POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Investigating Deep Learning Methods for Drug Repurposing

Tesi di Laurea Magistrale in
Computer Science and Engineering - Ingegneria Informatica

Author: **Ismail Fatih Gonen**

Student ID: 10755808
Advisor: Prof. Pietro Pinoli
Academic Year: 2022-23

Abstract

The thriving field of drug repurposing presents a unique opportunity to address the challenges of prolonged timelines and high costs associated with traditional drug discovery. This study introduces a novel approach that employs Long Short-Term Memory (LSTM) autoencoders for drug repurposing, focusing on drug-target interaction (DTI) predictions. Our methodology leverages the sequential learning capabilities of LSTM networks to analyze and interpret complex patterns in biochemical data, specifically targeting the interactions between drugs and their potential protein targets. The autoencoder architecture is adept at capturing essential features in high-dimensional drug and target data, facilitating more accurate predictions of DTI. We applied this framework to a comprehensive dataset of known drug-target interactions, using it to predict new interactions that suggest repurposing opportunities for existing drugs. The results demonstrate promising accuracy and specificity in identifying potential new uses for established drugs, highlighting the effectiveness of Deep Learning methods such as LSTM autoencoders in uncovering complex relationships within pharmacological data. This approach not only provides a powerful tool for drug repurposing but also offers insights into the mechanisms of drug action, potentially accelerating the identification of therapeutic applications for existing drugs and contributing to personalized medicine. This study paves the way for advanced computational strategies in drug discovery, underscoring the potential of machine learning models in revolutionizing pharmaceutical research.

Keywords: drug repurposing, drug-target interaction, deep learning, lstm, autoencoder, protein, ligand

Abstract in lingua italiana

Il fiorente campo del riposizionamento dei farmaci presenta un'opportunità unica per affrontare le sfide dei tempi prolungati e dei costi elevati associati alla scoperta di farmaci tradizionali. Questo studio introduce un approccio innovativo che impiega gli autoencoder LSTM (Long Short-Term Memory) per il riposizionamento dei farmaci, concentrandosi sulla previsione dell'interazione farmaco-target (DTI). La nostra metodologia sfrutta le capacità delle reti LSTM di apprendere sequenze per analizzare e interpretare modelli complessi nei dati biochimici, con particolare attenzione alle interazioni tra i farmaci e i loro potenziali bersagli proteici. L'architettura dell'autoencoder è in grado di ridurre la dimensionalità e catturare le caratteristiche essenziali nei dati ad alta dimensionalità di farmaci e target, facilitando previsioni più accurate di DTI. Abbiamo applicato questo framework a un set di dati completo di interazioni farmaco-target note, utilizzandolo per prevedere nuove interazioni che suggeriscono opportunità di riposizionamento per i farmaci esistenti. I risultati dimostrano un'accuratezza e una specificità promettenti nell'identificazione di potenziali nuovi usi per i farmaci esistenti, evidenziando l'efficacia delle tecniche di apprendimento profondo, come gli autoencoder LSTM, nello scoprire relazioni complesse all'interno dei dati farmacologici. Questo approccio non solo fornisce un potente strumento per il riposizionamento dei farmaci, ma offre anche approfondimenti sui meccanismi di azione dei farmaci, accelerando potenzialmente l'identificazione di applicazioni terapeutiche per i farmaci esistenti e contribuendo alla medicina personalizzata. Questo studio apre la strada a strategie computazionali avanzate nella scoperta di farmaci, sottolineando il potenziale dei modelli di apprendimento profondo nel rivoluzionare la ricerca farmaceutica.

Parole chiave: riposizionamento dei farmaci, interazione farmaco-target, deep learning, lstm, autoencoder, protein, ligand

Contents

1 | Introduction

Drug repurposing is an approach to discovering new uses for drugs that have already been approved for other indications. Instead of developing entirely new drugs, scientists and researchers are exploring the potential of drugs that have already been approved for one use to treat other diseases or conditions. This approach is gaining momentum due to several advantages, such as the ability to expedite the drug development process, reduced risk of toxicity, and lower costs. Figure 1.1 shows the comparison between the traditional way of drug discovery versus drug repurposing in a conventional way. It can be seen that even without using the edge given by AI to drug repurposing, it is a method that is much faster than classical methods.

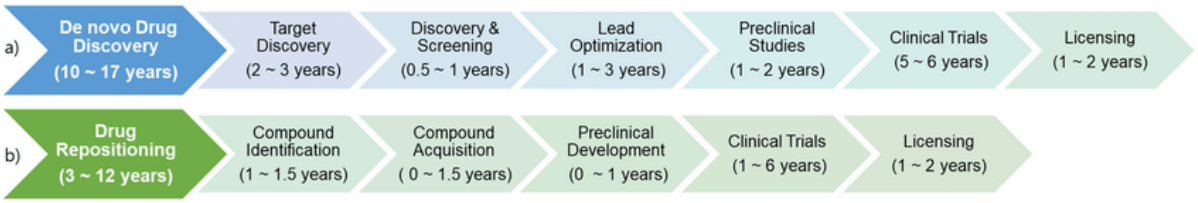

Figure 1.1: Traditional way of drug discovery versus drug repurposing using conventional methods [13]

One critical aspect of drug repurposing is understanding drug-target interactions, which refer to the relationship between a drug molecule and its intended target in the body. The target can be a protein, an enzyme, or a specific cell type involved in a disease or condition. When a drug molecule interacts with its target, it triggers a biochemical response that can either inhibit or activate the target, leading to a therapeutic effect.

Thus, a thorough understanding of drug-target interactions is essential for successful drug repurposing. It helps researchers identify new indications for existing drugs and predict potential side effects. With the aid of advanced technologies such as artificial intelligence and machine learning, scientists can analyze vast amounts of data to uncover novel drug-target interactions and accelerate drug development.

The use of artificial intelligence (AI) in drug repurposing and drug target interaction has

revolutionized the drug development process. AI algorithms can analyze vast amounts of biological and chemical data to predict new therapeutic uses for existing drugs, identify potential drug targets, and optimize drug design. One of the significant advantages of AI in drug repurposing is its ability to accelerate the identification of drug candidates for clinical trials. Traditionally, drug discovery and development involve a lengthy and costly process of trial and error, with researchers testing thousands of molecules to find a potential candidate. However, with AI, researchers can use predictive models to screen large databases of molecules and identify those with the highest likelihood of success, saving time and resources. Furthermore, AI can help researchers understand the complex interactions between drugs and their targets. By analyzing large datasets of biological and chemical information, AI algorithms can identify potential off-target effects and predict the safety and efficacy of a drug in different patient populations. Overall, the use of AI in drug repurposing and drug target interaction has the potential to transform the drug development process by increasing the speed and efficiency of drug discovery, reducing costs, and improving patient outcomes.

Deep learning, a subset of machine learning methods, is increasingly used in this field to analyze complex biological data and predict potential new uses for drugs.

In recent studies, researchers have utilized deep learning methods with inputs such as protein sequences, which are chains of amino acids that constitute proteins, and SMILES, a notation that encodes the structure of molecules in a string format. These data forms are crucial because they contain the information necessary to understand the drug's action at the molecular level.

For instance, in the context of COVID-19, various approaches have been adopted using SMILES strings and protein sequences to repurpose drugs to treat the disease. The methods often involve creating molecular graphs with nodes and edges that represent atoms and chemical bonds, respectively. This allows for a detailed analysis of the molecular interactions involved in drug-target binding.

DeepPurpose [10] is an example of a deep learning toolkit designed specifically for drug-target interaction (DTI) prediction. It utilizes encoding-based methods to process the information contained in drug molecules and protein sequences. The toolkit can facilitate the identification of potential binding affinities between drugs and biological targets, which is essential for repurposing efforts.

Another innovative model, known as DeepLPI [21], predicts protein–ligand interactions using the raw 1D sequences of proteins and ligands. Such models are beneficial because they can handle the simple formats of the data without the need for pre-processing or

feature extraction, which can often be a complex and error-prone process.

Moreover, deep learning-based drug repurposing studies have been shown to provide powerful tools for future research, particularly in the case of diseases that are not well understood. By extracting physical and chemical features from protein sequences and ligands, these methods help in elucidating the mechanisms of drug action and potential off-target effects.

Compared to traditional methods, deep learning approaches are advantageous because they depend on raw data, like the SMILES representations of ligands for drugs and protein sequences for targets. These methods can automatically extract molecular features by designing efficient algorithms, which potentially simplify and accelerate the repurposing process.

In summary, deep learning is transforming drug repurposing by efficiently analyzing protein sequences and ligands to uncover novel drug applications. This approach has the potential to significantly reduce the time and cost associated with drug development by repurposing existing drugs for new therapeutic uses.

In this thesis the focus is correctly classifying the pairs of protein sequences that consist of different amino acids and ligands, methods that have been used for this purpose will be explained in the following chapters. The aim of classification in drug repurposing is to categorize existing drugs based on their potential to be effective in treating new or different medical conditions than those for which they were originally developed. The primary goal is to predict novel interactions between protein sequences and ligands. This can lead to quicker and more cost-effective development compared to creating new drugs from scratch, as these drugs have already passed several safety and regulatory hurdles. By classifying drugs based on their action mechanisms, researchers can better understand how these drugs interact with biological systems. This can reveal new insights into the underlying mechanisms of diseases and potentially identify new treatment pathways. Repurposing may also aim to enhance the efficacy of a drug or reduce its side effects in treating a particular condition. Drugs that were not very effective or had adverse effects for their original purpose might be more suitable for other conditions. Some other use cases of drug repurposing with Machine Learning will be given here. Classification in drug repurposing can contribute to personalized medicine by matching specific drugs to the individual patient's profile. This includes genetic makeup, disease characteristics, and response to previous treatments, aiming to maximize efficacy and minimize adverse effects.

Drug repurposing can also be used for cost-effective healthcare solutions, it can reduce the time and cost associated with drug development, providing more affordable healthcare

solutions. This is particularly important for rare or neglected diseases, where new drug development might be economically unfeasible.

Another topic to consider is combating drug resistance. In the case of infectious diseases or cancer, drug repurposing can help in finding alternative treatments when resistance to standard treatments develops.

2 | Literature Review

2.1. History of Drug Repurposing

The history of drug repurposing, also known as drug repositioning, is a fascinating aspect of pharmaceutical development. This process involves finding new medical uses for existing drugs, a concept that has been around for several decades.

Historically, some of the most successful drug repurposing efforts were based on serendipity or retrospective clinical experience. Two notable examples are thalidomide and sildenafil citrate. The story of thalidomide is particularly striking. Initially synthesized in 1952 and marketed as a sedative and antiemetic for morning sickness, it was withdrawn due to its teratogenic effects, causing severe birth defects. However, it was later repurposed for the treatment of leprosy and multiple myeloma, illustrating its dramatic journey from disaster to a WHO-listed essential medicine [17].

Modern drug repurposing approaches now leverage an increasing wealth of drug- and disease-related data, computational hypothesis generation, and high-throughput screening methods, reflecting a shift from serendipitous discoveries to more systematic and data-driven strategies [17].

