POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Investigating Deep Learning Methods for Drug Repurposing Predictions

Laurea Magistrale in Computer Science and Engineering - Ingegneria Informatica

Author: Ismail Fatih Gonen

Advisor: Prof. Pietro Pinoli

Academic year: 2022-2023

1. Introduction

Drug repurposing is an approach in the medical field that involves finding new therapeutic uses for existing drugs. This strategy offers several advantages such as speeding up the drug development process, reducing toxicity risks, and lowering costs. A key aspect of drug repurposing is understanding drug-target interactions, which are the mechanisms through which drugs exert their effects on specific targets like proteins or cells in the body.

Advancements in artificial intelligence (AI) and machine learning, particularly deep learning, have greatly enhanced the process of drug repurposing. These technologies enable the analysis of vast amounts of biological and chemical data, leading to the prediction of new uses for existing drugs, the identification of drug targets, and the optimization of drug design. AI's predictive models facilitate the efficient screening of molecules, potentially saving time and resources in the drug discovery process. Furthermore, these models aid in understanding complex drug-target interactions, predicting off-target effects, and assessing drug safety and efficacy across different patient groups.

Deep learning, specifically, is crucial in analyzing complex biological data like protein sequences and ligands such as small molecules. Previous studies suggested models like DeepLPI [4] that use deep learning for drug-target interaction predictions. These technologies are instrumental in drug repurposing, offering a faster, more efficient alternative to traditional methods by leveraging raw data for automatic feature extraction. But there is still a lot of room for further studies in this area since it can transform the way to develop new drugs in a really innovative way.

The focus of this study is to predict novel interactions between proteins and ligands. Such tasks aid in understanding drugs' mechanism of action and in the development of personalized medicine strategies, potentially improving treatment efficacy and reducing side effects. Additionally, drug repurposing through AI and machine learning can contribute to cost-effective healthcare solutions, particularly in the treatment of rare diseases and in combating drug resistance in infectious diseases and cancer.

2. Dataset and Preprocessing

Throughout this thesis, three distinct datasets were utilized, with BindingDB [1] being the primary and initial one. BindingDB is a publicly accessible online database that compiles binding affinity data, focusing mainly on the interactions

between proteins, which are potential drug targets, and small, drug-like molecules.
The information in BindingDB is derived from a variety of measurement methods such as enzyme inhibition, kinetics, isothermal titration calorimetry, NMR, radioligand assays, and competition assays. This database contains data gathered from scientific publications, patents, select PubChem confirmatory BioAssays, and ChEMBL entries that offer clearly defined protein targets.
The other two datasets are used for pretraining data in different models that will be explained later. These two datasets were used with the names "Homosapiens Db" and "AllProDb". These datasets include more protein sequences in addition to BindingDb.
Proteins are expressed as sequences of 20 amino acids while ligands are encoded as smiles.

	BindingDb	HomosapiensDb	AllProDb
Total Proteins	84,840	20,598	79,006
Unique Proteins	2,483	20,528	75,948

Table 1: Number of Proteins in Datasets

2.1. Affinity and Drug Target Interaction

In the context of drug repurposing, "affinity" refers to the degree to which a drug can bind to a target, usually a protein, in the body. This concept is central to understanding how drugs interact with biological systems and how they can be repurposed for new therapeutic uses. This measures how strongly a drug binds to its target. A high affinity means the drug binds tightly to the target, which is often desirable for efficacy. In drug repurposing, researchers look for drugs that have a high affinity for new targets, which might be implicated in different diseases than the drug was originally developed for.
Measuring Affinity: Techniques like measuring the equilibrium dissociation constant (Kd) are used to quantify affinity. In drug repurposing, comparing the Kd values of a drug for different targets can provide insights into potential new uses.

2.2. Data Labeling

In BindingDb a labeling operation 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

$$K_d = \frac{([L][R])}{[LR]} \tag{1}$$

The Kd value is instrumental in comprehending the affinity between proteins and drugs. However, due to its continuous nature at the nanometer scale, it cannot be directly employed as a label for classification purposes. To address this, another value must be computed, which involves converting the Kd value into a logarithmic scale. As suggested in previous studies [2] pKd value was calculated. Which is the result of the transformation the Kd value into log space as

$$pK_d = -\log_{10}\left(\frac{K_d}{1e9}\right) \tag{2}$$

The labeling process is carried out based on the pKd value, with 7 sets as the cutoff point. If the pKd value is 7 or higher, the input is labeled as 1, indicating a match between the protein and drug, while values below 7 are labeled as 0.

