UNIVERSIDADE D
COIMBRA

Luis Henrique Magalhães Ramos Torres

EXPLORING A SIAMESE NEURAL NETWORK ARCHITECTURE FOR ONE-SHOT DRUG DISCOVERY

Projeto de dissertação no âmbito do Mestrado em Engenharia Biomédica orientado pelo Professor Doutor Joel Arrais e pela Professora Doutora Bernardete Ribeiro e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Setembro de 2020

UNIVERSIDADE D COIMBRA
FACULDADE DE CIÊNCIAS E TECNOLOGIA

Luis Henrique Magalhães Ramos Torres

Exploring a Siamese Neural Network Architecture for One-Shot Drug Discovery

Thesis submitted to the
University of Coimbra for the degree of
Master in Biomedical Engineering

Supervisors:
Prof. Dr. Joel P. Arrais
Prof. Dr. Bernardete Ribeiro

Coimbra, 2020

This work was developed in collaboration with:

Center for Informatics and Systems of the University of Coimbra

BSIM Therapeutics

Acknowledgments

Antes de mais queria agradecer ter tido a oportunidade de trabalhar com o Professor Doutor Joel Arrais, por ser capaz de motivar nos seus alunos e orientandos a capacidade de questionar, investigar e desenvolver o espírito crítico fundamental à formação científica e pessoal. Este contributo revelou-se muito importante e remonta às primeiras aulas da unidade curricular de Introdução à Bioinformática onde sempre cativou os seus alunos na procura das suas próprias respostas e soluções para os problemas e desafios colocados. Desta forma, agradeço a aposta nas minhas capacidades e o voto de confiança colocado nas minhas decisões ao longo de todo o projeto, pela liberdade na exploração das minhas ideias e pela ajuda imprescindível sem a qual a concretização do presente trabalho não seria possível. A aprendizagem sob a sua orientação permitiu-me crescer a nivel pessoal e intelectual e participar ativamente na comunicação e produção científica. Desta forma, quero deixar um agradecimento especial ao professor, pela oportunidade concedida, pela disponibilidade na resolução dos múltiplos obstáculos que surgiram e pelo seu papel preponderante em todas as etapas do projeto.

A conceção de uma abordagem inovadora para a resolução dos desafios identificados surgiu motivada pela contribuição determinante da Professora Doutora Bernardete Ribeiro. Assim, a inspiração para o desenvolvimento do presente trabalho surge a partir do seu valioso contributo que se revelou imprescindível e serviu de mote para o presente estudo. Desta forma, queria deixar os meus mais sinceros agradecimentos à Professora Doutora Bernardete pelos valiosos testemunhos, orientações e acompanhamento permanente que serviram de alicerce a todo o projeto, desde o primeiro momento. Método, capacidade de trabalho, espírito crítico e curiosidade são alguns dos valores estruturais que transmite aos seus alunos e orientandos e com os quais tive o privilégio de contactar, em particular na unidade curricular de Reconhecimento de Padrões e, presentemente no desenvolvimento do projeto final de mestrado. O seu acompanhamento e orientação dedicadas foram essenciais à minha aprendizagem no decurso de todo o projeto e servem de inspiração a todo o trabalho

desenvolvido e a futuras contribuições científicas. Desta forma, deixo um agradecimento especial à Professora Doutora Bernardete cujo contributo foi decisivo para a definição da ideia, estrutura e abordagens conduzidas e que, indubitavelmente, acrescenta rigor e robustez a todo o trabalho desenvolvido.

Queria agradecer ao meu amigo e colega de trabalho, Nelson Monteiro, pelo acompanhamento constante e dedicação demonstradas em todas as etapas do projeto e, cujo conhecimento e espírito crítico estão indelevelmente marcados no presente trabalho e contribuições associadas. A sua capacidade de trabalho, dedicação permanente e comprometimento com a produção científica são valores que, inequivocamente, transparecem no seu trabalho e que considero como um exemplo e referência a seguir. Para além do tempo e forças dispendidas em incontáveis sugestões, comentários construtivos e revisões detalhadas agradeço profundamente todos os momentos de discussão, debate e reflexão, assim como os também valiosos momentos de descontração que amenizaram todos os de maior ansiedade ou apreensão. Desta forma, agradeço o seu contributo descomprometido e dedicado, cujas impreteríveis apreciações e juízo críticos contribuíram para a validação de ideias e abordagens conduzidas.

Um agradecimento muito especial à minha família, a principal responsável por estar aqui hoje, tornando realidade o culminar de um longo percurso repleto de dificuldades, obstáculos e desafios, mas também de muitas alegrias e triunfos. Obrigado por me acompanharem sempre desde o primeiro ao último instante, por me apoiarem em todos os momentos de maior dificuldade e me acarinharem nas até então, pequenas, mas importantes vitórias. Obrigado porque sempre acreditaram em mim, por toda a força e por todo o esforço e sacrifícios que fazem todos os dias para que possa alcançar os meus objetivos e para que persiga os meus maiores sonhos. Obrigado por nunca me deixarem desistir e espero que algum dia vos consiga retribuir todo o vosso apoio sem o qual nada disto seria possível. Obrigado ao meu pai, porque nunca deixaste de acreditar em mim e nas minhas capacidades e, por comemorares comigo todas as vitórias e me tranquilizares nas derrotas. Obrigado à minha mãe, porque estás lá sempre quando mais preciso e me apoias em todos os momentos na esperança de ver os meus sonhos concretizados. Obrigado ao meu irmão gémeo, porque somos inseparáveis, por acompanhares sempre todo o meu percurso e por partilharmos todos os momentos mais importantes. Obrigado à minha irmã, por todas as brincadeiras, por todos os momentos de descontração e por acreditares em mim, sempre. Obrigado, porque sem vocês o meu percurso não faria sentido e é a vocês que dedico este trabalho na busca de que vos deixe orgulhosos.

Um último agradecimento a todos os amigos e "família" de Coimbra por estarem sempre lá e me apoiarem nos momentos mais importantes. Obrigado por acreditarem em mim e seguirem o meu percurso sempre, desde o princípio. Obrigado por todos os momentos inesquecíveis, por todas as conversas, por todas as gargalhadas, por todos os jantares, por todas as noitadas e por todas as histórias incríveis, memórias únicas e momentos inolvidáveis que levo comigo para sempre.

Financing

This research has been funded by the Portuguese Research Agency FCT, through D4 - Deep Drug Discovery and Deployment (CENTRO-01-0145-FEDER-029266).

"Look up at the stars and not down at your feet. Try to make sense of what you see, and wonder about what makes the universe exist. Be curious!"

Stephen Hawking

Resumo

As redes neuronais profundas oferecem um elevado poder preditivo ao ser capazes de inferir as propriedades farmacológicas e atividades biológicas de pequenas moléculas em aplicações para a descoberta de fármacos. No entanto, a quantidade de informação supervisionada disponível para treino é escassa e o passo de *lead-optimization* apresenta um problema de *low-data*, dificultando a descoberta de novos compostos com a atividade terapêutica pretendida e dos respetivos análogos farmacológicos. Um requisito fundamental é a necessidade de uma grande quantidade de exemplos de treino por classe, o que nem sempre é possível em aplicações para descoberta de fármacos. Estes desafios invalidam o uso de instâncias cujas classes são desconhecidas no treino ou de dados biológicos onde o número de classes é elevado e oscila dinamicamente.

O grande objetivo deste estudo é otimizar a descoberta de novos compostos partindo de um conjunto reduzido de moléculas candidatas. Assim, propomos uma arquitetura de redes neuronais paralelas implementando uma estratégia de *one-shot learning*, baseada num conjunto de redes neuronais convolucionais (CNNs) capazes de aprender a partir de um dado *score* de semelhança entre duas moléculas devolvido por uma dada função de similaridade.

Aplicando uma estratégia de *one-shot learning*, apenas necessitamos de um reduzido conjunto de instâncias por classe para treino e de um pequeno conjunto de dados e recursos computacionais para a construção de um modelo eficaz na previsão. Os resultados obtidos demonstram que o uso de um conjunto de redes neuronais convolucionais paralelas implementando uma estratégia de *one-shot learning* conduz à obtenção de desempenhos superiores na previsão de novos compostos comparando com os modelos *state-of-the-art*. Assim, o modelo proposto permite prever corretamente e com elevada eficácia, novos compostos e respetivos análogos farmacológicos, considerando a escassez de dados biológicos disponíveis para aplicações de descoberta e desenvolvimento de fármacos.

Palavras-Chave: Descoberta de Fármacos, Aprendizagem Profunda, *One-Shot Learning*, Redes Neuronais Paralelas, Rede Neuronal Convolucional.

Abstract

Deep neural networks offer a great predictive power when inferring the pharmacological properties and biological activities of small molecules in drug discovery applications. However, in the traditional drug discovery process, where supervised data is scarce, the lead-optimization step is a low-data problem, making it difficult to find molecules with the desired therapeutic activity and obtain accurate predictions for novel compounds and their pharmacological analogs. One major requirement to ensure the validity of the obtained neural network models is the need for a large number of training examples per class, which is not always feasible in drug discovery applications. This invalidates the use of instances whose classes were not considered in the training phase or in data where the number of classes is high and oscillates dynamically.

The main objective of the study is to optimize the discovery of novel compounds based on a reduced set of candidate drugs. We propose a Siamese neural network architecture for one-shot classification, based on Convolutional Neural Networks (CNNs), that learns from a similarity score between two input molecules according to a given similarity function.

Using a one-shot learning strategy, few instances per class are needed for training, and a small amount of data and computational resources are required to build an accurate model. The results achieved demonstrate that using a Siamese Deep Neural Network for one-shot classification leads to overall improved performance when compared to other state-of-the-art models. The proposed one-shot Siamese neural network architecture provides an accurate and reliable prediction of novel compounds considering the lack of biological data available for drug discovery tasks.

Keywords: Drug Discovery, Deep Learning, One-Shot Learning, Siamese Neural Network, Convolutional Neural Network.

Contents

List of Tables

List of Figures

Abbreviations

Adam Adaptive Moment Estimation

ADMET Absorption, Distribution, Metabolism, Excretion and Toxicity

ANN Artificial Neural Network

AUC Area Under the Curve

CADD Computer-Aided Drug Discovery

cAMP Cyclic AMP

CNN Convolution Neural Network

DMPK Drug Metabolism and Pharmacokinetic

EMA European Medicines Agency

FC Fully-Connected

FDA Food and Drug Administration

GPCR G-Protein-Coupled Receptors

GTP Guanosine Triphosphate

HBPL Hierarchical Bayesian Program Learning

HTS High Throughput Screening

KNN K-Nearest Neighbour

LBDD Ligand-Based Drug Discovery

LSTM Long Short-Term Memory

MANN Memory Augmented Neural Networks

MD Molecular Dynamic simulation

MHI Motion-History-Image

MLP Multi-Layer Perceptron

MSE Mean Squared Error

MUV Maximum Unbiased Validation

NLP Natural Language Processing

NME New Molecular Entity

NMR Nuclear Magnetic Resonance

QSAR Quantitative Structure-Activity Relationship

RBF Radial Basis Function

RBG Red-Blue-Green

ReLU Rectified Linear Unit

RF Random Forest

SBDD Structure-Based Drug Discovery

SMILES Simplified Molecular Input Line Entry System

SGD Stochastic Gradient Descent

SIDER Side Effect Resource

SVM Support Vector Machine

UGRNN Undirected Graph Recursive Neural Network

Nomenclature

Mathematical Concepts

α	Level of statistical significance
$argmax$	Argmax function
$H0$	Null hipothesis
$L2$	Euclidean distance
max	Max function
p	$p - value$ for statistical significance
$sigmoid$	Sigmoid function
T_c	Tanimoto coefficient
$tanh$	Hyperbolic tangent function

Parameters

α	Learning rate
β_1	Exponential decay rate for the first moment estimates
β_2	Exponential decay rate for the second-moment estimates
γ	Kernel coefficient
ϕ	Mapping function
C	Regularization parameter
K	Number of neighbours for KNN
N	Number of pairs per one-shot task
$score$	Similarity score

1

Introduction

1.1 Context

The Role of Drugs: Drug-Target Interaction

Ligands are molecules capable of binding to a specific receptor or enzyme and trigger an appropriate physiological response within the cell. Ligands bind to very specific regions of the receptor macromolecules called binding sites. The interaction with a binding site can be reversible and may activate or inactivate the receptor, increasing or decreasing its biological activity [1].

Drugs are ligands capable of binding specifically to these receptors, activating or inactivating them and, consequently, enhancing or inhibiting a given cellular function. Generally, drugs are highly selective by binding only to a specific receptor and its multiple subtypes. They generally bind to receptors but can also form complexes with specific enzymes [2].

The binding site in the receptor macromolecules may be different for different endogenous agonists in the production of a physiological response. Agonists are analog molecules which bind to the receptor and mimic the interaction of a natural ligand. On the other hand, antagonists bind without triggering the effect of the natural ligand leading to the blocking of the receptors activity [3]. They can interact reversibly or irreversibly, competing or not with an agonist [4]. Thus, drugs can act as agonists or antagonists, activating or inactivating the receptor in the signal transduction pathway that leads to the final response in organs and tissues.

The effectiveness of a given drug in producing the desired physiological response is determined by aspects that directly influence the activity of the drug-target complex. Thus, the probability that a given molecule binds to a specific receptor at a given instant (affinity), directly influences the formation of this complex. The cellular response to the stimulus is also determined by the intrinsic activity, that is, the

degree to which the drug forms the complex and triggers the desired effect. Both of these factors are determined by the molecule chemical structure which models the efficacy of the physiological response to the signal [5].

The mechanisms of drug-target interaction are a key aspect in the prediction of the biological function of the various classes of therapeutic targets and the consequent response on organs and tissues. The signal transduction mechanisms triggered by the binding of drugs to specific receptors and enzymes are highly sensitive and specific to this interaction pair.

The specificity of the interactions triggered by drug-target binding result from the affinity between the signal and its specific receptor in the target cells and tissues and is mediated by weak non-covalent forces identical to those that regulate the enzyme-substrate or antibody-antigen binding [6]. Complex organisms have an increased level of specificity in their transduction pathways due to the presence of specific receptors for a given signal in certain cell types, which determines the effectiveness in the cellular response to a given receptor-specific signal.

The transduction pathways activated by drug-target binding are extraordinarily sensitive to the stimulus that results from this interaction. This is due to the high molecular affinity of the receptors for a given specific signal, the cooperation of the receptors in the face of low concentration variations of the ligands and the amplification in the signal transduction through enzymatic cascade reactions [7].

The activation of a signal transduction pathway arises primarily from the interaction of a signal with its specific target or receptor. The signal interacts with its specific receptor which enables the production of a second signal. The signal triggers a transduction pathway that leads to the desired effect through a change in the metabolic activity of the cell by modifying the biological activity of a certain protein or enzyme [8].

Most drugs and their pharmacological analogs act in many transduction pathways mediated by proteins responsible for the physiological response to an intracellular or extracellular signal. These signals are divided into different types according to the stimulus and transduction mechanism: G protein-coupled receptors (GPCRs), enzymatic receptors, ionic channels and nuclear receptors.

The GPCRs are capable of indirectly activating a signal transduction pathway through the activity of GTP-binding proteins. These enzymes synthesize a secondary intracellular messenger, typically cAMP (cyclic AMP). The transduction mechanism is defined by 3 essential components: a transmembrane receptor, an en-

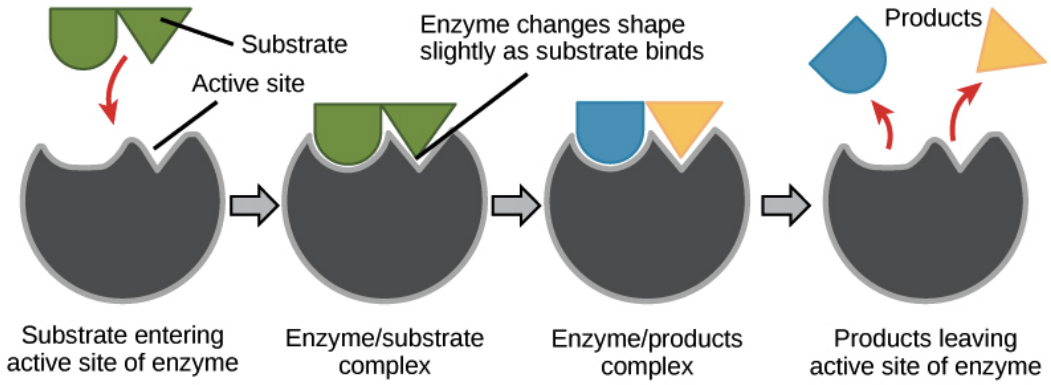


Figure 1.1: Enzime-substrate interaction as guide for the ligand-receptor binding. Image from `"www.khanacademy.org/science/ap-biology/cellular-energetics/enzyme-structure-and-catalysis/a/enzymes-and-the-active-site"`.

zyme that synthesizes a secondary messenger and finally a GTP-binding protein that dissociates itself from the receptor and activates an enzyme that triggers a cellular response to the stimulus. A model of this type of receptor is the beta-adrenergic receptor [9, 10].

Enzyme receptors are receptors that have an extracellular binding domain at the membrane surface and an intracellular active site. In the case of a tyrosine-kinase receptor, the intracellular active site phosphorylates tyrosine residues triggering a reactive cascade within the cell (a set of consecutive phosphorylations and dephosphorylations) [11].

Ion channels are permeable channels arranged transversely along the cell membrane. Their opening occurs through conformational changes mediated by the interaction of specific intracellular or extracellular ligands to the membrane surface binding sites or through changes in transmembrane potential [12].

Nuclear receptors interact with specific ligands capable of triggering a change in gene expression. The change in gene expression translates into a change in the concentration of a given protein or enzyme within the cell and hence a change in cell metabolism [13].

The discovery of drugs that interact with specific enzymes represents a major challenge in drug discovery tasks. The importance of the activation of transduction pathways coupled with enzyme receptors lies in their determining role in all cellular mechanisms. Conformational changes and modification of its biophysical and biochemical profile trigger important effects in the regulation of cell homeostasis,

signalling mechanisms, energy production and storage and also in the elimination of cytotoxic agents.

The high sensitivity and specificity of the drug-receptor interaction hinders the modelling of the physiological response to a stimulus. Thus, it is necessary to develop compound prediction models able to predict new classes of molecular structures leading to the discovery of novel compounds with improved therapeutic effects and increased biological activity.

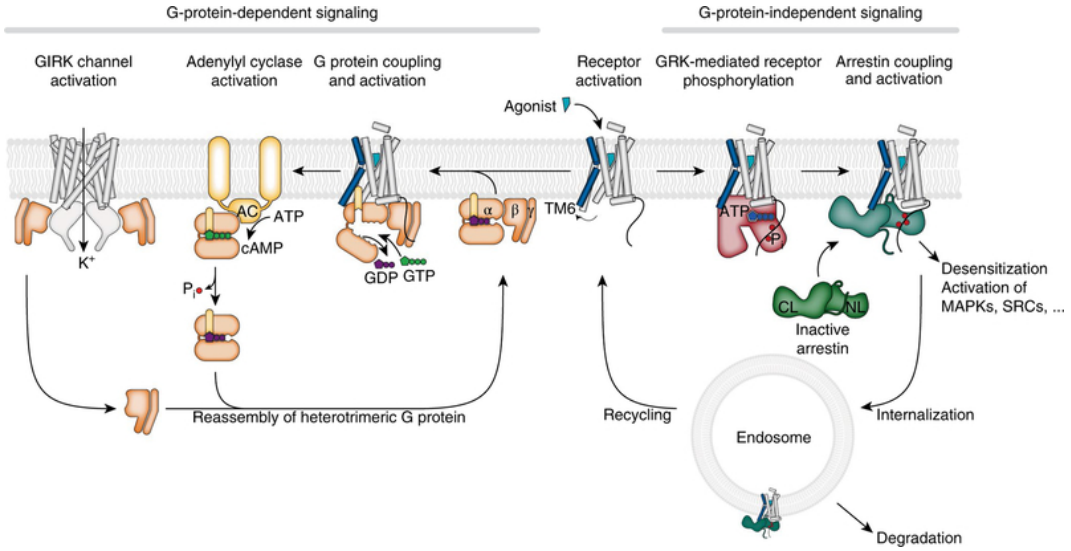


Figure 1.2: G-coupled receptor-signal transduction. Image from "Structure and dynamics of GPCR signaling complexes" [9].

Pharmacodynamics

Pharmacodynamics consists of the set of physiological and biochemical mechanisms triggered by a given drug in organs and tissues, involving the binding to its specific receptor, post-binding effects and chemical interactions [14].

The receptors are macromolecular structures involved in intracellular and intercellular signal transduction mechanisms. The ligand-receptor complex or drug-target complex directly or indirectly regulates the cell metabolism, ion flow through the cell membranes, protein phosphorylations, gene expression and cell enzymatic activity [15].

The efficacy of the response to a stimulus depends on multiple factors, including the degree of affinity of the drug-target complex, the intrinsic activity, the life span of the complex (residence time) and also the conformational changes that occur within an interaction. Longer residence times may explain a longer pharmacological effect and increased toxicity against a pharmacological target [16]. Many of these factors are determined by the molecular structure of the compounds. The ability to interact with a receptor still depends on factors external to the cell and on intracellular regulation mechanisms. Ageing, mutations, and the interaction with other drugs and molecules may increase (upregulate) or decrease (downregulate) the affinity of a compound to its specific receptor.