Early Instances and the concept of drug repurposing date back to the mid-20th century, although it wasn't formalized as a distinct strategy at the time. Drugs were often found to have multiple effects, some of which were initially considered side effects but later recognized as potential therapeutic benefits for other conditions.

In the traditional drug discovery pipeline, the path from concept to approved therapy is long, costly, and uncertain. Drug repurposing offers an alternative, allowing for the reduction of time and costs associated with pharmaceutical research. This is achieved by identifying new uses for drugs that are already approved or under investigation. In recent decades, there have been many successful examples of drug repurposing across various pathologies, showcasing its potential to accelerate the pace of discovery [4].

Historically, many successful drug repurposing ventures were based on serendipity or ret-

rospective clinical experience. Thalidomide and sildenafil citrate are two prominent examples of such unintentional discoveries. Thalidomide, initially a sedative and antiemetic, was later repurposed for treating leprosy and multiple myeloma after its withdrawal due to teratogenic effects. Sildenafil, developed for angina pectoris(heart pain due to coronary heart disease), was repurposed as Viagra after its unexpected effect on erectile dysfunction was discovered during trials [17].

Drug repurposing is an essential component of pharmaceutical research, marked by both serendipitous discoveries and systematic approaches. Over the last three decades, the pharmaceutical industry has faced a growing productivity gap, with high drug attrition rates, escalating development costs, and increased time to bring new chemical entities (NCEs) to market. These challenges have underscored the need for innovative strategies like drug repurposing [17].

With the rise of systematic approaches in the late 20th and early 21st centuries, the approach to drug repurposing became more systematic and deliberate [1]. The realization that developing new drugs was becoming increasingly costly and time-consuming led pharmaceutical companies and researchers to actively search for new uses of existing drugs.

Another factor in the advancement of drug repurposing was technological advancements. The advent of high-throughput screening, bioinformatics, and computational biology in the 21st century has significantly advanced the field of drug repurposing. These technologies allow for the systematic and rapid testing of large libraries of existing drugs against a wide array of targets and diseases.

Impact of Genomics and Personalized Medicine: The rise of genomics and personalized medicine has further fueled drug repurposing [15]. Understanding the genetic basis of diseases has enabled researchers to identify potential new uses for drugs based on their molecular mechanisms of action.

Today, the field of drug repurposing is characterized by a diverse range of techniques and targets. From machine learning-driven frameworks for kinase inhibitor repositioning to the use of natural products against viral infections, the methodologies are as varied as they are inventive. This diversity reflects the integration of multidisciplinary sciences, combining computational techniques, pharmacological insights, and molecular biology. Currently, drug repurposing is an integral part of pharmaceutical research and development. It is seen as a cost-effective, time-saving strategy that can complement traditional drug discovery processes. The use of AI and machine learning has further enhanced the ability to identify repurposing opportunities.

Artificial intelligence (AI) and machine learning have become crucial in drug repurposing. These advanced computational methods enable researchers to analyze vast amounts of data, uncover hidden patterns, and generate insights that would be challenging to achieve through traditional means. For instance, the KUALA framework automates the identification of kinase active ligands and prioritizes multi-target scores for repurposable molecules [5]. KUALA paper presents a novel approach for repositioning kinase inhibitors using a machine learning framework. The researchers developed the Kinase drUgs mAchine Learning framework (KUALA) to automatically identify kinase active ligands and provide a multi-target priority score to suggest the best repurposable molecules.

Scientists have addressed diverse therapeutic needs through drug repurposing, exploring treatments for conditions like COVID-19, Alzheimer's disease, and infectious diseases. For instance, leveraging single-cell RNA sequencing data from brain tissues of Alzheimer's disease patients, researchers constructed a multi-cellular disease molecular network to identify 54 candidate drugs for potential therapy. Moreover, innovative approaches are being developed to tackle antibiotic resistance, reflecting a broader shift in thinking where drug repurposing is viewed as a holistic strategy to respond to global health concerns.

COVID-19 Pandemic: The COVID-19 pandemic brought renewed attention to drug repurposing, as the urgent need for effective treatments led to the repurposing of existing drugs, such as remdesivir, originally developed for Ebola, and dexamethasone, a steroid, for treating severe cases of COVID-19.

However, the path to successful drug repurposing is not without challenges. Issues of selectivity, toxicity, and the balance between binding site similarity and target numbers are complex considerations. Despite these challenges, there's a long history of off-label use of pharmaceutical products in the clinic for indications other than the primary or listed case, which continues to be an avenue for development.

In summary, the history of drug repurposing is a testament to the field's evolution from serendipitous discoveries to a systematic, data-driven approach, underpinned by technological advancements and a multidisciplinary perspective. This evolution reflects the ongoing need for innovative, cost-effective strategies in pharmaceutical research and development.

2.2. Studies Using Machine Learning

In this part, previous studies that use different Machine Learning techniques for Drug Repurposing will be investigated.

The paper titled "A Novel Deep Neural Network Technique for Drug–Target Interaction", presents a novel method for predicting drug-target interactions (DTIs) using deep learning [20]. The key contributions of this research are two-fold:

Molecule and Protein Sequence to Image Transformation (MPS2IT-DTI): This technique involves transforming molecule and protein sequences into image-based representations. The authors developed a new method to encode these sequences onto images, which is a departure from traditional natural language processing (NLP) based techniques. This method does not require an embedding layer, unlike other models.

Convolutional Neural Network (CNN)-Based Architecture: The transformed images are then processed using a dual-CNN architecture, which is designed to predict the interactions between molecules (drugs) and proteins (targets). The CNNs process the images and output predictions of their drug-target interaction.

The study demonstrated the viability of MPS2IT-DTI through training results using the Davis and KIBA datasets [20]. Compared to other state-of-the-art approaches, this model promises competitive performance in terms of both accuracy and complexity. Specifically, with the Davis dataset, the model achieved a concordance index of 0.876 and a mean squared error (MSE) of 0.276. For the KIBA dataset, the concordance index was 0.836 and MSE was 0.226.

The advantage of MPS2IT-DTI is highlighted in representing molecule and protein sequences as images rather than treating them as text-based sequences. This novel approach offers a promising alternative to existing NLP-based techniques in drug-target interaction prediction.

The authors proposed a unique representation for both molecules and proteins, which was a key part of their drug-target interaction prediction model, named MPS2IT-DTI (Molecule and Protein Sequence to Image Transformation - Drug-Target Interaction)[20].

Molecule Representation SMILES Representation: The process starts with the SMILES (Simplified Molecular Input Line Entry System) representation of the molecule. K-mer Counting: The SMILES string is then processed to define a set of all possible k-mers (subsequences of k characters). Counting Vector: A counting vector is created, listing the occurrences of each possible k-mer in the SMILES sequence. Normalization: The counting vector is normalized, resulting in values between 0 and 1. Image Transformation: The normalized vector is reshaped into a 2D matrix, creating an image-like representation of the molecule. Protein Representation Amino Acid Sequence: It starts with the sequence of amino acids that make up the protein. K-mer Counting: Similar to molecules, the process

involves defining all possible k-mers from the amino acid sequence. Counting Vector: A counting vector for the k-mers in the protein sequence is generated. Normalization: This vector is normalized to have values between 0 and 1. Image Transformation: The normalized vector is then converted into a 2D matrix, forming an image representation of the protein. It has been claimed that this innovative approach of representing molecules and proteins as images, rather than as text sequences, allows for the application of image processing techniques, specifically convolutional neural networks (CNNs), in the drug-target interaction prediction process. This method is distinct from traditional natural language processing techniques used in other models and does not require an embedding layer.

Another approach in this topic is "DeepDTA: deep drug–target binding affinity prediction", which focuses on developing a deep learning-based model to predict drug-target interaction (DTI) binding affinities using only the sequence information of targets (proteins) and drugs (compounds) [16]. The key aspects of this study are:

Deep Learning Approach: The model, named DeepDTA, employs Convolutional Neural Networks (CNNs) to process the 1D representations of proteins and drugs derived from their sequences. This approach is distinct from traditional methods that use either 3D structures of protein-ligand complexes or 2D features of compounds. The model was compared against two baseline methodologies, KronRLS and SimBoost, using metrics like Concordance Index (CI) and Mean Squared Error (MSE).

Results from the paper: DeepDTA demonstrated effective performance in drug-target binding affinity prediction. It outperformed the baseline methods in the KIBA dataset, achieving better CI scores and lower MSE values. It is claimed that the results suggested CNNs could capture more information from the SMILES (drug) representations than traditional methods.

The paper highlights the potential of using deep learning to process raw sequence data of drugs and proteins for predicting DTI affinities. The authors suggest that their methodology could be extended to predict the affinity of known compounds/targets to novel targets/drugs and for the affinity prediction of novel drug-target pairs.

The primary contribution of this study is the demonstration of a novel deep learning-based model that successfully predicts drug-target affinities using only the sequence information of proteins and drugs, offering a promising alternative to traditional feature-based methods.

SMILES Sequences Representation: The SMILES sequences of compounds were encoded

using a set of 64 unique labels (letters) identified from approximately 2 million SMILES sequences gathered from Pubchem. Each label (letter) in a SMILES sequence was represented by a corresponding integer (e.g., 'C': 1, 'H': 2, 'N': 3, etc.). The maximum length for SMILES sequences was set at 85 characters for the Davis dataset and 100 characters for the KIBA dataset. Sequences longer than the maximum length were truncated, while shorter sequences were padded with zeros to maintain a consistent length. Protein Sequences Representation: Protein sequences were encoded in this paper using label encoding based on 25 unique categories (letters) extracted from 550,000 protein sequences from UniProt. The maximum length for protein sequences was determined as 1200 characters for the Davis dataset and 1000 characters for the KIBA dataset. Similar to SMILES sequences, longer protein sequences were truncated and shorter ones were zero-padded to achieve the fixed lengths.

Another study reviews the machine learning approaches in drug-target interaction. This article underscores the significance of DTIs in the selection of potential drugs and in providing insights into drug mechanisms and side effects. The review concentrates on machine learning methods that integrate chemical and genomic spaces, categorizing them into supervised and semi-supervised methods [7]. It points out that while machine learning shows promise in DTI prediction, there's substantial scope for improvement. The paper suggests focusing on ensemble approaches, semi-supervised learning, and new regression methods that consider binding affinities and dose-dependence for more accurate predictions. The paper concludes by acknowledging the need for further research in these areas, especially given the rapid growth of data from high-throughput biotechnology.

Another study presents a deep learning approach, DeepMHADTA, capable of predicting the binding affinity between proteins and drugs. This method was evaluated using two established benchmark datasets, Davis and KIBA, commonly used for protein-drug binding affinity prediction [6]. The DeepMHADTA model combines both sequence and structural information of proteins and drugs to extract relevant features.

Key components of the model include:

- Drug Representation: Drugs are represented using SMILES descriptors, converted into vectors using integer/label encoding, and molecular fingerprints, capturing the structural features of drugs.
- Protein Representation: Proteins are represented using n-gram methods and pretrained word2vec models to extract sequence information, and Spider3 is employed to predict protein secondary structure.

