and cytoskeleton [Pankonien, 2022].

To explain these seamlessly unrelated phenotypes, we propose to use a systems biology approach for CF, where the two aims are (1) to explore how the absence of CFTR can be functionally related to the signalling dysregulations that ultimately lead to CF cellular phenotypes; and (2) to suggest new therapeutic targets that may modulate these phenotypes.

Indeed, systems biology approaches provide tools for building network models to reason on complex systems. Subsequent topological analysis or dynamic mathematical models performed on these networks allow to study how different biological components of the networks interact to produce phenotypic properties, which is relevant to the questions at hand.

Systems biology approaches have seldom been implemented in monogenic diseases, but have been widely used in cancer, often referred to as a network disease [Hornberg, 2006], where intricate processes contribute to the emergence of unexpected and often non-intuitive phenotypes. Very few contributions have been devoted to systems biology approaches of CF. Previous studies have focused on the construction of the CFTR interactome that distinguishes PPI networks involving wt-CFTR and those involving the

most frequent mutant F508del-CFTR [Pankow, 2015; Pereira, 2021]. The latter led to the construction of a navigable knowledge map, the CyFi-MAP, that integrates all proteins known to be involved in the processing, maturation, retention and degradation of wt-CFTR and F508del-CFTR. Although the CyFi-MAP represents a key contribution for the problem of rescuing F508del-CFTR, this map does not tackle the questions of interest in the present paper. Other studies highlighted links between CFTR and signalling pathways involved in the disease (see [Pankonien, 2022] for a review), but they did not provide a global view of how CFTR is linked to dysregulated molecular mechanisms and to CF phenotypes. Recently, transcriptomic data have been produced to identify differentially expressed genes in CF. These genes were connected within a PPI network, based on information available in PPI databases [Trivedi, 2023]. Although this network comprises genes that are consistent with current knowledge in CF, it does not contain CFTR, which prevents understanding the functional link between CFTR and the differentially expressed genes, or with CF cellular phenotypes.

To overcome the limitation of previous studies, in the systems biology approach proposed here, we build a comprehensive signalling network, called the CF network in the following, that recapitulates CF pathway dysregulations, using transcriptomic data available for CF and control patients and information available in biological pathway databases. As detailed below, we connected CFTR to this network based on PPI information. Analysis of the CF network topology allows to formulate hypotheses on key proteins and molecular mechanisms that functionally link CFTR to major CF cellular phenotypes, and to highlight potential targets that may counteract these phenotypes.

5.2.2 Results

Global approach to building the CF network

In systems biology, various networks can be built to represent different types of biological information, such as gene regulatory networks, genetic interaction networks, signal transduction networks, metabolic networks, PPI networks, or disease networks. There is no universal technique that can be followed to construct networks, and the choice of their representation needs to be adapted to the question of interest. In the present study, we wish to establish a link between the absence of CFTR and the overall signalling dysregulations leading to the cellular phenotypes that characterize CF. Therefore, we chose to build a CF network focusing on the signalling pathways that are perturbed in the disease, and where dysregulations in one pathway may affect other pathways. In order to avoid potential bias in the CF literature, we adopted a data-driven approach based on publicly available transcriptomic studies. We are aware that some CF phenotypes might arise from biological events that are not detectable in the transcriptome of CF cells, but we considered that gene expression data had the potential to capture some of the major molecular dysregulations present in CF cells. Our study relies on a meta-analysis of public transcriptomic datasets for CF respiratory epithelial cells and their Non-Cystic Fibrosis (NCF) control counterparts, allowing the identification of the signalling pathways dysregulated in CF. Based on information available in pathway databases, these dysregulated pathways share many common proteins, which allowed to connect them into a network. As detailed below, CFTR was absent from this network, because it did not belong to any of the differentially expressed

signalling pathways. However, we observed that several proteins of the network were also present in the CFTR PPI interactome, either as direct interactors of CFTR, or as indirect interactors of CFTR via a single intermediate protein. This important result was consistent with the assumption that CFTR direct interactors may be perturbed in CF, and initiate the propagation of dysregulations within the CF network.

The figure 5.1 summarizes the global approach followed in the present study.

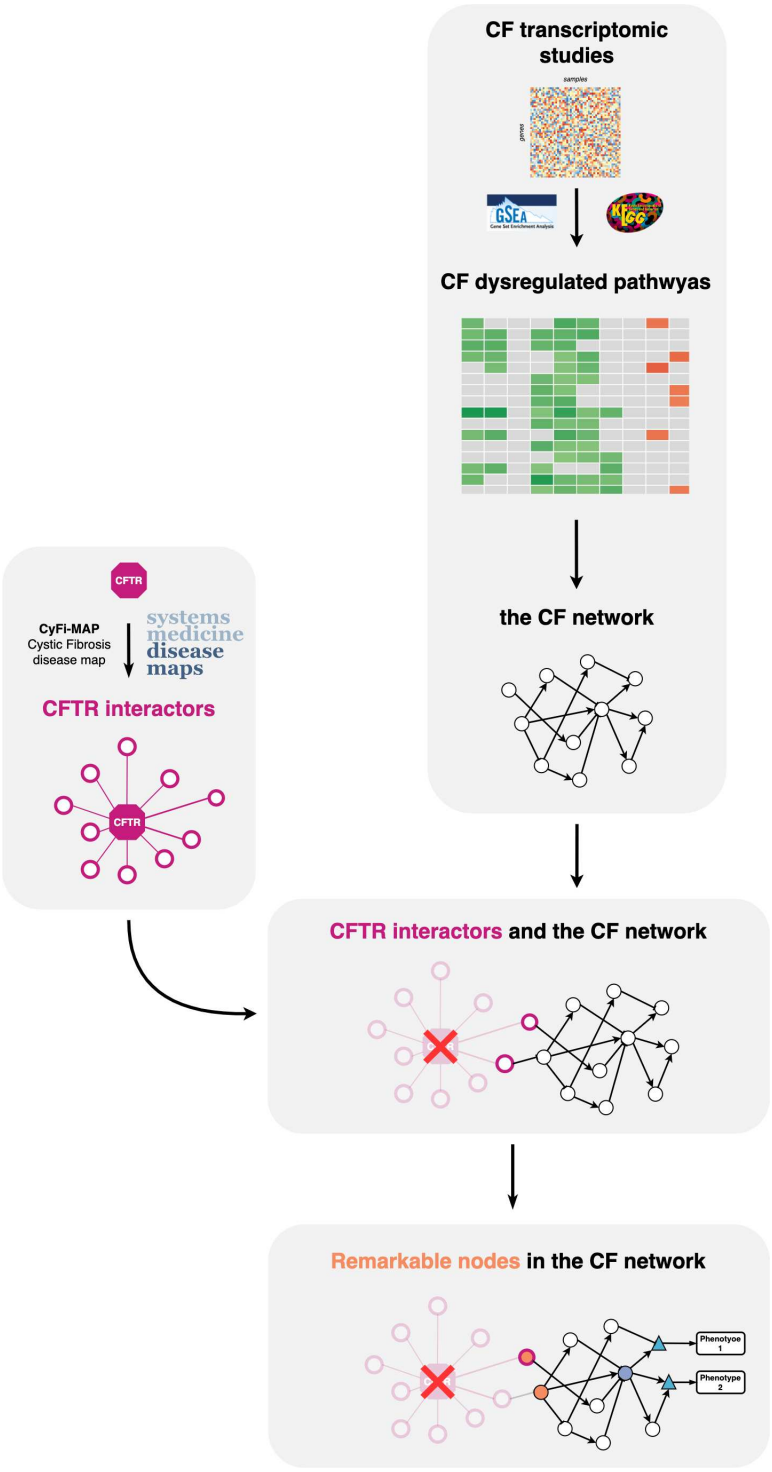


Figure 5.1 – Global approach followed to build the CF network.
A meta-analysis of CF transcriptomic data allowed the identification of dysregulated pathways and the construction of the corresponding CF network. This network comprises known CFTR interactors that can be viewed as source nodes initiating the propagation of dysregulations.

Selection of publicly available transcriptomics data

Many transcriptomic studies have been performed in CF over the last 15 years [Ideozu, 2019]. However, these data suffer from a few limitations that are obstacles to improve our understanding of CF. First, they consider a wide range of cell types, including native nasal or bronchial cells, primary cultures of these cells, whole blood, peripheral mononuclear cells, leukocytes, or immortalized cell lines. Therefore, comparison between studies to identify common key molecular determinants can lead to inconsistent results. Then, compared to studies on more common diseases such as cancer, most of CF transcriptomic studies have very few samples per condition (disease and control), decreasing the statistical power of these datasets when analyzed alone. Finally, these studies rely on various experimental biological models and transcriptomic technologies which rarely lead to consistent results between studies [Clarke, 2013], particularly when the analyses are performed at the gene level.

Study	Tissue sample	Cell culture	Technology	nb CF,NCF	Dataset	References
Verhaeghe	Tracheal	Cell line	Microarray	3, 3	E-MEXP-980	[Verhaeghe, 2007]
Ogilvie (Nasal)	Nasal	Primary culture	Microarray	27, 18	E-MEXP-436	[Ogilvie, 2011]
Ogilvie (Bronchial)	Bronchial	Primary culture	Microarray	8, 17	E-MEXP-436	[Ogilvie, 2011]
Voisin	Nasal	Primary culture	Microarray	5, 5	GSE40445	[Clarke, 2013]
Clarke	Bronchial	Cell line	Microarray	3, 3	GSE39843	[Voisin, 2014]
Balloy	Bronchial	Primary culture	RNA-Seq	4, 3	ERP010372	[Balloy, 2015]
Zoso	Bronchial	Primary culture	RNA-Seq	7, 6	GSE127696	[Zoso, 2019]
Ling	Bronchial	Primary culture	RNA-Seq	7, 5	GSE138167	[Ling, 2020]
Saint-Criq (UNC)	Bronchial	Primary culture	RNA-Seq	3, 2	GSE154905	[Saint-Criq, 2020]
Saint-Criq (SC)	Bronchial	Primary culture	RNA-Seq	3, 3	GSE154905	[Saint-Criq, 2020]

Table 5.1 – List of the 10 datasets considered in the meta-analysis, indicating the number of CF and NCF samples in each study.

To try and overcome these limitations, we focused on studies considering only samples from human Airway Epithelial Cells (hAEC hereafter), i.e., bronchial, tracheal, or nasal cells. Indeed, functional modifications in these cells are expected to reflect some of the most severe symptoms in the lung. We included studies of cell lines or primary cultures, in order to gather a statistically significant number of samples, because as shown in Table 5.1, each dataset comprises a very limited number of samples. We also focused on studies on the F508del mutation, for which most data are available. We discarded two studies ([Virella-Lowell, 2004] and [Rehman, 2021]) that provide transcriptomic data for other mutations, because the corresponding cells could display disparities with respect to F508del cells. We retrieved from the literature all the CF transcriptomic studies with publicly available data that matched these criteria (see Methods subsection), which led to 10 CF transcriptomic datasets shown in Table 5.1.

The studies are still heterogeneous in terms of tissue sample (bronchial, tracheal or nasal), cell culture type (cell-line or primary culture) and transcriptomic technology (micro-array or RNA-Seq). However, we kept the 10 studies in order to improve statistical significance, because the numbers of samples per condition are very small in all studies: the median number of samples was 5 for disease (CF) and control (NCF) conditions.

Meta-analysis of transcriptomic studies at the level of biological pathways

The most straightforward way to analyse transcriptomic data is to identify Differentially Expressed Genes (DEGs), and to search for biological pathways enriched in these DEGs. This approach failed in the present meta-analysis, because the number of DEGs common to at least 3 out of 7 studies was too small to be enriched in any pathway, even though many reference pathway databases were considered (the Hallmark gene sets from the the MSigDB Database [Liberzon, 2015], the Pathway Interaction Database (PID) [Schaefer, 2009], the KEGG database [Kanehisa, 2021]). In fact, it has become clear that, in complex diseases, identification of pathway dysregulations based on DEGs is not optimal and does not provide robust results [Wang, 2010].

Therefore, the meta-analysis was conducted at the pathway level. Many methods have been proposed to capture pathway dysregulations when they do not appear clearly based on enrichment from lists of DEGs [Martignetti, 2016; Landais, 2023; Schubert, 2018; Vaske, 2010]. In the present study, we used the Gene Set Enrichment Analysis (GSEA) [Subramanian, 2005] approach. GSEA was performed separately on each dataset identified as over-activated or under-activated signalling pathways in hAEC CF cells, based on the complete expression matrix of CF and NCF samples, and taking into account the expression level of all genes belonging to the same pathway. We used pathway definitions provided by the KEGG pathway database [Kanehisa, 2021], because this database provides graphical pathway representations that also include phenotypes, which helped the analysis of the CF network, as detailed in Section 5.2.2. We tested 131 KEGG biological pathways, and Differentially Expressed Pathways (DEPs, hereafter) were identified according to a adjusted p-value lower or equal to 0.25, as detailed in Section 5.2.4. The number of up- and down-regulated pathways for each dataset is provided in Table 5.2

Study	nb detected genes	nb CF,NCF	nb up-regulated pathways	nb down-regulated pathways
Verhaeghe	22880	3, 3	11	2
Ogilvie (Nasal)	19880	27, 18	0	0
Ogilvie (Bronchial)	19880	8, 17	33	0
Voisin	13144	3, 3	10	0
Clarke	14118	5, 5	40	2
Balloy	39430	4, 3	24	2
Zoso	18846	7, 6	2	0
Ling	28138	7, 5	1	6
Saint-Criq (UNC)	39434	3, 2	11	0
Saint-Criq (SC)	39432	3, 3	1	9

Table 5.2 – Number of detected genes, CF and NCF samples, tested pathways and dysregulated pathways per study with a corrected p-value < 0.25 and a $|\log2FC| > 1$ thresholds at the gene scale.

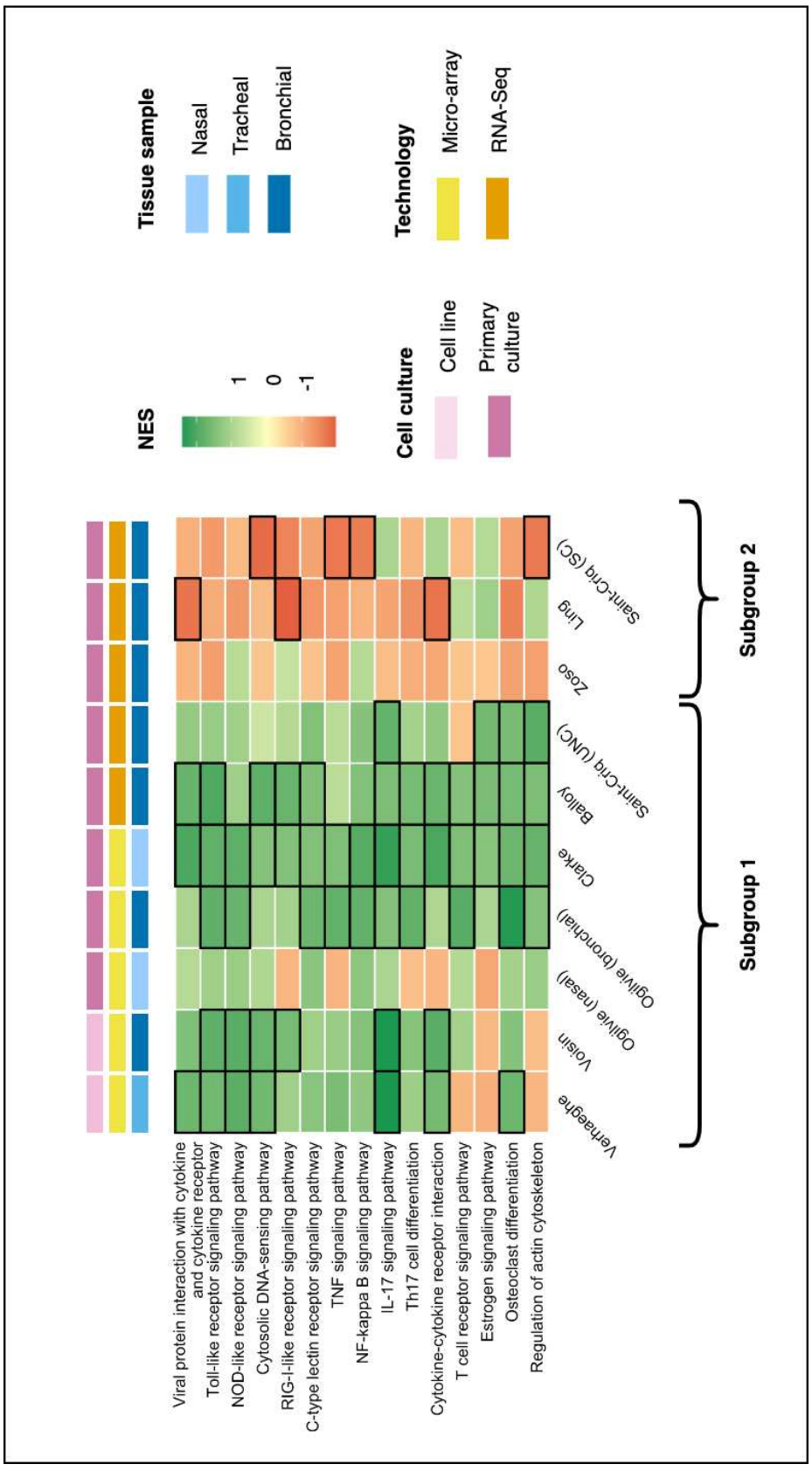


Figure 5.2 – Heatmap of the GSEA Normalized Enrichment Scores (NES) of the biological pathways differentially expressed in at least 3 studies.
The datasets can be clustered in two subgroups based on their NES: Subgroup 1 and Subgroup 2, respectively in agreement and in contradiction with CF physio-pathology. Black boxes around the tiles represent the pathways significantly differentially expressed in the corresponding dataset.

The analysis of DEPs showed that 15 of the 134 biological pathways tested were differentially expressed in at least 3 studies. However, a closer analysis highlighted discrepancies between studies. As shown in the heatmap presenting the GSEA Normalised Enrichment Score (NES) (Figure 5.2), for these 15 common DEPs, the 10 datasets can be gathered into 2 subgroups: subgroup 1 comprising 7 datasets in which common DEPs tend to be up-regulated, while they tend to be down-regulated in subgroup 2 comprising the 3 other datasets. This appeared somewhat puzzling. Our hypothesis is that datasets belonging to subgroups 1 or 2 arise from studies in which the differentiation media used for the primary cultures did not favor the same cell type, and therefore, should not be analyzed together.

This was confirmed by the Saint-Criq (UNC) and Saint-Criq (SC) datasets (see Table 5.1), belonging respectively to subgroups 1 and 2, where it was shown that the UNC and SC differentiation media (two common differentiation media used on CF and non-CF epithelia) significantly impact cell lineage in primary cultures of CF hAEC, and consequently, the resulting transcriptomic profiles [Saint-Criq, 2020]. In this study, it was shown that the UNC medium promoted differentiation into club and goblet cells, while the SC medium favored the growth of ionocytes and ciliated cells. Consistent with this result, the Ling transcriptomic dataset, which belongs to subgroup 2, was also obtained from primary cultures of CF and NCF airway epithelia that were differentiated into ciliated pseudo-stratified airway cells [Ling, 2020]. Datasets from subgroup 2 appeared in contradiction with the main CF phenotypes. In particular, the *TNF-α signalling pathway* or *NF-κB signalling pathway* are down-regulated in this subgroup, although the over-activation of these pathways is a well-known feature of CF disease. Therefore, we only considered the 7 datasets belonging to subgroup 1 for further analysis.

In this subgroup, the transcriptomic analysis appears to be highly consistent, since among the 15 DEPs common to at least 3 studies, 5 are up-regulated in CF vs NCF samples in 4 studies (*NOD-like receptor signalling pathway*, *Cytosolic DNA-sensing pathway*, *Cytokine-cytokine receptor signalling pathway*, and *Regulation of actin cytoskeleton*), 2 are up-regulated in CF vs NCF samples in 5 studies (*Osteoclast differentiation* and *Toll-like receptor signalling pathway*), and the *IL-17 signalling pathway* is up-regulated in CF vs NCF in 6 studies.

Overall, the 15 DEPs common to at least 3 studies are in agreement with various known aspects of CF disease, which confirms that our analysis did capture relevant information about CF. In particular, besides the TNF-α and NF-κB signalling pathways well known to be up-regulated in CF, the IL-17 pathway contributes to CF lung disease [Hsu, 2016], the differentiation of osteoclast is perturbed in CF [Dumortier, 2021], the Toll-like receptor signalling pathway modulates function, inflammation and infection of lung in CF [Kosamo, 2020; Curutiu, 2018; Fleurot, 2022], and CFTR plays a role in cell junction and actin cytoskeleton organization [Pankonien, 2022].

Building the CF network

The 15 individual DEPs of the KEGG database provide interesting information about what is dysregulated in CF, but a lot of these pathways are partially redundant and show a high overlap of genes and interactions, indicating that they are highly intertwined. A dysregulation in one of these pathways will have a consequence in

another pathway. To study the connection between them, we propose to merge them into a single network called the CF network.

The DEPs were extracted with the OmniPathR package [Türei, 2016] and curated, as described in Section 5.2.4. The rules that were used to build and clean this network are detailed in Section 5.2.4. The network, comprising 330 nodes and 529 interactions, is not fully connected: it contains one main component including 317 nodes connected by 515 interactions, and two small additional components that are non connected to the main component, and called unconnected components hereafter (See Figure 5.3). The overall network can be accessed as a Cytoscape session, in the sysbio-curie/CFnetwork_cystoscape repository for further analysis.

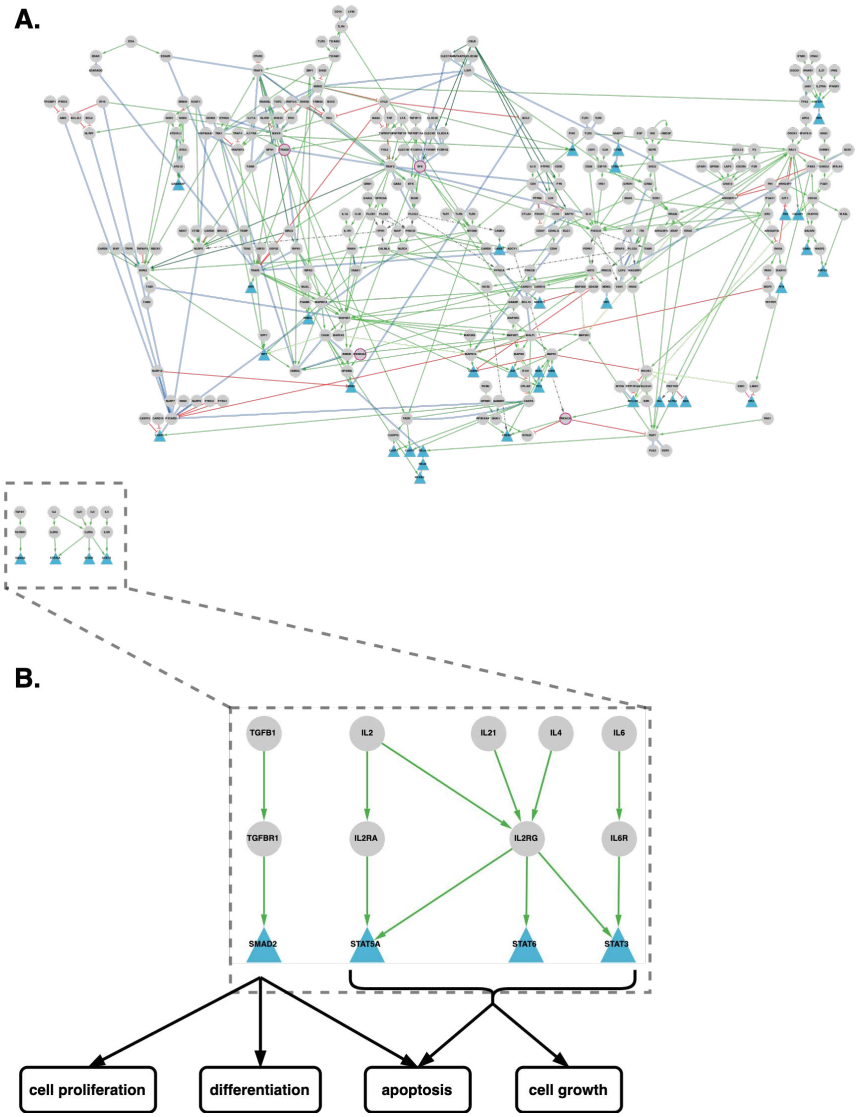


Figure 5.3 – The CF network.
(A): The main component comprises 317 nodes connected by 517 interactions and two small unconnected components shown in (B): the two unconnected components correspond to the TGFβ and the JAK-STAT signalling pathways. The cellular phenotypes triggered by the sink nodes of the two components are surrounded by black contours.

Identification of CFTR interactors in the CF network

It is striking to note that CFTR does not belong to any of the 15 DEPs, and therefore, is not part of the network. In fact, CFTR is present in only 7 biological pathways of the KEGG database (*ABC Transporters*, *cAMP signalling pathway*, *AMPK signalling pathway*, *tight junction*, *Gastric and acid secretion*, *pancreatic secretion* and *bile secretion*), but these pathways did not belong to the DEPs.

Therefore, we searched for the presence of CF network proteins in the network of

proteins reported to be involved in protein-protein interactions (PPI) with wt-CFTR or F508del-CFTR [Pereira, 2021]. Indeed, according to the CyFi-MAP, 4 direct interactors of wt-CFTR but not of F508del-CFTR (CSNK2A1, PRKACA, SYK and TRADD) belong to the CF network. Furthermore, 4 additional proteins (EZR, SRC, PLCB1/3) present in the network interact with wt-CFTR (but not with F508del-CFTR) through a single intermediate protein. Figure 5.4 shows these 8 proteins, their intermediates and their interactions with CFTR. The presence in the CF network of 8 first or second neighbours in the CFTR interactome is an interesting result in favour of our assumption that CFTR interactors may propagate functional dysregulations into the network.

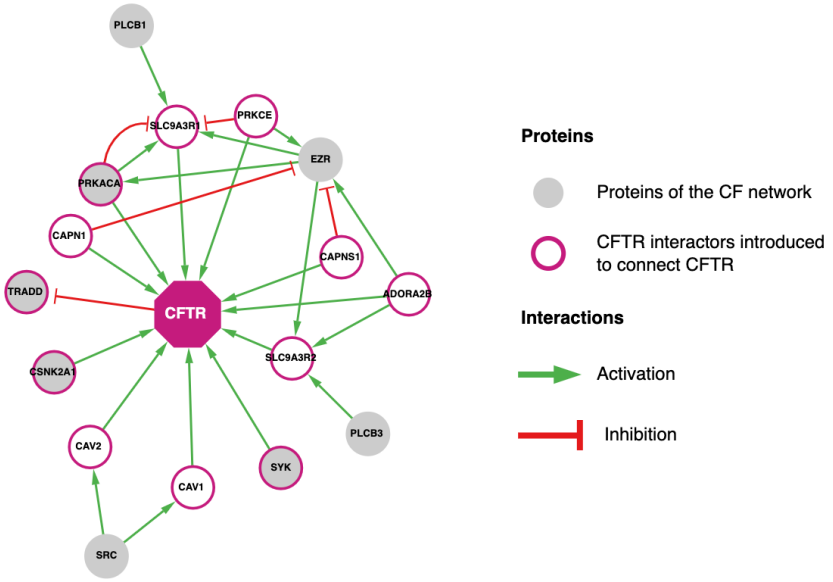


Figure 5.4 – CFTR interactors in the CF network: Known protein-protein interactions involving CFTR interactors in CFTR PPI.

Analysis of the CF network

Extensive interpretation of this large network, which contains rich but complex information, is beyond the scope of the present paper. However, we will investigate how analysis of its topology can help tackle the two questions of interest: how the absence of the CFTR protein at the membrane leads to CF cellular phenotypes, and how therapeutic targets can be suggested from this network.

Topological description of the CF network The final CF network comprises 330 proteins and 529 interactions. Interestingly, CFTR interactors are present only in the main component, because according to the CyFi-MAP, it would not have been possible to link CFTR to proteins of the two small unconnected components without adding a large number of intermediate nodes. One of the two unconnected components contains 10 proteins and 9 interactions, and corresponds to cascades of the JAK/STAT

signalling pathway. The other contains 3 proteins and 3 interactions, and corresponds to a cascade of the Transforming Growth Factor Beta (TGFβ) signalling pathway. In the present section, we will focus on the main component of the CF network, and the two unconnected components will be discussed in Section 5.2.2.

The topological description of the main component will be organized around three types of remarkable nodes: (1) the source nodes, i.e., CFTR first or second neighbours that were used to connect CFTR to the network, as described in Section 5.2.2; (2) the sink nodes, i.e., the nodes from which no edge leaves in the network, and whose activation finally triggers their associated phenotypes (for example, transcription factors are typical sink nodes); (3) the hubs, i.e. the nodes with high betweenness centrality, through which the flow of information that passes is high. Figure 5.5 illustrates where these remarkable nodes stand within the network's topology.

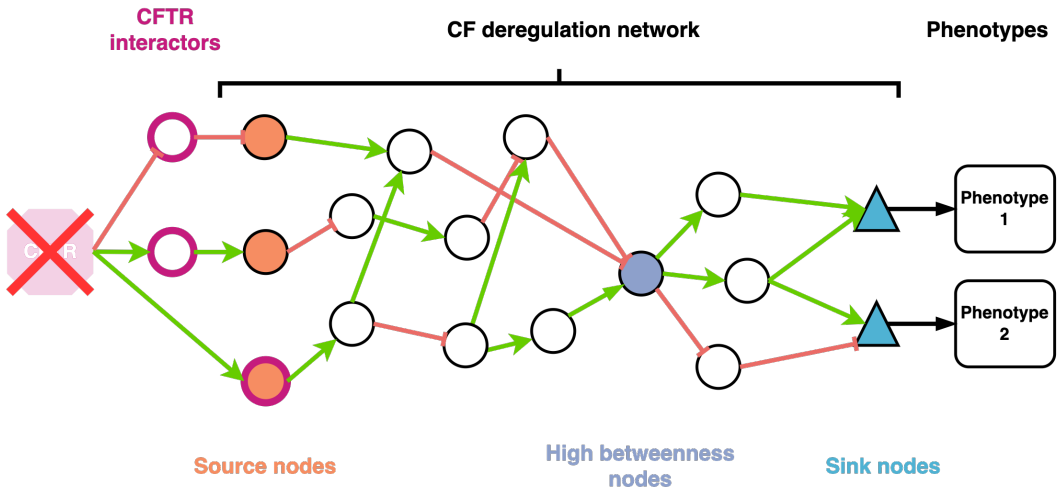


Figure 5.5 – Illustration of propagation of dysregulation and remarkable nodes in the CF network.

The source nodes (orange disks) are CFTR interactors or connected to CFTR interactors via a single intermediate protein (magenta circles). Nodes with high betweenness centrality (purple disks) are proteins through which much information flows within the network. Sink nodes (blue triangles) modulate their corresponding phenotypes.

Source nodes and initiation of dysregulations

According to the CyFi-MAP, 8 first or second neighbours of wt-CFTR interactors whose interactions are lost with F508del-CFTR are present in the CF network: **CSNK2A1**, **EZR**, **PLCB1**, **PLCB3**,**PRKACA**, **SRC**, **SYK** and **TRADD**. In the absence of CFTR, these 8 proteins can be viewed as source nodes that may initiate dysregulations that subsequently propagate within the network and finally reach the sink nodes (see Figure 5.4).

Perturbations of some of these source nodes in CF cells, or their role in CF cellular phenotypes, are sustained by various studies:

— **CSNK2A1**, also known as CK2 (casein kinase 2), is strongly overactivated in CF vs wild-type cells [Venerando, 2011].

— Cellular levels of **TRADD** are controlled by its lysosomal degradation in a wt-CFTR-dependant manner, and this regulation is lost with F508del-CFTR and G551D-CFTR [Wang, 2016].

— **SRC** was shown to be overexpressed and overactivated in CF cells [Massip Copiz, 2016].

— **PLCB3** is a known CF modifier gene, for which the loss of function S845L variant is associated with a mild progression of the pulmonary disease and a reduction of *Pseudomonas aeroginosa*-induced IL8 release. This indicates that PLCB3 plays a role in the inflammation phenotype in CF [Rimessi, 2018].

— The active form of ezrin (**EZR**) is mainly located in the apical region of wild type airway epithelial cells, while in their CF counterparts, it is diffusely expressed in its inactive state in the cytosol [Favia, 2010; Wu, 2019].

— The **SYK** and **PRKACA** kinases play key roles with respect to CFTR, since the former negatively regulates the amount of CFTR at the membrane through phosphorylation at Y512 [Mendes, 2011], while the latter is a well-known regulator of the CFTR chloride channel conductance [Egan, 1992], but their implication as propagators of dysregulations has not been investigated yet.

Sink nodes and CF phenotypes

There are 35 sink nodes in the main component of the CF network that are reached from each of the 8 source nodes. The full list of sink nodes and their associated phenotypes are given in the Supplementary file 2. Among them, we can cite:

— **NFKB1**, **NFKB2**, **RELA** and **RELB** are part of the NF-κB complex, a transcription factor that can be activated by various stimuli such as cytokines, oxidant radicals, bacterial or viral products. It controls the expression of pro-inflammatory genes, and is related to various phenotypes including inflammation and cell survival/proliferation.

— **FOS** and **JUN** are two sub-units of the AP-1 transcription factor activated by the MAPK signalling pathways, and are associated with inflammation and proliferation phenotypes.

— **CASP3** and **CASP7** caspases are the effectors of apoptosis.

— **CASP1** is a caspase known to be the effector of pyroptosis, a highly pro-inflammatory cell death mechanism.

— 10 sink nodes belong to the regulation of actin cytoskeleton pathway, including **ACTN4**, **ARPC5**, **PFN**, **MYL12B** and **VCL**. These nodes are associated to various phenotypes related to cytoskeleton, including focal adhesion, adherens junction, and actin polymerisation.

— **IRF1**, **IRF3**, **IRF5**, and **IRF7**, that are members of the IRF family of transcription factors involved in the innate immune response phenotype, and controlling expression of Type-1 interferons upon viral infection.

Importantly, the phenotypes associated to these sink nodes have already been described in the CF context. In particular: (1) The NFκB and AP-1 transcription factors are complexes of sink nodes that mediate inflammation, the most studied phenotype of CF disease. In addition to the well-known activation of NFκB in CF, AP-1 is one

of the downstream transcription factors of the MAPK pathway that was shown to be activated in CF [Bérubé, 2010; Wellmerling, 2022], as shown in Figure 5.6. (2) Controversial results were reported about apoptosis in CF epithelial cells. Some studies showed defective susceptibility of CF cells to pro-apoptotic stimuli [Cannon, 2003; Gottlieb, 1996], while others observed increased apoptosis [Chen, 2018; Voisin, 2014; Yalçin, 2009; Rottner, 2007]. All agree that apoptosis is dysregulated in CF. (3) The dysregulation of actin cytoskeleton in CF is well documented, with a disorganized actin cytoskeleton, absence of actin stress fibres [Favia, 2010; Lasalvia, 2016; Burat, 2022], and disrupted tight junctions [De Lisle, 2014; Castellani, 2012]. (4) Finally, various works indicate a dysfunction in the innate immune response of CF patients [Kosamo, 2020; Gillan, 2023; Dugger, 2020].

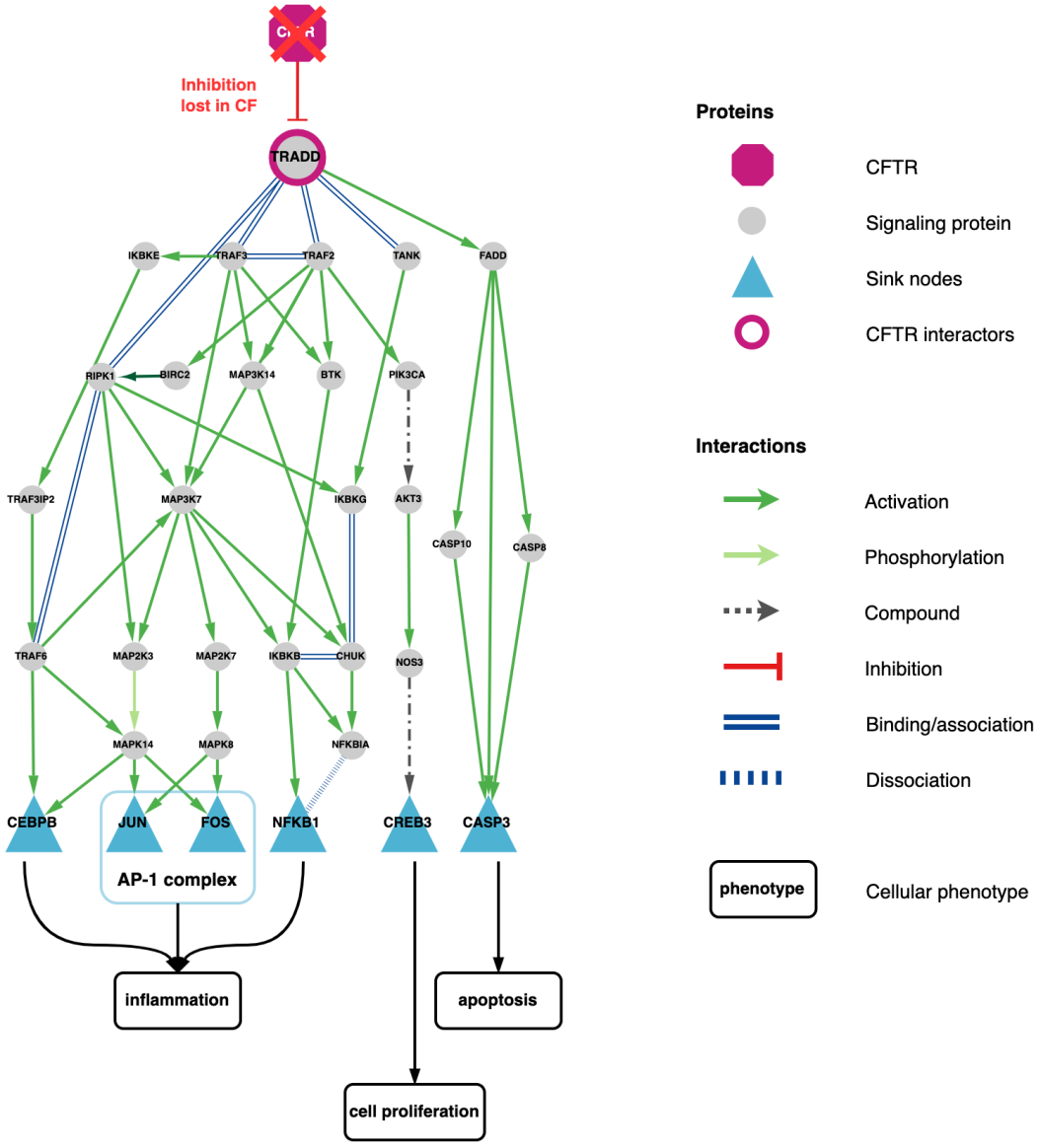


Figure 5.6 – Extract from the CF network showing the TRADD protein connected to the TNF-α signalling pathway, and to 5 other sink nodes, including FOS and JUN which form the AP-1 transcription factor, downstream of the MAPK cascade.

The cellular phenotypes triggered by the sink nodes are surrounded by black contours. Note that TRADD is connected to the 35 sink nodes, but only part of the nodes downstream of TRADD in the network are represented.

Betweenness centrality and flow of information

In a network, the betweeness centrality (BC) of a node is the number of shortest paths that pass through that node. This measure is a way of detecting the amount of influence a node has over the flow of information in a network. Nodes with high BC, referred to as hubs, may provide interesting therapeutic targets, because their inhibition may efficiently reduce the propagation of information within the network

[Durón, 2019]. Therefore, we calculated the BC for all nodes of the CF network, as detailed in the Methods section. All nodes were then ranked according to this measure, and Figure 5.7A displays the histogram of the BC score. Interestingly, most proteins have a BC score below 3000, and only a very limited number of proteins have a BC score above 6000 (ARHGEF12, IKBKE, LSP1, PIK3KC1, PYCARD, RAC1, TRAF2, TRAF3, TRAF6). The list of the top 30 proteins is provided in the Supplementary File 2. Among them, PI3KCA could be an interesting therapeutic target candidate and is discussed in the next section.

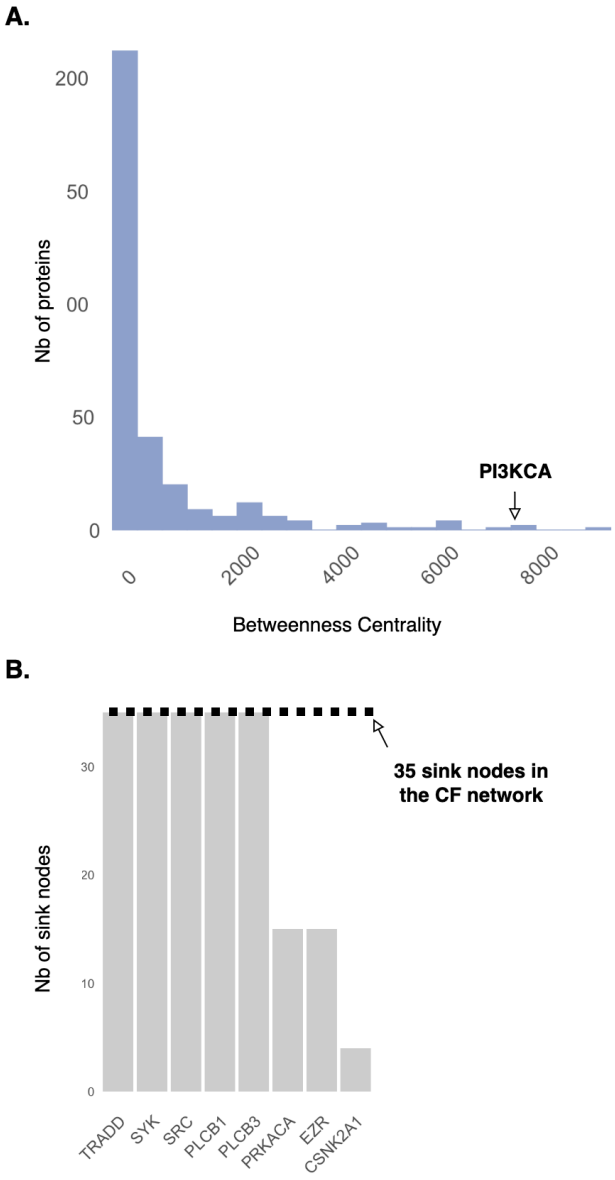


Figure 5.7 – (A) Histogram of the betweenness centrality measures for all nodes in the CF signalling network; (B) Number of sink nodes to which each of the 8 source nodes are connected.

Biological insights from the topological analysis A simple path analysis of the network shows that several source nodes may contribute collectively to the emergence of the CF phenotypes, which illustrates the complexity of the disease. Indeed, while the source nodes PRKACA, EZR and CSNK2A1 are upstream of a limited number of sink nodes, PLCB1/3, SRC, SYK and TRADD are upstream of the 35 sink nodes, i.e., there exists a path from each of these 6 source nodes to each of the 35 sink nodes (Figure 5.7B).

For example, TRADD is known to be up-regulated in CF [Ferenc Karpati, 2000] and to participate in the uncontrolled inflammation (See Figure 5.6). Interaction between wt-CFTR and TRADD enhances the degradation of TRADD, which controls the activity of this pathway, as demonstrated by Wang and colleagues [Wang, 2016]. This direct interaction is lost with F508del-CFTR, which may contribute to the dysregulation of TNF-α and NF-κB signalling pathways in CF. However, up-regulation of TRADD could also contribute to the inflammation phenotype through another route, by inducing over-activation of the MAPK pathway, and in particular of AP-1, one of its output transcription factors. In addition, as shown in Figure 5.8, our network suggests that other source nodes than TRADD could also initiate dysregulation of the inflammation phenotype because they are also connected to the NF-κB sink node. Among these sources, we can cite: (1) SYK, which would be consistent with its role in inflammation processes shown in other diseases [Riccaboni, 2010; Wong, 2004]; (2) PLCB1/3, which are consistent with previous studies reporting PLCB3 as a key modulator of IL8 expression in CF bronchial epithelial cells [Bezzerri, 2011]; (3) CSNK2A1, whose hyperactivity could contribute to activation of NF-κB by enhancing the phosphorylation and degradation of IKBKA.

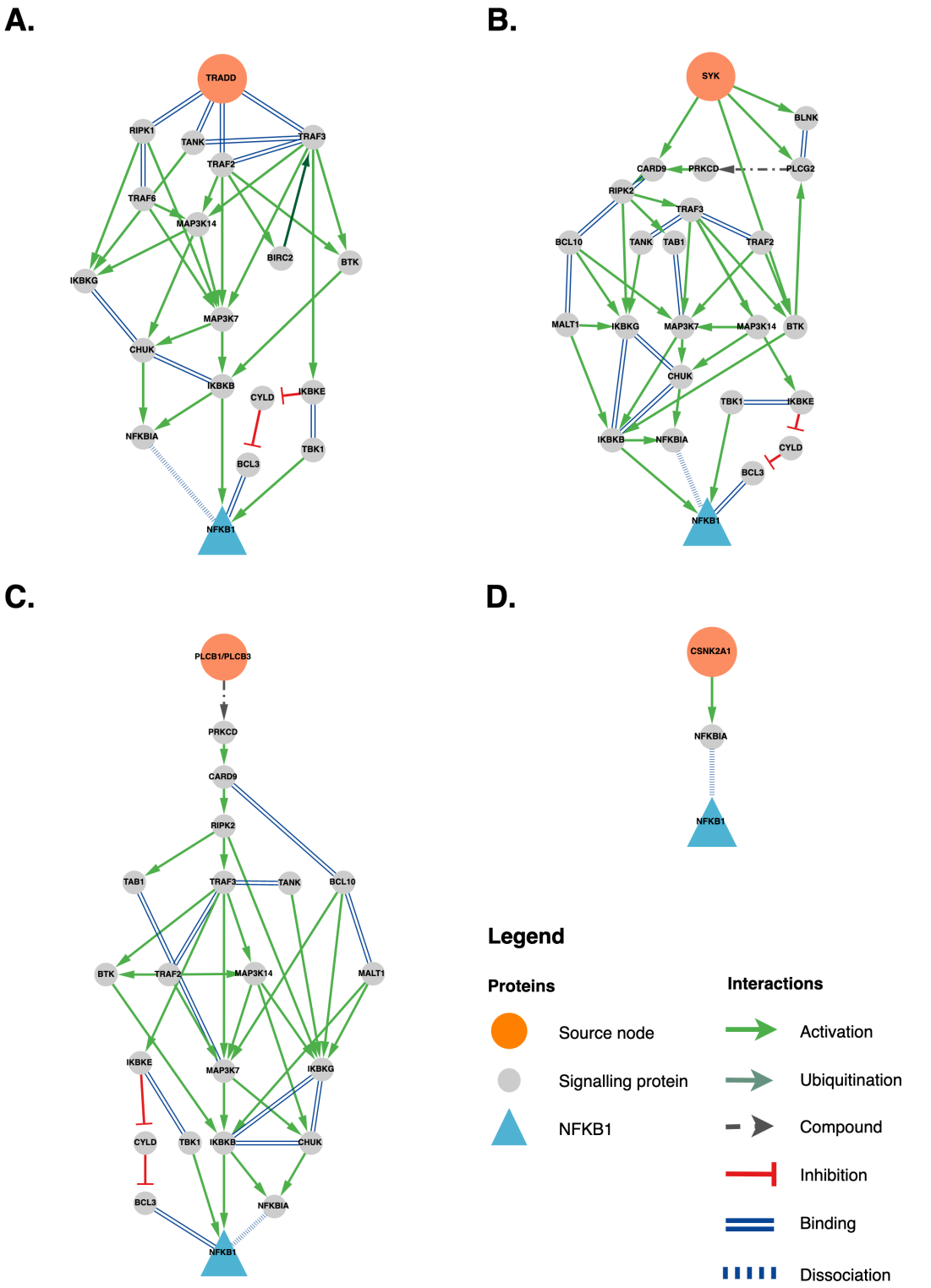


Figure 5.8 – Subnetworks of the CF network illustrating the connections between the source nodes TRADD, SYK, PLCB1/3, and CSNK2A1 and the sink node NFKB1.

Overall, the number of source nodes and routes that may contribute to inflammation in CF illustrates the challenge posed by its modulation, in order to reduce the related clinical symptoms. Various anti-inflammatory drugs have been recently evaluated in clinical trials [Bell, 2020], but none of them target the source nodes of the present study. Our hypothesis is that these source nodes could be interesting candidate targets in CF. In particular, SYK has emerged as a potential target for the treatment of numerous diseases. Many inhibitors are known for this kinase, which would allow to evaluate their potential anti-inflammatory effect in CF cells. These inhibitors include one marketed drug (Fostamatinib), but other inhibitors are currently under investigation in clinical trials for a range of indications [Cooper, 2023]. Interestingly, since SYK is connected to the 35 sink nodes, its inhibition may also contribute to the modulation of other CF phenotypes than inflammation. In particular, it could modulate CF phenotypes associated to the 35 sink nodes and mentioned in Section 5.2.2, such as dysregulations in apoptosis, cytoskeleton or innate immune response. Similarly, our network suggests that PLCB3 could be an interesting target for inflammation in CF. This is consistent with the fact that PLCB3 silencing in CF bronchial epithelial cells exposed to *Pseudomonas aeroginosa*, reduces the expression of IL-8 chemokine [Bezzerri, 2011]. The U73122 PLC inhibitor could be an interesting pharmacological tool to further evaluate this strategy. As in the case of SYK, PLCB3 is connected to the 35 sink nodes, which means that its inhibition may also improve other CF cellular phenotypes. Consistent with this idea, it was shown that treatment with a SRC inhibitor, another of the 6 source nodes upstream of the 35 sink nodes, decreased the inflammatory changes and improved cytoskeletal defects in F508del human cholangiocytes [Fiorotto, 2018].

Besides source nodes, candidate therapeutic targets can be searched among hubs in the network, i.e. among the best ranked proteins according to the BC score. Besides this score, additional arguments can be invoked to highlight the best candidates. In particular, the fact that a protein is known in the literature to play a role in the disease, and that pharmacological modulators (or even better, marketed drugs) are available to allow experimental validation, are important criteria. In line with these ideas, PI3KCA appears as an interesting candidate target. Indeed, several inhibitors are known for this kinase, including the marketed drug Alpelisib, which would allow experimental tests in CF models. It has been suggested as a candidate target in CF based on its role in many signalling pathways implicated in CF lung pathogenesis [Natarajan, 2020]. The fact that PI3KCA belongs to best ranked proteins with respect to the BC score (See Figure 5.7A) offers a quantitative argument in favor of this idea. In addition, PI3KCA is connected through the network to the 35 sink nodes, which means that its inhibition may modulate inflammation, but also other CF cellular phenotypes related to the sink nodes.

Analysis of the unconnected components in the CF network As mentioned in Section 5.2.2, Figure 5.3 shows that the CF network comprises two small unconnected components that are part of the TGFβ and JAK/STAT signalling pathways. Contrary to source nodes of the main component, dysregulation of the source nodes of these unconnected components (namely the 4 interleukins IL2, IL21, IL4 and IL6 for one component, and TGFβ for the other) cannot be explained by the absence of CFTR in a direct manner, because they are not linked to CFTR within a single network. However,

activation of a sink node of the main component may modulate the expression of a source node in an unconnected component, affecting the activity of this unconnected component. For example, activation of the AP-1 transcription factor (a sink node of the main component) due to activation of the MAPK pathway in the main component, regulates the expression of TGFβ. This example shows how dysregulations in one pathway may have consequences in other pathways of the CF network, even if they are not connected, again illustrating to the complexity of the disease. We propose that phenotypes arising from the two unconnected components could be defined as secondary phenotypes, as opposite to primary phenotypes arising from dysregulations of the main component (discussed in Section 5.2.2).

The JAK-STAT component mediates various cellular processes, including cell growth and apoptosis, but the role of these cascades has not been widely studied in CF. The TGFβ component leads to the activation of SMAD2, a transcriptional modulator that regulates multiple cellular phenotypes, including cell proliferation, apoptosis, and differentiation. High levels of TGFβ have been associated with the severity of lung disease [Dorfman, 2008; Sagwal, 2020], and this protein was proposed as a therapeutic target for CF [Kramer, 2018]. Our study suggests that therapeutic targets should be chosen among proteins closer to CFTR in the network, in particular among the source nodes of the main component (as discussed above), because they may more successfully limit the global propagation of molecular dysregulations within the overall network.

5.2.3 Discussion

Using a pathway-based meta-analysis of publicly available transcriptomic data, we built the CF network that provides a more global understanding of the molecular dysregulations in CF than the view of a CFTR-related channelopathy disease. Indeed, an important outcome of this work was to integrate data analyses to network reconstruction, while proposing a strategy to relate CFTR to proteins of the network, based on CFTR interactome. The CF network comprises a restricted number of source nodes that connect the absence of CFTR to the downstream sink nodes triggering CF cellular phenotypes. Another important contribution was to propose candidate therapeutic targets, based on the topological analysis of this network (namely, SYK, PI3KCA and PLCB1/3). The network provides a comprehensive view of how pathway interactions contribute to a given disease phenotype. It reveals unintuitive effects of targeting candidate proteins because of the complex interactions of the biological pathways in the network. Overall, the CF network can be seen as a tool to formulate hypotheses and interpret experimental observations.

Although several transcriptomic datasets were gathered, the total number of samples globally included remains modest (57 CF and 46 control samples). Additional data may refine the list of dysregulated pathways, and help to improve the proposed CF network.

To cope with the low number of samples per study, we opted for a meta-analysis combining various CF transcriptomic datasets, which highlighted that distinct differentiation media used for the primary cultures may favor different cell types, leading to inconsistent transcriptomic profiles and potential erroneous interpretations. This may explain why previous transcriptomic comparative studies reported incoherent signs of gene dysregulation (up- versus down-) between different datasets for many genes

[Clarke, 2013]. We observed the same phenomenon at the pathway level for datasets belonging to subgroup 1 or 2 (see Section 5.2.2). Clustering studies based on the heatmap of common DEPs appears to be a good tool to select consistent data in future meta-analysis.

Other types of dysregulations such as aberrant phosphorylations are not detectable in transcriptomic data. Including information from other types of omics data such as proteomic, phosphoproteomic, metabolomic,or volatilomic may help to refine the CF network. In particular, in the past three years, CF airway epithelial single-cell RNAseq (sc-RNAseq) datasets have been reported [Carraro, 2021; Thurman, 2022]. Such data allow the study of dysregulations at the cell type level, and could facilitate building of the CF networks for specific epithelial cell types. Furthermore, CFTR is expressed in cell types beyond airway epithelial cells. Thus, refining this network within the context of these cell types could enhance our understanding of the role of CFTR in these specific cells such as macrophages, where CFTR seems to have non-channel functions [Duan, 2021].

Prior knowledge gathered in the KEGG pathway database was used to identify and connect DEPs, but the proposed methodology can be followed using other pathways databases. Pathway names and definitions vary between databases, and therefore, the resulting network may slightly depend on the reference database that was used. Nevertheless, it would comprise globally the same interactions and proteins. Similarly, CFTR interactors present in the network were identified according to PPI information in the CyFi-MAP. If new CFTR interactors are identified, this information may help improve the content of the network, highlighting new source nodes or routes for the propagation of dysregulations. In particular, missing interactions, because they are not present in pathway databases, or have not been discovered yet, may explain the presence of unconnected components. If they exist, their discovery in the future may allow to link the two unconnected small components to the main component of the network. However, the proposed notion of targeting proteins as upstream as possible in the network, or among key hubs of the network, are still an interesting concept in order to prioritize candidate therapeutic targets.

