Cystic Fibrosis as a model use case for implementing cell based disease models in systems medicine

Dissertation

for the award of the degree

Doctor rerum naturalium (Dr. rer. nat.)

of the Georg-August-Universität Göttingen

within the doctoral program *Molecular Biology of Cells*
of the Georg-August University School of Science (GAUSS)

Submitted by

Liza Vinhoven

from Frankfurt am Main

Göttingen, 2022

Thesis Committee

Dr. Manuel Nietert

Department of Medical Bioinformatics
University Medical Center Göttingen

Prof. Dr. Wolfram-Hubertus Zimmermann

Institute of Pharmacology and Toxicology
University Medical Center Göttingen

Dr. Alexander Stein

Research Group Membrane Protein Biochemistry
Max Planck Institute for Biophysical Chemistry

Members of the Examination Board

1st Referee: Dr. Manuel Nietert

Department of Medical Bioinformatics
University Medical Center Göttingen

2nd Referee: Prof. Dr. Wolfram-Hubertus Zimmermann

Institute of Pharmacology and Toxicology
University Medical Center Göttingen

Further members of the Examination Board

Dr. Alexander Stein

Research Group Membrane Protein Biochemistry
Max Planck Institute for Biophysical Chemistry

Prof. Dr. Burkhard Morgenstern

Department of Bioinformatics
Georg-August University of Göttingen

Dr. Nico Posnien

Department of Developmental Biology
Johann-Friedrich-Blumenbach-Institute of Zoology and Anthropology
Georg-August University of Göttingen

Prof. Dr. Ulrich Sax

Department of Medical Informatics
University Medical Center Göttingen

Date of oral examination: 10th of November 2022

I hereby declare that the PhD thesis “Cystic Fibrosis as a model use case for implementing cell based disease models in systems medicine” is my own work that was prepared with no other sources and aids than quoted. This thesis, or parts thereof, have not been submitted elsewhere.

Liza Vinhoven, September 2022, Göttingen

ABSTRACT

In the last two decades, tremendous progress has been made in producing large amounts of biological and biomedical data in a time- and cost-effective manner. Along with the increasing amount of data has come the requirement for methods to analyze and interpret it effectively. This lead to a multitude of computational methods being developed, often rather sophisticated and highly specific to the datatype and -source. However, in order to utilize the wealth of data to its full potential in systems medicine, it is essential to bring the different data sources together and use approaches, that can easily be applied and adapted to different use cases.

Therefore, using the example of Cystic Fibrosis, this thesis focuses on applying generic systems medicine methods to integrate different kinds of data and thereby create a holistic overview of the disease and gain new insights. Cystic Fibrosis is one of the most common genetic diseases prevalent among the white European population. Its vast range of geno- and phenotypes makes the development of therapeutics especially challenging. During the last years, different small-molecule therapeutics have been developed that amplify CFTR function, but they are not effective for all patients. The latest research efforts, therefore, focus on developing combination therapies to target multiple defects at once.

To provide an overview of already tested compounds, I contributed to establishing the publicly available database CandActCFTR, where substances are listed and categorized according to their interaction with CFTR. It becomes apparent that for the majority of compounds it is unknown whether they affect CFTR directly or indirectly. To elucidate the mechanism of action for promising candidate substances and be able to predict possible synergistic effects of substance combinations, I created a systems medicine disease map of the CFTR biogenesis, function and interactions. In order to support the manual curation and upkeep of disease maps, a tool was developed to integrate text mining approaches into the disease map curation. The tool allows the user to iterate through the text mined interactions to validate the results, thereby bringing together the speed of text mining and the accuracy of scientific expert knowledge.

To bring together the chemical knowledge from the database and the biological pathways from the disease map, an interlinking tool was developed, to interactively

and computationally map compounds to their respective targets based on publically available interaction data.

This data, however, still leaves the mechanism of action for the majority of the active compounds unexplained. Therefore, in order to suggest possible modes of action for all active compounds in the database, I used two complementary *in silico* target identification approaches, namely target-based molecular docking and ligand-based similarity searches. I thereby identified possible targets for all active compounds, which will help to understand which compound classes affect CFTR at which stage of its life cycle and which compounds can be combined to alleviate different defects in its biogenesis.

All parts of the project can be seen individually as stand-alone resources and be combined to yield new findings. Since the approaches are generic and easily adaptable, they can also be applied to other disease besides Cystic Fibrosis in a similar manner, to give a holistic, systems medicine approach to answer research questions relevant to them.

ZUSAMMENFASSUNG

In den letzten zwei Jahrzehnten wurden enorme Fortschritte bei der zeit- und kosteneffizienten Erzeugung großer Mengen biologischer und biomedizinischer Daten erzielt. Mit der zunehmenden Datenmenge steigt auch der Bedarf an Methoden zu deren effektiver Analyse und Interpretation. Dies hat zur Entwicklung einer Vielzahl von Berechnungsmethoden geführt, die oft anspruchsvoll und sehr spezifisch für den jeweiligen Datentyp und die Datenquelle sind. Um jedoch die Fülle der Daten in der Systemmedizin voll ausschöpfen zu können, müssen die verschiedenen Datenquellen zusammengeführt und Ansätze verwendet werden, die sich leicht anwenden und an verschiedene Anwendungsfälle anpassen lassen.

Daher konzentriert sich diese Arbeit am Beispiel der Mukoviszidose auf die Anwendung generischer Methoden der Systemmedizin, um verschiedene Arten von Daten zu integrieren und dadurch einen ganzheitlichen Überblick über die Krankheit zu schaffen, und neue Erkenntnisse zu gewinnen. Mukoviszidose ist eine der häufigsten genetischen Krankheiten in der Bevölkerung nordeuropäischer Abstammung. Ihre große Bandbreite an Geno- und Phänotypen macht die Entwicklung von Therapeutika zu einer besonderen Herausforderung. In den letzten Jahren wurden verschiedene kleinmolekulare Therapeutika entwickelt, die die CFTR-Funktion verstärken, aber nicht bei allen Patienten wirksam sind. Die jüngsten Forschungsbemühungen konzentrieren sich daher auf die Entwicklung von Kombinationstherapien, die auf mehrere Defekte gleichzeitig abzielen.

Um einen Überblick über die bereits getesteten Wirkstoffe zu geben, habe ich an der Erstellung der öffentlich zugänglichen Datenbank CandActCFTR mitgewirkt, in der die Substanzen nach ihrer Interaktion mit CFTR aufgelistet und kategorisiert sind. Es zeigt sich, dass bei dem Großteil der Substanzen nicht bekannt ist, ob sie CFTR direkt oder indirekt beeinflussen. Um den Wirkmechanismus vielversprechender Wirkstoffkandidaten aufzuklären und mögliche synergistische Effekte von Substanzkombinationen vorhersagen zu können, habe ich ein systemmedizinisches Modell der CFTR-Biogenese, -Funktion und -Interaktionen erstellt. Um die manuelle Erstellung und Pflege solcher Modelle zu unterstützen, wurde ein Tool entwickelt, das Text-Mining-Ansätze in den Erstellungprozess integriert. Das Tool ermöglicht es dem

Benutzer, durch die im Text Mining identifizierten Interaktionen zu iterieren, um die Ergebnisse zu validieren, wodurch die Geschwindigkeit des Text Mining mit der Genauigkeit von wissenschaftlichem Expertenwissen kombiniert wird.

Um das chemische Wissen aus der Datenbank und die biologischen Zusammenhänge aus dem Modell zusammenzubringen, wurde ein Tool entwickelt, das die Substanzen auf Basis öffentlich zugänglicher Interaktionsdaten interaktiv ihren jeweiligen biologischen Zielen zuordnet.

Für einen Großteil der Wirkstoffe lassen sich durch diese Daten jedoch keine Rückschlüsse auf den Wirkmechanismus ziehen. Um potentielle Wirkmechanismen für alle Wirkstoffe in der Datenbank vorzuschlagen, habe ich daher zwei komplementäre *in silico* Ansätze zur Identifizierung von Targets verwendet: Struktur-basiertes molekulares Docking und Liganden-basierte Ähnlichkeitssuche. Auf diese Weise konnten mögliche Targets für alle Wirkstoffe identifiziert weden, was dazu beitragen wird, zu verstehen, welche Verbindungsklassen CFTR in welchem Stadium seines Lebenszyklus beeinflussen und welche Verbindungen kombiniert werden können, um verschiedene Defekte in seiner Biogenese zu lindern.

Alle Teile des Projekts können einzeln als eigenständige Ressourcen betrachtet werden, oder kombiniert werden, um neue Erkenntnisse zu gewinnen. Da die Ansätze generisch und leicht adaptierbar sind, können sie auch auf andere Krankheiten ähnlicher Weise angewandt werden, um einen ganzheitlichen, systemmedizinischen Ansatz zur Beantwortung der für sie relevanten Forschungsfragen zu bieten.

ACKNOWLEDGEMENTS

First and foremost, I to express my gratitude to my supervisor Dr. Manuel Nietert for the opportunity to work on this great project first as a research assistant and then as a PhD student. He was the best mentor I could have hoped for and truly taught me a lot over the past years, not only scientifically, but also about the inner workings of academia and work life in general. Thank you for all the long talks and discussions, all the encouragement, and the support in pursuing my own ideas.

Furthermore, I want to thank Prof. Dr. Michael Meinecke and Dr. Alexander Stein for being on my thesis committee and Prof. Dr. Wolfram Zimmermann for taking over as my second referee in the final year of my PhD. Thank you for the valuable input and discussions during the TAC meetings. Thanks also go out to Prof. Dr. Ulrich Sax, Prof. Dr. Burkhard Morgenstern, and Dr. Nico Posnien for taking the time to be part of my examination committee.

I am especially grateful to our project partners. Thanks to PD Dr. Frauke Stanke from the MHH for sharing her extensive knowledge on CF and the great impulses, for all the interesting discussions and advice on scientific writing, and taking the time for every question. Great thanks also go to Dr. Sylvia Hafkemeyer, for the insights into the CF research landscape and for providing the framework for this project in the community.

Special thanks go to Prof. Dr. Tim Beißbarth for the opportunity to work in his department and for the regular discussions.

I also want to thank the rest of the Department of Medical Bioinformatics for the great time and the pleasant working atmosphere. I especially want to thank Yvonne for being there for all administrative concerns, Torsten for all the technical help, and Daniela for the insights into academic project management and the interesting and entertaining Friday morning meetings. Special thanks also go to Hryhorii, for all the fun coffee- and tea breaks, for sharing his PhD experiences, and especially for proofreading my thesis. Thank you also to Niels, for being my university companion and accompanying me on this journey from the first day of my Bachelor's studies to the final day of my PhD. I am very grateful to my Master's student Malte for the great contributions, I truly enjoyed working with you.

Many thanks go to the btS Göttingen. I learned a lot of new things and really broadened my horizon over the years with you and I am proud of all the projects we could realize together. I also thoroughly enjoyed all our meetings, events, and other activities. I want to especially thank Malena, Krishya, Paul, Naomi, Julia, and Leon for the amazing past couple of years.

My most heartfelt thanks go to Tobias and my family. Special thanks go to Tobias, for proofreading my thesis and providing me with a much-needed outside view. I want to thank

you for all your patience, encouragement, reassurance, and support. I am deeply grateful to Leah and Lana for proofreading not only this thesis but every important piece of writing since my school days. Thank you for your unconditional support and for being my personal cheerleaders. Thank you to my grandfather for encouraging me to be inquisitive since my childhood and always supporting me. Finally, I am greatly indebted to my mother, without whom I would have never been able to pursue this course of education and field of study. Thank you for supporting me in every way possible in all my endeavours.

TABLE OF CONTENTS

LIST OF FIGURES

ABBREVIATIONS

2D	*2-dimensional*
3D	*3-dimensional*
ABC	*ATP binding cassette*
AF	*Activity Flow*
AFT	*Arginine-framed tripeptides*
AI	*Artificial intelligence*
ASL	*Airway surface liquid*
ATP	*Adenosine triphosphate*
BRENDA	*Braunschweig Enzyme Database*
CANX	*Calnexin*
CDK2	*Cyclin-dependent kinase 2*
CF	*Cystic Fibrosis*
CFTR	*Cystic fibrosis transmembrane conductance regulator*
COPI	*Coat protein I*
COVID-19	*Coronavirus disease 2019*
CRE	*cAMP response element*
CTD	*The Comparative Toxicogenomics Database*
DNA	*Deoxyribonucleic acid*
DNJA1	*DnaJ homolog subfamily A member 1*
DTP	*Developmental Therapeutics Program*
EBI	*European Bioinformatics Institute*
ELX-02	*Exaluren*
EMBL	*European Molecular Biology Laboratory*
ENaC	*Amiloride-sensitive sodium channel*
ER	*Endoplasmic reticulum*
F508del	*CFTR variant (deletion of phenylalanine at position 508)*
G542X	*CFTR variant (change of glycine at position 542, introduces a stop codon)*
G551D	*CFTR variant (substitution of glycine by aspartate at position 551)*
GERAD	*Glycoprotein endoplasmic reticulum-associated degradation*

H_2O	*Water*
HCO_3^-	*Bicarbonate ion*
HSP7C	*Heat shock cognate 71 kDa protein*
HT	*High throughput*
HTS	*High throughput screen*
MINERVA	*Molecular Interaction Network Visualization Platform*
MIRIAM	*Minimum information required in the annotation of models*
MSD	*Membrane spanning domain*
N1303K	*CFTR variant (substitution of asparagine by lysine at position 1303)*
NBD	*Nucleotide binding domain*
ODE	*ordinary differential equation*
PAINS	*Pan-assay interference compounds*
PD	*Process Description*
PM	*Plasma membrane*
PRKACA	*(Protein Kinase CAMP-Activated Catalytic Subunit Alpha*
PTC	*Premature termination codon*
R117H	*CFTR variant (substitution of arginine by histidine at position 117*
RNA	*Ribonucleic acid*
SBGN	*Systems Biology Graphical Notation*
SBML	*Systems Biology Markup Language*
SLC26A9	*Solute Carrier Family 26 Member 9*
TMEM16A	*Transmembrane member 16A*
W1282X	*CFTR variant (change of tryptophan at position 1282, introduces a stop codon)*
XML	*Extensible Markup Language*

Chapter 1 Introduction

1.1 Cystic Fibrosis

"Woe to the child who tastes salty from a kiss on the brow, for he is cursed and soon will die."

mid-17th century Northern European folklore

This medieval proverb from Northern European folklore is considered to be the first description of Cystic Fibrosis (CF), which today affects approximately one in 3000 new-borns amongst people of white European ancestry, making it the most prevalent monogenic autosomal recessive disorder in this population (Bobadilla *et al.*, 2002; Farrell, 2008; Bell *et al.*, 2020). The proverb refers to the elevated salt concentration in the sweat of people with CF, which is one of its major hallmarks and is still used for diagnostic purposes until today (Di Sant'Agnese *et al.*, 1953; Quinton, 1999; Elborn, 2016). CF is caused by mutation of the *cftr*-gene, which encodes the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) protein, a chloride- and bicarbonate ion channel expressed in the apical membrane of epithelial cells in exocrine glands throughout the body. This makes CF a multi-organ disease, affecting, apart from the sweat glands, the pancreas, liver, intestines and reproductive system (Riordan *et al.*, 1989; O'Sullivan and Freedman, 2009; Elborn, 2016). The most well-known hallmarks of CF, however, are its effects on the lungs and upper respiratory system. Normally, fully functional CFTR transports Cl^- ions from the inside of the cytoplasm to the lumen, thereby increasing the salt concentration in the airway surface liquid (ASL). As a consequence, the osmotic pressure increases, and water from the inside of the cell diffuses to the outside, keeping it liquid, which allows a process called mucociliary clearance. During mucociliary clearance, the ASL and the atop lying mucous are being moved across the airway surface by the beating cilia, which are hair-like membrane-bound organelles that extend from the cell surface. In the respiratory system, mucous is responsible for trapping pathogens of all kinds, which makes it possible to clear them out by e.g. coughing (Tarran *et al.*, 2005; Proesmans, Vermeulen and De Boeck, 2008). In CF the CFTR ion channel is defect

due to mutation, which impairs Cl^- secretion, hence no water diffuses to the ASL. Additionally, functional CFTR inhibits Na^+ absorption through the epithelial sodium channel ENaC. Therefore, defect CFTR further increases intracellular salt concentration and depletes water from the ASL through unregulated Na^+ absorbance. As a result, the now highly viscous mucous cannot be transported across the airway surface any longer and accumulates there. This not only severely obstructs the airways but gives rise to inflammation and bronchiectasis and allows pathogens to infest the lungs. This results in a vicious cycle of mucous obstruction, chronic infection and inflammation, which leads to fibrosis and ultimately respiratory failure (Tarran *et al.*, 2005; Proesmans, Vermeulen and De Boeck, 2008; Dechecchi, Tamanini and Cabrini, 2018; Lopes-Pacheco, 2020). Another hallmark with an early onset is exocrine pancreas insufficiency, which is also caused by mucus obstruction and the resulting inability to release digestive enzymes. This leads to malabsorption and low weight gain, which was the main cause of death in infants with CF before treatments became available (Busch, 1979; Quinton, 1999; Elborn, 2016). CF lung disease and pancreatic insufficiency were the main hallmarks of CF in past times, but with the advance in care and treatments available, life expectancy increased and other morbidities, such as CF-related diabetes and CF-related liver disease, came to the forefront (Bell *et al.*, 2020).

1.1.1 CFTR protein and mutations

The CFTR protein is located in the apical plasma membrane of epithelial cells. It is composed of one 1440 amino acid chain and complex glycosylation, resulting in a 170 kDa membrane-spanning glycoprotein (Riordan *et al.*, 1989; Cheng *et al.*, 1990; O'Riordan *et al.*, 2000). CFTR belongs to the superfamily of ABC (ATP-binding cassette) transporters, a ubiquitous family of integral membrane proteins (Higgins *et al.*, 1988; Higgins, 1989). Like most other ABC transporters, CFTR is composed of four domains, two membrane-spanning domains (MSD1 and MSD2), which make up the channel pore, and two cytosolic nucleotide binding domains (NBD1 and NBD2). MSD1 and NBD1, and MSD2 and NBD2 respectively, make up the two homologous halves of CFTR. The two halves are additionally connected by the highly flexible regulatory region (R-region) between NBD1 and MSD2, which controls channel opening and is unique to CFTR (Zhang, Liu and Chen, 2018; Csanády, Vergani and

Gadsby, 2019; Meng *et al.*, 2019). CFTR, as opposed to the remainder of the ABC transporter superfamily, acts as an ion channel instead of an active transporter. While it does require ATP for opening, it then forms a channel pore for the substrate to passively diffuse through instead of actively transporting it across the membrane (Aleksandrov, Aleksandrov and Riordan, 2007; Moran, 2017; Csanády, Vergani and Gadsby, 2019).

The two membrane-spanning domains are each made up of six α-helices, which pack compactly in the outer leaflet of the plasma membrane (PM) and separate into two bundles in its inner leaflet, which extend into the cytoplasm, and attach to the respective NBD (Callebaut, Chong and Forman-Kay, 2018; Zhang, Liu and Chen, 2018; Meng *et al.*, 2019).

More than 2000 mutations of the *cftr* gene have been reported to date, several hundred of which have been shown to be disease causing by affecting the CFTR protein. Disease causing mutations are not exclusive to specific areas of the protein, but are spread throughout the entire structure, and can cause different kinds of defects during the intricate and error-prone biogenesis of CFTR (*Cystic Fibrosis Mutation Database*, no date; *Welcome to CFTR2 | CFTR2*, no date; Sosnay *et al.*, 2013). As originally proposed by Welsh and Smith in 1993, to make working with the multitude of mutations easier, they have been traditionally categorized into six distinct classes, according the kind of basic defect they cause (Figure 1) (Welsh and Smith, 1993).

- *Class I* mutations cause reduced CFTR expression through premature termination codons caused by nonsense mutations, frameshifts, and splicing mutations.
- *Class II* mutations are folding mutations, that result in premature degradation through ER-associated degradation or other complications during the biogenesis of CFTR.
- *Class III* and *IV* mutations both lead to a reduced conductance. Class III mutations do so by hampering channel regulation and thereby decreasing its open probabilities, while class IV mutations obstruct the pore itself.
- *Class V* mutations encompass splicing or promoter mutations that reduce protein abundance.
- *Class VI* mutations reduce protein stability, thereby increasing degradation at the PM and also reducing the amount of functional protein.

This classification system has been in use for almost three decades and has since been universally accepted, reviewed extensively and adapted by the community (Zielenski and Tsui, 1995; Zielenski, 2000; Rowe, Miller and Sorscher, 2005). The biggest modification has been proposed by Veit *et al.* in 2016 (Veit *et al.*, 2016). Since many mutations can cause multiple defects that fall into different mutation classes, they suggested a combinatorial classification system, where categories are based on all combinations of the six original mutation classes. This way, the limits of the traditional classification system can be removed, especially with respect to the development of improved therapeutics for all genotypes.

The most prominent example for a mutation causing different molecular defects is the deletion of phenylalanine at position 508 (F508del), which is the most common CF causing mutation. The main defect caused by F508del is a folding defect, leading to premature degradation at the ER. Hence, it has traditionally been classified as a class II mutation. However, when rescued to the PM, F508del CFTR also exhibits reduced channel gating and a decrease in stability compared to the wild type, which are defects that belong to classes III and VI respectively. Thus, according to the new comprehensive classification system, F508del is a II-III-VI mutation, which makes it more apparent that a drug rescuing the trafficking defect is not sufficient, but has to be augmented by therapeutics that improve channel gating and membrane stability (Veit *et al.*, 2016).

According to CFTR2, a database with information collected by CF patient registries worldwide, F508del is by far the most common mutation, with an allele frequency of almost 70%. The next most frequent mutations are G542X (2.5%, class I), G551D (2.1%, class III), N1303K (1.6%, class II-III-VI), R117H (1.3%, class II-III) and W1282X (1.2%, class I-II-III-VI). The remaining 395 mutations listed in the database have an allele frequency of less than 1% (*Welcome to CFTR2 | CFTR2*, no date; Veit *et al.*, 2016).

1.1.2 CFTR biogenesis

CFTR expression is specific to cell type and tissue (Pranke and Sermet-Gaudelus, 2014a). While the highest amounts are expressed in the pancreas, followed by tissues of the endocrine and the digestive system, expression levels in the lung and other tissues of the respiratory system are relatively low (Ochoa *et al.*, 2021). The control of

the complex expression involves different transcription sites, multiple tissue selective cis-regulatory elements (CREs) and alternative splicing. The main airway selective enhancer elements are located at -44 kb and -35 kb, whereas the main intestinal enhancer elements are located at 185+10 kb (intron 1) and 1811+0.8 kb (intron 11) (Ott, Blackledge, *et al.*, 2009; Ott, Suszko, *et al.*, 2009; Gillen, 2012; Swahn and Harris, 2019). A range of transcription factors are known to bind to these and other regulatory elements, but a lot remains to be understood about CFTRs expression regulation. Translation of the *cftr*-transcript occurs at approximately 2.7 residues per second, leading to an overall translation time of nine minutes for the whole transcript (Ward and Kopito, 1994).

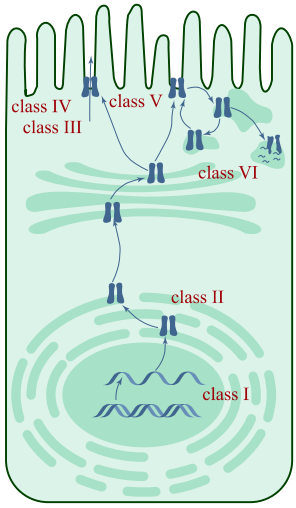


Figure 1. Cartoon representation of the CFTR biogenesis and the mutation classes. cftr is transcribed in the nucleus, which can be affected by class I mutations. The nascent peptide is folded co-translationally in the ER, which can be affected by class II mutations. After being fully glycosylated and transported to the PM through the secretory pathway, it transports Cl^- and HCO_3^- ions into the lumen. Class III and IV mutations affect ion conductance and class V mutations cause lower protein abundance. CFTR regularly undergoes endocytosis and either recycling or degradation, where class VI mutations decrease its stability at the membrane and cause premature lysosomal degradation.

As a large, multi-glycosylated protein, CFTR undergoes an intricate, delicately balanced and error-prone maturation pathway, the main steps of which are depicted in Figure 1. Even in wt-CFTR, only 20-40% of the transcripts produced are successfully integrated into the PM as fully functional protein (Pranke and Sermet-Gaudelus, 2014a). Folding of the newly synthesized peptide occurs co-translationally early on during translation at the ER membrane (Kim and Skach, 2012). Here, first of all

MSD1 is positioned in the ER membrane while NBD1 is being synthesized and then temporary bound by the chaperones HSP7C and DNJA1. The chaperones increase the stability and support folding of NBD1 while the R-region is being translated, which in turn reduces the affinity of DNJA1 to NBD1 (Lu *et al.*, 1998). Next, the synthesis of MSD2 stabilizes the interactions between NBD1 and the R-domain, causing the release of DNJA1 (Ostedgaard *et al.*, 1997). The complex formation between MSD1 and MSD2 in the ER membrane is facilitated by the chaperone calnexin (CANX), while NBD2 is being synthesized (Pind, Riordan and Williams, 1994; Okiyoneda *et al.*, 2004). The conformational maturation, which is then completed post-translationally, requires high energy in form of ATP and is assisted by different cytosolic and ER chaperones, which are also involved in ER-quality control (Lukacs *et al.*, 1994; Okiyoneda and Lukacs, 2012). During ER-quality control, CFTR passes through four quality check points (Farinha, Matos and Amaral, 2013; Farinha and Canato, 2017).

The first one occurs early in its biogenesis when the nascent polypeptide interacts with the chaperones HSP7C and DNJA1. In case of misfolding, the peptide is ubiquitinated by ligases and targeted for degradation at the proteasome (Meacham *et al.*, 1999; Farinha *et al.*, 2002).

The second checkpoint occurs during the calnexin cycle and the N-terminal glycosylation of CFTR. After folding of the MSD2 domain, a 14-unit oligosaccharide is attached to it and afterwards two of three glucose residues are removed by glucosidase I. When folded correctly, the third glucose residue is also removed by glucosidase II, which in turn leads to a decreased affinity to the calnexin chaperone, thereby releasing the core-glycosylated CFTR from the calnexin cycle. If, however, the protein is misfolded, it becomes re-glucosylated and retained in the calnexin cycle, and when retained too long (Hammond, Braakman and Helenius, 1994), degraded by glycoprotein ER-associated degradation (GERAD) (Farinha and Amaral, 2005).

The third checkpoint is constituted by arginine-framed tripeptides (AFTs), which are transient retention/retrieval motifs that CFTR has four of. When exposed due to misfolding, these AFTs result in CFTR being retained in the ER instead of being trafficked to the Golgi (Michelsen, Yuan and Schwappach, 2005).

The fourth and final checkpoint takes place when CFTR exits the ER through coat protein II coated vesicles, which relies on a specific exit motif being exposed in NBD1 (Nishimura and Balch, 1997; Wang *et al.*, 2004).

When successfully transported to the Golgi, the core-glycosylated CFTR undergoes its remaining glycosylation steps and becomes complex glycosylated and thereby functional. The mature protein is then trafficked and integrated into the plasma membrane by COPI vesicles (Yu *et al.*, 2007; Rennolds *et al.*, 2008). At the plasma membrane, CFTR interacts with a range of proteins with different functions, such as other ion channels, cytoskeletal proteins, and proteins directly involved in its activity.

Furthermore, ten percent of the CFTR protein at the PM become internalized every minute through endocytosis for quality control purposes (Prince *et al.*, 1999; Bertrand and Frizzell, 2003). Fully functional CFTR passes through the quality control mechanisms mediated in the endosomes and is recycled back to the PM, where native CFTR has a rather long half-life. Misfolded CFTR however, gets ubiquitinated and degraded in the lysosomes (Picciano *et al.*, 2003; Sharma *et al.*, 2004).

1.1.3 CF treatments

Until the approval of the first causative treatment in 2012, people with CF were solely treated with symptomatic medications. Early in the 1950s, long before the *cftr* gene was identified in 1989 by Kerem *et al.*, pancreatic enzymes were started being used to treat people with CF for their pancreatic insufficiency to allow them to absorb nutrients (Levy *et al.*, 1986). Around the same time, different mucus thinning enzymes were used in order to promote airway clearance. With the introduction of these first treatments, the average life expectancy of children with CF, who previously rarely lived through toddler age, started increasing. In the late 1950s, antibiotics against the most common CF pathogens, *Staphylococcus* and *Pseudomonas*, were used to combat the recurrent microbial infections in CF lungs (Ratjen, 2001). More than two decades later, in 1983, the first successful lung transplantation was performed, increasing the life expectancy of people, mainly adults, with end stage CF lung disease (Scott *et al.*, 1988; Adler *et al.*, 2009). The next major medication was introduced shortly after, in 1990, when recombinant human deoxyribonuclease I (rhDNase), also called dornase alpha, was first produced. rhDNase cleaves extracellular DNA, which accumulates in CF sputum due to its release from

degenerating neutrophils and contributes to its high viscoelasticity. By cleaving the DNA into shorter strands, the mucous becomes more liquid and easier to clear (Shak *et al.*, 1990). Afterwards, another important antibiotic, tobramycin, was introduced to target the main CF airway pathogen, *Pseudomonas aeruginosa*. Tobramycin is administered as an aerosol by inhalation, thereby directly delivering it to the site of infection (Ramsey *et al.*, 1999). All these symptomatic treatments, and many others, greatly improved the quality of life and life expectancy of people with CF and are still routinely used in patient care. Nonetheless, due to the vast amount of different geno- and resulting phenotypes, it has been difficult to find causative treatments for CF.

During the last years, however, different small-molecule therapeutics have been developed for clinical applications, which improve the CFTR function by directly targeting the CFTR protein – and not just alleviate symptoms of CF-patients. The first drug was clinically approved in 2012 and changed medical treatment of CF drastically. Currently, four pharmaceutical drugs, different combinations of four compounds, are approved and available as causative therapy to some CF patients (*Drug Development Pipeline | CFF Clinical Trials Tool*, no date; *Clinical Pipeline*, no date; Gentzsch and Mall, 2018; Zaher *et al.*, 2021). The first drug to be approved was Kalydeco, where the active compound Ivacaftor is a CFTR potentiator, which is approved mainly for gating mutations (Van Goor *et al.*, 2009; Ramsey *et al.*, 2011). Potentiators increase the open probability of CFTR, thereby leading to a higher chloride conductance, hence they are targeted at class III mutations. The newer approved drugs are combination therapies (*Drug Development Pipeline | CFF Clinical Trials Tool*, no date; *Clinical Pipeline*, no date), which contain more than one active compound and thereby target multiple defects. The second approved drug, Orkambi, contains Lumacaftor in addition to Ivacaftor. Lumcaftor is a CFTR corrector, which acts as small-molecule chaperone to correct the folding defect of class II mutations (Clancy *et al.*, 2012; Wainwright *et al.*, 2015). Similarly, Symdeco also contains Ivacaftor and an alternative CFTR corrector called Tezacaftor (Taylor-Cousar *et al.*, 2017). In addition to Ivacaftor and Tezacaftor, the most recent drug, Kaftrio (known as Trikafta in the US) contains a second CFTR corrector called Elaxacaftor, thereby making it the first triple combination (Voelker, 2019; Ridley and Condren, 2020). Additionally, there are currently eight other potentiators and

correctors in the clinical pipeline, of which two are again triple combinations (*Drug Development Pipeline | CFF Clinical Trials Tool*, no date).

Despite the great progress made with respect to causative treatments during the last decade, there is still no approved modulator for about 10% of people with CF. These are mainly people with mutations of class I, especially those with premature termination codons (PTC), which terminates translation and therefore no complete CFTR peptide is produced. Extensive research is thus being conducted on so-called readthrough agents, which suppress the termination of translation and PTCs. Several readthrough compounds have been handled as promising, some of which, such as Escin (Mutyam *et al.*, 2016) and Ataluren (Konstan *et al.*, 2020), are already clinically approved. One of the readthrough compounds, ELX-02 disulfate, is currently undergoing a phase II clinical trial (*Drug Development Pipeline | CFF Clinical Trials Tool*, no date).

In order to treat all mutation classes at once, great strides have been made in terms of mRNA and gene therapy. At the moment, one mRNA-based therapy is in a phase I clinical trial, and five other mRNA or gene therapies are in the pre-clinical stage. However, all of these therapies are being administered via inhalation, which brings them only into the lungs, but does not alleviate the morbidities in the other organs affected by CF (*Drug Development Pipeline | CFF Clinical Trials Tool*, no date).

Another genotype independent approach to CF treatment is targeting alternative chloride channels to circumvent CFTR entirely. Here, the Ca^{2+} activated chloride channel TMEM16A and the anion exchanger SLC26A9 have been of particular interest (Mall and Galietta, 2015; Amaral and Beekman, 2020).

Despite the immense progress made, the need for modulators and other causative treatments is still a great and research is continuously ongoing.

1.2 Systems biology disease maps

1.2.1 Systems Biology

Systems biology is a rather young, interdisciplinary field of study that was established in the late 1990s and early 2000s. By definition, it is the holistic study of a biological system, be it at the level of a single cell, tissue or entire organism, with all its entities and the interactions between them. This involves computational and mathematical

analysis and modelling to not only understand, but also predict and control the behaviour of complex biological systems, networks and processes (Ideker, Galitski and Hood, 2001; Kitano, 2002). Creating *in silico* models of these systems makes it possible to efficiently test different perturbations and manipulations and predict the most likely outcomes, which can then be translated into the real system. These models find application in a number of different disciplines and therefore vary widely in their nature and the methods used. To name just a few, there are ecological models, which study entire ecosystems, epidemic models of infectious diseases, and organ models, which aim to simulate a complete organ, such as the liver (Holzhütter *et al.*, 2012) or the brain (Markram, 2006). One of the largest applications of systems biology modelling, however, are cellular models, where the model components are on the molecular level. Systems biological cellular models in their most basic form can be displayed as mathematical graphs, which are structure that describe relations between objects. Graphs are made up of nodes (also known as vertices) and edges that connect them. In systems biology, one node stands for a biological entity, often molecules such as proteins or genes, and the edges represent the interactions between them. While a lot of information can already be derived from these graphs in their rudimentary form, they are mainly descriptive (Emmert-Streib and Dehmer, 2011; Najafi *et al.*, 2014). Different modelling techniques have been developed to make them predictive, the two major classes being qualitative and quantitative models. Qualitative, mainly logical models allow for representing and predicting qualitative relations and non-numerical information at discrete time points (Le Novère, 2015). The most common qualitative modelling techniques are Boolean network modelling, derived from Boolean algebra in mathematical logic, and Petri nets, which originate from the computer science field of distributed computing (Machado *et al.*, 2011; Najafi *et al.*, 2014; Koch, 2015). The main advantage of qualitative models is that they do not require the experimental determination of kinetic parameters, which is oftentimes not feasible (Machado *et al.*, 2011). For quantitative models, however, the nature and numerical properties of every interaction have to be known, which makes it possible to predict the transient behaviour of the system. Quantitative models are mostly based on systems of ordinary differential equations (ODEs), which contain the relevant rate equations of biochemical interactions. They therefore require a lot of prior experimental data to estimate all the different kinetic parameters (Chassagnole *et*

al., 2002; Machado *et al.*, 2011; Najafi *et al.*, 2014). Over the years, different hybrid methods, such as semi-quantitative models, have been developed to reap the benefits of both, qualitative and quantitative models.

In general, a distinction can be made between three major types of cellular models: 1. Metabolic pathways 2. signal transduction pathways, and 3. gene regulatory networks (Najafi *et al.*, 2014). A gene regulatory network simulates how gene expression is controlled by different factors and entities, often a complex network of transcription factors. Since they require little detail to make accurate predictions, and usually little kinetic information is available, they are most often simulated by Boolean models. Signal transduction pathways simulate signalling cascades and the response of the cell to them. As the amplification of signals starting from a low number of molecules relies heavily on diffusion, they are often modelled by stochastic simulations and probability functions. Metabolic pathways are often considered the most complex models as they contain many different metabolites and enzymes, with highly specific kinetics and rate laws, and can be modelled by quantitative ODE models (Machado *et al.*, 2011).

In order to standardize systems biology models and make them exchangeable and reusable, over the years, different data formats have been developed by the community. The most well established and widely used format is the *Systems Biology Markup Language* (SBML) (Hucka *et al.*, 2003). SBML is a machine-readable, XML-based format that was first introduced in 2000 and is continuously being developed by experts until today. To make systems biology models not only machine- but also human-readable, a standardized graphical representation, the *Systems Biology Graphical Notation* (SBGN), was developed (Novère *et al.*, 2009). SBGN visualizes models in the intuitive way pathways have traditionally been represented in, but in a comprehensive and standardized manner, that can also be reused and exchanged. An example of a molecular pathway represented in SBGN can be seen in Figure 2.

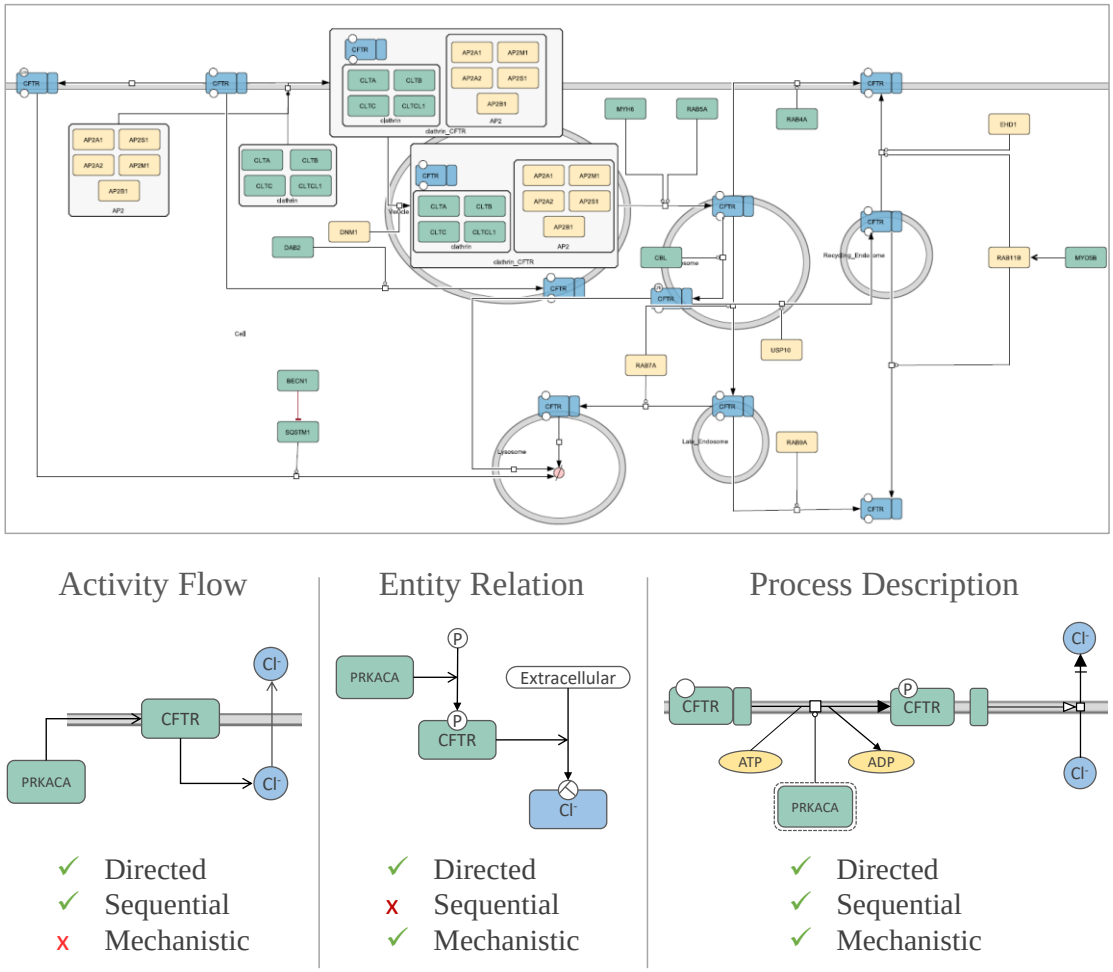


Figure 2. Exemplary SBGN model and different SBGN languages. The upper panel shows an example overview of an SBGN model. The lower panel shows one molecular process, the activation of CFTR by the protein kinase PRKACA and the subsequent transport of Cl^--ions across the plasma membrane, in the three different SBGN languages "Activity Flow", "Entity Relation" and "Process Description" (adapted from Le Novère, 2015).

Depending on the exact requirements and level of detail of the model, three different languages exist within SBGN (Figure 2). The first language is the *Activity Flow* (AF), which is directed and sequential, but not mechanistic, meaning that it does not show the mechanism by which two entities interact. The second language is the *Entity Relation* (ER), which in turn is directed and mechanistic, but not sequential, so it is not possible to draw a successive path through the model. *Process Description* (PD) is the third and most detailed language. It is directed, sequential and mechanistic and therefore describes all interactions and pathways in great detail but also requires the most information. In general, activity flows are most suitable for gene regulation networks, entity relationships are mostly used for signal transduction pathways and process description are the language of choice for metabolic networks (Mi *et al.*,

2015; Sorokin *et al.*, 2015; Rougny *et al.*, 2019). In SBGN, every kind of molecular entity is represented by an individual shape. For example, regular proteins are depicted as rounded rectangles, ions are shown as circles, and membranes are represented as thick double lines. The molecular entities are connected by different arrows, which stand for the various interactions and transitions. For example, a state transition is shown as a regular arrow with a square in the middle, inhibitions have a perpendicular line as an arrow head, and stimulations have a circle as an arrow head. This way, the graphical notation is intuitive for researchers, as it closely resembles the ways biochemical reactions have been traditionally depicted, but is overall much more comprehensive (Novère *et al.*, 2009).

One of the most important aspects in the reproducibility and reusability of systems biology models is consistency in their annotation and curation. This already starts with the naming of molecular entities, such as genes and proteins. For every protein, a range of identifiers exists, extended by a multitude of different abbreviations used by each individual researcher. It is therefore of utmost importance to correctly annotate biological models in a standardized manner. To help with this, a set of guidelines called the "Minimal Information Required In the Annotation of Models" (MIRIAM) has been established, which all systems biology models should adhere to (Le Novère *et al.*, 2005). The MIRIAM guidelines can be subdivided into three different parts: Reference correspondence, attribution annotation and external resource annotation. The reference correspondence includes guidelines on how a model should relate to a specific reference and comply to a certain format and standard. The attribute annotation refers to the model itself, for which a name, a reference, the model creators and the dates of creation and modification have to be provided. The external resource annotations describe the way all parts and elements of the model have to be annotated. Here, each piece of information has to be annotated with the collection (e.g. the database), the identifier (e.g. the UniProtID), and the qualifier, which links the element and the information (e.g. "is", "is described by" or "is homolog to").

1.2.2 Disease Maps

During the last decade, systems biology approaches have more and more found application in biomedical research, and the field of systems medicine evolved. Especially with the rise of personalized medicine, systems medicine is becoming

increasingly important, as it integrates the different omics data, which by now can be derived for individual patients. Hereby, systems medicine is on the one hand extremely customizable and on the other hand provides the bigger picture by considering the interplay between different physiological processes. One of the approaches to support biomedical research through systems medicine are disease maps, based on the systems biology models. Disease maps have been proposed as a community project in 2018 to link big data from health-care research and biomedical knowledge networks and are defined as "[…] comprehensive, knowledge-based representations of disease mechanisms." (Mazein *et al.*, 2018). They are mostly written in SBGN and adhere to one of the three languages within them. As opposed to traditional systems biology models, disease maps do not differentiate between metabolic, signal transduction and gene regulatory networks, but integrate all of them to provide a comprehensive picture of their relations. Other important additions to disease maps are physiological mechanisms and phenotypes, which are mostly excluded in biochemical systems biology models. Disease maps are often interdisciplinary community efforts, involving domain experts and computational biologists. At the moment, there are disease maps either already published or being created for 16 different diseases, ranging from Asthma (Mazein *et al.*, 2021) and Parkinson's disease (Fujita *et al.*, 2014), to Cancer (Kuperstein *et al.*, 2015) and Atherosclerosis (Parton *et al.*, 2019), to only name a few. During the COVID-19 pandemic, a huge community effort, involving 230 researchers from 120 institutions across 30 different countries, was launched to develop the COVID-19 disease map. This is the largest disease map created so far, consisting of 5499 elements connected by 1836 interactions from 617 publications (Ostaszewski *et al.*, 2021a).

In general, disease maps serve a range of different purposes. They can provide a backbone to structure large amounts of experimental data such as omics data, they can be used for modelling and simulation purposes, or to predict side effects of medications. Another important application of disease maps is drug repurposing, which has become increasingly significant during the past few years, to circumvent some of the time- and cost intensive processes of early clinical development. Most importantly, disease maps can be used to identify novel drug targets, for example by detecting specific key players in molecular pathways.

1.3 Virtual screening & target identification

On average, only one in 5000-10000 compounds tested in the early stage of drug discovery is approved as drug in the end. Therefore, these early stages, were thousands of assays have to be performed are extremely time and cost intensive (McInnes, 2007; Matthews, Hanison and Nirmalan, 2016). One way to reduce cost and speed up the process is virtual screening. During virtual screening, large compound libraries are tested *in silico* to identify potential lead substances to narrow down the lists and thereby identify candidates for further testing. In contrast to experimental high throughput screening (HTS), this requires no materials and no special equipment, except computational resources, and little hands-on time, making it much more economic (McInnes, 2007; Rester, 2008; Shaker *et al.*, 2021). As can be seen in Figure 3, virtual screening techniques can be subdivided into two main approaches: Target- and ligand-based methods (McInnes, 2007; Matthews, Hanison and Nirmalan, 2016). Traditional virtual screening aims at finding compounds that bind to one specific target, most often a protein. In recent years however, the inverse method has started to gain importance (Chen and Zhi, 2001; Paul *et al.*, 2004; Byrne and Schneider, 2019). During inverse virtual screening, multiple proteins are screened against one or more specific compound of interest to identify its target and elucidate the mechanism of action. Just like in classical virtual screening, target- and ligand-based methods exist (Huang *et al.*, 2018). The main application of reverse screening approaches has been to elucidate targets of natural compounds (Huang *et al.*, 2018; Xu, Huang and Zou, 2018). Lim *et al.* used a ligand-based approach to explain the cancer-preventive properties of the phytochemical curcumin and identified the cyclin dependent kinase (CDK2) as one of its targets (Lim *et al.*, 2014). By using inverse docking, i.e. a target based approach, Buendia-Atencio *et al.* studied different helicases in Zika viruses as targets of ligands from a flowering plant (Buendia-Atencio *et al.*, 2021). Furthermore, Ban *et al.* investigated the effect of thymol, which is derived from the thyme plant, on fat deposition (Ban *et al.*, 2021) and Lauro *et al.* shed light on the antitumor targets of a library of natural bioactive compounds (Lauro *et al.*, 2011).

In general, ligand-based approaches are much faster and require less computational power than target-based approaches. Both methods rely on prior knowledge, either known ligands or the protein structure are required for successful application. Hence, depending on the available data, it is case dependent which method will yield more comprehensive results and it can be beneficial to use both in a complementary manner, if applicable.

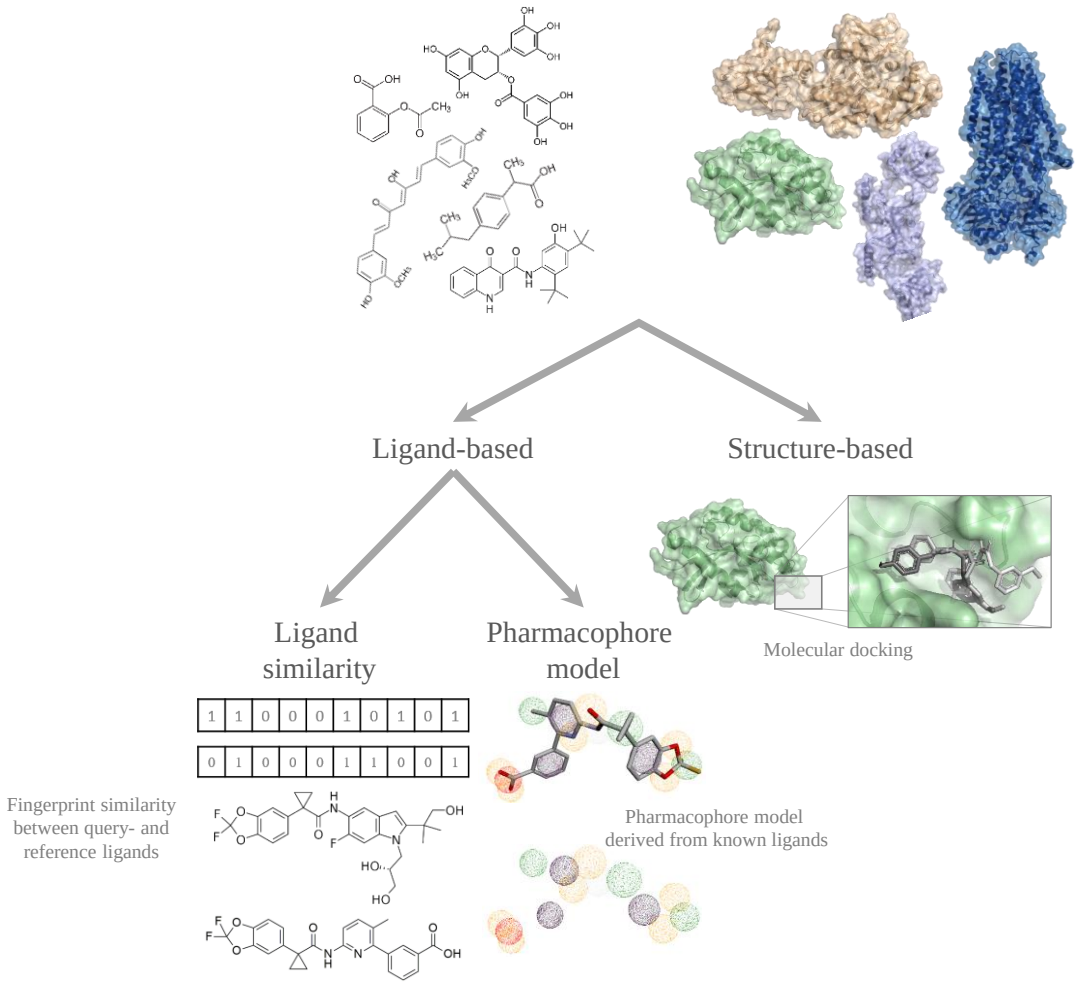


Figure 3. Overview of (reverse) virtual screening techniques. Virtual screening can be subdivided into target-based approaches, mainly docking approaches, and ligand-based approaches. Ligand-based approaches can either rely on pharmacophore models, which are based on steric and electronic features of known ligands, or ligand similarity, where binary fingerprints of the known ligands and query ligands are calculated and compared.

1.3.1 Target-based approach

As the name says, target-based methods, also called structure-based methods, centre around the structure of the target. The most common target-based method is molecular docking, which predicts the binding affinity of small molecules to the

target protein (Brooijmans and Kuntz, 2003; Batool, Ahmad and Choi, 2019). To do so, the ligand is placed in different positions and conformations on the protein and the binding affinity between them is calculated. Usually, a specific site of the protein where the ligand should bind, such as the active site or a known binding pocket, is defined beforehand. If unknown, however, it is also possible to conduct blind docking, where the entire target structure is sampled (Hetényi and Van Der Spoel, 2006; Hetényi and van der Spoel, 2009; Hassan *et al.*, 2017). Molecular docking consists of two main parts, the search algorithm and the scoring function (Maia *et al.*, 2020). The search algorithm predicts possible conformations of the ligand and the binding site. In order to do so, different approaches can be used, e.g. local shape feature matching, genetic algorithms, molecular dynamics or systematic searching (Halperin *et al.*, 2002; Yadava, 2018; Maia *et al.*, 2020). The scoring function is responsible for calculating the binding affinity of each predicted binding pose. Again, there are different approaches to calculate the binding affinity between the ligand and the protein (Guedes, Pereira and Dardenne, 2018; Li, Fu and Zhang, 2019). The first approach is physics based, which uses a force field that takes into account different energy contributions by e.g. electro static and van der Waals interactions, as well as solvation. The second approach is empirical, which uses known binding affinities to predict energetic factors, such as hydrogen bonds and steric clashes. Another class of scoring functions, knowledge-based approaches, uses the statistical analysis of atom pairs from existing protein-ligand pairs to calculate the binding potentials. The fourth kind of scoring functions are machine-learning-based. A range of different machine learning approaches have been applied for scoring functions. Although some of them have been shown to outperform classical scoring functions, they are not yet employed in docking programs as they rely heavily on the training dataset, which causes the results to be un-reproducible and inconsistent. (Guedes, Pereira and Dardenne, 2018; Li, Fu and Zhang, 2019).

