UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTMENTO DE BIOLOGIA ANIMAL


# A Machine Learning Based Drug Discovery Pipeline: Finding New Therapies for Cystic Fibrosis

Paulo Nuno Hilário Teixeira de Sousa

**Mestrado em Bioinformática e Biologia Computacional**

Dissertação orientada por:
Prof. Dr. André Osório e Cruz de Azerêdo Falcão
e co-orientada por Prof. Dr. Carlos Miguel Farinha

2019

# Acknowledgments

I would like to begin by expressing my gratitude to Dr André Falcão for the opportunity to develop this work, for his guidance, availability and sharing his knowledge and resources.

Thanks to Dr Carlos Farinha for his resourcefulness, for his guidance and on point instructions.

A special thanks to Carla Costa, for her hidden role in this work, without her persistence and infinite patience dealing with the endless bureaucracies and shipping nightmares I would still be waiting for the compounds.

Thanks to Sofia Correia, for making stuff happen and for opening many doors, both literally and figuratively.

Thanks to Madalena Pinto, the unacknowledged co-supervisor of this work. For always going the extra mile, always helpful and caring about what I was doing.

Thanks to Hugo Botelho, for his amazing work, for sharing his knowledge and for always going out of his way to help me.

To everyone at LASIGE and BioISI, thanks for the help and all the positive experiences.

Last but not least, I would like to thank FCT (*Fundação para a Ciência e a Tecnologia*) for financing the scholarship and this project in the context of the project *MIMED - Mining the Molecular Metric Space for Drug Design* (PTDC/EEI-ESS/4923/2014).

*For all*

# Resumo

O avanço tecnológico e a crescente disponibilidade de dados públicos levaram ao desenvolvimento de metodologias robustas de predição de atividade de compostos com base em aprendizagem automática. Estas metodologias apresentam maior rapidez, eficiência e menores custos que os métodos tradicionais de descoberta de fármacos.

Fibrose Quística (FQ) é uma doença autossómica progressiva para a qual existe urgente necessidade de surgimento de novas terapias. Mutações no gene *CFTR* nos pacientes de FQ levam à produção deficiente do canal de membrana de transporte de aniões CFTR, gerando desequilíbrios iónicos e transporte anormal de fluidos. FQ afeta vários órgãos, os pulmões com mais gravidade, sendo normalmente devido a problemas nestes a causa de morte prematura. A mutação mais prevalente e relevante em FQ é a deleção da fenilalanina 508 (F508del-CFTR). Por esta razão, os principais esforços de descoberta de novos fármacos são direcionados a corrigir ou amenizar os efeitos desta mutação.

Foi criada uma metodologia com recurso a modelos de aprendizagem automática de classificação e regressão baseada em máquinas de vetores de suporte e *Random Forests* para descoberta de compostos com potêncial terapêutico em FQ a partir de bases de dados de compostos de acesso público. Os compostos mais promissores foram selecionados e testados em laboratório através de ensaios de imunofluorescência com microscopia automatizada de triagem e análise de alto rendimento sobre o efeito na F508del-CFTR, com base na eficiência de tráfego da F508del-CFTR para a membrana plasmática. Os 10 compostos com melhores resultados neste ensaio foram validados com *Western Blot* e comparados com dois conhecidos compostos corretores da F508del-CFTR. 4 compostos foram identificados como promissores compostos terapêuticos para FQ.

**Palavras-chave:** Aprendizagem Automática; Químioinformática; Fibrose Quística; Predição de Fármacos; CFTR.

# Abstract

The recent technological advancements and growth in publicly available data led to the development of robust methodologies of prediction of biological activity based on machine learning models. These methodologies are typically faster, more efficient and less expensive than traditional drug discovery approaches.

Cystic Fibrosis (CF) is an progressive autosomal recessive disease for which there is an urgent need of new therapies. Mutations in the *CFTR* gene in CF patients lead to the production of a deficient CFTR anion transport membrane channel, resulting in ionic imbalances and abnormal fluid transport. CF affects many organs, but the typical cause of early death is due to lung complications. The most prevalent and relevant mutation in CF is the deletion of phenylalanine 508 (F508del-CFTR), and for this reason, the main efforts in drug discovery in CF are directed at correcting or reducing the effects of this mutation.

A pipeline for drug discovery in CF was developed, based on classification and regression machine learning models using Support Vector Machines and Random Forests and public access databases of compounds. The most promising compounds were selected and studied *in vitro* in a high-throughput screening immunofluorescence assay with automated microscopy. The traffic efficiency of F508del-CFTR to the plasma membrane was assessed and the 10 best compounds were validated with Western Blot. 4 compounds were identified as promising therapeutics for CF.



**Keywords:** Machine Learning; Chemoinformatics; Drug Discovery; Cystic Fibrosis; CFTR.

# Resumo Alargado

Com o avanço tecnológico e crescente disponibilidade de dados públicos, cada vez mais são procurados meios mais rápidos, eficientes e menos dispendiosos nos processos de descoberta de novos fármacos. Em paralelo, a crescente disponibilidade de dados e de recursos computacionais permite novas abordagens a problemas de difícil resolução em biologia e medicina. Uma possível abordagem baseia-se no uso de modelos de predição de atividade biológica de compostos. Estes podem ser feitos com recurso a modelos de aprendizagem automática e criação de espaços métricos de distâncias entre moléculas.

Uma patologia para a qual existe grande necessidade de descoberta de novas terapias é a Fibrose Quística (FQ). FQ é a doença autossómica recessiva progressiva com impacto negativo na esperança média de vida que mais afeta caucasianos. Esta doença é causada por mutações no gene *CFTR*, que levam à produção deficiente da proteína CFTR, um canal de membrana de transporte de aniões. Geram-se desequilíbrios iónicos e de transporte de fluidos, afetando vários órgãos, mas mais gravemente os pulmões.

A mutação do gene *CFTR* mais prevalente e com maior incidência nos doentes com FQ é a deleção da fenilalanina 508 (F508del-CFTR), gerando uma proteína com problemas de conformação, ficando em grande parte retida no retículo endoplasmático. Grande parte dos esforços de descoberta de novos fármacos são direcionados a corrigir ou amenizar os efeitos desta mutação.

Este projeto teve como objetivos identificar potenciais novos fármacos para FQ através de bases de dados públicas usando métodos computacionais e modelos de aprendizagem automática. Após identificação de compostos promissores, testar em laboratório através de ensaios de imunofluorescência de triagem e análise de alto rendimento sobre o efeito na F508del-CFTR. Os compostos com melhores resultados neste ensaio seriam validados com *Western Blot* (WB).

A tarefa inicial foi encontrar um conjunto de dados apropriado e tratar os dados. O ensaio escolhido foi um ensaio biológico funcional de supressão de fluorescência sobre o efeito de pequenos compostos na função da F508del-CFTR, disponível abertamente na base de dados PubChem com a referência PubChem AID #743267.

As estruturas destes compostos foram recolhidas nas notações digitais *simplified molecular-input line-entry system* (SMILES) e IUPAC International Chemical Identifier (InChi), acompanhadas de um indicador de atividade (ativo ou inativo) e de um valor numérico de atividade, baseado na concentração do fármaco que induz metade da resposta máxima ($AC_{50}$).

A informação das estruturas em SMILES e InChi foi convertida para *fingerprints* ("impressões digitais") moleculares, em formato Morgan e *Atom Pairs* ("pares de átomos") em várias configurações, em ambiente de programação Python 3. Todos os processos computacionais consequentes foram executados e programados em ambiente R. Para tarefas de previsão de atividade são usados modelos supervisiona-

dos, isto é, o conjunto de dados de treino contém a informação a modelar e o resultado obtido. Estes modelos podem ser de classificação, caso apresentem resultados em classes, neste caso, ativo ou inativo, ou de regressão caso apresentem resultados numéricos, neste caso valores de atividade.

Para validação dos modelos de aprendizagem automática, a sua performance deve ser testada com um conjunto de validação independente (CVI), de modo a que os resultados a prever sejam desconhecidos para os modelos. Esta validação só deverá ser aplicada aos melhores modelos, que serão avaliados com um processo de validação semelhante chamado de validação cruzada (VC). Em VC, o conjunto de dados de treino é repartido em várias partes que serão iterativamente usados para avaliar a performance dos modelos treinados com o somatório das restantes partes. Para avaliar a performance dos modelos de classificação foi usada a métrica Coeficiente de Correlação de Matthews (CCM) e para os modelos de regressão a raíz do erro quadrático médio (REQM).

O conjunto de dados inicial foi repartido 1/10 para CVI e 9/10 para o conjunto de treino. Os algoritmos de aprendizagem automática usados para criar os modelos foram máquinas de vetores de suporte (MVS) e *Random Forests* (RF, "florestas aleatórias"). As variáveis usadas para prever atividade são chamadas de preditores. Neste contexto, inicialmente, foram usados como preditores os *bits* dos *fingerprints*. O número de preditores usados afeta a performance dos modelos, especialmente MVS. Nem todos os *bits* têm a mesma importância. Para determinar qual a importância de cada preditor foi usada a função de importância do algoritmo de RF para todas as definições de *fingerprints* usadas. MVS foram escolhidas para abordagem principal de modelagem por obterem melhor performance em avaliações iniciais com seleção de preditores.

Foi escolhida uma abordagem de modelação em duas camadas, uma inicial de classificação e uma segunda camada de regressão. O processo de seleção dos melhores modelos de classificação passou por escolher quais as melhores definições de Morgan e Atom Pairs *fingerprints* e qual o número de preditores a usar com MSV. Foram criados modelos para cada definição de *fingerprints* com um número crescente de preditores. Os modelos que se destacaram foram 1024 *bits* with raio = 2, no interval entre 50 e 250 preditores, 1024 *bits* com raio = 3 no intervalo entre 50 e 500, 2048 *bits* e raio = e 2048 *bits* com raio = 3 no intervalo entre 50 e 500 preditores.

Foi criado um espaço métrico de distâncias entre compostos baseado nas distâncias de Tanimoto. Estas distâncias foram calculadas com base nos Morgan *Fingerprints*. Foi aplicada a técnica de redução de dimensionalidade de Análise de Coordenadas Principais, e as duas primeiras coordenadas principais foram projetadas num espaço métrico de duas dimensões com base em classificação. Não houve separação clara entre ativos e inativos para nenhuma das definições. Foram também usadas as distâncias como preditores em modelos de MSV de RF. Estes modelos apresentaram uma performance inferior aos modelos com Morgan *Fingerprints* como preditores.

Para selecionar os melhores modelos de regressão com MSV apenas foram usados os compostos ativos para treino e validação. Os restantes processos foram análogos aos da escolha dos modelos de classificação. Foi escolhido o modelo com Morgan *Fingerprints* em 1024 *bits* com raio = 2. Os melhores modelos foram validados e usados para fazer triagem dos compostos mais promissores na base de dados ZINC15, juntamente com outros passos de filtragem. 28 compostos foram selecionados para validação experimental.

O primeiro ensaio experimental consistiu numa triagem de alto rendimento de imunofluorescência

com microscopia automatizada. 3 concentrações de cada composto, juntamente com os moduladores conhecidos de F508del-CFTR VX-661, VX-809 e VX-770, foram aplicadas a células CFBE expressando F508del-CFTR acoplada com mCherry (uma proteína fluorescente) e FLAG-tag (antigénio). Foi medida a CFTR total expressa nas células através da fluorescência característica da mCherry e a fluorescência na membrana plasmática (MP) através da ligação de anticorpos primários ao FLAG-tag e consequente ligação de anticorpos secundários com fluorescência.

Os resultados foram obtidos automaticamente com o software CellProfiler e analisados com o script ShinyHTM. Foi aplicada uma correção de gradiente de fluorescência por placa. Os tratamentos mais promissores foram selecionados com base na mediana dos testes-Z do rácio entre a fluorescência da CFTR na MP e a fluorescência de CFTR total, sendo uma medida de eficiência de tráfego de F508del-CFTR comparativamente ao controlo.

10 tratamentos candidatos foram selecionados para validação com WB, juntamente com os moduladores VX-661 e VX-809 para avaliar o Processamento de CFTR, a quantidade de CFTR maturada e CFTR total. Neste ensaio todos os compostos apresentaram aumento de CFTR total comparativamente ao controlo e aos VX-661 e -809, dos quais o C14 com significância. Os compostos C07, C14 e C25 apresentaram um aumento significativo na quantidade de CFTR maturada. O composto C17 embora não tenha obtido significância estatística, apresentou aumento de CFTR total no WB em duas concentrações diferentes sendo por isso também considerado promissor.

Uma análise posterior foi feita aos resultados de fluorescência de CFTR total e de CFTR na MP, confirmando a existência de compostos com aumento promissor e significativo nestas métricas. Como conclusão, foi criada uma metodologia de descoberta de novos fármacos para FQ. Foram selecionados 4 compostos como especialmente promissores.

# Contents

**List of Figures** **20**

**List of Tables** **23**

**1 Introduction** **1**

1.1 Motivation . . . 1

1.2 Background and State of the Art . . . 1

1.2.1 Drug discovery and development . . . 1

1.2.2 Chemoinformatics and Machine Learning . . . 2

1.2.3 Cystic Fibrosis . . . 3

1.3 Objectives . . . 4

1.3.1 Specific Aims . . . 5

**2 Materials and Methods** **7**

2.1 Overview of Tasks . . . 7

2.2 Choosing Dataset . . . 8

2.3 Data treatment, Processing and Creating Final Datasets . . . 9

2.4 Creating and Evaluating the Machine Learning Models . . . 11

2.5 Screening and Choosing Candidate Drugs . . . 13

2.6 Compound Screening with Immunofluorescence F508del-CFTR Traffic Assay . . . 14

2.6.1 CFTR Constructs and Cell Line Generation . . . 14

2.6.2 Cell Culture . . . 15

2.6.3 Preparation of the screening compounds . . . 15

2.6.4 Seeding, Induction of CFTR expression and Adding the Compounds . . . 15

2.6.5 Immunostaining . . . 16

2.6.6 Image Acquisition . . . 16

2.6.7 Image Analysis . . . 16

2.7 Western Blot Immunocytochemistry Assay . . . 18

2.7.1 CFTR Constructs and Cell Line Generation . . . 18

2.7.2 Cell Culture, Seeding and Adding the Compounds . . . 18

2.7.3 Sample Extraction, Quantification and SDS-PAGE . . . 19

2.7.4 Immunostaining . . . 19

2.7.5 Imaging . . . 19

**3 Results and Discussion** **21**
3.1 Machine Learning Models – Determining Optimal Number of Predictors . . . . . 21
3.2 Choosing the Best Models and Architecture . . . . . 22
3.3 Classification - Choosing Best Setting for SVM with Morgan Fingerprints . . . . . 23
3.4 Creating a Chemical Metric Space and using Molecular Distances to Predict Activity . . 26
3.5 Regression - Choosing Best Models . . . . . 28
3.6 Validation of the Most Promising Models . . . . . 30
3.7 Screening the ZINC15 Database . . . . . 30
3.8 Compound Screening with Immunofluorescence F508del-CFTR Traffic Assay . . . . . 32
3.9 Western Blot Assay . . . . . 40
3.10 Revisiting the Immunofluorescence Assay . . . . . 43

**4 Conclusions** **45**

**5 Perspectives** **47**

**Glossary and Abbreviations** **49**

**Bibliography** **56**

**Supplementary Figures** **58**

**Supplementary Tables** **65**

**Annex A** **67**

# List of Figures


2.1 **Overview of the project design.** . . . 7
2.2 **Density Plot of the Activity Scores of the Molecules in the Assay "Broad Institute Identification of Small Molecule Correctors of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Delta508 Mutation Function in Human Bronchial Epithelial Cells. Probe Project".** . . . 9
2.3 **Overview of the Dataset Creation Process and Their Usage on the Machine Learning Tasks.** . . . 10
2.4 **Overview of Experimental Design of Immunofluorescence Assay.** . . . 14
2.5 **Double Tagged CFTR variant.** . . . 15
2.6 **Quality Control with shinyHTM.** . . . 17
2.7 **Overview of experimental design of Western Blot assay.** . . . 18

3.1 **Comparison of algorithms for determining predictor importance.** . . . 22
3.2 **Initial Comparison of RF and SVM.** . . . 23
3.3 **Scoring SVM Classification Models Using Molecular Fingerprints as Predictors.** . . . 24
3.4 **Projected Distances of the Molecules of the Dataset in a 2-Dimensional Plane.** . . . 26
3.5 **Comparison of RF and SVM in Distances Models.** . . . 27
3.6 **Scoring SVM Regression Models Using Molecular Fingerprints as Predictors.** . . . 29
3.7 **Immunostaining characterization under microscopy of the CFBE cell lines expressing mCherry-Flag-F508del-CFTR and automated image analysis using CellProfiler software.** . . . 32
3.8 **Total and Final Cell Count.** . . . 33
3.9 **Ratio of Fluorescence Between PM and Total Fluorescence by Plate.** . . . 34
3.10 **Ratio of Fluorescence Between PM and Total Fluorescence by Plate after Median Polish Normalization.** . . . 35
3.11 **Extensive Scores of Ratio of Fluorescence Between PM and Total CFTR in F508del-CFTR Immunofluorescence Assay.** . . . 36
3.12 **Z-scores of Fluorescence Ratio Between Plasma Membrane and Total Fluorescence by Treatment.** . . . 37
3.13 **Hits from the Compound Screening Assay with CFBE cells Expressing the mCherry-Flag-F508del-CFTR.** . . . 38
3.14 **Representative Image of Western Blot of Drug Candidate Compounds.** . . . 41
3.15 **Quantification of Western Blot results.** . . . 42

3.16 **Z-scores of Total CFTR Fluorescence, PM CFTR Fluorescence and Ratio Between Plasma Membrane and Total Fluorescence by Treatment.** . . . . . 44

S1 **Layout of Preparation of Stock Solutions, Intermediate Dilutions and of Each Treatment on the 96-well Plates of the Immunofluorescence assay for Recovery of F508del-CFTR.** . . . . . 58

S2 **Fluorescence Gradient Obtained from Median 5x5 Normalization.** . . . . . 59

S3 **Detailed Overview of Experimental Design of Preparation of Stock Solutions, Dilutions and Media composition for Induction of CFTR Expression in the Presence of Drug Candidate Compounds.** . . . . . 60

S4 **Presence of color in Preparations of Stock and Intermediated Solutions of Screening Compounds.** . . . . . 61

S5 **Total Cell Count by Well in 96-well Plates.** . . . . . 62

S6 **Final Cell Count by Well in 96-well Plates.** . . . . . 63

S7 **Extensive Z-Scores of Ratio of Fluorescence Between PM and Total CFTR, of Total and of PM Fluorescence in F508del-CFTR Immunofluorescence Assay.** . . . . . 64

# List of Tables

3.1 **Validation of the Best Models.** . . . . . 30

S1 **First Part of Summary of Results of the Ratio of Fluorescence Between PM and Total Fluorescence of the Assay for Recovery of F508del-CFTR.** . . . . . 65

S2 **Second Part of Summary of Results of the Ratio of Fluorescence Between PM and Total Fluorescence of the Assay for Recovery of F508del-CFTR.** . . . . . 66

# Section 1

# Introduction

## 1.1 Motivation

There is a clinical need to identify new candidate drugs for cystic fibrosis (CF). Even though a significant amount of data is available on this subject there is a still a lack of sufficiently efficient treatments or robust methodologies to discover or design them. These approaches should be fast and efficient, and a possible methodology is to use machine learning models on datasets of previously made assays regarding substances with potential for enhancing $Cl^-$ transport through CFTR.

The field of computational prediction of biological activity of molecules, also called quantitative structure–activity relationship (QSAR), is rapidly evolving and showing great promise and accuracy in drug discovery. There are no reports of QSAR pipelines or workflows being previously done in the context of CF. This project was developed to create and apply a Machine Learning-based workflow for discovery of candidate drugs for CF with experimental validation of their effect.

## 1.2 Background and State of the Art

### 1.2.1 Drug discovery and development

It is believed that the use of drugs for medicinal purposes started with prehistoric people, with the use of naturally occurring substances that were collected from living organisms, such as plants, animals, algae, fungi. Only in the late 17th century, theorizing in medicine started to be replaced with observation and experimentation of the effects of drugs in study of disease. Advances in chemistry and physiology in the late 18th century, 19th and early 20th century laid the foundation needed for isolating and identifying the active compounds and understanding how drugs work at organ and tissue levels (Katzung, 2018). It was only in the 1960's, with the advances in the understand of the functioning of receptors, ion channels and enzymes, that the process of drug discovery started to be more scientific and rational (Takenaka, 2008). The previous therapeutic claims started to be accurately evaluated, with the emergence of crucial concepts of rational therapeutics such as the controlled clinical trial (Katzung, 2018).

Many of the currently available drugs have been discovered through classical pharmacology (also called forward pharmacology), in which compound libraries are created and tested on cell cultures to look for phenotypical changes, and in later stages tested in animals (Hacker et al., 2009). A more recent approach, usually called reverse pharmacology, consists in testing these compound libraries directly against purified target proteins and to look for conformational changes on these proteins (Hacker et al.,

2009; Takenaka, 2008). The more promising compounds are then tested in cell cultures and later with animal testing. This latter approach is more common nowadays, being considerably faster, usually takes 2 years, while classical pharmacology takes approximately 5 years (Takenaka, 2008).

After candidate substances are selected, there are also other factors to optimize, such as affinity, potency, stability, bioavailability, forms of administration and whether it efficiently reaches the target, for example, if it crosses the blood brain barrier for central nervous system (CNS) drugs.

In order to solve specific problems in medicine, the drug discovery process requires ever faster and more efficient methods and with less nefarious consequences towards sentient beings. The current trend is to reduce animal suffering through testing, its associated time-consuming protocols and high costs (Rai and Kaushik, 2018).

Nowadays the array of synthesizable compounds available is vast. In order to solve more difficult biological problems, not only is it impractical to test all available compounds in different molecular and cellular conditions, it is also very costly and time and resource consuming. *in vitro* High-Throughput Screening (HTS) is an approach with much potential in drug discover and toxicity testing, where concentration-response data can be generated simultaneously for up to thousands of compounds and mixtures (Shockley, 2015). The most common measure for activity in pharmacological and toxicity research is the concentration for half-maximal activity ($AC_{50}$). It is derived from the Hill equation model (Hill, 1910) and is widely used to assess approximate estimates for compound potency. $AC_{50}$ is often used to prioritize chemicals for further studies and is commonly used as the basis for prediction modeling (Shockley, 2015). There is however a large uncertainty associated with the $AC_{50}$ parameter in many concentration-response relationships (Shockley, 2016).

Rational drug design is the process of developing medications using the known information about a molecular target (Katzung, 2018). The most common approaches are through computer-based modeling and relying on the knowledge of the three-dimensional structure of the target. These compounds are usually small molecules or peptides (Hacker et al., 2009).

### 1.2.2 Chemoinformatics and Machine Learning

Chemoinformatics is the use of computer science and information techniques in the field of chemistry. A common application of chemoinformatics is, for example, to model chemical substances into digital information. This information can be used to perform complex tasks of information retrieval and prediction analysis through machine learning. In recent years, great advancements have been made in computational methods in modelling the biological activity of compounds in an accurate manner.

To work with representations of chemical structure, it is of great importance to choose a nomenclature or notation that easily represents molecules in a clearly defined way. Two widely used such notations using character strings are the Simplified Molecular Input Line Entry System (SMILES) (Weininger, 1988) and the IUPAC International Chemical Identifier (InChI) (Heller et al., 2015).

A commonly used method for modeling chemical molecules is to use molecular fingerprints, which are bit maps that represent chemical structure (Rogers and Hahn, 2010). Morgan fingerprints (also called circular fingerprints) (Morgan, 1965; Rogers and Hahn, 2010) and Atom Pairs fingerprints (Carhart et al., 1985) are two such algorithms, both currently widely used as descriptors of molecular activity producing robust results (Kausar and Falcao, 2019). These fingerprints are saved as text data. Each fingerprint bit

corresponds to a fragment of the molecule and as such, it is safe to assume that molecules that are similar have a lot of bits/fragments in common. Using algorithms and statistic models, computer systems can learn how to improve their performance in a specific task. This method is called machine learning (ML) and requires a training data set, that should represent the data for which the models will be used (Kuhn and Johnson, 2013).

The digitally coded chemical data can be used by ML models to predict the activity of untested compounds. This concept is usually called quantitative structure–activity relationship (QSAR). QSAR applications and are rapidly evolving alongside the rise of large quantities of data from HTS studies in a way that properties and biological activities of novel compounds can be rapidly predicted *in silico*, with only a fraction of the costs, labor and resources of traditional lab-based approaches (Nantasenamat et al., 2010). Although computational approaches are a great way to discover new drug candidates, it is still required to validate their effect with experimental testing, first with *in vitro* studies, and in final stages, *in vivo*.

These advances in drug discovery and compound screening methodologies show great promise in the discovery of new therapies for challenging biological problems (Kausar and Falcao, 2018).

### 1.2.3 Cystic Fibrosis

One medical condition for which there is still no viable treatment is cystic fibrosis (CF). Although being classified as a rare disease, it is the most common life shortening monogenic disease in Caucasians (Bell et al., 2015). Cystic fibrosis is a progressive autosomal recessive disease, caused by mutations in the *CFTR* gene (Riordan et al., 1989), leading to a defective CFTR protein (cystic fibrosis transmembrane conductance regulator), a cAMP-regulated $Cl^-$ and $HCO_3^-$ channel located at the apical surface of epithelial cells (Amaral, 2015).

The hallmark of the disease is disrupted $Cl^-$ transport through CFTR across epithelia (Welsh and Smith, 1993). Despite being a disease that affects multiple organs, it primarily affects the lungs, being the typical cause of mortality (Amaral, 2015). Patients suffering from this disease have less cellular permeability to anions, resulting in disturbances in electrolyte and fluid transport. Typical symptoms of these patients are a poor reabsorption of NaCl in the sweat glands (the most common diagnostic test) and abnormalities in lung, pancreas and intestine function, caused by changes in the cells' membrane potential (Quinton, 1983). The ionic dysregulation leads to dehydration of the surface liquid of the airways, excessive thickening of the mucus and impaired mucociliary clearance. This results in difficulty in clearance of pathogens in the lungs, leading to a cycle of chronic pulmonary obstruction, infection, inflammation and lung damage (Flume et al., 2009; Amaral, 2015).

The most common disease-causing mutation, among the more than 2000 variants already reported (CFTR2.ORG, 2019), is the deletion of three nucleotides resulting in the deletion of phenylalanine residue 508 (Phe508del or F508del). Approximately 85 % of patients with cystic fibrosis have at least one allele for F508del-CFTR (Bell et al., 2015) and approximately 45 % to 70 % of patients with cystic fibrosis are homozygous for this allele (Wainwright et al., 2016; Guggino and Stanton, 2006). When initially described, CF was usually fatal in infancy or early childhood. More recently the median survival from CF has increased dramatically to approximately 40 years, while premature death before 50 years old remains the norm (Guggino and Stanton, 2006; Bell et al., 2015).

Since this mutation is so prevalent, research and drug development efforts are mainly focused on addressing F508del-CFTR. This mutation causes folding and processing defects in CFTR, leading to retention in the endoplasmic reticulum and rapid degradation and thus severely reducing the amount of this protein that correctly locates to the cell's surface (Mogayzel and Flume, 2010). The few channels that do reach the surface of the epithelium are functionally impaired, since the mutation disrupts the channel's opening (Wainwright et al., 2016).

One treatment strategy is to increase the amount of matured CFTR at the cell surface. Small molecules that can promote the correct folding of F508del-CFTR are called "correctors" (Mogayzel and Flume, 2010).

There is already FDA/EMA-approved medication to address this mutation, such as Lumacaftor (VX-809) or Tezacaftor (VX-661) combined with Ivacaftor (VX-770) (Lommatzsch and Taylor-Cousar, 2019). Lumacaftor and Tezacaftor aim at correcting the misprocessing and increasing cell surface localized protein. VX-770 is a "potentiator", which increases the probability of CFTR being open and of reaching the surface of the cell (Wainwright et al., 2016). VX-770, has been shown to potentiate chloride transport by both G551D- and F508del-CFTR proteins in vitro (Mogayzel and Flume, 2010). While alone it is debatable if these agents have meaningful effects on F508del-CFTR *in vivo*, when combined there is a small improvement of approximately 3 % in lung function and while it does not dramatically improve symptoms, it does seem to have clinical significance in some cases (Deeks, 2016). It is suggested that a combination of agents is necessary for full correction of F508del-CFTR (Farinha et al., 2013), and recently, combinations of three agents (new modulators combined with Tezacaftor and Ivacaftor) are also being considered for clinical use (Taylor-Cousar et al., 2019).

