Εθνικό Μετσόβιο Πολυτεχνείο

Εργαστήριο Βιοϊατρικών Συστημάτων

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ ΚΑΙ ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

Διπλωματική Εργασία

Platform for compound prioritization based on transcription factor activity

Φοιτητής: Αθανασόπουλος Κωνσταντίνος

Επιβλέπων Καθηγητής: Αλεξόπουλος Λεωνίδας

Αναπληρωτής Καθηγητής ΕΜΠ

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2020

Εθνικό Μετσόβιο Πολυτεχνείο

Εργαστήριο Βιοϊατρικών Συστημάτων

ΤΟΜΕΑΣ ΜΗΧΑΝΟΛΟΓΙΚΩΝ ΚΑΤΑΣΚΕΥΩΝ ΚΑΙ ΑΥΤΟΜΑΤΟΥ ΕΛΕΓΧΟΥ

Διπλωματική Εργασία

Platform for compound prioritization based on transcription factor activity

Φοιτητής: Αθανασόπουλος Κωνσταντίνος

Επιβλέπων Καθηγητής: Αλεξόπουλος Λεωνίδας

Αναπληρωτής Καθηγητής ΕΜΠ

Εγκρίθηκε την 16η Οκτωβρίου 2020 από την τριμελή επιτροπή:

Αλεξόπουλος Λεωνίδας	Προβατίδης Χριστόφορος	Κυριακόπουλος Κωνσταντίνος
Αναπληρωτής Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ	Καθηγητής ΕΜΠ

ΑΘΗΝΑ, ΟΚΤΩΒΡΙΟΣ 2020

Abstract

Prioritizing chemical structures based on their exhibition of desired biological behaviors could be of great significance in the early stages of drug discovery. In this study, a platform was developed that screens unknown compounds and selects chemical structures that are most likely to display sought after biological effects, based on their transcription factor activity. To make this possible, the compound differences were translated into compound distances on biological and structural levels and assigned a numerical value accordingly. The model incorporates several criteria to evaluate the known and unknown chemical structures and allocate the ones that approximate the required biological effect. Most importantly, the platform makes use of a deep learning model to predict the biological distances of the unknown compounds and map their effect. The proposed model was able to select unknown compounds that portrayed desired biological activity with high precision and accuracy.

Περίληψη

Στη πρώτα στάδια ανακάλυψης φαρμάκων, όταν και οι επιστήμονες έρχονται αντιμέτωποι με την επιλογή δραστικών ουσιών, η εναπόθεση προτεραιότητας σε ορισμένες χημικές ενώσεις, που θα έχουν την απαραίτητη βιολογική συμπεριφορά, θα μπορούσε να είναι καίρια. Στην παρούσα μελέτη, αναπτύχθηκε ένα μοντέλο το οποίο αξιολογεί την βιολογική συμπεριφορά άγνωστων χημικών ενώσεων και επιλέγει ενώσεις που πιθανότατα θα εμφανίζουν τα επιθυμητά βιολογικά χαρακτηριστικά, με γνώμονα τη δράση μεταγραφικών παραγόντων. Για την μοντελοποίηση του προβλήματος, οι διαφορές μεταξύ ενώσεων μαθηματικοποιήθηκαν ως αποστάσεις σε λειτουργικό επίπεδο και αναπαραστάθηκαν από μία τιμή αντίστοιχα. Στο μοντέλο έχει ενταχθεί πλήθος κριτηρίων αξιολόγησης των ουσιών με σκοπό τον ακριβέστερο εντοπισμό κατάλληλων χημικών ενώσεων. Ταυτόχρονα γίνεται χρήση ενός μοντέλου βαθιάς εκμάθησης για τον προσδιορισμό των βιολογικών αποστάσεων αγνώστων ουσιών από την επιθυμητή συμπεριφορά. Συνολικά, το μοντέλο αποδείχθηκε ικανό να αναγνωρίζει και να θέτει προτεραιότητα σε χημικές ενώσεις, οι οποίες προσεγγίζουν την απαιτούμενη βιολογική δράση, με μεγάλη ακρίβεια.

Acknowledgements

The project described in this thesis was the result of the contribution, support, and effort of many people.

First of all, I would like to express my gratitude towards the supervisor of the project, Associate Professor Leonidas Alexopoulos. He captivated my interest in class and his opinions and lessons motivated me throughout my studies. In the later years of my studies, he gave me the chance to work in a great research environment, with great people, while always watching over my progress and caring about my goals and hopes.

From the first day that I decided to work on this project, the man who lavishly offered his help and guidance was Chris Fotis. Chris, remarkably and single handedly manages every student and every project at the Biosystems Lab, all while being a PhD student. He was the first person I turned to when I faced problems, he guided me throughout my project and was by my side with compassion and helpfulness all along. For all the help he offered me, I feel the need to thank him wholeheartedly.

Moreover, I would like to thank my collaborators in this project, Nikos and Christina for their valuable contribution to the project. I would also like to thank all my friends at the lab, for welcoming me with open arms and creating a great work environment.

Finally, I feel I need to thank my family and my friends for being by my side throughout my studies, caring for me, offering me countless stimuli, pushing me to go further, and helping me become a better person and a better student.

Contents

1. Introduction

1.1. Current Drug Discovery Process and Possibilities

The current state of the world finds our species most resilient than ever. Since the dawn of ages, humankind has struggled to survive. And that survival has never been easy. Water and food shortage, predators, wars, and many other obstacles stood in the way of humankind's survival and prosperity. Evolution has been kind to our species though, and century by century we have come to outlast our predators, cure our weaknesses, and prove -mostly to ourselves- that humans can beat all obstacles. The proof for the human ascendance is all around us; over the last century we have quadrupled our population on the planet, while our life expectancy has steadily been on the rise. The one enemy that has proven impossible to exterminate, and is quite relevant today, is diseases.

Diseases are conditions that interfere with an organism's homeostasis and affect its structure or function. We cannot "see" them, but rather spot them out by the symptoms of the affected organism. Diseases can be caused by external (e.g. pathogens) or internal (e.g. immune system disorders) factors. Humankind has come to identify death by disease, as "death by natural cause." This statement of causality between disease and "natural death" suggests that the natural barrier that must be overcome for humankind to reach a longer and better life -even immortality- is diseases. And the 20th century has paved the way with great victories over this everlasting enemy. With the boom of the drug industry and the vaccination against the worst diseases, life expectancy nearly doubled in most parts of the world.

In the field of pharmacology, drugs are chemical substances which, when administered to an organism, produce a biological effect. Drugs for humans are used to cure or prevent diseases, alleviate pain, and set the table for longer and better lives [1]. The most fundamental process that allows continuous progress of drugs is drug discovery.

Drug discovery is the process that aims for the identification of new candidate medications that could be valuable for the cure and combat the symptoms of a specific disease. Throughout history, humankind has been using traditional remedies like plants and powders retrieved from nature. Though enterprising, this method lacked medical basis and was mostly ineffective. Thus, it was replaced by the identification of the active ingredients and the effort to connect these ingredients and specific diseases. Over the past century, the drug discovery process has evolved, and we can now reap its benefits. Early stage drug discovery involves the connection of the right compound to the right target (disease). The sequencing of the human genome has also allowed researchers to study the proteins of the human body and their role in every process [3]. This led to the construction of large compound libraries, most significant for the widely used method of High Throughput Screening (HTS). High Throughput Screening makes use of technology (data processing, robotics, chemical tests, ea.) to conduct millions of experiments at the same time and identify hits, compounds with strong binding affinity to the target. Hits from these in-vitro experiments are prioritized for further in-vivo testing to analyze their effects and discover their efficacy. It is easy to conclude that the millions of tests conducted in the HTS combine for a huge cost, while the results are not guaranteed [2].

The accumulating possibilities and computational strength of computers has forwarded the rise of a new method of drug discovery, called Computer Aided Drug Design (CADD), that uses computational systems to perform in-silico tests. Such methods tend to focus on the structural and biological effects (e.g. similarities and distances) of tested compounds against the target. This method is highly successful and can perform much faster and effortlessly, conducting billions of otherwise time-consuming experiments in the blink of an eye. The results are then used for further testing and optimization.

1.2. Computer Aided Drug Design

Computer Aided Drug Design (CADD) was incepted in 1981 and since then it has continuously been growing, following the fast-paced growth of technology, computer hardware and software. Lately, CADD has become a trustworthy tool to forward the selection, development and design of novel compounds that could potentially evolve into disease battling medication. As mentioned earlier, CADD was developed as an advancement to HTS, because it is cost-effective, timesaving and it requires little prior knowledge of the compounds. Typically, CADD is based on screening large compound libraries (e.g. their binding affinity, structural or biological effect distances to the tested function or characteristic) and can return multiple hit compounds. These compounds are amongst the most promising candidates and are to be selected for further testing and development. The hits are usually categorized into smaller clusters of predicted compounds (Figure 1). Those clusters are afterwards studied closer, and the lead compounds are optimized by improving their biological properties or combining whole compounds or parts of them to reach a desired function [4].

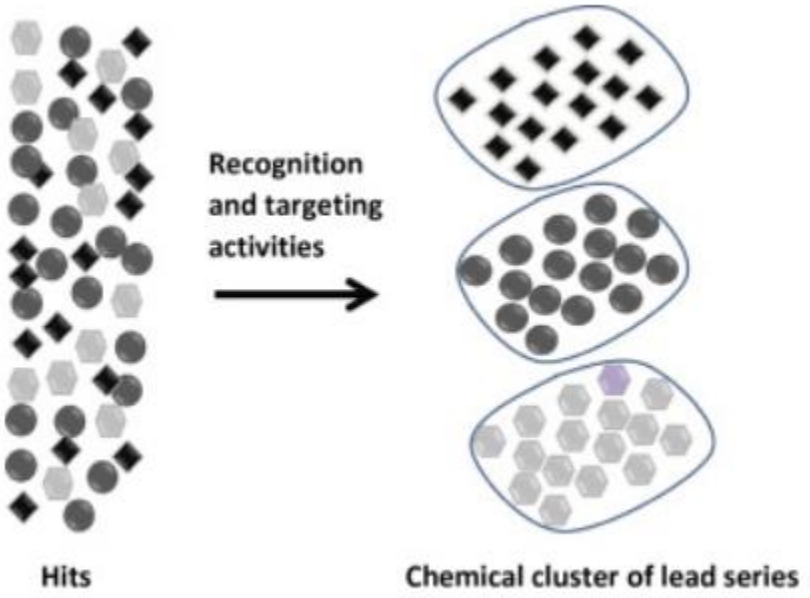

Figure 1: Hits being clustered to recognise their biological and structural parts [4]

The two main categories of CADD approaches are the structure-based and the ligand-based.

Structure-based CADD requests the finding of the target protein's structure to determine its interaction levels with the compounds found in the libraries. X-ray crystallography or NMR spectroscopy are usually used to attain the 3D structure of the target protein and simulate each protein-compound structure as a complex. Ligand-based CADD involves the chemical similarity of the target and the compounds. It is largely based on predictive models of quantitative structure–activity relationship (QSAR) models that are created in order to judge the activity strength (or inactivity) of the compounds. This type of models (QSAR) connect the structural elements of compounds to their biological activity and can be further used in the construction of compounds with optimal biological activity [4]. In the case of ligand-based virtual screening, the 3D structure of the target is not available, and the similarity is calculated between an active ligand and the compounds of a selected library [5]. The idea behind the ligand-based approach is that similarity in structures leads to similarity at the biological effect level. At this point, there needs to be a separation so that this hypothesis stands. Similar structures lead to similar biological effects, but the opposite does not stand. For example, two compounds may share completely different chemical structures, but have the same biological effect, due to the way they target the proteins or because of off target effects [6].

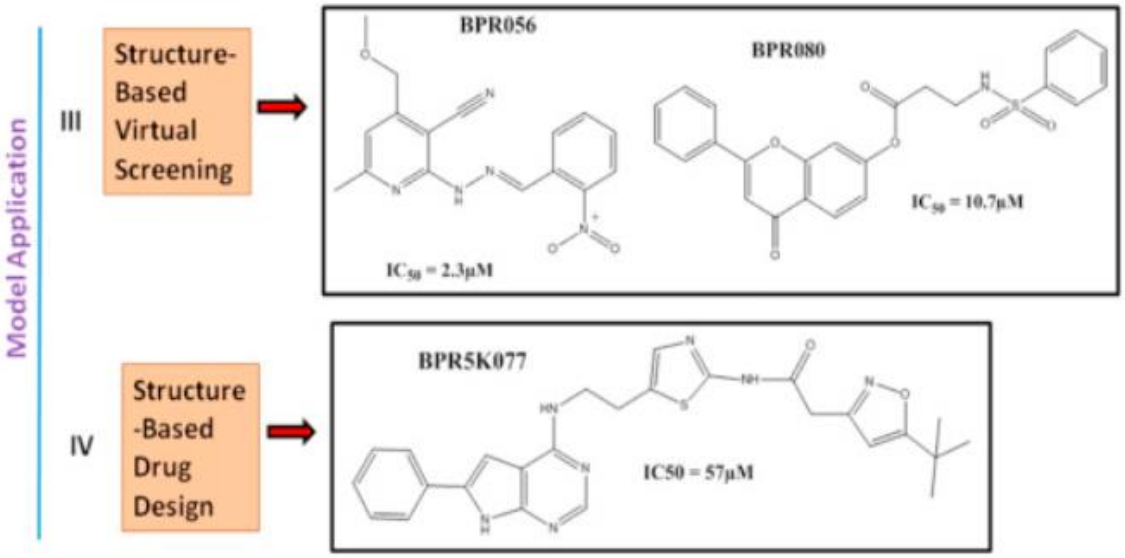

Figure 2: Combining two structures into one to optimize the desired effect [4]

1.3. Systems Biology

Systems biology is the scientific approach of complex biological systems through the means of mathematics and computational analysis. As systems biology evolves it has come to attract many more fields of study such as engineering, physics, and chemistry, all under the wide umbrella of biology. The result is a biology-focused interdisciplinary field with boundaries that are not quite defined. Such an approach examines every aspect and element of the complex biological systems under study, without

making mathematical admissions like in the traditional method of reductionism. The holism, that stems from the examination of every element of the system, contributes to the increase of both the complexity of the problem -thus the complexity of computations- and the accuracy of the results [7]. We mainly focus on the contribution of computer science to systems biology, that allows the analysis and use of large experimental data to simulate biological functions, understand complex systems and phenomena, and predict biological behaviors.

Systems biology offers new possibilities on the testing of simple or more complex hypotheses, with a new way of experimental validation and modelling. The quantitative description of systems and processes allows the control of interactions within the system and the formation of more dynamic models. A main field of interest for systems biology is the single cell and its biological behavior, with transcriptomics, proteomics and metabolomics playing a crucial role to the understanding and modelling the quantitative dynamics of the cell. This focus on the dynamic evolution within biological systems constitutes the main difference between bioinformatics and systems biology [8].

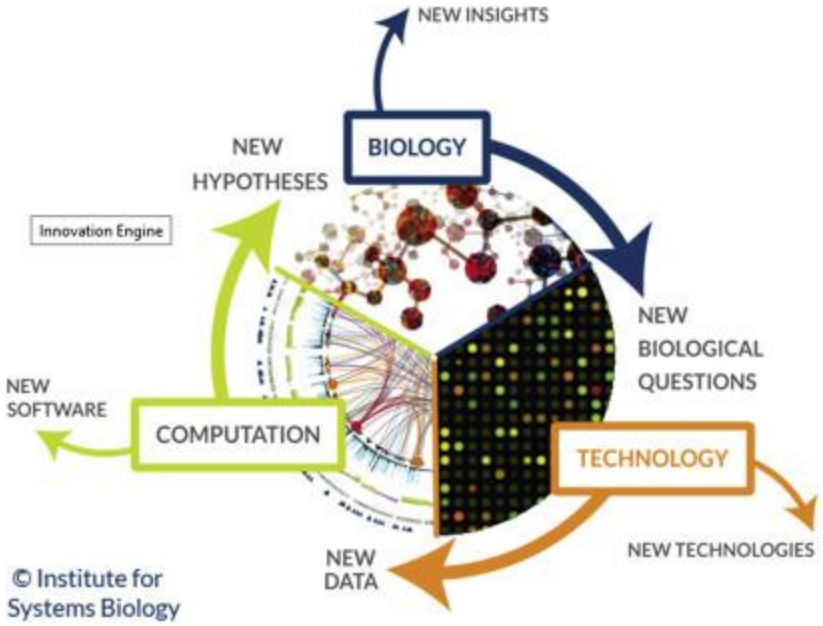

Figure 3: The Process of Systems Biology Analysis [9]

1.4. Gene Expression

Genes are sequences of nucleotides of DNA or RNA that include the genetic information to synthesize gene products (RNA or proteins). DNA and RNA serve as the mediators for the production of proteins, which in turn are the most significant mediators of the biological behavior of the cell and its functions. Most biological traits and cell functions are directly influenced by the genes along with the gene-environment interactions. That extends to other levels of biological organization, even as high as the organism level. For example, genes dictate the levels of production for a certain protein in a cell, but they also dictate the color of people's eyes, their blood type, sensitivity to certain substances and all their other traits [10].