One of the main limitations of (inverse) docking is the availability of suitable protein structures. Molecular docking requires complete structures with a resolution of ideally at least 2.5 Å (Li et al., 2010). For a lot of targets, especially larger or membrane proteins, the structure has yet to be elucidated experimentally. This particularly limits inverse docking for target identification approaches, since not all targets in question can be screened. One way to circumvent this limit is by using structures predicated by

artificial intelligence-based programmes, such as AlphaFold (Jumper *et al.*, 2021). Artificial intelligence based protein structure predictions have greatly improved and gained importance in the last years. In some cases, they generate very accurate results that can be readily used for docking purposes. Nonetheless, this is not a universal solution currently, since often parts of the structures predicted, especially in flexible regions of the protein, remain unfolded, which excludes them from being used for docking. As they rely heavily on the training dataset, AI-based structure determination methods additionally perform less well on protein classes where experimental data is scarcer, such as large membrane proteins.

1.3.2 Ligand-based approach

Ligand-based approaches rely on the prior knowledge on molecules that are known to bind to a specific target. Here, different properties are used to compare the query ligand to already known ligands to evaluate if it is likely to bind to the target or not. The two main methods used in ligand-based drug design are pharmacophore modelling and similarity-based virtual screening (Huang *et al.*, 2018; Shaker *et al.*, 2021), as shown in Figure 3. Pharmacophores are the functional and electronic features of a molecule and their spatial arrangement at the surface of a molecule, which play a role in its binding affinity to potential targets, and are therefore complementary to their respective binding pocket. For pharmacophore modelling, these molecular features, such as hydrogen bond donors and acceptors, are extracted from known active ligands. Properties that are shared amongst ligands that bind to the same protein pocket are then used to create pharmacophore models, which represent the features of likely ligands. By matching the query ligands to the pharmacophore models, ligands can be proposed for certain protein pockets and vice versa (Huang *et al.*, 2018; Shaker *et al.*, 2021). Pharmacophore modelling requires several known ligands that have been shown to interact with the same complementary binding site of a protein. Hence, an exhaustive data basis is essential in order to use it for target identification purposes, since it would involve a range of proteins with potentially multiple binding sites. It is therefore more suitable for classical virtual screening, where the focus is on one specific target and the aim is to effectively narrow down a large compound library.

Similarity-based screening uses the general shape of a molecule to match it to a protein. Like pharmacophore modelling, it does so by comparing it to known ligands. Instead of creating a model of features from multiple ligands, however, it simply compares how similar the structure of the query ligand is to that of a reference one (Huang *et al.*, 2018; Mathai and Kirchmair, 2020). Therefore, in theory, one reference molecule is enough to match a query ligand to a protein, and the exact binding positions of the reference ligands do not necessarily have to be elucidated. Similarity-based screening can be performed with either two- (2D) or three-dimensional (3D) structures. In the case of 3D structures, the similarity search is based on the geometries and volumes of the two molecules. For 2D methods, the common approach is to calculate a molecular fingerprint, which encodes the molecular characteristics as a simple bit vector, i.e. a sequence of 0's and 1's (Huang *et al.*, 2018; Kumar and Zhang, 2018). The vector representation allows easy and rapid comparison of molecules and requires very little computational power. A number of different 2D fingerprints have been developed over the years, which take different approaches to encode the molecular structures. In general, they can be categorized into either substructure key-based fingerprints or topological fingerprints (Cereto-Massagué *et al.*, 2015; Seo *et al.*, 2020). Substructure based fingerprints show whether certain predefined substructures are present or not (Cereto-Massagué *et al.*, 2015). The most basic example for substructures are chemical elements. In this case, a sequence of elements would be defined (e.g. hydrogen-carbon-oxygen-nitrogen-chlorine-phosphorus) and a "1" is assigned to elements present in the molecule, whereas a "0" is assigned to elements not present. In this example, the fingerprint of water (H_2O) would be "101000", since only hydrogen and oxygen are present. By using more complex substructures, such as functional groups, the fingerprint is made more specific and meaningful. These fingerprints are binary, as they only take into consideration whether or not a certain substructure is present, but ignores how many there are of each. Topological fingerprints take one starting atom in the molecule and analyse all molecular fragments that can be produced by going through (linearly or circularly) all possible paths from this atom up to a certain path length, i.e. number of bonds (usually seven steps). The resulting paths are then hashed to produce a binary vector (Cereto-Massagué *et al.*, 2015). In all cases, the fingerprints can then be

compared by different similarity measures, the most common one being the Tanimoto similarity (Bajusz, Rácz and Héberger, 2015).

Overall, while similarity-based screening can generally be less accurate than pharmacophore modelling, it is much more feasible for target identification purposes, since it makes do with less prior knowledge and the exact binding sites of the reference ligands do not have to be known. For pharmacophore modelling, multiple ligands and their binding poses in a specific binding site have to be known to overlay them and extract the relevant features. In contrast, for ligand similarity-based approaches, one reference compound of an unknown binding site is enough to identify potential query ligands. However, both methods rely on there being known ligands of the relevant targets for the query ligands to be compared to in the first place.

1.4 Thesis objectives and structure

1.4.1 Objectives

During the last two decades, systems medicine and computational biology have opened up a range of novel possibilities for biomedical research. Not only is it getting cheaper, faster and easier to produce large amounts of biomedical and biochemical data, but the variety of tools and methods available to utilize this data is ever increasing. Still, oftentimes the different kinds of data from diverse sources are analysed separately and with highly specific tools, so that a comprehensive view is missing and the data is not used to its full potential. This thesis aims at utilizing a set of generic methods and approaches to integrate different kinds of biological and chemical data from various sources. Its main objective is thereby creating an inclusive model of molecular disease mechanisms to use as basis for gaining new insights and developing treatment approaches. As a use case, this thesis looks at the genetic disorder CF. While some causative treatments for CF have been developed and clinically approved in the last decade, these therapeutics are not suitable for all people with CF and the search for active compounds, and especially compound combinations, is ongoing. Ultimately, the aim of this thesis is therefore to suggest novel, synergistically acting compound combinations as causative treatment for Cystic Fibrosis. Figure 4 gives an overview of the project's general structure. Overall, three main objectives can be defined:

Creating a chemistry centric data basis – Assemble knowledge on compounds that have been tested for activity in the CF context. This includes not only positive results, but also negative ones to exclude specific properties and areas of the chemical space.

Creating a systems medicine model as biology centric data basis – Integrate different data sources into a comprehensive systems biology model of the molecular mechanisms and pathways underlying CF. The model is human- as well as machine readable and adheres to the community standards. Apply text mining techniques to support the curation process where relevant.

Integration of chemistry- and biology centric data – Bring together both data sources to generate new knowledge on both sides. Using different *in silico* drug discovery methods, annotate the systems biology model with possible active compounds and suggest biological targets for active compounds. Ultimately, suggest compounds that modulate different parts of the molecular disease pathways to test for synergistic activity.

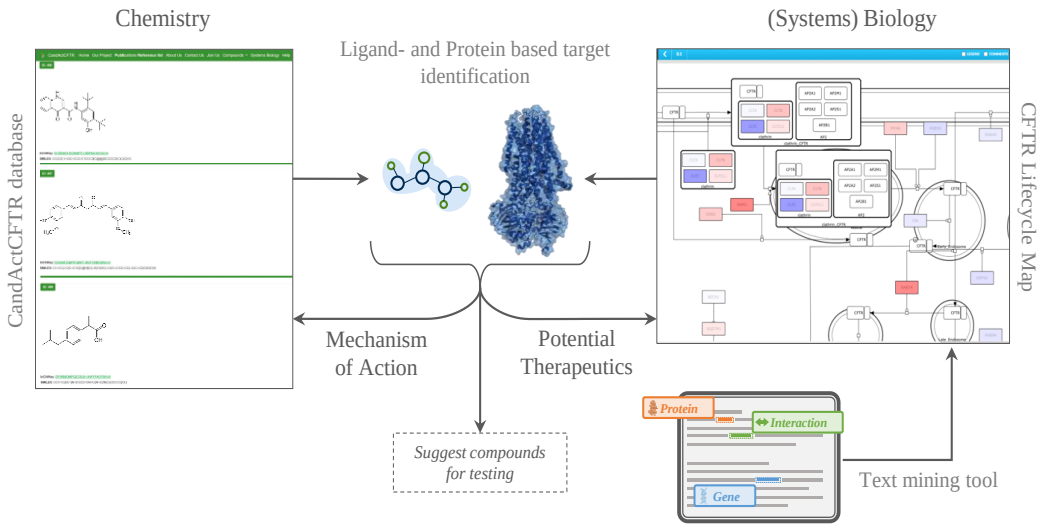


Figure 4. Project overview and structure. The project is based on two main resources, the chemistry-centric CandActCFTR compound database, and the (systems-)biological CFTR Lifecycle Map. Bringing the two resources together through target identification approaches allows the elucidation of the mechanism of action of active compounds in the database and the identification of potential therapeutics for important targets in the CFTR biogenesis. A software tool was developed to connect the two resources computationally and a second tool was developed to support the creation of disease maps through text mining.

1.4.2 Thesis structure

Chapter 2 of this thesis focusses on the first objective. It consists of a publication that describes the development of CandActCFTR, a publicly available database of compounds that have been tested as direct causative modulators in Cystic Fibrosis. The aim here was to collect the small molecules that have been shown to either modulate CFTR function, or to not exhibit any activity. The compounds are annotated and categorized according to their activity and their order of interaction with CFTR.

The second objective is covered in Chapter 3 and Chapter 4. The publication in Chapter 3 introduces the CFTR Lifecycle map, the disease map used as biological basis for the project. It describes the pathways involved in CFTR biogenesis and is written in SBGN, the community standard for systems biological pathway maps. It consists of two different data layers with different levels of detail and is subdivided into sub-maps according to the different parts and stages of the CFTR biogenesis. Chapter 4 describes a tool to support the time-consuming manual creation of systems biology models and disease maps through text mining. It provides the means to integrate text-mined molecular interaction data from any text mining algorithm or software into the curation process by displaying the results and letting the user iterate through and validate or reject the interactions identified.

Chapter 5 and Chapter 6 focus on the third objective. The publication in Chapter 5 describes a tool that forms the interface between CandActCFTR and the CFTR Lifecycle Map software-wise. The tool is integrated into the platform hosting the systems biology model and maps the compounds to their potential targets based on publicly available data. It provides the user with the possibility to either find known targets of specific compounds, or search for compounds associated to a specific target. As the interaction data in publicly available databases is scarce and does not cover the entire list of active compounds in CandActCFTR, the aim of the publication in Chapter 6 is to elucidate the mechanism of action of these compounds through target identification approaches. A dual approach, using target- and ligand- based *in silico* methods, is employed to suggest possible targets in the disease map for all active compounds in the database and rank the interactions according to the level of confidence. The results of the preceding chapters are discussed in Chapter 7 and conclusions are drawn in Chapter 8.

Chapter 2 Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR

This manuscript has originally been published in Frontiers in Pharmacology

Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR

Manuel Manfred Nietert [1,2], Liza Vinhoven[1], Florian Auer[3], Sylvia Hafkemeyer[4] and Frauke Stanke[5,6]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany, [3]Institute for Informatics, University of Augsburg, Augsburg,Germany
[4]Mukoviszidose Institut gGmbH, Bonn, Germany
[5]German Center for Lung Research (DZL), Partner Site BREATH, Hannover, Germany
[6]Clinic for Pediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Hannover, Germany

Authors contribution

M.M.N.: conception and design of the database; implementation of the software; acquisition, analysis, and interpretation of data; drafting the article or revising it critically for important intellectual content

L.V.: acquisition, analysis, and interpretation of the data

F.A.: supporting Grails implementation

S.H.: acquisition, analysis, and interpretation of the data; drafting the article or revising it critically for important intellectual content

F.S.: conception and design of the database; acquisition, analysis, and interpretation of data; drafting the article or revising it critically for important intellectual content

frontiers in Pharmacology

ORIGINAL RESEARCH
published: 08 December 2021
doi: 10.3389/fphar.2021.689205

Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR

Manuel Manfred Nietert[1,2], *Liza Vinhoven*[1], *Florian Auer*[3], *Sylvia Hafkemeyer*[4] *and Frauke Stanke*[5,6]*

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany, [2]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany, [3]Institute for Informatics, University of Augsburg, Augsburg, Germany, [4]Mukoviszidose Institut gGmbH, Bonn, Germany, [5]German Center for Lung Research (DZL), Partner Site BREATH, Hannover, Germany, [6]Clinic for Pediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Hannover, Germany

OPEN ACCESS

Edited by:
Viola Hélène Lobert,
Østfold University College, Norway

Reviewed by:
Antonio Recchiuti,
University of Studies G.d'Annunzio Chieti and Pescara, Italy
Andre Falcao,
Universidade de Lisboa, Portugal

***Correspondence:**
Frauke Stanke
mekus.frauke@mh-hannover.de

Specialty section:
This article was submitted to Inflammation Pharmacology, a section of the journal Frontiers in Pharmacology

Received: *01 April 2021*
Accepted: *08 November 2021*
Published: *08 December 2021*

Citation:
Nietert MM, Vinhoven L, Auer F, Hafkemeyer S and Stanke F (2021) Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR. Front. Pharmacol. 12:689205. doi: 10.3389/fphar.2021.689205

Background: Cystic fibrosis (CF) is a genetic disease caused by mutations in *CFTR*, which encodes a chloride and bicarbonate transporter expressed in exocrine epithelia throughout the body. Recently, some therapeutics became available that directly target dysfunctional CFTR, yet research for more effective substances is ongoing. The database CandActCFTR aims to provide detailed and comprehensive information on candidate therapeutics for the activation of CFTR-mediated ion conductance aiding systems-biology approaches to identify substances that will synergistically activate CFTR-mediated ion conductance based on published data.

Results: Until 10/2020, we derived data from 108 publications on 3,109 CFTR-relevant substances via the literature database PubMed and further 666 substances via ChEMBL; only 19 substances were shared between these sources. One hundred and forty-five molecules do not have a corresponding entry in PubChem or ChemSpider, which indicates that there currently is no single comprehensive database on chemical substances in the public domain. Apart from basic data on all compounds, we have visualized the chemical space derived from their chemical descriptors via a principal component analysis annotated for CFTR-relevant biological categories. Our online query tools enable the search for most similar compounds and provide the relevant annotations in a structured way. The integration of the KNIME software environment in the back-end facilitates a fast and user-friendly maintenance of the provided data sets and a quick extension with new functionalities, e.g., new analysis routines. CandActBase automatically integrates information from other online sources, such as synonyms from PubChem and provides links to other resources like ChEMBL or the source publications.

Conclusion: CandActCFTR aims to establish a database model of candidate cystic fibrosis therapeutics for the activation of CFTR-mediated ion conductance to merge data from publicly available sources. Using CandActBase, our strategy to represent data from

several internet resources in a merged and organized form can also be applied to other use cases. For substances tested as CFTR activating compounds, the search function allows users to check if a specific compound or a closely related substance was already tested in the CF field. The acquired information on tested substances will assist in the identification of the most promising candidates for future therapeutics.

Keywords: cystic fibrosis, substance database, compound database, therapeutic substances, high-throughput screening, library collection, chemical space annotation, search tool

INTRODUCTION

Cystic fibrosis (CF) is a genetic disease inherited in an autosomal recessive fashion (Elborn, 2016). The highest incidence is observed among people with northern European ancestry where it affects approximately one out of 3,000 newborns in populations who offer CF genetic testing to couples (SpringerMedizin, 2021). The disease-causing gene *CFTR* encodes a chloride and bicarbonate transporter expressed in exocrine epithelia throughout the body (Elborn, 2016). Manifestations of the generalized exocrinopathy encompass failure to thrive and recurrent pulmonary infections as the hallmarks of the two major affected organ systems, i.e., the gastrointestinal and the respiratory tracts (Elborn, 2016).

The clinical diagnosis of CF is assisted by bioassays that rely on the detection of CFTR dysfunction in the sweat gland (Elborn, 2016), in the nasal and in the intestinal epithelium (Wilschanski et al., 2016). Symptomatic therapy at centers specialized in CF care has increased the life span of CF patients considerably: while CF was once known as a devastating disease leading to death in infancy or early school age, the average survival of CF patients is now by 40 years in developed countries (Elborn, 2016). Hallmarks of therapy improvement were the supplementation of pancreatic enzymes and a consistent treatment of infections of the respiratory systems (Elborn, 2016). For a few years, therapeutics are available that directly target dysfunctional CFTR, developed for clinical application by Vertex Pharmaceuticals (Van Goor et al., 2006). While Kalydeco targets the CFTR mutant G551D-CFTR, Orkambi is licensed for the most frequent CFTR disease causing lesion F508del-CFTR (Martiniano et al., 2016). Surprisingly, the experience with these causal treatments that complement the successful symptomatic treatment falls behind the enthusiastic expectations leading to "efforts from the community to look for other therapeutics in spite of Orkambi" (Martiniano et al., 2016). The Cochrane Reviews conclude in summary that "Combination therapies (lumacaftor–ivacaftor and tezacaftor–ivacaftor) each result in similarly small improvements in clinical outcomes in people with CF; specifically, in improvements in quality of life (moderate-quality evidence), in respiratory function (high-quality evidence) and lower pulmonary exacerbation rates (moderate-quality evidence) (Southern et al., 2018). Taken together, CFTR mutation-specific therapeutics became available, but research for more effective substances is ongoing (Gentzsch and Mall, 2018).

The database CandActCFTR aims to provide detailed and comprehensive information on candidate therapeutics for the activation of CFTR-mediated ion conductance using a systems-biology approach to identify substances that will activate CFTR-mediated ion conductance in a synergistic fashion based on published data. There are several efforts to identify compounds that activate CFTR residual function in the community: Hit-CF Europe specializes to identify compounds that can be used to treat rare CF mutations in an organoid model (HIT-CF, 2021). Recently, Veit et al. (2018) and Phuan et al. (2018) have combined therapeutics derived from their screening efforts. In addition, several competing pharmaceutical companies develop CFTR therapeutics through their own screening data (CFF, 2021). Our approach differs from all of these efforts as it is not restricted to a particular CFTR mutation type and not restricted to a particular screening data set. In contrast, we have merged publicly available information in a meta-database enabling comprehensive data retrieval and analysis. To the best of our knowledge, our effort is the only holistic approach to use integrated data from multiple sources employing advanced digital technologies to provide unbiased criteria for selecting therapeutic substances. This strategy should be particularly suitable to successfully select substance combinations as several of the compounds published by academia showed small effects, albeit they were tested in many bioassays. This is in contrast to the high-throughput-screening strategy as here, effective compounds are selected based on one assay only. Interestingly, Veit et al. (2018) could recently show that a combination of compounds that perform poorly when considered isolated will improve mutant CFTR function to 50 or 100% of wild-type level, confirming that all substances that activate CFTR might be valuable therapeutics.

MATERIALS AND METHODS

The CandActCFTR project is organized into three project domains:

- The data sources regarding CFTR and ways to search and access these.
- Means to afterwards handle and organize the data coming from these sources.
- Data extracted, annotated, and stored for structured access.

Data Sources and Ways to Search and Access These

To facilitate that the developed workflows are adaptable by other research groups for other use cases, we focused on using open-

source resources and modularized our system as well as we could. For the project, we collected a pool of literature derived data regarding chemical compounds in the context of cystic fibrosis, especially focusing on interactions with the CFTR protein. The aim was to facilitate the collection, organization, and thereafter user-friendly representation of the collected data to a research community. To achieve this, we mined the PubMed service (PubMed, 2021) to find a list of relevant entries to inspect. We organized this shared literature list using the free to use Zotero web service (Zotero, 2021) and documented our search at https://www.zotero.org/groups/1632179/candactcftr_public_ references. We then used Zotero's application programming interface (API) to pull the citation data into our information system.

The chemical structure is the root of our data model and all other information is linked directly or indirectly with this root. We provide an entry form to enter structures in the database, either by simply providing identifiers or drawing the molecular graph. The system queries PubChem (Kim et al., 2021) with the provided identifiers for more information, for instance synonyms. Users can draw their molecule online using the Ketcher JavaScript plugin (Ketcher, 2021) and then translate this graph to an isomeric SMILES, which is used to look up the compound after being converted to an InChIKey (Southan, 2013) using OpenBabel (Open Babel, 2021) by the server. Alternatively, if the user provides a PubChem ID (Kim et al., 2021), this is used to collect the isomeric smiles and InChIKey (Southan, 2013) and synonyms into our database. In case of a match, the system can create links to the PubChem's web resources (Kim et al., 2021) for this compound, when shown in the web views. If no match is found querying PubChem (Kim et al., 2021) via InChIKey (Southan, 2013), the compound is stored without the PubChem (Kim et al., 2021) synonyms and saved annotated with just the name entered by the user. InChIKeys (Southan, 2013) are used to look up structures and prevent double entries.

Means to Handle and Organize the Data

To achieve our aim, we needed a classical server stack with storage (domain/databases), request & data handling (Controllers), and visualizing interfaces (views). The data system should be easily extendable and accessible by multiple means of loading and querying the data. Data table definitions should be adaptable without extensive training. Therefore, we chose the groovy-dialect-based Java implementation of a Grails (Grails Framework, 2021) system as our base for the project (https://candactcftr.ams.med.uni-goettingen.de/). The models/domain definition can be done using simple and structured human-readable text documents containing the declarations of the data variables to be stored and the interconnection of these tables, while the details to instantiate and update the actual data tables in the used database system of choice is taken care of by the Grails (Grails Framework, 2021) system itself. After the instantiation of the data structures within the database server of choice, one can either use the Grails (Grails Framework, 2021) controller system to query the data resources and ultimately display some of the data or process the data and display the results. Because the back-end database system in the Grails (Grails Framework, 2021) stack can be a generic Structured Query Language (SQL) server, any means to access and manipulate the data tables using APIs can be used, thus allowing the use of established data analysis pipeline tools like KNIME (Berthold et al., 2007; KNIME, 2021) to interface the data. Administrative tasks like backups can thus be handled by known tools like mysqldump (MariaDB KnowledgeBase, 2021) and phpMyAdmin (phpMyAdmin, 2021) or others, depending on the used backend system for storage. For small projects with few tables, KNIME (Berthold et al., 2007) can be used and later extended for batch updating data tables. To enable the storage of the literature data in a way compatible to a common exchange format as provided by Zotero (2021), we used the Citation Style Language schema of the Citation Style Language (CSL) project (Citation Style Language, 2021) to generate a generic data representation for literature meta-data (a domain model) to be used in our Grails (Grails Framework, 2021) web server. This definition is not CFTR specific and can thus be used in other projects. In our implementation, Grails Framework (2021) creates from this abstract definition a representation in the attached database MariaDB (MariaDB Foundation, 2021). Grails Framework (2021) has connectors for multiple target database systems and can be adapted to local infrastructure prerequisites. It contains supports of in-memory databases, thus reducing minimal installation requirements to Java and facilitates the start of the development phase as well as deployment. As data sets can be shipped as text files with the Grails Framework (2021) implementation to be loaded on startup of Grails Framework (2021), this is an easy way to provide a preset yet interactive view for the data tables or web APIs.

Data Extracted, Annotated, and Stored for Structured Access

Literature references are internally stored in the CSL format. Similarly to the entering of compounds, we integrate references with their meta-data via the PubMed API (Kim et al., 2021) querying their PubMed ID and receiving titles, authors and journal. After compounds and literature references have been entered, we provide web interfaces to help with linking compounds to literature. This web interface provides the remote working curator with the information of what has been linked to one substance before, and allows the curator to select a new reference guided by a search form-based workflow to pick a reference from the literature list. To allow easier batch processing however, likely done by curators who have direct working access to the backend of the database, we also provide similar maintenance workflow pipelines again using KNIME. We integrated so far three ways of providing molecular graph images to the web views: dynamic integration of a PubChem (Kim et al., 2021) derived Portable Network Graphics (PNG), by accessing PubChem's Web API, or independent of an internet connection using the Kekule JavaScript libraries (Jiang et al., 2016) in our web view pages to render the graph directly on the client machine, as our third option we now create the PNGs on the server side using Open Babel (2021).

In many instances in CandActCFTR, KNIME (Berthold et al., 2007) is used as a back-end controller and analysis pipeline. The literature search was enhanced by automatization: we used the PubMed (Kim et al., 2021) API via search nodes from KNIME (Kim et al., 2021) to perform further searches, improving our literature references by automatic annotations with data not stored previously. We also used the PubChem (Kim et al., 2021) API to convert large lists of compound names found in some paper supplements without structure information, to extend the structure data set, defining molecular compounds by drawing molecular graphs.

While interactive web pages can provide easy access to search perspectives of common interest, deeper analysis or rapid prototyping requires a more flexible way of interacting with the stored data not only in our system but also when cross-linking to other resources. Since we cannot implement all potential use-cases, but as we suggest that this tool is used for other applications, we provide access to the raw data tables that are accessed via pipeline tools like KNIME (Berthold et al., 2007) to allow rapid building of a prototype workflow for specific aspects to be analyzed within the data and to motivate to generate a specific web view. An example of how this might look like can be seen on https://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpaceInfluenceOnCFTRFunction, where we use the open source ECharts-JavaScript library ECharts first developed by Baidu (Li et al., 2018a) and now the Apache Foundation (Apache ECharts, 2021) to load a JavaScript Object Notation (JSON) formatted data set on the client side to depict our chemical space as an interactive scatterplot with CFTR annotations and links to the landing pages of the respective compounds, where one can investigate further annotations such as references.

Modeling a KNIME Workflow to Retrieve and Extend Spreadsheet-Defined Content

Starting with a list of papers extracted from a literature search (e.g., PubMed (Kim et al., 2021) search term "CFTR"), one is left with the task of organizing how to first get the papers and then extract the information. At first, we do not know what type of data we might extract (e.g., pictures, supplements, sketches), and it might differ very much between individual papers as a source. Thus, it is advantageous to have a folder with individual subfolders to collect the data for each paper, ideally named so that we can identify it quickly (e.g., first author and year). The goal is to copy the paper into the respective folder and then start a spreadsheet to summarize the excerpts (e.g., compound identifiers, assay). One can create and fill these folders automatically using KNIME (Berthold et al., 2007) and an input list of references coming from the citation manager, programming it to cycle through the list and creating the folders with names simply concatenated from first author, year, and maybe PubMed (PubMed, 2021) ID (PMID). Either automatically or if something similar was already done by hand, one can import folder names to guide later analysis and quality checking by annotating data found within the folders. Using this system allows to retrace where the data came from and in case of missing or faulty data one can go back to one particular folder and source spreadsheet file. In other words, this method requires accepting the use of spreadsheets as the first step in the processing of data but enables an improvement of quality by iterating over its content. Going from the unstructured list of papers mainly in PDF format, obtained from the initial PubMed search, sorting the data into such an organized folder structure is providing first means to organize the data further and the resulting data stack can be seen as a pseudo-database, which can be parsed by computational approaches and transformed further to answer specific question (see also **Figure 1**).

A KNIME (Berthold et al., 2007) pipeline looks for all xlsx files it encounters in a specified folder location and will also cycle through any subfolders it encounters within. After the list is defined, a loop will load all files and concatenate the columns where possible, additionally the path of the information is accessed, which can be broken down for folder and file names. To behave as a means to collect chemical information like structures, we defined a minimum column to be "smiles" (Weininger, 1988), which should contain an isomeric SMILES representation of the molecule "(C1=CC(=CC=C1C2=COC3=CC(=CC(=C3C2=O)O)O)O)O)"; for quality control the InChIKey (Southan, 2013), if given in the source, is recommended, as the SMILES encoded structure can be deterministically converted to an InChIKey. Another identifier is the PCID (PubChem Compound Identifier) used at the PubChem repository. For defining the literature source it came from, we recommend, if possible, to use the PubMed (Kim et al., 2021) ID (PMID) as we can extract all other information (title, authors, doi, journal, . . .) from PubMed (2021). These general information columns, which hold data that are valid for each entry in the table collecting the information retrieved by the pipeline during the loop, have to be filled only in the first row and the pipeline takes care of extending this information to each row, after the import of the xlsx file. All additional columns will be collected and attached one after the other, with columns with the same name occurring in multiple files being merged into one column. When the pipeline is executed again, it updates the mapping of the columns and joins the information. Analogously, for errors uncovered, one always has the option to alter the pipeline structure in KNIME to fix this in the workflow or in the source data. The benefit is that the current state of the data at execution time is preserved and one can then export the collected data in various formats or upload the data to the server instance.

Interfacing With PubChem-API

To realize CandActCFTR, we have used KNIME (Berthold et al., 2007) to interface the PubChem (Kim et al., 2021) service, the PubMed (PubMed, 2021) service, and chemical tool kits such as OpenBabel (2021). We use multiple services of the PubChem (Kim et al., 2021) web service to amend our data sets, where possible. Straight forward, to resolve structures of entries with reported PubChem Identifier (PCID), the linked information on the PubChem (Kim et al., 2021) servers, such as the isomeric SMILES (Weininger, 1988) and InChIKey (Southan, 2013; InChI Trust, 2021), can be pulled for entries that did not yet specify

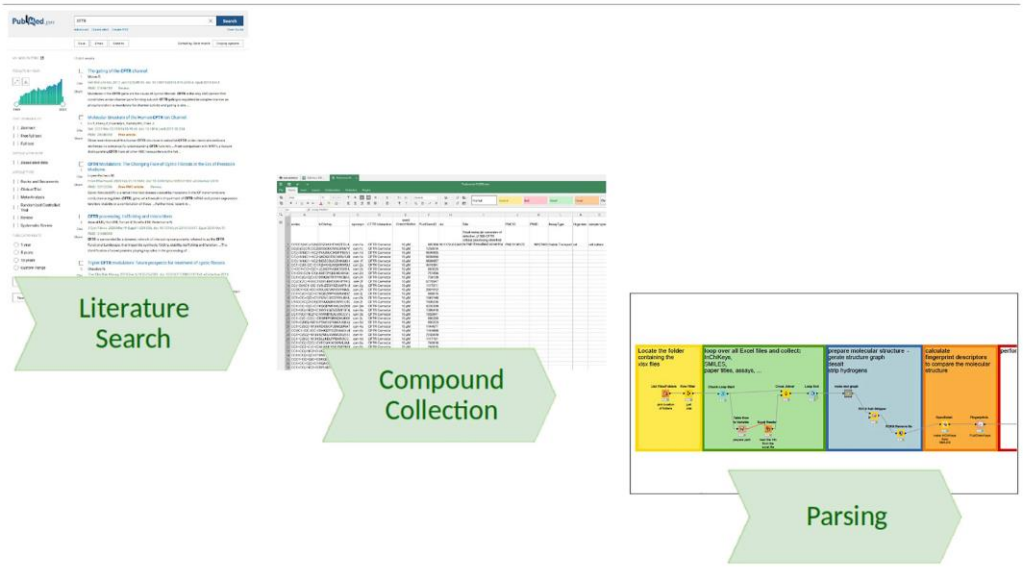


FIGURE 1 | Information flow from literature search to database. Beginning with a literature search (e.g., via PubMed), we collect the PDFs in a structured folder system and convert the information into a tabular format, adhering to a common minimal column requirement (e.g., SMILES, PubMedID, CFTR Interaction), allowing for any other relevant column definition, and finally parsing the file stack using KNIME pipelines, consolidating the tables, extending the information with annotations from other sources such as synonyms from PubChem, and calculating molecular descriptors for a subsequent similarity search.

them yet. *Vice versa*, check if we can resolve an InChIKey (InChI Trust, 2021) to a PubChem (Kim et al., 2021) entry, for linking our entry to this additional source. In both use cases, we can then query for synonyms of the compound if a PCID exists, thus the initial list that is checked for coverage in PubChem is amended by a synonyms list queried from this service. This list might be used for further direct matching with other data sources (e.g., papers, patents), which only contain names for the compounds but no chemical information.

Interfacing With PubMed-API

We also used the PubMed (PubMed, 2021) web services to query for additional information to amend our data set where having a PMID to directly get the meta-data for a specific literature resource and in batch mode for multiple entries at once. If only information like the "title" is known, we can use this to inquire PubMed (PubMed, 2021) for potential matches and in case of a defined match amend the meta-data. KNIME (Berthold et al., 2007) can be extended with multiple open-source chemistry-related tool kits (Chemistry Development Kit, 2021; Open Babel, 2021; RDKit, 2021). Thus, after importing a structure containing column from a data source, these toolkits are used to convert the molecules into different formats, for instance, to export Structure Data Format (SDF) files or create molecular descriptors for which CandActCFTR uses RDKit (2021).

KNIME-Based Similarity Search

The similarity search provided by our webserver is encapsulated in another KNIME workflow and uses a list of SMILES as input, which are then combined with the data stored in the SQL database. As KNIME enables the storage of data within a workflow, we can provide our preloaded data set as a reference to be used also in offline environments, or keeping specific archived versions. We are using the PubChem fingerprints computable by the chemoinformatics nodes in KNIME for similarity search/calculating distances. For the creation of the PCA coordinates, we use the available standard descriptors (e.g., MW, logP) in KNIME and use the correlation filter to reduce the dimensionality removing redundant information before calculating the PCA. The PCA serves mainly to help visualize the chemical space for the online site and serves to get a first glance of potential overlap regions. In the future, we might pass the descriptors through as well to be selected by the user and enable 3D views like with our similarity search results.

Web Resources

The CandActCFTR webpage is provided at: https://candactcftr.ams.med.uni-goettingen.de/.

The software, the CFTR data content, and documentation are provided at: https://gitlab.gwdg.de/mnieter1/CandActBase.

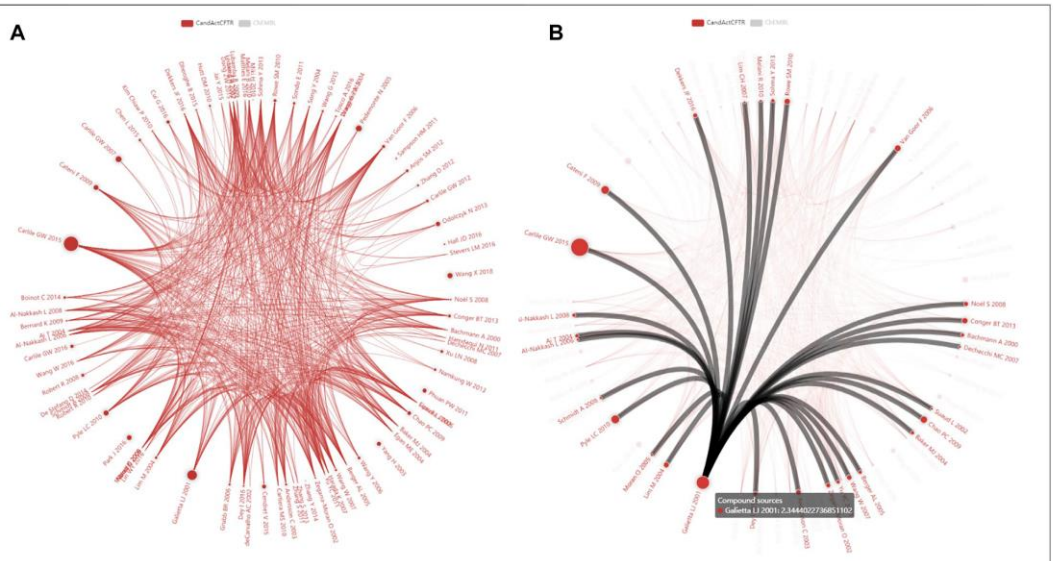


FIGURE 2 | Impact of individual publications on CandActCFTR content. The screenshots **(A, B)** in the diagram show each publication as a circle whereby the number of substances is visualized as the circle diameter (Log10 scaled). Connections between contributions describe substances that are listed in more than one publication. **(A)** represents the actual CandActCFTR publication dataset, hiding the also loaded ChEMBL reference data set and is displayed on our website at http://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpaceCompoundOverlaps BetweenPublications where it is provided with a zoom function, and **(B)** shows additional information on links and on publications upon mouse-over of connections or circles, e.g., all publications highlighted with at least one overlap compound to Gallietta et al. (2001). Interestingly, a contribution with a large content but only few shared substances exist (e.g., Pedemonte et al., 2005) and some authors describe a set of unique substances (e.g., Miki et al., 2010; Phuan et al., 2011; Sampson et al., 2011; Hall et al., 2016). Below this diagram on the webpage are tables summarizing the interactions with individual overlap counts, providing also links to the publications.

RESULTS

Within the last 3 years, we have collected data on molecules tested as CF therapeutics from several resources, screening the literature and chemical databases for substances that are reported to activate CFTR residual chloride secretion or elevate the expression of CFTR and compiled this information in a database. Until 10/2020, we could derive data from 108 publications (**Figure 2**) on 3,109 CFTR-relevant substances via the literature database PubMed (PubMed, 2021) and further 666 substances via ChEMBL (Gaulton et al., 2012) whereby only 19 substances were shared between these sources. This poor overlap was expected as databases such as ChEMBL, DrugBank (Wishart et al., 2018), and others recruit their content from different sources and thus share only a minority of structures (Southan et al., 2013). Strategies realized by the researchers to uncover CFTR therapeutic substances were 1) direct testing of a therapeutic substance that resembles an already known CFTR activating substance structurally (example: flavonoids) or influences a pathway that is known to be vital for CFTR (example: inhibitors of autophagy) or 2) high-throughput screening of a chemical substances bank with a CFTR-relevant assay.

Our database, described in detail in methods, is assembled from the following principle modules: The software framework uses open-source resources and integrates existing tools and resources, which allows CandActCFTR to be repurposed for adaptation to other use cases. The software is based on Grails Framework (2021) for which content is provided by the established data analysis pipeline tool KNIME (Berthold et al., 2007). PubMed (2021) was mined to find a list of relevant entries to further inspect, which were organized as a shared literature list using the Zotero web service (Zotero, 2021). Literature meta-data were incorporated into our system through the Citation Style Language schema of the Citation Style Language (CSL) project (Citation Style Language, 2021). The chemical structure is the root of our data model and all other information is linked directly or indirectly with this root. Direct incorporation of chemical structures was facilitated by the Ketcher (Ketcher, 2021) JavaScript plugin that provides the means to draw molecular structures and translates these graphs to isomeric SMILES (Weininger, 1988). CandActCFTR next converts isomeric SMILES to an InChIKey (Southan, 2013; InChI Trust, 2021) using OpenBabel (2021). Additional information about chemical structures were retrieved from information resources such as the shared literature and converted to database content by providing identifiers or drawing the molecular graph. Synonyms, isomeric SMILES, and InChIKey (Southan, 2013) are retrieved from PubChem's Web API (Kim et al., 2021). Molecular graph images on CandActCFTR were first provided by accessing

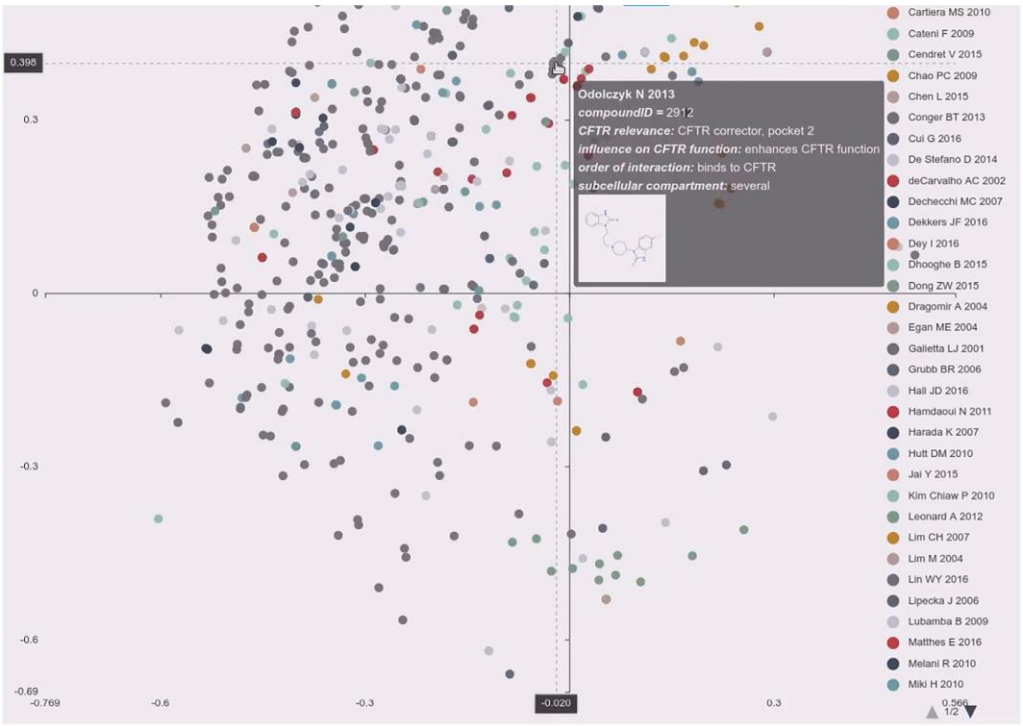


FIGURE 3 | Exploring the distribution of individual publications over the chemical space. The screenshot shows the principal component analysis plot of the chemical descriptors labeled by publication. The plot is interactive and using the legend the users can select/deselect certain publications. It supports zooming and upon mouse-over on a data point the quick summary is displayed consisting of the paper references short name, compound id, main annotations like *CFTR relevance*, *influence on CFTR function*, *order of interaction*, *subcellular compartment*, and the molecules graph. Actual clicking on the dot will load the specific compound page in a new tab in the browser (https://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpacePaperAssociationSortedBySize).

PubChem's (Kim et al., 2021) Web API or using the Kekule JavaScript libraries, but are now replaced by pre-generated images using OpenBabels "SMILES to PNG" function. Processed CandActCFTR content for the web page is provided, for instance, through the open source ECharts-JavaScript library (eCharts Baidu (Li et al., 2018a) is now part of the Apache Foundation (Apache ECharts, 2021)), which receives a JSON formatted data set, provided by Grails (Grails Framework, 2021), on the client side, to depict our chemical space as an interactive scatterplot with CFTR annotations.

Data on CandActCFTR substances are provided on our webpage at https://candactcftr.ams.med.uni-goettingen.de/. The ordering principle is the chemical structure of the compound, which we archive by its isomeric SMILES (Weininger, 1988) and its corresponding InChIKey (InChI Trust, 2021). We use the InChIKey (InChI Trust, 2021) as a unique identifier to retrieve corresponding entries from PubChem (Kim et al., 2021), and add links to those resources. Generic names and all available used synonyms are provided with each compound. Compounds are affiliated with all publications in which the compound is mentioned, and the key message of the publication is provided to the reader as a short reference-into-function (RIF) text. Apart from information about the project, we provide

- A site that allows searching for compounds by names, SMILES (Weininger, 1988), or InChIKey (InChI Trust, 2021). Using structural information encoded in SMILES format, we also provide the means for similarity search, also for lists of compounds. Hereby, novel structures can also be drawn into a window and directly converted to SMILES (Weininger, 1988) using Ketcher (Ketcher, 2021) (see https://candactcftr.ams.med.uni-goettingen.de/Compound/searchCompounds and **Figure 5** and **Supplementary Videos S2–S4**).
- Information/data on the individual compound summary pages containing chemical structure information, synonyms, affiliated publications with explanatory RIF text, classification information for its influence on CFTR, its order of interaction with CFTR, and the cellular compartment in which it works. As the InChIKey is a

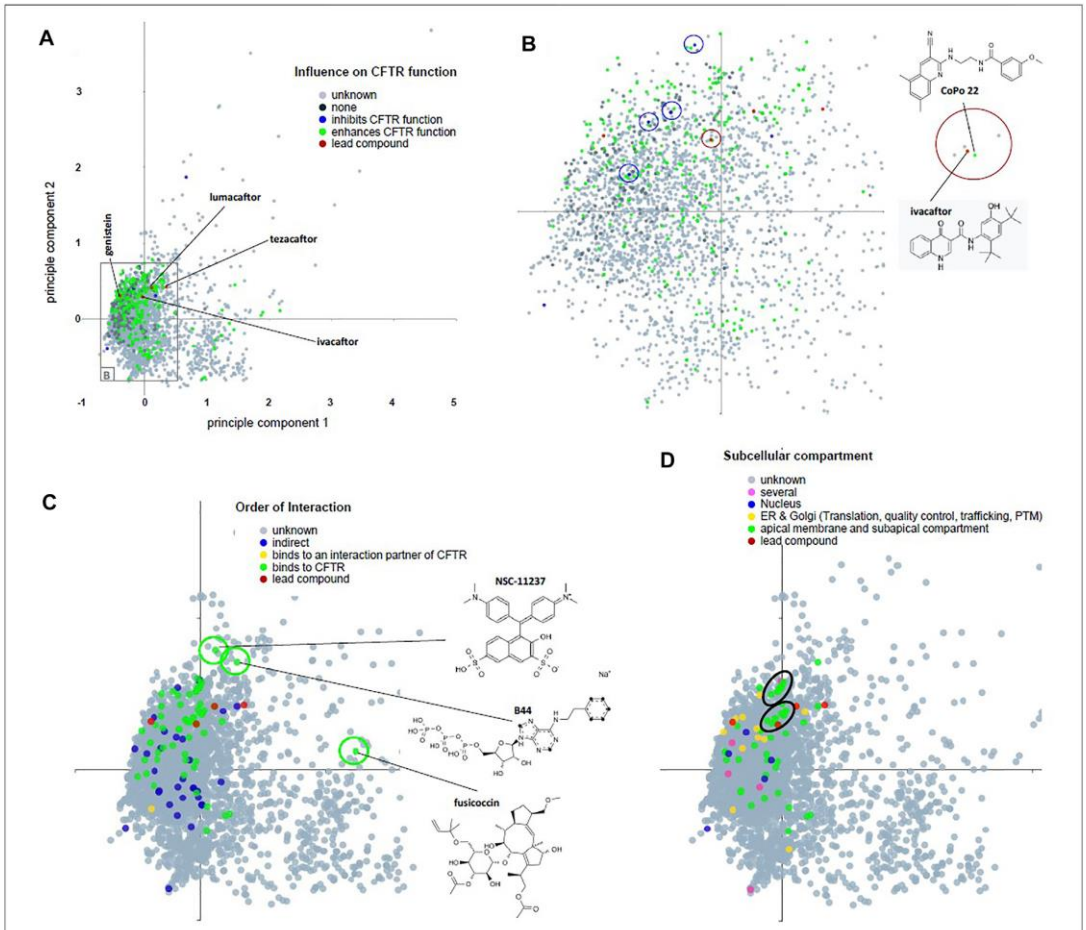


FIGURE 4 | Principal component analysis of descriptors for chemical attributes of CandActCFTR compounds and their classification in three categories. **(A–D)** Genistein and the Vertex molecules lumacaftor, ivacaftor, and tezacaftor are represented by red symbols (labeled in **A**) as lead compounds. **(A)** Dataset annotated for influence on CFTR function (see **Table 1**). Inconsistent assignment and unknown are summarized as one category (gray). Compounds that enhance CFTR function occupy the central space of the PC diagram whereby their occurrence roughly reflects the density of compounds tested. **(B)** Compound-rich central area as indicated in **(A)**. We are aware that the current PC analysis projects a multidimensional space into two dimensions; as a consequence, compounds in close proximity are unlikely to share all attributes. However, we have noticed that CFTR inhibitors and CFTR activators are closely located (blue circles) and that the lead compound ivacaftor has a nearby neighbor CoPo22 (InChI Trust, 2021). **(C)** Compounds stratified by order of interaction. Three structural outlier compounds are marked by a green circle, indicating an area of interest in the chemical space in a region where few tested compounds are within our database. Outliers are fusicoccin (Chemistry Development Kit, 2021), the non-hydrolyzable ATP analogon B44 (RDKit, 2021), and NSC-11237 (Gaulton et al., 2012). **(D)** Compounds stratified by subcellular compartment. Two clusters of substances that act at the apical membrane in the subapical compartment are encircled in black. Interactive PC diagrams are provided at https://candactcftr.ams.med.uni-goettingen.de/Compound/index.

universal identifier also used on many other sites, we linked a Google search using the InChIKey for convenience to identify other resources, e.g., potential vendors (see https://candactcftr.ams.med.uni-goettingen.de/Compound/cycleCompounds/1).

- Information/data on all compounds: an interactive depiction of the chemical space derived from their chemical descriptors via a principal component analysis (see, e.g., https://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpaceInfluenceOnCFTRFunction and **Figures 3**, **4**), showing as popups on mouse-over on a data point also the structure and main annotations like *CFTR relevance*, *Influence on CFTR function*, *Order of*

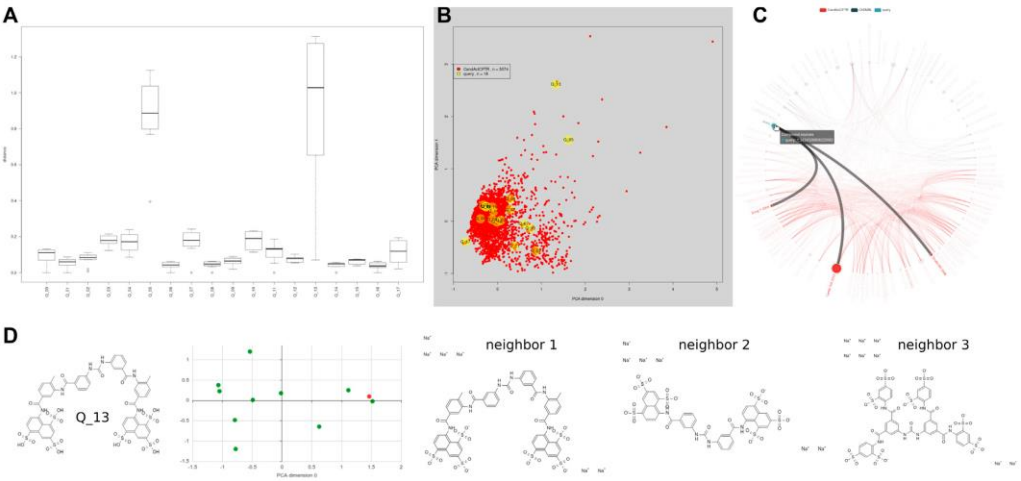


FIGURE 5 | Using the similarity search to investigate the overlap of new data sets of interest with the annotated chemical space, for example, a small ENaC data set. Forty-six entries from 18 different publications about ENaC, resulting in a set of 18 distinct structures, some with multiple entries distributed over publications, e.g., 11× amiloride. The resulting set was used for a similarity search using the form @ https://candactcftr.ams.med.uni-goettingen.de/Compound/searchCompounds. **(A)** Analysis identified six direct overlaps and shows the relative distance of the closest 10 compounds, **(B)** 16 out of 18 query structures cover generally the same area as the existing CandActCFTR point cloud and two rather distinct compounds can be seen, but one entry has a single close neighbor in CandActCFTR space. **(C)** Three listed papers contain an overlap, where Carlile et al. (2015), Grubb et al. (2006), and Song et al. (2004) all contain amiloride; while Carlile et al. (2015) also contained exact matches for benzamil, phenamil, LY-294002, pioglitazone, and rosiglitazone. **(D)** Using the individual structure similarity search, one can have a more distinct look at the structures surrounding the query structure; for example, looking at the query Q_14 suramin and its 10 closest neighbors, we can see that the closest neighbor is actually the query structure with another ionization state. Q_05 Vasopressin is a cyclic polypeptide and has no close by neighbors, but its assigned closest neighbor is another cyclic polypeptide called tyrothricin.

interaction, and *Subcellular compartment*. Upon mouse-click, the corresponding compound page is loaded.

- Information/data on all publications: an overview of relative size and overlaps between publications is provided, also including the ChEMBL-derived reference data set as connection graph and table view (see https://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpaceCompoundOverlapsBetweenPublications or alternatively using the chemical space depiction again but colored by paper association instead https://candactcftr.ams.med.uni-goettingen.de/Compound/showFullSetChemSpacePaperAssociationSortedBySize).

See also the short supplement video captures of using the webservice:

- Search_Ivacaftor_By_Name.mp4
- Search_Ivacaftor_By_Drawing_Exact_Structure.mp4
- Search_Ivacaftor_By_Drawing_Structure_One_MethylGroup_Off.mp4 - implicitely activating similarity search
- Exploring_PointCloud_TaggedByPaper_with_CompoundImagesAndAnnotationPopUps.mp4

The following examples may serve to illustrate that the effort undertaken by collecting data from different resources and through the joint analysis of these data has already revealed valuable insights into how CFTR-acting substances can be identified and understood:

1. Among the 3,109 substances listed in CandActCFTR, 145 molecules do not have a corresponding entry in PubChem (Kim et al., 2021) or ChemSpider (2021), which indicates that there currently is no single comprehensive database on chemical substances in the public domain.
2. We now have an overview of systematic screens for compounds undertaken for cystic fibrosis in academia. Verkman, University of California, San Francisco (Galietta et al., 2001; Yang et al., 2003; Pedemonte et al., 2005; Phuan et al., 2011; Namkung et al., 2013); Galietta, Genova, Italy (Galietta et al., 2001; Pedemonte et al., 2005; Cateni et al., 2009; Pedemonte et al., 2011); and Hanrahan & Thomas, McGill University, Montreal, Canada (Carlile et al., 2007; Robert et al., 2008; Carlile et al., 2012) contribute most data in that regard and dominate the field; in addition, Xu et al. have screened Chinese medicinal herbs for substances that act on CFTR (Xu et al., 2008). Moreover, Cui et al. (2016) have used a pharmacophore modeling approach to predict CFTR activating substances.