- Feature Extraction: The method utilizes a multi-head self-attention mechanism to identify and focus on important features, and employs a residual network for feature extraction layers. The extracted features of proteins and drugs are concatenated and fed into a fully connected layer for regression prediction of binding affinity.

The study highlights several advantages of the DeepMHADTA approach:

- Multi-Head Self-Attention Mechanism: Effectively focuses on important features.
- Word2Vec for Protein Sequence Features: Provides an efficient semantic representation compared to traditional encoding methods.
- Comprehensive Feature Extraction: Incorporates not only the sequence information but also the spatial structure of proteins and drugs.

Another study about Drug Repurposing provides a method called "CSatDTA". This was said to be a method for "Prediction of Drug–Target Binding Affinity Using Convolution Model with Self-Attention".It introduces the CSatDTA model, a novel approach for predicting drug-target interaction (DTI) affinity [8]. The key methods employed in this model are as follows:

Combining Convolutional Neural Networks (CNNs) with Self-Attention:

The model is designed to enhance traditional convolutional networks by integrating a self-attention mechanism. Self-attention is used to overcome the limitations of CNNs, particularly their inability to capture long-distance interactions between atoms in molecular structures. Attention Mechanism:

The attention mechanism in the model focuses on both spatial and feature subspaces, using a multi-head attention (MHA) mechanism. This allows the model to assign importance to different parts of the input [8]. The relative self-attention is extended to 2D inputs, systematically improving its representational capacity. Architecture Details:

The CSatDTA model consists of convolutional layers, max-pooling layers, and fully connected (FC) dense layers. The convolutional layers are designed to extract local dependencies, with the size and number of filters in these layers directly impacting the type of characteristics extracted from the input data. The model includes two self-attention-augmented convolutional blocks, each comprising five convolution layers and one attention layer. Key parameters in the model include the depth of keys, depth of values, and the number of heads in MHA. Learning Representations from Sequences:

The model aims to learn representations from the sequences of proteins and SMILES strings, which encode molecular structures. The approach is motivated by the similarity

of target sequences and drug structures to natural language texts, where understanding atomic, structural, and contextual information is crucial. Dynamic Modification and Design Evaluation:

The method allows for dynamic modification of the proportion of attentional channels. This flexibility enables the evaluation of a range of designs from fully convolutional to attentional models. This innovative approach combines the strengths of CNNs in local feature extraction with the global contextual awareness provided by self-attention mechanisms, enhancing the prediction accuracy of drug-target binding affinities.

In another study about drug repurposing SPVec model was introduced. "SPVec: A Word2vec-Inspired Feature Representation Method for Drug-Target Interaction Prediction" SPVec is a feature representation method for predicting drug-target interactions (DTIs) [22]. The paper addresses the challenge of accurately identifying DTIs, a crucial step in drug discovery. Traditional methods for DTI prediction are labor-intensive and require significant human expertise. To overcome these limitations, the authors propose SPVec, inspired by Word2vec, an unsupervised representation learning method [22]. SPVec is designed to automatically represent raw data like SMILES strings (for drugs) and protein sequences into continuous, information-rich, and lower-dimensional vectors. This method aims to avoid the sparsity and bit collisions of manually extracted features. The SPVec method combines two models: SMILES2Vec for drug compounds and ProtVec for target proteins. These models are trained using a revised Skip-gram model with negative sampling.

The SPVec method was evaluated using the BindingDB database and external validation with the DrugBank database. The performance of SPVec was compared against traditional feature representation methods like MACCS fingerprints and amino acid composition (AAC), using machine learning classifiers such as Gradient Boosting Decision Tree (GBDT), Random Forest (RF), and Deep Neural Network (DNN). Results and Significance:

SPVec shows good performance compared to traditional feature representation methods in DTI prediction. The method is also robust when tested on independent test sets and demonstrated potential in discovering reliable DTIs, which could be beneficial for drug re-profiling. The paper highlights the advantages of SPVec in terms of automatic learning and lower dimensionality, which could significantly speed up training and reduce memory requirements.

SPVec, by combining SMILES2Vec and ProtVec, effectively transforms SMILES strings and protein sequences into useful vectors for machine learning models. The proposed mod-

els achieved better performance than traditional methods, suggesting SPVec's potential utility in DTI prediction and drug discovery.

3 | Datasets

Three different datasets have been used throughout this thesis. The main one and first dataset used is BindingDB. BindingDB is an open, internet-accessible repository containing recorded binding affinities, primarily emphasizing interactions between proteins regarded as potential drug targets and small, drug-like compounds [3].

The data in BindingDB originates from various measurement techniques, including enzyme inhibition, kinetics, isothermal titration calorimetry, NMR, radioligand assays, and competition assays. The database includes information extracted from scientific literature, patents, selected PubChem confirmatory BioAssays, and ChEMBL entries that provide well-defined protein targets.

It's a dynamic database with ongoing curation, including the addition of recently identified targets and ligands, as evidenced by the inclusion of new data on influenza virus hemagglutinin and human N6-adenosine-methyltransferase non-catalytic subunit. Moreover, BindingDB ensures the availability of their archived data for reference and research continuity.

In response to emergent needs, such as the COVID-19 pandemic, BindingDB has accelerated the collection of related data, providing researchers with critical information to aid in the discovery of treatments for the coronavirus.

The database is also comprehensive in curating data from US Patents, with a vast collection of binding measurements, compounds, and target proteins. Additionally, BindingDB fills gaps left by other databases by curating a range of scientific journals, providing a broad spectrum of data not available elsewhere [3].

This wealth of information is used to facilitate the identification of potential drug-target interactions, which is essential for drug repurposing efforts and for advancing the field of pharmacology.

In addition to BindingDb, two datasets for protein sequences were used. These databases were used for pretraining purposes in the Autoencoder models, as will come up in the methods chapter.

3.1. Statistics About Data

In this section, some statistics about the 3 datasets that have been used throughout this thesis will be shared. To add to the introductory information about the main dataset BindingDb, after necessary preprocessing operations in total, there were 84840 drug-protein pairs. 24435 of them being labeled as 1 according to relevant labeling operation as explained meaning a match between drug and protein. As mentioned before in the context of protein and drug interaction, a "match" typically refers to the compatibility or affinity between a drug molecule and a specific protein target. This match or interaction is crucial for the drug to carry out its intended function within the body. Since it is not possible to calculate a definite match in this context high affinity is taken as a match. The number of pairs that do not match is 60405. As it can be seen from this data, number of zeros is larger than number of ones. This imbalanced data problem will be investigated again.

BindingDB is a significant resource for research in drug discovery and pharmacology, particularly for drug-target interactions. It contains data for over 1.2 million compounds and 9.2K targets, with a substantial portion curated by BindingDB's own curators. The database not only supports research but also education and practice in related fields [3].

The other two datasets include protein sequences and they have been used for pretraining purposes. Those will be mentioned as "HomosapiensDb" and "AllProDb". HomosapiensDb includes 20,598 protein sequences and AllProDb includes 79,006 protein sequences in total.

3.2. Affinity and Kd Value In Data

In this section, some background information related to the context of Drug Repurposing will be given so that the next parts will be easier to comprehend. Currently best option to use for labeling is the affinity of drug and target interaction. Here affinity refers to both the proportion and the strength with which a drug attaches to its receptors at a given concentration. Irving Langmuir Kenakin first developed a mathematical model to describe this concept in 2004 [12]. Affinity, which is inversely related to the drug's potency, is a key determinant of potency. This is represented by 1/Kd, where Kd is the dissociation constant. Essentially, affinity measures how strongly a ligand binds to its receptor. A higher Kd value indicates weaker binding and thus lower affinity, whereas a lower Kd suggests stronger binding.

Potency, on the other hand, is defined as the quantity of a drug needed to achieve a

specific level of effect. It is usually expressed as the median effective concentration or dose, represented by EC50/ED50/Kd [12].

Efficacy, or intrinsic activity, is the capacity of a drug to trigger a pharmacological or physiological response when it interacts with a receptor. Efficacy relies on how efficiently the receptor activates cellular responses and the number of drug-receptor complexes formed. This describes the relationship between the response and the occupancy of the receptor by the drug.

3.3. Data Labeling

In BindingDb labeling operations had to be done since they are needed for the classification task. In the original database, only some measurements about matching proteins and drugs exist but there is no label such as 1 or 0. In order to do this some options are considered and in the end Kd value is chosen for the labeling operation.

Kd value is called the dissociation constant which is a commonly utilized parameter to elucidate the degree of attachment between a ligand and its receptor. Essentially, Kd serves as a quantification of binding affinity, signifying how strongly a ligand attaches to a receptor. The interaction between a ligand and receptor can be symbolized as L + R LR, and the Kd value is computed as

$$K_d = \frac{([L][R])}{[LR]} \tag{3.1}$$

In the context of ligand-receptor complexes, Kd denotes the ligand concentration at which 50% of the receptors are bound to ligands. A lower Kd value indicates a tighter bond between the ligand and the receptor, reflecting a higher level of affinity between them. Kd value is useful to understand the affinity between proteins and drugs but since they are continuous values in the nanometer level, it can't be used directly as a label for classification.

Instead another value needed to be calculated for this. As it has been suggested in previous studies [9] pKd value was used for this purpose. Which is the result of the transformation the Kd value into log space as

$$pK_d = -\log_{10}\left(\frac{K_d}{1e9}\right) \tag{3.2}$$

Labeling operation is done using this pKd value. 7 was chosen as the threshold value. Inputs with pKd value greater than or equal to 7 are labeled as 1 and others as 0. 1

meaning that there is a match between protein and drug.

3.4. Additional Data For Pretraining

In future chapters, some models including pretraining will be introduced. In order to achieve this 2 more datasets, including protein sequences were used. These two datasets, referred to as "HomosapiensDb" and "AllProDb," were employed for pretraining various models, which will be detailed subsequently. In addition to the sequences found in BindingDb, these datasets encompass a broader range of protein sequences. The number of proteins in each dataset can be seen in Table 3.1.

	BindingDb	HomosapiensDb	AllProDb
Total Proteins	84,840	20,598	79,0068
Unique Proteins	2,483	20,528	75,948

Table 3.1: Number of Proteins in Datasets

3.5. Imbalanced Dataset Problem

In drug repurposing datasets, there are often many more negative instances (where a drug does not work for a specific condition) than positive ones (where it does). Models trained on such data can become biased towards predicting the majority class, leading to a high rate of false negatives (missed opportunities for repurposing). In BindingDb similar problem were faced since the number of matching pairs were significantly lower than the number of not matching pairs. The exact numbers can be seen in Table 3.2.

	Negative Pairs	Positive Pairs
BindingDb	60,405	24,435

Table 3.2: Number of Proteins in Datasets

Around 29% of our data was actually positive. A number of methods were used to challenge this problem. Some of them were in the data preparation phase and the rest were implemented during the evaluation phase. In the data preparation sampling techniques were tested as suggested in different studies [14]. After experiments with both under-sampling and over-sampling, it was seen that over-sampling leads to better results. Hence over sampling was chosen as the better solution to this problem.