HTS initiatives have proven that F508del-CFTR correctors are much more difficult to identify than potentiators (Farinha et al., 2015). Having these issues in mind, the main focus in drug development for cystic fibrosis is in finding correctors, this is, getting the channels to locate to their correct location.

Parallel to the medication therapies, a primary therapy for patients with CF has been the clearance of airway secretions, through a variety of clearance therapies. However, these are intrusive and require considerable time and effort (Flume et al., 2009).

## 1.3 Objectives

The main aim of this project was to identify new candidate drugs for cystic fibrosis, using machine learning and computational methods.

After training, learning, testing and *in silico* validation, it was expected to perform *in vitro* validation studies for the most promising candidate drugs for cystic fibrosis. The most promising compounds would be tested *in vitro* for their effect on F508del-CFTR in a HTS immunofluorescence assay. Based on the results of the immunofluorescence assay, the best scoring compounds would then be validated with a Western Blot (WB).

### 1.3.1 Specific Aims

- To use Support Vector Machines (SVM) and/or Random Forests ML models to predict activity of commercially available compounds.
- To identify new candidate drugs for CF based on the predictions of the ML models.
- To test their effect on the trafficking levels of F508del-CFTR expressed in human bronchial epithelial cells in several concentrations through HTS with immunofluorescence microscopy.
- To identify the most promising compounds in the immunofluorescent assay.
- To validate the effect of the most promising compounds on F508del-CFTR using Western Blot.

# Section 2

# Materials and Methods

## 2.1 Overview of Tasks

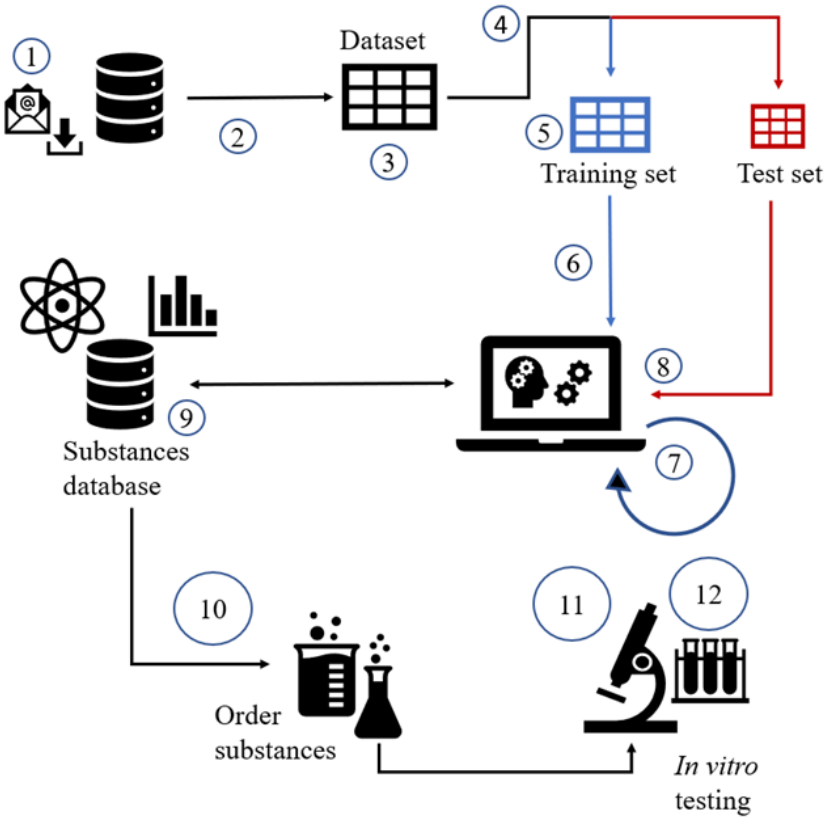



Figure 2.1: **Overview of the project design.**
(1) Finding and choosing datasets. (2) Data treatment and processing. (3) Creating Final datasets using Molecular fingerprints with bits as predictors and Tanimoto distances as predictors. (4) Creating training and testing partitions. (5) Determining importance of each predictor. (6) Determining optimal number of predictors by training SVM and RF models. (7) Optimizing architecture & and repeat steps 5 and 6. (8) Validating. (9) Screening. (10) Scoring and ranking those substances and order approximately the 30 most promising ones. (11) Testing substances' effects on CFTR cellular localization through High-throughput Microscopy Immunofluorescence assay in CFBE cells. (12) Testing substances' effect on F508del-CFTR through Western Blot assay in CFBE cells.

The initial task was to find a suitable CF compound screening assay from which to extract the data to use for modeling. This data was treated and processed into adequate formats in order to build appropriately structured datasets. The digital chemical structure and properties of the compounds was then created in several formats. These formats were all based in molecular fingerprints, both Morgan Fingerprints and Atom Pairs. From the Morgan Fingerprints, another type of dataset was created using Tanimoto molecular distances.

From the entire dataset, the Training Sets and Independent Validation Sets were created through random partitioning. The Training Sets were used in the process of training and choosing the best ML models and settings and architecture.

After the selecting the best models, these were validated with the IVS and used to screen the ZINC15 database, along with other filtering methods and criteria.

The most promising compounds were selected and studied *in vitro* with an immunofluorescence based HTS and with Western Blotting (WB).

An overview of the main tasks are represented in figure 2.1

## 2.2 Choosing Dataset

The chosen dataset was obtained from the functional cell-based confirmatory bioassay on defective CFTR correction "Broad Institute Identification of Small Molecule Correctors of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Delta508 Mutation Function in Human Bronchial Epithelial Cells. Probe Project" (The Broad Institute of MIT and Harvard, 2014), publicly available on PubChem (Kim et al., 2018) with the PubChem AID #743267.

As the name indicates, in this assay the effect of small molecules on correcting defective CFTR was studied, in Human Bronchial Epithelial Cells (CFBE). These cells expressed a halide (halogen element such as F, Cl, I, Br) sensitive YFP, a yellow fluorescent protein whose fluorescence is quenched (decreased fluorescence intensity) in the presence of sodium iodide (NaI). It was expected that compounds restoring F508del-CFTR function would allow mutated CFTR channels to be expressed at the cell surface resulting in an enhanced anion transport and subsequent fluorescent quenching ability of the CFBE cells.

The measurement used to determine the active concentration (AC) was the $AC_{50}$, which estimates the concentration at which a chemical produces the half-maximal response along a sigmoidal curve (Shockley, 2016). Compounds reducing the fluorescence with $AC_{50}$ lower than 5$\mu$M were considered as active.

pAC was set to equal 1*log10(AC). The assay Score is calculated by formula 2.1,

$$Score = 10 * pAC \tag{2.1}$$

and as such, the Scores relate to AC in the following manner; 120 = 1 pM, 90 = 1 nM, 60 = 1 $\mu$M, 30 = 1 mM and 0 = 1 M. The assay then attributed to the aggregation of individual tests, an outcome of Active if all tests were Active, Inactive when all tests were Inactive or Inconclusive if there were mixed test results. If the outcome was considered Inactive or Inconclusive, the Score was set to 0.

All the data from the initial dataset that was not relevant for this project was removed, such as inconclusive results and redundant assay score data, and a new dataset was created containing only the PubChem Substance ID (SID), activity outcome (active or inactive), activity score (a numerical value), the SMILES and InChI.

From the total 1700 molecules, after excluding 6 molecules classified as inconclusive, the final dataset contained 605 molecules classified as active and 559 as inactive, a final number of 1164 molecules. The activity values varied between 0 and 76, being that all inactive molecules were given a score of 0. A density plot of the distribution of the data points can be seen in figure 2.2. It can be observed that there is a big density of data points with a score of 0, the inactive molecules and the active molecules are distributed between 42 and 76.

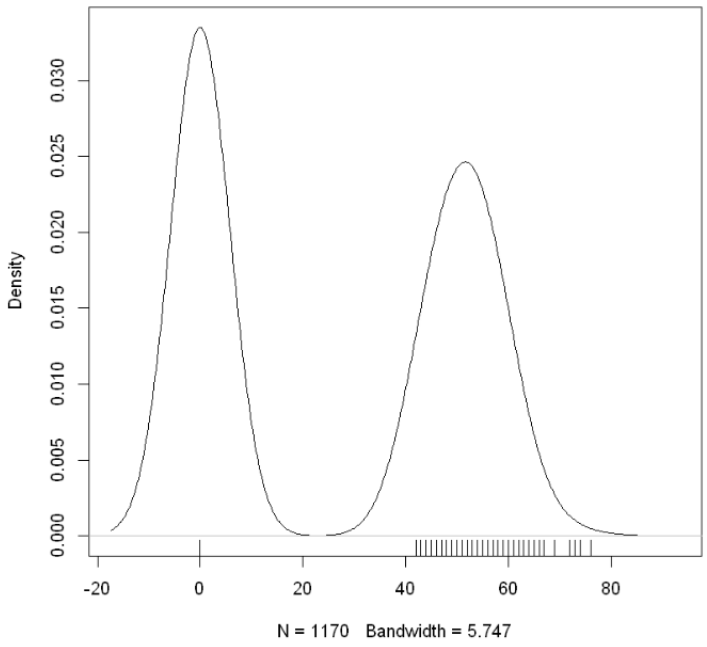



Figure 2.2: **Density Plot of the Activity Scores of the Molecules in the Assay "Broad Institute Identification of Small Molecule Correctors of the Cystic Fibrosis Transmembrane Conductance Regulator (CFTR) Delta508 Mutation Function in Human Bronchial Epithelial Cells. Probe Project".**
Vertical lines just above the x axis represent the activity scores of molecules in the assay. All Inactive molecules had a Score of 0 and the Active molecules between 42 and 76.

## 2.3 Data treatment, Processing and Creating Final Datasets

For an overview representation of the following procedures, see figure 2.3.

The chemical structure of the compounds was obtained in SMILES and InChi. SMILES is a line notation for encoding molecular structures in a human readable way. A possible problem of using SMILES is that there are several ways to represent the same molecules. An approach to this problem is the use of algorithms that canonize SMILES (Weininger et al., 1989), this is, that produce SMILES that are unique for each structure. InChI was designed to encode molecular information in a standard and unique way that can be read by humans. The InChI format and algorithms are nonproprietary and free to use and can be computed from structural information. InChI also express more information than SMILES, such as stereochemistry and electronic charge information. The InChIkey is a 27-character representation of the InChI that is not human-readable, however it can facilitate computation and web-searched because it is usually much smaller.

The chemical structure in SMILES or InChI was converted to molecular fingerprints to use as predictors and assess molecular distances in the following computational procedures.

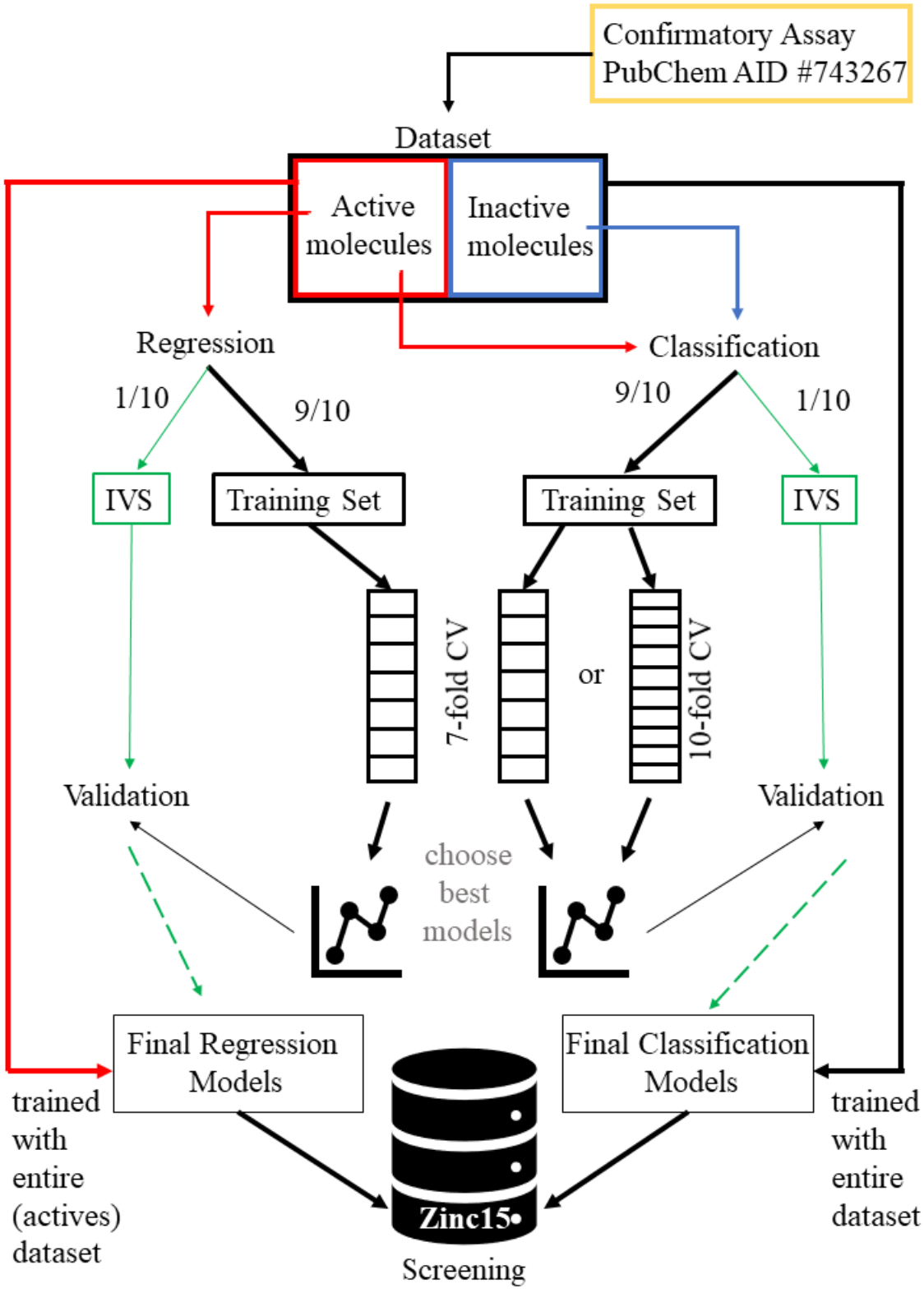



Figure 2.3: **Overview of the Dataset Creation Process and Their Usage on the Machine Learning Tasks.**
Data was obtained from PubChem and treated. The whole treated dataset was used for classification and for the regression models only the active molecules were used. Data was split 9/10 for training and 1/10 for Independent Validation Set (IVS). 7-fold or 10-fold Cross Validation was used to score the models. The best models were validated with the IVS and after validation were trained with the entirety of the treated dataset (only the active molecules for the regression models) and used to screen ZINC15 database of commercially available compounds.

Two approaches to making molecular fingerprints were used, Morgan Fingerprints and Atom Pair Fingerprints. The Morgan Molecular Fingerprints were made through a custom script in Python 3 environment, using the RDKit package (Landrum, 2006). For each substance, a Morgan Molecular Fingerprint was made in the settings 128 bits and radius 2, 128 bits and radius 3, 256 bits and radius 2, 256 bits and radius 3, 512 bits and radius 2, 512 bits and radius 3, 1024 bits radius 2, 1024 bits radius 3, 2048 bits radius 2 and 2048 bits radius 3. Atom Pairs fingerprints were made for each substance through a custom script in Python 3 environment using the RDKit package, 4415 bits both in binary and standard form, and also in 985 bits by using the knime® software (Berthold et al., 2007).

The following procedures were all made in R 3.4.3 environment. In order to have an independent validation process, the molecules used to validate the models were previously separated into an independent validation set (IVS). The remaining molecules formed the Training Set, the molecules used to train and score the models. The dataset was randomly split in a 1:10 ratio, between IVS and training set. This process was repeat using only the active models, to create a Training Set and IVS for the regression models (fig 2.3). The IVS was only used after the best models were chosen for validation.

## 2.4 Creating and Evaluating the Machine Learning Models

Machine learning (ML) can be supervised, if the training data contains both the input and the desired output or unsupervised, where data has only inputs and the models are used only to structure the data. Supervised learning can be used for either classification or regression tasks. In classification the output is discreet or in classes, for example classifying molecules as “Active” or “Inactive”. In regression the outputs are continuous, for example values of binding affinity in Ki or numerical values of functional activity in a fluorescence assay.

For validation of the models, their performance should be tested with an IVS, data for which the inputs and outputs are known but wasn’t used in the training of the models. In order to score the models’ performance on the training procedures, a common method is to cross-validate (CV) the training data. CV works by creating several partitions of the training data and iterate through each as the validation set, while the remaining partitions are used to train the model with the same parameters. This approach mimicks the use of an IVS for scoring how they perform.

There are many machine learning methods and algorithms. Two of the more widely used and versatile algorithms are Support Vector Machines (SVM) and Random Forests, both supervised methods for classification or regression tasks.

Support Vector Machines work by representing the data as points is a space of n-dimensions and map them in a way that examples of different categories or range of values are separated by a clear gap, that is as wide as possible. New examples are then predicted to belong to a certain category or to have a certain value based on where they are mapped in the model (Cortes and Vapnik, 1995).

Random forests operate by constructing a multitude of decision trees at training. Decision trees are a combination of mathematical and computational techniques for description, categorization and generalization of a given set of data. Random Forests are a type of aggregated decision trees, where multiple decision trees are built with random selection of features, by repeatedly re-sampling the training data with replacement and voting and scoring the trees for a consensus prediction (Gama et al., 2012).

ML models were created using the svm package and the randomForest package in R. The importance of each predictor (the bits of the fingerprints) was estimated through the mean decrease in accuracy parameter of the randomForest function of the randomForest package in R. 7-fold cross-validation was made for estimating the optimal number of predictors to include in each machine learning model for each fingerprint setting in the models using Morgan Fingerprints as predictors and 10-CV was used for the models with distances.

The main score used to evaluate the classification machine learning models was the Mathews Correlation Coefficient (MCC) using the classification of the predictions as true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) (formula 2.2).

$$MCC = \frac{(TP \times TN - FP \times FN)}{\sqrt{((TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN))}} \tag{2.2}$$

For training and testing the regression models, only the active molecules were used. The main score used to evaluate the regression machine learning models was the root-mean-square error (RMSE) (formula 2.3).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(obs - pred)^2} \tag{2.3}$$

For the models that used molecular distances as predictors, the distances were calculated through the Jaccard/Tanimoto Coefficient (formula 2.4), which is a measure of the similarity between finite sample sets (in this case, the bits of the molecular fingerprints). The distances were calculated between all molecules.

$$Distance_{A-B} = 1 - Similarity_{A-B} = 1 - \frac{|A \cap B|}{|A \cup B|} = 1 - \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{2.4}$$

To create a metric space with projected distances, multidimensional scaling (Gower, 1966) was performed by Principal Coordinates Analysis (PCoA) using the cmdscale function in R. The same transformations were made for all the molecules not present in the training set, in order to project them into the same metric space. This was done by calculating the linear projection (T), that can transform the distance matrix (D) into the projected space (P), according to formula 2.5.

$$D \times T = P => D^{-1} \times D \times T = D^{-1} \times P => T = D^{-1} \times P \tag{2.5}$$

This transformation was done by applying the Moore-Penrose Inverse on the distance matrix of the Training Set with the ginv() function from the MASS package.

## 2.5 Screening and Choosing Candidate Drugs

A screening dataset was created by obtaining the entirety of the in-stock commercially available compounds in the ZINC15 database (Sterling and Irwin, 2015) in SMILES format and converting to Morgan Fingerprints in 1024 bits with radius = 2 and = 3 and 2048 bits with radius = 2 and = 3 using the RDKit (Landrum, 2006) package in Python 3 programming environment.

A kriging-based filtering step was applied as described next. The molecular similarity between all molecules in the screening dataset and the initial dataset was calculated according to formula 2.4, with Morgan Fingerprints in 1024 bits and radius = 3. For each molecule in the screening dataset, the 20 most similar molecules in the initial dataset were gathered and used to train a RF classification model. The screening compound would then be classified and kept for further consideration if predicted Active. If all these 20 molecules from the initial dataset were Inactive, the screened compound would be automatically classified as inactive, and inversely, if all 20 molecules were active, the screened compound would be classified as Active.

Machine learning models using the SVM algorithm were created with the svm package in R programming environment. The settings used to create the classification models were Morgan Fingerprints as predictors in 1024 bits with radius = 2 and 132 predictors, 1024 bits with radius = 3 and 236 predictors, 2048 bits with radius = 2 and 226 predictors and molecular distance as predictors using Morgan Fingerprints in 1024 bits and radius = 3 and 65 predictors and 2048 bits and radius = 3 with 89 predictors. For the training of these classification models the entirety of the initial dataset after treatment was used. The combination of these 5 classification models was used to score the screening compounds according to the sum of the Active predictions (0 to 5). Compounds scoring less than 4 were excluded.

An SVM regression model was created using Morgan Fingerprints as predictors with 1024 bits and radius = 2 with 210 predictors. The entirety of the Active molecules in the initial dataset after treatment was used to train this model. Molecules scoring more than 48 were kept for further consideration.

The remaining screening compounds were individually studied in ZINC15 and ChEMBL (Gaulton et al., 2016) for the presence of pan-assay interference compounds (PAINS) (Baell and Walters, 2014), Lipinski's rule of 5 (RO5) (Lipinski, 2004). Compounds with PAINS associated structures and with more than 1 RO5 violation were excluded.

## 2.6 Compound Screening with Immunofluorescence F508del-CFTR Traffic Assay

This assay is based on the quantification of the amount of CFTR that reaches the plasma membrane through quantification of immunofluorescence. This assay was based on the work described in Botelho et al. (2015). A general overview of this experiment is described in figure 2.4.

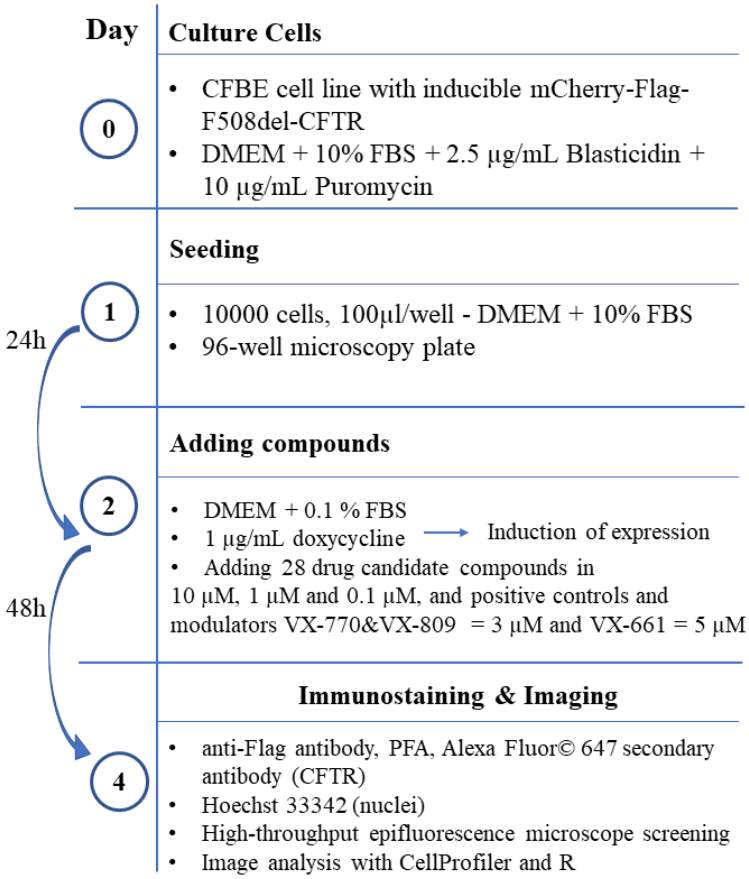



Figure 2.4: **Overview of Experimental Design of Immunofluorescence Assay.**
General pipeline of the immunofluorescence assay is described. Cells were cultured, seeded, and after 24h, F508del-CFTR expression was induced and the compounds were added. 48h after, the cells were immunostained and the plates were imaged. Main stock solutions, dilutions and media composition are described, along with the antibodies and cell stains used. DMEM - Dulbecco's modified Eagle's medium; FBS - fetal bovine serum; PFA – paraformaldehyde.

### 2.6.1 CFTR Constructs and Cell Line Generation

Cystic Fibrosis Bronchial Epithelial cells (CFBE 41o- cells, further referred to as CFBE) were used (Ehrhardt et al., 2006). These immortalized cells were developed from bronchial epithelial cells from an F508del-CFTR homozygous CF patient and allowed the development of other types of cells used on this work. For the immunofluorescence assay, the CFBE cells used were genetically engineered to express a double tagged CFTR construct: an mCherry-F508del-CFTR fusion molecule containing a Flag tag insertion (Fig. 2.5). This cell line has no expression of wt-CFTR which enables the assumption that all CFTR that localizes to the PM is F508del-CFTR. The CFTR traffic reporter construct had been previously built as described by Almaça et al. (2011), by fusing mCherry to the N-terminus of F508del-CFTR via a small linker (QISSSSFEFCSRRYRGPT). A Flag tag sequence was also inserted in the fourth extracellular loop of CFTR, between Asn901 and Ser902 (Fig.2.5). These CFBE cells were stably transduced with lentivirus encoding the previously described constructs under the regulation of a Tet-ON promoter (generated by ADV Bioscience LLC, Birmingham, AL, USA). The use of a Tet-ON promoter allows for an inducible expression of the construct upon addition of doxycycline in the culture medium.

### 2.6.2 Cell Culture

CFBE with mCherry-Flag-F508delCFTR were cultured in Dulbecco's modified Eagle's medium (DMEM) high glucose (Gibco #41965) supplemented with 10 % (v/v) heat inactivated fetal bovine serum (FBS) (Gibco #10106), 10 $\mu$g/mL blasticidin (Invivogen #ant-bl) and 2.5 $\mu$g/mL puromycin (Sigma-Aldrich #P8833) at 37ºC and 5 % CO2. Uncoated 10 cm plastic petri dishes were used.

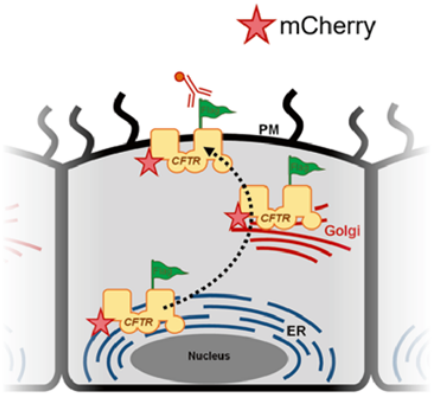



Figure 2.5: **Double Tagged CFTR variant.**
Representation of the F508del-CFTR construct contained in the used CFBE cell line, tagged with mCherry and Flag tag. The Flag tag resides in the fourth extracellular loop and only becomes extracellular if the protein successfully traffics to the membrane. Figure adapted from Amaral et al. (2016).

### 2.6.3 Preparation of the screening compounds

All the screening compounds are mostly non-polar and for this reason had to be solubilized in DMSO. Stock solutions of the screening compounds were prepared in DMSO in a 10 mM concentration. The stock solutions were placed in a 96-well polypropylene plate (Supplementary Figure S1-A) to facilitate pipetting with a Xplorer® multichannel pipette (Eppendorf #4861000112, #4861000139, #4861000155).

### 2.6.4 Seeding, Induction of CFTR expression and Adding the Compounds

CFBE cell line with mCherry-Flag-F508delCFTR were cultured to confluence and split 24h before the experiment. On the following day, the cells were trypsinized to antibiotic-free DMEM supplemented with 10 % (v/v) of FBS. 100$\mu$l of medium containing approximately 10000 cells in suspension were seeded on each well of a 96-well microscopy plate using a Multidrop Combi Reagent Dispenser (Thermo ScientificTM #5840300).

24h after seeding the screening compounds were added and CFTR expression was simultaneously induced (through addition of doxycycline). The medium was prepared by supplementing DMEM with 0.1 % FBS and 1 $\mu$g/mL doxycycline (Sigma #9891). Each of the 28 screening compounds were separately solubilized in the previously described medium to a 0.1, 1 and 10 $\mu$M concentration. 3 other compounds for which there is described activity in CFTR studies were also prepared. VX-661 at a 5 $\mu$M concentration, 3 $\mu$M VX-809 and 3$\mu$M VX-770.