Ultimately, the desired physiological response is mediated by multiple drug-targets complexes, and several steps can be interposed between the first ligand-receptor interaction and the final effect on organs and tissues. Multiple molecules may intervene in the production of the same pharmacological response which motivates the change from a 'one drug, one target' paradigm to a 'multi-drug, multi-target' model in the prediction of novel compounds in drug discovery tasks [17, 18].

The non-specific binding of a drug with a molecular binding site not designated as a receptor may lead to the inhibition of the response by preventing the formation of the drug-receptor complex, inactivating the drugs effect. Given the complexity of the drug-receptor interactions it is time, resource-consuming and extremely expensive to find new candidate molecules through traditional experimental procedures. Therefore, it is crucial to develop prediction models through *in silico* methods that start from a given set of molecular structures and allow the identification of novel compounds that interact with specific target enzymes [19].

Agonist or antagonist ligands are capable of producing or eliminating the final response to a signal. The discovery of analog agonists or antagonists for given pharma-

cological target may potentiate the discovery of drugs with less off-target activities, optimal therapeutic effects and improved physicochemical profile.

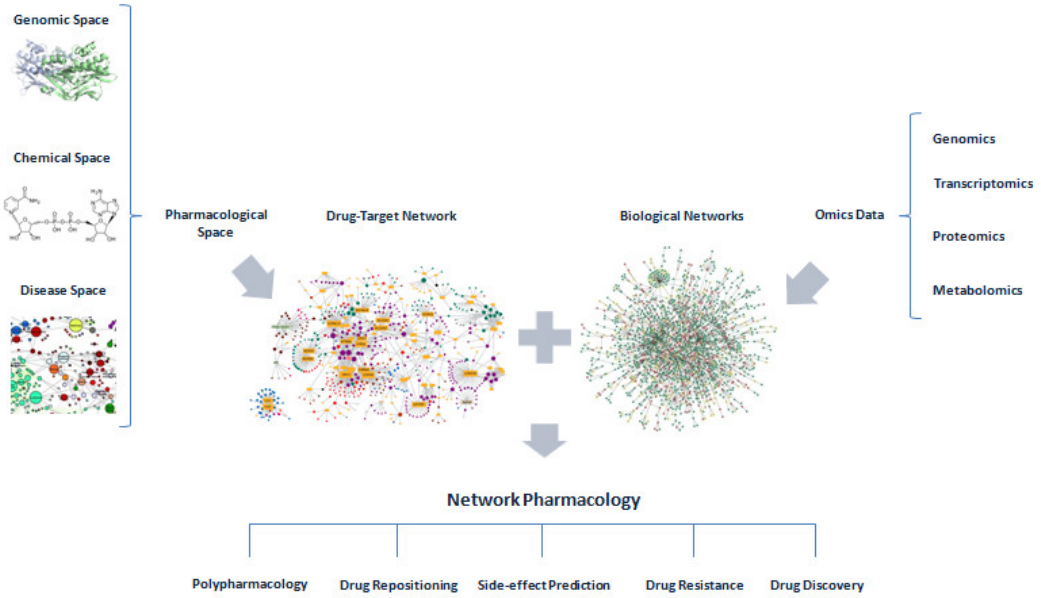


Figure 1.3: Polipharmacology using the pharmacological space following a "multi-target, multi-drug" paradigm. Image from "Drug-target and disease networks: polypharmacology in the post-genomic era" [19].

Drug Discovery and Development: Lead Optimization Step

The drug discovery process is a complex process ranging from the discovery of a lead molecule, approval through clinical trials and post-approval monitoring. The drug discovery process consists of the following steps: target discovery, lead discovery, lead optimization, pre-clinical trials, clinical trials and regulatory approval. The first 2 steps include the discovery of a new molecule involved in a given disease and a drug capable of interacting with the identified target (hit molecule) through a compound library screening against the identified target molecule (HTS - High Throughput Screening). The lead optimization step consists in changing the molecular properties of the hit to increase the effectiveness and safety of the therapy. Once a molecule has been identified (lead molecule), pre-clinical trials (*in vivo* tests) are performed by evaluating the conditions (in *vitro* and *in situ*) of adsorption , distribution, metabolism, excretion and toxicity (ADMET). The discovered lead molecule then passes through a set of clinical trials to be approved by a regulatory agency (by the FDA or EMA), evaluating the efficacy and safety of the therapy through 3 stages of human tests [20].

The lead optimization step follows the discovery of a set of hit molecules able to interact with the identified target. The objetive is to improve the effectiveness of the hits and identify a pharmacological profile consistent with their biological activity and desired therapeutic effect. Therefore, this hit-to-lead process aims to optimize the activity of the hit for a specific target and achieve the required safety conditions, affinity, ADMET and pharmacokinetic properties. The lead optimization step is crucial in the discovery and development of novel compounds with improved pharmacological activity on the identified target. This step intervenes in the adjustment of the molecular structure of the compound intending to introduce modifications that enhance its pharmacological activity. In this step, we seek to maintain the desired properties of the main components of the molecule, preventing any structural deviation that might compromise their biological activity while eliminating deficiencies in the compound structure. Typically it involves the application of multiple compound screenings (*in vivo* and *in vitro*) to characterize the metabolic and pharmacokinetic properties (DMPK) of multiple compounds and to find structural analogs with optimal pharmacological and pharmacokinetic activities for a given target molecule [21].

Thus, there are multiple requirements to be met for a drug to interact with its specific target and trigger the physiological response leading to the desired effect on organs and tissues. These fundamental properties include potency, bioavailability, safety,

duration and pharmaceutical acceptability. The intrinsic capacity of the drug to produce a given response, the capacity to overcome multiple physiological barriers, the time that remains in circulation, the ratio between the interaction with the target and the unspecific interactions with other targets are important properties, when predicting the appropriate compound structure for the desired therapeutic effect [22]. The hit-to-lead process also prevents the occurence of off-target binding with non-specific targets optimizing the interaction with the identified target molecule.

Therefore, the lead optimization process leads to the discovery of a molecule which produces the desired physiological response, as well as presents the optimal DMPK properties and a safety profile compatible with its therapeutic effect. The main objective is to identify analog compounds with optimal therapeutic effects, less toxicity, greater pharmacological activity, reduced risks for the organism, and better conditions of solubility and selectivity.

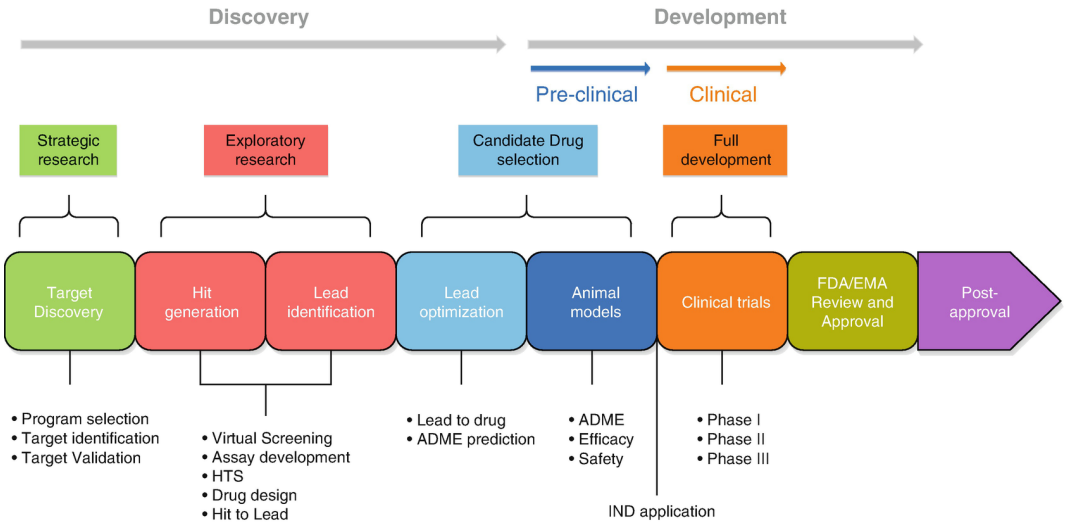


Figure 1.4: Stages of drug discovery and development. Image from "Medicinal Biotechnology for Disease Modeling, Clinical Therapy, and Drug Discovery and Development" [23].

1.2 Motivation

The hit-to-lead stage of drug discovery is an extremely long and expensive process. It is common for the initial drug screening to generate a set of at least tens of thousands of potential lead compounds. The lead molecules undergo a new screening that reduces the initial compound library to a small set of candidate molecules.

However, only a small fraction of the drug discovery projects survive and culminate in the identification of a small set of lead molecules ready to be tested and later distributed and commercialized.

Among the most common failures in the identification of lead molecules in drug discovery processes, there are: the inability to constitute a viable, consistent and reliable test; failure to obtain potential lead compounds through HTS screening; toxicity of molecules obtained *in vivo* or *in vitro*; abnormal behaviour in the target tissues; failure to obtain a good metabolic and pharmacodynamic profile (DMPK) that enhances their therapeutic function; failure to cross the blood-brain barrier in compounds intended to interact with elements of the central nervous system [24].

On the other hand, once identified a lead molecule the possibility of attrition with other compounds and the occurrence of side effects in the clinical testing phase is quite high, so that only 1 in every 10 candidate molecules is configured as a compound that can be effectively administered to a patient.

Across all areas of therapeutic research for drug discovery and development, approval by regulatory agencies is a process that can extend over more than 12 years and cost billions. The development of new molecular entities (NMEs), small molecular structures, is an extremely expensive process with a low probability of success considering all the constraints associated with the feasibility of their administration for human consumption [25].

Due to the costs and resources required for the experimental execution of the drug discovery and development process, *in silico* molecule prediction seems to be an efficient and notoriously more economical approach. This strategy saves resources but also provides important information to support the tests conducted, experimentally, in the laboratory [26].

However, the biological data available to find new analog molecules in the lead-optimization step of drug discovery tasks are often limited in size and are very expensive to obtain due to the scarcity of experimental data for drug discovery applications [27]. These datasets are the result of clinical trials, which might not

be repeatable due to ethical reasons. Therefore, it is extremely difficult to obtain accurate and reliable predictions for novel compounds given the lack of supervised biological data available to optimize drug discovery.

1.3 Objectives

The application of deep neural networks is an important asset to significantly increase the predictive power when inferring the properties and activities of small-molecules and those of their pharmacological analogs. However, in the traditional drug discovery process, the lead-optimization step is a low-data problem, which makes it difficult to find analog molecules with the desired therapeutic activity.

Ultimately, we aim to validate how a one-shot-based Siamese Neural Network allows us to outperform the state-of-the-art models in the accurate and reliable prediction of pharmacological analogs that could lead to the discovery of promising lead molecules. Therefore, there are four main objectives to fulfill:

1. Use of deep learning methods to tackle the challenges identified in the lead optimization step of drug discovery tasks;
2. Face the low-data problem of the lead optimization step of drug discovery tasks using an one-shot learning strategy;
3. Use of a Siamese Neural Network architecture to achieve strong results on one-shot classification tasks based on a similarity score between two input molecules;
4. Evaluate the performance of a one-shot Siamese neural network and compare it with other machine learning and deep learning approaches;
5. Optimization of the prediction of drug analogs with increased biological activity based on a reduced set of candidate molecules.

1.4 Workflow

The reduced amount of labelled data available for drug discovery makes it necessary to develop innovative *in silico* approaches capable of predicting a set of candidate leads with optimized biological activity for an identified pharmacological target.

Thus, by adopting a drug repositioning approach, it is possible to develop models capable of predicting analog compounds of pre-existing drugs. The identified lead molecules must be able to produce the intended therapeutic effect in different types of pharmacological targets involved in a given signal transduction pathway.

It is important to combine this strategy with high-performance predictive models to speed up the hit-to-lead process and to find the set of lead molecules that produces the most appropriate physiological response for the identified target molecules. Unlike traditional drug discovery methodologies, drug-repositioning iteratively optimizes pre-existing drugs resulting in a set of candidate lead molecules with higher potency, safety and efficacy for multiple pharmacological targets.

The application of *in silico* methods, like deep neural networks, enhances the learning of models capable of inferring the properties of molecular structures through the identification of patterns that best describe the optimal molecular structure for the interaction with the identified target molecule.

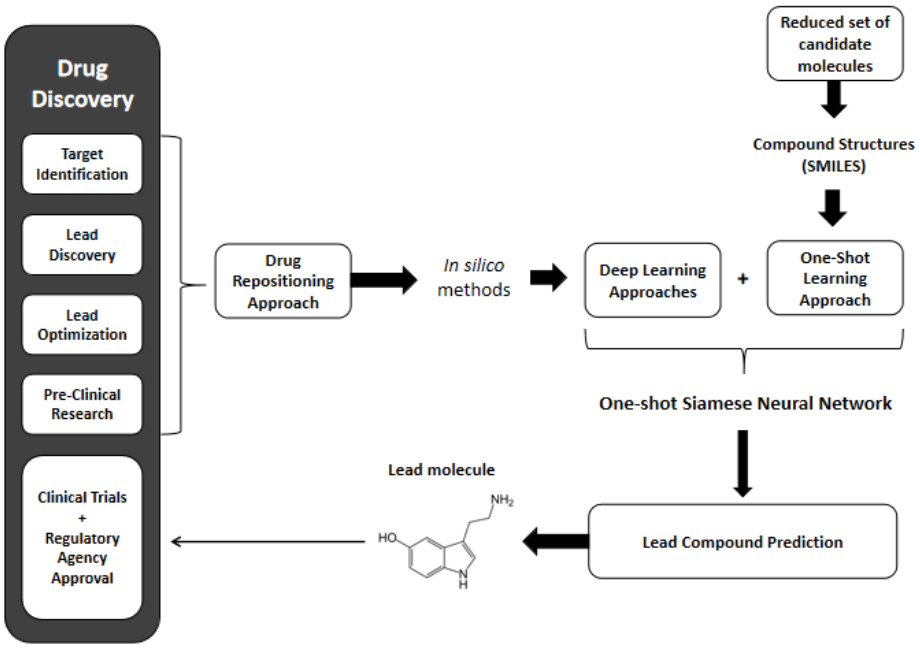


Figure 1.5: *In silico* methods in the optimization of drug discovery using a one-shot deep learning approach.

1.5 Research Contributions

The work developed during this thesis resulted in the following contributions:

Papers

Torres, L.H.M. and Nelson R. C. Monteiro and Oliveira, J.L. and Arrais, J.P. and Ribeiro, B. , "Exploring a Siamese Neural Network Architecture for One-Shot Drug Discovery". BIBE2020, 20th IEEE International Conference on Bioinformatics and Bioengineering. (Submitted on 13th July 2020 and accepted as a short paper).

Posters

Torres, L.H.M. and Nelson R. C. Monteiro and Ribeiro, B. and Arrais, J.P. , "Siamese Neural Networks for One-Shot Drug Discovery", in Bioinformatics Open Days (BOD) 2020, 2020 (Poster Presentation on February 2020).

1.6 Document Structure

The document is divided into 7 different chapters. The present chapter, Introduction, provides a brief introduction to the biochemical phenomena and pharmacodynamic principles of drug-receptor interactions as well as, the motivations, objectives and the approaches adopted in this study. The second chapter, State of the Art, presents a set of *in silico* methodologies and previous works for drug discovery and development applications, describing a wide range of computational approaches, including a one-shot learning strategy. The Chapter 3, Data Preparation, identifies a set of methods used in data collection, data processing and data grouping required for the implementation of the proposed model. Chapter 4, Model, presents the deep neural network used in the development of the proposed model and how the training and hyperparameter optimization was performed. Chapter 5, Experimental Setup, describes in more detail the type of architecture used in the proposed deep learning model, the set of hyperparameters used, the one-shot learning methodology implemented and the models used for the comparison with the proposed model. Chapter 6, Results and Discussion, presents the results obtained by different models providing a statistical significance analysis and discussion of the results regarding the performance, advantages and disadvantages of different model approaches. The last chapter 7, Conclusion, concludes on the proposed model indicating possible future applications and suggestions for improvement of the proposed work.

2

State of the Art

Currently, the process of research and development in drug discovery has faced an extensive set of challenges thanks to the scientific advances verified in the last decades. Innovative methods for experimental screening of drugs have made it possible to speed up and improve the process of drug discovery leading to the generation of promising lead molecules. These techniques combined with molecular modelling strategies, genomics and computational biology, provided a better understanding of drug-target interactions leading to more accurate predictions in drug discovery applications.

The increase in corporate investment led to the development of computer-aided drug discovery (CADD) methods able to predict the optimal DMPK profile and molecular structures capable of triggering the intended biological response in organs and tissues. Thus, *in silico* methods applied to drug discovery have enhanced the discovery of new leads with increased therapeutic potential and approval rates for distribution and commercialization [26].

In silico methods used in CADD can significantly reduce the amount of data that needs to be screened out in an HTS assay. These methods reduce the cost of drug discovery and drug design but also reduce the amount of time needed for a drug to be approved by the regulatory agencies and, consequently, reach the market. These tools identify a set of possible lead molecules suitable for testing and effectively predict the toxicity and bioavailability of the identified compound structures. CADD also modulates the prediction of novel compounds by identifying a set of possible analog molecules with increased pharmacological activity derived from high potency compound structures. CADD methods generally, combine approaches based on the knowledge of ligands that interact with the target of interest (ligand-based drug discovery, LBDD) and others based on the three-dimensional structure of the target molecule and the drug-target complex (structure-based drug discovery, SBDD) [28].

The application of a CADD method depends on the accessibility of the target

structural information. Therefore, when the information about the tridimensional structure is available, it is possible to apply a SBDD approach to predict the lead molecules compound structure. LBDD methods predict according to the structure of the active binders of the identified target molecule when there is no available information about its tridimensional structure.

The most widely used SBDD approaches include structure-based virtual screening, molecular docking, homology modelling or molecular dynamics. On the other hand, LBDD methods like pharmacophore modelling, quantitative structure-activity relationships (QSAR), similarity approaches and ligand-based virtual screening are important in the discovery of small molecules and, at identifying possible correlations between the molecular structures and their biological activity [29].

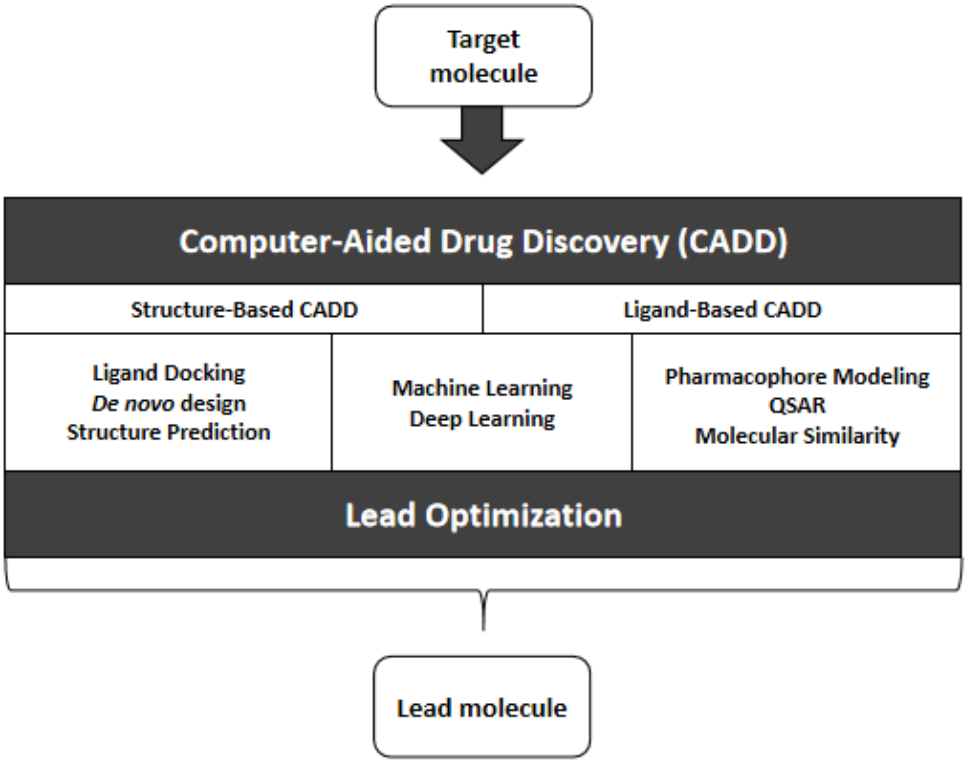


Figure 2.1: Computer-aided drug discovery methods for compound prediction.

2.1 Structure-Based Drug Discovery

Once the three-dimensional structure of a target molecule is known, the most correct approach to adopt is the use of a structure-based method. The drug design is carried out based on the knowledge of the tridimensional structure of its specific target. This is only feasible if it is possible to determine the three-dimensional structure of the target molecule either by X-ray crystallography or by NMR spectroscopy. Both techniques entail high costs and have serious limitations related to their experimental execution. X-ray crystallography is only possible if protein crystalization is achievable, which in the case of transmembrane proteins, is an extremely difficult task. On the other hand, NMR spectroscopy is limited by the size of the target molecules, being restricted only to small molecules [30]. Molecular dynamic simulation (MD) techniques are alternative methods often used to predict the structure of the binding site, modelling different signalling pathways [31]. Molecular docking and the novel design of ligands (agonists, antagonists and inhibitors) are among the most used methods in a structure-based approach.