Gene expression (GEx) is the process by which the encoded information hidden in a gene is converted into products with specific biologic functions. Of all these products, the most important for the regulation of a cell's functions are proteins. These proteins are the offspring of the transcription operation, as it is elaborately described in the Central Dogma of Biology (Figure 4) [11].

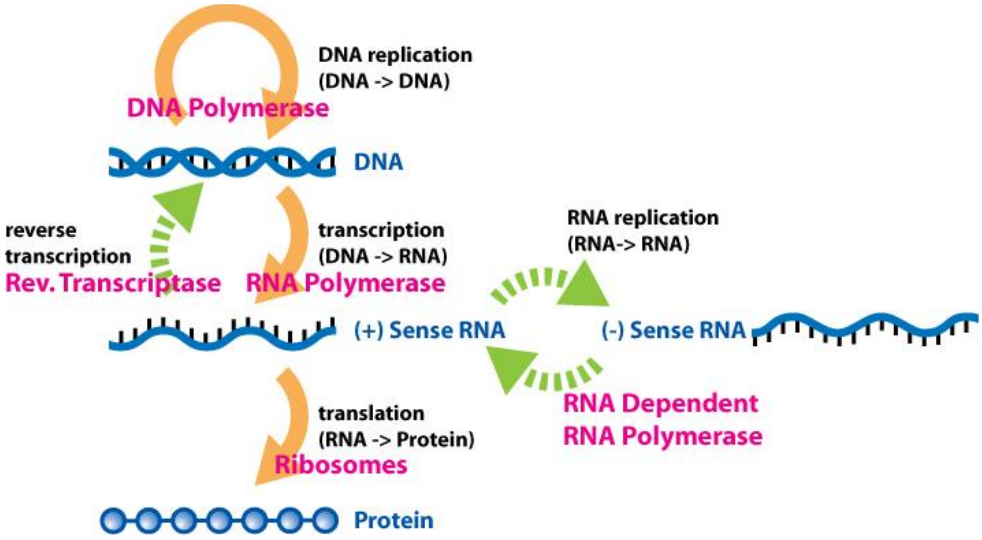

Figure 4: The Central Dogma of Biology [12]

Gene expression is the most important function of every living cell, since it connects the genotype (the genetic information within a cell) and the phenotype (every trait and function that can be observed when the cell is under study). The phenotype is often expressed by the synthesis of proteins. Such proteins control the structure, the growth, the development, and the total of functions of the cell. Therefore, it is only logical that anomalies in the gene expression of certain genes lead to differentiation of the cell's biological function.

In systems biology, gene expression data analysis focuses on the identification of genes that are over- or under- expressed on the phenotypic level or during an experiment. This search for anomalies of the expression level is called Differential Gene Expression Analysis. Such an analysis is the basic tool for in-silico and in-vitro experiments, targeting the detection of connections between genes and biological functions. The basis for the Differential Gene Expression Analysis is the fact that alterations in a cell's functions are either internally or externally caused. This causality of factors and biological perturbations, when studied, can create useful conclusions. For example, Differential Gene Expression after the administration of a drug on a cell could prove that the active ingredient of the drug creates perturbations on the expression of certain genes. Processes like the aforementioned have contributed to the integration of the human knowledge on topics like chemical intervention on cells through drugs or how the cell can regulate its functions when its biological processes are damaged.

1.5. Transcriptomics

Transcriptomics focuses on the transcriptome—the complete set of RNA transcripts produced by the genome, in a single cell or under specific circumstances—using high-throughput methods. After

comparing transcriptomes in distinct cell populations, one can identify the Differentially Expressed Genes in response to specific drugs or to the change of external factors (e.g. radiation, heat levels, etc.) [13].

The progress of high-throughput technologies and computer science has allowed the obtaining and tracking of the transcriptome and how it changes over time within cell populations. Data that is retrieved from the transcriptome, mainly through the techniques of DNA microarray and RNA-seq [13], has quite a large range of use. It can be used to gain knowledge on the structure and functions of the single cell, along with the problems it may face (cellular differentiation, transcription regulation, carcinogenesis, etc.). The transcriptome and its analysis have allowed a deeper understanding of the cell and its behavior and contributed greatly to the evolution of systems biology.

1.6. Transcription Factors & Transcriptomic Signatures

Two of the most significant elements that participate in the vital process of transcription are transcription factors and transcriptomic signatures. In molecular biology, transcription factors (TFs) are proteins that control the rate of transcription of genes from DNA to messenger RNA. That process is the result of the TFs (proteins) binding to specific parts of the DNA sequence, so that certain genes can be expressed. Consequently, transcription factors are the main regulators of the protein levels in a cell, as the over- and under- expression of these proteins is based on transcription factors turning on and off genes, affecting their over- or under- expression. Of course, transcription factors can trigger the transcription of proteins that can multiply or destroy the TF itself. As a result, groups of TFs act in a coordinated way, to regulate, promote or demote basic cellular functions, such as growth, movement, metabolism etc. Surprisingly enough, TFs are responsible for even the death of the cell. All in all, this regulation of protein production by the TFs is a basic weapon for the understanding of how the cell works. In human cells, about 1600 transcription factors can be found. These TFs are the point of focus in the medicinal field, because of the role they play in the cells, since TF mutations can cause regulatory problems and disrupt the homeostasis of the cell and consequently the organism resulting into diseases. The regulation of the TFs and the avoidance of those mutations offer a promising insight on the future of medicines [14][15].

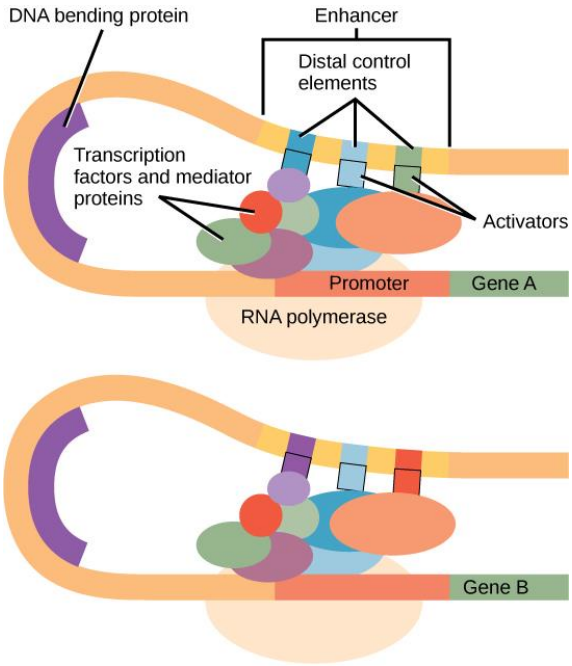

Figure 5: The function of the Transcription Factors during the transcription process [16]

On the other hand, transcriptomic signatures are perturbations in the process of transcription. Mutations in TFs, errors in the genome and small mistakes in the binding process of proteins to the genome can lead to such perturbations.

1.7. Systems Pharmacology

Pharmacology is the study of the action that certain medication induced to an organism. It is a significant branch of medical sciences since it provides vital information on drugs and their effect. Systems pharmacology uses the mechanisms and methods of systems biology to produce results in the field of pharmacology. Examining the human organism as a complex biological system and taking into account all its parameters, systems pharmacology can suggest new medication, reveal new actions of medications, and help understand the human-medication interactions better [17]. This broader understanding of the human organism and its complexity, drives this method away from the classic drug-protein interaction assumption, and creates new paths of perceiving the biological effects of drugs through a large base of interactions that they have with different factors that regulate the well-being of the human body. Therefore, systems pharmacology examines protein-protein, signaling, chemical or even genetic interactions on all levels of biological complexity – from the small atom system to the large organism.

The examination of such interactions has been continuously approached through gene expression methods and the perturbations that compounds have on the expression of genes. This examination is assisted by large-scale perturbation databases, like the Connectivity Map (CMap) and the Library of Integrated Network-based Cellular Signatures (LINCS). These databases make the processes of drug design and pharmacogenomics possible, by facilitating transcriptomics profiles, used on the pathway and networks levels through several systems biology approaches [18]. The basis of the CMap approach are the

transcriptomic signatures and their biological effects, under the assumption that similarity in transcriptomic signatures equals similarity in biological effect. Such a theory is short-sighted as it conscientiously lacks basic information concerning the biological structure of compounds, their binding affinity to the target and other factors that could counter its effect.

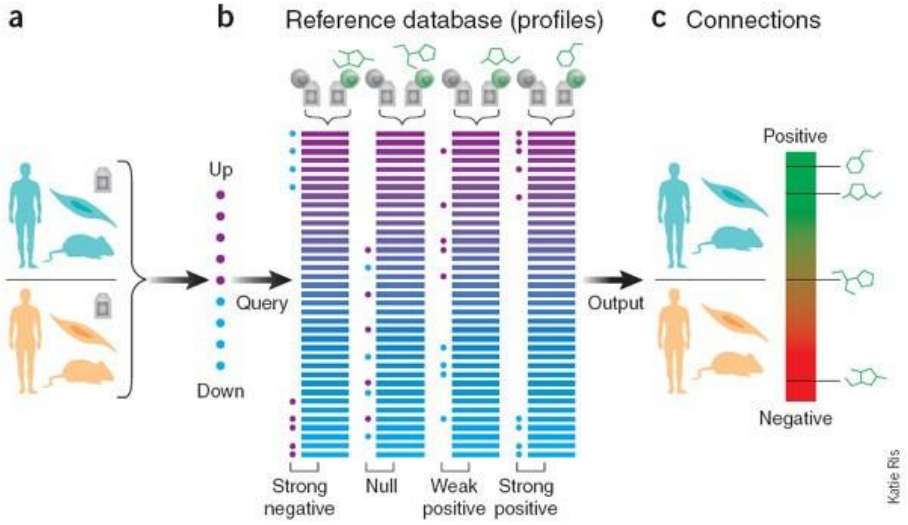

Figure 6: How the CMap method works [19]

On this matter, a helping hand has been lent to scientists by Artificial Intelligence (AI). More accurately, machine learning methods have managed to include the structure and binding affinity of compounds to the target as inputs, enhancing this way the predictive strength of models and enabling the drug discovery and drug repurposing branches of pharmacology.

1.8. Machine Learning in Drug Discovery

Machine learning is a broad scientific field that uses computer algorithms to gain knowledge from a set of data and optimize the predictions of new data. These algorithms try to adjust the weights of the proposed set of mathematical functions based on sample data (training set) and optimize them to acquire the best results over the training set. The whole purpose of machine learning is the implementation of these adjusted functions over new, raw data of the same kind to make the best possible predictions and help in any decision-making process. The most exciting machine learning branch is Deep Learning.

Deep Learning is a machine learning method based on artificial neural networks (ANNs); algorithms that assimilate the information processing and communication nodes in cells and biological systems. As shown in Figure 7, an artificial neural network can receive many inputs, process them, and determine the correct way (function) to connect these inputs into the desired output. This flexibility and ability to translate inputs of all different kinds into an output gives ANNs a respectable predicting power.

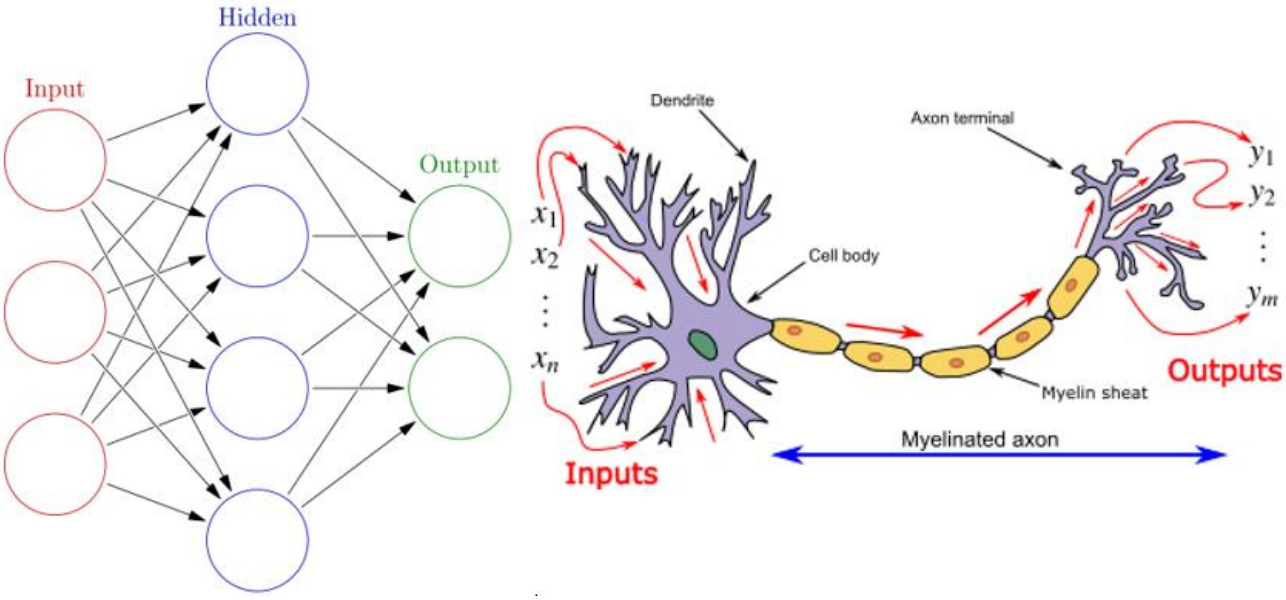

Figure 7: The parallelization of Artificial Neural Networks and a Human Neuron. Notice how both systems have the ability to receive a large number of input data and translate it into an output [20].

ANNs consist of layers that are densely connected. The layers contain several neurons that receive input from the former layer and, after adjusting the weights of the function sought after, give their output to the next layer. This process is dynamic, meaning that the true output regulates the weights of the neuron functions, so that the error of the true output and the prediction of the neural network is minimized. All in all, the importance of ANNs is vast, since they can handle big data, extract useful information, and help in predicting future interactions in any field [21].

Deep learning is a great weapon for the drug discovery process. Over the past decade, the rising abilities of computing systems, along with the multiplication of available data, has allowed that interaction to expand. Deep learning methods can utilize a big range of input of biological importance, encoded in the right way, and extract useful information or make predictions for the system under review. Deep learning libraries and architectures have provided scientists the ability to progress the drug discovery process, by predicting interactions between compounds and targets and selecting compounds for biological testing. Inputs such as chemical structures, biological effect, binding affinity, molecular interactions, and geometries et.al has helped both in the drug discovery and drug repurposing domains, minimizing the time and resources needed.

1.9. GO Terms and GO Term Enrichment

Gene Ontology (GO) is an effort to connect the representation of gene and gene product attributes. The GO project has its own vocabulary, with specific terms and ways to annotate data and genes. Moreover, GO provides the tools to understand the function of genes through experimental data, like enrichment analysis, through which the effects on compounds on cells are determined. The GO syntax is built using specific representations of either cellular components, molecular functions, or biological processes, called

GO-Terms. Each GO-Term within GO has a term name, which may be a word or string of words, and a unique alphanumeric identifier (e.g. id: GO:0000016, name: lactase activity). These GO-Terms are used instead of words, to indicate functional or structural characteristics of the cell and quantify them [22].

The GO-Term Enrichment Analysis is used to specify which GO terms are over-represented (or under-represented) in a gene set under certain conditions and is a great method to quantify the biological functions of a cell, along with the effects of compounds and make them more computationally trackable. Furthermore, such a controlled biological method can provide information on functional similarity or distance of gene products. For this thesis, the GO-Term Enrichment Analysis was performed using the FGSEA package in Bioconductor in R [23].