For consistency of results, all wells should have the same concentration of DMSO. The treatments with 10 $\mu$M concentration contained 0.1% (v/v) of DMSO, the highest value for all preparations. For this reason, the appropriate amount of DMSO was added to each treatment preparation with a concentration of compound lower than 10 $\mu$M, to increase the amount to 0.1 % (v/v) of DMSO. DMEM with 0.1 % FBS, 1 $\mu$g/mL doxycycline and 0.1 % (v/v) DMSO was prepared as control treatment.

The detailed design and description of the procedure is described in Supplementary Figure S3.

The medium in the cells was removed and the medium with the compounds and DMSO only control was added using a multichannel pipette. The layout of each treatment can is detailed in Supplementary Figure S1-C.

### 2.6.5 Immunostaining

48h after adding the compounds and inducing F508del-CFTR expression (72h after seeding), extracellular Flag-tags were immunostained in non-permeabilized cells. The present media was aspirated, and the cells were washed once with ice-cold phosphate buffered saline (PBS) and incubated with monoclonal anti-Flag antibody (2 $\mu$g/mL Sigma-Aldrich #F1804) for 1h at 4$^{o}$C.

Cells were then washed 3 times with ice cold PBS, incubated for 20 min with 3 % (w/v) paraformaldehyde (PFA) on ice. The remaining procedures were done at room temperature. Cells were washed 3 times with PBS and afterwards incubated with anti-mouse Alexa Fluor© 647 conjugated secondary antibody (2 $\mu$g/mL Molecular Probes #A31571). Cells were then washed 3 times with PBS and incubated with a Hoechst 33342 solution (200 ng/mL, Sigma #B2261) for 1h. The cells were washed 3 times for a last time with PBS and left immersed in PBS. The plates were kept at 4$^{o}$C overnight until imaging.

All previous solutions were prepared immediately before use in PBS supplemented with 0.7 mM CaCl2 and 1.1 mM MgCl2. Antibody solutions contained 1 % bovine serum albumin (BSA, Sigma-Aldrich #A9056). All manipulations of solutions on the plates were performed using a HydroSpeed$^{TM}$ plate washer (Tecan #INSTHS-02). Solution volumes were 30 $\mu$l/well for antibodies and 50 $\mu$l/well for PFA and Hoechst.

### 2.6.6 Image Acquisition

Imaging was made using an automated inverted widefield epifluorescence microscope for high-throughput screening at room temperature. The microscope used was a DMI6000 B (Leica) equipped with a mercury metal halide light source (EL6000), with an Orca-Flash4.0 camera (Hammamatsu) with 16 bit 2048 x 2048 pixel resolution, 6.5 $\mu$m x 6.5 $\mu$m pixel size and a HC PL APO objective (Leica) with a numerical aperture of 0.4. The Hoechst channel was used for contrast-based autofocus. Imaging was made on 96-well plates, in 5 positions per well.

### 2.6.7 Image Analysis

The quantification of CFTR was made by assessing the amount of CFTR localized to the PM and the amount of total CFTR in the cell, as previously described by Botelho et al. (2015). PM CFTR is proportional to the Alexa Fluor© 647 integrated fluorescence and total CFTR is proportional to the mCherry integrated fluorescence. By dividing the PM CFTR with the Total CFTR, a measurement of CFTR traffic efficiency is obtained (formula 2.6).

$$\text{CFTR traffic efficiency} = \frac{PM\ CFTR}{Total\ CFTR} = \frac{\text{AlexaFluor© 647 integrated fluorescence}}{mCherry\ integrated\ fluorescence} \tag{2.6}$$

Automatic image analysis was performed using the CellProfiler open source software (Kamentsky et al., 2011) and the shinyHTM custom script (Botelho et al., 2019) in R programming environment. Initially the cells with an undesired phenotype were excluded, by determining a minimum and a maximum radius for cell nucleus and abnormal nuclear shapes (e.g. apoptotic cells).

A flat-field/dark-frame background correction was performed. Initially images with background fluorescence and illumination were taken with the same microscope and image acquisition setup. These (flat-field and dark-frame) images were used for correcting the illumination and subtracting the fluorescence baseline for each image.

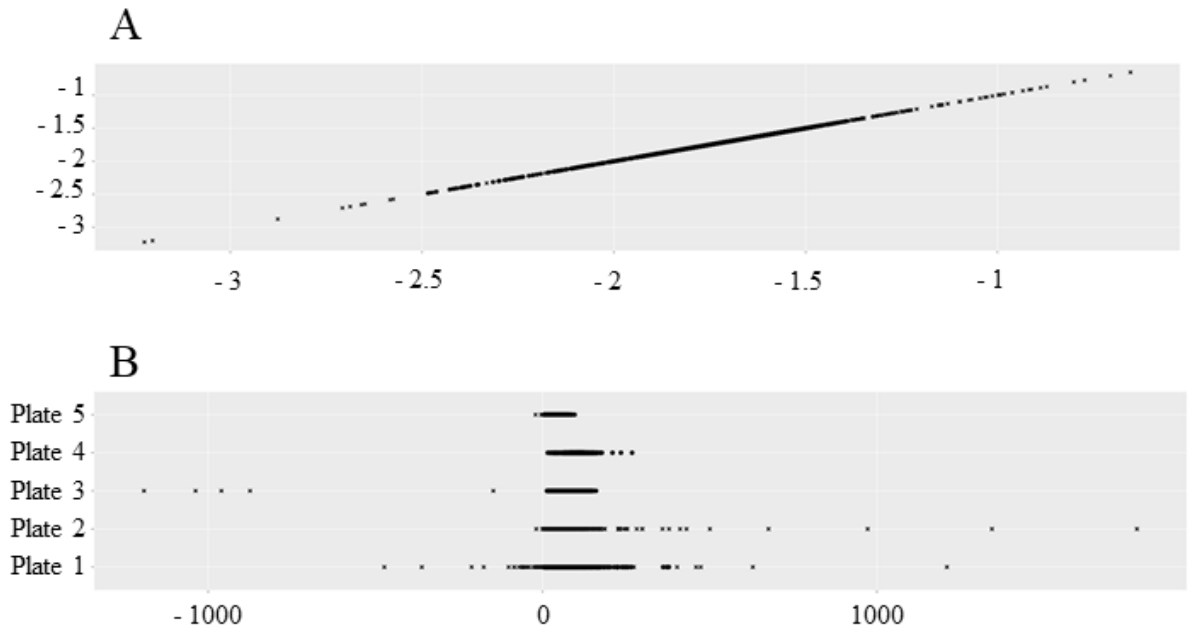



Figure 2.6: **Quality Control with shinyHTM.**
(A) Image quality and focus. Measure used was powerloglogSlope of nucleus images. Points with values under -2.5 were excluded. Note that no higher limit was chosen, all excluded points in the region with values superior to -1.25 belong to plate 2. (B) Intensity of plasma membrane fluorescence (PM intensity) by plate. Measure used was median of cell final integrated intensity of PM with background correction. Points with values under 0 were excluded. Note that no higher limit was chosen, still all outlier data points with positive values were excluded by other criteria.

Another QC filter included a maximum number of saturated pixels. For each image, the CFTR traffic efficiency was considered to be the median CFTR traffic efficiency for all cells in the image. After imaging the 5 image fields for each condition in 5 replicates, shinyHTM was used analyze the data and to exclude images based on QC filters (Fig. 2.6).

Out of focus images were excluded based on the value of the PowerLogLogSlope, the slope of the image log-log power spectrum. According to CellProfiler's resources, the power spectrum contains the frequency information of the image. The slope gives a measure of image blur, with higher slopes indicating more lower frequency components, and hence more blur, being a metric recommended for blur detection in most cases. Images with a PowerLogLogSlope $<$-2.5 were excluded (Fig. 2.6-A).

Also excluded were images not reaching a minimum of 20 identified cells (not shown) and not reaching a minimum level of PM CFTR fluorescence of 0 (Fig. 2.6-B).

Fluorescence gradient was corrected on a plate by plate basis, using the median polish methodology in shinyHTM. Median polish can be used as a normalization by utilizing the medians from the rows and the columns of a 2-way table to calculate the row effect and column effect on the data. The overall effect is calculated by finding the row medians for each row and calculating the median of the row medians. This is followed by a subtraction of the row median to each element in that row for all rows. The overall effect is then subtracted to each row median. The same procedure is then done for the columns, and the overall effect of the columns is added to the previous overall effect. These steps are repeated until the change within row or column medians is negligible.

For each replicate, after averaging the CFTR traffic efficiency for all images related to the same

treatment, the effect of each treatment was compared with the DMSO control using a Z-score (formula 2.7):

$$Z - Score = \frac{(Mean, i - Mean, ctrl)}{SDctl} \tag{2.7}$$

## 2.7 Western Blot Immunocytochemistry Assay

This assay is based on the quantification of the amount of total CFTR and F508del-CFTR that is properly folded by the action of a corrector (the drug candidate compounds) as detected after electrophoresis and transfer by an anti-CFTR specific antibody. The quantification is made by chemiluminescence. CFBE cells are used, which express either wt- or F508del-CFTR. An overview of this experiment is represented in figure 2.7.

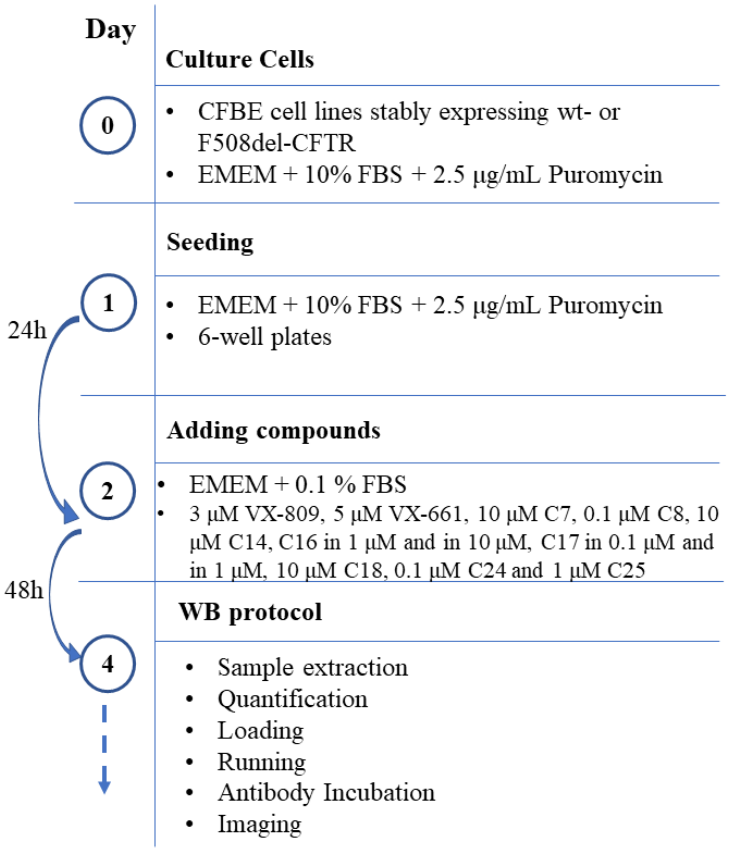



Figure 2.7: **Overview of experimental design of Western Blot assay.**
Stock solutions, dilutions and media composition for induction of CFTR expression in the presence of compound.

### 2.7.1 CFTR Constructs and Cell Line Generation

CFBE cells stably overexpressing wt-CFTR or F508del-CFTR (CFBE wt-CFTR or CFBE F508del-CFTR) were used.

### 2.7.2 Cell Culture, Seeding and Adding the Compounds

CFBE cells were cultured in EMEM supplemented with 10 % (v/v) FBS and 2.5$\mu$g/mL of puromycin (Sigma-Aldrich #P8833) at 37$^{o}$C and 5 % CO2. Cells were trypsinized and approxi-mately 200,000 cells were seeded into each well of 6-well plates in 2 mL of EMEM supple-mented with 10 % (v/v) FBS and 2.5 $\mu$g/mL of puromycin. 24h after seeding the screening compounds were prepared in antibiotic-free EMEM supplemented with 0.1 % (v/v) FBS and added to the cells. The compounds and concentrations prepared were 3 $\mu$M VX-809, 5 $\mu$M VX-661, 10 $\mu$M C7, 0.1 $\mu$M C8, 10 $\mu$M C14, C16 in 1 $\mu$M and in 10 $\mu$M, C17 in 0.1 $\mu$M and in 1 $\mu$M, 10 $\mu$M C18, 0.1 $\mu$M C24 and 1 $\mu$M C25. Similarly to what was done in the immunofluorescence assay, all wells had the same concentration of DMSO, for consistency of results. EMEM with 0.1 % FBS and 0.1 % (v/v) DMSO was prepared as control treatment, both for wt-CFTR CFBE and F508del-CFTR CFBE.

### 2.7.3 Sample Extraction, Quantification and SDS-PAGE

A detailed description of this protocol, including reagents, equipment and techniques is in Annex A. SDS-PAGE (sodium dodecyl sulfate–polyacrylamide gel electrophoresis) gels were made by stacking a 4 % (v/v) acrylamide for separation above a 7 % (v/v) acrylamide for resolving the protein samples. 48h after adding the screening compounds, the cells were washed 2 times with ice-cold PBS and lysed using a sample extraction buffer (SB, see Annex A), collected to tubes and kept on ice during the following procedures. Quantification of protein in the samples was assessed with a Bradford Assay (see Annex A). The loading samples were prepared with the same amount of protein in all samples for replicate (ideally more than 10 $\mu$g) and 1:4 (v/v) of loading buffer (see Annex A). Each gel was loaded with DMSO (wt- and F508del-CFTR) and VX661 and VX809 controls. Approximately 50 $\mu$L of sample were loaded onto the SDS-PAGE gel. The gels ran at 60-75V to concentrate the samples and separated with 100-120V, on ice. The content of the gels was transferred to PVDF (Polyvinylidene fluoride or polyvinylidene difluoride) membranes at 400 mA for 1h30min on ice.

### 2.7.4 Immunostaining

Membranes were blocked in 5 % (w/v) Non-fat-milk (powder) PBS-T for 30 min at room temperature. Membranes were then incubated with anti-CFTR 596 mouse primary antibody (1:3000 (v/v), Cystic Fibrosis Foundation Therapeutics, Bethesda, MD; USA, #A4) and with Calnexin mouse primary antibody (1:3000 (v/v), BD Transduction Laboratories$^{TM}$ #610523)(as loading control) in 5 % (w/v) Non-fat-milk (powder) PBS-T overnight at 4$^{o}$C. After primary antibody incubation the membrane was washed 3 times for 10 min with PBS-T with agitation. The membrane was then incubated with Goat Anti-Mouse IgG (H + L)-HRP Conjugate secondary antibody (1:3000 (v/v), Bio-Rad #1706516) in 5 % (w/v) Non-fat-milk (powder) PBS-T for 1h at room temperature. The membrane was washed 3 times for 10 min with PBS-T with agitation and kept on PBS-T at 4$^{o}$C.

### 2.7.5 Imaging

Signal was developed with the Immun-Star$^{TM}$ WesternC$^{TM}$ Chemiluminescence Kit (Bio-Rad #1705070). Detection was performed with the Chemidoc XRS+ system (Bio-Rad). Quantification was performed using the ImageLab$^{TM}$ software (Bio-Rad).

# Section 3

# Results and Discussion

## 3.1 Machine Learning Models – Determining Optimal Number of Predictors

In this machine learning context, the predictors (sometimes called features or attributes), the variables used to predict biological activity values, are the array of bits of the fingerprints. The number of predictors used affects the performance of the machine learning models (Carhart et al., 1985). Also important to note, is that not all predictors have the same importance for predicting a molecules activity. Parallel to what happens *in vivo* and *in vitro*, not all parts of a molecule have the same importance in exerting a certain activity or reaction. In machine learning modeling, this is called feature selection (Kuhn and Johnson, 2013). In most cases, some of the bits have variance = 0 for the entire dataset, offering no information for the modelling and being an added computational cost and in the same logic as before, reducing the accuracy of the machine learning models, and as such, should be removed.

The first approach to this part of the modeling was to choose a good and not exceedingly time-consuming algorithm to determine importance of predictors. The algorithms tested were the importance function within the randomForest function of the randomForest package, the varImp function from the Caret package (Kuhn, 2008) using both Random Forests and SVM models, and recursive feature elimination (RFE) with the rfe function of the caret package. RFE works by using a Random Forest algorithm on each iteration to evaluate the modeling in a way that explores all possible subsets of the attributes. These algorithms were tested on several different fingerprint settings on the dataset. A representation of their performance on 128 bits and 2 radius Morgan Fingerprints is on fig 3.1.

randomForest, varImp with RF and RFE all performed similarly, and markedly better than varImp with SVM. A control with the predictors with a random order of importance was used (Scrambled). By comparing the importance prediction algorithms with the scrambled control, it can be clearly seen how choosing an optimal number of predictors enables a more accurate prediction from the SVM models, while all the scores converge to the same values when the number of predictors approaches the complete set of bits of the fingerprints.

A similar performance was reported for all fingerprint settings. randomForest was chosen based on being the fastest algorithm and because of its familiarity, know reliability and good performance from previous studies within this research group (Teixeira et al., 2013). RFE is also a fast algorithm that performed remarkably but its use requires that the user be familiarized with Caret's particular methodology and use of functions, which could be an obstacle for future users of this methodology.

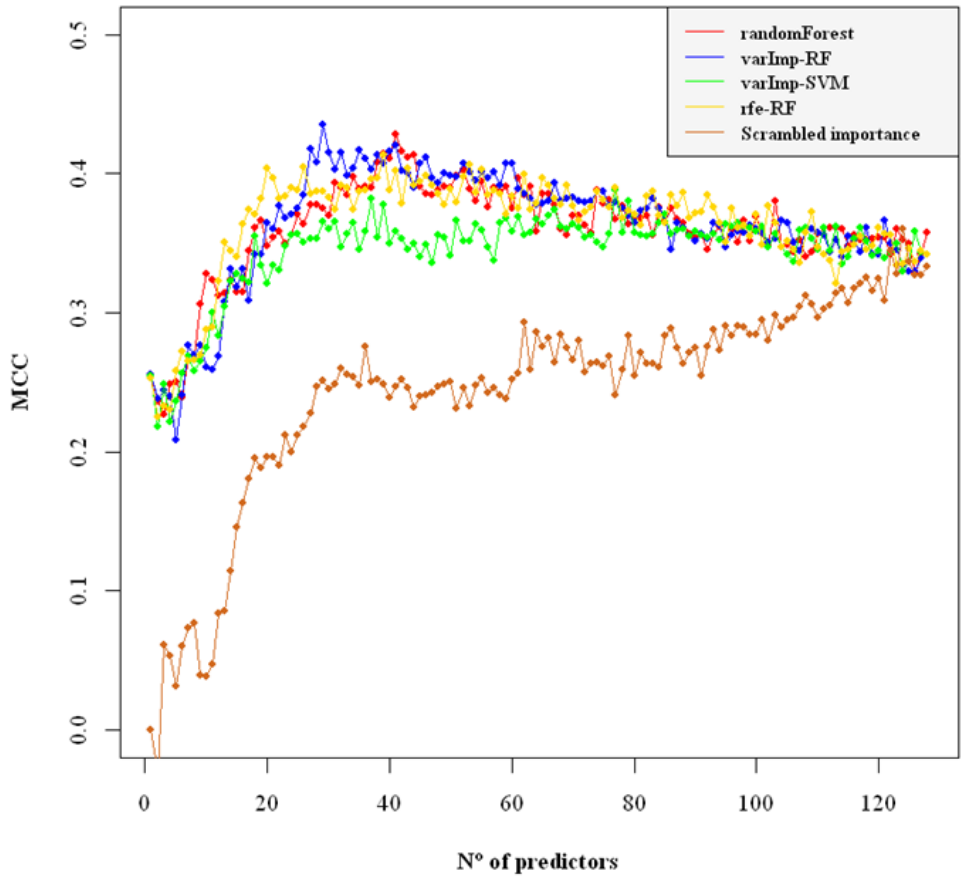



Figure 3.1: **Comparison of algorithms for determining predictor importance.** Models were constructed with SVM using Morgan Fingerprints with 128 bits and radius = 2. For each algorithm of determining predictor importance, models were created with all possible numbers of predictors (variables). Models were scored with the MCC metric, with 7-CVx2. Scrambled order of predictors was used as control.

## 3.2 Choosing the Best Models and Architecture

The two machine learning algorithms chosen to use were SVM and RF. Initial tests using a smaller dataset for predicting biological activity of small molecules on the Sigma1 receptor, using Morgan Fingerprints, indicated that while RF performed better without predictor selection, SVM with selection of number of predictors based on performance had better and more consistent results. It is also noteworthy that SVM are considerably faster than RF. Having these preliminary results in mind, the focus of this work was on mainly using SVM, while still testing the performance of RF on instances where it seemed appropriate. After choosing the algorithm for determining predictor importance, models were constructed for each fingerprint setting. An overall preliminary look comparing SVM against RF showed again that SVM seemed to perform better than RF when selecting an optimal number of predictors, which is especially noticeable in settings with higher number of bits (fig 3.2).

The initial approach to this project was to make regression machine learning models to predict activity for the entire dataset. This approach was proven to be flawed due to the nature of the dataset. The inactive molecules all had a score of 0, and the active molecules showed scores between 42 and 76 (fig 2.2). This is a big gap of data for a regression model that predicts activity based on continuous values.

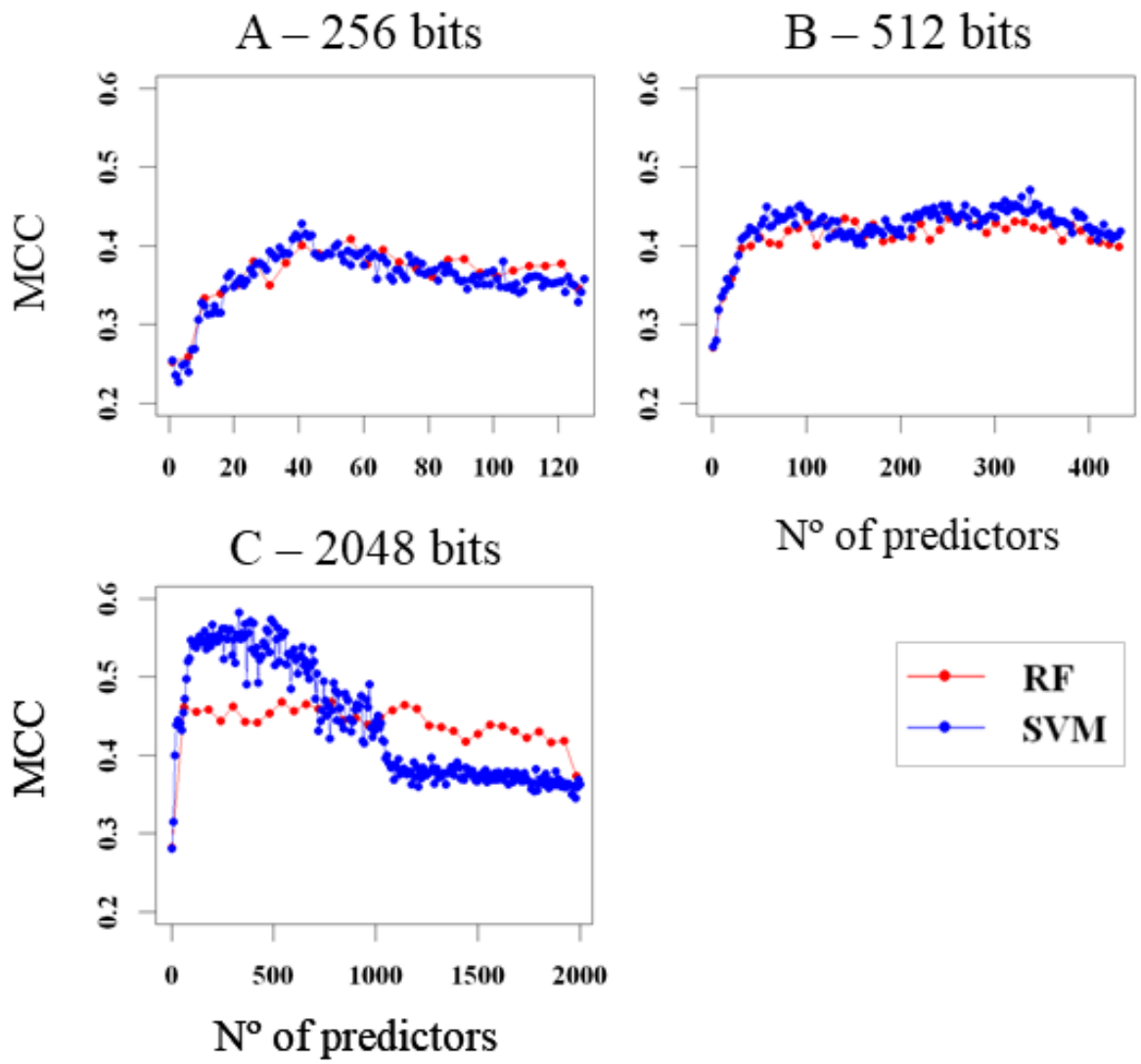



Figure 3.2: **Initial Comparison of RF and SVM.**
Models were constructed using Morgan Fingerprints with radius = 2, with 128 bits (A), 256 bits (B) and 2048 bits (C). Models were scored with the MCC metric, with 7-CVx2.

These models had unsatisfactory scores with a minimum RMSE of approximately 20. To address this issue, it was chosen to create a 2-layer model with a first layer consisting of one or several classification models and a second layer with a regression model to predict quantitative activity for only the molecules classified as active.

## 3.3 Classification - Choosing Best Setting for SVM with Morgan Fingerprints

Two types of molecular fingerprints were chosen as the most promising to model the chemical structure, Morgan Fingerprints and Atom Pairs. Atom Pairs can be created under two types of configurations, binary or "standard".

Atom Pairs fingerprints as created by RDKit, in a simplistic way, code chemical structure into bit

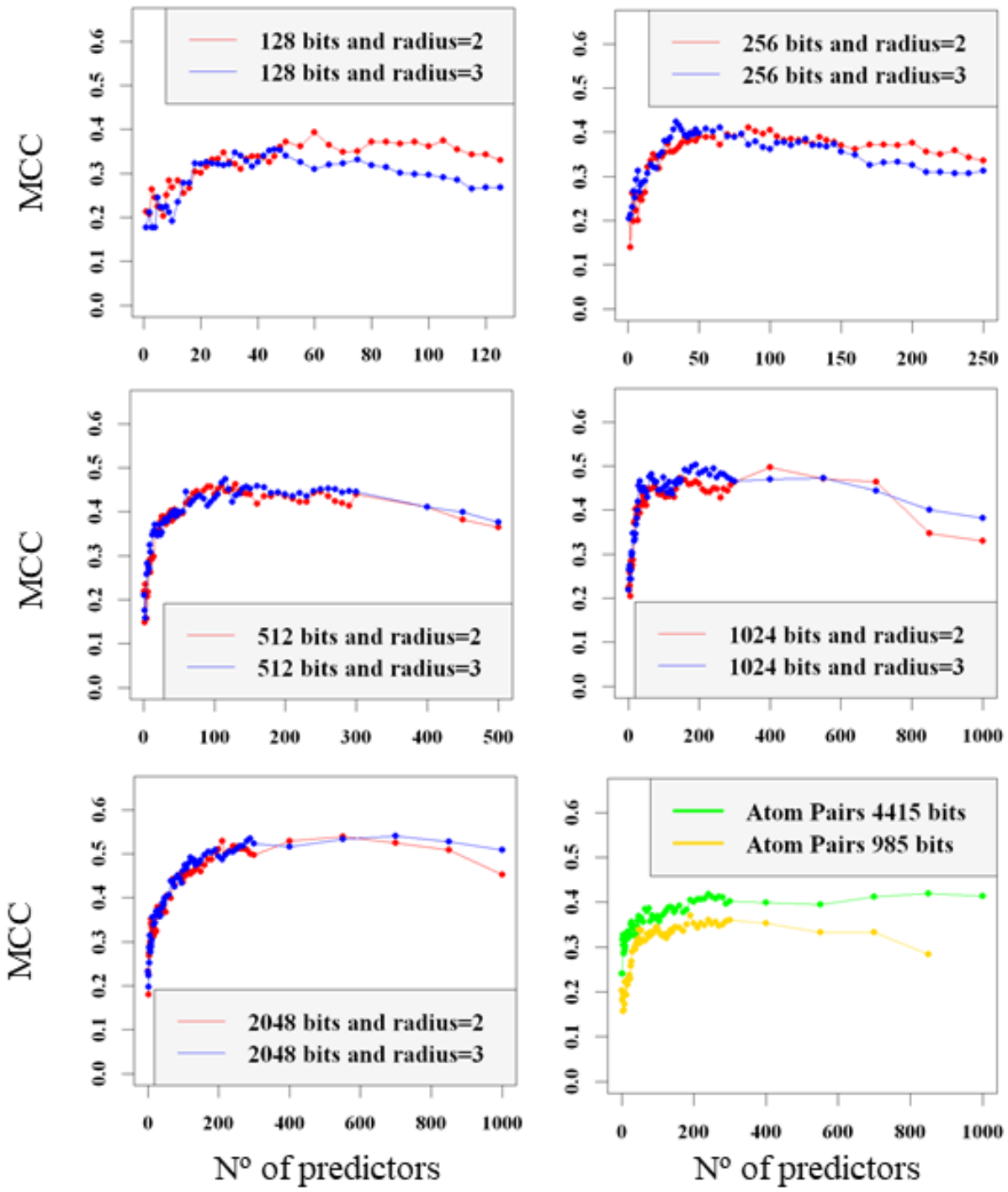



Figure 3.3: **Scoring SVM Classification Models Using Molecular Fingerprints as Predictors.** SVM was used to create the models and MCC was used as score metric. For each fingerprint setting, the importance of the predictors was estimated, and the models were trained with increasing numbers of predictors.

vectors by attributing a particular bit to a simple attribute, like a structure or feature (Carhart et al., 1985). If this attribute is absent in the structure, it will be coded as an off-bit, which is a 0, and if it is present it will be registered as an on-bit. An on-bit in binary is always a 1, while in "standard" Atom Pairs it can be any positive integer value, usually corresponding to the amount of times that feature is

present. This seems to imply that for most applications, Atom Pairs fingerprints in binary would have a loss of information. Initial training and scoring of the ML models confirmed this, showing that models trained with binary Atom Pairs fingerprints always scored worse than their not-binary equals, and for this reason were excluded from further analysis.