Wu et al. [32] (2003) developed a grid-based docking algorithm using a representative docking method called CDOCKER. The objective of this study was to compare the performance of this MD algorithm using grid-based approximations for the ligand-target interaction against an all-atom force field calculations. The best performance was achieved when combining both approaches, using a grid-based docking algorithm and an all-atom force field at the final minimization step to optimize the molecular poses at the docking sites. The use of a grid-based docking algorithm also provided a computational time reduction and statistically identical results when compared with an all-force field docking algorithm.

Claußen et al. [33] (2001) designed a docking software tool, FlexE, capable of addressing the problem of structural variations in protein sequence and the challenge of docking a flexible ligand into alternative models of a protein structure. These alternative models are superimposed into a single protein model capable of describing multiple conformations of its tridimensional structure. The docking results achieved were similar to the best results achieved by docking the ligand with the target sequence and provided a significantly lower computing time. The discovery of potential inhibitors is achieved by superimposing different conformations of the protein active site considering all the possible structures and conformational changes that may occur during an interaction.

Bhardwaj et al. [34] (2016) proposed a series of computational methods for the *de*

novo design of structurally restricted peptides. These methods allowed the design of 18-47 residue variations of the peptides and the design of disulfide-crosslinked peptides and heterochiral and cyclized variations. These computational methods served as a basis for the development of small-molecule-based drugs.

Pegg et al. [35] (2001) developed a genetic algorithm capable of reducing the search space of compounds using the information available of other known ligands. This genetic algorithm called ADAPT uses the interactions observed in docking calculations as a function capable of reducing the number of predictable structures. The algorithm uses a set of candidate molecules to iteratively build novel compounds without prior information about the target ligands and according to the function score achieved by the previous set of candidate molecules. The ability to perform local sampling and introduce compound diversity in each iteration of the design cycle provided results that outperform those of the known genetic algorithms.

2.2 Ligand-Based Drug Discovery

The absence of information about the target tridimensional structure leads to the application of a ligand-based approach. This methodology uses information about analog compounds with relevant biological activity over an identified target. Descriptors of the molecular structure establish relationships between the molecular structure and biological activity useful for compound prediction. Some ligand-based approaches include pharmacophore modelling, molecular similarity approaches and quantitative structure-activity relationship (QSAR) analysis [36].

Since the information about the three-dimensional structure is not available, target information is gathered from known active binders. If the target three-dimensional structure is available, all the information concerning the structure of the binding site is used in the development of a predictive model. Important features such as the presence of relevant functional groups, acid and basic amino acid residues or the existence of polar or apolar molecular regions are used in the development of pharmacophore models [37].

QSAR (quantitative structure-activity relationships) are predictive models capable of correlating the physical, chemical and structural properties of a molecule with the corresponding biological activity. This method assumes a relationship between the molecular structure and the biological activity. Thus, similar molecular structures will result in similar biological activities [38]. These models are based on the quantitative description of the physicochemical properties and molecular structure and, through the mathematical formulation of a given property relationship predict the biological activity of an identified target molecule.

Once this mathematical relationship between variables is found, it is possible to predict the response or activity of other chemical structures based on the identified relationship:

$$Biological\ Activity = function(Structural\ and\ Molecular\ Information) \quad (2.1)$$

However, the relationship between the activity and the molecule descriptors is not always linear, and different parameters and properties may present different weights in obtaining the desired biological activity through non-linear dependency relationships with the remaining variables. Thus, machine learning and deep learning techniques have been used to develop QSAR models capable of describing a non-linear relationship between the biological activity and the target molecule descriptors [39].

Mestres et al. [40] (1997) developed MIMIC, a program capable of identifying the similarities between molecular structures and return the relative orientation of the molecules that maximize compound similarity. These molecule alignments are the basis for the development of models capable of predicting molecule structures able to interact with protein binding sites and identify molecular regions important for target binding. This information is later used design compound structures that mimic the key characteristics of these binding domains. MIMIC identifies the set of structurally similar compounds and locates the global and local similarity maximums. The ability to calculate the atomic contributions of the molecular structure to the molecular similarity gives the means to build useful pharmacophore patterns.

Steindl et al. [41] (2006) proposed a parallel pharmacophore-based virtual screening method to develop a system capable of providing reliable *in silico* predictions for bioactivity profiles of virtual compounds and small molecules. The method was applied to 50 structure-based pharmacophore models for multiple targets and 100 antiviral molecules. The enrichment of the pharmacophore models resulted in a successful activity profiling for almost all input molecules.

Lima et al. [42] (2006) developed a QSAR approach to explore multiple combinations of modelling techniques and various types of descriptors followed by rigorous model validation. The modelling method used to perform the prediction was Support Vector Machine (SVM), binary QSAR, decision trees and K-nearest neighbour (KNN). These methods combined with different types of molecular descriptors (connectivity indices, atom pair descriptors, VolSurf descriptors, molecular operation environment descriptors) resulted in 16 different QSAR model approaches. This study demonstrated the increased performance of the combined models with concurrent methods and averaged predictions, over the individual model approaches highlighting the benefits of a combinatorial QSAR methodology.

2.3 Deep Learning in Drug Discovery

The large volume of data available for classification and the availability of high-performance computing resources has motivated the use of artificial intelligence techniques for drug discovery. These algorithms are able to predict relationships between the molecular structure and physicochemical parameters with the biological activity observed during their interaction with a given target molecule. These relationships help in the discovery of new candidate compounds starting from the identification of patterns to the extraction of important features that model the drug-target interaction. The high predictive power of these algorithms lies in their ability to identify patterns in the molecular structure with a pivotal role in the interaction with the active site of the identified target molecules [43].

Artificial neural networks (ANNs) are widely used in several research fields due to their ability to extract features and develop powerful predictive models. Deep learning methods model non-linear systems of compound prediction and can be applied as non-linear regression models used in QSAR. These systems learn patterns and local dependencies which model the relation between molecular and structural descriptors and compound biological activity through an iterative process of learning and prediction [44]. ANNs learn different representations to develop structural information that gives a general perception or belief about data. However, despite the high predictive power of these inner representations and the diversity of molecular structures to be learned, all neural networks require a large amount of training data to predict novel compounds with the desired pharmacological activity [45].

The neuron is a unique biological unit consisting of a cell body, an axon and multiple dendrites. Typically, neurons receive an input signal from the dendrites and send a response signal along the axon that is transmitted to the dendrites of the next neuron. The signal travels along the axon of the input neuron and communicates the output response by synapses with other neurons. Mathematically, a neuron sends a signal x to the next neuron through a specific synapse of weight w. Different synapses between two different neurons i and j have different weights w_{ij}. The weight of a synapse returns its influence in the signal transmission between neurons and indicates the direction of the signal. A positive signal means an excitatory synapse, while a negative signal corresponds to an inhibitory synapse. Once the signal is transmitted, all the signals are multiplied by the corresponding synaptic weights and get summed, returning an output signal y. The neuron returns a signal if the sum is above a certain threshold. An activation function $a(x)$ at the end

modulates the frequency of the signals transmitted [46].

The output y_i of a given neuron is given by,

$$y_i = a(\sum_j (w_{ij} * x_j) + b_i) \tag{2.2}$$

Neural networks are different sets of neurons organized in layers where the outputs serve as an input to another set of neurons in the next layer.

A neural network consists on the following components:

1. An input layer, X ;
2. A set of hidden layers H;
3. An output layer, Y;
4. A set of weights W describing the weight of the connection of each neuron in one layer with each one of the neurons in the next layer;
5. A set of biases B describing the bias terms of set of neurons.

The input layer X of the network receives an input vector x. Each element in x corresponds to the number of neurons in the input layer. The output layer Y returns the vector y of the output values of the network. The values of y represent a probability distribution over the k output classes. y is returned by propagating x through the hidden layers H between X and Y. W is the weight matrix representing the connections between neurons in different layers. The number of entries in W is equal to the number of connections between neurons.

Therefore, the value of a neuron in a layer results from the linear combination of the neuron values from the previous layers weighted by the values w_{ij} in W. The weights represent the strength of specific connections between neurons. The weight matrix W is updated by changing their values and, consequently, the strength of the connections between neurons. Non-linearity is introduced in the output layer through the activation function, which transforms the output of the weighted sum computed in the previous layer into the final prediction p [46].

Considering p the vector of predictions and l the vector of labels, the loss function is defined as a function which depends on the difference between both vectors. The smaller the difference between them, the lower the loss and, the higher the accuracy

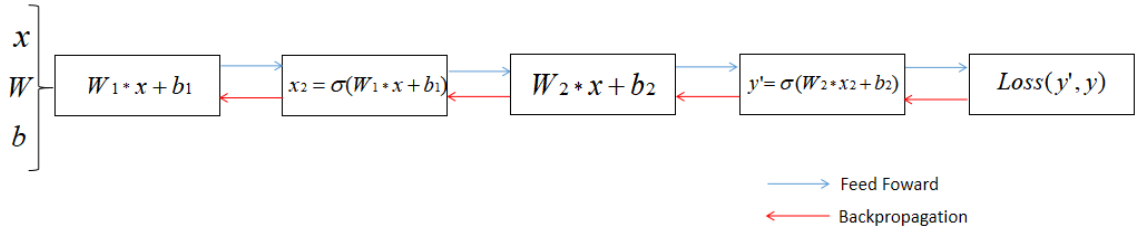
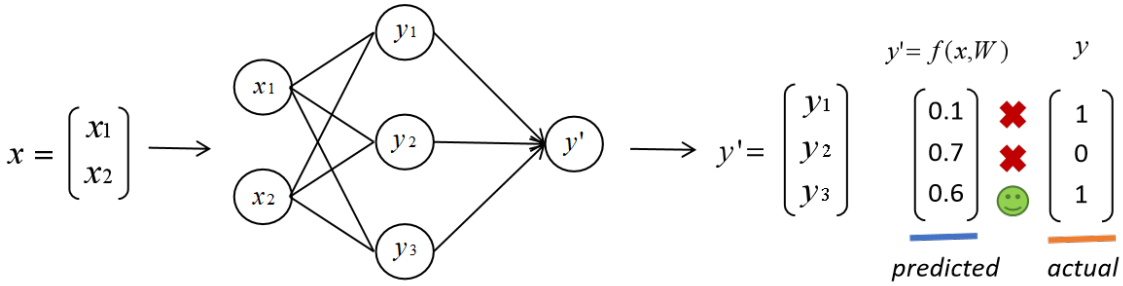


Figure 2.2: Neural network training: feedforward propagation and backpropagation.

of the prediction. An example is the quadratic loss:

$$Loss(x) = \frac{1}{2} \cdot (p(x) - l(x))^2 \tag{2.3}$$

The output is modulated by both weights in W and biases in B. The strength of the prediction is given by the right values of weights and biases which minimize the loss value. The training process consists of multiple iterations where the output vector y is calculated (feedforward propagation) and weights and biases are updated (backpropagation).



$$L(W, x) = \frac{1}{n} \sum_{i=1}^{n} y^{(i)} . \log(f(x^{(i)}, W)) + (1 - y^{(i)}) . \log(1 - f(x^{(i)}, W))$$

Figure 2.3: Neural network training: loss function

In the context of the standard classification, deep neural networks have become state-of-the-art methods in drug discovery. Deep neural networks are capable of learning multiple-level representations based on the diversity present in training data. Feature complexity is distributed throughout the layers, the feature space of a layer being the result of a combination of simpler features of the previous layers. This increase in complexity translates into an output feature space whose size is proportional to the size of the filters that connect a layer to those that precedes it. The addiction of complexity induces the high-level layers to classify more complex structures by learning patterns and local dependencies within data [47].

2.4 Deep Learning in Ligand-Based CADD

The ligand-based methodology takes into account analogous compounds with biological activity or physicochemical parameters relevant for predicting the drug-target interaction. Structure descriptors and the physicochemical profile are features used to establish relationships of interest between the molecular structure and its biological activity.

Lusci et al. [48] (2013) predicted the solubility of compounds of different datasets using a deep learning approach based on recursive neural network architecture. The prediction was based on the learning of graph representations of the compounds (UGRNN) that allowed the extraction of features during each learning iteration. This work achieved outstanding results outperforming the state-of-the-art methods in the accurate prediction of the aqueous solubility.

Under the Tox21 Data Challenge, Unterthiner et al. [49] (2015) designed a deep neural network architecture able to predict different toxic effects associated with the binding to nuclear receptors and related with stress response pathways from Tox21 dataset. The proposed models were built based on key-features as the similarity with existing analog compounds, toxicophores and other molecule descriptors. The deep learning models developed with different combinations of layers, hidden units and learning rates outperformed other model approaches.

Graph convolutional networks were developed by Ohue et al. [50] (2019) to represent, in a more consistent way, the differences associated with the interatomic distance between different molecule conformations. This was achieved by correcting the distances between atoms in a ring system and by correcting atom pairs which improved the performance of the proposed models in comparison with a graph convolutional network. This study showed that representing tridimensional effects from sequential structures may improve the model's performance in the accurate prediction of biological activities.

2.5 Deep Learning in Structure-Based CADD

The structure-based methodology considers the importance of the three-dimensional structure of the target molecules in modelling their biological activity and in predicting compounds capable of producing the response to the identified pharmacological target. This type of approach uses the knowledge of the interactions verified in the ligand-receptor complex through the use of scoring functions that rank a set of hit molecules and interaction fingerprints capable of describing the atomic interactions between pairs of molecules.

Wallach et al. [51] (2015) developed a deep learning model that outperformed the Smina scoring function by implementing a convolutional 3D layer model. Descriptors derived from the ligand-target interaction, as the presence or absence of certain molecular groups and atom types in the target's binding site, were used to generate the model.

DeepVS, a deep learning model approach developed by Pereira et al. [52] (2016), was proposed to distinguish different compounds from inespecific molecules that act as decoys in drug-target interaction. This model uses molecular docking results to train a neural network extracting features as types of functional groups, amino acid residues, intermolecular/interatomic distances, atom types, etc. The results achieved by the model outperformed the standard docking programs producing great AUC results which validated its use on virtual screening applications.

2.6 One-Shot Learning: A low-data deep learning approach in Ligand-Based CADD

Despite the high predictive power of deep learning methods one of its biggest limitations is the requirement of a huge amount of labelled data. The feasibility of recognizing new lead-like drugs with a reduced set of biological data available for training remains an important challenge in compound prediction for drug discovery applications. The applicability of these deep learning techniques has been limited by the scarcity of labelled information available for training and the costs associated with data collection and processing [53]. Moreover, the identification of the class whenever a new group of molecules is observed, without requiring periodic retraining and using only a few training examples per class, is crucial in drug discovery tasks.

One-shot learning is a deep learning strategy that aims to face the problem of low-data in the accurate and reliable prediction of novel compounds for drug discovery applications.

Humans learn multiple representations from a small number of examples, and then use the knowledge acquired to distinguishing new concepts, even if only observed once. This ability of one-shot learn different representations, by observing a concept and generate meaningful and diverse variations, is important when classifying new instances of unknown concepts. This idea of one-shot generalization gave rise to one-shot learning methods [54].

In deep learning standard classification, an input is propagated through a series of layers and finally, the output is a prediction containing the probability distribution over multiple classes. Therefore, if we try to classify an input in one of k distinct classes, k different probabilities are generated at the output indicating the probability of the input belonging to each one of the k possible labels. To perform an accurate prediction, a large number of examples of each class is needed for the network's training and, if the model is trained considering each one of this k classes it is not possible to test the prediction on any other class. To perform a prediction given this new label we must re-train the model considering a certain amount of examples of the new class.

In drug discovery applications, the size of the datasets is dynamically changing once a new compound is discovered and some classes of molecules are less-represented than others in training. Thus, to perform an accurate prediction we must periodically retrain the model and collect more examples of the less-represented classes which

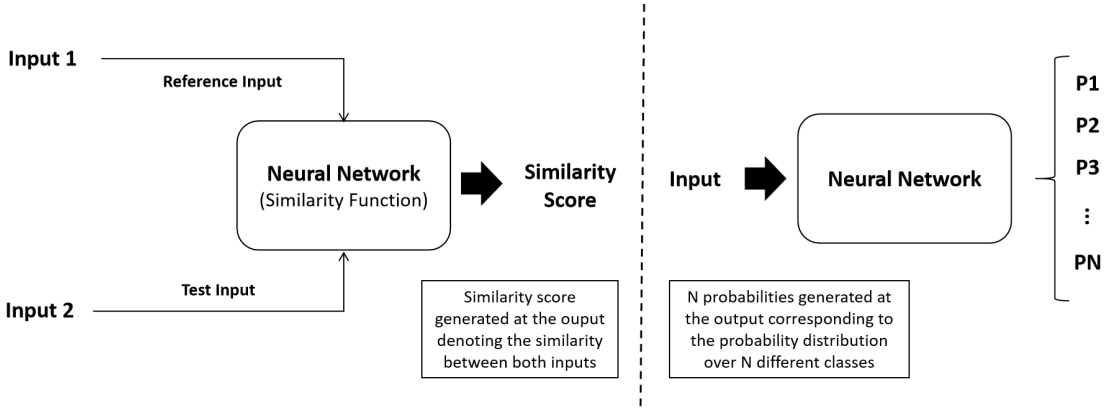


Figure 2.4: One-shot learning classification versus a deep learning standard classification.

might not be feasible.

Instead of directly classifying a given instance, a one-shot learning model learns a similarity function that accepts two inputs and returns a score that denotes the similarity between them. The learnt similarity rule predicts instances whose classes are unknown at the training stage. Therefore, instead of returning a probability distribution of a given test instance belonging to any of the output classes, it learns a distance metric capable of distinguishing two different inputs and highlight the similarities between them [55].

In the context of drug discovery, the application of a one-shot classification strategy improves the prediction of novel compounds whose classes are less-represented, and only requires few examples per class for training. Despite the size of the training set, a single molecule per class is needed for training. This molecule is used as a reference instance to compute the distance with any other molecule to predict a novel compound in one-shot, according to the output similarity score generated between them. This similarity measure is the probability of both inputs belonging to the same class of molecular structures.

One of the first applications of one-shot learning was focused on the identification of characters. The objective was to find meaning for the observed images' structure by decomposing them into successively smaller components, using a method called Hierarchical Bayesian Program Learning (HBPL) (Lake et al. [56] (2013)). This novel approach by considering one training example and decomposing it can classify new examples outperforming the standard models in character prediction and modelling the structures captured by the human perception in the identification of visual concepts.

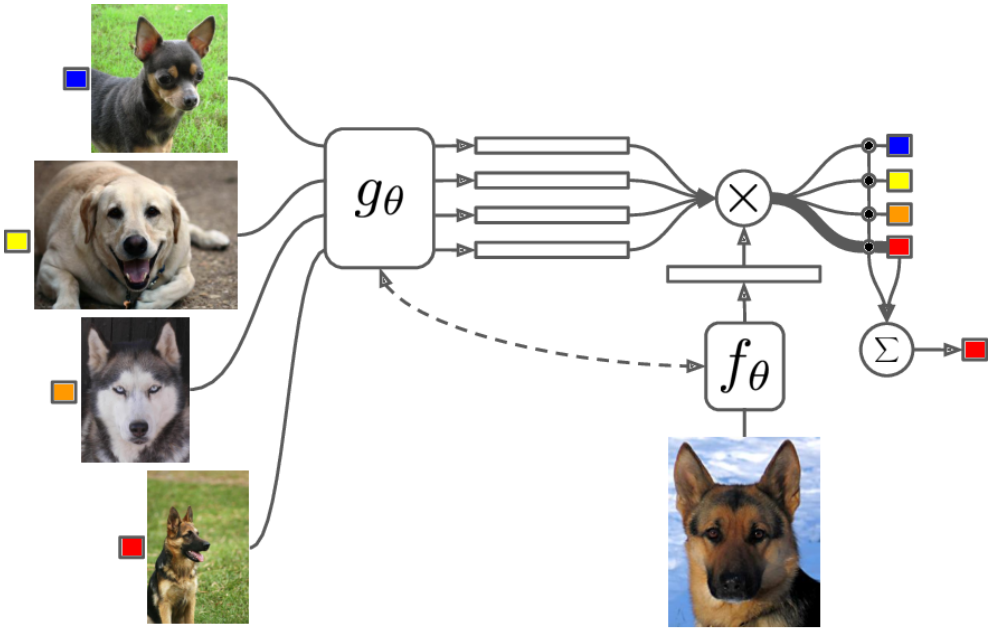


Figure 2.5: One-shot learning approach. Image from "Matching networks for one shot learning" [57].

Wu et al. [58] (2012) proposed a one-learning based approach capable of classifying human gestures using only one example per class for training using RGB and depth information collected from various sensors. Multi-view Spectral Embedding was used to merge gesture modalities in a physically meaningful manner and an extended-MHI (Extended-Motion-History-Image) acted as a motion descriptors of the gestures. Under the CHALEARN Challenge, this approach allowed the classification of human gestures with a single learning example per class and achieving good performances (less than 0.3 in Levenshtein distance achieved in the challenge). One-shot learning and a maximum correlation coefficient approach reduced the overfitting problem related to the application of complex models with an extensive amount of parameters to learn.