An important issue of the present paper was to explore the link between absence of the CFTR protein, and more global pathway dysregulations that lead to CF cellular phenotypes. However, the precise definition of a diseased cellular phenotype is not clearly defined yet, and we used key words provided in the KEGG database or in the Gene Cards database [Stelzer, 2016]. The present work proposes an answer this question in the context of systems biology studies. Associating phenotypes to the activity of outputs of the signalling cascades, referred to here as sink nodes, could be a first step towards the definition of the disease read-outs. This is of particular interest for *in vitro* evaluation of drug candidates, because we expect that drugs active in CF would reduce the activity of these sink nodes.

The methodological approach proposed in our study was settled based on transcriptomic data from hAEC cells homozyguous for F508del, because publicly available data are more abundant for this most frequent mutation. Therefore, our CF network characterizes the disease caused by this mutation. It would be interesting to study to which extent the CF network would differ for other mutations. A recent paper indicates that DEGs in human bronchial epithelial cell lines bearing mutations from different classes

share about 30% DEGs, while 70% of the DEGs are class specific [Santos, 2023]. It would be interesting to study if this still holds at the level of biological pathways, as they are defined in the present work, and to study whether the resulting network is strongly modified, or not. The methodology proposed in the present paper and based on network topology could still be applied in order to search for new, and potentially class-specific, therapeutic targets.

The candidate therapeutic targets proposed based on our CF network could also be tested on CF cellular models for other mutations, because these targets may belong to biological pathways that are also dysregulated with other mutations. If this was the case, it would help to extend the therapeutic arsenal available for CF patients who are not eligible for CFTR modulators.

In the same line, it is now clear that CF patients bearing the same mutation may present diseases of different severity. Although many factors can modulate disease severity, including environmental factors, it would be interesting to explore the contribution of patients molecular profiles. In particular, building a "personalized" CF network based on patients' transcriptomic profiles would be an interesting tool to answer this question.

Beyond CF, reduced amounts of functional CFTR have also been observed in other diseases like chronic obstructive pulmonary disease (COPD) [Saint-Criq, 2017; Simões, 2021], cigarette smoke [Valdivieso, 2018] , or cancer [Duan, 2021; Wang, 2022]. The network could provide a basis to explore the consequences of reduced CFTR activity in these diseases.

Finally, an important contribution of the present work is that the adopted global methodology of the CFTR context, although perfectible, did provide interesting results for CF, and can be used as a common framework for other monogenic diseases.

5.2.4 Methods

Datasets selection

Based on the search engines of the National Center for Biotechnology Information (NCBI) and the European Nucleotide Archive (ENA), we selected 10 datasets from 8 studies published between 2007 and 2021. The selection criteria to include CF transcriptomic datasets were the following: (1) they should correspond to human Airway Epithelial Cells (hAEC); (2) the cells should be homozygous for the most common mutation F508del; (3) the transcriptomic data should be publicly available. Therefore, studies including samples heterozygous for the F508del mutation ([Virella-Lowell, 2004] and [Rehman, 2021]), studies with no data available [Zabner, 2005] and [Wright, 2006]) were not included. In addition, studies with less than two samples were excluded ([Bampi, 2020] and [Veltman, 2021]), as the subsequent statistical analyses require several samples per condition. The list of selected transcriptomic studies is provided in Table 5.1.

Biological pathways databases

We initially considered a total of 380 gene sets corresponding to 380 biological pathways: 50 Hallmark gene sets from the the Molecular Signatures Database (MSigDB) [Liberzon, 2015], 196 from the Pathway Interaction Database (PID) [Schaefer, 2009] and

134 from the KEGG database, restricted to the *Genetic Information Processing*, *Environmental Information Processing*; *Cellular Processes* and *Organismal systems* subdivision. However, most of the analyses were performed using only KEGG database. Indeed, in the Hallmark and the PID databases, gene sets are defined as gene signatures rather than as biological pathways. Thus, the genes are not necessarily connected to each other through functional interactions. Conversely, gene sets retrieved from the KEGG database correspond to biological pathways defined as genes corresponding to proteins that participate in oriented molecular cascades. They are available in the form of maps on the KEGG website. In addition, the structure of the KEGG database allows to build a network that provides mechanistic interpretation. Therefore, gene set enrichment algorithms required to build the signalling network was performed based on the KEGG database. All interactions and nodes from each biological pathway of the KEGG database were retrieved thanks to the OmnipathR R package [Türei, 2016].

Preprocessing of RNA-Seq data

Limma was originally developed for differential expression analysis of microarray data, which values are assumed to be normally distributed, and the variance independent of the mean. This is not the case for log2-counts per million (log-CPM) values in RNA-Seq data: expression distributions may vary across samples and methods modelling counts assume a quadratic mean-variance relationship. Therefore, for the RNA-Seq data, 3 steps of pre-processing are necessary before applying the statistical tests [Law, 2018]: (1) low expressed genes are filtered (i.e. genes with less than 10 read counts in at least one sample in the condition with the minimum sample size); (2) normalisation using the method of trimmed mean of M-values (TMM) is performed [Robinson, 2010b]; (3) raw counts are converted to log-CPM and the mean-variance relationship is estimated with the *voom* method.

Identification of Differentially Expressed Pathways (DEPs)

For each of the 10 transcriptomic datasets, identification of DEPs was performed using the *fgseaSimple* function of the Bioconductor package *fgsea* [Korotkevich, 2021], for fast preranked Gene Set Enrichment Analsyis (GSEA) [Subramanian, 2005].

The *fgseaSimple* method takes two inputs: a gene-level signed statistics and a defined list of genes known as *gene set*. The method ranks the genes in descending order based on the chosen statistics, and then computes the Enrichment Score (ES) for the gene set. The ES reflects how often members of that gene set occur at the top (e.g., upregulated) or the bottom (e.g., downregulated) of the ranked gene list. To account for differences in gene sets size, a normalisation step is performed to obtain the Normalised Enrichment Score (NES). Besides, random gene sets are generated and their NES computed. These NES are then used to create a null distribution from which the significance of the NES of the tested gene set is estimated. In our study, we used the t-statistics from the differential expression analysis comparing gene expression levels of CF sample to NCF samples as the control condition. In order to compare all the studies together, all the microarray and RNA-Seq datasets were processed using the same pipeline, involving the limma [Ritchie, 2015] and edgeR [Robinson, 2010a] packages. After removing technical outlier samples and retrieving gene symbols using the

biomaRt package [Durinck, 2009], differential expression analysis at the gene level was performed by fitting a linear model using weighted least squares for each gene.

Gene sets with size larger than 500 were excluded for statistical testing. The p-values of the gene sets were adjusted for multiple testing error with Benjamini-Hochberg (BH) procedure. Differentially Expressed Pathways (DEP)s were considered with a corrected p-value lower or equal to 0.25. If the NES is positive, the DEP is categorized as up-regulated, and if it is negative, the DEP is categorized as down-regulated.

Up-dating Omnipath DEPs pathways

The CF network was built from DEPs among pathways in the KEGG database, as extracted with the OmniPathR package. We observed a few inconsistencies between the corresponding list of genes and interactions downloaded with OmnipathR R package, and those in the 'up to date' pathways maps, as they are displayed on KEGG website. Therefore, we updated the OmnipathR version of the KEGG pathways by adding (or removing) a few nodes or interactions, in order to map the OmnipathR pathways with their corresponding pathways in KEGG. For each modification, bibliographic references were manually checked into other databases stored in Omnipath, in particular in the high confident databases SignorDB [Lo Surdo, 2022], and the Human Reference Interactome [Drew, 2021]. In addition, in a few pathways, some interactions are labelled as "indirect" in KEGG database. They involve part of signalling cascades belonging to other biological pathways, and they are not detailed in the considered pathway. For example, part of the PI3K-AKT pathway belongs to the Toll-like receptor signalling pathway but is not detailed in this pathway (See KEGG map for Toll-like receptor signalling pathway). In such cases, in order to build the network based on complete cascades involving only direct interactions, we added the missing nodes and interactions.

All the pathways modifications and the corresponding codes used to perform these modifications are available in the following repository: sysbio-curie/CFnetwork.

Network building and pruning

In the KEGG database, most of the 15 common DEPs display the same overall topology: some cell-surface receptor proteins activate one or more intra-cellular signalling cascades that in turn activate downstream transcription factors, thus triggering corresponding phenotypes. For example, the NF-κB pathway leads to the "inflammation" or "cell survival" phenotypes. However, 2 of the common DEPs, *Cytokine-cytokine receptor interaction* and *Viral protein interaction with cytokine and cytokine receptor*, are pathways that do not consist in such functional cascades. The *Cytokine-cytokine receptor interaction* pathway consists in a list of interactions between extra-cellular signal molecules and cell-surface receptors (see KEGG database to visualise this pathway's topology). These interactions are also part of larger biological pathways that comprise their corresponding downstream cascades. This means that KEGG pathways are partially redundant (i.e. small pathways are part of larger pathways), which is also found in all commonly used pathway databases. In the case of the *Cytokine-cytokine receptor interaction* pathway, this DEP is dysregulated in the meta-analysis because some of the interactions between extracellular molecules and cell surface receptors are

dysregulated, but not necessarily all of them. For example, interactions between TNF-α and its receptors, or IL17 and its receptors are dysregulated, but this information is also present in the DEPs containing the complete corresponding cascades, i.e. the *TNF-α signalling pathway* and the *IL-17 signalling pathway*. The same type of analysis also holds for the *Viral protein interaction with cytokine and cytokine receptor* DEP. Overall, from these 2 DEPs, we only retained the cell-surface receptors that are sources of downstream dysregulated cascades in our network. Overall, 25 cell surface receptors without downstream dysregulations in our CF transcriptomic data were removed from the network.

Finally, we also removed from the pathways all the interactions corresponding to genes targeted by transcription factors, downstream of the pathways' cascades, because these target genes do not define the pathways themselves.

Network building and pruning were performed using the R packages tidyr v.1.2.1, and dplyr v.1.0.10. Transcription factors were identified using the R packages dorothea v.1.4.2 and hgnc v.0.1.2, which give access to the Dorothea [Garcia-Alonso, 2019] and HUGO collections [Seal, 2023], respectively.

Betweenness centrality score

The betweenness centrality (BC) score of node n is defined by

$$\sum_{i \neq j, i \neq n, j \neq n} p_{inj}/p_{ij}$$

where p_{ij} is the total number of shortest paths between nodes i and j while p_{inj} is the number of those shortest paths which pass though vertex n.

BC scores were computed using the *betweenness* function of the R package igraph v.1.3.4 [Csardi, 2005]. This pacakge was also used for the other network topology analyses.

Network Visualization and Figure Generation

The networks, generated as dataframes in R, were imported into Cytoscape v.3.9.0 [Shannon, 2003] for visualization. We designed a custom style for nodes and edges, which is available in the Cytoscape session and also saved as an independent XML file, available in the sysbio-curie/CFnetwork_cystoscape repository. The hierarchical layout was used to emphasize the information flow from the source nodes to the sink nodes.

Barplots were generated using the R package ggplot2 v.3.3.6, and didactic figures were created using the open-source platform diagrams.net.

Data Availability

The codes and datasets supporting the conclusions of this article are available in the folowing repository: sysbio-curie/CFnetwork.
The Cytoscape session of the CF network, the TSV files of the nodes and the edges of the CF network and the XML file of the custom style of the Cytoscape session

required to reproduce the Cytoscape session are available in the following repository: sysbio-curie/CFnetwork_cystoscape.

Supplementary Material

Table 1. The 35 sink nodes of the CF network and their corresponding cellular phenotypes.

HGNC	Cellular phenotypes
CASP1	pyroptosis, cell death, inflammation
CASP3	apoptosis
CASP7	apoptosis
CYBA	ROS/oxidative stess
CYBB	ROS/oxidative stess
DNM1L	Necroptosis/Cell Death
GABARAP	Autophagy
ACTN4	Regulation of actin polymerisation
ARPC5	Regulation of actin polymerisation
CFL1	Regulation of actin polymerisation
ENAH	Regulation of actin polymerisation
GSN	Regulation of actin polymerisation
IQGAP1	Regulation of actin polymerisation
MYL12B	Regulation of actin polymerisation
PFN	Regulation of actin polymerisation
PXN	Regulation of actin polymerisation
VCL	Regulation of actin polymerisation
CEBPB	inflammation
CREB1	cell cycle, apoptosis, inflammation
CREB3	proliferation, migration, differentiation, inflammation
ESR1	Regulation of cell cycle, apoptosis, cell adhesion
ESR2	Regulation of cell cycle, apoptosis, cell adhesion
FOS	innflammation, proliferation
IRF1	innate immune response
IRF3	innate immune response
IRF5	innate immune response
IRF7	innate immune response
IRF9	innate immune response
JUN	inflammation, proliferation
NFATC1	cellular differentiation, immune response
NFKB1	inflammation, cell survival/proliferation
NFKB2	inflammation, cell survival/proliferation
RELA	inflammation, cell survival/proliferation
RELB	inflammation, cell survival/proliferation
STAT1	innate immune response

Table 5.3 – The 35 sink nodes of the CF network and their corresponding cellular phenotypes.

Cellular phenotypes were retrieved from the KEGG database or from the Gene Cards database when no phenotype was associated with the sink node in any of the KEGG pathways.

Table 2. The top 30 proteins in the CF network according to their betweenness centrality score

HGNC	BC score
TRAF2	9541.17
LSP1	7987.59
PYCARD	7845.23
PIK3CA	7710.44
IKBKE	6734.15
TRAF3	6691.04
TRAF6	6365.37
ARHGEF12	6268.20
RAC1	6018.87
MAVS	5527.33
STING1	5000.21
IFI16	4890.92
PAK3	4813.73
TBK1	4456.73
MAP3K7	4345.00
SYK	3622.03
VAV1	3347.09
ZAP70	3309.76
PLCG2	3295.18
TRAF3IP2	3086.74
LCP2	2974.65
CLEC7A	2944.23
RHOA	2830.65
NLRP3	2814.41
AKT3	2784.62
CASP8	2748.34
IKBKG	2723.52
HSP90AA1	2705.86
MAPK1	2657.41
CYLD	2628.59

Table 5.4 – The top 30 proteins in the CF network according to their betweenness centrality score

5.3 Discussion of the methodological choices made when building the CF network

In this paper, we presented building of **the CF network**, a signalling network gathering the molecular dysregulations caused by the absence of CFTR. This network allowed us to propose relevant mechanistic hypotheses with respect to various CF cellular phenotypes. This work is unique both in the object of study and the methodological choices in the field of CF research. Indeed, we used a data-driven approach to study dysregulated signalling pathways of the CF HAEC, without any a priori hypothesis. It is to our knowledge the first study of this type applied to CF. This network can be used as a basis to tackle disease heterogeneity and to model other systems such as other cell types of the respiratory tract or cell types of other organs, to study signalling dysregulations in other CF mutations.

The adopted methodology is based on three pillars: the omics data, the computational method, and the prior knowledge database (i.e. biological database). In the present discussion, I detail each of these components and discuss how the network could be improved. In addition, I discuss possible approaches to validate candidate therapeutic targets proposed in the paper.

5.3.1 The omics data

The first criticism could obviously be the use of gene expression data to infer pathway activities. This assumes a strong correlation between gene expression, protein abundance, and protein activity. This assumption is clearly limited and remains an area of investigation [Buccitelli, 2020]. Whole-cell proteomics (or whole-cell phosphoproteomics) is the type of choice for describing pathway activity [Szalai, 2020], and integrating them in the analysis would have been ideal. The combined analysis of paired proteomic and transcriptomic data would have been the most relevant in our opinion, but this kind of study on the same samples are not yet available in CF studies.

5.3.2 The computational method

As mentioned in the discussion of the article, we believe that adopting a different method to analyse omics data at the scale of the biological pathways would have led to a very similar network. Indeed, the pathways identified as dysregulated, using other PB methods, would have overlapped those we found, potentially only with a few exceptions.

Nonetheless, it is also possible to explore alternative approaches besides PB methods for extracting dysregulated signalling cascades. These alternatives include network-based approaches and inference of TF activities. In this section, we explore these two potential avenues.

Network-based approaches

We used PB approaches rather than network-based approaches to extract dysregulated protein cascades. These recent network-based approaches, reviewed in 2022 [Garrido-Rodriguez, 2022], appear promising. However, using them to build a network

with functional interactions would have faced several limitations in the context of the present project.

First, when we started this project in 2019, very few studies had applied network-based methods to study biological systems of diseases, conversely to PB methods. In addition, there was a lack of benchmarks to compare and validate these methods, and this limitation is still valid today [Garrido-Rodriguez, 2022].

Besides this pragmatic reason, it is important to highlight that these methods rely on the choice of Prior Knowledge Network (PKN) as input in the algorithm, while PB approaches rely on prior knowledge of biological pathways. As presented in chapter 2, biological databases store signed interactions in two different ways: resources such as SIGNOR [Lo Surdo, 2022] or STRING [Szklarczyk, 2019] store interactions, whereas KEGG [Kanehisa, 2012] or REACTOME [Gillespie, 2022] store biological pathways maps as separate graphs, in which interactions are directed and signed. Adopting a network-based approach to extract a CF network of signalling dysregulations would first require to combine these resources into a single and large signalling network that is used as the PKN input. However, the information gathered in pathway databases arise from studies undertaken in different biological contexts, so that the resulting merged PKN would not correspond to a consistent biological system. This limitation also applies to the method chosen in the article, as we have also combined different pathways from different pathway graphs. Nevertheless, in our case, it is mitigated by the fact that we combine far fewer pathways (fifteen compared to hundreds), so that we can supervise beforehand which pathways are relevant to our study. As an example, we excluded disease pathways from the KEGG database, because they did not correspond to pathways that are relevant to the building of a the CF signalling network.

MOGAMUN

We investigated this type of approaches with the MOGAMUN algorithm, which stands for *A Multi-Objective Genetic Algorithm to find active modules in MUltiplex biological Networks* [Novoa-del-Toro, 2021]. Multiplex networks are networks composed of different layers, where each node is present in the different layers and each layer describes all the edges of a specific type, such as physical interactions, functional interaction or co-expression [Battiston, 2014]. MOGAMUN identifies subnetworks of interest by optimizing a score based both on the density of interactions and on a score of the nodes (e.g. the genes t-test in their differential expression). The output of the algorithm is a list of subnetworks with the highest scores.

We applied MOGAMUN on three transcriptomic datasets considered in the meta-analysis [Clarke, 2013; Ogilvie, 2011; Zoso, 2019]. We tested the algorithm with a single layer of PKN, using the PPI network of the reference article. This PPI network was established by merging interactions from various databases from the PSICQUIC portal [del-Toro, 2013], and from the Center for Cancer Systems Biology (CCSB) Interactome database [Rolland, 2014]. The resulting PPI network comprises 12621 nodes and 66971 interactions. For the dataset from Ogilvie et al. [Ogilvie, 2011], we obtained 17 subnetworks comprising 110 unique proteins, for the dataset from Clarke et al. [Clarke, 2013], we obtained 20 subnetworks comprising 110 unique proteins and for the dataset from Zoso et al. [Zoso, 2019], we obtained 21 subnetworks comprising 107 unique proteins.

For each study, we merged all the subnetworks into a single network, and compared the three resulting networks: they shared 25 proteins, i.e. almost the fifth of the size of each subnetwork. Some of these proteins, such as JUN, ESR1, EP300, MDM2, UBE21, STAT3 also belong to our CF network based on the PB approach, and are consistent with CF pathophysiology. Among the 25 proteins, 11 proteins belong to the Keratin Associated Protein Family (Gene code beginning with KRTAP), which may be explained by the fact that some keratin proteins control the surface expression of CFTR, such as the Keratin 18 (K18) protein [Stanke, 2011]. However, the enrichment of this family may be a statistical artefact and may be caused by the density of the interactions of these proteins in the PPI network.

Now that the meta-analysis with PB methods is completed, I plan on re-running MOGAMUN, but also trying other network-based approaches, such as CARNIVAL [Liu, 2019] or CausalR [Bradley, 2017], on the 10 datasets of the meta-analysis, and compare the networks obtained with each approach.

Inference of TF activity

Using the same omics data, other types of information about CF molecular dysregulations could have also been extracted. For instance, we could have retrieved a list of over-activated or under-activated transcription factors (TF) in CF HAEC cells. The principle is the same as to retrieve dysregulated biological pathways: we need a database of TFs and their targets (also called *regulon*), and a computational method called in this case *footprint method* (See chapter 2 for the distinction between *pathway-based* methods and *footprint-based* methods). Expression levels of TF regulons can be viewed as footprints of TF activity [Szalai, 2020].

We explored this direction applying fgsea [Korotkevich, 2021] with the DoroThea database [Garcia-Alonso, 2019] on the 10 transcriptomic datasets (see Table 5.1). The heatmap of GSEA Normalized Enrichment Scores (NES) for each dataset of the TF significantly dysregulated in at least 3 studies is presented in Figure 5.9. Most of the significantly over-activated TF in at least 3 studies belong to the CF network: HIF1A, IRF2, IRF9, JUN, NFKB1, RELA, SP1, SPI1 and STAT1.

In principle, we could have then subtracted from the network the signalling cascades upstream to the TFs that are not found over-activated (or under-activated). I decided not to include these steps to the construction of the network, because I would have liked to try other algorithms, such as viper [Alvarez, 2016], specially developed for the inference of TF activity, as well as other databases, such as the new one CollecTRI meta-resource [Müller-Dott, 2023] which outperforms all the other public collections of regulatory interactions. I would consider these steps as a way to refine the network.

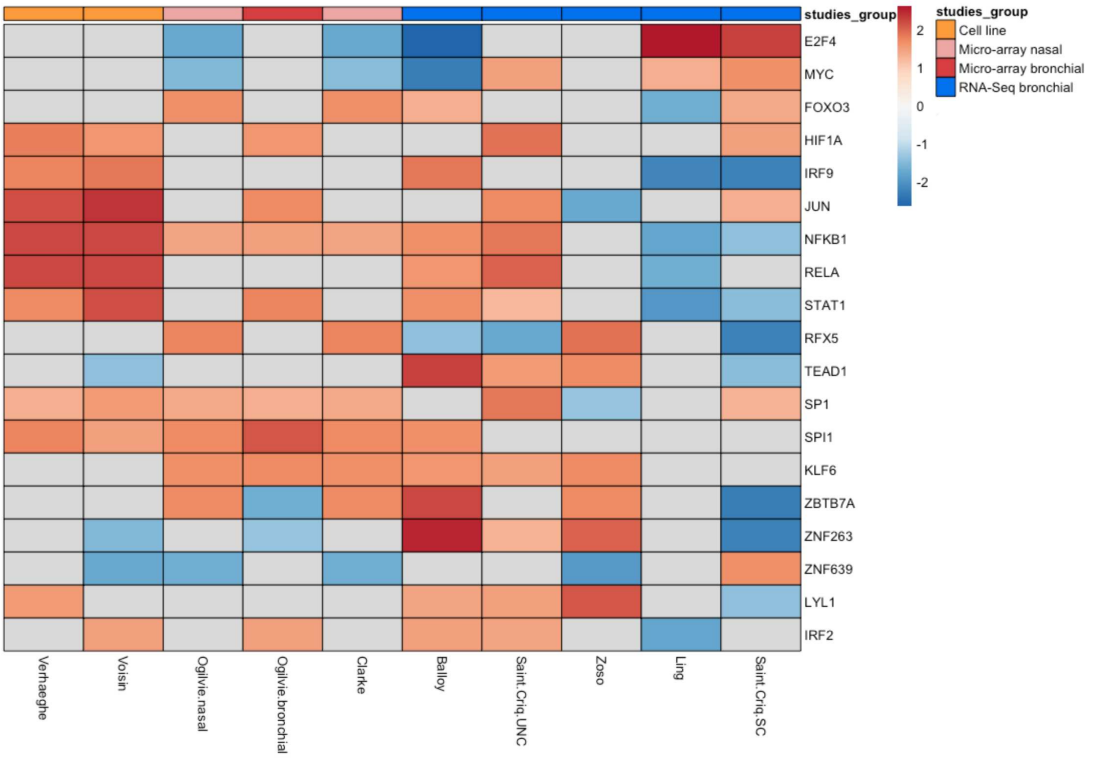


Figure 5.9 – Heatmap of TF activity scores for each dataset.

5.3.3 The prior knowledge database

PB approaches on other pathway databases

Results of analyses of transcriptomics studies are much more sensitive to the reference pathway database used rather than to the computational method applied to extract differentially expressed pathway [Garcia-Alonso, 2019]. In this project, we also apply fGSEA algorithm with three other pathway databases: the Molecular Signatures Database (MSigDB) Hallmark gene set collection [Liberzon, 2015], the PID (Pathway Interaction Database) [Schaefer, 2009] and the Ingenuity Pathway (IPA) database. Overall, we tested 831 gene sets from these databases, and 16 gene sets were found significantly over-expressed, among which many overlap with the results obtained with the KEGG database (e.g. the PID AP-1 Pathway, the Hallmark Inflammatory Response, the Hallmark IL6 JAK STAT6, the Hallmark Interferron Gamma Response and the Hallmark $TNF\alpha$ via $NF\kappa$B pathway). The heatmap of the GSEA Normalized Enrichment Scores (NES) of the pathways significantly dysregulated in at least 3 studies for each dataset is presented in Figure 5.10.

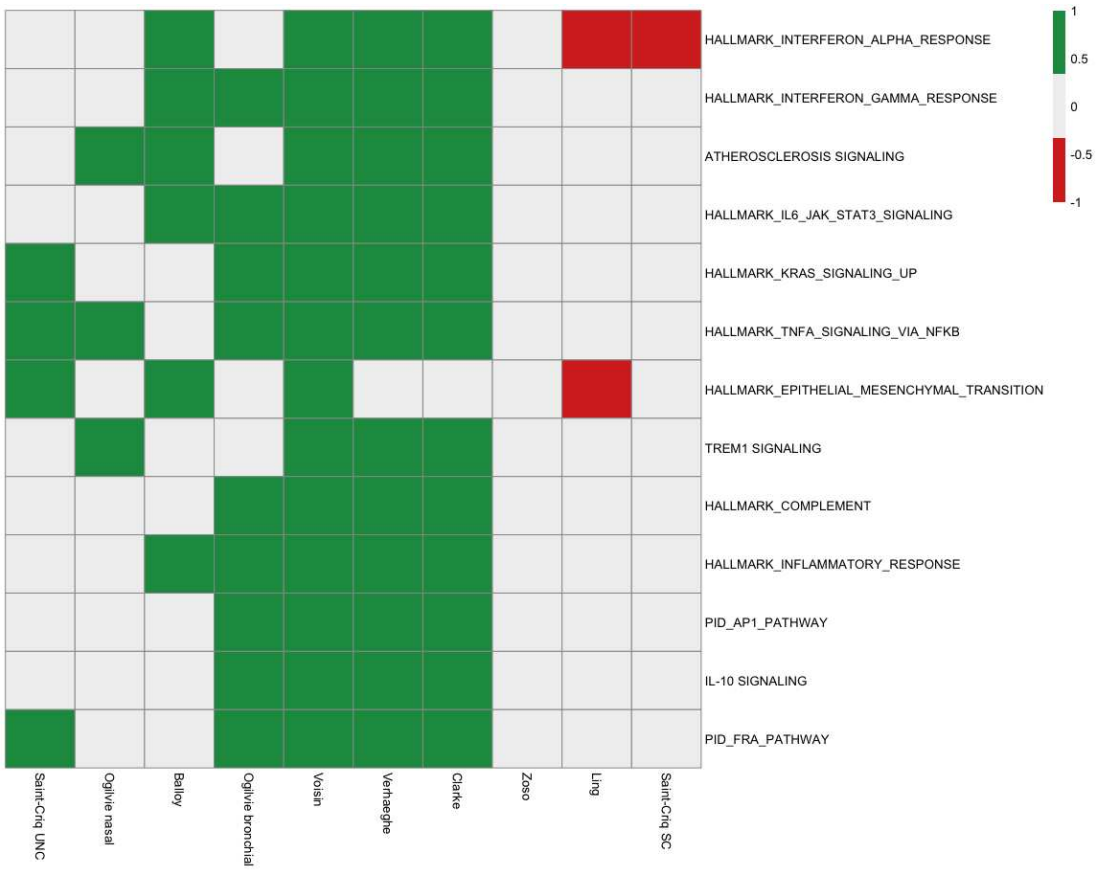


Figure 5.10 – Heatmap of the differentially expressed pathways (DEP) for 3 gene set databases and for each dataset.
A green tile indicates an over-expressed pathway in the study and a red tile indicates an under-expressed pathway in the study.

These results show that the molecular mechanisms identified as dysregulated do not actually depend on the prior knowledge database. Nonetheless, the choice of the database is crucial to the methodology, as it is used for biological interpretation of the results. The aim of the project is to connect CFTR to the CF cellular phenotypes via functional interactions, and model the molecular dysregulations at the scale of the signalling cascades. KEGG pathway maps detail molecular interactions implicated in various signalling pathways and are a good resource for interpreting our data. In contrast, the three databases mentioned above (the MsigDB, the PID and the IPA databases) define pathways as gene signatures, but the information about how these genes are connected and cooperate in the pathway is not provided.

Additional interactions between proteins of different pathway graphs

We used KEGG to retrieve the dysregulated pathways, and merge them to have one single network. However, some proteins of the network may also interact by functional interactions but would not appear because of the content of KEGG pathway maps. It

would therefore be interesting to explore functional databases, such as Signor database [Lo Surdo, 2022], to search for all possible additional interactions between the proteins of the network. A method has recently been developed in Curie's systems biology group to extract functional interactions from a list of proteins. It could be used to search for additional interactions of the CF network proteins.

CFTR interactome databases

In this study, the use of prior knowledge was also essential to connect CFTR to the molecular dysregulations. We used the CyFi-MAP, a disease map repository comprising proteins interacting with CFTR during all its life cycle [Pereira, 2021]. CFTR interactors are included in the disease map if they are confirmed in a minimum of two published references that considered airway epithelial cells, and if the interactions between the components are physical.

However, CFTR interactors may not still be all elucidated. Datasets from HT technologies depicting protein interactomes (also called *interactomics*) can also be very useful in this case. For instance, a very recent study focusing on CFTR and rescued F508del-CFTR interactomes at the PM [Matos, 2019] could be used to search for potential newly detected CFTR interactors that could bridge CFTR and the dysregulated pathways, offering new routes for the propagation of dysregulations.

The technologies of these experimental approaches are evolving fast. Recently, new approaches have been developed based on the proximity tagging of protein partners or nearby proteins and their subsequent identification by mass spectrometry. During my PhD, I took part in the statistical analyses for a study exploring two proximity labeling techniques for WT-CFTR and two CFTR mutants (G551D and W1282X). The study identified additional CFTR protein partners, which do not appear in the CyFi-MAP. These partners were identified in kidney cell lines (HEK293). We wanted a consistent model focusing on HAEC so we decided not to include them. The research paper, published in *International Journal of Molecular Sciences (IJMS)* is presented in appendix B.

5.3.4 Potential therapeutic targets?

Finally, an important contribution of this work is the proposal of candidate therapeutic targets, in particular SYK, PLCB1/3 and PIK3CA. Although only experimental approaches could validate or invalidate these candidates, while waiting for such experiments, an interesting question is how computational methods could further evaluate these candidates.

First, one could simulate the dynamic behavior of the network upon inhibition of these candidates using logic modelling. Such approach would be possible with MaBoSS, a software developed by the systems biology group of Institut Curie [Stoll, 2017]. The simulation with MaBoSS enables to show how biological information flows through the network towards the output nodes that trigger the phenotypes. It is possible to simulate the pharmacologic inhibition of a candidate therapeutic target by maintaining its value at 0 over the whole simulations, and explore how this *in silico* intervention modulates the flow of information through the network, possibly moderating some of the cellular phenotypes.

The actual network of dysregulations is too big for logic modelling (330 proteins and 529 interactions), but this could be performed on a subnetwork centred on a subset of cellular phenotypes. For instance, studying the link between the phenotypes related to cytoskeleton, such as focal adhesion or actin polymerisation, and the phenotype of inflammation, would be of particular interest as there is increasing evidence of the entanglement of these signalling pathways [Di Pietro, 2017; Ding, 2022; Papa, 2021].

Finally, these candidates should be validated experimentally by conducting *in vitro* and *in vivo* experiments of inhibition/activation of these proteins. As mentioned in the discussion of the article, we could monitor the evolution of the cellular phenotypes after these perturbations by checking the activity of the corresponding network sink nodes. This would require to identify relevant experimental readouts that could be followed upon inhibition of the candidate targets.

Part III

Chemogenomic approaches to study CF

Chapter 6

Prediction of Drug Target Interaction in brief

Contents

Abstract

This chapter provides a short introduction to in silico approaches to target identification, and especially to machine-learning (ML) chemogenomics approaches. The aim of this chapter is not to present exhaustively all the current methods and algorithms to predict drug target interactions. It is rather to define the vocabulary and the tools to the non familiar reader, and to set the stage to the third contribution of this thesis presented in the next chapter.

Résumé

Ce chapitre présente brièvement les approches in silico de l'identification des cibles, et en particulier les approches de chémogénomiques basées sur l'apprentissage automatique. L'objectif de ce chapitre n'est pas de présenter de manière exhaustive toutes les méthodes et tous les algorithmes actuels pour prédire les interactions molécules-cibles mais de définir le vocabulaire et les outils pour le lecteur non familier afin de préparer le terrain pour la troisième contribution de cette thèse présentée dans le chapitre suivant.

6.1 Purpose

We presented in the first part of this thesis a systems biology approach to study CF. This approach allowed a better understanding of the molecular dysregulations of CF and enabled the identification of proteins which inhibition could potentially reverse some CF cellular phenotypes.

Another original way to explore CF molecular dysregulations is to understand the MoA of CFTR modulators. These molecules seem very efficient, improving the lung function and quality of life for the majority of the CF patients. Particularly, the the new combination that includes elexacaftor appears to offer the greatest benefits to patients. These molecules discovered based on phenotypic screens aiming at improving the processing, maturation and function of mutated CFTR. However the heterogeneity of patients' response and the evolution of some biomarkers indicate that they could involve off-target proteins (see chapter 1 for a more detailed description). Understanding the MoA of CFTR modulators is necessary to better understand the heterogeneity in patients' response, or suggest other therapeutic strategies. More precisely, identification of potential off-targets could point at key biological pathways involved in the disease and modulated by these treatments, and could help to better understand CF molecular basis. Some ligands or even drugs might have been specifically designed for these unknown off-targets, and might be more efficient for these off-targets than CFTR modulators because they have not been optimised for this purpose. This could suggest new therapeutic solutions in CF, on their own or in synergy with other drugs, such as the CFTR modulators. Besides, the identification of off-targets could also highlight proteins which are not dysregulated due to the disease, but whose modulation by CFTR modulators could lead to adverse drug reaction (ADR), also called side-effects. The quality of life of patients can then be further improved by finding molecules that do not target these proteins.

Understand the MoA of the CFTR modulators boils down to the search for "unexpected" off-target proteins for the four CFTR modulators, i.e. targets that are not CFTR. We do not make any assumption about the potential off-targets that will be searched at the scale of the entire "druggable" proteome, i.e. the proteins against which at least one drug is known.

We searched for CFTR modulators off-targets with computational approaches, because identification of drug targets at the druggable proteome level is not feasible based on experiments alone. The idea was to search for the most probable targets based on computational methods, in order to reduce the number of experiments to be performed by focusing on a limited number of high-probability protein targets. These methods use Drug Target Interaction (DTI) databases or/and physico-chemical properties to predict new DTIs.

In this chapter 6, we will give a short introduction to *in silico* approaches to target identification. We will focus on machine-learning (ML) chemogenomics approaches, that we applied to predict off-targets for CFTR modulators. The aim of this chapter is not to present exhaustively all the current methods and algorithms to predict DTI. It is rather to define the vocabulary and the tools to the non familiar reader, and to set the stage to the third contribution of this thesis presented in the next chapter.

One can see this as prerequisites to understand a short, well-defined chemogenomic machine-learning (ML) problem.

6.2 *In silico* approaches to DTI prediction

6.2.1 Drug-Target Interaction (DTI)

The problem of target identification can be addressed computationally in the form of the prediction of Drug-Target Interaction (DTI) for the drug of interest, and for all the proteins in the space of the human proteome. Proteins are then ranked according to their binding probability scores, and the top ranked proteins are considered as potential targets.

By DTI, we mean **direct binding** between a small molecule and a protein whose 3D structure presents a pocket into which the molecule can bind. The binding trigger a modulation of the protein activity: an **inhibition** or an **activation**. In the general case, the small molecule is called a **ligand**.

6.2.2 Various approaches

There are three main categories of approaches for predicting DTIs:

— *Ligand-based methods*, such as *Quantitative Structure Activity Relationship (QSAR)*, create a model to predict if a molecule will bind to a given target, based on the binding affinities of known ligands for this target. They are efficient for the prediction of new ligands for a given protein. However, for the reverse problem of predicting all targets for a given molecule, using QSAR would require training a model for each protein. This is out of reach at the proteome level, because many proteins have only few, or even no, known ligand to train such predictor.

— *Docking* is a molecular modeling technique that predicts the binding affinity between a molecule and a protein by estimating their interaction energy. Docking methods rely on the 3D structures of proteins, which restricts their application for large-scale predictions, because many proteins have unknown 3D structures.

— *Chemogenomic approaches* are mathematical and computational frameworks that are suitable to predict DTI at large scales both in the protein and in chemical spaces. These approaches are based on the assumption that the prediction of a given DTI may benefit from all interactions known between other proteins and other molecules, even if they do not involve the protein or the molecule under study. This is in contrast to ligand-based methods which predict whether a protein p interacts with a molecule m, based on all ligands known for protein p. They can be viewed as an attempt to fill a binary interaction matrix where rows are molecules and columns are proteins, partially filled with known protein-ligand interactions. (Read [Playe, 2018] for a review of the three approaches).

Chemogenomic algorithms are the most appropriate for predicting DTI on a large scale. They seem best suited to our problem since we are looking for CFTR modulators off-targets in the entire proteome space.

6.2.3 Regression or classification problem

Depending on the models, DTI prediction can be formulated in two ways: one can predict the strength of the interaction, which would correspond to a **regression problem**, or just distinguish the pairs of molecules and proteins that bind from those that do not. It is then formulated as a **classification problem**.

The first problem requires the measurement of affinities of pairs or the 3D structures of the two partners, in order to measure quantitatively the interaction energies. Binding affinities are accessible for a few families of proteins, but not for all of them, which prevents us from formulating DTI prediction as a regression problem on a large scale. Therefore, we will consider DTI prediction as a classification problem in the next chapters.

6.2.4 Rule-based vs supervised ML algorithms

Regardless of the algorithm, we can summarize the prediction of DTI as to predict whether a pair of a molecule m and a protein p, interacts or not. The input is the pair, referred to as (molecule m, protein p) or (m, p), and the output a Boolean value: True, or 1, if the pair (m, p) interacts, and False, or 0, if the pair (m, p) does not interact. Finally, we have an algorithm, also called **classifier**, that takes the pair as input and provides the prediction as output.

These classifiers can be categorised into two broad classes: **rule-based algorithms** and **supervised machine-learning (ML) algorithms**.

Rule-based algorithms make decisions based on a set of predefined knowledge and rules, trying to resemble the decisions made by a human expert in the field. This is typically the case for QSAR methods, which define rules based on the physico-chemical and/or structural properties of the protein and molecule. This is also the case for docking approaches, whose prediction is based on calculating the binding energy from the 3D structures of the protein and the molecule.

Conversely, **supervised ML algorithms** do not know the rules underlying decision making. The algorithms learn mathematical models from available data, i.e. from known interactions in our case, to separate pairs that interact and from pairs that do not. Known interactions are generally compiled in **DTI databases**, like the PubChem database at NCBI [Bolton, 2008] or chEMBL [Gaulton, 2012].

Finally, the classifier is used to predict interactions on unknown data.

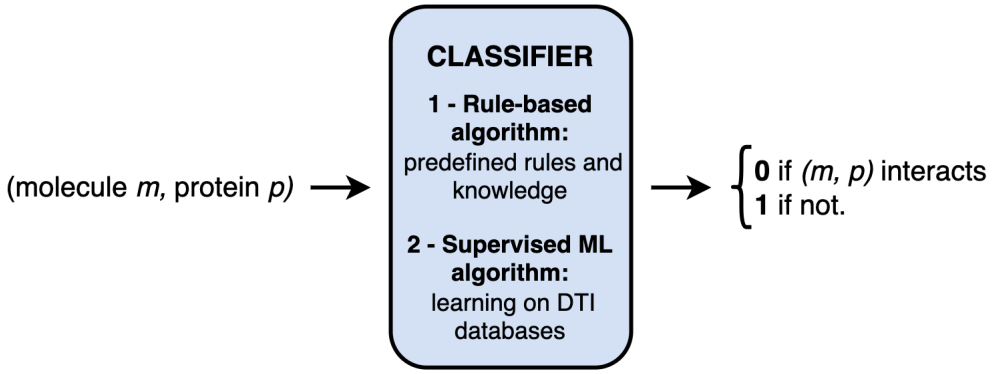


Figure 6.1 – Rule-based vs supervised ML algorithms in DTI prediction.

Knowing the interaction rules for all possible pairs (molecule m, protein p) is no easy task. ML algorithms enable large-scale predictions by overcoming this limitation.

6.3 Brief formalisation of DTI prediction

Let us briefly formalise the problem of DTI prediction using a ML algorithm.

The known (m, p) interactions represent the **samples** of study from which the algorithm will learn. This requires an encoding of the (m, p) pairs based on molecular and protein characteristics, which are called **features** (or **descriptors**). We can describe the pairs based on features that describe the molecules on the one side and on the proteins on the other side, and concatenate these features into one single vector describing the pairs.

This corresponds to a matrix $X \in \mathbb{R}^{n \times p}$ describing n interactions $x^i = (m, p)$ in the form of p features. Thus, $X_{ij} = x^i_j$ corresponds to the j-th feature of the i-th interaction. For each interaction x^i, we also know the **label** y^i, i.e. the value to be predicted. As we consider a classification problem, all the labels $\{y^1, y^2, ..., y^n\}$ belong to $\{0, 1\}$: 1 if the corresponding pair interacts, and 0 otherwise. Finally, the set of interactions and their associated labels $\mathcal{D} = \{(\vec{x}^i, y^i)\}_{i=1,...,n}$ forms the **training set**.

The goal of supervised learning is to find a function $\hat{f}$ such that $\hat{f}(\vec{x}) \approx y$, not only for the n observed interactions in the training set, but more generally for all possible pairs (m, p). Let's assume there exists a function f that assigns the labels to every interaction (m, p). The aim of the learning algorithm is to approximate the unknown function f with $\hat{f}$, referred to as the **predictive model**. The learning algorithm uses the training set $\mathcal{D}$ to determine (or "learn") $\hat{f}$, which is why it is also called **supervised** learning. Training the prediction model is therefore akin to an optimisation problem.

Once the predictive model is learned, i.e. once the function $\hat{f}$ is found such that the predictions $\hat{f}(x)$ are closest to the labels y, the prediction of a new interaction $x_{\vec{new}}$ is done by evaluating $f(x_{\vec{new}})$, in the same way as a "rule-based" algorithm.

It's important to note that in our case, the predictive model $\hat{f}$ does not directly outputs values in $0, 1$. Instead, it assigns an output in the interval of $[0, 1]$, representing

the probability of binding between the molecule and the protein. The higher the probability, the more likely the interaction exists. To determine whether the interaction is predicted positive, we need to set a threshold value, typically 0.5. If the probability score is above this threshold, the prediction is considered positive, if less, it is considered negative.

6.4 Chemogenomic ML algorithms

Various chemogenomics ML methods have been proposed in the last decade. They differ mainly by :

— the features used to encode the pairs of molecules and proteins.

— the learning algorithm used to learn the predictive model.

The methods can be categorized into two broad classes depending on the input data of the learning algorithm: the **feature-based** approaches and the **similarity-based** approaches.

In **feature-based** methods, samples (i.e. pairs of (molecule m, protein p)) are represented as feature vectors. The features reflect non exhaustively various types of physio-chemical, structural, topological or geometrical properties. Molecules are generally described with fingerprint vectors that encode the presence and absence of structural properties and proteins by their sequences or their physical or chemical properties [Playe, 2019]. The learning algorithm learns on these vectors of features to find the predictive model. For these approaches, one can use as learning algorithm, a **Random Forest** (RF) algorithm or a **Feed-Forward Neural Network** (FNN) algorithm, among others.

Similarity-based methods consider the similarity between samples as input data for the learning algorithm. They rely on the assumption that similar molecules, regarding structure, topology or physical properties etc., have similar functions and bioactivities and, therefore, have similar targets and vice-versa [Playe, 2019]. In fact, the similarities between pairs are calculated on the features calculated for each pair. There is therefore an additional step compared to feature-based methods: the matrix X_{ij} of features of all pairs is transformed into a similarity matrix S between each pair x^i using a predefined similarity measure. The learning model then learns to find a hyperplane that separates interacting pairs from non-interacting pairs in the pairwise similarity space. In this case, the learning algorithm are often **Multi-task Support Vector Machine** (SVM) or **Matrix Factorization** algorithms.

6.5 Performance criteria

The last step before using a classifier, learned using a learning algorithm, on new unknown data, is to evaluate the performance of this classifier. This evaluation is generally carried out on a **test set** $\mathcal{D}_{te} = \left\{ \left(\vec{x_{te}}^{\,i}, y_{te}^i \right) \right\}_{i=1,\dots,m}$, never seen by the classifier, i.e. a group of interactions x which have not been used to build the classifier and whose labels are known.

Once the predictions have been made for all the pairs of the test set $\vec{x^i_{te}}$, the predictions $\hat{f}(\vec{x^i_{te}})$ are compared to their labels y^i_{te}. In the case of a binary classifier, as in our case, a prediction is classified into 4 categories:

— True Positive (TP): predicted as a positive example when the actual label is indeed positive.

— False Positive (FP): predicted as a positive example when the actual label is indeed negative.

— True Negative (TN): predicted as a negative example when the actual label is indeed negative.

— False Negative (FN): predicted as a negative example when the actual label is indeed positive.

Performance scores are generally computed based on the number of TPs, FPs, TNs and FNs for the test set. But it is important to choose the score to optimise that is the most appropriate to the project, in order to select the best model.

If we return to our initial problem of target identification, our aim is to find new possible interactions, i.e. to predict new positive interactions. We are not interested in predicting negative interactions. In addition, the aim of *in silico* DTI prediction is to reduce the number of experimental tests to conduct. Therefore, we want to have as few false positives as possible among the top predictions.

In the article presented in the next chapter, we show that the scores generally used to evaluate the performance of ML algorithms in chemogenomics, such as the AUC-ROC or the AUPR, do not reflect the fact that there can be many false positives among the top predictions, due to the biases in DTI databases. We therefore propose to look at another score, the False Positive Rate (FPR), representing the fraction of negative among predicted positives. We show that this score is more adapted to the problem of target identification.

Chapter 7

Drug target identification with Machine Learning: How to choose negative examples.

Contents

Abstract

Identification of the protein targets of hit molecules is essential in the drug discovery process. Target prediction with machine learning algorithms can help accelerate this search, limiting the number of required experiments. However, Drug-Target Interactions databases used for training present high statistical bias, leading to a high number of false positives, thus increasing time and cost of experimental validation campaigns. To minimize the number of false positives among predicted targets, we propose a new scheme for choosing negative examples, so that each protein and each drug appears an equal number of times in positive and negative examples. We artificially reproduce the process of target identification for three specific drugs, and more globally for 200 approved drugs. For the detailed three drug examples, and for the larger set of 200 drugs, training with the proposed scheme for the choice of negative examples improved target prediction results: The average number of false positives among the top ranked predicted targets decreased and overall, the rank of the true targets was improved. Our method corrects databases' statistical bias and reduces the number of false positive predictions, and therefore the number of useless experiments potentially undertaken.

Résumé

L'identification des cibles protéiques des molécules à succès est essentielle dans le processus de découverte de médicaments. La prédiction des cibles à l'aide d'algorithmes d'apprentissage automatique peut contribuer à accélérer cette recherche, en limitant le nombre d'expériences nécessaires. Cependant, les bases de données d'interactions entre médicaments et cibles utilisées pour l'apprentissage présentent un biais statistique important, ce qui conduit à un nombre élevé de faux positifs, augmentant ainsi le temps et le coût des campagnes de validation expérimentale. Afin de minimiser le nombre de faux positifs parmi les cibles prédites, nous proposons un nouveau schéma de sélection des exemples négatifs, de sorte que chaque protéine et chaque médicament apparaissent un nombre égal de fois dans les exemples positifs et négatifs. Nous reproduisons artificiellement le processus d'identification des cibles pour trois médicaments spécifiques, et plus globalement pour 200 médicaments approuvés. Pour les trois exemples détaillés de médicaments, et pour l'ensemble plus large de 200 médicaments, l'entraînement avec le schéma proposé pour le choix des exemples négatifs a amélioré les résultats de la prédiction des cibles : Le nombre moyen de faux positifs parmi les cibles prédites les mieux classées a diminué et, dans l'ensemble, le classement des vraies cibles a été amélioré. Notre méthode corrige les biais statistiques des bases de données et réduit le nombre de prédictions faussement positives, et donc le nombre d'expériences inutiles potentiellement entreprises.

7.1 Preface

The aim of this second part of the project is to predict the "unexpected" off-targets of CFTR modulators from DTI databases. This problem can be formulated as a classification problem in which for each modulator, all (m, p) pairs are predicted as "interacting" or "not interacting". To solve this problem, we used a chemogenomics ML algorithm developed in the CBIO team.

For most chemogenomics methods, both **positive examples** of (m, p) pairs known to interact and **negative examples** of (m, p) known not to interact are required in the training set. However DTI databases store only positive examples and do not record negative ones. Training a chemogenomics algorithm is then considered as a Positive-Unlabelled (PU) learning problem. Nevertheless, molecules are expected to interact with a restricted number of proteins compared to the overall protein diversity. Thus, the majority of unknown interactions are usually considered as negative examples. In this context, it is then classical to randomly sample negative examples among the unknown interactions [Playe, 2018] to recover a balanced Positive-Negative (PN) learning problem. This strategy assumes that the ratio of "interacting" to "non-interacting" (m, p) pairs is so low that random selection would yield a high quality set of negative examples, that is to say, very few negative examples used for training would in fact be unknown positive ones.

Unfortunately, just like pathway databases discussed in the first part of this thesis, DTI databases present high bias towards a few proteins that have been more extensively studied than others. A small number of proteins have a large number of known ligands, while the majority of proteins have very few, if any, known interactions. This bias is undoubtedly due to drug discovery studies, which have focused on certain diseases rather than others, and on certain mechanisms rather than others. With a random sampling of negative interactions, we observed that all the top ranked predictions were enriched in "frequent hitters", i.e. proteins for which the highest number of ligands were recorded in the training dataset. The good ranking of these frequent hitters is due to the database bias, so that the top of the ranking is enriched with false positive predictions. There is thus a need to develop new training schemes that take database bias into account, in order to reduce the number of frequent hitters among the top predictions, thus reducing the number of false positives.

In this chapter we investigate how to best choose negative examples to minimize the number of false positives in DTI predictions. This work was made in collaboration with Chloé-Agathe Azencott, Benoit Playe and Véronique Stoven. Our research resulted in a publication in the *International Journal of Molecular Sciences* in June 2021. In the following section, the article is reproduced as published in the scientific journal.

Finally, in the last section of this chapter, we apply the chemogenomic algorithm with the newly developed negative examples selection scheme to predict the off-targets of the CFTR modulators. We also discuss the predictions with respect to the CF network, and to the CF scientific literature.

7.2 Drug target identification with Machine Learning: How to choose negative examples.

7.2.1 Introduction

Drug discovery often relies on the identification of a therapeutic target, usually a protein playing a role in a disease. Then, small molecular drugs that interact with the protein target to alter disease development are designed or searched for among large molecular databases. However, there has been a renewed interest in recent years for phenotypic drug discovery, which does not rely on prior knowledge of the target. In particular, the pharmaceutical industry has invested more efforts in poorly understood rare diseases, and for which therapeutic targets have not been discovered yet. While phenotypic drug discovery has made possible the identification of a few first-in class drugs [Swinney, 2011], once a phenotypic hit is identified, not knowing its mechanism of action is a strong limitation to fill the gap between the hit and a drug that can reach the market [Moffat, 2017]. More fundamentally, the target points at key biological pathways involved in the disease, helping to better understand its molecular basis.

Our work aims at helping determination of the protein targets for hit molecules discovered in phenotypic screens. Identification of a drug target based solely on experiments is out of reach because it would require to design biological assays for all possible proteins. In that context, *in silico* approaches can reduce number of experimental tests by focusing on a limited number of high probable protein targets. Among them, Quantitative Structure-Activity Relationship (QSAR) methods were developed for that purpose [Martinez-Lopez, 2017]. They are efficient methods for the inverse problem of finding new molecules against a given target, when ligands are already known for this target. However, using them to identify the targets of a given molecule would require training a model for each protein across the protein space, which is not possible because many proteins have only few, or even no, known ligand.

Docking approaches can address this question [Xu, 2018], but they are restricted to proteins with known 3D structures, which is far from covering the human proteome.

In the present paper, we tackle target identification in the form of Drug-Target Interaction (DTI) prediction based on machine learning (ML) chemogenomic algorithms [Vert, 2008]. These approaches can be viewed as an attempt to complete a matrix of binary interactions relating molecules to proteins (1 if the protein and molecule interact, 0 otherwise). This matrix is partially filled with known interactions reported in the literature and gathered in large databases such as the PubChem database at NCBI [Bolton, 2008]. They can be used to train ML chemogenomic algorithms by formulating the problem of DTIs prediction as a binary classification task, where the goal is predict the probability for pairs (m, p) of molecules and proteins to interact. They can be used both to predict drugs against protein targets, or protein targets for a drug, the latter being relevant to our topic.

Various ML algorithms have been proposed for DTI predictions. They include similarity-based (or kernel-based) methods such as kernel ridge linear regression, Support Vector Machines (SVM) [Jacob, 2008], or Neighborhood Regularized Logistic Matrix Factorization (NRLMF) that decompose the interaction matrix into the product of two matrices of lower ranks that operate in two latent spaces of proteins and

molecules [Liu, 2016]. Other ML algorithms are featured-based, which means that they rely on explicit descriptors for molecules and proteins, such as Random Forests (RF) [Svetnik, 2003], or Sparse Canonical Correspondence Analysis (SCCA) [Yamanishi, 2011]. Their prediction performances are usually very high when the training data are not too far from the (m, p) pairs in the test set [Playe, 2018]. Deep learning approaches relying on protein and molecule descriptors have also been proposed, but their prediction performances outperforms those of shallow learning methods only when the training data are very abundant, or when various heterogeneous sources of information are used in the context of transfer learning [Playe, 2020].

However, whatever the algorithm used, training a good ML chemogenomic model is hindered by biases in the DTI databases, such as whether the molecule for which one wishes to make predictions has known interactions or not [Pahikkala, 2015]. An additional issue arises when the databases only contain positive examples of (m, p) pairs known to interact, but no negative examples of (m, p) known not to interact. In this context, it is classical to assume that most unlabeled interactions are negatives, and to randomly sample negative examples among them [Playe, 2018]. In this work, we explore how to best choose negative examples to correct the statistical bias of databases, and reduce the number of false positive predictions, which is essential to reduce the number of biological experiments required for validation of the true protein targets. While the goal of the present paper was not to compare the prediction performances of various ML algorithms, we first compared the performances of two algorithms, namely SVM and RF, on the DrugBank dataset considered in the present study. We found that overall, SVM displayed the best results, and therefore, this algorithm was further kept to study how to correct learning bias.

7.2.2 Materials and Methods

Datasets

ML algorithms for DTI predictions need to be trained on datasets of known DTIs in which proteins and molecules are similar to those for which predictions will be performed. Hit molecules in phenotypic screens for drug discovery are mostly drug-like molecules [Lipinski, 2001], and proteins will be human proteins. We used the DrugBank database (version 5.1.5) [Law, 2014] to build our training dataset, because although much smaller than other databases like PubChem or ChEMBL, it provides high quality bio-activity information regarding approved and experimental drugs, including their targets, and contains around 17,000 curated Drug-Target Interactions (DTIs). Therefore, we built a dataset called DB-Database hereafter, that comprises all (m, p) DTIs reported in DrugBank involving a human protein and a small molecular drug. Overall, the DB-Database comprises 14,637 interactions between 2670 human proteins and 5070 drug-like molecules, which make up our positive DTIs. Because training a ML algorithm also requires negative examples, we added an equal number of negative DTIs to the DB-Database following two strategies:

— Random sampling: Negative examples were randomly chosen among the pairs (m, p) that are not labeled as a DTI but such that both m and p are in the DB-Database, under the assumption that most of the unlabeled interactions are expected to be negative. This process was repeated 5 times, leading to 5 training

datasets called RN-datasets (for Random Negatives-datasets) hereafter, differing only by their negative examples.