Relying solely on accuracy as an evaluation metric can be deceptive when analyzing imbalanced datasets. A model might achieve high accuracy by accurately predicting the more prevalent class (majority), yet it might struggle to identify the less frequent and more crucial category (successful repurposings), rendering its overall performance less reliable. To fight this issue different metrics other than accuracy were implemented. These include precision, recall, f1 score, AUC and Matthews correlation. Some of these metrics are particularly beneficial in imbalanced datasets such as f1 score and Matthews correlation. More explanations about these metrics will be given in future chapters.

4 | Methods

4.1. Drug Representation

Since in the BindingDb dataset proteins and drugs are represented in different ways. For ligands representation, a method called SMILES is used. SMILES, which stands for Simplified Molecular Input Line Entry System, is a widely used notation for representing chemical structures, including those of drug molecules. It is a textual representation that encodes the structural information of a molecule in a concise and human-readable format. In a SMILES notation, atoms and bonds are represented using specific characters and symbols.

SMILES notation allows chemists and researchers to easily communicate and store chemical structures in a compact and standardized format. It is commonly used in cheminformatics, drug discovery, and computational chemistry for tasks such as database storage, structure searching, and predictive modeling of molecular properties. As an example molecule structure of aspirin can be examined in Figure 4.1.

Figure 4.1: Molecule Structure of Aspirin

When it is shown in SMILES notation it becomes:

$$CC(=O)OC1 = CC = CC = C1C(=O)O \tag{4.1}$$

As previously noted this representation consists of characters and symbols. What is needed for a Machine Learning model is a number. So a transformation of this representation was needed. For this transformation labeling each character or symbol with a number is chosen. For example 'C' corresponds to 4 and 'O' to 15. After this normalization is applied to these values. For the normalization, Min-Max scaling is chosen to scale values in a range between 1 and 0. Without normalization, the coefficients of features with larger scales might have not provided meaningful insights into their actual impact on the model's output. In order to test this both possibilities were used in experiments and improvements were seen after the normalization operation. In order to decide the proper length of the SMILES vectors, related statistics were used. The length distribution of drugs can be seen in Figure 4.1.

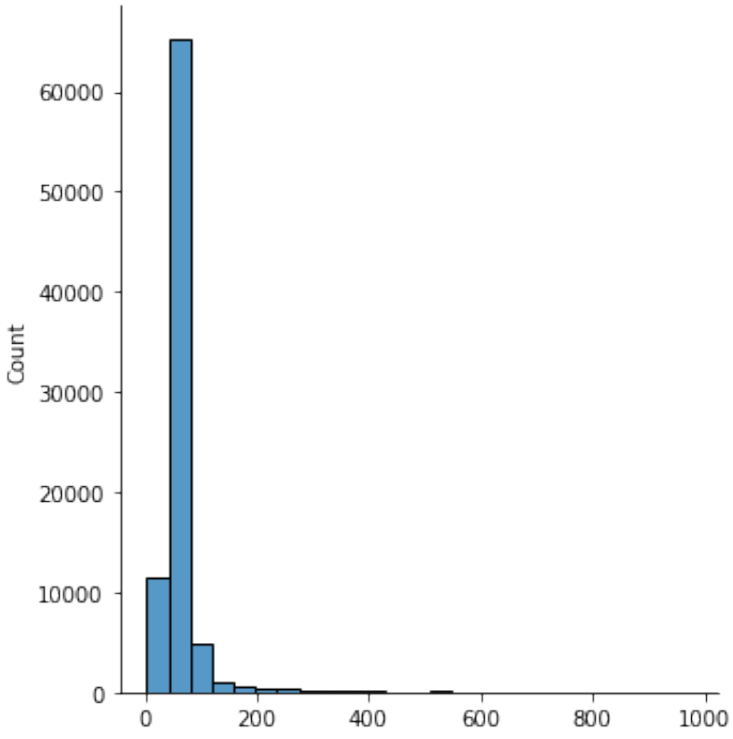

Figure 4.2: Drugs Length Distribution

4.2. Protein Representation

From a chemical standpoint, a protein can be essentially described as a linear chain made up of the 20 major amino acids, and its chemical makeup is primarily determined by the sequence in which these amino acids are arranged. The length of protein sequences can vary significantly, ranging from tens to thousands of amino acids, with the most common length likely being in the hundreds. The immense diversity in protein functions arises from the vast number of possible combinations of amino acid sequences. For instance, there are approximately 10^{260} potential proteins that are 200 amino acids in length. To put this in perspective, it's worth noting that the estimated number of atoms in the observable universe is around 10^{80}. Clearly, only an exceedingly small fraction of all conceivable proteins can exist at any given time or may have ever existed throughout the history of life on Earth. As an example amino acid sequence of hemoglobin can be given as:

- VHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRF FESFGDLST-PDAVMGNPKVKAHGKKVLGAFSDGLAHLDNL KGTFALSELHCDKLHVDPEN-FRLLGNVLVCV LAHHFGK EFTPPVQAAYQKVVAGVANALAHKYH

And the 3D shape of its representation is given in figure 4.3.

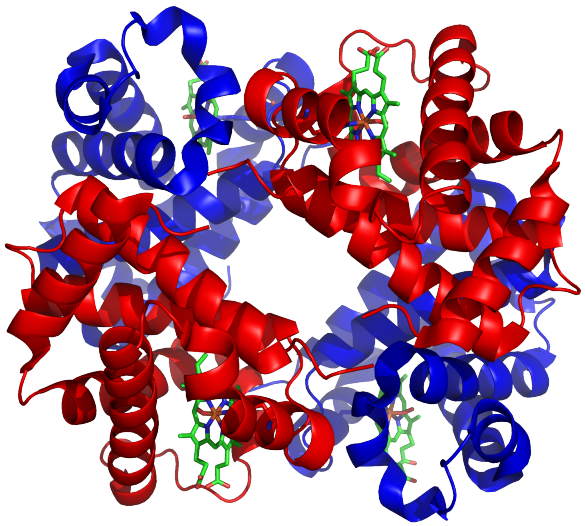

Figure 4.3: 3D structure of hemoglobin

In the databases used in this thesis protein sequences were given in a way to show each amino acid with one letter. Again a similar approach was chosen to convert this represen-

tation into a numerical one. Numbers were given to each amino acid and later these were scaled in a range between 0 and 1 using Min-Max scaling. After further research on this, a Word2Vec approach was chosen for protein sequences. For each amino acid, a vector with 3 lengths was calculated. For example, one vector for the amino acid alanine (short version 'A') was calculated as [0.6454009 , 0.4708575 , 0.37278453]. The distribution of lengths of protein sequences is given in Figure 4.2. This data was used to decide the length of the protein vector.

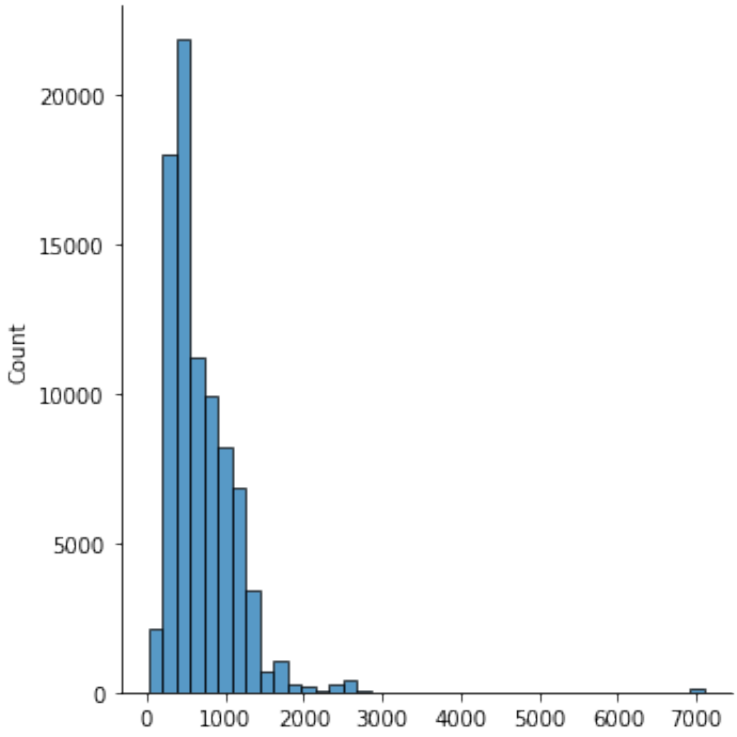

Figure 4.4: Proteins Length Distribution

Proteins length was chosen between 100 and 300 depending on the model. These values were chosen according to the given statistics here and with the experiments made.

4.3. Model Selection

Incorporating the use of cross-validation into the model selection process for the drug repurposing task adds an essential layer of rigor and reliability to the evaluation of the machine learning models. Here a revised outline of the selection process will be given that is used during this thesis.

Objective-oriented model exploration was used throughout. The primary goal remained to

identify effective drug repurposing opportunities. This requires models capable of accurate predictions regarding drug-target interactions. The selection of various model architectures like LSTM, LSTM Autoencoders, and Autoencoders is indicative of an exploratory approach tailored to meet this objective.

Diverse architectures and configurations were used during this study. The inclusion of different models, such as LSTM with and without attention mechanisms and various Autoencoder configurations, suggests a comprehensive approach to evaluating how different architectures perform in the context of drug repurposing.

Cross-Validation for Robust Evaluation: The use of cross-validation, a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set, is critical. It involves partitioning the data into subsets, training the model on some subsets (training set) and testing it on others (validation set). This approach helps in understanding the model's performance across different data samples, reducing the risk of overfitting, and ensuring that the performance metrics are reliable and consistent across various scenarios. The cross-validation method used will be explained later.

Balancing Different Performance Aspects: The focus on a diverse range of performance metrics, evaluated through cross-validation, indicates an effort to select models that are not just accurate but also reliable and generalizable. This is crucial in drug repurposing where the cost of false predictions can be significant. For this reason, many different performance metrics were used.

Also in this model selection process, baseline models were selected, since including simpler models or baselines such as a Random Predictor helps in setting a comparative standard and appreciating the value added by more complex models.

In summary, the integration of cross-validation into the model selection process for drug repurposing tasks ensures a more thorough and reliable evaluation of the models. This approach underscores the commitment to selecting the best possible models based on their ability to consistently perform well across different data samples, a crucial factor in the high-stakes domain of drug discovery and repurposing.

In the following sections, different models that have been explored will be investigated.

4.3.1. Feedforward Neural Network

As a start to experiments with Deep Learning models simple feedforward neural networks were chosen. It was thought to be a good starting point since it is simple to interpret.

Feedforward neural networks (FNNs) represent a category of artificial neural networks where information moves in a single direction. These networks are structured with input, hidden, and output layers, with each hidden layer comprising multiple artificial neurons. The training process of FNNs involves fine-tuning the weights of the connections between neurons. They find wide-ranging uses in numerous fields, such as image recognition, natural language processing, and predicting time series data.