1.10. Gene and TF-Knockdowns

Gene knockdown is the process by which the expression of an organism's gene is reduced or even completely blocked. Such a perturbation to the gene expression can happen either through genetic modification of the organism, or through the treatment with a compound designed to have gene specific knockdown abilities [24]. As mentioned earlier, transcription factors (TFs) help with the quantification of the gene expression process as well as the function of the single cell through proteins, since they regulate the production of proteins. The conclusion that can be drawn, that directly affects the thought process of drug discovery, is that sustaining or changing the levels of the TFs in a single cell could be the way to keeping the cell healthy and well-functioning.

Since transcription factors are proteins that bind on specific parts of the genome, their production is either triggered or suspended in turn by the expression of respective genes. A basic term used in this diploma thesis is "TF-knockdowns", meaning compounds responsible for the reduction or suspension of the expression of the gene that will later produce the transcription factor we want to eliminate. Essentially, TF-knockdowns are compounds that suspend the production of a TF. Since the connection between protein production and transcription factors is direct, we focus on blocking TFs and not on blocking the effects of proteins. Therefore, TF-knockdowns are important in the efforts to monitor and control the cell's interior.

1.11. Current study

In this diploma thesis, a model is developed that aims to prioritize unknown compounds for further testing based on their transcription factor activity. Since researchers possess large libraries of compounds and compounds' interactions, when in need to create a new drug, they refer to these libraries to detect the compounds that could work as active ingredients in the drug under development. To replace the costly and time-consuming process of high throughput screening, that tests a large number of compounds in parallel, we created a model that can virtually perform screening of the compounds and propose a number of compounds that biologists are to run tests on first. It is most probable that the proposed compounds contain a desired chemical structure that would have the sought-after biological effects.

The model that was developed takes as input a transcription factor that needs to be suspended or knocked-down, along with a library of data in which to pursue the search, and returns a set of compounds that could perform this knockdown. This process essentially consists of four stages: knockdown availability, neighbor selection, inference selection and evaluation. In the first step, we indicate which transcription factor knockdowns are available to work with in the given data library. Secondly, we try to find "neighbors" of the knockdown in the known data, meaning compounds that are virtually close to the knockdown at the biological effect level. Later, we use a set of criteria and a deep learning model to screen unknown compounds and select the ones closest to the neighbors that came up in the former step. Finally, we conclude the model with an overall evaluation of the process, by proposing several compounds and judging their performance.

To justify the selection of the criteria used in this model, we present the results of the criteria selection and optimization, along with the weight of each criterion and the role it plays in the boosting of the model's performance. Finally, we prove the statistical significance of our model and confirm that our model should indeed be used to prioritize compounds for further testing.

2. Data

2.1. Data Preprocessing

The latest version of CMAP (GSE92742) provided data for transcriptomic signatures after compound application and the L1000 assay the 978 landmark genes of which the differential expression was examined. The L1000 assay is a high-throughput gene expression assay that measures the mRNA transcript quantity of 978 landmark genes in human cells. The name L1000 derives from L for landmark and 1000 for the approximate number of landmark genes. Proceeding with these signatures, a quality score was calculated for each one. Only the best quality signatures (quality 1) were considered.

Biological functions and their significance in perturbations are represented in whole gene sets. For this specific analysis, GO-Terms for biological processes of the landmark genes of the L1000 assay were used [25]. Bioconductor's package *TopGo* was used to recover the data [26]. Afterwards, GO-Term enrichment was applied for every signature through the FGSEA package of Bioconductor in R [23]. The outcome of the enrichment was a vector of Normalized Enrichment Scores (NES) at the GO-Term level (meaning representing to the biological function of the signature) for each signature.

Given that the proposed model is based on differences on the biological function level, that difference was to be quantified through some form. We chose to quantify these differences as *GO-Term level distances of pairs,* in accordance with Iorio et al. [27]. More information on the subject of distance is offered in the Methods chapter.

2.2. Training Sets and Test Sets

The structure of the model under construction strongly depends on the data that is being used. On that front, we split the available data into a training set and a test set for each cell line used. The training set essentially consists of pairs of drugs with their distance. We chose to operate on a GO-Term distance level and a transcription factor distance level. To create a test set, we excluded a number of drugs from the training set and moved all their pairs to the test set. In conclusion, the test set comprises of pairs of drugs with one of them being "cold," meaning a compound not included in the training set. The application of the model on the test set, provides useful information on the effectiveness and the robustness of the model when coming across totally unknown structures.

To create efficient test sets and training sets for the data, we took into consideration the distribution of the distances and the chemical structure of the compounds. Firstly, to check the behavior of the model to the data provided, we tried to make the test set as difficult for the model as possible, by excluding drugs that don't have chemical similarity to the compounds in the training set, so that the test set comprises of drugs with low structural similarity to the training set. Moreover, to test the model's performance in a just way, we prohibited the creation of totally random test sets, meaning we tried to have matching distributions of distances in the test set and the training set. An example of these similar distance distributions for the a375 cell line is available in Figure 8.

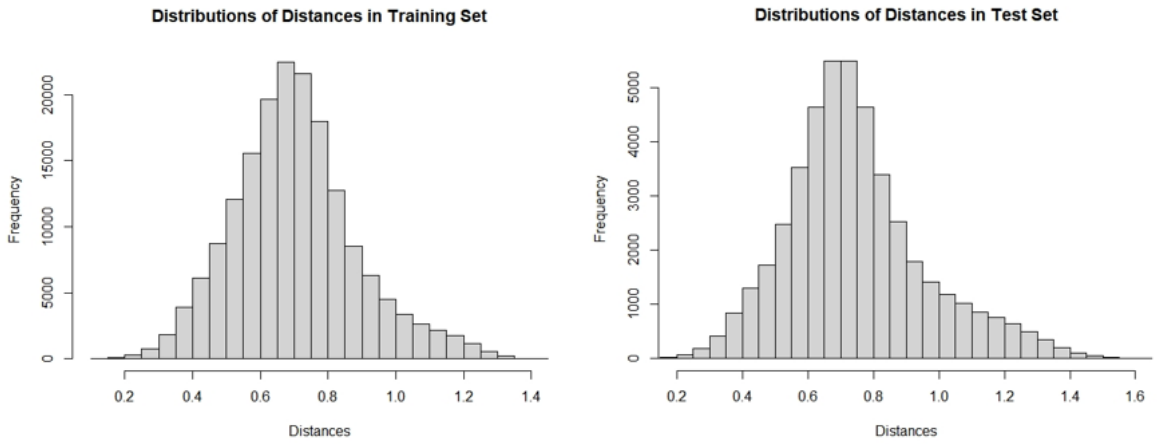

Figure 8: Similarity between the distributions of the test set and the training distances. In particular, these sets are representative of the distances in the a375 cell line. The training and test split was produced heavily on the premise that the these distributions should be as similar as possible.

Finally, the results of the splitting process produced the following results for the three cell lines that are of interest, shown in Tables 1&2. The most useful information on each split is the number of pairs and the number of unique compounds in each set. It is important to mention that the data that was kept was of the highest quality, as described in the previous chapter.

Cell Line	Number of Pairs	Number of Compounds
A375	174936	592
PC3	184528	608
MCF7	253828	713

Table 1:Training Set pairs and compounds quantity for the three cell lines

Cell Line	Number of Pairs	Number of Compounds
A375	45584	77
PC3	44992	74
MCF7	49910	70

Table 2: Test Set pairs and compounds quantity for the three cell lines

In parallel, we worked the same way by utilizing the transcription factor distance of pairs of compounds. These distances were calculated with a Gene Set Enrichment method, similarly to the GO-Term enrichment method, by substituting the GO-Terms with transcription factors and considering the enrichment of the transcription factors. As a result, one more training set and test set was created for each cell line, including the transcription factor distances of the pairs of compounds.

2.3. Complementary Data

For the development of the model described in this thesis, there were plenty of complementary data used to connect and be able to transition through different fields of interest. The most important of these data were:

- The knockdown-signature connection data frame: This dataframe provided useful information about which transcription factors knockdowns are included in each cell line and what their respective transcriptomic signature is.

- The knockdown-transcription factor enrichment scores in each cell line: The enrichment scores were calculated using the FGSEA package in R. Every transcriptomic signature respective to the known transcription factor knockdowns in each cell line resulted in an enrichment score for each transcription factor, to be used later, to extract information about which knockdowns would be useable. For example, the transcriptomic signatures of the knockdown named *MYC* caused quite different enrichment scores throughout several cell lines.

- GO-Term enrichment scores for all the available compounds: This dataframe contains the enrichment scores of the 2105 GO-Terms examined after the application of 90,000 compounds.

- GO-Term enrichment scores for all the available knockdowns: This dataframe contains the enrichment score of the 2105 GO-Terms examined connected with the knockdown signatures in each cell line.

- The enrichment ranks of all the transcription factors for each compound: This dataframe was created through a simple enrichment analysis and a ranking procedure. More accurately, for each compound, a enrichment score was calculated with FGSEA for each transcription factor. Later on, these enrichment scores were ranked signature-wise, meaning that each transcription factor got a rank (from 1 to 175, since the included transcription factors are 175) for each compound signature.

A more in-depth analysis of the aforementioned data, along with their use in the model will be given in the *Methods chapter.*

3. Methods

The model that was created in this thesis is a combination of logical hypotheses and biological prerequisites. To back our hypotheses, we used analytical data and performance evaluation, labelling the project and the central hypothesis successful. The model was built in four stages briefly: knockdown availability, neighbor selection, inference selection, evaluation. As most models that are built on training and test data, it makes use of a deep learning model to learn relationships between data and implement them on other data to predict missing values. Through the predicted data, and with a series of biological thresholds in mind, the finally selected compounds confirm the success of the model.

The main target of this thesis is to identify compounds that approximate the biological effect of certain transcription factor knockdowns, in order to be able to create drugs that contain these compounds and suspend the effect of specific transcription factors. Such a method could be vital in the field of drug discovery and play an important role in developing target-specific drugs.

On this front, the logical steps of the model along with the methods that were used to draw logical conclusions and proceed from one step to the next were the following:

3.1. Knockdown Availability

As mentioned earlier, the prime goal of the model is to find compounds with close biological effects to certain knockdowns to suspend the effect of desired transcription factors. When dealing with big data libraries where not all relations between variables are comprehensible, it is important to try to understand the data essence and decipher its meaning through deliberate analysis and clustering. In this case, the

large data libraries that were used are three cell lines (A375, MCF7, PC3). To be able to run our model and detect drugs with close effects to certain knockdowns in these cell lines, there first needed to be proof that these cell lines included these particular knockdowns, or to be more accurate, their transcriptomic signatures. This knowledge was provided through the knockdown-signature connection data frame (See Complementary Data chapter) that pointed which knockdowns were included in each cell line. Of the knockdowns that were included in the three cell lines we decided to work only with the ones that were of quality 1, to build a more robust and accurate platform, through working with the best quality data. The results of this enquiry are shown in Table 3:

Cell Line	Number of knockdowns included	Number of knockdowns with quality 1
A375	125	32
MCF7	117	20
PC3	123	28

Table 3: Number of knockdowns included in each cell line

Moreover, the three cell lines that were used had 2 common knockdowns of quality 1– meaning knockdowns that were included in every cell line. These were knockdowns of transcription factors that are quite important to the functions of the organism and thus they are more frequent in cell lines. In total, there were 130 unique knockdowns throughout the three selected cell lines, of which 61 were of quality 1. It is important to note that data of another cell line was available (the VCAP cell line), but since it included only one knockdown, it was decided that it was of no computational importance and it was better if it were excluded from the platform.

Quality aside, it was decided that there needed to be a more complex criterion to determine which knockdowns would be acceptable to work with. On this front, we decided to use the enrichment calculations of the transcription factors over the respective knockdown signatures in the cell lines and decipher which knockdowns' enrichment score was tolerable. Using the FGSEA package in R, we created the knockdown-TF enrichment scores in each cell line dataframe (See more in Complementary Data). We then ranked the enrichment scores of each transcription factor over the knockdowns' transcriptomic signatures in the cell lines available and kept only the pairs of knockdowns and transcription factor of said knockdown in every cell line. This diagonalization of the enrichment scores ranks allowed deeper knowledge on which knockdowns of the available were over – or under – represented in the three cell lines.

In general, low rank meant that the enrichment score was negative, and that the transcription factor was downregulated. This, of course, meant that the respective knockdown was over-represented in the cell line, while high rank indicated the opposite. This over-representation was what we were seeking to complete a first screening on the given quality 1 knockdowns of each cell line. In this context, a threshold was introduced, that allowed only a limited number of the best (over-represented) knockdowns to pass. That threshold was determined to keep only the best 20% of knockdowns in each cell line, which were the knockdowns that we decided to keep working with to reach the best possible results.

All in all, for this first step of the model, a function was created named *available_tfs_in_cell_line*. Its main purpose was to interpret the data given and work out which knockdowns would be acceptable to use in the next steps. Its inputs are:

- The knockdown-signature connection data frame
- The knockdown-transcription factor enrichment scores in each cell line
- The cell line under review
- The threshold for the enrichment scores ranks, which will determine which knockdown we can work with

In accordance with the inputs, the output of the function created is a vector with the names of the knockdowns that are guaranteed to give the most accurate results.

3.2. Neighbor Selection

The second part of the model, which is called neighbor selection, can overall be described as the function to allocate the compounds in the training set that have the closest biological effect to the selected knockdown. This step utilizes the known data of the training set to find the relations between the transcriptomic signature of the knockdown and those of the rest compounds. Moreover, it makes use of different criteria to determine which compounds are the neighbors of the knockdown. These criteria derive mainly by the understanding of biological and structural differences of compounds, as well as implementing strict and targeted computational processes.

Having already run the first step of the model, we have determined which knockdowns are available to work with in each cell line. By selecting one of the available knockdowns in a cell line we will try to find which of the known compounds in the training set shares biological activity with this particular knockdown. This process takes into consideration the following aspects of the compounds:

3.2.1. GO-Term level distance of knockdown and compounds

The earliest concept that one naturally refers to when trying to decipher relations between compounds is a way to quantify their similarity or dissimilarity on the structural or biological level. Since we care to find compounds that share similar biological effect with the knockdown whatever their structure may be, we concentrated on the biological distance of the compounds and the knockdown. This biological distance was derived as distance between GO-Term-level vectors for every compound in the training set, calculated similar to Iorio et al. The data for the GO-Term enrichment scores of the compounds were available by the respective dataframe, described in the Complementary Data chapter. In parallel, the GO-Term enrichment scores of the knockdowns were available in the respective dataframe, described in the same chapter. By using the Gene Expression Signature package in Bioconductor in R we were able to compute every distance between the training set's compounds and the knockdown of interest.

For every pair of knockdown and compound, two feature vectors ranked by NES are considered, called A and B, GSEA is used to calculate the enrichment scores of a number of top and bottom GO-Terms of A in B and vice versa. The calculation of the distance between the two vectors is given by the averaging of the two enrichment scores, a method that gives a result in range of 0 to 2. To have a more precise calculation of distances we repeated the process by creating an ensemble of 5 different ES, respective to 5 different numbers of top and bottom GO-Terms considered (10, 20, 30, 40, 50). The ensemble distance was a result of the averaging of these 5 ES and normalized between 0 and 1. The outcome of this process was a vector of the GO-Term distances of the selected knockdown and all the compounds in the training set.

The following Figure 9 depicts an example of the output of this process. The vector of the biological distances shows the relevance of each compound of the training set to the knockdown on the biological effect level. From such a histogram one can draw many conclusions about the quality of the data and the proximity of compounds to the knockdown. For example, in Figure 9 the MYC knockdown has many compounds with similar biological effect (small distance), since it is quite a common knockdown.