In a novel one-shot learning approach, Vynials et al. [57] (2016) developed a deep learning model based on the nearest neighbour classifier, training directly on each one-shot task and consequently, applying one-shot learning in NLP tasks. The proposed neural network architecture outperforms the baseline models improving the predictions made on Omniglot and ImageNet datasets. This study proposes the learning of a model that maps a support set of known class labels and an unseen input example whose class is predicted by a KNN-based classifier through a one-shot learning classification strategy.

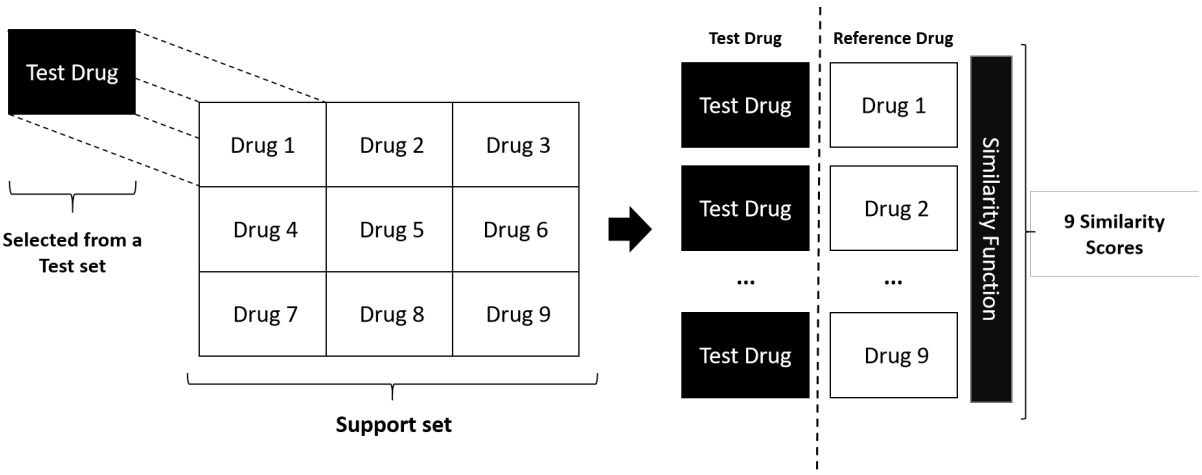


Figure 2.6: N-way one-shot learning example: 9-way one-shot learning.

To explore the relation between one-shot learning and a meta-learning approach, Santoro et al. [59] (2016) trained a memory augmented network for character classification on Omniglot dataset. This study proposed a set of deep learning models, Memory Augmented Neural Networks (MANN), with flexible memory resources that store important information about newly observed tasks and encode background information about previously seen tasks. This model can store new data and manipulate the assimilated data to make accurate predictions considering only a few examples per class for training.

To learn a certain task according to a human demonstration Duan et al. [60] (2017) developed a neural net trained with pairs of human demonstrations to reproduce one task of the pair as a function of the other, setting a promising path to the development of learn-based robots. The network considers one demonstration of the pair as an input for training, and a state of the second demonstration and can predict that same state for the first demonstration of the pair. A one-shot learning approach maximizes the performance of the model in reproducing a learned task when facing a new and previously unseen demonstration of that task.

More recently, Altae-Tran et al. [61] (2017) developed a new LTSM model called Iteratively Refined LSTM (IterRefLSTM), capable of optimizing the prediction of novel compounds on the lead-optimization step of drug discovery tasks. This method removes data dependencies by iteratively co-evolving compound structures in an iterative process. The study made a comparison between 4 convolutional neural network architectures with a random forest model using circular fingerprints as an input in the prediction of multiple datasets (SIDER, Tox21 and MUV). The convolutional models performed better predictions for the SIDER and Tox21 datasets, however

for the MUV dataset random forest model performed better. On the other hand, the authors tried to apply a transfer learning approach with the models trained on Tox21 dataset to predict a set of patient observations from SIDER dataset. This study highlighted the high predictive power of convolutional models using a one-shot learning approach given the low-data available in the lead-optimization step of drug discovery tasks (Figure 2.7).

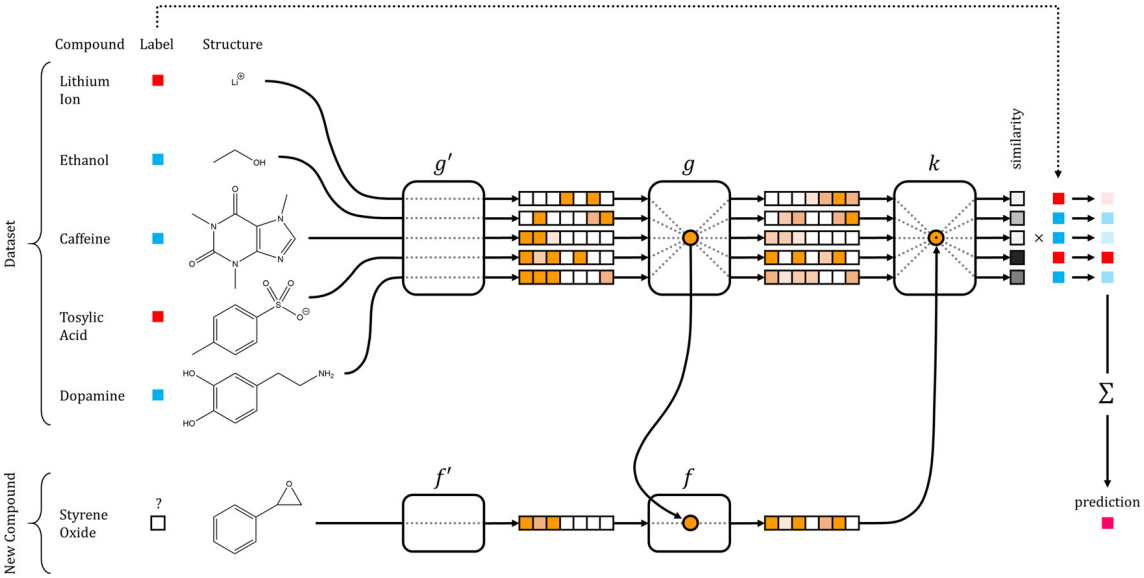


Figure 2.7: One-shot learning approach for compound prediction. Image from "Low-Data Drug Discovery with One-Shot Learning" [61].

3

Data Preparation

3.1 Data Processing

The molecular structure of drugs consists of a set of atoms, chemical bonds and functional groups important for the interaction with specific target molecules. Each drug has a molecular structure that varies either in size or in the molecular arrangement of its atoms. SMILES are representative sequences of characters which directly describe the molecular structure and the chemical bonds between atoms [62]. This data structure contains important features that represent specific patterns of the molecular sequences.

In this study, the SMILES will serve as input to a set of predictive models capable of inferring the compound structural analogs. These representations identify patterns related to multiple compound classes according to the differences in the molecular structures. The SMILES processing is performed in such a way that the strings are treated as images representing the molecular structure of drugs.

Since different drugs correspond to different character sequences, it was established a threshold for the minimum and the maximum number of characters. The minimum size of the sequences was considered to be the size of the shortest SMILES sequence and the maximum as the size of the longest sequence of characters. SMILES are encoded into binary matrices with an equal number of rows and columns as binary matrices with the same amount of features and dimensions. For sequences smaller than the maximum threshold, a padding operation is performed. This operation represents the missing characters in the SMILES, as an additional line of '0s' in the corresponding binary matrix. The padding operation produces equivalent representations of the molecular structures prepared for the classification by the predictive models.

The SMILES processing method is described in Figure 3.1.

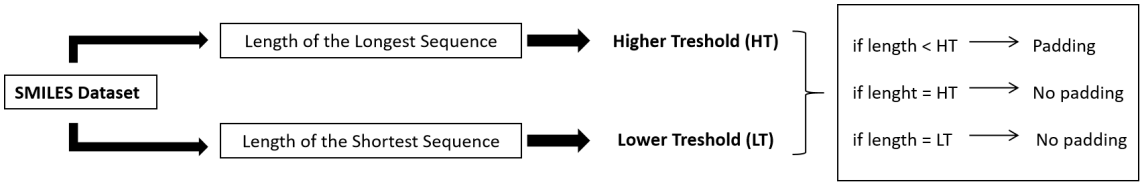


Figure 3.1: Processing methodology applied for SMILES based on a length threshold.

3.2 Data Representation and Encoding

SMILES are representations not well suited to serve as an input for a convolutional model. Instead of using the characters sequences directly, the SMILES are converted into numeric representations (integer-based encoding). The integer-based representations of the SMILES are encoded into binary matrices capable of acting as consistent and unambiguous representations of the molecular structures prepared to be used by the predictive models (one-hot encoding) [63].

Integer-based encoding (Figure 3.2) performs the transformation between the character sequence and the numerical representation where each character corresponds to a distinct integer value.

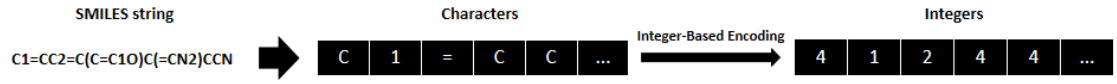


Figure 3.2: Integer-based encoding.

The one-hot encoding was carried out by converting the character sequences into binary matrices with the number of lines equal to the maximum threshold and the number of columns equal to the number of distinct characters identified in the SMILES.

Drugs are treated as images in the form of binary matrices by assigning a binary variable to each distinct character. Thus, to each distinct character used for the representation, a distinct integer is assigned using a dictionary. Each integer value is converted into a binary vector that assigns '1' to the corresponding integer and '0' to the others.

In the one-hot encoding operation, we consider each one of the features in the molecule sequences as a categorical variable, whose corresponding integer value in the dictionary is converted into a binary vector representative of each one of the characters observed in the SMILES.

3.2.1 SMILES Encoding

The one-hot encoding was performed in order to match each of the different characters with an integer value which is later converted into a binary vector. This binary vector integrates a binary matrix which is used to represent the whole sequence of characters.

A dictionary consisting of 54 different categories was used to represent the different characters identified in the SMILES (Table 3.1). Binary matrices establish a consistent and uniform representation of compound structure and the preservation of structural information suitable for classification by the predictive models.

Table 3.1: SMILES char-integer dictionary.

Character	**Integer**
I	1
...	...
C	9
...	...
t	54

3.3 Data Grouping

The model used to predict different compounds through the learning of a given similarity function is trained with a set of examples of different classes. Nevertheless, the dataset used does not allow to identify classes according to a common structure.

Therefore, it was necessary to process the SMILES with the objective of forming groups according to their chemical structure. This procedure enables the identification of categories that could be predicted by a model designed for that purpose.

Siti Asmah Bero et al. (2018) [64] identified distance metrics (Tanimoto distance and Euclidean distance) capable of inferring the dissimilarities observed between classes of molecules as an inter-class differentiation, based on the common characteristics within molecular fingerprints. These distance metrics applied to drug fingerprints allow the identification of multiple classes of compounds according to the structural differences observed between sequences. Thus, molecular structures belonging to the same class should present a smaller Tanimoto distance. Thus, it is possible to infer that structurally identical compounds tend to present identical biological activities and pharmacological properties. Zhang et al. (2015) [65] and Maggiora et al. (2014) [66] state that similarity value distributions such as Tanimoto depend on the classes of bioactive compounds as well as on the chosen metrics and descriptors. Zwierzyna et al. (2015) [67] revealed class-specific differences between active compound classes. These structural differences identify compound categories given the properties of the molecule chemical structures.

RDKit was used to access the SMILES fingerprint representation [68] in order to form clusters according to the Tanimoto distance [69].

Considering two molecular fingerprints A and B, the Tanimoto coefficient T_c is given by the following equation:

$$T_c(A, B) = \frac{N_{AB}}{N_A + N_B - N_{AB}} \ , \tag{3.1}$$

with N_{AB} the number of 'on' bits in A and B, N_A the number of 'on' bits in A and N_B the number of 'on' bits in B [70]. Overall, the data was organized into groups according to the common chemical structures among the SMILES fingerprints. On that account, it was possible to establish categories of predictable compounds.

4

Model

4.1 Encoding Layer

The set of SMILES representing the molecular structure of drugs underwent a series of processing operations which converted the characters into numerical sequences (character-integer encoding) unambiguously representing the compound structures. After this transformation, the SMILES are treated as images represented by matrices encoding the molecular sequences. The one-hot encoding operation normalizes the contribution of each categorical value, each of the different characters in the SMILES, in the network's training. In particular, this is illustrated in Figure 4.1 with respect to serotonin.

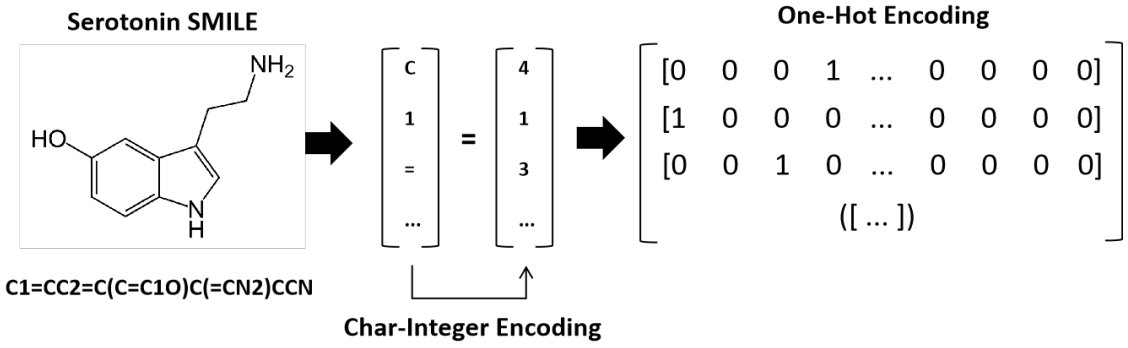


Figure 4.1: One-hot encoding of SMILES.

4.2 Convolutional Neural Networks (CNNs)

Convolutional neural networks inherit many of the properties of artificial neural networks (ANNs) and consist of a set of layers containing multiple neurons and a set of learnable parameters, representing the strength of the connections between neurons. A CNN treats the input as an image, returning a score that denotes the probability of a given input belonging to each one of the output classes.

Like a conventional neural network, CNNs consist of a set of hidden layers that propagate a given input through the network. These layers are made up of neurons, each one connected to each neuron of the previous layer. The output is given by a fully-connected layer that returns a probability distribution over each of the different output classes (Figure 4.2).

Height, width and depth are the 3 dimensions of the neurons. A convolutional neural network transforms an input image into an output feature map applied to a given activation function. A convolutional neural network can stack up to different types of layers, in particular, convolutional layers, pooling layers, ReLU layers and fully-connected layers [71].

Thus, the architecture of CNN may contain:

1. Input Layer: Displays the neural network input as an array of pixels of dimension $(width, height, depth)$;
2. Convolutional Layers (Conv): Returns the output of the convolution operation between the weight matrices of each neuron and the regions to which they are connected in the input. If the convolutional layer has a total of n filters the size of the output corresponds to: $(width$,$height, n)$
3. ReLU Layer (ReLU): Performs an operation that transforms the values of a given input to non-negative values. The operation $max(0, input)$ does not change the size of the output;
4. Pooling Layer (Pool): The pooling layer performs an operation that reduces the dimension of the input in the two-dimensional space (downsampling). This operation changes the size of the output.
5. Fully-Connected Layer (FC): Returns the probability distribution of the input belonging to each of the k output classes.

The convolutional layer is a set of convolutional filters capable of extracting local features from an input. The input is propagated through the convolutional layers

(forward propagation) while the convolutional filters slide through the spatial dimensions of the input. The two-dimensional output feature map returns the result of the convolution operation of the filter for each position in the two-dimensional space $(width, height)$. Convolutional neural networks learn a set of convolutional filters capable of extracting important features for the network prediction. As we progress to deeper convolutional layers, the greater the ability to distinguish more complex patterns by converting low-level features into more complex abstract concepts. Each convolutional layer presents a given set of filters and each one of them is responsible for returning a two-dimensional feature map which produces the output of that same layer [72].

Convolutional layers have the ability to restrict the connections with neurons to a small region of the input. The connection defines the size of the filters used in the convolution operation and the length of the connections corresponds to the depth of the input of the convolutional layer. Thus, in the spatial dimension, there is a local connection, which is not the case with the input depth.

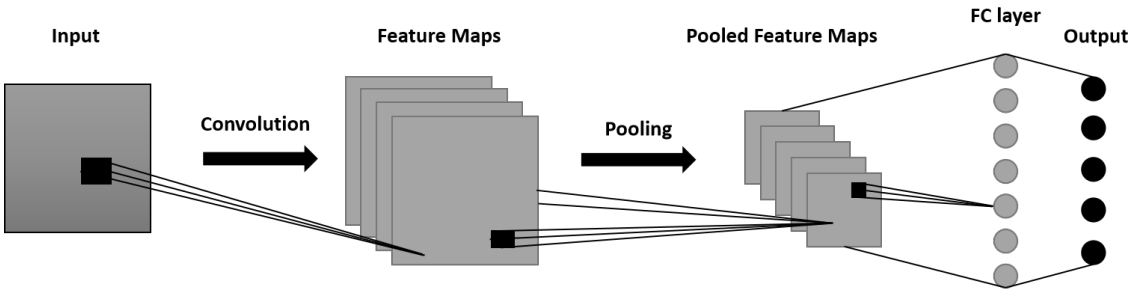


Figure 4.2: Convolutional neural network.

Three fundamental parameters determine the size of the output of a convolutional layer and they are defined as hyperparameters of the network architecture:

1. Depth: Number of filters chosen for the convolution operation. In each filter, the neurons of the convolutional layer operate on different regions of the input.

2. Stride: Number of positions in the input image over which the filter slides through in the convolution operation. The application of a larger stride produces a smaller output feature map;

3. Zero-Padding: Operation that adds zeros at the edges of the spatial dimension of the input. This parameter preserves the spatial dimension of the output feature map generated in the convolution operation. This property allows the existence of inputs and outputs with the same spatial dimension after the convolution.

The output size of the output feature map is a function of the input size, the number of filters, the amount of padding performed and the stride applied in the convolution operation:

$$output\_size = \frac{input\_size - filter\_size + 2 * padding}{stride} + 1 \tag{4.1}$$

This value must be an integer value so that the neurons in the next convolutional layer can fit into the output feature map.

The convolution operation (Figure 4.3) consists in a set of multiplication operations and later sum between the elements of the filter and the elements of an input image over which this filter slides, resulting in a feature map of convolutional outputs.

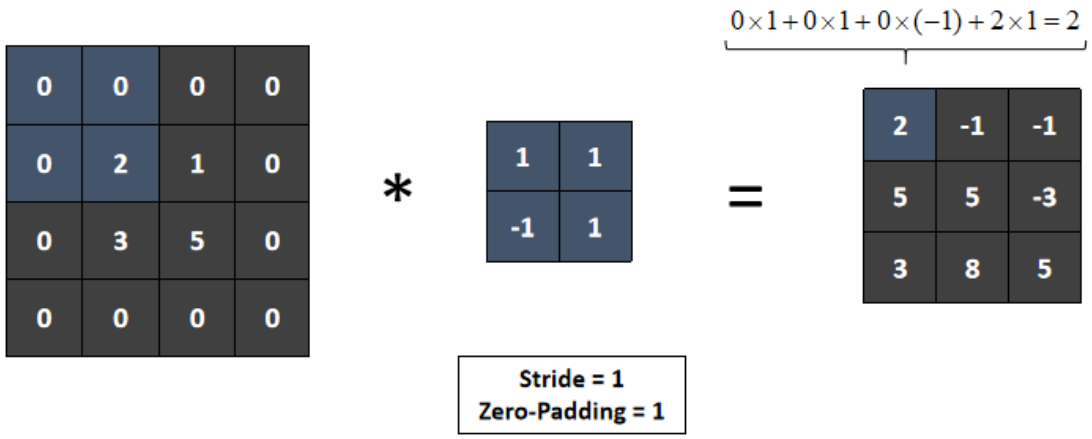


Figure 4.3: Convolution operation.

The ReLU (Rectified Linear Unit) Layer applies a function that normalizes the values present in a given input volume to non-negative values: $max(0,x)$. This function sets a threshold at 0, where the negative values are nullified. This layer accelerates the process of convergence of the predictive model when compared to functions like the *tanh* or *sigmoid* functions due to the linear profile of the $max(0,x)$ function. On the other hand, ReLU layers involve less computationally demanding operations and less time for the model to converge [73].

A very common strategy is the placement of pooling layers interspersed with convolutional layers by inserting them between consecutive convolutional layers. Pooling layers (Figure 4.4) are responsible for reducing the spatial dimension of the outputs [74]. This operation reduces the computational cost associated with convolutional operations and prevents overfitting. These layers act independently along the depth of the input, reducing the size of each depth slice in the two-dimensional space.

Convolutional networks develop structure in a feature space where the complexity stratifies along with the different layers. The lower layers allow the detection of small motifs and the extraction of low-level features. The increase in the number of filters with an increase of the combinatorial size of the feature space results in the learning of representations of increasing complexity converting low-level features into more complex abstract concepts.