3. Feature Representations

3.1. Drug Representation

Since in the BindingDb dataset proteins and drugs are represented in different ways. For drugs 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.
When it is shown in SMILES notation it becomes:

$$CC(=O)OC1=CC=CC=C1C(=O)O \tag{3}$$

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 Carbon 'C' corresponds to 4 and Oxygen '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 and different values between 100 and 300 were tried in different experiments.

3.2. Protein Representation

Chemically, proteins are essentially long chains composed of 20 primary amino acids. Their chemical properties hinge on the sequence of these amino acids. Protein sequences can vary greatly in length, from a few tens to several thousands of amino acids, with the more typical lengths being in the hundreds. The extraordinary diversity in protein functions stems from the countless combinations of amino acid sequences. For example, there are roughly 10^{260} potential proteins that are 200 amino acids long, a number far exceeding the estimated 10^{80} atoms in the observable universe. This highlights that only a minuscule fraction of all possible proteins may ever exist or have existed on Earth.
In the databases referenced in this thesis, protein sequences were denoted with a single letter for each amino acid. Similar methods were used previously in other studies [3]. A similar method was used to convert this representation into a numerical format. Each amino acid was assigned a number, which was then normalized to a range between 0 and 1 using Min-Max scaling. Further research led to the use of a Word2Vec approach which includes the calculation of a three-dimensional vector for each amino acid. For instance, the vector for the amino acid alanine (abbreviated as 'A') was determined to be [0.6454009, 0.4708575, 0.37278453].
Calculating different statistics such as median and average informed the decision on the length of the protein vectors used in the models. Depending on the model, protein lengths varied between 100 and 300 amino acids. This decision was based on the statistics provided and the results of various experiments conducted.

4. Experiments

Throughout this study, different deep learning models were tried and compared with each other. The first method that was tried was simple Neural Networks. Since in this study starting point is text data and the sequence of the input is important, it has been thought that LSTMs can be useful. For this purpose simple LSTM models, LSTM models with attention mechanisms were tested. Along with these models in order to help the model to extract information from the features better models with pretraining are created. The architecture chosen for this was Autoencoders. After experiments with both standard Autoencoders and LSTM autoencoders, it was seen that LSTM Autoencoders were the best models for this drug repurposing task.
First LSTM Autoencoders were tested with only BindingDb and then other dataset including more proteins were added in the pretraining phase and the results were compared. In this LSTM Autoencoder model along with BindingDb for drug-target interaction, more data including different proteins were used for pretrain-

ing.
In the table 1 configuration of the best performing LSTM Autoencoder model can be seen.

Feature/Model	Autoencoder2	Autoencoder1	Combined Model
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 2: Configuration of the LSTM Autoencoder model

The results of these experiments will take place in the next section.

5. Results

The results that have been acquired from the experiments are given in Table 1.

Model	Accuracy	Precision	Recall	F1 Score
Feedforward Neural Network	88.94	**95.51**	87.02	91.06
Simple LSTM	84.30	92.45	84.16	88.18
LSTM and Attention Mechanism	90.3	87,15	94,28	90.57
Autoencoders BindingDb	88.27	95.39	87.98	91.53
Autoencoders HomosapiensDb	88.84	95.42	86.7	90.85
Autoencoders AllProDb	87.32	86	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	58.1			