TABLE 1 | Classification of CandActCFTR substances in 2019/2010

Category: influence on CFTR function	
Enhances CFTR function	354
Likely enhances CFTR function	31
Inhibits CFTR function	14
None	217
Inconsistent assignment	18
Unknown	2,441
Category: order of interaction	
Binds to CFTR	83
Binds to an interaction partner of CFTR	1
Indirect	68
Unknown	2,923
Category: subcellular compartment	
Apical membrane & subapical compartment	88
ER & Golgi (translation, quality control, trafficking, PTM[a])	50
Nucleus (transcription)	8
More than one compartment	71
Unknown	2,858

[a]*Post-translational modification, e.g., glycosylation.*

3. Ibuprofen and glafenine are both identified as CFTR-activating substances (Robert et al., 2010; Carlile et al., 2015). Both are nonsteroidal anti-inflammatory drugs (NSAIDs). It is interesting to note that these two NSAIDs both target and partially correct the CF-typical, proinflammatory status as genes that determine immunology and inflammation, having been uncovered as CF modifying genes, target the basic defect of impaired ion conductance in cystic fibrosis epithelia as well (Stanke et al., 2011). In other words, the two NSAIDs suggested as CFTR activating substances and the identified CF modifying genes both emphasize the weight of the inflammatory pathway for the manifestation of the CF basic defect.
4. Some substances have been tested extensively in several biosystems by different groups. For instance, resveratrol has been published as a CFTR activating compound in five cell lines and primary cells by CFTR protein visualization, by CFTR function in transepithelial current measurements in primary cells as well as animal tissues and by patch clamp in two heterologous expression systems and *in vivo* in a mouse model (Hamdaoui et al., 2011; Zhang et al., 2013; Zhang et al., 2014; Dhooghe et al., 2015; Jai et al., 2015). Even though these bioassays appear to cover the entire spectrum of CFTR-relevant assays, no conclusion is reached by the field on the application of resveratrol as a therapeutic agent as two research groups conclude that resveratrol does not work or even inhibits CFTR.
5. A similar controversy is seen for miglustat, tested in 10 different bioassays encompassing cell lines and data from mouse models (Norez et al., 2006; Lubamba et al., 2009; Norez et al., 2009; Jenkins and Glenn, 2013; Leonard et al., 2013; Europe PMC, 2021; Noël, 2021). A follow-up investigation on whether or not the substance activates CFTR in nasal epithelium in CF patients reports no effect. However, recently it was noticed that even approved drug Orkambi does not improve NPD in all cases (Graeber et al., 2018), and moreover, that the correction of the basic defect by Orkambi did not predict the influence of Orkambi on the clinical parameters assessed by Graeber and colleagues, suggesting that a weak performance of the substance miglustat might not be contradictory to the primary data obtained in the preclinical phase.

We conclude from our survey that even if a substance has been confirmed as a CFTR activating agent in several bioassays, the field does not rely on these data for choosing such a substance as a chemical scaffold for future drug development. Thus, relevant information can be overlooked. While CandActCFTR collects data from CFTR-related screens, its structure can easily be adapted to other data collections. As an example, we here provide our second use case for ENaC-activating substances, and show the similarity search for overlaps of CFTR-tested compounds with ENaC-tested compounds, depicted in **Figure 5**. The query contains suramin, which is also present in our CandActCFTR data set, derived from Carlile et al. (2015), where it is part of a screening library. Looking further into this entry, we realized we are missing detailed annotation for this compound and started to investigate broader from more sources for this entry point. As suramin has many different targets (72 potential targets according to Wiedemar et al. (2020) it seems at first glance like a good candidate for initial screening, yet Bachmann et al. published already in 1999 (Bachmann et al., 1999), the "*potent inhibition of the CFTR chloride channel by suramin*." Thus, merging of available resources can help to plan experiments and interpret results by using additional annotations, in this case to extend the annotation of suramin to CFTR inhibitor.

Substances can activate CFTR at several steps during its maturation pathway from a nascent polypeptide chain to a fully functional membrane protein (Lukacs and Verkman, 2012; Veit et al., 2016): CFTR mRNA is transcribed in the nucleus and translated into a polypeptide chain in the endoplasmatic reticulum. If misfolded, CFTR is recognized by the ER quality control, and the protein is degraded by the proteasome (ER-associated degradation, ERAD pathway). Correctly folded CFTR is promoted to the Golgi apparatus by the ER-associated folding pathway ERAF. In the Golgi apparatus, CFTR is complex glycosylated by MGAT enzymes and transported via vesicles to the apical membrane. In the subapical compartment, CFTR can be degraded via lysosomes (LY). Only 30% of wild-type CFTR are processed to the apical membrane and many CF-causing mutations such as F508del undergo much lower maturation rates (Lukacs and Verkman, 2012). Thus, we sought to annotate the substances according to three distinct functional categories: according to their influence on CFTR (active/inactive/inhibitory), their order of interaction (direct = binds to CFTR, indirect = influences a pathway that is important for CFTR), and the cellular compartment in which the substance causes the CFTR-relevant effect (nucleus, ER and Golgi compartment, apical membrane, and subapical compartment) (**Table 1**). The functional categories have been retrieved from the assessment of the authors and concatenated for all publications that report on one substance. This sum of author-guided statements, retrieved by conventional text mining, on how the

substance acts on CFTR has next been used to assign an entry in each category to the substance.

We next have analyzed the data set annotated for influence on CFTR function, order of interaction, and subcellular compartment by principal component analysis (**Figure 4**). Compounds that enhance CFTR function occupy the central space of the PC diagram whereby their occurrence roughly reflects the density of compounds tested. Three structural outliers were seen, indicating an area of interest in the chemical space in a region, where few tested compounds are within our database, which are fusicoccin (Stevers et al., 2016), the non-hydrolyzable ATP analogon B44 (Miki et al., 2010), and NSC-11237 (Odolczyk et al., 2013). When stratified by subcellular compartment, we could observe two clusters of substances that act at the apical membrane in the subapical compartment. However, some CFTR inhibitors and activators were located closely together, suggesting that the PC analysis needs to be adapted with respect to its dimensions if a segregation of such contrasting functionalities in distinct clusters has to be achieved. In summary, CandActCFTR collects data from different resources such as 12 systematic screens for compounds undertaken for cystic fibrosis in academia (Galietta et al., 2001; Yang et al., 2003; Pedemonte et al., 2005; Carlile et al., 2007; Robert et al., 2008; Xu et al., 2008; Cateni et al., 2009; Pedemonte et al., 2011; Phuan et al., 2011; Carlile et al., 2012; Namkung et al., 2013; Cui et al., 2016). Through the joint analysis of these data, the chemical space used by the compounds becomes accessible and can be employed for compound selection: when chemical structures of the tested compounds are displayed in a principal component analysis in their chemical space, structural similarities between compounds that share annotated features such as "mode of action" or "subcellular compartment" are visualized as clusters of substances in the chemical space. These clusters suggest a highly attractive area of the chemical space for substance optimization.

DISCUSSION

In the digital age where a huge amount of information is available, it is advantageous to organize such knowledge in a meta-database for retrieval and analysis of data. While in theory a good data structure design comes first, there is the issue of first knowing what the content of the meta-database is going to be (items and linkage between items). Under this condition, the entry forms can be designed first and the data can be directly entered into the structured database. In reality, one does not necessarily know in advance of starting screening publications for their information, which of the content is deemed to be interesting or needs be accumulated for analysis. To transform rather unstructured data into uniformly structured data, we looked at the process applied by the researcher using multiple screening and rescreening rounds to refine the data collection with each cycle. Many people rely on spreadsheets to organize this data. We identified a need regarding the handling of such a plethora of changing input tables, and joining the data from different tables with varying column titles, as well as scaling up from dozens to hundreds of publications. To tackle this goal of joining various sets of information extracted from literature texts, transforming them into tabular organized information excerpts, and to furthermore enable the organization of the content, we defined rules for these transformations. Our project can be used as an exemplary case on how to proceed when spreadsheets become the main entry mask format for a literature excerpt-based information organization and aggregation system. We propose using tools like the graphical workflow manager KNIME (Berthold et al., 2007), which can be taught to people untrained in IT processing within a very short time period to organize their data collection and ultimately clean their data, so that it is fit for analysis or distribution using web tools like a Grails (Grails Framework, 2021) server. Thus, we also provide our KNIME workflows accompanying the webserver, which can also be used independently.

We provide the software tool CandActCFTR, which can be repurposed for adaptation to other use cases and applications where chemical compounds with the structure, synonyms, InChIKey (Southan, 2013; InChI Trust, 2021), and literature are of interest and we provide our seed content dataset on CFTR-relevant substances (GitLab, 2021). CandActCFTR is a comprehensive research tool combining information on a growing amount of CFTR acting substances from different sources, mainly retrieved from publications in scientific journals, abstracts, and presentations on scientific meetings. CandActCFTR in its current form can be installed and operated at other sites. We have implemented a principal component analysis to visualize the similarity of substances and we have handled requests from the CF community to answer whether a certain substance is similar in structure to a CFTR activator listed in CandActCFTR (**Figures 2** and **5**).

Chemical properties of all substances in CandActCFTR have been assessed using principal component analyses (**Figure 4**) to identify those areas within the chemical space that are occupied by true-active substances. Other substance-related databases are available for medicinal and aromatic plant's aroma molecules (Kumar et al., 2018), on therapeutic targets in *Campylobacter jejuni* (Hossain et al., 2018), a therapeutic targets database (Li et al., 2018b), and a database of structurally annotated therapeutic peptides (Singh et al., 2016). These are specialized databases like CandActCFTR, but among these examples, only CandActCFTR is provided as a generic tool that can be adapted by interested researchers to collect and analyze their data for other diseases and other therapeutic targets. Furthermore, the data set provided by CandActCFTR is, to the best of our knowledge, unique in that it compiles comprehensively the literature on CFTR activating substances and thus enables meta-analysis (**Figure 4**).

It was shown for CFTR-targeting cystic fibrosis molecular therapeutics that combination therapies are superior to approaches that build on single substances (Phuan et al., 2018; Southern et al., 2018; Veit et al., 2018), and thus our category definition considering subcellular compartmentation and the order of interaction will assist

in selecting candidate therapeutics for that act on CFTR at several steps vital for CFTR gene expression, protein maturation, and activation. This phenomenon reflects that mutant CFTR is functionally deficient in several aspects (Veit et al., 2016): F508del is known as a processing mutant, failing to mature properly and at the same time, F508del-CFTR is functionally impaired if it reaches the apical membrane (Veit et al., 2016). To correct both properties of F508del-CFTR, the current therapeutic Orkambi combines a substance that promotes CFTR maturation and a substance that activates F508del-CFTR. Veit et al. (2018) and Phuan et al. (2018) have recently combined CFTR acting compounds to correct mutant CFTR whereby, interestingly, the individual substances had only a minor influence on CFTR (Veit et al., 2018). We have noticed that most substances in CandActCFTR could not yet be categorized (influence on CFTR—80% unknown, order of interaction—95% unknown, cellular compartment—93% unknown; **Table 1**) and envisage that future data will enable us to assign more substances to specific categories.

CONCLUSION

CandActCFTR is a pilot project to merge data from publicly available sources and establish a database of candidate cystic fibrosis therapeutics for the activation of CFTR-mediated ion conductance. The acquired information on tested substances will assist in the identification of the most promising candidates for future therapeutics. Besides its specific application to identify CFTR therapeutics, we provide the software base of CandActCFTR as a tool for other chemoinformatics applications where properties of chemical molecules are at the core of interest; https://gitlab.gwdg.de/mnieter1/CandActBase. By not only providing a web service but also distributing the KNIME workflows used to prepare the loading of the data into our databases backbone, as well as the processing recipes for similarity searching, we hope to provide the tools for the community with the necessary flexibility to be of use in the future.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

AUTHOR CONTRIBUTIONS

Conception and design of the database: MN and FS. Implementation of the software: MN. Supporting Grails implementation: FA. Acquisition, analysis, and interpretation of data: MN, LV, SH, and FS. Drafting the article or revising it critically for important intellectual content: MN, SH, and FS.

FUNDING

This work was funded by the Deutsche Forschungsgemeinschaft DFG (gepris: 315063128). https://gepris.dfg.de/gepris/projekt/315063128.

ACKNOWLEDGMENTS

The CandActCFTR project is supported by a scientific advisory board (SAB), consisting of well-known CF researchers, which we like to thank here for their input and feedback to our project: Prof. Dr. Luis Galietta (Istituto G. Gaslini, Genua, Italy), Prof. Dr. Frederic Becq (University Poitiers, France), Prof. Dr. Ulrich Martin (Medical University of Hannover, Germany), Prof. Dr. Bertrand Kleizen (University of Utrecht, Netherlands), and PD Dr. Nico Derichs (Charité Berlin, Germany).

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fphar.2021.689205/full#supplementary-material

REFERENCES

Apache ECharts (2021). Apache ECharts. Available at: https://echarts.apache.org/en/index.html (Accessed March 31, 2021).

Bachmann, A., Russ, U., and Quast, U. (1999). Potent Inhibition of the CFTR Chloride Channel by Suramin. *Naunyn Schmiedebergs Arch. Pharmacol.* 360 (4), 473–476. doi:10.1007/s002109900096

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., et al. (2007). "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization* (Springer).

Carlile, G. W., Keyzers, R. A., Teske, K. A., Robert, R., Williams, D. E., Linington, R. G., et al. (2012). Correction of F508del-CFTR Trafficking by the Sponge Alkaloid Latonduine Is Modulated by Interaction with PARP. *Chem. Biol.* 19 (10), 1288–1299. doi:10.1016/j.chembiol.2012.08.014

Carlile, G. W., Robert, R., Goepp, J., Matthes, E., Liao, J., Kus, B., et al. (2015). Ibuprofen Rescues Mutant Cystic Fibrosis Transmembrane Conductance Regulator Trafficking. *J. Cyst. Fibros.* 14 (1), 16–25. doi:10.1016/j.jcf.2014.06.001

Carlile, G. W., Robert, R., Zhang, D., Teske, K. A., Luo, Y., Hanrahan, J. W., et al. (2007). Correctors of Protein Trafficking Defects Identified by a Novel High-Throughput Screening Assay. *ChemBioChem* 8 (9), 1012–1020. doi:10.1002/cbic.200700027

Cateni, F., Zacchigna, M., Pedemonte, N., Galietta, L. J., Mazzei, M. T., Fossa, P., et al. (2009). Synthesis of 4-Thiophen-2'-yl-1,4-Dihydropyridines as Potentiators of the CFTR Chloride Channel. *Bioorg. Med. Chem.* 17 (23), 7894–7903. doi:10.1016/j.bmc.2009.10.028

CFF (2021). Drug Development Pipeline. Available at: https://www.cff.org/Trials/Pipeline (Accessed March 31, 2021).

Chemistry Development Kit (2021). Chemistry Development Kit. Available at: https://cdk.github.io/ (Accessed March 31, 2021).

ChemSpider (2021). Search and Share Chemistry. Available at: http://www.chemspider.com/ (Accessed March 31, 2021).

Citation Style Language (2021). Citation Style Language. Available at: https://citationstyles.org/ (Accessed March 31, 2021).

Cui, G., Khazanov, N., Stauffer, B. B., Infield, D. T., Imhoff, B. R., Senderowitz, H., et al. (2016). Potentiators Exert Distinct Effects on Human, Murine, and Xenopus CFTR. *Am. J. Physiol. Lung Cel Mol Physiol* 311 (2), L192–L207. doi:10.1152/ajplung.00056.2016

Dhooghe, B., Bouckaert, C., Capron, A., Wallemacq, P., Leal, T., and Noel, S. (2015). Resveratrol Increases F508del-CFTR Dependent Salivary Secretion in Cystic Fibrosis Mice. *Biol. Open* 4 (7), 929–936. doi:10.1242/bio.010967

Elborn, J. S. (2016). Cystic Fibrosis. *Lancet* 388 (10059), 2519–2531. doi:10.1016/s0140-6736(16)00576-6

Europe PMC (2021). A Randomized Placebo-Controlled Trial of Miglustat in Cystic Fibrosis Based on Nasal Potential Difference. Available at: https://europepmc.org/article/MED/22281182 (Accessed March 31, 2021).

Galietta, L. J., Springsteel, M. F., Eda, M., Niedzinski, E. J., By, K., Haddadin, M. J., et al. (2001). Novel CFTR Chloride Channel Activators Identified by Screening of Combinatorial Libraries Based on Flavone and Benzoquinolizinium lead Compounds. *J. Biol. Chem.* 276 (23), 19723–19728. doi:10.1074/jbc.M101892200

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: A Large-Scale Bioactivity Database for Drug Discovery. *Nucleic Acids Res.* 40, D1100–D1107. doi:10.1093/nar/gkr777

Gentzsch, M., and Mall, M. A. (2018). Ion Channel Modulators in Cystic Fibrosis. *Chest* 154 (2), 383–393. doi:10.1016/j.chest.2018.04.036

GitLab (2021). Home Wiki Manuel Nietert/CandActBase. Available at: https://gitlab.gwdg.de/mnieter1/CandActBase/wikis/home (Accessed November 21, 2021).

Graeber, S. Y., Dopfer, C., Naehrlich, L., Gyulumyan, L., Scheuermann, H., Hirtz, S., et al. (2018). Effects of Lumacaftor-Ivacaftor Therapy on Cystic Fibrosis Transmembrane Conductance Regulator Function in Phe508del Homozygous Patients with Cystic Fibrosis. *Am. J. Respir. Crit. Care Med.* 197 (11), 1433–1442. doi:10.1164/rccm.201710-1983OC

Grails Framework (2021). Grails® Framework. Available at: https://grails.org (Accessed March 31, 2021).

Grubb, B. R., Gabriel, S. E., Mengos, A., Gentzsch, M., Randell, S. H., Van Heeckeren, A. M., et al. (2006). SERCA Pump Inhibitors Do Not Correct Biosynthetic Arrest of ΔF508 CFTR in Cystic Fibrosis. *Am. J. Respir. Cell Mol. Biol.* 34, 355–363. doi:10.1165/rcmb.2005-0286OC

Hall, J. D., Wang, H., Byrnes, L. J., Shanker, S., Wang, K., Efremov, I. V., et al. (2016). Binding Screen for Cystic Fibrosis Transmembrane Conductance Regulator Correctors Finds New Chemical Matter and Yields Insights into Cystic Fibrosis Therapeutic Strategy. *Protein Sci.* 25, 360–373. doi:10.1002/pro.2821

Hamdaoui, N., Baudoin-Legros, M., Kelly, M., Aissat, A., Moriceau, S., Vieu, D. L., et al. (2011). Resveratrol Rescues cAMP-Dependent Anionic Transport in the Cystic Fibrosis Pancreatic Cell Line CFPAC1. *Br. J. Pharmacol.* 163 (4), 876–886. doi:10.1111/j.1476-5381.2011.01289.x

HIT-CF (2021). What is HIT-CF Europe? Available at: https://www.hitcf.org/ (Accessed March 31, 2021).

Hossain, M. U., Omar, T. M., Alam, I., Das, K. C., Mohiuddin, A. K. M., Keya, C. A., et al. (2018). Pathway Based Therapeutic Targets Identification and Development of an Interactive Database CampyNIBase of Campylobacter Jejuni RM1221 through Non-redundant Protein Dataset. *PLoS One* 13 (6), e0198170. doi:10.1371/journal.pone.0198170

InChI Trust (2021). Developing the InChI Chemical Structure Standard. Available at: https://www.inchi-trust.org/ (Accessed March 31, 2021).

Jai, Y., Shah, K., Bridges, R. J., and Bradbury, N. A. (2015). Evidence Against Resveratrol as a Viable Therapy for the rescue of Defective ΔF508 CFTR. *Biochim. Biophys. Acta* 1850 (11), 2377–2384. doi:10.1016/j.bbagen.2015.08.020

Jenkins, B. A., and Glenn, L. L. (2013). Miglustat Effects on the Basal Nasal Potential Differences in Cystic Fibrosis. *J. Cyst Fibros* 12 (1), 88. doi:10.1016/j.jcf.2012.06.003

Jiang, C., Jin, X., Dong, Y., and Chen, M. (2016). Kekule.js: An Open Source JavaScript Chemoinformatics Toolkit. *J. Chem. Inf. Model.* 56 (6), 1132–1138. doi:10.1021/acs.jcim.6b00167

Ketcher (2021). Ketcher. Available at: https://lifescience.opensource.epam.com/ketcher/ (Accessed March 31, 2021).

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., et al. (2021). PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* 49 (D1), D1388–D1395. doi:10.1093/nar/gkaa971

KNIME (2021). KNIME Analytics Platform. Available at: https://www.knime.com/knime-analytics-platform (Accessed March 31, 2021).

Kumar, Y., Prakash, O., Tripathi, H., Tandon, S., Gupta, M. M., Rahman, L-U., et al. (2018). AromaDb: A Database of Medicinal and Aromatic Plant's Aroma Molecules With Phytochemistry and Therapeutic Potentials. *Front. Plant Sci.* 9, 1081. doi:10.3389/fpls.2018.01081

Leonard, A., Lebecque, P., Dingemanse, J., and Leal, T. (2013). Miglustat Effects on the Basal Nasal Potential Differences in Cystic Fibrosis. *J. Cyst Fibros.* 12 (1), 89. doi:10.1016/j.jcf.2012.06.004

Li, D., Mei, H., Shen, Y., Su, S., Zhang, W., Wang, J., et al. (2018). ECharts: A Declarative Framework for Rapid Construction of Web-Based Visualization. *Vis. Inform.* 2 (2), 136–146. doi:10.1016/j.visinf.2018.04.011

Li, Y. H., Yu, C. Y., Li, X. X., Zhang, P., Tang, J., Yang, Q., et al. (2018). Therapeutic Target Database Update 2018: Enriched Resource for Facilitating Bench-To-Clinic Research of Targeted Therapeutics. *Nucleic Acids Res.* 46, D1121. doi:10.1093/nar/gkx1076

Lubamba, B., Lebacq, J., Lebecque, P., Vanbever, R., Leonard, A., Wallemacq, P., et al. (2009). Airway Delivery of Low-Dose Miglustat Normalizes Nasal Potential Difference in F508del Cystic Fibrosis Mice. *Am. J. Respir. Crit. Care Med.* 179 (11), 1022–1028. doi:10.1164/rccm.200901-0049OC

Lukacs, G. L., and Verkman, A. S. (2012). CFTR: Folding, Misfolding and Correcting the ΔF508 Conformational Defect. *Trends Mol. Med.* 18 (2), 81–91. doi:10.1016/j.molmed.2011.10.003

MariaDB Foundation (2021). MariaDB Foundation. Available at: https://mariadb.org/ (Accessed March 31, 2021).

MariaDB KnowledgeBase (2021). mysqldump. Available at: https://mariadb.com/kb/en/mysqldump/ (Accessed March 31, 2021).

Martiniano, S. L., Sagel, S. D., and Zemanick, E. T. (2016). Cystic Fibrosis: A Model System for Precision Medicine. *Curr. Opin. Pediatr.* 28 (3), 312–317. doi:10.1097/MOP.0000000000000351

Miki, H., Zhou, Z., Li, M., Hwang, T. C., and Bompadre, S. G. (2010). Potentiation of Disease-Associated Cystic Fibrosis Transmembrane Conductance Regulator Mutants by Hydrolyzable ATP Analogs. *J. Biol. Chem.* 285 (26), 19967–19975. doi:10.1074/jbc.M109.092684

Namkung, W., Park, J., Seo, Y., and Verkman, A. S. (2013). Novel Amino-Carbonitrile-Pyrazole Identified in a Small Molecule Screen Activates Wild-type and ΔF508 Cystic Fibrosis Transmembrane Conductance Regulator in the Absence of a cAMP Agonist. *Mol. Pharmacol.* 84 (3), 384–392. doi:10.1124/mol.113.086348

Noël, S., Wilke, M., Bot, A. G., De Jonge, H. R., and Becq, F. (2021). Parallel Improvement of Sodium and Chloride Transport Defects by Miglustat (n-Butyldeoxynojyrimicin) in Cystic Fibrosis Epithelial Cells. *J. Pharmacol. Exp. Ther.* 325, 1016–1023. doi:10.1124/jpet.107.135582

Norez, C., Antigny, F., Noel, S., Vandebrouck, C., and Becq, F. (2009). A Cystic Fibrosis Respiratory Epithelial Cell Chronically Treated by Miglustat Acquires a Non-Cystic Fibrosis-Like Phenotype. *Am. J. Respir. Cell Mol. Biol.* 41 (2), 217–225. doi:10.1165/rcmb.2008-0285OC

Norez, C., Noel, S., Wilke, M., Bijvelds, M., Jorna, H., Melin, P., et al. (2006). Rescue of Functional delF508-CFTR Channels in Cystic Fibrosis Epithelial Cells by the Alpha-Glucosidase Inhibitor Miglustat. *FEBS Lett.* 580 (8), 2081–2086. doi:10.1016/j.febslet.2006.03.010

Odolczyk, N., Fritsch, J., Norez, C., Servel, N., da Cunha, M. F., Bitam, S., et al. (2013). Discovery of Novel Potent ΔF508-CFTR Correctors that Target the Nucleotide Binding Domain. *EMBO Mol. Med.* 5 (10), 1484–1501. doi:10.1002/emmm.201302699

Open Babel (2021). Open Babel. Available at: http://openbabel.org/wiki/Main_Page (Accessed March 31, 2021).

Pedemonte, N., Lukacs, G. L., Du, K., Caci, E., Zegarra-Moran, O., Galietta, L. J., et al. (2005). Small-molecule Correctors of Defective DeltaF508-CFTR Cellular Processing Identified by High-Throughput Screening. *J. Clin. Invest.* 115 (9), 2564–2571. doi:10.1172/JCI24898

Pedemonte, N., Zegarra-Moran, O., and Galietta, L. J. (2011). High-throughput Screening of Libraries of Compounds to Identify CFTR Modulators. *Methods Mol. Biol.* 741, 13–21. doi:10.1007/978-1-61779-117-8_2

phpMyAdmin (2021). phpMyAdmin. Available at: https://www.phpmyadmin.net/ (Accessed March 31, 2021).

Phuan, P. W., Son, J. H., Tan, J. A., Li, C., Musante, I., Zlock, L., et al. (2018). Combination Potentiator ('Co-Potentiator') Therapy for CF Caused by CFTR Mutants, Including N1303K, that Are Poorly Responsive to Single Potentiators. *J. Cyst. Fibros.* 17 (5), 595–606. doi:10.1016/j.jcf.2018.05.010

Phuan, P. W., Yang, B., Knapp, J. M., Wood, A. B., Lukacs, G. L., Kurth, M. J., et al. (2011). Cyanoquinolines with Independent Corrector and Potentiator

Activities Restore ΔPhe508-cystic Fibrosis Transmembrane Conductance Regulator Chloride Channel Function in Cystic Fibrosis. *Mol. Pharmacol.* 80 (4), 683–693. doi:10.1124/mol.111.073056

PubMed (2021). PubMed. Available at: https://pubmed.ncbi.nlm.nih.gov/.

RDKit (2021). RDKit. Available at: http://www.rdkit.org/.

Robert, R., Carlile, G. W., Liao, J., Balghi, H., Lesimple, P., Liu, N., et al. (2010). Correction of the Delta Phe508 Cystic Fibrosis Transmembrane Conductance Regulator Trafficking Defect by the Bioavailable Compound Glafenine. *Mol. Pharmacol.* 77 (6), 922–930. doi:10.1124/mol.109.062679

Robert, R., Carlile, G. W., Pavel, C., Liu, N., Anjos, S. M., Liao, J., et al. (2008). Structural Analog of Sildenafil Identified as a Novel Corrector of the F508del-CFTR Trafficking Defect. *Mol. Pharmacol.* 73 (2), 478–489. doi:10.1124/mol.107.040725

Sampson, H. M., Robert, R., Liao, J., Matthes, E., Carlile, G. W., Hanrahan, J. W., et al. (2011). Identification of a NBD1-Binding Pharmacological Chaperone that Corrects the Trafficking Defect of F508del-CFTR. *Chem. Biol.* 18, 231–242. doi:10.1016/j.chembiol.2010.11.016

Singh, S., Chaudhary, K., Dhanda, S. K., Bhalla, S., Usmani, S. S., Gautam, A., et al. (2016). SATPdb: a Database of Structurally Annotated Therapeutic Peptides. *Nucleic Acids Res.* 44 (Database issue), D1119–D1126. doi:10.1093/nar/gkv1114

Song, Y., Sonawane, N. D., Salinas, D., Qian, L., Pedemonte, N., Galietta, L. J. V., et al. (2004). Evidence Against the Rescue of Defective ΔF508-CFTR Cellular Processing by Curcumin in Cell Culture and Mouse Models. *J. Biol. Chem.* 279, 40629–40633. doi:10.1074/jbc.M407308200

Southan, C. (2013). InChI in the Wild: An Assessment of InChIKey Searching in Google. *J. Cheminform* 5, 10. doi:10.1186/1758-2946-5-10

Southan, C., Sitzmann, M., and Muresan, S. (2013). Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol. Inform.* 32 (11–12), 881–897. doi:10.1002/minf.201300103

Southern, K. W., Patel, S., Sinha, I. P., and Nevitt, S. J. (2018). Correctors (Specific Therapies for Class II CFTR Mutations) for Cystic Fibrosis. *Cochrane Database Syst. Rev.* Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6513216/.

SpringerMedizin (2021). Evidence for Decline in the Incidence of Cystic Fibrosis: A 35-Year Observational Study in Brittany, France. Available at: https://www.springermedizin.de/evidence-for-decline-in-the-incidence-of-cystic-fibrosis-a-35-ye/9642196 (Accessed March 31, 2021).

Stanke, F., Becker, T., Kumar, V., Hedtfeld, S., Becker, C., Cuppens, H., et al. (2011). Genes that Determine Immunology and Inflammation Modify the Basic Defect of Impaired Ion Conductance in Cystic Fibrosis Epithelia. *J. Med. Genet.* 48 (1), 24–31. doi:10.1136/jmg.2010.080937

Stevers, L. M., Lam, C. V., Leysen, S. F., Meijer, F. A., van Scheppingen, D. S., de Vries, R. M., et al. (2016). Characterization and Small-Molecule Stabilization of the Multisite Tandem Binding between 14-3-3 and the R Domain of CFTR. *Proc. Natl. Acad. Sci. U S A.* 113 (9), E1152–E1161. doi:10.1073/pnas.1516631113

Van Goor, F., Straley, K. S., Cao, D., González, J., Hadida, S., Hazlewood, A., et al. (2006). Rescue of DeltaF508-CFTR Trafficking and Gating in Human Cystic Fibrosis Airway Primary Cultures by Small Molecules. *Am. J. Physiol. Lung Cel Mol Physiol* 290 (6), L1117–L1130. doi:10.1152/ajplung.00169.2005

Veit, G., Avramescu, R. G., Chiang, A. N., Houck, S. A., Cai, Z., Peters, K. W., et al. (2016). From CFTR Biology toward Combinatorial Pharmacotherapy: Expanded Classification of Cystic Fibrosis Mutations. *Mol. Biol. Cel* 27 (3), 424–433. doi:10.1091/mbc.E14-04-0935

Veit, G., Xu, H., Dreano, E., Avramescu, R. G., Bagdany, M., Beitel, L. K., et al. (2018). Structure-guided Combination Therapy to Potently Improve the Function of Mutant CFTRs. *Nat. Med.* 24 (11), 1732–1742. doi:10.1038/s41591-018-0200-x

Weininger, D. (1988). SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* 28 (1), 31–36. doi:10.1021/ci00057a005

Wiedemar, N., Hauser, D. A., and Mäser, P. (2020). 100 Years of Suramin. Antimicrob Agents Chemother. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7038244/.

Wilschanski, M., Yaakov, Y., Omari, I., Zaman, M., Martin, C. R., Cohen-Cymberknoh, M., et al. (2016). Comparison of Nasal Potential Difference and Intestinal Current Measurements as Surrogate Markers for CFTR Function. *J. Pediatr. Gastroenterol. Nutr.* 63 (5), e92. doi:10.1097/MPG.0000000000001366

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46 (Database issue), D1074–D1082. doi:10.1093/nar/gkx1037

Xu, L. N., Na, W. L., Liu, X., Hou, S. G., Lin, S., Yang, H., et al. (2008). Identification of Natural Coumarin Compounds that rescue Defective DeltaF508-CFTR Chloride Channel Gating. *Clin. Exp. Pharmacol. Physiol.* 35 (8), 878–883. doi:10.1111/j.1440-1681.2008.04943.x

Yang, H., Shelat, A. A., Guy, R. K., Gopinath, V. S., Ma, T., Du, K., et al. (2003). Nanomolar Affinity Small Molecule Correctors of Defective Delta F508-CFTR Chloride Channel Gating. *J. Biol. Chem.* 278 (37), 35079–35085. doi:10.1074/jbc.M303098200

Zhang, S., Blount, A. C., McNicholas, C. M., Skinner, D. F., Chestnut, M., Kappes, J. C., et al. (2013). Resveratrol Enhances Airway Surface Liquid Depth in Sinonasal Epithelium by Increasing Cystic Fibrosis Transmembrane Conductance Regulator Open Probability. *PLoS One.* 8, e81589. doi:10.1371/journal.pone.0081589

Zhang, Y., Yu, B., Sui, Y., Gao, X., Yang, H., and Ma, T. (2014). Identification of Resveratrol Oligomers as Inhibitors of Cystic Fibrosis Transmembrane Conductance Regulator by High-Throughput Screening of Natural Products from Chinese Medicinal Plants. *PLoS One.* 9, e94302. doi:10.1371/journal.pone.0094302

Zotero (2021). Your Personal Research Assistant. Available at: https://www.zotero.org/.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Chapter 3 CFTR Lifecycle Map – A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations

This manuscript has originally been published in the International Journal of Molecular Sciences

CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations

Liza Vinhoven[1], Frauke Stanke[2,3], Sylvia Hafkemeyer[4] and Manuel Manfred Nietert[1,5]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2]German Center for Lung Research (DZL), Partner Site BREATH, Hannover, Germany
[3]Clinic for Pediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Hannover, Germany
[4]Mukoviszidose Institut gGmbH, Bonn, Germany
[5]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany

Authors contribution

L.V.: methodology; data curation; investigation; formal analysis; visualization; writing – original draft preparation,

F.S.: conceptualization; writing – review and editing; funding acquisition

S.H.: writing – review and editing

M.M.N.: conceptualization; writing – review and editing; funding acquisition

International Journal of *Molecular Sciences*

MDPI

Article

CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations

Liza Vinhoven [1], Frauke Stanke [2,3], Sylvia Hafkemeyer [4] and Manuel Manfred Nietert [1,5,*]

[1] Department of Medical Bioinformatics, University Medical Center Göttingen, Goldschmidtstraße 1, 37077 Göttingen, Germany; liza.vinhoven@med.uni-goettingen.de
[2] Clinic for Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany; mekus.frauke@mh-hannover.de
[3] Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), The German Center for Lung Research, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
[4] Mukoviszidose Institut gGmbH, In den Dauen 6, 53117 Bonn, Germany; shafkemeyer@muko.info
[5] CIDAS Campus Institute Data Science, Goldschmidtstraße 1, 37077 Göttingen, Germany
* Correspondence: manuel.nietert@med.uni-goettingen.de; Tel.: +49-551-39-14920

Abstract: Different causative therapeutics for CF patients have been developed. There are still no mutation-specific therapeutics for some patients, especially those with rare CFTR mutations. For this purpose, high-throughput screens have been performed which result in various candidate compounds, with mostly unclear modes of action. In order to elucidate the mechanism of action for promising candidate substances and to be able to predict possible synergistic effects of substance combinations, we used a systems biology approach to create a model of the CFTR maturation pathway in cells in a standardized, human- and machine-readable format. It is composed of a core map, manually curated from small-scale experiments in human cells, and a coarse map including interactors identified in large-scale efforts. The manually curated core map includes 170 different molecular entities and 156 reactions from 221 publications. The coarse map encompasses 1384 unique proteins from four publications. The overlap between the two data sources amounts to 46 proteins. The CFTR Lifecycle Map can be used to support the identification of potential targets inside the cell and elucidate the mode of action for candidate substances. It thereby provides a backbone to structure available data as well as a tool to develop hypotheses regarding novel therapeutics.

Keywords: cystic fibrosis; CFTR; CFTR maturation; systems medicine model; trafficking; CFTR modulators

Citation: Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations. *Int. J. Mol. Sci.* **2021**, *22*, 7590. https://doi.org/10.3390/ijms22147590

Academic Editor: Martina Gentzsch

Received: 18 June 2021
Accepted: 13 July 2021
Published: 15 July 2021

1. Introduction

Cystic fibrosis (CF) is an inherited disorder prevalent among the white European population, where, with an incidence of approximately 1 in 3000 newborns, it is one of the most common monogenic autosomal recessive diseases [1]. CF is caused by mutations of the cystic fibrosis transmembrane conductance regulator (*CFTR*) gene [2], which encodes a membrane protein that serves as a chloride and bicarbonate channel in exocrine epithelia of various organs, and thereby regulates the viscosity of the mucus lining [3]. Defective CFTR, therefore, has severe implications throughout the body, its major hallmarks being recurrent pulmonary infections and pancreatic insufficiency [2,3].

CFTR is an approximately 170 kDa membrane-spanning glycoprotein composed of 1440 amino acid residues and complex glycosylation [4–6]. It belongs to the ATP binding cassette (ABC) Transporter Superfamily [7,8], but, unlike most of them, acts as an ion channel as opposed to an active transporter. As an ion channel, CFTR only requires ATP for opening, whereas other ABC transporters also require ATP for the active transport of substrates across membranes [9]. In accordance with the other ABC transporters, CFTR consists of two transmembrane domains (TM1 and TM2) and two cytosolic nucleotide-binding

domains (NBD1 and NBD2) with an additional highly flexible regulatory region (R-region) at its center [4,9–11]. Several in-depth reviews cover the structure and opening mechanism of CFTR in great detail [9,12–14]. CFTR undergoes an intricate maturation pathway with complex folding and core glycosylation at the endoplasmic reticulum (ER), before being trafficked through the secretory pathway and further glycosylated at the Golgi apparatus. At the ER, CFTR is subject to extensive quality control mechanisms, often resulting in premature degradation through the ER-associated degradation pathway (ERAD), which leads to only 20–40% of nascent peptides of wild-type (wt) CFTR being correctly folded and trafficked to the apical plasma membrane (PM). After being integrated into the membrane, CFTR undergoes continuous endocytosis, recycling and, when misfolded, degradation in the lysosome [15].

To date, more than 2100 mutations of the *CFTR* gene are known, several hundred of which are known to be disease-causing [16–18]. They can cause defects anywhere in CFTR's complicated and sensitive lifecycle, which is why they were traditionally subdivided into six different classes by the effect they have on the CFTR protein: (I) no protein synthesis, (II) CFTR trafficking defect, (III) dysregulation of CFTR, (IV) defective chloride conductance or channel gating, (V) reduced CFTR transcription and synthesis and (VI) less stable CFTR [19–21]. The traditional classification system has been proposed by Welsh and Smith in 1993 [19] and since has been reviewed and adapted in a range of publications [20,22,23]. However, since many mutations exhibit more complex phenotypes, nowadays, a modified and expanded classification system finds use, where all combinations of the six original mutation classes are regarded as possible classes [21]. For example, the most prevalent CF-causing mutation is the deletion of phenylalanine at position 508 (F508del), which accounts for almost 70% of CF chromosomes worldwide [1,24]. F508del causes a folding defect which results in premature degradation at the ER, making it a class II mutation [6,25]. However, even when rescued and trafficked to the membrane, its channel gating is reduced and it is less stable than wt-CFTR, meaning it exhibits mutation class III/IV and VI characteristics as well [26,27]. According to the expanded classification system, it is therefore classified as a class II-III-VI mutation [21]. Several other mutations also display different defects. Therefore, the wide range of *CFTR* gene mutations, resulting in different defects in the CFTR protein, makes causative treatments for CF difficult to find, and the recently available CF modulators do not target all mutations. As a result, there is still no mutation-specific therapeutic for some patients, especially those with rare *CFTR* mutations. The latest research efforts, therefore, focused on developing combination therapies to target multiple defects at once [28,29]. For this purpose, high-throughput (HT) screens have been performed [30–42], where thousands of substances have been tested in different cell models [43]. These result in a plethora of data and various candidate compounds, often with an unclear mode of action. In order to provide an overview of already tested compounds, we previously established the publicly available database CandActCFTR (https://candactcftr.ams.med.uni-goettingen.de/ (accessed on 6 June 2021)), where substances from 90 publications are listed and categorized according to their influence on CFTR function (manuscript submitted for publication).

In order to support the elucidation of the mechanism of action for promising candidate substances and to be able to predict possible synergistic effects of substance combinations, we used a systems biology approach to create a model of the CFTR maturation pathway in cells. Systems biology modeling aims to gather knowledge on biological systems and translate it into a human- and machine-understandable format in order to analyze its behavior and interactions. To make models reproducible and reusable, there are certain standards and formats to adhere to. The most well-established, standardized format in the graphical representation of biological processes is the Systems Biology Graphical Notation (SBGN), consisting of three languages [44]. Of the three, the SBGN Process Description (PD) language allows the most detailed representation of molecular mechanisms using nodes and directed edges. Molecular entities are shown as nodes and have different shapes depending on their molecular species. For example, proteins are represented as rounded

rectangles, RNAs as parallelograms and small molecules as ovals. The reactions (e.g., transcription, translation and state transition) and reaction regulations (e.g., catalysis and inhibition) between the molecular entities are represented by edges, shaped as arrows with differently shaped heads depending on the type of interaction [44]. A glossary of systems biology and bioinformatics terms used here can be found in the Supplementary Materials.

Here, we present a systems biological model of the CFTR lifecycle in a standardized, explorable and tractable format. The model is composed of two datasets, a core map manually curated from small-scale experiments in human cells, and a coarse map including interactors identified in high-throughput (HT) efforts [45–48]. Interactors are here defined as molecular entities that influence CFTR directly or indirectly. Both data layers are divided into submaps focusing on different stages of the CFTR lifecycle and different processes the ion channel is involved in. Overall, the manually curated core map includes 170 different molecular entities and 156 reactions from 221 experimental publications. The high-throughput data layer encompasses 1384 unique interactors from four publications by Wang et al., 2006, Pankow et al., 2015, Santos et al., 2019 and Matos et al., 2019 [45–48]. The CFTR Lifecycle map provides a tool to structure and exploit existing knowledge and data, as well as develop a hypothesis regarding synergistic drug targets and novel therapeutics.

2. Results

2.1. CFTR Map

The CFTR map encompasses information from small-scale experiments as well as high-throughput efforts, leading to differences in the degree of detail and confidence. It was therefore split into different data layers. The first data layer, the core map, was manually curated and only includes high-confidence interactors, confirmed by at least two independent small-scale experiments or acknowledged by two reviews from different research groups. As a result, the number of molecular interactions in the core map is limited but each is described with a high level of detail. The second data layer in the coarse map represents the high-throughput interactome of wt-CFTR (the interactome of F508del was excluded) as published by Wang et al., 2006, Pankow et al., 2015, Santos et al., 2019 and Matos et al., 2019 [45–48] in a structured cell layout. Here, the large-scale experimental method does not allow for conclusions regarding the nature of the interactions. Therefore, a small level of detail, but a high number of interactors, is included in the map. Overall, the manually curated core map comprises 170 different interactors and the coarse map from large-scale efforts contains 1384 interactors; 46 interactors could be found in both (Figure 1). A list of all interactors in both maps, as well as their overlap, can be found in Supplementary Table S1. To prevent redundancies, the overlapping interactors were only kept in the manually curated core map and excluded from the coarse map.

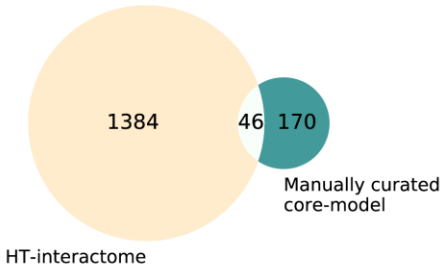


Figure 1. Venn diagram of the interactors in the manually curated core map vs. coarse map. The manually curated list of interactors within the core map comprises 170 interactors, the coarse map derived from high-throughput data contains 1384 interactors, and 46 interactors occur in both lists. The overlap was subtracted from the high-throughput interactome for the visualization to avoid redundancies, resulting in 1338 interactors in the coarse map.

2.2. Representation of the CFTR Lifecycle in the CFTR Core Map

The CFTR core map represents the molecular mechanisms affecting wt-CFTR during its lifecycle. It is the product of an exhaustive literature curation process and the manual integration of different data sources. As the whole model is represented and written in the standardized SBGN format, it is human understandable as well as computationally tractable. It was created in the editor CellDesigner4.4.2 [49,50], adhering to the Process Description language of the SBGN format [44]. At the moment, it encompasses 262 different molecular entities and 156 reactions in 6 main cellular compartments. The biomolecules are categorized into 149 proteins, 58 complexes, 28 simple molecules, 13 ions, 6 genes, 5 RNAs, and 3 pools of degraded protein, amino acids, or nucleotides. Proteins can be subdivided into generic proteins, truncated proteins, ion channels, and receptor proteins. The color of the molecular entity indicates whether it was identified in at least one polarized cell line (green) or non-polarized cell lines only (yellow). Reactions are specified as state transitions, inhibitions, catalysis, transports, and heterodimer associations and dissociations. Each interaction is supported by at least two independent publications, leading to an overall number of 221 publications. A complete list of all interactors and references can be found in Supplementary Tables S2–S6.

The CFTR core map (Figure 2) has a roughly cell-shaped layout, with CFTR making its way from the nucleus at the bottom all the way up to its site of action at the plasma membrane. It covers the molecular interactions CFTR undergoes on its way from being transcribed in the nucleus to being a functional ion channel at the apical plasma membrane, including its activity and regulation there, as well as endocytosis, recycling and degradation of the mature protein. The map can be subdivided into five submaps (Table 1) to enable the user to either look at its whole or individual processes, depending on their focus or interest. The five submaps are guided by the subcellular location and process they focus on.

1. Transcription—Nucleus: The Nucleus submap covers the transcriptional regulation of the *CFTR* gene into its mRNA.
2. Translation, Folding and ER Quality Control—ER: The ER submap step-by-step describes the translation of the mRNA into the CFTR peptide and its integration into the membrane as well as folding steps modulated by chaperones, core glycosylation, and the calnexin cycle involved in ER quality control. Depending on the folding success, CFTR may progress through the secretory pathway or be degraded through ER-associated degradation.
3. Secretory Pathway—ER, Golgi Apparatus, Plasma Membrane: The Secretory Pathway submap covers COPII vesicle-mediated trafficking between the ER, Golgi, and the plasma membrane and the full glycosylation at the Golgi apparatus. It also describes unconventional trafficking of the protein between the ER and plasma membrane, which has been found to be an alternative route CFTR may take.
4. Activity and Regulation—Plasma Membrane: The Activity submap covers the phosphorylation-dependent activation of CFTR through the cAMP signaling cascade, channel opening, closing, and ion conductance as well as regulatory interactions with other ion channels and stabilization through interactions with the cytoskeleton.
5. Endocytosis, Recycling and Degradation—Plasma Membrane, Endosomes, Lysosomes (Figure 2c): The final submap describes the endocytosis of the mature CFTR protein from the plasma membrane, which can be recycled back to the membrane or degraded in lysosomes.

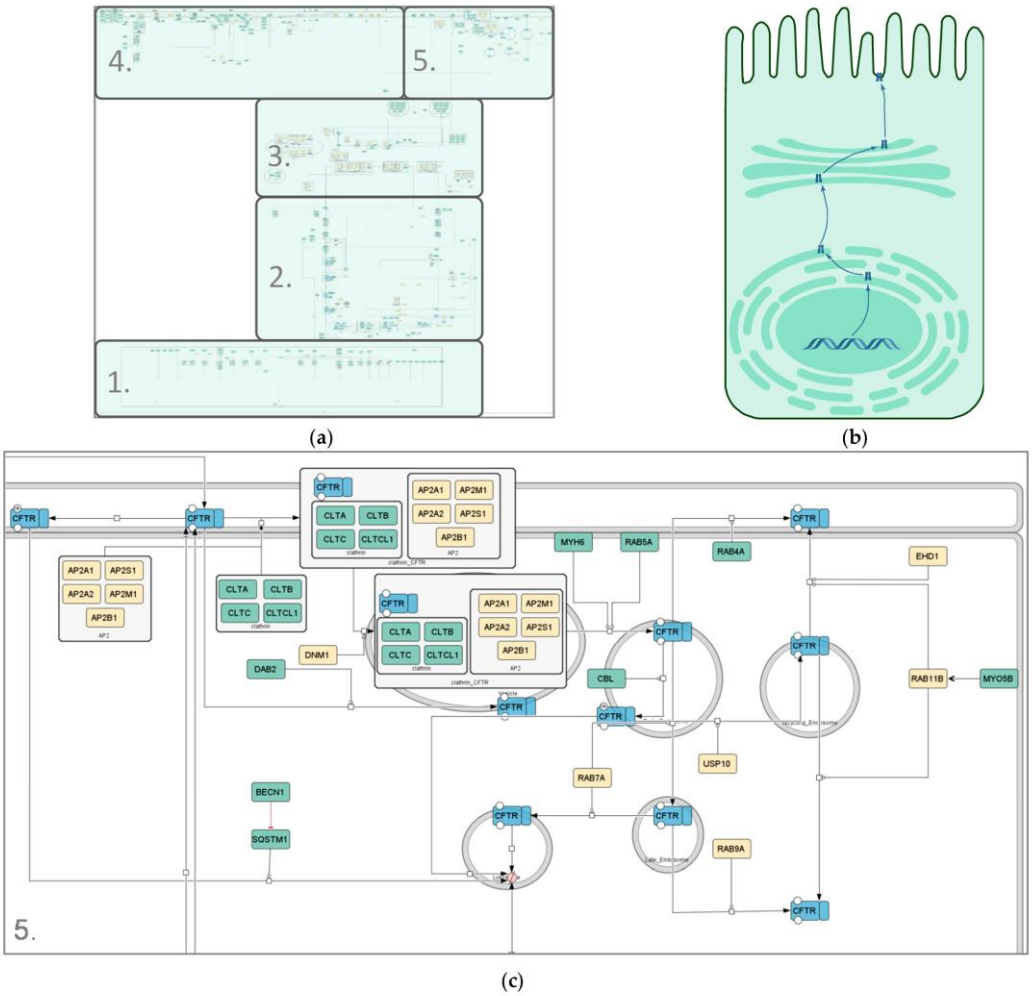


Figure 2. Different representations of the CFTR lifecycle. (**a**) Image of the SBGN-compliant manually curated CFTR lifecycle core map; (**b**) cartoon representation of the CFTR lifecycle in an apical epithelial cell; (**c**) zoomed-in section of the endocytosis pathway in the manually curated CFTR lifecycle core map (submap 5). Compartments are depicted in grey, CFTR in blue and interactors in differing shades of green and yellow. The green color scheme represents interactors identified in at least one polarized cell line; yellow interactors were identified in non-polarized cell lines only. State transitions, catalysis and positive influences are shown in black; negative influences and inhibitions are displayed in red. Different shapes represent different kinds of interactors. Rounded rectangles correspond to proteins, ovals and circles to small molecules and ions, respectively, rectangles correspond to genes, rhomboids to RNA molecules and chevron shapes to receptors. The map was created using CellDesigner4.4.2.

Table 1. Description of the five submaps of the CFTR core map.