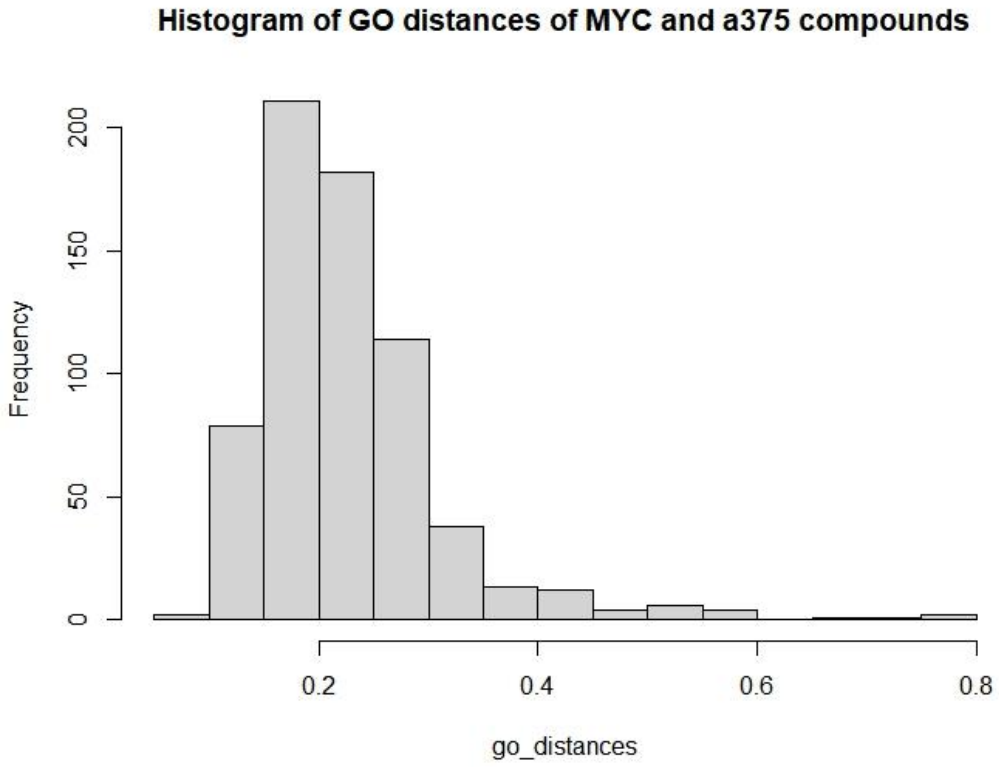

Figure 9: Histogram of biological distances between the MYC knockdown and the compounds of the a375 cell line. Notice how the distances of the compounds are strongly shifted to the left, meaning that many compounds show similar biological effects to the knockdown. This happens because the MYC knockdown is a quite common knockdown relevant to the survival of the organism.

The whole premise that smaller GO distances equal similarity in biological effects deeply influenced our search. Therefore, by observing the histogram in Figure 9, a decision had to be made on which distances would be considered close enough, to label a compound as a neighbor. After a lot of tampering and parameter exploration, which is best described in the Results chapter, we came to the conclusion that the threshold for the distance would be 0.15. It is important to note that the platform analyzed in this thesis

tries to assign the right priority to compounds to be used in further research. Since every knockdown has a different distribution of distances - most of them being shifted to the right – this threshold had to be judged accordingly. It was decided that at this point, the best threshold parameter for the go distances would be 0.15. Thus, by applying this threshold to the compounds of the training set, we excluded most of the compounds and kept only the best and closest to the knockdown compounds to continue our search for neighbors.

3.2.2. Transcriptional signature distance of knockdown and compounds

For the compounds that had quite low GO-Term distance and passed the aforementioned threshold, there needed to be a better and more precise screening. To exact the first screening, we utilized the GO-Term enrichment scores of the compounds to create vectors of GO-Term distances. In this second layer of screening, we chose to make use of the knockdowns enrichment score of each signature and create a similar kind of method to calculate compound-knockdown distances and make use of a second threshold to close in on better neighbors.

The process followed at this step was similar to the last one. The key difference was that the enrichment scores for the signatures of the compounds and the knockdowns were expressed over transcription factors and not GO-Terms. The data for these processes were available from the enrichment ranks of all the transcription factors for each compound and the knockdown-transcription factor enrichment scores in each cell line as described in the Complementary Data chapter. After using the Gene Expression Signature package again, the ensemble provided distances at the transcription factor level for every knockdown and every compound. To distinguish the two kinds of distances we named the new ones tf_distances, since they were derived not from GO-Terms, but from knockdowns.

These distances presented a kind of similarity to the GO-Term distances of the last step, dew to them both being indicators of the biological distance between compounds. The following Figure 10 depicts a typical distribution of the distances of the MYC signature and the signatures of the training set of the a375 cell line. Notice that the new distances follow virtually the same distribution as the go_distances, but with a visible shift to the right, meaning our second threshold must be more relaxed.

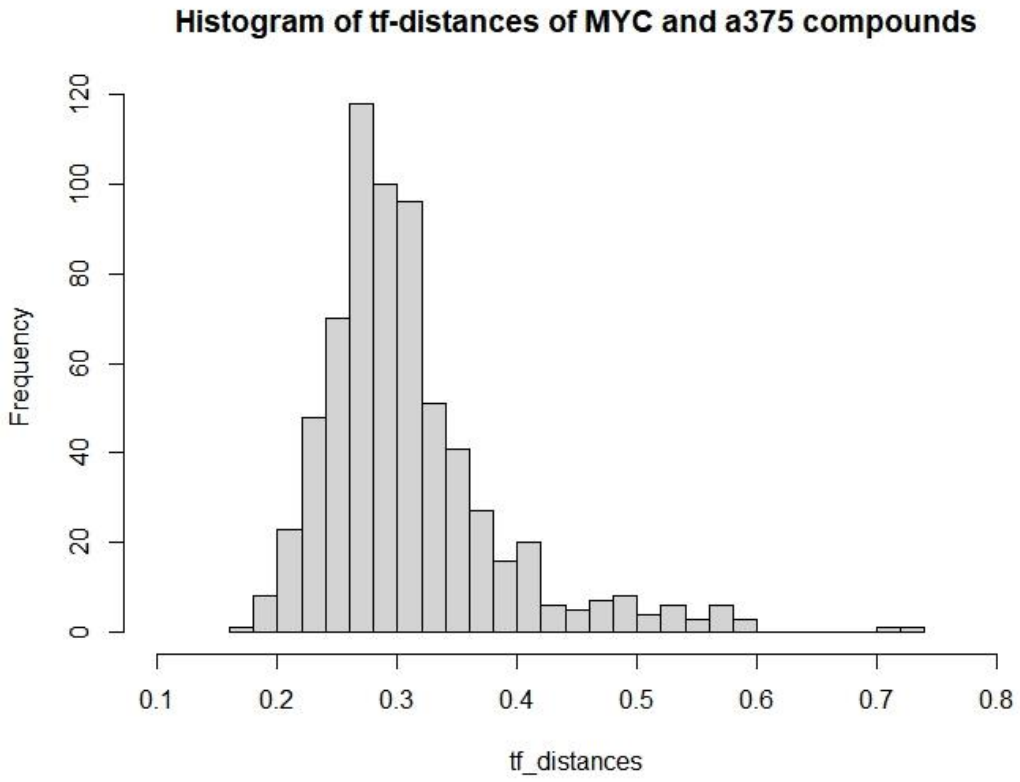

Figure 10: The distances between the transcriptomic signature of the MYC knockdown and the transcriptomic signatures of the compounds in the a375 training set. Notice how the distribution follows the same pattern as before but is shifted to the right.

To extract a threshold for the tf_distances of the compounds there needed to be a closer examination of how the threshold affects the accuracy of the model and the number of neighbors. This analysis will be presented in the Results chapter. It was clear though, that running the two thresholds in parallel and keeping strict values for both would result in picking no neighbors for the knockdown. In this regard, we decided to apply a looser threshold for the tf_distances at 0.25. Of course, this threshold could change depending on the circumstances of the prioritization and the quality of the data.

All in all, the neighbor finding process was strongly based on the biological understanding behind computational data. By utilizing computational processes, we managed to find distances on the biological level and show the relations between knockdowns and compounds. For this whole purpose, a function was built that had as inputs:

- The selected knockdown and its transcriptomic signature which should be included in the available knockdowns presented by the first step (See former chapter)
- The cell line on which we want to work
- The training set of the cell line containing pairs of compounds and their biological distance
- The dataframes with the GO-Term enrichment scores for all the available compounds and the knockdowns

- The dataframe with the enrichment ranks of all the transcription factors for each compound
- The threshold for the go_distance
- The threshold for the tf_distance

The outcome of this function is a vector of compounds that are considered *neighbors of the knockdown.* These are the compounds of the training set that have passed all the thresholds and are believed to have the greatest biological similarity to the knockdown of interest. These neighbors will be utilized in the next steps.

3.3. Inference Selection

The third step – and the most important one – in this model is the selection of compounds from an unknown library to prioritize for further testing in the drug discovery process. More accurately, the model can process unknown compounds with high precision and propose several compounds that approximate the effect of the selected knockdown as closely as possible. As expected, the current step of the platform is strongly dependent on the neighbors that were selected in previous step. We will be trying to utilize the predictions of a deep learning model on the test set, which includes all the unknown compounds, to locate chemical structures that show great similarity with the neighbors we have already suggested. Therefore, we are using the neighbors as a medium to detect unknown compounds with biological similarities to the selected knockdown.

3.3.1. The deepSIBA Model & GO-Term distance

The most significant predictive model that was used during this diploma thesis was the deepSIBA model (Fotis, Meimetis, et al.) [28]. DeepSIBA stands for Deep Learning for Chemical Structure-based Inference of Biological Alterations. This deep learning model was constructed by my fellow colleagues at the BioSysLab and can predict the biological similarity of chemical structures with great precision.

DeepSIBA uses representations of the compound structures given as graphs and then draws links to their biological effect distances. These compound differences were connected to their subsequent biological effect differences through the use of Siamese Graph Convolutional Neural Networks (GCNNs).(Figure 12) GCNNs in this case worked on molecules of chemical structures represented as graphs, containing nodes and bonds. The nodes represented the atoms and the edges represented the bonds between these atoms. (Figure 11). It is important to note that the present model succeeds in learning neighborhood level representations of the input graphs and link them to respective biological effects. Moreover, the model managed to learn and identify structurally dissimilar compounds that share biological effects. The application of model ensembles provided both greater accuracy and precision, while calculating the uncertainty of the predictions.

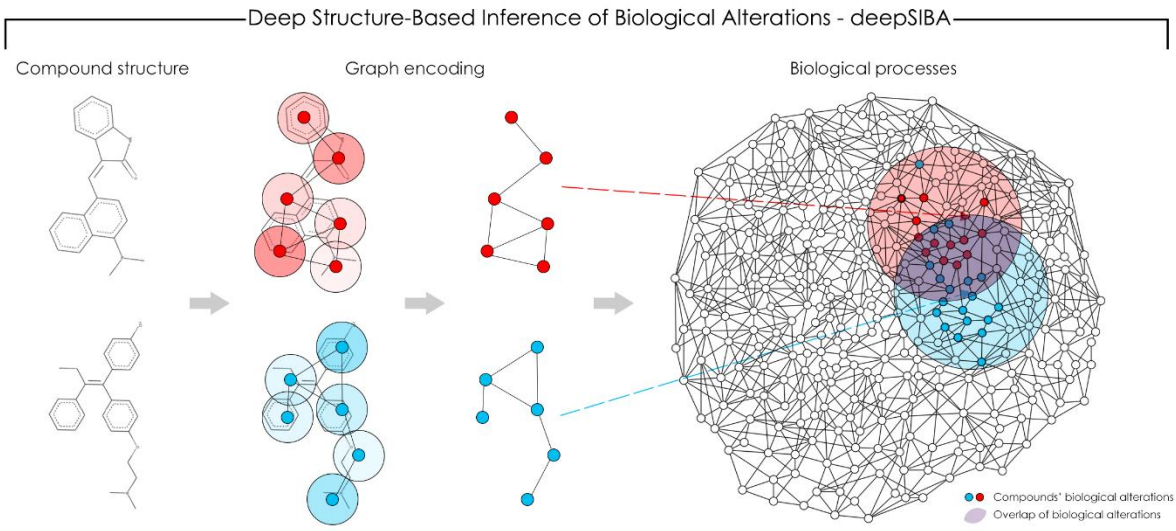

Figure 31: Graph encoding of compounds pairs as nodes and edges and connection of structures to biological function distance

The model is trained on pairs of compound structures and their biological effect distance. Predictions are made only by supplementing pairs of compounds, similar or completely dissimilar to the training set's structures. The reliability and flexibility of the model make possible its use for any set of structural pairs and a value connecting that pair. To achieve maximal precision and ensure the best performance of the model, we used training sets and test sets whose distributions of distances was similar.

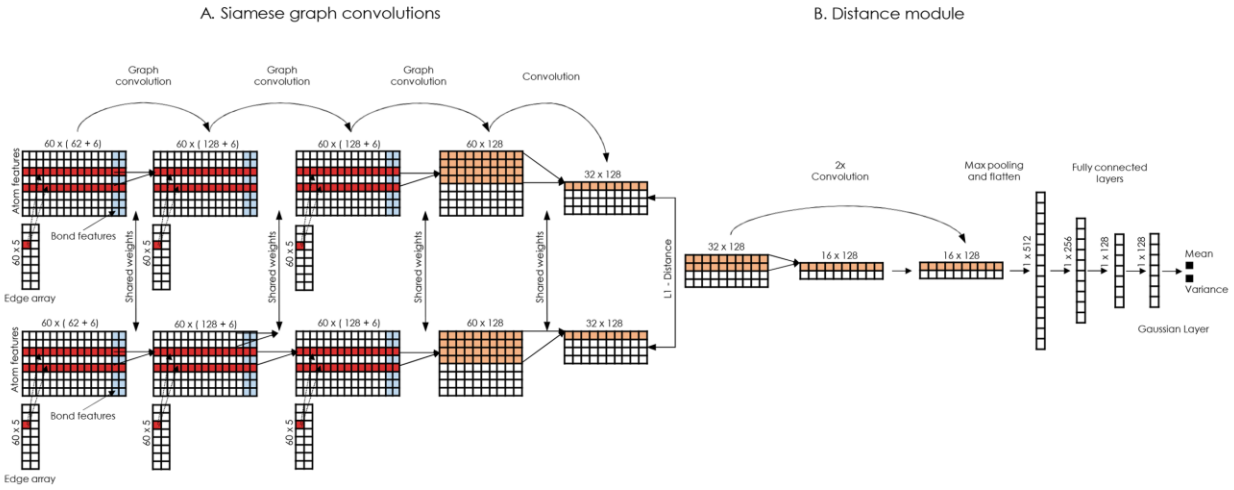

Figure 14: The architecture of the deepSIBA model

In our case, the deepSIBA was used to learn representations from the compounds in the training set and predict the biological distances of the compounds in the test set based on their structure. Such a predictive tool was quite useful, since it made it possible to predict, with high precision, the biological distances of all the unknown compounds and the neighbors that were suggested in the former step. For example, on the a375 cell line, the model was used to predict the distance of the neighbors and 77 cold compounds, giving us the opportunity to pass judgement on totally unknown compounds up to that point.

The deepSIBA model presents the results with both the average value of the prediction and the standard deviation of that value. For our purposes we used the average predicted value to compare the biological effects of the neighbors and the cold compounds. Through that average value was it possible to screen the unknown compounds and choose the ones that approximate the neighbors on a biological level. On that front, we introduced a threshold to keep only compounds with low distances to the suggested neighbors. This threshold was decided to equal 0.22. This way, there was a first screening of the compounds that resulted in keeping the ones that presented low distance to at least one neighbor (they were neighbors of a neighbor of the knockdown). That first screening, although it eliminated many compounds, was not enough to conclude the model accurately.

3.3.2. Majority

Since the predictions of the deepSIBA model along with the last threshold provided a number of cold compounds that could potentially be neighbors to the selected knockdown, the next step was to elaborate on these compounds and provide a more precise way in order to pick the best of them.

The logical conclusion that was drawn from the relations between the cold compounds and the neighbors is that the more the compounds resemble the biological effects of most of the neighbors, the more they are bound to resemble the biological effect of the knockdown. In short, strong relations to more neighbors means stronger relations to the knockdown.

In this respect, we were concerned with the number of the neighbors each compound had a close distance to. Therefore, the concept of majority was born, meaning the percentage of neighbors each cold compound resembled closely enough. For example, if a knockdown had 5 neighbors and a compound resembled the effect of 3 of them (they passed the threshold of the last step for 3 out of the 5 neighbors), then the majority of this compound would be 0.6. Consequently, we tried to screen the cold compounds deeper, by introducing a new threshold concerning the majority of each compound, as it was defined above.

This majority threshold was the way to select the best compounds, the ones that showed resemblance to the most neighbors. Therefore, after a parameter exploration, as presented in the Results section, we determined that the majority threshold should allow only the top 30% of the compounds concerning their majority. This way we deepened the screening and made it more robust.

3.3.3. Transcription factor distance

To add another layer to the inference selection and make it even more precise, we decided to expand the model based on the predictive power of the deepSIBA model. In this context, we created a new training set and a test set by replacing the GO-Term distances with the transcription factor distances, as they were calculated in the neighbor selection chapter. A new deepSIBA model was trained based on the new data and new transcription factor distances were predicted for the test set.