4.3.2. LSTM

Long Short-Term Memory (LSTM) models, a variant of Recurrent Neural Networks (RNNs), have proven to be particularly effective for classification tasks involving sequential data, such as time series analysis or natural language processing. LSTM models are designed to remember patterns over time and are thus well-suited for classifying, predicting, and generating sequences. LSTMs, given their capability to handle sequence data, can be used in drug repurposing, especially considering the sequential nature of biological and chemical data, like genetic sequences, chemical structures, and temporal patient data. LSTM-based models for drug repurposing can provide valuable insights by identifying patterns and associations in sequential data. For this reason, LSTM models were implemented with implemented with different configurations and tested.

4.3.3. LSTM with Attention Mechanism

Attention layers have emerged as a key innovation in deep learning, particularly when integrated with Long Short-Term Memory (LSTM) networks. These layers boost the proficiency of LSTM in processing sequential data by allowing the model to selectively concentrate on certain segments of the input sequence during prediction. This feature proves invaluable in applications such as language translation, speech recognition, and text summarization.

- Exploring the Attention Mechanism: The attention mechanism empowers the model to dynamically prioritize specific sections of the input sequence while constructing each segment of the output sequence. It assesses the significance or relevance of each input timestep in relation to the present output.
- Context Vector Creation: For every output timestep, the attention mechanism creates a context vector. This vector is a weighted aggregate of the input sequence's hidden states, where the weights denote the pertinence of each input timestep to the ongoing output.

- Calculation of Alignment Scores: The weights are derived using an alignment function, which evaluates the compatibility of inputs near a certain position 'i' with the output at position 'j'.

To encapsulate, attention layers significantly enhance LSTM networks by facilitating a dynamic engagement with different portions of the input sequence. This enhancement substantially improves the model's learning capacity and generalization, especially in intricate sequence modeling tasks.

Implementation

Envision an LSTM (Long Short-Term Memory) network as an adept memory specialist, tasked with comprehending and condensing extensive narratives. Within this framework, the attention mechanism functions akin to an advanced, selective highlighter, meticulously guiding the LSTM to concentrate on the narrative's pivotal elements.

As the LSTM processes each word of the text sequentially, it meticulously catalogs and integrates key information. Upon reaching the conclusion, it synthesizes a summary, encapsulating the essence of the story. Here, the attention mechanism plays a crucial role by enabling the LSTM to prioritize and emphasize the segments of the text that are most pertinent, thereby enhancing the precision and succinctness of the summary.

The attention mechanism operates akin to a meticulous evaluator, allocating significance scores to each word in the text, and gauging their relative importance within the overall narrative. The LSTM, leveraging these evaluations, judiciously selects the words that should be featured in the summary.

Furthermore, the efficacy of the attention mechanism evolves through continuous training on diverse texts, progressively refining the LSTM's capability to interpret and summarize new material more effectively. This ongoing learning process fortifies the LSTM's proficiency in distilling the core message from complex and varied narratives.

Feature/Component	Model Configuration
Number of Inputs	2 (Input1 and Input2, both with shape (1, 100))
LSTM Layers for Input1	
- Number of Layers	3
- Units per Layer	256 (LSTM_1), 128 (LSTM_12), 64 (LSTM_13)
Attention Layer for Input1	
- Units	128
LSTM Layers for Input2	
- Number of Layers	3
- Units per Layer	256 (LSTM_2), 128 (LSTM_22), 64 (LSTM_23)
Attention Layer for Input2	
- Units	128
Concatenation Layer	
- Name	Concatenate_layer
Dense and Dropout Layers	
- Dense Layer Units	64 (activation: relu)
- Dropout Rate	0.2
Output Layer	
- Units	1 (sigmoid activation)

Table 4.1: Configuration of the LSTM model

4.3.4. AutoEncoders

Autoencoders are a type of artificial neural network used primarily for unsupervised learning tasks, particularly for the purpose of dimensionality reduction or feature learning. The basic idea of an autoencoder is to learn a compressed representation of the input data, typically for the purpose of data reconstruction [2]. Structure of Autoencoders An autoencoder typically consists of two main parts:

- Encoder: This part of the network compresses the input into a latent-space representation. It encodes the input data as a compressed representation in a reduced dimension [19]. The encoder layer transforms the input into a smaller, dense representation, which is a lower-dimensional space than the input data.
- Decoder: This part of the network reconstructs the input data from the compressed representation. The decoder layer takes the encoded data and expands it back to the original input shape — ideally, the output of the decoder is a close match to the original input data.

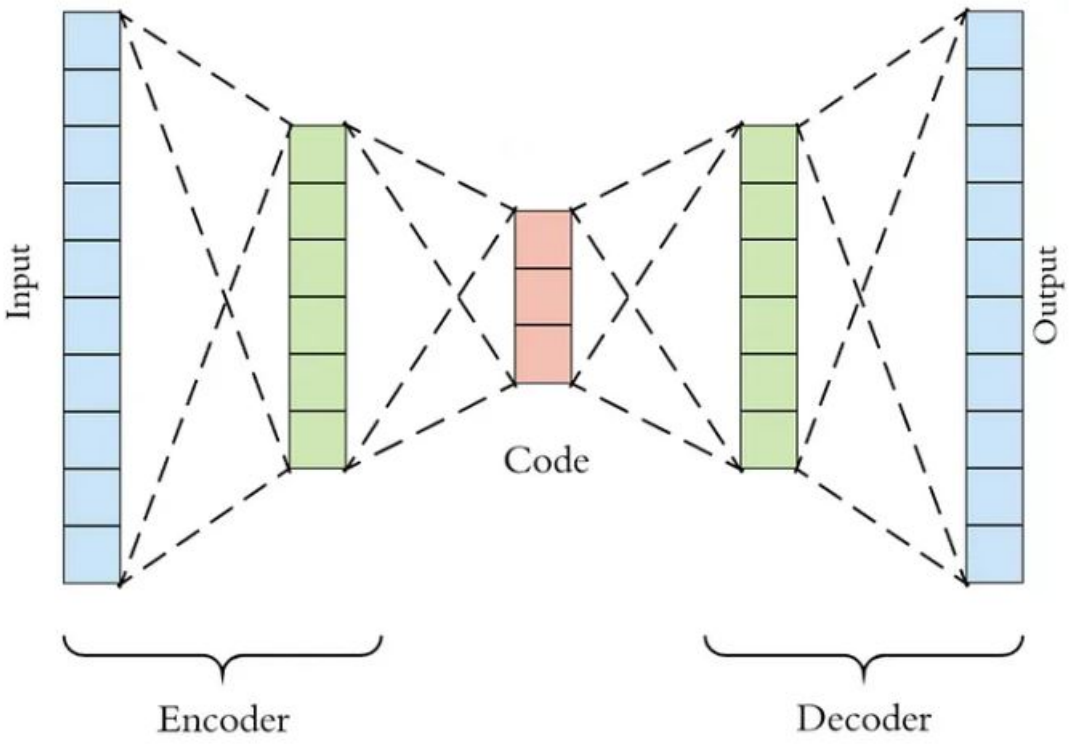

Figure 4.5: Autoencoder Architecture

Autoencoders have been extensively utilized in various fields, including biology, where they can be applied to extract informative features from protein sequences and molecular structures. Using autoencoders for pre-training in a binary classification model where the data consists of protein sequences and molecules involves several steps, which will be explored later on. Autoencoders can be effectively utilized for pre-training in binary classification models, serving to learn a robust representation of input data in an unsupervised manner before fine-tuning the model for classification. This strategy is particularly helpful in scenarios where labeled data for classification is scarce or expensive to obtain.

Ensuring the complexity of the autoencoder and classifier is suitable for the size and nature of the dataset is key to preventing overfitting and underfitting. For this reason, different models have been tried and evaluated with cross validation. Thorough evaluation using metrics like precision, recall, F1-score, and AUC-ROC, especially in the context of imbalanced datasets, is crucial for assessing model performance in biological classifications. For this reason throughout this thesis, different evaluation metrics were used as they will be introduced in the next chapter. Autoencoders can serve as powerful tools for unsupervised feature learning from biological data like protein sequences and molecular

structures, subsequently enhancing the performance of binary classification models in predictive tasks. Balancing the complexity of the model, ensuring robust pre-processing, and rigorously evaluating model performance are key to successfully employing autoencoders in this domain.

Implementation

During the execution phase, various dataset combinations underwent experimentation. Initially, BindingDb was the sole source for both pretraining and classification tasks. Subsequently, two alternative datasets were explored for pretraining purposes: HomosapiensDb and AllProDb, both containing a broader range of protein sequences. Similar to other models, a variety of configurations, varying in the number of layers and neurons, were also subjected to testing. The configuration of the model that has the best performance is given in Table 4.2.

Feature/Model	**Autoencoder 1**	**Autoencoder 2**	**Final Combined Model**
Number of Input Features	100	200	100 & 200
Encoder Layers			
- Number of Layers	3	3	3 each (total 6)
- Neurons (per Layer)	64, 40, 25	100, 50, 25	64, 40, 25 & 100, 50, 25
Decoder Layers			
- Number of Layers	3	3	1
- Neurons (per Layer)	40, 64, 100	50, 100, 200	50

Table 4.2: Configuration of the Autoencoder model

4.3.5. LSTM AutoEncoders

Before, use cases for LSTM and autoencoders were seen. Here the two will be used in a combined manner. When LSTMs and autoencoders are combined, we get LSTM autoencoders. These are particularly useful for sequence data. This model consists of a pre-training phase just like the previous autoencoder model. LSTM Autoencoders have been used in medical research in previous studies [11] but not in the drug repurposing context. In the LSTM Autoencoder, there is an additional layer called RepeatVector which is used to create a bridge between the encoder and autoencoder while using LSTM layers. The RepeatVector layer in Keras is a utility layer that repeats its input a specific number of times. It's often used in sequence processing models, particularly in sequence-to-sequence models like autoencoders and Recurrent Neural Networks (RNNs).

Functionality: The RepeatVector layer takes a 2D input (a single vector) and converts it into a 3D output (a sequence of vectors). It repeats its input vector n times, where n is

a parameter you specify. For instance, if the input to RepeatVector(n) is a vector [a, b, c], and n=3, the output will be [[a, b, c], [a, b, c], [a, b, c]]. Usage in Autoencoders:

In the context of an LSTM autoencoder, the RepeatVector layer is crucial for bridging the gap between the encoder and decoder. The encoder typically processes the input sequence and compresses it into a lower-dimensional representation (a single vector). This is done by LSTM layers that do not return sequences. The decoder, on the other hand, is designed to process sequences. To transform the compressed representation back into a sequence, RepeatVector is used. It repeats the single vector output from the encoder to create a sequence that can be fed into the decoder's LSTM layers.