Process	Localization	Molecular Entities Present in the Model	N [1]	Proportion of the Interactors Identified in Polarized Cells
Transcription	Nucleus	Proteins RNAs and gene elements Small molecules and ions	28 16 1	97%
Translation, Folding and ER quality control	ER	Proteins Small molecules and ions	45 13	69%
Secretory pathway	ER, Golgi apparatus, Plasma Membrane	Proteins Small molecules and ions	27 8	52%
Activity and Regulation	Plasma Membrane	Proteins Small molecules and ions	44 20	82%
Endocytosis, Recycling and Degradation	Plasma Membrane, Endosomes, Lysosomes	Proteins	23	74%

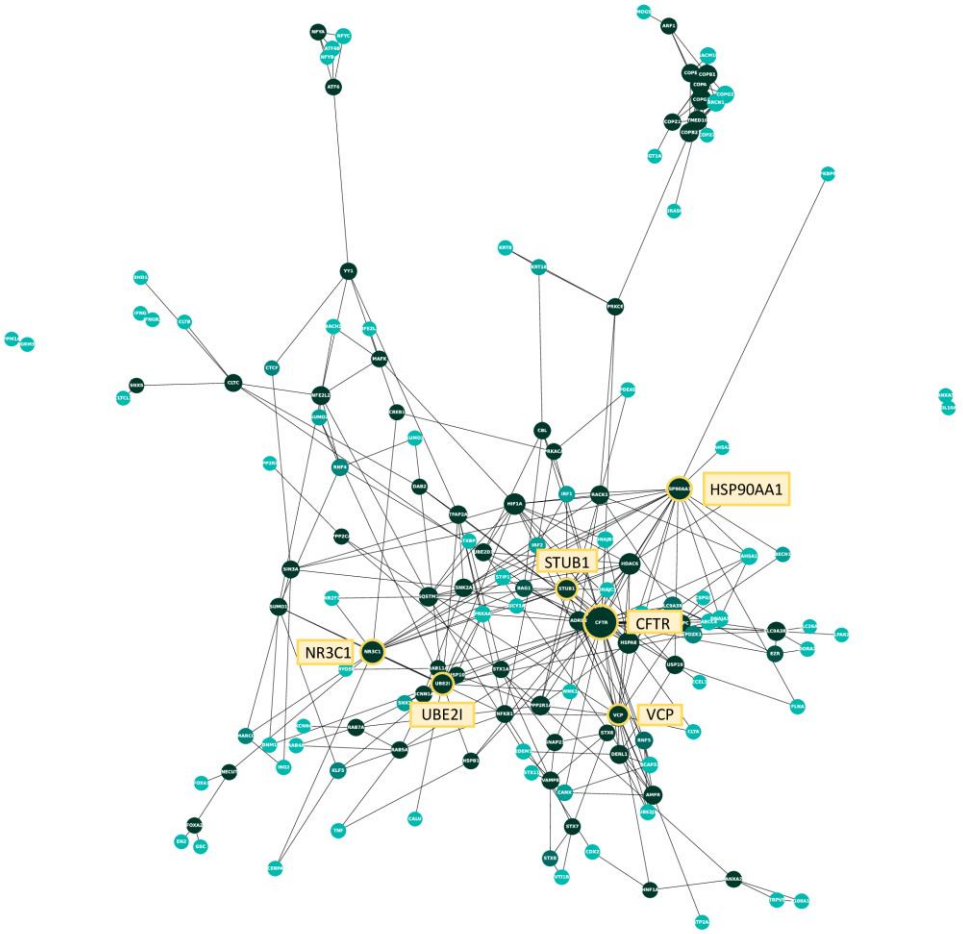
[1] Number of different molecular entities present in the respective submap.

As different cell lines can have different effects on the interactome of proteins [43,51], the cell lines used in the small-scale experiments were identified from each reference and categorized into polarized and non-polarized cells. As the question whether cells are to be characterized as polarized or non-polarized can often be a matter of debate when grown under experimental conditions, polarized cells are here defined as cells with the general ability to polarize. For each interactor, it is indicated by color whether or not they were identified in polarized (green color scheme) or non-polarized cell lines (yellow color scheme), whereby the specific cell lines for each interactor can be found in Supplementary Tables S2–S6. The percentage of interactors identified in at least one polarized cell line was calculated for each submap, amounting to 97% in the Transcription submap, 69% and 52% in the ER and Golgi submap, respectively, 82% in Activity and Regulation and 74% in Endocytosis, Recycling and Degradation.

2.3. Protein–Protein Interaction Network and Topological Analysis of the CFTR Core Map

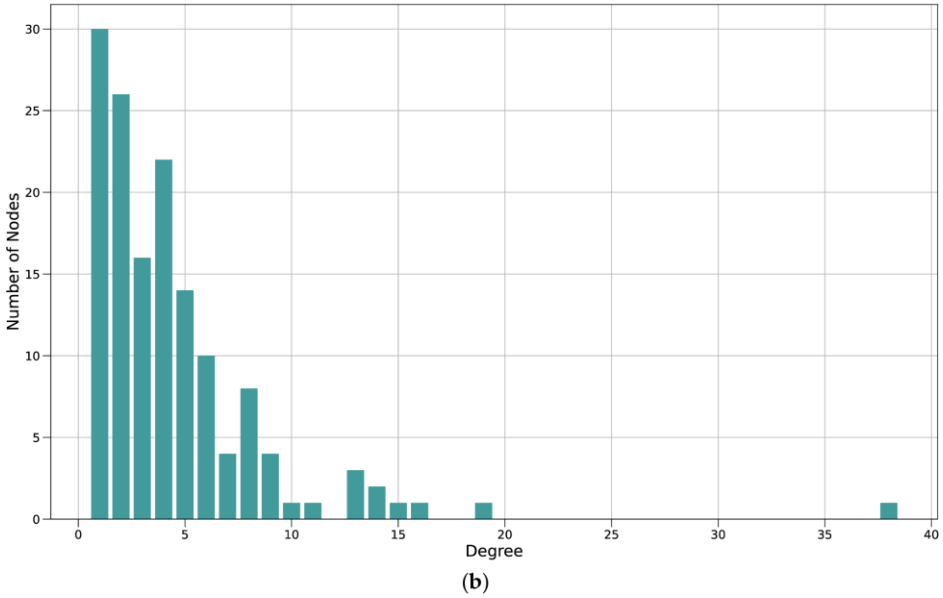
To analyze the manually curated core map with regard to interactions between the proteins included, a protein–protein interaction network was created using the list of genes present in the model (Figure 3a). All proteins (nodes) of the protein–protein interaction network were identified as CFTR interactors through the manual literature curation, whereas all interactions (edges) between them were identified through the BioGrid database. A list of all interactions present in the protein–protein interaction network can be found in Supplementary Table S7. There are 145 nodes and 326 edges present in the network, and the average number of neighbors amounts to approximately 4.5. Another important property to assess when analyzing protein–protein interaction networks is its degree distribution. Here, the degree of a protein (node) is the number of interactions (edges) it shares with another protein. Most biological networks are considered 'scale-free', meaning that the majority of nodes have a low degree with only a few highly interconnected hubs, represented by a large diameter in Figure 3a. In order to analyze whether the CFTR protein–protein interaction network is scale-free, the degree distribution was calculated (Figure 3b). As can be seen, it follows a power law, confirming that it is, indeed, scale-free. Their scale-free character lends biological networks features important in biological systems. On the one hand, they are quite stable, as a random failure is likely to affect a protein with a low degree due to their high prevalence. Therefore, random failures and changes are unlikely to have a large effect on the overall network. On the other hand, targeted interventions at one of the few hubs will have a large effect on the whole network, which can be important when treating diseases and considering side effects. Here, CFTR is, as expected, the node with the highest degree, which can also be seen in the visualization in Figure 3a, where CFTR is the

largest node. Apart from CFTR, the five nodes with the next highest degree are HSP90AA1 (heat shock protein 90), STUB1 (ubiquitin–protein ligase CHIP), UBE2I (SUMO-conjugating enzyme), NR3C1 (Glucorticoid receptor), and VCP (Transitional ER ATPase).



(**a**)

Figure 3. *Cont.*



(**b**)

Figure 3. Protein–protein interaction network and degree distribution of the manual CFTR map. (**a**) Each node represents one protein, each edge between them a physical interaction shown in a small-scale study and reported on BioGrid. The larger a node, the higher its degree (i.e., the more interactions it shares with other proteins). CFTR and the five proteins with the next-highest degree are marked in yellow. The color of the protein represents its betweenness centrality, which is a measure of how important the node is to the flow of information through the CFTR Lifecycle Map. The betweenness centrality of a protein is the number of times it lies on the shortest path between two other proteins. The darker the node, the higher its betweenness centrality; (**b**) bar plot of the degree distribution of the protein–protein interaction network in A. The x-axis shows the degree of a protein; the degree is the number of other proteins a protein interacts with. The y-axis shows the number of proteins in the network with a certain degree. For example, the number of proteins that interact with only one other protein in the network (i.e., have a degree of 1) is 30, the number of proteins that interact with six other proteins is ten. The node with the highest degree (38) is CFTR.

2.4. Visualization of the wt-CFTR Interactome as Coarse Model

In addition to the extensive manually curated interaction pathways, which include all high-confidence interactors and detailed interactions, a second data layer was included for interactors with lower confidence and detail. The second data layer represents information from large-scale experiments, namely the wt-CFTR core interactomes published by Wang et al., 2006, Pankow et al., 2015, Santos et al., 2019 and Matos et al., 2019 [45–48]. As the data are stored in long gene lists, in contrast to the detailed, text-based description of interactions in the manual model, different tools were used to represent the high-throughput interactome. The coarse map is also written in the SBGN format, but was created using libsbgnpy [52] and CellDesigner [49,50] and is represented in the SBGN Activity Flow notation, which lacks mechanistic information. The use of the python library libsbgnpy allowed for an automated construction of the maps. In order to structure the information into an intuitive, cell-based layout, the interactors were grouped according to their function and subcellular localization. Again, the model was divided into several submaps, based on the functional categorization (Table 2). The functional category of each interactor was solely based on information from the respective publications and not inferred from other sources. Interactors for which no function was indicated were assigned to the

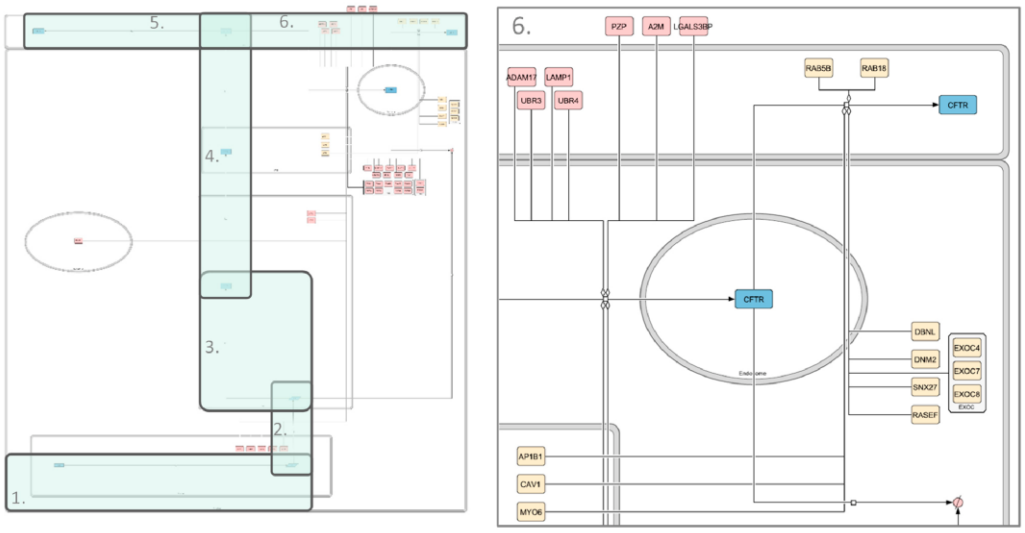
'other/unknown' category. Figure 4 shows an exemplary image of one of the submaps, which is focused on Endocytosis, Recycling and Degradation. Each submap focuses on one main step or area of function in the CFTR-lifecycle, abstracted into a state-transition reaction. Six of them correspond to those from the core map, with an extra map for mRNA processing between the "Transcription" and "Translation, Folding and ER Quality control". Additional maps focus on interactions with the cytoskeleton, as well as immune-related and other interactors.

1. Transcription—Nucleus: The Transcription submap focuses on the *CFTR* gene and the production of pre-mRNA. All interactors are divided into two functional categories, those that affect the gene directly, e.g., "DNA repair" and "replication", and those that affect the transcription, such as "transcription" and chromatin structure". Apart from the CFTR entities, it includes 17 nodes, seven affecting the gene and ten affecting the state transition.
2. RNA processing: The additional RNA processing map describes the conversion of pre-mRNA to mature mRNA. It includes interactors with functional categorizations, such as nuclear export and RNA splicing, but also RNA degradation, and contains 36 nodes apart from CFTR.
3. Translation, Folding and ER Quality Control—ER: The third submap summarizes the processes taking place in the ER in two state transitions. One is the processing from mature mRNA to folded, core-glycosylated CFTR peptide and degradation at any stage during ER quality control, resulting in an overall number of 45 interactors. The interactors are color-coded depending on whether they affect folding (57 interactors, green), degradation (one interactor, red), both (three interactors, red), or the interaction is unspecified (653 interactors, yellow). The 653 unspecified interactors are mainly from the data published by Santos et al. [47], where the authors characterize the interactome of CFTR prior to its exit from the ER.
4. Secretory Pathway: In accordance with the core map, the Secretory Pathway submap shows the trafficking of the CFTR peptide between the ER, Golgi and PM after folding and core glycosylation. For reasons of simplicity and a lack of information, all 22 interactors were depicted as influencing CFTR trafficking between the ER and Golgi, even though they might be affecting different steps.
5. Activity: Here, all reactions involved in the activity and regulation of and by mature CFTR within the plasma membrane PM are summarized as channel opening, influenced by 38 different entities. It also includes 145 unspecified interactors, that were reported to interact with CFTR at the PM [48], but for which the nature of the interaction is unclear.
6. Recycling and Degradation (Figure 4b): This submap is split into the recycling and degradation of mature CFTR. Endocytosis-regulating interactors are included in the recycling category, resulting in 12 interactors affecting recycling and 32 influencing degradation.
7. Cytoskeleton: An additional submap is designated for interactors with an influence on the anchoring of CFTR in the cell, including 62 entities apart from CFTR.
8. Immunity: A separate submap shows nine interactors playing a role in immunity (10 interactors).
9. Other Functions: In order to represent the whole datasets published, another submap includes all interactors that fall into none of the categories above. These include, for example, proteins involved in metabolism and those for which no function regarding CFTR could be specified (250 interactors).

Table 2. Description of the five submaps of the CFTR high-throughput model.

Process	Localization	Functional Category	N [1]
Transcription	Nucleus	DNA Replication Transcription	7 10
RNA Processing	Nucleus–Cytoplasm		36
Translation, Folding and ER quality control	ER	Folding ER-associated degradation both unspecified	57 1 3 653
Secretory pathway	ER, Golgi apparatus, Plasma Membrane		22
Activity and Regulation	Plasma Membrane	Activity unspecified	38 145
Endocytosis, Recycling and Degradation	Plasma Membrane, Endosomes, Lysosomes	Recycling Degradation	12 32
Cytoskeleton	Cytoplasm, Plasma Membrane		62
Immunity			10
Other/Unknown			250

[1] Number of interactors present in the respective submap.



(**a**) (**b**)

Figure 4. Image of the high-throughput CFTR Endocytosis, Recycling and Degradation map. (**a**) Whole map; (**b**) zoomed-in section of the map. Compartments are depicted in grey, CFTR in blue and interactors in shades of red (degradation associated) and yellow (recycling associated). State transitions and modulations are shown in black. Different shapes represent different kinds of interactors. Rounded rectangles correspond to proteins, rectangles correspond to genes and rhomboids to RNA molecules. The map was created using libsbgnpy 0.2.2 and CellDesigner4.4.2.

2.5. Systemic Interpretation and Comparison of Manually Curated Model and Large-Scale Interactome

The lists of interactors from the core map and coarse map were compared for overlaps. Interestingly, of the 170 manually curated interactors and the 1384 interactors from the high-throughput screens, only 46 interactors could be found in both datasets (Figure 1).We next wanted to see whether the interactors in both datasets belong to similar functional categories, as this would indicate that similar pathways of importance for CFTR folding and maturation have been identified by the targeted analysis of interactors (core map) or by hypothesis-free high-throughput analysis (coarse map).

Datasets of the core and the coarse map were analyzed using the BioInfoMiner web application [53], which performs a biological interpretation based on a list of genes, resulting in prioritized lists of systemic processes and genes, similar to a gene enrichment analysis. Here, the gene ontology (GO) [54,55] terms and Reactome Pathway Database [56] terms were assigned to all interactors. Reactome is a manually curated and peer-reviewed database for cellular pathways on a molecular level, whereas the gene ontology knowledgebase provides a model of biological systems to represent the current knowledge on the function of genes from the molecular to organism level. Therefore, while Reactome and Gene Ontology share a certain overlap, they mainly complement each other, as Reactome associates genes with specific molecular processes, whereas Gene Ontology also takes broader biological processes into account.

Of the top 20 prioritized Gene Ontology terms of the core model-derived gene list and of the coarse model-derived high-throughput gene list, 12 were shared between core and coarse model, including *cellular localization*, *macromolecule localization*, *protein localization* and *vesicle-mediated transport* (Figure 5). Furthermore, the heat shock protein *HSP90AA1*, as well as the ER ATPase *VCP* were found in the top 45 prioritized genes of both gene sets.

In addition to the analysis using Gene Ontology terms, the BioInfoMiner analysis was repeated with terms from the Reactome Pathway Database [56]. Here, however, there was no overlap between Reactome terms assigned to the two gene lists derived from the core model and the coarse model, respectively. While the top-ranked terms for the manually curated gene list derived from Reactome were mostly intracellular transport-related, the top-ranked terms for the high-throughput gene list mostly related to the cellular response to stress and infection. The complete results of the analysis can be found in Supplementary Tables S8–S12.

The diverging results between the Gene Ontology-based analysis and Reactome-based analysis most likely stem from the different levels of biological activity that are defined by these two databases. While the analyzed Gene Ontology terms mainly include the broad biological process terms, such as *intracellular transport*, the Reactome pathway terms are on a smaller, more detailed level, such as *SRP-dependent cotranslational protein targeting to membrane*. Consequently, the same major biological processes appear to be overrepresented in the manually curated core model and the high-throughput interactome represented in the coarse model, while different specific molecular pathways from Reactome are recognized in both layers of the CFTR Lifecycle Map.

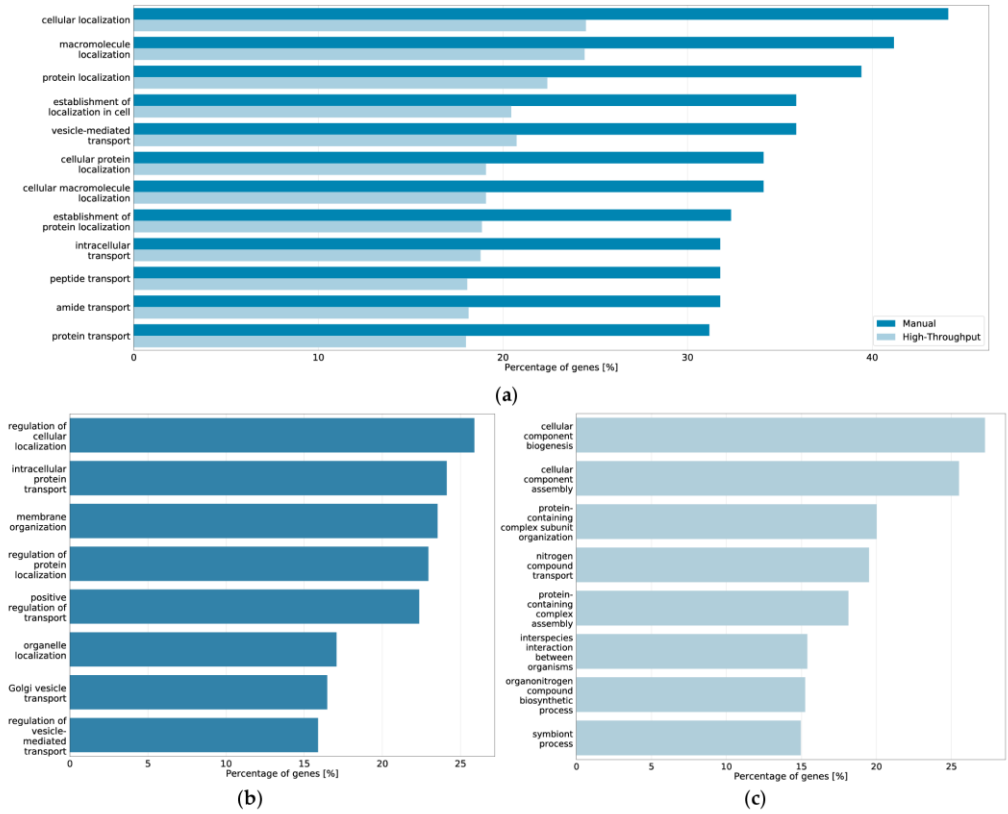


Figure 5. Top 20 prioritized Gene Ontology processes from BioInfoMiner analysis of the CFTR Lifecycle Map. (**a**) Bar plot of the processes prioritized among the top 20 in both the core map and the coarse map datasets and percentage of genes from the respective map associated with the processes; (**b**) bar plot of the processes prioritized among the top 20 in only the manually curated gene list of the core map and percentage of genes from the core map associated with the processes; (**c**) bar plot of the processes prioritized among the top 20 in only the high-throughput interactome of the coarse map and percentage of genes from the coarse map associated with the processes.

3. Discussion

Extensive research has been dedicated towards the elucidation of the processes CFTR undergoes from its transcription on the way to becoming a folded, complex glycosylated and fully functional ion channel at the plasma membrane of apical epithelial cells. In the CFTR Lifecycle Map, we collected the knowledge accumulated by researchers over the course of three decades and represented it in a human-and machine-readable way. In doing so, we adhered to the standards established by the systems biology community. We applied common literature curation criteria and followed the well-established SBGN format, as well as MIRIAM guidelines to create and annotate a core map of the CFTR lifecycle. Additionally, we included data from large-scale efforts to identify the CFTR interactome in a coarse map as a second data layer.

When comparing the manually curated data from small-scale experiments within the core model with the data from high-throughput screens by Wang et al., 2006, Pankow et al., 2015, Santos et al., 2019 and Matos et al., 2019 [45–48] embedded into the coarse

model, it becomes evident that the overlap between them is surprisingly small. To detect whether genes from the core and the coarse model belong to shared functional categories, BioInfoMiner was also used to prioritize genes in the core model and in the high-throughput generated interactome of the coarse model. BioInfoMiner is a tool used for the analysis of semantic networks and prioritizes key systemic processes and related genes present in a set of genes based on different databases [53]. Here, functionalities were assigned to the CFTR interaction partners based on Gene Ontology biological process terms [54,55] and the terms from the Reactome Pathway Database [56]. Gene Ontology provides information on the function of genes and gene products, which is derived from the scientific literature by the Gene Ontology Consortium. The Reactome Pathway Database is a manually curated and peer-reviewed database for molecular pathways by an international multidisciplinary team. The overlap between core model and coarse model genes detected by the Gene Ontology categories indicate that similar pathways of importance for CFTR folding and maturation have been identified by the targeted analysis of interactors (core map) or by hypothesis-free high-throughput analysis (coarse map). Using the Gene Ontology term analysis, two genes were prioritized among the top ten for both datasets. Both of them, HSP90AA1, better known as heat shock protein 90, and the ER ATPase VCP were also found amongst the hubs in the protein–protein interaction network of the manually curated interactors. Both VCP and HSP90AA1 are involved in a multitude of different cellular processes. It is therefore not surprising that they are highly connected in the protein–protein interaction network and are ranked as important genes in the Gene Ontology term analysis. They most likely also play an important role in CFTR maturation; however, they are very unspecific and likely influence the folding and maturation of proteins other than CFTR as well.

In contrast to the overlap in Gene Ontology categories, where entries of the core and the cores model were shared, the pathway terms from the Reactome Pathway Database differ substantially between the manually curated core model and the high-throughput-derived coarse model gene lists, which reinforces the poor overlap of only 46 interactors present in both layers of the CFTR Lifecycle Map. These differences on the small-scale level may stem from the different experimental approaches used to identify the interactors. While the manual curation from small-scale experiments includes a lot of information on the immature CFTR, the high-throughput screens use large-scale co-immunoprecipitation-based approaches to identify the interactome of the mature, or at least folded, wt-CFTR. These results, however, also indicate that, although a lot is known about the maturation and processing of CFTR, there are still substantial knowledge gaps. This is especially true for the interaction with other ion channels at the plasma membrane. In the last decade, other ion channels that may be in regulatory interaction with CFTR, such as the calcium-activated chloride channel Anoctamin-1 (also known as ANO1, TMEM16A and ORAOV2), or other chloride channels which are being discussed as CFTR alternatives, have been brought to the center of attention [57–59]. However, there is still work to be carried out in the elucidation of the CFTR regulatory network and its role in ion homeostasis at the plasma membrane, especially considering that potentially interacting ion channels, such as CLCA2 (also known as CACC3) and KCNJ1 (also known as ROMK) had to be rejected during the construction of the CFTR map due to a lack of experimental evidence. Knowledge gaps also seem to exist with regard to the intracellular trafficking pathway. It is known that CFTR is usually trafficked through conventional Golgi-mediated exocytosis, but may also circumvent the Golgi via an unconventional route [60–62]. This may be relevant for the rescue of misfolded variants such as F508del-CFTR [62–65]. Differences in the observed pathways and interactions may also arise from the variety of cell models used, which we addressed by color-coding interactors from polarized and non-polarized cells in our model. However, it becomes clear that the molecular mechanisms and interactors of both pathways are not fully elucidated yet, as can be seen in the lack of detail in the CFTR core map.

The CFTR Lifecycle Map is part of the CandActCFTR project, which established a publicly available database of candidate cystic fibrosis therapeutics, combining data from different sources, such as high-throughput- and small scale screens, data from relevant

databases and unpublished primary data (candactcftr.ams.med.uni-goettingen.de/). The CFTR Lifecycle Map, as a second part of the project, aims to provide the means to identify promising drug targets and elucidate the mode of action for candidate substances. Furthermore, it will be used to predict possible additive effects of different substance combinations. It thereby simultaneously provides a backbone to structure available data as well as a tool to develop hypotheses regarding novel therapeutics.

4. Materials and Methods

4.1. Creation of the Core Map

The map was drawn using the Process Description (PD) language of the Systems Biology Graphical Notation (SBGN) syntax [44] and the diagram editor CellDesigner4.4.2 [49,50]. The PD language of SBGN allows the representation of molecular biological models in a standardized manner using nodes and edges. The nodes, i.e., the molecular entities, of the network are represented by different symbols according to their molecular species. This map includes entities specified as genes, mRNA, proteins, truncated proteins, protein complexes, receptors, ion channels, simple molecules, and ions. The edges between molecular entities serve to specify reactions and reaction regulations. Reaction types present in the map include transcription, translation, state transition, complex formation and dissociation and transport, and reaction regulations are of the type catalysis, inhibition, physical stimulation, or modulation, where the exact nature of the interaction is unknown. Each molecular entity is named according to its HUGO-approved gene symbol [66] (https://www.genenames.org/ (accessed on 26 January 2021)) for genes and gene products, and ChEBI name [67] (www.ebi.ac.uk/chebi/ (accessed on 26 January 2021)) for simple molecules and ions. The model is split into six main cellular compartments (cytoplasm, plasma membrane, extracellular space, nucleus, endoplasmic reticulum (ER), and Golgi apparatus) and smaller compartments (vesicles and endosomes). Following the Minimal Information Required for the Annotation of Models (MIRIAM) Guidelines [68], an established standard for annotating systems biology models, all components are annotated by unique identifiers. The HUGO ID [66] (https://www.genenames.org/ (accessed on 26 June 2021)) is used for genes, and gene products, the UniProt ID [69] (www.uniprot.org (accessed on 26 January 2021)) for proteins, the PubChem CID [70] (pubchem.ncbi.nlm.nih.gov/ (accessed on 26 January 2021)) and ChEBI ID [67] (www.ebi.ac.uk/chebi/ (accessed on 26 January 2021)) for simple molecules. Furthermore, all interactors are annotated by the PMID (pubmed.ncbi.nlm.nih.gov (accessed on 6 June 2021)) of the references they are derived from. We define interactors as molecular entities that influence CFTR either through direct physical interactions or indirect regulatory interactions.

In order to provide additional information, all interactors in the map are color-coded. Blue entities represent CFTR or versions thereof (gene, mRNA, protein). All interactors that could be confirmed in at least one polarized cell line are depicted using a green color scheme, and interactors only identified in non-polarized cell lines are shown in yellow.

4.2. Literature Curation for the Core Map

In order to assess the current state of knowledge for CFTR maturation and activity, a literature overview was created, starting from 23 relevant main reviews of the last two decades, which can be found in Supplementary Table S13. Highly cited examples of these reviews are Riordan, 2008 [71], Lukacs and Verkman, 2012 [25] and Farinha and Canato, 2017 [72], which cover the expression, folding, maturation and function of CFTR in great detail and were published over the course of nearly one decade, thereby providing extensive descriptions of recent as well as earlier findings. From there, a preliminary consensus network of CFTR-relevant pathways and a list of resulting interaction partners was created. To validate the list of interactors, the literature database PubMed [73] was searched for experimental publications that confirmed the interaction through small-scale experiments using methods such as co-immunoprecipitation, pull-down assays, two-hybrid assays, NMR, X-ray crystallography, surface plasmon resonance, and functional

(mutational) assays. A complete list of the 221 references published between 1991 and 2020, which were used for validation, can be found in Supplementary Table S14. An interactor was accepted for the model when it could be identified in at least two small-scale experiments conducted in human cells from independent references or was considered as acknowledged by the research community when described in at least two reviews from different research groups. For example, the interaction of CFTR with the gene product of *SLC9A3R1* (also known as NHRF1 or EBP50) at the plasma membrane is well documented in experimental publications as well as reviews [72,74–76], and it was therefore accepted for the CFTR core map. On the other hand, the serum response factor (SRF) has been reported to act as a transcription factor for CFTR by René et al. [77] but, to our knowledge, has not yet been confirmed by other research groups. Consequently, it was not included in the CFTR core map for now. This does not mean that SRF is not considered a CFTR interactor; it is merely a quality control check to ensure a high data quality and high evidence of the interactions compiled in the core map.

4.3. Integration of Protein–Protein Interaction Databases

To ensure that no high-confidence interactors were missed, the manually curated list of interactors was then complemented with data from small-scale experiments from existing protein–protein interaction databases (SIGNOR2.0 [78], BioGRID [79], the Human Protein Reference Database [80], String DB [81], MINT [82], InnateDB [83], APID [84] and IntAct [85]). The same literature criteria were applied. References to large-scale experiments were excluded from the database searches for the core map and later added as a second data layer in the coarse map (see Section 4.5).

4.4. Consideration of Cell Polarity

The proper maturation and integration of functional CFTR into the plasma membrane is highly dependent on the polarization of the cell [51]. In order to take this into account in our model, the experimental method was specified and the cell lines used were listed for each interaction. Cell lines were divided into non-polarized and polarized cells and it was specified for each interaction whether or not it was shown in at least one polarized cell line. Under experimental conditions, the question whether cells are to be characterized as polarized or non-polarized can often be a matter of debate, which is beyond the scope of this study. Therefore, polarized cells are here defined as cells with the general ability to polarize. A complete list of interactions and the respective publications with the experimental method and cell line used can be found in Supplementary Tables S2–S6. During the literature curation, interactors were successively added to the list and model, which was visualized using the diagram editor CellDesigner4.4.2 [49,50] and the cell-polarity was indicated by color. Therefore, our model rudimentarily takes into account that a range of different model systems was used to study the CFTR lifecycle.

4.5. Visualization of the High-Throughput Interactome as Coarse Map

In the last 15 years, many potential interactors of CFTR have been identified through high-throughput methods [45–48]. While these methods are able to generate high amounts of data, they provide less information on the nature of the interaction and the confidence of the interaction is lower than through the identification in several small-scale studies, as more false-positives may occur. In order to address this discrepancy in data quality and still provide an extensive interaction map, the wt-CFTR interactomes published by Wang et al., 2006, Pankow et al., 2015, Santos et al., 2019 and Matos et al., 2019 [45–48] were added as separate data layer from large-scale experiments. Only the wt-specific interactome was considered, and interactors unique to the F508del-variant were excluded. Missing information on the subcellular localization of the interactors reported was gathered from UniProt and the Human Protein Atlas, using the respective application programming interface, to update to the knowledge on subcellular localization in 2021. The subcellular localizations specified in the published data were combined into more general main

localizations (nucleus, endoplasmic reticulum (ER), Golgi apparatus, endosome, plasma membrane, mitochondrion, cytoplasm, and extracellular space). Based on the functional categorization specified in the published data, the interactors were grouped into general functional categories (DNA replication, transcription, RNA processing, translation and folding, ER-associated degradation (ERAD), trafficking pathway, cytoskeleton and stabilization, activity and regulation, recycling, degradation, immune response, and other). The functional categorization was solely based on the information provided in the publications and not inferred from other sources. Interactors already present in the core map (46 interactors) were excluded in the coarse map to avoid redundancies between the data layers. A complete list of the interactors from the large-scale experiments, together with their subcellular localization and functional categorization, can be found in Supplementary Table S15. The list of interactors was split according to their functional categorizations and visualized in individual maps using the SBGN Activity Flow notation through the Python library libsbgnpy 0.2.2 [52] and CellDesigner4.4.2 [49,50].

Through this procedure, the findings from high-throughput efforts are available in the model, while the different data-quality, compared to the manually validated interaction partners, becomes visible for the user. While the high-throughput data have a lower confidence, they are also free from prior assumptions, whereas the small-scale experiments offer a high confidence but stem from restricting hypothesis-driven approaches.

4.6. Analysis of the Protein–Protein Interaction Network within the CFTR Core Map

To analyze the manually curated core map with regard to cross-interactions between the proteins included, a protein–protein interaction network was created using the list of genes present in the core map. Physical interactions between the manually curated interactors of CFTR were identified in GeneMania [86] using only physical interactions reported by BioGrid-small scale studies [79]. The complete list of interactions can be found in Supplementary Table S7. The network was visualized using the Python plotting library Matplotlib [87] and analyzed as a weighted, undirected graph using the Python packages NetworkX [88] and python-louvain [89].

4.7. Comparison of Content Provided to the Model by Small-Scale and Large-Scale Experiments

BioInfoMiner [53] was used to analyze and compare the gene lists derived from the manually curated core map and from the coarse map containing the interactomes published by Wang et al. [45], Pankow et al. [46], Santos et al. [47] and Matos et al. [48]. The BioInfoMiner tool provides a topological analysis of semantic networks and prioritizes key systemic processes and related genes present in a set of genes. Here, analysis based on Gene Ontology [54,55] and the Reactome Pathway Database [56] were conducted. The Reactome Pathway Database represents molecular pathways that are part of human biological processes, while Gene Ontology also assigns broader biological process terms to genes and gene products. Hence, the Reactome Pathway Database mainly associates genes and gene products to rather specific terms (e.g., *N-glycan trimming in the ER and Calnexin/Calreticulin cycle*), whereas they might be assigned to more general terms in Gene Ontology (e.g., *Protein Folding*).

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10.3390/ijms22147590/s1, Glossary, Table S1: gene lists of core map and coarse map, Tables, S2–S6: lists of interactors from core map, split by submaps, Table S7 protein–protein interactions between proteins in the core map, Tables S8–S12: comparison and systemic analysis of core map and coarse map (Sheet 1: overrepresented systemic processes from Gene Ontology and Reactome Pathway Database, Sheet 2: overrepresented genes from core map from Gene Ontology-based analysis, Sheet 3: overrepresented genes from core map from Reactome Pathway Database-based analysis, Sheet 4: overrepresented genes from coarse map from Gene Ontology-based analysis, Sheet 5: overrepresented genes from coarse map from Reactome Pathway Database-based analysis), Tables S13 and S14: lists of references for primary literature overview (Table S13) and validation of interactors (Table S14), Table S15: list

of interactors from coarse map, CoreMaps.zip: PDF and .xml files of whole cell core maps and individual submaps, CoarseMaps.zip: PDF and .xml files of coarse map submaps.

Author Contributions: Conceptualization, M.M.N. and F.S.; methodology, L.V.; formal analysis, L.V.; investigation, L.V.; data curation, L.V.; writing—original draft preparation, L.V.; writing—review and editing, F.S., S.H. and M.M.N.; visualization, L.V.; funding acquisition, F.S. and M.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft DFG, grant number 315063128. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the Supplementary Material. Future versions of the CFTR Lifecycle map will be made available here https://candactcftr.ams.med.uni-goettingen.de/SystemsBiology/ (accessed on 26 January 2021).

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

References

1. Bobadilla, J.L.; Macek, M.; Fine, J.P.; Farrell, P.M. Cystic fibrosis: A worldwide analysis of CFTR mutations—Correlation with incidence data and application to screening. *Hum. Mutat.* **2002**, *19*, 575–606. [CrossRef]
2. O'Sullivan, B.P.; Freedman, S.D. Cystic fibrosis. *Lancet* **2009**, *373*, 1891–1904. [CrossRef]
3. Elborn, J.S. Cystic fibrosis. *Lancet* **2016**, *388*, 2519–2531. [CrossRef]
4. Riordan, J.R.; Rommens, J.M.; Kerem, B.S.; Alon, N.O.A.; Rozmahel, R.; Grzelczak, Z.; Zielenski, J.; Lok, S.I.; Plavsic, N.; Chou, J.L.; et al. Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA. *Science* **1989**, *245*, 1066–1073. [CrossRef]
5. O'Riordan, C.R.; Lachapelle, A.L.; Marshall, J.; Higgins, E.A.; Cheng, S.H. Characterization of the oligosaccharide structures associated with the cystic fibrosis transmembrane conductance regulator. *Glycobiology* **2000**, *10*, 1225–1233. [CrossRef]
6. Cheng, S.H.; Gregory, R.J.; Marshall, J.; Paul, S.; Souza, D.W.; White, G.A.; O'Riordan, C.R.; Smith, A.E. Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis. *Cell* **1990**, *63*, 827–834. [CrossRef]
7. Higgins, C.F.; Gallagher, M.P.; Mimmack, M.L.; Pearce, S.R. A family of closely related ATP-binding subunits from prokaryotic and eukaryotic cells. *BioEssays* **1988**, *8*, 111–116. [CrossRef] [PubMed]
8. Higgins, C. Export-import family expands. *Nature* **1989**, *340*, 342. [CrossRef] [PubMed]
9. Csanády, L.; Vergani, P.; Gadsby, D.C. Structure, gating, and regulation of the CFTR anion channel. *Physiol. Rev.* **2019**, *99*, 707–738. [CrossRef]
10. Zhang, Z.; Liu, F.; Chen, J. Molecular structure of the ATP-bound, phosphorylated human CFTR. *Proc. Natl. Acad. Sci. USA* **2018**. [CrossRef] [PubMed]
11. Meng, X.; Clews, J.; Ciuta, A.D.; Martin, E.R.; Ford, R.C. CFTR structure, stability, function and regulation. *Biol. Chem.* **2019**, *400*, 1359–1370. [CrossRef]
12. Aleksandrov, A.A.; Aleksandrov, L.A.; Riordan, J.R. CFTR (ABCC7) is a hydrolyzable-ligand-gated channel. *Pflugers Arch. Eur. J. Physiol.* **2007**, *453*, 693–702. [CrossRef]
13. Moran, O. The gating of the CFTR channel. *Cell. Mol. Life Sci.* **2017**, *74*, 85–92. [CrossRef]
14. Callebaut, I.; Chong, P.A.; Forman-Kay, J.D. CFTR structure. *J. Cyst. Fibros.* **2018**, *17*, S5–S8. [CrossRef]
15. Pranke, I.M.; Sermet-Gaudelus, I. Biosynthesis of cystic fibrosis transmembrane conductance regulator. *Int. J. Biochem. Cell Biol.* **2014**, *52*, 26–38. [CrossRef]
16. Cystic Fibrosis Mutation Database. Available online: http://www.genet.sickkids.on.ca/ (accessed on 26 January 2021).
17. Welcome to CFTR2 | CFTR2. Available online: https://www.cftr2.org/ (accessed on 26 January 2021).
18. Sosnay, P.R.; Siklosi, K.R.; Van Goor, F.; Kaniecki, K.; Yu, H.; Sharma, N.; Ramalho, A.S.; Amaral, M.D.; Dorfman, R.; Zielenski, J.; et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet.* **2013**, *45*, 1160–1167. [CrossRef] [PubMed]
19. Welsh, M.J.; Smith, A.E. Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell* **1993**, *73*, 1251–1254. [CrossRef]
20. Rowe, S.M.; Miller, S.; Sorscher, E.J. Cystic Fibrosis. *N. Engl. J. Med.* **2005**, *352*, 1992–2001. [CrossRef] [PubMed]
21. Veit, G.; Avramescu, R.G.; Chiang, A.N.; Houck, S.A.; Cai, Z.; Peters, K.W.; Hong, J.S.; Pollard, H.B.; Guggino, W.B.; Balch, W.E.; et al. From CFTR biology toward combinatorial pharmacotherapy: Expanded classification of cystic fibrosis mutations. *Mol. Biol. Cell* **2016**, *27*, 424–433. [CrossRef]
22. Zielenski, J.; Tsui, L.C. Cystic fibrosis: Genotypic and phenotypic variations. *Annu. Rev. Genet.* **1995**, *29*, 777–807. [CrossRef]
23. Zielenski, J. Genotype and Phenotype in Cystic Fibrosis. *Respiration* **2000**, *67*, 117–133. [CrossRef]

24. Kerem, B.S.; Rommens, J.M.; Buchanan, J.A.; Markiewicz, D.; Cox, T.K.; Chakravarti, A.; Buchwald, M.; Tsui, L.C. Identification of the cystic fibrosis gene: Genetic analysis. *Science* **1989**, *245*, 1073–1080. [CrossRef]
25. Lukacs, G.L.; Verkman, A.S. CFTR: Folding, misfolding and correcting the ΔF508 conformational defect. *Trends Mol. Med.* **2012**, *18*, 81–91. [CrossRef]
26. 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**. [CrossRef]
27. Wang, W.; Okeyo, G.O.; Tao, B.; Hong, J.S.; Kirk, K.L. Thermally unstable gating of the most common cystic fibrosis mutant channel (ΔF508): "Rescue" by suppressor mutations in nucleotide binding domain 1 and by constitutive mutations in the cytosolic loops. *J. Biol. Chem.* **2011**, *286*, 41937–41948. [CrossRef]
28. Martiniano, S.L.; Sagel, S.D.; Zemanick, E.T. Cystic fibrosis: A model system for precision medicine. *Curr. Opin. Pediatr.* **2016**, *28*, 312–317. [CrossRef]
29. Southern, K.W.; Patel, S.; Sinha, I.P.; Nevitt, S.J. Correctors (specific therapies for class II CFTR mutations) for cystic fibrosis. *Cochrane Database Syst. Rev.* **2018**. [CrossRef]
30. Pedemonte, N.; Lukacs, G.L.; Du, K.; Caci, E.; Zegarra-Moran, O.; Galietta, L.J.V.; Verkman, A.S. Small-molecule correctors of defective ΔF508-CFTR cellular processing identified by high-throughput screening. *J. Clin. Investig.* **2005**. [CrossRef]
31. Van Goor, F.; Hadida, S.; Grootenhuis, P.D.J.; Burton, B.; Cao, D.; Neuberger, T.; Turnbull, A.; Singh, A.; Joubran, J.; Hazlewood, A.; et al. Rescue of CF airway epithelial cell function in vitro by a CFTR potentiator, VX-770. *Proc. Natl. Acad. Sci. USA* **2009**. [CrossRef]
32. Berg, A.; Hallowell, S.; Tibbetts, M.; Beasley, C.; Brown-Phillips, T.; Healy, A.; Pustilnik, L.; Doyonnas, R.; Pregel, M. High-Throughput Surface Liquid Absorption and Secretion Assays to Identify F508del CFTR Correctors Using Patient Primary Airway Epithelial Cultures. *SLAS Discov.* **2019**. [CrossRef]
33. De Wilde, G.; Gees, M.; Musch, S.; Verdonck, K.; Jans, M.; Wesse, A.S.; Singh, A.K.; Hwang, T.C.; Christophe, T.; Pizzonero, M.; et al. Identification of GLPG/ABBV-2737, a novel class of corrector, which exerts functional synergy with other CFTR modulators. *Front. Pharmacol.* **2019**, *10*. [CrossRef]
34. Merkert, S.; Schubert, M.; Olmer, R.; Engels, L.; Radetzki, S.; Veltman, M.; Scholte, B.J.; Zöllner, J.; Pedemonte, N.; Galietta, L.J.V.; et al. High-Throughput Screening for Modulators of CFTR Activity Based on Genetically Engineered Cystic Fibrosis Disease-Specific iPSCs. *Stem Cell Reports* **2019**. [CrossRef]
35. Van Goor, F.; Hadida, S.; Grootenhuis, P.D.J.; Burton, B.; Stack, J.H.; Straley, K.S.; Decker, C.J.; Miller, M.; McCartney, J.; Olson, E.R.; et al. Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proc. Natl. Acad. Sci. USA* **2011**. [CrossRef]
36. Phuan, P.W.; Veit, G.; Tan, J.A.; Finkbeiner, W.E.; Lukacs, G.L.; Verkman, A.S. Potentiators of defective DF508-CFTR gating that do not interfere with corrector action. *Mol. Pharmacol.* **2015**. [CrossRef]
37. Carlile, G.W.; Robert, R.; Goepp, J.; Matthes, E.; Liao, J.; Kus, B.; Macknight, S.D.; Rotin, D.; Hanrahan, J.W.; Thomas, D.Y. Ibuprofen rescues mutant cystic fibrosis transmembrane conductance regulator trafficking. *J. Cyst. Fibros.* **2015**. [CrossRef]
38. Liang, F.; Shang, H.; Jordan, N.J.; Wong, E.; Mercadante, D.; Saltz, J.; Mahiou, J.; Bihler, H.J.; Mense, M. High-Throughput Screening for Readthrough Modulators of CFTR PTC Mutations. *SLAS Technol.* **2017**. [CrossRef]
39. Giuliano, K.A.; Wachi, S.; Drew, L.; Dukovski, D.; Green, O.; Bastos, C.; Cullen, M.D.; Hauck, S.; Tait, B.D.; Munoz, B.; et al. Use of a High-Throughput Phenotypic Screening Strategy to Identify Amplifiers, a Novel Pharmacological Class of Small Molecules That Exhibit Functional Synergy with Potentiators and Correctors. *SLAS Discov.* **2018**. [CrossRef]
40. Van Der Plas, S.E.; Kelgtermans, H.; De Munck, T.; Martina, S.L.X.; Dropsit, S.; Quinton, E.; De Blieck, A.; Joannesse, C.; Tomaskovic, L.; Jans, M.; et al. Discovery of N-(3-Carbamoyl-5,5,7,7-tetramethyl-5,7-dihydro-4H-thieno[2,3-c]pyran-2-yl)-lH-pyrazole-5-carboxamide (GLPG1837), a Novel Potentiator Which Can Open Class III Mutant Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Channels to a High Extent. *J. Med. Chem.* **2018**, *61*, 1425–1435. [CrossRef] [PubMed]
41. Veit, G.; Xu, H.; Dreano, E.; Avramescu, R.G.; Bagdany, M.; Beitel, L.K.; Roldan, A.; Hancock, M.A.; Lay, C.; Li, W.; et al. Structure-guided combination therapy to potently improve the function of mutant CFTRs. *Nat. Med.* **2018**. [CrossRef] [PubMed]
42. Wang, X.; Liu, B.; Searle, X.; Yeung, C.; Bogdan, A.; Greszler, S.; Singh, A.; Fan, Y.; Swensen, A.M.; Vortherms, T.; et al. Discovery of 4-[(2R,4R)-4-({[1-(2,2-Difluoro-1,3-benzodioxol-5-yl)cyclopropyl]carbonyl}amino)-7-(difluoromethoxy)-3,4-dihydro-2H-chromen-2-yl]benzoic Acid (ABBV/GLPG-2222), a Potent Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Corrector for the Treatment of Cystic Fibrosis. *J. Med. Chem.* **2018**, *61*, 1436–1449. [CrossRef] [PubMed]
43. Pedemonte, N.; Tomati, V.; Sondo, E.; Galietta, L.J.V. Influence of cell background on pharmacological rescue of mutant CFTR. *Am. J. Physiol. Cell Physiol.* **2010**, *298*. [CrossRef]
44. Novère, N.L.; Hucka, M.; Mi, H.; Moodie, S.; Schreiber, F.; Sorokin, A.; Demir, E.; Wegner, K.; Aladjem, M.I.; Wimalaratne, S.M.; et al. The Systems Biology Graphical Notation. *Nat. Biotechnol.* **2009**, *27*, 735–741. [CrossRef]
45. 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**. [CrossRef]
46. 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]

47. Santos, J.D.; Canato, S.; Carvalho, A.S.; Botelho, H.M.; Aloria, K.; Amaral, M.D.; Matthiesen, R.; Falcao, A.O.; Farinha, C.M. Folding Status Is Determinant over Traffic-Competence in Defining CFTR Interactors in the Endoplasmic Reticulum. *Cells* **2019**, *8*, 353. [CrossRef]
48. Matos, A.M.; Pinto, F.R.; Barros, P.; Amaral, M.D.; Pepperkok, R.; Matos, P. Inhibition of calpain 1 restores plasma membrane stability to pharmacologically rescued Phe508del-CFTR variant. *J. Biol. Chem.* **2019**, *294*, 13396–13410. [CrossRef]
49. Funahashi, A.; Morohashi, M.; Kitano, H.; Tanimura, N. CellDesigner: A process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO* **2003**. [CrossRef]
50. Funahashi, A.; Matsuoka, Y.; Jouraku, A.; Morohashi, M.; Kikuchi, N.; Kitano, H. CellDesigner 3.5: A versatile modeling tool for biochemical networks. *Proc. IEEE* **2008**. [CrossRef]
51. Hollande, E.; Fanjul, M.; Chemin-Thomas, C.; Devaux, C.; Demolombe, S.; Van Rietschoten, J.; Guy-Crotte, O.; Figarella, C. Targeting of CFTR protein is linked to the polarization of human pancreatic duct cells in culture. *Eur. J. Cell Biol.* **1998**. [CrossRef]
52. König, M. matthiaskoenig/libsbgn-python: 0.2.2. *Zenodo* **2020**. [CrossRef]
53. Pilalis, E.; Valavanis, I.; Chatziioannou, A. e-NIOS BioInfoMiner. Available online: https://bioinfominer.com/login (accessed on 26 January 2021).
54. Ashburner, M.; Ball, C.A.; Blake, J.A.; Botstein, D.; Butler, H.; Cherry, J.M.; Davis, A.P.; Dolinski, K.; Dwight, S.S.; Eppig, J.T.; et al. Gene ontology: Tool for the unification of biology. *Nat. Genet.* **2000**, *25*, 25–29. [CrossRef]
55. The Gene Ontology resource: Enriching a GOld mine. *Nucleic Acids Res.* **2021**. [CrossRef]
56. Jassal, B.; Matthews, L.; Viteri, G.; Gong, C.; Lorente, P.; Fabregat, A.; Sidiropoulos, K.; Cook, J.; Gillespie, M.; Haw, R.; et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **2020**. [CrossRef]
57. Mall, M.A.; Galietta, L.J.V. Targeting ion channels in cystic fibrosis. *J. Cyst. Fibros.* **2015**, *14*, 561–570. [CrossRef] [PubMed]
58. Mall, M.A.; Danahay, H.; Boucher, R.C. Emerging concepts and therapies for mucoobstructive lung disease. *Ann. Am. Thorac. Soc.* **2018**, *15*, S216–S226. [CrossRef] [PubMed]
59. Danahay, H.L.; Lilley, S.; Fox, R.; Charlton, H.; Sabater, J.; Button, B.; McCarthy, C.; Collingwood, S.P.; Gosling, M. TMEM16A Potentiation: A Novel Therapeutic Approach for the Treatment of Cystic Fibrosis. *Am. J. Respir. Crit. Care Med.* **2020**. [CrossRef]
60. Rennolds, J.; Boyaka, P.N.; Bellis, S.L.; Cormet-Boyaka, E. Low temperature induces the delivery of mature and immature CFTR to the plasma membrane. *Biochem. Biophys. Res. Commun.* **2008**. [CrossRef]
61. Luo, Y.; McDonald, K.; Hanrahan, J.W. Trafficking of immature ΔF508-CFTR to the plasma membrane and its detection by biotinylation. *Biochem. J.* **2009**. [CrossRef]
62. Gee, H.Y.; Noh, S.H.; Tang, B.L.; Kim, K.H.; Lee, M.G. Rescue of Δf508-CFTR trafficking via a GRASP-dependent unconventional secretion pathway. *Cell* **2011**. [CrossRef]
63. Yoo, J.S.; Moyer, B.D.; Bannykh, S.; Yoo, H.M.; Riordan, J.R.; Balch, W.E. Non-conventional trafficking of the cystic fibrosis transmembrane conductance regulator through the early secretory pathway. *J. Biol. Chem.* **2002**. [CrossRef]
64. Gee, H.Y.; Kim, J.Y.; Lee, M.G. Analysis of conventional and unconventional trafficking of CFTR and other membrane proteins. *Methods Mol. Biol.* **2015**. [CrossRef]
65. Piao, H.; Kim, J.; Noh, S.H.; Kweon, H.S.; Kim, J.Y.; Lee, M.G. Sec16A is critical for both conventional and unconventional secretion of CFTR. *Sci. Rep.* **2017**. [CrossRef] [PubMed]
66. Braschi, B.; Denny, P.; Gray, K.; Jones, T.; Seal, R.; Tweedie, S.; Yates, B.; Bruford, E. Genenames.org: The HGNC and VGNC resources in 2019. *Nucleic Acids Res.* **2019**, *47*, D786–D792. [CrossRef]
67. Hastings, J.; Owen, G.; Dekker, A.; Ennis, M.; Kale, N.; Muthukrishnan, V.; Turner, S.; Swainston, N.; Mendes, P.; Steinbeck, C. ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**. [CrossRef] [PubMed]
68. Le Novère, N.; Finney, A.; Hucka, M.; Bhalla, U.S.; Campagne, F.; Collado-Vides, J.; Crampin, E.J.; Halstead, M.; Klipp, E.; Mendes, P.; et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **2005**, *23*, 1509–1515. [CrossRef]
69. Bateman, A. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [CrossRef]
70. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef]
71. Riordan, J.R. CFTR Function and Prospects for Therapy. *Annu. Rev. Biochem.* **2008**. [CrossRef]
72. 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]
73. PubMed. Available online: https://pubmed.ncbi.nlm.nih.gov/ (accessed on 26 January 2021).
74. Naren, A.P.; Cobb, B.; Li, C.; Roy, K.; Nelson, D.; Heda, G.D.; Liao, J.; Kirk, K.L.; Sorscher, E.J.; Hanrahan, J.; et al. A macromolecular complex of beta 2 adrenergic receptor, CFTR, and ezrin/radixin/moesin-binding phosphoprotein 50 is regulated by PKA. *Proc. Natl. Acad. Sci. USA* **2003**, *100*, 342–346. [CrossRef]
75. Li, J.; Dai, Z.; Jana, D.; Callaway, D.J.E.; Bu, Z. Ezrin Controls the Macromolecular Complexes Formed between an Adapter Protein Na+/H+ Exchanger Regulatory Factor and the Cystic Fibrosis Transmembrane Conductance Regulator. *J. Biol. Chem.* **2005**, *280*, 37634–37643. [CrossRef]
76. Loureiro, C.A.; Matos, A.M.; Dias-Alves, Â.; Pereira, J.F.; Uliyakina, I.; Barros, P.; Amaral, M.D.; Matos, P. A molecular switch in the scaffold NHERF1 enables misfolded CFTR to evade the peripheral quality control checkpoint. *Sci. Signal.* **2015**, *8*, ra48. [CrossRef]

77. René, C.; Taulan, M.; Iral, F.; Doudement, J.; L'Honoré, A.L.; Gerbon, C.; Demaille, J.; Claustres, M.; Romey, M.C. Binding of serum response factor to cystic fibrosis transmembrane conductance regulator CArG-like elements, as a new potential CFTR transcriptional regulation pathway. *Nucleic Acids Res.* **2005**. [CrossRef] [PubMed]
78. Licata, L.; Lo Surdo, P.; Iannuccelli, M.; Palma, A.; Micarelli, E.; Perfetto, L.; Peluso, D.; Calderone, A.; Castagnoli, L.; Cesareni, G. SIGNOR 2.0, the SIGnaling Network Open Resource 2.0: 2019 update. *Nucleic Acids Res.* **2020**, *48*, D504–D510. [CrossRef] [PubMed]
79. Oughtred, R.; Rust, J.; Chang, C.; Breitkreutz, B.J.; Stark, C.; Willems, A.; Boucher, L.; Leung, G.; Kolas, N.; Zhang, F.; et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Protein Sci.* **2021**, *30*, 187–200. [CrossRef] [PubMed]
80. Keshava Prasad, T.S.; Goel, R.; Kandasamy, K.; Keerthikumar, S.; Kumar, S.; Mathivanan, S.; Telikicherla, D.; Raju, R.; Shafreen, B.; Venugopal, A.; et al. Human Protein Reference Database—2009 update. *Nucleic Acids Res.* **2009**, *37*. [CrossRef] [PubMed]
81. Szklarczyk, D.; Gable, A.L.; Lyon, D.; Junge, A.; Wyder, S.; Huerta-Cepas, J.; Simonovic, M.; Doncheva, N.T.; Morris, J.H.; Bork, P.; et al. STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **2019**. [CrossRef]
82. Licata, L.; Briganti, L.; Peluso, D.; Perfetto, L.; Iannuccelli, M.; Galeota, E.; Sacco, F.; Palma, A.; Nardozza, A.P.; Santonico, E.; et al. MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Res.* **2012**. [CrossRef] [PubMed]
83. Breuer, K.; Foroushani, A.K.; Laird, M.R.; Chen, C.; Sribnaia, A.; Lo, R.; Winsor, G.L.; Hancock, R.E.W.; Brinkman, F.S.L.; Lynn, D.J. InnateDB: Systems biology of innate immunity and beyond—Recent updates and continuing curation. *Nucleic Acids Res.* **2013**. [CrossRef] [PubMed]
84. Alonso-López, D.; Campos-Laborie, F.J.; Gutiérrez, M.A.; Lambourne, L.; Calderwood, M.A.; Vidal, M.; De Las Rivas, J. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. *Database* **2019**. [CrossRef]
85. 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**. [CrossRef]
86. Warde-Farley, D.; Donaldson, S.L.; Comes, O.; Zuberi, K.; Badrawi, R.; Chao, P.; Franz, M.; Grouios, C.; Kazi, F.; Lopes, C.T.; et al. The GeneMANIA prediction server: Biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.* **2010**, *38*. [CrossRef] [PubMed]
87. Hunter, J.D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
88. Hagberg, A.A.; Schult, D.A.; Swart, P.J. Exploring Network Structure, Dynamics, and Function using NetworkX. In Proceedings of the 7th Python in Science Conference, Pasadena, CA USA, 19–24 August 2008; Varoquaux, G., Vaught, T., Millman, J., Eds.; Los Alamos National Lab.: Walnut Creek, CA, USA, 2008; pp. 11–15.
89. Aynaud, T. Python-Louvain x.y: Louvain Algorithm for Community Detection 2020. Available online: https://github.com/taynaud/python-louvain (accessed on 26 January 2021).