Accordingly, based on these predictions we eliminated a few more compounds from the finally inferred ones. This happened exactly in the same manner as when handling the GO-Term distances. We introduced a threshold and when exploring it we settled on the threshold value of 0.25 for the transcription factor distances.

In parallel with the majority threshold for the GO-Term distances, we tried to integrate a similar majority threshold for the transcription factor distances, but the more the thresholds, the more difficult it was to locate any compounds that fulfilled them all. Thus, we propose a majority threshold of 0.7 for the transcription factor distances, the same way as before, in libraries that include much more data.

All in all, to select the "cold" compounds that had the smallest distance to the desired knockdown, a function was created that had the following inputs:

- The neighbors of the selected knockdown as they were calculated in the former step
- The predictions of the deepSIBA model on the test set for the GO-Term distances of pairs of compounds
- The predictions of the deepSIBA model on the test set for the transcription factor distances of pairs of compounds
- A GO-Term distance threshold to keep "cold" compounds that have close go_distance to neighbors of the knockdown
- A TF distance threshold to keep "cold" compounds that have close tf_distance to neighbors of the knockdown
- A majority threshold to keep the "cold" compounds that show similarity to most neighbors

The output of this function is a list of the "cold" compounds that are most likely to approximate the biological effect of the selected knockdown. The whole model is a method to prioritize certain compounds in the drug discovery process.

3.4. Evaluation

To conclude a model like the one developed in this thesis, there needs to be an evaluation of its performance and its accuracy. Having already selected a number of "cold" compounds through the rest of the model, there needs to be testing to determine whether the selected compounds actually have a biological proximity to the desired knockdown. For this purpose, we developed the following evaluation metrics:

3.4.1. GO-Term distance accuracy

This evaluation metric tests whether the proposed compounds' GO-Term distances to the knockdown are actually little enough so that they would be described as neighbors. It is based on a proposed threshold largely affected by the biological perception of the data distribution. The go_distance accuracy of the model is defined as the percentage of the GO-Term distances of the proposed compounds and the knockdown are below the threshold. We decided that for the model to be characterized as accurate over this metric, the evaluation threshold should ensure that the cold compounds are only of the top performing compounds as far as distance to the knockdown is concerned. For this purpose, this evaluation threshold was set to be 0.23 for the MYC knockdown, from the distribution shown in Figure 13. The go_distances were calculated as before, through the GO-Term Gene Expression Signature package of Bioconductor. Naturally, every knockdown has its own distribution of go_distances and the threshold should be adjusted accordingly.

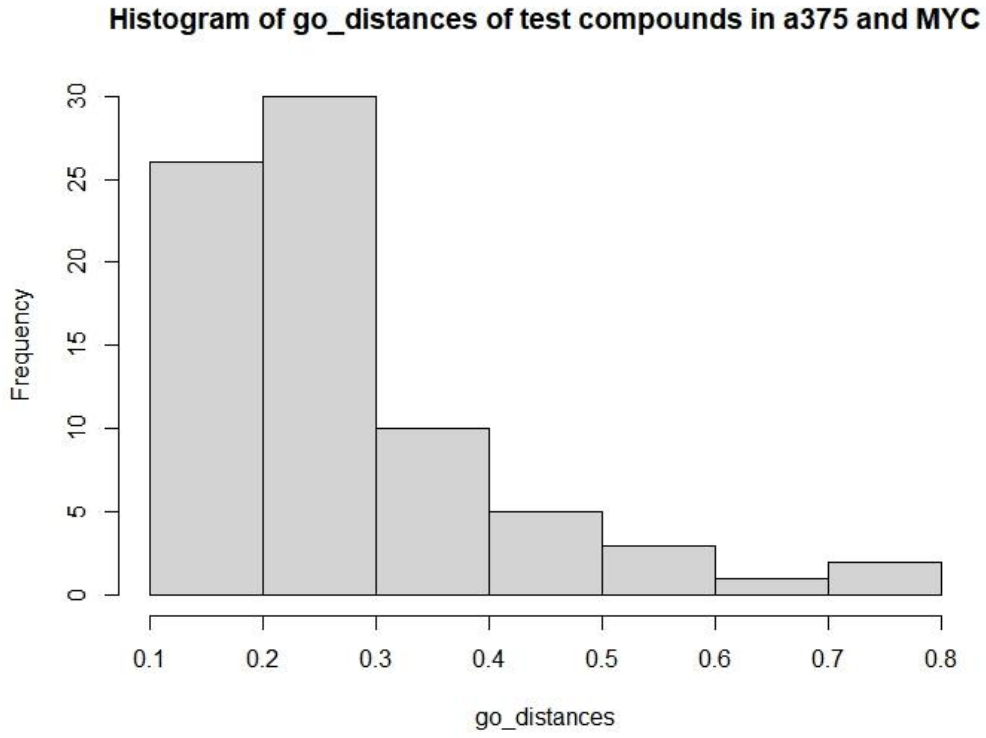

Figure 53: Histogram of the GO-Term distances of the cold compounds in the test set of the a375 cell line and the MYC knockdown. We can see that the distribution of the values is highly shifted to the left, as they were in the training set for this knockdown. For this purpose, we chose to set the evaluation threshold at 0.23. Of course, this evaluation threshold should be different for each knockdown.

3.4.2. TF distance accuracy

In parallel to the last accuracy metric, the TF distance accuracy measure the proposed compounds' TF distances to the knockdown and deciphers if they are actually neighbors based on a threshold. This

threshold is suggested from the distribution of the tf_distances of the cold compounds, to judge whether the selected compounds are included in the best possible ones (have the lowest tf_distance to the knockdown). The tf_distance accuracy of the model is defined as the percentage of the tf_distances of the proposed compounds and the knockdown are below this threshold. For this purpose, this evaluation threshold was set to be 0.27 for the MYC knockdown, from the distribution shown in Figure 14. The tf_distances were calculated as before. Of course, every knockdown has its own distribution of tf_distances and the thresholds should be adjusted accordingly.

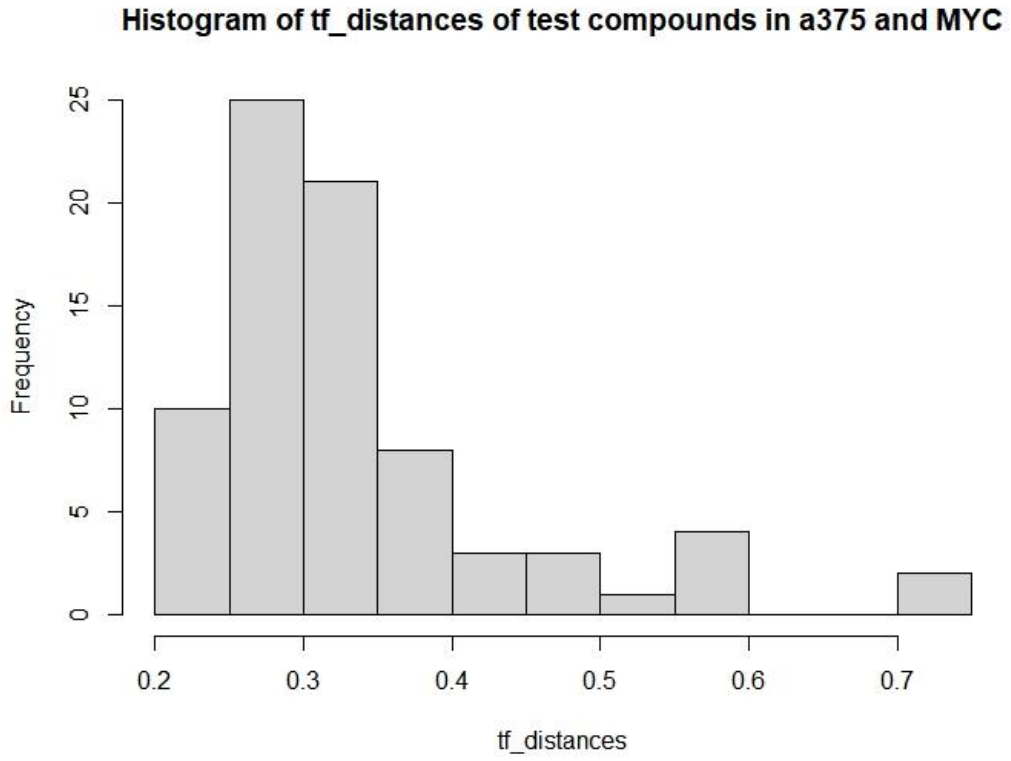

Figure 14: Histogram of the tf_distances of the cold compounds of the a375 test set and the MYC knockdown. Notice that the distribution resembles the respective distribution of the GO-Term distances. This distribution suggests that the evaluation metric for the tf_distances should be around 0.25-0.27

3.4.3. Combined Accuracy

This was a general accuracy metric that stemmed from the last two metrics. It is a way to generalize the go_distance accuracy and the tf_distance accuracy. In short, it returns the percentage of the compounds that fulfill both the evaluation thresholds that were described above.

4. Results

Throughout this thesis our goal was to develop a valid model that would effectively prioritize unknown compounds that resemble the biological effect of a transcription factor knockdown. This prioritization would offer a helping hand to biologists' efforts to discover new drugs and combat diseases through transcription factor activity. The performance of such a model is complicated to map, but this chapter offers proof that the method that was followed produces high quality results.

Judging the effectiveness of a method is not solely based on its results. To conclude that a series of computational steps have biological basis and play a role in the compounds' screening, there needs to be proof that the criteria which performed the screening are strict and produce results accordingly. The success on the selection of these criteria will be discussed through a parameter exploration. Moreover, the optimal values of the thresholds performing the screening and how we arrived at them will be discussed. In addition, to prove the methods generalization, its performance over different cell lines is considered along with the behavior of the parameters in parallel. Finally, we will present evidence that the model has strong computational value and that the prioritization of compounds helps select compounds with the desired characteristics.

4.1. Available knockdowns per cell line

Working with knockdowns in cell lines prerequisites that the knockdown is over-represented in that particular cell line. As described in the respective chapter in methods, this knowledge stems from the enrichment scores of transcription factors over their knockdown's signature in the cell line. Low enrichment score means that the transcription factor is downregulated and therefore the knockdown has a strong effect in that cell line. Keeping only the 20% of the best represented knockdowns in each cell line to make our model more robust and dependable we found that the cell lines provided the following numbers of knockdowns to work with. (Table 4)

Cell Line	Number of Knockdowns Available
A375	6
MCF7	4
PC3	3

Table 4: The number of knockdowns available to work with in each cell line. The most common knockdown present in the cell lines is the knockdown of the MYC transcription factor

Provided that the most common knockdown in these cell lines is the one controlling the MYC transcription factor, we decided to present most of the results focusing on MYC.

4.2. Parameter Exploration

Creating a model that utilizes biological data prerequisites knowledge and understanding of the biology behind the data, but most importantly it requires a clear path to connect the computational data and the biology. In this compound screening platform we developed a model that prioritizes chemical compounds over others with the main criterion being the compounds' similarity of biological effect to a transcription factor's knockdown (a structure that suspends the transcription of certain genes). To create the platform, there needed to be a connection between the data available and the biology. This connection was attempted with a series of parameters-thresholds throughout the span of the platform, that essentially were the ones to perform the screening.

The parameters that were included in the model were:

- The GO-Term distance threshold between the training compounds and the knockdown in the neighbor selection
- The TF-distance threshold between the training compounds and the knockdown in the neighbor selection
- The GO-Term distance threshold between the cold compounds and the neighbors of the knockdown in the training set
- The majority threshold to select only cold compounds that had similarity to most of the neighbors of the knockdown in the training set
- The TF-distance threshold between the cold compounds and the neighbors of the knockdown in the training set

The aforementioned parameters performed the neighbor selection and the inference selection. One could argue though, that although the motives and the biology of the reasoning behind the selection of these thresholds were correct, the whole computational process and the intervention on the data through several functions would alter the outcome of the platform and make it ineffective. To cast away any arguments of disbelief that the thresholds chosen are random or dysfunctional, we will present evidence that shows how each parameter regulates the results of the platform and how we worked to determine the optimal value of each threshold. This parameter exploration and evaluation will have to be conducted in a way that clearly cements the role of each threshold in the successful outcome of the model. Therefore, we will pass judgement on every threshold by linking its value to the accuracy metrics of the evaluation. To perform this task, we will be freezing every time the rest of the parameters at logical values and presenting how the accuracy metrics differentiate over different values of said threshold. We will consider a threshold successful only if there is clear evidence that its relaxation results in worse model performance.

4.2.1. GO-Term distance threshold in neighbor selection

The first case where we made use of a threshold was in the neighbor selection process. The neighbor selection is the process of screening training set compounds that are considered neighbors to the knockdown. This consideration of a compound being a neighbor of the compound stems from both the biological perception of the data and the threshold we apply for this purpose. Having already settled on the evaluation parameter for the GO-Term distance of the inferred compounds, there needs to be an obvious link between the neighbor selection GO distance threshold and the GO accuracy of the model, to be able to conclude that the threshold indeed has a positive effect on the model. While keeping the rest of the parameters frozen at normal values, we investigated the effect of the GO distance threshold of the neighbor selection on the GO accuracy of the model. This investigation is pictured in the following Figure 15:

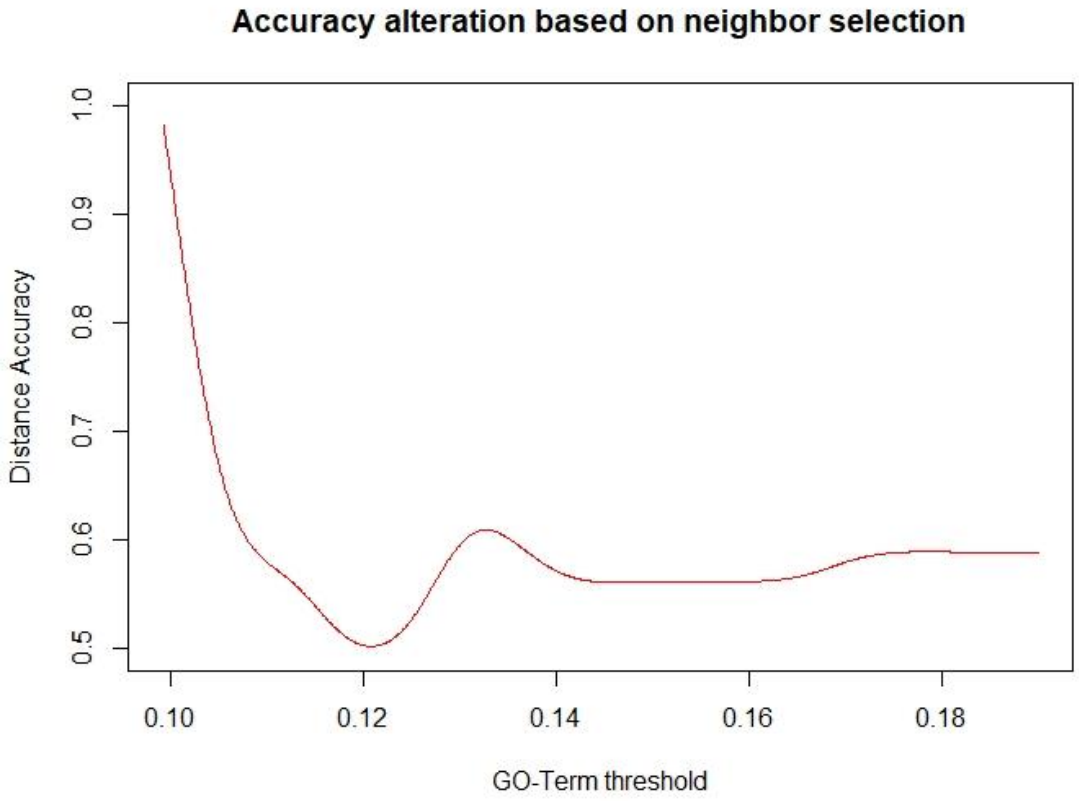

Figure 15: Diagram of the effect of the GO-Term threshold in the neighbor selection to the GO-distance accuracy of the model. It is obvious that the stricter this threshold is, the better the model performs. Thus, we conclude that the threshold is indeed working.