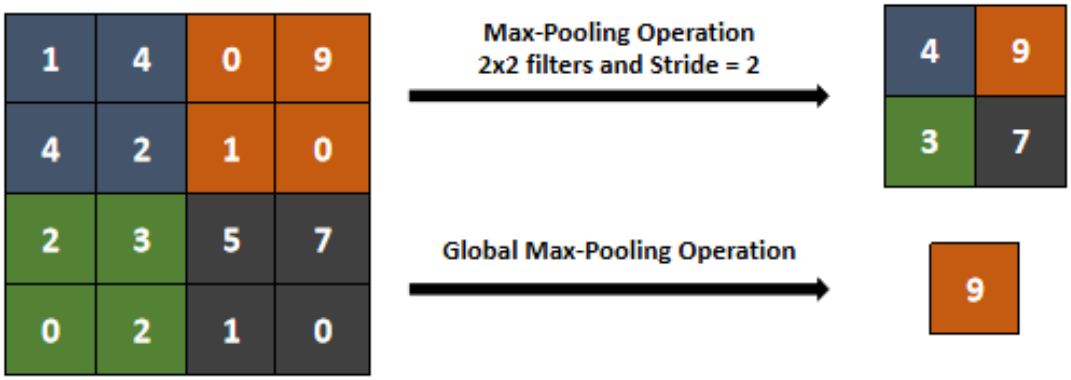


Figure 4.4: Max-pooling and global max-pooling operations.

4.3 Model Overview

In this study, starting from a reduced set of candidate molecules, a set of parallel convolutional neural networks is adapted to predict novel compounds, according to the structural similarities observed between molecules.

SMILES are processed based on the length of the character sequence, as previously mentioned in Section 3.1. A char-integer encoding is performed, transforming the SMILES characters in the corresponding integer values, as mentioned in Section 3.2.

The integer values are encoded into binary vectors, converting the SMILES into binary matrices representing the compound molecular structures. The one-hot encoding operation normalizes the contribution of each character identified in the SMILES.

The proposed model accepts different pairs of molecules and learns a similarity function, which returns a similarity score between two input molecules. Thus, according to the learned similarity rule, the network predicts the similarity score in one shot. A Siamese neural network built upon two parallel and identical convolutional neural networks is introduced as the proposed model approach. This network is compatible with a set of pairs of compounds provided for training. The model learns a similarity function and returns a distance metric applied to the output feature vectors from both Siamese twins. This similarity measure allows the model to predict novel compounds based on a reduced set of candidate molecules available for training. Both Siamese twins are indistinguishable since they are two copies of the same network and share the same set of parameters [75]. Parallel neural networks present a dense layer structure where the output feature vector of the last layer serves as an input to a sigmoid function, which condenses the prediction into a probability value between 0 and 1.

These parallel convolutional neural networks reduce their respective inputs to increasingly smaller tensors as we progress to the high-level layers. The absolute difference between the output feature vectors is used as an input to the learnt similarity function [76]. This metric returns a similarity score between two input molecules.

In a one-shot learning approach, one compound is established as a reference molecule and compared with different compounds expressing the probability of both belonging to the same class, according to a given similarity score *score*. The Siamese twins are symmetric neural networks, which means that the probability P of a given instance d_1 belonging to the same class as d_2 is equal to the probability d_2 belonging to the

same class as d_1. Thus, if d_1 and d_2 are compounds of the same class, we represent this relation as $d_1 * d_2$. If the network is symmetric, $d_1 * d_2$ and $d_2 * d_1$ represent the same, which means that if we switch the order of the inputs of the Siamese network, the returned output prediction would be the same:

$$P(d_1 * d_2) = P(d_2 * d_1) \tag{4.2}$$

This symmetry property is very important when learning a similarity metric:

$$score\ (d_1 * d_2) = score\ (d_2 * d_1) \tag{4.3}$$

If we chose to concatenate the Siamese inputs, the final input is concatenated binary matrix which is propagated through the layers of a single convolutional neural network. Each element is convoluted with a different set of filters, which breaks the symmetry required to compute a similarity score between two molecular sequences.

An architecture based on two parallel neural networks propagates two inputs through the same set of weights and the difference between the output feature vectors serves as an input to a similarity metric. This symmetry-based approach is less expensive and leads to a pairwise training which improves the model prediction accuracy.

4.4 Pairwise Training

A training set in which half are pairs of the same class and another half of different classes was considered. Since the Siamese neural network accepts pairs of molecules, the dataset size increases, given the number of possible combinations for the pairs of molecules available for training. This number quadratically increases which prevents the Siamese model to overfit.

If we have D examples of drugs of C classes of compounds, there are $D.C$ molecules in total which leads to a number of pairs of:

$$number\ of\ pairs = \binom{D\ .\ C}{2} = \frac{(D\ .\ C)!}{2!\ .\ (D\ .\ C-2)!} \tag{4.4}$$

However, we consider half of the pairs of the same class and half of the different classes for training. Therefore, the maximum number of possible combinations for compound pairs is the total number of possible pairs with compounds of the same class.

If there are L examples each of Q classes, the total number of possible pairs of the same class is given by,

$$max\ number\ of\ pairs = L\ .\ \binom{Q}{2} = \frac{L\ .\ Q!}{2!\ .\ (Q-2)!} \tag{4.5}$$

Thus, the number of training instances increases in Q of a square factor and in L of a linear factor. The increase in the size of the training set reduces the effect of overfitting [77].

The increase in the maximum number of possible combinations for the pairs of molecules available for training is important in the learning of a similarity function. Thus, the maximum number of combinations for the pairs of molecules is calculated in order to obtain a ratio of 1:1 for these two types of pairs. This approach is consistent with the training of a Siamese model which learns a distance metric capable of discriminating very identical compounds and compounds with very high disparities in an equally effective and accurate manner.

4.5 Hyperparameter Optimization Approach

To optimize the network performance, the one-shot learning accuracy of the model on the validation set was established as a guiding criteria. Additionally, a set of 500 one-shot tasks was considered for each of the epochs that took place at the training stage. Therefore, instead of the standard k-fold cross-validation, two methods were used to determine the best model, early stopping and model checkpoint. When we verified that there was no significant increase in the accuracy values over a certain number of epochs, we stopped training and saved the parameters for which the model returns the best value for one-shot learning accuracy on the validation set.

For a higher number of iterations, the model tends to adjust to the training data and memorizing it, losing its generalization capacity. Early stopping and model checkpoint allow to determine the best set of weights that lead to the best accuracy values in the validation set. Thus, it is possible to stop the training of the model before it has the chance to overfit, identifying the inflexion point where the loss in the testing data starts to increase over the loss in the training data (Figure 4.5).

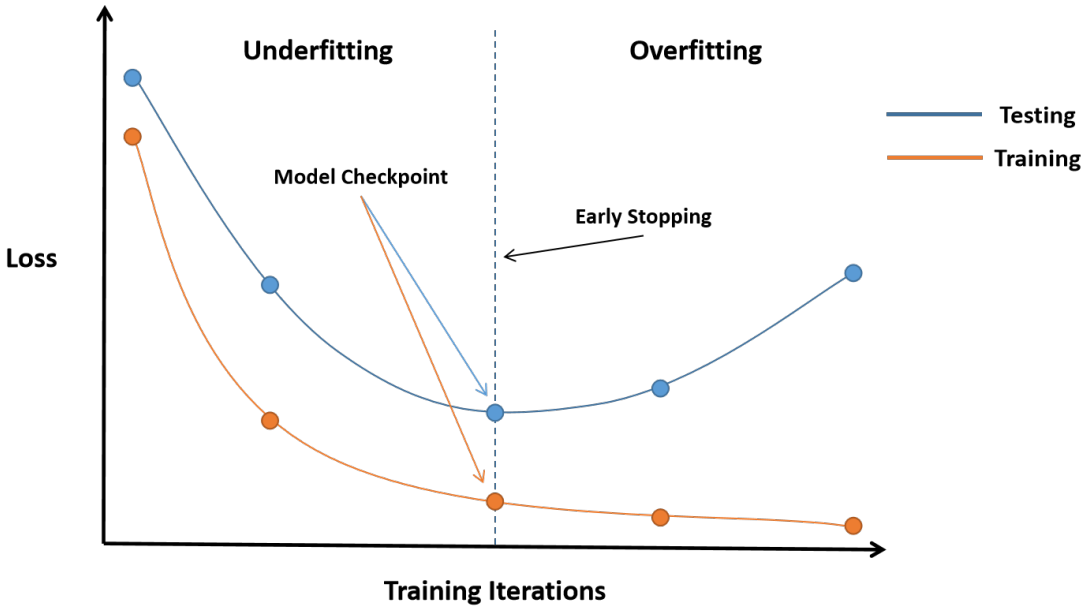


Figure 4.5: Hyperparameter optimization: model checkpoint and early stopping.

amsmath systeme

5

Experimental Setup

5.1 Dataset

5.1.1 SMILES Dataset

MoleculeNet [78] is a benchmark commonly used in the application of machine learning methods in the study of the molecular structures. It was used to extract SMILES (Simplified Molecular Input Line Entry System) for compound representation and encoding. Tox21 was the dataset used, which contains, approximately, 8000 compounds [79, 80].

The canonical form of the SMILES was considered as a consistent representation of the molecular structures, in which each character is considered as feature useful for the model prediction.

Originally, the dataset contained some missing and duplicated values and, was processed so that all these entries were removed completely.

In order to restrict the study to small molecules, the dataset was reduced according to a length threshold based on the character distribution of the SMILES (less than 100 characters). Therefore, the dataset was reduced from 7831 valid SMILES to 7074 (Table 5.2). The reduction of the dataset based on this criteria allows not only to carry out this study on small molecules, but also to evaluate the performance and behavior of the proposed model with a reduced set of training data, which is one of the main objectives of this study.

Table 5.1: Training and testing datasets: number of samples.

	Total	**Reduced**
Training	5873	5305
Testing	1958	1769
	7831	7074

5.2 Model Architecture

The proposed Siamese Neural Network model aims to predict novel compounds by learning a similarity function capable of identifying compound analogs with increased biological activity. The network's training is based on SMILES representing the compound molecular structures. The optimization of a convolutional model involves the fine-tuning of several network parameters. Five different network parameters were selected in order to optimize the model such as: the number of filters in each convolutional layer, the size of the convolutional filters, the number of units in the dense layers and the learning rate.

The parallel neural networks intercalate 4 convolutional layers with 3 pooling layers. The use of a higher number of dense layers does not necessarily result in a significant improvement of the model performance. Conversely, a higher number of dense layers may increase the tendency to overfit due to the memorization of local patterns learned from the training data. The optimization of the model allows the adjustment of the network parameters without the need of adding more convolutional layers. On the other hand, the selection of the activation and optimization functions is an important step when optimizing the model prediction.

The activation function allows the transformation of the output of the weigthed sum of a given set of neurons. Some activation functions are capable of introducing the non-linearity that best describes the complexity of patterns and local dependencies present in deep representations extracted from the input sequences of a deep neural network [81].

The sigmoid activation function accepts a given input x and returns a score representing a probability value between 0 and 1. The function allows the cancellation of highly negative values and returns a unit value for highly positive inputs, representing the behavior of a biological neuron that changes from an inactive state ($sigmoid(x) = 0$) to a saturated state ($sigmoid(x) = 1$), according to different stimulus.

$$sigmoid\ (x) = \frac{1}{1 + e^{-x}} \tag{5.1}$$

The ReLU activation function, as referred to in previous sections, sets the activation threshold to 0, applying the function $f(x) = max(0,x)$. This function accelerates the model convergence through a stochastic gradient descent optimization (SGD) due to its unsaturated profile and linear shape. However, ReLU neurons are fragile

and susceptible to inactivation, when a high gradient crosses a given set of ReLU units. The inactivation of neurons happens in such a way that the update of the weights in each iteration prevents their reactivation leading to the 'death' of the ReLU neurons [82].

$$R(x) = max(0,x) \tag{5.2}$$

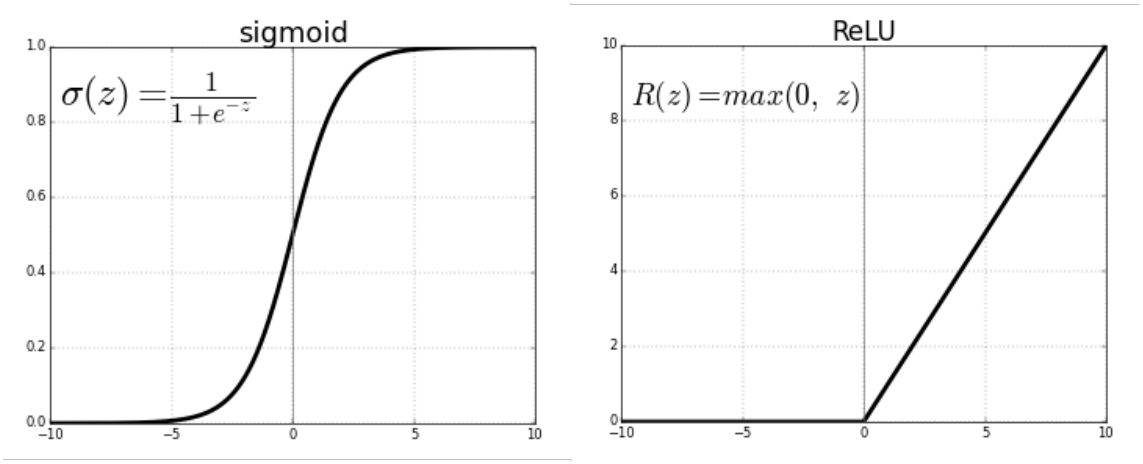


Figure 5.1: Activation functions: sigmoid and ReLU funtions. Image from "www.saugatbhattarai.com.np/what-is-activation-functions-in-neural-network-nn/reluvssigmoid".

The loss function returns a value that translates the difference between the prediction and the label which contains the values to be predicted by the model. The objective is to minimize the loss function which depends on the difference between the prediction and the vector of labels in order to maximize the performance of the model and obtain a more accurate prediction.

The vector of predictions $f(x,W)$ depends on a given input x and a set of weights W. The loss function depends on the difference between two vectors, $f(x,W)$ and the vector of labels y. In the proposed model a binary cross-entropy loss is applied between the target and the prediction:

$$L(W,x) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}.log(f(x^{(i)},W)) + (1 - y^{(i)}).log(1 - f(x^{(i)},W)) \tag{5.3}$$

Thus, the smaller the difference between both distributions $f(x,W)$ e y, the lower the value returned by the loss function and the higher the accuracy of the prediction.

In order to reduce the noise of the weights throughout the learning process and

improve the model generalization capacity, a regularization factor was added to our classifier.

$$L(x_1, x_2, t) = t.log(p) + (1 - t).log(1 - p) + \lambda.||w||.2, \tag{5.4}$$

where t is $y(x_1,x_2)$, that is equal to 1 if x_1 and x_2 are compounds of the same class and equal to 0 if they are from different classes. p is $p(x_1,x_2)$, the predicted similarity score between x_1 and x_2. λ is a regularization term that allows the network to learn increasingly less noisy weights w, leading to improved generalization.

The mathematical method that allows to minimize the loss function and to optimize the model prediction is the gradient descendent method. This method uses the variation of the loss function derivative, that is, its gradient, in the optimization of the weights of a neural network. Thus, there is a certain set of weights W for which the loss function reaches a global minimum corresponding to the point where the model prediction is equal to its label. The objective of the gradient descent method is to update the set of weights of the neural network in order to minimize the value returned by the loss function. This method finds the set of weights that results in a higher loss value and calculates the opposite gradient, since it returns the direction of the loss function decrease towards its global minimum. The size of the steps required for the convergence towards the global minimum is determined by the learning rate parameter α [83].

As gradient descendent method, we chose the Adam optimizer [84] which combines the effectiveness of two other extensions of gradient descendent methods, AdaGrad and RMSProp. This method, instead of adapting the learning rate according to the mean of the first moment of the gradient, makes use of the averages of the second moments. On the other hand, this method calculates the moving average of the gradient and the β_1 and β_2 parameter values, in order to control the decay of the moving average.

The construction of a model capable of discriminating different classes of molecules and predicting novel compounds starting from a very small set of molecular structures is, as previously mentioned, the main objective of this study. The main model consists of a Siamese Neural Network based on convolutional neural networks. The inputs consist of two twin input vectors with a corresponding hidden vector in each convolutional layer. The model architecture that maximizes the network performance is the one whose number of convolutional layers is 4 and the number of filters in each layer is a multiple of 16. A ReLu activation function thresholds the output

feature maps at zero and a maxpooling layer reduces the spatial dimension of the outputs of each convolutional layer (Figure 5.3).

The output feature vectors returned by the last convolutional layer of the Siamese twins serves as an input to a fully connected layer with 1024 units. This layer learns a similarity function between two feature vectors by applying a distance metric to the learned feature maps. It is followed by a dense layer that computes the absolute difference between the two output feature vectors. This value serves as input to a sigmoid function in the last layer.

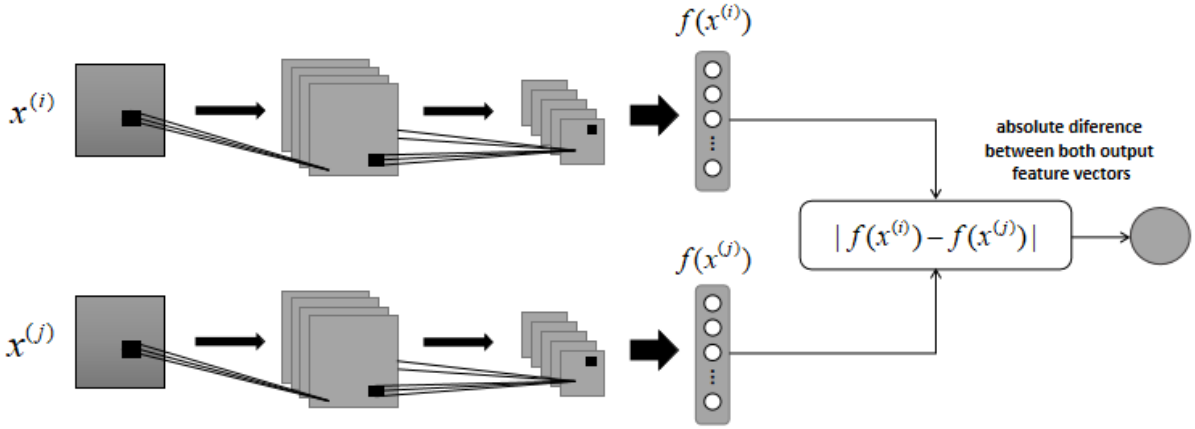


Figure 5.2: Siamese neural network model: absolute difference between the output feature vectors.

The predicted similarity score is given by,

$$score = sigmoid\ (\sum_{i} |\ v^{i}_{1,l} - \ v^{i}_{2,l}\ |\ ), \tag{5.5}$$

v_1 and v_2 are the output feature vectors of the last convolutional layer of each Siamese twin, l the index representing the dense layer, i the index i in each output feature vector and *sigmoid* the activation function. This defines a fully-connected layer which joins the two Siamese twins and computes a distance metric returning the similarity score between both feature vectors.

The first Siamese twin returns the output feature vector for a given query molecule and the other returns an output feature vector for a molecule representing each one of the compound classes. The differences observed between them are proportional to the dissimilarity between both compounds. If both compounds belong to the same class then the feature vectors must be similar, otherwise they must have significant

differences. For both cases, the differences vectors are quite different, which results in a similarity score that is also different.

This similarity measure is a probability, assuming a value between 0 and 1. If *score* is equal to 1, the probability of both compounds belonging to the same class is maximum. If $score = 0$, this probability is minimum according to the learnt similarity rule.

Python 3.7.3 and Keras with Tensorflow back-end [85] were used to develop the proposed model.

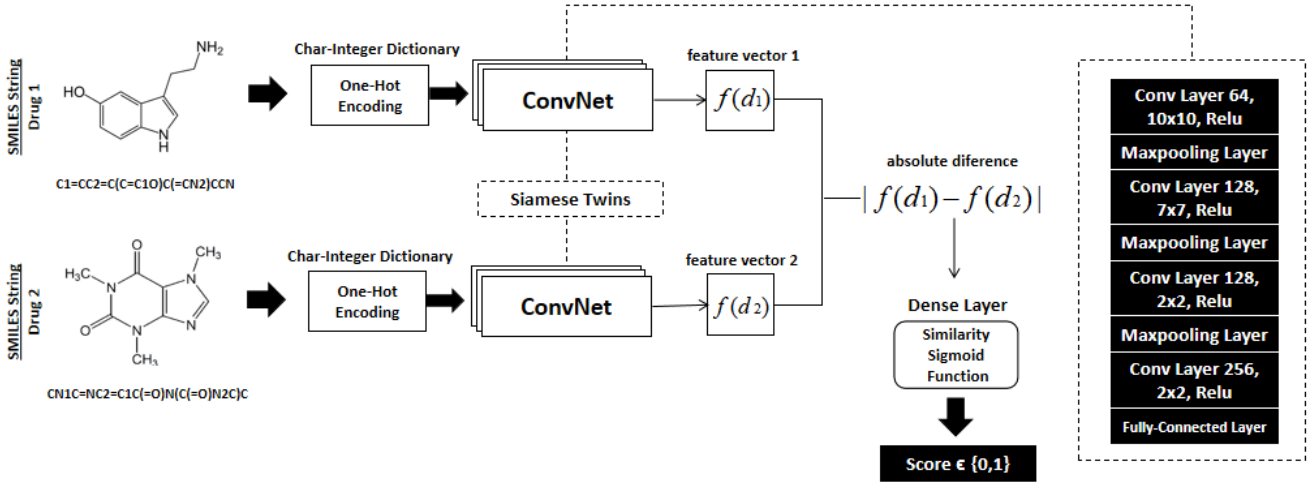


Figure 5.3: Siamese neural network architecture of the proposed model.