Chapter 4 Integrating text mining into the curation of disease maps

This manuscript has originally been published in the Journal Biomolecules

Integrating text mining into the curation of disease maps

Malte Voskamp[1,†], Liza Vinhoven[1,†], Frauke Stanke[2,3], Sylvia Hafkemeyer[4] and Manuel Manfred Nietert[1,5]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2]German Center for Lung Research (DZL), Partner Site BREATH, Hannover, Germany
[3]Clinic for Pediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Hannover, Germany
[4]Mukoviszidose Institut gGmbH, Bonn, Germany
[5]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany
† These authors contributed equally to this work

Authors contribution

L.V.: conceptualization; methodology; data curation; visualization; writing – original draft preparation; supervision

M.V.: methodology; software; data curation; writing – original draft preparation

F.S.: writing – review and editing; funding acquisition

S.H.: writing – review and editing

M.M.N.: conceptualization; writing – review and editing; supervision; funding acquisition

biomolecules

MDPI

Article

Integrating Text Mining into the Curation of Disease Maps

Malte Voskamp [1,†], Liza Vinhoven [1,†], Frauke Stanke [2,3], Sylvia Hafkemeyer [4] and Manuel Manfred Nietert [1,5,*]

[1] Department of Medical Bioinformatics, University Medical Center Göttingen, Goldschmidtstraße 1, 37077 Göttingen, Germany
[2] Clinic for Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
[3] Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), the German Center for Lung Research, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
[4] Mukoviszidose Institut gGmbH, In den Dauen 6, 53117 Bonn, Germany
[5] CIDAS Campus Institute Data Science, Goldschmidtstraße 1, 37077 Göttingen, Germany
* Correspondence: manuel.nietert@med.uni-goettingen.de; Tel.: +49-551-39-14920
† These authors contributed equally to this work.

Abstract: An adequate visualization form is required to gain an overview and ultimately understand the complex and diverse biological mechanisms of diseases. Recently, disease maps have been introduced for this purpose. A disease map is defined as a systems biological map or model that combines metabolic, signaling, and physiological pathways to create a comprehensive overview of known disease mechanisms. With the increase in publications describing biological interactions, efforts in creating and curating comprehensive disease maps is growing accordingly. Therefore, new computational approaches are needed to reduce the time that manual curation takes. Test mining algorithms can be used to analyse the natural language of scientific publications. These types of algorithms can take humanly readable text passages and convert them into a more ordered, machine-usable data structure. To support the creation of disease maps by text mining, we developed an interactive, user-friendly disease map viewer. The disease map viewer displays text mining results in a systems biology map, where the user can review them and either validate or reject identified interactions. Ultimately, the viewer brings together the time-saving advantages of text mining with the accuracy of manual data curation.

Keywords: text mining; disease maps; systems biology

Citation: Voskamp, M.; Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. Integrating Text Mining into the Curation of Disease Maps. *Biomolecules* **2022**, *12*, 1278. https://doi.org/10.3390/biom12091278

Academic Editor: Jiangning Song

Received: 19 August 2022
Accepted: 7 September 2022
Published: 10 September 2022

1. Introduction

Every day, more and more data and knowledge on different diseases and their underlying biological pathways are being acquired. Thus, it is becoming increasingly important to develop methods of data and knowledge integration, storage, and representation in ways that can be interpreted and analysed by humans and computers alike. One of these approaches is systems medicine disease maps, which has been proposed by Mazein et al. in 2018. The authors define disease maps as a "comprehensive, knowledge-based representation of disease mechanisms" [1]. They evolved from and are comparable to metabolic and signaling pathways, stored and represented in standardized formats such as the Systems Biology Graphical Notation (SBGN) [2] or Systems Biology Markup Language (SBML) [3]. A major difference between metabolic or signaling pathways and disease maps is that the latter are not limited to biochemical or regulatory relations between entities but can also include physiological ones. Disease maps can be used for a multitude of purposes, such as identifying disease biomarkers and drug targets, drug repositioning, structuring omics data, and developing improved diagnostics [1,4]. Most recently, a large, interdisciplinary community of over 230 researchers launched a project to create a COVID-19 disease map [5]. This resulted in what, to the best of our knowledge, is the largest disease map to date, currently

consisting of 5499 elements, which are connected by 1836 interactions across 42 diagrams. The data for this enormous knowledge resource were curated from 617 publications and preprints, highlighting the sheer time and manpower required to create these manually curated disease maps. One way to support the construction of disease maps is text mining, the automated annotation of texts that produces a condensed keyword list, which can then be formatted into machine- and human-readable media and to consist of the core information of that text. In principle, text mining means the extraction of information from textual data, thereby speeding up the curation and annotation process of human-written text [6]. To do so, many possible information technologies are applicable, for example, machine learning, pattern matching, or the processing of natural, human-readable language [7].

In general, a text mining algorithm will follow the steps below. As an input, the algorithm will take a human-readable sentence, in this case from a biological paper. It will then first highlight the named entities (NE), which are terms that are then normalized and transformed into identifiers. These NEs can be proteins, genes, diseases, or any other biologically relevant term, taken from an underlying database that contains NEs that the system should be able to identify. This recognition (named entity recognition (NER)) is crucial for the success and effectiveness of text mining and is therefore a focus of refinement to increase the specificity and sensitivity of the algorithm. The entities are then assigned to unique identifiers, which are then organized into an identifier scheme. Afterward, the extracted relationships from the input text data are included between named entities. The resulting network of nodes and relationships can then be compared and expanded with additional text data. With the help of this network, new hypotheses can be formed and these can then be the subject of further research [7].

One of the main challenges in NER is the multitude of different identifiers for almost anything in biology or chemistry, sometimes varying greatly between different publications and databases. This variation in names for the same biological entity needs to be recognized and normalized by the algorithms. In addition to these intended differences in nomenclature, there are more variations that need normalization: for example, variations in orthography ("amino acid" vs. "amino-acid"), abbreviations, and spelling variations, such as upper/lower case or American vs. British English wording [8]. All these variations must be taken into account, and the term needs to be assigned to the same biological identifier, which then results in a list of possible terms all referring to the same ID. When it comes to interactions, even more words can be used to describe similar relations between entities. The system needs to recognize the buzzwords for relations and the entity terms to create the desired entity-relationship model. Moreover, differences in the structure of the sentence in combination with the wording can be challenging to the system.

Text mining has been gaining more and more applications in scientific projects over the last two decades. The principle and technique of data mining have been known since the late 1990s but have not been widely used by the scientific community [9]. In particular, in systems biology and biomedicine, the use of text mining can be of essential value. Even if those scientific fields rely heavily on data stored in unified formats and databases to ensure cross-author usability, a substantial proportion of essential information is still only available as text in human-written publications. As of now, there are many algorithms that are specialized for biological terms that are implemented as NER. In order to establish, compare, and evaluate common standards challenges such as BioCreative (http://www.biocreative.org, accessed on 1 September 2022) have been put into place, which aims to compare methods and critically assess scientific progress in text mining [10]. Currently, biological text mining and NER specifically already find applications in the curation of different databases. For example, the BRENDA database (BRaunschweigENzymeDatabase; http://www.brenda-enzymes.org, accessed on 1 September 2022) [11], which collects enzyme functional data, employs text mining approaches to extract kinetic data from PubMed abstracts [11]. Furthermore, the protein interaction database STRING (https://string-db.org/, accessed on 1 September 2022) uses text-mined data to identify protein–

protein interactions [12]. A more extensive overview of many more examples of existing text mining applications with a focus on cancer research can be found in Zhu et al., 2013 [7].

Nonetheless, even though great strides have been made in the development of text mining algorithms with high sensitivity and specificity, they cannot yet replace a human expert curator. We, therefore, developed a tool to bring together the advantages of text mining and the expert knowledge and experience of scientists to support the creation of systems biology disease maps. Our tool consists of an interactive disease map viewer, which takes the output of an independent text mining system, translates it to the required format, and displays it in a disease map-like cellular layout. This allows the user to utilize the text mining approach they find most suitable for their use case or even include results from more than one system. The user then has the possibility to examine the interactions identified by the text mining algorithm and evaluate them based on the text passage they are based on. In the end, this results in a list of automatically parsed but expert-validated interactions, which can then be used as a basis for a disease map. Ultimately, this simplifies and significantly speeds up the curation step during the construction of disease maps.

2. Materials and Methods

2.1. Data Preparation

To bring text mining results into a format appropriate for further use, the results from an independent text mining algorithm were brought into a simple, reproducible format, consisting of two tables in CSV format. One table consists of all mined entities and their subcellular localization, and the other includes derived interactions observed between them. The tables were then parsed into the JavaScript Object Notation (JSON) format. The JSON format is a very storage-efficient way to save and interchange data between different JavaScript applications. Currently, it is widely used for providing data to the user from a server or a web service, where data can be parsed via a host's API (Application Programming Interface). The conversion of the tables into JSON format was performed using Python with the libraries: pandas [13,14], numpy [15], libsbgnpy [16], and json.

The resulting JSON file was then further used as an exemplary disease map for the disease map viewer.

2.2. Disease Map Viewer Implementation

The disease map viewer tool was implemented with the Cytoscape.js library [17]. Cytoscape.js is the JavaScript variant of the Cytoscape software [18]. Cytoscape itself is an open-source project for accessing and viewing graphical networks inside a downloadable instance. This software can be programmatically accessed and therefore personalized and implemented into our tool via the JS library Cytoscape.js. Another big advantage of the Cytoscape.js variant is the capability of loading data dynamically while the user is browsing the graphical map. Furthermore, it is possible to load big maps in a memory-efficient manner into the Cytoscape.js instance using the JSON format.

To make our Cytoscape.js instance accessible, we used Grails and our previously developed CandActBase [19] as an underlying framework. We used the AJAX (Asynchronous JavaScript and XML) protocol to dynamically load data into a JavaScript application from a web server [20]. AJAX is capable of loading data dynamically based on the input of the user, even if the website has already been loaded completely. The AJAX call will access a defined URL (in this case, a local file) and load the data into the JS script. This data can then be processed, altered, and presented by the rest of the code. This asynchronous behaviour makes AJAX valuable for our purpose and improves the speed of the script significantly.

3. Results

In order to support the creation of disease maps, we developed a tool capable of displaying text mining results as disease maps and validating them through the integration of expert domain knowledge.

For this purpose, we used an independent, exchangeable text mining algorithm to parse molecular interactions between biological entities' data from publicly available scientific text. The results are output in two simple, reproducible CSV files, one containing the interactions between the entities themselves and the other specifying their subcellular localization. A flowchart of the input data, software, and output data of the systems can be seen in Figure 1.

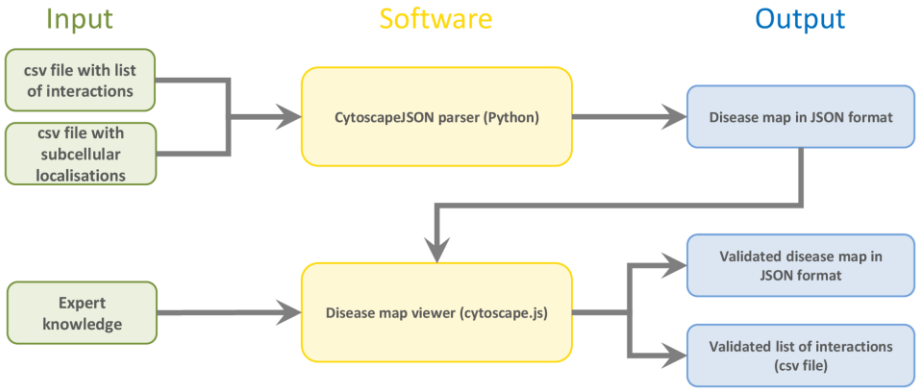


Figure 1. Flowchart of the processes included in the tool. Input knowledge and data are shown in green on the right, the software modules are shown in yellow, and the output files are shown in blue on the right. Two CSV files, one containing the list of interactions and one containing the subcellular localisation of the entities, serve as input for the CytoscapeJSON parser implemented in Python. The resulting JSON file serves as input for the disease map viewer, where the interactions are validated by expert knowledge. The validated interactions can then be exported in a cellular layout in a JSON file or as a list of interactions in a CSV file.

To prepare text mining results that are easy to store, share, and use, we used a Python script to convert them from a simple CSV file to JSON format. Simply put, the JSON data structure of the text mining results is a list of every element (nodes, compartments, and edges) in the disease map. Depending on the element, the structure differs slightly. Each element has three basic properties: "data", "position", and "group". The "group" specifies if the element is a node or an edge, i.e., a molecular entity or an interaction. The "position" property, which is automatically created by the python script, sets the x and y parameters to assign it to a specific location on the map. The most advanced property of each element is the "data" property, where all associated data are stored. Additionally, edge-type entities have the property "classes", where the category of the interaction is defined ("neutral", "inhibit", "activate", and "undefined"). Further properties are the unique identifier, and cytoscape.js-specific parameters (For more external information visit https://js.cytoscape.org/, accessed on 1 September 2022). The following additional parameters are important for representation in the SBGN format: For nodes, the "label" is the name specified, and the "parent" is the cellular compartment in which the gene is active. For edges, the start and end nodes are defined by the respective identifiers. Furthermore, all edges have a parameter called "references", which lists the PubMed IDs of all references this edge is based on. For each reference, the PubMed ID is given together with the sentence where the interaction was identified. Moreover, all verbs found in those sentences as well as the categorization of those verbs are stored.

This SBML-based JSON format is used by the Cytoscape.js library to create the graphical SBGN map from it.

The interface is built around the Cytoscape.js instance that renders and displays disease maps to help the user annotate and review the text-mined disease map conveniently.

Figure 2 shows the interface with exemplary data. The main graph is shown in a cell-like layout, where the user can zoom in and out. The rectangular nodes represent the molecular entities and are localized in the subcellular compartment specified in the JSON file. The arrow-shaped edges represent molecular interactions between them. All entities (genes/proteins and compartments), as well as their respective edges, can be moved freely by dragging to improve structure and visibility to fit the user's needs.

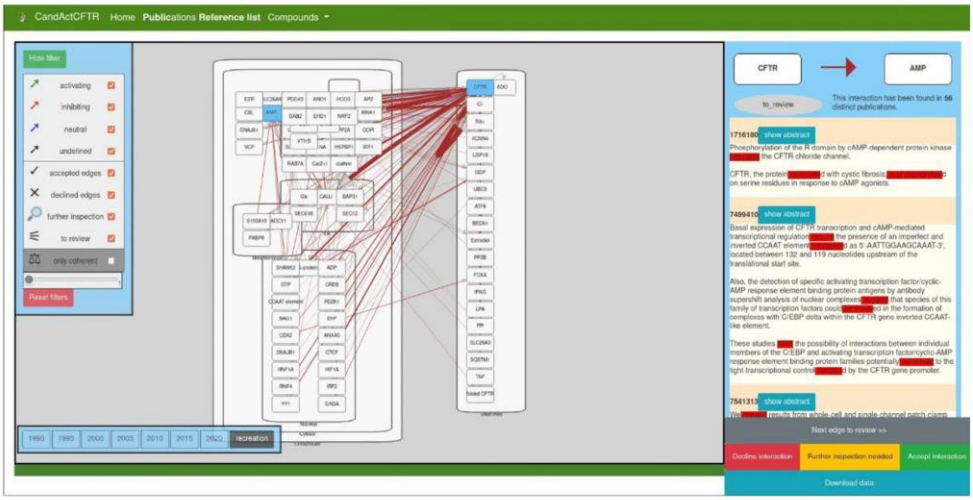


Figure 2. Interface of the disease map viewer. The large window in the middle shows the text mining data as a coarse disease map in a cellular layout. The left sidebar shows the legend and filter options, and the right sidebar shows the review function, where the supporting sentences from the parsed publications are displayed and the user can validate or reject an interaction. The buttons on the bottom left show the timeline option, where the interaction data can be filtered by date of publication.

The colouring is the colour of categorization of found verbs. All "activating" edges are coloured green, "inhibiting" edges are coloured red, "neutral" edges are coloured blue, and "undefined" edges have a grey colour, while incoherent interactions are shown in brown.

The left sidebar shows the legend and filter options for the edges in the graph. As a default, all edges are displayed, but the user can uncheck types of edges to hide them and thus obtain a better overview of the remaining categories of edges. This legend can be opened and closed by clicking the top button "hide/show filter".

Another way the data from the text mining are categorized is by the thickness of the edges in the graph. The more distinct publications have been found to have both connected nodes mentioned in the same sentence, the thicker the edge between them. In the bottom-left corner of the filter window, the user can filter the edges depending on the number of supporting publications. The slider can be moved to define a minimum number of publications an edge needs to have to display it. Moreover, below the slider is a button that will reset the filter and reload the map.

Another feature of the disease map/SBGN map viewer is the timeline function. As an interesting use case of our text mining workflow, we chose to create a timeline made from SBGN maps from publications published in different years and, thus, show the focus of research in the past. To obtain biological interactions that are associated with the query subject over time, we categorized texts by their years of publication. Thus, we created exemplary momentary snapshots over the years. The user can choose which disease map from which year they would like to access between the years 1990 and 2020 in 5-year steps.

In order to integrate expert knowledge and validate text-mined data, we included a review function, as observed in the right-hand panel of the interface. The user can examine all interactions with two methods: by clicking the "Next edge" button to iterate all interactions that need to be reviewed or by directly selecting a specific edge from the graph. The review panel will then display the two nodes connected by the clicked edge and the colour of the edge between both, as well as the current review status of the interaction. Below this, a list of PubMed IDs is displayed together with the sentences that have been used to identify the interaction in each reference. The verbs that have been used to categorize the interaction are coloured in red. The user can then load the entire text to obtain more context for the sentence. The user can then review the interaction with all available data on hand and assign a status to the interaction. If the expert approves the text-mined interaction, the "accept" status can be selected. If the text-mined interaction is a false positive, the "decline" status is appropriate, and if more research needs to be conducted to approve the interaction, the "further inspection needed" status can be assigned.

To view the status of the review process, the data can be downloaded either as a CSV file with all interactions, their current review status, and the PubMed ID from with the interaction, which was text mined from the disease map, or as a JSON file with the entire disease map in a JSON object that can be saved for reloading in a later session or to share with other users.

4. Discussion

With more and more biological and biomedical data being published, more knowledge is available and needs to be processed and structured. One way to do so in biomedicine is by using disease maps that visualize and describe disease pathways in a human- and machine-readable medium [1]. However, with the increasing number of publications every year, new computational approaches are needed to support researchers and clinicians in filtering and annotating large data sets to extract scientifically meaningful knowledge. Here, we propose a tool to (re)view text-mined data and display it appropriately to accelerate the curation process of textual data significantly. It spares the researcher from having to manually read large sets of publications to construct or curate disease maps but allows them to conveniently iterate text-mined interactions and preprocessed publications to verify found interactions.

Our tool can be used in combination with text mining software to preprocess large textual data sets and review the text mining results easily to ultimately combine the advantages of the speed of text mining with the accuracy of manual data curation reviewed by experts in the scientific field in question. The interface is kept clearly laid out and is easy to use; thus, researchers with limited experience in computational software can use it intuitively.

To ensure maximum transparency, the text-mined data can be reviewed in a very detailed manner. Every text-mined interaction can be examined to see which terms in which sentences from which publications were used to identify an interaction. In this way, the reviewer can closely inspect if the interaction is a true or false positive and mark the interaction accordingly. Moreover, all data can be downloaded at every step of the curation process. In this way, the data can be shared with co-workers and peer-reviewed easily. For this purpose, standardized data formats are used to ensure the exchangeability of the input text mining data. Therefore, the viewer can use interchangeable external text mining software just by making little adjustments to the input data. This is important with respect to the rapid improvement of text mining algorithms. The tool can be used to display results from all kinds of text mining software and can be employed for comparison purposes.

To the best of our knowledge, this is the first tool that integrates text mining directly into the disease map curation process. Several different tools have been developed to extract interactions between biological entities and can create protein–protein Interaction (PPI) networks [21–23]. The HPIminer, for example, uses NER to identify interactions from sentences and then adds information from PPI databases and additionally extracts, overlays,

and displays KEGG (http://www.kegg.jp, accessed on 1 September 2022) pathways from the two interacting proteins [21]. All these tools come with their own highly sophisticated text mining algorithms and include different data sources to produce extensive networks. In contrast, our tool does not focus on text mining itself but on making the results from the already existing, high-quality text mining tools usable and integratable. Users can use their preferred text mining tool or algorithm and visualize the results in the disease map viewer so domain experts can verify the data and then further utilize it, e.g., in a disease map.

To show how the viewer operates, we used an individualized text mining workflow to create a sample data set with the use case of cystic fibrosis, based on the CFTR Lifecycle Map we previously curated [24].

The disease map viewer, installation instructions, and the exemplary cystic fibrosis data set are available under https://s.gwdg.de/8bK6f5, accessed on 2 September 2022 (Supplementary Materials).

5. Conclusions

We developed a tool to create an interface between biological text mining and the creation of systems medicine disease maps. Our disease map viewer takes the interaction data extracted by a text mining algorithm of choice and displays it in a cellular layout and interactive manner. Domain experts can then intuitively examine individual interactions and validate or reject them, and the verified interactions can be exported for further use. This supports the creation of disease maps and systems biological models, as it brings together the speed of automated text mining and the high accuracy of human expert knowledge, thereby using the benefits of both without sacrificing quality or time effectiveness.

Supplementary Materials: The code for the disease map viewer plugin as well as installation instructions can be found on https://s.gwdg.de/8bK6f5, accessed on 2 September 2022.

Author Contributions: Conceptualization, L.V. and M.M.N.; methodology, M.V. and L.V.; software, M.V.; data curation, M.V. and L.V.; writing—original draft preparation, M.V. and L.V.; writing—review and editing, S.H., F.S. and M.M.N.; visualization, M.V. and L.V.; supervision, L.V. and M.M.N.; funding acquisition, F.S. and M.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft DFG, grant number 315063128. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Mazein, A.; Ostaszewski, M.; Kuperstein, I.; Watterson, S.; Le Novère, N.; Lefaudeux, D.; De Meulder, B.; Pellet, J.; Balaur, I.; Saqi, M.; et al. Systems medicine disease maps: Community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl.* **2018**, *4*, 21. [CrossRef] [PubMed]
2. Novère, N.L.; Hucka, M.; Mi, H.; Moodie, S.; Schreiber, F.; Sorokin, A.; Demir, E.; Wegner, K.; Aladjem, M.I.; Wimalaratne, S.M.; et al. The Systems Biology Graphical Notation. *Nat. Biotechnol.* **2009**, *27*, 735–741. [CrossRef] [PubMed]
3. Hucka, M.; Finney, A.; Sauro, H.M.; Bolouri, H.; Doyle, J.C.; Kitano, H.; Arkin, A.P.; Bornstein, B.J.; Bray, D.; Cornish-Bowden, A.; et al. The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics* **2003**, *19*, 524–531. [CrossRef] [PubMed]
4. Ostaszewski, M.; Gebel, S.; Kuperstein, I.; Mazein, A.; Zinovyev, A.; Dogrusoz, U.; Hasenauer, J.; Fleming, R.M.T.; Le Novère, N.; Gawron, P.; et al. Community-driven roadmap for integrated disease maps. *Brief. Bioinform.* **2019**, *20*, 659–670. [CrossRef] [PubMed]
5. Ostaszewski, M.; Niarakis, A.; Mazein, A.; Kuperstein, I.; Phair, R.; Orta-Resendiz, A.; Singh, V.; Aghamiri, S.S.; Acencio, M.L.; Glaab, E.; et al. COVID-19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms. *Mol. Syst. Biol.* **2021**, *17*. [CrossRef] [PubMed]

6. Harmston, N.; Filsell, W.; Stumpf, M.P.H. What the papers say: Text mining for genomics and systems biology. *Hum. Genom.* **2010**, *5*, 17–29. [CrossRef] [PubMed]
7. Zhu, F.; Patumcharoenpol, P.; Zhang, C.; Yang, Y.; Chan, J.; Meechai, A.; Vongsangnak, W.; Shen, B. Biomedical text mining and its applications in cancer research. *J. Biomed. Inform.* **2013**, *46*, 200–211. [CrossRef] [PubMed]
8. Ananiadou, S.; Kell, D.B.; Tsujii, J. Text mining and its potential applications in systems biology. *Trends Biotechnol.* **2006**, *24*, 571–579. [CrossRef] [PubMed]
9. Hearst, M.A. Untangling text data mining. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*; Association for Computational Linguistics: Morristown, NJ, USA, 1999; pp. 3–10.
10. BioCreative—Latest 3 News Items. Available online: https://biocreative.bioinformatics.udel.edu/ (accessed on 12 July 2022).
11. Chang, A.; Jeske, L.; Ulbrich, S.; Hofmann, J.; Koblitz, J.; Schomburg, I.; Neumann-Schaal, M.; Jahn, D.; Schomburg, D. BRENDA, the ELIXIR core data resource in 2021: New developments and updates. *Nucleic Acids Res.* **2021**, *49*, D498–D508. [CrossRef] [PubMed]
12. Szklarczyk, D.; Gable, A.L.; Nastou, K.C.; Lyon, D.; Kirsch, R.; Pyysalo, S.; Doncheva, N.T.; Legeay, M.; Fang, T.; Bork, P.; et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **2021**, *49*, D605–D612. [CrossRef] [PubMed]
13. *The Pandas Development Team Pandas-Dev./Pandas: Pandas, Zenodo*, 2020.
14. McKinney, W. *Data Structures for Statistical Computing in Python*, 2010; 56–61.
15. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
16. König, M. *Matthiaskoenig/Libsbgn-Python: Libsbgn-Python-V0.2.0, Zenodo*, 2020.
17. Franz, M.; Lopes, C.T.; Huck, G.; Dong, Y.; Sumer, O.; Bader, G.D. Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics* **2015**, *32*, btv557. [CrossRef] [PubMed]
18. Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N.S.; Wang, J.T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**, *13*, 2498–2504. [CrossRef] [PubMed]
19. Nietert, M.M.; Vinhoven, L.; Auer, F.; Hafkemeyer, S.; Stanke, F. Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR. *Front. Pharmacol.* **2021**, *12*, 689205. [CrossRef] [PubMed]
20. Goll, M. Asynchronous JavaScript and XML. In *JavaServer Faces*; Springer Fachmedien Wiesbaden: Wiesbaden, Germany, 2020; pp. 189–196.
21. Subramani, S.; Kalpana, R.; Monickaraj, P.M.; Natarajan, J. HPIminer: A text mining system for building and visualizing human protein interaction networks and pathways. *J. Biomed. Inform.* **2015**, *54*, 121–131. [CrossRef] [PubMed]
22. Raja, K.; Natarajan, J.; Kuusisto, F.; Steill, J.; Ross, I.; Thomson, J.; Stewart, R. Automated extraction and visualization of protein–protein interaction networks and beyond: A text-mining protocol. In *Methods in Molecular Biology*; Humana Press Inc.: Totowa, NJ, USA, 2020; Volume 2074, pp. 13–34.
23. He, M.; Wang, Y.; Li, W. PPI Finder: A Mining Tool for Human Protein-Protein Interactions. *PLoS ONE* **2009**, *4*, e4554. [CrossRef] [PubMed]
24. Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations. *Int. J. Mol. Sci.* **2021**, *22*, 7590. [CrossRef] [PubMed]

Chapter 5 Mapping Compound Databases to Disease Maps – A MINERVA Plugin for CandActBase

This manuscript has originally been published in the Journal of Personalized Medicine

Mapping Compound Databases to Disease Maps – A MINERVA Plugin for CandActBase

Liza Vinhoven[1,†], Malte Voskamp[1,†] and Manuel Manfred Nietert[1,2]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany
† These authors contributed equally to this work

Authors contribution

L.V.: conceptualization; methodology; visualization; writing – original draft preparation,

M.V.: methodology; software; writing – review and editing

M.M.N.: conceptualization; writing – review and editing; supervision; funding acquisition

Journal of Personalized Medicine

MDPI

Article

Mapping Compound Databases to Disease Maps—A MINERVA Plugin for CandActBase

Liza Vinhoven [1,†], Malte Voskamp [1,†] and Manuel Manfred Nietert [1,2,*]

[1] Department of Medical Bioinformatics, University Medical Center Göttingen, Goldschmidtstraße 1, 37077 Göttingen, Germany; liza.vinhoven@med.uni-goettingen.de (L.V.); malte.voskamp@stud.uni-goettingen.de (M.V.)
[2] CIDAS Campus Institute Data Science, Goldschmidtstraße 1, 37077 Göttingen, Germany
* Correspondence: manuel.nietert@med.uni-goettingen.de; Tel.: +49-551-39-14920
† These authors contributed equally to this work.

Abstract: The MINERVA platform is currently the most widely used platform for visualizing and providing access to disease maps. Disease maps are systems biological maps of molecular interactions relevant in a certain disease context, where they can be used to support drug discovery. For this purpose, we extended MINERVA's own drug and chemical search using the MINERVA plugin starter kit. We developed a plugin to provide a linkage between disease maps in MINERVA and application-specific databases of candidate therapeutics. The plugin has three main functionalities; one shows all the targets of all the compounds in the database, the second is a compound-based search to highlight targets of specific compounds, and the third can be used to find compounds that affect a certain target. As a use case, we applied the plugin to link a disease map and compound database we previously established in the context of cystic fibrosis and, herein, point out possible issues and difficulties. The plugin is publicly available on GitLab; the use-case application to cystic fibrosis, connecting disease maps and the compound database CandActCFTR, is available online.

Keywords: systems medicine; disease maps; drug targets; drug repurposing; knowledge repository; data integration

Citation: Vinhoven, L.; Voskamp, M.; Nietert, M.M. Mapping Compound Databases to Disease Maps—A MINERVA Plugin for CandActBase. *J. Pers. Med.* **2021**, *11*, 1072. https://doi.org/10.3390/jpm11111072

Academic Editors: Ali Salehzadeh-Yazdi and Mohieddin Jafari

Received: 16 September 2021
Accepted: 22 October 2021
Published: 24 October 2021

1. Introduction

One of the main aims of systems biology and systems medicine is to understand and model molecular mechanisms in diseases, which can support the development of novel therapeutics. For this purpose, disease maps are being developed to represent existing knowledge on disease mechanisms in a computationally readable and comprehensive manner [1]. These disease maps can then be used by clinicians and experimental scientists as well as computational scientists for different purposes, such as structuring high-throughput data, identifying disease biomarkers, developing better diagnostics and also identifying potential drug targets and drug repositioning [1,2].

To make disease maps publicly available in a comprehensive way, an open source visualization platform for disease maps and molecular interaction networks was developed in 2016 under the name of Molecular Interaction Network Visualization (MINERVA) [3]. MINERVA provides the means to visualize disease maps and make them accessible and explorable for the public, as well as individualize and extend the platform through plugins to suit one's purposes. As described above, disease maps are especially useful for structuring high-throughput data and assisting in the identification of drug targets and potential therapeutics [1]. A common method in the field of drug discovery is high-throughput screens, where hundreds of small molecules are tested in a certain disease context. These often lead to promising lead substances; however, the means by which the compounds achieve their effects remain unclear in most cases. In this context, disease maps can be used to elucidate the mechanism of action of potential therapeutics that have been tested

in high-throughput screens and also identify potential side-effects or adverse reactions that cannot be detected during the screening process.

For this purpose, this project aimed to create a MINERVA plugin to link disease maps and application-specific databases for drug candidates from the literature and high-throughput screens. It provides the means to highlight targets of promising compounds, but also to search for compounds that target a specific protein in the disease map. As a use case, we applied the plugin to link our compound database CandActCFTR (https://candactcftr.ams.med.uni-goettingen.de/; last accessed 23 October 2021) and our CFTR Lifecycle Map [4], which aim at supporting drug discovery and data structuring in cystic fibrosis research.

The MINERVA platform [3] comes with its own drug and chemical searches, which can be found directly on a tab in the user panel. MINERVA's own drug and chemical searches provide an interface to the DrugBank [5], ChEMBL [6] and the Comparative Toxicogenomics Database (CTD) [7]. Here, the user can search for drugs or chemicals, and different relevant databases will be queried for known targets in the map, which are then highlighted by a pin and displayed in the panel. The drug and chemical searches are separated from each other and use differently shaped pins for differentiation between different compounds. Whereas drugs can be searched by name or brand name and the databases DrugBank [5] and ChEMBL [6] will be queried, chemicals can be searched via their synonyms, querying the CTD [7]. The chemical search only displays interactions with direct evidence for the respective disease. Potential drugs that have not been proven confirmed in this particular disease context are not shown, which restricts the plugin's use for, for example, drug-repurposing applications.

Furthermore, in 2019, the MINERVA *Drug reactions* plugin [8] was developed, which is based on MINERVA's own drug search and aims at exploring adverse reactions of drugs that are interacting with entities in a given disease map. It connects to an external data file [9] source and uses MINERVA's drug search to find the targets of any of the drugs in the database map. The results are displayed as a table in the panel, which shows the drug, the entity with which it interacts and any known drug reactions, such as adverse reactions, warnings and precautions. Additionally, all the targets are highlighted by a pin in the map.

While the available plugins described above allow querying for targets of specific drugs and chemicals, they do not support the reverse case, i.e., the search for drugs or chemicals that interact with a specific target in the map. Moreover, our plugin was developed to integrate data from custom databases of candidate therapeutics. This makes it possible for researchers to integrate their own experimental data or compound collections from different sources for a specific use case. One advantage of using custom databases instead of or on top of generic databases such as PubChem [10] is that all the compounds tested for a specific disease are bundled in one place and information is stored and presented in a way and format that are convenient for its users. Furthermore, if compounds were synthesized specifically for testing in a study, they often cannot be found in generic databases.

Mapping custom collections of candidate compounds to disease maps can provide a means for elucidating drug mechanisms. This is especially important in light of high-throughput drug screens, where hundreds or thousands of compounds are tested in certain settings, but the mechanisms of action of promising lead substances often remain unclear. It can, therefore, be useful to display and query whole custom lists of chemicals that have been tested with regard to a specific disease context in a disease map in MINERVA. The current MINERVA plugins do not allow for the query of large custom datasets, which would be a valuable application for making MINERVA more utilizable for the interpretation of one's own experimental data from high-throughput screens. Furthermore, it can also be useful to search for compounds that affect a certain target in the disease map, a functionality not yet supported by the available MINERVA plugins. We previously developed the generic IT solution CandActBase [11] for the collection and organization of data on drug candidates tested in a certain context (manuscript submitted for publication). The MINERVA-CandActBase plugin shown here offers a linkage between these

application-specific databases of potential drug candidates and pathway data encoded as disease maps on the MINERVA platform [3]. For this purpose, we extended the data from the CandActBase database with data on compound–gene interactions from the ChEMBL database [6] and the Comparative Toxicogenomics Database (CTD) [7].

2. Materials and Methods

2.1. Implementation

The prototype of our tool was implemented as a MINERVA plugin that runs on the same tomcat instance as the MINVERVA web app [3]. From there, it provides an interface to the databases via their web APIs [6,7,11].

The concept and data flow of the plugin are shown in Figure 1. In order to identify interactions between the compounds in the database and genes and gene products in the disease map, in a first step, the properties of each compound in CandActBase were extracted from PubChem [10] via the PubChem CID. These data were then filtered to obtain the names of the targets connected to the respective compounds from the ChEMBL Protein Target Tree within PubChem. These target names were then searched directly in the ChEMBL database, where the official names of corresponding genes for searched targets can be found. Secondly, the CTD [7] was called via its API to collect more compound–gene interactions. To identify the compounds from the CandActBase correctly, the CAS registry number from the already downloaded PubChem data was used as a query. With the CAS number, we could then call the CTD API and extract chemical–gene interactions for each compound. The resulting data from both databases were then restructured, so they can be compared and filtered for overlaps with the given disease map. The data are stored in the easy-to-use and storage-saving JSON format directly on the hosting server to ensure fast loading times. The first JSON file consists of the compound identifier used in the CandActBase and the name of the genes targeted by each compound. Conversely, the second JSON file includes every gene or protein name from the graph and a list of compound IDs that target the given gene or protein.

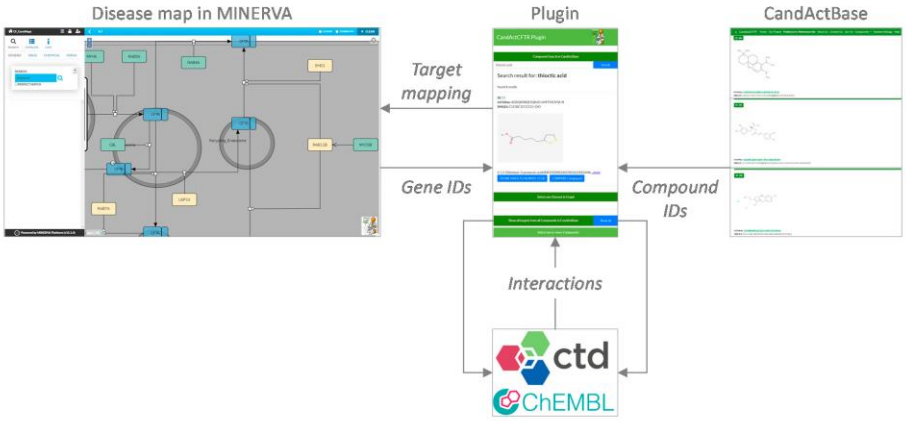
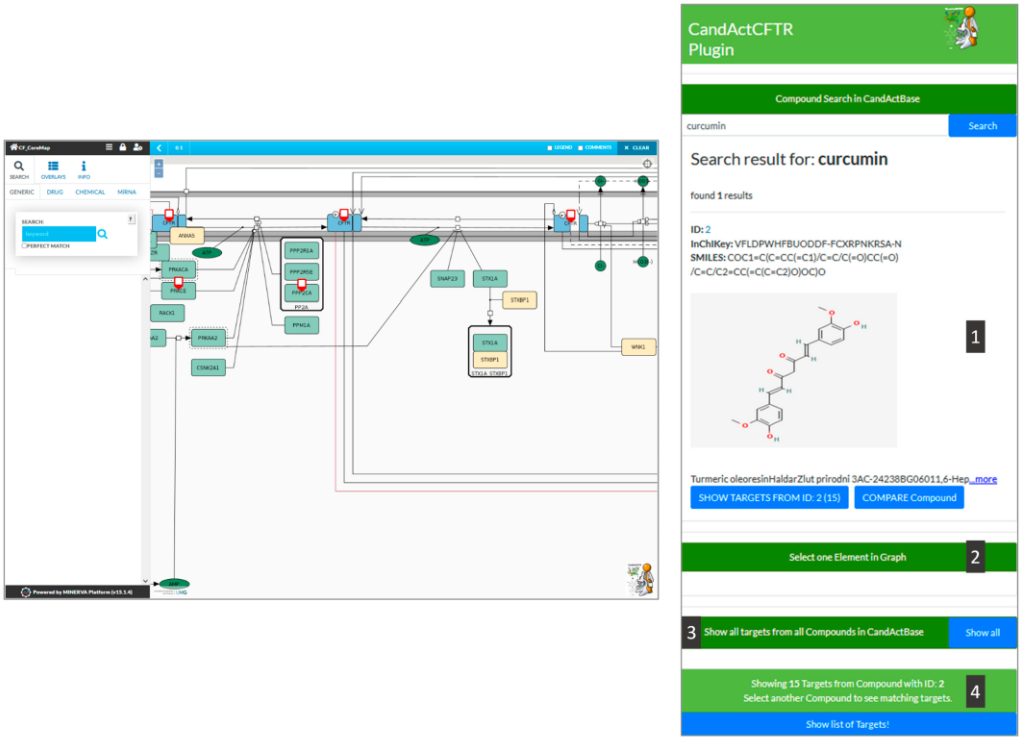


Figure 1. Concept and data flow of the plugin. Gene IDs and Chemical IDs are extracted from the disease map on MINERVA [3] (© Université du Luxembourg, 2021) and the CandActCFTR database, respectively. Using these unique identifiers, gene-chemical interactions are retrieved from the databases ChEMBL [6] (© EMBL-EBI 2018) and CTD [7] (© 2012–2021 NC State University) via the plugin. The results are then listed in the plugin and displayed in the disease map on MINERVA.

The code structure of the plugin is based on the existing MINERVA plugin starter kit (https://git-r3lab.uni.lu/minerva/plugins/starter-kit; last accessed 18 October 2021) [8].

We adjusted the existing methods to fit our purposes, as well as writing new methods to expand the functionality.

We designed our graphical user interface (GUI) to be well structured and visually matching with the color scheme of the CandActCFTR website (https://candactcftr.ams.med.uni-goettingen.de/; last accessed 23 October 2021) as shown in Figure 2 with an exemplary search for the compound "Curcumin" and the identified targets in the graph.



(a) (b)

Figure 2. Screenshot from the MINERVA instance showing (**a**) a disease map and highlighted targets, and (**b**) a zoomed-in feature of the UI of the CandActBase plugin. The plugin interface allows the use of all functions from one window: 1. display results from compound query (ID, InChIKey, SMILES Code, chemical structure and synonyms), 2. reverse search by target from disease map, 3. button to highlight all targets of all database compounds in the disease map and 4. button to show targets of selected compound as list.

2.2. Exemplary Queries

Figure 2 shows the graphical user interface of the plugin (Figure 2b) integrated into the CFTR-disease map hosted on MINERVA (Figure 2a). In the following, the functionalities of the plugin are explained via different queries.

2.2.1. Compound Search

As an example, we searched for the compound "Curcumin", which is typed into the search bar at the top of the plugin on the right side of the window. The connected database, in this case, CandActCFTR, is then queried for the search term. If a match is

found, the search results are then displayed in the box below the search bar, where the ID, the InChIKey, the SMILES Code, the chemical structure and synonyms of the relevant compounds are shown (1). By clicking the "Show Targets" button below a search result, the targets of the desired compound can be highlighted in the disease map (A). The currently selected compounds can be seen in the status bar at the bottom of the plugin interface, which also displays the number of targets highlighted in the disease map.

2.2.2. Compare Two Compounds

To search for common targets of two compounds, the button "Compare Compound" under the initial compound in question can be clicked; a second compound can then be selected, and the common targets are highlighted in the disease map. Furthermore, a list of all the highlighted targets in the graph can be called by clicking the "Show list of targets" button (4).

2.2.3. Reverse Target Search

In the reverse search, any entity in the disease map can be used by selecting a target in the disease map and clicking on the "Select one element in graph" button in the lower half of the plugin interface (2). As a result, all compounds found to have an interaction with the chosen entity are listed in the plugin interface on the side in a style similar to the synonym search results (1).

2.2.4. Show Database-Disease Map Coverage

In order to show the database coverage on the disease map, all targets from all compounds in the CandActBase database can be highlighted in the graph (3).

The plugin application connecting CandActCFTR and the disease maps is available at https://cf-map.uni-goettingen.de (last accessed 23 October 2021); an independent version can be downloaded from GitLab (Supplementary) and used in any MINERVA instance.

3. Results

We developed a MINEVRA plugin to link disease maps and application-specific compound databases, using the CandActCFTR database designed for cystic fibrosis as a use case. Our plugin offers three main functionalities:

The show all function allows the user to highlight all the genes and gene products in the disease map that are targeted by one or more chemicals in the custom database. This makes it possible to see which compartments and pathways of the disease map are already covered by potential drugs.

The compound search allows the user to search a specific compound in the compound database by name and synonym, and will be extended to also search by unique identifiers such as SMILES, InChIKey or PubChem CID. The database is then queried for matching compounds, and, upon selection, the targets deposited in the database for the individual compound are highlighted in the map and displayed in the plugin panel. Additionally, the database can be queried for two compounds, and the entities targeted by both compounds are highlighted and listed, which allows the user to identify similarities by inspecting the target overlaps and thus compare the sites of action for different chemicals.

Furthermore, the target search offers the reverse of the compound search, where the user can select a gene or protein from the displayed disease map and the compound database is searched for all the compounds that target this specific entity. This allows users to explore specific pathways and targets in the disease map with regard to potential therapeutics.

4. Discussion

In recent years, great efforts have gone into developing systems biology models detailing the molecular, cellular and physiological interactions for different diseases. Not only do disease maps represent current knowledge in a comprehensive manner, but they

can also be used to identify drug targets, propose potential drugs, elucidate the mechanisms of action for active compounds and detect possible adverse effects. In order to support these endeavors, we have created a plugin linking disease maps displayed on the well-established MINERVA platform and application-specific compound databases, such as ones based on the CandActBase database solution.

The main advantage of using custom compound databases over generic ones is that the respective data structure and information on each compound can be tailored to the users and the use case in question. Furthermore, all the compounds are available from one place, and it is guaranteed that all the compounds have an entry, which is not always the case in generic databases, for example, due to newly synthesized compounds.

Using cystic fibrosis as a use case, we, here, linked our CF disease map and CandActCFTR, our database for compounds tested as cystic fibrosis therapeutics. The plugin uses gene–chemical interactions parsed from ChEMBL and the CTD to map compounds from the database to the disease map, which is displayed in MINERVA. We implemented three main functions: the first to show all the targets of all the compounds in the database, the second to retrieve the targets of a specific compound, and the third to retrieve all the compounds interacting with a specific target.

In order to ensure fast and responsive loading times, even when handling chemical–gene interactions from multiple queries, we decided to download the interaction data from the databases beforehand. Having the interaction data in local files instead of parsing them directly from the databases also has the advantage that one's own interaction data can be used or added. The downside of such a system is the lack of recentness and completeness. This could be overcome by creating a refreshing script that downloads, formats and updates changes in entries received from the called databases. Thus, integrating the goal of fast responsiveness in the system, as well as integrating the latest data, would be achieved.

For our use case, only entirely freely accessible databases were used to adhere to the project's open-source notion. Here, ChEMBL [6] and CTD [7] were used as example databases. Data from other gene–compound interaction databases or own experiments can readily be included for other use cases. Our plugin is easily adaptable to other databases, if their data can be transformed to the common JSON data format. Furthermore, compounds as well as targets should be stored using a common unique identifier, such as the official gene symbols for targets and the InChIKey for compounds.

The lack of consistent identifiers and the huge variety in the data structure of different databases were the biggest hurdles in this project. There is no standardized data format on which to rely. Even basic information, such as common and unique identifiers for chemicals, differs across databases. This is naturally due to the variable use cases of databases, but it would simplify the development of cross-database applications and interdisciplinary research endeavors considerably if standardized structures and identifiers existed.

5. Conclusions

We were able to visualize chemical–gene interactions on the MINERVA platform and built a plugin that serves as a template for connecting new, customized database sources to be used in this interactive visualization solution. We thus extended the functionality of the MINERVA platform and showed what a customized plugin could look like.

It is now also possible for clinicians and biologists to view and compare chemicals from the CandActCFTR project with respect to the genes or proteins with which they interact. Additionally, possible side effects resulting from mutual targets of drug combinations can be taken into account based on previous research. The tool can also be applied to drug-repurposing efforts by searching for possible targets of known drugs in different disease maps and, conversely, by searching for drugs through specific targets in a disease map.

The tool can be applied in other disease contexts and to other disease maps. The code and our example data are available on GitLab with instructions on how to implement it for new disease maps (https://s.gwdg.de/SU0rqd, accessed on 23 October 2021).

Supplementary Materials: The code for the MINERVA plugin as well as installation instructions can be found on GitLab: https://s.gwdg.de/SU0rqd.

Author Contributions: Conceptualization, L.V. and M.M.N.; methodology, L.V. and M.V.; software, M.V.; writing—original draft preparation, L.V.; writing—review and editing, M.V. and M.M.N.; visualization, L.V.; supervision, M.M.N.; funding acquisition, M.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by the Deutsche Forschungsgemeinschaft DFG, grant number 315063128. We acknowledge support by the Open Access Publication Funds of the Göttingen University.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: https://candactcftr.ams.med.uni-goettingen.de/, https://www.ebi.ac.uk/chembl/, http://ctdbase.org/, accessed on 23 October 2021.