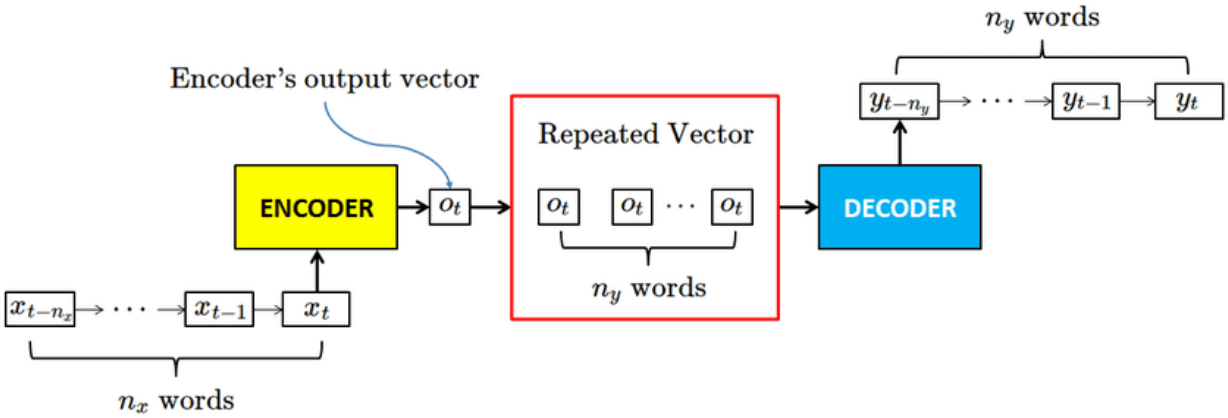

Figure 4.6: LSTM Autoencoder Architecture

Imagine an LSTM autoencoder that processes input sequences with a length of 10. The encoder part compresses these inputs into a solitary vector. However, for the purpose of reconstructing the initial sequence, the decoder requires more than just this single vector - it needs a full sequence. This is where RepeatVector comes into play; it duplicates the condensed vector 10 times, creating a sequence that mirrors the length of the original input. This allows the decoder to work towards rebuilding the initial sequence.

To put it briefly, RepeatVector is a tool in sequence-to-sequence models that transforms a singular vector into a series of vectors. This transformation is crucial for moving from a compressed form back to a sequence structure that the decoder can work with.

This model has 2 parts like classical autoencoders:

- Encoder: The encoder processes the input sequence and returns its own internal state. For this, LSTM layers are used, and only the final LSTM state (after processing the sequence) is passed on, which serves as the compressed representation

of the input.

- Decoder: This part aims to reproduce the input sequence from the internal state. Again, LSTM layers are used. The decoder takes the final state of the encoder as its initial state and tries to generate the original sequence. The RepeatVector layer in Keras can be used to convert the encoder's final state to the initial sequence for the decoder.

The LSTM autoencoder is trained by feeding a sequence into the encoder, which then produces a compressed representation. This representation is then fed into the decoder to produce the output sequence. The model is trained to minimize the difference between the input and output sequences.

In LSTM Autoencoders, the encoder's output can be used for feature extraction for other tasks. This use case will was used in this thesis.

Implementation

In this context, using an LSTM autoencoder for protein and molecule sequences means we're attempting to capture the intrinsic patterns and structures within those sequences in a compressed manner. Once trained, the encoder's representation can then be utilized as meaningful features for further tasks, like classification. In Table 4.3 configuration of the best-performing LSTM Autoencoder model can be seen.

Feature/Model	**Autoencoder2**	**Autoencoder1**	**Final Combined Model**
Maximum Sequence Length	1	1	1
Number of Input Features	300	100	100 & 300
Encoder LSTM Layers			
Number of Layers	2	2	2 each (total 4)
Neurons (per Layer)	128, 64	64, 32	64, 32 & 128, 64
Decoder LSTM Layer			
Neurons	300	100	N/A
Dense Layers			
Neurons	N/A	N/A	96

Table 4.3: Configuration of the LSTM Autoencoder model

During the implementation, different combinations with different datasets were tested. The first option was using only BindingDb for both pretraining and classification. After that two other datasets were tested as the pretraining data. These are HomosapiensDb and AllProDb which include many additional protein sequences. Like other models, different configurations with different numbers of layers and neurons were put into tests.

5 | Experiments

5.1. Experiment design

5.1.1. Cross Validation

Cross-validation is a robust statistical technique used to assess the performance of machine learning models, including deep learning models in classification tasks. It's particularly valuable because it provides a more generalized performance metric than a single train/test split. For these reasons cross validation was chosen to evaluate the models that has been used. Here's a detailed explanation:

The definition of cross-validation is it involves partitioning the original dataset into a training set to train the model and a test set to evaluate it. However, unlike a simple split, it does this multiple times in different ways. The most common type of cross validation is k-fold cross-validation. The data is partitioned into 'k' segments, and the model is trained on 'k-1' segments, with the remaining segment serving as the test set. This process is repeated 'k' times, cycling through the segments used for testing. Here for the 'k' value 5 was chosen since it shows a balance between complexity and evaluation power. Choosing a value higher would be computationally expensive in this drug repurposing task. Why Use Cross-Validation in Deep Learning for Classification?

- Model Generalization: It helps in assessing how well the deep learning model will generalize to an independent dataset.
- One benefit of cross validation is mitigating overfitting. Since deep learning models are prone to overfitting, especially with limited data, cross-validation ensures that the model's performance is not just a result of the specific way the data was split.
- It can also be used in hyperparameter tuning. It's a reliable method for tuning hyperparameters. By evaluating different hyperparameters across the folds, one can choose the set that performs best on average.
- Cross validation also provides a more robust and less biased estimate of the model's

performance, especially important in classification tasks where the balance of classes can vary.

Performance Evaluation: After training and validating all folds, the performance metrics (like accuracy, precision, recall, F1-score) are averaged out. This average performance is a more reliable estimate of how the model will perform on unseen data.

Cross validation was also used to optimize hyperparameters. Choose the configuration that yields the best average performance across all folds.

First thing to consider while applying cross validation is computational cost.Cross-validation can be computationally expensive, especially with large datasets and complex deep learning models. Each fold essentially requires training a new model from scratch. For this reason 5 was the most feasible option since values more than 5 would be computationally quite expensive for this use case.

The second aspect to consider is data representativeness: It's crucial that each fold is representative of the overall dataset, especially in terms of class distribution in classification tasks.

Also, k value selection is quite important in application. The choice of 'k' (e.g., 5 or 10) can impact the balance between bias and variance in the model assessment. More folds typically provide a more accurate estimate but at a higher computational cost.

Cross-validation is a powerful tool in the machine learning workflow, especially for deep learning models in classification tasks. It helps in rigorously assessing a model's ability to generalize beyond the training data, guiding decisions about model architecture and hyperparameters. Despite its computational intensity, the benefits it offers in terms of robust model evaluation are often worth the extra time and resources [18].

5.2. Evaluation Metrics

In this part, different evaluation metrics that have been used throughout this thesis will be discussed in terms of their advantages and disadvantages in certain conditions. This will be a useful guide to evaluating different methods that have been used in the later chapters.

5.2.1. Accuracy

Accuracy is a performance metric that measures the overall correctness of a model's predictions. It tells you the percentage of correctly classified instances out of the total

number of instances in the dataset. Accuracy is defined as:

$$Accuracy = \frac{NumberofCorrectPredictions}{TotalNumberofPredictions} \tag{5.1}$$

- Correct Predictions: In the context of drug repurposing, "correct predictions" refer to cases where the model correctly identifies whether a drug is suitable for repurposing or not. If the model accurately classifies drugs based on their potential for repurposing, these are considered correct predictions.
- Total Predictions: This is the sum of all the predictions the model makes, including both true positives (correctly identified repurposable drugs), true negatives (correctly identified non-repurposable drugs), false positives (non-repurposable drugs mistakenly identified as repurposable), and false negatives (repurposable drugs mistakenly identified as non-repurposable).

Accuracy provides an overall view of how well the model is performing in terms of correctly classifying drugs for repurposing. It represents the proportion of correct predictions out of all predictions, regardless of whether they are positive (repurposable) or negative (non-repurposable).

While accuracy is a widely used metric, it's important to consider its limitations, especially in imbalanced datasets. In drug repurposing, where the number of potential repurposable drugs may be much smaller than non-repurposable drugs, a high accuracy can be achieved by simply predicting all drugs as non-repurposable. This would not be useful in practice, as it would miss potential repurposing opportunities. Therefore, accuracy should be considered alongside other metrics like precision, recall, and F1 score to get a more comprehensive understanding of the model's performance, especially in situations with imbalanced classes. For this reason, more metrics to evaluate the performances of the models will be introduced.

5.2.2. Precision

Precision is a fundamental performance metric used in classification tasks, including machine learning and deep learning. It measures the accuracy of positive predictions made by a model. In the context of classification, we often have two classes: positive (the class of interest) and negative (everything not in the positive class). Precision is defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{5.2}$$

Here's what each term means:

- True Positives (TP): These are the instances that are actually in the positive class, and the model correctly predicted them as positive.
- False Positives (FP): These are the instances that are not in the positive class, but the model incorrectly predicted them as positive.

Precision essentially tells us: out of all the things the model labeled as "positive," how many were truly positive? Some key points about precision include:

It focuses on the accuracy of positive predictions. It tells you how reliable the model's positive predictions are.

High precision means that when the model predicts something as positive, it is usually correct. It indicates a low rate of false positives.

Precision is particularly important when false positives are costly or undesirable. For example, in medical diagnosis, you want a model with high precision to avoid unnecessary treatments or surgeries.

they often have an inverse relationship, meaning that improving one metric typically comes at the expense of the other. Increasing precision may lead to a decrease in recall, and vice versa.

In summary, precision is a critical metric for assessing the quality of a model's positive predictions. It helps you evaluate how well the model performs when it claims that something belongs to the positive class.

In the context of drug repurposing, precision plays a crucial role in assessing the performance of classification models. Here's what can be said about precision in this context: Precision is essential because you want to be highly confident that the drugs recommended by the model are indeed effective for the target disease. High precision means that when the model suggests a drug, it's likely to be a genuinely promising candidate for repurposing.

Also minimizing false positives is an important part of drug repurposing. False positives in drug repurposing can be costly and potentially harmful. Recommending a drug that is not effective for the target disease could waste resources and time and potentially harm patients. Therefore, high precision in drug repurposing models is desirable to reduce the likelihood of false positive drug recommendations.

Balancing precision and recall is essential in drug repurposing. While high precision is

desirable to ensure the safety and efficacy of recommended drugs, you also don't want to miss out on potential candidates. Therefore, finding the right balance between precision and recall is critical, as overly stringent criteria for precision might lead to missing out on promising repurposing opportunities.

Drug repurposing often requires domain expertise in pharmacology and biology. High precision in a model can be achieved by incorporating domain knowledge and careful data curation. Domain experts can help validate and refine the model's predictions to ensure they align with the current state of scientific understanding.

In summary, precision is a vital metric in drug repurposing classification models because it directly relates to the reliability and safety of drug recommendations. High precision is desirable to reduce the risk of false positive predictions, but it should be balanced with other considerations, such as recall and domain expertise, to ensure that valuable repurposing opportunities are not missed.

5.2.3. Recall

Recall, also known as Sensitivity or True Positive Rate, is a fundamental performance metric used in classification tasks, including machine learning and deep learning. Recall measures the ability of a model to correctly identify all relevant instances in a dataset. In the context of classification, we often have two classes: positive (the class of interest) and negative (everything not in the positive class). Recall is defined as:

$$Precision = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5.3}$$

In simple terms, recall answers the question: "Of all the instances that are actually in the positive class, how many did the model correctly identify as positive?"