— Balanced sampling: To avoid biasing our algorithms towards proteins with many interactions, negative examples were randomly chosen among unlabeled DTIs, although in such a way that each protein and each drug appeared an equal number of times in positive and negative interactions. This process was also repeated 5 times, leading to 5 training datasets again differing only by their negative examples called hereafter BN-datasets (for Balanced Negatives-datasets). Building this set of negative DTIs is not trivial, and we propose the following algorithm:

1. Each protein and molecule in the DB-Database has a counter corresponding initially to its number of known ligands or targets, respectively;
2. For each protein, starting from those with the highest counter to those with a counter equal to 1, molecules are randomly chosen among those not known to interact with this protein and whose counter is greater or equal to 1;
3. Each time a negative DTI is chosen, the counter of the corresponding protein and of the molecule is decreased by one unit;
4. The process is repeated until all proteins and molecules counters are equal to 0.

Overall, the RN-datasets and the BN-datasets share the same set of positive DTIs, which are those in the DB-Database, and their total number of negative DTIs are the same and equal to that of positive DTIs. The construction of one RN-dataset (or one BN-dataset) is summarized in Figure 7.1.

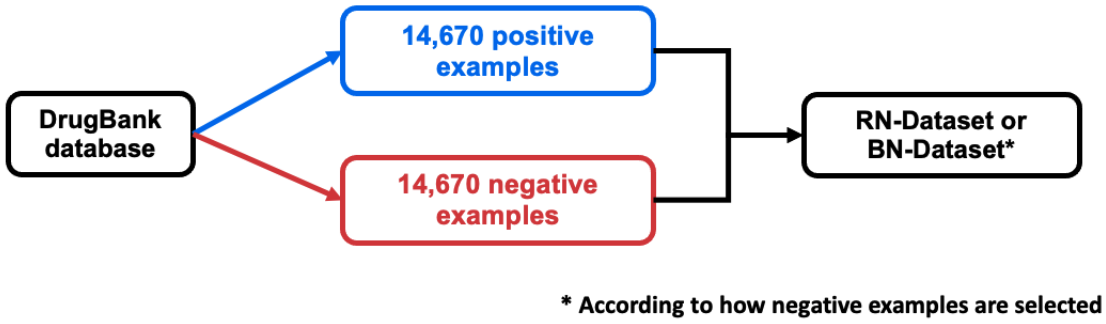


Figure 7.1 – Method for building one RN-dataset (or one BN-dataset).

Finally, to compare the performance of the algorithm trained on the RN-datasets or the BN-datasets when predicting targets for “difficult” molecules (hit molecules will generally be “difficult” molecules, in the sense that they will have no or few known targets), we considered a small dataset of DTIs involving 200 drugs that have few known targets. We built this dataset as follows: from the 5070 molecules in the DB-Database, we kept approved drugs that do not have more than 4 targets. This leads to 560 drugs involved in 851 interactions, among which we randomly selected 200 of these positive DTIs, involving 200 different drugs, defining the so-called 200-positive-dataset. 200 negative DTIs were also randomly chosen among all unlabeled DTIs involving theses

drugs and not belonging to the training RN- or BN-datasets, defining the so-called 200-negative-dataset.

All datasets are provided in the github repository mentioned under "Data Availability Statement".

Machine Learning Algorithms

Throughout the paper, the main algorithm we use to address target identification through a chemogenomics approach for DTI prediction is based on the Support Vector Machines (SVM) ML algorithm [Cortes, 1995]. Briefly, the SVM is trained on a dataset of known DTIs and learns the optimal hyperplane that separates the (m, p) pairs that interact from those that do not. While SVM can use vector representations of the data (i.e., descriptors for proteins and molecules), thanks to the so-called "kernel trick" [Schölkopf, 2004], they can also find this hyperplane based on particular similarity measures between (m, p) pairs of training dataset, and called kernel functions K, without requiring explicit representation of the data.

A general method to build a kernel on (m, p) pairs is to use the Kronecker product of molecule and protein kernels [Erhan, 2006]. Given a molecule kernel $K_{molecule}$ and a protein kernel $K_{protein}$, the Kronecker kernel K_{pair} is defined by:

$$K_{pair}((m, p), (m', p')) = K_{molecule}(m, m') \times K_{protein}(p, p') \tag{7.1}$$

For proteins, we used a centred and normalized Local Alignment kernel (*LAkernel*), which mimics the Smith–Waterman alignment score between two proteins [Smith, 1981]. For the molecules, we used a centred and Tanimoto kernel, that uses molecular descriptors based on the number of fragments of a given length on the molecular graph [Swamidass, 2005].

The *LAkernel* has three hyperparameters: the penalties for opening (o) and extending (e) a gap, and the β parameter which controls the contribution of non-optimal local alignments to the final score. The Tanimoto kernel has one hyperparameter: the length d of the paths up to which paths on the molecule structure are considered. According to [Playe, 2018], we used the following values for these hyperparameters: $o = 20$, $e = 1$, and $\beta = 1$ for the *LAkernel*, and $d = 14$ for the Tanimoto kernel. The SVM also requires a regularisation parameter classically called C, which controls the trade-off between maximising the margin (i.e., the distance separating the hyperplane and the two classes distributions) and minimizing classification error on the training points. This parameter was set to $C = 10$ for both RN- and BN-datasets, based on the nested cross-validation (CV) scheme, as described in Subsection 7.2.2.

SVM is a kernel-based ML algorithm. In the context of chemogenomics, it relies on similarity (or kernel) matrices between (m, p) pairs. Other algorithms, such as RF, are feature-based, and rely on explicit descriptors of proteins and ligands. To compare the performance of the kernel-based SVM to a feature-based approach, we compared our SVM to a RF on the RN-datasets. For the RF algorithm, we considered Extended-Connectivity Fingerprints (ECFP) [Rogers, 2010] as molecular descriptors, and 1920-dimensional feature vectors summarizing physicochemical properties as protein descriptors, as in [Ong, 2007]. We considered four hyperparameters for RF: the number of trees; the minimum number of samples required at a leaf node; the mini-

mum number of samples required to split an internal node; and the maximum depth of a tree. These hyperparameters were optimized based on a nested cross-validation scheme, as described in Subsection 7.2.2.

Performance Evaluation and Hyperparameters Optimisation

We used a nested cross-validation (CV), which allows to combine model selection and model evaluation without overfitting the dataset, as classically observed with a simple CV scheme ([Hastie, 2009; Cawley, 2010]). In the nested CV scheme, the CV procedure for hyperparameter optimization (called "the inner CV") is nested inside the CV procedure for performance evaluation (called "the outer CV"). The dataset is split into N folds: in each outer split, one fold is separated to form a test set. The N-1 remaining folds define an inner split. The hyperparameters are optimized on this inner split, based on a simple CV scheme. The set of hyperparameters providing the best inner CV prediction performance is then used on the test set of the corresponding outer split to evaluate the prediction scores. Thus, the model is tuned on the inner split, and performance of the model is evaluated on the test set of the outer split that was never used for model tuning. This procedure is repeated N times for each of the N outer splits, providing a mean and a variance for the performance scores. Figure S1 in Supplementary file presents a workflow chart describing a 5-fold nested CV used in the present study.

We used the following scores to quantify prediction performance of the classifiers:

— the Area Under the Receiver Operating Characteristic curve (ROC-AUC) [Hanley, 1982]. The ROC curve represents true positive rate as a function of false positive rate, for all thresholds on the prediction score. Intuitively, the ROC-AUC score can be interpreted as the probability that the classifier assigns a higher score to a positive interaction than to a negative interaction.

— fthe Area Under the Precision-Recall curve (AUPR) [Raghavan, 1989], which indicates how far the scores of true positives are from those of true negatives, on average;

— the Recall, representing the fraction of positive examples that are retrieved;

— the Precision, representing the fraction of true positives retrieved among predicted positives;

— the False Positive Rate (FPR), representing the fraction of true negatives among predicted positives.

More precisely, we used a $N = 5$ fold nested CV scheme to select the hyperparameter C of the SVM algorithm: RN-datasets (or BN-datasets) are split into $N = 5$ folds. Each fold comprises the same number of positive and negative DTIs. For the BN datasets, all molecules and all proteins appear in the same number of positive and negative DTIs, in each fold, as described in Subsection 7.2.2. Among the values {0.1, 1, 10, 100, 1000}, $C = 10$ consistently leads to the best performance across folds, both in terms of ROC-AUC and AUPR, and both on the RN- and BN-datasets.

We used the same nested CV scheme to optimize the hyperparameters of the RF algorithm (listed in Subsection 7.2.2) and to evaluate its performance on the RN-datasets. The number of trees was selected to be 600, chosen from {200, 400, 600};

the minimum number of samples required to be at a leaf node was selected to be 1, chosen from {1, 2, 5, 10}; the minimum number of samples required to split an internal node was selected to be 5, chosen from {2, 5}; and the maximum depth of the tree was selected to be 20, chosen from {10, 20}. The prediction scores were determined as for the SVM algorithm.

Flowcharts of DTI Prediction and Target Identification

In the present paper, we discuss two types of problems that we solve using ML algorithms: first, the prediction of new pairs (m, p) of interacting molecules and proteins, which we call DTI (Drug-Target Interaction) prediction, and second, the identification of new targets for a given drug. The former is only discussed in Subsection 7.2.3, where DTI prediction is used to evaluate the overall prediction capabilities of ML algorithms, and to determine the distribution of the prediction scores of positive and negative DTI, respectively. We used these distributions to determine thresholds for the latter problem, i.e., target identification for a given drug, which is the central topic of the paper. Figure 7.2 illustrates the pipeline for DTI prediction: 5 ML models are trained on 5 RN-datasets (or 5 BN-datasets), providing 5 interaction scores for each new (m, p) pair. These 5 scores are averaged to provide a final score. Figure 7.3 illustrates the pipeline for target identification: for each new drug d, 2670 (d, p) pairs are formed between this drug and each of the 2670 proteins p present in the DB-Database. DTI prediction is performed for each pair, as described above and illustrated on Figure 7.2. This provides a mean score of interaction with this drug for each of the 2670 proteins, which are then ranked accordingly. The candidate targets for this drug are the top ranked proteins with a score above a given threshold.

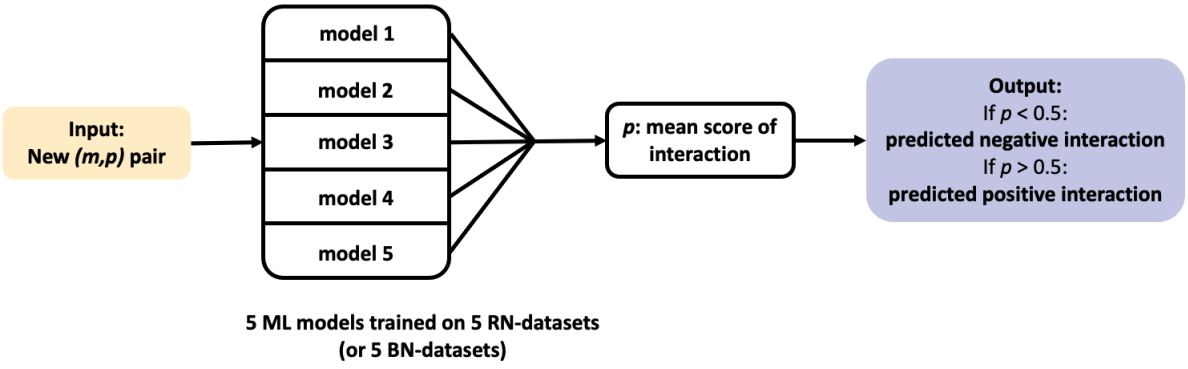


Figure 7.2 – Flowchart of the Drug-Target Interaction (DTI) prediction pipeline.

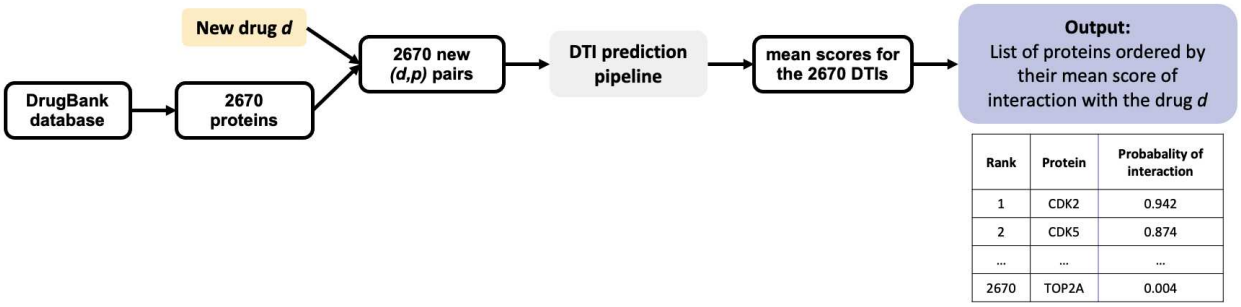


Figure 7.3 – Flowchart of the target identification pipeline.

7.2.3 Results

Performance of the SVM and RF Algorithms on the RN-Datasets

In the present paper, we focus on using ML chemogenomics approaches to identify target candidates for phenotypic hit molecules. The first step is to train the ML algorithms. More precisely, training a ML chemogenomics algorithm from a large DTIs database is an example of Positive-Unlabelled (PU) learning problem. Indeed, in practice, most databases only contain positive examples (that is to say, known DTIs), while all other possible interactions between molecules and proteins present in the data are unlabeled, whether because they have never been tested, or because they are negative interactions that have not been published or included in the database. Most of the unlabeled interactions are usually considered as true negatives. Therefore, in chemogenomics, the classical approach is to label as negatives a randomly chosen subset of the unlabeled interactions. This allows to convert a PU learning problem a into Positive-Negative (PN) learning problem for which many efficient ML algorithms are available.

We considered a ML algorithm based on SVM, with the *LAkernel* [Saigo, 2004] and the Tanimoto kernel [Swamidass, 2005] for proteins and molecules, respectively, because these methods displayed good prediction performances in previous chemogenomic studies, on average ([Wang, 2011; Meslamani, 2011; Playe, 2018]). The *LAkernel* is related to the Smith–Waterman score [Smith, 1981], but while the latter only keeps the contribution of the best local alignment between two sequences to quantify their similarity, the *LAkernel* sums up the contributions of all possible local alignments, which proved to be efficient for detecting remote homology.

While the purpose of this paper is not to discuss the choice of the ML algorithm, but rather to study how best to train it for the particular task of target identification, we also include a comparison of the SVM with a feature-based ML algorithm, i.e., Random Forests (RF) [Cao, 2014; Breiman, 2001].

The two algorithms were trained on the 5 RN-datasets described in Subsection 7.2.2, using a 5-fold nested cross-validation scheme, as detailed in Subsections 7.2.2 and 7.2.2. A threshold of 0.5 on the output score was chosen to discriminate between positive and negative predictions.

Table 7.1 shows the mean performance scores of the SVM and RF algorithms, when cross-validated on the RN-datasets. In the context of target identification, it is

important to limit the FPR, to avoid unnecessary experimental validation. However, a threshold of 0.5 over the probability scores was used to separate predicted positive interactions from predicted negative interactions, as classically, although in practical cases, a higher threshold would be chosen to select target candidates, in order to reduce the number of experimental tests to the predictions with the highest confidence. The results in Table 7.1 show that the SVM clearly outperforms RF across all performance scores, including FPR. We therefore retained the SVM for the rest of the paper.

Table 7.1 – Performance of the SVM and RF algorithms for DTI predictions on the RN-datasets.

Algorithm	AUPR	ROC-AUC	Recall	Precision	FPR
SVM	85.5 ± 0.2	88.0 ± 0.1	82.0 ± 0.4	93.3 ± 0.4	5.9 ± 0.4
RF	73.5 ± 0.8	79.1 ± 0.7	76.8 ± 1.0	80.6 ± 0.8	18.5 ± 1.0

We studied the distributions of the probability scores for positive and unlabeled (presumably, mainly negative) interactions for the SVM algorithm, according to the nested CV scheme. Figure 7.4 shows that these two distributions are well separated, and also suggests that on the RN-dataset, a threshold of 0.7 over the prediction score can be used to predict positive interactions with high confidence. In addition, the rank of a predicted interaction is also an important criterion to consider, because the goal of virtual screens is to drastically reduce the number of experiments to perform. When the goal is to identify hit molecules against a given therapeutic target, typically, the top 5% percent of the best-ranked molecules are screened [Adeshina, 2020]. Usually, an experimental assay with a simple readout has been set up for the target of interest, which allows to evaluate relatively high numbers of candidate molecules selected in the virtual screen. The inverse problem of target identification is more difficult because validation requires to test the phenotypic hit molecule in a different biological assay for each predicted target considered for experimental evaluation. This obviously requires much more time and effort, because these assays may not all be available, and therefore, may have to be designed. This can be a real challenge if the function of a candidate target is not suitable to design a simple biological test. Therefore, we added the stringent but realistic threshold of top 1% in rank. In other words, in the following, we will consider as candidate targets proteins with a predicted score above 0.7 and ranked among the top 1% of the tested proteins, to simulate a realistic experimental setting. We discuss how to best train the algorithm in order to minimize the number of useless biological experiments that would be undertaken for false positive targets satisfying these two criteria, because this represents a real bottleneck for real-case studies. Consequently, in what follows, since the DB-Database comprises 2 670 proteins, we will consider as candidate targets only proteins with a probability score above 0.7 and rank smaller than or equal to 27.

Statistical Analysis of the DrugBank Database

The DrugBank database [Law, 2014] is a widely used bio-activity database. While much smaller than PubChem or ChEMBL, it provides high-quality information for approved and experimental drugs along with their targets. It contains around 15,000 curated

DTIs involving 2670 human proteins (this set of proteins can be viewed as the "druggable" human proteome), and 5070 druglike molecules, corresponding to the DB-Database described in Subsection 7.2.2. This database is relevant for training of ML models for DTI predictions involving human proteins and drug-like molecules. However, Figure 7.6 shows that there is a strong discrepancy between the number of known ligands per protein, or known protein targets per molecule.

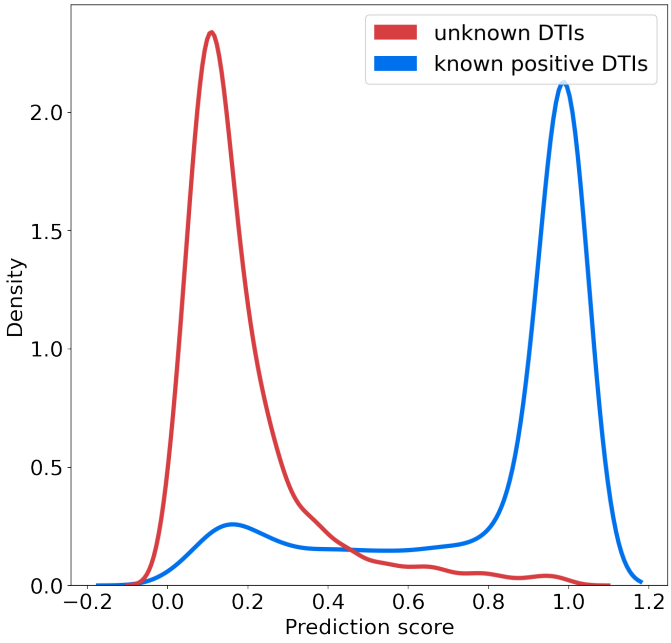
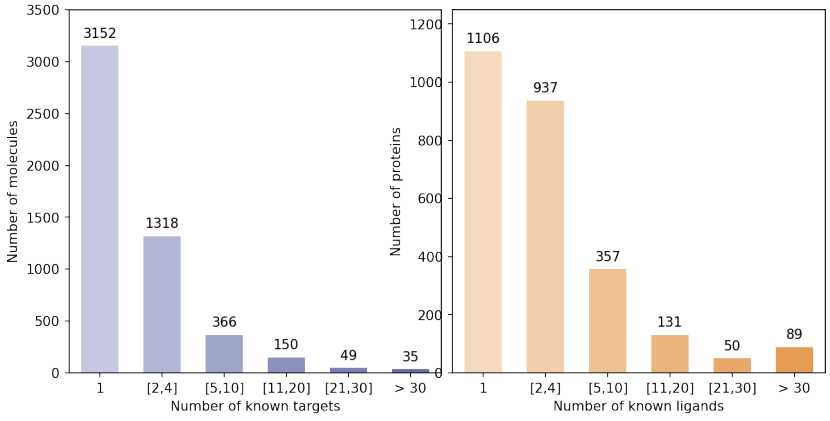


Figure 7.4 – Distribution of the probability scores predicted for known positive DTIs and randomly chosen negative DTIs among unlabeled DTIs.



(a) (b)

Figure 7.5 – Statistical bias in the DB-Database.
(**a**) Distribution of the molecules according to their number of targets in the DB-Database. (**b**) Distribution of the proteins according to their number of ligands in the DB-Database.

Indeed, the majority of proteins have 4 or fewer known ligands, while around 140 proteins have more than 21 ligands. We defined categories of proteins, depending on their number of known ligands (1, 2 to 4, 5 to 10, 11 to 20, 21 to 30, more than 30), and calculated the number of DTIs in the DB-Database in each category. Overall, according to Table 7.2, 5.2% of the proteins are involved in 44% of DB-Database DTIs.

This bias arises from the fact that a few diseases like cancer or inflammatory diseases have attracted most research efforts, and many ligands have been identified against related therapeutic targets, compared to other less studied human proteins. For example, Prostaglandin G/H synthase 2, a well-known protein involved in inflammation, has 109 drugs reported at DrugBank. This statistical bias affects training of the SVM and is expected to perturb identification of targets for hit molecules, potentially by enriching top ranked proteins in false positive targets that have many known ligands.

Table 7.2 – Distribution in the DB-Database of the number of DTIs involving proteins from various categories, according to their number on known ligands.

Protein nb of Ligands	**nb of Interactions**
1	1106
2 to 4	2527
5 to 10	2404
11 to 20	1920
21 to 30	1238
> 30	5442

Examples Illustrating the Impact of Learning Bias for Target Identification

Once trained, a ML algorithm identifies targets for a hit molecule by providing a list of proteins ranked by decreasing order of the estimated probability score of all (protein, hit) pairs. Candidate targets are chosen based on their probability score, their rank, and on potential prior biological knowledge that would highlight their relation to the considered disease. For example, a top ranked protein involved in cell cycle would be considered as a realistic candidate target for a hit identified in a cell proliferation screen in cancer research. The presence of many false positive targets among the top ranked proteins will not only lead to undertake useless experiments, but also potentially to discard true predicted targets pushed further down the list. Let us illustrate this problem in the case of 3 molecules, randomly chosen among marketed drugs with only one known target in DrugBank. Assuming that their targets have been well characterized because they are marketed molecules, most of the other top ranked predicted targets will be false positive predictions. The 3 considered molecules are: alectinib (DrugBank ID DB11363, target: ALK), lasmiditan (DrugBank ID DB11732, target: HTR1F), and doxapram also known as angiotensin II (DrugBank ID DB11842, target: AGTR1). We orphanized these 3 molecules (i.e., we suppressed their single known target from the train set), as if they were hits from phenotypic screens, and used the SVM algorithm presented in Subsection 7.2.2 on the RN-datasets to predict their targets. For each molecule, the results consist in a list of the 2670 proteins in the DB-Database, ranked by decreasing order of score.

As shown in the RN-datasets columns of Table 7.3, none of the known targets for

those drugs are among the candidate targets as defined in Subsection 7.2.3. More precisely, for DB11363 and DB11842, although the probability scores of their known targets are above 0.7 (values of 0.8 and 0.76 respectively), their rank is 31 in both cases, above the threshold of 27. For DB11732, the probability score of HTR1F is 0.67, with a rank of 107, and HTR1F would not either have been classified among the candidate targets for testing.

Analysis of the results highlighted that some of the best ranked candidate targets are frequent targets. For example, prothrombin F2 (120 ligands), cyclin dependant kisase CDK2 (137 ligands), and dopamine receptor 2 DRD2 (109 ligands) are top ranked predicted targets respectively for DB11842 (score of 0.97, rank 2), DB11732 (score 0.98, rank 1) and DB11363 (score 0.94, rank 5). The three ranked lists are provided in full in the github repository mentioned under "Data Availability Statement".

Table 7.3 – DTI prediction results for 3 marketed drugs, when the algorithm is trained on the RN-datasets or the BN-datasets: number of False Positive predicted targets, score and rank of the true target.

	RN-Datasets			**BN-Datasets**		
Drug	**FP**	**Target Score**	**Target Rank**	**FP**	**Target Score**	**Target Rank**
DB11363	27	0.8	31	16	0.8	3
DB11842	27	0.76	31	26	0.85	18
DB11732	27	0.67	107	26	0.83	17

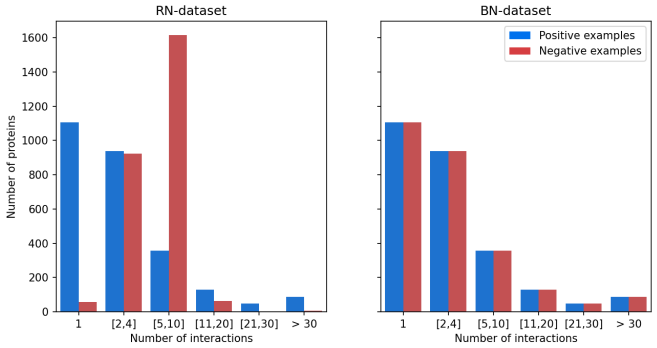
These examples illustrate the impact of false positive predictions for target identification, because they can lead to discard even high-scoring true targets as for DB11363 and DB11842.

Choice of Negative Examples to Correct Statistical Bias

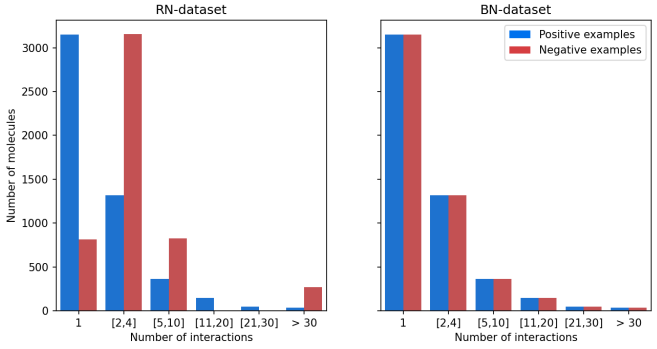
Our observation that high-scoring false positives tend to have a large number of known ligands led us to make the assumption that the model trained using randomly sampled negative interactions is biased towards proteins with many known ligands, as well as possibly drugs with many known targets. This suggested us to choose negative DTIs in such a way that the training dataset contains, for each molecule and for each protein, as many positive than negative DTIs. The corresponding so-called BN-datasets (for Balanced Negatives-datasets) are detailed in Subsection 7.2.2. Note that what we mean by "balanced" in the BN-dataset is that negative examples present the same bias as the positive examples: all molecules and all proteins appear in the same number of positive and negative DTIs. As shown in Figure 7.6: (1) in the positive examples, the distribution of known protein targets per molecule is similar to that of proteins known (chosen, in fact) not to interact per molecule in the negative examples; (2) in the positive examples, the distribution of known ligands per protein is similar to that of molecules known (chosen, in fact) not to interact per protein in the negative examples. This prevents proteins with many known ligands to be viewed by the algorithm as statistically much more probable targets, leading to many false positive predictions among this category of proteins. We recall that the BN-datasets contains the same positive DTIs as the RN-datasets, the former differing from the latter only by the

negative DTIs.

The SVM algorithm presented in Subsection 7.2.2 was trained on the BN-datasets. As discussed above, for the problem of target identification, reducing the number of false positives among the top-ranked proteins is critical. Table 7.4 reports, for prediction score thresholds of 0.5 (usually considered) and 0.7 (considered in the present paper), the cross-validated FPR scores on these two training sets. It shows a strong statistical bias in FPR for the RN-datasets between proteins with few or with many known ligands, and it illustrates that training on the BN-datasets greatly reduced this bias.



(a)



(b)

Figure 7.6 – Balancing the BN-datasets.
(**a**) Distribution of the proteins according to the number of positive examples or negative examples in which they are involved. (**b**) Distribution of the molecules according to the number of positive examples or negative examples in which they are involved.

Table 7.4 – Rate of false positives for proteins with various numbers of known ligands.

	FPR (Threshold = 0.5)		**FPR (Threshold = 0.7)**	
Prot in Category	**RN-Datasets**	**BN-Datasets**	**RN-Datasets**	**BN-Datasets**
0	2.2 ± 0.4	3.1 ± 0.5	0.5 ± 0.4	0.7 ± 0.5
1	3.7 ± 0.5	3.1 ± 0.8	1.5 ± 0.1	1.1 ± 0.7
2 to 4	5.1 ± 0.9	6.4 ± 0.8	2.4 ± 0.8	2.2 ± 0.8
5 to 10	9.9 ± 0.9	8.3 ± 0.6	4.4 ± 0.9	3.3 ± 0.5
11 to 20	13.8 ± 1.7	10.6 ± 0.5	7.3 ± 1.9	3.9 ± 1.1
21 to 30	23.0 ± 4.9	12.0 ± 3.0	11.4 ± 2.7	5.6 ± 2.0
> 30	18.6 ± 2.8	9.0 ± 0.4	11.0 ± 2.1	4.5 ± 0.3

To highlight the impact of this bias correction in terms of target prediction, we show in Table 7.3 the prediction results for the 3 molecules discussed in Subsection 7.2.3, when the algorithm is trained with the RN-datasets or with the BN-datasets. When trained on the RN-datasets, none of the true targets would have been considered as a positive candidate target for testing, because of a score below 0.7 or a rank above 27, as discussed above. Training on the BN-datasets greatly improved the ranks and scores of the three true targets, and reduced the number of false positives, allowing the 3 corresponding true targets to fulfill the rank and score criteria defined in Subsection 7.2.3 to become candidate target for testing.

To better illustrate the interest of the proposed scheme for the choice of negative DTIs on a larger number of drugs we considered the 200-positive-dataset consisting of 200 DTIs involving 200 marketed drugs with 4 of less known targets, as described in Subsection 7.2.2. This "difficult" test set was chosen because the aim was to mimic newly identified phenotypic hits, for which known targets are expected to be scarce. For each drug, we artificially reproduced the process of target identification: the corresponding DTI was removed from the train set, a new SVM classifier was trained and used to score 2670 DTIs involving this drug and all proteins of the DB-Database. We compared the top-ranked predicted targets obtained when the algorithm is trained on the RN-datasets versus on the BN-datasets, as well as the number of removed false positive DTIs that would have been retrieved as candidates for testing (i.e., with a score above 0.7 and a rank lower or equal to 27).

Overall, training with the BN-datasets improved the predictions: the number of false positive DTIs decreased for 106 drugs, remained unchanged in 85 drugs, and increased in 9 drugs, as compared to training with the RN-datasets. In particular, this improvement allowed one additional true positive interaction to reach a score above 0.7 and a rank below 27: 104 true targets were retrieved as candidates when training with BN-datasets, compared to 103 when training with RN-datasets. For the corresponding 104 drugs, the number of false positives decreased by 2.9 in average, and the rank of the true interactions decreased by 1.8 in average, bringing them even closer to the top ranked predicted proteins, and more likely to be chosen for experimental validation. Consistent with the results in Subsection 7.2.3 for the 3 example molecules, on average over the 200 considered molecules, the number of useless experiments potentially undertaken would have decreased when training with the BN-datasets.

We also made predictions for the 200 negative DTIs of the corresponding 200-negative-dataset, involving the same molecules as the 200-positive-dataset. Predictions

were made by the classifier trained on the RN- or BN- datasets. Overall the distributions of the prediction scores were very similar in both cases, centred around 0.2, and similar to that shown for the RN-dataset in Figure 7.4. Among the 200 negative pairs, only 2 pairs were predicted as positives, for the two RN- and BN- datasets. This can be viewed as a sanity check indicating that the proposed method did not introduce bias in the prediction of negative DTIs, while it globally improved the predictions of positive DTIs.

7.2.4 Discussion

The goal of the present paper was to tackle the question of protein target identification for new drug candidates, using ML-based chemogenomics. Indeed, these approaches can be run at a large scale in the protein space, including in their scope proteins with no known 3D structures, or proteins for which few, or even no ligands are known. Another key asset is that they can be applied to drugs with few, or even no known targets, as illustrated on the 200-positive-dataset. This is of particular importance because new phenotypic drugs are often orphan (i.e. have no known protein target) when they are identified. No other computational method presents these advantages. However, before making predictions, ML chemogenomic algorithms need to be trained on a database of known DTIs, which raises a few issues.

First, these databases are biased in terms of the number of protein targets per molecule, or of ligand molecules per protein, as shown for the DrugBank database used in our study. While we are aware that other and larger DTIs databases could have been used, the purpose of our study was not to discuss the choice of training set, in particular because other databases will also present the same type of bias as the DrugBank, for the same reasons. This point is rarely discussed in ML chemogenomic studies.

Second, the performance of ML algorithms in chemogenomics are usually evaluated based on AUPR and ROC-AUC scores in cross-validation procedures. However, the identification of true protein targets for phenotypic hit molecules in real case studies may become a challenge when the algorithm is trained on a biased dataset. Indeed, despite very high AUPR and ROC-AUC scores, false positive targets can be found among top-ranked proteins, and correspond to proteins with many known ligands. In target identification studies, biological experiments are guided by the predicted scores and ranks of candidate proteins. Training on a biased dataset may lead not only to conduct useless experiments, but also to discard true positive targets because their scores are below the considered threshold, or because their rank is too high due to the presence of false positives among the top-ranked proteins. This point is also rarely discussed in ML chemogenomic studies, usually focusing on cross-validation schemes that does not correspond to real case applications.

Third, training databases such as the DrugBank only contain positive examples, and therefore, negative examples are usually randomly chosen among unlabeled DTIs in order to train the ML algorithms. It is however unclear that this is an optimal choice for target identification.

The key result of the present paper was to show that choosing an equal number of positive and negative DTIs per molecule and per protein helps decrease the FPR in biased datasets, and improves the identification of true targets for a given drug. Three striking examples are given for the case study of three drugs (DB11363, DB11842,

and DB11732) that were "orphanized" (all their known DTIs were removed from the training set) to illustrate the most difficult situation encountered in the case of new phenotypic drugs: training with the BN-datasets allowed to recover the true target in all cases, while none of them would have been retrieved when training with the RN-datasets. To illustrate the advantage of the proposed scheme for the choice of negative interactions, we used a threshold of 0.7 over the probability scores to identify candidate targets for experimental testing, although proteins with scores above 0.5 are classified as positives. This threshold of 0.7 was guided by the results in Figure 7.4, in order to select highly probable positive targets. It can be adjusted to a different value if the algorithm is trained with other databases, whether through the same kind of plot, or through a ROC-curve in order to correspond to a predefined false positive rate.

We added the stringent threshold of 1% on the ranks of proteins to define which targets would be tested. This threshold could also be adjusted depending on available resources for experimental validation. The issue we identified and addressed in this paper does not depend on the scores and rank thresholds used, and choosing equal numbers of positive and negative DTIs per molecule and per protein for the training set will limit the number of false positives independently of the choice of thresholds, as shown in Table 7.4 in the case of the threshold on the prediction score. Finally, while the proposed scheme for the choice of negative examples was presented here in the context of target identification for hit molecules, it is of general interest and should be applicable to other types of PU learning problems when bias is present in the training set, which is a very common situation, in particular in many biological databases.

Data availabity Datasets and results, presented in this study, are available at `https://github.com/njmmatthieu/dti_negative_examples_data.git`, included a README.md file describing them.

Supplementary Materials Flowchart of Nested Cross Validation

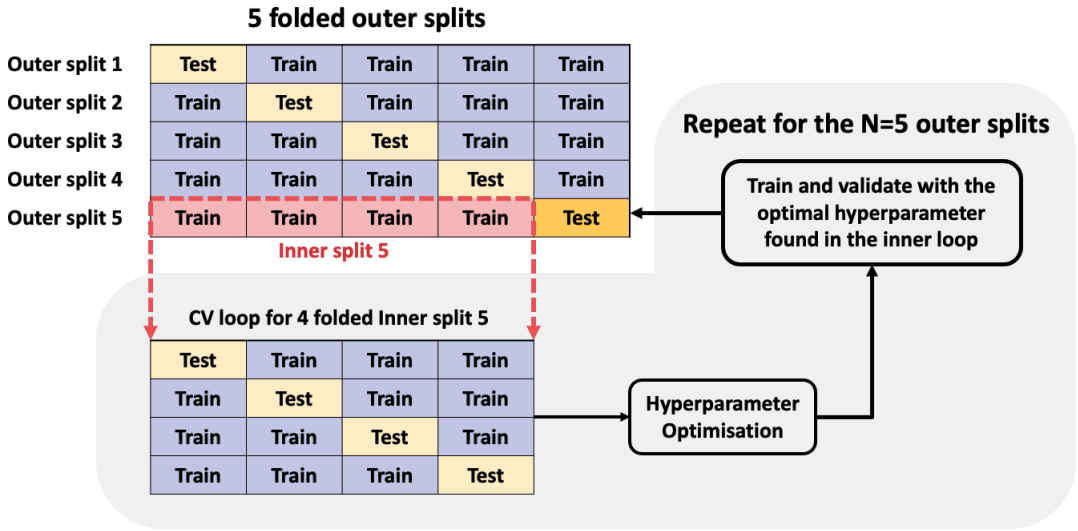


Figure 7.7 – Nested Cross Validation Workflow with N=5 outer splits.

7.3 Application to the DTI predictions on the CFTR modulators

7.3.1 General comments

The major contribution of this article lies in the discussion of two points rarely discussed in ML chemogenomic studies:

— Bias of DTI interaction databases, which leads to a high number of false positives among top ranked predicted interactions.

— Validation of DTI algorithms based on AUPR and ROC-AUC, which does not correspond to real applications of target identification. We propose to look at the False Positive Rate (FPR) instead.

To overcome these two challenges, we propose to choose an equal number of positive and negative DTIs per molecule and per protein for the training dataset. We tested this new scheme on a small dataset of DTIs involving 200 drugs that have few known targets and we observed a decrease of the FPR among the top targets compared to a random selection of negative DTIs for the training dataset.

This project highlightled the importance of including true, i.e. tested, negative examples of (m, p) interactions in public databases, particularly in the evolving landscape of AI approaches. ML algorithms require both positive and negative examples for effective predictions. Developing schemes in which negative examples are selected primarily from published "non-interacting" pairs, and subsequently from the pool of "unknown" pairs, could significantly improve predictions.

The aim of this paper was to address the challenge of identifying protein targets for new drug candidates. The pipeline can also be used reciprocally, i.e. identifying potential drugs ligand for proteins. Indeed, once a protein target is identified, one might want to find drugs that bind to the protein and that alter the disease progression. Even if a ligand is known for a protein target, a lot of work remains to be done to optimize the molecule in order to meet the ADME (Absorption, Distribution, Metabolism, Elimination), toxicity, and industrial synthesis requirements. Besides, some drugs are known for a small number of proteins. We could leverage this information to predict which drugs, already meeting the various prerequisites of ADME, would be most likely to bind the protein target.

7.3.2 Predictions of the CFTR modualtors targets

A key asset of chemogenomic algorithms is that they can be applied to drugs with few, or even no known targets. It is of particular interest in the CF project because CFTR modulators have been discovered thanks to HT phenotypic screens and their MoA has not yet been fully deciphered (see chapter 1). According to the DrugBank database, the four molecules (ivacaftor - VX-770, lumacaftor - VX-809, tezacaftor - VX-661, elexacaftor - VX-445) interact only with CFTR, although the corresponding interactions were removed from the train set.

We used the SVM algorithm presented in section 7.2.2, to predict the targets of the 4 modulators, using the same learning scheme as for the three examples of marketed drugs discussed in the article.

For each modulator, the pipeline provides a list of proteins ranked by decreasing order of the estimated probability score for all $(modulator, protein)$ pairs. This score is in fact the average of the prediction scores of the five classifiers used in the pipeline. We considered interactions with an average predicted score above 0.7 as predicted positive, as recommended in the article. The frequency plots of the average predicted scores for each modulator are presented in figure 7.8.

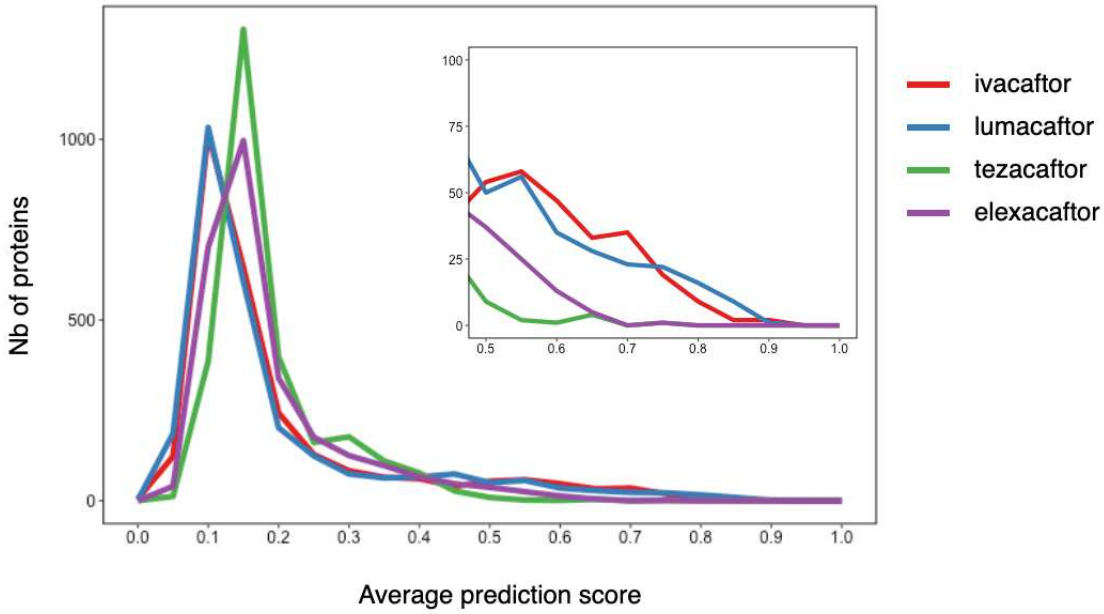


Figure 7.8 – Frequency plot of CFTR modulators prediction scores.
The inset shows a zoom of prediction score frequencies between 0.5 and 1.

The figure 7.9 shows the top 20 predicted targets for each of the 4 modulators. The rank is shown in Figure A and the prediction scores in Figure B. The list of these proteins and their associated scores are given in the appendix C.

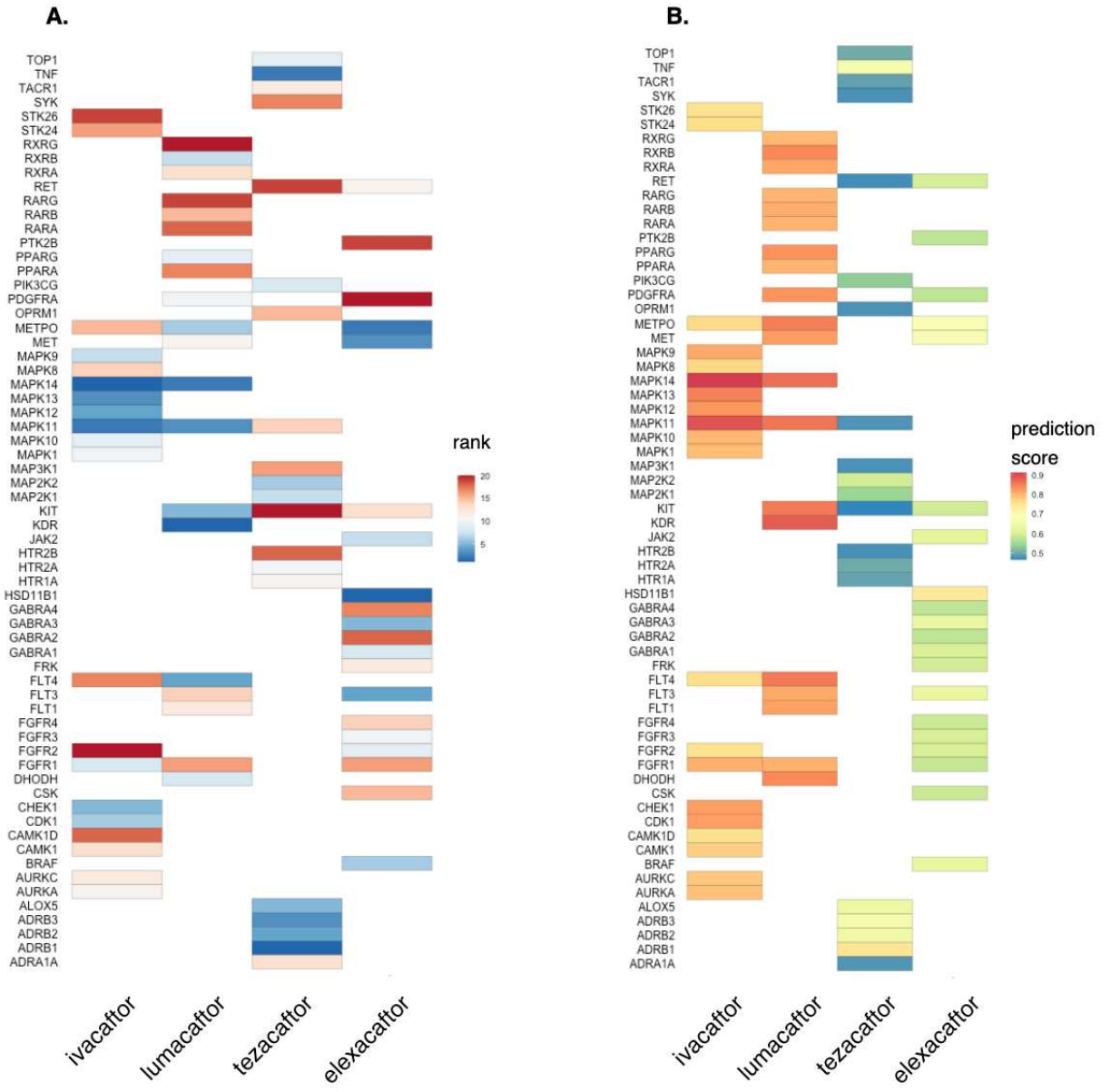


Figure 7.9 – Heatmap of the top 20 predicted proteins of the 4 CFTR modulators. Their ranks of prediction are given in A. and their predicted scores in B.

For tezacaftor and elexacaftor, only one protein is predicted positive: respectively ADRB1 for tezacaftor and HSD11B1 for elexacaftor. Conversely, for ivacaftor and lumacaftor, more than 50 proteins have an average predicted score above 0.7, including many kinases. The mitogen-activated protein kinases (MAPK or MAP kinases) and tyrosine kinases, are among the highest predicted targets (See Figure 7.9).

The kinase family attracted our attention for several reasons: (1) various kinases have been reported to play a role in CF. In particular MAPKs, are over-expressed or over-activated in CF cells. Verhaeghe et al. [Verhaeghe, 2007] showed over-phosphorylation (i.e. over-activation) of the ERK1/2 MAPKs in CF cells. Bérubé et al. [Bérubé, 2010]

also found increased immunoreactivity for the p38 MAPKs activity markers in CF lung biopsies. Finally, it has been shown that kinase inhibitors could improve the correction of F508del-CFTR function [Trzcińska-Daneluti, 2012]; (2) many kinases belong to the CF network, including the MAPKs; (3) Importantly, this family of proteins has been widely studied, and commercial tests are available as services to assay molecules. In particular, the Eurofins company offer such services. Therefore, based on commercial services at Eurofins, we tested the CFTR modulators against a panel of 50 kinases, among kinases with high probability scores in our predictions or known to play a role in CF, according to the literature. Lumacator, Ivacaftor and Tezacaftor did not present any inhibition property against the tested kinases. However, Elexacaftor was found to be a low, but significant, inhibitor for SYK, GSK3B, CSNK1A1, and MAPK1/ERK2, with a Ki value in the micro-molar range. These results will be discussed in the Conclusion section of the manuscript, in the more global context of the project.

With respect to the prediction of the algorithm, these results suggest two directions to improve the performances, by improving the training dataset, or by improving the algorithm itself.

7.3.3 Improving the prediction performances of the algorithm

Improving the training set

In the project, we trained the algorithm on a Drug-Bank-derived dataset, which contains around only 15.000 (m, p) pairs for 2.670 proteins. In addition, DrugBank dataset, which contains many indirect interactions stored as direct ones, which impacts the predictions.

Other larger training sets can be used, for example derived from larger databases such as PubChem [Bolton, 2011] or the Binding Database [Liu, 2007]. This database stores the Half maximal inhibitory concentration (IC_{50}) for each (m, p) pair tested. We can therefore consider interactions with IC_{50} of less than 10^{-4} Mol as negative interactions. A larger training dataset has been recently built in our laboratory, and it will be used to make new predictions for CFTR modulators.

We opted to predict across the entire druggable proteome, i.e. accross proteins belonging to many protein families. As a result, the classifier is less accurate than a classifier trained on a specific protein family. However, our classifier lacks the precision required for high confident predictions at the protein level within a given family of proteins (such as kinases, CPGRs etc...). It could be interesting to develop a dedicated classifier trained exclusively on the kinase family, using databases that contain a larger number of DTI, in order to improve the predictions in this family of proteins. In particular, this new training set could be modified to include the results from the Eurofins tests.

Improving the algorithm

Another way to improve the predictions is to refine the model. Playe et al. [Playe, 2018] showed that the type of the model chosen for the classifier (e.g. SVM, Matrix Factorisation, Random Forest, etc) does not have so much influence on the score of the prediction. However, one could play on the descriptors to improve the predictions. In the model presented in this chapter, we used the 2D structure of the molecules to

compute the molecules similarities and the nucleotide sequences to compute the proteins similarities. However, the binding of a ligand in a protein is a phenomenon occurring in the 3D space so 3D descriptors are expected to be more relevant. Many 3D descriptors have already been developed for the molecules but this is not the case for all the proteins. Indeed, access the 3D structure of all the proteins of the human proteome is not feasible. However, in recent years, more high-quality 3D crystallographic structures became available, and thanks to projects such as AlfaFold [Jumper, 2021], more 3D protein structures could be predicted for proteins with unknown structures. This could help the development of similarity measures based on the predicted 3D structures of the proteins.

Although 3D descriptors are more refined, they still have some limitations. For molecules, this would require to know the active conformation of the molecule, i.e. the conformation into which it binds to the protein. For proteins, it would require to consider the protein as a rigid structure, thus neglecting the induced fit often observed upon ligand binding. Despite these limitations, it would be interesting to explore to which extent 3D descriptors would improve DTI predictions.

In relation with this topic, during my PhD, I took part in a project where the 3D structures were used in ligand-based approaches. More precisely, this project was part of a study aimed at evaluating the performance of computational methods for solving scaffold hopping problems. In drug discovery, a *scaffold hopping problem* corresponds to the identification of novel chemotypes with biological activity similar to a known active molecule. In this study, we compared a few classical 2D and 3D ligand-based methods to the chemogenomic pipeline we developed, for the specific problem of scaffold hopping. The results showed that the similarities computed with 3D approaches performed better than the ones with 2D approaches. Moreover, the chemogenomic algorithm outperforms the ligand-based methods thanks to the information coming from the additional (m, p) pairs provided to the algorithm. The article was published in *Molecular Informatics* in January 2023 and is transcribed as published in appendix D.

Part IV

Conclusion

Chapter 8

Conclusion and perspectives

Contents

One objective of my PhD thesis is to better understand CF overall molecular dysregulations, by relating the absence of the CFTR protein to the CF cellular phenotypes. At a more applied level, this was expected to help identification of new therapeutic strategies in CF, particularly for patients who are not eligible to CFTR modulators, or who do not respond to these therapies. We propose to combine two fields of computational biology to approach these questions: systems biology methods and chemogenomics algorithms.

Both approaches provide insights into the molecular complexity of the disease. We built a single network of the signalling dysregulations of the CF cell, homozygous for the F508del mutation. However, it is essential to admit that a unique network is not sufficient to model the disease, as the precise molecular mechanisms may differ depending on the patient genetic background, the mutation, or the cell type. The proposed approaches provide opportunities to go one step further and to study these differences.

8.1 Results of the thesis

In this thesis, I first investigate systems biology approaches using CF transcriptomic data (**part II**).

This study required the development of tools that allow the detection of differentially expressed pathways. Therefore, in **chapter 3**, I present the R package, rROMA, for the representation and quantification of Module activity from omics data. rROMA is a sample-wise method dedicated to the analysis of bulk omics data at the level of the biological pathways. It assigns pathways activities to each sample, without requiring prior labels (such as "disease" or "control" labels). In the new version of the algorithm, I introduced the detection of shifted pathways, i.e pathways genes that are significantly differentially expressed in one direction, in addition to the detection of overdispersed pathways. rROMA stands out from the other methods because it detects both types of dysregulated pathways while providing a statistical assessment of the dysregulations. Furthermore, the numerous visualisation functions, and fine analysis of outlier genes inside the pathways make rROMA a user-friendly tool to the exploratory analysis of bulk omics data, without a priori hypotheses.

In **chapter 4**, I review the studies on systems biology approaches applied to cystic fibrosis. These studies have mainly focused on molecular mechanisms involved in CFTR processing, stability, and recycling. In **chapter 5**, I adopted a systems biology approach with a different goal, aiming at understanding how the absence of CFTR can be functionally linked to CF cellular phenotypes. I carried out a meta-analysis of 10 CF airway transcriptomic studies, focusing on the F508del mutation, at the level of the biological pathways. This allowed me to retrieve a list of 15 differentially expressed pathways. I used these pathways to build a signalling network, called **the CF network**, recapitulating the dysregulated signalling cascades that flow from the source nodes (proteins directly connected to CFTR) to the sink nodes (proteins that trigger CF cellular phenotypes). These phenotypes are consistent with those described in the CF literature, which indicates that our global approach did capture relevant biological information about CF. Five of the source nodes are upstream of all the sink nodes in the CF network: PLCB1/3, TRADD, SRC,and SYK . These proteins may collectively

initiate the emergence of CF phenotypes (together with the 3 other source nodes EZR, CSNK2A1, and PRKCA), illustrating the complexity of the disease. The topological analysis of the network also highlighted nodes with a high degree of betweenness centrality, which are other important players in the propagation of the dysregulations, including PI3KCA. Among these key source nodes and nodes with high degree centrality, SYK, SRC, PLCB1/3 and PIK3CA appeared as interesting candidate therapeutic targets. These proteins have already been discussed in the CF context. Interestingly, specific inhibitors are known for these proteins, and even marketed drugs in the case of SYK and PI3KCA. They stand out as potential therapeutic candidates for drug repositioning, potentially allowing the modulation of various CF phenotypes.

The methodology adopted for the F508del mutation, although perfectible, provided relevant and biologically interpretable results. However, we believe that the approach could be used for the study of other CFTR mutations. More generally, it could also apply to other monogenic diseases, particularly to rare diseases, in order to help understanding their biological determinants.

In parallel, I undertook a machine-learning study to explore the mechanisms of action of CFTR modulators. More precisely, in **part III**, I used machine-learning chemogenomics algorithms to search for potential off-targets. Because these drugs are active in CF, these targets give an indication of which proteins may belong to dysregulated pathways in CF cells. In practice, ML algorithms need to be trained on pairs of $(molecule, protein)$ known to interact and pairs known not to interact. However, "non interacting" pairs are not recorded in drug-target interaction (DTI) databases, so that "negative" interactions need to be chosen randomly among the unknown interactions. In addition, the DTI databases usually present a high bias towards a few proteins that have been extensively studied, and for which many ligands are known. In this context, random selection of negative interactions among unknown interactions in these databases lead to over-representation of false positive targets among the top ranked predicted proteins. I proposed a new scheme to select negative interactions in order to take into account bias observed in DTI databases, which allowed to reduce the number of the false positive rate in the predictions. The ML chemogenomic algorithm was trained with this new scheme, to predict the potential off-targets of the CFTR modulators.