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

References

1. Mazein, A.; Ostaszewski, M.; Kuperstein, I.; Watterson, S.; Le Novère, N.; Lefaudeux, D.; De Meulder, B.; Pellet, J.; Balaur, I.; Saqi, M.; et al. Systems medicine disease maps: Community-driven comprehensive representation of disease mechanisms. *NPJ Syst. Biol. Appl.* **2018**, *4*, 1–10. [CrossRef] [PubMed]
2. Ostaszewski, M.; Gebel, S.; Kuperstein, I.; Mazein, A.; Zinovyev, A.; Dogrusoz, U.; Hasenauer, J.; Fleming, R.M.T.; Le Novère, N.; Gawron, P.; et al. Community-driven roadmap for integrated disease maps. *Brief. Bioinform.* **2019**, *20*, 659–670. [CrossRef] [PubMed]
3. Gawron, P.; Ostaszewski, M.; Satagopam, V.; Gebel, S.; Mazein, A.; Kuzma, M.; Zorzan, S.; McGee, F.; Otjacques, B.; Balling, R.; et al. MINERVA—A platform for visualization and curation of molecular interaction networks. *NPJ Syst. Biol. Appl.* **2016**, *2*, 1–6. [CrossRef]
4. Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations. *Int. J. Mol. Sci.* **2021**, *22*, 7590. [CrossRef] [PubMed]
5. Wishart, D.S. DrugBank: A comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res.* **2006**, *34*, D668–D672. [CrossRef] [PubMed]
6. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A.P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L.J.; Cibrián-Uhalte, E.; et al. The ChEMBL database in 2017. *Nucleic Acids Res.* **2017**, *45*, D945–D954. [CrossRef]
7. Davis, A.P.; Grondin, C.J.; Johnson, R.J.; Sciaky, D.; Wiegers, J.; Wiegers, T.C.; Mattingly, C.J. Comparative Toxicogenomics Database (CTD): Update 2021. *Nucleic Acids Res.* **2021**, *49*, D1138–D1143. [CrossRef] [PubMed]
8. Hoksza, D.; Gawron, P.; Ostaszewski, M.; Smula, E.; Schneider, R. MINERVA API and plugins: Opening molecular network analysis and visualization to the community. *Bioinformatics* **2019**, *35*, 4496–4498. [CrossRef] [PubMed]
9. Demner-Fushman, D.; Shooshan, S.E.; Rodriguez, L.; Aronson, A.R.; Lang, F.; Rogers, W.; Roberts, K.; Tonning, J. A dataset of 200 structured product labels annotated for adverse drug reactions. *Sci. Data* **2018**, *5*, 180001. [CrossRef] [PubMed]
10. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem in 2021: New data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395. [CrossRef] [PubMed]
11. CandActBase | ToolPool Gesundheitsforschung. Available online: https://www.toolpool-gesundheitsforschung.de/produkte/candactbase (accessed on 26 March 2021).

Chapter 6 Complementary dual approach for *in silico* target identification of potential pharmaceutical compounds in Cystic Fibrosis

This manuscript has originally been published in the International Journal of Molecular Sciences

Complementary dual approach for in silico target identification of potential pharmaceutical compounds in Cystic Fibrosis

Liza Vinhoven[1], Frauke Stanke[2,3], Sylvia Hafkemeyer[4] and Manuel Manfred Nietert[1,5]

[1]Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany
[2]German Center for Lung Research (DZL), Partner Site BREATH, Hannover, Germany
[3]Clinic for Pediatric Pneumology, Allergology, and Neonatology, Hannover Medical School, Hannover, Germany
[4]Mukoviszidose Institut gGmbH, Bonn, Germany
[5]CIDAS Campus Institute Data Science, Georg-August-University, Göttingen, Germany

Authors contribution

L.V.: conceptualization; methodology, formal analysis, investigation, data curation, visualisation, writing-original draft preparation

F.S.: conceptualization; writing – review and editing; funding acquisition

S.H.: writing – review and editing

M.M.N.: conceptualization; writing – review and editing; funding acquisition

International Journal of *Molecular Sciences*

MDPI

Article

Complementary Dual Approach for In Silico Target Identification of Potential Pharmaceutical Compounds in Cystic Fibrosis

Liza Vinhoven [1], Frauke Stanke [2,3], Sylvia Hafkemeyer [4] and Manuel Manfred Nietert [1,5,*]

[1] Department of Medical Bioinformatics, University Medical Center Göttingen, Goldschmidtstraße 1, 37077 Göttingen, Germany
[2] Clinic for Pediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
[3] Biomedical Research in Endstage and Obstructive Lung Disease Hannover (BREATH), German Center for Lung Research, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany
[4] Mukoviszidose Institut, In den Dauen 6, 53117 Bonn, Germany
[5] CIDAS Campus Institute Data Science, Goldschmidtstraße 1, 37077 Göttingen, Germany
* Correspondence: manuel.nietert@med.uni-goettingen.de; Tel.: +49-551-39-14920

Abstract: Cystic fibrosis is a genetic disease caused by mutation of the CFTR gene, which encodes a chloride and bicarbonate transporter in epithelial cells. Due to the vast range of geno- and phenotypes, it is difficult to find causative treatments; however, small-molecule therapeutics have been clinically approved in the last decade. Still, the search for novel therapeutics is ongoing, and thousands of compounds are being tested in different assays, often leaving their mechanism of action unknown. Here, we bring together a CFTR-specific compound database (CandActCFTR) and systems biology model (CFTR Lifecycle Map) to identify the targets of the most promising compounds. We use a dual inverse screening approach, where we employ target- and ligand-based methods to suggest targets of 309 active compounds in the database amongst 90 protein targets from the systems biology model. Overall, we identified 1038 potential target–compound pairings and were able to suggest targets for all 309 active compounds in the database.

Keywords: cystic fibrosis; docking; ligand-based drug design; target-based drug design/target identification; virtual screening

check for updates

Citation: Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. Complementary Dual Approach for In Silico Target Identification of Potential Pharmaceutical Compounds in Cystic Fibrosis. *Int. J. Mol. Sci.* **2022**, *23*, 12351. https://doi.org/10.3390/ijms232012351

Academic Editors: Gabriella Guerrini and Maria P. Giovannoni

Received: 5 September 2022
Accepted: 12 October 2022
Published: 15 October 2022

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

1. Introduction

Cystic fibrosis (CF) is one of the most common genetic diseases prevalent among the population of Caucasian ancestry, where it affects approximately 1 in 3000 newborns [1–3]. It is caused by mutations of the cystic fibrosis transmembrane conductance regulator (CFTR) gene [4], which encodes a membrane protein that serves as a chloride and bicarbonate channel in the exocrine epithelia of various organs, thereby regulating the viscosity of the mucus lining [5]. Defective CFTR, therefore, has severe implications throughout the body, its major hallmarks being recurrent pulmonary infections and pancreatic insufficiency [4,5]. Of the currently known 2100 mutations of the CFTR gene, several hundred have been shown to be disease causing [6–8]. These mutations cause disturbances throughout CFTR's intricate and delicately balanced biogenesis, which, even in its wild-type (wt) form, only converts 20–40% of the transcripts into a fully functional protein [9]. In order to make handling and working with the multitude of mutations easier, they are categorized into different classes, depending on what kind of defect they cause. Originally, four major mutation classes were proposed, and over the years, this has been expanded to seven classes (I: no CFTR synthesis; II: CFTR trafficking defect; III: CFTR dysregulation; IV: defective gating; V: reduced CFTR transcription; VI: less stable protein; and VII: no CFTR mRNA) [10–14]; however, as many mutations cause multiple defects, an expanded, combinatorial classification system

was proposed [15]. Here, mutations are assigned to groups of all possible combinations of the single-defect mutation classes, resulting in a more comprehensive grouping. In addition to the multi-defect mutations, patients often carry more than one mutation. These factors lead to a vast range of geno- and phenotypes, which makes the development of effective causative therapeutics especially challenging. In recent years, different small-molecule therapeutics have been developed for clinical applications, which improve the CFTR function by directly targeting the CFTR protein and not just alleviating symptoms of CF-patients. Currently, four pharmaceutical drugs, different combinations of four compounds, are approved and available as causative therapy to some CF patients [16–19]. The active compound in the first approved drug, Kalydeco, is Ivacaftor, a CFTR potentiator which is approved mainly for gating mutations [20,21]. The other drugs are combination therapies [18,19], i.e., they contain two or more active compounds and thereby target multiple defects. Orkambi contains Ivacaftor and Lumcaftor, a CFTR corrector, which acts as small-molecule chaperone to correct the folding defect of some class II mutations [22,23], and Symdeco, which contains Ivacaftor and Tezacaftor, an alternative CFTR corrector [24]. The most recent drug, Kaftrio (known as Trikafta in the US), is a triple combination of Ivacaftor, Tezacaftor and Elaxacaftor, which also acts as CFTR corrector [25,26]. Still, for about 10% of patients, especially those with rare mutations, there is no causative medication available [27]. In an effort to find effective treatments for all patients, the search for CFTR modulators, especially synergistic compound combinations, is ongoing [28,29]. Thousands of compounds are being tested in different cell-assays, often in high-throughput screens, leading to large amounts of data and various candidate substances [20,30–41]. In order to structure and collect the compounds tested as CFTR modulators, we previously developed the publicly available database CandActCFTR (candactcftr.ams.med.uni-goettingen.de), where compounds are annotated and categorized according to various characteristics, including their mode of action and order of interaction with CFTR [42,43]. When analysing the data, it becomes apparent that for about 70% of the active compounds, it is unknown whether they affect CFTR directly through physical interaction, or indirectly through its interactome. To support the elucidation of their mode of action, we previously developed the CFTR Lifecycle Map (cf-map.uni-goettingen.de), where we used a systems biology approach to create a human- and machine-readable model of the CFTR maturation pathway in cells [44,45]. The CFTR Lifecycle Map is written in the standardized SBGN Process Description format and comprises detailed representations of the molecular interactions and pathways CFTR undergoes during its entire biogenesis. It contains 156 reactions with 262 different molecular entities, including 170 biomacromolecules, mainly proteins. Here, we now connect the two resources in order to shed light on the mechanism of action of active compounds by identifying their targets. For this purpose, we are using and combining two different reverse screening approaches. Traditionally, in drug discovery, virtual screening approaches are applied to find bioactive compounds that bind to a specific target protein. In reverse screening, the opposite approach is employed to identify the target proteins of active compounds [46,47]. This is becoming increasingly important in order to predict drug side effects and for drug repositioning, where existing drugs are repurposed for other disorders. Several computational methods exist for reverse screening, which, similarly to traditional virtual screening, can generally be divided into three categories [47]. The first class is target-based approaches, mainly docking, which requires high-quality protein structures and has high computational costs. The two other classes are ligand based, either on their shape or their pharmacological features. These require a solid data foundation of known ligand–target interactions as reference databases. In the last decade, reverse screening approaches have been applied to a range of use cases, especially to find targets of natural compounds [46,47]. For example, a molecular shape-based method was employed to identify the cyclin dependent kinase (CDK2) as target of the phytochemical curcumin as possible explanation for its cancer-preventive properties [48]. Furthermore, inverse docking, i.e., a target-based approach, has been used to, amongst others, study different helicases in Zika viruses as targets of ligands from a flowering plant [49], inves-

tigate the effect of thyme derived thymol on fat deposition [50], and shed light on the antitumor targets of a library of natural bioactive compounds [51]. While most studies use either one or the other approach, we here use both a target- and ligand-based approach independently and combined to identify possible targets in the CFTR Lifecycle Map of the active compounds from CandActCFTR. By using the two approaches, we were able to include more potential protein targets than by exclusively using docking or shape-based approaches. Ultimately, we tested 309 active compounds against 90 targets, resulting in almost 30,000 possible combinations. Of these, 1038 unique target–compound pairings were identified with graduated confidence levels in the range of 1–5. Importantly, we could suggest at least one target for each active compound in the database, thereby bringing the elucidation of their mechanism of action one step closer and serving as a basis for finding novel compounds (classes) and predicting synergistic compound combinations. This application shows how systems medicine disease maps, such as the CFTR Lifecycle Map, can be utilized for in silico target identification and to provide the means to fill knowledge gaps and support drug design.

2. Results

2.1. PDB Targets

In order to suggest possible mechanisms of actions of active compounds in the CandActCFTR database [43], possible protein targets that are involved in CFTR biogenesis were selected from the CFTR Lifecycle Map [45]. For this purpose, the Protein Data Bank [52] was queried for experimental X-ray or CryoEM structures of all proteins within the CFTR Lifecycle Map. The list was then filtered according to different criteria such as structure completeness, resolution and experimental method. This list was narrowed down to 35 PDB structures, including that of wt-CFTR, which are listed in Supplementary Table S1, together with their experimental specifications and their role in the CFTR biogenesis. Overall, for 35 of the 170 proteins present in the CFTR Lifecycle Map, appropriate PDB structures could be found, which cover a diverse range of functions and stages in the CFTR biogenesis. This will be especially important to develop combination therapies with synergistic effects that influence CFTR at different stages of its lifecycle. However, for some steps in the CFTR lifecycle, specifically at the transcription stage, no docking-appropriate PDB structures of CFTR interactors could be found. To remedy this situation, structure predictions from the AlphaFold database (https://alphafold.ebi.ac.uk/, accessed on 9 September 2022) [53–55] were considered as alternatives. Unfortunately, when comparing the results of the blind cross-docking between a set of reference protein structures with co-crystallized ligands and their predicted counterparts, the predicted structures resulted in much less accurate docking results. Since the aim of this study was to find potential targets for our query ligands, rather than of putative ligands for specific proteins, the inaccurate docking results from predicted structures could potentially bias the overall results. It was therefore decided not to use the predicted structures from the AlphaFold database and continue with the subset of 35 experimental structures from the PDB database. Nonetheless, except for the transcription stage, the 35 targets are well distributed across the CFTR Lifecycle Map. More precisely, in addition to CFTR itself, 9 potential targets are involved in translation, folding and ER quality control, 7 are associated with the secretory pathway, 5 with endocytosis, and 13 are involved in CFTR activity.

2.2. Docking

All active compounds from the CandActCFTR [42,43] database were docked against all 35 targets using the Smina [56] and QuickVina-W (qvina-W) [57] docking programs. After docking, two methods (2.1) were used for post-processing. The first method [51] (method I) filters out false positives and ultimately results in a list of the overall best-scoring target–ligand pairings, while the second method [58] (method II) also filters out false positives but then calculates the most likely target for each ligand. A comparison of the results from all four approaches can be seen in Figure 1A. Method I resulted in a list of

21 target–ligand pairings for the docking results from the qvina-w docking program. Of the 48 high-scoring pairings, 19 were also among the ones identified by method II.

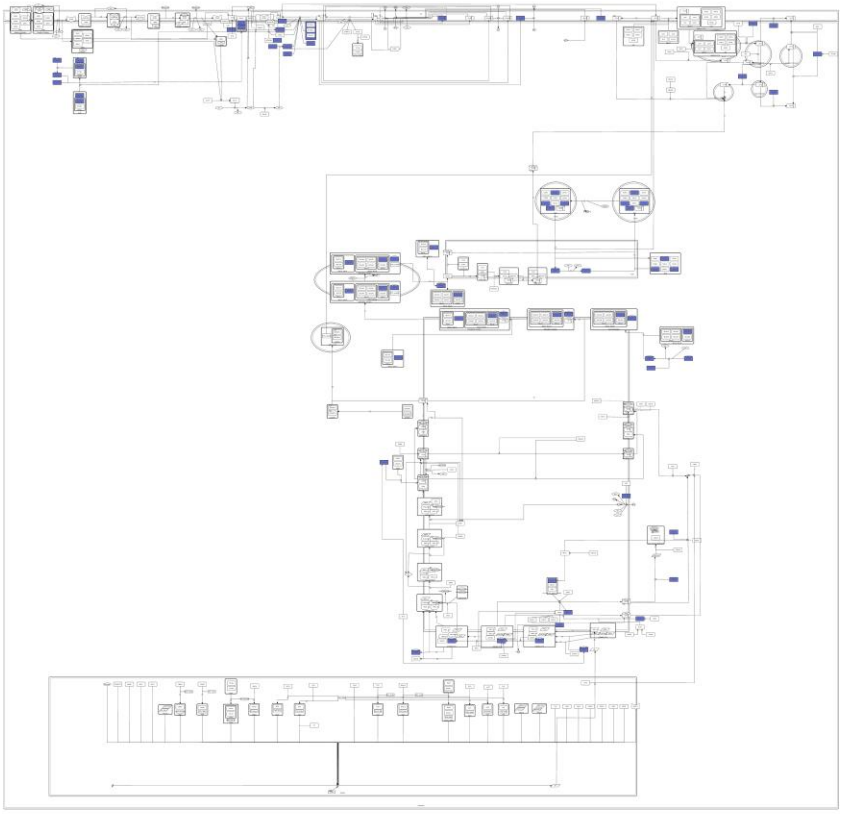


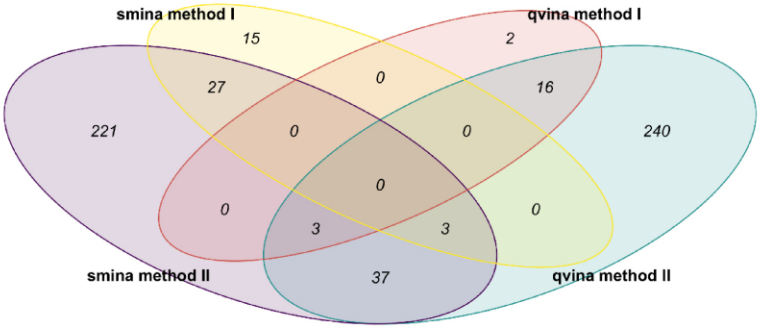
Figure 1. Available protein structures of targets in the CFTR Lifecycle Map. Highlighted in blue are all proteins of which an appropriate PDB structure could be found and which were used as targets in the target-based approach.

Using the results from the Smina docking program, 45 target–ligand pairings were identified using the method I, 30 of which were also identified by the second approach.

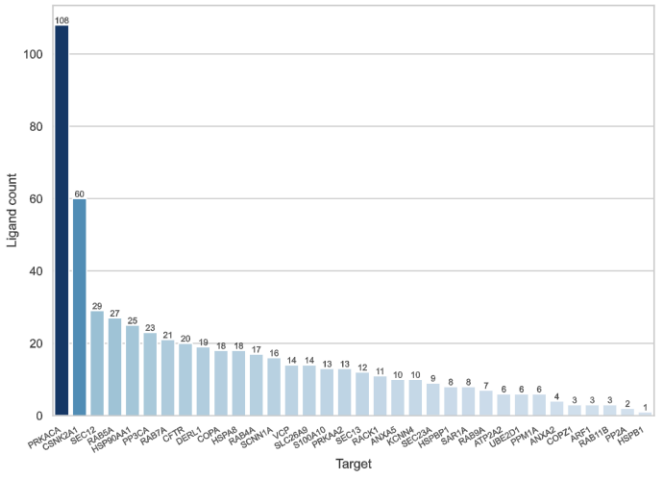
When comparing the results from the two docking programs, method II identified 43 common target–ligand pairings, while there were no common pairings found by both method I approaches.

Figure 1 shows the number of ligands attributed to the targets by all methods combined. On average, the mean Tanimoto similarity amongst ligands associated with the same target was 0.25 on average, indicating that there is no bias towards a specific compound class for one target. As can be seen, PRKACA (the catalytic subunit α of protein kinase A) (108 ligands) and CSNK2A1 (Casein kinase II subunit α) (60 ligands) have a lot of compounds ascribed to them by at least one approach. As both analysis methods take into account average scores for each ligand over all targets, the docking results for PRKACA and CSNK2A1 were removed from the data, and the entire analysis of the docking results was repeated, in order to eliminate bias from these two targets.

Figure 2B shows that the number of ligands identified per target are much more evenly distributed after removing the PRKACA and CSNK2A1 from the docking data. The target with the most ligands associated with it here is RAB5A (Ras-related protein Rab-5A) (47 ligands). On average, the mean Tanimoto similarity amongst ligands associated with the same target was 0.21, so they appear to be structurally different.



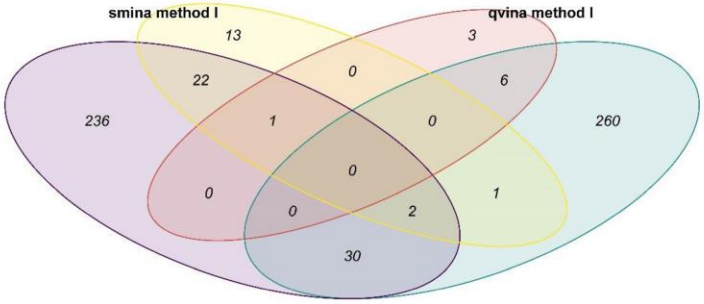
(**A**)



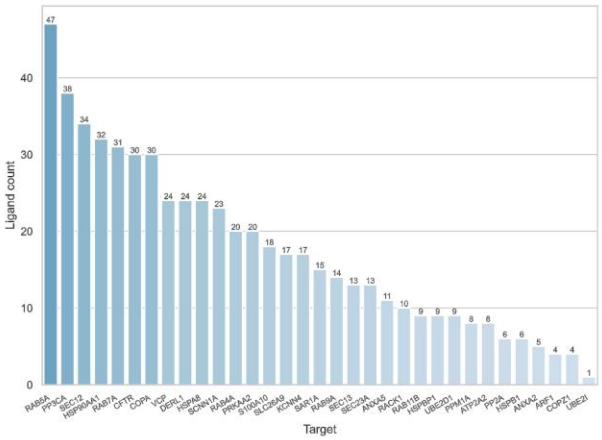
(**B**)

Figure 2. Results of the target-based approach using all targets. (**A**): Venn diagram representing the number of target–ligand pairings identified by both methods with both docking programs and the overlap in results. The yellow ellipse shows the pairings identified by method I using the docking results by Smina, the red ellipse shows the results of method I using qvina-w docking results. Shown in purple and teal are the pairings identified by method II using the Smina and qvina-w docking results, respectively. (**B**): The number of ligands associated with each target by all four approaches combined.

The number of targets per ligand of all pairings identified by method I (including and excluding the results with PRKACA and CSNK2A1) can be seen in Figure 3. As can be seen, the number of potential targets for each ligand is relatively low, indicating specific interactions. Most ligands are only associated with one or two targets, one is associated with three and four targets, respectively, two to five targets, and only one ligand is associated with six targets. The ligand with the most targets associated (ligand 2144) is the deoxyribose dATP, which is predicted to bind to the Ras-related proteins RAB4A, RAB5A, RAB7A, RAB9A, and RAB11B and the secretion-associated Ras-related GTPase 1A (SAR1A), all of which are GTP-binding proteins.



(**A**)



(**B**)

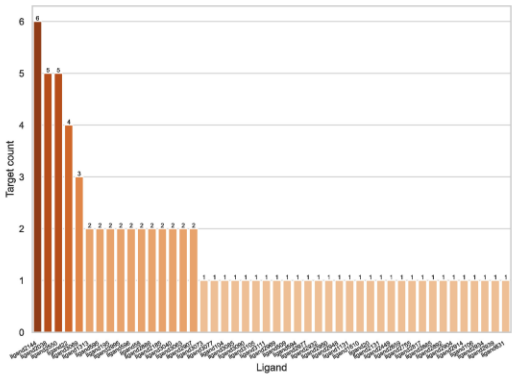
Figure 3. Results of the target-based approach using excluding PRKACA and CSNK2A1. (**A**): Venn diagram representing the number of target–ligand pairings identified by both methods with both docking programs and the overlap in results. The yellow ellipse shows the pairings identified by method I using the docking results by Smina, the red ellipse shows the results of method I using qvina-w docking results. Shown in purple and teal are the pairings identified by method II using the Smina and qvina-w docking results, respectively. (**B**): The number of ligands associated with each target by all four approaches combined.

Figure 3B shows the Protein–Ligand-Interaction network for all pairings identified by method I (including and excluding the results with PRKACA and CSNK2A1). Each edge between two targets stands for a compound found for both of them, the width of the

edge representing the number of compounds and the size of the node (targets) representing its degree, i.e., how many other nodes it is connected to. Most targets have only one or two potential compounds in common, except PRKACA and CSNK2A1, which share four compounds.

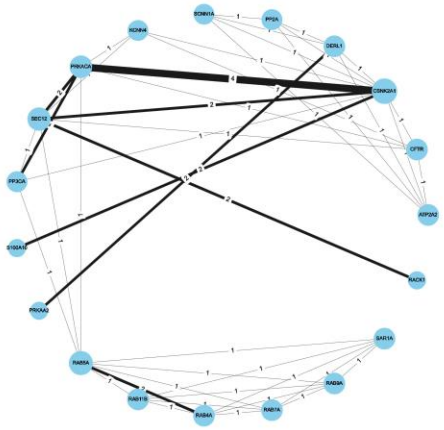
As can be seen in the Venn diagram in Figure 2A, the consensus amongst the different approaches was similar compared to when all targets were used. For the qvina-w docking program, method I led to ten target–ligand parings, six of which were also identified by method II. The results from the Smina docking program led to 39 pairings using method I, 25 of which were also found with method II. Comparing the two docking programs, method II identified 32 common pairings, and method I identified only one common pairing. Interestingly, the ligand in this pairing is Lumacaftor (InChI Key: UFSKUSARDNFIRC-UHFFFAOYSA-N), also known as VX-809, a well-known and clinically approved CF-drug, known to bind to CFTR directly. Here, however, it is predicted to bind to the Ras-related protein Rab-7a (RAB7A), a protein involved in endocytosis.

Two different binding pockets and binding poses were identified by both docking programs, Smina and qvina-w, in accordance. The binding pockets can be seen in Figure 4, where the poses predicted by Smina are coloured in blue, and the ones predicted by qvina-w are green. Binding pocket A corresponds to the GTP binding site of RAB7A. The two docking poses calculated by Smina and qvina-w, respectively, have an RMSD of 2.30 Å in pocket A and 2.24 Å in pocket B. Figure 5 shows the 2D poseview of Lumacaftor docked into both pockets by both docking programs. The black dotted lines represent hydrogen bonds, the green spline sections represent hydrophobic interactions and the green dotted lines show π-π stacking or π–cation interactions. As can be seen in Figure 5, in pocket A, for both predicted binding poses, Lumacaftor interacts with various residues via hydrogen bonds, and it undergoes hydrophobic interactions with tyrosine-37 and an π–cation interaction with lysine-126. However, while the binding pose calculated by Smina shows a π–cation interaction with the Magnesium ion in the GTP binding site, the binding pose predicted by qvina-w suggests a π-π stacking interaction with tyrosine-37. In binding pocket B, both binding poses suggest Lumacaftor undergoes hydrogen bonding with serine-72. The binding pose predicted by qvina-w suggests an additional hydrogen bond with glutamine-71, while the one predicted by Smina shows π-π stacking with tryptophane-102, as well as hydrophobic interactions. Overall, both docking programs predict a higher binding affinity for pocket A (Smina 13.4 kcal/mol, qvina-w 12.4 kcal/mol) than for pocket B (both 10.8 kcal/mol).



(**A**)

Figure 4. *Cont.*



(**B**)

Figure 4. Results of method I using both docking programs combined, including and excluding PRKACA and CSNK2A1. (**A**): Number of targets associated with each ligand. (**B**): Protein–Ligand-Interaction network. Each nodes (circle) stands for one target, while the edges (lines) connecting them represent the number of common ligands (=number on edge) associated with them by either method. The thicker the edge, the more ligands two targets share. The size of the nodes represents their degree, i.e., the number of other nodes they are connected to.

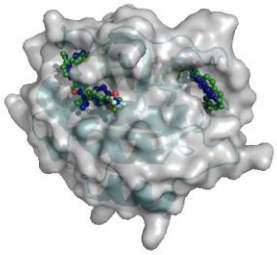


Figure 5. Docking results of Lumcaftor to RAB7A (PDB 1T91). Pocket A is shown on the right side, pocket B on the left. The docking poses of Lumacaftor predicted by qvina-w are coloured in green, the ones by Smina in blue.

Two other target–ligand pairings were identified by three of the four approaches after the exclusion of PRKACA and CSNK2A1 from the docking data. The Ras-related protein Rab-4a (RAB4A) was predicted to interact with ligand2995 (InChI Key: PSPRNONTFBJUDQ-SCFUHWHPSA-N), an ATP analogue. Figure 6A shows the ligand at the GTP binding site of RAB4A. There it is suggested to be coordinated via hydrogen bonding with leucine-155, serine-158, alanine-154 and lysine-124, as well as a π-π stacking interaction with phenylalanine-35. For this pairing, qvina-w calculated a binding affinity of -10 kcal/mol, and Smina calculated a binding affinity of -12.7 kcal/mol.

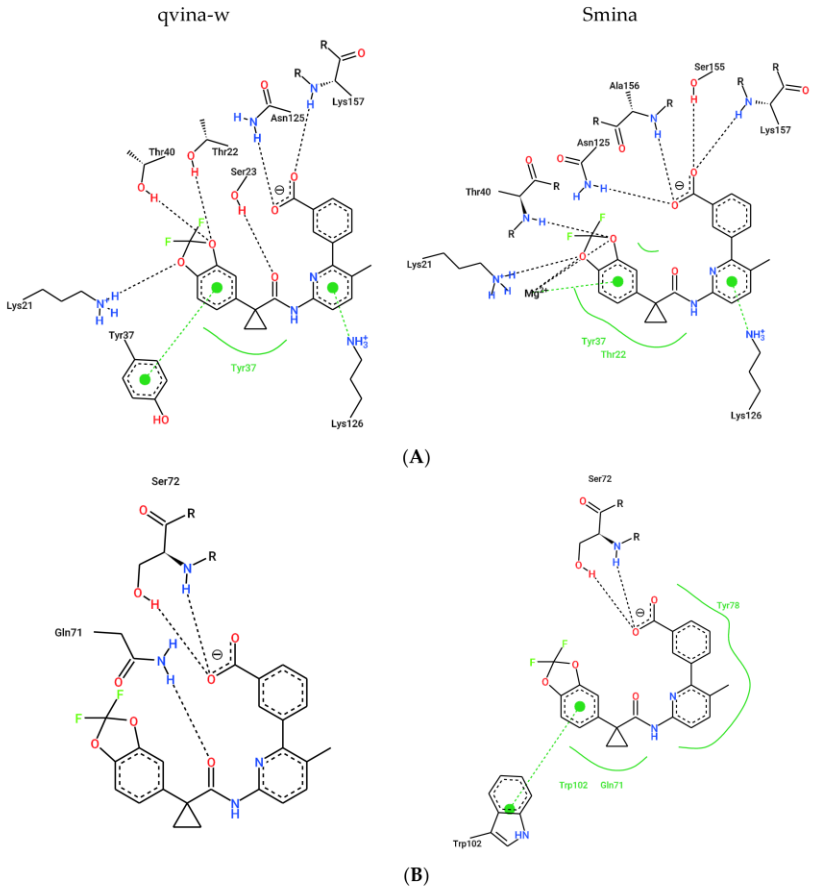


Figure 6. Two-dimensional representation of the docking poses of Lumacaftor to RAB7A and predicted interactions between them. Black dashed lines represent hydrogen bonds, green spline sections show hydrophobic interactions and green dashed lines represent π-π stacking or π−cation interactions. (**A**) Docking poses calculated by qvina-w and Smina for binding pocket A. (**B**) Docking poses calculated by qvina-w and Smina for binding pocket B.

Furthermore, Derlin-1 (DERL1), also known as degradation in endoplasmic reticulum protein 1, was predicted to interact with Equol (ligand104, InChIKey: ADFCQWZHKCXPAJ-GFCCVEGCSA-N). Figure 6B shows the predicted binding pocket of Equol in DERL1, where it is predicted to undergo π-π stacking interactions with tryptophane-106 and 18. Here, qvina-w calculated a binding affinity of −8.9 kcal/mol, while Smina calculated a binding affinity of −11.1 kcal/mol.

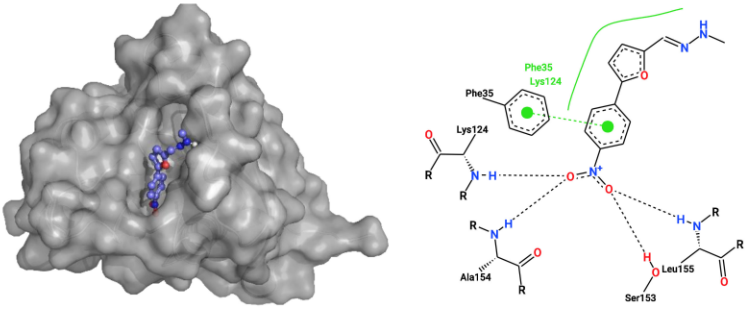
2.3. Ligand Similarity Approach

For the ligand-based similarity approach, known ligands for the 35 selected targets were collected from the databases BindingDB [59] (https://www.bindingdb.org, accessed on 9 September 2022) and ChEMBL (https://www.ebi.ac.uk/chembl/, accessed on 9 September 2022) [60–62]. Overall, 6787 target–ligand interactions could be found in the

BindingDB and 8536 were collected from ChEMBL. When merging the two datasets, 3631 of target–ligand interactions could be found in both databases, resulting in a combined dataset of 11,692 unique interactions. From ChEMBL, ligands could be found for 22 targets, of which 15 also had ligands listed in the BindingDB. The targets and the number of ligands per target from each database are listed in Supplementary Table S2. The most ligands could be found for the chaperone HSP90AA1 (3112 ligands), followed by the protein kinases PRKACA (1730 ligands), CSNK2A1 (1710 ligands) and PRKAA2 (1551 ligands), which are well-known pharmacological targets. For CFTR, 1012 ligands could be found, while less than 1000 compounds were found for the remaining targets. Hence, ligand-based similarity comparisons were conducted for 22 of the 35 targets.

For this purpose, the molecular fingerprints of the reference compounds and the query compounds were calculated and compared pairwise to each other by calculating the Tanimoto similarity. Using a similarity cut-off of 0.75, similar compounds between reference and query ligands could be found for eight targets (Supplementary Table S3). Figure 7A shows that the most compounds were found for CFTR (52 compounds), followed by the catalytic subunit of the Serine/threonine-protein phosphatase 2B (PPP3CA; 34 compounds), and the kinases PRKACA (11 compounds), PRKAA2 (11 compounds) and CSNK2A1 (10 compounds). Two compounds could be found for each of the chaperones HSPB1 (also known as Hsp27) and HSP90AA1, and one for ATPase 2 (ATP2A1). Of the 52 compounds identified for CFTR, 13 were previously known to interact with CFTR directly. Reversely, not all compounds reported as directly interacting with CFTR in the CandActCFTR database were also identified via ligand similarity, showing that the reference ligands from ChEMBL and the BindingDB do not exhaustively include all known ligands.

Figure 7B shows the Protein–Ligand-Interaction network of the results. As can be seen, the kinases PRKACA, PRKAA2 and CSNK2A1 share the most compounds with each other. All 11 compounds associated with them are shared between PRKACA and PRKAA2, 9 of which they also share with CSNK2A1. All three kinases share the same seven compounds with CFTR and the same two with HSPB1. PRKACA and PRKAA2 additionally share two identical compounds with PPP3CA, which also shares one compound with each CFTR and HSP90AA1, who share another compound amongst them.



(A) (B)

Figure 7. *Cont.*

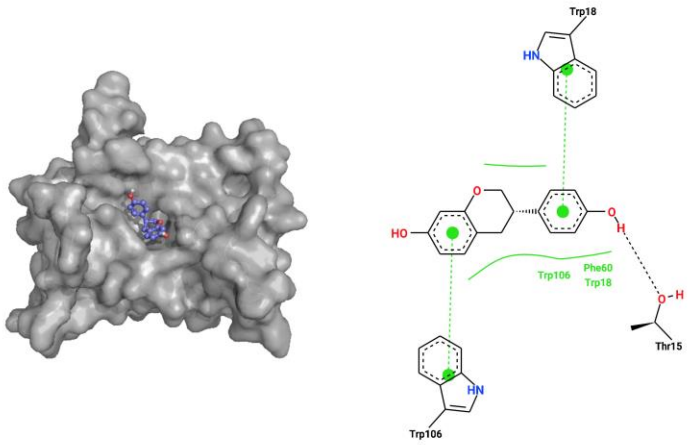
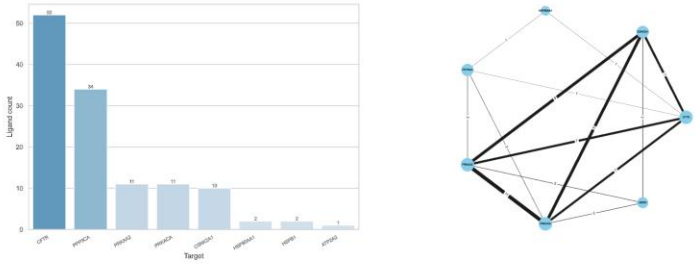


Figure 7. Docking results for ligand2995 to RAB4A and Equol to DERL1. The upper panel shows the docking results of ligand2995 to RAB4A (PDB 2BME). (**A**) Ligand2995 (coloured in purple) docked to RAB4A (PDB:2BME) crystal structure at the GTP (coloured in grey) biding site. (**B**) Two-dimensional representation of the binding pose of ligand2995 to RAB4A and predicted interactions between them. Black dashed lines represent hydrogen bonds, green spline sections show hydrophobic interactions and green dashed lines represent π-π stacking or π–cation interactions. (**C**): Equol (coloured in purple) docked to DERL1 (PDB:7CZB) crystal structure. (**D**) Two-dimensional representation of the binding pose of Equol to DERL1 and predicted interactions between them. Black dashed lines represent hydrogen bonds, green spline sections show hydrophobic interactions and green dashed lines represent π-π stacking or π–cation interactions.

As the ligand-based similarity approach does not depend on protein structures, the dataset was extended by including all remaining potential targets from the CFTR Lifecycle Map, resulting in a list of 168 proteins. The same procedure used for the small dataset was used here to identify potential targets of active compounds in the CandActCFTR database. For all targets combined, 36,851 target–ligand interactions could be found in the BindingDB, and 29,020 were found in ChEMBL. There was an overlap of 15,984 target–ligand interactions, resulting in a combined dataset of 49,887 unique interactions. Ligands for 75 targets were collected from ChEMBL and ligands for 54 targets were collected from the BindingDB, with an overlap of 52 targets between them. More than 5000 ligands could be found for the histone deacetylase 6 (HDAC6, 6815 ligands), the phosphodiesterase 4D (PDE4D, 6769 ligands), the glucocorticoid receptor NR3C1 (5063 ligands) and the adenosine A2B receptor (ADORA2B, 5032 ligands). Again, the high number of ligands present in the database indicates that they are pharmacological interesting targets. For 8 different targets, more than 1000 ligands could be collected, and less than 100 ligands were found for 46 targets.

The ligand similarity comparisons resulted in potential targets for 108 compounds, distributed across 25 targets. Figure 8A shows the number of compounds identified per target. Again, most compounds were found for CFTR (52 compounds), due to the general bias towards CFTR-relevant compounds amongst the query ligands from CandActCFTR. More than 40 compounds were linked to NR3C1 (42 compounds), the beta-2 adrenergic receptor (ADRB2, 40 compounds), followed again by the catalytic subunit of the Serine/threonine-protein phosphatase 2B (PPP3CA; 34 compounds). For the remainder of the targets, less than 15 compounds were found to be similar to the known ligands.



(**A**) (**B**)

Figure 8. Results from the ligand-based approach using the 35 selected targets. (**A**) Number of ligands associated with each target. (**B**) Protein–Ligand-Interaction network. Each nodes (circle) stands for one target, while the edges (lines) connecting them represent the number of common ligands (=number on edge) associated with them by either method. The thicker the edge, the more ligands two targets share. The size of the nodes represents their degree, i.e., the number of other nodes they are connected to.

Figure 8B shows the Protein–Ligand-Interaction network of the results. As can be seen, the targets with the most shared compounds are NR3C1, ADRB2 and PPP3CA. PPP3CA shares all of its 34 compounds with ADRB2 and NR3C1. The next targets with the most shared compounds are again PRKACA, PRKAA2 and CSNK2A1, which, as already described above, have 11 and 9 shared compounds, respectively. Additionally, CFTR and NR3C1 share an overlap on 9 compounds. Overall, these seven targets (ADRB2, CFTR, CSNK2A1, NR3C1, PPP3CA, PRKAA2 and PRKACA) are the most connected between each other. The remaining targets share a smaller number of compounds with any of the other targets.

2.4. Combined Approach

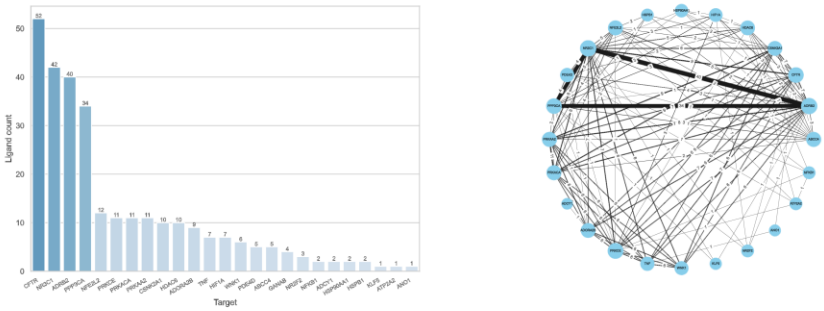
In order to obtain the most comprehensive overview of potential target–compound interactions, the results from both the structure-based docking and the ligand-based approach were combined.

Overall, a total of 1038 unique target–compound pairings were found, 757 via the structure-based approach and 290 via the ligand-based approach (Supplementary Table S4). The pairings were assigned confidence scores in the range of 1–5, depending on the number of approaches they were identified by. Hence, a score of 5 is assigned when pairings are identified by all methods, i.e., all four target-based approaches as well as the ligand-based approach. The score therefore represents the consensus amongst the different approaches used. The highest score reached was 4, by the RAB7A–Lumacaftor pairing. Overall, 8 pairings have a score of 3, 117 pairings have a score of 2, and the remaining 912 pairings were identified by only 1 approach. No significant structural similarities could be found between the compounds with higher confidence levels, indicating that there is no bias present amongst them.

Furthermore, nine target–ligand pairings were identified by both the target- and ligand-based approach. Of the nine pairings, five ligands were associated with CFTR and two to each of PRKACA and PRKAA2. Four of the five compounds *(Corr4a, InChIKey RDOBOPJBMQURAT-UHFFFAOYSA-N; CHEMBL4471507, InChIKey DBGTUHVUTOSJOJ-UHFFFAOYSA-N; CHEMBL4574818, InChIKey KHNUPLUQJFLSLN-UHFFFAOYSA-N; CHEMBL4435663, InChIKey XTWGHYUHSFTZLL-UHFFFAOYSA-N)* associated with CFTR are similar in structure, but the fifth (InChIKey NOGRVEQYQUCZIW-UHFFFAOYSA-N) differs substantially. One compound, apigenin, was associated with both PRKACA and PRKAA2. Apart from apigenin, PRKAA2 was associated with Kaempferol (InChIKey

IYRMWMYZSQPJKC-UHFFFAOYSA-N) and PRKACA was associated with Biochanin (InChIKey WUADCCWRTIWANL-UHFFFAOYSA-N).

Potential targets were found for all 309 active compounds from the CandActCFTR database (Supplementary Table S5). The distribution of compounds across the targets is visualized in Figure 9. Targets which were associated with at least one compound are displayed in colour, the colour itself representing how many compounds they were associated to, ranging from 1 compound (yellow) to 120 compounds (red). The target with by far the most compounds associated with it was PRKACA (120 compounds), followed by CFTR with 76 compounds, CSNK2A1 with 72 compounds and RAB5A with 52 compounds. Of the remaining targets, 27 had between 10 and 50 ligands associated with them, and 20 targets had less than 10 compounds. When looking at the binding sites of the potential CFTR ligands identified by the target-based approach, mostly five main binding sites were identified (Figure 10). Two of these correspond to the ATP-binding sites in the nucleotide binding domains 1 and 2. One was close to the experimentally resolved binding site of Lumacaftor (coloured in blue), one was close to the one of Ivacaftor (coloured in purple) [63], and there was one additional binding site in the transmembrane domain 2.



(**A**) (**B**)

Figure 9. Results from the ligand-based approach using all targets. (**A**) Number of ligands associated with each target. (**B**) Protein–Ligand-Interaction network. Each nodes (circle) stands for one target, while the edges (lines) connecting them represent the number of common ligands (=number on edge) associated with them by either method. The thicker the edge, the more ligands two targets share. The size of the nodes represents their degree, i.e., the number of other nodes they are connected to.

Conversely, for the majority of compounds, 2–4 targets were identified, while a single target was identified for only 15 compounds. Between 5 and 10 targets were identified for 49 compounds, 1 compound has 11 targets associated with it and 6 compounds have 12 targets associated with them. Interestingly, all of these 6 compounds are almost identical in structure and vary only in stereochemistry or side chain. These compounds were all predicted to target the kinases CSNK2A1, PRKAA2, PRKACA, PRKCE and WNK1, but also other targets, namely, ADOR2B, ADRB2, CFTR, NR3C1 and TNF.

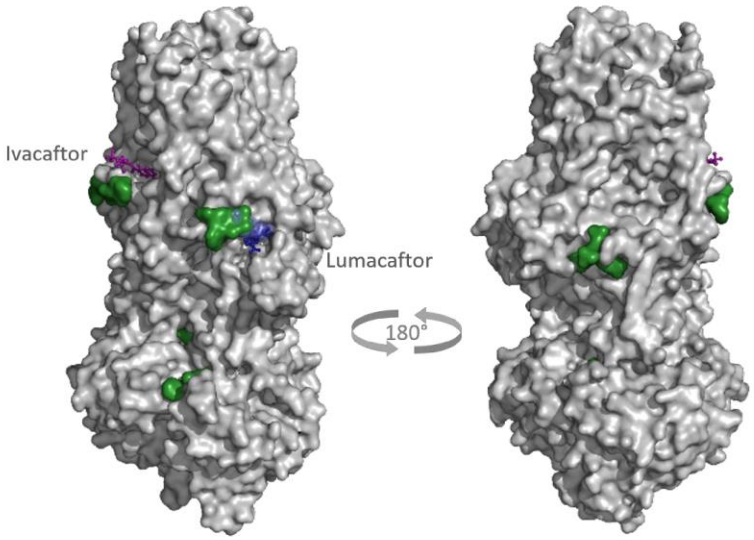


Figure 10. Most predicted binding sites in CFTR by target-based approach (PDB 7SVD and 6O2P). Coloured in green are the potential binding sites. The positions of Lumacaftor (blue) and Ivacaftor (purple) were shown experimentally by Fiedorczuk and Chen, 2022 [63]. *Shown in light blue is the binding site of Lumacaftor predicted by blind docking in the target-based approach, which is in agreement with the experimentally predicted binding site.*

3. Discussion

When searching for new drugs experimentally, especially in high-throughput screens, where thousands of compounds are tested simultaneously, the mechanism of action of promising compounds often remains unclear. Knowing the mechanism of action, however, is helpful and important for a number of reasons. It is useful in order to identify novel drug targets, find new classes of potentially active compounds, and to aid lead optimization. Furthermore, it can be helpful for the early identification of potential side effects. Elucidating the mechanism of action of promising compounds is especially important in the drug discovery for cystic fibrosis, as it is crucial for the development of combination therapies. Here, we bring together two previously developed resources of community-derived knowledge on cystic fibrosis, namely, the CandActCFTR database and the CFTR Lifecycle Map, to aid in the elucidation of modes of action for known active compounds. For this purpose, we use two different approaches to in silico target identification, a target-based approach and a ligand-based approach.

Both approaches have different advantages and disadvantages when it comes to prerequisites. The ligand-based approach is much faster and less computationally expensive. However, in order to be most effective, it requires comprehensive ligand libraries of the targets in question. The target-based approach, on the other hand, requires complete and high-resolution structures of the target proteins. From the 170 potential targets in the CFTR Lifecycle Map, quality-sufficient protein structures could be found for 35 targets, and ligand collections could be found for 77 targets, with an overlap of 22 targets. By combining the approaches, it was thus possible to cover more than 50% of the potential targets in the CFTR Lifecycle Map.

For the target-based approach, all active compounds from the CandActCFTR database were docked blindly into the 35 target structures, meaning that no prior binding site was defined, but the whole protein was searched. In order to produce comprehensive results,

two different docking programs, QuickVina-W (qvina-w) and Smina, were used, which differ in their method to calculate binding affinities. Due to their different scoring functions, qvina-w and Smina produced different results, but with overall similar trends. Generally, the binding affinities calculated by Smina were slightly higher than those calculated by qvina-w. This is most likely caused by their different approaches to estimating ligand–receptor-based affinity. While qvina-w uses the empirical Vina scoring function [57], which is based on machine-learning [64], Smina uses a more physics-based approach [65]. However, the exact binding affinities are of secondary importance in this case, as the aim in target identification is not to calculate the most precise scores, but rather compare the likeliness of different matches to find potential target–ligand pairings. Therefore, in order to normalize the results and remove false positives, two post-processing methods were applied to the docking data independently. Method I results in a list of high-ranking target–ligand pairings, while method II identifies the most likely target for each ligand. Overall, however, the two methods produce similar results. Due to the high number of ligands, method II results in a lot of more potential pairings; however, a high portion of pairings identified by method I are also suggested by method II, underpinning both their validity, while still remaining non-redundant. Interestingly, two proteins, PRKACA and CSNK2A1, were identified in pairings significantly more often than the other ones. Both proteins are protein kinases, which have been previously found to be promiscuous targets [66,67]. With respect to the high number of ligands associated with the PRKACA and CSNK2A1, the results from the ligand-based approach support the results from the target-based approach. Again, the two kinases were amongst those with the most compounds associated with them. Interestingly, a majority of these compounds could be associated with both proteins, and additionally the kinase PRKAA2, which suggests a common, possibly promiscuous, binding motif amongst these kinases. In order to remove bias from PRKACA and CSNK2A1 from the docking calculations, the analysis was repeated without the data for these two kinases.

Overall, the majority of the predicted interactions (506 of the 1038 predicted interactions) involve proteins that play a role in the activity and regulation of CFTR at the plasma membrane. One explanation for this are the readouts of the experimental assays used to identify the active compounds in their original publications. Most assays use the ion conductance directly as a readout, so compounds that affect the activity of CFTR directly at the membrane are readily detected by these methods. Of the remaining predicted interactions, 161 involve proteins that play a role in CFTR translation and folding, 120 involve proteins of endocytosis and 108 involve protein of the secretory pathway. At total of 76 interactions involve CFTR directly, and only 67 interactions involve transcriptions factors or other proteins that influence CFTR transcription.

Remarkably, one of the protein-ligand pairings with the highest consensus amongst the different approaches was for the compound Lumacaftor (VX-809) with RAB7A. Lumacaftor is a well-known, clinically approved drug for patients with F508del mutations, as it acts as a small-molecule chaperone to correct the folding defect of F508del-CFTR. It is known to bind directly to CFTR, with its exact binding site elucidated by CryoEM (PDBs 7SVD and 7SVR) [63]. Here, however, it was suggested to potentially also bind to the Ras-related protein RAB7A, a GTP-binding protein involved in endocytosis, including that of CFTR [68–70]. When looking close at the predicted binding mode, two different binding sites are suggested for Lumacaftor in the RAB7A protein by both docking programs. In both pockets, Lumacaftor is predicted to interact with the protein via different hydrogen bonds, hydrophobic interactions and π-interactions. When looking at the binding mode for Lumacaftor to CFTR as shown in the CryoEM structure (7SVD) [63], the compound is coordinated mainly by hydrophobic interactions and only one hydrogen bond (Figure 11). Hydrophobic interactions are rather weak intermolecular interactions, which is probably why the binding affinity for Lumacaftor was calculated to be higher for RAB7A than CFTR by the docking programs. However, when looking at the raw docking results for CFTR, Lumacaftor is nonetheless amongst the top 10 highest ranking active compounds calculated by both docking programs. Furthermore, Hou et al. previously showed how

RAB7A inhibition increases apical CFTR stability [70]. Hence, while certainly intriguing and requiring further investigation before confirmation, the pairing of Lumacaftor with RAB7A, rather than with CFTR, therefore does not undermine the potential validity of the results, as the scores calculated for the CFTR–Lumacaftor (Figure 12) pairing are only slightly lower than the one for RAB7A–Lumacaftor.

Figure 11. Combined results of all target identification approaches visualized in CFTR Lifecycle Map. The CFTR Lifecycle Map is an SBGN (Systems Biology Graphical Notation) representation of CFTR biogenesis in the cell. All proteins are shown as rounded rectangles. The colour represents the number of compounds associated with each target by all four approaches combined, ranging from one compound (yellow) to 120 compounds (red).

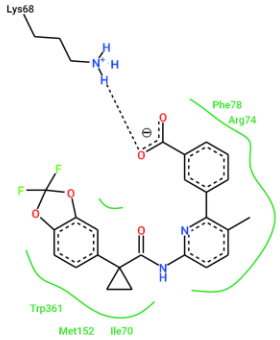


Figure 12. Two-dimensional representation of the experimentally shown binding site of Lumacaftor to CFTR (PDB 7SVD) [63] and predicted interactions between them. Black dashed lines represent hydrogen bonds, green spline sections show hydrophobic interactions and green dashed lines represent π-π stacking or π–cation interactions.

Another high-consensus protein–ligand pairing from the target-based approach was for the compound Equol with DERL1. Equol is an isoflavandiol oestrogen, known to bind to oestrogen receptor β [71]. It was also shown to be on the one hand a potent activator of F508del-CFTR, but on the other hand, it affects its misprocessing [72]. DERL1 is an ER membrane protein mediating the degradation of misfolded proteins such as misfolded CFTR [73,74]. Thus, the effect Equol has on F508del-CFTR misprocessing might be mediated by its binding to DERL1.