The Figure 15 above really does credit to our selection of the GO distance threshold as a neighbor selection mechanism. It is obvious that the threshold is working, since the accuracy of the model is deeply affected by it. We can see that the stricter the value of the threshold is the better the model performs. The simple explanation behind this fact is that a more relaxed GO-distance threshold in the neighbor selection process leads to more and worse performing neighbors, as shown in the next Figure 16. These new

neighbors later, in the inference selection, will result in more cold compounds being screened, and since the neighbors are of lower quality, the inferred cold compounds are also of low worse quality than before.

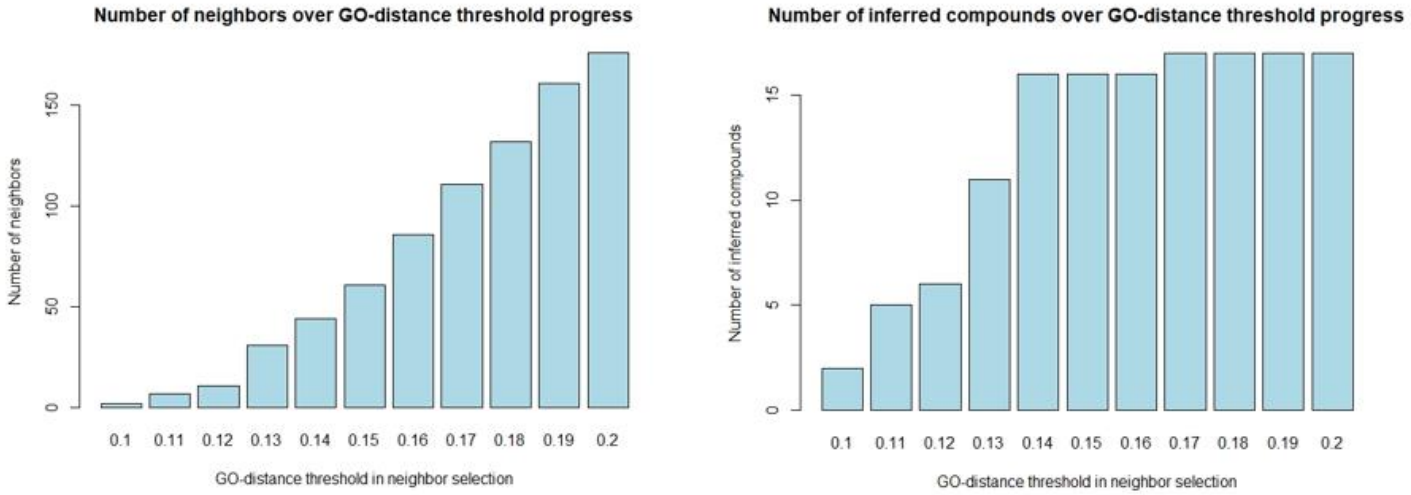

Figure 16: Progression of the number of neighbors and the number of inferred compounds as the GO-distance threshold in neighbor selection is loosened. We can see that the number of inferred compounds rises significantly, triggered by the enlarged number of neighbors. Of course, it is topped at 17 compounds, since the rest of the thresholds' effect somehow confines this number.

4.2.2. Transcription Factor distance threshold in neighbor selection

The second threshold used in the neighbor selection process integrates the distance of the compounds from the knockdown based on transcription factors. This time around, we made use of the TF-distance accuracy of the evaluation and tried to prove that after the GO-distance, the TF-distance threshold plays an important role in the neighbor selection. The same way as before, this strong link between the threshold and the biology of the problem can be proven only through the examination of the effect of the loosening of the parameter on the performance of the model.

Thus, by following a similar approach as with the GO-distance threshold, we focused on the effect of the parameter shift on the TF-accuracy in the evaluation. By freezing the rest of the parameters, we managed to create the chart in Figure 17. This diagram strongly proves that the stricter the threshold for the TF-distance of the compounds and the knockdown in the neighbor selection is, the better the model performs.

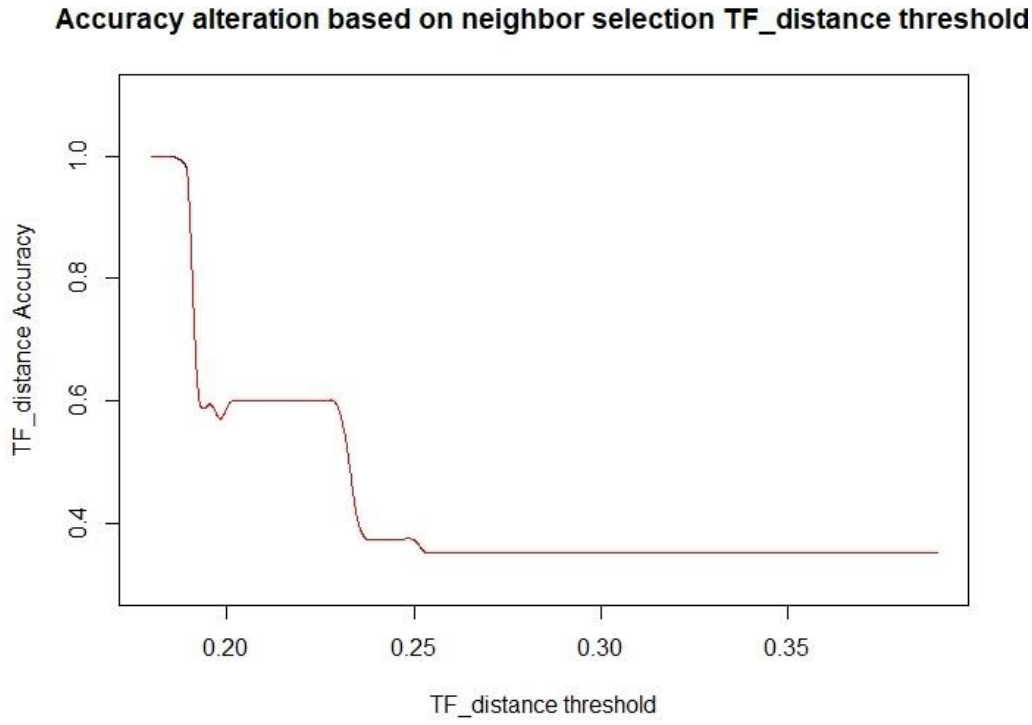

Figure 17: Chart of the effect of the TF-distance threshold in the neighbor selection to the TF-distance accuracy of the model. One can notice that the more we loosen the threshold the worse the model performs.

Figure 17 strongly suggests that the threshold selection is successful, since looser parameter results in worse performance for the model. As before, this occurs due to the bigger sample of neighbors' compounds that are being screened which will in turn lead to a bigger sample of inferred compounds, which will undoubtedly be of lower quality (Figure 18).

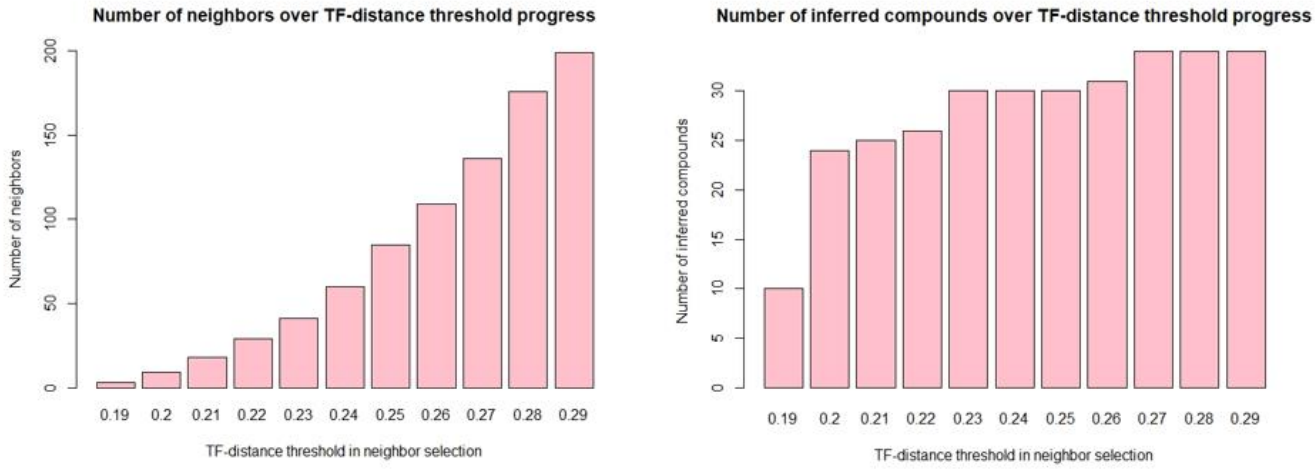

Figure 18: Progression of the number of neighbors and the number of inferred compounds as the TF-distance threshold in neighbor selection is loosened. The same way as with the GO-distance threshold, stricter threshold means lower number of cold compounds finally inferred, which of course have the best possible quality.

The results advocate that the TF-distance threshold in the neighbor selection process creates the basis for the screening of quality compounds. The steps in Figure 17 are due to the cold compounds passing through the “filtering” in a quantized form.

All in all, we proved that both the thresholds used in the neighbor selection have a computational and biological basis, that is pictured in the model performance. These thresholds screen compounds with high precision and present the neighbors of the knockdown to be used for the inference selection.

4.2.3. GO-Term distance threshold in inference selection

After the neighbor selection process, the model proceeds with the selection of the cold compounds that approximate the biological effect of the knockdown. For that purpose, we followed the same approach as before and introduced a GO-Term distance threshold and a TF-distance threshold. The same as the GO-Term distance threshold in the neighbor selection, here, the GO-Term distance threshold appears to be facilitating the model with the required accuracy. Figure 19 points out the change in the models’ accuracy based solely on that GO-distance threshold.

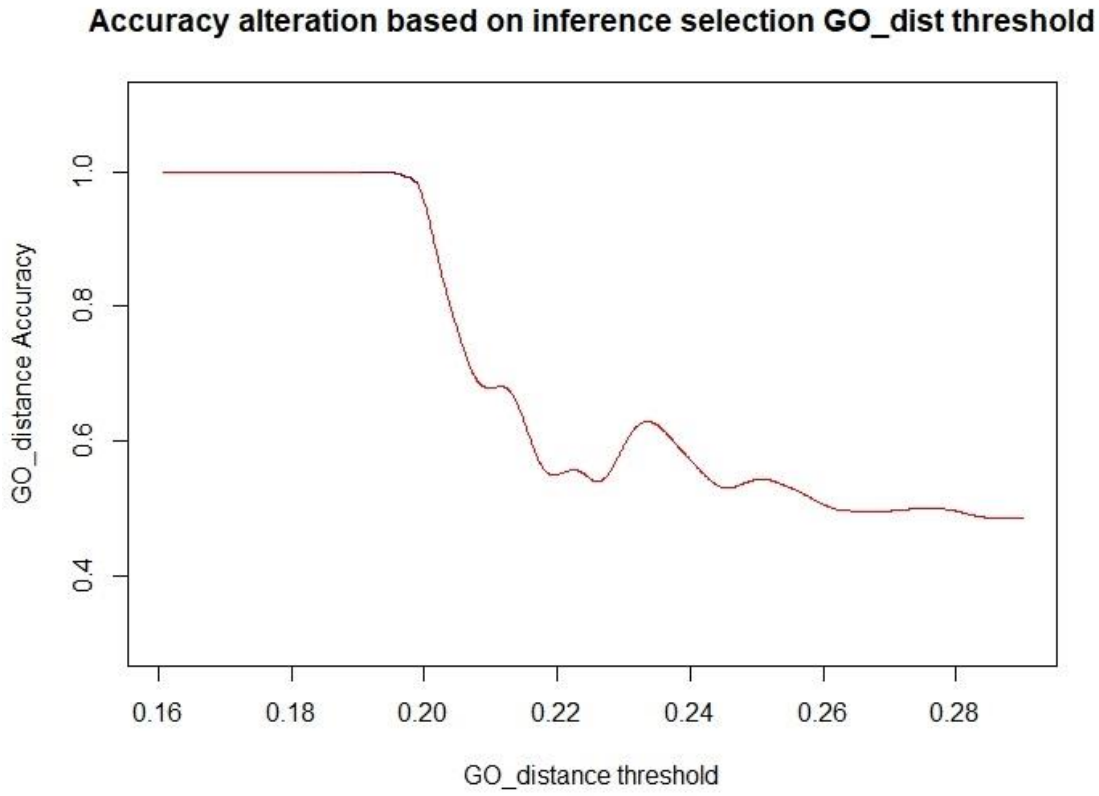

Figure 19: Diagram of the relation between the GO-distance threshold and the GO-distance accuracy of the model. It is obvious that the threshold is working, since loosening it leads to worse model performance. To investigate this relation, the rest of the parameters were kept stable.

The observation of Figure 19 allows us to draw the same conclusions as before. Lower and more strict GO-Term distance thresholds in the inference selection lead to better results. Moreover, we can see that there is an area in the left part of the chart where the accuracy remains steady at 100%. This means that the threshold has strong biological basis, since it screens the best possible cold compounds out of the test set.

4.2.4. TF-distance threshold in inference selection

To cement the performance of the model, we introduced the TF-distance threshold in the inference selection, the same as we did in the neighbor selection. Following the same mindset as before, we conclude that the threshold has actual importance in the model, since it affects its performance. The TF-distance threshold in the inference selection improves the models' accuracy as it becomes stricter. This proof of concept for the TF-distance threshold is clearly shown in Figure 20.

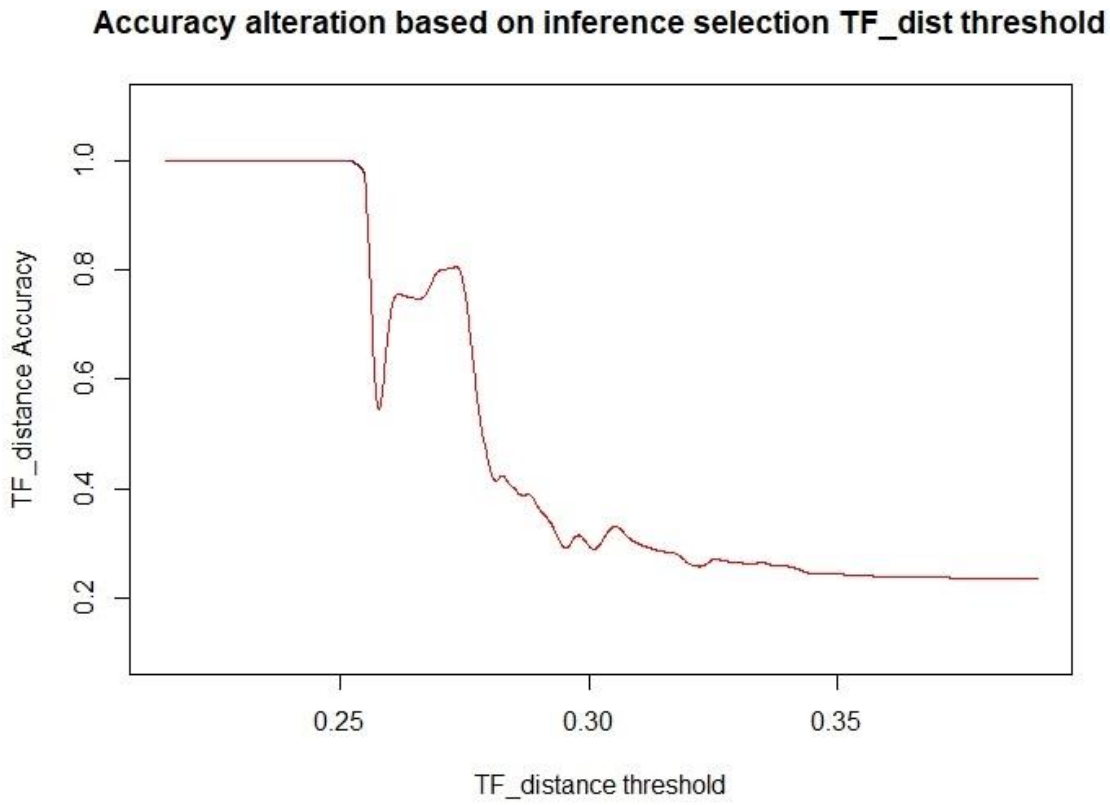

Figure 20: Following the same trend as in the cases before, the TF-Term distance threshold proves its value for the model. Low TF-distance threshold values in the inference selection process lead to better model performance, providing evidence for the successful selection of a concept threshold.