The chemogenomic approach still needs to be improved because: (1) predicting targets for CFTR modulators, considered here as orphan molecules, is a difficult task, and (2) the DrugBank database used to train the algorithm is of modest size, compared to other databases such as BindingBD. In addition, it stores numerous indirect interactions, resulting in bias in predictions. Despite these challenges, the predictions provided valuable insights into the family of proteins of interest, namely, the kinase family. Experimental tests showed that elexacaftor was a modest inhibitor of SYK, MAPK1/ERK2, CSNK1A1, GSK3B. These four proteins belong to the CF network, and therefore, although elexacaftor is only a modest inhibitor of these proteins, one could hypothesize that part of its mechanism of action could involve these proteins. In such case, one could distinguish two possible mechanisms:

— elexacaftor targets CFTR and the kinases independently: it will improve CFTR processing and function, but also target these kinases. Both targets (i.e. CFTR

and the kinases) would contribute to reduce CF cellular phenotypes, leading to the clinical benefits observed in patients.

— the kinases targeted by elexacaftor may improve (directly or indirectly) CFTR processing, and targeting these kinases would participate to CFTR rescue at the PM.

Several experimental results support this second hypothesis. In particular, inhibition of kinases was shown to improve F508del-CFTR function [Trzcińska-Daneluti, 2012]. Similarly, the inhibition of MAPKs were shown to improve CFTR expression and has mainly been mentioned in contexts other than CF. Only very recently, it was acknowledged in the CF context. In particular, Xu et al. [Xu, 2015] revealed that the ERK pathway contributes to the degradation of CFTR in cells exposed to cigarette smoke, and reported that pharmacological inhibition of the MEK/ERK1/2 MAPK pathway prevented the loss of PM CFTR. Chang et al. [Chang, 2018] described that THC exposure downregulates the expression and function of CFTR in airway epithelial cells, resulting in the activation of the Epidemial Growth Factor Receptor (EGFR) protein and the ERK MAPK pathway. The inhibition of EGFR or the MEK/ERK pathway prevented the THC-induced regulation of CFTR. Very recently, Wellmerling et al. [Wellmerling, 2022] showed that ERK phosphorylation was increased in CF HBE cells compared to controls. The decrease of ERK phosphorylation by ectoine in combination with tezacaftor (VX-661) increased CFTR processing and function.

However, the analysis of the CF network showed that direct modulation of CF cellular phenotypes by inhibition of kinases should also be considered. This hypothesis is consistent with a recent paper underlying clinical benefits observed for CF patients receiving CFTR modulators, although bearing "unrescuable" mutations and in principle, not eligible to these therapies [Burgel, 2023].

8.2 Perspectives: Cystic fibrosis, a heterogeneous disease

In the discussions following each contribution, I mentioned how the project could be improved using newly developed methods or new data. I would like to suggest some broader perspectives on the differences in the molecular mechanisms of the disease on several scales.

In section II, I presented the construction of a CF signalling network caused by the absence of CFTR at the PM. We chose to build the CF network for the F508del mutation, but we are aware that reducing the CF dysregulations with only data corresponding to this mutation is simplistic. Indeed, the heterogeneity of CF patients' symptoms and the heterogeneity of CF patients' response to treatment demonstrate that molecular mechanisms caused by *CFTR* mutations may differ. Patients with different mutations may not have the same symptoms, and some patients with the same mutation sometimes exhibit various disease severity. Besides, the development of single-cell RNA sequencing technologies showed that different airway cell types do not express the same level of CFTR, which may trigger different cellular phenotypes depending on the cell type. Even if they relate to different scales of biology, we will refer to these differences as **heterogeneity**: patient heterogeneity or cellular heterogeneity. Systems

biology approaches offer computational frameworks to gain a deeper understanding of the complexity and the heterogeneity of CF biology.

8.2.1 CF patients bearing different mutations

Classes of *CFTR* mutations have been defined based on their primary biological defect on the CFTR protein, i.e. a default in mRNA expression, protein synthesis, maturation, or function. However, these classes have not been described in terms of cellular phenotypes, namely disturbed biological processes observed at the cellular level, as defined in this thesis. Therefore, it would be very interesting to apply the methodology used in the present thesis to datasets with patients with different mutations to characterise the mutation-specific cellular phenotypes.

Recent studies enabled to generate these data but they have been including very few samples, such as in the Bampi et al. [Bampi, 2020] study, or they have not provided the exact mutation for each sample, such as in Rehman et al. [Rehman, 2021]. A very recent paper analyses the differentially expressed genes (DEG) of human bronchial epithelial cells bearing different mutations [Santos, 2023]. These data were generated from cell line samples. Data from primary culture samples bearing different mutations should be published soon, and would be the most suitable datasets for this kind of analysis.

8.2.2 CF patients bearing the same mutation

CF patients bearing the same mutation exhibit various degrees of symptoms' severity and response to treatments [Cornet, 2022b]. As already discussed, it is difficult to find studies with many samples, in order to explore such heterogeneity. There is only one dataset with 124 CF patients, homozygous for the F508del mutation, but the data is not publicly available. In addition, in such dataset, each sample would need to be associated with severity of symptoms or response to treatment, but this type of data is not yet available.

Despite these limitations, it would be interesting to study large datasets and cluster samples according to their dysregulated pathways, thus exploring whether the obtained clusters correspond to disease severity. A network could be built for each cluster, or even a network for each patient, as it is increasingly being done for other diseases in the context of personalised medicine [Béal, 2021; Montagud, 2022]. Analysis of these networks would allow to study, in terms of molecular mechanisms, why patients present different symptoms' severity. The presence or absence of potential off-targets of CFTR modulators in the networks could help to understand why some patients respond well, and others poorly. This more detailed analysis would make it possible to optimise the therapeutic strategy at patient level.

8.2.3 Cellular heterogeneity of dysregulations

Finally, with the increasing development of single-cell RNA-Sequencing and Fluorescence-activated cell sorting (FACS) studies, the differences in signal transduction between different cell types could be studied in the context of the disease.

The study at the level of biological pathways (see chapter 5) enabled to classify the datasets considered in the meta-analysis into two subgroups according to their DEP: a subgroup comprising studies with dysregulated pathways in agreement with the pathophysiology of CF, and a second subgroup comprising studies with very few dysregulated pathways or even some in opposition to the pathophysiology of CF (see figure 5.2). We hypothesized that the datasets of the first subgroup would come from studies in which the cell differentiation media favoured secretory cells, whereas the datasets in the second subgroup would come from studies in which the media favoured ciliated cells over secretory cells. This hypothesis would lead to the conclusion that CF secretory cells would better account for signalling dysregulation than ciliated cells.

This would also be in agreement with a recent single-cell RNA sequencing study on airway epithelial cells [Okuda, 2021]. The analysis showed that CFTR expression is higher in secretory cells than in ciliated cells, where it is infrequent and low. Building CF network specific to secretory cells may provide a refined network that better models CF molecular dysregulations that could then be used to identify therapeutic targets specifically designed for this cell type.

These studies have also highlighted a new rare type of epithelial cells, called ionocytes, which expresses the highest level of CFTR. However, the proportion of ionocytes is around 0.3% while the proportion of secretory cells is around 20%. Hence, secretory cells dominate CFTR expression and function in human airway epithelia. Building a specific network for this cell type would facilitate the search for new therapeutic targets.

These hypotheses are very preliminary and deserve further investigation. A new study used single-cell RNA sequencing to compare CF to healthy samples [Carraro, 2021]. These data would help build cell-type-specific networks following the approach presented in this manuscript. One would expect from such studies that networks of the secretory cell type would be similar to our CF network, and that a much smaller network would be obtained for the ciliated cell type.

Part V

Appendices

Appendix A

Publications and communications

A.1 Publications

First author or co-first author

— **Najm, M.**, Azencott, C.-A., Playe, B. & Stoven, V. *Drug Target Identification with Machine Learning: How to Choose Negative Examples.* International Journal of Molecular Sciences 22, 5118. doi:10.3390/ijms22105118 (June 2021).

— **Najm, M.**, Cornet, M., Albergante, L., Zinovyev, A., Sermet-Gaudelus, I., Stoven, V., Calzone, L., Martignetti, L. *Representation and quantification Of Module Activity from omics data with rROMA* bioRxiv 2022.10.24.513448; doi: 10.1101/2022.10.24.513448 (May 2023)

— **Najm, M.**, Martignetti, L., Cornet, M., Kelly-Aubert, M., Sermet-Gaudelus, I., Calzone, L.* Stoven, V. *From CFTR to a CF signalling network: a systems biology approach to study Cystic Fibrosis* (to be published)

Co-author

— Chevalier, B., Baatallah, N., **Najm, M.**, Castanier, S., Jung, V., Pranke, I., Golec, A., Stoven, V., Marullo, S., Antigny, F., Guerrera, I. C., Sermet-Gaudelus, I., Edelman, A. & Hinzpeter, A. *Differential CFTR-Interactome Proximity Labeling Procedures Identify Enrichment in Multiple SLC Transporters.* International Journal of Molecular Sciences 23, 8937. issn: 1422-0067. doi:10.3390/ijms23168937 (Aug. 2022).

— Pinel, P., Guichaoua, G., **Najm, M.**, Labouille, S., Drizard, N., Gaston-Mathé, Y., Hoffmann, B. & Stoven, V. *Exploring isofunctional molecules: Design of a benchmark and evaluation of prediction performance.* Molecular Informatics, 2200216. issn: 1868-1751. doi:10.1002/minf.202200216 (Jan. 2023).

A.2 Communications

Conferences and seminars

— *Identification of new therapeutic targets by machine learning and systems biology approaches.* Conférence des Jeunes Chercheurs sur la Mucoviscidose, February

2020 Paris, France.

— *Construction of the cystic fibrosis biological network based on the meta-analysis of transcriptomic studies.*, European Young Investigators Meeting in Cystic Fibrosis (EYIM 2022), March 2022, online.

— *Representation and quantification of Module Activity from omics data with rROMA.* Statistical Methods for Post Genomic Data (SMPGD 2023), February 2023, Gent, Belgium.

Posters

— *Construction of the cystic fibrosis biological network based on the meta-analysis of transcriptomic studies.* European Conference of Computational Biology (ECCB 2022), September 2022, Barcelona, Spain.

— *Systems biology approach identifies dysregulated mechanisms in Cysitc Fibrosis*, Interdisciplinary Signalling Workshop (ISW 2023), July 2023, Visegrad, Hungary.

Appendix B

Differential CFTR-Interactome Proximity Labeling Procedures Identify Enrichment in Multiple SLC Transporters

International Journal of
Molecular Sciences

MDPI

Article

Differential CFTR-Interactome Proximity Labeling Procedures Identify Enrichment in Multiple SLC Transporters

Benoît Chevalier [1,2], Nesrine Baatallah [1,2], Matthieu Najm [3,4,5], Solène Castanier [1,2], Vincent Jung [6], Iwona Pranke [1,2], Anita Golec [1,2], Véronique Stoven [3,4,5], Stefano Marullo [7], Fabrice Antigny [8,9], Ida Chiara Guerrera [6], Isabelle Sermet-Gaudelus [1,2,10], Aleksander Edelman [1,2] and Alexandre Hinzpeter [1,2,*]

1 INSERM, U1151, Institut Necker-Enfants Malades, 75015 Paris, France
2 CNRS UMR 8253, Université Paris Cité, 75015 Paris, France
3 Center for Computational Biology, Mines Paris-PSL, PSL Research University, 75006 Paris, France
4 Institut Curie, 75248 Paris, France
5 INSERM U900, 75428 Paris, France
6 INSERM US24/CNRS UAR3633, Proteomic Platform Necker, Universit Paris Cité—Federative Research Structure Necker, 75015 Paris, France
7 Institut Cochin, Université Paris Cité, INSERM, U1016, CNRS UMR 8104, 75014 Paris, France
8 Faculté de Médecine, Université Paris-Saclay, 94210 Le Kremlin-Bicêtre, France
9 INSERM UMR_S 999, Hôpital Marie Lannelongue, 92350 Le Plessis-Robinson, France
10 Centre de Référence Maladies Rares Mucoviscidose et Maladies de CFTR, Hôpital Necker Enfants Malades, European Reference National (ERN) Lung Center, 75015 Paris, France
* Correspondence: alexandre.hinzpeter@inserm.fr

Abstract: Proteins interacting with CFTR and its mutants have been intensively studied using different experimental approaches. These studies provided information on the cellular processes leading to proper protein folding, routing to the plasma membrane, recycling, activation and degradation. Recently, new approaches have been developed based on the proximity labeling of protein partners or proteins in close vicinity and their subsequent identification by mass spectrometry. In this study, we evaluated TurboID- and APEX2-based proximity labeling of WT CFTR and compared the obtained data to those reported in databases. The CFTR-WT interactome was then compared to that of two CFTR (G551D and W1282X) mutants and the structurally unrelated potassium channel KCNK3. The two proximity labeling approaches identified both known and additional CFTR protein partners, including multiple SLC transporters. Proximity labeling approaches provided a more comprehensive picture of the CFTR interactome and improved our knowledge of the CFTR environment.

Keywords: proximity labeling; cystic fibrosis; CFTR; SLC transporters; KCNK3; interactome

check for updates

Citation: Chevalier, B.; Baatallah, N.; Najm, M.; Castanier, S.; Jung, V.; Pranke, I.; Golec, A.; Stoven, V.; Marullo, S.; Antigny, F.; et al. Differential CFTR-Interactome Proximity Labeling Procedures Identify Enrichment in Multiple SLC Transporters. *Int. J. Mol. Sci.* **2022**, *23*, 8937. https://doi.org/10.3390/ijms23168937

Academic Editors: Carlos M Farinha and Martina Gentzsch

Received: 12 July 2022
Accepted: 8 August 2022
Published: 11 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Cystic fibrosis (CF), the most common monogenic life-threatening disease, is caused by mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene [1], which encodes for a chloride channel located at the apical membrane of respiratory epithelial cells [2].

CFTR protein partners have been intensively studied, enabling a better understanding of the cellular processes leading to proper protein folding, its transport to the plasma membrane, recycling and degradation. Numerous protein partners implicated in these different steps have been identified (reviewed in this special edition [3]). They have often been identified based on the comparison between WT-CFTR and the CFTR-F508del mutant [4,5], the most frequent CF-causing mutation, or other misfolded mutants [6,7]. These interactions occur in different cellular compartments, which correspond to different steps in the CFTR biogenesis route. The first set of protein partners locates within the endoplasmic reticulum and is mainly implicated in CFTR synthesis and folding (reviewed in Refs. [8,9]). Some of them are implicated in the ER quality control (ERQC) of CFTR,

recognizing misfolded channels and targeting them to proteasomal degradation. ERQC includes different checkpoints involving both chaperones, e.g., calnexin, calreticulin, Hsp70 and their co-chaperones, and specific motifs located on CFTR, such as RXR motifs implicated in ER retention and a diacidic exit code (DAD), which is involved in the recruitment of CFTR cargo into vesicles budding form ER exit sites [10]. It has been proposed that proper folding of CFTR reduces the accessibility to RXR motifs, favoring the ER exit of correctly folded channels [10–12]. After complex glycosylation in the Golgi apparatus, CFTR is exported to the plasma membrane where it associates with different types of proteins, such as membrane anchoring proteins, which link the channel to the cytoskeleton, or endosomal proteins implicated in the vesicular recycling of CFTR [3,13]. As in the ER, a peripheral quality control system monitors protein quality and targets altered channels to lysosomal degradation [14,15]. Finally, once at the cell surface, CFTR channel activity is mainly regulated by phosphorylation of its regulatory domain [16]. Several kinases participate in this regulation, mainly PKA [17,18] and, to a lesser extent, PKC [19,20] and tyrosine kinases [21]. Recently, Mihalyi et al. showed that CFTR association with PKA initiated conformational changes leading to channel activation, the phosphorylation of specific residues being necessary to maintain the effect over time [22]. Similarly, a specific protein–protein interaction between CFTR and WNK1 was recently shown to modulate channel selectivity toward bicarbonate versus chloride ions [23], an effect independent of the kinase activity of WNK1.

CFTR has also been shown to modulate the cell surface activity of other channels and transporters, such as ENaC [24], ORCC [25], SLC26A9 [26–28], SLC26A3 [29] or SLC26A6 [30]. Co-activation of CFTR and SLC26 transporters was associated with direct interactions between the STAS domain of SLC26 transporters and the R domain of CFTR [27,29]. The CFTR C-terminal PDZ domain also plays a key role in protein–protein interactions at the plasma membrane, anchoring CFTR to the cytoskeleton and enabling interactions with other PDZ containing proteins via PDZ-binding proteins, such as NHERF1 [9].

The CFTR interactome appears to be location specific and highly dynamic, affecting several steps in CFTR biogenesis, turnover and activity. Several approaches have been used to identify CFTR partners, such as yeast two-hybrid screens and CFTR immunoprecipitation coupled to mass spectrometry. These strategies have provided detailed CFTR interactome maps and are constantly improving. While yeast two-hybrid screens remain challenging for transmembrane proteins. These strategies have provided detailed CFTR interactome maps and are constantly improving. While, yeast two-hybrid screens remain challenging for transmembrane proteins leading to the use of CFTR fragments as baits, technological advances now enable to screen full-length CFTR in mammalian cells [31] Immunoprecipitation approaches require cell lysis with detergents that may modify the interactome compared to interactions taking place in a living cell [32–35]. Furthermore, the specificity of the obtained interactomes depends on the availability and specificity of antibodies as well as the experimental conditions.

To address these limiting aspects, new techniques have recently been developed to label protein partners in a native environment. These include, among others, BioID, TurboID and APEX2 proximity labeling enzymes [32–35], which are fused to the protein of interest. BioID is an Escherichia coli biotin ligase, which biotinylates proteins on lysine residues in a radius of approximately 10 nm. While the low activity of BioID usually requires between 18 h to 24 h of labeling, the sequence optimization of the enzyme resulted in a mutant ligase, called TurboID, characterized by enhanced enzymatic activity, reducing the labeling time to 10 min [33]. Another strategy is based on APEX2, a peroxidase allowing the labeling of protein partner electron-rich amino acid residues with a biotin derivative (Biotin-Phenol) at a spatial resolution of approximately 20 nm [33,35]. The labeling reaction is induced by adding H_2O_2 for a short period of time in living cells (1 min), providing a snapshot of the proximal interactome.

In this study, we explored and compared CFTR interactomes using three proximity labeling approaches, i.e., APEX2, BioID and TurboID. Experiments were performed in

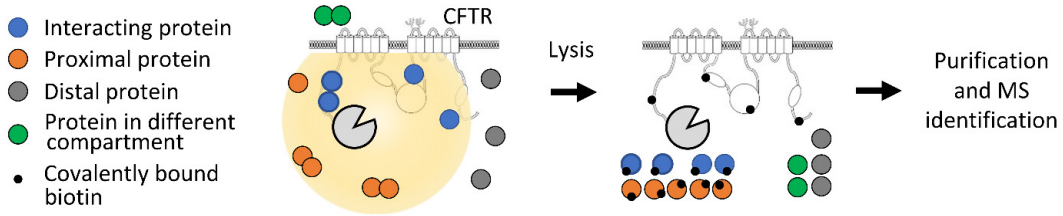
transiently transfected HEK293 cells to achieve high expression levels of fusion proteins and facilitate mass spectrometry identification.

2. Results

2.1. Proximity Labeling Approaches

The coding sequences of BioID, TurboID and APEX2 were subcloned upstream of that coding for CFTR to generate fusion proteins containing the enzymes at the N-terminus of CFTR. A linker region consisting of five glycine-serine repeats (GS5) motifs was introduced between BioID/APEX2/TurboID and CFTR to improve flexibility and to decrease CFTR near-end crowding. The activity of these fusion proteins enabled the labeling of proteins interacting with CFTR or proximal proteins within a radius of 10–20 nm, while distal proteins or proteins separated by a membrane were not labeled (Figure 1A). The covalent labeling of interacting and proximal proteins with biotin was not affected by cell lysis procedures under denaturing conditions, such as with a RIPA buffer. Biotinylated proteins were purified using streptavidin-coated beads, washed and eluted in denaturing Laemmli buffer [35]. Samples were then digested and analyzed by mass spectrometry for protein identification (Figure 1A). While the overall procedure was similar for the three fusion proteins, some specificities exist, such as the length of labeling times (from 1 min to 18 h) and the targeted amino acids (Lys versus Tyr/Trp/Cys, His) (Figure 1B). The peroxidase activity of APEX2 requires biotin-phenol as a substrate and H2O2 addition to activate the enzyme, while BioID and TurboID require a biotin pulse for labeling (Figure 1B).

A



B

Enzyme	Type	Size (kDa)	Labelling time	Substrate	Target
APEX2	Peroxidase	28	1 min	Biotin-phenol + H_2O_2	Tyr, Trp, Cys, His
BioID	Biotin ligase	35	18 h	Biotin	Lys
TurboID	Biotin ligase	35	10 min	Biotin	Lys

Figure 1. Characterization of fusion proteins. (**A**) Schematic representation of the proximity labeling strategy. Labeling enzymes were fused to the N-terminus of CFTR to biotin tag interacting and proximal proteins, while distal proteins and proteins separated by a membrane were not tagged. Labeled proteins were purified with streptavidin-coated beads and identified with mass spectrometry. (**B**) Characteristics of the fused enzymes used in the study, indicating the type of activity, the size, the recommended labeling time, the substrate used and the targeted amino acid.

2.2. Characterization of Fusion Proteins

N-terminal fusions of BioID, TurboID and APEX2 with CFTR WT were first analyzed by Western blot. Untagged CFTR expressed in HEK293 cells was detected under the form of two bands, one corresponding to core-glycosylated CFTR (band B), the second, more diffuse band corresponding to fully glycosylated CFTR (band C). Fusion proteins showed the same pattern and intensity, with a size shift increase of approximately 30 kDa compared to WT CFTR due to the fusion of BioID, TurboID or APEX2 (Figure 2A). In addition to their conserved maturation pattern, fusion proteins were active, as shown by a halide

sensitive fluorescent assay following cAMP stimulation, in the presence of the VX-770 potentiator (Figure 2B). These results are consistent with previous studies, indicating that the fusion of a GFP tag to the N-terminus of CFTR preserves the functional CFTR chloride channels [36]. As BioID labeling requires a much longer incubation period (18–24 h), the activity of BioID-CFTR fusion protein was measured after this longer biotin labeling. The results revealed enhanced channel activity (Figure 2C), concomitant with greater amounts of fully glycosylated BioID-CFTR in the Western blot analysis (Figure 2A). These results indicate that biotinylation of CFTR or CFTR partners enhanced channel stability at this prolonged time point.

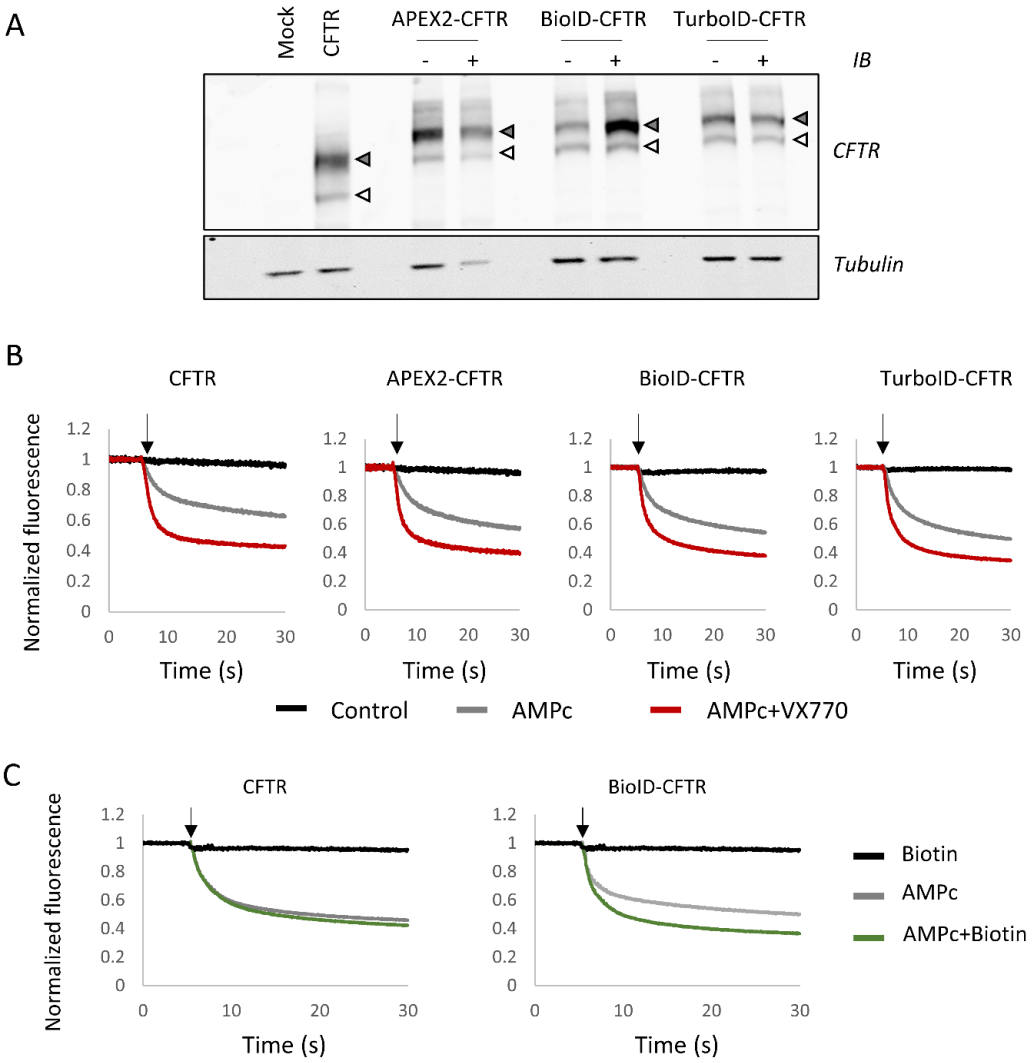


Figure 2. Characterization of CFTR fusion proteins. (**A**) HEK293 cells are transfected with either an empty vector (Mock), CFTR without fusion (CFTR) or CFTR fused with proximity labeling enzymes (APEX2-CFTR, BioID-CFTR and TurboID-CFTR). For CFTR fusions with proximity labeling enzymes, cells are either untreated (indicated as −) or treated under biotinylation-inducing conditions (indicated as +) (see Materials and Methods). The upper panel corresponds to the detection of CFTR with band B (white arrow head) and band C (gray arrow head). In the bottom panel, Tubulin was used to assess equal loading. (**B**,**C**) Halide-sensitive YFP assay of fusion proteins measured in HEK293 transfected cells. The arrow indicates the time point of PBS NaI injection. Measures were performed in control conditions (Control, black) after a 30 min incubation with cpt-AMPc/IBMX to activate CFTR (AMPc, gray) or with cpt-AMPC/IBMX and 1 µM VX-770 (AMPc + VX-770, red). Cells were also incubated with 10 mM biotin for 18 h (**C**) prior to measurements (Biotin, black and AMPc + Biotin, green).

2.3. Mass Spectrometry Identification

The labeling capacity of the fusion proteins was assessed in Western blots using fluorescent streptavidin. Biotinylation by APEX2 was initiated in living cells pre-incubated with Biotin-Phenol by the addition of H2O2 in the cell media for 1 min, while BioID and TurboID labeling required a biotin pulse of 18–24 h and 10 min, respectively. Upon the activation of APEX2, BioID or TurboID, a smear was visible, corresponding to CFTR partners that were biotin labeled in transiently transfected HEK293 cells (Figure 3A). The overall biotinylation intensity was similar between the different conditions, with notably higher labeling of CFTR with BioID and TurboID compared to APEX2 and some differences observed in the patterns (Figure 3A). Immunohistochemistry showed that the fusion protein was enriched at the cell surface, where the highest level of biotinylation was visible (Figure 3B). The diffusion of some biotinylated proteins within the cytoplasm reflected most probably the mobility of the proteins within the cell.

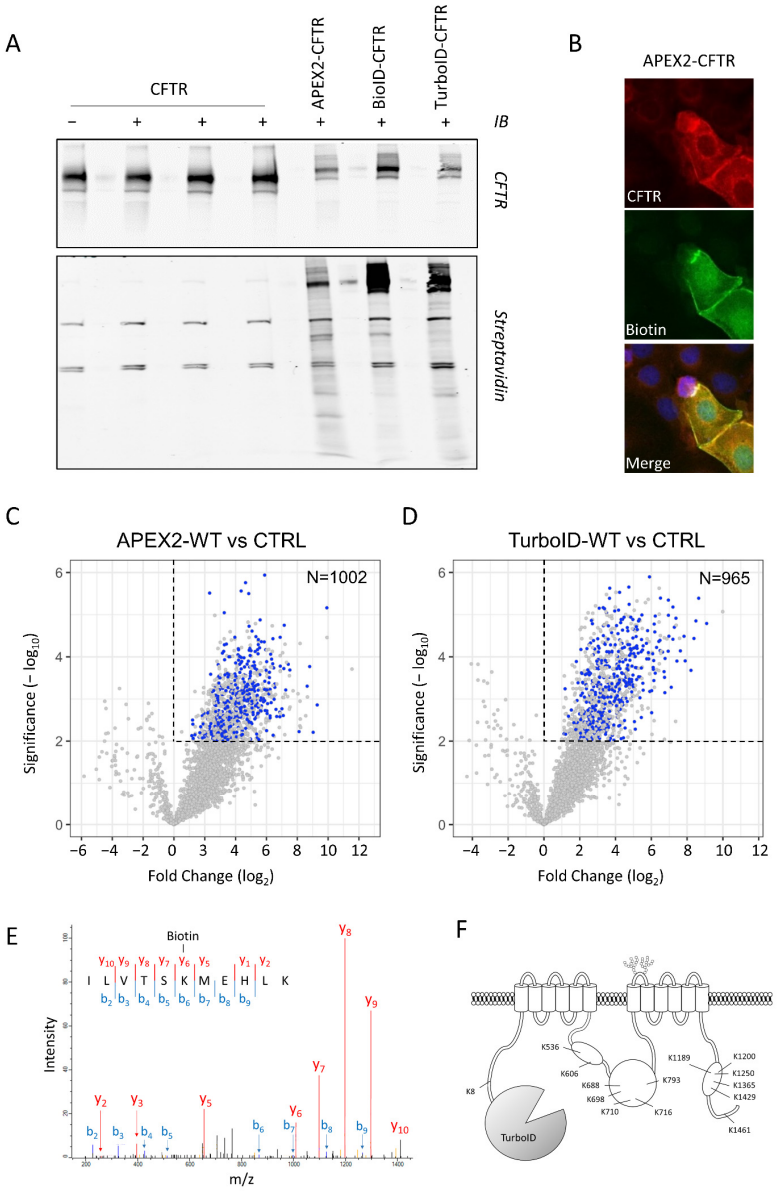


Figure 3. MS identification of CFTR partners. (**A**) Western blot analysis of CFTR fusion proteins and biotinylated proteins. The upper inlay was probed with CFTR antibody, while the lower was probed

with streptavidin. Each labeling procedure was performed on both untagged and tagged CFTR, in the same order. (**B**) Localization of APEX2-CFTR transfected CFBE cells. CFTR was identified using CFTR antibody 24.1, biotinylated proteins using streptavidin-Alexa488 and nuclei using Hoechst dye. (**C**,**D**) Volcano plots of APEX2-CFTR (C, $n = 3$ replicates) and TurboID-CFTR (D, $n = 4$ replicates) versus matched non-transfected HEK293 cells ($n = 3$ and $n = 4$ replicates). APEX2-CFTR identified a total of 3088 proteins and TurboID-CFTR a total of 3054 proteins, of which 1002 and 965, respectively, are enriched compared to non-transfected controls. Blue dots indicate 433 proteins identified in both sets (Student's *t*-test; *p*-value < 0.01). (**E**) Fragmentation mass spectrum of CFTR peptide 1184–1199 aa with biotinylation located on K1189 (one example of the 14 biotinylated peptides found for CFTR). (**F**) Position of the 14 biotinylated CFTR lysines identified in the TurboID-CFTR samples.

CFTR protein partners were identified using both APEX2 and TurboID labeling procedures (see Materials and Methods). After labeling, cells were lysed and the biotinylated proteins purified using streptavidin-coated beads. Mass spectrometry analysis of purified biotinylated preys identified more than one peptide in 3088 proteins for the APEX2 and in 3054 proteins for the TurboID procedure (Supplementary Table). Among them, 1002 and 965, respectively, were enriched in positive samples as compared to the non-transfected negative control (Student's *t*-test, *p*-value < 0.01), and 433 proteins were found enriched with both procedures (Figure 3C,D, blue dots).

Moreover, CFTR was found to be biotinylated as well, and multiple peptides carried the modification (Figure 3E), confirming the specificity of the procedure and validating the approach. Specifically, biotinylated lysine residues were located in the different cytoplasmic regions of CFTR: N-terminus, NBD1, R-domain and NBD2 (Figure 3F), without any labeling within or across the membranes.

2.4. Analysis of Proximal Datasets and Comparison to Biogrid

The two total datasets were then analyzed using the Significance Analysis of INTeractome (SAINT) probabilistic scoring tool [37] to identify the high confident proximal partners (FDR < 1%) for APEX2 ($n = 1091$) and TurboID ($n = 939$) (Figure 4A,B). The comparison of these groups of high confident CFTR proximal partners identified 435 common proteins, representing 39.7% of the APEX2 group and 46.3% of the TurboID group (Figure 4C). We then compared the datasets to the Biogrid database (Figure 4D,E), which categorizes interactants reported in the literature from low-throughput studies (LTPs, in green) on specific CFTR interactants and from high-throughput experiments (HTPs, in dark gray) corresponding to immunoprecipitation studies followed by mass spectrometry [4,38]. The APEX2 and TurboID procedures identified a similar proportion of CFTR interactants found by LTPs and HTPs. Interestingly, some interactants from LTPs that were not identified previously by HTPs were detected by proximity labeling—10/48 proteins for APEX2 and 7/48 proteins for TurboID (Figure 4D,E). Of note, a large number of partners reported in the LTPs (Figure 4A,B, in green) and LTP + HTP (Figure 4A,B, in orange) datasets were part of the high confident proximal partners (FDR < 1%), while proteins in the HTP dataset showed more dispersion (Figure 4A,B, in dark gray).

We then performed a gene ontology (GO) enrichment analysis of the 435 proteins identified in both the APEX2 and TurboID datasets to delineate the cellular functions and biological processes involved in CFTR biogenesis function or regulation. We observed enrichments with terms associated with protein localization and intracellular vesicular transport of CFTR (Figure 4F and Supplementary Figure S1A,B). One of the strongest enrichments corresponded to SNAP receptor activity (Supplementary Figure S1B,C), which is involved in membrane fusion during vesicular transport. Some of these SNAREs were reported to interact with CFTR and to impact CFTR biogenesis [39]. Proteins associated with small GTPases were also highly enriched, including regulators of the Rab small GTPases (RAB11FIP1 and RAB3GAP2) involved in vesicular transport and effectors of the RhoA signaling pathway (ROCK1/2) (Supplementary Figure S1B,C). This finding is in agreement with previous studies reporting cross talks between the CFTR function or

processing and the RhoA/ROCK pathway, including Refs. [40–42]. APEX2 and TurboID proximity labeling also enabled the identification of interaction partners of CFTR folding, including chaperones and proteins involved in the ubiquitination process (Supplementary Figure S1B,C).

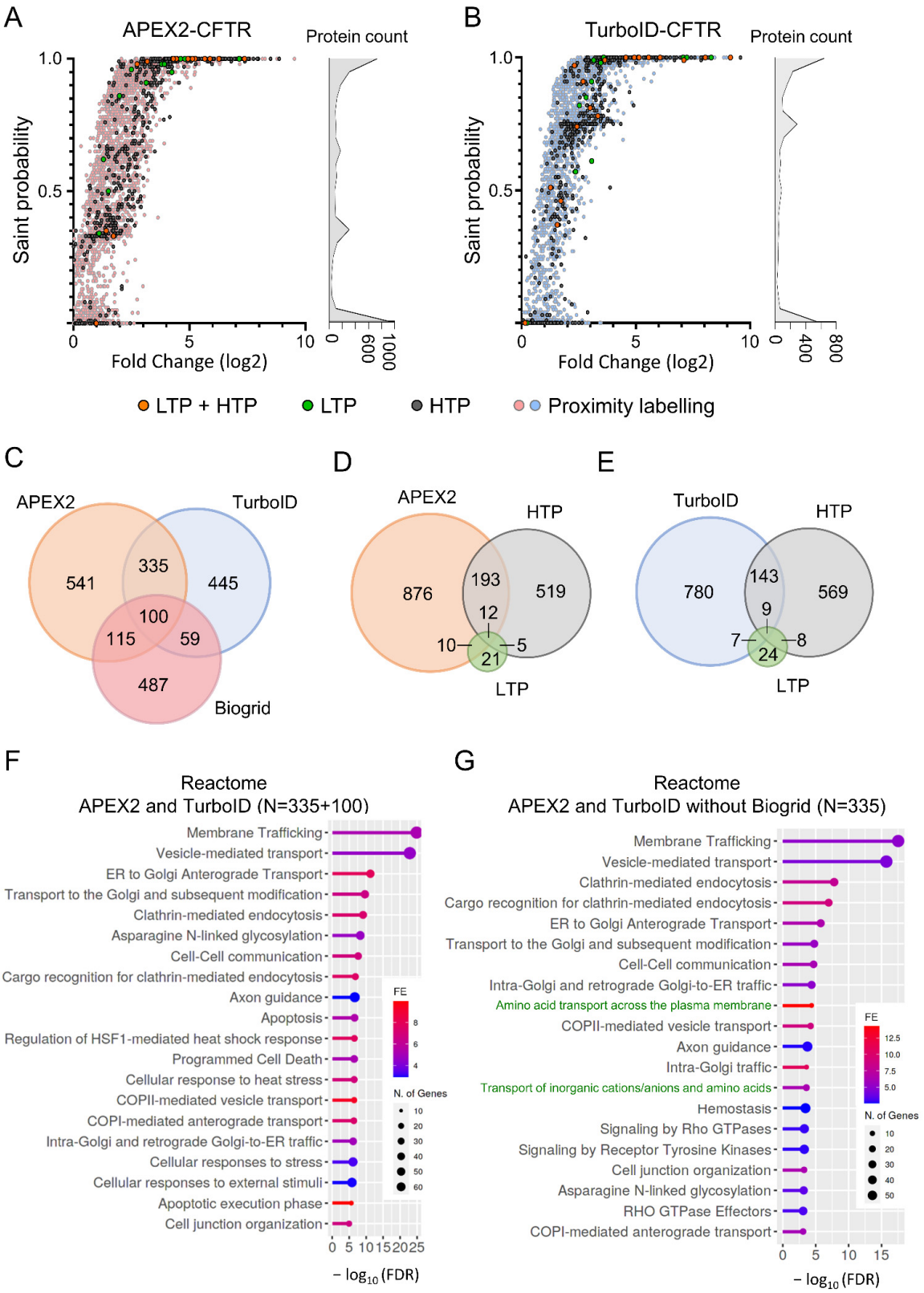


Figure 4. Analysis of proximal datasets and comparison to Biogrid. (**A**,**B**). The computational tool

SAINT assigns confidence scores to protein–protein interaction. Analysis using SAINT of the datasets obtained with the APEX2 ((**A**), $n = 3$ replicates, 4490 total proteins) or the TurboID ((**B**), $n = 4$, 3356 total proteins) procedure. The X axis indicates the fold change of intensities for each individual interaction compared to control purifications. CFTR partners also referenced in Biogrid database, either from low-throughput (LTP, in green), high-throughput (HTP, in dark gray) or both (LTP + HTP, in orange) studies are also indicated. (**C**) Venn diagram performed between APEX2 and TurboID datasets. A FDR < 1% was used to identify high confident proximal partners ($n = 1091$ for APEX2 and $n = 939$ TurboID). (**D**,**E**) Venn diagram performed between APEX2 (**D**) and TurboID (**E**) datasets and CFTR partners referenced in Biogrid database, either from low-throughput (LTP) or high-throughput (HTP) studies. (**F**,**G**) Reactome enrichment analysis of the 435 proteins identified with both APEX2 and TurboID procedures as high confident proximal partners (FDR < 1%) (**F**) or 335 specifically detected using both APEX2 and TurboID but not referenced in Biogrid with, in green, terms associated with solute transport (**G**).

We finally evaluated whether these approaches could reveal novel biological signaling pathways or CFTR functions. We therefore searched for enrichments within a set of proteins common to both APEX2 and TurboID but not yet described as interacting with CFTR in the Biogrid database (N = 335 proteins). The Reactome database identified enrichment for biological pathways associated with CFTR trafficking, as described with the previous set (Figure 4F,G). Multiple terms associated with solute transport were also found to be enriched, each of them containing SLC transporters (Figure 4G, in green). Detection of SLC transporters was more efficient using APEX2 ($n = 18$) and especially TurboID ($n = 34$) as compared to methods referenced in Biogrid ($n = 11$) (Supplementary Figure S1D). This is also true for the global detection of transmembrane proteins, as TurboID detected almost twice as many (30.6%) transmembrane proteins compared to APEX2 (16.2%) and the Biogrid set (16.7%) (Supplementary Figure S1E).

2.5. Comparison of CFTR-WT Versus Mutant CFTR-G551D and -W1282X

In order to evaluate the impact of mutations on the dynamics of the CFTR network, we compared the interactome of CFTR-WT with mutant CFTR-G551D and CFTR-W1282X. These two mutations induce distinct functional defects. CFTR-G551D alters channel gating by affecting ATP binding and/or NBD dimerization while preserving the global architecture and localization of the channel. Proximity labeling of CFTR-G551D showed important similarities with CFTR-WT. A comparison of proteins identified in each replicate showed few changes in the proximal interactome obtained with both APEX2 (22 out of 1966 proteins) and TurboID (9 out of 1654 proteins) (Student's *t*-test, *p*-value < 0.1, Supplementary Figure S2A,B). However, the gene set enrichment analysis (GSEA) indicated an enrichment for some GO terms, indicative of enhanced proximity of CFTR to the endoplasmic reticulum membrane (APEX2 dataset) or other GO terms associated with the plasma membrane as well as the actin cytoskeleton (TurboID dataset). Nonetheless, among the identified interacting proteins, no proximal partners appeared lost or gained for G551D (Supplementary Figure S2C,D), which suggests that the CFTR interactome is minimally perturbed for this mutant.

The W1282X mutation truncates part of NBD2 and the end C-terminus, which contains the PDZ domain of the protein, leading to both protein instability and abrogation of channel function. The differential interactome between WT and W1282X showed several differences between the APEX2 and TurboID assays, with 101 and 280 proteins enriched (Student's *t*-test, *p*-value < 0.1) (Supplementary Figures S2E and Figure 5A). The gene enrichment analysis for APEX2 indicated an enrichment of W1282X with mitochondria-associated terms (Supplementary Figure S2F). For TurboID, the analysis indicated enrichment of proteins related to the misfolding protein response (Supplementary Figure S3A,B) in addition to the expected significant loss of terms related to the plasma membrane (Figure 5B). Additionally, the Interprot domain enrichment analysis showed the expected drastic drop in the number of proteins with a PDZ domain (Figure 5C). Among these proximal CFTR interacting

partners, scaffolding proteins, such as SLC9A3R1 and SLC9A3R2 (NHERF1 and NHERF2), were strongly reduced or lost with W1282X (Figure 5D). This loss of interaction was particularly observed with the TurboID approach and, to a lesser extent, with the APEX2 approach (Figure 5D).

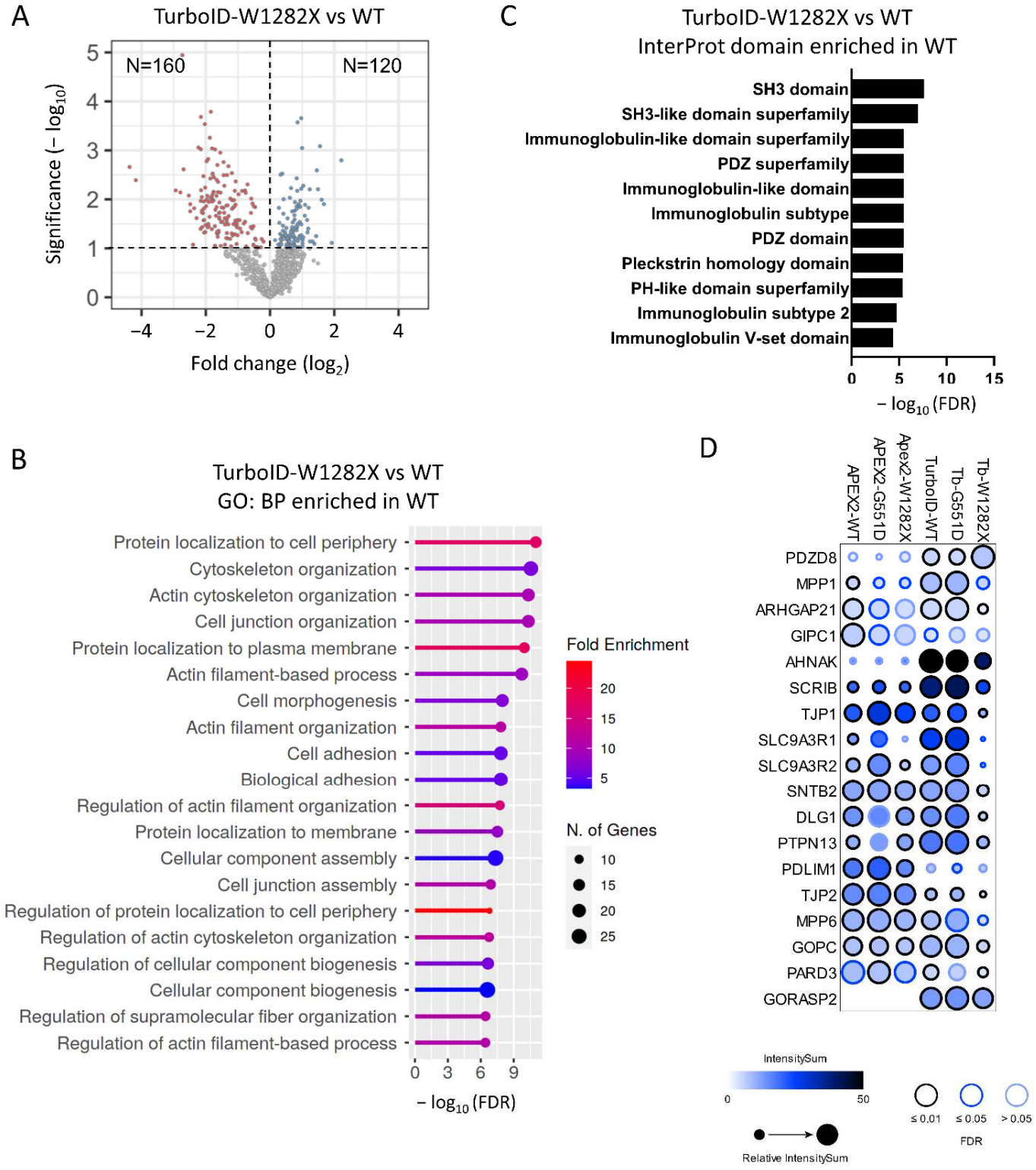


Figure 5. Analysis of TurboID-CFTR-W1282X proximal dataset. (**A**) Volcano plot of TurboID-CFTR-

W1282X versus TurboID-CFTR-WT transfected HEK293 cells (n = 4 replicates, 1654 total proteins). Blue dots indicate proteins identified as enriched in the W1282X sample (Student's t-test, p-value < 0.1, n = 120) and red dots in the WT sample (Student's t-test, p-value < 0.1, n = 160). (**B**) GO enrichment terms identified in TurboID-CFTR-W1282X. (**C**) Enrichment analysis of Pfam domain in proteins enriched in TurboID-CFTR-WT compared to TurboID-CFTR-W1282X. (**D**) PDZ domain proteins enriched in TurboID-CFTR high confident proximal partners (FDR < 1%), shown as dot plots with ProHits-viz [43]. The color of each circle represents the intensity; the circle size indicates the relative value of the intensity across APEX2 and TurboID and confidence in the measurement via colored edge.

2.6. Comparison of TurboID-CFTR and KCNK3-TurboID Interactomes

In order to highlight the protein partners, which specifically and selectively bind to CFTR, we next compared the CFTR interactome with that of KCNK3. KCNK3 is a pH-dependent, voltage-insensitive, background outward potassium channel, structurally unrelated to CFTR but with similar biogenesis and final cell localization. It is formed by protomers of four transmembrane domains, which dimerize to form the pore (Figure 6A). A C-terminal KCNK3-TurboID fusion protein was generated, which maintained the same activity as the untagged channel in whole-cell patch-clamp recordings performed in transiently transfected HEK293 cells (data not shown). Proximity labeling was performed with KCNK3-TurboID in HEK293 cells, and, as observed for TurboID-CFTR, a biotinylated peptide was identified, corresponding to the modification of the amino acid residue K320 (Figure 6A). Comparison with the TurboID-CFTR dataset (Figure 6B,C) showed the presence of both common protein partners and proteins more specific to either CFTR or KCNK3 (Student's t-test, p-value < 0.01). Pathway enrichment analysis showed that common partners were mainly involved in protein biogenesis with enriched terms associated with Golgi vesicle transport and intracellular protein transport (Supplementary Figure S4A). The analysis of proteins interacting specifically with CFTR indicated a strong enrichment in terms associated with the plasma membrane and proteins containing a PDZ binding domain (Figure 6B,D, orange, and Supplementary Figure S4B). However, the strongest enrichments corresponded with terms associated with the activity of transporters (Figure 6B,D, green), among which we found proteins belonging to three large families of transporters: SLC transporters (Figure 6E), ATP transporters (Supplementary Figure S4B) and ABC transporters (Supplementary Figure S4B). The 53 SLC proteins detected with CFTR were completely absent in the KCNK3 interactome and partially lost with CFTR-W1282X (Figure 6E). However, the effect was much smaller on ATP and ABC transporters (Supplementary Figure S4B), suggesting a close proximity of CFTR with multiple SLCs.

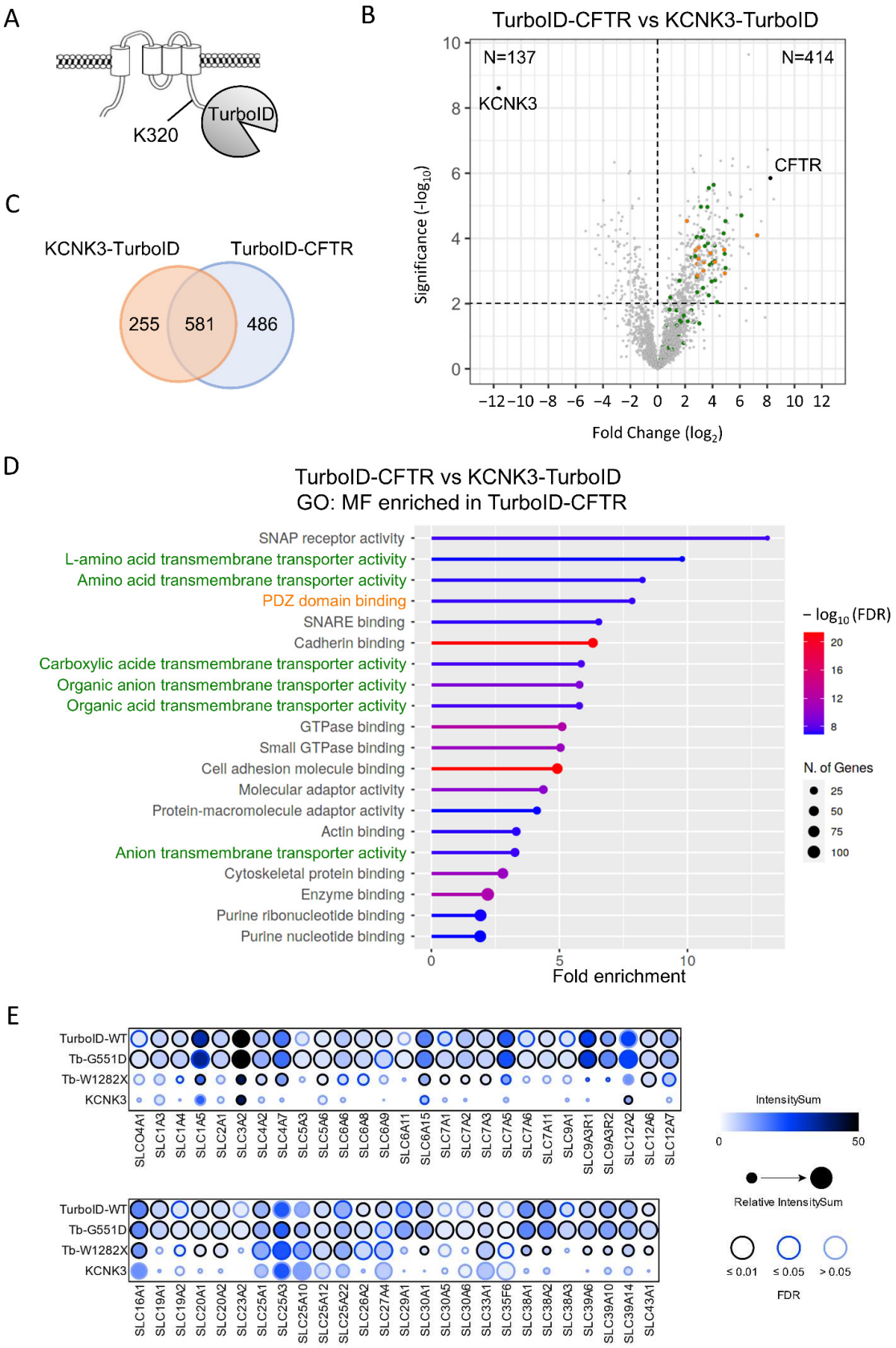


Figure 6. Comparison of TurboID-CFTR and KCNK3-TurboID proximity labeling. (**A**) Schematic representation of KCNK3-TurboID topology. (**B**) Volcano of proteins identified in TurboID-CFTR and KCNK3-TurboID conditions (n = 4 and n = 5 replicates, 3041 total proteins). Proteins enriched in the CFTR sample (Student's t-test, p-value < 0.01, n = 414) are in top right panel, and proteins enriched

in KCNK3 sample (Student's *t*-test, *p*-value < 0.1, *n* = 137) are in top left panel. Green dots indicate proteins identified as enriched in TurboID-CFTR samples associated with SLC transporters in the GO: Molecular Function (**D**) and in orange enriched in PDZ/PDZ binding domain in the Interpro domain analysis (Supplementary Figure S4B). (**C**) Venn diagram performed on TurboID-CFTR and KCNK3-TurboID partners. (**D**) GO enrichment terms identified in TurboID-CFTR Molecular Function (GO: MF). (**E**) SLC members detected in at least one condition as high confident partners (FDR < 1%) are shown as dot plots with ProHits-viz [43]. The color of each circle represents the intensity; the circle size indicates the relative value of the intensity across APEX2 and TurboID and confidence in the measurement via colored edge.

3. Discussion

Novel techniques based on proximity biotin labeling provide a snapshot of the CFTR environment with both direct binding partners and proximal non-interacting proteins.

Covalent biotin binding to specific amino acid residues can affect their post-transcriptional modification and/or their conformation. This could explain the results obtained with BioID-CFTR, where 18 h labeling led to the increase in both CFTR activity and concentration (Figure 2C). It is possible that biotinylation on specific lysine residues prevents their ubiquitination and, consequently, CFTR trafficking and degradation. CFTR ubiquitination on multiple lysine residues was reported by several teams [44,45]. Some ubiquitinated lysines were identified by mass spectrometry after TurboID assays (e.g., K536, K698, K710, K716, K793, K1250). The stabilization of both the bait and interacting partners upon biotinylation has been reported in other contexts [46]. This probably also occurs during the 10 min labeling with TurboID, possibly affecting CFTR behavior but to a lesser extent. CFTR biotinylation and/or stabilization could alter the interactome, preventing some interactions or inducing interactions that do not normally occur.