For Morgan fingerprints this is not an issue, since its algorithm always codes chemical structure as binary vectors. The main choice with this type of fingerprints is which number of bits to code the structure in and with which radius. While it seems logical that higher bit vectors would code the structure with a higher resolution, this does not always translate into better predictions with the machine learning algorithms. As it was explained before, with some machine learning algorithms, such as SVM, prediction accuracy tends to decrease as the number of predictors used to construct the model (in this case the fingerprint bits) is increased past a certain optimal region.

Before scoring the models, an Independent Validation Set (IVS) was created, by randomly sorting 1/10 of the dataset for IVS and the remaining 9/10 substances of the dataset for the training data (Training Set). Only the Training Set was used for training and scoring the models. The IVS was only used after the best models were chosen, for validation. This ratio was chosen so that it was high enough for a reliable validation while minimizing the risk of excluding a particular cluster of activity present in the molecules of the dataset.

The training set was randomly sorted into 7 parts, used for a 7-fold Cross Validation (7-CV), the procedure used to score the models. This procedure works by iterating through each of the 7 partitions of the Training Set as the validation set, while the remaining 6 partitions are used to train the model. In each iteration, the activity of the molecules in the validation set was predicted by the model created as Active (positive) or Inactive (negative). This result is then categorized as a either a true positive (TP, if the molecule is active and was predicted as being active), false positive (FP, if the molecule is inactive and was predicted as being active), true negative (TN, if the molecule is inactive and was predicted as such) or false negative (FN, if the molecule is active and was predicted as inactive).

All the results throughout the 7-CV were combined and the Mathews Coefficient Correlation metric was used to score the models, which rates the performance of the model by weighing all the previously described type of results (TP, TN, FP and FN), being a more balanced metric than the accuracy (% of correct predictions).

The focus was then directed to choosing the best combinations between fingerprint setting and number of predictors to use for that setting. To analyze the general performance of the ML models using SVM and Morgan Fingerprints, a graphical approach was chosen. For each setting, a ML model was created with an increasing number of predictors, for which the importance was previously calculated and sorted from highest to lowest. Each score was then plotted, enabling a view of how the models performed by number of predictors, in each setting (Fig 3.3).

The highest scoring settings were 1024 bits with radius = 2, in the range between 50 and 250 predictors used, 1024 bits with radius = 3 in the range between 50 and 500, 2048 bits and radius = 2 in the range between 50 and 500 predictors and 2048 bits and radius = 3 in the same range of number of predictors.

It is important to note that it is not critical to be precise in choosing the absolute best number of predictors to use for each setting, since there is an inherent random component associated with the

IVS and cross-validation procedure, which means that if the process is repeated, the scores would vary slightly. The focus was then to identify the region with the highest consistency of best results, create models for all numbers of predictors within that range, score them and choose a number of predictors within that seems best and validate with the IVS.

## 3.4 Creating a Chemical Metric Space and using Molecular Distances to Predict Activity

An alternative approach to using molecular fingerprints directly as predictors of activity is to use Tanimoto distances to create a metric space for the substances. In these models, the fingerprints are treated as vectors, and the distance (or dissimilarity) between each molecule of the dataset is calculated in a matrix and mapped into an n-dimension abstract cartesian metric space.

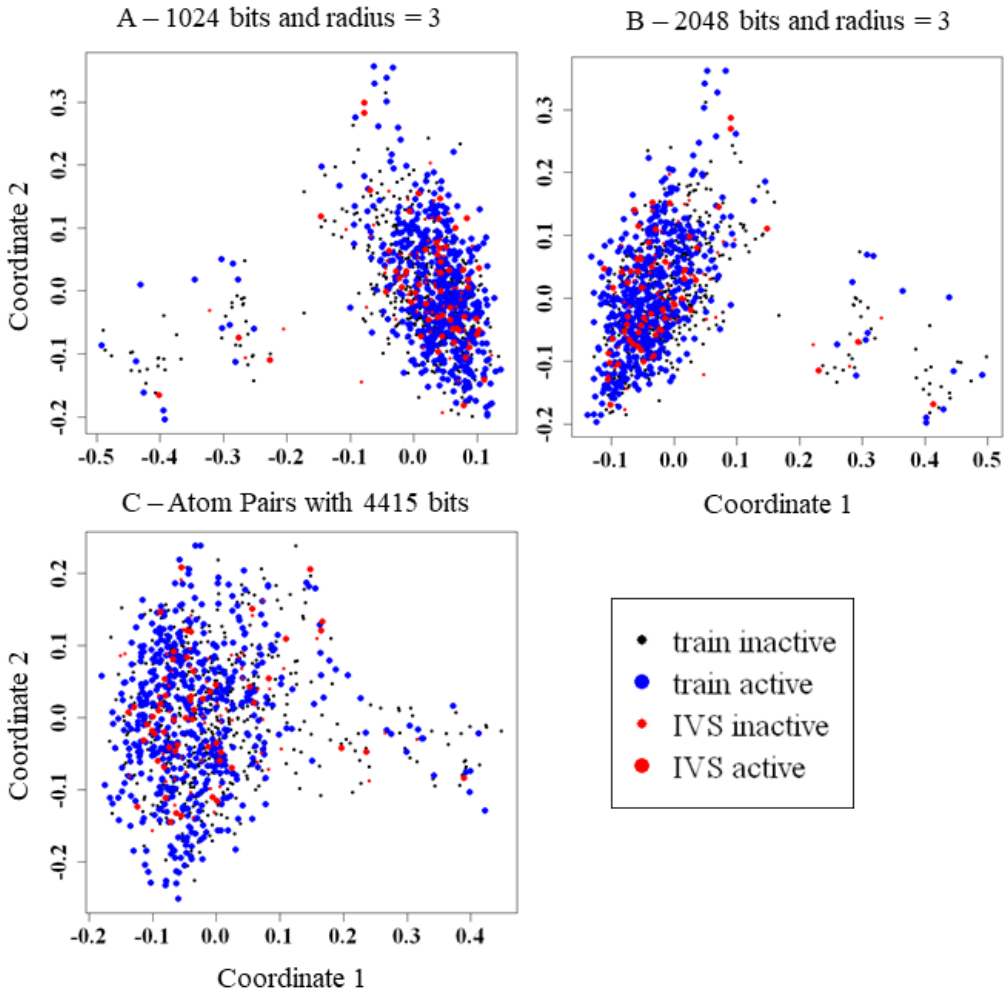



Figure 3.4: **Projected Distances of the Molecules of the Dataset in a 2-Dimensional Plane.** Tanimoto distances were calculated between all molecules of the Training Set using with Morgan Fingerprints with radius = 3, with 1024 bits (A), 2043 bits (B) and Atom Pairs with 4415 bits (C). Distances between molecules are shown as coordinates in the first two Principal Coordinates (PCoA). Blue points represent the Active molecules of the Training Set and black points the Inactive. Molecules of the IVS were projected onto the same metric space and are shown as red points.

One advantage of using distance models is that through multidimensional scaling it is possible to map or project the distance between molecules on a 1-, 2- or 3-dimensional plane, using for example Principal Coordinate Analysis (PCoA) (Zuur et al., 2007). The algorithm maximizes the correlation of the distances with the number of dimensions in a way that the first Principal Coordinate explains the most variation of the data, the second Principal Coordinate explains the most variation of the data after the first dimension, so on and so forth. This means that by projecting the data on a 2-dimension plane metric space, one can observe in a simplistic way how the data is distributed (Fig. 3.4). Ideally in these 2-dimensional projections of data points, there would be a clear separation between the different classes (for example Active vs Inactive). This does not seem to be the case. In figure 3.4, it is shown the distance projections in 2-dimensions for the higher bit resolution fingerprint settings that better resolved the distances. Except for the Atom Pairs Fingerprints, the molecules grouped in a generally similar way, having a big cluster of molecules, and 1 or 2 smaller ones.

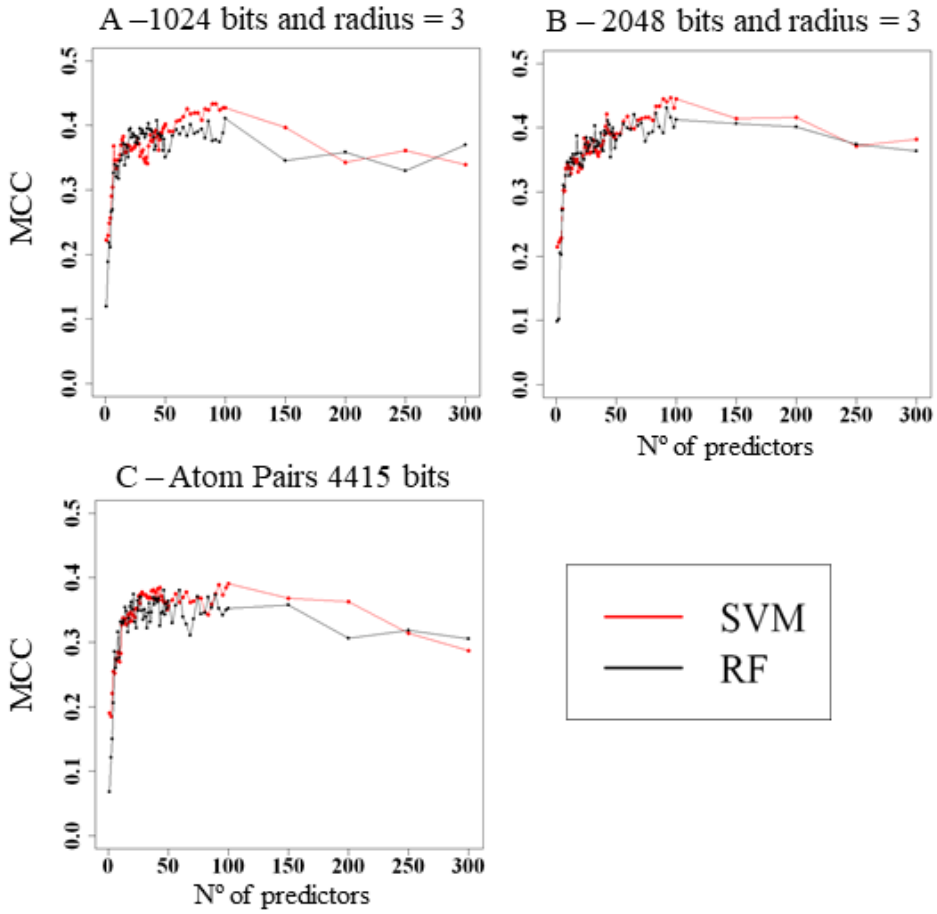



Figure 3.5: **Comparison of RF and SVM in Distances Models.**
Models were constructed using SVM and RF algorithms with distances calculated with Morgan Fingerprints with radius = 3, with 1024 bits (A), 2043 bits (B) with Atom Pairs with 4415 bits (C). Principal Coordinates (PCoA) were used as predictors. Models were scored with the MCC metric, with 10-CV.

If the dataset has n molecules, the number of dimensions on which the vectors are mapped can be increased up to n-1, which is the number of dimensions for which the distances between the n molecules

are faithfully represented. While for visually observing the projected molecules an increasing the number of dimensions makes understanding the data more difficult, it usually enhances the predicting capability of the ML learning models. In this context, the dimensions, or Principal Coordinates, are the predictors, and as such, there is an optimal range of dimensions that maximizes the prediction capability of the models, similarly to what was described previously for the SVM models with Morgan Fingerprints as predictors. To assess which type of model, the fingerprint setting to calculate the distances and the number of predictors to use, SVM and RF models were created with an increasing number of Principal Coordinates as predictors for distances between the molecules in the Data Set. The classification models were scored with the MCC metric in a 10-CV, and representative plots can be seen in (fig 3.5).

For all settings, SVM scored better than RF. For both 1024 and 2048 bits, radius = 3 scored better than radius = 2 (not shown), and Atom Pairs scored slightly lower than these settings. The region with more consistent MCC seems to be between 50 and 100 predictors and after that region there is a decreasing trend. While it cannot be said that these models performed poorly, most models using Morgan Fingerprints as predictors have clearly scored better.

## 3.5 Regression - Choosing Best Models

Before making the regression models, the Inactive molecules were removed from the initial dataset. An analogous process was made to the modeling of the classification models. This dataset was split 9/10 and 1/10 between Training Set and IVS respectively and 7-CV was used to score the models. The metric used for scoring regression models was the Root Mean Square Error (RMSE).

RMSE is a measure of the distance between predicted values and the observed values. It is the squared root of the average of squared errors (formula 2.3).

This measures serves to aggregate the magnitude of the errors in many predictions into a single measure (prediction errors can be described as the distance of the observed points from the regression line or curve). In other words it is a measure of how spread out the data points are from the line of best fit of the model. An ideal RMSE of 0 would mean that there would be a perfect fit between the predictions and observed values, and usually a lower value indicates a better performance on a model. This measure is sensitive to ouliers, since each error's effect on the score is proportional to its squared value. RMSE should only be used to compare models within the same dataset, as it is scale dependent.

A representation of the performance of the models using Morgan Fingerprints as predictors is shown in figure 3.6 Morgan Fingerprints with 1024 bits was chosen because for both radius = 2 and = 3, as their scores were among the best in a most consistent way and for a wider range of number of predictors, suggesting some optimization of SVM in regression tasks for this bit rate, in the region around 200 predictors. Radius = 2 was chosen because between both radius settings it had a region with the lowest RMSE of approximately 5.5.

The distance models didn't perform as well on regression, all scoring above approximately 5.9 RMSE in the best performance settings.

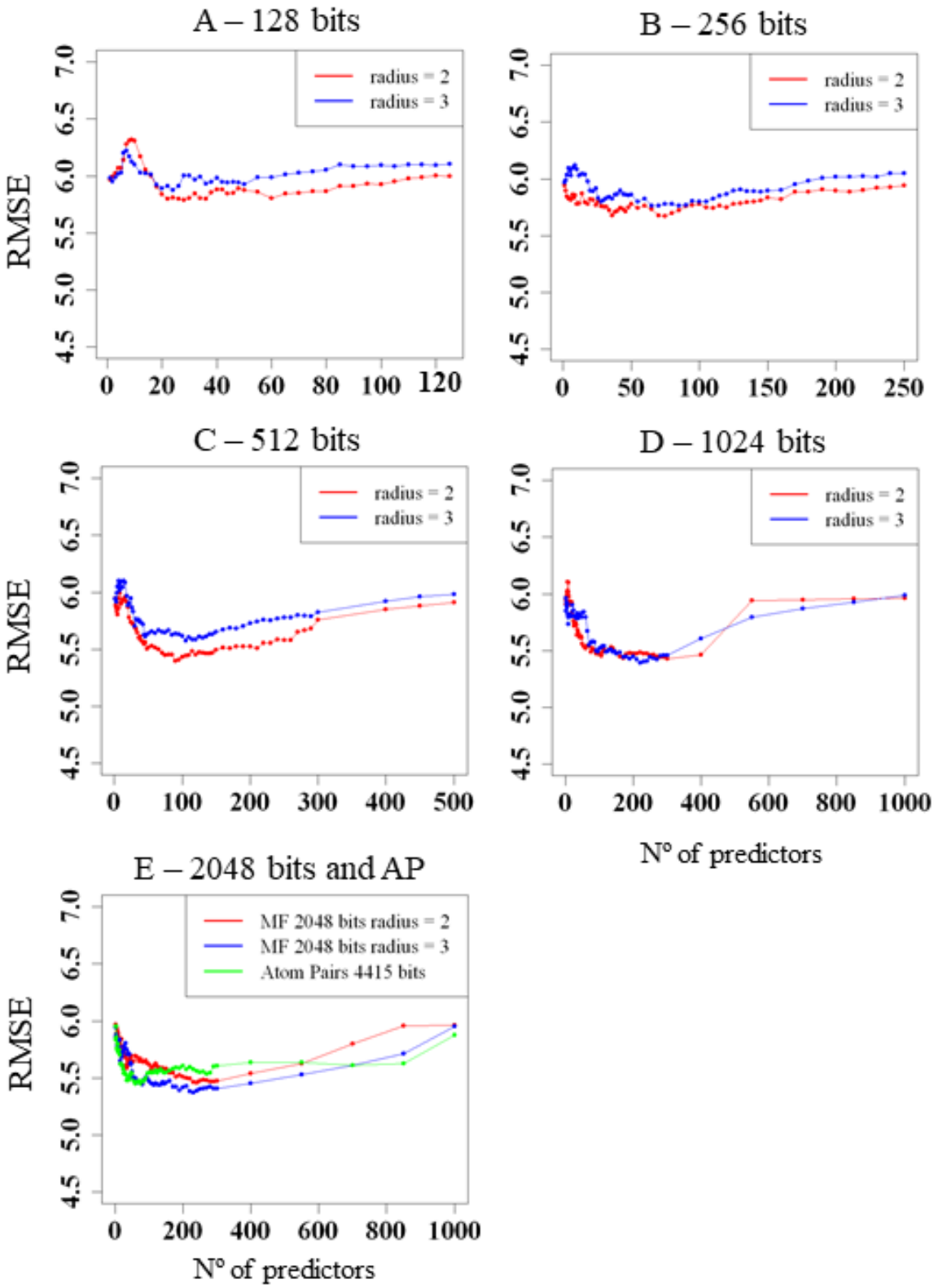



Figure 3.6: **Scoring SVM Regression Models Using Molecular Fingerprints as Predictors.** SVM was used to create the models and RMSE was used as score metric. For each fingerprint setting, the importance of the predictors was estimated, and the models were trained with increasing numbers of predictors.

## 3.6 Validation of the Most Promising Models

After identifying the most promising models, their performance was tested against molecules in the IVS for validations. The final models chosen, their correspondent settings and scores can be seen in table 3.1.

Table 3.1: **Validation of the Best Models.**
Description of the type, settings and scores of the chosen models. CV is the type of cross-validation used to score the models in training and IVS is the score with independent validation set.

| | | Data Settings | | | | Scores | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Models | Type of Predictors | Fingerprints | | N° of Predictors | CV | | IVS | Metric |
| Classification | SVM | Fingerprints | Morgan | 1024 bits and radius = 2 | 132 | 7-CV | 0.49 | 0.48 | MCC |
| | SVM | Fingerprints | Morgan | 1024 bits and radius = 3 | 236 | 7-CV | 0.49 | 0.48 | MCC |
| | SVM | Fingerprints | Morgan | 2048 bits and radius = 2 | 226 | 7-CV | 0.50 | 0.48 | MCC |
| | SVM | Distances | Morgan | 1024 bits and radius = 3 | 65 | 10-CV | 0.41 | 0.36 | MCC |
| | SVM | Distances | Morgan | 2048 bits and radius = 3 | 89 | 10-CV | 0.44 | 0.31 | MCC |
| Regression | SVM | Fingerprints | Morgan | 1024 bits and radius = 2 | 210 | 7-CV | 5.47 | 5.33 | RMSE |

The validation of the distance models wielded a bigger difference from the CV scores. An in depth look at the results showed that this different was not sufficient to exclude these models, as it was only evident in specific ranges of numbers of predictors. It does however reinforce the notion that the distance models do not perform as well as the models using Morgan Fingerprints in this dataset.

## 3.7 Screening the ZINC15 Database

Following the task of choosing the general architecture and choosing the best models, the aim was to use these models to screen the free database for commercially available compounds ZINC15 (Sterling and Irwin, 2015) for the most promising drug candidates for CF.

For screening, the models were trained with the entire dataset, instead of only the Training Set. Once the best settings were determined, training the models with the complete dataset provides the whole available chemical information of the dataset from which to predict activity.

The initial task step was to convert the entire dataset of compounds in-stock from SMILES format to Morgan Fingerprints. This task was considerably resource consuming, since this dataset contained 13,123,788 substances and the models used 4 different settings of Morgan Fingerprints, 1024 bits with both radius = 2 and = 3 and 2048 bits with radius = 2 and = 3. The resulting data accounted for approximately 200 GB of text data. Managing the size of the files and how the data was processed by the R scripts was taken into account, since it was easy to occupy the entire RAM of the available servers if this aspect was overlooked.

Preliminary results on fractions of the screening dataset indicated that additional filtering steps were

necessary. The sheer number of compounds classified as Active by a single ML models was too high, around 30-45%. This suggested two problems; first was that the amount of results was inherently too high, and second, the models were classifying too many compounds as actives. One explanation for this is that ML algorithms are designed to always provide results. The training data used for modelling corresponds only to a tiny fraction of all the structures of available compounds. This implies that models must decide how to classify molecules that are very different from what was modelled as either an Active or Inactive molecule. Even if these molecules are bad fits to the model, they will be classified as Actives, if they happen to fit the model closer to what is modeled as an Active molecule.

There were several possible approaches to these problems. Since the number of drug candidates needed to be drastically reduced, to under 100 at least, it was decided to include several filtering steps to the screening process.

One possibly approach could be to create a set of dummy structures with Inactive classification. This could also be accompanied with generating a projection of possible values for the regression tasks in Inactive molecules, between 0 and 42. This would allow the regression models to fit the inactive molecules along a much wider range of scores, filling the gap of scores in the initial dataset (see chapter 2.3). While interesting and likely result wielding, this approach wasn't followed, since a problem with this approach is that it introduces a lot of artifacts into the modelling and could reduce the capacity of the models to correctly identify true positives.

A first filtering step was introduced, with molecular structure-based kriging, based on the approach described in Teixeira and Falcao (2014). Similar structured molecules tend to have similar properties. The procedure relied on structure Jaccard/Tanimoto similarity, with Morgan Fingerprints with 1024 bits and radius = 3.

For each molecule in the screening dataset, the 20 most similar molecules in our dataset were gathered. These 20 molecules were used to train a RF model. The screened model would then be classified with this model, and if predicted Active, was kept for further steps. RF was chosen because they inherently perform better without predictor selection, due to the inclusion of bootstrapping in their algorithm. This step greatly reduced the number of screened substances classified as Active, to approximately 100.

The remaining compounds were then subjected classification by the 5 ML models previously validated (table 3.1), and a score of 0 to 5 was attributed based on the sum of the results classified as active. Molecules scoring under 4 were excluded.

The regression model (in table 3.1) was then applied and the compounds scoring over 48 were selected for final analysis. A total of 59 compounds remained, 39 of which were equal to compounds present in our initial dataset. Although 59 compounds were already a feasible amount for *in vitro* screening, manual curation of these compounds was still advised. Each compound was manually checked in ZINC15 and ChEMBL (a manually curated database for bioactive drug-like small molecules) (Gaulton et al., 2016) for a combination of known or predicted properties. The presence of structures attributed to pan-assay interference compounds (PAINS) (Baell and Walters, 2014), was a criterion for exclusion. PAINS are compounds that are typically present in assay screens, that contain structures that tend to produce false positive results. These structures are well characterized. ChEMBL has an online tool that identifies or predicts their presence and compounds containing them are best left out of screening assays.

Another aspect that was assessed was the likeness of the compound for being an orally administered drug with the Lipinski's rule of 5 (RO5) (Lipinski, 2004). RO5 is a rule of thumb for the chemical and physical properties that most bioactive oral drugs share, such as size and lipophilicity. Is it usually applied by assessing the number of violations to the criteria of not having more than 5 hydrogen bond, not having more than 10 hydrogen bond acceptors, having a molecular mass less than 500 Daltons and an octanol-water partition coefficient (log P) that does not exceed 5. Compounds with more than 1 violation were also excluded. There was a special interest on compounds already FDA approved, such as Colchicine and Mestranol. Since these compounds have already been extensively tested for human consumption, they were especially attractive for repurposing.

The final step was choosing the suppliers, amounts and in which form to order. Preference was given to compounds already in solution, for reduction of human manipulations, which induce uncertainty in the final concentrations and increase the risk of contamination. Not all these compounds were available, and some seemed to correspond to the same compounds, due to redundancies or presence or salts. 28 compounds were selected and ordered.

## 3.8 Compound Screening with Immunofluorescence F508del-CFTR Traffic Assay

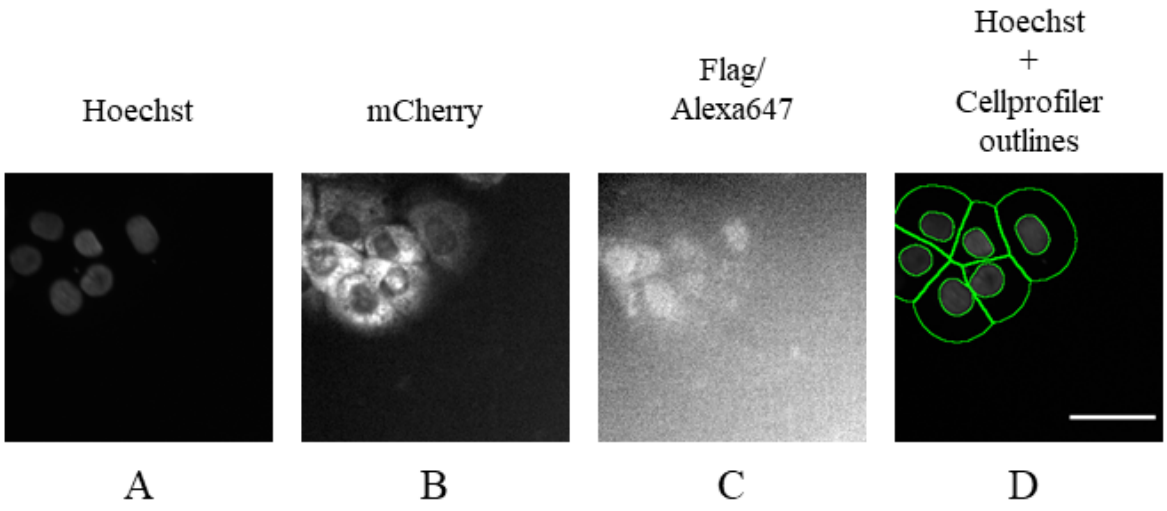



Figure 3.7: **Immunostaining characterization under microscopy of the CFBE cell lines expressing mCherry-Flag-F508del-CFTR and automated image analysis using CellProfiler software.** Cells were grown in the presence of 1 $\mu$g/mL Dox to induce expression of CFTR with C16 compound at 1 $\mu$M concentration. (A) Nuclei stained with Hoechst 33342. (B) mCherry fluorescence is proportional to the total amount of expressed CFTR. (C) Alexa Fluor© 647 immunofluorescence is proportional to the amount of Flag tags exposed extracellularly (i.e. CFTR localized to the PM). (D) Representation of outline of cells and nuclei by CellProfiler software. Scale bar = 53 $\mu$m.

28 drug candidate compounds were obtained to perform a high-throughput screening with 3 different concentrations, on the levels of CFTR expressed in CFBE cells stably transduced with Flag-mCherry-

F508del-CFTR CFBE cells. It was also chosen to include the FDA approved CF therapeutics VX-770, VX-661 and VX-809 in their optimal concentrations (Botelho et al., 2015; Matthes et al., 2016; Awatade et al., 2019).