5.3 Hyperparameter Optimization

As previously mentioned, in order to perform the hyperparameter optimization, instead of the standard k-fold cross-validation, we used other strategies in order to determine the best set of parameters that resulted in the best model performance, early stopping and model checkpoint [86]. Therefore, to optimize the network performance over the one-shot tasks, we established the accuracy of the model on the validation set as a guiding criteria. We considered a set of 500 one-shot tasks for each of the epochs that took place at training stage. Thus, when there was no significant increase in the accuracy values, we stopped training and saved the parameters for which the model returns the best value for one-shot learning accuracy on the validation set. If an increase in this value wasn't verified during the entire learning procedure, we accepted the parameters of the final state of the model generated during the entire learning schedule.

We chose to vary the size of the filters from 2x2 to 20x20, considering a number of filters that allowed us to optimize performance. Moreover, we found that a number of filters multiple of 16 returned the best performance, so we considered values between 64 and 256 for the convolutional layers and between 256 and 1024 for the fully connected layer. Since we found that a higher learning rate would result in a longer time for the model to converge, this value was kept low in an interval between 10^{-6} and 0.1.

Table 5.2: Network parameter settings for the Siamese model * Initial number of epochs, however early stopping and model checkpoint were applied.

Parameter	Value
Number of convolutional layers	4
Number of dense layers	2
Number of filters	[64,128,128,256]
Filter length	[10,7,2,2]
Epochs*	100
Batch size	50
Optimizer	Adam
Learning rate	0.0001
Number of neurons (FC)	1024
Activation function (CNN)	ReLU
Activation function (FC)	ReLU
Activation function (Output)	Sigmoid Function
Loss Function	Binary Cross Entropy

5.4 N-way One-Shot Learning Approach for Classification

The reduced amount of biological data available for training led us to adopt a new strategy to train and predict novel compounds using the proposed model. A N-way one-shot classification strategy is applied to demonstrate the discriminating power of the learned features.

The Siamese network earlier described accepts pairs of compounds from a small set of molecules D with a given number of N examples of encoded matrices of equal dimension and label l:

$$D = (d_1, l_1), ... , (d_N, l_N) \tag{5.6}$$

The data for classification is organized in pairs, one example from a support set and the other from a test set. The support set consists of set of molecules representing each class selected at random whenever a one-shot learning task is performed. The support set has compounds representing each one of the compound categories and the test set has the test molecule provided for classification. In order to access the ability to make accurate predictions, a test instance d_i of unknown class is selected. Knowing that only one instance d_j in our support set corresponds to that same class, the objective is to predict that class l belonging to D as the label l_i of an instance d_i.

Note that for every pair of input twins, the model generates a similarity score between 0 and 1 in one shot. Therefore, to evaluate whether the model is really able to recognize similar molecules and distinguish dissimilar ones, an N-way one shot learning strategy is used (Figure 5.4). The same molecule is compared to N different ones and only one of those matches the original input. Thus, the model returns N different similarity scores p_1, ..., p_N denoting the similarity between the test molecules and the support set molecules. If the pair of compounds of the same class gets the maximum score, the model prediction is correct. This process is repeated across multiple trials, the network accuracy being determined as the percentage of correct predictions.

In practice, for model validation a test instance is selected and compared to each one of the molecules in the support set.

Subsequently, the test instance is paired with each one of the compounds in the

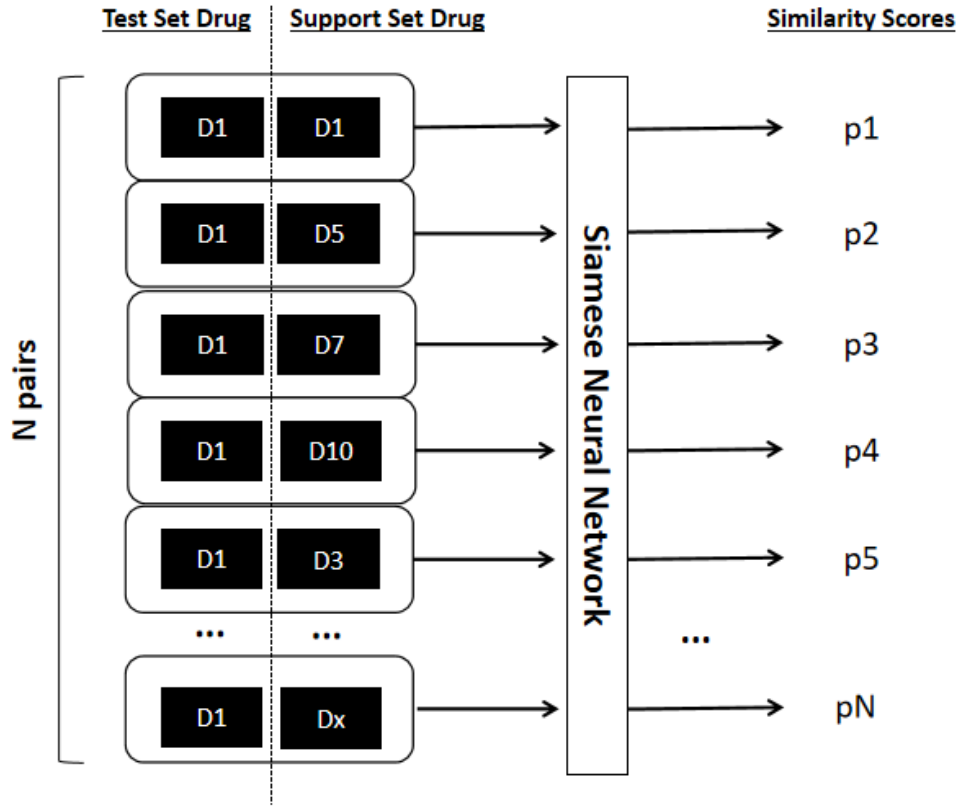


Figure 5.4: N-way one-shot learning approach.

support set and the compound with the highest similarity score according to the learned similarity rule is selected. The prediction is correct, if the maximum score corresponds to the pair of molecules of the same class, that is, if the test instance and the support set instance are from the same class. Thus, in each trial, the first pair corresponds to a pair of molecules of the same class, with the remaining pairs formed by molecules of different classes, as illustrated in Figure 5.5.

At each time a one-shot task is performed, the model verifies which of the molecules of a given class is more similar to a given test compound and predicts a similarity score according to the learnt similarity metric. Over multiple trials, in each one-shot task the Siamese network predicts which of the compounds present in the support set S most closely resembles the given test molecule in the test set T. Thus, the learned similarity function uses an $argmax$ function, since the model returns the maximum similarity score $score(d_i,d_j)$ of the pairs of compounds d_i,d_j of the test set T and support set S, respectively. The output prediction is given by,

$$pred(d_i,S) = argmax(score(d_i,d_j)) \ , \ d_j \in S \tag{5.7}$$

The prediction corresponds to the highest similarity score between the pairs of compounds in comparison in a given trial, in each of the N one-shot tasks performed. The value N refers to the number of comparisons (test molecule, support set molecule) at each trial. It is possible to verify that increasing this value, more challenging it becomes to obtain a correct prediction and lower is the accuracy of the model. This is due to the fact, that it is more difficult to obtain the maximum similarity score for the first pair due to the presence of a greater number of pairs in comparison at each trial.

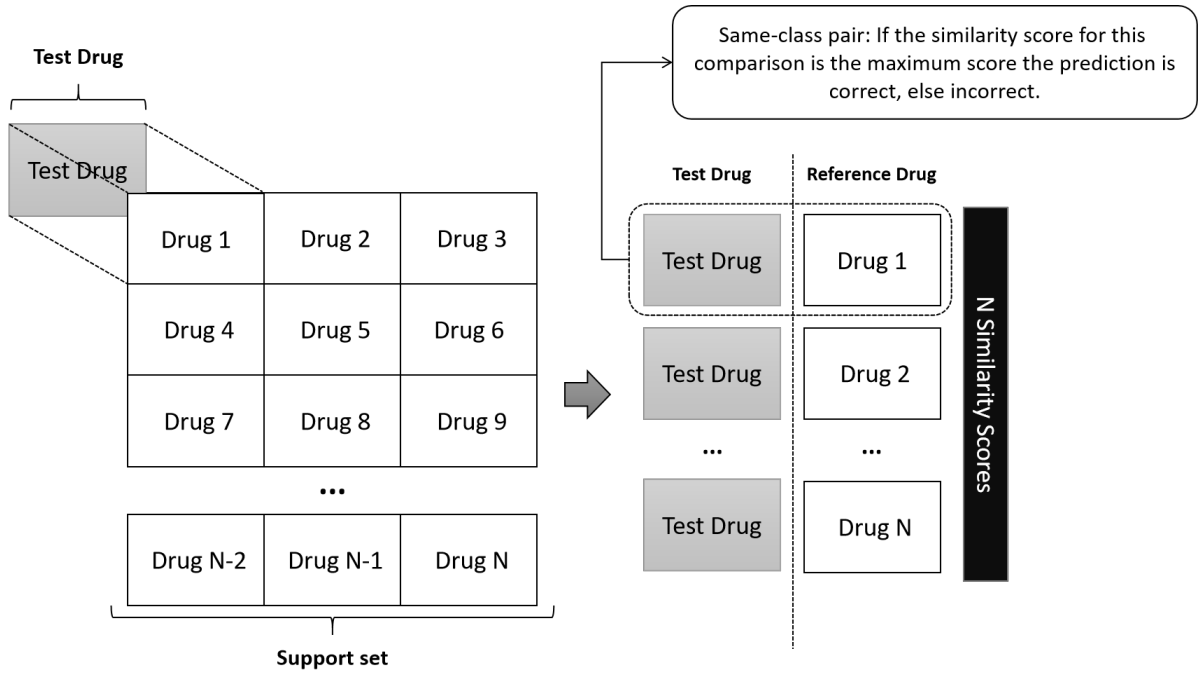


Figure 5.5: N-way one shot learning strategy for classification.

5.5 Model Comparison

The comparison of a given complex model with a set of simpler base models is a common strategy when assessing performance. Since methods such as Random Forest or Support vector machine are used across many machine learning applications and are the most popular approaches to solve both classification and regression problems, it was crucial to compare the proposed model with these traditional machine learning approaches. On the other hand, the model comparison allowed to evaluate the discrepancies between the application of deep learning techniques and traditional machine learning methods. This approach highlighted the capacity of deep neural networks in extracting features with high discriminating power and learning deep representations capable of describing the complexity of hidden patterns and local dependencies relevant to the prediction of novel compounds in drug discovery applications.

A training set in which half are pairs of the same class and another half of different classes was considered. To ensure a consistent and meaningful comparison, the accuracy was determined using a N-way one-shot learning strategy. A set of N-way one-shot tasks was established to compare N concatenated pairs across 500 trials.

Python 3.7.3 and Keras with Tensorflow [85] back-end were used to develop the standard convolutional model. Scikit-learn was used to implement some of the other models for comparison purposes [87].

5.5.1 K-Nearest Neighbour Classifier

The first model suggested for comparison is the K-Nearest Neighbour classifier [88]. KNN takes a test instance as an input and compares it with the k nearest neighbours. Each of these neighbors votes according to its label, returning the most voted label as the final prediction.

However, in the proposed model approach the compounds are organized in pairs and each test instance in each one-shot task has only 1 neighbor, its pair selected from the support set representing a given class of compounds. Therefore, in an N-way one shot learning approach, the test compound is compared with N different compounds selected from a support set, returning the pair with the highest similarity score as the final prediction. This type of classification is identical to the application of a KNN classifier, where the number of neighbors is equal to 1 ($K = 1$, the pair of the test instance in a one-shot task.

The Euclidean distance d, or L2 distance, was used to compute the distance between two input vectors a and b [89]:

$$d(a, b) = ||a - b|| = \sqrt{\sum_{i}^{n}(a_i - b_i)^2} \tag{5.8}$$

However, since the inputs of the classifier are binary matrices, encoded representations of the molecules, it is necessary to convert these representations into vectors on which a distance metric can be applied.

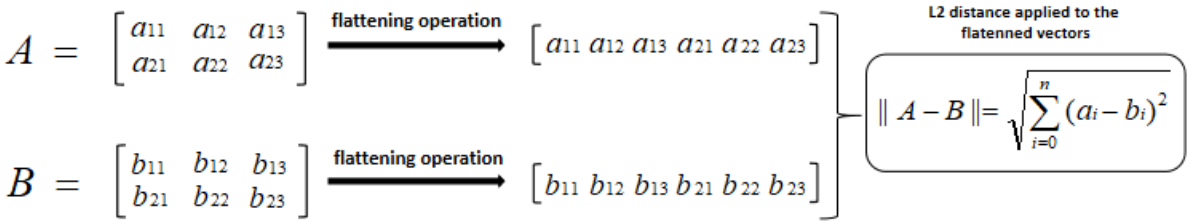


Figure 5.6: K-nearest neighbour: L2 distance applied to a pair of flattened encoded vectors A and B.

To compute the L2 distance between compounds the matrices are converted into flattened vectors and the distance metric is applied to each pair in comparison (Figure 5.6).

The Euclidean distances $d_1, d_2, ..., d_N$ between pairs of test set and support set

molecules are compared. Similarity is inversely proportional to a distance measure thus, a shorter distance corresponds to a greater similarity. That is, when the distance between a molecule and its pair is minimum, the similarity between both is maximum, and the class of the molecule pair is the classifier prediction. If a given compound in the support set belongs to the same class as the test molecule and, the distance obtained between them is the minimum, then the prediction is correct, as illustrated in Figure 5.7. Otherwise, it will be incorrect. This procedure was repeated over multiple trials with the accuracy determined as the average of correct predictions.

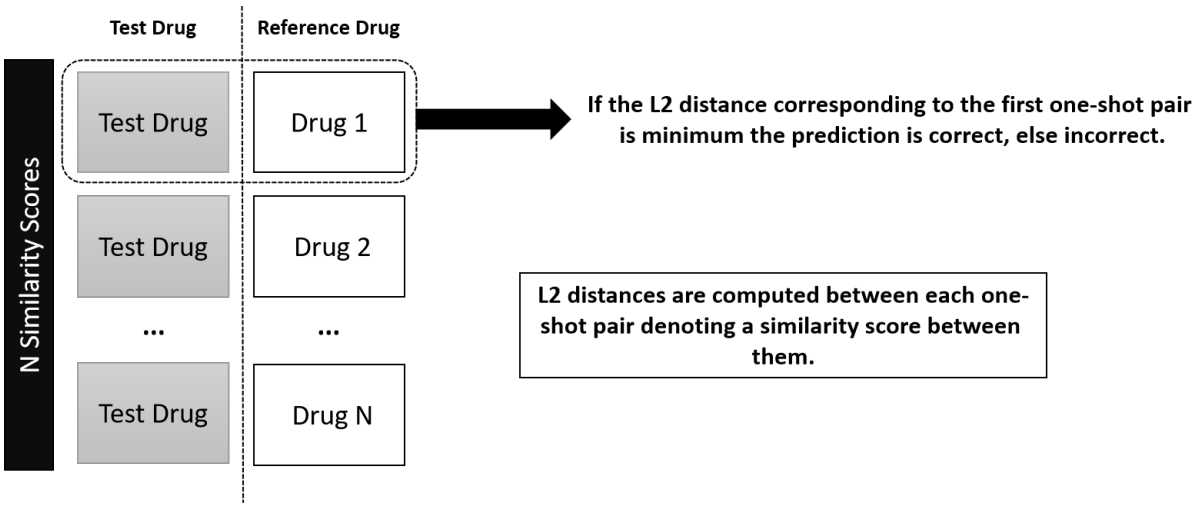


Figure 5.7: K-nearest neighbour model prediction.

5.5.2 Support Vector Machine

The purpose of Support Vector Machine (SVM) is to define a set of hyperplanes in a multidimensional space that can be easily used for classification purposes [90]. Thus, the objective is to find a hyperplane in an n-dimensional space that maximizes the separation margin between data points of different classes. For that purpose, a set of support vectors is defined. The support vectors influence the orientation and position of the hyperplane and maximize the separation margin of the classifier. The increase of the separation margin returns a higher probability of a correct prediction and a higher degree of confidence in the classification. The predicted label is the class next to the separation margin on which the test instances fall. In addition to the linear classification, an SVM model uses kernel functions to perform a non-linear classification, mapping a given input to an high-dimensional feature space. Therefore, SVM represents a given set of molecules by mapping them across different sides of a separation margin of a linear decision surface. (Figure 5.8).

The regularization parameter C represents a tolerance to misclassifications and establishes the number of misclassifications allowed. A higher value of C establishes a small separation margin that reduces the number of misclassifications, increasing the number of correct predictions. The gamma parameter establishes the set of data points that are involved in the calculation of the separation line. High values of gamma consider the points closest to the line in the calculation, low values consider also the points far away from the separation line. Kernel functions represent complex decision surfaces efficiently, calculating the separation lines in high-dimensional feature spaces. Thus, non-linearly separable features are converted to linearly separable features when mapped to a high-dimensional space.

Table 5.3 describes the parameter settings for the SVM model.

Table 5.3: Parameter values for the SVM Model.

Parameters	Value
C	1.0
kernel	RBF
gamma	scale
tol	0.001

The Radial Basis Function [91], an exponential kernel function, was used in the

implementation of the SVM model:

$$K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T.\phi(x^{(j)}) = \\ = exp(-\gamma.||x^{(i)} - x^{(j)}||^2), \gamma > 0 \tag{5.9}$$

Thus, it is not necessary to determine the feature mapping $\phi(x^{(i)})$ of the input $x^{(i)}$ directly, since the kernel trick maps to a larger dimensional space where features are linearly separable.

In this case, Scikit-learn [87] was used to implement the SVM classifier.

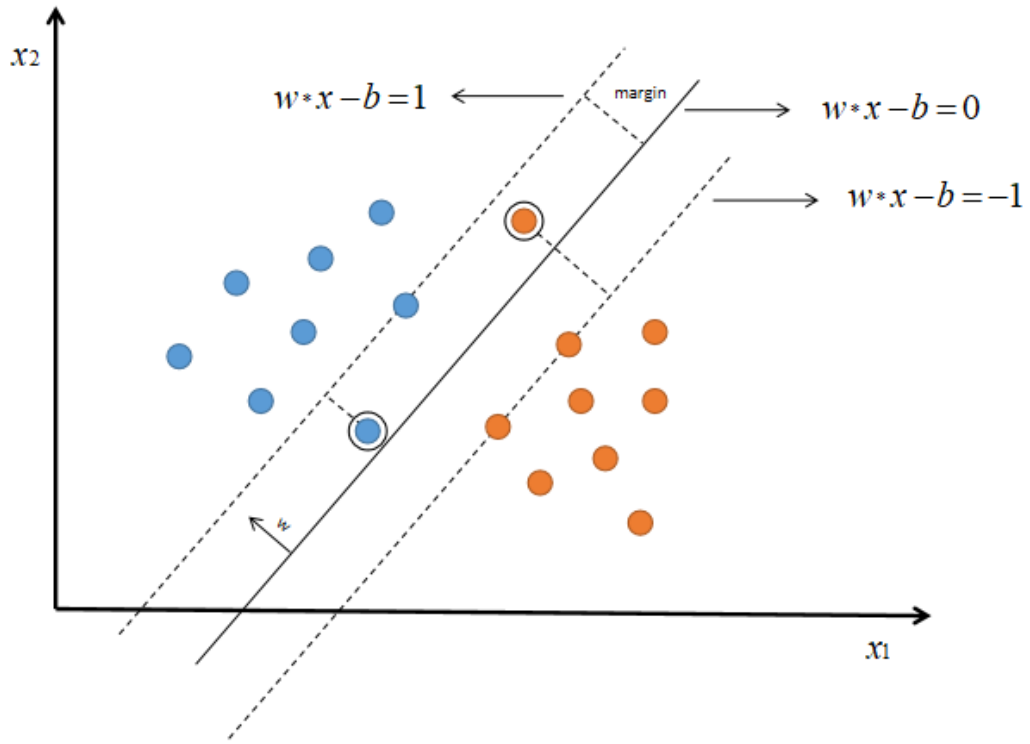


Figure 5.8: Support vector machine: hyperplanes and separation margin.

5.5.3 Random Forest

Random forest (RF) is an ensemble learning method that generates a set of individual classifiers returning the mode of the predictions of each individual decision tree [92]. Each individual tree returns a prediction that follows a set of 'if-then' rules through multiple decision splits. The selected splits generate less impurity and more information gain. The algorithm starts by randomly selecting a given subset of features, calculating the nodes by defining a split point. The node is divided into daughter nodes in a process that is repeated multiple times until we obtain a given set of nodes. The previous steps are repeated a number of k times, until we get a total number of k individual decision trees. Each decision tree corresponds to a different classifier, returning a prediction for the given set of test features. The final prediction of the algorithm will correspond to the most voted label given all the predictions made by each one of the individual trees (Figure 5.9).