Table 3: Results of each model

Overall, the best-performing models are LSTM Autoencoders. Autoencoder models follow them in this. Particularly those with pretraining and Homosapiens data. These models exhibit a good balance between accuracy, precision, recall, and F1 score, indicating robust and reliable performance.
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 (58,1%) shows it's not a good predictive model, as expected. This shows that other models that have been used do not just depend on the bias of the dataset. Feedforward Neural Network was the first deep learning method that has been used. Even this simple neural network model shows some promising values. The model has high precision (95,51%) and a good F1 score (91.06%), indicating effective identification of true positives. To analyze the other values from the results table: LSTM Models: These models (Long Short-Term Memory) vary in configuration and performance. LSTM with Attention Layer resulted in being the best among them, with the highest accuracy (90,3%) and a balanced F1 score (90.57%). These LSTM models with Attention mechanisms perform a little bit higher compared to the simple Neural Networks. Autoencoder Models: These models are used for learning efficient data codings in an unsupervised manner. The "Autoencoders Pretraining with BindingDb" shows the Autoencoder model with only BindingDb as the dataset. And "Autoencoder with Homosapiens data" shows the model with additional proteins for pretraining. The second one has high accuracy and F1 scores, indicating robust performance.
LSTM autoencoder variant 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 96.37. This suggests that this model is not only accurate overall but is 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 Table 3 additional metrics for the top-performing models were given. These were AUC and Matthews Correlation(MCC) values. The Matthews correlation coefficient is an indicator of the effectiveness of binary and multiclass classifications. It considers both true and false positives and negatives, making it a balanced metric suitable for classes of varying sizes. This MCC is a correlation coefficient ranging from -1 to +1. A value of +1 indicates a flawless prediction, 0 signifies a prediction no better than random, and -1 denotes a completely inverse prediction. This statistic can also be referred to as the phi coefficient.
The AUC metric is calculated using the area under the ROC curve. It is a single number giving the summary of how well the model discriminates between the two classes (positive and neg-

ative). The value of AUC ranges from 0 to 1. Closer to 1 meaning the model discriminates well between classes.
These AUC and MCC metrics show correlating results with previous metrics, which is an

Model	AUC	Matthews Correlation
Autoencoders with AllProDb	94.47	74.56
LSTM Autoencoders BindingDb	96.51	82.78
LSTM Autoencoders HomosapiensDb	96.5	83.32
LSTM Autoencoders AllProDb	96.37	83.42

Table 4: AUC and Matthews Correlation for Top Models

In summary, the LSTM Autoencoder models, particularly the "LSTM Autoencoders with AllproDb," demonstrate 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. As a result, it was seen that using more data on proteins for the purpose of pretraining helps to achieve better results. Especially when the Autoencoder architecture is combined with LSTMs.

6. Conclusions and Future Work

The transformative potential of machine learning (ML) in revolutionizing drug repurposing has been delved into here, particularly through the identification of promising candidates for repurposing. This study meticulously evaluates a range of ML models for drug repurposing and unveils LSTM Autoencoder models as the frontrunners, particularly when configured with many additional protein data for pretraining. These models showcase exceptional predictive accuracy and reliability, as corroborated by their high AUC values, robust MCC scores, and balanced performance across accuracy, precision, recall, and F1 scores. These compelling findings suggest that LSTM Autoencoder models could significantly accelerate the drug development process, paving the way for quicker and more cost-effective therapeutic solutions.
Future research might focus on further optimizing these models, expanding their applicability to a broader spectrum of datasets, and integrating them seamlessly into a holistic drug discovery framework. Here it was seen that using more data for the purpose of pretraining with Autoencoder models improves the performance, so this is something to consider in future studies. Adding more pretraining data for the molecules can be helpful. Additionally, enhancing model interpretability to gain a nuanced understanding of the rationale behind predictions could foster trust and provide valuable insights for researchers and clinicians. Furthermore, integrating these models into existing drug discovery pipelines could empower pharmaceutical companies with powerful tools to identify repurposing candidates more efficiently. Fostering collaborative efforts with experts in bioinformatics, pharmacology, and clinical sciences could pave the way for more holistic and interdisciplinary approaches to drug repurposing.
In conclusion, this research unequivocally demonstrates the immense potential of ML in revolutionizing drug repurposing through accurate and reliable drug classification. These promising results lay the foundation for more efficient, cost-effective, and innovative approaches to therapeutic discovery, underscoring the pivotal role of ML in shaping the future of pharmaceutical research and development.

References

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

[2] 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.

[3] Jackson Souza, Marcelo Fernandes, and Raquel De Melo Barbosa. A novel deep neural network technique for drug–target interaction. *Pharmaceutics*, 14:625, 03 2022.

[4] 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.