For a detailed description of the procedure, see Chapter 2.6.

The overall characterization of these cells under the epifluorescence microscopy settings used can be seen in figure 3.7.

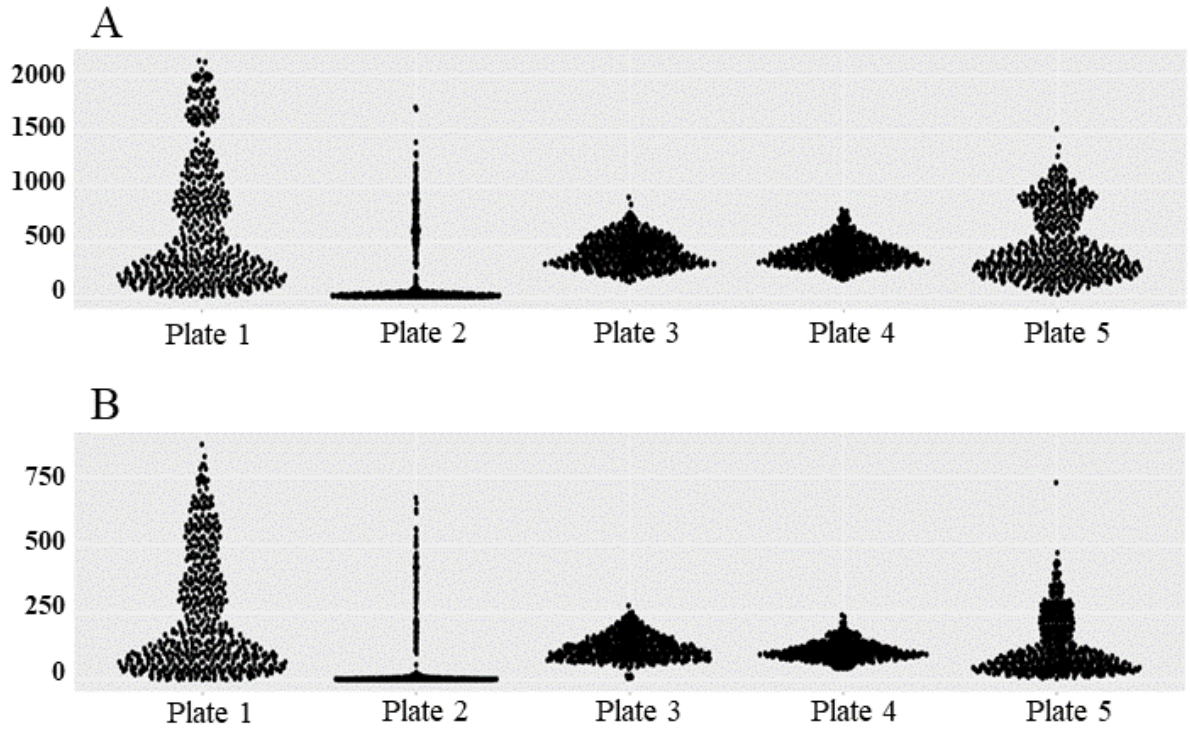



Figure 3.8: **Total and Final Cell Count.**
Automated cell count on 96-well plates by replicate (plates) through cell profiler and analyzed using shinyHTM. (A) total cell count after initial nucleus size filtering. (B) final cell count, after all quality control filters.

CellProfiler software was used for automated analysis of the microscope images. For each image field, 3 images were taken, one of the nuclei (Fig. 3.7-A), one of the total CFTR fluorescence (Fig. 3.7-B) and one of the PM CFTR fluorescence (Fig. 3.7-C). From the nuclei images, the CellProfiler software would outline each identified cell nucleus and expand a second area selection, corresponding to the each cell's cytoplasm delimited by its plasma membrane (PM) (Fig. 3.7-D).

After fluorescence background correction and several QC steps, the data of the total CFTR fluorescence and PM fluorescence was exported and analyzed through the shinyHTM script. Standard QC control steps were performed in shinyHTM, such as Image focus, minimum number of cells and minimum PM fluorescence. An analysis of the total and final cell counts by plate confirmed what was previously observed during the experiment. DMSO has some toxicity towards animal cells (Galvao et al., 2014), so all wells required the same concentration of DMSO. This toxicity is clearly exemplified in plate 2 (Fig. 3.8). The most likely cause for the markedly low number of cells in this replicate is due to pipetting

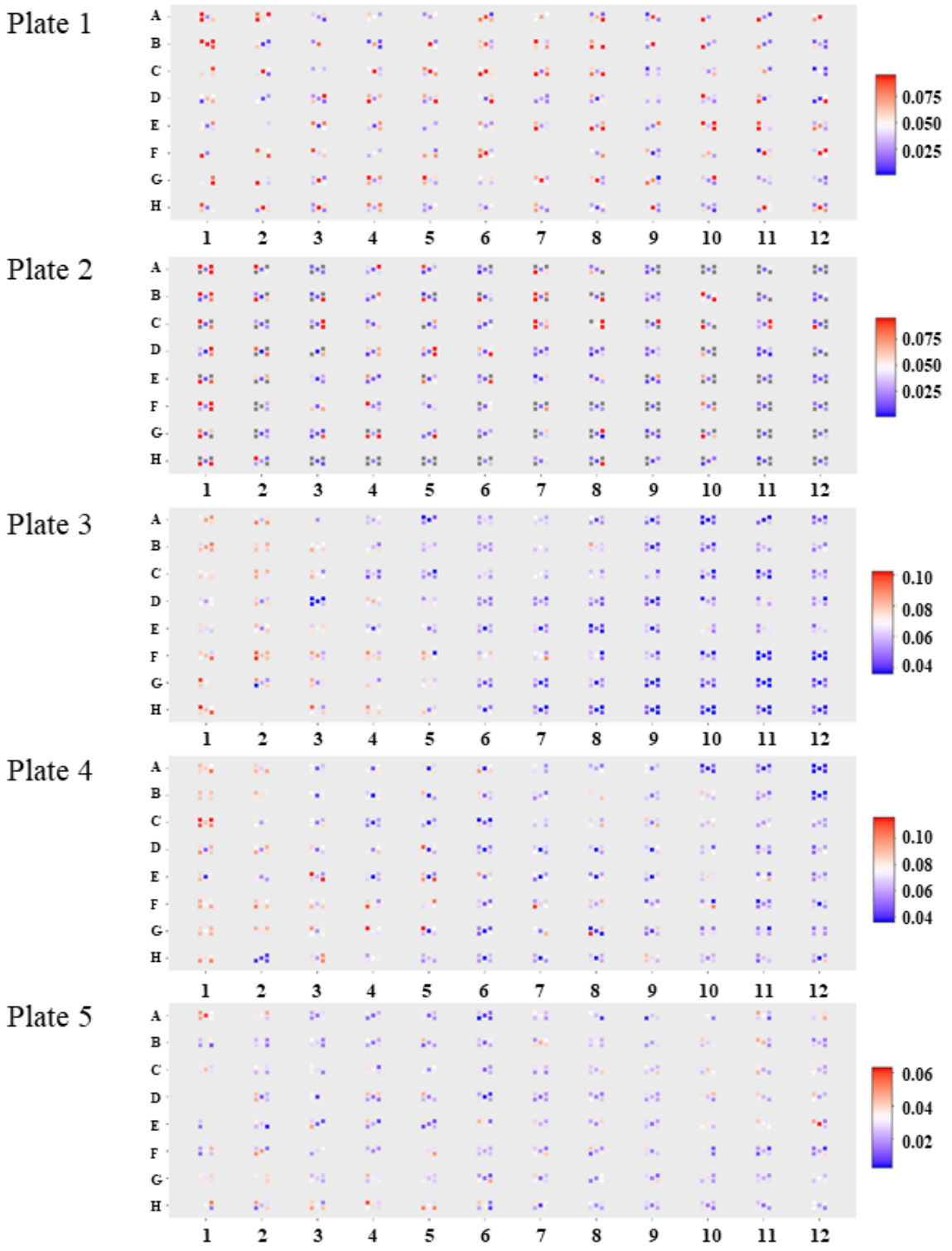



Figure 3.9: **Ratio of Fluorescence Between PM and Total Fluorescence by Plate.**
Each point represents the ratio of fluorescence between PM and Total fluorescence (traffic efficiency) of an image field in a well of a 96-well plate. The range of data points of the heatmap is not the same for each plate. It was adapted in order to see how the fluorescence values vary according to the location of the wells for each plate.

errors, specifically, increased DMSO concentration, having a 1 % (v/v) DMSO instead of 0.1 %. For this reason, plate 2 was excluded from further analysis.

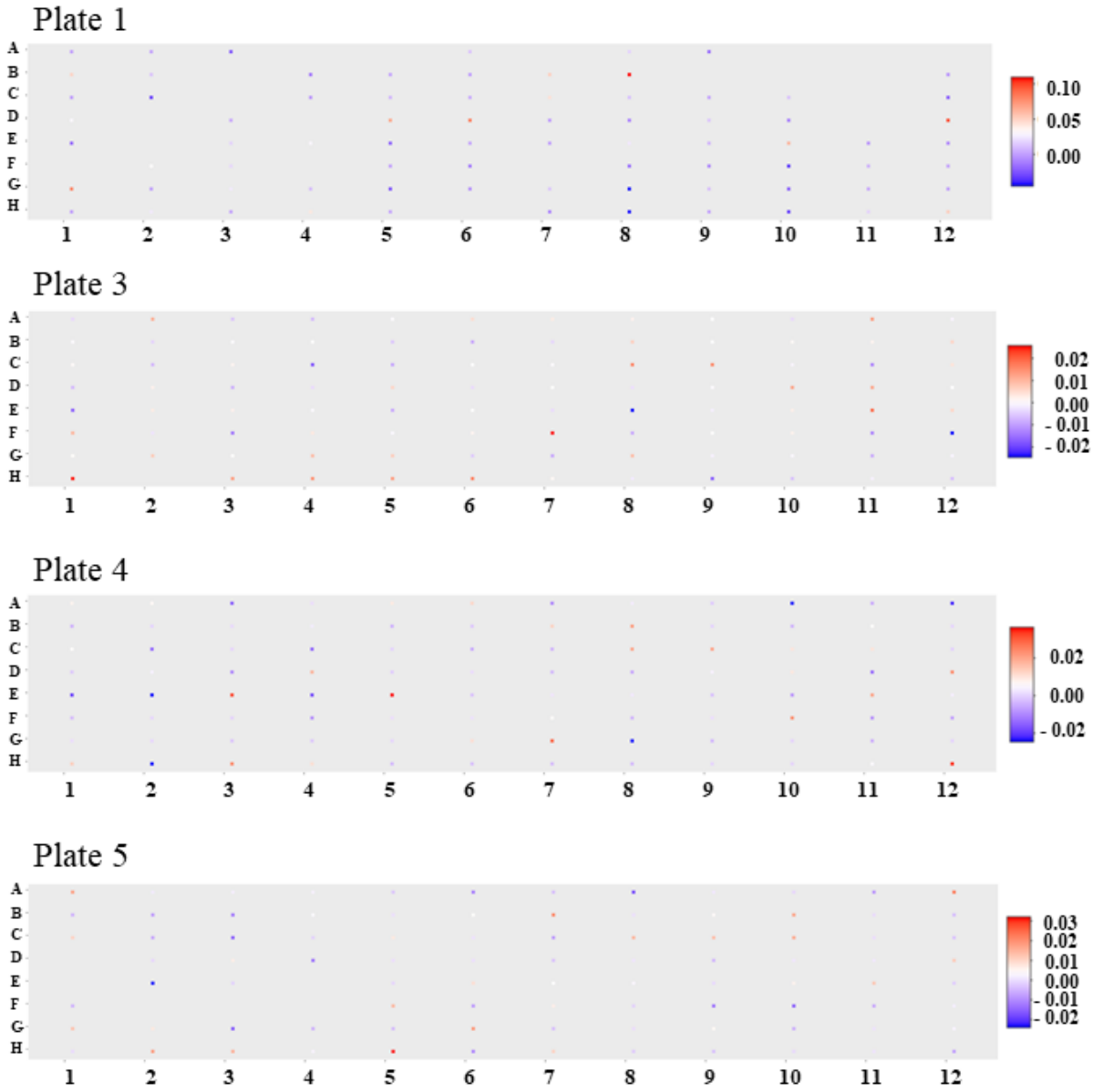



Figure 3.10: **Ratio of Fluorescence Between PM and Total Fluorescence by Plate after Median Polish Normalization.**
Each point represents the ratio of fluorescence between PM and Total fluorescence (traffic efficiency) in a well of a 96-well plate after median polish normalization. The range of data points of the heatmap is not the same for each plate. It was adapted in order to see how the fluorescence values vary according to the location of the wells for each plate.

A preliminary summary of results was made, with the Median Z-score of the traffic efficiency (ratio between PM and total CFTR fluorescence) as the main score for the treatments. Upon comparison of the results with the disposition of the treatments on the 96-well plates (Supplementary Fig. S1), there seemed to be a bias of higher scores towards the left edges of the plates. An initial look upon the fluorescence results had been made through heatmaps of fluorescence by plate, however, without careful

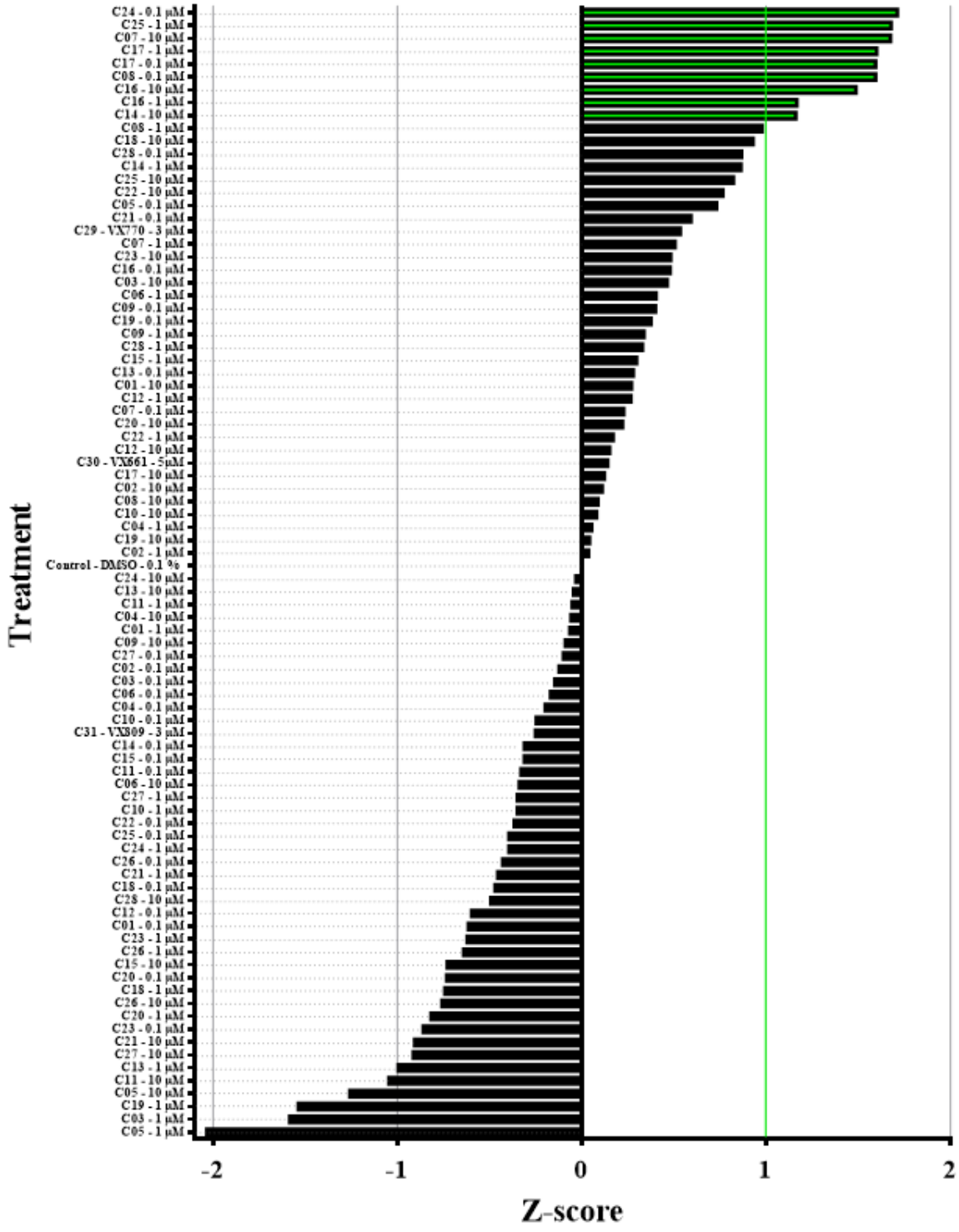



Figure 3.11: **Extensive Scores of Ratio of Fluorescence Between PM and Total CFTR in F508del-CFTR Immunofluorescence Assay.**
Median of Z-scores of fluorescence ratio between Plasma Membrane and Total Fluorescence (traffic efficiency) by treatment. Z-score of 1 is marked as a green line, being the threshold above which, compounds with median Z-scores were considered promising (compounds marked as green). DMSO only treatment was used as control. Ordered by median Z-score of traffic efficiency.

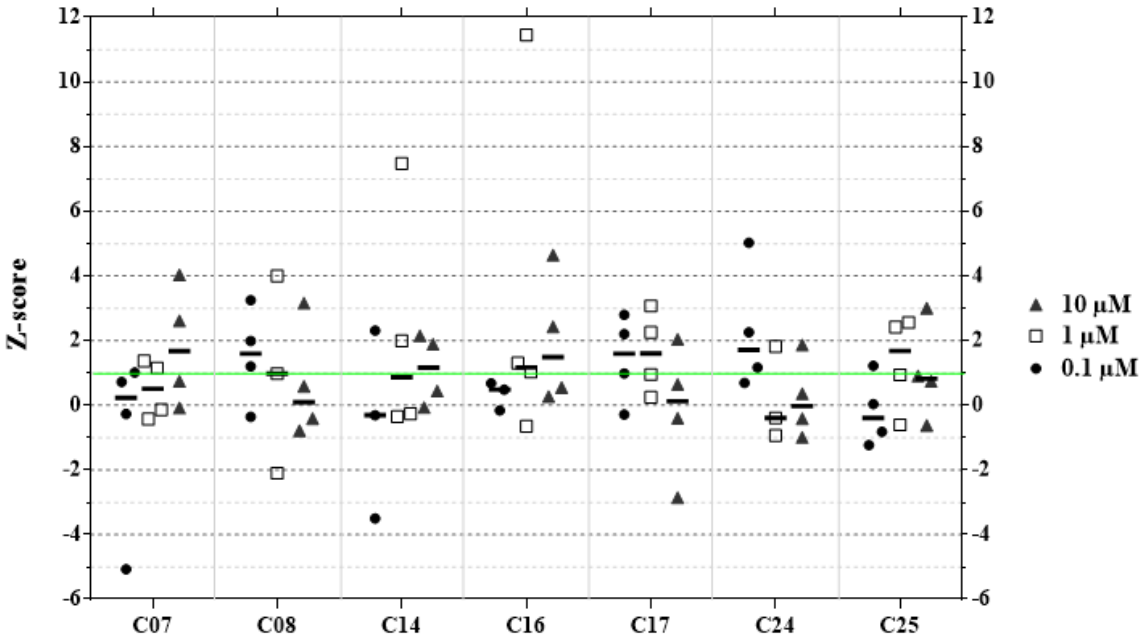



Figure 3.12: **Z-scores of Fluorescence Ratio Between Plasma Membrane and Total Fluorescence by Treatment.**
Z-score of fluorescence ratio between Plasma Membrane and Total Fluorescence (traffic efficiency) for each compound is represented by concentration for each replicate, and the median Z-score for that concentration is represented as a horizontal bar. Z-score of 1 is marked as a green line, being the threshold above which compounds with median Z-scores were considered promising. DMSO only treatment was used as control.

adjustment of the range of data points represented in the color gradient (or LUT, look up table), the fluorescence gradient can easily be overlooked. Upon adjusting the heatmap representations of the ratios of fluorescence for each plate, it was noticed that there was in fact a gradient of fluorescence, according to the disposition of the wells on each plate, in most cases with a bias of higher values towards the left side and the edges of the plates (Fig. 3.9). The gradient of fluorescence by plate was also directly observable, confirming the initial impression, as the result function of applying a 5x5 median normalization (which was not the final normalization used) (Supplementary Fig. S2). Plate 2 was especially useful to confirm the existence of this gradient, since it had such few cells, it would be hard to explain this fluorescence gradient through biological activity.

To address this issue, a median polish normalization was applied to each plate individually, resulting in what was thought as a compression of the 5 imaging positions per each well into one data value. The corrected fluorescence ratio values can be visualized in a heatmap in figure 3.10.

Unfortunately, only after the whole pipeline of this project was performed, it was discovered that a "bug" existed in the median polish normalization algorithm in shinyHTM, resulting not in a compression of values (by a median or mean), but in excluding all image positions except one per well. This has resulted in a loss of statistical significance in the analysis of this results, which is not critical, since the "bug" has since been fixed and the fluorescence results were not lost and can be re-analyzed. What is more unfortunate is that some hits might have been lost in this analysis, resulting in a WB assay not

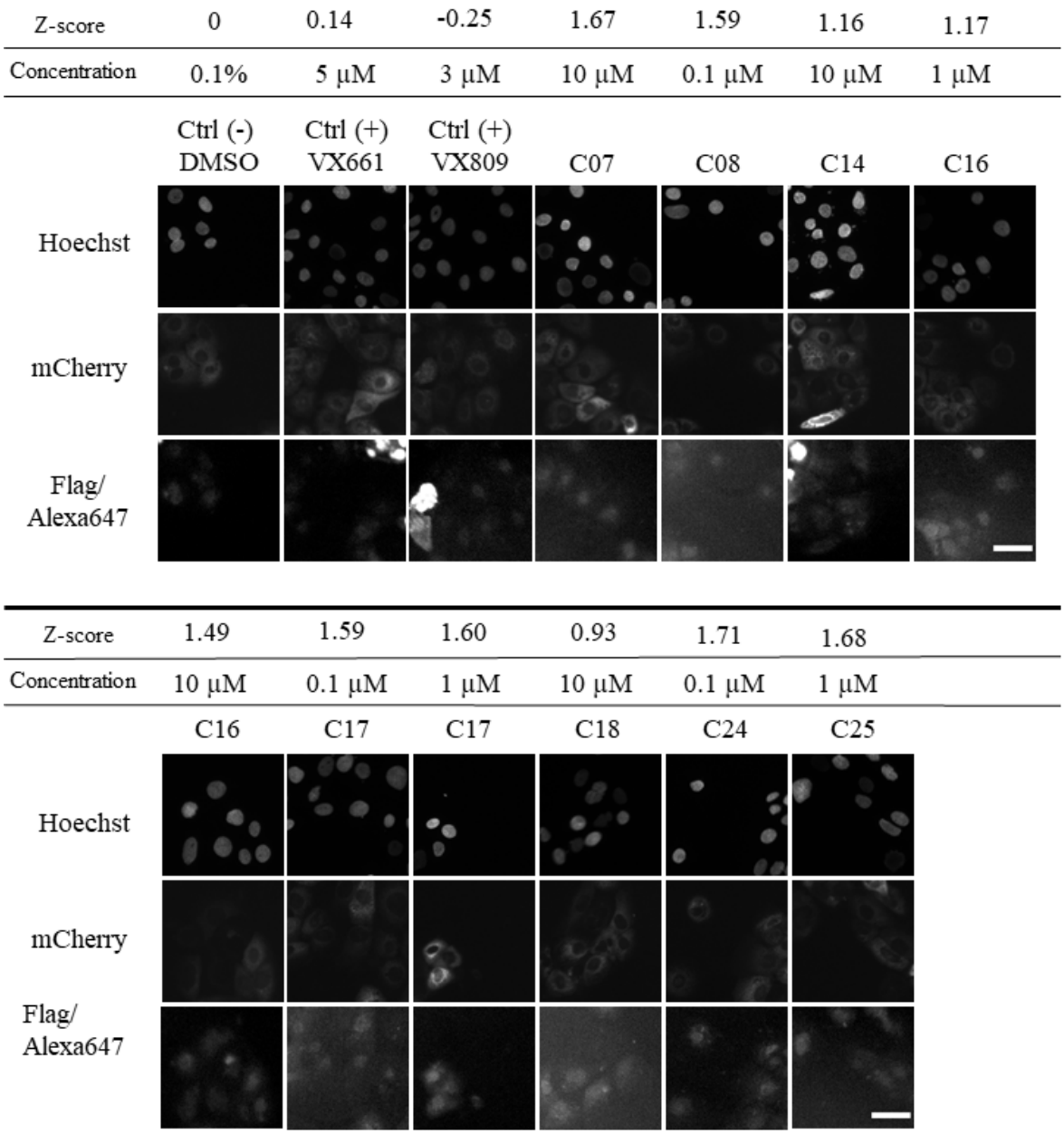



Figure 3.13: **Hits from the Compound Screening Assay with CFBE cells Expressing the mCherry-Flag-F508del-CFTR.**
Cells were treated with DMSO as negative control, and 31 compounds, including VX-809 and VX-661 as positive controls. Representative images are shown of the compounds with a Z-score $>1$ or very close to 1. Shown Z-scores are the median of the Z-scores of the traffic efficiency (PM fluorescence/Total fluorescence) of each replicate for each treatment. Concentration is the concentration of each compound. Nuclei stained with Hoechst 33342, mCherry fluorescence is proportional to the total amount of expressed CFTR and Alexa Fluor© 647 immunofluorescence is proportional to the amount of Flag tags exposed extracellularly (i.e. CFTR localized to the PM). Scale bar = 53 $\mu$m.

containing all promising compounds.

After this correction the Z-scores were calculated for the traffic efficiency for all treatments, compared to DMSO only as control. The extensive results can be seen in supplementary tables S1 and S2.

A plot of the median of the traffic efficiency Z-scores shows the treatments, ordered by Z-score (Fig. 3.11). The main criteria chosen to determine which compounds were chosen as hits, was having a traffic efficiency Z-score of 1 or above, and n >= 3.

Some treatments didn't reach a Z-score of 1 but were very close, such as C08 – 1 $\mu$M with a Z-score of 0.98 (but only n=3), and it's worth to note that C08 at 0.1 $\mu$M was considered hit with a Z-score of 1.59 (Fig. 3.11, descriptive and tratamentos) and the case of C18 – 10 $\mu$M with a Z-score of 0.93 (Fig. 3.13) and a Z-score of PM CFTR of 0.73. For this reason, it was decided to also include C18 at 10 $\mu$M as a hit, and it was also included in the WB assay. C28 – 0.1 $\mu$M would also have been an interesting compound to test, with a Z-score of traffic efficiency of 0.87 and a Z-score of PM CFTR of 1.15, but for practical and efficiency reasons in the design of the WB experiment only 10 screening treatments were chosen. The treatments chosen for the WB assay were, 10 $\mu$M of C7, 0.1 $\mu$M C8, 10 $\mu$M C14, C16 in 1 $\mu$M and in 10 $\mu$M, C17 in 0.1 $\mu$M and in 1 $\mu$M, 10 $\mu$M C18, 0.1 $\mu$M C24 and 1 $\mu$M C25. For positive controls 3 $\mu$M VX-809, 5 $\mu$M VX-661 and for negative control DMSO only 0.1 % (v/v), for both wt-CFTR and F508del-CFTR CFBE cells (Fig. 3.14). Even though C16 at 1 $\mu$M has a traffic efficiency Z-score outlier value of 11.46, this treatment was still considered a hit, given that 2 other replicates had values above 1.

Both assays were targeted at estimating the localization of F508del-CFTR to the membrane, either directly (immunofluorescence) or indirectly though assessment of proper glycosylation (WB). However, the assay for which the dataset on which the modeling of the machine learning models was made was a functional assay. This means that the effects on $Cl^-$ transport observed could have been due to different approaches to overcome CFTR dysfunction and not necessarily through increased processing (e.g. increased total CFTR, increased CFTR on the membrane through overactivation of other proteins in CFTR's interactome, ENac inhibition (to decrease the characteristic sodium hyperabsorption) and also activation of alternative chloride channels (Farinha and Matos, 2015). There can be many ways by which there was an increased anion transport, not all directly related to CFTR function. It cannot be excluded then, the possibility that there could still be a correction of the $Cl^-$ transport in a manner that could be clinically significant, especially if combined with other correctors. Another possibility for false positives in the immunofluorescence assays is the overestimation of the mCherry (total CFTR) and Alexa© 647 fluorescence (PM CFTR) due to innate fluorescence of the screening compounds. Many of these compounds had color in solution (Supplementary Fig. S4), one of them being fluorescein (compound C28), a known and widely used dye for biological assays, and another compound being molecularly closely related (carboxyfluorescein, compound C10). Interestingly, carboxyfluorescein (C10) is noted on ChEMBL as a Targets Solute carrier family 22 member 6, although neither of these compounds was deter-mined a hit in the screening (this is, a traffic efficiency Z-score of 1 or higher). Fluorescein (C28) at 0.1 $\mu$M scored fairly positively with a Z-score of 0.87 (Supplementary Table S2). However, as the concentration was increased, there was a negative trend in the fluorescence values, so the data to support the speculation that fluorescein's innate fluorescence could significantly alter the results was not convincing.