Both TurboID and APEX2 approaches identified multiple proteins associated with CFTR. In our study, APEX2 and TurboID identified a similar total number of proteins, and around 50% of the proteins were identified by the two methods. Label-free quantifications (LFQ) of proteins identified with both procedures were comparable and the CFTR protein intensity similar (Supplementary Figure S5A). Principal component analysis (PCA) of the APEX2 datasets showed experiment-dependent clustering, a feature not observed with the TurboID dataset (Supplementary Figure S5B) or when performing analysis between the CFTR mutants (Supplementary Figure S5C). Differences between the two approaches could be related to the reactivity of the Biotin-phenol, its diffusion radius or the usage of H_2O_2 (Figure 1B). It has been reported that different organelles with distinct pH, redox environments and endogenous nucleophile concentrations may influence the proximity ligation activity [34]. The subcellular distribution of the different datasets was explored using SubcellulaRVis, which calculates enrichment for 14 subcellular compartments (Supplementary Figure S6). Differences between APEX2 and TurboID showed that APEX2 had a stronger enrichment for proteins associated with the cytoplasm and the cytoskeleton, while TurboID showed higher enrichments for proteins associated with the endoplasmic reticulum and the Golgi apparatus. Few differences were observed between the two methods for plasma membrane proteins or intracellular vesicles. However, these observations cannot explain the totality of the differences between the two methods. Additionally, labeling is based on the covalent binding of biotin to specific amino acid residues, which may be more or less accessible under native conditions. Of note, the 10 min biotinylation pulse in the TurboID procedure can be prolonged to increase the number of proteins identified but can possibly affect protein synthesis or degradation and enhance non-specific background.

The CFTR interactome has been extensively studied by co-immunoprecipitation (co-IP) coupled with mass spectrometry [4]. Differences between biotin labeling approaches and co-IPs include the necessity to fuse the biotin ligase (or peroxidase) to the protein of interest. The fusion procedure can alter protein expression, folding and function, parameters that need to be evaluated. Even if tagged and untagged CFTR showed similar maturation and

function (Figure 2A), the accessibility to the N-terminal region could be affected due to steric hindrances caused by the fused protein. In CFTR, the N-terminal region was shown to be engaged in interactions with Filamin A [47], syntaxins [39,48–50] and WNK1 [23]. Only around half of the partners identified by co-IP were also identified with TurboID or APEX2 approaches. A major difference between the two approaches is the labeling of both transient partners and non-interacting but proximal proteins, which might not be co-immunoprecipitated. Of note, around half of the identified proteins were unique to each method (547 out of 1095 for APEX2, 446 out of 941 for TurboID and 466 out of 729 for coIP). In the same line, the recently developed MaMTH-HTS method, which used full-length CFTR-WT as a bait to screen a library of around 10,000 ORF [31], only marginally overlapped with proximity labeling or co-IP. Another important issue is the cell type used. In this study, transiently transfected HEK293 cells were used to achieve high expression levels. It has to be kept in mind that overexpression can affect CFTR interactome and that partners specific to lung epithelial cells or pancreatic duct cells where CFTR is endogenously expressed may be absent in HEK293 cells. It appears that combining different approaches and cell types is necessary to obtain a full picture of CFTR interactome, to feed the databases and provide a better understanding of the CFTR environment.

Compared to co-IP (HTS dataset), proximity labeling (especially with TurboID) identified a greater proportion of transmembrane proteins, such as transporters, and more specifically, SLC transporters (Supplementary Figure S1D,E). This probably reflects both the difficulty in preserving membrane protein complexes using detergents in the co-IP procedure and the labeling of non-interacting but proximal proteins, which might not be co-immunoprecipitated. Functional co-regulation between CFTR and members of the SLC26 subfamily was found to involve a direct protein–protein interaction between the STAS domain of SLC26 transporters and the R domain of CFTR [27,29]. In this "special issue", CFTR biogenesis and stability were shown to be affected by SLC26A9 expression levels [26]. As many of the SLC transporters identified (Figure 6E) do not contain a STAS domain, it can be speculated that other domains could be involved, which still need to be identified. Alternatively, proximity could be driven by localization in particular subcellular compartments or the sharing of common pathways during protein biogenesis. While structurally unrelated, it is plausible that these large transmembrane transporters require specific protein complexes for their proper folding and expression at the cell surface. Finally, both CFTR [51,52] and some SLCs [53,54] have been identified within and outside sphingolipid- and cholesterol-rich lipid nanodomains (or lipid rafts), raising the possibility of defining specific regions where these transporters are in close proximity. Taken together, the close proximity between CFTR and SLC transporters should be further explored.

Finally, the comparisons between WT and mutant CFTR showed that multiple interacting partners were nevertheless preserved. CFTR-G551D showed very few differences with CFTR-WT, consistent with a global preservation of the channel structure and localization. As reported in a previous study [55], some GO enrichments were identified, suggesting a higher affinity of G551D for the actin network compared to CFTR WT. Another report showed multiple differences between WT and G551D [7] not found here. These discrepancies will need further studies and could be linked, in part, to different cell types used, CFBE41o- [7] versus HEK293 in our study, or HeLa cells [55]. CFTR-W1282X, which lacks part of NBD2 and the C-terminus of the protein, showed more differences, including, as expected, the PDZ binding proteins and proteins associated with protein misfolding. Protein misfolding is consistent with the stabilizing effect of the corrector VX-445 observed on CFTR-W1282X upon inhibition of nonsense-mediated decay [56].

The importance of the identified proximal proteins to CFTR biogenesis and function needs to be functionally evaluated, as their labeling could only reflect common cellular processes and localization within the same cellular subdomain. When comparing CFTR and KCNK3 interactomes, many proteins were identified in both sets, revealing common biogenesis pathways and localization. Nonetheless, clear differences were observed for sets of proteins that were enriched in CFTR samples, such as PDZ containing proteins (KCNK3

lacks a PDZ binding domain) and SLC transporters. For the latter, CFTR was found to be in the vicinity of multiple SLC transporters, some of which have been found to be functionally co-regulated with CFTR [27,29] or influence CFTR biogenesis [26]. This result indicates that while undergoing the same biogenesis pathways and localizing in the same compartments, these two channels are associated preferentially with distinct protein sets. The comparison of interactomes from specific channels or channel families could enable identifying signatures associated with their localization, function or regulatory pathways.

In conclusion, our study provides evidence that the various approaches developed for interactomic studies can each identify unique proteins and therefore should be combined to obtain a more complete picture of the CFTR interactome.

4. Materials and Methods

4.1. Plasmid Constructs

CFTR fusion constructs were obtained by PCR assembly (NEBuilder HiFi DNA Assembly, NEB, Évry-Courcouronnes, France) in the expression vector pLenti-III digested with NheI and XbaI. The assembly was performed for each construct to the digested pLIII plasmid with 3 fragments: a fragment obtained by PCR amplification (Q5® High-Fidelity, NEB, Évry-Courcouronnes, France) of proximity labeling enzymes (APEX2, BioID or TurboID), a synthetic fragment corresponding to the GS5 linker and the N-terminal part of CFTR (Eurofins Genomics, Les Ulis, France), a second fragment obtained by PCR amplification of CFTR-WT or its mutants (amplification of the BlpI site up to the CFTR stop codon). The template BioID plasmid was obtained from Morgan Gallazzini, (Institut Necker Enfants Malades Paris France), APEX2 from Jacques Camonis, (Institut Curie Paris France); V5-TurboID-NES_pcDNA3 and C1(1-29)-TurboID-V5_pCDNA3 were a gift from Alice Ting (Addgene plasmid #107173 and #107173).

All plasmids obtained were entirely sequenced (Eurofins Genomic, Les Ulis, France). KCNK3 were obtained by PCR assembly using as template C1(1-29)-TurboID-V5_pCDNA3.

4.2. Cell Culture and Transfection

HEK293 cells were purchased from ATCC and cultivated in DMEM medium supplemented with 10% fetal calf serum (Thermo Fisher Scientific, Illkirch-Graffenstaden, France). Cells were maintained at 37 °C, 5% CO_2. For the functional assay, cells were co-transfected with equal amount of halide-sensitive YFP and CFTR plasmids using Turbofect (Thermo Fisher Scientific, Illkirch-Graffenstaden, France). For Western blot analysis and mass spectrometry analysis, cells were transfected with CFTR plasmids using Lipofectamine 3000, following instructions (Thermo Fisher Scientific, Illkirch-Graffenstaden, France).

4.3. Western Blot Analysis

The transfected cells were lysed in RIPA buffer containing protease inhibitors (Roche Life Science, Basel, Switzerland), and protein concentration was assessed using RcDc assay (BioRad, Marnes-la-Coquette, France). Western blot analysis was performed using 60 µg of protein from each sample separated on a 7% acrylamide gel. After transfer onto nitrocellulose membranes, CFTR was probed using antibody 660 (Cystic Fibrosis Foundation, Chapel Hill, NC, USA), and equal loading was assessed using anti-tubulin (SantaCruz, Dallas, TX, USA).

4.4. Halide-Sensitive Functional Assay

CFTR activity was measured in transiently transfected HEK293 cells using the halide-sensitive yellow fluorescent protein YFP-H148Q/I152L [57]. The day after transfection, cells were transferred to poly-L-lysine-coated 96-well black/clear bottom microplates. After 24 h, plates were washed with PBS, and each well was incubated for 30 min with 100 µL of PBS containing cpt-AMPc (100 µM) and IBMX (100 µM) (Sigma-Aldrich, Saint-Quentin-Fallavier, France). Plates were then transferred to a ClarioStar plate reader (BMG Labtech, Ortenberg, Germany) equipped with an injector, which enabled the continuous recording

of fluorescence during the injection. After 5 s, 200 µL of PBS-NaI (PBS solution where NaCl is replaced with NaI) was injected.

4.5. Proximity Labeling

HEK293 cells were transiently transfected for 48 h in 10 cm diameter poly-L-lysine-coated dishes. Biotinylation was induced by adding biotin (Thermo Fisher Scientific, Illkirch-Graffenstaden, France) to the medium at 37 °C for 10 min (TurboID, 500 µM biotin) or 18 h (BioID, 50 µM biotin). For APEX2, the cells were pre-incubated for 30 min at 37 °C with Biotin-Phenol (500 µM, Iris Biotech, Marktredwitz, Germany), and the peroxidase activity was activated by the addition of H2O2 at a final concentration of 1 mM. The reaction was then quenched by the addition of quenching buffer (10 mM sodium azide, 10 mM sodium ascorbate and 5 mM Trolox, Sigma-Aldrich, Saint-Quentin-Fallavier, France). Cells were washed several times with PBS+ at 4 °C before being harvested and centrifuged. Lysis and streptavidin pull-down steps were performed, as previously described by Hung et al. [35].

4.6. NanoLC-MS/MS Protein Identification and Quantification

S-TrapTM micro spin column (Protifi, Farmingdale, NY, USA) digestion was performed on streptavidin eluates in 4× Laemmli buffer according to the manufacturer's protocol but with 2 extra washing steps for thorough SDS elimination. Samples were digested with 2 µg of trypsin (Promega, Charbonnières-les-Bains, France) at 47 °C for 1 h 30 min. After elution, peptides were finally vacuum dried down and resuspended in 35 µL of 10% ACN and 0.1% TFA in HPLC-grade water prior to MS analysis. For each run, 5 µL was injected in a nanoRSLC-Q Exactive PLUS (RSLC Ultimate 3000) (Thermo Scientific, Illkirch-Graffenstaden, France). Peptides were loaded onto a µ-precolumn (Acclaim PepMap 100 C18, cartridge, 300 µm i.d. × 5 mm, 5 µm) (Thermo Scientific, Illkirch-Graffenstaden, France) and were separated on a 50 cm reversed-phase liquid chromatographic column (0.075 mm ID, Acclaim PepMap 100, C18, 2 µm) (Thermo Scientific, Illkirch-Graffenstaden, France). The chromatography solvents were (A) 0.1% formic acid in water and (B) 80% acetonitrile, 0.08% formic acid. Peptides were eluted from the column with the following gradients: 5% to 40% B (120 min), 40% to 80% (1 min). At 121 min, the gradient stayed at 80% for 5 min, and at 127 min, it returned to 5% to re-equilibrate the column for 20 min before the next injection. One blank was run between each series to prevent sample carryover. Peptides eluting from the column were analyzed by data-dependent MS/MS, using the top-10 acquisition method. Peptides were fragmented using higher-energy collisional dissociation (HCD). Briefly, the instrument settings were as follows: the resolution was set to 70,000 for MS scans and 17,500 for the data-dependent MS/MS scans in order to increase speed. The MS AGC target was set to 3.106 counts with a maximum injection time set to 200 ms, while the MS/MS AGC target was set to 1.105 with a maximum injection time set to 120 ms. The MS scan range was from 400 to 2000 m/z.

4.7. Data Processing Following LC-MS/MS Acquisition

The MS files were processed with the MaxQuant software version 2.0.1.0 and searched with the Andromeda search engine against the database of Homo sapiens from Swiss-Prot 04/2020. To search for parent mass and fragment ions, we set an initial mass deviation of 4.5 ppm and 20 ppm, respectively. The minimum peptide length was set to 7 amino acids, and strict specificity for trypsin cleavage was required, allowing up to two missed cleavage sites. Carbamidomethylation (Cys) was set as fixed modification, whereas oxidation (Met) and N-term acetylation were set as variable modifications. For APEX2, biotinylation (H23C18N3O3S) was set as variable modification on any tyrosine, tryptophane and histidine, and for TurboID, biotinylation (H14C10N2O2S) was set as variable modification on any lysine. A match between the runs was allowed. LFQ minimum ratio count was set to 2. The false discovery rates (FDRs) at the protein and peptide levels were set to 1%. Scores were calculated in MaxQuant, as described previously [58]. The reverse and

common contaminants hits were removed from the MaxQuant output. Proteins were quantified according to the MaxQuant label-free algorithm using LFQ intensities [58,59]. Fragmentation visualization spectra were also extracted using the MQviewer integrated in the Maxquant software.

4.8. Data Processing and Statistical Analysis

Three to five independent experiments of HEK293 cells transfected with untagged CFTR, APEX2-CFTR-WT, TurboID-CFTR-WT, TurboID-CFTR-G551D, TurboID-CFTR-W1282X and KCNK3-TurboID were analyzed with Perseus software (version 1.6.15.0) freely available at www.perseus-framework.org [59]. The label-free quantification (LFQ) data were transformed in log2, and the Significance Analysis of INTeractome (SAINT [37]; https://reprint-apms.org/) was used for the identification of the proximal partners of CFTR on the raw MS files. Comparisons between CFTR mutants and the WT condition were performed with R software (version 4.1.0) based on the label-free quantification (LFQ) log2-transformed data. All proteins identified in all replicates of all conditions were subjected to Student's *t*-test without correction for multiple testing. Where applicable, LogFC were shown as means, and *p*-values of less than 0.01 or 0.1 were considered. For CFTR comparison with KCNK3, the *p*-value was set at <0.01, since large differences were expected. For mutant comparison, since the differences were smaller, we lowered the stringency to a *p*-value of <0.1. The gene ontology enrichment calculations and lollipop graphs were generated with the ShinyGO v0.741 tool (http://bioinformatics.sdstate.edu/go74/ (accessed on February 2022)) [60].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms23168937/s1.

Author Contributions: Conceptualization, B.C., A.E. and A.H.; Data curation, B.C., M.N., V.J., F.A. and I.C.G.; Formal analysis, B.C., M.N., V.J., I.P., I.C.G. and A.H.; Funding acquisition, V.S., S.M., I.S.-G. and A.H.; Investigation, B.C., N.B., S.C., V.J., I.P., A.G. and F.A.; Methodology, B.C., N.B., S.C., V.J., A.G., I.C.G. and A.H.; Project administration, A.H.; Software, M.N.; Supervision, V.S., I.C.G., I.S.-G., A.E. and A.H.; Validation, B.C., F.A. and A.H.; Visualization, B.C., I.C.G. and A.H.; Writing—original draft, B.C. and A.H.; Writing—Review and editing, B.C., M.N., I.P., V.S., S.M., F.A., I.C.G., I.S.-G., A.E. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by institutional grants from the INSERM, the CNRS and the Université de Paris, by the ANR ANR-18-CE14-0004-02 grant to S.M. and A.H., the "Vaincre la Mucoviscidose" RF20190502488 to M.N., RF20180502264 and RF20210502867 to A.H., Fondation Dassault Systèmes to B.C., the "Association pour l'Aide à la Recherche contre la Mucoviscidose (AARM)" and the "Mucoviscidose: ABCF2" to I.S.-G.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The mass spectrometry proteomics data were deposited to the ProteomeXchange Consortium via the PRIDE [61] partner repository with the dataset identifier PXD035184, and the protein interactions were submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct [62] and assigned the identifier IM-29540.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Riordan, J.R.; Rommens, J.M.; Kerem, B.; Alon, N.; Rozmahel, R.; Grzelczak, Z.; Zielenski, J.; Lok, S.; Plavsic, N.; Chou, J.L. Identification of the Cystic Fibrosis Gene: Cloning and Characterization of Complementary DNA. *Science* **1989**, *245*, 1066–1073. [CrossRef]
2. Kreda, S.M.; Mall, M.; Mengos, A.; Rochelle, L.; Yankaskas, J.; Riordan, J.R.; Boucher, R.C. Characterization of Wild-Type and DeltaF508 Cystic Fibrosis Transmembrane Regulator in Human Respiratory Epithelia. *Mol. Biol. Cell* **2005**, *16*, 2154–2167. [CrossRef] [PubMed]

3. Farinha, C.M.; Gentzsch, M. Revisiting CFTR Interactions: Old Partners and New Players. *Int. J. Mol. Sci.* **2021**, *22*, 13196. [CrossRef] [PubMed]
4. Pankow, S.; Bamberger, C.; Calzolari, D.; Martínez-Bartolomé, S.; Lavallée-Adam, M.; Balch, W.E.; Yates, J.R. ΔF508 CFTR Interactome Remodelling Promotes Rescue of Cystic Fibrosis. *Nature* **2015**, *528*, 510–516. [CrossRef]
5. Davezac, N.; Tondelier, D.; Lipecka, J.; Fanen, P.; Demaugre, F.; Debski, J.; Dadlez, M.; Schrattenholz, A.; Cahill, M.A.; Edelman, A. Global Proteomic Approach Unmasks Involvement of Keratins 8 and 18 in the Delivery of Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)/DeltaF508-CFTR to the Plasma Membrane. *Proteomics* **2004**, *4*, 3833–3844. [CrossRef] [PubMed]
6. Ramalho, S.S.; Silva, I.A.L.; Amaral, M.D.; Farinha, C.M. Rare Trafficking CFTR Mutations Involve Distinct Cellular Retention Machineries and Require Different Rescuing Strategies. *Int. J. Mol. Sci.* **2021**, *23*, 24. [CrossRef] [PubMed]
7. Hutt, D.M.; Loguercio, S.; Campos, A.R.; Balch, W.E. A Proteomic Variant Approach (ProVarA) for Personalized Medicine of Inherited and Somatic Disease. *J. Mol. Biol.* **2018**, *430*, 2951–2973. [CrossRef]
8. Estabrooks, S.; Brodsky, J.L. Regulation of CFTR Biogenesis by the Proteostatic Network and Pharmacological Modulators. *Int. J. Mol. Sci.* **2020**, *21*, 452. [CrossRef]
9. Farinha, C.M.; Canato, S. From the Endoplasmic Reticulum to the Plasma Membrane: Mechanisms of CFTR Folding and Trafficking. *Cell. Mol. Life Sci.* **2017**, *74*, 39–55. [CrossRef]
10. Wang, X.; Matteson, J.; An, Y.; Moyer, B.; Yoo, J.-S.; Bannykh, S.; Wilson, I.A.; Riordan, J.R.; Balch, W.E. COPII-Dependent Export of Cystic Fibrosis Transmembrane Conductance Regulator from the ER Uses a Di-Acidic Exit Code. *J. Cell Biol.* **2004**, *167*, 65–74. [CrossRef]
11. Roxo-Rosa, M.; Xu, Z.; Schmidt, A.; Neto, M.; Cai, Z.; Soares, C.M.; Sheppard, D.N.; Amaral, M.D. Revertant Mutants G550E and 4RK Rescue Cystic Fibrosis Mutants in the First Nucleotide-Binding Domain of CFTR by Different Mechanisms. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 17891–17896. [CrossRef] [PubMed]
12. Hegedus, T.; Aleksandrov, A.; Cui, L.; Gentzsch, M.; Chang, X.-B.; Riordan, J.R. F508del CFTR with Two Altered RXR Motifs Escapes from ER Quality Control but Its Channel Activity Is Thermally Sensitive. *Biochim. Biophys. Acta* **2006**, *1758*, 565–572. [CrossRef] [PubMed]
13. Fukuda, R.; Okiyoneda, T. Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Ubiquitylation as a Novel Pharmaceutical Target for Cystic Fibrosis. *Pharmaceuticals* **2020**, *13*, 75. [CrossRef]
14. Okiyoneda, T.; Barrière, H.; Bagdány, M.; Rabeh, W.M.; Du, K.; Höhfeld, J.; Young, J.C.; Lukacs, G.L. Peripheral Protein Quality Control Removes Unfolded CFTR from the Plasma Membrane. *Science* **2010**, *329*, 805–810. [CrossRef]
15. Apaja, P.M.; Xu, H.; Lukacs, G.L. Quality Control for Unfolded Proteins at the Plasma Membrane. *J. Cell Biol.* **2010**, *191*, 553–570. [CrossRef] [PubMed]
16. Csanády, L.; Vergani, P.; Gadsby, D.C. Structure, Gating, and Regulation of the CFTR Anion Channel. *Physiol. Rev.* **2019**, *99*, 707–738. [CrossRef]
17. Della Sala, A.; Prono, G.; Hirsch, E.; Ghigo, A. Role of Protein Kinase A-Mediated Phosphorylation in CFTR Channel Activity Regulation. *Front. Physiol.* **2021**, *12*, 690247. [CrossRef]
18. Chin, S.; Hung, M.; Bear, C.E. Current Insights into the Role of PKA Phosphorylation in CFTR Channel Activity and the Pharmacological Rescue of Cystic Fibrosis Disease-Causing Mutants. *Cell. Mol. Life Sci.* **2017**, *74*, 57–66. [CrossRef]
19. Seavilleklein, G.; Amer, N.; Evagelidis, A.; Chappe, F.; Irvine, T.; Hanrahan, J.W.; Chappe, V. PKC Phosphorylation Modulates PKA-Dependent Binding of the R Domain to Other Domains of CFTR. *Am. J. Physiol. Cell Physiol.* **2008**, *295*, C1366–C1375. [CrossRef]
20. Chappe, V.; Hinkson, D.A.; Howell, L.D.; Evagelidis, A.; Liao, J.; Chang, X.-B.; Riordan, J.R.; Hanrahan, J.W. Stimulatory and Inhibitory Protein Kinase C Consensus Sequences Regulate the Cystic Fibrosis Transmembrane Conductance Regulator. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 390–395. [CrossRef]
21. Billet, A.; Luo, Y.; Balghi, H.; Hanrahan, J.W. Role of Tyrosine Phosphorylation in the Muscarinic Activation of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR). *J. Biol. Chem.* **2013**, *288*, 21815–21823. [CrossRef] [PubMed]
22. Mihályi, C.; Iordanov, I.; Töröcsik, B.; Csanády, L. Simple Binding of Protein Kinase A Prior to Phosphorylation Allows CFTR Anion Channels to Be Opened by Nucleotides. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 21740–21746. [CrossRef] [PubMed]
23. Kim, Y.; Jun, I.; Shin, D.H.; Yoon, J.G.; Piao, H.; Jung, J.; Park, H.W.; Cheng, M.H.; Bahar, I.; Whitcomb, D.C.; et al. Regulation of CFTR Bicarbonate Channel Activity by WNK1: Implications for Pancreatitis and CFTR-Related Disorders. *Cell. Mol. Gastroenterol. Hepatol.* **2020**, *9*, 79–103. [CrossRef] [PubMed]
24. Reddy, M.M.; Quinton, P.M. Functional Interaction of CFTR and ENaC in Sweat Glands. *Pflug. Arch.* **2003**, *445*, 499–503. [CrossRef]
25. Schwiebert, E.M.; Benos, D.J.; Egan, M.E.; Stutts, M.J.; Guggino, W.B. CFTR Is a Conductance Regulator as Well as a Chloride Channel. *Physiol. Rev.* **1999**, *79*, S145–S166. [CrossRef]
26. Pinto, M.C.; Quaresma, M.C.; Silva, I.A.L.; Railean, V.; Ramalho, S.S.; Amaral, M.D. Synergy in Cystic Fibrosis Therapies: Targeting SLC26A9. *Int. J. Mol. Sci.* **2021**, *22*, 13064. [CrossRef]
27. Bakouh, N.; Bienvenu, T.; Thomas, A.; Ehrenfeld, J.; Liote, H.; Roussel, D.; Duquesnoy, P.; Farman, N.; Viel, M.; Cherif-Zahar, B.; et al. Characterization of SLC26A9 in Patients with CF-like Lung Disease. *Hum. Mutat.* **2013**, *34*, 1404–1414. [CrossRef]

28. Bertrand, C.A.; Mitra, S.; Mishra, S.K.; Wang, X.; Zhao, Y.; Pilewski, J.M.; Madden, D.R.; Frizzell, R.A. The CFTR Trafficking Mutation F508del Inhibits the Constitutive Activity of SLC26A9. *Am. J. Physiol. Lung Cell. Mol. Physiol.* **2017**, *312*, L912–L925. [CrossRef]
29. Ko, S.B.H.; Zeng, W.; Dorwart, M.R.; Luo, X.; Kim, K.H.; Millen, L.; Goto, H.; Naruse, S.; Soyombo, A.; Thomas, P.J.; et al. Gating of CFTR by the STAS Domain of SLC26 Transporters. *Nat. Cell Biol.* **2004**, *6*, 343–350. [CrossRef]
30. Wang, Y.; Soyombo, A.A.; Shcheynikov, N.; Zeng, W.; Dorwart, M.; Marino, C.R.; Thomas, P.J.; Muallem, S. Slc26a6 Regulates CFTR Activity in Vivo to Determine Pancreatic Duct HCO3− Secretion: Relevance to Cystic Fibrosis. *EMBO J.* **2006**, *25*, 5049–5057. [CrossRef]
31. Lim, S.H.; Snider, J.; Birimberg-Schwartz, L.; Ip, W.; Serralha, J.C.; Botelho, H.M.; Lopes-Pacheco, M.; Pinto, M.C.; Moutaoufik, M.T.; Zilocchi, M.; et al. CFTR Interactome Mapping Using the Mammalian Membrane Two-Hybrid High-Throughput Screening System. *Mol. Syst. Biol.* **2022**, *18*, e10629. [CrossRef] [PubMed]
32. Roux, K.J.; Kim, D.I.; Raida, M.; Burke, B. A Promiscuous Biotin Ligase Fusion Protein Identifies Proximal and Interacting Proteins in Mammalian Cells. *J. Cell Biol.* **2012**, *196*, 801–810. [CrossRef] [PubMed]
33. Hung, V.; Zou, P.; Rhee, H.-W.; Udeshi, N.D.; Cracan, V.; Svinkina, T.; Carr, S.A.; Mootha, V.K.; Ting, A.Y. Proteomic Mapping of the Human Mitochondrial Intermembrane Space in Live Cells via Ratiometric APEX Tagging. *Mol. Cell* **2014**, *55*, 332–341. [CrossRef] [PubMed]
34. Branon, T.C.; Bosch, J.A.; Sanchez, A.D.; Udeshi, N.D.; Svinkina, T.; Carr, S.A.; Feldman, J.L.; Perrimon, N.; Ting, A.Y. Efficient Proximity Labeling in Living Cells and Organisms with TurboID. *Nat. Biotechnol.* **2018**, *36*, 880–887. [CrossRef] [PubMed]
35. Hung, V.; Udeshi, N.D.; Lam, S.S.; Loh, K.H.; Cox, K.J.; Pedram, K.; Carr, S.A.; Ting, A.Y. Spatially Resolved Proteomic Mapping in Living Cells with the Engineered Peroxidase APEX2. *Nat. Protoc.* **2016**, *11*, 456–475. [CrossRef]
36. Moyer, B.D.; Loffing, J.; Schwiebert, E.M.; Loffing-Cueni, D.; Halpin, P.A.; Karlson, K.H.; Ismailov, I.I.; Guggino, W.B.; Langford, G.M.; Stanton, B.A. Membrane Trafficking of the Cystic Fibrosis Gene Product, Cystic Fibrosis Transmembrane Conductance Regulator, Tagged with Green Fluorescent Protein in Madin-Darby Canine Kidney Cells. *J. Biol. Chem.* **1998**, *273*, 21759–21768. [CrossRef]
37. Choi, H.; Larsen, B.; Lin, Z.-Y.; Breitkreutz, A.; Mellacheruvu, D.; Fermin, D.; Qin, Z.S.; Tyers, M.; Gingras, A.-C.; Nesvizhskii, A.I. SAINT: Probabilistic Scoring of Affinity Purification-Mass Spectrometry Data. *Nat. Methods* **2011**, *8*, 70–73. [CrossRef]
38. Wang, X.; Venable, J.; LaPointe, P.; Hutt, D.M.; Koulov, A.V.; Coppinger, J.; Gurkan, C.; Kellner, W.; Matteson, J.; Plutner, H.; et al. Hsp90 Cochaperone Aha1 Downregulation Rescues Misfolding of CFTR in Cystic Fibrosis. *Cell* **2006**, *127*, 803–815. [CrossRef]
39. Tang, B.L.; Gee, H.Y.; Lee, M.G. The Cystic Fibrosis Transmembrane Conductance Regulator's Expanding SNARE Interactome. *Traffic* **2011**, *12*, 364–371. [CrossRef]
40. Zhao, L.; Yuan, F.; Pan, N.; Yu, Y.; Yang, H.; Liu, Y.; Wang, R.; Zhang, B.; Wang, G. CFTR Deficiency Aggravates Ang II Induced Vasoconstriction and Hypertension by Regulating Ca2+ Influx and RhoA/Rock Pathway in VSMCs. *Front. Biosci.* **2021**, *26*, 1396–1410. [CrossRef]
41. Huang, W.; Tan, M.; Wang, Y.; Liu, L.; Pan, Y.; Li, J.; Ouyang, M.; Long, C.; Qu, X.; Liu, H.; et al. Increased Intracellular Cl- Concentration Improves Airway Epithelial Migration by Activating the RhoA/ROCK Pathway. *Theranostics* **2020**, *10*, 8528–8540. [CrossRef] [PubMed]
42. Castellani, S.; Guerra, L.; Favia, M.; Di Gioia, S.; Casavola, V.; Conese, M. NHERF1 and CFTR Restore Tight Junction Organisation and Function in Cystic Fibrosis Airway Epithelial Cells: Role of Ezrin and the RhoA/ROCK Pathway. *Lab. Investig.* **2012**, *92*, 1527–1540. [CrossRef] [PubMed]
43. Knight, J.D.R.; Choi, H.; Gupta, G.D.; Pelletier, L.; Raught, B.; Nesvizhskii, A.I.; Gingras, A.-C. ProHits-Viz: A Suite of Web Tools for Visualizing Interaction Proteomics Data. *Nat. Methods* **2017**, *14*, 645–646. [CrossRef] [PubMed]
44. Pankow, S.; Bamberger, C.; Yates, J.R. A Posttranslational Modification Code for CFTR Maturation Is Altered in Cystic Fibrosis. *Sci. Signal* **2019**, *12*, eaan7984. [CrossRef]
45. Lee, S.; Henderson, M.J.; Schiffhauer, E.; Despanie, J.; Henry, K.; Kang, P.W.; Walker, D.; McClure, M.L.; Wilson, L.; Sorscher, E.J.; et al. Interference with Ubiquitination in CFTR Modifies Stability of Core Glycosylated and Cell Surface Pools. *Mol. Cell. Biol.* **2014**, *34*, 2554–2565. [CrossRef]
46. Freitas, F.C.; Maldonado, M.; Oliveira Junior, A.B.; Onuchic, J.N.; de Oliveira, R.J. Biotin-Painted Proteins Have Thermodynamic Stability Switched by Kinetic Folding Routes. *J. Chem. Phys.* **2022**, *156*, 195101. [CrossRef]
47. Thelin, W.R.; Chen, Y.; Gentzsch, M.; Kreda, S.M.; Sallee, J.L.; Scarlett, C.O.; Borchers, C.H.; Jacobson, K.; Stutts, M.J.; Milgram, S.L. Direct Interaction with Filamins Modulates the Stability and Plasma Membrane Expression of CFTR. *J. Clin. Investig.* **2007**, *117*, 364–374. [CrossRef]
48. Cormet-Boyaka, E.; Di, A.; Chang, S.Y.; Naren, A.P.; Tousson, A.; Nelson, D.J.; Kirk, K.L. CFTR Chloride Channels Are Regulated by a SNAP-23/Syntaxin 1A Complex. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 12477–12482. [CrossRef]
49. Sabirzhanova, I.; Boinot, C.; Guggino, W.B.; Cebotaru, L. Syntaxin 8 and the Endoplasmic Reticulum Processing of ΔF508-CFTR. *Cell. Physiol. Biochem.* **2018**, *51*, 1489–1499. [CrossRef]
50. Arora, K.; Liyanage, P.; Zhong, Q.; Naren, A.P. A SNARE Protein Syntaxin 17 Captures CFTR to Potentiate Autophagosomal Clearance under Stress. *FASEB J.* **2021**, *35*, e21185. [CrossRef]
51. Abu-Arish, A.; Pandžić, E.; Luo, Y.; Sato, Y.; Turner, M.J.; Wiseman, P.W.; Hanrahan, J.W. Lipid-Driven CFTR Clustering Is Impaired in Cystic Fibrosis and Restored by Corrector Drugs. *J. Cell Sci.* **2022**, *135*, jcs259002. [CrossRef] [PubMed]

52. Dudez, T.; Borot, F.; Huang, S.; Kwak, B.R.; Bacchetta, M.; Ollero, M.; Stanton, B.A.; Chanson, M. CFTR in a Lipid Raft-TNFR1 Complex Modulates Gap Junctional Intercellular Communication and IL-8 Secretion. *Biochim. Biophys. Acta* **2008**, *1783*, 779–788. [CrossRef] [PubMed]
53. Hilgemann, D.W.; Fine, M.; Linder, M.E.; Jennings, B.C.; Lin, M.-J. Massive Endocytosis Triggered by Surface Membrane Palmitoylation under Mitochondrial Control in BHK Fibroblasts. *eLife* **2013**, *2*, e01293. [CrossRef] [PubMed]
54. Reilly, L.; Howie, J.; Wypijewski, K.; Ashford, M.L.J.; Hilgemann, D.W.; Fuller, W. Palmitoylation of the Na/Ca Exchanger Cytoplasmic Loop Controls Its Inactivation and Internalization during Stress Signaling. *FASEB J.* **2015**, *29*, 4532–4543. [CrossRef] [PubMed]
55. Trouvé, P.; Kerbiriou, M.; Teng, L.; Benz, N.; Taiya, M.; Le Hir, S.; Férec, C. G551D-CFTR Needs More Bound Actin than Wild-Type CFTR to Maintain Its Presence in Plasma Membranes. *Cell Biol. Int.* **2015**, *39*, 978–985. [CrossRef]
56. Venturini, A.; Borrelli, A.; Musante, I.; Scudieri, P.; Capurro, V.; Renda, M.; Pedemonte, N.; Galietta, L.J.V. Comprehensive Analysis of Combinatorial Pharmacological Treatments to Correct Nonsense Mutations in the CFTR Gene. *Int. J. Mol. Sci.* **2021**, *22*, 11972. [CrossRef]
57. Galietta, L.J.; Haggie, P.M.; Verkman, A.S. Green Fluorescent Protein-Based Halide Indicators with Improved Chloride and Iodide Affinities. *FEBS Lett.* **2001**, *499*, 220–224. [CrossRef]
58. Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372. [CrossRef]
59. Tyanova, S.; Temu, T.; Sinitcyn, P.; Carlson, A.; Hein, M.Y.; Geiger, T.; Mann, M.; Cox, J. The Perseus Computational Platform for Comprehensive Analysis of (Prote)Omics Data. *Nat. Methods* **2016**, *13*, 731–740. [CrossRef]
60. Ge, S.X.; Jung, D.; Yao, R. ShinyGO: A Graphical Gene-Set Enrichment Tool for Animals and Plants. *Bioinformatics* **2020**, *36*, 2628–2629. [CrossRef]
61. Perez-Riverol, Y.; Bai, J.; Bandla, C.; García-Seisdedos, D.; Hewapathirana, S.; Kamatchinathan, S.; Kundu, D.J.; Prakash, A.; Frericks-Zipper, A.; Eisenacher, M.; et al. The PRIDE Database Resources in 2022: A Hub for Mass Spectrometry-Based Proteomics Evidences. *Nucleic Acids Res.* **2022**, *50*, D543–D552. [CrossRef] [PubMed]
62. Orchard, S.; Ammari, M.; Aranda, B.; Breuza, L.; Briganti, L.; Broackes-Carter, F.; Campbell, N.H.; Chavali, G.; Chen, C.; del-Toro, N.; et al. The MIntAct Project–IntAct as a Common Curation Platform for 11 Molecular Interaction Databases. *Nucleic Acids Res.* **2014**, *42*, D358–D363. [CrossRef] [PubMed]

Appendix C

Predictions of CFTR modulators targets

Ivacaftor

Table C.1 – Top 20 predictions of proteins interacting with ivacaftor (VX-770)

Rank	UniProt ID	Gene Name	Name	Prediction Score
1	Q16539	MAPK14	Mitogen-activated protein kinase 14	0.909813527
2	Q15759	MAPK11	Mitogen-activated protein kinase 11	0.892286204
3	O15264	MAPK13	Mitogen-activated protein kinase 13	0.847776496
4	P53778	MAPK12	Mitogen-activated protein kinase 12	0.825312485
5	O14757	CHEK1	Serine/threonine-protein kinase Chk1	0.821140137
6	P06493	CDK1	Cyclin-dependent kinase 1	0.821038156
7	P45984	MAPK9	Mitogen-activated protein kinase 9	0.810185387
8	P11362	FGFR1	Fibroblast growth factor receptor 1	0.804103458
9	P53779	MAPK10	Mitogen-activated protein kinase 10	0.798419301
10	P28482	MAPK1	Mitogen-activated protein kinase 1	0.793260891
11	O14965	AURKA	Aurora kinase A	0.789339875
12	Q9UQB9	AURKC	Aurora kinase C	0.78480879
13	Q14012	CAMK1	Calcium/calmodulin-dependent protein kinase type 1	0.776144628
14	P45983	MAPK8	Mitogen-activated protein kinase 8	0.768880996
15	Q59EB3	METPO	Met proto-oncogene variant	0.766007475
16	Q9Y6E0	STK24	Serine/threonine-protein kinase 24	0.761798072
17	P35916	FLT4	Vascular endothelial growth factor receptor 3	0.760975174
18	Q8IU85	CAMK1D	Calcium/calmodulin-dependent protein kinase type 1D	0.755742446
19	Q9P289	STK26	Serine/threonine-protein kinase 26	0.754947055
20	P21802	FGFR2	Fibroblast growth factor receptor 2	0.753457753

Lumacaftor

Table C.2 – Top 20 predictions of proteins interacting with lumacaftor (VX-809)

Rank	UniProt ID	Gene Name	Name	Prediction score
1	P35968	KDR	Vascular endothelial growth factor receptor 2	0.8819896
2	Q16539	MAPK14	Mitogen-activated protein kinase 14	0.86577077
3	Q15759	MAPK11	Mitogen-activated protein kinase 11	0.85832436
4	P35916	FLT4	Vascular endothelial growth factor receptor 3	0.85433318
5	P10721	KIT	Mast/stem cell growth factor receptor Kit	0.85395575
6	Q59EB3	METPO	Met proto-oncogene variant	0.85044356
7	P28702	RXRB	Retinoic acid receptor RXR-beta	0.8406796
8	Q02127	DHODH	Dihydroorotate dehydrogenase (quinone), mitochondrial	0.83880829
9	P37231	PPARG	Peroxisome proliferator-activated receptor gamma	0.83142857
10	P16234	PDGFRA	Platelet-derived growth factor receptor alpha	0.82972663
11	P08581	MET	Hepatocyte growth factor receptor	0.82233972
12	P17948	FLT1	Vascular endothelial growth factor receptor 1	0.81758213
13	P19793	RXRA	Retinoic acid receptor RXR-alpha	0.81251173
14	P36888	FLT3	Receptor-type tyrosine-protein kinase FLT3	0.81027904
15	P10826	RARB	Retinoic acid receptor beta	0.8072451
16	P11362	FGFR1	Fibroblast growth factor receptor 1	0.80376057
17	Q07869	PPARA	Peroxisome proliferator-activated receptor alpha	0.80218581
18	P10276	RARA	Retinoic acid receptor alpha	0.80139136
19	P13631	RARG	Retinoic acid receptor gamma	0.80061832
20	P48443	RXRG	Retinoic acid receptor RXR-gamma	0.79760288

Tezacaftor

Table C.3 – Top 20 predictions of proteins interacting with tezacaftor (VX-661)

Rank	UniProt ID	Gene Name	Name	Prediction Score
1	P08588	ADRB1	Beta-1 adrenergic receptor	0.74769924
2	P01375	TNF	Tumor necrosis factor	0.657412
3	P13945	ADRB3	Beta-3 adrenergic receptor	0.65625447
4	P07550	ADRB2	Beta-2 adrenergic receptor	0.64676234
5	P09917	ALOX5	Arachidonate 5-lipoxygenase	0.63307618
6	P36507	MAP2K2	Dual specificity mitogen-activated protein kinase kinase 2	0.59480746
7	Q02750	MAP2K1	Dual specificity mitogen-activated protein kinase kinase 1	0.54321257
8	P48736	PIK3CG	Phosphatidylinositol 4,5-bisphosphate 3-kinase catalytic subunit gamma isoform	0.53157856
9	P11387	TOP1	DNA topoisomerase 1	0.50231082
10	P28223	HTR2A	5-hydroxytryptamine receptor 2A	0.50031088
11	P08908	HTR1A	5-hydroxytryptamine receptor 1A	0.49273511
12	P25103	TACR1	Substance-P receptor	0.49108715
13	P35348	ADRA1A	Alpha-1A adrenergic receptor	0.4799034
14	Q15759	MAPK11	Mitogen-activated protein kinase 11	0.47887549
15	P35372	OPRM1	Mu-type opioid receptor	0.47691449
16	Q13233	MAP3K1	Mitogen-activated protein kinase kinase kinase 1	0.47611187
17	P43405	SYK	Tyrosine-protein kinase SYK	0.47569695
18	P41595	HTR2B	5-hydroxytryptamine receptor 2B	0.47489231
19	P07949	RET	Proto-oncogene tyrosine-protein kinase receptor Ret	0.47262364
20	P10721	KIT	Mast/stem cell growth factor receptor Kit	0.46576152

Elexacftor

Table C.4 – Top 20 predictions of proteins interacting with elexacaftor (VX-445)

Rank	UniProt ID	Gene Name	Name	Prediction Score
1	P28845	HSD11B1	Corticosteroid 11-beta-dehydrogenase isozyme 1	0.74085683
2	Q59EB3		Met proto-oncogene variant	0.67139369
3	P08581	MET	Hepatocyte growth factor receptor	0.66692742
4	P36888	FLT3	Receptor-type tyrosine-protein kinase FLT3	0.63152203
5	P34903	GABRA3	Gamma-aminobutyric acid receptor subunit alpha-3	0.63020414
6	P15056	BRAF	Serine/threonine-protein kinase B-raf	0.62695787
7	O60674	JAK2	Tyrosine-protein kinase JAK2	0.6150386
8	P14867	GABRA1	Gamma-aminobutyric acid receptor subunit alpha-1	0.60464625
9	P21802	FGFR2	Fibroblast growth factor receptor 2	0.60233647
10	P22607	FGFR3	Fibroblast growth factor receptor 3	0.60061471
11	P07949	RET	Proto-oncogene tyrosine-protein kinase receptor Ret	0.59901981
12	P42685	FRK	Tyrosine-protein kinase FRK	0.5952845
13	P10721	KIT	Mast/stem cell growth factor receptor Kit	0.59286403
14	P22455	FGFR4	Fibroblast growth factor receptor 4	0.58879764
15	P41240	CSK	Tyrosine-protein kinase CSK	0.58677948
16	P11362	FGFR1	Fibroblast growth factor receptor 1	0.5825777
17	P48169	GABRA4	Gamma-aminobutyric acid receptor subunit alpha-4	0.57733962
18	P47869	GABRA2	Gamma-aminobutyric acid receptor subunit alpha-2	0.57535979
19	Q14289	PTK2B	Protein-tyrosine kinase 2-beta	0.57522896
20	P16234	PDGFRA	Platelet-derived growth factor receptor alpha	0.57464986

Appendix D

Exploring isofunctional molecules: Design ofabenchmark andevaluation ofprediction performance

Received: 30 August 2022 | Revised: 19 December 2022 | Accepted: 11 January 2023

DOI: 10.1002/minf.202200216

RESEARCH ARTICLE

molecular informatics

Exploring isofunctional molecules: Design of a benchmark and evaluation of prediction performance

Philippe Pinel[1, 2, 3, 4] | **Gwenn Guichaoua**[1, 2, 3] | **Matthieu Najm**[1, 2, 3] | **Stéphanie Labouille**[4] | **Nicolas Drizard**[4] | **Yann Gaston-Mathé**[4] | **Brice Hoffmann**[4] | **Véronique Stoven**[1, 2, 3]

[1]Center for Computational Biology, Mines Paris-PSL, PSL Research University, Paris, France

[2]Institut Curie, Paris, France

[3]INSERM U900, Paris, France

[4]Iktos SAS, Paris, France

Correspondence
Philippe Pinel, Center for Computational Biology, Mines Paris-PSL, PSL Research University, 75006 Paris, France.
Email: philippe.pinel@iktos.com

Abstract

Identification of novel chemotypes with biological activity similar to a known active molecule is an important challenge in drug discovery called 'scaffold hopping'. Small-, medium-, and large-step scaffold hopping efforts may lead to increasing degrees of chemical structure novelty with respect to the parent compound. In the present paper, we focus on the problem of large-step scaffold hopping. We assembled a high quality and well characterized dataset of scaffold hopping examples comprising pairs of active molecules and including a variety of protein targets. This dataset was used to build a benchmark corresponding to the setting of real-life applications: one active molecule is known, and the second active is searched among a set of decoys chosen in a way to avoid statistical bias. This allowed us to evaluate the performance of computational methods for solving large-step scaffold hopping problems. In particular, we assessed how difficult these problems are, particularly for classical 2D and 3D ligand-based methods. We also showed that a machine-learning chemogenomic algorithm outperforms classical methods and we provided some useful hints for future improvements.

KEYWORDS
benchmark, chemogenomics, ligand-based, molecular interactions, scaffold hopping

1 | INTRODUCTION

Identification of novel chemotypes with biological activity similar to a known active molecule is a critical and recurrent challenge in drug discovery called 'scaffold hopping' [1]. Indeed, once a hit molecule has been identified against a therapeutic target, it may not be a proper drug candidate because of poor selectivity or ADME profile, unacceptable toxicity, or complex, inefficient, or expensive synthesis routes. The hit compound's chemotype may also be protected by patents, which restrains the downstream development process. Various strategies are available to solve scaffold hopping problems.

The "scaffold hopping" term has been used in different ways in the literature and remains ambiguous [2]. One definition relies on keeping substituents (R-groups)

Brice Hoffmann and Véronique Stoven have contributed equally to project management.

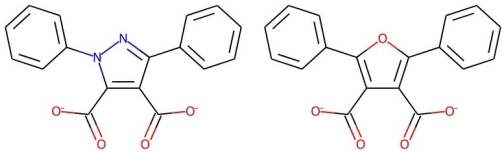
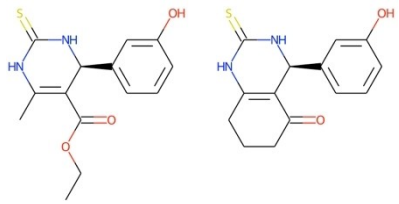
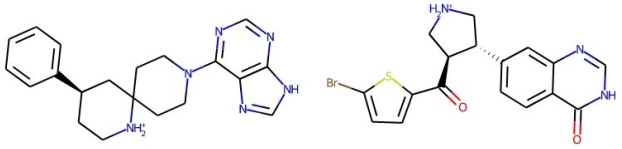
that form interactions with the protein pocket while changing the core of the molecule. This usually means substitution of ring systems and linker fragments between rings by other molecular moieties. Medicinal chemists have designed different strategies to identify such new scaffolds from a parent molecule: swapping of carbons and heteroatoms in heterocycles, heterocycles ring opening or closure. They lead to new compounds that retain some degree of similarity with the parent compound and have been classified as small- to medium-scaffold hopping strategies [3]. Small- to medium-step hopping problems are tractable by a trained medicinal chemist, but a variety of efficient ligand-based methods have also been proposed to help solve these cases [3,4]. These methods are usually referred to as QSAR methods. Other contributions to the field include identification of completely different molecules that are unrelated to the parent compound and with which no common R-group or core structures can be defined. Indomethacin and Etoricoxib are two examples of such structurally unrelated COX2 inhibitors [5]. Such examples usually arise from topology-based (or 3D) approaches. New molecules identified by such methods can display greater chemical novelty with respect to the parent compound (i.e. the two molecules have very dissimilar chemical structures), sharing no common R-groups, while still forming the same key interactions with a protein pocket. Such pairs of molecules can be seen as "isofunctional" molecules, and have been referred to as large-step scaffold-hopping cases [3], and this term will be used throughout the paper. Table 1 shows examples of small-, medium- and large-step scaffold hopping examples, and their associated Murcko-based or molecular Morgan similarities.

In the present paper, we focus on the problem of large-step scaffold hopping, involving pairs of active molecules with very low structure similarity, because this corresponds to the most difficult problems in scaffold-hopping that typically require computational approaches [2].

When a high-quality 3D crystallographic structure is available for the protein target, or when a reliable protein structure model can be built, docking methods can help identify ligands with new scaffolds [6]. However, docking is not applicable when the structure of the targeted protein is unknown. This is often the case for transmembrane proteins, a category of protein targets for many marketed drugs [7]. Hence, in this paper, we focus on computational methods that do not require knowledge of the target's 3D structure, so that the methods are applicable to all cases.

In this context, most studies reporting success cases using 2D and 3D approaches considered a single protein or a very small number of proteins [8–11]. Their performance in the general cases is essentially unpredictable [2], and there is a crucial need to design benchmarks that span a variety of proteins, to evaluate the performances of computational methods for solving large-step scaffold hopping problems.

TABLE 1 Examples for the three degrees of scaffold hopping: from the small-step scaffold hopping cases, to the large-step scaffold hopping cases, characterized by their respective Morgan and Generic Murcko similarities, as described in section 2.2.3. For each pair, the molecules bind similarly to the same protein.

Scaffold hopping degree	Description	Molecules	Murcko-based Morgan similarity	Molecular Morgan similarity
Small-step	Change of atoms in hetero-cycles		1.0	0.31
Medium-step	Ring opening and closure		0.36	0.43
Large-step	Novel core structure		0.21	0.15

As a matter of fact, a few large-scale benchmark studies have been reported, comparing the performances of various topology-based (or other ligand-based) methods [12,13]. However, they only considered proteins with a relatively large number of known ligands, so that these ligands can be used to train prediction models. In addition, they evaluated the performances based on the chemical diversity of known ligands retrieved among the top ranked molecules. This does not clearly specify to which extent the structures of retrieved ligands were distant from those of molecules in the training set. In other words, it does not characterize the sizes of the corresponding hops and prevents from drawing conclusions about the ability of these methods to specifically solve large-step scaffold hopping problems. Finally, because of their design, these benchmarks do not mimic real-life applications, where an active compound is known, and a new active with very different chemical structure is searched. Therefore, it is difficult to anticipate the performances of the proposed methods in a real-life setting.

In addition to the degree of chemical novelty searched, one must distinguish 'easy' targets with many known ligands, and 'hard' targets with only one known ligand. Most available computational methods may fail on the latter [14].

Overall, one of the main challenges in the field of computational methods for scaffold hopping is the lack of appropriate benchmarks to evaluate those methods on ′hard′ targets and large-step hops, because these settings are typically encountered in the design of new drugs and correspond to the most difficult cases. In this paper, we precisely address this challenge. Our main contribution is to provide a flowchart to build a high-quality and well characterized large-step scaffold hopping benchmark for 'hard' targets, which is a prerequisite to develop and test new methods dedicated to these problems. We also illustrate how to use this benchmark to compare the performance of a few classical 2D and 3D ligand-based methods and of an alternative approach that relies on a machine-learning chemogenomic algorithm. This process allows us to evaluate the difficulty of large-step scaffold hopping problems in a setting that corresponds to real-case studies.

2 | RESULTS

In the following sections, we present the global approach adopted to build the Large-Hops benchmark ($\mathcal{LH}$), designed for problems of large-step scaffold hopping for 'hard′ targets, and propose criteria to evaluate computational methods using the benchmark. Then, we detail how this benchmark is built by gathering a dataset of well characterized large-step scaffold hopping cases with their corresponding decoy molecules. Finally, we compare the performance of different computational methods on this benchmark.

2.1 | Overall design of the Large-Hops benchmark and criteria for performance evaluation of computational methods

In order to go beyond previous benchmark studies, we build the Large-Hops benchmark $\mathcal{LH}$ of well characterized large-step scaffold hopping cases, as detailed in the next section. It comprises pairs of active molecules illustrating large-step scaffold hopping cases, and their corresponding 499 decoy molecules, to reach meaningful active/inactive ratio of 1/500, which is well below the frequently used ratio of 1/50 [15].

To evaluate computational methods on this benchmark, we follow a scheme that mimics real-world settings for 'hard' targets: for each pair ($molecule_1$, $molecule_2$) of known active molecules against a given target, one molecule is set apart as the only known active (for example $molecule_1$), while the other (query molecule $molecule_2$, called the unknown active) is added to the 499 decoys. Then, given the known active molecule, computational methods are used to rank the unknown active and the 499 decoys. The higher the rank of the unknown active, the more efficient is the method to solve this particular scaffold hopping case (best rank being 1, worst being 500). Note that for each pair of actives, one molecule or the other can be used as the known active, which leads to twice more scaffold hopping problems as pairs of active molecules in the benchmark. We propose to compare the methods based on three criteria: (1) We draw Cumulative Histogram Curves (CHC), representing the number of cases for which the considered method ranked the unknown active below a given rank, as detailed in Materials and Methods. The curves of the best performing methods will stand above those of the other methods. (2) This will be quantitatively assessed by the Area Under the Curve (AUC) of the CHC curves, to provide a global comparison of the methods. (3) In real-life screening campaigns, only the best ranked molecules are usually considered as candidate molecules for experimental tests. Thus, we also compare the relative positions of the CHC curves at high ranks and determine the proportion of cases where the unknown active is retrieved in the top 5% best ranked molecules [12], which can be seen as the success rate of the methods. The global principle of the benchmark design is illustrated in Figure 1.

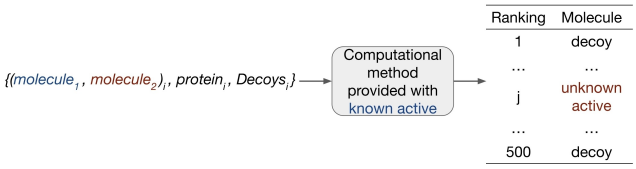


FIGURE 1 Principle of performance evaluation on the LargeHops benchmark. For instance *i*, one molecule of the pair is set as the known active, and the other as the unknown active. The known active is provided to the computational method that ranks the unknown active and the decoys. The rank *j* of the unknown active is used to evaluate the considered computational method.