Of the 170 different proteins in the map, either protein structures or ligand data could be found for 90 of them, which amounts to a coverage of 53%. This means that 80 proteins (47% of proteins in the map) could not be considered as targets, due to the lack of available data. As is often the case in in silico approaches, this effort to elucidate the mechanism of action of candidate compounds was hampered by the lack of a comprehensive data corpus. Nonetheless, by combining the ligand- and the target-based approach, potential targets were suggested for all 309 active compounds in the CandActCFTR database, distributed across 53 of the potential targets from the CFTR Lifecycle map.

It has to be noted that not all proteins involved in the CFTR biogenesis are equally suitable as targets for CFTR modulators. For example, chaperones and other proteins that are important for CFTR folding, ER quality control, and trafficking play a quite general role in cells and are not specific to CFTR alone. Targeting them can therefore potentially lead to unwanted side effects. Therefore, when selecting compounds and targets for experimental testing, more research is needed on which proteins can be targeted and which should rather not be interfered with.

Of course, experimental validation of protein–ligand interactions is always necessary and preferred to docking- and ligand-similarity approaches. However, where experimental data is not available or its generation not feasible, these in silico approaches are a good way to fill knowledge gaps, suggest potential interactors, and thereby narrow down the candidates for testing in the wet lab. These methods should therefore not be viewed as replacement for wet-lab experiments, but as preceding step to support experimental work and, in the meantime, at least provide indications of possible modes of action.

This project is part of the CandActCFTR project and serves as an essential connecting piece between the chemistry centric CandActCFTR database, which collects compounds tested as CFTR modulators, and the cell biology centric CFTR Lifecycle map, which is a systems biology model of the CFTR biogenesis. By bringing the two parts together here, we not only use and repurpose data from both resources, but are able to generate new knowledge. The structured knowledge in systems medicine disease maps can therefore directly be applied to drug design approaches. By using this comprehensive approach,

we can now suggest mechanisms of actions for active compounds, propose potential drug targets and predict possible additive effects of different substance combinations.

4. Materials and Methods

4.1. Reverse Docking

In order to identify targets for the active compounds of the CF-specific compound database CandActCFTR [42,43], 37 potentially relevant protein targets were used for reverse docking. The proteins were selected using the CFTR lifecycle map [45], a systems biology model of the CFTR lifecycle. For this purpose, a KNIME [75] workflow was used to search all PDB [52] entries for all proteins in the map and then filter them according different criteria such as structure completeness, resolution and experimental method. This list was narrowed down to 35 PDB structures, including that of wt-CFTR, belonging to targets evenly distributed across the CFTR lifecycle. All target structures were prepared for docking using AutoDockTools 1.5.7 [76].

The active compounds from the CandActCFTR database [43] were used as ligand library. Structures were obtained in SMILES notation and converted and prepared for docking using Open Babel [77,78].

Docking of all compounds against all ligands was performed using Virtual Flow [79] with two different docking programs. Specifically, it was carried out using the Smina docking program [56], with the Vinardo scoring function [65] to calculate binding scores between ligands and targets, and the QuickVina-W [57] program, with the AutoDock Vina scoring function [64]. Docking was carried out as blind docking, meaning that no specific binding pocket was defined for each protein, but the entire protein was searched in order to account for structures without known binding site and targets with several, potentially unknown, binding sites. The search space in blind docking is significantly higher than when using pre-defined binding sites, which leads to higher probabilities of not finding the optimal conformation. Nevertheless, to obtain comprehensive results, the exhaustiveness, i.e., the number of runs calculated per ligand and target was increased to 100.

All docking calculations were run on the Scientific Compute Cluster at the joint data centre of Max Planck Society for the Advancement of Science (MPG) and University of Göttingen with the SLURM Workload Manager [80].

Post-processing of the docking was performed using KNIME [75,81] and Pyrhon. In order to remove false positives and obtain more reliable consensus results, two different approaches were used to evaluate target–ligand pairings.

The first approach proposed by Lauro et al. [51] normalizes the binding energies of each target–ligand pairing in a way that eliminates systematic errors and exclude false positives. To do so, the docking results are written in a matrix format and equation (1) was applied, where V is the new value, V_0 is the binding energy from the docking calculate, M_L is the average binding energy of each ligand, and M_T is the average binding energy of each target.

$$V = \frac{V_0}{[M_L + M_T]/2} \quad (1)$$

Promising pairings were then selected by setting a lower threshold at V = + 3σ, where M is the average of the whole matric, and σ is its standard deviation.

The second approach by Kim et al. [58] uses a 2-directional Z-transformation on the docking-score matrix. Here, a target- and ligand-specific Z-score is calculated using Equations (2) and (3), where is x_i the docking score specific to the query ligand i, $\overline{x}_t$ and $\overline{x}_l$ are the average scores of all targets and ligands, respectively, and SD_T and SD_L are the respective standard deviations.

$$Z_T = \frac{(x_i - \overline{x}_T)}{SD_T} \quad (2)$$

$$Z_L = \frac{(x_i - \overline{x}_L)}{SD_L} \quad (3)$$

The combined Z_{comb} is then calculated through $Z_{comb} = 0.7 * Z_T + 0.3 * Z_L$. In order to select the most likely target, the receptor with the lowest Z_{comb} is selected for each ligand. The 2-directional Z-transformation was carried out using the Python script provided by Kim et al. [58].

Results from both filtering approaches were then compared to check for consensus and find the most promising target–ligand pairings.

4.2. Ligand Based Approach

For the ligand-based approach, known ligands of the 35 proteins also used for the reverse docking were collected from different databases. Databases used were the BindingDB [59] (https://www.bindingdb.org, accessed on 9 September 2022) and ChEMBL (https://www.ebi.ac.uk/chembl/, accessed on 9 September 2022) [60–62]. Interaction data from BindingDB were collected using their API via the KNIME workflow provided by the BindingDB team, and from ChEMBL using the ChEMBL websource client [60] for Python. All ligands were collected in the SMILES notation [82], which was converted to the InChIKey [83] using OpenBabel [78].

In order to perform similarity comparisons according to the structural properties of the molecules, for each reference ligand from the databases, as well as each active compound from the CandActCFTR database [42,43], the Morgan fingerprint [84] was computed using the RDKit [85]. Next, the similarity between each reference ligand and each query compound from CandActCFTR was calculated using the Tanimoto coefficient. If a query compound shared a similarity of $\geq$0.75 with at least one of the target-specific reference compounds, it was considered a potential ligand of the respective target. The threshold of 0.75 was chosen to be less restrictive than the commonly used threshold of 0.85.

5. Conclusions

This study used a complementary approach of target- and ligand-based in silico target identification methods to predict potential targets of compounds that were shown to affect CFTR activity. By using a dual approach, we were able to reap the benefits of both methods and increase the number of potential target proteins to get a better coverage of the CFTR biogenesis. Overall, the number of potential pairings could be narrowed down to 1038 predicted interactions, which are assigned a score depending on how high the consensus is amongst the methods employed. We were thereby able to predict potential targets for all 309 active compounds from the CandActCFTR database. These results help elucidate the mechanism of action of promising compounds and can be used to select compounds and targets to predict synergistically acting compound combinations for testing in the wet lab.

Supplementary Materials: The supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijms232012351/s1.

Author Contributions: Conceptualization, L.V. and M.M.N.; methodology, L.V.; formal analysis, L.V.; investigation, L.V.; data curation, L.V.; writing—original draft preparation, L.V.; writing—review and editing, F.S., S.H. and M.M.N.; visualization, L.V.; funding acquisition, F.S. and M.M.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft DFG, grant number 315063128.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data on compounds tested as CFTR modulators is available under https://candactcftr.ams.med.uni-goettingen.de/ (accessed on 9 September 2022), the CFTR Lifecycle Map is available under https://cf-map.uni-goettingen.de/ (accessed on 9 September 2022) and the data generated during this study can be found in the supplementary materials.

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

References and Note

1. Bobadilla, J.L.; Macek, M.; Fine, J.P.; Farrell, P.M. Cystic fibrosis: A worldwide analysis of CFTR mutations—Correlation with incidence data and application to screening. *Hum. Mutat.* **2002**, *19*, 575–606. [CrossRef] [PubMed]
2. Farrell, P.M. The prevalence of cystic fibrosis in the European Union. *J. Cyst. Fibros.* **2008**, *7*, 450–453. [CrossRef] [PubMed]
3. Bell, S.C.; Mall, M.A.; Gutierrez, H.; Macek, M.; Madge, S.; Davies, J.C.; Burgel, P.R.; Tullis, E.; Castaños, C.; Castellani, C.; et al. The future of cystic fibrosis care: A global perspective. *Lancet Respir. Med.* **2020**, *8*, 65–124. [CrossRef]
4. O'Sullivan, B.P.; Freedman, S.D. Cystic fibrosis. *Lancet* **2009**, *373*, 1891–1904. [CrossRef]
5. Elborn, J.S. Cystic fibrosis. *Lancet* **2016**, *388*, 2519–2531. [CrossRef]
6. Cystic Fibrosis Mutation Database. Available online: http://www.genet.sickkids.on.ca/ (accessed on 26 January 2021).
7. Welcome to CFTR2|CFTR2. Available online: https://www.cftr2.org/ (accessed on 26 January 2021).
8. Sosnay, P.R.; Siklosi, K.R.; van Goor, F.; Kaniecki, K.; Yu, H.; Sharma, N.; Ramalho, A.S.; Amaral, M.D.; Dorfman, R.; Zielenski, J.; et al. Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene. *Nat. Genet.* **2013**, *45*, 1160–1167. [CrossRef]
9. Pranke, I.M.; Sermet-Gaudelus, I. Biosynthesis of cystic fibrosis transmembrane conductance regulator. *Int. J. Biochem. Cell Biol.* **2014**, *52*, 26–38. [CrossRef]
10. Welsh, M.J.; Smith, A.E. Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis. *Cell* **1993**, *73*, 1251–1254. [CrossRef]
11. Rowe, S.M.; Miller, S.; Sorscher, E.J. Cystic fibrosis. *N. Engl. J. Med.* **2005**, *352*, 1992–2001. [CrossRef]
12. Zielenski, J.; Tsui, L.C. Cystic fibrosis: Genotypic and phenotypic variations. *Annu. Rev. Genet.* **1995**, *29*, 777–807. [CrossRef]
13. Zielenski, J. Genotype and Phenotype in Cystic Fibrosis. *Respiration* **2000**, *67*, 117–133. [CrossRef] [PubMed]
14. De Boeck, K. Cystic fibrosis in the year 2020: A disease with a new face. *Acta Paediatr.* **2020**, *109*, 893–899. [CrossRef] [PubMed]
15. Veit, G.; Avramescu, R.G.; Chiang, A.N.; Houck, S.A.; Cai, Z.; Peters, K.W.; Hong, J.S.; Pollard, H.B.; Guggino, W.B.; Balch, W.E.; et al. From CFTR biology toward combinatorial pharmacotherapy: Expanded classification of cystic fibrosis mutations. *Mol. Biol. Cell* **2016**, *27*, 424–433. [CrossRef] [PubMed]
16. Gentzsch, M.; Mall, M.A. Ion Channel Modulators in Cystic Fibrosis. *Chest* **2018**, *154*, 383–393. [CrossRef]
17. Zaher, A.; ElSaygh, J.; Elsori, D.; ElSaygh, H.; Sanni, A. A Review of Trikafta: Triple Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Modulator Therapy. *Cureus* **2021**, *13*, e16144. [CrossRef]
18. Drug Development Pipeline: CFF Clinical Trials Tool. Available online: https://www.cff.org/Trials/Pipeline (accessed on 26 January 2021).
19. Clinical Pipeline. Available online: https://www.glpg.com/clinical-pipelines (accessed on 28 June 2022).
20. Van Goor, F.; Hadida, S.; Grootenhuis, P.D.J.; Burton, B.; Cao, D.; Neuberger, T.; Turnbull, A.; Singh, A.; Joubran, J.; Hazlewood, A.; et al. Rescue of CF airway epithelial cell function in vitro by a CFTR potentiator, VX-770. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 18825–18830. [CrossRef]
21. Ramsey, B.W.; Davies, J.; McElvaney, N.G.; Tullis, E.; Bell, S.C.; Dřevínek, P.; Griese, M.; McKone, E.F.; Wainwright, C.E.; Konstan, M.W.; et al. A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation. *N. Engl. J. Med.* **2011**, *365*, 1663–1672. [CrossRef]
22. Clancy, J.P.; Rowe, S.M.; Accurso, F.J.; Aitken, M.L.; Amin, R.S.; Ashlock, M.A.; Ballmann, M.; Boyle, M.P.; Bronsveld, I.; Campbell, P.W.; 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* **2012**, *67*, 12–18. [CrossRef]
23. Wainwright, C.E.; Elborn, J.S.; Ramsey, B.W.; Marigowda, G.; Huang, X.; Cipolli, M.; Colombo, C.; Davies, J.C.; de Boeck, K.; Flume, P.A.; et al. Lumacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR. *N. Engl. J. Med.* **2015**, *373*, 220–231. [CrossRef]
24. Taylor-Cousar, J.L.; Munck, A.; McKone, E.F.; van der Ent, C.K.; Moeller, A.; Simard, C.; Wang, L.T.; Ingenito, E.P.; McKee, C.; Lu, Y.; et al. Tezacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del. *N. Engl. J. Med.* **2017**, *377*, 2013–2023. [CrossRef]
25. Voelker, R. Patients with Cystic Fibrosis Have New Triple-Drug Combination. *JAMA* **2019**, *322*, 2068. [CrossRef] [PubMed]
26. Ridley, K.; Condren, M. Elexacaftor-tezacaftor-ivacaftor: The first triple-combination cystic fibrosis transmembrane conductance regulator modulating therapy. *J. Pediatr. Pharmacol. Ther.* **2020**, *25*, 192–197. [CrossRef] [PubMed]
27. Goetz, D.M.; Savant, A.P. Review of CFTR modulators 2020. *Pediatr. Pulmonol.* **2021**, *56*, 3595–3606. [CrossRef] [PubMed]
28. Martiniano, S.L.; Sagel, S.D.; Zemanick, E.T. Cystic fibrosis: A model system for precision medicine. *Curr. Opin. Pediatr.* **2016**, *28*, 312–317. [CrossRef] [PubMed]
29. Southern, K.W.; Patel, S.; Sinha, I.P.; Nevitt, S.J. Correctors (specific therapies for class II CFTR mutations) for cystic fibrosis. *Cochrane Database Syst. Rev.* **2018**, *2018*, CD010966. [CrossRef]
30. Pedemonte, N.; Lukacs, G.L.; Du, K.; Caci, E.; Zegarra-Moran, O.; Galietta, L.J.V.; Verkman, A.S. Small-molecule correctors of defective ΔF508-CFTR cellular processing identified by high-throughput screening. *J. Clin. Investig.* **2005**, *115*, 2564–2571. [CrossRef]

31. Berg, A.; Hallowell, S.; Tibbetts, M.; Beasley, C.; Brown-Phillips, T.; Healy, A.; Pustilnik, L.; Doyonnas, R.; Pregel, M. High-Throughput Surface Liquid Absorption and Secretion Assays to Identify F508del CFTR Correctors Using Patient Primary Airway Epithelial Cultures. *SLAS Discov.* **2019**, *24*, 724–737. [CrossRef]
32. De Wilde, G.; Gees, M.; Musch, S.; Verdonck, K.; Jans, M.; Wesse, A.S.; Singh, A.K.; Hwang, T.C.; Christophe, T.; Pizzonero, M.; et al. Identification of GLPG/ABBV-2737, a novel class of corrector, which exerts functional synergy with other CFTR modulators. *Front. Pharmacol.* **2019**, *10*, 514. [CrossRef]
33. Merkert, S.; Schubert, M.; Olmer, R.; Engels, L.; Radetzki, S.; Veltman, M.; Scholte, B.J.; Zöllner, J.; Pedemonte, N.; Galietta, L.J.V.; et al. High-Throughput Screening for Modulators of CFTR Activity Based on Genetically Engineered Cystic Fibrosis Disease-Specific iPSCs. *Stem Cell Rep.* **2019**, *12*, 1389–1403. [CrossRef]
34. Van Goor, F.; Hadida, S.; Grootenhuis, P.D.J.; Burton, B.; Stack, J.H.; Straley, K.S.; Decker, C.J.; Miller, M.; McCartney, J.; Olson, E.R.; et al. Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 18843–18848. [CrossRef]
35. Phuan, P.W.; Veit, G.; Tan, J.A.; Finkbeiner, W.E.; Lukacs, G.L.; Verkman, A.S. Potentiators of defective DF508-CFTR gating that do not interfere with corrector action. *Mol. Pharmacol.* **2015**, *88*, 791–799. [CrossRef] [PubMed]
36. Carlile, G.W.; Robert, R.; Goepp, J.; Matthes, E.; Liao, J.; Kus, B.; Macknight, S.D.; Rotin, D.; Hanrahan, J.W.; Thomas, D.Y. Ibuprofen rescues mutant cystic fibrosis transmembrane conductance regulator trafficking. *J. Cyst. Fibros.* **2015**, *14*, 16–25. [CrossRef] [PubMed]
37. Liang, F.; Shang, H.; Jordan, N.J.; Wong, E.; Mercadante, D.; Saltz, J.; Mahiou, J.; Bihler, H.J.; Mense, M. High-Throughput Screening for Readthrough Modulators of CFTR PTC Mutations. *SLAS Technol.* **2017**, *22*, 315–324. [CrossRef]
38. Giuliano, K.A.; Wachi, S.; Drew, L.; Dukovski, D.; Green, O.; Bastos, C.; Cullen, M.D.; Hauck, S.; Tait, B.D.; Munoz, B.; et al. Use of a High-Throughput Phenotypic Screening Strategy to Identify Amplifiers, a Novel Pharmacological Class of Small Molecules That Exhibit Functional Synergy with Potentiators and Correctors. *SLAS Discov.* **2018**, *23*, 392–399. [CrossRef] [PubMed]
39. Van der Plas, S.E.; Kelgtermans, H.; de Munck, T.; Martina, S.L.X.; Dropsit, S.; Quinton, E.; de Blieck, A.; Joannesse, C.; Tomaskovic, L.; Jans, M.; et al. Discovery of N-(3-Carbamoyl-5,5,7,7-tetramethyl-5,7-dihydro-4H-thieno[2,3-c]pyran-2-yl)-lH-pyrazole-5-carboxamide (GLPG1837), a Novel Potentiator Which Can Open Class III Mutant Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Channels to a High Extent. *J. Med. Chem.* **2018**, *61*, 1425–1435. [CrossRef] [PubMed]
40. Veit, G.; Xu, H.; Dreano, E.; Avramescu, R.G.; Bagdany, M.; Beitel, L.K.; Roldan, A.; Hancock, M.A.; Lay, C.; Li, W.; et al. Structure-guided combination therapy to potently improve the function of mutant CFTRs. *Nat. Med.* **2018**, *24*, 1732–1742. [CrossRef] [PubMed]
41. Wang, X.; Liu, B.; Searle, X.; Yeung, C.; Bogdan, A.; Greszler, S.; Singh, A.; Fan, Y.; Swensen, A.M.; Vortherms, T.; et al. Discovery of 4-[(2R,4R)-4-({[1-(2,2-Difluoro-1,3-benzodioxol-5-yl)cyclopropyl]carbonyl}amino)-7-(difluoromethoxy)-3,4-dihydro-2H-chromen-2-yl]benzoic Acid (ABBV/GLPG-2222), a Potent Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Corrector for the Treatment of Cystic Fibrosis. *J. Med. Chem.* **2018**, *61*, 1436–1449. [CrossRef]
42. Welcome to CandActCFTR. Available online: https://candactcftr.ams.med.uni-goettingen.de/ (accessed on 26 January 2021).
43. Nietert, M.M.; Vinhoven, L.; Auer, F.; Hafkemeyer, S.; Stanke, F. Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR. *Front. Pharmacol.* **2021**, *12*, 689205. [CrossRef]
44. CF-Map. Available online: https://cf-map.uni-goettingen.de/ (accessed on 29 June 2022).
45. Vinhoven, L.; Stanke, F.; Hafkemeyer, S.; Nietert, M.M. CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations. *Int. J. Mol. Sci.* **2021**, *22*, 7590. [CrossRef]
46. Xu, X.; Huang, M.; Zou, X. Docking-based inverse virtual screening: Methods, applications, and challenges. *Biophys. Reports* **2018**, *4*, 1–16. [CrossRef]
47. Huang, H.; Zhang, G.; Zhou, Y.; Lin, C.; Chen, S.; Lin, Y.; Mai, S.; Huang, Z. Reverse screening methods to search for the protein targets of chemopreventive compounds. *Front. Chem.* **2018**, *6*, 138. [CrossRef] [PubMed]
48. Lim, T.G.; Lee, S.Y.; Huang, Z.; Lim, D.Y.; Chen, H.; Jung, S.K.; Bode, A.M.; Lee, K.W.; Dong, Z. Curcumin suppresses proliferation of colon cancer cells by targeting CDK2. *Cancer Prev. Res.* **2014**, *7*, 466–474. [CrossRef] [PubMed]
49. Buendia-Atencio, C.; Pieffet, G.P.; Montoya-Vargas, S.; Martínez Bernal, J.A.; Rangel, H.R.; Muñoz, A.L.; Losada-Barragán, M.; Segura, N.A.; Torres, O.A.; Bello, F.; et al. Inverse Molecular Docking Study of NS3-Helicase and NS5-RNA Polymerase of Zika Virus as Possible Therapeutic Targets of Ligands Derived from Marcetia taxifolia and Its Implications to Dengue Virus. *ACS Omega* **2021**, *6*, 6134–6143. [CrossRef] [PubMed]
50. Ban, F.; Hu, L.; Zhou, X.H.; Zhao, Y.; Mo, H.; Li, H.; Zhou, W. Inverse molecular docking reveals a novel function of thymol: Inhibition of fat deposition induced by high-dose glucose in Caenorhabditis elegans. *Food Sci. Nutr.* **2021**, *9*, 4243–4253. [CrossRef] [PubMed]
51. Lauro, G.; Romano, A.; Riccio, R.; Bifulco, G. Inverse virtual screening of antitumor targets: Pilot study on a small database of natural bioactive compounds. *J. Nat. Prod.* **2011**, *74*, 1401–1407. [CrossRef]
52. RCSB Research Collaboratory for Structural Bioinformatics (RCSB).
53. AlphaFold Protein Structure Database. Available online: https://alphafold.ebi.ac.uk/ (accessed on 3 June 2022).
54. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [CrossRef]

55. Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; et al. AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **2022**, *50*, D439–D444. [CrossRef]
56. Koes, D.R.; Baumgartner, M.P.; Camacho, C.J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904. [CrossRef]
57. Hassan, N.M.; Alhossary, A.A.; Mu, Y.; Kwoh, C.K. Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration. *Sci. Rep.* **2017**, *7*, 1–13. [CrossRef]
58. Kim, S.S.; Aprahamian, M.L.; Lindert, S. Improving inverse docking target identification with Z-score selection. *Chem. Biol. Drug Des.* **2019**, *93*, 1105–1116. [CrossRef]
59. Gilson, M.K.; Liu, T.; Baitaluk, M.; Nicola, G.; Hwang, L.; Chong, J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*, D1045–D1053. [CrossRef] [PubMed]
60. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J.P. ChEMBL web services: Streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **2015**, *43*, W612–W620. [CrossRef] [PubMed]
61. Mendez, D.; Gaulton, A.; Bento, A.P.; Chambers, J.; de Veij, M.; Félix, E.; Magariños, M.P.; Mosquera, J.F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940. [CrossRef]
62. CHEMBL Database Release 30. 2022. Available online: http://chembl.blogspot.com/2022/03/chembl-30-released.html (accessed on 9 September 2022).
63. Fiedorczuk, K.; Chen, J. Mechanism of CFTR correction by type I folding correctors. *Cell* **2022**, *185*, 158.e11–168.e11. [CrossRef] [PubMed]
64. Trott, O.; Olson, A.J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **2009**, *31*, 455–461. [CrossRef] [PubMed]
65. Quiroga, R.; Villarreal, M.A. Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening. *PLoS ONE* **2016**, *11*, e0155183. [CrossRef] [PubMed]
66. Hanson, S.M.; Georghiou, G.; Thakur, M.K.; Miller, W.T.; Rest, J.S.; Chodera, J.D.; Seeliger, M.A. What Makes a Kinase Promiscuous for Inhibitors? *Cell Chem. Biol.* **2019**, *26*, 390.e5–399.e5. [CrossRef]
67. Cerisier, N.; Petitjean, M.; Regad, L.; Bayard, Q.; Réau, M.; Badel, A.; Camproux, A.C. High impact: The role of promiscuous binding sites in polypharmacology. *Molecules* **2019**, *24*, 2529. [CrossRef]
68. Ameen, N.; Silvis, M.; Bradbury, N.A. Endocytic trafficking of CFTR in health and disease. *J. Cyst. Fibros.* **2007**, *6*, 1–14. [CrossRef]
69. Farinha, C.M.; Matos, P. Rab GTPases regulate the trafficking of channels and transporters—A focus on cystic fibrosis. *Small GTPases* **2018**, *9*, 136–144. [CrossRef]
70. Hou, X.; Wu, Q.; Rajagopalan, C.; Zhang, C.; Bouhamdan, M.; Wei, H.; Chen, X.; Zaman, K.; Li, C.; Sun, X.; et al. CK19 stabilizes CFTR at the cell surface by limiting its endocytic pathway degradation. *FASEB J.* **2019**, *33*, 12602–12615. [CrossRef]
71. Muthyala, R.S.; Ju, Y.H.; Sheng, S.; Williams, L.D.; Doerge, D.R.; Katzenellenbogen, B.S.; Helferich, W.G.; Katzenellenbogen, J.A. Equol, a natural estrogenic metabolite from soy isoflavones: Convenient preparation and resolution of R- and S-equols and their differing binding and biological activity through estrogen receptors alpha and beta. *Bioorganic Med. Chem.* **2004**, *12*, 1559–1567. [CrossRef] [PubMed]
72. Pyle, L.C.; Fulton, J.C.; Sloane, P.A.; Backer, K.; Mazur, M.; Prasain, J.; Barnes, S.; Clancy, J.P.; Rowe, S.M. Activation of the cystic fibrosis transmembrane conductance regulator by the flavonoid quercetin: Potential use as a biomarker of ΔF508 cystic fibrosis transmembrane conductance regulator rescue. *Am. J. Respir. Cell Mol. Biol.* **2010**, *43*, 607–616. [CrossRef] [PubMed]
73. Younger, J.M.; Chen, L.; Ren, H.-Y.; Rosser, M.F.N.; Turnbull, E.L.; Fan, C.-Y.; Patterson, C.; Cyr, D.M. Sequential quality-control checkpoints triage misfolded cystic fibrosis transmembrane conductance regulator. *Cell* **2006**, *126*, 571–582. [CrossRef]
74. Grove, D.E.; Fan, C.-Y.; Ren, H.Y.; Cyr, D.M. The endoplasmic reticulum-associated Hsp40 DNAJB12 and Hsc70 cooperate to facilitate RMA1 E3-dependent degradation of nascent CFTRDeltaF508. *Mol. Biol. Cell* **2011**, *22*, 301–314. [CrossRef] [PubMed]
75. Berthold, M.R.; Cebron, N.; Dill, F.; Gabriel, T.R.; Kötter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME—The Konstanz information miner. *ACM SIGKDD Explor. Newsl.* **2009**, *11*, 26–31. [CrossRef]
76. Morris, G.M.; Ruth, H.; Lindstrom, W.; Sanner, M.F.; Belew, R.K.; Goodsell, D.S.; Olson, A.J. Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791. [CrossRef]
77. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3*, 33. [CrossRef]
78. Open Babel. Available online: http://openbabel.org/wiki/Main_Page (accessed on 20 May 2022).
79. Gorgulla, C.; Boeszoermenyi, A.; Wang, Z.F.; Fischer, P.D.; Coote, P.W.; Padmanabha Das, K.M.; Malets, Y.S.; Radchenko, D.S.; Moroz, Y.S.; Scott, D.A.; et al. An open-source drug discovery platform enables ultra-large virtual screens. *Nature* **2020**, *580*, 663–668. [CrossRef]
80. SLURM. Available online: https://slurm.schedmd.com (accessed on 9 September 2022).
81. Data Analytics Platform: Open Source Software Tools | KNIME. Available online: https://www.knime.com/knime-analytics-platform (accessed on 31 May 2022).
82. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28*, 31–36. [CrossRef]

83. Southan, C. InChI in the wild: An assessment of InChIKey searching in Google. *J. Cheminform.* **2013**, *5*, 10. [CrossRef] [PubMed]
84. Morgan, H.L. The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service. *J. Chem. Doc.* **1965**, *5*, 107–113. [CrossRef]
85. RDKit. Available online: https://www.rdkit.org/ (accessed on 2 June 2022).

Chapter 7 Discussion

With the advance of high-throughput technologies to generate large amounts of biological and biomedical data in a time- and cost-efficient manner, systems medicine has become of tremendous importance over the last decades. A great variety of sophisticated and highly specialized techniques and tools have been developed to analyse, interpret and utilize this great wealth of data. Yet, oftentimes the tools used are only applicable to a certain type of data and a specific use case, making it difficult to integrate different data sources and thereby realizing their full potential. The aim of this thesis was therefore to use a set of generic and interchangeable tools to demonstrate the added value of integrating and combining different types of data from open access sources. This was carried out on the use case of Cystic Fibrosis, a genetic disorder still lacking causative treatment for all patients due to its plethora of different geno- and phenotypes. The main emphasis of this project was on using generic methods that can later be easily adapted to other use cases. These were applied to already published open-access data to use what is already accessible and generate new value from it. However, due to the versatile nature of the tools and methods used, there are no specific data requirements, hence other data sources can also be easily integrated if available.

The project is carried by two main pillars making up the data basis: The chemistry centric compound database CandActCFTR and the systems biology model CFTR Lifecycle Map, which contains information on the biological pathways and biomolecules involved in the disease. The two sides of the projects are then brought together by *in silico* experiments to, on the one hand, identify the targets of the chemical compounds and, on the other hand, suggest potential therapeutics for specific targets in the CFTR biogenesis.

7.1 CandActCFTR database

The CandActCFTR database is based on the software framework CandActBase developed in our work group. In general, CandActBase consists of different software tools to collect, store and link literature data on a specific topic, and make it publicly

available in a structured and consistent manner. Overall, many different compound databases have been developed for a range of purposes. With more than 112 Million compounds, PubChem is by far the largest collection of chemical compounds (Kim *et al.*, 2016, 2021). It contains chemical and physical properties of all compounds, as well as information on their (bio)activity, along with listing all synonyms the compounds are known under. PubChem is a general database not focussed on one specific topic, but also includes results from bioassays and information on all kinds of biological targets with tested compounds and the bioassays used. Apart from the compound pages, which provide details on the compounds' properties, there are also target pages, which summarize information on and list interacting compounds of genes and proteins. Another similar and somewhat related database is ChEMBL (Davies *et al.*, 2015; Mendez *et al.*, 2019). Like PubChem, ChEMBL is a generic database, but it is manually curated and puts a larger focus on information on bioactivity. There is a range of other universal compound databases, such as ChemSpider (Pence and Williams, 2010) and ChEBI (Hastings *et al.*, 2016), but also databases focussed on either specific compound classes or use cases. Databases on defined compound classes could for example collect only natural product compounds (Banerjee *et al.*, 2015), human metabolites (Wishart *et al.*, 2007, 2022), lipids (Sud *et al.*, 2007), or compounds of a specific size (Koleti *et al.*, 2018; Stathias *et al.*, 2020). Other databases collect compounds for specific applications, such as pesticides (Liesner, 2003), cancer (*Databases & Tools | Developmental Therapeutics Program (DTP)*, no date) and COVID-19 (Martin *et al.*, 2020). Most databases include data from the others, so while some information is redundant, others are exclusive to only one database. On top of that, of course, all databases are structured and constructed differently. CandActBase, the underlying construct of CandActCFTR, therefore aims at providing a general software basis that can be adapted to individual use cases e.g. collecting information for a specific disease. The required information and their primary sources can be changed according to their relevance, but the overall structure stays the same and can be expanded with new tools. In our use case of CF, the aim of the CandActCFTR database was to provide a collection of all already tested substances with context annotations. This was done to give researchers the possibility to get an easier overview of the current research status, to find compounds and compound classes that are of interest to them, and additionally provide insights into

who is working on which compounds and in which experimental environment. Research communities for different applications are often large and diverse, which is why it can be challenging to stay up to date with all recent developments. Therefore, CandActCFTR links the compounds tested directly to the researchers, which can foster research exchange, as well as facilitate the entry to the field for newcomers.

Currently, CandActCFTR contains 3114 small molecules collected from 111 scientific publications. Of them, 309 compounds have been shown to modulate CFTR activity, while the remaining ones are inactive in this context. While seemingly less important than the active compounds, knowledge of inactive compounds is just as crucial. The chemical space, defined as the entirety of potential pharmacologically active compounds, is thought to be in the order of 10^{60} molecules (Bohacek, McMartin and Guida, 1996; Kirkpatrick and Ellis, 2004). Hence, it is impossible to test all of this inconceivable number of molecules. It is therefore important to exclude areas of the chemical space and thereby narrow down the list of molecules. This can be done by considering inactive compounds, as they indicate which compound classes or areas of the chemical space might be overall inactive in the disease context. However, as often the case in science, it is still rather uncommon to publish negative results, which was also evident when collecting the compounds for the CandActCFTR database. Only 23 of the 111 publications reported their negative result, and for only three of the 19 high-throughput screens lists of the tested compounds were published. As a result, almost 89% of all compounds and 98% of the inactive ones come from only two publications. According to conversations with our peer researchers, the main reason to keep inactive chemical structures unpublished is if they are from newly synthesized compounds, which they will also test for other use cases. In this case, researchers keep the compound structure undisclosed so they do not lose the possibility to file a patent for the compounds for a different application, or disclose the structure to potential competitors. However, this is not the case for high-throughput screens, were thousands of known compounds are used, oftentimes from commercially available compound libraries. The compounds in CandActCFTR were categorized by different properties, which were suggested by our wet-lab collaboration partners. The coarsest categorization is their *influence on CFTR*, which separates the compounds into inactive, inhibiting, enhancing, likely enhancing and

unknown. The other properties categorized are their *order of interaction* with CFTR, the *subcellular compartment* they are active in and the *CFTR relevance*, which give more information on the mechanism of action. The *CFTR relevance* specifies the exact effect the compound has on CFTR or its interactome, e.g. whether it corrects folding or affects the channel activity. The *subcellular localisation* subdivides their site of action into the different compartments, which can also give indications as to which pathway they affect. The *order of interaction* describes whether the compound interacts with CFTR directly or via an intermediary. For almost 70% of the 309 active compounds in the database, the subcellular compartment they are active in, as well as their order of interaction with CFTR are not reported and therefore viewed as unknown. Hence, their mechanism of action is completely unclear. Knowing the mechanism of action of (potential) drugs is essential for a number of purposes. By knowing which pathways and targets are affected by the compound, off-target effects can be predicted and assessed. Furthermore, it facilitates the development of combination therapies, which is of particular interest in CF. Due to the many possible defects mutated CFTR can exhibit at once, several compounds that target different parts of the CFTR biogenesis are required to effectively treat CF by increasing the amount of functional CFTR at the PM. To do so, the mechanism of action of each compound has to be elucidated to efficiently combine them and achieve synergistic effects.

Lastly, it makes finding other compounds that achieve similar effects easier, especially with respect to drug repurposing. If a compound is known to affect a specific pathway or target which are also involved in other disorders, drugs that are already approved can potentially be effective for the disorder in question. This method significantly reduces time and cost in drug discovery processes and has therefore been shifted into focus especially for rare diseases, which are often neglected as they are less economically profitable due to smaller markets. In CF, drug repurposing has so far been considered for delivering antibiotics to the lung as symptomatic treatment (Newman, 2018) and only one study has been conducted on drug repurposing to rescue one variant of CFTR (Orro *et al.*, 2021). Here, 846 clinically approved small-molecule drugs where tested for binding to the F508del CFTR, the most common disease-causing variant of CFTR. This, however, excludes compounds that might modulate CFTR indirectly via a mediator.

7.2 CFTR Lifecycle Map

Most experimental assays to assess the effect of potential CFTR modulators rely on readouts that focus purely on CFTRs Cl^- conductance, such as electrophysiological tools like patch-clamp, or halide-sensitive fluorescence assays. None of these give an indication on whether the modulators achieved their effect by directly influencing channel activity, or by influencing CFTR more early on in its biogenesis. As a result, there are no experimental methods and readouts for the steps prior to channel opening at the membrane.

In order to be able to also predict compounds that achieve their effect indirectly, and thereby target other steps in the biogenesis, as well as to elucidate the mechanism of action for the active compounds in CandActCFTR, we developed a systems biological disease map, which can be seen as a virtual, human- and machine-readable cell model. The CFTR Lifecycle Map makes up the second pillar of the project by covering the biological information important for drug target identification. It contains the pathways CFTR undergoes in its biogenesis as well as the interactions it takes part in. Since there are various kinds of data describing molecular interactions, the CFTR Lifecycle Map is split into different data layers, to neither compromise quality nor quantity. The first layer contains only interactors from small-scale experiments, which offer great detail on the interaction mechanism. This layer is at the core of the model and maps the pathways of the CFTR biogenesis, from its transcription, through its maturation and activity, to its recycling. It is written in the SBGN language *Process Description*, which is directed, sequential and mechanistic (Novère *et al.*, 2009). The second data layer includes interactors of CFTR that have been identified via high-throughput (HT) methods such as immuno-precipitation based proteomic profiling. These methods provide considerably less information on the manner in which the proteins interact, but a lot more interactors can be detected. This second data layer is written in the SBGN *Activity Flow* language, which is non-mechanistic (Novère *et al.*, 2009), since the HT-studies only show that a protein interacts with CFTR, but give no mechanistic insights.

Both model layers are currently static, i.e. no dynamic behaviour can be simulated with them. As described above (1.2.1), in general it is possible to create quantitative ODE-based models if the interaction rate laws are known or can be determined and if

there is a lot of data to support the kinetic parameters. This was not the case for the CFTR Lifecycle Map. At the moment, there are 225 different molecular entities, 170 of which are proteins, involved in 156 reactions in the first data layer. Kinetic data could be found for only very few of them, and if so, it was recorded in very different settings. It was therefore not possible to collect parameter data for a quantitative model, and at the moment it does not appear feasible to study the different stages of the CFTR biogenesis in such high detail and resolution to perform parameter estimation methods due to technical and experimental limitations. In its current state, the interactive CFTR Lifecycle Map, can be used to plan such experiments to generate kinetic data, since it provides a detailed human- and machine-readable overview of the reactions and pathways, as well as the reactants involved.

Qualitative modelling, such as logical models, would be slightly more realistic. However, these also require data for model validation, which again is not available for the CFTR biogenesis at this point. Even so, when in the future such data will be produced, it will be easy to adapt and expand the CFTR Lifecycle Map. This is supported on the one hand by the standardized format and annotation of the model, and on the other hand by the division into individual sub maps, which each focus on a specific part and pathways of the lifecycle.

In general, it can be challenging to create dynamic models from disease maps due to their high number and variety of entities and interactions. Currently, only one dynamic model exists. The Atherosclerosis disease map (Parton *et al.*, 2019) was published in 2019 and is composed of 89 different biological entities, including not only proteins and small molecules, but also different cell types, lipids and platelets. In order to predicts plaque formation in atherosclerosis, the authors developed a quantitative, ODE based model using mass action laws, Michaelis-Menten equations and rate parameters from the literature and the BRENDA enzyme database (Chang *et al.*, 2021b). The authors then estimated the remaining parameters and validated the model using clinical and laboratory data. The atherosclerosis disease map was used to identify five different drugs that might work synergistically to reverse plaque formation. This shows that disease maps can readily be translated to dynamic models and can be used to predict potential drugs. The biggest limitation is the availability of appropriate data to verify the model. Accordingly, a dynamic model will be made out of the CFTR Lifecycle Map, when the required data can be collected.

In November 2021, a second CF centric disease map, called the CyFi-Map, was published by Pereira *et al.* (Pereira *et al.*, 2021). Despite being rather similar to the CFTR Lifecycle Map overall, there are some unique features to both disease maps. While our CFTR Lifecycle Map is composed of two data layers, the detailed core map and the second layer with HT-data, the CyFi-Map does not include data from HT-experiments. However, in contrast to the detailed core map written in the SBGN PD language, the CyFi-Map is written in the non-mechanistic Activity Flow (AF) language. This makes the interactions in the CFTR Lifecycle Map more detailed, while the influence of the interactors on CFTR is more immediately visible in the CyFi-Map. Structure wise, while the CFTR Lifecycle Map can be split into different sub-maps based on the biogenesis stage and the subcellular compartment, the CyFi-Map is available as a whole map. However, the CyFi-Map is composed of two sub-maps based on different CFTR variants. While the main map describes the biogenesis of wt-CFTR, the other one focusses on F508del-CFTR. The CFTR Lifecycle Map, on the other hand, is more generic and shows interactions that have been studied for any variant. Both approaches have advantages and disadvantages. In our generic approach, users have to be more careful to also look at the original publication since not all interactors interact with all variants. However, it is possible to colour code the entities in the disease maps by individual properties. Hence, the mutation specific entities could be displayed in different colours to make it more obvious which interact exclusively with one or the other CFTR variant. Furthermore, the variant specific approach used by Pereira *et al.* excludes other mutations aside from F508del, and is therefore currently not applicable for other CFTR variants. Another difference between the two disease maps is the data basis. For the CFTR Lifecycle Map, we used only interactions that were shown in human cell lines and differentiated between polarizable and non-polarized cells by a colour code in the map. The CyFi-Map also uses studies conducted in non-human cell lines and does not differentiate between polarized and non-polarized cells, which allows them to represent more interactions overall. While the core map of the CFTR Lifecycle Map is based on 221 scientific publications, the CyFi-Map includes data from 297 studies, resulting in 200 interactors in the wt-CFTR submap and approximately 30 additional F508del-CFTR specific ones. Overall, of all genes and proteins in the two maps, 87 occur in both, 83 occur only in the CFTR Lifecycle Map and 104 occur only in the CyFi-Map. Thus,

while the overall structure and pathways are generally the same for both maps, they differ in several points, can be used for different applications and can therefore be viewed as complementary.

7.3 Text mining

The creation of disease maps is very time- and labour intensive. First, one has to get a broad overview of the topic and the pathways involved. Then more detailed studies have to be collected and read to extract the molecular entities and the interactions between them, which can then be written into the model and annotated with further information. This can be made especially challenging by inconsistent naming, the large variety of identifiers and contradictory results. As described in 1.2.2, the largest disease map is currently the COVID-19 disease map, which includes data from 617 publications and to which 230 researchers contributed (Ostaszewski *et al.*, 2021a). This highlights the huge effort involved in creating disease maps. One way this is being addressed in the community is text mining. During text mining the biological entities and their interactions can be computationally extracted from natural language texts written by humans. There is a number of different approaches to text mining and a multitude of algorithms is being developed. In order to support the generic usage of text mining to create disease maps, we developed a tool to integrate any text mining algorithm into the curation process for disease maps. The tool displays the text mining results in a cellular layout similar to a disease map. Each biological entity identified is displayed as a node in their respective subcellular compartment and connected to other nodes by arrows representing the interactions identified from the texts. The user can then iterate through all entities and interactions and display the text passage from which they were extracted. Based on this, they can make the decision to accept and thereby verify the interaction or reject it as falsely identified. Ultimately, the interactions can be exported, shared with project partners and used for creating the final disease map. Since many different text mining algorithms exist, and different researchers have different preferences or develop their own algorithms, we decided to create a generic tool that can theoretically use data from all of them to support the creation of disease maps, when the results can be parsed to the common underlying format.

7.4 In silico *target identification*

The third main objective of this thesis was to bring the two pillars, the chemical database with the active compounds and the systems biological disease map with the potential drug targets, together to gain new insights on both sides and support hypothesis generation.

In order to connect the two sides of the project, we developed another tool to map the compounds from the compound database onto the disease map. The disease map is hosted on a publicly available server using the MINERVA platform. The MINERVA platform was developed by Gawron *et al.* in 2016 specifically for displaying and hosting disease maps (Gawron *et al.*, 2016). It comes with a built-in drug and chemical search, which allows users to search for a drug or chemical by the name of a compound. The compound name is then used to query the databases DrugBank, ChEMBL and the Comparative Toxicogenomic Database (CTD) (Davis *et al.*, 2021) and when interactions with a target in the disease maps are found, the respective targets will be highlighted in the map. In order to integrate custom databases like CandActCFTR, we wrote a new plugin to extend MINERVAs functionality and thus make it more customizable and expandable to new resources. To connect the disease map in MINERVA to the data in the compound database, we queried publicly available databases for known interactions between the molecular entities in the disease map and the chemical compounds. The interaction data is stored in the plugin and the user can either search for compounds that interact with a query target or vice versa. This plugin extends the functionality of the MINERVA built-in compound search in two main ways. First, it offers the reverse search, were compounds that bind to specific targets can be searched. Second, it can be extended with customized databases and further interaction data, be it from other databases or own experimental data.

Integrating interaction data between compounds and targets can give indications on the mechanism of action of the active compounds. By using the publicly available interaction data, possible targets in disease map could be identified for 38 of the compounds with an unknown order of interaction. This corresponds to 18% of the 213 compounds without known order of interaction, leaving 82% and therefore 175 compounds for which potential targets remain unknown.

In order to shed some light on the mechanism of action of all active compounds, *in silico* target identification approaches were used to identify potential interaction partners of all active compounds. As described in chapter 1.3, there are two approaches to *in silico* drug design and target identification, the target- and ligand-based approach. Both of them have certain advantages and disadvantages, and both of them rely on a solid data foundation. Therefore, a complementary dual approach was employed here, to take advantage of both methods and compensate for gaps in the available data. For this purpose, a computational pipeline was set up to extract the targets from the disease maps and the ligands from the database and prepare them for two *in silico* experiments. For the target-based docking approach, appropriate protein structures with a high enough resolution were found for 35 of the proteins in the core map of the CFTR Lifecycle Map. For the ligand-based approach, ligands were collected from publicly available databases. Ligands could be found for 90 proteins, 22 of which overlapping with the list of proteins for the target-based approach. An automated pipeline was set up to first analyse the resulting data from both approaches separately and subsequently compare them. By doing so, potential targets for all 309 active compounds could be suggested by at least one of the approaches, and overall 1038 unique protein-compound pairings could be suggested. The compounds were assigned confidence scores from one to five according to how consistent the results were amongst the different approaches.

The inverse screening approaches used here are adaptations and combinatorial extensions of the classical *in silico* screening approaches, which are much more common in drug discovery. Usually, a specific target of interest is already known and large compound libraries are tested against it to see which compounds might exhibit the desired activity. In this approach, there was a number of potential targets and a small library of compounds that were already known to be active. This brings about new challenges and possible sources for errors and inaccuracies. In the case of the target-based approach, the high number of targets and not knowing the binding pockets in advance make it more difficult to find the correct binding site. To address this, blind docking was used, where the entire protein structure is sampled for likely binding sites and poses. To yield meaningful results, it was therefore necessary to use a higher exhaustiveness than in traditional focussed docking, i.e. a higher number of poses is tested for each target-ligand pairing. This results in much higher

computational costs, hence it becomes unfeasible if the amount of ligands and target exceeds a certain number. The number of approximately 300 ligands and 35 targets used here was well within the limits, however if there are hundreds of potential targets and thousands of ligands, this approach cannot be applied without huge computation resources. Another source of inaccuracies of the target-based approach are promiscuous interactors. One challenge of traditional drug discovery are the so-called Pan-assay interference compounds (PAINS), which are compounds that bind to many targets non-specifically. In reverse docking, this can also apply to the targets. Certain proteins, or rather protein binding sites, can also exhibit high promiscuity and bind non-specifically to a range of compounds (Cerisier *et al.*, 2019; Ehrt, Brinkjost and Koch, 2019; Hanson *et al.*, 2019). This can lead to compounds being assigned to promiscuous protein structures rather than their actual target. However, there are different statistical methods to filter unspecific pairings caused by both, PAINS and promiscuous binding sites, from the docking data, which were employed here. Nonetheless, it is important to note that even specific protein-ligand interactions do not automatically mean that a compound is active, as it can also bind to a protein without affecting its activity in a measurable way.

The main limitation in target-based inverse screening is the availability of protein structures. Since not all possible targets from the disease maps have a high-quality resolved structure, this restricts the number of targets that can be used for the docking. This restriction can be theoretically addressed by the recent advances in machine learning based structure prediction such as DeepMinds AlphaFold (Jumper *et al.*, 2021; Varadi *et al.*, 2022). However, when testing the structures predicted by AlphaFold against their respective experimental structures with our docking approach, the results for the experimental structures were significantly better and more accurate. This is possible due to using blind docking rather that focussed docking. Especially the more flexible (terminal) regions of the proteins are often not resolved correctly by AlphaFold (Jumper *et al.*, 2021). This may be irrelevant for focussed docking, where the binding site is well defined, but can pose a problem for blind docking, as the flexible regions can interfere with searching the entire protein structure. Furthermore, AlphaFold does not predict all protein structures equally well, since it relies on a data basis of experimentally resolved structures, where certain protein classes, such as the difficult to handle membrane proteins, are underrepresented due them being

challenging to resolve structurally. Predicted structures were therefore not included in the protein library for the target-based approach. Hence, it is possible that the target of some compounds is not amongst those tested here, or that compounds are assigned to the wrong target.

The same limitation applies also to the ligand-based approach, where not for all targets a list of known ligands is available. This excludes proteins, especially those which have not been extensively studied in the past, from being considered as potential targets. Additionally, some proteins may only have very few known ligands, which makes it unlikely for them to be identified as a target for a specific compound, since there are too few reference ligands to compare the query ligands to. On the other hand, this results in a large bias towards very well studied proteins, for example those that have been considered as targets for other disorders. As these often have a high number of ligands already associated to them, it is more likely to also find ligands that share a higher similarity with the query compounds. Additionally, again, promiscuous proteins and ligands can lead to non-specific pairings being identified. Hence, both the target- and the ligand-based approach can lead to unspecific pairings and the results should be viewed as potential interactions to narrow down the possible targets of a compound. Interactions of interest should be studied in depth and ideally confirmed experimentally. Furthermore, binding does not automatically translate to activity, so activity assays have to be performed to confirm proposed mechanisms of action. Nonetheless, these *in silico* methods can greatly reduce time and cost of target identification and support hypothesis development. Especially when combining the target- and ligand-based approach, which, to the best of our knowledge, this study is the first to do, shortcomings of both methods can be circumvented and limitations reduced.

Again, this target identification approach was applied to the example of CF, bringing together the chemical data from the chemical CandActCFTR database and the biological CFTR Lifecycle Map. However, this approach can be applied in the same manner for other use cases and diseases as well. Depending on the research question and the data available, the methods can be adapted to fit the conditions. For example, if the structures of most of the potential targets are resolved, the emphasis could be placed on the target-based identification approach. On the other hand, if the number of ligands and proteins were to large, one might focus on the ligand-based approach

and only use the target-based approach where no ligand data is available to reduce computational costs. Interactions identified that appear promising in the disease context could also be studied further by more detailed and extensive docking or molecular dynamics to refine the results. Additionally, if available, the *a priori* data and knowledge generated *in silico* could always be extended by experimental data, which would greatly improve the data basis.

Chapter 8 Summary & Conclusion

The ever-increasing amounts of biological data and knowledge being produced, and the sophisticated bioinformatics tools to analyse them offer a wide range of possibilities in the field of systems medicine. Nonetheless, oftentimes, the data is not used to its full potential, as it is being analysed separately and with highly specific methods. However, integrating different kinds of data in a reproducible way is imperative to make full use of the potential of systems medicine.

For this purpose, it is necessary to generate an overview over the different aspects of a topic to then derive novel insights.

Therefore, in this thesis, the main objective was to bring together various data sources in a modular way to gain a holistic view on the molecular biology underlying a disease and thereby support treatment development. This was developed and demonstrated at the use case of Cystic Fibrosis. Cystic Fibrosis is a hereditary disease with a wide range of geno- and phenotypes, which make it difficult to treat causatively. To predict compounds and synergistic compound combinations as candidate therapeutics, a set of generic tools and methodologies was developed and applied. The project is built on two main data sources, chemical data from compounds that have been tested as modulators in the CF context, and biological interaction and pathway data from detailed small-scale experiments and high-throughput efforts. These two data collections were prepared and made available as independent resources and then linked and brought together to gain new insights on drug targets in CF and the mechanism of action of active compounds. Specifically, the tested substances were collected in a compound database and the molecular pathways and interactions underlying the disease were modelled in a computational disease map. To support the otherwise time-consuming manual creation and upkeep of disease maps, a tool was developed that provides a computational link between text mining methods and disease maps. In order to then link the compound database and the disease map, an interactive plugin was created to map the compounds to their potential targets based on available data. Since the available data on their bioactivity did not cover all active molecules, target- and ligand-based inverse screening were used complementary to identify potential targets of the active compounds. This data will

serve as basis to suggest novel compounds and compound combinations for testing in the wet-lab.