Some key points about recall include, Recall focuses on the ability of the model to capture all relevant instances of the positive class. It tells you how well the model avoids missing positive cases. High recall means that the model is good at finding most of the relevant instances of the positive class. It indicates a low rate of false negatives.

Recall is particularly important when it's crucial not to miss any positive instances, even at the cost of some false positives. For example, in medical diagnosis, you want a model with high recall to ensure that potentially life-threatening conditions are not overlooked.

In summary, recall is a critical metric for assessing the completeness of a model's predictions regarding the positive class. It helps you evaluate how well the model identifies

and includes all relevant instances in the positive class, which is important in various classification tasks, including drug repurposing.

5.2.4. F1 Score

The F1 score is a performance metric commonly used in classification problems, including those related to drug repurposing. It is especially valuable when dealing with imbalanced datasets, where one class is significantly smaller than the other class. The F1 score is the harmonic mean of precision and recall, and it balances the trade-off between these two metrics. It is defined as:

$$F1Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5.4}$$

The F1 score is the balance between precision and recall. It is particularly useful when you want to find a trade-off between minimizing false positives (precision) and ensuring that all positive instances are captured (recall).

In drug repurposing, the F1 score helps you evaluate the model's ability to identify potential repurposable drugs while maintaining a reasonable level of precision. It addresses the challenge of imbalanced datasets, where the majority of drugs may be non-repurposable, and it ensures that the model doesn't overly bias predictions toward the majority class. Since the dataset used also includes some imbalance this metric is useful for the context of this thesis. A high F1 score indicates that the model is performing well in terms of both precision and recall, striking a balance between correctly identifying repurposable drugs and avoiding false positives. It is a valuable metric when the goal is to discover promising candidates for drug repurposing while minimizing the risk of recommending ineffective or unsafe drugs.

5.2.5. AUC

AUC refers to "Area Under the (Receiver Operating Characteristic) Curve." When applied to drug repurposing, or any other domain using classification models, AUC is a widely used evaluation metric to understand the performance of a binary classifier.

AUC-ROC Explained: ROC Curve: The Receiver Operating Characteristic (ROC) curve is a graphical representation of the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across different thresholds for classifying an instance.

AUC: The Area Under the ROC Curve quantifies the overall performance of the binary

classification model, representing the likelihood that the model will rank a randomly chosen positive instance higher than a randomly chosen negative one.

In the context of drug repurposing, using machine learning and classification models, it is possible to utilize AUC as a metric to evaluate how well the models distinguish between two classes: for example, effective and non-effective drug candidates for a new therapeutic application. Here's how:

- Positive Class: Compounds (or drugs) that are effective or show a desirable effect against a particular disease or condition.
- Negative Class: Compounds that are not effective or don't show the desired effect.

The AUC provides a single scalar value representing the overall model performance. A model that predicts classes perfectly has an AUC of 1.0, while a model that predicts classes no better than random has an AUC of 0.5. In the scope of drug repurposing, a high AUC indicates that the model is capable of effectively distinguishing between effective and non-effective drugs, which can be immensely useful for identifying promising drug candidates for further experimental validation.

5.2.6. Matthews Correlation

Since imbalanced databases were included in this project, more evalution metrics were used in order to be able to assess the performance of the models. Matthews correlation coefficient (MCC) is a good example of this since it provides a balanced measure even if the classes are of very different sizes, making it particularly useful in cases where precision and recall may give misleading results. Here's an overview:

The Matthews correlation coefficient is a correlation coefficient between the observed and predicted binary classifications. It returns a value between -1 and +1. A coefficient of +1 represents a perfect prediction. 0 indicates no better than random prediction. -1 indicates total disagreement between prediction and observation.The MCC is calculated using the formula:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.5}$$

There are some characteristics of MCC to consider which includes: Being balanced: It is regarded as a balanced measure which can be used even if the classes are of very different sizes. Interpretability: The value of MCC is easy to interpret. Applicability: It is used in various fields, especially in bioinformatics for validation of protein structure prediction

methods. Robustness: MCC is generally regarded as a robust metric that provides a more truthful representation of the model's performance than other metrics like F1-score, especially in imbalanced datasets. Different advantages of MCC can be listed. Effectiveness in Imbalanced Datasets: MCC is effective for evaluating classifiers on imbalanced datasets, whereas other metrics like accuracy can be misleading. Comprehensive: It takes into account true and false positives and negatives, providing a more comprehensive measure than accuracy alone. If disadvantages are considered they can be listed as: Less Intuitive: For those unfamiliar with it, MCC can be less intuitive than other metrics like accuracy, precision, and recall. Complexity: The calculation is more complex than simpler metrics like accuracy. MCC is particularly useful in medical, biological, or any field where binary classification tasks are performed. MCC is also useful in imbalanced data. It is ideal for datasets where one class is much larger than the other. Comparisons with Other Metrics Unlike accuracy, MCC considers all four quadrants of the confusion matrix (TP, TN, FP, FN). MCC is generally a more reliable statistical rate than precision, recall, or F1 score when dealing with imbalanced datasets.

In summary, the Matthews correlation coefficient is a valuable metric for evaluating the performance of binary classification models, particularly when dealing with imbalanced datasets. It provides a more nuanced and accurate measure of performance compared to more traditional metrics. For this reason, it was selected as one of the metrics to evaluate different models in this drug repurposing task.

6 | Results and Discussion

Model	Accuracy	Precision	Recall	F1 Score
Feedforward Neural Network	88.94	**95.51**	87.02	91.06
LSTM v.1	83.49	92.30	83.29	87.45
LSTM v.2	84.30	92.45	84.16	88.18
LSTM with Attention Mechanism	84.42	93.19	86.25	89.44
LSTM with Attention Mechanism v.2	90.30	87.15	94.28	90.57
Autoencoders BindingDb	81.88	88.97	85.21	87.04
Autoencoders BindingDb v.2	88.27	95.39	87.98	91.53
Autoencoders HomosapiensDb v.1	88.84	95.42	86.70	90.85
Autoencoder HomosapiensDb v.2	88.68	95.54	86.71	90.91
Autoencoder AllProDb v.1	83.64	84.24	82.76	83.49
Autoencoders AllProDb v.2	87,32	86.02	89,16	87.55
LSTM Autoencoders BindingDb	90.60	88.39	93.48	90.86
LSTM Autoencoders HomosapiensDb	91.61	88.56	95.56	91.92
LSTM Autoencoders AllProDb	**91.62**	88.25	**96.04**	**91.98**
Random Predictor with Bias	58.10			

Table 6.1: Results of each model

In table 6.1 results of different models that have been tested throughout this study can be seen. Here the metrics shown are Accuracy, Precision, Recall and F1 score.

To analyze the values from the results table 6.1: Random Predictor with Bias is the baseline model for comparison. It was created by generating random values while giving bias to popular values in order to measure the effect of imbalance in the dataset. Its low accuracy (58,10%) shows it's not a good predictive model, as expected.

Feedforward Neural Network was the beginning point here. This model has a high precision (95,51%) and a good F1 score (91.06%), indicating effective identification of true positives, even though it does not include any complicated structure.

LSTM Models: These models (Long Short-Term Memory) vary in configuration and performance. LSTM with Attention Mechanism v.2 is the best among them, with the highest accuracy (90,3%) and a balanced F1 score (90.57%).

Autoencoder Models: These models are used for learning efficient data codings in an unsupervised manner. The "Autoencoders BindingDb v.2" and "Autoencoders HomosapiensDb" have high accuracy and F1 scores, indicating robust performance.

Overall, the best-performing models appear to be the LSTM Autoencoders and various configurations of Autoencoders, particularly those with pretraining and using HomosapiensDb. These models exhibit a good balance between accuracy, precision, recall, and F1 score, indicating robust and reliable performance.

The variant with AllProDb shows even higher performance metrics with an accuracy of 91.62%, precision at 88.25%, and an exceptional recall of 96.04%. The F1 score is significantly high at 91.98%. This suggests that this model is not only accurate overall but particularly strong in identifying true positive cases (as indicated by the high recall). Its F1 score suggests an excellent balance between precision and recall, making it potentially the most effective model in the table.

In summary, the LSTM Autoencoder models, particularly the "LSTM Autoencoders AllproDb," demonstrates outstanding performance across all metrics. The high recall rates are especially notable, indicating these models are very effective in identifying positive cases, which is often a critical aspect in many machine learning applications.

The results presented in the table are metrics for various models used in this task. Drug repurposing involves finding new uses for existing drugs, which requires accurate and reliable models to predict drug-target interactions, efficacy, or suitability for new diseases. Let's interpret each metric and how the models perform:

AUC (Area Under the Curve): This mentioned metric is derived from the Receiver Operating Characteristic (ROC) curve and measures the ability of the model to distinguish between the classes (effective vs. non-effective drugs for new purposes). A higher AUC

Model	AUC	Matthews Correlation
Autoencoder with AllProDb v.1	91.19	67.16
Autoencoders with AllProDb v.2	94.47	74.56
LSTM Autoencoders AllProDb	96.3	81.19
LSTM Autoencoders BindingDb	96.51	82.78
LSTM Autoencoders HomosapiensDb	**96.52**	83.32
LSTM Autoencoders AllProDb	96.37	**83.42**

Table 6.2: AUC and Matthews Correlation for Top Models

indicates better model performance. An AUC close to 1.0 suggests excellent model performance, while an AUC closer to 0.5 suggests no discriminative power.

Matthews Correlation Coefficient (MCC): This is a more informative metric than accuracy, especially for imbalanced datasets, which are common in drug discovery. It considers true and false positives and negatives, providing a balanced measure even if the classes are of very different sizes. A coefficient of +1 represents a perfect prediction, 0 is no better than a random prediction, and -1 indicates total disagreement between prediction and observation.

Looking at the results:

Autoencoder Models: These models have high AUC values (above 90), indicating good predictive capabilities. Their MCC values are also relatively high, suggesting that the predictions made by these models are reliable and not skewed by class imbalance.

LSTM Autoencoder Models demonstrate even higher AUC values, nearing 96, indicating excellent predictive performance. The MCC values are also robust (above 80), suggesting a high level of reliability in the predictions, considering the balance of true and false positives and negatives, while LSTM Autoencoder with AllProDb shows the best results.

So values in this table show results that match with previous ones. This gives the model more reliability. Having a high matthews correlation value is especially important here since it gives good results with imbalanced datasets. Since imbalanced datasets are common in drug repurposing and the dataset that has been here has similar characteristics, these results imply that this model is promising with different types of data.

In summary, different models show promising results in the drug repurposing task, with LSTM Autoencoder models showing powerful performance. These models are likely effective in distinguishing between drugs that can be repurposed and those that cannot, and their predictions are balanced and reliable, as indicated by the high MCC values.

This makes them very promising for use in drug repurposing research, where accuracy and reliability are paramount.

In the pretrained LSTM Autoencoder models, even the beginner model with only BindingDb showed good results. Having more protein sequences in HomosapiensDb gives better, more balanced predictions. Since AllProDb has more diverse protein sequences when it is added to the used datasets it results in slight increases in almost all metrics. These factors show us the potential of the LSTM Autoencoder models in drug repurposing.