2.2 | Building a dataset of large-step scaffold hopping cases

Our goal is to build a benchmark of large-step scaffold hopping cases that can be used by ligand-based computational methods. This allows to compare the performances of methods to solve scaffold hopping problems in the general case where the 3D structure of the considered protein is unknown. Nevertheless, we built this benchmark from examples extracted from the PDBbind database to ensure that the selected pairs of molecules are 'true' large-step scaffold hopping cases, i.e., highly dissimilar compounds that share similar binding modes with the same protein, as identified by the same UNIPROT ID. Indeed, for example, there would not be any rationale to relate two inhibitors of an enzyme binding to two distinct and distant binding sites, and such 'false' cases must not be present in the benchmark. Note that the molecules binding modes are only used to select 'true' scaffold hopping cases when building the benchmark. They are not further used or provided to the considered computational methods, in order to remain in a ligand-based framework.

Identification of such examples is not straightforward: some examples presented below show that it is not possible to use only one criterion based on a single molecular similarity measure, in order to build a reliable dataset where 'false' large-step cases and 'false' scaffold hopping cases are not present. The next subsections present the subsequent steps that are used, and more details are given in Materials and Methods.

2.2.1 | Filtering the PDBbind database

To identify scaffold hopping cases, we need to search for pairs of highly different molecules that bind to the same protein pocket with similar binding modes, because molecules that would present totally different binding modes in the same pocket, or bind to different pockets of the protein, do not meet the definition of scaffold hopping. To enable the selection of such pairs, we use the PDBbind database [16] that contains 17.652 PDB files (2019) of 3D crystallographic structures of protein-ligand complexes. We only keep structures with a resolution below 2.8 Å, to ensure that the binding modes of the ligands can be analysed with confidence. Second, as some compounds can be crystallized by soaking experiments even with unspecific affinities in the millimolar range, we remove all complexes with affinity above 10 μM. This allows to only select 'successful' scaffold hops, for which both molecules present specific activities against the same target. Finally, we discard proteins for which only one ligand is available in PDBbind, since scaffold hopping examples cannot be searched for these proteins. This leads to 181.635 pairs of ligands. Examples of large-step scaffold hopping cases are further searched among these pairs.

2.2.2 | Selecting pairs of drug-like molecules

Because our study stands in the context of drug design, the selected molecules need to represent molecular characteristics encountered in drug-like molecules. Otherwise, the performances of computational methods on our benchmark may not be representative of those expected in drug design applications. We only keep pairs involving ligands of molecular weight between 200 and 900 g/mol, to discard salts, solvent or other molecules present in crystallisation buffers, and large interacting partners like peptides. This leads to 6.494 PDB files, corresponding to 148.002 pairs of ligands and involving 856 different proteins. Among the 148.002 pairs, we select those in which both molecules have a quantitative estimate of drug-likeness (QED) above 0.5 [17], which allows to remove molecules with unwanted physical-chemical properties, leading to 49.686 pairs involving 449 proteins.

2.2.3 | Selecting pairs of large-step hops ligands

Among the 49.686 pairs of molecules, we use several criteria to exclude those corresponding to small- or medium-step hops cases. First, we determine the generic Murcko scaffolds of molecules because they characterize the core structure of molecules [18]. These scaffolds are obtained by removal of all substituents while retaining ring systems and linker moieties between rings, and converting all bonds to single bonds. To remove small-step hops, we exclude pairs of ligands whose generic Murcko

scaffolds have Morgan fingerprints [19] Tanimoto similarities above 0.6 (in the following, this similarity is called Murcko-based Morgan similarity), which selects 45.534 pairs. This single criterion does not always guarantee that the two molecules are highly dissimilar: in a few cases, they still display significant similarities, as illustrated in Figure 2 for one pair. To discard these cases, pairs of molecules with an overall Morgan fingerprints Tanimoto similarity above 0.3 are removed (in the following, this similarity is called molecular Morgan similarity), as they may represent medium-step hops rather than large-step hops. This leads to 44.386 pairs of molecules.

2.2.4 | Selecting pairs with similar binding modes

Among the 44.386 pairs of highly dissimilar molecules, we need to identify those that correspond to a scaffold hopping case, i.e., to select those in which two molecules have similar binding modes within the same protein pocket. As explained in section 2.2.1, this step is necessary to ensure that all the cases retrieved share the same rationale and thus could be identified either by a chemist or a computational method in the benchmark. Various binary target-focused protein-ligand interaction fingerprints (IFP) have been proposed to perform such tasks, because they are an easily interpretable way to encode binding modes [20–23]. As detailed in Materials and Methods, we develop our own IFPs that include usual interactions and a few types of additional interactions that are missed by classical IFPs [24–28].

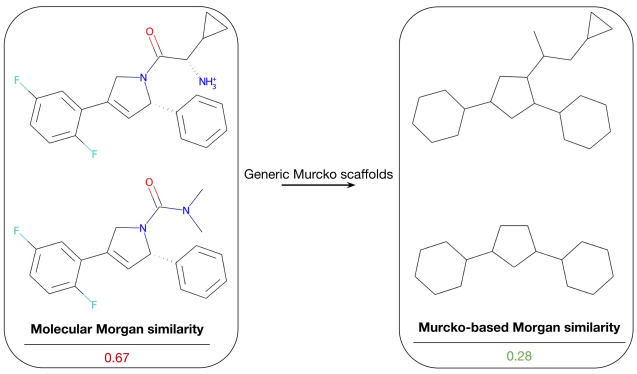


FIGURE 2 Example of a pair of molecules with low Murcko-based Morgan similarity but similar structures, leading to a higher molecular Morgan similarity. On the left the pair of molecules (PDBs: '2fl2' and '2fl6') is displayed and, on the right, their corresponding generic Murcko scaffolds are shown. This pair should not be present in the $\mathcal{LH}$ benchmark. It is excluded based on the molecular Morgan similarity between the molecules greater than the chosen threshold.

A Tanimoto similarity between IFPs is used to compare the binding modes of ligands and remove 'false' scaffold hopping cases, as illustrated in Figure 3. Ligands forming only few interactions with the protein (less than five) are removed, as the computation of Tanimoto similarities would not be reliable. We keep pairs of ligands with IFPs similarities above 0.6. This leads to 821 pairs of molecules with highly dissimilar chemical structures, but similar binding modes.

2.2.5 | Discarding pairs based on Maximum Common Substructures

Among the 821 pairs, visual analysis allows us to observe cases where the two molecules share a common substructure forming most of the interactions with the protein. These cases cannot be considered as scaffold hops if the common substructure is responsible for most of their interactions with the protein pocket since these substructures can then be viewed as a common scaffold that drives binding to the protein. To remove these instances, we use the Maximum Common Substructure (MCS) concept, because it has been shown to help identify scaffold hopping cases [29]. For each pair of molecules, we search for their MCS and compute the ratio between the number of common interactions arising from chemical groups in the MCS, and the total number of common interactions to the two molecules. A high ratio means that the MCS is responsible for most of the common interactions, and the corresponding pair should not be considered as a large-step scaffold hopping case, as described in Figure 4.

Concretely, the MCS between two molecules is searched based on three different types of MCS, as defined in RDKit [30]: MCS with matching of complete

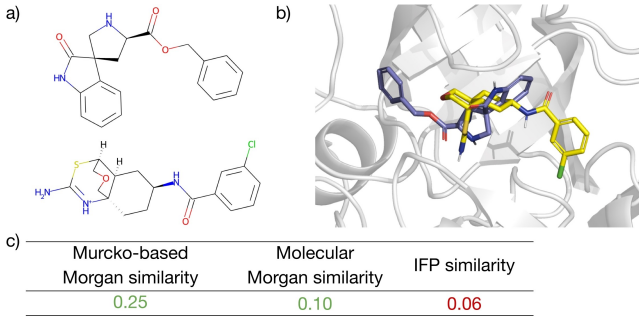
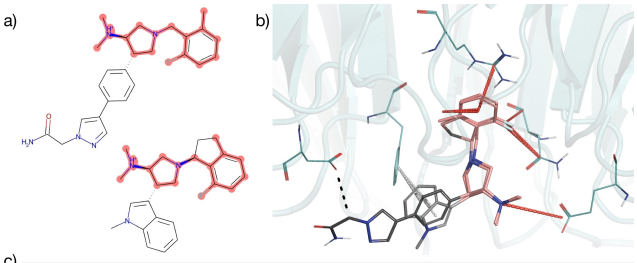
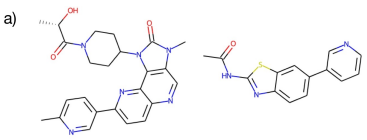


FIGURE 3 Example of a pair of dissimilar ligands for Beta-Secretase 1 (PDBs: '3udm' and '4zsq') occupying different areas of the binding site of the protein. The molecules are shown in a). The crystallographic conformations are displayed in b). Table c) compares the two molecules: they share little common binding modes and cannot be considered as a scaffold hopping case.



Murcko-based Morgan similarity	Molecular Morgan similarity	IFP similarity	MCS interaction Ratio
0.23	0.22	0.86	0.83

FIGURE 4 Example for Polycomb protein EED of molecules (PDBs: '5u6d' and '5u8f') with similar binding modes explained by a common substructure. The two ligands are displayed in a) with their common substructure highlighted in light red. Their crystallographic conformations are shown in b) along with their interactions with the protein. The red interactions corresponds to common interactions arising from the common substructure (colored in light red in the molecules), while the light grey interaction is the only common interaction arising from dissimilar parts of the molecules. Table c) compares the two molecules. As 5 out of the 6 common interactions are explained by the MCS, such a case cannot be considered as a scaffold hopping example.



b)

	MCS complete rings	MCS partial rings	MCS ring breaking
Properties	Only matches complete rings	Allows to match partial rings	Can break rings to match ring bonds to non ring bonds
SMARTS			
MCS interaction ratio	0.4	0.6	0.2

FIGURE 5 Illustration of the three different MCS searched. A pair of molecules (PDBs: '4hvb' and '4 ps7') is displayed in a), and the table describing the three different MCS searched on this pair along with their ratios of common interactions is shown in b).

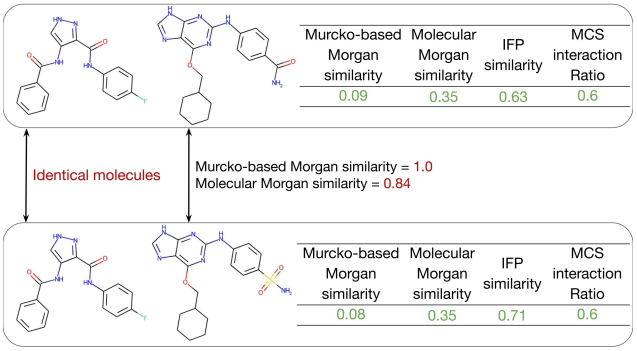


FIGURE 6 Example of redundant scaffold hopping cases for the cell division protein kinase 2: the two pairs are highly similar, since the second pair (PDBs: ('2vto', '4eok')) can be obtained from the first pair (PDBs: ('2vto', '1oiy')) by replacing the amide group on one of the molecule by a sulfonamide. In such cases, one of the two pairs was discarded.

rings, MCS with partial matching of rings and MCS with allowed ring breaking. In particular, the first MCS searches for complete ring matches, allowing to discard pairs of molecules that would correspond to small-step scaffold hops.

The maximum ratio of the number of common interactions formed by MCSs and the total of common interactions between the two molecules of a pair is computed as detailed in Materials and Methods. Pairs with a maximum ratio above 0.8 were discarded, resulting in 531 pairs for 79 proteins.

Overall, the three types of substructure search are complementary, and the maximum ratio of common interactions formed by the MCSs over the total number of common interactions ensures to retrieve only large-step scaffold hopping cases. An example of MCS search between two molecules is given in Figure 5.

2.2.6 | Discarding redundant pairs

We observe that for some of the 79 proteins, the selected pairs are strongly redundant and represent only slightly different examples of scaffold hopping cases: they involve two molecules that belong to the same chemical series (for instance, they differ by the addition of a small group not involved in the binding). A compelling example is given in Figure 6.

To avoid redundancy in our dataset, which may lead to bias for performance evaluation of computational methods, we remove pairs in which both ligands are similar to both ligands of another pair, using a threshold of 0.5 on both their molecular Morgan similarities as detailed in Material and Methods, which finally leads to 178 pairs.

2.2.7 | Resulting large-step scaffold hopping dataset

The global selection flowchart is shown in Figure 7. Overall, 178 large-step scaffold hopping cases of drug-like pairs binding to 79 different proteins are selected. On average, each protein is involved in 2.3 large-step scaffold hopping cases in the dataset. For the most represented protein, cell division protein kinase 2, 10 cases are selected. The most represented family of proteins is the kinases family, with 61 pairs involving 21 different kinases. This can be

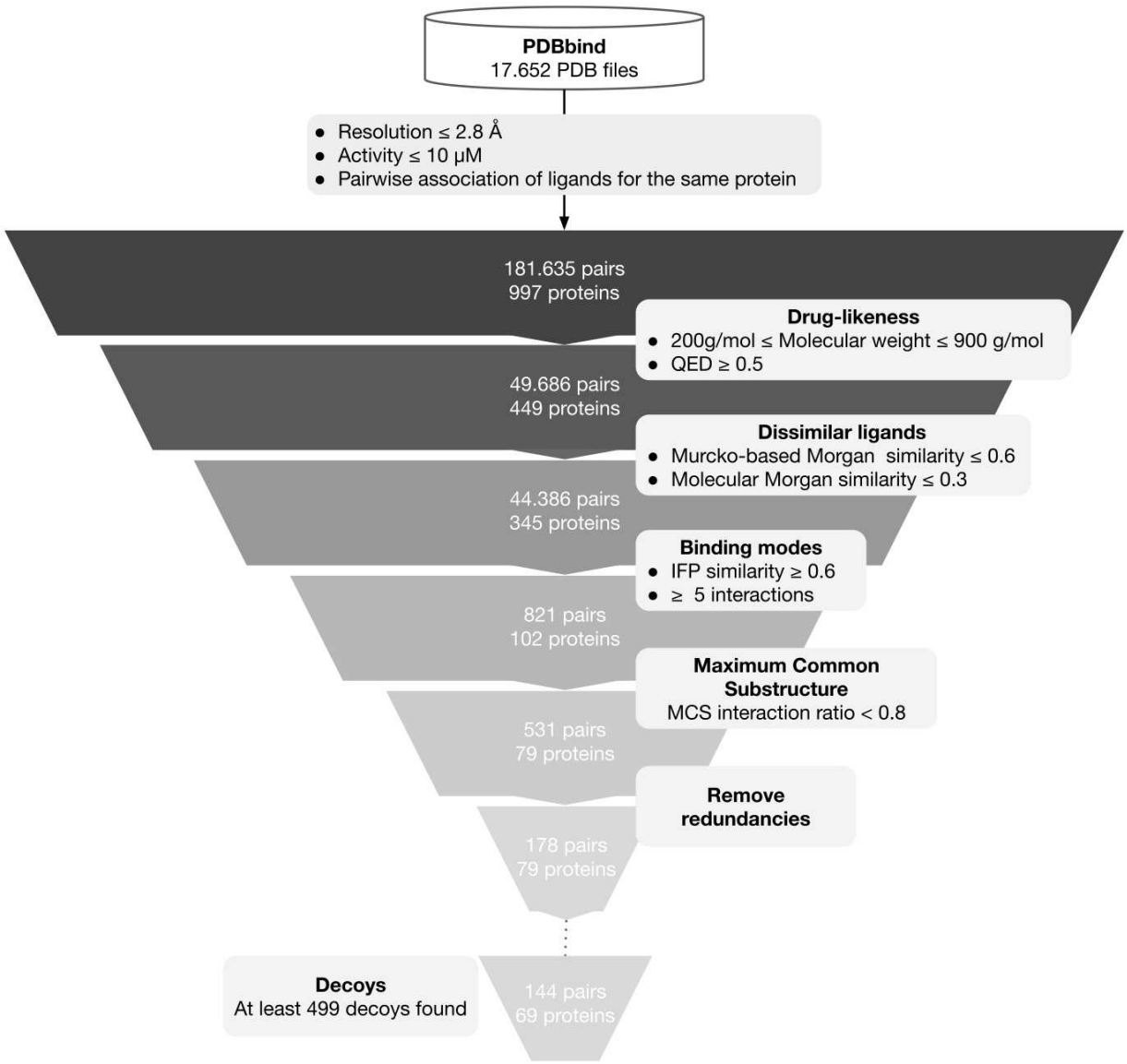
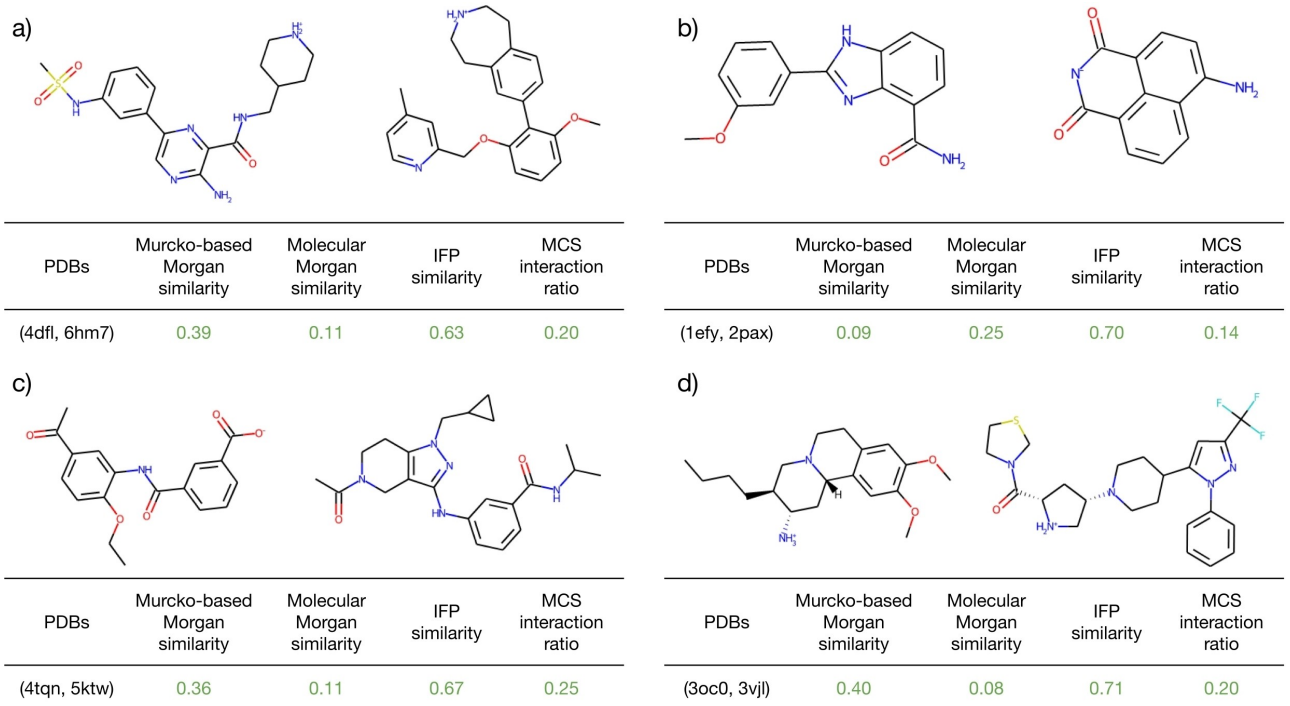


FIGURE 7 Flowchart describing the successive filters applied to identify large-step scaffold hopping cases. Starting from PDBbind crystal structures with good resolutions of proteins in complex with at least two ligands (181.635 pairs), we keep those involving drug-like molecules of dissimilar structures but similar binding modes. We removed pairs containing a common substructure responsible for most common interactions. We then discarded redundant pairs, leading to 178 large-step scaffold hopping cases. Among these cases, we keep those for which 499 decoy molecules could be found (see below). The chosen thresholds are arbitrary but ensured us to retrieve only confident large-step scaffold hopping cases, as detailed in the Results section.

explained by the fact that kinases belong to a highly studied family of proteins, with many therapeutic targets against which many drug design projects have been devoted [31]. However, the dataset still contains significant protein diversity, since the 79 proteins belong to 35 different super-families of the SCOP protein family's hierarchy database [32]. On average, each super-family is involved in 5.1 scaffold hops. Pairwise sequence identities between proteins have been computed using the Needleman-Wunsch algorithm, which shows that they share modest similarity: on average, proteins have pairwise sequence identities of 8%, and only 12 pairs (over a total of 3081 pairs of proteins) display identities above 30%.

Each selection step involves selection criteria with threshold values. The above sections show that one must be careful to avoid 'false' large-step hops, or 'false' scaffold hopping cases, which has scarcely been discussed in previous benchmark studies. In the present work, the thresholds are chosen arbitrarily and somewhat stringently, to build a highly reliable dataset, as judged by visual analysis of the selected pairs. Of course, these thresholds can be changed, as detailed in the Discussion



PDBs	Murcko-based Morgan similarity	Molecular Morgan similarity	IFP similarity	MCS interaction ratio
(4dfl, 6hm7)	0.39	0.11	0.63	0.20

PDBs	Murcko-based Morgan similarity	Molecular Morgan similarity	IFP similarity	MCS interaction ratio
(1efy, 2pax)	0.09	0.25	0.70	0.14

PDBs	Murcko-based Morgan similarity	Molecular Morgan similarity	IFP similarity	MCS interaction ratio
(4tqn, 5ktw)	0.36	0.11	0.67	0.25

PDBs	Murcko-based Morgan similarity	Molecular Morgan similarity	IFP similarity	MCS interaction ratio
(3oc0, 3vjl)	0.40	0.08	0.71	0.20

FIGURE 8 Examples of selected large-step scaffold hopping for a) Tyrosine-protein kinase SYK, b) Poly (ADP-ribose) polymerase, c) Creb-binding protein and d) Dipeptidyl peptidase 4.

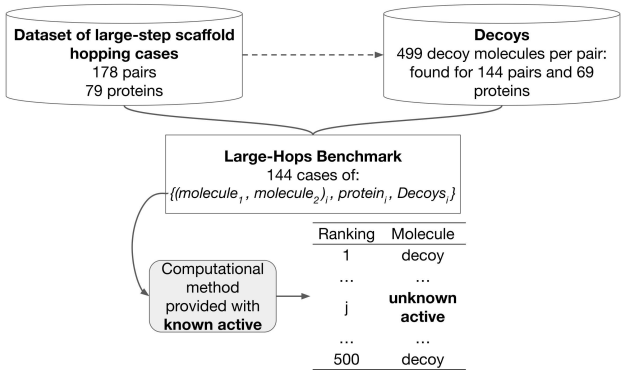


FIGURE 9 The Large-Hops benchmark was built from a dataset of large-step scaffold hopping cases extracted from PDBbind for which 499 decoy molecules were gathered. Overall, it comprises 144 cases defined by a pair of active molecules against the same protein target, and their corresponding decoys.

section. Examples of selected large-step scaffold hopping are provided in Figure 8.

2.3 | Choice of decoy molecules

As illustrated in Figure 9, our benchmark consists in a dataset of pairs of molecules representing large-step scaffold hopping cases for the associated protein, and their corresponding decoy molecules. Hence, for each of the 178 selected pairs, a set of decoys needs to be defined. In fact, all studies devoted to benchmarking virtual screening methods require the selection of decoy compounds, which is not an easy problem as we want to avoid statistical bias with respect to the active molecules [33]. In particular, when decoys stand in regions of chemical space that are very distant from the two active molecules, the resulting benchmark may suffer from 'analogous bias' [34], and the success rate may be artificially overestimated. In addition, since we also want to mimic real-life applications, the decoys must be realistic scaffold hopping candidates that, in practice, would be searched among molecules sharing some physicochemical characteristics with the known active.

It has been shown that random selection of decoys from large chemical databases does not prevent the occurrence of statistical bias [33]. Therefore, as detailed in Materials and Methods, decoys are picked in the ZINC database [35], among molecules that have similar physicochemical and complexity properties [36] with respect to the two molecules of their corresponding pair of actives. More precisely, for each pair, the decoys are selected in such a way that they are as distant to both molecules of the pair, as these two molecules are distant from each other, according to the Murcko-based and molecular Morgan similarities. For 144 pairs out of the 178 pairs selected in the previous sections, we are able to find 499 decoy molecules that satisfy these criteria. Note that we cannot rule out the possibility of a few false negatives [37,38], because we may have accidentally picked

decoys that bind to the same protein as the molecules in the pair. However, we assume that such cases are rare and that their potential presence does not change the overall conclusions of the paper.

2.4 | The final Large-Hops benchmark

The resulting $\mathcal{LH}$ benchmark finally consists in 144 pairs of molecules associated with their corresponding proteins and their 499 decoys (the 499 decoys are different for each of the 144 pairs). These 144 pairs of active molecules involve 69 different proteins, belonging to 31 different super-families of the SCOP [32]. This benchmark contains scaffold hops cased within protein families that have been more extensively studied than others. Nevertheless this panel of proteins is wide enough to set apart from a case study, and to get a broader glance at how well a computational method performs to solve large-step scaffold hopping problems. Taken together, the rules used to select the active pairs and their decoys are stringent, but this ensures to build a realistic, high-quality, and well characterized benchmark dedicated to the problem of large-step scaffold hopping.

2.5 | Performance evaluation of classical 2D and 3D ligand-based methods on the Large-Hops benchmark

As introduced in section 2.1 and displayed in Figure 9, the goal of our benchmark is to evaluate ligand-based methods' ability to solve large-step scaffold hopping problems. For each pair of active molecules in the benchmark, each method is given one active, called the known active, while the other, called the unknown active, is joined to 499 decoys. Each method ranks the unknown active and the decoys. As either of the two active molecules can be set as the known active, this leads to twice as many experiments as scaffold hopping cases. No other ligand information for the corresponding protein is provided to the computation methods.

Ligand-based methods usually rely on molecular descriptors to encode molecules. This encoding is used to train QSAR models that predict a property for molecules, based on prior knowledge of molecules that present this property, or not. This principle was applied in previous ligand-based scaffold hopping benchmark studies [12,13]. Such QSAR models cannot be trained on the $\mathcal{LH}$ benchmark, because a single active molecule is provided to the algorithm, i.e., one molecule from each active pair. However, encoding of molecules with descriptors can also be viewed as the representation of a chemical space in which similarity measures between molecules can be defined. This can be used to rank the unknown active and the decoys according to their similarity with respect to the known active.

Many types of 2D descriptors have been proposed in the literature, and perform very well for the prediction of various molecular properties [39]. Since scaffold hopping relates to ligand binding to a protein, a phenomenon occurring in the 3D space, 3D descriptors are expected to be more relevant in this context. We first explore the performance of classical 2D similarity measures (usually employed for small- to medium- scaffold hops), and then study classical 3D similarity measures. Finally, we also consider a more original chemogenomic algorithm, to show how the benchmark could be used for the development of new methods dedicated to large-step scaffold hops.

2.6 | 2D similarity-based methods

Because they neither require the 3D structure of the target, nor the 3D conformations of the molecules, we first consider 2D structure descriptors. Although they are not meant to best encode ligand binding properties, it is interesting to see whether these simple methods capture some valuable information to solve scaffold hopping problems.

We consider three types of 2D representations: (1) Morgan fingerprints that encode 2D molecular structures. These descriptors are not expected to perform well on our benchmark, because solving large-step scaffold hopping problems requires to search for molecules with highly dissimilar chemical structures. In addition, the molecular Morgan Tanimoto similarity was used to select pairs of dissimilar active molecules, so that ranking the unknown active and the decoys according to this similarity is doomed to fail. Testing 2D Morgan fingerprints encoding is a kind of internal control to confirm that our benchmark is an interesting tool for the development of original methods dedicated to large-step scaffold hopping. (2) MACCS keys fingerprints [40], that in principle should suffer from the same limitations as the Morgan fingerprints. In fact, since the former encodes the presence or absence of particular chemical groups rather than the molecular graph itself, it is interesting to test if this can be beneficial to the current problem. (3) 2D pharmacophore fingerprints, that encode for the presence and relative positions in the 2D graph of the molecular structure of chemical groups able to drive different types of interactions with the protein, as defined in RDKit [30]. Although these descriptors implement a notion of 2D (but not 3D) topology, they may improve

over MACCS keys fingerprints. These three types of 2D representations lead to a binary vector encoding for the molecules, allowing the definition of corresponding similarity measures based on Tanimoto coefficients. Thus, for each pair of active molecules in the benchmark, the unknown active and the decoys are ranked according to their molecular Morgan, MACCS keys and 2D pharmacophore Tanimoto similarity with respect to the known active, as detailed in Materials and Methods.

As shown in Figure 11, the molecular Morgan similarity displays overall very poor performances in terms of AUC (random ranking corresponds to an AUC of 0.5). At high ranks, its CHC curve stands only slightly above those random rankings, and the unknown active is retrieved in the top 5% in only 11.5% of the cases. Global failure of the Morgan similarity confirms that our benchmark mainly comprises large-step scaffold hopping cases. The MACCS keys and 2D pharmacophore similarities both improve over the molecular Morgan similarity. The 2D pharmacophore similarity was expected to perform better than the MACCS keys similarity, but their AUC score, relative positions of CHC curves at high ranks, and success rates in the top 5% best ranked molecules are comparable. In fact, in our benchmark, since the two active molecules have very dissimilar chemical structures, the inter-distances between pharmacophoric groups measured on a planar representation of the molecule and encoded in the 2D pharmacophore descriptors may not bring additional information for the problem at hand than the simple presence or absence of specific chemical groups encoded in the MACCS keys descriptors. Overall, the performances of these two methods remain modest since their success rate in the top 5% is below 15%.

2.7 | 3D similarity-based methods

We also tested 3D approaches, since they capture molecular features that can be better related to ligand binding. We consider two types of representations: 3D molecular shape, and 3D pharmacophores, because both approaches have been described as useful tools to help solving scaffold hopping problems [41,42]. We study the general case where the 3D structure of the protein is unknown, so that the 'active' conformations (i.e. the ligand conformation when bound to the protein pocket) of the active molecules are unknown. For each pair of active molecules and their 499 decoys in the $\mathcal{LH}$ benchmark, we consider up to ten conformers of low energy, and the unknown active is ranked among the decoys based on their 3D shape or pharmacophore similarity with respect to the known active. All details about conformers calculation, and 3D similarities are given in Materials and Methods.

The performance of the 3D pharmacophore and shape methods are presented in Figure 11. 3D pharmacophore similarity performs better than shape similarity on all criteria: AUC score, relative positions of the CHC curves, and success rate at 5%. This may be explained by the fact that 3D pharmacophore descriptors encode key information about chemical groups able to form interactions with a protein that are not present in the solely molecular shape. This result is in agreement with previous studies where 3D pharmacophore is depicted as a reference method for scaffold hopping [42]. The performances of 3D pharmacophore remain above those of the shape similarity, or those of 2D methods. This is an interesting result, because some studies have reported that when the active conformations are unknown, performances of 2D methods might outperform those of 3D methods with calculated conformers, in the context of ligand binding prediction [43].

According to the results observed on the Large-Hops benchmark, a classical 3D pharmacophore appears as a good default similarity measure to solve large-step scaffold hopping problems. Note however that the achieved success rate at 5% lies around 20%. This allows to quantify the range in performance that can be expected, in general, with classical approaches on these types of problems, thus answering to the question raised by Bajorath [14]. This leaves much room for the development of approaches more specifically designed to solve large-step scaffold hopping problems.

2.8 | Performance evaluation of a machine-learning chemogenomic algorithm on the Large-Hops benchmark

With the $\mathcal{LH}$ benchmark, we tackle large-step scaffold hopping cases with 'hard' targets, for which only one active ligand is known. This setting prevents from training ligand-based prediction models, and the associated computational methods are restricted to similarity measures, as above. Chemogenomics can overcome this limitation if bindings involving other molecules and other proteins are known (these molecule-protein pairs are noted (m, p) pairs in the following). Such (m, p) pairs can be collected from many public databases, such as the PubChem at NCBI [44]. Basically, the main difference between ligand-based and chemogenomic methods is that the former predict ligands for a query protein given its known ligands (one known active in our case), while the latter predict ligands for a query protein given its known ligands and those known for other proteins. In the case of

the $\mathcal{LH}$ benchmark, chemogenomic algorithms can be trained with the (*known active, query protein*) binding pair and additional (*ligand, protein*) pairs known to bind, or not, gathered from an external database. Once trained, the prediction model provides a binding probability for the (*decoys, query protein*) and (*unknown active, query protein*) pairs, and the unknown active can be ranked among the 499 decoys according to these probabilities.

We use a training set derived from the DrugBank database [45] to provide the additional (*m, p*) pairs. Indeed, DrugBank provides high quality bio-activity information regarding approved and experimental drugs, including their targets. It contains around 15.000 curated drug-target (*m, p*) pairs for 2.670 proteins. Although much smaller than other databases like PubChem or ChEMBL, DrugBank appears relevant to the $\mathcal{LH}$ benchmark because it contains drug-like ligands. More precisely, for each pair of active ligands in the $\mathcal{LH}$ benchmark, the ML chemogenomic algorithm is trained with the same information regarding the query protein as the ligand-based methods tested above: a single known pair between the query protein and the known active. This pair is added to the DrugBank dataset, if not already present. All other pairs involving the query protein and any other molecule, if present in the DrugBank dataset, are removed. This allows to avoid redundancies between the training set and the tested pairs, and to compare the prediction performances of the chemogenomic method with those of the ligand-based methods tested in the previous sections because both types of algorithms are provided the same ligand information for the query protein. Figure 10 summarizes the difference between the ligand-based and the chemogenomic setups. All details about the training scheme and the ML algorithm are given in Materials and Methods.

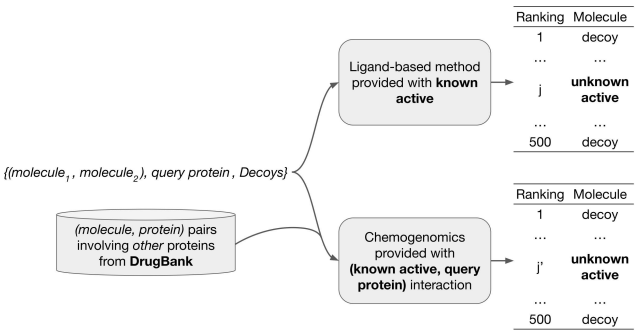
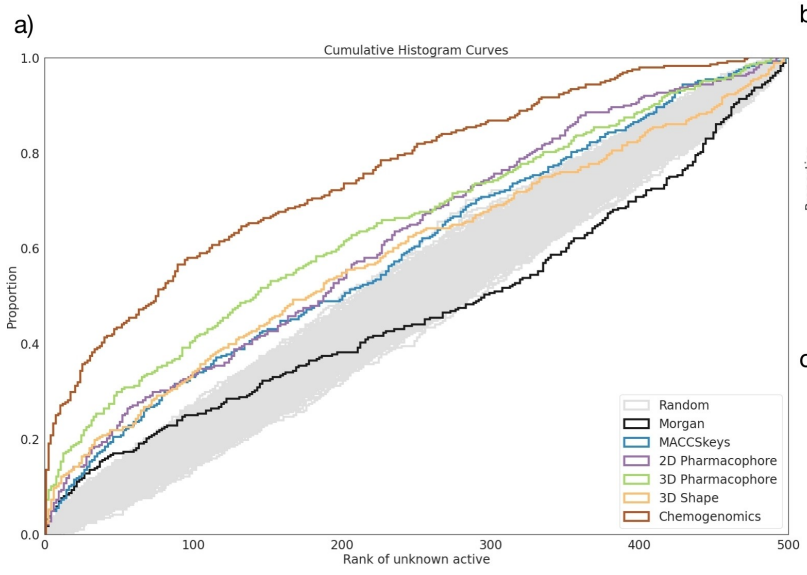


FIGURE 10 Illustration of the schemes followed by ligand-based methods and by the chemogenomics approach to solve a scaffold hopping case.

The performances of the chemogenomic algorithm are shown in Figure 11. It outperforms all tested ligand-based methods on all considered criteria: its CHC curve stands above all curves, at all ranks, leading to the highest AUC score, and to the best success rate at 5%. These performance improvements arise from the additional (*ligand, protein*) pair provided to the chemogenomic algorithm, besides the (*known active, query protein*) pair. Ligand-based methods cannot process this additional



c)

Method	Normalised AUC	Percentage in Top 5%
Average Random	0.50	5.1
Morgan	0.47	11.5
MACCSkeys	0.59	12.2
2D Pharmacophore	0.60	14.2
3D Shape	0.59	14.9
3D Pharmacophore	0.63	20.1
Chemogenomics	0.71	34.4

FIGURE 11 Results of the Large-Hops benchmark. The cumulative histogram curves of each method are plotted in a). A zoom of the same graph is provided in b). Table c) displays the Area Under the Curve (normalized i.e., divided by 500 to be between 0 and 1), and the percentage of scaffold hopping problems for which the unknown active was ranked in the top 5%.

information (nor would a medicinal chemist), and the ML chemogenomic algorithm provides a computational method to consider and profit from such otherwise accessible prior knowledge. Note that the performances inside families of proteins are heterogeneous: on average, the families' success rate is about 37.8%, and for the most represented one, the kinases, the success rate is 35.1%. This means that the method depends little on the family of the proteins. However, the general success rate of 34.4% still leaves space for improvements. In particular, the kernel Support Vector Machine (SVM) algorithm used in the present study should be better adjusted to the scaffold hopping problem. Indeed, as detailed in Materials and Methods, the SVM use a Tanimoto kernel for molecules that is calculated based on the molecular Morgan similarity [46]. Considering the results of the ligand-based methods in Figure 11c, a straightforward improvement would be to build a new topology-based kernel calculated from a 3D pharmacophore similarity measure.

3 | DISCUSSION

The scope of the present paper is essentially: (1) to propose a flowchart to cover the need for a publicly available and well-characterized large-step scaffold hopping benchmark for the community; (2) to provide a general assessment of the relative performances of classical 2D and 3D ligand-based methods for solving large-step scaffold hopping problems with 'hard' targets, in a setting that mimics real-life drug discovery applications [14].

To our knowledge, the $\mathcal{LH}$ benchmark is the first public high-quality benchmark dataset for large-step scaffold hopping. Starting from PDBbind, the proposed flowchart requires threshold values for various criteria. These thresholds were chosen in an expert-based manner to exclude irrelevant scaffold hopping cases. Some criteria enable the selection of pairs of highly different molecules, while others ensure that molecules in the same pair share similar binding modes, i.e., correspond to 'true' scaffold hopping cases. We use stringent thresholds for both types of criteria, because our goal is to build a high quality large-step scaffold hopping dataset. The resulting size for the $\mathcal{LH}$ benchmark is smaller than that reported for other less characterized benchmarks [12,13], but this illustrates that the number of large-step scaffold hopping cases reported is much smaller than that of small- to medium-step scaffold hops. Note however that available benchmarks are not comparable to the $\mathcal{LH}$ benchmark, because they were not conceived in a comparable setting. Should a large-step scaffold hopping benchmark of larger size be desired, the same flowchart could be followed with a lower drug-likeliness threshold, a larger range in molecular weights, or a higher threshold for redundancy between the pairs of molecules. Should an easier benchmark be designed, including medium-step hops, the thresholds in Murcko-based and molecular Morgan similarities could be increased. However, we advise not to relax the IFPs and MCS thresholds, to avoid selecting pairs of molecules that could correspond to 'false' scaffold hopping cases. An important contribution of the present study is to underline that building a reliable scaffold hopping benchmark must be a well-controlled multi-step process and cannot be achieved with the blind use of a few criteria. This important point, illustrated by the 'false' scaffold hopping examples shown in Figure 2, Figure 3 and Figure 4, has not been discussed in previous work reporting the construction of scaffold hopping benchmarks.

Based on the $\mathcal{LH}$ benchmark, all computational methods tested display modest performances, which confirms that solving large-step scaffold hopping for 'hard' targets is a difficult problem. This was expected, but our study allows to quantify how difficult these problems are, in general. Among the classical ligand-based methods that are tested, the 3D pharmacophore similarity performs best, on average, even when the active conformations are unknown, which is an interesting result. We are aware that many other 2D and 3D descriptors are available in the literature [47]. However, our goal was not to test all encoding methods, but to provide an overview of expected performances, and to globally rank the 2D, 3D, and chemogenomic approaches when they are run with classical and widely available descriptors, i.e., in settings that can be easily implemented by people in the community. Other promising topology-based descriptors have been recently proposed [11–13], and future work could be to evaluate their performance on the Large-Hops benchmark.

The chemogenomic algorithm leads to the best performances, although the kernel SVM algorithm can be improved. Because our benchmark contained drug-like molecules for proteins belonging to diverse families, we trained the chemogenomic algorithm based on a DrugBank-derived dataset. However, other larger training sets can be used, for example derived from larger databases such as PubChem. For more focused problems like scaffold hopping problems involving a protein belonging to a specific well studied family, such as kinases or GPCRs, one can also use other training databases that gather (*ligand, protein*) molecular interactions known within these families of proteins [48,49]. As an illustration, although chemogenomics has been hardly explored in the field of large-step scaffold hopping, this approach was

used in one study within the GPCR family, reporting identification of a new scaffold for an antagonist of Vasopressin 1 A [50]. This underlines the interest to further explore these strategies in the field of scaffold hopping.

Strategies based on descriptors that encode the bioactivity profiles of molecules have also been proposed [2,51–53]. This is an interesting idea, because it allows to abstract from the chemical structure and address scaffold hopping issues. However, some of these profiles are not publicly available, but descriptors based on public domain HTS studies [52] are interesting starting points to test their implementation in computational methods. In this context, we hope that the Large-Hops benchmark will be a convenient tool provided to the community, in order to test new strategies for the difficult but important problem of large-step scaffold hopping.

4 | MATERIALS AND METHODS

4.1 | Performance evaluation of computational methods

For all the considered computational methods, Cumulative Histogram Curves corresponding (CHC) to the rank of the unknown active molecules are plotted. The CHC curves of the most efficient methods stand above the others. The x-axis represents the rank, and the y-axis represents the proportion of cases (i.e., the proportion of scaffold hopping cases, among the $144 \times 2 = 288$ scaffold hopping problems in the Large-Hops benchmark) where the method recovers the unknown active at a rank below the x-axis value. For instance, for the chemogenomic approach, the unknown active was ranked in top 50 molecules for 45% of cases, as seen in Figure 11. Methods are also compared to random ranking: we perform one thousand random rankings for the unknown active for the 438 scaffold hopping problems. This leads to 1000 CHC curves plotted in grey in Figure 11.

The Area Under the Curve (AUC) score of the CHC curves quantifies the global performances of methods: the higher the better. It ranges from 0 (unknown actives are ranked 500 in all scaffold hopping problems) to 500 (unknown actives are ranked 1 in all scaffold hopping problems). In the Results section, normalized AUC are provided: dividing the AUC by 500 leads to values between 0 and 1. A random ranking corresponds to an AUC of 0.5, as illustrated with the grey curves in Figure 11.

For each method, we also compute the percentage of cases where the unknown active is ranked in the top 5%, i.e., in the first 25 molecules. This metric is complementary to the AUC, since it can be viewed as the percentage of successful cases.

4.2 | Building a dataset of large-step scaffold hops cases

4.2.1 | Selecting pairs of drug-like molecules

The molecules quantitative estimation of drug-likeness (QED) score [17] is calculated with RDKit [30]: all pairs with a QED below 0.5 are discarded. To discard redundant pairs, which differ by only slight molecular modifications, the Tanimoto similarities based on the Morgan fingerprints of the generic Murcko scaffolds and of the whole molecules are calculated. When two pairs of ligands have one of these Tanimoto coefficient above 0.5, only one pair is kept in the dataset.

4.2.2 | Selecting pairs of large-scale hops ligands

The RDKit package [30] was used to provide generic Murcko scaffold of molecules, where all atoms are replaced by carbons, all bonds are switched to single bonds and only the linkers between the rings are conserved. RDKit [30] is also used to compute Morgan fingerprints for the Murcko scaffolds and for the whole structures of molecules, and to calculate the resulting Tanimoto similarities. Pairs with Murcko and Morgan Tanimoto similarities respectively below 0.6 and 0.3 are selected.

4.2.3 | Selecting pairs with similar binding modes

Interaction fingerprints (IFPs) are used to encode the binding modes of ligands. These fingerprints are target-focused binary vectors that incorporate, for each protein residue in the binding site, its interactions with the ligand. Bits are allocated for each residue, each encoding for the presence of one type of interaction with the ligand. To construct those IFPs, we need to detect the protein-ligand interactions. Starting from PLIP [54], a freely available algorithm that detects such interactions,

TABLE 2 Table summarizing the criteria to detect protein-ligand interactions.

Interaction	Ligand	Protein	Distance	Angle
Hydrophobic	C, S, F, Cl	C, S	$d \leq 3.5$ Å	–
Pi-stacking	Aromatic ring	Aromatic ring	$d \leq 5.5$ Å offset ≤ 2.5 Å	$0 \leq \theta \leq 30$ -> parallel $60 \leq \theta \leq 90$ -> T-shaped $30 < \theta < 60$ -> face-to-face, face-to-edge
Pi-amide	Aromatic ring Polar C sp2	Polar C sp2 Aromatic ring	$d \leq 4.5$ Å offset ≤ 2.5 Å	$0 \leq \theta \leq 30$
Pi-cation	Aromatic ring Cation	Cation Aromatic ring	$d \leq 4.5$ Å offset ≤ 2.5 Å	–
Pi-hydrophobic	Aromatic ring C sp3, S sp3, F, Cl	C sp3, S sp3 Aromatic ring	$d \leq 4.5$ Å offset ≤ 2.0 Å	–
Multipolar	X-bond donor Polar C sp2	Polar C sp2 X-bond donor	$d_{X...Csp2} \leq 4.5$ Å	$\theta \geq 70$ where θ: angle between C–X and X...Csp2 $\theta \leq 60$ where θ: angle between X...Csp2 and normal to amide plan
H-bond	H-bond donor H-bond acceptor	H-bond acceptor H-bond donor	$d_{D...A} \leq 4.2$ Å	$\theta_{A...H-D} \geq 130$ $\theta_{R-A...D} \geq 90$ where R = A's neighbours
Weak H-bond	Weak H-bond donor Weak H-bond acceptor	Weak H-bond acceptor Weak H-bond donor	$d_{D...A} \leq 4.0$ Å	$\theta_{A...H-D} \geq 140$ $\theta_{R-A...D} \geq 90$ where R = A's neighbours
Salt bridge	Anion Cation	Cation Anion	$d_{cation...anion} \leq 5.5$ Å	–
Halogen bond	X-bond donor Polar C sp2	Polar C sp2 X-bond donor	$d_{X...A} \leq 4.5$ Å	$\theta_{AXD} \geq 120$ $\theta_{RAX} \geq 90$ where R = A's neighbours

including hydrogen bond, weak hydrogen bond, halogen bond, salt bridge, hydrophobic, pi-cation, and pi-stacking, we built an extended version adding several interactions [24–28] that are missed by classical IFPs: Pi-hydrophobic, Pi-amide and Multipolar. The detection criteria are described in Table 2. The thresholds used here are less restrictive than those of the original package as we want to avoid missing the detection because of lower resolution of some PDB structures.

Then, the similarity in binding mode between pairs of ligands for the same protein is calculated based on the Tanimoto similarities between the ligands IFPs. Pairs of ligands with IFPs similarities above 0.6 were selected.

We observe that, in a few cases, even though the two molecules of the pair bind similarly to the protein, the IFP similarity is 0, because in the corresponding PDB files, the residues have different numberings. In these cases, comparison of interacting residues between the two molecules is doomed to fail. We do not correct the numbering, because those cases are rare and case-specific, and automatic rectification of residues IDs may lead to mistakes.

4.2.4 | Discarding pairs based on Maximum Common Substructures

Each MCS type is matched on both ligands, and the ratio of the number of common interactions formed by the considered MCS and the total of common interactions is computed as following:

$$\boldsymbol{ratio}_{\boldsymbol{MCS\ interaction}} = \frac{|\boldsymbol{Interactions}_{\boldsymbol{Common}} \cap \boldsymbol{Interactions}_{\boldsymbol{MCS}}|}{|\boldsymbol{Interactions}_{\boldsymbol{Common}}|}$$

When several MCS matches are possible on a molecule, the match with the highest ratio is kept. The final ratio is defined as the highest of the ratios for all MCS types. Pairs with a final ratio above 0.8 are discarded.

4.3 | Choice of decoy molecules

To avoid statistical bias between molecules in the active pairs and their corresponding decoys, decoys are selected

from the ZINC database, among molecules with physical and chemical properties similar to those of the active molecules, as detailed below. The considered physical and chemical descriptors are:

- Number of hydrogen bond donor and acceptor
- Number of aromatic and aliphatic rings
- Number of consecutive rotatable bonds
- Molecular weight
- Lipophilicity
- Topological polar surface area

More precisely, molecules are selected if their physical and chemical descriptors fulfil the following criteria:

$$\boldsymbol{descriptor}_{\boldsymbol{molecule}} \in [\mathbf{min}(\boldsymbol{descriptor}_{\boldsymbol{ligands}}) - \boldsymbol{c}, \mathbf{max}(\boldsymbol{descriptor}_{\boldsymbol{ligands}}) + \boldsymbol{c}]$$

where $\boldsymbol{c} = 1$ for integer descriptors, and

$$\boldsymbol{c} = \frac{10}{100}\left|\boldsymbol{descriptor}_{\boldsymbol{ligand\ 1}} - \boldsymbol{descriptor}_{\boldsymbol{ligand\ 2}}\right|$$

for continuous descriptors.

As the decoys need to be realistic large-step scaffold hopping candidates, they are chosen at a Murcko-based Morgan similarity from the molecules in the active pairs below 0.6, since this threshold is used to select pairs of active molecules.

In addition, the decoys should not either be too distant from the ligands, in order to avoid analogous bias, and to mimic real-life screens for the search of scaffold hop candidates. Thus, the decoys selected from the ZINC also have to be as similar to the ligands as the ligands are similar to each other, according to their overall structure Morgan fingerprints:

$$\boldsymbol{similarity}_{\boldsymbol{molecule,ligands}} \in \left[\boldsymbol{similarity}_{\boldsymbol{ligand\ 1,ligand\ 2}} - \boldsymbol{c}^{'},\ \boldsymbol{similarity}_{\boldsymbol{ligand\ 1,ligand\ 2}} + \boldsymbol{c}^{'}\right]$$

where $\boldsymbol{c}^{'} = 0.15$ to have an interval of size 0.3 and capture enough decoy molecules (i.e., a number of 499 decoys), not too distant nor too close, with respect to the molecules in the active pair.

These criteria were successively applied to molecules in the ZINC database, and 499 decoys satisfying these criteria could be found for 144 pairs of active molecules.

4.4 | 2D Similarity-based methods

Three different 2D molecular representations are calculated with the RDKit package [30]:

- The classical Morgan fingerprints, that implements the ECFP extended connectivity fingerprint [19] with radius 4 as a 4096-bit binary vector.
- MACCS keys fingerprints, a binary 166-bit vector that encodes the presence of SMARTS-based strings in the molecular structure.
- 2D pharmacophore fingerprints calculated using the distance separating 2- and 3-point pharmacophores defined as SMARTS strings, in a planar representation of molecules.

For these three types of binary fingerprints, the similarity measure between two molecules is calculated based on their Tanimoto coefficient.

4.5 | 3D Similarity-based methods

We used two types of 3D similarity methods, based on shape or 3D pharmacophore representations of the molecules. In both cases, this first requires generating 3D molecular conformers. For all pairs of active molecules and their 499 decoys, a pool of 500 conformers is calculated, from which we keep up to ten conformers of local minimal energy that differ from a RMSD value of at least 1.5 Å, under MMFF94 force field using RDKit [55]. Then, all conformers of the known active are aligned pairwise with those of the unknown active or those of the decoy molecules, to maximize their overlap. For the 3D-pharmacophore similarity, for each pair of active molecules and their decoys, the freely available Pharao software [56] is used to detect the pharmacophore groups for conformers. The Tanimoto coefficient quantifying the overlap between aligned conformers of the known active and those of the unknown active or of the decoys is calculated pairwise. The largest Tanimoto 3D pharmacophore coefficient observed is used to define the 3D-pharmacophore similarity between the corresponding known active and the unknown active or the decoys. The same method is used for the shape similarity [56], where the largest Tanimoto shape coefficient observed between conformers is used to define the shape similarity between the known active and the unknown active or the decoys. Finally, for each pair of active

molecules, the unknown active and the decoys are ranked according to their Tanimoto 3D pharmacophore, or shape similarity.

4.6 | ML chemogenomic algorithm

4.6.1 | Kernel SVM

The chemogenomic approach used in the present study recasts the problem as a supervised learning binary classification over the space of pairs $(\boldsymbol{m}, \boldsymbol{p})$ of molecules and proteins, to separate binding pairs from a carefully selected set of non-binding pairs. We rely on a kernelized Support Vector Machines (SVM) classifier [57] to perform this classification. Briefly, the SVM is trained on a dataset of $(\boldsymbol{m}, \boldsymbol{p})$ pairs and learns the optimal hyperplane that separates pairs that bind from those that do not. The kernel SVM leverages a kernel $\mathcal{K}$ encoding similarities between $(\boldsymbol{m}, \boldsymbol{p})$ pairs [58]. A general method to build a kernel on $(\boldsymbol{m}, \boldsymbol{p})$ pairs is to use the Kronecker product of molecule and protein kernels as done in [58] and [59]. Given a molecule kernel $\mathcal{K}_{\boldsymbol{molecule}}$ and a protein kernel $\mathcal{K}_{\boldsymbol{protein}}$, the Kronecker kernel $\mathcal{K}_{\boldsymbol{pair}}$ is defined by:

$$\mathcal{K}_{\boldsymbol{pair}}((\boldsymbol{m}, \boldsymbol{p}), \ (\boldsymbol{m}', \boldsymbol{p}')) = \mathcal{K}_{\boldsymbol{molecule}}(\boldsymbol{m}, \boldsymbol{m}') \times \mathcal{K}_{\boldsymbol{protein}}(\boldsymbol{p}, \boldsymbol{p}')$$

where $\boldsymbol{m}$ and $\boldsymbol{m}'$ are molecules and $\boldsymbol{p}$ and $\boldsymbol{p}'$ are proteins. We choose the Local Alignment kernel for proteins [60] and the Tanimoto kernel with Morgan fingerprints for molecules [46], whose hyperparameters are validated by cross validation in [61]. The Local Alignment kernel for proteins sums up the contributions of all possible local alignments with gaps of the sequences which is efficient for detecting remote homology [60]. The Tanimoto kernel between two molecules is calculated as the Tanimoto similarity of their Morgan fingerprints [46]. Protein and molecular kernels are centered and normalized.

The SVM algorithm also requires a regularisation parameter classically called C, which controls the trade-off between maximising the margin (the distance separating the hyperplane and the two classes' distributions) and minimizing classification error on the training points. To implement this algorithm, we use the sklearn [62] function *SVC* with the parameter $\boldsymbol{C} = 10$ validated by cross validation in [45]. Once the SVM is trained, it can be applied to any pair $(\boldsymbol{m}, \boldsymbol{p})$ to give a binding probability. This probability is computed by applying a sigmoid function to the SVM outputs, where the parameters are trained by cross validation as explained in [63]. It is implemented in the *predict proba* method of *SVC*.

4.6.2 | Training dataset

To build our training set, we use the DrugBank v1.5.1 [64] which defines a set of $(\boldsymbol{m}, \boldsymbol{p})$ pairs which bind together (i.e. $\boldsymbol{m}$ targets $\boldsymbol{p}$). We choose this dataset because it is well curated and composed of FDA-approved drugs. We kept molecules with molecular weights between 100 and 800 g/mol which is in the range of drug-like molecules [65]. Among these molecules, we select those which target at least one human protein. Thus, the train dataset comprises 5.071 molecules, 2.670 proteins and 14.638 positive bindings. To complete the dataset, we need to select negative pairs. This selection should be designed with care to correct potential statistical bias in the database and reduce the number of false positive predictions. We use the greedy algorithm in [45], which randomly chooses the same number (14.638) of negative pairs so that each molecule and each protein have the same number of positive and negative pairs in the training dataset.