Overall, each module of the project can be viewed independently as their own resource or, as demonstrated here, brought together to provide new insights. Thereby, each part can be customized according to the specific application and different data sources can be integrated, depending on what is available. Generally speaking, the better the data foundation, the greater the added value. However, this project relies solely on already published open-source data. This shows that it is possible to also gain knowledge if the data foundation is scarce in certain areas and different kinds of data can be substituted by others. For example, missing experimental protein-ligand interaction data can be substituted by virtual screening data, and interactive, expandable systems biology models can be used to understand biological mechanisms, generate hypotheses and plan experiments accordingly. This is possible through the use of generic methodologies, which are flexible and adaptable to the data at hand. While complex and sophisticated bioinformatics methods certainly provide great possibilities for systems medicine, they are also often highly specific and require certain data types and formats to produce meaningful results, which excludes them from many applications. To integrate different data sources and benefit from the added value, it is therefore essential to employ replicable and adaptable tools and methods and adhere to community data standard. These approaches can also be used and adjusted to other diseases.

In conclusion, the generic and adaptable systems medicine tools developed and used within this thesis allow researchers a thorough and holistic view on the molecular basis of diseases via integration and recycling of different sources of data and knowledge. The applicability of the tools was shown on CF as an exemplary disease, demonstrating how this approach can be used for other diseases and use cases.

Chapter 9 References

Adler, F. R. *et al.* (2009) 'Lung transplantation for cystic fibrosis.', *Proceedings of the American Thoracic Society*, 6(8), pp. 619–33. doi: 10.1513/pats.2009008-088TL.

Aleksandrov, A. A., Aleksandrov, L. A. and Riordan, J. R. (2007) 'CFTR (ABCC7) is a hydrolyzable-ligand-gated channel', *Pflugers Archiv European Journal of Physiology*. doi: 10.1007/s00424-006-0140-z.

AlphaFold Protein Structure Database (no date). Available at: https://alphafold.ebi.ac.uk/ (Accessed: 3 June 2022).

Amaral, M. D. and Beekman, J. M. (2020) 'Activating alternative chloride channels to treat CF: Friends or Foes?', *Journal of Cystic Fibrosis*, 19(1), pp. 11–15. doi: 10.1016/j.jcf.2019.10.005.

Ameen, N., Silvis, M. and Bradbury, N. A. (2007) 'Endocytic trafficking of CFTR in health and disease.', *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 6(1), pp. 1–14. doi: 10.1016/j.jcf.2006.09.002.

Ananiadou, S., Kell, D. B. and Tsujii, J. (2006) 'Text mining and its potential applications in systems biology.', *Trends in biotechnology*, 24(12), pp. 571–9. doi: 10.1016/j.tibtech.2006.10.002.

Bajusz, D., Rácz, A. and Héberger, K. (2015) 'Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?', *Journal of Cheminformatics*. BioMed Central Ltd., 7(1), p. 20. doi: 10.1186/s13321-015-0069-3.

Ban, F. *et al.* (2021) 'Inverse molecular docking reveals a novel function of thymol: Inhibition of fat deposition induced by high-dose glucose in Caenorhabditis elegans', *Food Science and Nutrition*. John Wiley and Sons Inc, 9(8), pp. 4243–4253. doi: 10.1002/fsn3.2392.

Banerjee, P. *et al.* (2015) 'Super Natural II-a database of natural products', *Nucleic Acids Research*. Oxford University Press, 43(D1), pp. D935–D939. doi: 10.1093/nar/gku886.

Batool, M., Ahmad, B. and Choi, S. (2019) 'A structure-based drug discovery paradigm', *International Journal of Molecular Sciences*. MDPI AG. doi: 10.3390/ijms20112783.

Bell, S. C. *et al.* (2020) 'The future of cystic fibrosis care: a global perspective', *The Lancet Respiratory Medicine*. Lancet Publishing Group, pp. 65–124. doi: 10.1016/S2213-2600(19)30337-6.

Berg, A. *et al.* (2019) 'High-Throughput Surface Liquid Absorption and Secretion Assays to Identify F508del CFTR Correctors Using Patient Primary Airway Epithelial Cultures', *SLAS Discovery*. doi: 10.1177/2472555219849375.

Berthold, M. R. *et al.* (2009) 'KNIME - the Konstanz information miner', *ACM SIGKDD*

Explorations Newsletter, 11(1), pp. 26–31. doi: 10.1145/1656274.1656280.

Bertrand, C. A. and Frizzell, R. A. (2003) 'The role of regulated CFTR trafficking in epithelial secretion', *American Journal of Physiology - Cell Physiology*. doi: 10.1152/ajpcell.00554.2002.

BioCreative - Latest 3 News Items (no date). Available at: https://biocreative.bioinformatics.udel.edu/ (Accessed: 12 July 2022).

Bobadilla, J. L. *et al.* (2002) 'Cystic fibrosis: A worldwide analysis of CFTR mutations - Correlation with incidence data and application to screening', *Human Mutation*. doi: 10.1002/humu.10041.

Bohacek, R. S., McMartin, C. and Guida, W. C. (1996) 'The art and practice of structure-based drug design: A molecular modeling perspective', *Medicinal Research Reviews*. Med Res Rev, pp. 3–50. doi: 10.1002/(SICI)1098-1128(199601)16:1<3::AID-MED1>3.0.CO;2-6.

Brooijmans, N. and Kuntz, I. D. (2003) 'Molecular recognition and docking algorithms', *Annual Review of Biophysics and Biomolecular Structure*. Annu Rev Biophys Biomol Struct, pp. 335–373. doi: 10.1146/annurev.biophys.32.110601.142532.

Buendia-Atencio, C. *et al.* (2021) 'Inverse Molecular Docking Study of NS3-Helicase and NS5-RNA Polymerase of Zika Virus as Possible Therapeutic Targets of Ligands Derived from Marcetia taxifolia and Its Implications to Dengue Virus', *ACS Omega*. American Chemical Society, 6(9), pp. 6134–6143. doi: 10.1021/acsomega.0c04719.

Busch, R. (1979) 'Zur Frühgeschichte der zystischen Pankreasfibromatose.(Sur les débuts de l'histoire du fibrome cystique du pancréas)', *pascal-francis.inist.fr*. Available at: https://pascal-francis.inist.fr/vibad/index.php?action=getRecordDetail&idt=12661286 (Accessed: 18 August 2022).

Byrne, R. and Schneider, G. (2019) 'In silico target prediction for small molecules', in *Methods in Molecular Biology*. Humana Press Inc., pp. 273–309. doi: 10.1007/978-1-4939-8891-4_16.

Callebaut, I., Chong, P. A. and Forman-Kay, J. D. (2018) 'CFTR structure', *Journal of Cystic Fibrosis*, 17(2), pp. S5–S8. doi: 10.1016/j.jcf.2017.08.008.

Carlile, G. W. *et al.* (2015) 'Ibuprofen rescues mutant cystic fibrosis transmembrane conductance regulator trafficking', *Journal of Cystic Fibrosis*. doi: 10.1016/j.jcf.2014.06.001.

Cereto-Massagué, A. *et al.* (2015) 'Molecular fingerprint similarity search in virtual screening', *Methods*. Academic Press Inc., 71(C), pp. 58–63. doi: 10.1016/j.ymeth.2014.08.005.

Cerisier, N. *et al.* (2019) 'High impact: The role of promiscuous binding sites in polypharmacology', *Molecules*. MDPI AG, 24(14). doi: 10.3390/molecules24142529.

CF-Map (no date). Available at: https://cf-map.uni-goettingen.de/ (Accessed: 29 June 2022).

Chang, A. *et al.* (2021a) 'BRENDA, the ELIXIR core data resource in 2021: new developments and

updates.', *Nucleic acids research*, 49(D1), pp. D498–D508. doi: 10.1093/nar/gkaa1025.

Chang, A. *et al.* (2021b) 'BRENDA, the ELIXIR core data resource in 2021: New developments and updates', *Nucleic Acids Research*. Oxford University Press, 49(D1), pp. D498–D508. doi: 10.1093/nar/gkaa1025.

Chassagnole, C. *et al.* (2002) 'Dynamic modeling of the central carbon metabolism of Escherichia coli.', *Biotechnology and bioengineering*, 79(1), pp. 53–73. doi: 10.1002/bit.10288.

CHEMBL database release 30 (2022). doi: 10.6019/CHEMBL.database.30.

Chen, Y. Z. and Zhi, D. G. (2001) 'Ligand - Protein inverse docking and its potential use in the computer search of protein targets of a small molecule', *Proteins: Structure, Function and Genetics*, 43(2), pp. 217–226. doi: 10.1002/1097-0134(20010501)43:2<217::AID-PROT1032>3.0.CO;2-G.

Cheng, S. H. *et al.* (1990) 'Defective intracellular transport and processing of CFTR is the molecular basis of most cystic fibrosis', *Cell*. Cell, 63(4), pp. 827–834. doi: 10.1016/0092-8674(90)90148-8.

Clancy, J. P. *et al.* (2012) '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*. BMJ Publishing Group, 67(1), pp. 12–18. doi: 10.1136/thoraxjnl-2011-200393.

Clinical Pipeline (no date). Available at: https://www.glpg.com/clinical-pipelines (Accessed: 28 June 2022).

Csanády, L., Vergani, P. and Gadsby, D. C. (2019) 'Structure, Gating, and Regulation of the CFTR Anion Channel', *Physiological Reviews*, 99(1), pp. 707–738. doi: 10.1152/physrev.00007.2018.

Cystic Fibrosis Mutation Database (no date). Available at: http://www.genet.sickkids.on.ca/ (Accessed: 26 January 2021).

Data Analytics Platform: Open Source Software Tools | KNIME (no date). Available at: https://www.knime.com/knime-analytics-platform (Accessed: 31 May 2022).

Databases & Tools | Developmental Therapeutics Program (DTP) (no date). Available at: https://dtp.cancer.gov/databases_tools/default.htm (Accessed: 11 August 2022).

Davies, M. *et al.* (2015) 'ChEMBL web services: streamlining access to drug discovery data and utilities', *Nucleic Acids Research*, 43(W1), pp. W612–W620. doi: 10.1093/nar/gkv352.

Davis, A. P. *et al.* (2021) 'Comparative Toxicogenomics Database (CTD): Update 2021', *Nucleic Acids Research*. doi: 10.1093/nar/gkaa891.

Dechecchi, M. C., Tamanini, A. and Cabrini, G. (2018) 'Molecular basis of cystic fibrosis: from bench to bedside', *Annals of Translational Medicine*. AME Publishing Company, 6(17), pp. 334–334. doi: 10.21037/atm.2018.06.48.

Drug Development Pipeline | CFF Clinical Trials Tool (no date). Available at:

https://www.cff.org/Trials/Pipeline (Accessed: 26 January 2021).

Ehrt, C., Brinkjost, T. and Koch, O. (2019) 'Binding site characterization-similarity, promiscuity, and druggability', *MedChemComm*. Royal Society of Chemistry, 10(7), pp. 1145–1159. doi: 10.1039/c9md00102f.

Elborn, J. S. (2016) 'Cystic fibrosis', *The Lancet*. Lancet Publishing Group, pp. 2519–2531. doi: 10.1016/S0140-6736(16)00576-6.

Emmert-Streib, F. and Dehmer, M. (2011) 'Networks for systems biology: Conceptual connection of data and function', *IET Systems Biology*. IET Syst Biol, 5(3), pp. 185–207. doi: 10.1049/iet-syb.2010.0025.

Farinha, C. M. *et al.* (2002) '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*. doi: 10.1042/BJ20011717.

Farinha, C. M. and Amaral, M. D. (2005) 'Most F508del-CFTR Is Targeted to Degradation at an Early Folding Checkpoint and Independently of Calnexin', *Molecular and Cellular Biology*. doi: 10.1128/mcb.25.12.5242-5252.2005.

Farinha, C. M. and Canato, S. (2017) 'From the endoplasmic reticulum to the plasma membrane: mechanisms of CFTR folding and trafficking', *Cellular and Molecular Life Sciences*. doi: 10.1007/s00018-016-2387-7.

Farinha, C. M. and Matos, P. (2018) 'Rab GTPases regulate the trafficking of channels and transporters - a focus on cystic fibrosis.', *Small GTPases*, 9(1–2), pp. 136–144. doi: 10.1080/21541248.2017.1317700.

Farinha, C. M., Matos, P. and Amaral, M. D. (2013) 'Control of cystic fibrosis transmembrane conductance regulator membrane trafficking: not just from the endoplasmic reticulum to the Golgi', *FEBS Journal*. John Wiley & Sons, Ltd, 280(18), pp. 4396–4406. doi: 10.1111/febs.12392.

Farrell, P. M. (2008) 'The prevalence of cystic fibrosis in the European Union.', *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 7(5), pp. 450–3. doi: 10.1016/j.jcf.2008.03.007.

Fiedorczuk, K. and Chen, J. (2022) 'Mechanism of CFTR correction by type I folding correctors', *Cell*. Elsevier B.V., 185(1), pp. 158-168.e11. doi: 10.1016/j.cell.2021.12.009.

Fluck, J. and Hofmann-Apitius, M. (2014) 'Text mining for systems biology.', *Drug discovery today*, 19(2), pp. 140–4. doi: 10.1016/j.drudis.2013.09.012.

Franz, M. *et al.* (2015) 'Cytoscape.js: a graph theory library for visualisation and analysis', *Bioinformatics*, p. btv557. doi: 10.1093/bioinformatics/btv557.

Fujita, K. A. *et al.* (2014) 'Integrating pathways of Parkinson's disease in a molecular interaction

map.', *Molecular neurobiology*, 49(1), pp. 88–102. doi: 10.1007/s12035-013-8489-4.

Gawron, P. *et al.* (2016) 'MINERVA—A platform for visualization and curation of molecular interaction networks', *npj Systems Biology and Applications*. doi: 10.1038/npjsba.2016.20.

Gentzsch, M. and Mall, M. A. (2018) 'Ion Channel Modulators in Cystic Fibrosis', *Chest*. Elsevier Inc, pp. 383–393. doi: 10.1016/j.chest.2018.04.036.

Gillen, A. E. (2012) 'Transcriptional regulation of CFTR gene expression', *Frontiers in Bioscience*. Front Biosci (Elite Ed), E4(1), p. 587. doi: 10.2741/401.

Gilson, M. K. *et al.* (2016) 'BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D1045–D1053. doi: 10.1093/nar/gkv1072.

Giuliano, K. A. *et al.* (2018) 'Use of a High-Throughput Phenotypic Screening Strategy to Identify Amplifiers, a Novel Pharmacological Class of Small Molecules That Exhibit Functional Synergy with Potentiators and Correctors', *SLAS Discovery*. doi: 10.1177/2472555217729790.

Goetz, D. M. and Savant, A. P. (2021) 'Review of CFTR modulators 2020', *Pediatric Pulmonology*. John Wiley and Sons Inc, pp. 3595–3606. doi: 10.1002/ppul.25627.

Goll, M. (2020) 'Asynchronous JavaScript and XML **', in *JavaServer Faces*. Wiesbaden: Springer Fachmedien Wiesbaden, pp. 189–196. doi: 10.1007/978-3-658-31803-1_20.

Van Goor, F. *et al.* (2009) 'Rescue of CF airway epithelial cell function in vitro by a CFTR potentiator, VX-770', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0904709106.

Van Goor, F. *et al.* (2011) 'Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1105787108.

Gorgulla, C. *et al.* (2020) 'An open-source drug discovery platform enables ultra-large virtual screens', *Nature*. Nature Research, 580(7805), pp. 663–668. doi: 10.1038/s41586-020-2117-z.

Grove, D. E. *et al.* (2011) 'The endoplasmic reticulum-associated Hsp40 DNAJB12 and Hsc70 cooperate to facilitate RMA1 E3-dependent degradation of nascent CFTRDeltaF508.', *Molecular biology of the cell*, 22(3), pp. 301–14. doi: 10.1091/mbc.E10-09-0760.

Guedes, I. A., Pereira, F. S. S. and Dardenne, L. E. (2018) 'Empirical Scoring Functions for Structure-Based Virtual Screening: Applications, Critical Aspects, and Challenges', *Frontiers in Pharmacology*, 9. doi: 10.3389/fphar.2018.01089.

Halperin, I. *et al.* (2002) 'Principles of docking: An overview of search algorithms and a guide to scoring functions', *Proteins: Structure, Function and Genetics*, pp. 409–443. doi: 10.1002/prot.10115.

Hammond, C., Braakman, I. and Helenius, A. (1994) 'Role of N-linked oligosaccharide recognition, glucose trimming, and calnexin in glycoprotein folding and quality control', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 91(3), pp. 913–917. doi: 10.1073/pnas.91.3.913.

Hanson, S. M. *et al.* (2019) 'What Makes a Kinase Promiscuous for Inhibitors?', *Cell Chemical Biology*. Elsevier Ltd, 26(3), pp. 390-399.e5. doi: 10.1016/j.chembiol.2018.11.005.

Harmston, N., Filsell, W. and Stumpf, M. P. H. (2010) 'What the papers say: text mining for genomics and systems biology.', *Human genomics*, 5(1), pp. 17–29. doi: 10.1186/1479-7364-5-1-17.

Harris, C. R. *et al.* (2020) 'Array programming with NumPy', *Nature*, 585(7825), pp. 357–362. doi: 10.1038/s41586-020-2649-2.

Hassan, N. M. *et al.* (2017) 'Protein-Ligand Blind Docking Using QuickVina-W with Inter-Process Spatio-Temporal Integration', *Scientific Reports*. Nature Publishing Group, 7(1), pp. 1–13. doi: 10.1038/s41598-017-15571-7.

Hastings, J. *et al.* (2016) 'ChEBI in 2016: Improved services and an expanding collection of metabolites', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1031.

Hearst, M. A. (1999) 'Untangling text data mining', in *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics* -. Morristown, NJ, USA: Association for Computational Linguistics, pp. 3–10. doi: 10.3115/1034678.1034679.

Hetényi, C. and van der Spoel, D. (2009) 'Efficient docking of peptides to proteins without prior knowledge of the binding site', *Protein Science*. Wiley, 11(7), pp. 1729–1737. doi: 10.1110/ps.0202302.

Hetényi, C. and Van Der Spoel, D. (2006) 'Blind docking of drug-sized compounds to proteins with up to a thousand residues', *FEBS Letters*. FEBS Lett, 580(5), pp. 1447–1450. doi: 10.1016/j.febslet.2006.01.074.

Higgins, C. (1989) 'Export-import family expands', *Nature*. Nature, p. 342. doi: 10.1038/340342a0.

Higgins, C. F. *et al.* (1988) 'A family of closely related ATP-binding subunits from prokaryotic and eukaryotic cells', *BioEssays*. Bioessays, pp. 111–116. doi: 10.1002/bies.950080406.

Hoksza, D. *et al.* (2019) 'MINERVA API and plugins: opening molecular network analysis and visualization to the community', *Bioinformatics*. Edited by L. Cowen, 35(21), pp. 4496–4498. doi: 10.1093/bioinformatics/btz286.

Holzhütter, H. *et al.* (2012) 'The virtual liver: a multidisciplinary, multilevel challenge for systems biology', *WIREs Systems Biology and Medicine*, 4(3), pp. 221–235. doi: 10.1002/wsbm.1158.

Hou, X., Wu, Q., Rajagopalan, C., Zhang, C., Bouhamdan, M., Wei, H., Chen, X., Zaman, K., Li,

C., Sun, X., Chen, S., Frizzell, Raymond A, *et al.* (2019) 'CK19 stabilizes CFTR at the cell surface by limiting its endocytic pathway degradation.', *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 33(11), pp. 12602–12615. doi: 10.1096/fj.201901050R.

Hou, X., Wu, Q., Rajagopalan, C., Zhang, C., Bouhamdan, M., Wei, H., Chen, X., Zaman, K., Li, C., Sun, X., Chen, S., Frizzell, Raymond A., *et al.* (2019) 'CK19 stabilizes CFTR at the cell surface by limiting its endocytic pathway degradation', *The FASEB Journal*. John Wiley and Sons Inc., 33(11), pp. 12602–12615. doi: 10.1096/fj.201901050R.

Huang, H. *et al.* (2018) 'Reverse screening methods to search for the protein targets of chemopreventive compounds', *Frontiers in Chemistry*. Frontiers Media S. A, p. 138. doi: 10.3389/fchem.2018.00138.

Hucka, M. *et al.* (2003) 'The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models', *Bioinformatics*. Bioinformatics, 19(4), pp. 524–531. doi: 10.1093/bioinformatics/btg015.

Ideker, T., Galitski, T. and Hood, L. (2001) 'A new approach to decoding life: systems biology.', *Annual review of genomics and human genetics*, 2, pp. 343–72. doi: 10.1146/annurev.genom.2.1.343.

Jumper, J. *et al.* (2021) 'Highly accurate protein structure prediction with AlphaFold', *Nature*. Nature Research, 596(7873), pp. 583–589. doi: 10.1038/s41586-021-03819-2.

Kim, S. *et al.* (2016) 'PubChem substance and compound databases', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D1202–D1213. doi: 10.1093/nar/gkv951.

Kim, S. *et al.* (2021) 'PubChem in 2021: new data content and improved web interfaces', *Nucleic Acids Research*, 49(D1), pp. D1388–D1395. doi: 10.1093/nar/gkaa971.

Kim, S. J. and Skach, W. R. (2012) 'Mechanisms of CFTR Folding at the Endoplasmic Reticulum', *Frontiers in Pharmacology*. doi: 10.3389/fphar.2012.00201.

Kim, S. S., Aprahamian, M. L. and Lindert, S. (2019) 'Improving inverse docking target identification with Z-score selection', *Chemical Biology and Drug Design*. Blackwell Publishing Ltd, 93(6), pp. 1105–1116. doi: 10.1111/cbdd.13453.

Kirkpatrick, P. and Ellis, C. (2004) 'Chemical space', *Nature*, 432(7019), pp. 823–823. doi: 10.1038/432823a.

Kitano, H. (2002) 'Computational systems biology.', *Nature*, 420(6912), pp. 206–10. doi: 10.1038/nature01254.

Koch, I. (2015) 'Petri nets in systems biology', *Software and Systems Modeling*. Springer Verlag, 14(2), pp. 703–710. doi: 10.1007/s10270-014-0421-5.

Koes, D. R., Baumgartner, M. P. and Camacho, C. J. (2013) 'Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise', *Journal of Chemical Information and Modeling*. American Chemical Society, 53(8), pp. 1893–1904. doi: 10.1021/ci300604z.

Koleti, A. *et al.* (2018) 'Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: Integrated access to diverse large-scale cellular perturbation response data', *Nucleic Acids Research*. Oxford University Press, 46(D1), pp. D558–D566. doi: 10.1093/nar/gkx1063.

König, M. (2020) 'matthiaskoenig/libsbgn-python: libsbgn-python-v0.2.0'. Zenodo. doi: 10.5281/zenodo.3712285.

Konstan, M. W. *et al.* (2020) 'Efficacy and safety of ataluren in patients with nonsense-mutation cystic fibrosis not receiving chronic inhaled aminoglycosides: The international, randomized, double-blind, placebo-controlled Ataluren Confirmatory Trial in Cystic Fibrosis (ACT CF).', *Journal of cystic fibrosis : official journal of the European Cystic Fibrosis Society*, 19(4), pp. 595–601. doi: 10.1016/j.jcf.2020.01.007.

Kumar, A. and Zhang, K. Y. J. (2018) 'Advances in the development of shape similarity methods and their application in drug discovery', *Frontiers in Chemistry*. Frontiers Media S.A., p. 315. doi: 10.3389/fchem.2018.00315.

Kuperstein, I. *et al.* (2015) 'Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps.', *Oncogenesis*, 4, p. e160. doi: 10.1038/oncsis.2015.19.

Lauro, G. *et al.* (2011) 'Inverse virtual screening of antitumor targets: Pilot study on a small database of natural bioactive compounds', *Journal of Natural Products*. American Chemical Society and American Society of Pharmacognosy, 74(6), pp. 1401–1407. doi: 10.1021/np100935s.

Levy, L. *et al.* (1986) 'Prognostic factors associated with patient survival during nutritional rehabilitation in malnourished children and adolescents with cystic fibrosis.', *Journal of pediatric gastroenterology and nutrition*, 5(1), pp. 97–102. doi: 10.1097/00005176-198601000-00018.

Li, J., Fu, A. and Zhang, L. (2019) 'An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking', *Interdisciplinary Sciences: Computational Life Sciences*, 11(2), pp. 320–328. doi: 10.1007/s12539-019-00327-w.

Li, X. *et al.* (2010) 'Evaluation of the performance of four molecular docking programs on a diverse set of protein-ligand complexes', *Journal of Computational Chemistry*. John Wiley & Sons, Ltd, 31(11), pp. 2109–2125. doi: 10.1002/jcc.21498.

Liang, F. *et al.* (2017) 'High-Throughput Screening for Readthrough Modulators of CFTR PTC Mutations', *SLAS Technology*. doi: 10.1177/2472630317692561.

Liesner, M. (2003) ‘Compendium of Pesticide Common Names’, *Chemie in unserer Zeit*. John Wiley & Sons, Ltd, 37(2), pp. 155–155. doi: 10.1002/ciuz.200390027.

Lim, T. G. *et al.* (2014) ‘Curcumin suppresses proliferation of colon cancer cells by targeting CDK2’, *Cancer Prevention Research*. American Association for Cancer Research Inc., 7(4), pp. 466–474. doi: 10.1158/1940-6207.CAPR-13-0387.

Lopes-Pacheco, M. (2020) ‘CFTR Modulators: The Changing Face of Cystic Fibrosis in the Era of Precision Medicine’, *Frontiers in Pharmacology*. Frontiers Media S.A. doi: 10.3389/fphar.2019.01662.

Lu, Y. *et al.* (1998) ‘Co- and posttranslational translocation mechanisms direct cystic fibrosis transmembrane conductance regulator N terminus transmembrane assembly’, *Journal of Biological Chemistry*. doi: 10.1074/jbc.273.1.568.

Lukacs, G. L. *et al.* (1994) ‘Conformational maturation of CFTR but not its mutant counterpart (delta F508) occurs in the endoplasmic reticulum and requires ATP.’, *The EMBO Journal*. doi: 10.1002/j.1460-2075.1994.tb06954.x.

Machado, D. *et al.* (2011) ‘Modeling formalisms in Systems Biology.’, *AMB Express*, 1, p. 45. doi: 10.1186/2191-0855-1-45.

Maia, E. H. B. *et al.* (2020) ‘Structure-Based Virtual Screening: From Classical to Artificial Intelligence’, *Frontiers in Chemistry*. Frontiers Media S.A., p. 343. doi: 10.3389/fchem.2020.00343.

Mall, M. A. and Galietta, L. J. V. (2015) ‘Targeting ion channels in cystic fibrosis’, *Journal of Cystic Fibrosis*. doi: 10.1016/j.jcf.2015.06.002.

Markram, H. (2006) ‘Biology---The blue brain project’, in *Proceedings of the 2006 ACM/IEEE conference on Supercomputing - SC ’06*. New York, New York, USA: ACM Press, p. 53. doi: 10.1145/1188455.1188511.

Martin, R. *et al.* (2020) ‘CORDITE: The Curated CORona Drug InTERactions Database for SARS-CoV-2’, *iScience*. Elsevier Inc., 23(7). doi: 10.1016/j.isci.2020.101297.

Martiniano, S. L., Sagel, S. D. and Zemanick, E. T. (2016) ‘Cystic fibrosis: A model system for precision medicine’, *Current Opinion in Pediatrics*. Lippincott Williams and Wilkins, pp. 312–317. doi: 10.1097/MOP.0000000000000351.

Mathai, N. and Kirchmair, J. (2020) ‘Similarity-based methods and machine learning approaches for target prediction in early drug discovery: Performance and scope’, *International Journal of Molecular Sciences*. MDPI AG, 21(10). doi: 10.3390/ijms21103585.

Matthews, H., Hanison, J. and Nirmalan, N. (2016) ‘“Omics”-Informed Drug and Biomarker Discovery: Opportunities, Challenges and Future Perspectives’, *Proteomes*, 4(3), p. 28. doi:

10.3390/proteomes4030028.

Mazein, A. *et al.* (2018) 'Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms', *npj Systems Biology and Applications*. doi: 10.1038/s41540-018-0059-y.

Mazein, A. *et al.* (2021) 'AsthmaMap: An interactive knowledge repository for mechanisms of asthma.', *The Journal of allergy and clinical immunology*, 147(3), pp. 853–856. doi: 10.1016/j.jaci.2020.11.032.

McInnes, C. (2007) 'Virtual screening strategies in drug discovery', *Current Opinion in Chemical Biology*, 11(5), pp. 494–502. doi: 10.1016/j.cbpa.2007.08.033.

McKinney, W. (2010) 'Data Structures for Statistical Computing in Python', in, pp. 56–61. doi: 10.25080/Majora-92bf1922-00a.

Meacham, G. C. *et al.* (1999) *The Hdj-2/Hsc70 chaperone pair facilitates early steps in CFTR biogenesis*, *The EMBO Journal*.

Mendez, D. *et al.* (2019) 'ChEMBL: towards direct deposition of bioassay data', *Nucleic Acids Research*, 47(D1), pp. D930–D940. doi: 10.1093/nar/gky1075.

Meng, X. *et al.* (2019) 'CFTR structure, stability, function and regulation', *Biological Chemistry*, 400(10), pp. 1359–1370. doi: 10.1515/hsz-2018-0470.

Merkert, S. *et al.* (2019) 'High-Throughput Screening for Modulators of CFTR Activity Based on Genetically Engineered Cystic Fibrosis Disease-Specific iPSCs', *Stem Cell Reports*. doi: 10.1016/j.stemcr.2019.04.014.

Mi, H. *et al.* (2015) 'Systems Biology Graphical Notation: Activity Flow language Level 1 Version 1.2.', *Journal of integrative bioinformatics*, 12(2), p. 265. doi: 10.2390/biecoll-jib-2015-265.

Michelsen, K., Yuan, H. and Schwappach, B. (2005) 'Hide and run', *EMBO reports*, 6(8), pp. 717–722. doi: 10.1038/sj.embor.7400480.

Moran, O. (2017) 'The gating of the CFTR channel', *Cellular and Molecular Life Sciences*, 74(1), pp. 85–92. doi: 10.1007/s00018-016-2390-z.

Morgan, H. L. (1965) 'The Generation of a Unique Machine Description for Chemical Structures—A Technique Developed at Chemical Abstracts Service', *Journal of Chemical Documentation*. American Chemical Society, 5(2), pp. 107–113. doi: 10.1021/c160017a018.

Morris, G. M. *et al.* (2009) 'Software news and updates AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility', *Journal of Computational Chemistry*. NIH Public Access, 30(16), pp. 2785–2791. doi: 10.1002/jcc.21256.

Muthyala, R. S. *et al.* (2004) 'Equol, a natural estrogenic metabolite from soy isoflavones: Convenient preparation and resolution of R- and S-equols and their differing binding and biological

activity through estrogen receptors alpha and beta', *Bioorganic and Medicinal Chemistry*. Pergamon, 12(6), pp. 1559–1567. doi: 10.1016/j.bmc.2003.11.035.

Mutyam, V. *et al.* (2016) 'Discovery of Clinically Approved Agents That Promote Suppression of Cystic Fibrosis Transmembrane Conductance Regulator Nonsense Mutations.', *American journal of respiratory and critical care medicine*, 194(9), pp. 1092–1103. doi: 10.1164/rccm.201601-0154OC.

Najafi, A. *et al.* (2014) 'Genome Scale Modeling in Systems Biology: Algorithms and Resources', *Current Genomics*. Bentham Science Publishers Ltd., 15(2), pp. 130–159. doi: 10.2174/1389202915666140319002221.

Newman, S. P. (2018) 'Delivering drugs to the lungs: The history of repurposing in the treatment of respiratory diseases', *Advanced Drug Delivery Reviews*. Elsevier B.V., 133, pp. 5–18. doi: 10.1016/j.addr.2018.04.010.

Nietert, M. M. *et al.* (2021a) 'Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR.', *Frontiers in pharmacology*, 12, p. 689205. doi: 10.3389/fphar.2021.689205.

Nietert, M. M. *et al.* (2021b) 'Comprehensive Analysis of Chemical Structures That Have Been Tested as CFTR Activating Substances in a Publicly Available Database CandActCFTR', *Frontiers in Pharmacology*, 12. doi: 10.3389/fphar.2021.689205.

Nishimura, N. and Balch, W. E. (1997) 'A di-acidic signal required for selective export from the endoplasmic reticulum.', *Science (New York, N.Y.)*, 277(5325), pp. 556–8. doi: 10.1126/science.277.5325.556.

Le Novère, N. *et al.* (2005) 'Minimum information requested in the annotation of biochemical models (MIRIAM)', *Nature Biotechnology*. Nat Biotechnol, pp. 1509–1515. doi: 10.1038/nbt1156.

Novère, N. Le *et al.* (2009) 'The Systems Biology Graphical Notation', *Nature Biotechnology*. Nat Biotechnol, pp. 735–741. doi: 10.1038/nbt.1558.

Le Novère, N. (2015) 'Quantitative and logic modelling of molecular and gene networks.', *Nature reviews. Genetics*, 16(3), pp. 146–58. doi: 10.1038/nrg3885.

O'Boyle, N. M. *et al.* (2011) 'Open Babel: An open chemical toolbox', *Journal of Cheminformatics*, 3(1), p. 33. doi: 10.1186/1758-2946-3-33.

O'Riordan, C. R. *et al.* (2000) 'Characterization of the oligosaccharide structures associated with the cystic fibrosis transmembrane conductance regulator', *Glycobiology*. Oxford University Press, 10(11), pp. 1225–1233. doi: 10.1093/glycob/10.11.1225.

O'Sullivan, B. P. and Freedman, S. D. (2009) 'Cystic fibrosis', *The Lancet*, pp. 1891–1904. doi: 10.1016/S0140-6736(09)60327-5.

Ochoa, D. *et al.* (2021) 'Open Targets Platform: Supporting systematic drug-target identification

and prioritisation', *Nucleic Acids Research*. Oxford University Press, 49(D1), pp. D1302–D1310. doi: 10.1093/nar/gkaa1027.

Okiyoneda, T. *et al.* (2004) 'ΔF508 CFTR Pool in the Endoplasmic Reticulum Is Increased by Calnexin Overexpression', *Molecular Biology of the Cell*. doi: 10.1091/mbc.E03-06-0379.

Okiyoneda, T. and Lukacs, G. L. (2012) 'Fixing cystic fibrosis by correcting CFTR domain assembly', *Journal of Cell Biology*. doi: 10.1083/jcb.201208083.

Open Babel (no date). Available at: http://openbabel.org/wiki/Main_Page (Accessed: 20 May 2022).

Orro, A. *et al.* (2021) 'In silico drug repositioning on F508del-CFTR: A proof-of-concept study on the AIFA library', *European Journal of Medicinal Chemistry*. Elsevier Masson s.r.l., 213, p. 113186. doi: 10.1016/j.ejmech.2021.113186.

Ostaszewski, M. *et al.* (2019) 'Community-driven roadmap for integrated disease maps.', *Briefings in bioinformatics*, 20(2), pp. 659–670. doi: 10.1093/bib/bby024.

Ostaszewski, M. *et al.* (2021a) 'COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms', *Molecular Systems Biology*. EMBO, 17(10), p. e10387. doi: 10.15252/msb.202110387.

Ostaszewski, M. *et al.* (2021b) 'COVID19 Disease Map, a computational knowledge repository of virus–host interaction mechanisms', *Molecular Systems Biology*, 17(10). doi: 10.15252/msb.202110387.

Ostedgaard, L. S. *et al.* (1997) *Association of Domains within the Cystic Fibrosis Transmembrane Conductance Regulator †*. Available at: https://pubs.acs.org/sharingguidelines.

Ott, C. J., Suszko, M., *et al.* (2009) 'A complex intronic enhancer regulates expression of the *CFTR* gene by direct interaction with the promoter', *Journal of Cellular and Molecular Medicine*. John Wiley & Sons, Ltd, 13(4), pp. 680–692. doi: 10.1111/j.1582-4934.2008.00621.x.

Ott, C. J., Blackledge, N. P., *et al.* (2009) 'Intronic enhancers coordinate epithelial-specific looping of the active CFTR locus', *Proceedings of the National Academy of Sciences of the United States of America*. Proc Natl Acad Sci U S A, 106(47), pp. 19934–19939. doi: 10.1073/pnas.0900946106.

Parton, A. *et al.* (2019) 'New models of atherosclerosis and multi-drug therapeutic interventions', *Bioinformatics*. Oxford University Press, 35(14), pp. 2449–2457. doi: 10.1093/bioinformatics/bty980.

Paul, N. *et al.* (2004) 'Recovering the True Targets of Specific Ligands by Virtual Screening of the Protein Data Bank', *Proteins: Structure, Function and Genetics*, 54(4), pp. 671–680. doi: 10.1002/prot.10625.

Pedemonte, N. *et al.* (2005) 'Small-molecule correctors of defective ΔF508-CFTR cellular

processing identified by high-throughput screening', *Journal of Clinical Investigation*. doi: 10.1172/JCI24898.

Pence, H. E. and Williams, A. (2010) 'Chemspider: An online chemical information resource', *Journal of Chemical Education*. American Chemical Society and Division of Chemical Education, Inc., pp. 1123–1124. doi: 10.1021/ed100697w.

Pereira, C. *et al.* (2021) 'CyFi-MAP: an interactive pathway-based resource for cystic fibrosis', *Scientific Reports*. Nature Research, 11(1), pp. 1–17. doi: 10.1038/s41598-021-01618-3.

Phuan, P. W. *et al.* (2015) 'Potentiators of defective DF508-CFTR gating that do not interfere with corrector action', *Molecular Pharmacology*. doi: 10.1124/mol.115.099689.

Picciano, J. A. *et al.* (2003) 'Rme-1 regulates the recycling of the cystic fibrosis transmembrane conductance regulator', *American Journal of Physiology - Cell Physiology*. doi: 10.1152/ajpcell.00140.2003.

Pind, S., Riordan, J. R. and Williams, D. B. (1994) 'Participation of the endoplasmic reticulum chaperone calnexin (p88, IP90) in the biogenesis of the cystic fibrosis transmembrane conductance regulator', *Journal of Biological Chemistry*.

Van Der Plas, S. E. *et al.* (2018) 'Discovery of N-(3-Carbamoyl-5,5,7,7-tetramethyl-5,7-dihydro-4H-thieno[2,3-c]pyran-2-yl)-lH-pyrazole-5-carboxamide (GLPG1837), a Novel Potentiator Which Can Open Class III Mutant Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Channels to a High Extent', *Journal of Medicinal Chemistry*. American Chemical Society, 61(4), pp. 1425–1435. doi: 10.1021/acs.jmedchem.7b01288.

Pranke, I. M. and Sermet-Gaudelus, I. (2014a) 'Biosynthesis of cystic fibrosis transmembrane conductance regulator', *International Journal of Biochemistry and Cell Biology*. Elsevier Ltd, pp. 26–38. doi: 10.1016/j.biocel.2014.03.020.

Pranke, I. M. and Sermet-Gaudelus, I. (2014b) 'Biosynthesis of cystic fibrosis transmembrane conductance regulator', *International Journal of Biochemistry and Cell Biology*. Elsevier Ltd, pp. 26–38. doi: 10.1016/j.biocel.2014.03.020.

Prince, L. S. *et al.* (1999) 'Efficient endocytosis of the cystic fibrosis transmembrane conductance regulator requires a tyrosine-based signal.', *The Journal of biological chemistry*, 274(6), pp. 3602–9. doi: 10.1074/jbc.274.6.3602.

Proesmans, M., Vermeulen, F. and De Boeck, K. (2008) 'What's new in cystic fibrosis? from treating symptoms to correction of the basic defect', *European Journal of Pediatrics*. Springer, pp. 839–849. doi: 10.1007/s00431-008-0693-2.

Pyle, L. C. *et al.* (2010) 'Activation of the cystic fibrosis transmembrane conductance regulator by the flavonoid quercetin: Potential use as a biomarker of ΔF508 cystic fibrosis transmembrane

conductance regulator rescue', *American Journal of Respiratory Cell and Molecular Biology*. American Thoracic Society, 43(5), pp. 607–616. doi: 10.1165/rcmb.2009-0281OC.

Quinton, P. M. (1999) 'Physiological Basis of Cystic Fibrosis: A Historical Perspective', *Physiological Reviews*, 79(1), pp. S3–S22. doi: 10.1152/physrev.1999.79.1.S3.

Quiroga, R. and Villarreal, M. A. (2016) 'Vinardo: A scoring function based on autodock vina improves scoring, docking, and virtual screening', *PLoS ONE*. Public Library of Science, 11(5). doi: 10.1371/journal.pone.0155183.

Ramsey, B. W. *et al.* (1999) 'Intermittent administration of inhaled tobramycin in patients with cystic fibrosis. Cystic Fibrosis Inhaled Tobramycin Study Group.', *The New England journal of medicine*, 340(1), pp. 23–30. doi: 10.1056/NEJM199901073400104.

Ramsey, B. W. *et al.* (2011) ' A CFTR Potentiator in Patients with Cystic Fibrosis and the G551D Mutation ', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 365(18), pp. 1663–1672. doi: 10.1056/nejmoa1105185.

Ratjen, F. (2001) 'Changes in strategies for optimal antibacterial therapy in cystic fibrosis.', *International journal of antimicrobial agents*, 17(2), pp. 93–6. doi: 10.1016/s0924-8579(00)00333-2.

RCSB (1998) *Research Collaboratory for Structural Bioinformatics (RCSB), http://www.rcsb.org/pdb/home/home.do*.

RDKit (no date). Available at: https://www.rdkit.org/ (Accessed: 2 June 2022).

Rennolds, J. *et al.* (2008) 'Cystic fibrosis transmembrane conductance regulator trafficking is mediated by the COPI coat in epithelial cells.', *The Journal of biological chemistry*, 283(2), pp. 833–9. doi: 10.1074/jbc.M706504200.

Rester, U. (2008) 'From virtuality to reality - Virtual screening in lead discovery and lead optimization: a medicinal chemistry perspective.', *Current opinion in drug discovery & development*, 11(4), pp. 559–68. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18600572.

Ridley, K. and Condren, M. (2020) 'Elexacaftor-tezacaftor-ivacaftor: The first triple-combination cystic fibrosis transmembrane conductance regulator modulating therapy', *Journal of Pediatric Pharmacology and Therapeutics*. Pediatric Pharmacy Advocacy Group, Inc., 25(3), pp. 192–197. doi: 10.5863/1551-6776-25.3.192.

Riordan, J. R. *et al.* (1989) 'Identification of the cystic fibrosis gene: Cloning and characterization of complementary DNA', *Science*, 245(4922), pp. 1066–1073. doi: 10.1126/science.2475911.

Rougny, A. *et al.* (2019) 'Systems Biology Graphical Notation: Process Description language Level 1 Version 2.0', *Journal of Integrative Bioinformatics*, 16(2). doi: 10.1515/jib-2019-0022.

Rowe, S. M., Miller, S. and Sorscher, E. J. (2005) 'Cystic fibrosis.', *The New England journal of*

medicine, 352(19), pp. 1992–2001. doi: 10.1056/NEJMra043184.

Di Sant'Agnese, P. A. *et al.* (1953) 'ABNORMAL ELECTROLYTE COMPOSITION OF SWEAT IN CYSTIC FIBROSIS OF THE PANCREAS', *Pediatrics*, 12(5), pp. 549–563. doi: 10.1542/peds.12.5.549.

Scott, J. *et al.* (1988) 'Heart-lung transplantation for cystic fibrosis.', *Lancet (London, England)*, 2(8604), pp. 192–4. doi: 10.1016/s0140-6736(88)92290-8.

Seo, M. *et al.* (2020) 'Development of Natural Compound Molecular Fingerprint (NC-MFP) with the Dictionary of Natural Products (DNP) for natural product-based drug development', *Journal of Cheminformatics*. BioMed Central Ltd., 12(1), p. 6. doi: 10.1186/s13321-020-0410-3.

Shak, S. *et al.* (1990) 'Recombinant human DNase I reduces the viscosity of cystic fibrosis sputum.', *Proceedings of the National Academy of Sciences of the United States of America*, 87(23), pp. 9188–92. doi: 10.1073/pnas.87.23.9188.

Shaker, B. *et al.* (2021) 'In silico methods and tools for drug discovery', *Computers in Biology and Medicine*, 137, p. 104851. doi: 10.1016/j.compbiomed.2021.104851.

Shannon, P. *et al.* (2003) 'Cytoscape: a software environment for integrated models of biomolecular interaction networks.', *Genome research*, 13(11), pp. 2498–504. doi: 10.1101/gr.1239303.

Sharma, M. *et al.* (2004) 'Misfolding diverts CFTR from recycling to degradation: Quality control at early endosomes', *Journal of Cell Biology*. doi: 10.1083/jcb.200312018.

SLURM (no date). Available at: https://slurm.schedmd.com.

Sorokin, A. *et al.* (2015) 'Systems Biology Graphical Notation: Entity Relationship language Level 1 Version 2.', *Journal of integrative bioinformatics*, 12(2), p. 264. doi: 10.2390/biecoll-jib-2015-264.

Sosnay, P. R. *et al.* (2013) 'Defining the disease liability of variants in the cystic fibrosis transmembrane conductance regulator gene', *Nature Genetics*. Nat Genet, 45(10), pp. 1160–1167. doi: 10.1038/ng.2745.

Southan, C. (2013) 'InChI in the wild: an assessment of InChIKey searching in Google', *Journal of Cheminformatics*, 5(1), p. 10. doi: 10.1186/1758-2946-5-10.

Southern, K. W. *et al.* (2018) 'Correctors (specific therapies for class II CFTR mutations) for cystic fibrosis', *Cochrane Database of Systematic Reviews*. John Wiley and Sons Ltd. doi: 10.1002/14651858.CD010966.pub2.

Stathias, V. *et al.* (2020) 'LINCS Data Portal 2.0: Next generation access point for perturbation-response signatures', *Nucleic Acids Research*. Oxford University Press, 48(D1), pp. D431–D439. doi: 10.1093/nar/gkz1023.

Sud, M. *et al.* (2007) 'LMSD: LIPID MAPS structure database', *Nucleic Acids Research*. Nucleic

Acids Res, 35(SUPPL. 1). doi: 10.1093/nar/gkl838.

Swahn, H. and Harris, A. (2019) 'Cell-selective regulation of CFTR gene expression: Relevance to gene editing therapeutics', *Genes*. MDPI AG, p. 235. doi: 10.3390/genes10030235.

Tarran, R. *et al.* (2005) 'Normal and cystic fibrosis airway surface liquid homeostasis: The effects of phasic shear stress and viral infections', *Journal of Biological Chemistry*. J Biol Chem, 280(42), pp. 35751–35759. doi: 10.1074/jbc.M505832200.

Taylor-Cousar, J. L. *et al.* (2017) 'Tezacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del', *New England Journal of Medicine*. New England Journal of Medicine (NEJM/MMS), 377(21), pp. 2013–2023. doi: 10.1056/nejmoa1709846.

The pandas development Team (2020) 'pandas-dev/pandas: Pandas'. Zenodo. doi: 10.5281/zenodo.3509134.

Trott, O. and Olson, A. J. (2009) 'AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading', *Journal of Computational Chemistry*. Wiley, p. NA-NA. doi: 10.1002/jcc.21334.

Varadi, M. *et al.* (2022) 'AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models', *Nucleic Acids Research*. Oxford University Press, 50(D1), pp. D439–D444. doi: 10.1093/nar/gkab1061.

Veit, G. *et al.* (2016) 'From CFTR biology toward combinatorial pharmacotherapy: Expanded classification of cystic fibrosis mutations', *Molecular Biology of the Cell*. American Society for Cell Biology, 27(3), pp. 424–433. doi: 10.1091/mbc.E14-04-0935.

Veit, G. *et al.* (2018) 'Structure-guided combination therapy to potently improve the function of mutant CFTRs', *Nature Medicine*. doi: 10.1038/s41591-018-0200-x.

Vinhoven, L. *et al.* (2021) 'CFTR Lifecycle Map—A Systems Medicine Model of CFTR Maturation to Predict Possible Active Compound Combinations', *International Journal of Molecular Sciences*, 22(14), p. 7590. doi: 10.3390/ijms22147590.

Voelker, R. (2019) 'Patients With Cystic Fibrosis Have New Triple-Drug Combination', *JAMA*. NLM (Medline), 322(21), p. 2068. doi: 10.1001/jama.2019.19351.

Wainwright, C. E. *et al.* (2015) ' Lumacaftor–Ivacaftor in Patients with Cystic Fibrosis Homozygous for Phe508del CFTR ', *New England Journal of Medicine*. Massachusetts Medical Society, 373(3), pp. 220–231. doi: 10.1056/nejmoa1409547.

Wang, X. *et al.* (2004) 'COPII-dependent export of cystic fibrosis transmembrane conductance regulator from the ER uses di-acidic exit code', *Journal of Cell Biology*. doi: 10.1083/jcb.200401035.

Wang, X. *et al.* (2018) 'Discovery of 4-[(2R,4R)-4-({[1-(2,2-Difluoro-1,3-benzodioxol-5-

yl)cyclopropyl]carbonyl}amino)-7-(difluoromethoxy)-3,4-dihydro-2H-chromen-2-yl]benzoic Acid (ABBV/GLPG-2222), a Potent Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Corrector for the Treatment of Cystic Fibrosis', *Journal of Medicinal Chemistry*. American Chemical Society, 61(4), pp. 1436–1449. doi: 10.1021/acs.jmedchem.7b01339.

Ward, C. L. and Kopito, R. R. (1994) 'Intracellular turnover of cystic fibrosis transmembrane conductance regulator. Inefficient processing and rapid degradation of wild-type and mutant proteins', *Journal of Biological Chemistry*.

Weininger, D. (1988) 'SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules', *Journal of Chemical Information and Modeling*, 28(1), pp. 31–36. doi: 10.1021/ci00057a005.

Welcome to CandActCFTR (no date). Available at: https://candactcftr.ams.med.uni-goettingen.de/ (Accessed: 26 January 2021).

Welcome to CFTR2 | CFTR2 (no date). Available at: https://www.cftr2.org/ (Accessed: 26 January 2021).

Welsh, M. J. and Smith, A. E. (1993) 'Molecular mechanisms of CFTR chloride channel dysfunction in cystic fibrosis', *Cell*, 73(7), pp. 1251–1254. doi: 10.1016/0092-8674(93)90353-R.

De Wilde, G. *et al.* (2019) 'Identification of GLPG/ABBV-2737, a novel class of corrector, which exerts functional synergy with other CFTR modulators', *Frontiers in Pharmacology*. Frontiers Media S.A., 10(MAY). doi: 10.3389/fphar.2019.00514.

Wishart, D. S. *et al.* (2007) 'HMDB: The human metabolome database', *Nucleic Acids Research*. Oxford University Press, 35(SUPPL. 1). doi: 10.1093/nar/gkl923.

Wishart, D. S. *et al.* (2022) 'HMDB 5.0: The Human Metabolome Database for 2022', *Nucleic Acids Research*. Oxford University Press, 50(D1), pp. D622–D631. doi: 10.1093/nar/gkab1062.

Xu, X., Huang, M. and Zou, X. (2018) 'Docking-based inverse virtual screening: methods, applications, and challenges', *Biophysics Reports*. Springer Science and Business Media LLC, 4(1), pp. 1–16. doi: 10.1007/s41048-017-0045-8.

Yadava, U. (2018) 'Search algorithms and scoring methods in protein-ligand docking', *Endocrinology&Metabolism International Journal*. MedCrave Group, LLC, 6(6). doi: 10.15406/emij.2018.06.00212.

Younger, J. M. *et al.* (2006) 'Sequential quality-control checkpoints triage misfolded cystic fibrosis transmembrane conductance regulator.', *Cell*, 126(3), pp. 571–82. doi: 10.1016/j.cell.2006.06.041.

Yu, Y. *et al.* (2007) 'Cystic fibrosis transmembrane conductance regulator (CFTR) functionality is dependent on coatomer protein I (COPI).', *Biology of the cell*, 99(8), pp. 433–44. doi: 10.1042/BC20060114.

Zaher, A. *et al.* (2021) ‘A Review of Trikafta: Triple Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Modulator Therapy’, *Cureus*. Cureus, Inc., 13(7). doi: 10.7759/cureus.16144.

Zhang, Z., Liu, F. and Chen, J. (2018) ‘Molecular structure of the ATP-bound, phosphorylated human CFTR’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1815287115.

Zhu, F. *et al.* (2013) ‘Biomedical text mining and its applications in cancer research’, *Journal of Biomedical Informatics*, 46(2), pp. 200–211. doi: 10.1016/j.jbi.2012.10.007.

Zielenski, J. (2000) ‘Genotype and Phenotype in Cystic Fibrosis’, *Respiration*, 67(2), pp. 117–133. doi: 10.1159/000029497.

Zielenski, J. and Tsui, L. C. (1995) ‘Cystic fibrosis: Genotypic and phenotypic variations’, *Annual Review of Genetics*. Annual Reviews Inc., pp. 777–807. doi: 10.1146/annurev.ge.29.120195.004021.