In model training, each individual tree is trained using a randomly generated subset of data instead of the entire dataset. On the other hand, in each decision split only a randomly generated subset of features is considered. These techniques reduce the correlation and variance of the model while decreasing noise sensitivity and preventing overfitting, making the model more robust to features that greatly influence the model prediction. By injecting randomness into forests, it allows to decouple errors when considering an average of the predictions of the individual trees as the final result of the model.

Table 5.4 describes the parameter settings for the RF model.

Table 5.4: Parameter settings for the random forest model.

Parameters	**Value**
n_estimators	150
max_features	100
imp_metric	Gini
max_depth	None
min_split	2
min_node	1
limit_leaf_nodes	None

Parameters as the number of individual decisions trees (n_estimators), minimum number of samples needed for splitting and to form a node (min_split and min_node, respectively), maximum number of nodes (limit_leaf_nodes), the maximum number of features selected randomly at each decision split (max_features), evaluation crite-

ria to measure the impurity (imp_metric) or the maximum depth of the individual trees (max_depth) are some of the RF main hyperparamaters to fine-tune before training.

Scikit-learn [87] was used for the implementation of the RF classifier.

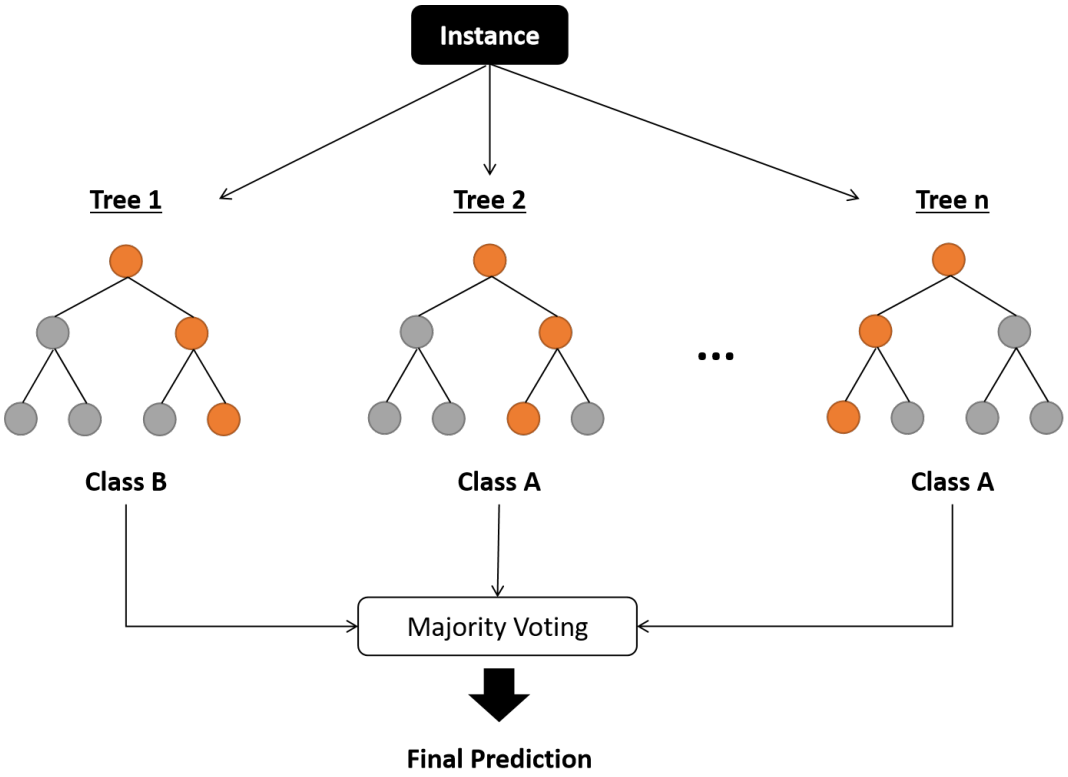


Figure 5.9: Random forest model.

5.5.4 Naïve Model

The Naïve model was implemented in order to demonstrate that the results obtained outperform a random prediction. Thus, random numbers were generated for each pair of molecules in each trial of a one-shot task (Figure 5.10).

This random number represents a randomly generated similarity score that denotes a similarity score between the test molecules and the support set molecules. As previously mentioned, the first pair of a one-shot task represents a pair of molecules of the same class. If the number generated for the first pair is maximum then the prediction is correct, else it is incorrect.

This procedure was also repeated over multiple trials with the accuracy determined as the average of correct predictions.

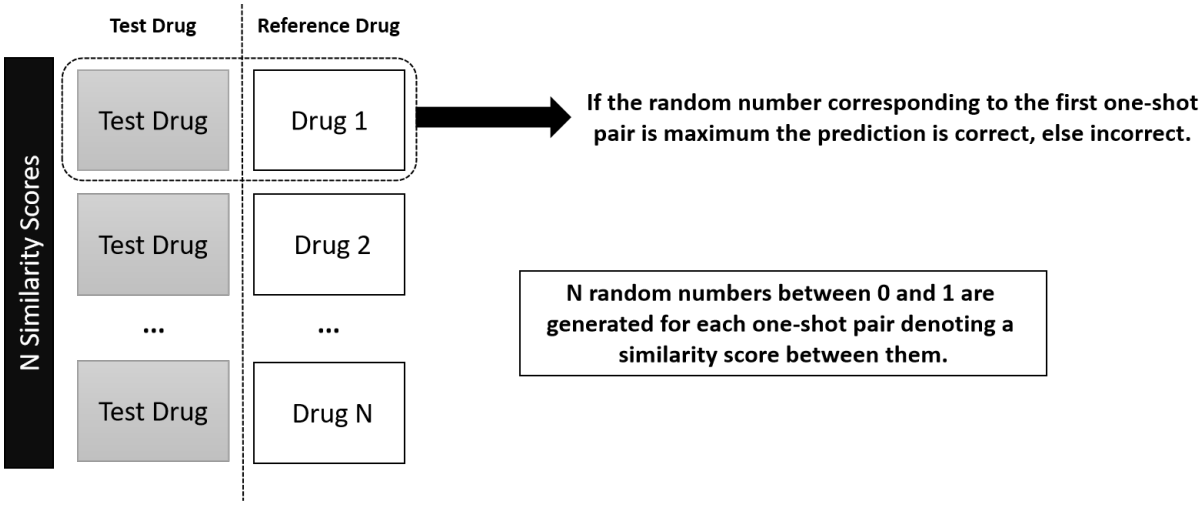


Figure 5.10: Naïve model prediction.

5.5.5 Multi-Layer Perceptron

Multi-Layer Perceptron (MLP) is a simple neural network with the following structure: input layer, a set of hidden layers and an output layer [93] (Figure 5.11). This model is a simple neural network model capable of modeling input-output dependencies and learning useful correlation relationships. Thus, an MLP model with only one hidden layer between the input and output layer allows to model any continuous function and is the basis for the development of more complex models such as CNNs.

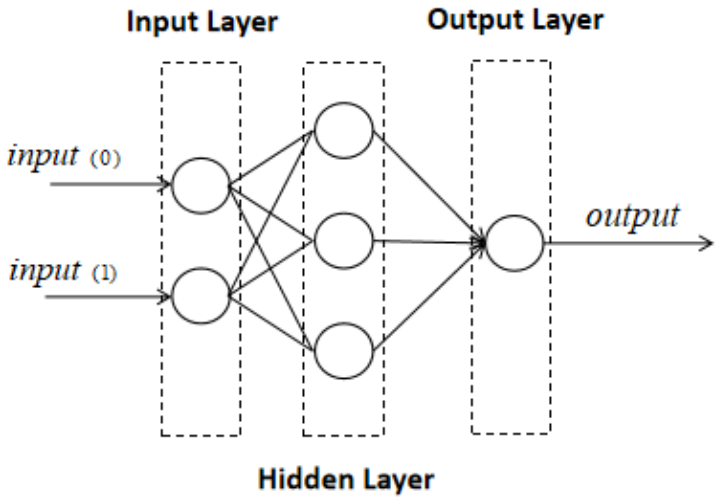


Figure 5.11: Multi-layer perceptron: simple perceptron model with 1 hidden layer.

The implementation of MLP allowed the comparison between the proposed model and a simpler deep learning model with a smaller set of hidden layers. For the MLP architecture 3 hidden layers with 100, 50 and 30 neurons were considered.

Table 5.5 describes the parameter settings for the MLP model.

Table 5.5: Network parameter settings for the MLP model.

Parameter	Value
Number of hidden layers	3
Number of neurons	[100,50,30]
Epochs	100
Optimizer	Adam
Learning rate	0.001
L2 penalty parameter	0.0001
Activation function	ReLU

Scikit-learn [87] was used for the implementation of the MLP.

5.5.6 Convolutional Neural Network (CNN)

Finally, the performance of a standard convolutional neural network is evaluated in order to analyse the differences when compared to a pair of identical parallel convolutional neural networks with the same structure. In this case, a neural network architecture identical to each one of the Siamese twins of the proposed Siamese model is considered (Figure 5.6). The convolutional model is prepared to receive a concatenated pair of molecules for classification. The compound pairs are propagated through a single network in order to extract deep representations from SMILES useful for compound prediction.

The comparison of the proposed model with a convolutional model highlights the differences when learning a similarity metric applied to pairs of compounds. The proposed model propagates two input molecules through identical neural networks with the same set of parameters and learns a distance metric applied to both inputs. However, the convolutional model learns a single set of weights applied to a concatenated input, where each encoded molecule is convolved with a different set of filters and applied to a different set of neurons in each iteration of the training process, breaking the symmetry property of the Siamese model.

Table 5.6 describes the parameter settings for the CNN model.

Table 5.6: Network parameter settings for the CNN model * Initial number of epochs, however early stopping and model checkpoint were applied.

Parameter	Value
Number of convolutional layers	4
Number of dense layers	2
Number of filters	[64,128,128,256]
Filter length	[10,7,2,2]
Epochs*	100
Batch size	50
Optimizer	Adam
Learning rate	0.0001
Number of neurons (FC)	1024
Activation function (CNN)	ReLU
Activation function (FC)	ReLU
Activation function (Output)	Sigmoid Function
Loss Function	Binary Cross Entropy

5.6 Statistical Significance Analysis

The choice of a statistical hypothesis test to compare the results of multiple classifiers is a challenging issue in machine learning. Since it was very expensive and time-consuming to train multiple deep learning models, we chose the McNemar's Chi-Square test [94] to analyse the statistical significance of the results as recommended by Dietterich et al. (1998) [95]. The proposed Siamese model considers multiple combinations of pairs of molecules to train a similarity metric learnt by a set of parallel convolutional neural networks requiring a lot of training time for the network to converge to an accurate and reliable prediction. McNemar's Chi-Square test is a statistical hypothesis test that operates over the results of multiple classifiers models considering only one test dataset.

Normality and statistical correlation tests were conducted in order to validate the use of a Chi-Square non-parametric test to evaluate the model results. Shapiro-Wilk test [96] was used to evaluate normality and Spearman correlation coefficient [97] to test correlation. Both were applied to the model prediction results in comparison. The $p - value$ results ($p - value \leq 0.05$) for Shapiro and ($p - value > 0.05$) for Spearman coefficient, indicate that a non-parametric test must be used to perform the statistical significance analysis of the classifier model results. McNemar Chi-Square test is a paired non-parametric test that operates upon a contingency table (Table 5.7).

The contingency table operates over 2 categorical variables (binary variables) which, in this case, correspond to the correct and incorrect predictions in each N-way one-shot learning task over a number of k trials. For each set of paired molecules in each trial the classifier prediction is evaluated returning a correct of incorrect prediction. The number of k predictions of the classifier model in each one-shot task over k trials is used to build the contingency table.

Table 5.7: Mcnemar's contingency table.

	Classifier 2 Correct	**Classifier 2 Incorrect**
Classifier 1 Correct	Yes/Yes	Yes/No
Classifier 1 Incorrect	No/Yes	No/No

The contingency table is a result of the comparison between two distinct classifier results given the number of correct predictions per an N-way one shot task done over multiple trials. Since we are comparing pairs of independent and mutually exclusive predictions from two different classifiers for each N-way one shot classification task

and the pairs selected from the test set and the support set are chosen at random, this was the most appropriate statistical test selected for the significance analysis of the model results.

The McNemar's Chi-Square test analyses the contingency table values and evaluates the homogeneity of the prediction results for both classifier models in comparison, inferring about the disagreement between the results. The test statistic is based on a Chi-Square distribution with only 1 degree of freedom.

Mcnemar's Chi-Square test statistic is given by,

$$statistic = \frac{(Yes/No - No/Yes)^2}{(Yes/No + No/Yes)} \tag{5.10}$$

The test statistic is based on the disagreement between both classifier results and on the difference between the correct and incorrect predictions returned, capturing the differences in the proportion of the errors. Therefore, the accuracy and the error of the individual classifier does not influence the disagreement or the statistical test assumptions. The null hypothesis $H0$ is that the two cases disagree to the same amount. If we reject the null hypothesis, it suggests that there is evidence that the cases disagree in different ways and there is a statistical significant difference between the model results. The $p - value$ can be interpreted as follows:

1. $p > \alpha$: fail to reject $H0$, no difference in the disagreement, the models have a similar proportion of errors.

2. $p \leq \alpha$: reject $H0$, significant difference in the disagreement, there is a different proportion of errors between the models.

In Table 6.3 in the next section, we display the $p - value$ results for the statistical significance of the proposed Siamese model results considering $N <= 10$. A level of significance of 1% is considered with $\alpha = 0.01$. The statistical hypothesis tests for the significance analysis were performed using SciPy [98] and Statsmodels [99] libraries.

graphicx

6

Results and Discussion

In the context of the prediction of promising lead molecules, identifying different classes of compounds with a small set of candidate molecules should be the main focus of this study. Nonetheless, the model must be able to discriminate new classes of compounds unknown at the training stage, eliminating the need for periodic re-training of the model and leading to the prediction of novel compounds for drug discovery applications. Additionally, the proposed model which learns a one-shot similarity metric must be able to discriminate very identical compounds and compounds with very high disparities in an equally effective and accurate manner. Thus, the model should be able to distinguish between organized pairs of compounds according to the dissimilarities observed in the structure and molecular descriptors and predict a given set of structural analogs with increased biological activity and therapeutic function.

In this study, a convolutional Siamese model is proposed to predict novel compounds according to a degree of similarity between molecules. The similarity rule is computed by a similarity function learnt by a one-shot Siamese neural network built upon two parallel and identical convolutional neural networks capable of extracting deep representations from SMILES.

As highlighted in Section 5.5, the proposed model is compared with the traditional machine learning approaches most commonly used in classification and regression problems. A comparison with a set of simple deep learning models and standard machine learning approaches highlights the discriminating power of the one-shot structural features extracted by the Siamese model.

To measure the model performance, the accuracy was determined using a N-way one-shot learning strategy, described in the previous sections.

$$accuracy\ (\%) = \frac{number\ of\ correct\ predictions}{number\ of\ trials\ per\ one-shot\ task}\ .\ 100 \tag{6.1}$$

The differences in performance between the traditional machine learning models, standard deep learning models and the proposed Siamese model can be interpreted as a result of the reduced training set, the attempt to classify novel compounds whose classes are unknown at training stage and the ability to propagate a set of compound pairs along two parallel neural networks. Therefore, in the context of the prediction of promising lead compounds using a one-shot Siamese model, there are 4 different questions that need to be answered:

1. **How useful is the application of a set of parallel convolutional neural networks using a one-shot learning strategy?**

In this study, we validate the potential of learning a similarity metric capable of classifying a set of compound pairs using a one-shot Siamese deep neural network. This type of neural network accepts data organized in pairs, increasing the size of the dataset, given the total number of possible combinations for the pairs of molecular structures available for training, preventing overfitting. Since the Siamese model accepts a set of compound pairs, the dataset size increases, given the total number of possible combinations for the pairs of molecules. This number quadratically increases with the number of classes and linearly with the number of examples per class available for training.

The performance results of the Siamese model highlight its ability to propagate a pair of input molecules along two parallel and identical neural networks with the same structure (achieved the highest accuracy (94%) for $N = 2$ and the accuracy does not fall bellow 60% for $N \leq 10$) (Table 6.1). Therefore, rather than propagating a concatenated and denser input across a single network, the same set of weights is propagated along two identical twin neural networks, making the learning task more effective and less computationally expensive. The Siamese model propagates two inputs across the same set of weights and the difference between the output feature vectors serves as an input to a similarity metric. This symmetry-based approach is less expensive and leads to a pairwise training which improves the model prediction accuracy. Therefore, the model generalization capacity increases when assuming the existence of a higher number of examples since there is a large and more diverse set of structures to be learned.

Conversely, considering a single convolutional neural network for the training of a similarity metric, there is a concatenated and denser input, which is propagated through the layers, each element is convoluted with a different set of filters, which breaks the symmetry required to compute a similarity score between two molecular structures. It is possible to verify the effect of symmetry breakdown on final accuracy

values obtained by the standart convolutional neural network (above 50% only for $N \leq 3$ and dropping abruptly for $N > 3$).

Table 6.1: N-way one-shot learning accuracy results for the Siamese model.

	N					
	2	**3**	**4**	**5**	**7**	**10**
Siamese Neural Network (validation)	**94%**	**90%**	**84%**	**78%**	**70%**	**65%**
Siamese Neural Network (training)	**95%**	**92%**	**86%**	**84%**	**72%**	**70%**

2. **What is the main reason behind the drop in the accuracy results for higher values of N?**

The final prediction corresponds to the pair of compounds with the highest similarity score in a one-shot trial. The value N refers to the number of pairs of molecules at each trial in a N-way one-shot task. It is possible to verify that increasing this value, more challenging it becomes to make a correct prediction and lower is the accuracy of the learnt similarity rule. This is due to the fact that it is more difficult to obtain the maximum similarity score for the first pair of compounds in comparison due to the presence of a greater number of pairs at each trial.

Therefore, if the number of comparisons N in a one-shot trial increases, there is a greater chance that the model fails the prediction. This is due to the fact that a higher number of classes is assigned to each test instance. This behavior increases the probability of a failed prediction which is visible in the accuracy results as we progress to higher values of N. As a consequence, we observe a drop in the network's performance for higher values of N. However, the Siamese Network achieved the highest accuracy (94%) for $N = 2$ and it does not fall below 60% for $N \leq 10$ (Table 6.4).

3. **How discriminating is the proposed model in comparison with the traditional machine learning models and simple deep learning methods?**

Traditional machine learning as SVM and Random Forest models present accuracy results that highlight the effectiveness of deep learning models in extracting useful features for compound prediction over these conventional machine learning approaches. The accuracy achieved by these methods remains above 50% for lower values of N. However, it drops abruptly when N increases since it is more difficult to make a correct prediction (Table 6.2). These machine learning algorithms consider the sequence as a whole, disregarding specific regions in the molecule structure and structural inter-dependency relations between them. These molecular segments are crucial to recognize different classes of compounds and to identify local dependencies

related with class-specific segments of the compound chemical structure.

CNN and MLP return a more consistent performance among the models built for comparison, achieving higher accuracy results for $N \leq 10$. The multi-layer perceptron achieves an accuracy above 60% for $N \leq 3$. However, it drops abruptly for higher values of N. These performance results highlight the inability to outperform the proposed set of parallel convolutional neural networks compatible with the set of compound pairs provided for training (Table 6.2).

The proposed one-shot Siamese model presents a set of parallel convolutional neural networks that propagate 2 compound SMILES and extract features from deep representations whose complexity increases as we progress to deeper layers. The extraction of low-level features supports the extraction of patterns and local dependencies of increasing complexity. These high-level representations identify patterns of the molecular structures shared among compounds and certain classes of molecules. The local patterns identified in the deep representations support the prediction of pharmacological analogs outperforming the standart machine learning methods and simple deep neural networks (Table 6.4).

Table 6.2: N-way one-shot learning accuracy results for standart machine learning methods and simple deep learning approaches.

	N					
	2	**3**	**4**	**5**	**7**	**10**
KNN	70%	55%	49%	43%	36%	30%
Naïve Model	61%	43%	34%	31%	22%	19%
SVM	56%	42%	30%	24%	16%	12%
Random Forest	71%	58%	60%	44%	34%	20%
Multi-Layer Perceptron	76%	60%	36%	34%	22%	13%
Convolutional Neural Network	81%	70%	58%	46%	41%	39%

4. **What does the statistical significance analysis conducted demonstrate and to what extent are the results obtained statistically significant?**

The analysis of the statistical significance of the model results in comparison (Table 6.3) proves that there is evidence of a statistically significant difference between model performances. The p-value results indicate a different proportion of errors on the test set, which corroborates the validity of the results achieved. Considering 1% as the level of significance ($\alpha = 0.01$), the results show a highly statistically significant difference in the disagreements between the Siamese model against the other model performances.

Therefore, the results achieved prove that the one-shot Siamese neural network

outperforms the state-of-the-art models in the accurate and reliable prediction of novel compounds for drug discovery applications.

Table 6.3: $p - value$ results for statistical significance of the Siamese model.