## 3.9 Western Blot Assay

After selecting the most promising drug candidates in the fluorescence assay, the effect of these compounds in the appropriate concentration was assessed with a Western Blot (WB) assay. CFTR is synthesized and folded in the endoplasmic reticulum (ER), undergoing traffic and processing through the Golgi complex, and becoming functional at the PM (Amaral, 2015). wt-CFTR is detected by WB as both a mature, fully glycosylated (band C) and an immature, core-glycosylated (band B) forms. In F508del-CFTR, only the immature core-glycosylated form of this protein, consistent with its ER retention, is present. In a WB, the band B with approximately 135 KDa and the band C is detected at a higher molecular weight of approximately 180 KDa (Farinha et al., 2013). This is visible in figure 3.14, where wt-CFTR is clearly shown as two separate bands. In this assay two types of cells were used, wt-CFTR CFBE and F508del-CFBE. wt-CFTR CFBE was included for comparative analysis (as it evidences both bands B and C) but not used in quantification.

Calnexin was used as internal control, for normalizing protein levels between treatments. Treatment of F508del-CFBE with DMSO only was used as negative control, for normalizing the CFTR quantity levels. Three replicates were made under the same conditions.

The previous fluorescence results indicated that the positive controls used on this assay, VX-661 and VX-809, were not being as effective as expected in increasing CFTR processing. The WB assay results however showed otherwise. While not forming a clear band in the region expected to form CFTR's band C, there is some effect, visible as a smear in figure 3.14. This diffuse pattern is typical of band C CFTR, and the appearance as smear when compared to the strong band visible in wt-CFTR probably derives from the fact that the amount of processed protein is much lower.

It is also visible that some compounds present a smear comparable to the positive controls, in some cases even more noticeable, such as C14 in 10 $\mu$M, C16 in 1 and 10 $\mu$M, C17 in 0.1 and 1 $\mu$M and C25 $\mu$M in figure 3.14. Also noteworthy is that some of the treatments seem to have increased expression of total CFTR. To better understand these results, the CFTR and Calnexin levels were quantified and presented in different metrics, Processing (ratio of band C over total CFTR), Matured CFTR (ratio of band C to loading control) and Total CFTR (ratio of bands B and C to loading control), normalized to DMSO only treatment as negative control (Fig. 3.15). The results were presented as the mean percentage of effect compared to the control in all replicates. Processing is a measure similar to the traffic efficiency in the fluorescence assay, that gives a ratio between mature CFTR (band C) and the total CFTR in that condition (band C + band B). This measure estimates how much of the total F508del-CFTR was processed to the PM. Matured CFTR estimates the presence of mature CFTR (band C), normalized to Calnexin as an internal control of quantity of protein in the cells for that treatment. Total CFTR is the sum of the band B with the band C, normalized to Calnexin as internal control of quantity of protein in the cells for that treatment.

In the quantification the effects of VX-661 and VX-809 on CFTR (Fig. 3.15), even though the effects are not as evident as expected, they are clearly observable. VX-661 increased the processing to 18 % and VX-809 in 59 %, confirming the expected effect on CFTR processing described on the literature (ref). All compounds seem to increase total CFTR, not only comparing to DMSO but also comparing to VX-661 and VX-809. Of all the compounds, the only one that increased total CFTR with significance

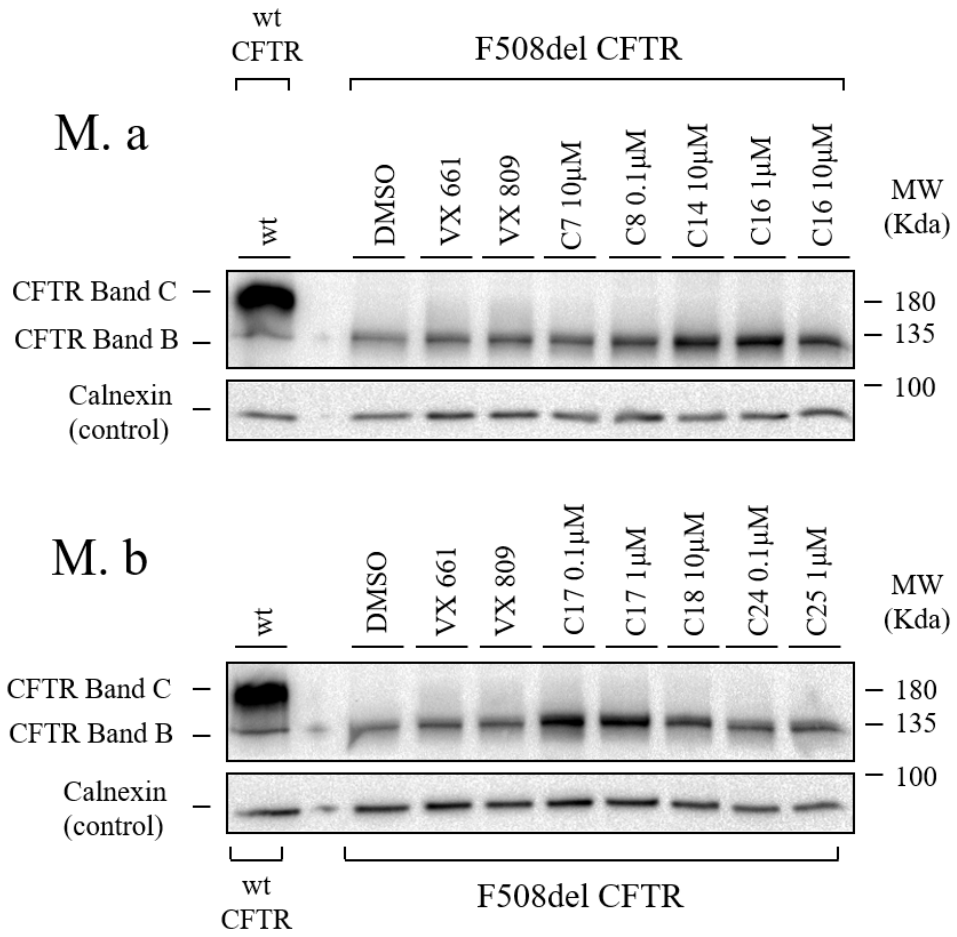



Figure 3.14: **Representative Image of Western Blot of Drug Candidate Compounds.** F508del-CFBE cells were treated with 5 $\mu$M VX-661, 3 $\mu$M VX-809, 10 $\mu$M C7, 0.1 $\mu$M C8, 10 $\mu$M C14, C16 in 1 $\mu$M and in 10 $\mu$M, C17 in 0.1 $\mu$M and in 1 $\mu$M, 10 $\mu$M C18, 0.1 $\mu$M C24 and 1 $\mu$M C25 (DMSO was used as a negative control) for 48 h. Calnexin was used as internal control and wt-CFBE cells were used for molecular weight reference (n=3).

was C14 at 10 $\mu$M.

The only compound with a significant effect on the processing of CFTR was VX-809. The remaining compounds either had a small increase, or a decrease in this metric. VX-809 is well characterized, and one of its main effects is to increase the stability of the immature form of CFTR, leading to an increase in total CFTR, and increased efficacy of rescue (Farinha et al., 2015). The initial focus of the fluorescence analysis was mainly on traffic efficiency, however the WB results let to a rethinking on the importance of relying solely on this measure on a CF drug screening assay. Compounds that increase total CFTR or increase just the F508del-CFTR that reaches the membrane could be equally effective in a clinical context. The reason for the initial criteria was that our main effort was to discover candidate drugs through increase in traffic efficiency, through rescue of function of F508del-CFTR, and this reflects on the proportion between total CFTR and CFTR that reaches the PM. Even if the traffic efficiency/processing aren't significantly increased, the fact that more CFTR localizes to the membrane with some compounds is already a desirable outcome, and in those cases a small increase in CFTR processing can be misleading, since there is an increase in the total amount of CFTR. Even in cases

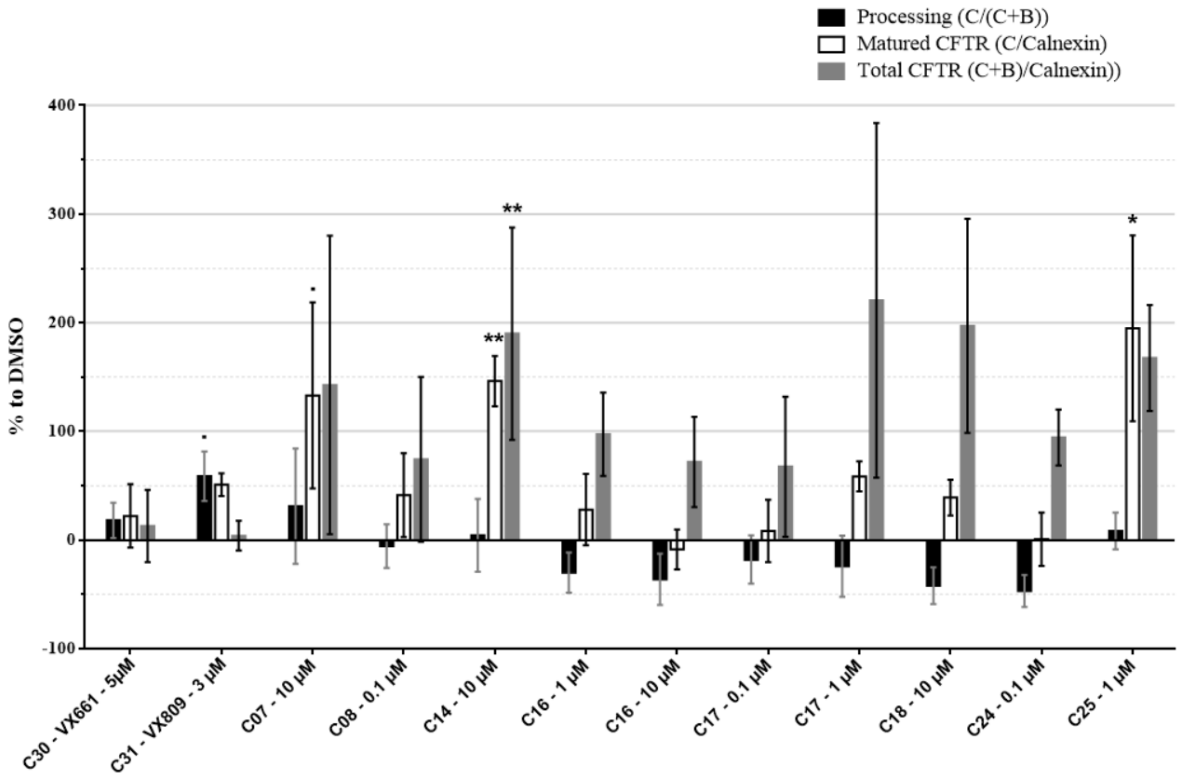



Figure 3.15: **Quantification of Western Blot results.**
F508del-CFBE cells were treated with 5 $\mu$M VX-661, 3 $\mu$M VX-809, 10 $\mu$M C7, 0.1 $\mu$M C8, 10 $\mu$M C14, C16 in 1 $\mu$M and in 10 $\mu$M, C17 in 0.1 $\mu$M and in 1 $\mu$M, 10 $\mu$M C18, 0.1 $\mu$M C24 and 1 $\mu$M C25 (DMSO was used as a negative control) for 48 h. CFTR levels were assessed by western blotting. Results are presented as CFTR processing (amount of band C normalized to total amount CFTR (band C+B)), Matured CFTR (amount of band C normalized to Calnexin as internal control), and total CFTR (amount of band C+B normalized to Calnexin as internal control). Data shown as a percentage of variation to DMSO only treated cells. Data represent mean $\pm$ s.e.m. (n=6 for VX-661, VX-809 and DMSO and n=3 for the remaining treatments). . $p < 0.1$, * $p < 0.05$ and ** $p < 0.01$ (p-value of unpaired t-test).

where there isn't a significant increase in CFTR localized to the PM, compounds that increase total CFTR can also be interesting and useful. Increase in total CFTR can be due to effect on one or more of many different levels of transcription and signal transduction pathways and protein interactions. Some examples are increase in the production of CFTR, increase in stability and inhibition of degradation. These results show a trend with borderline significance. It is fair to say that some of these compounds have an effect that is comparable to VX-661 and VX-809, or even more prominent. The compounds C07, C14 and C25 had significant effects on the quantity of mature CFTR and C14 had significant effects on total CFTR. Even if separately the effects of these compounds is not enough *in vivo*, a combination of these compounds with other correctors could prove to make a significant improvement in CF pathology, such as been described before (Taylor-Cousar et al., 2019; Farinha et al., 2013).

## 3.10 Revisiting the Immunofluorescence Assay

The importance of evaluating the total CFTR and PM CFTR was overlooked in the analysis of the immunofluorescence results. Due to the functional nature of the data used to create the dataset, it was not possible to know which molecular mechanism led to the increase on $Cl^-$ transport. For that reason, one should not exclude the search for potential therapeutics that act on alternative processes to increasing F508del-CFTR processing.

It is then suggested that on future assays following this pipeline, a more comprehensive criteria for determining hits in the high-throughput screening is used, one that besides traffic efficiency, also takes into account the total CFTR and PM CFTR. For a comprehensive representation of the fluorescence results as median Z-scores of traffic efficiency, PM CFTR and total CFTR, see supplementary figure S7.

Interestingly, in this revised analysis of the fluorescence results, C07 in 10 $\mu$M does not seem like a very effective treatment (Fig. 3.16). This treatment was chosen as a hit due to this compound having positive traffic efficiency median Z-scores in all treatments and one surpassing the threshold value. However, upon observing the decrease in values of total CFTR and the slightly increased values of PM CFTR, the traffic efficiency value might seem a bit misleading. It is then interesting to note that in the WB assay, this treatment faired fairly well in increasing the total CFTR (+143% $\pm$ 137) and also the band C which corresponds to the PM CFTR (+133% $\pm$ 86), maintain a favourable traffic efficiency, here represented as Processing (+30% $\pm$ 53), being the highest scoring drug candidate in this measure.

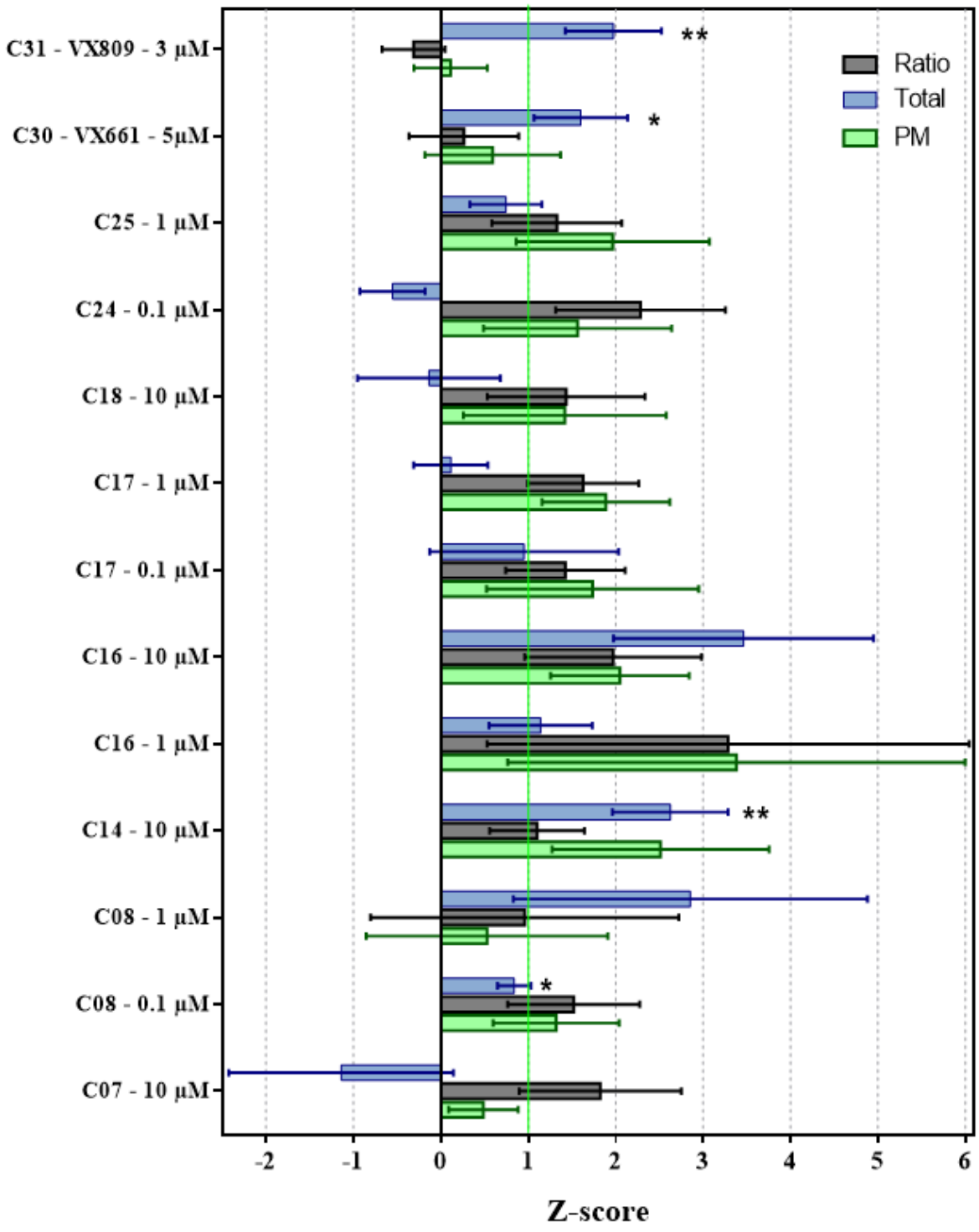



Figure 3.16: **Z-scores of Total CFTR Fluorescence, PM CFTR Fluorescence and Ratio Between Plasma Membrane and Total Fluorescence by Treatment.**
Data represented as Mean Z-scores ± SEM of total CFTR, PM CFTR and fluorescence ratio between Plasma Membrane and Total Fluorescence (traffic efficiency) for each compound is represented by treatment. Threshold Z-score of 1 is marked as a green line, above which compounds with median ratio of fluorescence Z-scores were considered hits. DMSO only treatment was used as control for Z-score normalization. * $p < 0.05$ and ** $p < 0.01$ (adjusted p-value t-test).

# Section 4

# Conclusions

A pipeline for discovery of new candidate drugs through chemoinformatics and machine learning activity prediction models with *in vitro* testing was created.

The SVM machine learning algorithm performed better than RF for tasks of prediction of activity, with estimation of predictor importance and choosing an optimal number of predictors to use. Morgan Fingerprints with 1024 bits and 2048 bits seem to be the best type of predictors to use by these ML models. To overcome inherent limitations of the ML models and of the dataset used to train them, additional filtering steps were required when screening databases for candidate compounds.

Regarding the validation of the HTS fluorescence results with the WB (Fig. 3.15), all drug candidate compounds seem to increase total CFTR, indicating that all have some effect on CFTR. Drug candidate compounds were identified. The most promising compounds seem to be the C07, C14, C17 and C25. C07, C14 and C25 were the only compounds with significant effects on the quantity of mature CFTR. C14 was also the only compound with significant effects on total CFTR in the WB. All these compounds seemed to increase total CFTR and matured CFTR more than VX-661 and VX-809, suggesting a comparable or better effect. It is then recommended that C07, C14, C17 and C25 are further analyzed as potential correctors of F508del-CFTR. These results suggest that the predictive chemoinformatics processes used had the ability to identify compounds with activity.

A greater effect was expected from VX-661 and VX-809. Even though VX-809 was the only compound with significant effects in CFTR processing in the WB (Fig. 3.15) and also had significant effects on total CFTR fluorescence (Fig. 3.16), it is not clear why there is not a distinct band C in the WB in any of the replicates and why both performed poorly in the traffic efficiency metric in the immunofluorescence assay. As such, it is suggested that this assay is repeated with a new batch of cells and new preparations of the compounds.

Re-analysis of the immunofluorescence is also highly advised. None of the imaging data was lost, however the "bug" in the median polish algorithm removed 4 out of every 5 image fields per well from the analysis shown in this report. Although the HTS immunofluorescence pipeline was mostly optimized from previous works, the fluorescence gradient correction was not commonly applied. Even though this is an automated pipeline for the most part, every assay must be careful analyzed with critical thinking for possible problems and overlooked details.

Although this project was directed at the pathology of CF, this pipeline is easily adjustable to other contexts. The aspects to take into careful consideration when applying this pipeline to different biolog-

ical contexts should be the choice and treatment of the initial dataset, the choice of the appropriate cell models with adaptation of the fluorescence constructs and their imaging, tuning the automated analysis pipeline for the size, shape and fluorescence of the cells and to the particular characteristics of the imaging devices. The automated imaging and analysis is easily adaptable to plates with different well numbers (i.e. 384 wells), and as such, can be used for testing higher amounts of treatment.

# Section 5

# Perspectives

An interesting alternative approach could to cluster molecules by presence of certain structures or properties and compare them to their effects. This approach could bring insights on the mechanisms of action and open possibilities of integration new criteria for searching candidate drugs for CF.

Obvious limitations of this approach in the ability for discovery of new drug candidates are tied to the dataset chosen to train the ML models. It is then suggested to build a dataset of known potentiators or correctors of CFTR and of $Cl^-$ transport with different metrics of activity, including also the results of the compounds in this assay and those found in publicly available data. The molecules known to not have positive effects could also be highly valuable to avoid overfitting of the models to the data.

Other layers of retrieving screening hits could be included in this pipeline, such as similarity-based approaches for molecules whose activity values do not fit the dataset chosen for modeling. These should include all known substances that positively affect levels of matured CFTR, F508del-CFTR processing and total CFTR.

There was a "bug" in the algorithm of correction of fluorescence and 4/5 of the imaging data was discarded. It is highly advised to repeat the analysis of the fluorescence results. This would likely increase the statistical significance of the results in this work and probably new hits will be found. If that is the case, the WB analysis should also be repeated. It would also be of interest to test all promising compounds at different concentrations with a WB assay, to estimate an ideal concentration.

To further advance the validation of the drug candidate compounds, future studies should assess potential side effects of the most promising drug candidates, for example on cell survival.

In parallel, the most promising compounds should be tested in functional assays, to assess if the drugs are increasing F508del-CFTR function. These assays could be for example using an Ussing Chamber, a device that replicates epithelial function, and can be used to measure net ion transport (Clarke, 2009).

As previously mentioned, a common approach is to combine correctors and/or potentiators. The effects on F508del-CFTR of the most promising compounds here described should be tested when combined with each other and with the known correctors and potentiators, such as VX-770, VX-661 and VX-809.

# Glossary and Abbreviations

10-CV – 10-fold cross-validation
7-CV – 7-fold cross-validation
CF – Cystic Fibrosis
Ctrl – Control
CV – Cross-validation
Del – deletion (mutation)
DMEM - Dulbecco's modified Eagle's medium
Dox – Doxycycline
EMEM - Eagle's Minimum Essential Medium
F – Phenylalanine
FBS – fetal bovine serum (heat inactivated, Gibco #10106)
HTS - High-Throughput Screening
InChI - IUPAC International Chemical Identifier
IVS – Independent Validation Set
MCC – Mathews Correlation Coefficient
ML – Machine Learning
MW – Molecular Weight
PBS - Phosphate Buffered Saline
PBS-T - Phosphate Buffered Saline with Tween 20
PFA - paraformaldehyde
Phe – Phenylalanine
QC – quality control
QSAR – Quantitative Structure-Activity Relationship
RMSE – Root mean square error
SDS - sodium dodecyl sulfate
SDS-PAGE - sodium dodecyl sulfate–polyacrylamide gel electrophoresis
SMILES - Simplified molecular-input line-entry system
wt – wild-type

# Bibliography


Almaça, J., Dahimène, S., Appel, N., Conrad, C., Kunzelmann, K., Pepperkok, R., and Amaral, M. D. (2011). Functional genomics assays to study cftr traffic and enac function. *Methods in Molecular Biology*, pages 249–264.

Amaral, M. D. (2015). Novel personalized therapies for cystic fibrosis: treating the basic defect in all patients. *Journal of Internal Medicine*, 277(2):155–166.

Amaral, M. D., Farinha, C. M., Matos, P., and Botelho, H. M. (2016). Investigating alternative transport of integral plasma membrane proteins from the er to the golgi: Lessons from the cystic fibrosis transmembrane conductance regulator (cftr). *Unconventional Protein Secretion*, pages 105–126.

Awatade, N. T., Ramalho, S., Silva, I. A., Felício, V., Botelho, H. M., de Poel, E., Vonk, A., Beekman, J. M., Farinha, C. M., and Amaral, M. D. (2019). R560s: A class ii cftr mutation that is not rescued by current modulators. *Journal of Cystic Fibrosis*, 18(2):182–189.

Baell, J. and Walters, M. A. (2014). Chemistry: Chemical con artists foil drug discovery. *Nature*, 513(7519):481–483.

Bell, S. C., De Boeck, K., and Amaral, M. D. (2015). New pharmacological approaches for cystic fibrosis: Promises, progress, pitfalls. *Pharmacology & Therapeutics*, 145:19–34.

Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. (2007). KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer.

Botelho, H., Tischer, C., Halavatyi, A., Amaral, M., and Pepperkok, R. (2019). *shinyHTM - Interactive High-Throughput Microscopy Analysis*. Zenodo.

Botelho, H. M., Uliyakina, I., Awatade, N. T., Proença, M. C., Tischer, C., Sirianant, L., Kunzelmann, K., Pepperkok, R., and Amaral, M. D. (2015). Protein traffic disorders: an effective high-throughput fluorescence microscopy pipeline for drug discovery. *Scientific Reports*, 5(1).

Carhart, R. E., Smith, D. H., and Venkataraghavan, R. (1985). Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Modeling*, 25(2):64–73.

Clarke, L. L. (2009). A guide to ussing chamber studies of mouse intestine. *American Journal of Physiology-Gastrointestinal and Liver Physiology*, 296(6):G1151–G1166.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Deeks, E. D. (2016). Lumacaftor/ivacaftor: A review in cystic fibrosis. *Drugs*, 76(12):1191–1201.

Ehrhardt, C., Collnot, E.-M., Baldes, C., Becker, U., Laue, M., Kim, K.-J., and Lehr, C.-M. (2006). Towards an in vitro model of cystic fibrosis small airway epithelium: characterisation of the human bronchial epithelial cell line cfbe41o-. *Cell and Tissue Research*, 323(3):405–415.

Farinha, C. M., King-Underwood, J., Sousa, M., Correia, A. R., Henriques, B. J., Roxo-Rosa, M., Da Paula, A. C., Williams, J., Hirst, S., and Gomes, C. M. e. a. (2013). Revertants, low temperature, and correctors reveal the mechanism of f508del-cftr rescue by vx-809 and suggest multiple agents for full correction. *Chemistry & Biology*, 20(7):943–955.

Farinha, C. M. and Matos, P. (2015). Repairing the basic defect in cystic fibrosis - one approach is not enough. *FEBS Journal*, 283(2):246–264.

Farinha, C. M., Sousa, M., Canato, S., Schmidt, A., Uliyakina, I., and Amaral, M. D. (2015). Increased efficacy of vx-809 in different cellular systems results from an early stabilization effect of f508del-cftr. *Pharmacology Research & Perspectives*, 3(4):e00152.

Flume, P. A. ., Robinson, K. A., O'Sullivan, B. P., Finder, J. D., Vender, R. L., Willey-Courand, D.-B., White, T. B., Marshall, B. C., and for Pulmonary Therapies Committee, C. P. G. (2009). Cystic fibrosis pulmonary guidelines: Airway clearance therapies. *Respiratory Care*, 54(4):522–537.

Galvao, J., Davis, B., Tilley, M., Normando, E., Duchen, M. R., and Cordeiro, M. F. (2014). Unexpected low-dose toxicity of the universal solvent dmso. *The FASEB Journal*, 28(3):1317–1330.