4.2.5. Majority Threshold

The last parameter that must be evaluated is the majority threshold introduced in the inference selection process. As stated earlier, the majority threshold screens the top performing cold compounds based on the number of neighbors they show similarity to (similarity stems from the GO-Term distance threshold in the inference selection). The majority threshold selects the top tier of cold compounds that show similarity to most neighbors. It is natural, that since the GO-Term distance thresholds work, so will the majority threshold. In short, this parameter screens the compounds that are the closest to most of the knockdown neighbors. The higher the value of the parameter, the better top percent compounds are selected.

To prove the usefulness of the majority threshold, we will once again demonstrate its effect on the optimization of the model's performance. The next Figure 21 depicts the relation between the majority threshold and the GO-Term accuracy of the platform. It is obvious that the majority threshold improves the overall performance of the model. The GO-Term distance accuracy of the model is enhanced, as expected, through eliminating more cold compounds that show less similarity to most of the neighbors of the knockdown. At the high values area of Figure 21, only a few of the cold compounds are screened, thus optimizing the accuracy of the model. This analysis of the majority threshold was conducted by keeping the rest of the inference selection parameters generally idle, just to view the influence of the majority to the performance of the model.

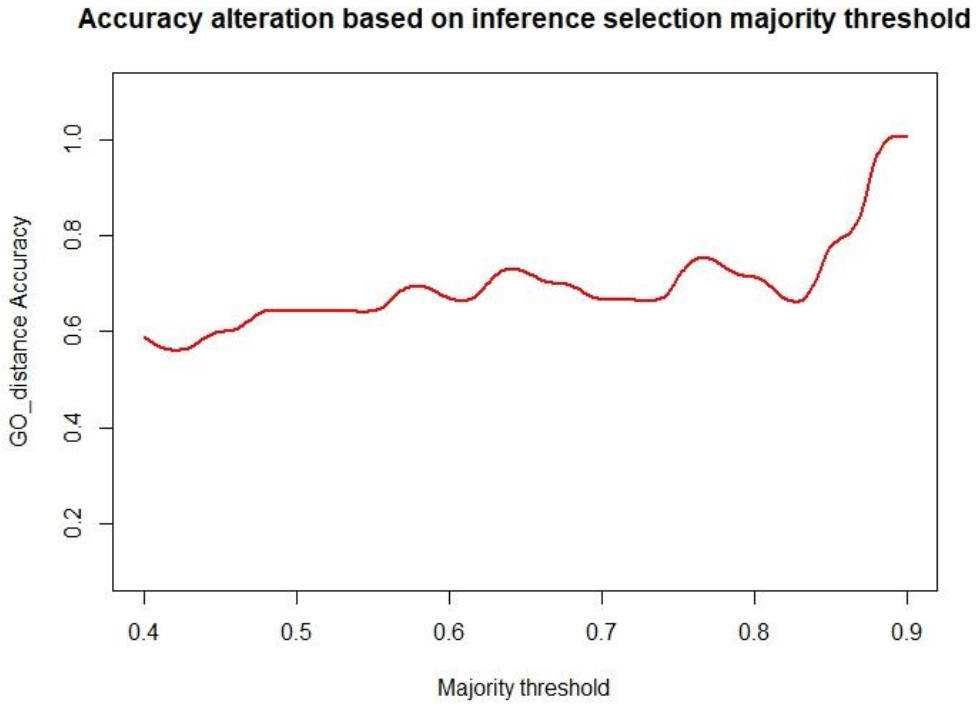

Figure 21: The representation of the relation between the majority threshold and the GO-Term distance accuracy of the model proves that the majority threshold is indeed a successful parameter for the cold compound screening process. Notice how the accuracy of the model reaches 100% as the majority threshold becomes strict. At around 0.85, the model selects only the top 15% of compounds with the best relations to most of the neighbors of the knockdown.

This effect of the majority threshold on the number of compounds being inferred is clearly pictured in Figure 22, since higher and stricter majority threshold values lead to less and better performing compounds.

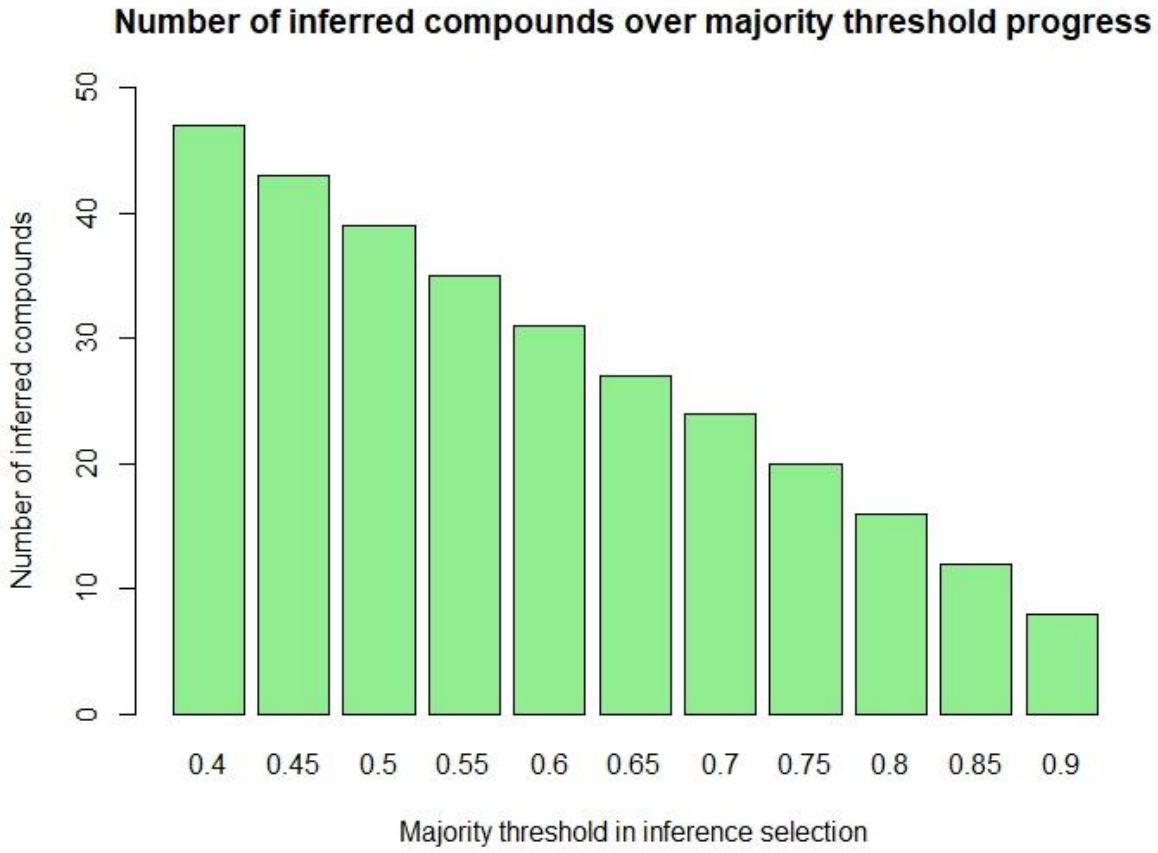

Figure 22: How the majority threshold affects the number of cold compounds being inferred by the model, strongly influences the performance of the model. The majority threshold selects the top tier compounds. The majority threshold organizes this "top tier" and selects only the exceptionally performing compounds.

4.3. Parameter Optimization

In the last chapter we explored every parameter of the model individually and proved the effectiveness and importance of each threshold to the performance of the model in general. The figures provided pictured the successful selection of each threshold and helped determine the best value for each threshold.

To make the platform more robust and cement its validity and effectiveness, we decided to not only use a single threshold in every run of the model, but better create an ensemble based on these thresholds, as described in the methods chapter. By using the individually optimal value for each threshold, the model

struggled to find hits, and most of the times it did not return any results. For our model to have a statistical importance, there needed to be as many hits as possible while retaining maximal accuracy. On this front, after determining the individual worth of each parameter, we proceeded to examine the combined effect of all the thresholds, on which the model's performance depends.

Combining the thresholds and optimizing their combined value to extract the best possible results from was a challenge that required working on many different levels at the same time. The basic parameters were 5 (GO-distance threshold and TF-distance threshold in the neighbor selection, GO-distance threshold, TF-distance threshold, and Majority threshold in the inference selection). Judging the success of the model was based not only on the combined accuracy of GO- and TF-distances in the evaluation, but also on the number of compounds proposed. This interest for high number of compounds arose from the need to test the statistical importance of the model and the desire to propose several compounds that researchers could potentially work with.

On this regard, we utilized a form of grid optimization on a 7-dimensional space defined by the 5 thresholds, the combined accuracy, and the number of inferred compounds. By shifting the threshold values, we tried to maximize both the combined accuracy and the number of inferred compounds.

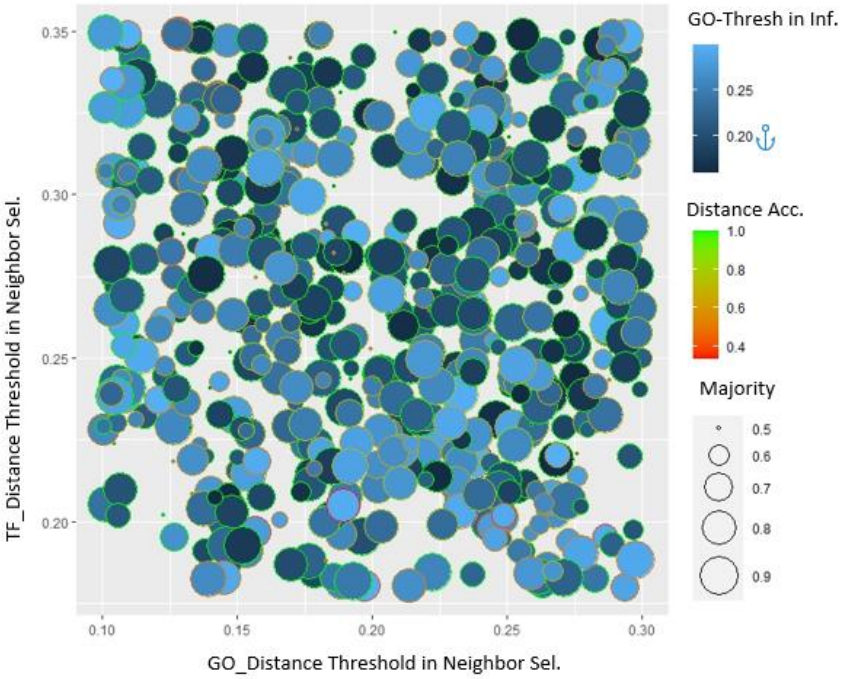

Figure 23: Parameter evaluation in 5D space. The 5 dimensions considered in this diagram are the GO-Term distance and TF-distance thresholds in neighbor selection (x and y axis), the GO-Term distance (lighter or darker blue) and the Majority (size of circle) thresholds in inference selection, and the GO-Distance Accuracy of the model (ranging from red to green as the circle perimeter).

In Figure 23 above, we can clearly spot the existence of clusters of high performing combinations of parameters. The trend that stems from the examination of Figure 23 is that smaller circles (lower majority) tend to perform quite worse than the larger ones. For this purpose, we decided to go with the majority threshold of 0.7, to select the top percentage of compounds, while attaining a number of hits. Having deciphered the majority threshold, we demoted the level of the optimization from 7 dimensions to 6.

Following that, Figures 24 and 25 may suggest that we deliberately examined the effect of the inference selection parameters separately, but in truth, we optimized the 5 parameters simultaneously, along with

the majority threshold we have seemingly already selected. The following Figures 24 and 25 were the only possible way of picturing in a comprehensible manner the relations of the parameters and the results of their selection. Figure 24 depicts the effect of the GO-Term distance in the inference selection, while Figure 25 the effect of the TF-distance.

In Figure 24 one can more clearly point out the areas of interest and determine the effect of every threshold. We see, that in the lower values of the thresholds of the neighbor selection the only way to find any neighbors is by loosening the GO-distance threshold in the inference selection. For that reason, the combined accuracy remains somewhat low in those areas, since large numbers of cold compounds are added to the inferred. At higher thresholds, the performance of the model is lowered significantly, and the number of inferred compounds is expanded.

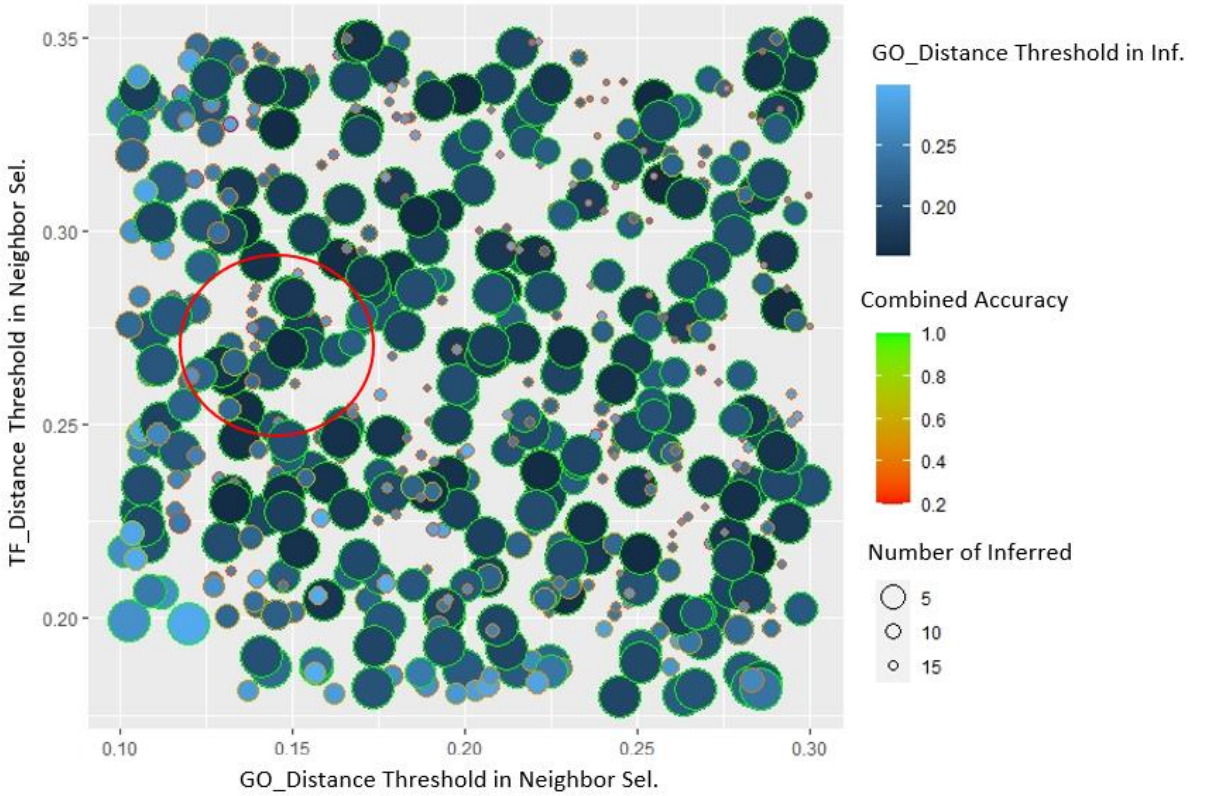

Figure 64: Parameter evaluation in 5D space. The 5 dimensions considered in this diagram are the GO-Term distance and TF-distance thresholds in neighbor selection (x and y axis), the GO-Term distance in the inference selection(lighter or darker blue), and the combined accuracy(red to green scale) and number of inferred compounds(size of circles) from the evaluation. The best performing cluster is circled with red.

The cluster that seems to have the best performance is shown in Figure 24 within the red circle. There, the values of the thresholds have combined optimally, and that area of the chart has the best overall performance, while screening several compounds.

In parallel, the same chart with the effect of the TF-distance threshold in the inference selection is shown in Figure 25 depicts how the model performance is changed based on the TF-distance threshold of the inference selection.