4.6.3 | Training scheme

The ML chemogenomic algorithm is trained for each of the 288 scaffold hopping cases in the $\mathcal{LH}$ benchmark as follows: one molecule of the pair is considered as the only known active for the query protein. If this pair is not already present in the DrugBank database, it is added to the training set. All other pairs involving the query protein that are present in DrugBank are removed from the training set. This allows us to exclude the $(\boldsymbol{m}, \boldsymbol{p})$ pair between the unknown active of the pair with the query protein if this pair is in DrugBank. Hence, for each pair of ligands in the $\mathcal{LH}$ benchmark, the chemogenomic algorithm is trained with the same information about ligands of the query protein than the ligand-based algorithms: a single known active ligand. Once trained, the algorithm predicts the binding probabilities of (*molecule, query protein*) pairs involving the 499 decoys and the unknown active molecule. In order to have a more robust score, this scheme is repeated 5 times for different sets of negative examples in the training set and the binding probabilities are averaged over these 5 versions. We observe that the variance across these repetitions is low (below 10^{-2}) which highlights the stability of the method. Finally, the unknown active molecule and the 499 decoys are ranked according to their averaged binding probabilities.

ACKNOWLEDGMENTS

The authors are grateful to Christopher Housseman, Vincent Mallet, Quentin Perron and Sree Vadlamudi for fruitful discussions about the project, carefully reading and providing us with constructive feedback about the manuscript. Financial support for Matthieu Najm was provided by Vaincre La Mucoviscidose.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data may be requested via the authors.

ORCID

Philippe Pinel http://orcid.org/0000-0002-6010-1853

REFERENCES

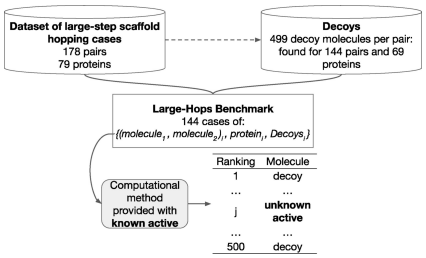
1. G. Schneider, W. Neidhart, T. Giller, G. Schmid, *Angewandte Chemie International Edition* **1999**, *38*, 2894–2896.
2. Y. Hu, D. Stumpfe, J. Bajorath, *Journal of Medicinal Chemistry* **2016**, *60*, 1238–1246.
3. H. Sun, G. Tawa, A. Wallqvist, *Drug Discovery Today* **2012**, *17*, 310–324.
4. E. N. Muratov, J. Bajorath, R. P. Sheridan, I. V. Tetko, D. Filimonov, V. Poroikov, T. I. Oprea, I. I. Baskin, A. Varnek, A. Roitberg, O. Isayev, S. Curtalolo, D. Fourches, Y. Cohen, A. Aspuru-Guzik, D. A. Winkler, D. Agrafiotis, A. Cherkasov, A. Tropsha, *Chemical Society Reviews* **2020**, *49*, 3525–3564.
5. J. B. Roque, Y. Kuroda, L. T. Göttemann, R. Sarpong, *Nature* **2018**, *564*, 244–248.
6. J. Pang, S. Gao, Z. Sun, G. Yang, *Struct Chem* **2021**, *32*, 879–886.
7. E. Gulezian, C. Crivello, J. Bednenko, C. Zafra, Y. Zhang, P. Colussi, S. Hussain, *Trends in Pharmacological Sciences* **2021**, *42*, 657–674.
8. A. Dick, S. Cocklin, *Pharmaceuticals* **2020**, *13*, 36.
9. A. Lovrics, V. F. S. Pape, D. Szisz, A. Kalászi, P. Heffeter, C. Magyar, G. Szakács, *Journal of Cheminformatics* **2019**, *11*, 67.
10. F. Grisoni, D. Merk, V. Consonni, J. A. Hiss, S. G. Tagliabue, R. Todeschini, G. Schneider, *Commun Chem* **2018**, *1*, 1–9.
11. H. Nakano, T. Miyao, J. Swarit, K. Funatsu, *Journal of Chemical Information and Modeling* **2021**, *61*, 3348–3360.
12. F. Grisoni, D. Merk, R. Byrne, G. Schneider, *Sci Rep* **2018**, *8*, 16469.
13. H. Nakano, T. Miyao, K. Funatsu, *J. Chem. Inf. Model.* **2020**, *60*, 2073–2081.
14. Bajorath, in *IB Chemistry Revision Guide*, Anthem Press, **2019**, pp. 222–238.
15. N. Lagarde, J.-F. Zagury, M. Montes, *J. Chem. Inf. Model.* **2015**, *55*, 1297–1307.
16. R. Wang, X. Fang, Y. Lu, S. Wang, *Journal of Medicinal Chemistry* **2004**, *47*, 2977–2980.
17. G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan, A. L. Hopkins, *Nature Chemistry* **2012**, *4*, 90–98.
18. G. W. Bemis, M. A. Murcko, *J. Med. Chem.* **1996**, *39*, 2887–2893.
19. D. Rogers, M. Hahn, *Journal of Chemical Information and Modeling* **2010**, *50*, 742–754.
20. G. Marcou, D. Rognan, *Journal of Chemical Information and Modeling* **2006**, *47*, 195–207.
21. V. Chupakhin, G. Marcou, H. Gaspar, A. Varnek, *Computational and Structural Biotechnology Journal* **2014**, *10*, 33–37.
22. C. Da, D. Kireev, *Journal of Chemical Information and Modeling* **2014**, *54*, 2555–2561.
23. S. Salentin, V. J. Haupt, S. Daminelli, M. Schroeder, *Progress in Biophysics and Molecular Biology* **2014**, *116*, 174–186.
24. C. Bissantz, B. Kuhn, M. Stahl, *Journal of Medicinal Chemistry* **2010**, *53*, 5061–5084.
25. R. F. de Freitas, M. Schapira, *MedChemComm* **2017**, *8*, 1970–1981.
26. N. K. Shinada, A. G. de Brevern, P. Schmidtke, *Journal of Medicinal Chemistry* **2019**, *62*, 9341–9356.
27. B. Kuhn, E. Gilberg, R. Taylor, J. Cole, O. Korb, *Journal of Medicinal Chemistry* **2019**, *62*, 10441–10455.
28. E. Nittinger, T. Inhester, S. Bietz, A. Meyder, K. T. Schomburg, G. Lange, R. Klein, M. Rarey, *Journal of Medicinal Chemistry* **2017**, *60*, 4245–4257.
29. E. J. Barker, D. Buttar, D. A. Cosgrove, E. J. Gardiner, P. Kitts, P. Willett, V. J. Gillet, *J. Chem. Inf. Model.* **2006**, *46*, 503–511.
30. G. Landrum, P. Tosco, B. Kelley, Ric, sriniker, gedeck, R. Vianello, NadineSchneider, A. Dalke, E. Kawashima, D. N, B. Cole, M. Swain, S. Turk, D. Cosgrove, AlexanderSavelyev, A. Vaucher, G. Jones, M. Wójcikowski, D. Probst, V. F. Scalfani, guillaume godin, A. Pahl, F. Berenger, JLVarjo, strets123, JP, DoliathGavid, G. Sforna, J. H. Jensen, **2021**, DOI 10.5281/zenodo.5242603.
31. R. H. Advani, R. T. Hoppe, D. Baer, J. Mason, R. Warnke, J. Allen, S. Daadi, S. A. Rosenberg, S. J. Horning, *Annals of Oncology* **2013**, *24*, 1044–1048.
32. A. G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia, *Journal of Molecular Biology* **1995**, *247*, 536–540.
33. M. Réau, F. Langenfeld, J.-F. Zagury, N. Lagarde, M. Montes, *Frontiers in Pharmacology* **2018**, *9*, DOI 10.3389/fphar.2018.00011.
34. A. C. Good, T. I. Oprea, *Journal of Computer-Aided Molecular Design* **2008**, *22*, 169–178.
35. J. J. Irwin, B. K. Shoichet, *Journal of Chemical Information and Modeling* **2004**, *45*, 177–182.
36. D. Stumpfe, J. Bajorath, in *Methods and Principles in Medicinal Chemistry*, Wiley-VCH Verlag GmbH & Co. KGaA, **2011**, pp. 291–318.
37. S. M. Vogel, M. R. Bauer, F. M. Boeckler, *Journal of Chemical Information and Modeling* **2011**, *51*, 2650–2665.

38. M. R. Bauer, T. M. Ibrahim, S. M. Vogel, F. M. Boeckler, *Journal of Chemical Information and Modeling* **2013**, *53*, 1447–1462.
39. A. M. Helguera, R. D. Combes, M. P. Gonzalez, M. N. D. S. Cordeiro, *Current Topics in Medicinal Chemistry* **2008**, *8*, 1628–1655.
40. J. L. Durant, B. A. Leland, D. R. Henry, J. G. Nourse, *Journal of Chemical Information and Computer Sciences* **2002**, *42*, 1273–1280.
41. T. S. Rush, J. A. Grant, L. Mosyak, A. Nicholls, *J. Med. Chem.* **2005**, *48*, 1489–1495.
42. G. Hessler, K.-H. Baringhaus, *Drug Discovery Today: Technologies* **2010**, *7*, e263–e269.
43. L. Jacob, B. Hoffmann, V. Stoven, J.-P. Vert, *BMC Bioinformatics* **2008**, *9*, 363.
44. E. E. Bolton, Y. Wang, P. A. Thiessen, S. H. Bryant, in *Annual Reports in Computational Chemistry* (Eds.: R. A. Wheeler, D. C. Spellmeyer), Elsevier, **2008**, pp. 217–241.
45. M. Najm, C.-A. Azencott, B. Playe, V. Stoven, *International Journal of Molecular Sciences* **2021**, *22*, 5118.
46. S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, P. Baldi, *Bioinformatics* **2005**, *21*, i359–i368.
47. S. Kausar, A. O. Falcao, *Molecules* **2019**, *24*, 1698.
48. F. Carles, S. Bourg, C. Meyer, P. Bonnet, *Molecules* **2018**, *23*, 908.
49. Y. Okuno, J. Yang, K. Taneishi, H. Yabuuchi, G. Tsujimoto, *Nucleic Acids Research* **2006**, *34*, D673–D677.
50. H. Ratni, M. Rogers-Evans, C. Bissantz, C. Grundschober, J.-L. Moreau, F. Schuler, H. Fischer, R. Alvarez Sanchez, P. Schneider, *J. Med. Chem.* **2015**, *58*, 2275–2289.
51. P. M. Petrone, B. Simms, F. Nigsch, E. Lounkine, P. Kutchukian, A. Cornett, Z. Deng, J. W. Davies, J. L. Jenkins, M. Glick, *ACS Chem. Biol.* **2012**, *7*, 1399–1409.
52. K. Y. Helal, M. Maciejewski, E. Gregori-Puigjané, M. Glick, A. M. Wassermann, *J. Chem. Inf. Model.* **2016**, *56*, 390–398.
53. G.-L. Xiong, Y. Zhao, L. Liu, Z.-Y. Ma, A.-P. Lu, Y. Cheng, T.-J. Hou, D.-S. Cao, *J. Med. Chem.* **2021**, *64*, 7544–7554.
54. S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, M. Schroeder, *Nucleic Acids Research* **2015**, *43*, W443–W447.
55. P. Tosco, N. Stiefl, G. Landrum, *J Cheminform* **2014**, *6*, 37.
56. J. Taminau, G. Thijs, H. D. Winter, *Journal of Molecular Graphics and Modelling* **2008**, *27*, 161–169.
57. C. Cortes, V. Vapnik, *Mach Learn* **1995**, *20*, 273–297.
58. B. Schölkopf, K. Tsuda, J.-P. Vert, *Kernel Methods in Computational Biology*, MIT Press, **2004**.
59. J.-P. Vert, L. Jacob, *Combinatorial Chemistry & High Throughput Screening* **2008**, *11*, 677–685.
60. H. Saigo, J.-P. Vert, N. Ueda, T. Akutsu, *Bioinformatics* **2004**, *20*, 1682–1689.
61. B. Playe, C.-A. Azencott, V. Stoven, *PLOS ONE* **2018**, *13*, e0204999.
62. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, *MACHINE LEARNING IN PYTHON* **n.d.**, 6.
63. J. Platt, *Advances in large margin classifiers* **1999**, *10*, 61–74.
64. D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, *Nucleic Acids Research* **2018**, *46*, D1074–D1082.
65. C. A. Lipinski, *Journal of Pharmacological and Toxicological Methods* **2000**, *44*, 235–249.

How to cite this article: P. Pinel, G. Guichaoua, M. Najm, S. Labouille, N. Drizard, Y. Gaston-Mathé, B. Hoffmann, V. Stoven, *Molecular Informatics* **2023**, *42*, e202200216. https://doi.org/10.1002/minf.202200216

Graphical Abstract

The contents of this page will be used as part of the graphical abstract of html only.
It will not be published as part of main.



Bibliography

[Adeshina, 2020] Yusuf O. Adeshina, Eric J. Deeds, and John Karanicolas. « Machine learning classification can reduce false positives in structure-based virtual screening ». *Proceedings of the National Academy of Sciences* 117.31 (2020), pp. 18477–18488 (cit. on p. 122).

[Alvarez, 2016] Mariano J. Alvarez, Yao Shen, Federico M. Giorgi, Alexander Lachmann, B. Belinda Ding, B. Hilda Ye, et al. « Functional characterization of somatic mutations in cancer using network-based inference of protein activity ». *Nature Genetics* 48.8 (2016), pp. 838–847 (cit. on p. 97).

[Amaral, 2007] Margarida D. Amaral and Karl Kunzelmann. « Molecular targeting of CFTR as a therapeutic approach to cystic fibrosis ». *Trends in Pharmacological Sciences* 28.7 (2007), pp. 334–341 (cit. on p. 10).

[Ashburner, 2000] Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. « Gene Ontology: tool for the unification of biology ». *Nature Genetics* 25.1 (2000), pp. 25–29 (cit. on p. 29).

[Baglama, 2005] James Baglama and Lothar Reichel. « Augmented Implicitly Restarted Lanczos Bidiagonalization Methods ». *SIAM Journal on Scientific Computing* 27.1 (2005), pp. 19–42 (cit. on p. 40).

[Balloy, 2015] Viviane Balloy, Hugo Varet, Marie-Agnès Dillies, Caroline Proux, Bernd Jagla, Jean-Yves Coppée, et al. « Normal and Cystic Fibrosis Human Bronchial Epithelial Cells Infected with Pseudomonas aeruginosa Exhibit Distinct Gene Activation Patterns ». *PLOS ONE* 10.10 (2015), e0140979 (cit. on pp. 55, 71).

[Balough, 1995] K. Balough, M. McCubbin, M. Weinberger, W. Smits, R. Ahrens, and R. Fick. « The relationship between infection and inflammation in the early stages of lung disease from cystic fibrosis ». *Pediatric Pulmonology* 20.2 (1995), pp. 63–70 (cit. on p. 12).

[Bampi, 2020] Giovana B. Bampi, Robert Rauscher, Sebastian Kirchner, Kathryn E. Oliver, Marcel J. C. Bijvelds, Leonardo A. Santos, et al. « Global assessment of the integrated stress response in CF patient-derived airway and intestinal tissues ». *Journal of Cystic Fibrosis* 19.6 (2020), pp. 1021–1026 (cit. on pp. 55, 89, 142).

[Barabasi, 2004] Albert-Laszlo Barabasi and Zoltan N. Oltvai. « Network biology: understanding the cell's functional organization ». *Nature reviews genetics* 5.2 (2004), pp. 101–113 (cit. on p. 25).

[Barbie, 2009] David A. Barbie, Pablo Tamayo, Jesse S. Boehm, So Young Kim, Susan E. Moody, Ian F. Dunn, et al. « Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1 ». *Nature* 462.7269 (2009), pp. 108–112 (cit. on p. 36).

[Bardin, 2018] Pauline Bardin, Emmeline Marchal-Duval, Florence Sonneville, Sabine Blouquit-Laye, Nathalie Rousselet, Philippe Le Rouzic, et al. « Small RNA and transcriptome sequencing reveal the role of miR-199a-3p in inflammatory processes in cystic fibrosis airways: miR-199a-3p regulates the NF-ϰB pathway in the lungs of CF patients ». *The Journal of Pathology* 245.4 (2018), pp. 410–420 (cit. on p. 55).

[Barillot, 2012] Emmanuel Barillot, Laurence Calzone, Philippe Hupe, Jean-Philippe Vert, and Andrei Zinovyev. *Computational Systems Biology of Cancer*. CRC Press, 2012 (cit. on pp. 25, 27).

[Bartlett, 2016] Jennifer A. Bartlett, Shyam Ramachandran, Christine L. Wohlford-Lenane, Carrie K. Barker, Alejandro A. Pezzulo, Joseph Zabner, et al. « Newborn Cystic Fibrosis Pigs Have a Blunted Early Response to an Inflammatory Stimulus ». *American Journal of Respiratory and Critical Care Medicine* 194.7 (2016), pp. 845–854 (cit. on p. 13).

[Battiston, 2014] Federico Battiston, Vincenzo Nicosia, and Vito Latora. « Structural measures for multiplex networks ». *Physical Review E* 89.3 (2014), p. 032804 (cit. on p. 96).

[Béal, 2021] Jonas Béal, Lorenzo Pantolini, Vincent Noël, Emmanuel Barillot, and Laurence Calzone. « Personalized logical models to investigate cancer response to BRAF treatments in melanomas and colorectal cancers ». *PLOS Computational Biology* 17.1 (2021), e1007900 (cit. on p. 142).

[Bell, 2020] Scott C. Bell, Marcus A. Mall, Hector Gutierrez, Milan Macek, Susan Madge, Jane C. Davies, et al. « The Lancet Respiratory Medicine Commission on the Future of Care of Cystic Fibrosis ». *The Lancet. Respiratory medicine* 8.1 (2020), pp. 65–124 (cit. on p. 86).

[Bérubé, 2010] Julie Bérubé, Lucie Roussel, Leila Nattagh, and Simon Rousseau. « Loss of Cystic Fibrosis Transmembrane Conductance Regulator Function Enhances Activation of p38 and ERK MAPKs, Increasing Interleukin-6 Synthesis in Airway Epithelial Cells Exposed to Pseudomonas aeruginosa ». *Journal of Biological Chemistry* 285.29 (2010), pp. 22299–22307 (cit. on pp. 81, 133).

[Bezzerri, 2011] Valentino Bezzerri, Pio d'Adamo, Alessandro Rimessi, Carmen Lanzara, Sergio Crovella, Elena Nicolis, et al. « Phospholipase C-β3 Is a Key Modulator of IL-8 Expression in Cystic Fibrosis Bronchial Epithelial Cells ». *The Journal of Immunology* 186.8 (2011), pp. 4946–4958 (cit. on pp. 84, 86).

[Bild, 2006] Andrea H Bild, Guang Yao, Jeffrey T Chang, Quanli Wang, Anil Potti, Dawn Chasse, et al. « Oncogenic pathway signatures in human cancers as a guide to targeted therapies ». *Nature* 439.7074 (2006), pp. 353–357 (cit. on p. 36).

[Bodas, 2010] M. Bodas and N. Vij. « The NF-kappaB signaling in cystic fibrosis lung disease: pathophysiology and therapeutic potential. » *Discovery medicine* 9.47 (2010), pp. 346–356 (cit. on p. 12).

[Bodas, 2019] Manish Bodas and Neeraj Vij. « Adapting Proteostasis and Autophagy for Controlling the Pathogenesis of Cystic Fibrosis Lung Disease ». *Frontiers in Pharmacology* 10 (2019) (cit. on p. 13).

[Bolton, 2011] Evan E. Bolton, Jie Chen, Sunghwan Kim, Lianyi Han, Siqian He, Wenyao Shi, et al. « PubChem3D: a new resource for scientists ». *Journal of Cheminformatics* 3.1 (2011), p. 32 (cit. on p. 134).

[Bolton, 2008] Evan E. Bolton, Yanli Wang, Paul A. Thiessen, and Stephen H. Bryant. « PubChem: Integrated Platform of Small Molecules and Biological Activities ». *Annual Reports in Computational Chemistry*. Vol. 4. Elsevier, 2008, pp. 217–241 (cit. on pp. 108, 115).

[Borisov, 2014] Nikolay M. Borisov, Nadezhda V. Terekhanova, Alexander M. Aliper, Larisa S. Venkova, Philip Yu Smirnov, Sergey Roumiantsev, et al. « Signaling pathways activation profiles make better markers of cancer than expression of individual genes ». *Oncotarget* 5.20 (2014), pp. 10198–10205 (cit. on pp. 35, 36).

[Bozoky, 2013] Zoltan Bozoky, Mickael Krzeminski, Ranjith Muhandiram, James R. Birtley, Ateeq Al-Zahrani, Philip J. Thomas, et al. « Regulatory R region of the CFTR chloride channel is a dynamic integrator of phospho-dependent intra- and intermolecular interactions ». *Proceedings of the National Academy of Sciences* 110.47 (2013), E4427–E4436 (cit. on p. 7).

[Braccia, 2019] Clarissa Braccia, Valeria Tomati, Emanuela Caci, Nicoletta Pedemonte, and Andrea Armirotti. « SWATH label-free proteomics for cystic fibrosis research ». *Journal of Cystic Fibrosis* 18.4 (2019), pp. 501–506 (cit. on p. 58).

[Bradley, 2017] Glyn Bradley and Steven J Barrett. « CausalR: extracting mechanistic sense from genome scale data ». *Bioinformatics* 33.22 (2017), pp. 3670–3672 (cit. on p. 97).

[Breiman, 2001] Leo Breiman. « Random Forests ». *Machine Learning* 45.1 (2001), pp. 5–32 (cit. on p. 121).

[Breuer, 2013] Karin Breuer, Amir K. Foroushani, Matthew R. Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, et al. « InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation ». *Nucleic Acids Research* 41.D1 (2013), pp. D1228–D1233 (cit. on p. 60).

[Brodlie, 2015] Malcolm Brodlie, Iram J. Haq, Katie Roberts, and J. Stuart Elborn. « Targeted therapies to improve CFTR function in cystic fibrosis ». *Genome Medicine* 7.1 (2015), p. 101 (cit. on p. 9).

[Buccitelli, 2020] Christopher Buccitelli and Matthias Selbach. « mRNAs, proteins and the emerging principles of gene expression control ». *Nature Reviews Genetics* 21.10 (2020), pp. 630–644 (cit. on p. 95).

[Bucki, 2015] Robert Bucki, Katrina Cruz, Katarzyna Pogoda, Ashley Eggert, LiKang Chin, Marianne Ferrin, et al. « Enhancement of Pulmozyme activity in purulent sputum by combination with poly-aspartic acid or gelsolin ». *Journal of Cystic Fibrosis* 14.5 (2015), pp. 587–593 (cit. on p. 46).

[Burat, 2022] Bastien Burat, Audrey Reynaerts, Dominique Baiwir, Maximilien Fléron, Sophie Gohy, Gauthier Eppe, et al. « Sweat Proteomics in Cystic Fibrosis: Discovering Companion Biomarkers for Precision Medicine and Therapeutic Development ». *Cells* 11.15 (2022), p. 2358 (cit. on p. 81).

[Burgel, 2023] Pierre-Régis Burgel, Isabelle Sermet-Gaudelus, Isabelle Durieu, Reem Kanaan, Julie Macey, Dominique Grenet, et al. « The French Compassionate Program of elexacaftor-tezacaftor-ivacaftor in people with cystic fibrosis with advanced lung disease and no F508del CFTR variant ». *European Respiratory Journal* (2023) (cit. on pp. 11, 141).

[Canato, 2018] Sara Canato, João D. Santos, Ana S. Carvalho, Kerman Aloria, Margarida D. Amaral, Rune Matthiesen, et al. « Proteomic interaction profiling reveals KIFC1 as a factor involved in early targeting of F508del-CFTR to degradation ». *Cellular and Molecular Life Sciences* 75.24 (2018), pp. 4495–4509 (cit. on p. 58).

[Cannon, 2003] Carolyn L. Cannon, Michael P. Kowalski, Kimberly S. Stopak, and Gerald B. Pier. « Pseudomonas aeruginosa–Induced Apoptosis Is Defective in Respiratory Epithelial Cells Expressing Mutant Cystic Fibrosis Transmembrane Conductance Regulator ». *American Journal of Respiratory Cell and Molecular Biology* 29.2 (2003), pp. 188–197 (cit. on p. 81).

[Cantiello, 1996] Hf Cantiello. « Role of the actin cytoskeleton in the regulation of the cystic fibrosis transmembrane conductance regulator ». *Experimental Physiology* 81.3 (1996), pp. 505–514 (cit. on pp. 45, 48).

[Cantini, 2017] Laura Cantini, Laurence Calzone, Loredana Martignetti, Mattias Rydenfelt, Nils Blüthgen, Emmanuel Barillot, et al. « Classification of gene signatures for their information value and functional redundancy ». *npj Systems Biology and Applications* 4.1 (2017), pp. 1–11 (cit. on p. 30).

[Cao, 2014] Dong-Sheng Cao, Liu-Xia Zhang, Gui-Shan Tan, Zheng Xiang, Wen-Bin Zeng, Qing-Song Xu, et al. « Computational Prediction of DrugTarget Interactions Using Chemical, Biological, and Network Features ». *Molecular Informatics* 33.10 (2014), pp. 669–681 (cit. on p. 121).

[Carraro, 2021] Gianni Carraro, Justin Langerman, Shan Sabri, Zareeb Lorenzana, Arunima Purkayastha, Guangzhu Zhang, et al. « Transcriptional analysis of cystic fibrosis airways at single-cell resolution reveals altered epithelial cell states and composition ». *Nature Medicine* 27.5 (2021), pp. 806–814 (cit. on pp. 57, 88, 143).

[Castellani, 2012] Stefano Castellani, Lorenzo Guerra, Maria Favia, Sante Di Gioia, Valeria Casavola, and Massimo Conese. « NHERF1 and CFTR restore tight junction organisation and function in cystic fibrosis airway epithelial cells: role of ezrin and the RhoA/ROCK pathway ». *Laboratory Investigation* 92.11 (2012), pp. 1527–1540 (cit. on p. 81).

[Cawley, 2010] Gavin C. Cawley and Nicola L.C. Talbot. « On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation ». *The Journal of Machine Learning Research* 11 (2010), pp. 2079–2107 (cit. on p. 119).

[Chang, 2018] Sheng-Wei Chang, Jack Wellmerling, Xiaoli Zhang, Rachael E. Rayner, Wissam Osman, Sara Mertz, et al. « The psychoactive substance of cannabis Δ9-tetrahydrocannabinol (THC) negatively regulates CFTR in airway cells ». *Biochimica et Biophysica Acta (BBA) - General Subjects* 1862.9 (2018), pp. 1988–1994 (cit. on p. 141).

[Chen, 2018] Qiwei Chen, Sudha Priya Soundara Pandi, Lauren Kerrigan, Noel G. McElvaney, Catherine M. Greene, J. Stuart Elborn, et al. « Cystic fibrosis epithelial cells are primed for apoptosis as a result of increased Fas (CD95) ». *Journal of Cystic Fibrosis* 17.5 (2018), pp. 616–623 (cit. on p. 81).

[Chesné, 2014] Julie Chesné, Richard Danger, Karine Botturi, Martine Reynaud-Gaubert, Sacha Mussot, Marc Stern, et al. « Systematic Analysis of Blood Cell Transcriptome in End-Stage Chronic Respiratory Diseases ». *PLOS ONE* 9.10 (2014), e109291 (cit. on p. 55).

[Choi, 2021] Soon H. Choi and John F. Engelhardt. « Gene Therapy for Cystic Fibrosis: Lessons Learned and Paths Forward ». *Molecular Therapy* 29.2 (2021), pp. 428–430 (cit. on p. 10).

[Ciavardelli, 2013] Domenico Ciavardelli, Melania D'Orazio, Luisa Pieroni, Ada Consalvo, Claudia Rossi, Paolo Sacchetta, et al. « Proteomic and ionomic profiling reveals significant alterations of protein expression and calcium homeostasis in cystic fibrosis cells ». *Molecular BioSystems* 9.6 (2013), pp. 1117–1126 (cit. on p. 58).

[Clancy, 2012] J. P. Clancy, Steven M. Rowe, Frank J. Accurso, Moira L. Aitken, Raouf S. Amin, Melissa A. Ashlock, et al. « Results of a phase IIa study of VX-809, an investigational CFTR corrector compound, in subjects with cystic fibrosis homozygous for the F508del-CFTR mutation ». *Thorax* 67.1 (2012), pp. 12–18 (cit. on p. 11).

[Clarke, 2013] Luka A. Clarke, Lisete Sousa, Celeste Barreto, and Margarida D. Amaral. « Changes in transcriptome of native nasal epithelium expressing F508del-CFTR and intersecting data from comparable studies ». *Respiratory Research* 14 (2013), p. 38 (cit. on pp. 55, 65, 71, 88, 96).

[Conese, 2021] Massimo Conese and Sante Di Gioia. « Pathophysiology of Lung Disease and Wound Repair in Cystic Fibrosis ». *Pathophysiology* 28.1 (2021), pp. 155–188 (cit. on p. 67).

[Consortium, 1993] The Cystic Fibrosis Genotype-Phenotype Consortium. « Correlation between Genotype and Phenotype in Patients with Cystic Fibrosis ». *New England Journal of Medicine* 329.18 (1993), pp. 1308–1313 (cit. on p. 13).

[Cooper, 2023] Nichola Cooper, Waleed Ghanima, Quentin A Hill, Phillip LR Nicolson, Vadim Markovtsov, and Craig Kessler. « Recent advances in understanding spleen tyrosine kinase (SYK) in human biology and disease, with a focus on fostamatinib ». *Platelets* 34.1 (2023), p. 2131751 (cit. on p. 86).

[Corey, 1997] Mary Corey, Lloyd Edwards, Henry Levison, and Michael Knowles. « Longitudinal analysis of pulmonary function decline in patients with cystic fibrosis ». *The Journal of Pediatrics* 131.6 (1997), pp. 809–814 (cit. on p. 10).

[Cornet, 2022a] Matthieu Cornet. « Etude informatique des effets cliniques et omiques des modulateurs de CFTR dans la mucoviscidose et recherche de nouvelles cibles ». PhD thesis. Université Paris sciences et lettres, 2022 (cit. on pp. 6, 8, 10).

[Cornet, 2022b] Matthieu Cornet, Geneviève Robin, Fabiana Ciciriello, Tiphaine Bihouee, Christophe Marguet, Valérie Roy, et al. « Profiling the response to lumacaftor-ivacaftor in children with cystic between fibrosis and new insight from a French-Italian real-life cohort ». *Pediatric Pulmonology* n/a.n/a (2022) (cit. on pp. 10, 12, 51, 142).

[Cortes, 1995] Corinna Cortes and Vladimir Vapnik. « Support-vector networks ». *Machine Learning* 20.3 (1995), pp. 273–297 (cit. on p. 118).

[Crites, 2015] Karoline St-Martin Crites, Geneviève Morin, Valérie Orlando, Natacha Patey, Catherine Cantin, Judith Martel, et al. « CFTR Knockdown induces proinflammatory changes in intestinal epithelial cells ». *Journal of Inflammation* 12.1 (2015), p. 62 (cit. on pp. 13, 67).

[Csabai, 2022] Luca Csabai, Dávid Fazekas, Tamás Kadlecsik, Máté Szalay-Bekő, Balázs Bohár, Matthew Madgwick, et al. « SignaLink3: a multi-layered resource to uncover tissue-specific signaling networks ». *Nucleic Acids Research* 50.D1 (2022), pp. D701–D709 (cit. on p. 23).

[Csardi, 2005] Gabor Csardi and Tamas Nepusz. « The Igraph Software Package for Complex Network Research ». *InterJournal* Complex Systems (2005), p. 1695 (cit. on p. 92).

[Curutiu, 2018] Carmen Curutiu, Florin Iordache, Veronica Lazar, Aurelia Magdalena Pisoschi, Aneta Pop, Mariana Carmen Chifiriuc, et al. « Impact of Pseudomonas aeruginosa quorum sensing signaling molecules on adhesion and inflammatory markers in endothelial cells ». *Beilstein Journal of Organic Chemistry* 14 (2018), pp. 2580–2588 (cit. on p. 75).

[David, 2023] Sharoon David and Christopher W. Edwards. « Forced Expiratory Volume ». *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2023 (cit. on p. 10).

[Davies, 2007] Jane C. Davies, Eric W. F. W. Alton, and Andrew Bush. « Cystic fibrosis ». *BMJ* 335.7632 (2007), pp. 1255–1259 (cit. on p. 10).

[De Lisle, 2014] Robert C. De Lisle. « Disrupted tight junctions in the small intestine of cystic fibrosis mice ». *Cell and Tissue Research* 355.1 (2014), pp. 131–142 (cit. on p. 81).

[del-Toro, 2013] Noemi del-Toro, Marine Dumousseau, Sandra Orchard, Rafael C. Jimenez, Eugenia Galeota, Guillaume Launay, et al. « A new reference implementation of the PSICQUIC web service ». *Nucleic Acids Research* 41.W1 (2013), W601–W606 (cit. on p. 96).

[Della Sala, 2021] Angela Della Sala, Giulia Prono, Emilio Hirsch, and Alessandra Ghigo. « Role of Protein Kinase A-Mediated Phosphorylation in CFTR Channel Activity Regulation ». *Frontiers in Physiology* 12 (2021), p. 690247 (cit. on p. 7).

[Deprez, 2020] Marie Deprez, Laure-Emmanuelle Zaragosi, Marin Truchi, Christophe Becavin, Sandra Ruiz García, Marie-Jeanne Arguel, et al. « A Single-Cell Atlas of the Human Healthy Airways ». *American Journal of Respiratory and Critical Care Medicine* 202.12 (2020), pp. 1636–1645 (cit. on p. 57).

[Di Pietro, 2017] Caterina Di Pietro, Ping-xia Zhang, Timothy K. O'Rourke, Thomas S. Murray, Lin Wang, Clemente J. Britto, et al. « Ezrin links CFTR to TLR4 signaling to orchestrate anti-bacterial immune response in macrophages ». *Scientific Reports* 7.1 (2017), p. 10882 (cit. on p. 101).

[Di Santagnese, 1953] P. A. Di Sant'agnese, R. C. Darling, G. A. Perera, and E. Shea. « Abnormal electrolyte composition of sweat in cystic fibrosis of the pancreas; clinical significance and relationship to the disease ». *Pediatrics* 12.5 (1953), pp. 549–563 (cit. on p. 6).

[Ding, 2022] Ning Ding, Pibao Li, Huiqing Li, Yunlong Lei, and Zengzhen Zhang. « The ROCK-ezrin signaling pathway mediates LPS-induced cytokine production in pulmonary alveolar epithelial cells ». *Cell Communication and Signaling* 20.1 (2022), p. 65 (cit. on p. 101).

[Donaldson, 2018] Scott H. Donaldson, Joseph M. Pilewski, Matthias Griese, Jon Cooke, Lakshmi Viswanathan, Elizabeth Tullis, et al. « Tezacaftor/Ivacaftor in Subjects with Cystic Fibrosis and F508del/F508del-CFTR or F508del/G551D-CFTR ». *American Journal of Respiratory and Critical Care Medicine* 197.2 (2018), pp. 214–224 (cit. on pp. 11, 12).

[Dorfman, 2008] Ruslan Dorfman, Andrew Sandford, Chelsea Taylor, Baisong Huang, Daisy Frangolias, Yongqian Wang, et al. « Complex two-gene modulation of lung disease severity in children with cystic fibrosis ». *The Journal of Clinical Investigation* 118.3 (2008), pp. 1040–1049 (cit. on p. 87).

[Drew, 2021] Kevin Drew, John B Wallingford, and Edward M Marcotte. « hu.MAP 2.0: integration of over 15,000 proteomic experiments builds a global compendium of human multiprotein assemblies ». *Molecular Systems Biology* 17.5 (2021), e10016 (cit. on p. 91).

[Duan, 2021] Yuanyuan Duan, Guangqiang Li, Miaomiao Xu, Xiaofei Qi, Mingxia Deng, Xuejia Lin, et al. « CFTR is a negative regulator of γδ T cell IFN-γ production and antitumor immunity ». *Cellular & Molecular Immunology* 18.8 (2021), pp. 1934–1944 (cit. on pp. 16, 88, 89).

[Dugger, 2020] Daniel T. Dugger, Monica Fung, Lorna Zlock, Saharai Caldera, Louis Sharp, Steven R. Hays, et al. « Cystic Fibrosis Lung Transplant Recipients Have Suppressed Airway Interferon Responses during Pseudomonas Infection ». *Cell Reports Medicine* 1.4 (2020), p. 100055 (cit. on p. 81).

[Dumortier, 2021] Claire Dumortier, Soula Danopoulos, Frédéric Velard, and Denise Al Alam. « Bone Cells Differentiation: How CFTR Mutations May Rule the Game of Stem Cells Commitment? » *Frontiers in Cell and Developmental Biology* 9 (2021) (cit. on p. 75).

[Durinck, 2009] Steffen Durinck, Paul T. Spellman, Ewan Birney, and Wolfgang Huber. « Mapping Identifiers for the Integration of Genomic Datasets with the R/Bioconductor package biomaRt ». *Nature protocols* 4.8 (2009), pp. 1184–1191 (cit. on p. 91).

[Durón, 2019] Christina Durón, Yuan Pan, David H. Gutmann, Johanna Hardin, and Ami Radunskaya. « Variability of Betweenness Centrality and Its Effect on Identifying Essential Genes ». *Bulletin of Mathematical Biology* 81.9 (2019), pp. 3655–3673 (cit. on p. 83).

[Egan, 1992] Marie Egan, Terence Flotte, Sandra Afione, Rikki Solow, Pamela L. Zeitlin, Barrie J. Carter, et al. « Defective regulation of outwardly rectifying Cl channels by protein kinase A corrected by insertion of CFTR ». *Nature* 358.6387 (1992), pp. 581–584 (cit. on p. 80).

[Erhan, 2006] Dumitru Erhan, Pierre-Jean L'Heureux, Shi Yi Yue, and Yoshua Bengio. « Collaborative Filtering on a Family of Biological Targets ». *Journal of Chemical Information and Modeling* 46.2 (2006), pp. 626–635 (cit. on p. 118).

[Esther, 2019] Charles R. Esther, Marianne S. Muhlebach, Camille Ehre, David B. Hill, Matthew C. Wolfgang, Mehmet Kesimer, et al. « Mucus accumulation in the lungs precedes structural changes and infection in children with cystic fibrosis ». *Science Translational Medicine* 11.486 (2019), eaav3488 (cit. on p. 13).

[Fan, 2016] Jean Fan, Neeraj Salathia, Rui Liu, Gwendolyn E. Kaeser, Yun C. Yung, Joseph L. Herman, et al. « Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis ». *Nature Methods* 13.3 (2016), pp. 241–244 (cit. on p. 36).

[Faria Poloni, 2021] Joice de Faria Poloni, Thaiane Rispoli, Maria Lucia Rossetti, Cristiano Trindade, and José Eduardo Vargas. « Cystic Fibrosis: Systems Biology Analysis from Homozygous p.Phe508del Variant Patients' Samples Reveals Perturbations in Tissue-Specific Pathways ». *BioMed Research International* 2021 (2021), e5262000 (cit. on p. 60).

[Farinha, 2017] Carlos M. Farinha and Sara Canato. « From the endoplasmic reticulum to the plasma membrane: mechanisms of CFTR folding and trafficking ». *Cellular and Molecular Life Sciences* 74.1 (2017), pp. 39–55 (cit. on p. 7).

[Farinha, 2021] Carlos M. Farinha and Martina Gentzsch. « Revisiting CFTR Interactions: Old Partners and New Players ». *International Journal of Molecular Sciences* 22.24 (2021), p. 13196 (cit. on p. 67).

[FARINHA, 2002] Carlos M. FARINHA, Paulo NOGUEIRA, Filipa MENDES, Deborah PENQUE, and Margarida D. AMARAL. « The human DnaJ homologue (Hdj)-1/heat-shock protein (Hsp) 40 co-chaperone is required for the in vivo stabilization of the cystic fibrosis transmembrane conductance regulator by Hsp70 ». *Biochemical Journal* 366.3 (2002), pp. 797–806 (cit. on p. 13).

[Farrell, 2008] Philip M. Farrell, Beryl J. Rosenstein, Terry B. White, Frank J. Accurso, Carlo Castellani, Garry R. Cutting, et al. « Guidelines for Diagnosis of Cystic Fibrosis in Newborns through Older Adults: Cystic Fibrosis Foundation Consensus Report ». *The Journal of pediatrics* 153.2 (2008), S4–S14 (cit. on p. 5).

[Favia, 2010] Maria Favia, Lorenzo Guerra, Teresa Fanelli, Rosa Angela Cardone, Stefania Monterisi, Francesca Di Sole, et al. « Na+/H+ Exchanger Regulatory Factor 1 Overexpression-dependent Increase of Cytoskeleton Organization Is Fundamental in the Rescue of F508del Cystic Fibrosis Transmembrane Conductance Regulator in Human Airway CFBE41o- Cells ». *Molecular Biology of the Cell* 21.1 (2010), pp. 73–86 (cit. on pp. 80, 81).

[Ferenc Karpati, 2000] Lena Hjelte Ferenc Karpati Bengt Wretlind. « TNF-A and IL-8 in Consecutive Sputum Samples from Cystic Fibrosis Patients During Antibiotic Treatment ». *Scandinavian Journal of Infectious Diseases* 32.1 (2000), pp. 75–79 (cit. on p. 84).

[Fiorotto, 2018] Romina Fiorotto, Mariangela Amenduni, Valeria Mariotti, Luca Fabris, Carlo Spirli, and Mario Strazzabosco. « Src kinase inhibition reduces inflammatory and cytoskeletal changes in ΔF508 human cholangiocytes and improves cystic fibrosis transmembrane conductance regulator correctors efficacy: Fiorotto, Amenduni, et al. » *Hepatology* 67.3 (2018), pp. 972–988 (cit. on p. 86).

[Fleurot, 2022] Isabelle Fleurot, Raquel López-Gálvez, Pascal Barbry, Antoine Guillon, Mustapha Si-Tahar, Andrea Bähr, et al. « TLR5 signalling is hyper-responsive in porcine cystic fibrosis airways epithelium ». *Journal of Cystic Fibrosis* 21.2 (2022), e117–e121 (cit. on pp. 13, 67, 75).

[Garcia-Alonso, 2019] Luz Garcia-Alonso, Christian H. Holland, Mahmoud M. Ibrahim, Denes Turei, and Julio Saez-Rodriguez. « Benchmark and integration of resources for the estimation of human transcription factor activities ». *Genome Research* 29.8 (2019), pp. 1363–1375 (cit. on pp. 29, 92, 97, 98).

[Garrido-Rodriguez, 2022] Martin Garrido-Rodriguez, Katharina Zirngibl, Olga Ivanova, Sebastian Lobentanzer, and Julio Saez-Rodriguez. « Integrating knowledge and omics to decipher mechanisms via large-scale models of signaling networks ». *Molecular Systems Biology* 18.7 (2022), e11036 (cit. on pp. 24, 29, 31, 95, 96).

[Gaulton, 2012] Anna Gaulton, Louisa J. Bellis, A. Patricia Bento, Jon Chambers, Mark Davies, Anne Hersey, et al. « ChEMBL: a large-scale bioactivity database for drug discovery ». *Nucleic Acids Research* 40.Database issue (2012), pp. D1100–1107 (cit. on p. 108).

[Gentzsch, 2018] Martina Gentzsch and Marcus A. Mall. « Ion Channel Modulators in Cystic Fibrosis ». *Chest* 154.2 (2018), pp. 383–393 (cit. on pp. 8, 10).

[Gillan, 2023] Jonathan L. Gillan, Mithil Chokshi, Gareth R. Hardisty, Sara Clohisey Hendry, Daniel Prasca-Chamorro, Nicola J. Robinson, et al. « CAGE sequencing reveals CFTR-dependent dysregulation of type I IFN signaling in activated cystic fibrosis macrophages ». *Science Advances* 9.21 (2023), eadg5128 (cit. on p. 81).

[Gillespie, 2022] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, et al. « The reactome pathway knowledgebase 2022 ». *Nucleic Acids Research* 50.D1 (2022), pp. D687–D692 (cit. on pp. 30, 96).

[Gorban, 2010] A. N. Gorban and A. Y. Zinovyev. *Principal Graphs and Manifolds*. 2010 (cit. on p. 41).

[Gottlieb, 1996] R A Gottlieb and A Dosanjh. « Mutant cystic fibrosis transmembrane conductance regulator inhibits acidification and apoptosis in C127 cells: possible relevance to cystic fibrosis. » *Proceedings of the National Academy of Sciences* 93.8 (1996), pp. 3587–3591 (cit. on p. 81).

[Guo, 2022] Jonathan Guo, Anna Garratt, and Andrew Hill. « Worldwide rates of diagnosis and effective treatment for cystic fibrosis ». *Journal of Cystic Fibrosis* 21.3 (2022), pp. 456–462 (cit. on p. 67).

[Hajj, 2007] R Hajj, P Lesimple, B Nawrocki-Raby, P Birembaut, E Puchelle, and C Coraux. « Human airway surface epithelial regeneration is delayed and abnormal in cystic fibrosis ». *The Journal of Pathology* 211.3 (2007), pp. 340–350 (cit. on p. 13).

[Hampton, 2010] Thomas H. Hampton and Bruce A. Stanton. « A novel approach to analyze gene expression data demonstrates that the ΔF508 mutation in CFTR downregulates the antigen presentation pathway ». *American Journal of Physiology-Lung Cellular and Molecular Physiology* 298.4 (2010), pp. L473–L482 (cit. on p. 55).

[Hanley, 1982] J A Hanley and B J McNeil. « The meaning and use of the area under a receiver operating characteristic (ROC) curve. » *Radiology* 143.1 (1982), pp. 29–36 (cit. on p. 119).

[Hanssens, 2021] Laurence S. Hanssens, Jean Duchateau, and Georges J. Casimir. « CFTR Protein: Not Just a Chloride Channel? » *Cells* 10.11 (2021), p. 2844 (cit. on pp. 13, 67).

[Hao, 2020] Shuyu Hao, Erica A. Roesch, Aura Perez, Rebecca L. Weiner, Leigh C. Henderson, Linda Cummings, et al. « Inactivation of CFTR by CRISPR/Cas9 alters transcriptional regulation of inflammatory pathways and other networks ». *Journal of Cystic Fibrosis* 19.1 (2020), pp. 34–39 (cit. on pp. 13, 67).

[Hasin, 2017] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. « Multi-omics approaches to disease ». *Genome Biology* 18.1 (2017), p. 83 (cit. on p. 27).

[Hastie, 2009] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. 2nd ed. Springer Series in Statistics. New York: Springer-Verlag, 2009 (cit. on pp. 42, 119).

[Hawkins, 2010] R. David Hawkins, Gary C. Hon, and Bing Ren. « Next-generation genomics: an integrative approach ». *Nature Reviews Genetics* 11.7 (2010), pp. 476–486 (cit. on p. 35).

[Haynes, 2013] Winston A. Haynes, Roger Higdon, Larissa Stanberry, Dwayne Collins, and Eugene Kolker. « Differential Expression Analysis for Pathways ». *PLOS Computational Biology* 9.3 (2013), e1002967 (cit. on p. 29).

[Heijerman, 2019] Harry G M Heijerman, Edward F McKone, Damian G Downey, Eva Van Braeckel, Steven M Rowe, Elizabeth Tullis, et al. « Efficacy and safety of the elexacaftor plus tezacaftor plus ivacaftor combination regimen in people with cystic fibrosis homozygous for the F508del mutation: a double-blind, randomised, phase 3 trial ». *The Lancet* 394.10212 (2019), pp. 1940–1948 (cit. on p. 11).

[Hodos, 2020] Rachel A. Hodos, Matthew D. Strub, Shyam Ramachandran, Li Li, Paul B. McCray, and Joel T. Dudley. « Integrative genomic meta-analysis reveals novel molecular insights into cystic fibrosis and ΔF508-CFTR rescue ». *Scientific Reports* 10 (2020), p. 20553 (cit. on p. 59).

[Hornberg, 2006] Jorrit J. Hornberg, Frank J. Bruggeman, Hans V. Westerhoff, and Jan Lankelma. « Cancer: A Systems Biology disease ». *Biosystems*. 5th International Conference on Systems Biology 83.2 (2006), pp. 81–90 (cit. on p. 67).

[Hsu, 2016] Daniel Hsu, Patricia Taylor, Dave Fletcher, Rolf van Heeckeren, Jean Eastman, Anna van Heeckeren, et al. « Interleukin-17 Pathophysiology and Therapeutic Intervention in Cystic Fibrosis Lung Infection and Inflammation ». *Infection and Immunity* 84.9 (2016), pp. 2410–2421 (cit. on p. 75).

[Ideozu, 2019] Justin Ideozu, Xi Zhang, Susanna McColley, and Hara Levy. « Transcriptome Profiling and Molecular Therapeutic Advances in Cystic Fibrosis: Recent Insights ». *Genes* 10.3 (2019), p. 180 (cit. on pp. 34, 54, 59, 64, 71).

[Jacob, 2012] Laurent Jacob, Pierre Neuvial, and Sandrine Dudoit. « More power via graph-structured tests for differential expression of gene networks ». *The Annals of Applied Statistics* 6.2 (2012), pp. 561–600 (cit. on p. 29).

[Jacob, 2008] Laurent Jacob and Jean-Philippe Vert. « Protein-ligand interaction prediction: an improved chemogenomics approach ». *Bioinformatics* 24.19 (2008), pp. 2149–2156 (cit. on p. 115).

[Jacquot, 2008] Jacky Jacquot, Olivier Tabary, Philippe Le Rouzic, and Annick Clement. « Airway epithelial cell inflammatory signalling in cystic fibrosis ». *The International Journal of Biochemistry & Cell Biology* 40.9 (2008), pp. 1703–1715 (cit. on p. 67).

[Jeanson, 2012] L. Jeanson, M. Kelly, A. Coste, I. C. Guerrera, J. Fritsch, T. Nguyen-Khoa, et al. « Oxidative stress induces unfolding protein response and inflammation in nasal polyposis ». *Allergy* 67.3 (2012), pp. 403–412 (cit. on pp. 13, 67).

[Jeanson, 2014] Ludovic Jeanson, Ida Chiara Guerrera, Jean-François Papon, Cerina Chhuon, Patricia Zadigue, Virginie Prulière-Escabasse, et al. « Proteomic Analysis of Nasal Epithelial Cells from Cystic Fibrosis Patients ». *PLOS ONE* 9.9 (2014), e108671 (cit. on p. 58).

[Jensen, 1995] Timothy J. Jensen, Melinda A. Loo, Steven Pind, David B. Williams, Alfred L. Goldberg, and John R. Riordan. « Multiple proteolytic systems, including the proteasome, contribute to CFTR processing ». *Cell* 83.1 (1995), pp. 129–135 (cit. on p. 7).

[Jiang, 2019] Kaiyu Jiang, Kerry E. Poppenberg, Laiping Wong, Yanmin Chen, Drucy Borowitz, Danielle Goetz, et al. « RNA sequencing data from neutrophils of patients with cystic fibrosis reveals potential for developing biomarkers for pulmonary exacerbations ». *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* 18.2 (2019), pp. 194–202 (cit. on p. 55).

[Jumper, 2021] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, et al. « Highly accurate protein structure prediction with AlphaFold ». *Nature* 596.7873 (2021), pp. 583–589 (cit. on p. 135).

[Kamei, 2019] Shunsuke Kamei, Kasumi Maruta, Haruka Fujikawa, Hirofumi Nohara, Keiko Ueno-Shuto, Yukihiro Tasaki, et al. « Integrative expression analysis identifies a novel interplay between CFTR and linc-SUMF1-2 that involves CF-associated gene dysregulation ». *Biochemical and Biophysical Research Communications* 509.2 (2019), pp. 521–528 (cit. on p. 55).

[Kanehisa, 2021] Minoru Kanehisa, Miho Furumichi, Yoko Sato, Mari Ishiguro-Watanabe, and Mao Tanabe. « KEGG: integrating viruses and cellular organisms ». *Nucleic Acids Research* 49.D1 (2021), pp. D545–D551 (cit. on p. 72).

[Kanehisa, 2012] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. « KEGG for integration and interpretation of large-scale molecular data sets ». *Nucleic Acids Research* 40.Database issue (2012), pp. D109–D114 (cit. on pp. 24, 30, 44, 96).

[Keating, 2018] Dominic Keating, Gautham Marigowda, Lucy Burr, Cori Daines, Marcus A. Mall, Edward F. McKone, et al. « VX-445–Tezacaftor–Ivacaftor in Patients with Cystic Fibrosis and One or Two Phe508del Alleles ». *New England Journal of Medicine* 379.17 (2018), pp. 1612–1620 (cit. on p. 11).

[Kelly-Aubert, 2011] Mairead Kelly-Aubert, Stéphanie Trudel, Janine Fritsch, Thao Nguyen-Khoa, Maryvonne Baudouin-Legros, Sandra Moriceau, et al. « GSH monoethyl ester rescues mitochondrial defects in cystic fibrosis models ». *Human Molecular Genetics* 20.14 (2011), pp. 2745–2759 (cit. on p. 13).

[Khan, 1995] Talat Z. Khan, Jeffrey S. Wagener, Thomas Bost, Jose Martinez, Frank J. Accurso, and David W. H. Riches. « Early Pulmonary Inflammation in Infants with Cystic Fibrosis ». *American Journal of Respiratory and Critical Care Medicine* 151.4 (1995), pp. 1075–1082 (cit. on p. 12).

[Khatri, 2012] Purvesh Khatri, Marina Sirota, and Atul J. Butte. « Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges ». *PLOS Computational Biology* 8.2 (2012), e1002375 (cit. on p. 28).

[Kitano, 2002] Hiroaki Kitano. « Computational systems biology ». *Nature* 420.6912 (2002), pp. 206–210 (cit. on p. 23).

[Konstan, 2017] Michael W Konstan, Edward F McKone, Richard B Moss, Gautham Marigowda, Simon Tian, David Waltz, et al. « Assessment of safety and efficacy of long-term treatment with combination lumacaftor and ivacaftor therapy in patients with cystic fibrosis homozygous for the F508del-CFTR mutation (PROGRESS): a phase 3, extension study ». *The Lancet Respiratory Medicine* 5.2 (2017), pp. 107–118 (cit. on p. 11).

[Kopp, 2020] Benjamin T. Kopp, James Fitch, Lisa Jaramillo, Chandra L. Shrestha, Frank Robledo-Avila, Shuzhong Zhang, et al. « Whole-blood transcriptomic responses to lumacaftor/ivacaftor therapy in cystic fibrosis ». *Journal of Cystic Fibrosis: Official Journal of the European Cystic Fibrosis Society* 19.2 (2020), pp. 245–254 (cit. on pp. 55, 56).

[Kormann, 2017] Michael S. D. Kormann, Alexander Dewerth, Felizitas Eichner, Praveen Baskaran, Andreas Hector, Nicolas Regamey, et al. « Transcriptomic profile of cystic fibrosis patients identifies type I interferon response and ribosomal stalk proteins as potential modifiers of disease severity ». *PLOS ONE* 12.8 (2017). Ed. by Sanjay Haresh Chotirmall, e0183526 (cit. on p. 55).

[Korotkevich, 2021] Gennady Korotkevich, Vladimir Sukhov, Nikolay Budin, Boris Shpak, Maxim N. Artyomov, and Alexey Sergushichev. *Fast gene set enrichment analysis*. 2021 (cit. on pp. 90, 97).

[Kosamo, 2020] Susanna Kosamo, Katherine B. Hisert, Victoria Dmyterko, Catherine Nguyen, R. Anthony Black, Tarah D. Holden, et al. « Strong toll-like receptor responses in cystic fibrosis patients are associated with higher lung function ». *Journal of Cystic Fibrosis* 19.4 (2020), pp. 608–613 (cit. on pp. 75, 81).

[Kramer, 2018] Elizabeth L. Kramer and John P. Clancy. « TGFβ as a therapeutic target in cystic fibrosis ». *Expert Opinion on Therapeutic Targets* 22.2 (2018), pp. 177–189 (cit. on p. 87).

[Kustatscher, 2022] Georg Kustatscher, Tom Collins, Anne-Claude Gingras, Tiannan Guo, Henning Hermjakob, Trey Ideker, et al. « Understudied proteins: opportunities and challenges for functional proteomics ». *Nature Methods* 19.7 (2022), pp. 774–779 (cit. on p. 64).