Gama, J., Carvalho, A., Faceli, K., Lorena, A., and Oliveira, M. (2012). *Extração de Conhecimento de Dados - Data Mining*. Edições Sílabo, LDA, Lisboa, 1st edition.

Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., and Cibrián-Uhalte, E. e. a. (2016). The chembl database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3/4):325.

Guggino, W. B. and Stanton, B. A. (2006). New insights into cystic fibrosis: molecular switches that regulate cftr. *Nature Reviews Molecular Cell Biology*, 7(6):426–436.

Hacker, M., Messer, W., and Bachmann, K. (2009). *Pharmacology*. Elsevier, Amsterdam.

Heller, S. R., McNaught, A., Pletnev, I., Stein, S., and Tchekhovskoi, D. (2015). Inchi, the iupac international chemical identifier. *Journal of Cheminformatics*, 7(1).

Hill, V. (1910). The possible effects of the aggregation of the molecules of haemoglobin on its dissociation curves. *Journal of Physiology*, 40:Proceedings iv–vii.

Kamentsky, L., Jones, T. R., Fraser, A., Bray, M.-A., Logan, D. J., Madden, K. L., Ljosa, V., Rueden, C., Eliceiri, K. W., and Carpenter, A. E. (2011). Improved structure, function and compatibility for cellprofiler: modular high-throughput image analysis software. *Bioinformatics*, 27(8):1179–1180.

Katzung, B. G. (2018). *Basic & Clinical Pharmacology, 14e*. McGraw-Hill Education LLC., New York, NY.

Kausar, S. and Falcao, A. O. (2018). An automated framework for qsar model building. *Journal of Cheminformatics*, 10(1).

Kausar, S. and Falcao, A. O. (2019). Analysis and comparison of vector space and metric space representations in qsar modeling. *Molecules*, 24(9):1698.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., and Yu, B. e. a. (2018). Pubchem 2019 update: improved access to chemical data. *Nucleic Acids Research*, 47(D1):D1102–D1109.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software, Articles*, 28(5):1–26.

Kuhn, M. and Johnson, K. (2013). *Applied predictive modeling*. Springer, New York, NY, 1st edition.

Landrum, G. (2006). *RDKit: Open-source cheminformatics*. Retrieved from http://www.rdkit.org.

Lipinski, C. A. (2004). Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies*, 1(4):337–341.

Lobo, M. J., Amaral, M. D., Zaccolo, M., and Farinha, C. M. (2016). Epac1 activation by camp stabilizes cftr at the membrane by promoting its interaction with nherf1. *Journal of Cell Science*, 129(13):2599–2612.

Lommatzsch, S. T. and Taylor-Cousar, J. L. (2019). The combination of tezacaftor and ivacaftor in the treatment of patients with cystic fibrosis: clinical evidence and future prospects in cystic fibrosis therapy. *Therapeutic Advances in Respiratory Disease*, 13:175346661984442.

Matthes, E., Goepp, J., Carlile, G. W., Luo, Y., Dejgaard, K., Billet, A., Robert, R., Thomas, D. Y., and Hanrahan, J. W. (2016). Low free drug concentration prevents inhibition of f508del cftr functional expression by the potentiator vx-770 (ivacaftor). *British Journal of Pharmacology*, 173(3):459–470.

Mogayzel, P. J. and Flume, P. A. (2010). Update in cystic fibrosis 2009. *American Journal of Respiratory and Critical Care Medicine*, 181(6):539–544.

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*, 5(2):107–113.

Nantasenamat, C., Isarankura-Na-Ayudhya, C., and Prachayasittikul, V. (2010). Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery*, 5(7):633–654.

O'Toole, M. (2003). *Encyclopedia & dictionary of medicine, nursing & allied health*. Saunders, Philadelphia, Pa.

Quinton, P. M. (1983). Chloride impermeability in cystic fibrosis. *Nature*, 301(5899):421–422.

Rai, J. and Kaushik, K. (2018). Reduction of animal sacrifice in biomedical science & research through alternative design of animal experiments. *Saudi Pharmaceutical Journal*, 26(6):896–902.

Riordan, Rommens, J., Kerem, B., Alon, N., Rozmahel, R., Grzelczak, Z., Zielenski, J., Lok, S., Plavsic, N., and Chou, J. e. a. (1989). Identification of the cystic fibrosis gene: cloning and characterization of complementary dna. *Science*, 245(4922):1066–1073.

Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.

Shockley, K. R. (2015). Quantitative high-throughput screening data analysis: challenges and recent advances. *Drug Discovery Today*, 20(3):296–300.

Shockley, K. R. (2016). Estimating potency in high-throughput screening experiments by maximizing the rate of change in weighted shannon entropy. *Scientific Reports*, 6(1).

Sterling, T. and Irwin, J. J. (2015). Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337.

Takenaka, T. (2008). Classical vs reverse pharmacology in drug discovery. *BJU International*, 88:7–10.

Taylor-Cousar, J. L., Mall, M. A., Ramsey, B. W., McKone, E. F., Tullis, E., Marigowda, G., McKee, C. M., Waltz, D., Moskowitz, S. M., and Savage, J. e. a. (2019). Clinical development of triple-combination cftr modulators for cystic fibrosis patients with one or two f508del alleles. *ERJ Open Research*, 5(2):00082–2019.

Teixeira, A. L. and Falcao, A. O. (2014). Structural similarity based kriging for quantitative structure activity and property relationship modeling. *Journal of Chemical Information and Modeling*, 54(7):1833–1849.

Teixeira, A. L., Leal, J. P., and Falcao, A. O. (2013). Random forests for feature selection in qspr models - an application for predicting standard enthalpy of formation of hydrocarbons. *Journal of Cheminformatics*, 5(1).

US CF Foundation, Johns Hopkins University, The Hospital for Sick Children (2019). *The Clinical and Functional Translation of CFTR (CFTR2)*. Available at https://www.cftr2.org/.

Wainwright, C., Elborn, J., Ramsey, B., Marigowda, G., Huang, X., Cipolli, M., Colombo, C., Davies, J., De Boek, K., and Flume, P. e. a. (2016). Lumacaftor/ivacaftor combination for cystic fibrosis patients homozygous for phe508del-cftr. *Drugs of Today*, 54(4):229.

Wallace, T. L., Ballard, T. M., Pouzet, B., Riedel, W. J., and Wettstein, J. G. (2011). Drug targets for cognitive enhancement in neuropsychiatric disorders. *Pharmacology Biochemistry and Behavior*, 99(2):130–145.

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):31–36.

Weininger, D., Weininger, A., and Weininger, J. (1989). Smiles. 2. algorithm for generation of unique smiles notation. *Journal of Chemical Information and Modeling*, 29(2):97–101.

Welsh, M. J. and Smith, A. E. (1993). Molecular mechanisms of cftr chloride channel dysfunction in cystic fibrosis. *Cell*, 73(7):1251–1254.

Zuur, A., Ieno, E., and Smith, G. (2007). Principal coordinate analysis and non-metric multidimensional scaling. In *Analysing Ecological Data*. Springer New York.

# Supplementary Figures

A

Stock (10mM) and intermediate concentrations (100µM) for replicates 1 and 2

| | 1 (10mM stock, 100µL) | 2 (int 1, 100µM, 200µL) | 3 (int 2, 100µM, 200µL) | 4 (10mM stock, 100µL) | 5 (int 1, 100µM, 200µL) | 6 (int 2, 100µM, 200µL) | 7 (10mM stock, 100µL) | 8 (int 1, 100µM, 200µL) | 9 (int 2, 100µM, 200µL) | 10 (10mM stock, 100µL) | 11 (int 1, 100µM, 200µL) | 12 (int 2, 100µM, 200µL) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C01 | C01 | C01 | C09 | C09 | C09 | C15 | C15 | C15 | C21 | C21 | C21 |
| B | C02 | C02 | C02 | C10 | C10 | C10 | C16 | C16 | C16 | C22 | C22 | C22 |
| C | C03 | C03 | C03 | C11 | C11 | C11 | C17 | C17 | C17 | C23 | C23 | C23 |
| D | C04 | C04 | C04 | Empty | Empty | Empty | Empty | Empty | Empty | C24 | C24 | C24 |
| E | C05 | C05 | C05 | Empty | Empty | Empty | Empty | Empty | Empty | C25 | C25 | C25 |
| F | C06 | C06 | C06 | C12 | C12 | C12 | C18 | C18 | C18 | C26 | C26 | C26 |
| G | C07 | C07 | C07 | C13 | C13 | C13 | C19 | C19 | C19 | C27 | C27 | C27 |
| H | C08 | C08 | C08 | C14 | C14 | C14 | C20 | C20 | C20 | C28 | C28 | C28 |

Max 300µL per well

B

intermediate concentrations (100µM) for replicates 3,4 & 5

| | 1 (int 3, 100µM, 200µL) | 2 (int 4, 100µM, 200µL) | 3 (int 5, 100µM, 200µL) | 4 (int 3, 100µM, 200µL) | 5 (int 4, 100µM, 200µL) | 6 (int 5, 100µM, 200µL) | 7 (int 3, 100µM, 200µL) | 8 (int 4, 100µM, 200µL) | 9 (int 5, 100µM, 200µL) | 10 (int 3, 100µM, 200µL) | 11 (int 4, 100µM, 200µL) | 12 (int 5, 100µM, 200µL) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C01 | C01 | C01 | C09 | C09 | C09 | C15 | C15 | C15 | C21 | C21 | C21 |
| B | C02 | C02 | C02 | C10 | C10 | C10 | C16 | C16 | C16 | C22 | C22 | C22 |
| C | C03 | C03 | C03 | C11 | C11 | C11 | C17 | C17 | C17 | C23 | C23 | C23 |
| D | C04 | C04 | C04 | Empty | Empty | Empty | Empty | Empty | Empty | C24 | C24 | C24 |
| E | C05 | C05 | C05 | Empty | Empty | Empty | Empty | Empty | Empty | C25 | C25 | C25 |
| F | C06 | C06 | C06 | C12 | C12 | C12 | C18 | C18 | C18 | C26 | C26 | C26 |
| G | C07 | C07 | C07 | C13 | C13 | C13 | C19 | C19 | C19 | C27 | C27 | C27 |
| H | C08 | C08 | C08 | C14 | C14 | C14 | C20 | C20 | C20 | C28 | C28 | C28 |

C

Layout Dilutions for each replicate

| | 1 (10µM) | 2 (1µM) | 3 (0,1µM) | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | C01 | C01 | C01 | C09 | C09 | C09 | C15 | C15 | C15 | C21 | C21 | C21 |
| B | C02 | C02 | C02 | C10 | C10 | C10 | C16 | C16 | C16 | C22 | C22 | C22 |
| C | C03 | C03 | C03 | C11 | C11 | C11 | C17 | C17 | C17 | C23 | C23 | C23 |
| D | C04 | C04 | C04 | VX770 | VX770 | VX661 | VX661 | VX809 | VX809 | C24 | C24 | C24 |
| E | C05 | C05 | C05 | DMSO | DMSO | DMSO | DMSO | DMSO | DMSO | C25 | C25 | C25 |
| F | C06 | C06 | C06 | C12 | C12 | C12 | C18 | C18 | C18 | C26 | C26 | C26 |
| G | C07 | C07 | C07 | C13 | C13 | C13 | C19 | C19 | C19 | C27 | C27 | C27 |
| H | C08 | C08 | C08 | C14 | C14 | C14 | C20 | C20 | C20 | C28 | C28 | C28 |

Control DMSO concentration = 0.1 %= biggest DMSO concentration for all compounds

[VX661]=5µM

[VX770/809]=3µM

Figure S1: **Layout of Preparation of Stock Solutions, Intermediate Dilutions and of Each Treatment on the 96-well Plates of the Immunofluorescence assay for Recovery of F508del-CFTR.**
(A) Preparation of stock solutions and intermediate dilutions. (B) Intermediate dilutions for replicates 3, 4 and 5. (C) Layout of disposition of treatments on the 96-well microscopy plate for the immunofluorescence assay.

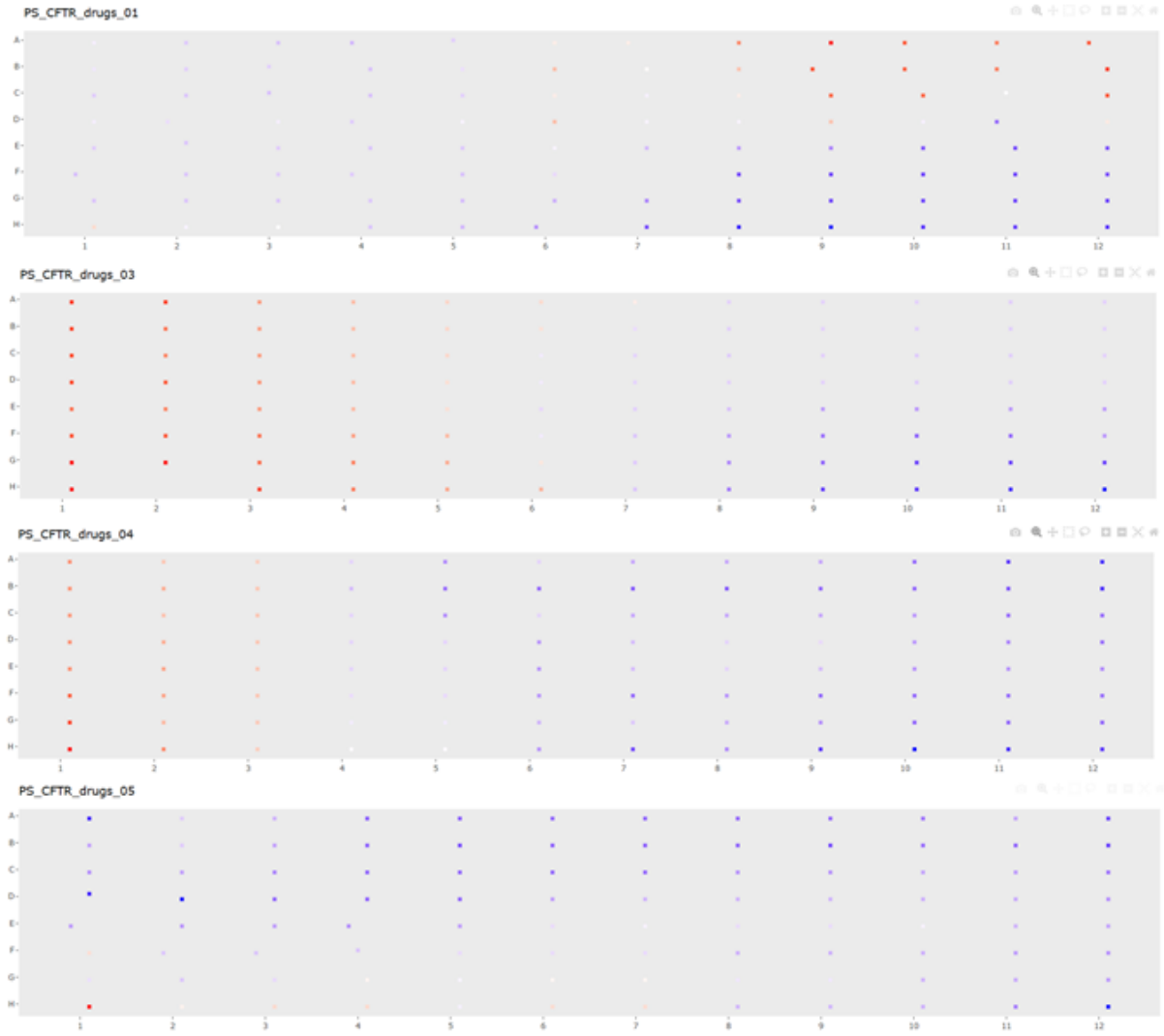



Figure S2: **Fluorescence Gradient Obtained from Median 5x5 Normalization.**

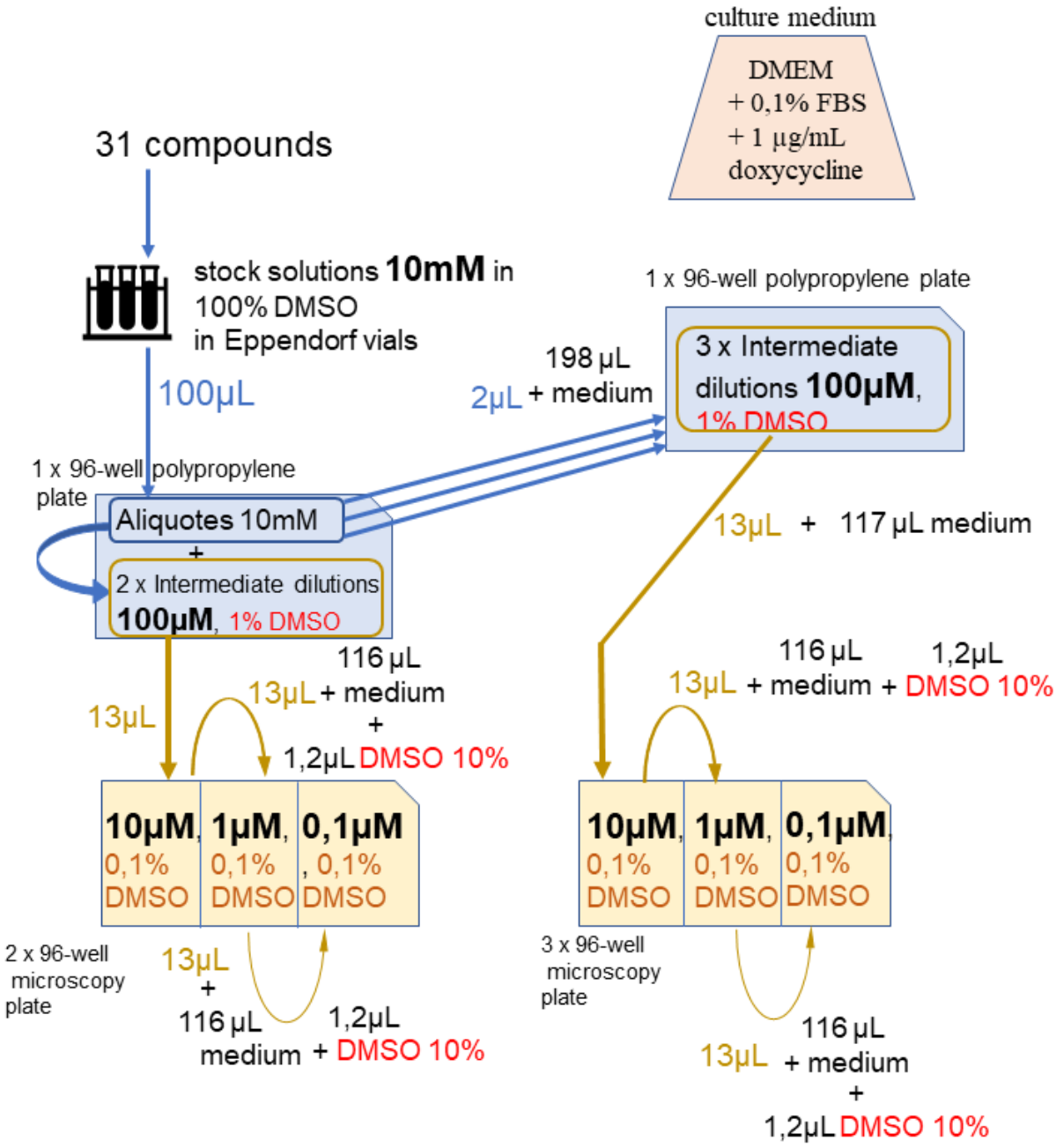



Figure S3: **Detailed Overview of Experimental Design of Preparation of Stock Solutions, Dilutions and Media composition for Induction of CFTR Expression in the Presence of Drug Candidate Compounds.**


Figure S4: **Presence of color in Preparations of Stock and Intermediated Solutions of Screening Compounds.**
Stock solutions prepared in DMSO and intermediate dilutions in DMEM medium. Many of the compounds presented color.

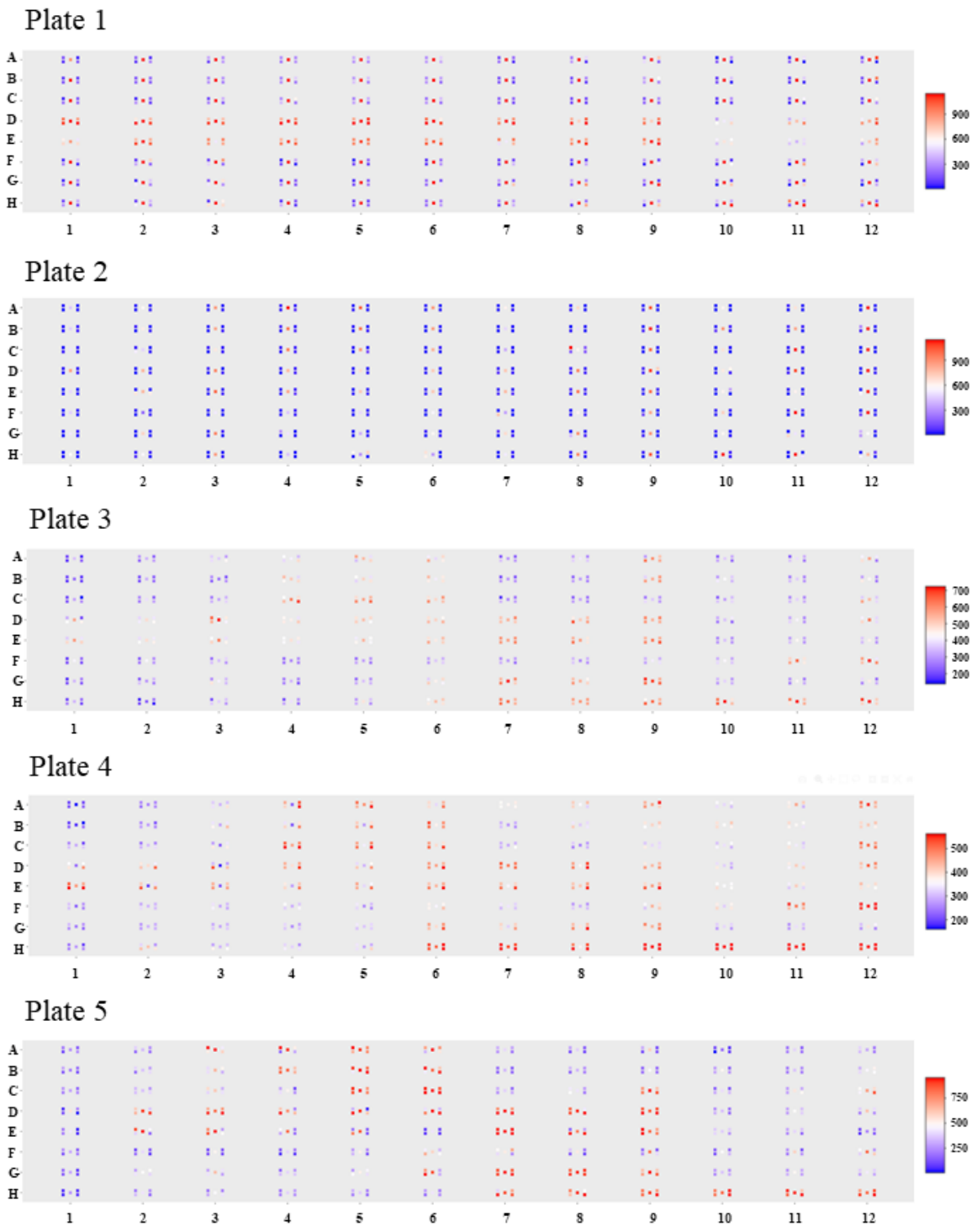



Figure S5: **Total Cell Count by Well in 96-well Plates.**
Total cell count is represented by well for each replicate. The scale of the heatmap is not the same for each plate. It was adapted in order to see how the cell number varies according to plate disposition for each plate.

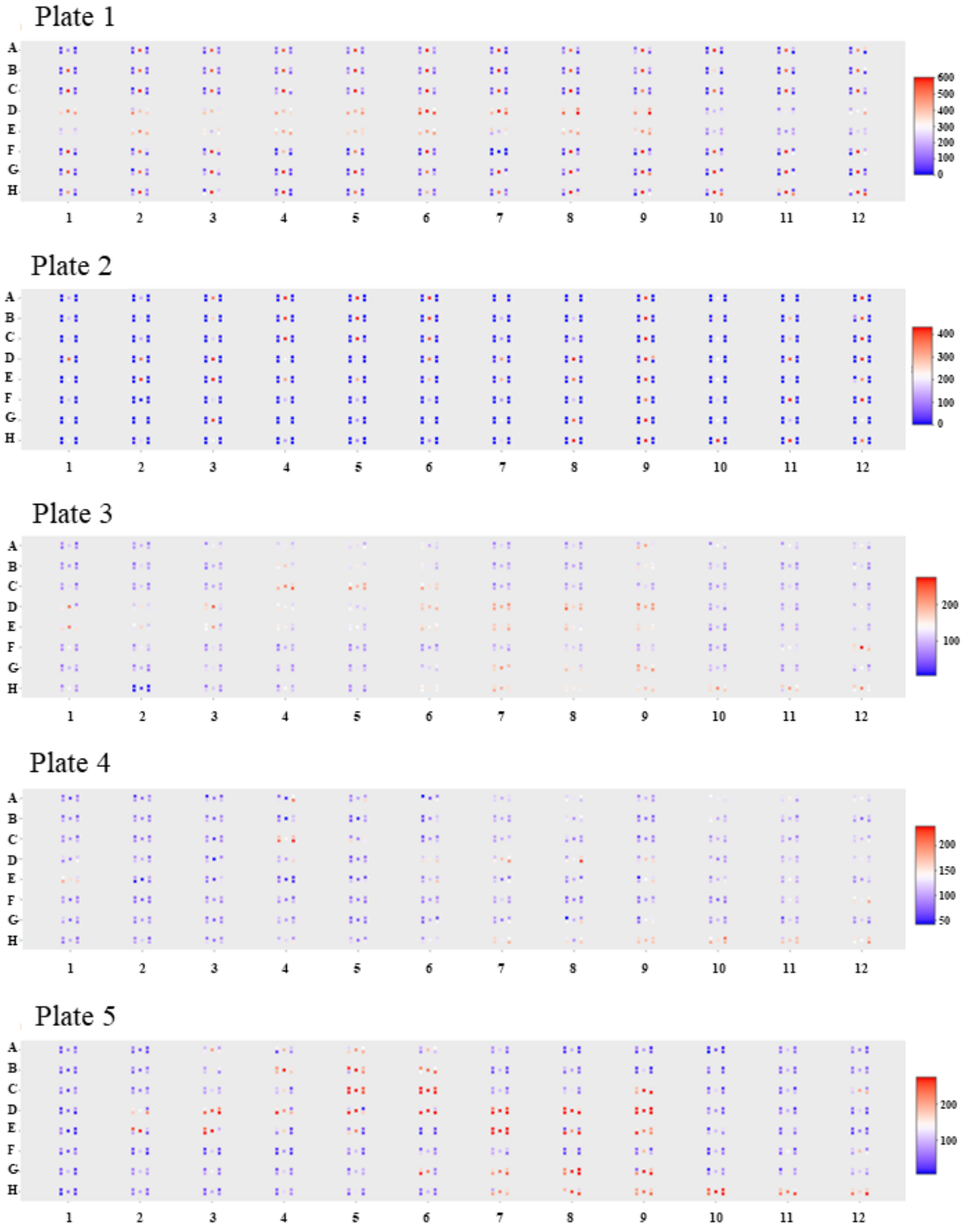



Figure S6: **Final Cell Count by Well in 96-well Plates.**
Final cell count is represented by well for each replicate, after CellProfiler and ShinyHTM quality control filters. The scale of the heatmap is not the same for each plate. It was adapted in order to see how the cell number varies according to plate disposition for each plate.