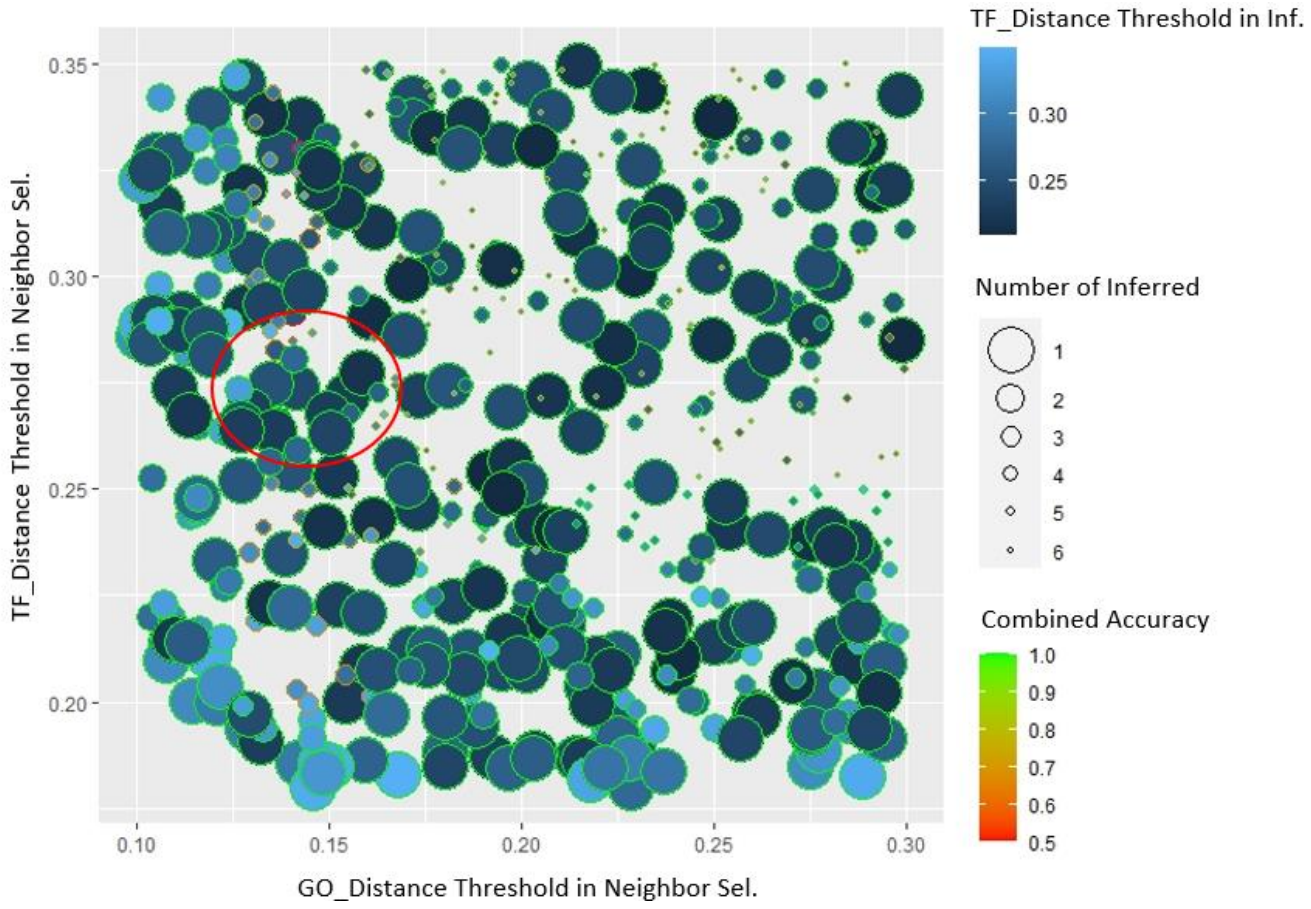

Figure 25: Parameter evaluation in 5D space. The 5 dimensions considered in this diagram are the GO-Term distance and TF-distance thresholds in neighbor selection (x and y axis), the TF-Term distance in the inference selection(lighter or darker blue), and the combined accuracy(red to green scale) and number of inferred compounds(size of circles) from the evaluation. The best performing cluster is again circled with red.

In this last figure, we focused on the effect of the threshold that we had not shown earlier, the TF- distance threshold in the inference selection. It is obvious, that in order to have a large number of inferred compounds, along with satisfactory combined accuracy, this threshold needs to be higher, at around 0.28.

All in all, by discovering the effects of different threshold values and implementing a grid optimization method, we arrived at the conclusion that the optimal values for the four thresholds are:

- 0.15 for the GO-Term distance threshold in neighbor selection

- 0.28 for the TF-Term distance threshold in neighbor selection
- 0.22 for the GO-Term distance threshold in inference selection
- 0.28 for the TF-Term distance threshold in inference selection
- 0.7 for the Majority threshold in inference selection

Of course, there was a small area around these values where the model performed at the best possible way, but we decided to use the rounded values of the thresholds. With these parameters in mind, we managed to get 100% combined accuracy for 4 cold compounds out of a total of 77.

4.4. Model Generalization

To test the performance of the model in the most challenging way, we decided to apply the model to the other cell lines. The results of this test were approximately the same. Our model predicted compounds with combined accuracy of 0.8-1 for each cell line for different compounds, at least where it could find any neighbors or cold compounds that fulfilled the parameters. By adjusting the thresholds around the selected area, we managed to boost the performance of the model over the two other cell lines and for the rest of the knockdowns in each cell line.

Moreover, the performance of the DeepSIBA model over the rest of the cell lines is pictured in Table 5:

Cell-Line	MSE	MSE @1%	Pearson's r	Precision (%)
a375	0.008	0.006	0.59	98.22
pc3	0.011	0.007	0.53	89.29
mcf7	0.012	0.007	0.56	61.03

Table 5: The performance of the DeepSIBA model over the data of the three cell lines. The MSE is the mean squared error of the predictions of the distances for the pairs of compounds in the test set, and the real values of these distances. Furthermore, the MSE 1% is an MSE form that only considered a random 1% of the test samples. Finally, Pearson's r is the correlation co-efficient used.

From the Table 5 above, one can notice that the model handles the data of the test set for each of the cell lines performing exceptionally. For most of the data, the deepSIBA maps the relations between known compounds and applies the same rules over the cold compounds with great precision. Additionally, the mean squared error (MSE) proves that we can work on every cell line. The only problem we faced during the implementation of the model over other cell lines were the absence of low distance compounds for some transcription factor knockdowns, meaning the model returned no results.

4.5. Statistical Importance

One of the most important benchmarks, when developing a model that processes data and makes suggestions based on data characteristics and imposed criteria, is the statistical importance of the output.

For this thesis, we developed a model that is able to prioritize several compounds in a data library, based on their transcription factor activity. More accurately, in cases where there is a large number of compounds that needed to be tested to determine whether they can knockdown a transcription factor, our model can be used to assess these compounds and propose which compounds should be prioritized for further testing, thus saving hours of worthless and inefficient processing.

To examine whether our model is actually worth implementing on data, instead of randomly selecting compounds to test, there needed to be proof that it would save time and resources. This comparison between random selection and screening with the use of the model happened with a use of a statistical test.

More precisely, after determining the optimal parameters for each threshold, the model came up with 4 compounds that had a combined accuracy of 100%, meaning that they all were within the GO-Term and TF-distances allowed. To show the value of the model, we compared these four compounds with a random selection of 4 compounds from the 77 cold compounds of the a375 test set. After running 10,000 random selections of 4-compound-sets, only 3 to 8 performed as well as the selected ones, meaning that the p-value of the statistical test was at 0.03-0.08%.

The fact that by choosing randomly from the test set, one has 0.03% to 0.08% chance to select as good a set of compounds as proposed through this model, is by itself a huge success. Further enhancements on the model discussed in the Future Works chapter, could possibly improve the p-value of the model and make it more robust to even bigger data.

5. Discussion and Limitations

The drug discovery process has made some great leaps forward over the past decades. With the introduction of computers, robotics and systems pharmacology, drug discovery has acquired weapons able to perform multi-dimensional analyses with the main target being selecting the active ingredients that could combat diseases. High throughput screening of compounds and their biological effect against diseases or conditions, though effective, has always proved to be a costly and time-consuming method. Instead of conducting the experiments in vitro, the new method for conducting high volume experiments is through computer simulations (in silico). The rise of computers has allowed the execution of millions of simulations within a matter of seconds and helped lower the costs and research duration for new drugs.

On this front, we developed a model to assist with the screening of compounds used to discover new drugs or find new purposes for already existing ones. By using the transcription factor activity of compounds in experiments, we essentially created a model that can propose compounds for biologists to focus on during their research. More accurately, when given a large library of compounds, our model can select the compounds with close enough biological effect distance to a target and set them as high-priority compounds to focus on.

Incorporating a realistic virtual high throughput screening allowed us to test our model's performance based on known and unknown compounds. Across different cell lines, our model was able to select sets of chemical structures whose biological effect was close enough to the transcription factor knockdown,

meaning compounds that could halt the transcription of certain important gene sets. This prioritization feature of our model could facilitate drug discovery with a simple and easy to use platform to select compounds to focus on, during the first steps of the drug development process. The model's worth is proven through its statistical significance. We found that the sets of compounds proposed by the model surpass the random selection of compounds from the same data library by a large factor. This means that selecting compounds through our platform greatly increases the probability to find compounds suitable for further testing, thus saving both resources and time.

The fact that within the model we calculate every distance needed through functions we created or through packages in R, gives credence to our model and makes it more robust. Moreover, this independence in the distance calculations allows our model to be used by anyone, in any case, on any given library of data containing chemical structures. The only important limitation we faced during our efforts was the lack of a large number of training sets and test sets, for other cell lines. The training sets for each cell line contained on average 320,000 samples of pairings of around 750 compounds, thus providing a low coverage of the chemical space. To solve this problem one needs to facilitate larger databases of compounds that spread all over the chemical space. Such an expansion of the data would result in better performance of the deepSIBA model along with a greater diversification of the output of the platform.

All in all, our model can make use of any given structures, to evaluate compounds and select ones with close biological effect to a target transcription factor knockdown. The prominent performance along with its statistical significance and its practicality render the model a useful tool for compound prioritization in the field of drug discovery.

6. Conclusions

In this study, we developed a model for the prioritization of compounds, within large libraries of chemical structures, to be used for further testing in the drug development process. Our model takes as inputs the transcription factor whose effect we aim to deactivate and uses a number of biological and logical criteria to screen compounds and select ones with close biological effect to transcription factor knockdowns. Our perception of proximity of the biological effects of compounds was heavily based on the distances at the GO-term and the transcription factor levels. By incorporating targeted thresholds and criteria, aiming to select the best possible compounds for each case, we managed to create a model that proposes sets of compounds having the desired effects, with high precision. We explored whether the criteria chosen had indeed significant impact on the compound selection process and the model's performance, thus proving their worth and cementing the screening process.

The evaluation of the performance of the model demonstrates its ability to screen large libraries of compounds and select ones with desired biological effects, with high precision, at low computational cost and within little time. Finally, the model's statistical significance manifests that the use of the platform would benefit researchers at their efforts to detect desired compounds.

7. Future Work

After the problems discussed in the sections above, it is apparent that there is room for improvement for our model. Firstly, there needs to be greater generalization of the model, by implementing it on more demanding and diverse datasets, defined by larger areas of the chemical space.

To boost the performance of the model and make it even more robust and dependable, there have been thoughts for alternate criteria for compounds' selection both in the neighbor selection and the inference selection steps. The one criterion that has been tampered with the most is the enrichment rank of the desired transcription factor over the compounds.

In such a criterion, low rank would imply that the compound has acted in an antagonizing way toward the transcription factor, since the said transcription factor was down-regulated. For such a purpose, a triplet loss model could be developed, incorporating machine learning and the given compounds to extract useful results for the rank of the compounds.

Finally, validating the results of our model with further testing could cement the model's worth and give a metric for measuring the biological significance of the output of the platform.

8. References

[1] H.P., Rang; M.M, Dale; J.M., Ritter; R.J., Flower; G., Henderson (2011). "What is Pharmacology". Rang & Dale's pharmacology (7th ed.). Edinburgh: Churchill Livingstone. p. 1. ISBN 978-0-7020-3471-8.

[2] Inglese J and Auld DS. (2009) Application of High Throughput Screening (HTS) Techniques: Applications in Chemical Biology in Wiley Encyclopedia of Chemical Biology (Wiley & Sons, Inc., Hoboken, NJ) Vol 2, pp 260–274.

[3] "The drug development process: Step 1: Discovery and development". US Food and Drug Administration. 4 January 2018. Retrieved 18 December 2019.

[4] Odilia Osakwe, "The Significance of Discovery Screening and Structure Optimization Studies." Social Aspects of Drug Discovery, Development and Commercialization (2016).

[5] Sliwoski, Gregory, et al. "Computational methods in drug discovery." Pharmacological reviews 66.1 (2014): 334-395.

[6] Sirci, Francesco, et al. "Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses." NPJ systems biology and applications 3.1 (2017): 25-32.

[7]Tavassoly, Iman; Goldfarb, Joseph; Iyengar, Ravi (2018-10-04). "Systems biology primer: the basic methods and approaches". Essays in Biochemistry. 62 (4): 487–500.

[8] Voit, Eberhard (2012). A First Course in Systems Biology. Garland Science. ISBN 9780815344674.

[9] Villa A., Sonis S. (2020). "Translational Systems Medicine and Oral Disease". pp. 9-16. ISBN

978-0-12-813762-8

[10] Elston RC, Satagopan JM, Sun S (2012). "Genetic terminology". Statistical Human Genetics. Methods in Molecular Biology. 850. Humana Press. pp. 1–9. ISBN 978-1-61779-554-1.

[11] Crick F (August 1970). "Central dogma of molecular biology". Nature. 227(5258): 561–3. Bibcode:1970Natur.227.561C.

[12] St. Petersburg Institute of Gerontology. "Peptide Regulation of Ageing." Peptides Store, St. Petersburg Institute of Gerontology, 2017, www.peptidesstore.com/blogs/articles/9236639-peptide-regulation-of-ageing.

[13] Wang, Zhong; Gerstein, Mark; Snyder, Michael (January 2009). "RNA-Seq: a revolutionary tool for transcriptomics". Nature Reviews Genetics. 10(1): 57–63.

[14] Latchman DS (December 1997). "Transcription factors: an overview". The International Journal of Biochemistry & Cell Biology. 29 (12): 1305–12.

[15] Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (June 2004). "Structure and evolution of transcriptional regulatory networks"(PDF). Current Opinion in Structural Biology. 14 (3): 283–91.

[16] Learning, Lumen. “Eukaryotic Transcription Gene Regulation.” Module 11: Gene Expression, Lumen, courses.lumenlearning.com/wm-biology1/chapter/reading-eukaryotic-transcription-gene-regulation/.

[17] *Vallance P, Smart TG (January 2006).* "The future of pharmacology"*. British Journal of Pharmacology. 147 Suppl 1 (S1): S304–7.*

[18] Musa, Aliyu, et.al. “A Review of Connectivity Map and Computational Approaches in Pharmacogenomics.” Briefings in Bioinformatics, Oxford University Press, 1 May 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC5952941/.

[19] Michnick, Stephen W. “The Connectivity Map.” Nature News, Nature Publishing Group, www.nature.com/articles/nchembio1206-663.

[20] Hodgkin, A. L.; Huxley, A. F. (1952-04-28). "Currents carried by sodium and potassium ions through the membrane of the giant axon of Loligo". The Journal of Physiology. 116: 449–472.

[21] Schmidhuber, J. (2015). "Deep Learning in Neural Networks: An Overview". Neural Networks. 61: 85–117. arXiv:1404.7828.

[22] The Gene Ontology Consortium (January 2008). "The Gene Ontology project in 2008". Nucleic Acids Research. 36 (Database issue): D440–4. doi:10.1093/nar/gkm883. PMC 2238979. PMID 17984083.

[23] Sergushichev, Alexey. "An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation." BioRxiv (2016): 060012.

[24] *Summerton JE (2007). "Morpholino, siRNA, and S-DNA compared: impact of structure and mechanism of action on off-target effects and sequence specificity". Current Topics in Medicinal Chemistry. 7 (7): 651–60.* doi*:*10.2174/156802607780487740*.* PMID 17430206*.* S2CID 12241724*.*

[25] Subramanian A, et al. A Next Generation Connectivity Map: L1000 Platform And The First 1,000,000 Profiles. Cell. 2017/12/1. 171(6):1437–1452.

[26] Alexa, Adrian, and Jörg Rahnenführer. "Gene set enrichment analysis with topGO." Bioconductor Improv 27 (2009).

[27] Sirci, Francesco, et al. "Comparing structural and transcriptional drug networks reveals signatures of drug activity and toxicity in transcriptional responses." NPJ systems biology and applications 3.1 (2017): 1-12.

[28] Fotis, Chris & Meimetis, N. & Sardis, A. & Alexopoulos, Leonidas. (2020). DeepSIBA: Chemical Structure-based Inference of Biological Alterations.