In conclusion, the task of drug repurposing within the BindingDb dataset was successfully undertaken using LSTM autoencoders, as evidenced by key performance metrics. The effectiveness of this approach is clearly reflected in the substantial improvements across various metrics, including accuracy, precision, recall, F1 score, and Matthews correlation coefficient. These results underscore the capability of LSTM autoencoders to adeptly navigate and interpret the complexities of imbalanced datasets, providing a robust and efficient solution to predict novel interactions between proteins and ligands in challenging data environments.

7 | Conclusions and Future Developments

In conclusion, the comprehensive analysis of various machine learning models, as detailed in the previous chapters, offers insightful revelations for a drug repurposing task. The evaluation spans a range of performance metrics, including Accuracy, Precision, Recall, F1 Score, Area Under the Curve (AUC), and Matthews Correlation Coefficient (MCC), providing a multifaceted view of each model's capabilities.

The LSTM Autoencoder models, particularly those configured with AllProDb, BindingDb, and HomosapiensDb data, emerged as the top performers. These models not only demonstrated excellent AUC values, nearing 96, indicative of their superior ability to discriminate effectively between suitable and unsuitable drugs for repurposing but also showcased robust MCC values above 80. This high MCC metric is especially crucial as it signifies that the models' predictions are not only accurate but also reliable and well-balanced, considering the true and false positives and negatives — a critical aspect in the context of drug repurposing where the cost of false predictions can be high.

As seen in the results, the LSTM Autoencoder models and their variations stand out. These models exhibit exceptional balance across all key metrics. The "LSTM Autoencoders with all the protein data configuration, in particular, display a notable blend of high precision and recall, culminating in a remarkable F1 score. This suggests its proficiency in accurately identifying drugs suitable for repurposing while minimizing false positives and negatives. Also in the other metrics, the LSTM Autoencoders demonstrate superior performance, as indicated by their high AUC values and robust MCC scores. These high scores are indicative of the models' excellent predictive power and reliability, crucial in drug repurposing where precision is paramount.

This extensive analysis highlights the immense potential of machine learning in drug repurposing. The LSTM Autoencoder models, in particular, demonstrate outstanding predictive accuracy and reliability, making them highly suitable for this task. Their ability to deliver balanced predictions across multiple metrics suggests strong applicability in the

complex domain of drug repurposing.

These findings underscore the potential of advanced machine learning techniques in revolutionizing drug repurposing. By efficiently identifying promising repurposing opportunities, these models can significantly accelerate the drug development process, potentially leading to quicker and more cost-effective therapeutic solutions. This is particularly valuable in the pharmaceutical industry, where the traditional drug discovery and development processes are lengthy and expensive.

In light of these results, future work should focus on further optimizing these models, exploring their applicability to a broader range of datasets, and integrating them into a holistic drug discovery framework. The integration of machine learning into drug repurposing not only promises to enhance the efficiency of the drug development process but also opens new avenues for discovering therapeutic options for unmet medical needs.

For the real-world application and validation of this study, applying these models in real-world scenarios, such as in clinical trials or in silico screenings, would provide valuable feedback on their practical utility and areas for improvement. Also enhancing model interpretability to understand the rationale behind predictions can build trust and provide valuable insights for researchers and clinicians.

Another possible approach is integrating these models into existing drug discovery pipelines. This could provide pharmaceutical companies with powerful tools to identify repurposing candidates more efficiently.

Collaboration with Bioinformatics and Pharmacology experts could also be beneficial in this kind of study. Collaborative efforts with experts in bioinformatics, pharmacology, and clinical sciences could lead to more holistic and interdisciplinary approaches to drug repurposing.

Looking forward, these results pave the way for a more integrated and data-driven approach in the pharmaceutical industry. Optimizing these models and exploring their applicability to diverse datasets could revolutionize drug discovery, leading to more efficient and cost-effective therapeutic solutions. The integration of robust machine learning models like LSTM Autoencoders in drug repurposing signifies a promising step toward addressing the urgent need for effective and accessible treatments in various medical domains. Here it was seen that with the number of protein sequences in the dataset for proteins increasing, the performance of the model was increasing. So for the possible future works on these adding more data for the pretraining phase might increase the performance of these models. At the same time, it is possible to extend the number of layers

and neurons in the models, of course, these are limited by the computation power of the system. Also, different architectures that are combined with LSTM Autoencoders might prove useful. So Future work could explore the integration of additional datasets, including more diverse molecular and clinical data, to further enhance the models' predictive capabilities.

In conclusion, the study showcases the potential of advanced machine learning models in revolutionizing drug repurposing through accurate and reliable classification using protein sequences and ligands. The promising results pave the way for more efficient, cost-effective, and innovative approaches to therapeutic discovery, highlighting the significant role of machine learning in the future of pharmaceutical research and development.

Bibliography

[1] N. C. Baker, S. Ekins, A. J. Williams, and A. Tropsha. A bibliometric review of drug repurposing. *Drug Discovery Today*, 23(3):661–672, 2018. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2018.01.018. URL `https://www.sciencedirect.com/science/article/pii/S1359644617302878`.

[2] D. H. Ballard. Modular learning in neural networks. In *Proceedings of the Sixth National Conference on Artificial Intelligence - Volume 1*, AAAI'87, page 279–284. AAAI Press, 1987. ISBN 0934613427.

[3] Binding Database. Bindingdb. `https://www.bindingdb.org`, 2023.

[4] M. De Rosa, R. Purohit, and A. García-Sosa. Drug repurposing: a nexus of innovation, science, and potential. *Scientific Reports*, 13:17887, 2023. doi: 10.1038/s41598-023-44264-7. URL `https://doi.org/10.1038/s41598-023-44264-7`.

[5] G. De Simone, D. Sardina, M. Gulotta, and U. Perricone. Kuala: a machine learning-driven framework for kinase inhibitors repositioning. *Scientific Reports*, 12(1):17877, 2022. doi: 10.1038/s41598-022-22324-8. URL `https://doi.org/10.1038/s41598-022-22324-8`. PMID: 36284125; PMCID: PMC9595087.

[6] L. Deng, Y. Zeng, H. Liu, Z. Liu, and X. Liu. Deepmhadta: Prediction of drug-target binding affinity using multi-head self-attention and convolutional neural network. *Current Issues in Molecular Biology*, 44(5):2287–2299, 2022. doi: 10.3390/cimb44050155. URL `https://doi.org/10.3390/cimb44050155`. PMID: 35678684; PMCID: PMC9164023.

[7] S. D'Souza, K. Prema, and S. Balaji. Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discovery Today*, 25(4):748–756, 2020. ISSN 1359-6446. doi: https://doi.org/10.1016/j.drudis.2020.03.003. URL `https://www.sciencedirect.com/science/article/pii/S1359644620301033`.

[8] A. Ghimire, H. Tayara, Z. Xuan, and K. Chong. Csatdta: Prediction of drug-target binding affinity using convolution model with self-attention. *International Journal*

of Molecular Sciences, 23(15):8453, 2022. doi: 10.3390/ijms23158453. URL https://doi.org/10.3390/ijms23158453. PMID: 35955587; PMCID: PMC9369082.

[9] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester. Simboost: a read-across approach for predicting drug-target binding affinities using gradient boosting machines. *Journal of Cheminformatics*, 9(1):24, 2017. doi: 10.1186/s13321-017-0209-z. URL https://doi.org/10.1186/s13321-017-0209-z.

[10] K. Huang, T. Fu, L. M. Glass, M. Zitnik, C. Xiao, and J. Sun. DeepPurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 36(22-23): 5545–5547, 12 2020. ISSN 1367-4803. doi: 10.1093/bioinformatics/btaa1005. URL https://doi.org/10.1093/bioinformatics/btaa1005.

[11] M. R. Ibrahim, J. Haworth, A. Lipani, N. Aslam, T. Cheng, and N. Christie. Variational-lstm autoencoder to forecast the spread of coronavirus across the globe. *PLOS ONE*, 16(1):1–22, 01 2021. doi: 10.1371/journal.pone.0246120. URL https://doi.org/10.1371/journal.pone.0246120.

[12] T. P. Kenakin. *A pharmacology primer: theory, application and methods*. Academic Press, 2009.

[13] Y. Ko. Computational drug repositioning: Current progress and challenges. *Applied Sciences*, 10:5076, 07 2020. doi: 10.3390/app10155076.

[14] B. Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. doi: 10.1007/s13748-016-0094-0. URL https://doi.org/10.1007/s13748-016-0094-0.

[15] Y. Y. Li and S. S. Jones. Drug repositioning for personalized medicine. *Genome Medicine*, 4(3):27, 2012. doi: 10.1186/gm326. URL https://doi.org/10.1186/gm326.

[16] H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: deep drug-target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018. doi: 10.1093/bioinformatics/bty593. URL https://doi.org/10.1093/bioinformatics/bty593. PMID: 30423097; PMCID: PMC6129291.

[17] S. Pushpakom. Introduction and Historical Overview of Drug Repurposing Opportunities. In *Drug Repurposing*. The Royal Society of Chemistry, 02 2022. ISBN 978-1-78801-903-3. doi: 10.1039/9781839163401-00001. URL https://doi.org/10.1039/9781839163401-00001.

[18] C. Schaffer. Selecting a classification method by cross-validation. *Machine Learning*,

13(1):135–143, 1993. doi: 10.1007/BF00993106. URL https://doi.org/10.1007/BF00993106.

[19] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, Jan. 2015. ISSN 0893-6080. doi: 10.1016/j.neunet.2014.09.003. URL http://dx.doi.org/10.1016/j.neunet.2014.09.003.

[20] J. Souza, M. Fernandes, and R. De Melo Barbosa. A novel deep neural network technique for drug–target interaction. *Pharmaceutics*, 14:625, 03 2022. doi: 10.3390/pharmaceutics14030625.

[21] B. Wei, Y. Zhang, and X. Gong. Deeplpi: a novel deep learning-based model for protein–ligand interaction prediction for drug repurposing. *Scientific Reports*, 12: 18200, 2022. doi: 10.1038/s41598-022-23014-1. URL https://doi.org/10.1038/s41598-022-23014-1.

[22] Y.-F. Zhang, X. Wang, A. C. Kaushik, Y. Chu, X. Shan, M.-Z. Zhao, Q. Xu, and D.-Q. Wei. Spvec: A word2vec-inspired feature representation method for drug-target interaction prediction. *Frontiers in Chemistry*, 7, 2020. ISSN 2296-2646. doi: 10.3389/fchem.2019.00895. URL https://www.frontiersin.org/articles/10.3389/fchem.2019.00895.

List of Figures

List of Tables

Acknowledgements

I extend my deepest gratitude to my family and friends, whose unwavering love and support were instrumental in my success. Without their encouragement, this endeavor would have been insurmountable. I am also immensely thankful to my professor and supervisor, who provided invaluable guidance and constructive feedback throughout my research journey. Their expertise and mentorship were invaluable in shaping my understanding and facilitating my progress.