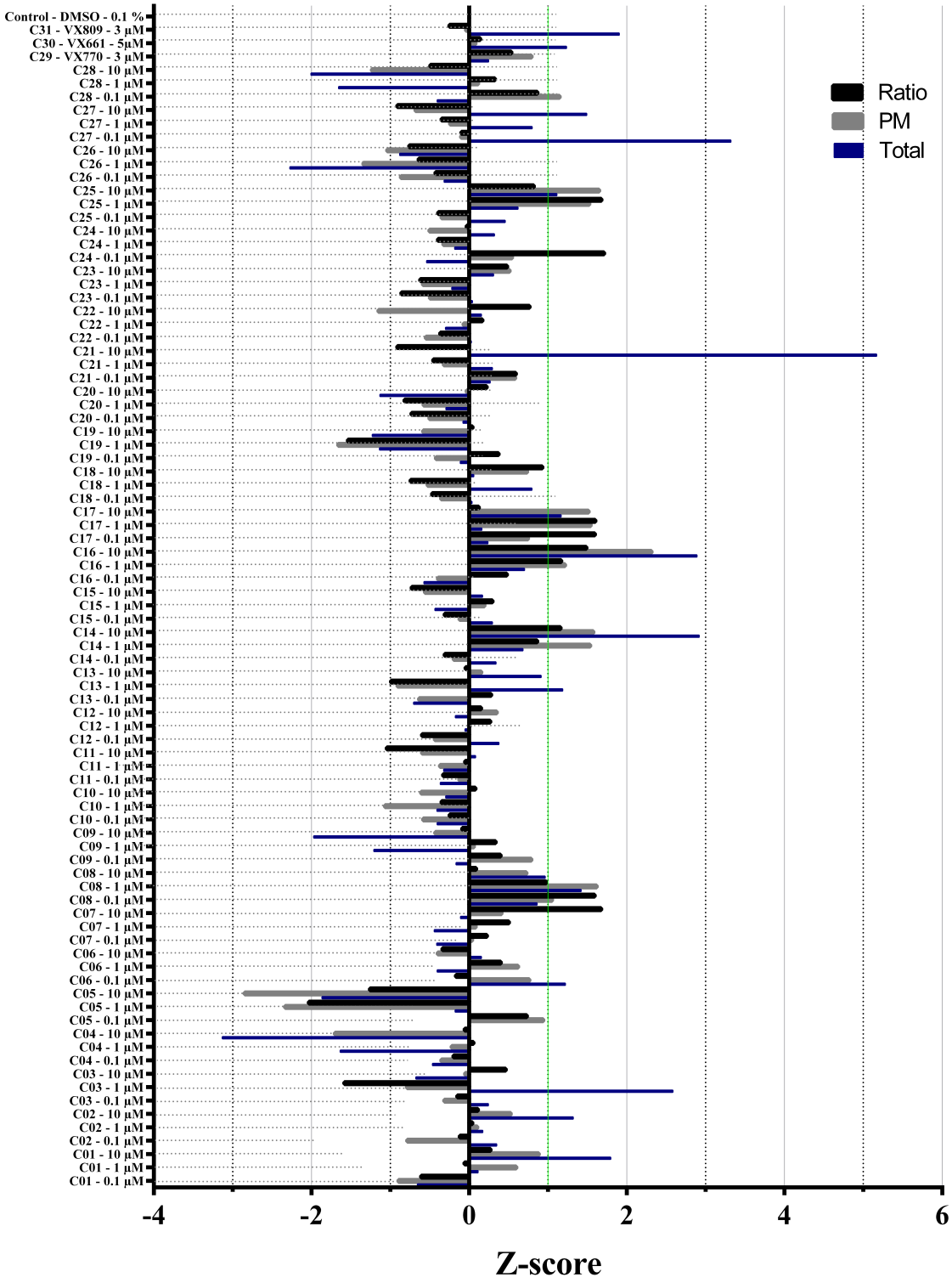



Figure S7: **Extensive Z-Scores of Ratio of Fluorescence Between PM and Total CFTR, of Total and of PM Fluorescence in F508del-CFTR Immunofluorescence Assay.**
Median of Z-scores of fluorescence ratio between Plasma Membrane and Total Fluorescence (traffic efficiency), of Total CFTR fluorescence and of PM CFTR fluorescence by treatment. Z-score of 1 is marked as a green line, being the threshold above which, compounds with median Z-scores were considered promising. DMSO only treatment was used as control. Ordered alphabetical by treatment.

# Supplementary Tables

Table S1: **First Part of Summary of Results of the Ratio of Fluorescence Between PM and Total Fluorescence of the Assay for Recovery of F508del-CFTR.**
Treatment is the compounds and respective concentrations; Batches/Plates is the replicates for each treatment that were not excluded by quality control; R 1-4 is the Z-score value for each replicate; SD is the standard deviation for Z-score values; Median is the Median Z-score value for the treatments (used as main score); T test with P value adjusted is the statistical test to assess statistical significance of the Z-scores against DMSO only treatment as control. Sig. is significance of T-test,* $p < 0.05$, ** $p < 0.01$.

| | | Z-score | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Treatment | Batches/Plates | R 1 | R 2 | R 3 | R 4 | SD | Median | T test with p value adjusted | Sig. |
| C01 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.82 | 0.06 | -1.06 | -0.17 | 0.86 | -0.61 | 0.708 | |
| C01 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.33 | 1.75 | 0.20 | -0.34 | 0.98 | -0.06 | 0.852 | |
| C01 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.36 | 0.29 | 0.25 | 3.81 | 1.90 | 0.27 | 0.791 | |
| C02 - 0.1 µM | Plate 03;Plate 04;Plate 05 | 0.65 | -0.12 | -3.97 | | 2.48 | -0.12 | 0.809 | |
| C02 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.30 | 0.24 | -0.16 | -2.82 | 1.49 | 0.04 | 0.809 | |
| C02 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 2.30 | 0.64 | -0.42 | -2.07 | 1.84 | 0.11 | 0.963 | |
| C03 - 0.1 µM | Plate 03;Plate 04;Plate 05 | 0.82 | -0.15 | -4.81 | | 3.01 | -0.15 | 0.809 | |
| C03 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -2.18 | -0.08 | -0.98 | -2.44 | 1.10 | -1.58 | 0.453 | |
| C03 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.41 | 0.75 | 0.18 | 1.94 | 1.00 | 0.46 | 0.791 | |
| C04 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.22 | -0.18 | -0.75 | 0.96 | 0.72 | -0.20 | 0.963 | |
| C04 - 1 µM | Plate 03;Plate 04;Plate 05 | 0.88 | 0.05 | -0.90 | | 0.89 | 0.05 | 0.997 | |
| C04 - 10 µM | Plate 01;Plate 03;Plate 04 | 1.26 | -0.06 | -0.25 | | 0.82 | -0.06 | 0.852 | |
| C05 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.59 | 0.88 | 1.65 | -1.07 | 1.15 | 0.73 | 0.809 | |
| C05 - 1 µM | Plate 03;Plate 04;Plate 05 | 0.94 | -2.03 | -6.87 | | 3.94 | -2.03 | 0.791 | |
| C05 - 10 µM | Plate 01;Plate 03;Plate 04 | -1.41 | -1.08 | -1.26 | | 0.17 | -1.26 | 0.004 | ** |
| C06 - 0.1 µM | Plate 01;Plate 03;Plate 04 | 0.67 | -0.84 | -0.17 | | 0.75 | -0.17 | 0.959 | |
| C06 - 1 µM | Plate 01;Plate 03;Plate 04 | 1.39 | 0.40 | -0.14 | | 0.78 | 0.40 | 0.791 | |
| C06 - 10 µM | Plate 03;Plate 04;Plate 05 | 1.64 | -0.34 | -2.06 | | 1.85 | -0.34 | 0.959 | |
| C07 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 1.01 | 0.72 | -0.27 | -5.08 | 2.83 | 0.23 | 0.852 | |
| C07 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.43 | 1.37 | -0.14 | 1.15 | 0.90 | 0.51 | 0.791 | |
| C07 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 4.03 | 0.74 | -0.08 | 2.61 | 1.86 | 1.67 | 0.506 | |
| C08 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.36 | 1.99 | 1.20 | 3.25 | 1.51 | 1.59 | 0.506 | |
| C08 - 1 µM | Plate 01;Plate 04;Plate 05 | 0.98 | -2.10 | 4.00 | | 3.05 | 0.98 | 0.873 | |
| C08 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.41 | 3.16 | 0.59 | -0.79 | 1.78 | 0.09 | 0.841 | |
| C09 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.30 | 1.14 | 0.50 | -3.75 | 2.23 | 0.40 | 0.916 | |
| C09 - 1 µM | Plate 03;Plate 04;Plate 05 | 0.65 | 0.34 | -1.44 | | 1.12 | 0.34 | 0.959 | |
| C09 - 10 µM | Plate 03;Plate 04;Plate 05 | -0.07 | -0.09 | -0.13 | | 0.03 | -0.09 | 0.916 | |
| C10 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.41 | -0.24 | -0.25 | 0.38 | 0.35 | -0.25 | 0.883 | |
| C10 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.28 | 0.10 | -0.42 | -0.66 | 0.32 | -0.35 | 0.791 | |
| C10 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.13 | 0.70 | 0.01 | 0.15 | 0.77 | 0.08 | 0.963 | |
| C11 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.19 | 0.69 | -0.47 | -0.57 | 0.57 | -0.33 | 0.916 | |
| C11 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.06 | -0.27 | -0.16 | 1.13 | 0.64 | -0.05 | 0.883 | |
| C11 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.54 | -1.19 | -1.01 | -1.08 | 0.28 | -1.05 | 0.028 | * |
| C12 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.13 | 0.83 | -0.06 | -2.85 | 1.58 | -0.60 | 0.791 | |
| C12 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.35 | 0.62 | -0.09 | 2.99 | 1.52 | 0.27 | 0.791 | |
| C12 - 10 µM | Plate 03;Plate 04 | 0.98 | -0.68 | | | 1.18 | 0.15 | 0.963 | |
| C13 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.71 | 0.13 | 0.43 | 4.22 | 2.19 | 0.28 | 0.809 | |
| C13 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.85 | 1.34 | -0.14 | -1.86 | 1.54 | -0.99 | 0.809 | |
| C13 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.15 | 1.57 | -0.24 | -2.18 | 1.55 | -0.04 | 0.959 | |
| C14 - 0.1 µM | Plate 03;Plate 04;Plate 05 | 2.31 | -0.31 | -3.50 | | 2.91 | -0.31 | 0.943 | |
| C14 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.26 | 1.99 | -0.35 | 7.48 | 3.67 | 0.86 | 0.791 | |
| C14 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 1.88 | 2.15 | 0.44 | -0.07 | 1.08 | 1.16 | 0.506 | |
| C15 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.32 | 0.69 | -0.22 | -0.40 | 0.82 | -0.31 | 0.841 | |
| C15 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.58 | 0.86 | 0.01 | -5.36 | 2.94 | 0.30 | 0.852 | |
| C15 - 10 µM | Plate 03;Plate 04;Plate 05 | 0.94 | -0.73 | -1.66 | | 1.32 | -0.73 | 0.852 | |

Table S2: **Second Part of Summary of Results of the Ratio of Fluorescence Between PM and Total Fluorescence of the Assay for Recovery of F508del-CFTR.**
Treatment is the compounds and respective concentrations; Batches/Plates is the replicates for each treatment that were not excluded by quality control; R 1-4 is the mean Z-score value for each replicate; SD is the standard deviation for Z-score values; Median is the Median Z-score value for the treatments; T test with P value adjusted is the statistical test to assess statistical significance of the Z-scores against DMSO only treatment as control.

| | | Z-score | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Treatment | Batches | R 1 | R 2 | R 3 | R 4 | SD | Median | T test with p value adjusted |
| C16 - 0.1 µM | Plate 03;Plate 04;Plate 05 | 0.68 | -0.16 | 0.48 | | 0.44 | 0.48 | 0.791 |
| C16 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 11.46 | 1.31 | 1.03 | -0.65 | 5.52 | 1.17 | 0.791 |
| C16 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 2.43 | 0.27 | 0.55 | 4.64 | 2.02 | 1.49 | 0.506 |
| C17 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.29 | 2.20 | 0.98 | 2.80 | 1.37 | 1.59 | 0.506 |
| C17 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.24 | 2.25 | 0.95 | 3.07 | 1.27 | 1.60 | 0.453 |
| C17 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 2.04 | 0.65 | -0.40 | -2.86 | 2.07 | 0.12 | 0.963 |
| C18 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.86 | 0.71 | -0.08 | -4.21 | 2.17 | -0.47 | 0.791 |
| C18 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.33 | -0.25 | -0.39 | -1.10 | 0.53 | -0.74 | 0.453 |
| C18 - 10 µM | Plate 03;Plate 04;Plate 05 | 3.18 | 0.19 | 0.93 | | 1.56 | 0.93 | 0.716 |
| C19 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.21 | 0.54 | -0.38 | 0.54 | 0.43 | 0.37 | 0.809 |
| C19 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -2.84 | 1.57 | -2.35 | -0.72 | 1.99 | -1.54 | 0.791 |
| C19 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.34 | -0.26 | 1.46 | -1.69 | 1.31 | 0.04 | 0.988 |
| C20 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.33 | -1.13 | -0.15 | -1.66 | 0.70 | -0.73 | 0.506 |
| C20 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -2.81 | 0.47 | -0.35 | -1.28 | 1.41 | -0.82 | 0.786 |
| C20 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.94 | 0.78 | -0.34 | 1.94 | 1.27 | 0.22 | 0.873 |
| C21 - 0.1 µM | Plate 03;Plate 04;Plate 05 | 0.59 | -1.43 | 5.09 | | 3.34 | 0.59 | 0.833 |
| C21 - 1 µM | Plate 03;Plate 04;Plate 05 | 1.97 | -0.46 | -2.97 | | 2.47 | -0.45 | 0.943 |
| C21 - 10 µM | Plate 03;Plate 04;Plate 05 | 0.30 | -2.21 | -0.91 | | 1.25 | -0.91 | 0.791 |
| C22 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.59 | 1.23 | -0.15 | -1.71 | 1.21 | -0.37 | 0.891 |
| C22 - 1 µM | Plate 03;Plate 04;Plate 05 | 0.83 | 0.17 | -0.78 | | 0.81 | 0.17 | 0.963 |
| C22 - 10 µM | Plate 03;Plate 04;Plate 05 | 0.77 | -0.41 | 3.71 | | 2.12 | 0.77 | 0.791 |
| C23 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.63 | 1.04 | -0.18 | -1.54 | 1.27 | -0.86 | 0.809 |
| C23 - 1 µM | Plate 03;Plate 04;Plate 05 | -0.62 | 0.39 | -0.73 | | 0.62 | -0.62 | 0.821 |
| C23 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.30 | 0.59 | 0.38 | 3.51 | 1.55 | 0.48 | 0.723 |
| C24 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 5.02 | 0.69 | 1.17 | 2.26 | 1.94 | 1.71 | 0.453 |
| C24 - 1 µM | Plate 03;Plate 04;Plate 05 | 1.81 | -0.94 | -0.39 | | 1.46 | -0.39 | 0.963 |
| C24 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.00 | 1.86 | 0.36 | -0.42 | 1.24 | -0.03 | 0.943 |
| C25 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.83 | 1.22 | 0.04 | -1.23 | 1.09 | -0.39 | 0.943 |
| C25 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.61 | 2.56 | 0.94 | 2.42 | 1.48 | 1.68 | 0.573 |
| C25 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 3.00 | 0.90 | -0.63 | 0.74 | 1.50 | 0.82 | 0.791 |
| C26 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.31 | -3.94 | -0.55 | -0.21 | 1.80 | -0.43 | 0.786 |
| C26 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.08 | -0.64 | -0.64 | -2.40 | 1.01 | -0.64 | 0.573 |
| C26 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -2.38 | 0.86 | 1.13 | -4.85 | 2.85 | -0.76 | 0.809 |
| C27 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.44 | 0.59 | -0.15 | -0.05 | 0.43 | -0.10 | 0.988 |
| C27 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.27 | -0.21 | -0.43 | -0.58 | 0.17 | -0.35 | 0.573 |
| C27 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -1.67 | 0.61 | -0.16 | -2.21 | 1.31 | -0.91 | 0.791 |
| C28 - 0.1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 2.43 | -0.02 | 1.76 | -2.73 | 2.31 | 0.87 | 0.943 |
| C28 - 1 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 0.58 | 0.55 | 0.11 | -0.52 | 0.51 | 0.33 | 0.873 |
| C28 - 10 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -2.24 | 0.01 | -0.16 | -0.83 | 1.02 | -0.49 | 0.723 |
| C29 - VX770 - 3 µM | Plate 01;Plate 03;Plate 04;Plate 05 | 3.33 | 0.77 | 0.30 | -2.44 | 2.36 | 0.53 | 0.969 |
| C30 - VX661 - 5µM | Plate 01;Plate 03;Plate 04;Plate 05 | 1.87 | 0.51 | -0.23 | -1.10 | 1.26 | 0.14 | 0.916 |
| C31 - VX809 - 3 µM | Plate 01;Plate 03;Plate 04;Plate 05 | -0.28 | 0.50 | -0.22 | -1.25 | 0.72 | -0.25 | 0.809 |
| Control - DMSO - 0.1% | Plate 01;Plate 03;Plate 04;Plate 05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 1.000 |

# Annex A

## Western Blot Protocol

Detailed protocol made in the context of studies using CFTR and F508Del-CFTR in CFBE cell lines, with BioISI's resources.

Paulo Sousa, 2019

## Making the SDS-PAGE Gels:

Each gel contains 2 mini-gels; one for concentration and one for resolving (see Recipes for detailed description).

- Start by setting up the apparatus. You will need the main support, a rubber band for sealing, a glass holder and two glass panels (for 1,5mm gels), a bigger and a smaller one.
- Place the glasses on the holder, bigger one facing the support and smaller one facing the user, with both glasses firmly against table surface to guarantee stability.
- Place the holder+glasses on top of rubber bands on the support and confirm the lack of leakage by pipetting water inside the glass panels. Remove the water by pouring upside down, and soak remaining water with paper.

Prepare the resolving gel first -> 7% acrylamide gel.

- prepare the recipe according to its order, top-down, in an erlenmeyer flask.
- mix gently.
- Pour gently against the side of the glass until the first mark on the plastic of the holder.
- Pour approximately 1 mL of isopropanol on top of the gel, to ensure that it polymerizes evenly.
- Wait 30-60 min until polymerization.

Preparing the 4% gel:

- pour the isopropanol out of the gel.
- prepare the recipe according to its order, top-down, in an erlenmeyer flask.
- mix gently.
- place on top of the 7% gel until the border of the glass.
- place “colm” on the gel with the letters facing you, carefully, avoid making bubbles (it’s expected to leak some gel from the top).
- wait until polymerization.
- after polymerization, wrap in moistened paper and then in aluminum foil, refrigerate until use.

*Preparing the samples*

## Extraction:

All the following steps are made on ice and with cold reagents.

Prepare 1 mL of Sample Buffer (SB) from 2x Stock (in the freezer, see detailed recipe in Recipes):

1. add 500 µL to the already present 500 µL
2. add 40 µL of protease inhibitor (pic)
3. add 2 or 3 µL of benzonase (in the ladder's box)
4. for each µL of benzonase, add 6.25 mL of $MgCl_2$ 0.5M (in the fridge) (18.75 µL)

To collect samples (6-well plate), start by washing 2x with ice-cold PBS, then aspirate the wells and add ~150 µL of SB to each well, one plate at a time so the wells don't dry. You can either use a P1000 pipette, scratching the bottom of the well to lysate the cells and collect by pipetting up & down a few times. If sample is too viscous, add another 50 µL of SB. Alternatively you can use a cell scraper to lysate the cells and collect the sample with a P1000 pipette.

Keep samples on ice until use. If you don't want to use them on the same day store at -20ºC, or -80ºC for long term storage.

## Quantification:

Quantify approximately 30 µg of protein and the correspondent volumes from each sample using the Bradford Assay.

Start by preparing the tubes with the contents of the calibration curve, using BSA 1,4 µg/mL (bovine serum albumin), Bradford Reagent and $H_2O$ as described on the right, and a sample of desired protein of unknown concentration from your assay.

Wait for 5 min.

Use the same cuvette for measuring the Abs595nm of all samples, start with the less concentrated ones. If there are outlier points, discard them before assessing your proteins' concentration.

| Diluting sample | | |
|---|---|---|
| BSA (µl) | H2O (µl) | Bradford (µl) |
| 0 | 800 | 200 |
| 1 | 799 | 200 |
| 2 | 798 | 200 |
| 3 | 797 | 200 |
| 6 | 794 | 200 |
| 10 | 790 | 200 |
| 15 | 785 | 200 |
| 20 | 780 | 200 |
| Quantification of unknown sample (below) | | |
| 10 | 790 | 200 |

All samples should have same volume in the end, calculate how much loading buffer and water/PBS each sample needs to even them out. Note that each sample will receive ¼ of the total volume in loading buffer.

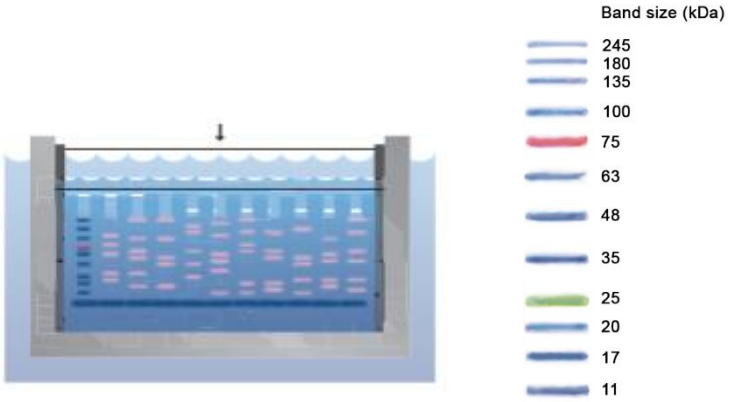
Note that the wells typically only hold about 50 µL of volume, so if the samples are not concentrated enough you might have to use less total amount of protein.

## Loading & Running:

Loading Buffer is a 10:90 ratio of β-mercaptoethanol to Laemmli 4x stock solution (see detailed in Recipes).
Prepare Loading Buffer for all samples in one tube in enough volume for 1:4 (V/V) with combined sample volumes (approximately the volume of desired protein / 3 or final volume/4 - different methodologies), count for a few extra volume to account for error.

Start by preparing all the tubes and add the water to the tubes.

Add loading buffer to all tubes 1:4 (V/V) and then the samples.




Assemble the gel in the WB tank and fill it with running buffer (RB) 1X. Gels are placed in the holder with the larger glass facing the outside.
Fill space between gels with RB until it overflows, blow bubbles away, fill the rest of the tank with RB until holder is submerged.
Take note of the number and position of each gel if you are using more than one.

(optional) Place “training wheels” adaptor on top of the gel, to facilitate the loading on the correct wells without leakage.

Load 8 µL of ladder NZYColour Protein Marker II with a P20 pipette with a P100 tip.

For all the remaining samples, use a P100 pipette, and before loading, mix the Loading sample and then apply against the glass of the gel apparatus, feeling the resistance of the tip against both glass surfaces.

Load all the samples on an SDS-PAGE gel. After loading remove “training wheels” if it was used.

Make sure the gels are submersed on RB.

After loading the samples run the gel at 60-75V in order to concentrate the samples and then change it to 100-120V to separate the proteins according to their molecular weight.

- for CFTR, run until the 35 KDa marker
- Calnexin has approximately 90 KDa
- CTRF can produce 2 bands, B-band with approximately 110 KDa (DeltaF, misfolded) and C-band (properly glycosylated and folded) has approximately 135 KDa

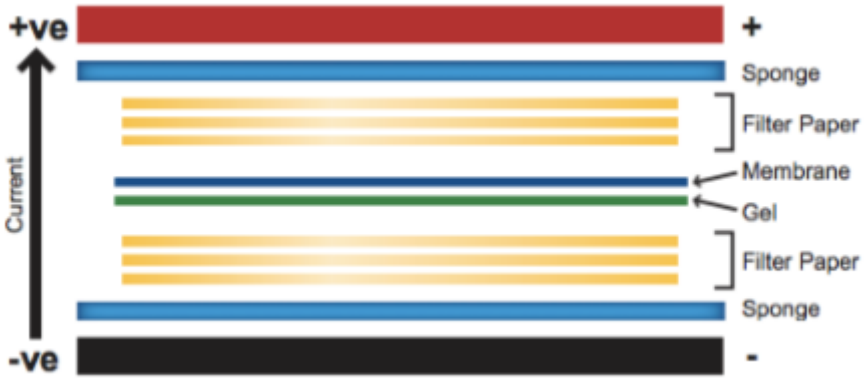
## Transfer:

Always use gloves and tweezers when contacting the membrane.

Start preparing the transfer by cutting a PVDF membrane with the proper dimensions (according to the gel, for a 1,5mm gel it is advised to cut 6x9cm).
Using a pencil, mark on the upper and bottom corner of the membrane the number of the gel on the side of the ladder, which is also the side that will contact the gel.

Prepare a recipient large enough to place the transfer cassette by putting enough transfer buffer (TB) to submerge it, at least partially.

Remove the gel from the glasses with the appropriate tool and cut and remove the concentration part of the gel. Take care to keep the gel hydrated with TB throughout the procedure so it




doesn't tear.

Activate the membrane by submerging it in methanol for a few seconds.

Assemble the sandwich:

- 2 sponges
- 4 papers
- 1 membrane
- 1 gel

Make sure the gel is placed against the membrane according to the marks previously made.
If you place membrane side underneath (red), you can see if there are bubbles between the gel and the membrane.

Use the appropriate roll to ensure there are no bubbles between the membrane and the gel.

Transfer at 400 mA (constant current) for 1h30 on ice.

## Antibody Incubation:

Incubate membrane (blocking) in 5% PBS-T Non-fat-milk (NFM) for about 30 min.

Cut each membrane in appropriate position to separate control from target protein

- for CFTR assays with calnexin as control, cut right above the 100 KDa marker

Incubate the membrane with primary antibody in 5% PBS-T Non-fat-milk with agitation, overnight at 4ºC or for 2h at RT.

- in assays with CFTR, use 1:3000 for both primary and secondary antibodies.

(note: some antibodies require TBST instead of PBS-T)

Next day remove the primary antibody and wash the membrane 3x (every 10-15 min) on PBS-T with agitation. NOTE: the primary antibody can be re-used (2-4 times) if kept at -20ºC.

Incubate the membrane with secondary antibody on 5% PBS-T Non-fat-milk with agitation for 1h at RT.

- Prepare about 9 mL (3µL of secondary anti-mouse-a.b.)

Remove the secondary antibody and wash the membrane 3x (every 15 min) with PBS-T with agitation.

To store the membranes, keep refrigerated on PBS-T.

## Developing:

To develop the membrane use ChemiDOC software.

Turn on equipment and PC, and put filter on “no-filter” posittion.

Place a plastic sheet on top of the visualization glass.

Mix the BioRad reagents 1:1 in a tube (don’t forget which one you put first).

On ChemiDOC, choose gel imaging, go to Protocol -> Blots -> chemiHiResolution -> signal accumulation mode
Choose the the time of exposure of first and last image, and the number of images to take.

Put some reagent on top of the part of the membrane to analyse and remove excess reagent and spread evenly by turning membrane against the plastic sheet.

Position the membrane adequately and acquire a colorimetric image.

## Quantification:

To quantify your western blots, use ChemiDOC software.

## Safety

Everything contaminated with acrylamide and β-mercaptoethanol should go into the red waste basket.
Throw immediately away immediately after use everything contaminated with β-mercaptoethanol.

## Recipes

### Laemmli Buffer 4x (for 10mL):

- 62.5mM trisHCl pH 6.8 -> 625 µL 1M stock (fridge)
- 10% glycerol -> 1 mL 100% stock
- 2% SDS -> 2 mL 10% stock
- bromophenol blue (to taste, just for color) -> ~200 µL 1% stock (shelf)
- $H_2O$ to 10 mL

### Sample Buffer 2x (blue):

- 1.25 mL stacking buffer
- 3 mL 10 % SDS
- 1 mL 100% glycerol
- 0.2 mL 1% Bromophenol
- 4.55 mL dH2O
- 154 mg DTT

**Western Blot SDS-PAGE mini gels:**
(all volumes are for 2 gels)

| | 4% | 7% |
|---|---|---|
| Distilled $H_2O$ | 7.4 mL | 9.75 mL |
| Separating Buffer 1.5M pH 8.8 | - | 4.5 mL |
| Stacking Buffer 0.5M pH 6.8 | 3 mL | - |
| Acrylamide 40% | 1.2 mL | 3.15 mL |
| Glycerol 10% | 120 µL | 180 µL |
| SDS 10% | 120 µL | 180 µL |
| PSA (or APS) 10% | 90 µL | 135 µL |
| TEMED 100% | 20 µL | 20 µL |

**5% (m/v) Non-fat-milk (NFM):**

For 100 mL in small flask

- weight 5 g of non-fat-milk powder and add to flask
- add PBS-T to the 100 mL mark

**PBS-T:**

For 1 L

- 100 mL PBS 10x
- 900 mL $H_2O$
- 1 mL Tween20 (use a P1000 and cut off the tip before pipetting)