[Landais, 2023] Yuna Landais and Céline Vallot. « Multi-modal quantification of pathway activity with MAYA ». *Nature Communications* 14.1 (2023), p. 1668 (cit. on p. 72).

[Lasalvia, 2016] Maria Lasalvia, Stefano Castellani, Palma D'Antonio, Giuseppe Perna, Annalucia Carbone, Anna Laura Colia, et al. « Human airway epithelial cells investigated by atomic force microscopy: A hint to cystic fibrosis epithelial pathology ». *Experimental Cell Research* 348.1 (2016), pp. 46–55 (cit. on p. 81).

[Law, 2018] Charity W. Law, Monther Alhamdoosh, Shian Su, Xueyi Dong, Luyi Tian, Gordon K. Smyth, et al. « RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR ». *F1000Research* 5 (2018), ISCB Comm J–1408 (cit. on p. 90).

[Law, 2014] Vivian Law, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, et al. « DrugBank 4.0: shedding new light on drug metabolism ». *Nucleic Acids Research* 42.D1 (2014), pp. D1091–D1097 (cit. on pp. 116, 122).

[Levine, 2006] David M. Levine, David R. Haynor, John C. Castle, Sergey B. Stepaniants, Matteo Pellegrini, Mao Mao, et al. « Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways ». *Genome Biology* 7.10 (2006), R93 (cit. on p. 35).

[Levy, 2012] H. Levy, X. Wang, M. Kaldunski, S. Jia, J. Kramer, S. J. Pavletich, et al. « Transcriptional signatures as a disease-specific and predictive inflammatory biomarker for type 1 diabetes ». *Genes & Immunity* 13.8 (2012), pp. 593–604 (cit. on p. 55).

[Levy, 2018] Hara Levy, Shuang Jia, Amy Pan, Xi Zhang, Mary Kaldunski, Melodee L. Nugent, et al. « Identification of molecular signatures of cystic fibrosis disease status with plasma-based functional genomics ». *Physiological Genomics* 51.1 (2018), pp. 27–41 (cit. on pp. 55, 56).

[Liberzon, 2015] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. « The Molecular Signatures Database (MSigDB) hallmark gene set collection ». *Cell systems* 1.6 (2015), pp. 417–425 (cit. on pp. 29, 44, 72, 89, 98).

[Lim, 2022] Sang Hyun Lim, Jamie Snider, Liron Birimberg-Schwartz, Wan Ip, Joana C Serralha, Hugo M Botelho, et al. « CFTR interactome mapping using the mammalian membrane two-hybrid high-throughput screening system ». *Molecular Systems Biology* 18.2 (2022), e10629 (cit. on p. 58).

[Ling, 2020] Kak-Ming Ling, Luke W. Garratt, Erin E. Gill, Amy H. Y. Lee, Patricia Agudelo-Romero, Erika N. Sutanto, et al. « Rhinovirus Infection Drives Complex Host Airway Molecular Responses in Children With Cystic Fibrosis ». *Frontiers in Immunology* 11 (2020), p. 1327 (cit. on pp. 55, 60, 71, 75).

[Liou, 2001] Theodore G. Liou, Frederick R. Adler, Stacey C. FitzSimmons, Barbara C. Cahill, Jonathan R. Hibbs, and Bruce C. Marshall. « Predictive 5-Year Survivorship Model of Cystic Fibrosis ». *American Journal of Epidemiology* 153.4 (2001), pp. 345–352 (cit. on p. 10).

[Lipinski, 2001] Christopher A Lipinski, Franco Lombardo, Beryl W Dominy, and Paul J Feeney. « Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings 1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25. 1 ». *Advanced Drug Delivery Reviews* 46.1-3 (2001), pp. 3–26 (cit. on p. 116).

[Liu, 2019] Anika Liu, Panuwat Trairatphisan, Enio Gjerga, Athanasios Didangelos, Jonathan Barratt, and Julio Saez-Rodriguez. « From expression footprints to causal pathways: contextualizing large signaling networks with CARNIVAL ». *npj Systems Biology and Applications* 5.1 (2019), pp. 1–10 (cit. on p. 97).

[Liu, 2007] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. « BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities ». *Nucleic Acids Research* 35.Database issue (2007), pp. D198–D201 (cit. on p. 134).

[Liu, 2016] Yong Liu, Min Wu, Chunyan Miao, Peilin Zhao, and Xiao-Li Li. « Neighborhood Regularized Logistic Matrix Factorization for Drug-Target Interaction Prediction ». *PLOS Computational Biology* 12.2 (2016), e1004760 (cit. on p. 116).

[Lo Surdo, 2022] Prisca Lo Surdo, Marta Iannuccelli, Silvia Contino, Luisa Castagnoli, Luana Licata, Gianni Cesareni, et al. « SIGNOR 3.0, the SIGnaling network open resource 3.0: 2022 update ». *Nucleic acids research* (2022), gkac883 (cit. on pp. 23, 91, 96, 100).

[Lopes-Pacheco, 2016] Miquéias Lopes-Pacheco. « CFTR Modulators: Shedding Light on Precision Medicine for Cystic Fibrosis ». *Frontiers in Pharmacology* 7 (2016) (cit. on p. 5).

[Loureiro, 2019] Cláudia Almeida Loureiro, João D. Santos, Ana Margarida Matos, Peter Jordan, Paulo Matos, Carlos M. Farinha, et al. « Network Biology Identifies Novel Regulators of CFTR Trafficking and Membrane Stability ». *Frontiers in Pharmacology* 10 (2019) (cit. on p. 60).

[Luck, 2020] Katja Luck, Dae-Kyum Kim, Luke Lambourne, Kerstin Spirohn, Bridget E. Begg, Wenting Bian, et al. « A reference map of the human binary protein interactome ». *Nature* 580.7803 (2020), pp. 402–408 (cit. on p. 23).

[Magalhães, 2022] João Pedro de Magalhães. « Every gene can (and possibly will) be associated with cancer ». *Trends in Genetics* 38.3 (2022), pp. 216–217 (cit. on p. 64).

[Marson, 2016] Fernando Augusto Lima Marson, Carmen Sílvia Bertuzzo, and José Dirceu Ribeiro. « Classification of CFTR mutation classes ». *The Lancet Respiratory Medicine* 4.8 (2016), e37–e38 (cit. on p. 8).

[Martens, 2021] Marvin Martens, Ammar Ammar, Anders Riutta, Andra Waagmeester, Denise N. Slenter, Kristina Hanspers, et al. « WikiPathways: connecting communities ». *Nucleic Acids Research* 49.D1 (2021), pp. D613–D621 (cit. on p. 30).

[Martignetti, 2016] Loredana Martignetti, Laurence Calzone, Eric Bonnet, Emmanuel Barillot, and Andrei Zinovyev. « ROMA: Representation and Quantification of Module Activity from Target Expression Data ». *Frontiers in Genetics* 7 (2016) (cit. on pp. 16, 34, 38, 72).

[Martinez-Lopez, 2017] Yoan Martinez-Lopez, Yaile Caballero, Stephen J. Barigye, Yovani Marrero-Ponce, Reisel Millan-Cabrera, Julio Madera, et al. « State of the Art Review and Report of New Tool for Drug Discovery ». *Current Topics in Medicinal Chemistry* 17.26 (2017) (cit. on p. 115).

[Martini, 2013] Paolo Martini, Gabriele Sales, M. Sofia Massa, Monica Chiogna, and Chiara Romualdi. « Along signal paths: an empirical gene set approach exploiting pathway topology ». *Nucleic Acids Research* 41.1 (2013), e19 (cit. on p. 29).

[Massip Copiz, 2016] María Macarena Massip Copiz and Tomás Antonio Santa Coloma. « c-Src and its role in cystic fibrosis ». *European Journal of Cell Biology* 95.10 (2016), pp. 401–413 (cit. on p. 80).

[Matos, 2018] Ana M. Matos, Andreia Gomes-Duarte, Márcia Faria, Patrícia Barros, Peter Jordan, Margarida D. Amaral, et al. « Prolonged co-treatment with HGF sustains epithelial integrity and improves pharmacological rescue of Phe508del-CFTR ». *Scientific Reports* 8.1 (2018), p. 13026 (cit. on pp. 58, 61).

[Matos, 2019] Ana M. Matos, Francisco R. Pinto, Patrícia Barros, Margarida D. Amaral, Rainer Pepperkok, and Paulo Matos. « Inhibition of calpain 1 restores plasma membrane stability to pharmacologically rescued Phe508del-CFTR variant ». *Journal of Biological Chemistry* 294.36 (2019), pp. 13396–13410 (cit. on p. 100).

[Mayer, 2012] Matt Mayer, Christoph Blohmke, Reza Falsafi, Christopher Fjell, Laurence Madera, Stuart Turvey, et al. « Rescue of Dysfunctional Autophagy Attenuates Hyperinflammatory Responses from Cystic Fibrosis Cells ». *Journal of immunology (Baltimore, Md. : 1950)* 190 (2012) (cit. on pp. 55, 60).

[Mazein, 2018] Alexander Mazein, Marek Ostaszewski, Inna Kuperstein, Steven Watterson, Nicolas Le Novère, Diane Lefaudeux, et al. « Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms ». *npj Systems Biology and Applications* 4.1 (2018), pp. 1–10 (cit. on p. 61).

[McKiernan, 2014] Paul J. McKiernan, Kevin Molloy, Sally A. Cryan, Noel G. McElvaney, and Catherine M. Greene. « Long noncoding RNA are aberrantly expressed in vivo in the cystic fibrosis bronchial epithelium ». *The International Journal of Biochemistry & Cell Biology*. Cystic Fibrosis: From o-mics to cell biology, physiology, and therapeutic advances 52 (2014), pp. 184–191 (cit. on p. 55).

[Mendes, 2011] Ana Isabel Mendes, Paulo Matos, Sónia Moniz, Simão Luz, Margarida D. Amaral, Carlos M. Farinha, et al. « Antagonistic Regulation of Cystic Fibrosis Transmembrane Conductance Regulator Cell Surface Expression by Protein Kinases WNK4 and Spleen Tyrosine Kinase ». *Molecular and Cellular Biology* 31.19 (2011), pp. 4076–4086 (cit. on p. 80).

[Meng, 2017] Xin Meng, Jack Clews, Vasileios Kargas, Xiaomeng Wang, and Robert C. Ford. « The cystic fibrosis transmembrane conductance regulator (CFTR) and its stability ». *Cellular and Molecular Life Sciences* 74.1 (2017), pp. 23–38 (cit. on p. 7).

[Meslamani, 2011] Jamel Meslamani and Didier Rognan. « Enhancing the Accuracy of Chemogenomic Models with a Three-Dimensional Binding Site Kernel ». *Journal of Chemical Information and Modeling* 51.7 (2011), pp. 1593–1603 (cit. on p. 121).

[Moffat, 2017] John G. Moffat, Fabien Vincent, Jonathan A. Lee, Jörg Eder, and Marco Prunotto. « Opportunities and challenges in phenotypic drug discovery: an industry perspective ». *Nature Reviews Drug Discovery* 16.8 (2017), pp. 531–543 (cit. on p. 115).

[Montagud, 2022] Arnau Montagud, Jonas Béal, Luis Tobalina, Pauline Traynard, Vigneshwari Subramanian, Bence Szalai, et al. « Patient-specific Boolean models of signalling networks guide personalised treatments ». *eLife* 11 (2022). Ed. by Jennifer Flegg, e72626 (cit. on p. 142).

[Montoro, 2018] Daniel T. Montoro, Adam L. Haber, Moshe Biton, Vladimir Vinarsky, Brian Lin, Susan Birket, et al. « A revised airway epithelial hierarchy includes CFTR-expressing ionocytes ». *Nature* 560.7718 (2018), pp. 319–324 (cit. on p. 57).

[Müller-Dott, 2023] Sophia Müller-Dott, Eirini Tsirvouli, Miguel Vázquez, Ricardo O. Ramirez Flores, Pau Badia-i-Mompel, Robin Fallegger, et al. *Expanding the coverage of regulons from high-confidence prior knowledge for accurate estimation of transcription factor activities*. 2023 (cit. on p. 97).

[Mutyam, 2017] Venkateshwar Mutyam, Emily Falk Libby, Ning Peng, Denis Hadjiliadis, Michael Bonk, George M. Solomon, et al. « Therapeutic benefit observed with the CFTR potentiator, ivacaftor, in a CF patient homozygous for the W1282X CFTR nonsense mutation ». *Journal of Cystic Fibrosis* 16.1 (2017), pp. 24–29 (cit. on p. 12).

[Natarajan, 2020] Viswanathan Natarajan. « Is PI3K a Villain in Cystic Fibrosis? » *American Journal of Respiratory Cell and Molecular Biology* 62.5 (2020), pp. 552–553 (cit. on p. 86).

[Nichols, 2015] David P. Nichols and James F. Chmiel. « Inflammation and its genesis in cystic fibrosis ». *Pediatric Pulmonology* 50.S40 (2015), S39–S56 (cit. on p. 12).

[Novère, 2009] Nicolas Le Novère, Michael Hucka, Huaiyu Mi, Stuart Moodie, Falk Schreiber, Anatoly Sorokin, et al. « The Systems Biology Graphical Notation ». *Nature Biotechnology* 27.8 (2009), pp. 735–741 (cit. on pp. 30, 61).

[Novoa-del-Toro, 2021] Elva María Novoa-del-Toro, Efrén Mezura-Montes, Matthieu Vignes, Morgane Térézol, Frédérique Magdinier, Laurent Tichit, et al. « A multi-objective genetic algorithm to find active modules in multiplex biological networks ». *PLOS Computational Biology* 17.8 (2021), e1009263 (cit. on p. 96).

[ONeal, 2015] Wanda K. O'Neal, Paul Gallins, Rhonda G. Pace, Hong Dang, Whitney E. Wolf, Lisa C. Jones, et al. « Gene Expression in Transformed Lymphocytes Reveals Variation in Endomembrane and HLA Pathways Modifying Cystic Fibrosis Pulmonary Phenotypes ». *The American Journal of Human Genetics* 96.2 (2015), pp. 318–328 (cit. on p. 55).

[Ogilvie, 2011] Varrie Ogilvie, Margaret Passmore, Laura Hyndman, Lisa Jones, Barbara Stevenson, Abigail Wilson, et al. « Differential global gene expression in cystic fibrosis nasal and bronchial epithelium ». *Genomics* 98.5 (2011), pp. 327–336 (cit. on pp. 55, 71, 96).

[Okuda, 2021] Kenichi Okuda, Hong Dang, Yoshihiko Kobayashi, Gianni Carraro, Satoko Nakano, Gang Chen, et al. « Secretory Cells Dominate Airway CFTR Expression and Function in Human Airway Superficial Epithelia ». *American Journal of Respiratory and Critical Care Medicine* (2021) (cit. on pp. 47, 57, 143).

[Ong, 2007] Serene AK Ong, Hong Huang Lin, Yu Zong Chen, Ze Rong Li, and Zhiwei Cao. « Efficacy of different protein descriptors in predicting protein functional families ». *BMC Bioinformatics* 8.1 (2007), p. 300 (cit. on p. 118).

[Oughtred, 2021] Rose Oughtred, Jennifer Rust, Christie Chang, Bobby-Joe Breitkreutz, Chris Stark, Andrew Willems, et al. « The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions ». *Protein Science : A Publication of the Protein Society* 30.1 (2021), pp. 187–200 (cit. on p. 23).

[Pahikkala, 2015] Tapio Pahikkala, Antti Airola, Sami Pietilä, Sushil Shakyawar, Agnieszka Szwajda, Jing Tang, et al. « Toward more realistic drug-target interaction predictions ». *Briefings in Bioinformatics* 16.2 (2015), pp. 325–337 (cit. on p. 116).

[Pankonien, 2022] Ines Pankonien, Margarida C. Quaresma, Cláudia S. Rodrigues, and Margarida D. Amaral. « CFTR, Cell Junctions and the Cytoskeleton ». *International Journal of Molecular Sciences* 23.5 (2022), p. 2688 (cit. on pp. 67, 68, 75).

[Pankow, 2015] Sandra Pankow, Casimir Bamberger, Diego Calzolari, Salvador Martínez-Bartolomé, Mathieu Lavallée-Adam, William E. Balch, et al. « F508 CFTR interactome remodelling promotes rescue of cystic fibrosis ». *Nature* 528.7583 (2015), pp. 510–516 (cit. on pp. 58, 61, 68).

[Papa, 2021] Riccardo Papa, Federica Penco, Stefano Volpi, and Marco Gattorno. « Actin Remodeling Defects Leading to Autoinflammation and Immune Dysregulation ». *Frontiers in Immunology* 11 (2021) (cit. on p. 101).

[Pereira, 2021] Catarina Pereira, Alexander Mazein, Carlos M. Farinha, Michael A. Gray, Karl Kunzelmann, Marek Ostaszewski, et al. « CyFi-MAP: an interactive pathway-based resource for cystic fibrosis ». *Scientific Reports* 11.1 (2021), p. 22223 (cit. on pp. 61, 67, 68, 78, 100).

[Picciotto, 1992] M. R. Picciotto, J. A. Cohn, G Bertuzzi, P Greengard, and A. C. Nairn. « Phosphorylation of the cystic fibrosis transmembrane conductance regulator. » *Journal of Biological Chemistry* 267.18 (1992), pp. 12742–12752 (cit. on p. 7).

[Pineau, 2020] Fanny Pineau, Davide Caimmi, Milena Magalhães, Enora Fremy, Abdillah Mohamed, Laurent Mely, et al. « Blood co-expression modules identify potential modifier genes of diabetes and lung function in cystic fibrosis ». *PLOS ONE* 15.4 (2020), e0231285 (cit. on p. 60).

[Plasschaert, 2018] Lindsey W. Plasschaert, Rapolas Žilionis, Rayman Choo-Wing, Virginia Savova, Judith Knehr, Guglielmo Roma, et al. « A single-cell atlas of the airway epithelium reveals the CFTR-rich pulmonary ionocyte ». *Nature* 560.7718 (2018), pp. 377–381 (cit. on pp. 46, 57).

[Playe, 2019] Benoit Playe. « Méthodes d'apprentissage statistique pour le criblage virtuel de médicament ». These de doctorat. Paris Sciences et Lettres (ComUE), 2019 (cit. on p. 110).

[Playe, 2018] Benoit Playe, Chloé-Agathe Azencott, and Véronique Stoven. « Efficient multi-task chemogenomics for drug specificity prediction ». *PLOS ONE* 13.10 (2018). Ed. by Alexandre G. de Brevern, e0204999 (cit. on pp. 107, 114, 116, 118, 121, 134).

[Playe, 2020] Benoit Playe and Veronique Stoven. « Evaluation of deep and shallow learning methods in chemogenomics for the prediction of drugs specificity ». *Journal of Cheminformatics* 12.1 (2020), p. 11 (cit. on p. 116).

[Polineni, 2018] Deepika Polineni, Hong Dang, Paul J. Gallins, Lisa C. Jones, Rhonda G. Pace, Jaclyn R. Stonebraker, et al. « Airway Mucosal Host Defense Is Key to Genomic Regulation of Cystic Fibrosis Lung Disease Severity ». *American Journal of Respiratory and Critical Care Medicine* 197.1 (2018), pp. 79–93 (cit. on pp. 55, 56).

[Pollard, 2006] Harvey B. Pollard, Ofer Eidelman, Catherine Jozwik, Wei Huang, Meera Srivastava, Xia D. Ji, et al. « De Novo Biosynthetic Profiling of High Abundance Proteins in Cystic Fibrosis Lung Epithelial Cells*S ». *Molecular & Cellular Proteomics* 5.9 (2006), pp. 1628–1637 (cit. on p. 58).

[Puglia, 2018] Michele Puglia, Claudia Landi, Assunta Gagliardi, Loretta Breslin, Alessandro Armini, Jlenia Brunetti, et al. « The proteome speciation of an immortalized cystic fibrosis cell line: New perspectives on the pathophysiology of the disease ». *Journal of Proteomics* 170 (2018), pp. 28–42 (cit. on p. 58).

[Raghavan, 1989] Vijay Raghavan, Peter Bollmann, and Gwang S. Jung. « A critical investigation of recall and precision as measures of retrieval system performance ». *ACM Transactions on Information Systems* 7.3 (1989), pp. 205–229 (cit. on p. 119).

[Ramos-Rodriguez, 2012] Roberto-Rafael Ramos-Rodriguez, Raquel Cuevas-Diaz-Duran, Francesco Falciani, Jose-Gerardo Tamez-Peña, and Victor Trevino. « COMPADRE: an R and web resource for pathway activity analysis by component decompositions ». *Bioinformatics* 28.20 (2012), pp. 2701–2702 (cit. on p. 35).

[Ramsey, 2011] Bonnie W. Ramsey, Jane Davies, N. Gerard McElvaney, Elizabeth Tullis, Scott C. Bell, Pavel Dřevínek, et al. « A CFTR Potentiator in Patients with Cystic Fibrosis and the *G551D* Mutation ». *New England Journal of Medicine* 365.18 (2011), pp. 1663–1672 (cit. on p. 11).

[Rauniyar, 2014] Navin Rauniyar, Vijay Gupta, William E. Balch, and John R. III Yates. « Quantitative Proteomic Profiling Reveals Differentially Regulated Proteins in Cystic Fibrosis Cells ». *Journal of Proteome Research* 13.11 (2014), pp. 4668–4675 (cit. on p. 58).

[Rehman, 2021] Tayyab Rehman, Philip H. Karp, Ping Tan, Brian J. Goodell, Alejandro A. Pezzulo, Andrew L. Thurman, et al. « Inflammatory cytokines TNF-α and IL-17 enhance the efficacy of cystic fibrosis transmembrane conductance regulator modulators ». *The Journal of Clinical Investigation* 131.16 (2021), p. 150398 (cit. on pp. 12, 44, 55, 71, 89, 142).

[Reilly, 2017] R. Reilly, M. S. Mroz, E. Dempsey, K. Wynne, S. J. Keely, E. F. McKone, et al. « Targeting the PI3K/Akt/mTOR signalling pathway in Cystic Fibrosis ». *Scientific Reports* 7.1 (2017), p. 7642 (cit. on p. 58).

[Ribeiro, 2009] Carla Maria P. Ribeiro, Harry Hurd, Yichao Wu, Mary E. B. Martino, Lisa Jones, Brian Brighton, et al. « Azithromycin Treatment Alters Gene Expression in Inflammatory, Lipid Metabolism, and Cell Cycle Pathways in Well-Differentiated Human Airway Epithelia ». *PLoS ONE* 4.6 (2009). Ed. by Dominik Hartl, e5806 (cit. on p. 55).

[Riccaboni, 2010] Mauro Riccaboni, Ivana Bianchi, and Paola Petrillo. « Spleen tyrosine kinases: biology, therapeutic targets and drugs ». *Drug Discovery Today* 15.13 (2010), pp. 517–530 (cit. on p. 84).

[Rimessi, 2018] Alessandro Rimessi, Valentino Bezzerri, Francesca Salvatori, Anna Tamanini, Federica Nigro, Maria Cristina Dechecchi, et al. « PLCB3 Loss of Function Reduces Pseudomonas aeruginosa–Dependent IL-8 Release in Cystic Fibrosis ». *American Journal of Respiratory Cell and Molecular Biology* 59.4 (2018), pp. 428–436 (cit. on p. 80).

[Ritchie, 2015] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, et al. « limma powers differential expression analyses for RNA-sequencing and microarray studies ». *Nucleic Acids Research* 43.7 (2015), e47 (cit. on p. 90).

[Robinson, 2010a] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. « edgeR: a Bioconductor package for differential expression analysis of digital gene expression data ». *Bioinformatics* 26.1 (2010), pp. 139–140 (cit. on p. 90).

[Robinson, 2010b] Mark D. Robinson and Alicia Oshlack. « A scaling normalization method for differential expression analysis of RNA-seq data ». *Genome Biology* 11.3 (2010), R25 (cit. on p. 90).

[Rogers, 2010] David Rogers and Mathew Hahn. « Extended-Connectivity Fingerprints ». *Journal of Chemical Information and Modeling* 50.5 (2010), pp. 742–754 (cit. on p. 118).

[Rolland, 2014] Thomas Rolland, Murat Taşan, Benoit Charloteaux, Samuel J. Pevzner, Quan Zhong, Nidhi Sahni, et al. « A Proteome-Scale Map of the Human Interactome Network ». *Cell* 159.5 (2014), pp. 1212–1226 (cit. on p. 96).

[Rommens, 1989] J. M. Rommens, M. C. Iannuzzi, B. Kerem, M. L. Drumm, G. Melmer, M. Dean, et al. « Identification of the cystic fibrosis gene: chromosome walking and jumping ». *Science (New York, N.Y.)* 245.4922 (1989), pp. 1059–1065 (cit. on p. 6).

[Rosen, 2018] Bradley H. Rosen, T. Idil Apak Evans, Shashanna R. Moll, Jaimie S. Gray, Bo Liang, Xingshen Sun, et al. « Infection Is Not Required for Mucoinflammatory Lung Disease in CFTR-Knockout Ferrets ». *American Journal of Respiratory and Critical Care Medicine* 197.10 (2018), pp. 1308–1318 (cit. on p. 13).

[Rottner, 2007] Mathilde Rottner, Corinne Kunzelmann, Martine Mergey, Jean-Marie Freyssinet, and María Carmen Martínez. « Exaggerated apoptosis and NF-kappaB activation in pancreatic and tracheal cystic fibrosis cells ». *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 21.11 (2007), pp. 2939–2948 (cit. on p. 81).

[Ruiz García, 2019] Sandra Ruiz García, Marie Deprez, Kevin Lebrigand, Amélie Cavard, Agnès Paquet, Marie-Jeanne Arguel, et al. « Novel dynamics of human mucociliary differentiation revealed by single-cell RNA sequencing of nasal epithelial cultures ». *Development* 146.20 (2019), dev177428 (cit. on p. 57).

[Saavedra, 2008] Milene T. Saavedra, Grant J. Hughes, Linda A. Sanders, Michelle Carr, David M. Rodman, Christopher D. Coldren, et al. « Circulating RNA Transcripts Identify Therapeutic Response in Cystic Fibrosis Lung Disease ». *American Journal of Respiratory and Critical Care Medicine* 178.9 (2008), pp. 929–938 (cit. on p. 56).

[Sagel, 2021] Scott D. Sagel, Umer Khan, Sonya L. Heltshe, John P. Clancy, Drucy Borowitz, Daniel Gelfond, et al. « Clinical Effectiveness of Lumacaftor/Ivacaftor in Patients with Cystic Fibrosis Homozygous for F508del-CFTR. A Clinical Trial ». *Annals of the American Thoracic Society* 18.1 (2021), pp. 75–83 (cit. on p. 12).

[Sagwal, 2020] Swati Sagwal, Anil Chauhan, Jyotdeep Kaur, Rajendra Prasad, Meenu Singh, and Manvi Singh. « Association of Serum TGF-β1 Levels with Different Clinical Phenotypes of Cystic Fibrosis Exacerbation ». *Lung* 198.2 (2020), pp. 377–383 (cit. on p. 87).

[Sahrawat, 2013] Tammanna R Sahrawat and Sherry Bhalla. « Identification of Critical Target Protein for Cystic Fibrosis using Systems Biology Network Approach » (2013), p. 14 (cit. on p. 61).

[Saigo, 2004] Hiroto Saigo, Jean-Philippe Vert, Nobuhisa Ueda, and Tatsuya Akutsu. « Protein homology detection using string alignment kernels ». *Bioinformatics (Oxford, England)* 20.11 (2004), pp. 1682–1689 (cit. on p. 121).

[Saint-Criq, 2020] Vinciane Saint-Criq, Livia Delpiano, John Casement, Jennifer C. Onuora, JinHeng Lin, and Michael A. Gray. « Choice of Differentiation Media Significantly Impacts Cell Lineage and Response to CFTR Modulators in Fully Differentiated Primary Cultures of Cystic Fibrosis Human Airway Epithelial Cells ». *Cells* 9.9 (2020), p. 2137 (cit. on pp. 38, 46, 55, 71, 75).

[Saint-Criq, 2017] Vinciane Saint-Criq and Michael A. Gray. « Role of CFTR in epithelial physiology ». *Cellular and Molecular Life Sciences* 74.1 (2017), pp. 93–115 (cit. on pp. 16, 89).

[Santos, 2023] Lúcia Santos, Rui Nascimento, Aires Duarte, Violeta Railean, Margarida D. Amaral, Patrick T. Harrison, et al. « Mutation-class dependent signatures outweigh disease-associated processes in cystic fibrosis cells ». *Cell & Bioscience* 13.1 (2023), p. 26 (cit. on pp. 89, 142).

[Schaefer, 2009] Carl F. Schaefer, Kira Anthony, Shiva Krupa, Jeffrey Buchoff, Matthew Day, Timo Hannay, et al. « PID: the Pathway Interaction Database ». *Nucleic Acids Research* 37.Database issue (2009), pp. D674–D679 (cit. on pp. 29, 72, 89, 98).

[Schölkopf, 2004] Bernhard Schölkopf, Koji Tsuda, and Jean-Philippe Vert, eds. *Kernel methods in computational biology*. Computational molecular biology. Cambridge, Mass: MIT Press, 2004 (cit. on p. 118).

[Schreiber, 2007] Andreas W. Schreiber and Ute Baumann. « A framework for gene expression analysis ». *Bioinformatics* 23.2 (2007), pp. 191–197 (cit. on p. 35).

[Schubert, 2018] Michael Schubert, Bertram Klinger, Martina Klünemann, Anja Sieber, Florian Uhlitz, Sascha Sauer, et al. « Perturbation-response genes reveal signaling footprints in cancer gene expression ». *Nature Communications* 9.1 (2018), p. 20 (cit. on p. 72).

[Seal, 2023] Ruth L Seal, Bryony Braschi, Kristian Gray, Tamsin E M Jones, Susan Tweedie, Liora Haim-Vilmovsky, et al. « Genenames.org: the HGNC resources in 2023 ». *Nucleic Acids Research* 51.D1 (2023), pp. D1003–D1009 (cit. on p. 92).

[Seibert, 1997] Fabian S. Seibert, Tip W. Loo, David M. Clarke, and John R. Riordan. « Cystic Fibrosis: Channel, Catalytic, and Folding Properties of the CFTR Protein ». *Journal of Bioenergetics and Biomembranes* 29.5 (1997), pp. 429–442 (cit. on p. 67).

[Sermet-Gaudelus, 2002] Isabelle Sermet-Gaudelus, Benoit Vallée, Ilse Urbin, Tania Torossi, Rémi Marianovski, Anne Fajac, et al. « Normal Function of the Cystic Fibrosis Conductance Regulator Protein Can Be Associated with Homozygous ΔF508 Mutation ». *Pediatric Research* 52.5 (2002), pp. 628–635 (cit. on p. 8).

[Shannon, 2003] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, et al. « Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks ». *Genome Research* 13.11 (2003), pp. 2498–2504 (cit. on p. 92).

[Simões, 2021] Filipa B. Simões, Arthur Kmit, and Margarida D. Amaral. « Cross-talk of inflammatory mediators and airway epithelium reveals the cystic fibrosis transmembrane conductance regulator as a major target ». *ERJ Open Research* 7.4 (2021), pp. 00247–2021 (cit. on p. 89).

[Smith, 1981] T.F. Smith and M.S. Waterman. « Identification of common molecular subsequences ». *Journal of Molecular Biology* 147.1 (1981), pp. 195–197 (cit. on pp. 118, 121).

[Stanke, 2014] Frauke Stanke, Andrea van Barneveld, Silke Hedtfeld, Stefan Wölfl, Tim Becker, and Burkhard Tümmler. « The CF-modifying gene EHF promotes p.Phe508del-CFTR residual function by altering protein glycosylation and trafficking in epithelial cells ». *European Journal of Human Genetics* 22.5 (2014), pp. 660–666 (cit. on p. 55).

[Stanke, 2011] Frauke Stanke, Silke Hedtfeld, Tim Becker, and Burkhard Tümmler. « An association study on contrasting cystic fibrosis endophenotypes recognizes KRT8 but not KRT18 as a modifier of cystic fibrosis disease severity and CFTR mediated residual chloride secretion ». *BMC Medical Genetics* 12.1 (2011), p. 62 (cit. on p. 97).

[Stark, 2019] Rory Stark, Marta Grzelak, and James Hadfield. « RNA sequencing: the teenage years ». *Nature Reviews Genetics* 20.11 (2019), pp. 631–656 (cit. on p. 57).

[Stelzer, 2016] Gil Stelzer, Naomi Rosen, Inbar Plaschkes, Shahar Zimmerman, Michal Twik, Simon Fishilevich, et al. « The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses ». *Current Protocols in Bioinformatics* 54.1 (2016), pp. 1.30.1–1.30.33 (cit. on p. 88).

[Stoll, 2017] Gautier Stoll, Barthélémy Caron, Eric Viara, Aurélien Dugourd, Andrei Zinovyev, Aurélien Naldi, et al. « MaBoSS 2.0: an environment for stochastic Boolean modeling ». *Bioinformatics* 33.14 (2017). Ed. by Jonathan Wren, pp. 2226–2228 (cit. on p. 100).

[Stoltz, 2015] David A Stoltz, David K Meyerholz, and Michael J Welsh. « Origins of Cystic Fibrosis Lung Disease ». *The New England journal of medicine* 372.4 (2015), pp. 351–362 (cit. on p. 6).

[Stoltz, 2010] David A. Stoltz, David K. Meyerholz, Alejandro A. Pezzulo, Shyam Ramachandran, Mark P. Rogan, Greg J. Davis, et al. « Cystic Fibrosis Pigs Develop Lung Disease and Exhibit Defective Bacterial Eradication at Birth ». *Science Translational Medicine* 2.29 (2010), 29ra31–29ra31 (cit. on p. 12).

[Strandvik, 2010] Birgitta Strandvik. « Fatty acid metabolism in cystic fibrosis ». *Prostaglandins, Leukotrienes and Essential Fatty Acids* 83.3 (2010), pp. 121–129 (cit. on p. 48).

[Strimbu, 2010] Kyle Strimbu and Jorge A. Tavel. « What are Biomarkers? » *Current opinion in HIV and AIDS* 5.6 (2010), pp. 463–466 (cit. on p. 9).

[Strub, 2021] Matthew D. Strub, Long Gao, Kai Tan, and Paul B. McCray. « Analysis of multiple gene co-expression networks to discover interactions favoring CFTR biogenesis and ΔF508-CFTR rescue ». *BMC Medical Genomics* 14.1 (2021), p. 258 (cit. on p. 60).

[Subramanian, 2005] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, et al. « Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles ». *Proceedings of the National Academy of Sciences* 102.43 (2005), pp. 15545–15550 (cit. on pp. 28, 50, 72, 90).

[Sun, 2019] Tao Sun, Zhe Sun, Yale Jiang, Annabel A. Ferguson, Joseph M. Pilewski, Jay K. Kolls, et al. « Transcriptomic Responses to Ivacaftor and Prediction of Ivacaftor Clinical Responsiveness ». *American Journal of Respiratory Cell and Molecular Biology* 61.5 (2019), pp. 643–652 (cit. on pp. 55, 56).

[Svetnik, 2003] Vladimir Svetnik, Andy Liaw, Christopher Tong, J. Christopher Culberson, Robert P. Sheridan, and Bradley P. Feuston. « Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling ». *Journal of Chemical Information and Computer Sciences* 43.6 (2003), pp. 1947–1958 (cit. on p. 116).

[Swamidass, 2005] S. J. Swamidass, J. Chen, J. Bruand, P. Phung, L. Ralaivola, and P. Baldi. « Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity ». *Bioinformatics* 21.Suppl 1 (2005), pp. i359–i368 (cit. on pp. 118, 121).

[Swinney, 2011] David C. Swinney and Jason Anthony. « How were new medicines discovered? » *Nature Reviews Drug Discovery* 10.7 (2011), pp. 507–519 (cit. on p. 115).

[Szalai, 2020] Bence Szalai and Julio Saez-Rodriguez. « Why do pathway methods work better than they should? » *FEBS Letters* 594.24 (2020), pp. 4189–4200 (cit. on pp. 27–30, 65, 95, 97).

[Szklarczyk, 2019] Damian Szklarczyk, Annika L Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, et al. « STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets ». *Nucleic Acids Research* 47.D1 (2019), pp. D607–D613 (cit. on pp. 23, 60, 96).

[Tang, 2016] Xiao Xiao Tang, Lynda S. Ostedgaard, Mark J. Hoegger, Thomas O. Moninger, Philip H. Karp, James D. McMenimen, et al. « Acidic pH increases airway surface liquid viscosity in cystic fibrosis ». *The Journal of Clinical Investigation* 126.3 (2016), pp. 879–891 (cit. on p. 5).

[Tarran, 2005] Robert Tarran, Brian Button, Maryse Picher, Anthony M. Paradiso, Carla M. Ribeiro, Eduardo R. Lazarowski, et al. « Normal and Cystic Fibrosis Airway Surface Liquid Homeostasis: THE EFFECTS OF PHASIC SHEAR STRESS AND VIRAL INFECTIONS* ». *Journal of Biological Chemistry* 280.42 (2005), pp. 35751–35759 (cit. on p. 5).

[Teng, 2012] Ling Teng, Mathieu Kerbiriou, Mehdi Taiya, Sophie Le Hir, Olivier Mignen, Nathalie Benz, et al. « Proteomic Identification of Calumenin as a G551D - CFTR Associated Protein ». *PLOS ONE* 7.6 (2012), e40173 (cit. on p. 58).

[Thurman, 2022] Andrew L. Thurman, Xiaopeng Li, Raul Villacreses, Wenjie Yu, Huiyu Gong, Steven E. Mather, et al. « A Single-Cell Atlas of Large and Small Airways at Birth in a Porcine Model of Cystic Fibrosis ». *American Journal of Respiratory Cell and Molecular Biology* 66.6 (2022), pp. 612–622 (cit. on p. 88).

[Tomfohr, 2005] John Tomfohr, Jun Lu, and Thomas B Kepler. « Pathway level analysis of gene expression using singular value decomposition ». *BMC Bioinformatics* 6 (2005), p. 225 (cit. on p. 36).

[Toro, 2022] Noemi del Toro, Anjali Shrivastava, Eliot Ragueneau, Birgit Meldal, Colin Combe, Elisabet Barrera, et al. « The IntAct database: efficient access to fine-grained molecular interaction data ». *Nucleic Acids Research* 50.D1 (2022), pp. D648–D653 (cit. on p. 23).

[Trezise, 1991] A. E. Trezise and M. Buchwald. « In vivo cell-specific expression of the cystic fibrosis transmembrane conductance regulator ». *Nature* 353.6343 (1991), pp. 434–437 (cit. on p. 6).

[Trivedi, 2023] Tithi S. Trivedi, Kinjal P. Bhadresha, Maulikkumar P. Patel, Archana U. Mankad, Rakesh M. Rawal, and Saumya K. Patel. « Identification of hub genes associated with human cystic fibrosis: A Meta-analysis approach ». *Human Gene* 35 (2023), p. 201139 (cit. on p. 68).

[Trzcińska-Daneluti, 2012] Agata M. Trzcińska-Daneluti, Leo Nguyen, Chong Jiang, Christopher Fladd, David Uehling, Michael Prakesch, et al. « Use of Kinase Inhibitors to Correct ΔF508-CFTR Function ». *Molecular & Cellular Proteomics : MCP* 11.9 (2012), pp. 745–757 (cit. on pp. 134, 141).

[Türei, 2016] Dénes Türei, Tamás Korcsmáros, and Julio Saez-Rodriguez. « OmniPath: guidelines and gateway for literature-curated signaling pathway resources ». *Nature Methods* 13.12 (2016), pp. 966–967 (cit. on pp. 24, 76, 90).

[Valdivieso, 2018] Ángel G. Valdivieso, Andrea V. Dugour, Verónica Sotomayor, Mariángeles Clauzure, Juan M. Figueroa, and Tomás A. Santa-Coloma. « N-acetyl cysteine reverts the proinflammatory state induced by cigarette smoke extract in lung Calu-3 cells ». *Redox Biology* 16 (2018), pp. 294–302 (cit. on pp. 16, 89).

[Vasconcellos, 1994] Carol A. Vasconcellos, Philip G. Allen, Mary Ellen Wohl, Jeffrey M. Drazen, Paul A. Janmey, and Thomas P. Stossel. « Reduction in Viscosity of Cystic Fibrosis Sputum in Vitro by Gelsolin ». *Science* 263.5149 (1994), pp. 969–971 (cit. on pp. 45, 48).

[Vaske, 2010] Charles J. Vaske, Stephen C. Benz, J. Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, et al. « Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM ». *Bioinformatics* 26.12 (2010), pp. i237–i245 (cit. on p. 72).

[Veit, 2016] Gudio Veit, Radu G. Avramescu, Annette N. Chiang, Scott A. Houck, Zhiwei Cai, Kathryn W. Peters, et al. « From CFTR biology toward combinatorial pharmacotherapy: expanded classification of cystic fibrosis mutations ». *Molecular Biology of the Cell* 27.3 (2016), pp. 424–433 (cit. on p. 67).

[Veltman, 2021] Mieke Veltman, Juan B. De Sanctis, Marta Stolarczyk, Nikolai Klymiuk, Andrea Bähr, Rutger W. Brouwer, et al. « CFTR Correctors and Antioxidants Partially Normalize Lipid Imbalance but not Abnormal Basal Inflammatory Cytokine Profile in CF Bronchial Epithelial Cells ». *Frontiers in Physiology* 12 (2021) (cit. on pp. 58, 89).

[Venerando, 2011] Andrea Venerando, Mario A. Pagano, Kendra Tosoni, Flavio Meggio, Diane Cassidy, Michelle Stobbart, et al. « Understanding protein kinase CK2 mis-regulation upon F508del CFTR expression ». *Naunyn-Schmiedeberg's Archives of Pharmacology* 384.4 (2011), pp. 473–488 (cit. on p. 79).

[Vergani, 2005] Paola Vergani, Steve W. Lockless, Angus C. Nairn, and David C. Gadsby. « CFTR channel opening by ATP-driven tight dimerization of its nucleotide-binding domains ». *Nature* 433.7028 (2005), pp. 876–880 (cit. on p. 7).

[Verhaeghe, 2007] Catherine Verhaeghe, Caroline Remouchamps, Benoît Hennuy, Alain Vanderplasschen, Alain Chariot, Sebastien P. Tabruyn, et al. « Role of IKK and ERK pathways in intrinsic inflammation of cystic fibrosis airways ». *Biochemical Pharmacology* 73.12 (2007), pp. 1982–1994 (cit. on pp. 55, 71, 133).

[Vert, 2008] Jean-Philippe Vert and Laurent Jacob. « Machine Learning for In Silico Virtual Screening and Chemical Genomics: New Strategies ». *Combinatorial Chemistry & High Throughput Screening* 11.8 (2008), pp. 677–685 (cit. on p. 115).

[Vieira Braga, 2019] Felipe A. Vieira Braga, Gozde Kar, Marijn Berg, Orestes A. Carpaij, Krzysztof Polanski, Lukas M. Simon, et al. « A cellular census of human lungs identifies novel cell states in health and in asthma ». *Nature Medicine* 25.7 (2019), pp. 1153–1163 (cit. on p. 57).

[Vinhoven, 2021] Liza Vinhoven, Frauke Stanke, Sylvia Hafkemeyer, and Manuel Manfred Nietert. « CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations ». *International Journal of Molecular Sciences* 22.14 (2021), p. 7590 (cit. on p. 61).

[Virella-Lowell, 2004] Isabel Virella-Lowell, John-David Herlihy, Barry Liu, Cecilia Lopez, Pedro Cruz, Chris Muller, et al. « Effects of CFTR, interleukin-10, and Pseudomonas aeruginosa on gene expression profiles in a CF bronchial epithelial cell Line ». *Molecular Therapy* 10.3 (2004), pp. 562–573 (cit. on pp. 55, 71, 89).

[Voisin, 2014] Grégory Voisin, Guillaume F. Bouvet, Pierre Legendre, André Dagenais, Chantal Massé, and Yves Berthiaume. « Oxidative stress modulates the expression of genes involved in cell survival in ΔF508 cystic fibrosis airway epithelial cells ». *Physiological Genomics* 46.17 (2014), pp. 634–646 (cit. on pp. 55, 71, 81).

[Wainwright, 2015] Claire E. Wainwright, J. Stuart Elborn, Bonnie W. Ramsey, Gautham Marigowda, Xiaohong Huang, Marco Cipolli, et al. « Lumacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR ». *New England Journal of Medicine* 373.3 (2015), pp. 220–231 (cit. on p. 11).

[Wang, 2016] Hua Wang, Liudmila Cebotaru, Ha Won Lee, QingFeng Yang, Bette S. Pollard, Harvey B. Pollard, et al. « CFTR Controls the Activity of NF-ϰB by Enhancing the Degradation of TRADD ». *Cellular Physiology and Biochemistry* 40.5 (2016), pp. 1063–1078 (cit. on pp. 80, 84).

[Wang, 2010] Kai Wang, Mingyao Li, and Hakon Hakonarson. « Analysing biological pathways in genome-wide association studies ». *Nature Reviews Genetics* 11.12 (2010), pp. 843–854 (cit. on pp. 28, 35, 72).

[Wang, 2006] Xiaodong Wang, John Venable, Paul LaPointe, Darren M. Hutt, Atanas V. Koulov, Judith Coppinger, et al. « Hsp90 Cochaperone Aha1 Downregulation Rescues Misfolding of CFTR in Cystic Fibrosis ». *Cell* 127.4 (2006), pp. 803–815 (cit. on pp. 58, 61).

[Wang, 2011] Yong-Cui Wang, Chun-Hua Zhang, Nai-Yang Deng, and Yong Wang. « Kernel-based data fusion improves the drug–protein interaction prediction ». *Computational Biology and Chemistry* 35.6 (2011), pp. 353–362 (cit. on p. 121).

[Wang, 2022] Yue Wang, Lu Tang, Liangliang Yang, Peiyun Lv, Shixiong Mai, Li Xu, et al. « DNA Methylation-Mediated Low Expression of CFTR Stimulates the Progression of Lung Adenocarcinoma ». *Biochemical Genetics* 60.2 (2022), pp. 807–821 (cit. on p. 89).

[Ward, 1995] Cristina L. Ward, Satoshi Omura, and Ron R. Kopito. « Degradation of CFTR by the ubiquitin-proteasome pathway ». *Cell* 83.1 (1995), pp. 121–127 (cit. on p. 7).

[Wellmerling, 2022] Jack Wellmerling, Rachael E. Rayner, Sheng-Wei Chang, Elizabeth L. Kairis, Sun Hee Kim, Amit Sharma, et al. « Targeting the EGFR-ERK axis using the compatible solute ectoine to stabilize CFTR mutant F508del ». *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* 36.5 (2022), e22270 (cit. on pp. 81, 141).

[Wong, 2004] Brian R Wong, Elliott B Grossbard, Donald G Payan, and Esteban S Masuda. « Targeting Syk as a treatment for allergic and autoimmune disorders ». *Expert Opinion on Investigational Drugs* 13.7 (2004), pp. 743–762 (cit. on p. 84).

[Worgall, 2005] S. Worgall, A. Heguy, K. Luettich, T. P. O'Connor, B.-G. Harvey, L. E. N. Quadri, et al. « Similarity of Gene Expression Patterns in Human Alveolar Macrophages in Response to Pseudomonas aeruginosa and Burkholderia cepacia ». *Infection and Immunity* 73.8 (2005), pp. 5262–5268 (cit. on p. 55).

[Wright, 2011] Fred A. Wright, Lisa J. Strug, Vishal K. Doshi, Clayton W. Commander, Scott M. Blackman, Lei Sun, et al. « Genome-wide association and linkage identify modifier loci of lung disease severity in cystic fibrosis at 11p13 and 20q13.2 ». *Nature Genetics* 43.6 (2011), pp. 539–546 (cit. on p. 13).

[Wright, 2006] Jerry M. Wright, Christian A. Merlo, Jeffrey B. Reynolds, Pamela L. Zeitlin, Joe G. N. Garcia, William B. Guggino, et al. « Respiratory epithelial gene expression in patients with mild and severe cystic fibrosis lung disease ». *American Journal of Respiratory Cell and Molecular Biology* 35.3 (2006), pp. 327–336 (cit. on pp. 55, 89).

[Wu, 2019] Qun Wu and Oliver Eickelberg. « Ezrin in Asthma: A First Step to Early Biomarkers of Airway Epithelial Dysfunction ». *American Journal of Respiratory and Critical Care Medicine* 199.4 (2019), pp. 408–410 (cit. on p. 80).

[Xia, 2017] Xian Xia, Jie Wang, Yuan Liu, and Ming Yue. « Lower Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Promotes the Proliferation and Migration of Endometrial Carcinoma ». *Medical Science Monitor* 23 (2017), pp. 966–974 (cit. on p. 16).

[Xu, 2018] Xianjin Xu, Marshal Huang, and Xiaoqin Zou. « Docking-based inverse virtual screening: methods, applications, and challenges ». *Biophysics Reports* 4.1 (2018), pp. 1–16 (cit. on p. 115).

[Xu, 2015] Xiaohua Xu, Robert Balsiger, Jean Tyrrell, Prosper N. Boyaka, Robert Tarran, and Estelle Cormet-Boyaka. « Cigarette smoke exposure reveals a novel role for the MEK/ERK1/2 MAPK pathway in regulation of CFTR ». *Biochimica et Biophysica Acta (BBA) - General Subjects* 1850.6 (2015), pp. 1224–1232 (cit. on p. 141).

[Yalçin, 2009] Ebru Yalçin, Beril Talim, Uğur Özçelik, Deniz Doğru, Nazan Çobanoğlu, Sevgi Pekcan, et al. « Does Defective Apoptosis Play A Role in Cystic Fibrosis Lung Disease? » *Archives of Medical Research* 40.7 (2009), pp. 561–564 (cit. on p. 81).

[Yamada, 2021] Ryo Yamada, Daigo Okada, Juan Wang, Tapati Basak, and Satoshi Koyama. « Interpretation of omics data analyses ». *Journal of Human Genetics* 66.1 (2021), pp. 93–102 (cit. on p. 27).

[Yamanishi, 2011] Yoshihiro Yamanishi, Edouard Pauwels, Hiroto Saigo, and Véronique Stoven. « Extracting Sets of Chemical Substructures and Protein Domains Governing Drug-Target Interactions ». *Journal of Chemical Information and Modeling* 51.5 (2011), pp. 1183–1194 (cit. on p. 116).

[Zabner, 2005] Joseph Zabner, Todd E. Scheetz, Hakeem G. Almabrazi, Thomas L. Casavant, Jian Huang, Shaf Keshavjee, et al. « CFTR ΔF508 mutation has minimal effect on the gene expression profile of differentiated human airway epithelia ». *American Journal of Physiology-Lung Cellular and Molecular Physiology* 289.4 (2005), pp. L545–L553 (cit. on pp. 55, 89).

[Zeitlin, 2017] Pamela L. Zeitlin, Marie Diener-West, Karen A. Callahan, Seakwoo Lee, C. Conover Talbot, Bette Pollard, et al. « Digitoxin for Airway Inflammation in Cystic Fibrosis: Preliminary Assessment of Safety, Pharmacokinetics, and Dose Finding ». *Annals of the American Thoracic Society* 14.2 (2017), pp. 220–229 (cit. on p. 55).

[Zhang, 2013] Jie Ting Zhang, Xiao Hua Jiang, Chen Xie, Hong Cheng, Jian Da Dong, Yan Wang, et al. « Downregulation of CFTR promotes epithelial-to-mesenchymal transition and is associated with poor prognosis of breast cancer ». *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* 1833.12 (2013), pp. 2961–2969 (cit. on p. 16).

[Zoso, 2019] Alice Zoso, Aderonke Sofoluwe, Marc Bacchetta, and Marc Chanson. « Transcriptomic profile of cystic fibrosis airway epithelial cells undergoing repair ». *Scientific Data* 6.1 (2019), pp. 1–7 (cit. on pp. 55, 71, 96).

RÉSUMÉ

La mucoviscidose est la maladie autosomale grave la plus fréquente dans la population caucasienne. Elle est causée par des mutations du gène codant pour la protéine CFTR (Cystic Fibrosis Transmembrane Regulator), qui agit comme un canal de chlorure (Cl^-) à la membrane des cellules épithéliales. La mucoviscidose est principalement délétère pour les poumons, où l'infection chronique et les lésions tissulaires provoquent progressivement une insuffisance respiratoire. Plus de 2000 mutations sont connues pour le gène CFTR, mais 70% des patients sont homozygotes pour la délétion du résidu F508 (F508del). Le traitement de la mucoviscidose est resté longtemps symptomatique, mais des modulateurs pharmacologiques de CFTR sont disponibles depuis peu. Cependant, ils ont un effet limité chez les patients homozygotes F508del et n'arrêtent pas l'évolution de la maladie. De plus, ils restent spécifiques à certaines mutations, et environ 15% des patients ne peuvent pas en bénéficier. Enfin, leurs cibles protéiques, leurs mécanismes d'action et leurs effets secondaires à long terme sont encore inconnus. Par ailleurs, la pathophysiologie globale de la mucoviscidose ne peut être expliquée uniquement par la perte de la fonction du canal chlorure CFTR. Notre hypothèse est que CFTR appartient à un réseau de protéines qui n'ont pas encore été toutes identifiées et dont les fonctions sont perturbées par l'absence de CFTR, participant ainsi à certains des phénotypes cellulaires anormaux qui caractérisent la maladie. En utilisant des approches de biologie des systèmes et des méthodes d'apprentissage automatique chémogénomique, les objectifs du projet sont les suivants : (1) identifier in-silico des cibles thérapeutiques candidates en construisant le réseau des dérégulations moléculaires de la mucoviscidose causées par l'absence de CFTR à l'aide de données transcriptomiques; (2) identifier les cibles protéiques des modulateurs de CFTR afin de déchiffrer leurs mécanismes d'action. À terme, le projet devrait permettre d'identifier de nouvelles stratégies thérapeutiques combinant des médicaments ciblant la restauration de la maturation et de la fonction de CFTR, à des médicaments ciblant le réseau de dérégulations de la maladie. Cette approche systémique pourrait apporter des solutions thérapeutiques aux patients présentant des mutations pour lesquelles il n'existe actuellement aucune thérapie.

MOTS CLÉS

mucoviscidose, biologie des systèmes, chemogenomics, machine-learning, cibles thérapeutiques, transcriptomics

ABSTRACT

Cystic Fibrosis (CF) is the most frequent life-limiting autosomal disease in the Caucasian population. It is caused by mutations in the gene coding the Cystic Fibrosis Transmembrane Regulator (CFTR) protein, acting as a chloride (Cl^-) channel at the membrane of epithelial cells. CF is mainly deleterious for the lung where chronic infection and tissue damage progressively cause respiratory insufficiency. More than 2000 mutations are known in the CFTR gene, but 70% of the patients are homozygous for the deletion of residue F508 (F508del). CF treatment remained symptomatic for a long time, but pharmacologic CFTR modulators became recently available. However, they have a limited effect in F508del homozygous patients, and do not stop disease evolution. They remain mutation specific, and around 15% of CF patients cannot benefit. Moreover, their protein targets, mechanisms of action and long-term side effects are still unknown. In addition, CF overall physiopathology cannot be solely explained by the loss of the CFTR chloride channel function. Our hypothesis is that CFTR belongs to a yet not fully deciphered network of proteins, whose functions are disrupted by the absence of CFTR, thus participating in some of the abnormal cellular phenotypes that characterise CF. Using systems biology approaches and machine-learning chemogenomics methods, the aims of the project are to: (1) identify in-silico candidate therapeutic targets by building the network of CF molecular dysregulations caused by the absence of CFTR based on transcriptomic data; (2) identify protein targets of CFTR modulators to decipher their mechanisms of action. At term, the project should help identify new therapeutic strategies combining drugs targeting restoration of CFTR maturation and function, to drugs targeting the network of CF molecular dysregulations. This systemic approach may provide therapeutic solutions for CF patients with mutations for which there is currently no specific therapy.

KEYWORDS

cystic fibrosis, systems biology, chemogenomics, machine-learning, therapeutic targets, transcriptomics