		N					
		2	**3**	**4**	**5**	**7**	**10**
Siamese Model	**CNN**	5.684e-10	1.058e-14	2.700e-19	1.034e-23	2.024e-19	2.448e-15
	MLP	9.052e-15	1.414e-27	2.416e-50	2.285e-44	1.343e-46	3.063e-59
	KNN	1.00e-23	1.720e-35	3.748e-30	1.855e-27	5.620e-26	6.621e-28
	Naïve	4.282e-37	1.580e-52	1.907e-55	2.764e-44	5.189e-49	9.457e-42
	RF	8.098e-22	5.888e-30	2.709e-17	5.374e-28	9.322e-29	3.268e-42
	SVM	1.172e-40	2.547e-53	8.606e-61	3.762e-58	1.519e-62	1.366e-60

The presence of a reasonable number of layers provides the ability of convolutional neural networks to retain information and extract important features for the detection of patterns relevant for the prediction. The use of parallel convolutional neural networks implementing a one-shot learning approach is compatible with the pairs of compounds in comparison. The learnt similarity rule leads to the discovery of compound analogs with increased biological activity. This strategy enhances the learning of a similarity function capable of inferring a set of promising lead molecules from the structure and molecular descriptors of a small set of candidate molecules.

The discovery of new analogous compounds lies in the ability to identify patterns and extract features capable of modelling the relationship between certain segments of the molecular structure and the compound classes and, consequently, their biological activity. Therefore, the high predictive power of one-shot Siamese neural networks lies in their ability to identify patterns or local dependencies related with certain segments of the molecular structure which play a pivotal role in the identification of certain classes of molecular structures. The deep representations of compounds

Table 6.4: Final N-way one-shot learning accuracy results.

	N					
	2	**3**	**4**	**5**	**7**	**10**
Siamese Neural Network (validation)	**94%**	**90%**	**84%**	**78%**	**70%**	**65%**
Siamese Neural Network (training)	**95%**	**92%**	**86%**	**84%**	**72%**	**70%**
KNN	70%	55%	49%	43%	36%	30%
Naïve Model	61%	43%	34%	31%	22%	19%
SVM	56%	42%	30%	24%	16%	12%
Random Forest	71%	58%	60%	44%	34%	20%
Multi-Layer Perceptron	76%	60%	36%	34%	22%	13%
Convolutional Neural Network	81%	70%	58%	46%	41%	39%

model non-linear relationships between the molecular structure and its biological activity important for the prediction of structural analogs with high therapeutic potential and increased pharmacological activity. The results highlight the efficiency of using deep representations in compound prediction extracted from a set of parallel convolutional neural networks using a one-shot learning strategy.

7

Conclusion

7.1 Conclusion

Humans are able to one-shot learn complex concepts related to a given object, image or symbol from the previous observation of multiple representations of other similar and equally complex concepts. This representation learning helps in the classification of newly observed instances of those concepts. However, it is difficult to find a deep learning model capable of simulating this similarity learning from very few training examples per class for training and in the absence of a previous observation that leads to a correct prediction.

Deep neural networks are able to learn a wide range of functions and possible representations from a large set of parameters. Deep learning methods are capable of adjusting the designed model to complex data-structures while extracting multiple sets of features valuable for the prediction. However, this set of parameters induces the learning of a huge number of possible representations and mappings for the observed instance. Thus, how is it possible to find a representation that generalizes in order to correctly classify a given test instance when there is only one previously observed example available?

A one-shot learning approach learns a small set of useful features from the simple similarity measure between two distinct inputs. Thus, it is possible to learn a similarity function from a set of pairs of instances by extracting deep representations related with the dissimilarity between compounds according to the similarity score returned by the learnt metric.

All deep neural networks develop structure and extract deep representations capable of improving the model prediction. The set of layers of a neural network stratifies the complexity of the extracted features, with the simplest features extracted from low-level layers and the most complex from the high-level layers. In the context of

drug-discovery, low-level layers emerge as detectors of small segments of the molecular structures which characterizes the compound primary structure. The layers that follow built upon the previous layers, are able to detect structural differences between portions of the previously identified segments, as well as relevant combinations of those segments. The increase in feature complexity results in the detection of patterns and local dependencies related with the active binding sites or capable of describing different compound classes.

One-shot deep neural networks extract relevant features for the prediction and extrapolate to different classes of compounds according to a degree of similarity with a given reference molecule representing each compound class, starting from a small set of similar drugs.

In this study, we validate the potential of learning a similarity metric capable of classifying a set of compound pairs using a one-shot Siamese deep neural network. This type of neural network accepts data organized in pairs, increasing the size of the dataset and preventing overfitting, given the total number of possible combinations for the pairs of molecular structures available for training.

The discriminating power of the similarity rule improves the ability to extract features related to specific regions of molecular structures useful for compound prediction. The learnt one-shot structural features improves the ability to extract information and to infer about complex concepts of the molecular structures. The high discriminating power of the learnt features extracted from a reduced set of candidate compounds is a property that results from the combination of a one-shot learning approach and a model built as a set of parallel convolutional neural networks.

The performance of the proposed model outperforms the traditional classification approaches such as KNN or a naïve model, the traditional machine learning methods such as SVM and Random Forest and simple deep learning methods such as CNN or MLP. Additionally, we validate the effectiveness of a one-shot learning approach in the accurate and reliable prediction of novel compounds in low-data drug discovery applications.

One-shot learning exploits the application of deep learning strategies into solving problems where with only a few examples per class for training. In the context of drug discovery, where high-quality data is scarce, the implementation of this strategy represents a vital asset to significantly increase the predictive power when inferring the properties and activities of small molecules and those of their pharmacological analogs.

7.2 Future Work

In future research, improvements could be made if the compounds in the test set were compared to the support set as a whole, instead of considering its elements individually. Therefore, it might be feasible trying to understand the influence of the presence of a significant number of similar compounds in a support set while predicting novel compounds in a test set, increasing the predictive power of the proposed model.

Therefore, in a one-shot parallel neural network architecture, the vector embeddings of both test and support set molecules, in each one-shot task takes into account only the structure of each pair and not the support set as a whole. However, it would be important to understand the influence that every support set element has on the classification of each pair of compounds. A possible solution would be the application of a full-context embedding, where the structure of the test molecule is influenced by each one of the elements of the support set. This could result in an improvement of the performance of the model, where for different one-shot tasks, the vector embeddings of the test molecule change according to the structure of the support set elements.

The one-shot learning approach used to train a similarity function capable of discriminating between different classes of molecules is compatible with the development of a model able to predict novel compounds and structural analogs with increased biological activity. Thus, it is possible to insert the proposed model in the context of the multi-class classification of compounds. Accepting different pairs of molecules and, according to the score returned by the learnt similarity rule, it is possible to predict the corresponding class and, consequently, its biological activity on an identified target molecule. The application of a parallel neural network architecture capable of accepting different pairs of molecules can be implemented in the prediction of structural analogs. Thus, starting from a set of reference molecules it may be possible to predict, according to the dissimilarity between pairs of molecular structures, the pharmacological properties and biological activities of a set of compounds from unknown classes. Additionally, it may be feasible trying to predict the optimal molecular structure for a previously identified target molecule, leading to the discovery of promising lead compounds capable of triggering the intended physiological response in organs and tissues.

Bibliography

[1] P. Bongrand, "Ligand-receptor interactions," *Reports on Progress in Physics*, 1999.

[2] C. E. Tinberg, S. D. Khare, J. Dou, L. Doyle, J. W. Nelson, A. Schena, W. Jankowski, C. G. Kalodimos, K. Johnsson, B. L. Stoddard, and D. Baker, "Computational design of ligand-binding proteins with high affinity and selectivity," *Nature*, 2013.

[3] B. J. Pleuvry, "Receptors, agonists and antagonists," *Anaesthesia and Intensive Care Medicine*, 2004.

[4] T. Kenakin, "Overview of receptor interactions of agonists and antagonists," 2008.

[5] N. X. Wang and H. A. Von Recum, "Affinity-Based Drug Delivery," 2011.

[6] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert, "A lock-and-key model for protein-protein interactions," *Bioinformatics*, 2006.

[7] J. Boyle, "Lehninger principles of biochemistry (4th ed.): Nelson, D., and Cox, M.," *Biochemistry and Molecular Biology Education*, 2005.

[8] P. van der Geer, "Signal Transduction," in *Brenner's Encyclopedia of Genetics: Second Edition*, 2013.

[9] D. Hilger, M. Masureel, and B. K. Kobilka, "Structure and dynamics of GPCR signaling complexes," *Nature Structural and Molecular Biology*, 2018.

[10] R. Zhang and X. Xie, "Tools for GPCR drug discovery," 2012.

[11] M. A. Lemmon and J. Schlessinger, "Cell signaling by receptor tyrosine kinases," 2010.

[12] V. Suppiramaniam, J. Bloemer, M. Reed, and S. Bhattacharya, "Ion Channels," in *Comprehensive Toxicology: Third Edition*, 2018.

[13] M. K. Bates and R. M. Kerr, *Nuclear receptors*. 2011.

[14] J. J. Lertora and K. M. Vanevski, "Introduction to pharmacokinetics and pharmacodynamics," in *Small Molecule Therapy for Genetic Disease*, 2010.

[15] T. Kenakin, "Principles: Receptor theory in pharmacology," 2004.

[16] H. Lu and P. J. Tonge, "Drug-target residence time: Critical information for lead optimization," 2010.

[17] A. L. Hopkins, "Network pharmacology: The next paradigm in drug discovery," 2008.

[18] G. R. Zimmermann, J. Lehár, and C. T. Keith, "Multi-target therapeutics: when the whole is greater than the sum of the parts," 2007.

[19] A. Masoudi-Nejad, Z. Mousavian, and J. H. Bozorgmehr, "Drug-target and disease networks: polypharmacology in the post-genomic era," *In Silico Pharmacology*, 2013.

[20] J. P. Hughes, S. S. Rees, S. B. Kalindjian, and K. L. Philpott, "Principles of early drug discovery," 2011.

[21] W. L. Jorgensen, "Efficient drug lead discovery and optimization," *Accounts of Chemical Research*, 2009.

[22] K. H. Bleicher, H. J. Böhm, K. Müller, and A. I. Alanine, "Hit and lead generation: Beyond high-throughput screening," 2003.

[23] R. Duelen, M. Corvelyn, I. Tortorella, L. Leonardi, Y. C. Chai, and M. Sampaolesi, "Medicinal Biotechnology for Disease Modeling, Clinical Therapy, and Drug Discovery and Development," in *Introduction to Biotech Entrepreneurship: From Idea to Business*, 2019.

[24] H. van de Waterbeemd and E. Gifford, "ADMET in silico modelling: Towards prediction paradise?," 2003.

[25] K. I. Kaitin, "Deconstructing the drug development process: The new face of innovation," 2010.

[26] S. Ekins, J. Mestres, and B. Testa, "In silico pharmacology for drug discovery: Applications to targets and beyond," 2007.

[27] J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, and S. Zhao, "Applications of machine learning in drug discovery and development," 2019.

[28] I. M. Kapetanovic, "Computer-aided drug discovery and development (CADDD): In silico-chemico-biological approach," *Chemico-Biological Interactions*, 2008.

[29] H. Jhoti and A. R. Leach, *Structure-based drug discovery*. 2007.

[30] H. Kubinyi, "Structure-based drug design," *Chimica Oggi*, 1998.

[31] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, "Molecular dynamics simulations: Advances and applications," 2015.

[32] G. Wu, D. H. Robertson, C. L. Brooks, and M. Vieth, "Detailed analysis of grid-based molecular docking: A case study of CDOCKER - A CHARMm-based MD docking algorithm," *Journal of Computational Chemistry*, 2003.

[33] H. Claußen, C. Buning, M. Rarey, and T. Lengauer, "FLEXE: Efficient molecular docking considering protein structure variations," *Journal of Molecular Biology*, 2001.

[34] G. Bhardwaj, V. K. Mulligan, C. D. Bahl, J. M. Gilmore, P. J. Harvey, O. Cheneval, G. W. Buchko, S. V. Pulavarti, Q. Kaas, A. Eletsky, P. S. Huang, W. A. Johnsen, P. J. Greisen, G. J. Rocklin, Y. Song, T. W. Linsky, A. Watkins, S. A. Rettie, X. Xu, L. P. Carter, R. Bonneau, J. M. Olson, E. Coutsias, C. E. Correnti, T. Szyperski, D. J. Craik, and D. Baker, "Accurate de novo design of hyperstable constrained peptides," *Nature*, 2016.

[35] S. C. Pegg, J. J. Haresco, and I. D. Kuntz, "A genetic algorithm for structure-based de novo design," *Journal of Computer-Aided Molecular Design*, 2001.

[36] S. J. Y. Macalino, V. Gosu, S. Hong, and S. Choi, "Role of computer-aided drug design in modern drug discovery," 2015.

[37] S. Y. Yang, "Pharmacophore modeling and applications in drug discovery: Challenges and recent advances," 2010.

[38] R. Perkins, H. Fang, W. Tong, and W. J. Welsh, "Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology," 2003.

[39] C. Nantasenamat, C. Isarankura-Na-Ayudhya, T. Naenna, and V. Prachayasittikul, "A practical overview of quantitative structure-activity relationship," 2009.

[40] J. Mestres, D. C. Rohrer, and G. M. Maggiora, "MIMIC: A molecular-field matching program. Exploiting applicability of molecular similarity approaches," *Journal of Computational Chemistry*, 1997.

[41] T. M. Steindl, D. Schuster, C. Laggner, and T. Langer, "Parallel screening: A novel concept in pharmacophore modeling and virtual screening," *Journal of Chemical Information and Modeling*, 2006.

[42] P. De Cerqueira Lima, A. Golbraikh, S. Oloff, Y. Xiao, and A. Tropsha, "Combinatorial QSAR modeling of P-glycoprotein substrates," *Journal of Chemical Information and Modeling*, 2006.

[43] X. Yang, Y. Wang, R. Byrne, G. Schneider, and S. Yang, "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," 2019.

[44] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: progress in machine intelligence for rational drug discovery," 2017.

[45] Z. R. Yang and Z. Yang, "Artificial Neural Networks," in *Comprehensive Biomedical Physics*, 2014.

[46] "Cs231n: Convolutional neural networks for visual recognition," 2020. `http://cs231n.stanford.edu`.

[47] K. Preuer, G. Klambauer, F. Rippmann, S. Hochreiter, and T. Unterthiner, "Interpretable Deep Learning in Drug Discovery," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[48] A. Lusci, G. Pollastri, and P. Baldi, "Deep architectures and deep learning in chemoinformatics: The prediction of aqueous solubility for drug-like molecules," *Journal of Chemical Information and Modeling*, 2013.

[49] T. Unterthiner, A. Mayr, G. Klambauer, and S. Hochreiter, "Toxicity Prediction using Deep Learning," *arXiv*, 2015.

[50] M. Ohue, Ryôta, K. Yanagisawa, and Y. Akiyama, "Molecular activity prediction using graph convolutional deep neural network considering distance on a molecular graph," *ArXiv*, vol. abs/1907.01103, 2019.

[51] I. Wallach, M. Dzamba, and A. Heifets, "Atomnet: A deep convolutional neural network for bioactivity prediction in structure-based drug discovery," 2015.

[52] J. C. Pereira, E. R. Caffarena, and C. N. Dos Santos, "Boosting Docking-Based Virtual Screening with Deep Learning," *Journal of Chemical Information and Modeling*, 2016.

[53] B. Raymer and S. K. Bhattacharya, "Lead-like Drugs: A Perspective," 2018.

[54] D. J. Rezende, S. Mohamed, I. Danihelka, K. Gregor, and D. Wierstra, "One-Shot generalization in deep generative models," in *33rd International Conference on Machine Learning, ICML 2016*, 2016.

[55] R. R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba, "One-Shot Learning with a Hierarchical Nonparametric Bayesian Model," *JMLR Workshop and Conference Proceedings*, 2012.

[56] B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum, "One-shot learning by inverting a compositional causal process," in *Advances in Neural Information Processing Systems*, 2013.

[57] O. Vinyals, C. Blundell, T. Lillicrap, K. Kavukcuoglu, and D. Wierstra, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, 2016.

[58] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 2012.

[59] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-Learning with Memory-Augmented Neural Networks," in *33rd International Conference on Machine Learning, ICML 2016*, 2016.

[60] Y. Duan, M. Andrychowicz, B. Stadie, J. Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," in *Advances in Neural Information Processing Systems*, 2017.

[61] H. Altae-Tran, B. Ramsundar, A. S. Pappu, and V. Pande, "Low Data Drug Discovery with One-Shot Learning," *ACS Central Science*, 2017.

[62] D. Weininger, "SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules," *Journal of Chemical Information and Computer Sciences*, 1988.

[63] J. Brownlee, "Why One-Hot Encode Data in Machine Learning?," 2017.

[64] S. A. Bero, A. K. Muda, Y.-H. Choo, N. A. Muda, and S. F. Pratama, "Weighted tanimoto coefficient for 3d molecule structure similarity measurement," 2018.

[65] B. Zhang, M. Vogt, G. M. Maggiora, and J. Bajorath, "Design of chemical space networks using a Tanimoto similarity variant based upon maximum common substructures," *Journal of Computer-Aided Molecular Design*, 2015.

[66] G. Maggiora, M. Vogt, D. Stumpfe, and J. Bajorath, "Molecular similarity in medicinal chemistry," 2014.

[67] M. Zwierzyna, M. Vogt, G. M. Maggiora, and J. Bajorath, "Design and characterization of chemical space networks for different compound data sets," *Journal of Computer-Aided Molecular Design*, 2015.

[68] G. Landrum, "RDKit: Open-source Cheminformatics," 2006.

[69] D. Butina, "Unsupervised data base clustering based on daylight's fingerprint and Tanimoto similarity: A fast and automated way to cluster small and large data sets," *Journal of Chemical Information and Computer Sciences*, 1999.

[70] D. Bajusz, A. Rácz, and K. Héberger, "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?," *Journal of Cheminformatics*, 2015.

[71] N. Milosevic, *Introduction to Convolutional Neural Networks*. 2020.

[72] J. Teuwen and N. Moriakov, "Convolutional neural networks," in *Handbook of Medical Image Computing and Computer Assisted Intervention*, 2019.

[73] R. Arora, A. Basu, P. Mianjy, and A. Mukherjee, "Understanding deep neural networks with rectified linear units," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[74] F. Saeedan, N. Weber, M. Goesele, and S. Roth, "Detail-Preserving Pooling in Deep Networks," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[75] G. Koch and G. Koch, "Siamese Thesis," *Cs.Toronto.Edu*, 2015.

[76] I. Melekhov, J. Kannala, and E. Rahtu, "Siamese network features for image matching," in *Proceedings - International Conference on Pattern Recognition*, 2016.

[77] S. H. Park and J. Fürnkranz, "Efficient pairwise classification," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2007.

[78] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande, "MoleculeNet: A benchmark for molecular machine learning," *Chemical Science*, 2018.

[79] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity prediction using deep learning," *Frontiers in Environmental Science*, 2016.

[80] R. Huang, M. Xia, D. T. Nguyen, T. Zhao, S. Sakamuru, J. Zhao, S. A. Shahane, A. Rossoshek, and A. Simeonov, "Tox21 challenge to build predictive models of nuclear receptor and stress response pathways as mediated by exposure to environmental chemicals and drugs," *Frontiers in Environmental Science*, 2016.

[81] "Searching for activation functions," in *6th International Conference on Learning Representations, ICLR 2018 - Workshop Track Proceedings*, 2018.

[82] J. He, L. Li, J. Xu, and C. Zheng, "Relu deep neural networks and linear finite elements," *Journal of Computational Mathematics*, 2020.

[83] A. C. Wilson, R. Roelofs, M. Stern, N. Srebro, and B. Recht, "The marginal value of adaptive gradient methods in machine learning," in *Advances in Neural Information Processing Systems*, 2017.

[84] O. Konur, "Adam Optimizer," *Energy Education Science and Technology Part B: Social and Educational Studies*, 2013.

[85] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: A system for large-scale machine learning," in *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 2016.

[86] L. Prechelt, "Early stopping - But when?," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012.

[87] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Pas-

sos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, 2011.

[88] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of Translational Medicine*, 2016.

[89] G. P. Rédei, "Euclidean Distance," in *Encyclopedia of Genetics, Genomics, Proteomics and Informatics*, 2008.

[90] B. Schölkopf, "An Introduction to Support Vector Machines," in *Recent Advances and Trends in Nonparametric Statistics*, 2003.

[91] H. M. Gutmann, "A Radial Basis Function Method for Global Optimization," *Journal of Global Optimization*, 2001.

[92] L. Breiman, "Random forests," *Machine Learning*, 2001.

[93] T. Marwala, "Multi-layer Perceptron," in *Handbook of Machine Learning*, 2018.

[94] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, 1947.

[95] T. G. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural Computation*, 1998.

[96] M. Lewis-Beck, A. Bryman, and T. Futing Liao, "Shapiro-Wilk Test," in *The SAGE Encyclopedia of Social Science Research Methods*, 2012.

[97] James Lani, "Correlation (Pearson, Kendall, Spearman)," *Statistics Solutions*, 2010.

[98] C. Hill and C. Hill, "SciPy," in *Learning Scientific Programming with Python*, 2016.

[99] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python," *Proc. of the 9th Python in Science Conf*, 2010.

[100] N. R. Monteiro, B. Ribeiro, and J. P. Arrais, "Deep Neural Network Architecture for Drug-Target Interaction Prediction," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2019.

[101] N. R. C. Monteiro, B. Ribeiro, and J. Arrais, "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2020.