Exploring Novel Targets for *Mycobacterium tuberculosis* through Artificial Intelligence Methods

Rodrigo Esperança[1], Nuno Alves[1], Miguel Rocha[1]
[1]Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
pg50923@alunos.uminho.pt, id10075@alunos.uminho.pt, mrocha@di.uminho.pt

June 19, 2024

1 Introduction

1.1 Context and motivation

Tuberculosis (TB) is an infectious disease caused by the *Mycobacterium tuberculosis* (Mtb) bacillus [1]. Historically, this disease has been one of the deadliest for humans. Every year, there are still more than 10 million new cases of active tuberculosis worldwide and an estimated 1.3 million deaths [2]. Given this, it is essential to find new drugs that act quickly, with a broader efficacy and minimal side effects. Despite decades of research on immune responses that determine protection against tuberculosis, there is still no clear idea of the set of immune responses needed to prevent infection or the progression of the disease [3].

The rise of multidrug-resistant strains of Mtb has further compounded the problem of TB control, highlighting the need for new and effective treatments. For these reasons, computational methods, such as ML, can be potential alternative approaches to improve the accuracy and effectiveness of TB diagnosis and treatment, while simultaneously reducing the costs, time and resources required to manage the disease.

Machine learning (ML) is a sub-field of artificial intelligence. This subfield aims to find useful representations of some input data, within a predefined space of possibilities, using the orientation of a feedback signal [4]. ML enables the analysis of drug resistance patterns using Mtb genetic data, facilitating the selection of optimal antibiotics for treatment.

1.2 Objectives

This study attempts to identify novel proteins that could serve as potential targets for combating Mtb. The methodology starts with data compilation from different databases that contain information on drug-target interactions (DTI).

Then, leveraging state-of-the-art classifiers for DTI prediction, new interactions will be classified and subsequently clustered with the entire dataset. These clustered data will then undergo a selection of few points, which will be validated against knowledge from literature. The sequential steps of this process are delineated below:

1. Gather data from diverse sources encompassing known drug-target interactions;
2. Standardize and preprocess the collected data to ensure uniformity and compatibility across different datasets;
3. Employ state-of-the-art machine learning models, such as deep learning-based architectures or ensemble methods, for DTI prediction
4. Apply the trained models to predict potential drug-target interactions for compounds targeting Mtb;
5. Utilize clustering algorithms, such as k-means or hierarchical clustering, to group the predicted drug-target interactions based on similarity;
6. Validate the predicted interactions and selected target proteins by consulting existing literature, experimental databases, and relevant research studies

2 Background

2.1 Tuberculosis

TB is one of the leading infectious diseases in the world [5]. This disease is caused by a bacterium of the phylum Actinobacteria known as *Mycobacterium tuberculosis* [6]. The most common mechanism of transmission of Mtb is through airborne particles that are transmitted from individual to individual through coughing and sneezing [7]. The main symptoms are hunger, night sweats, fever, weight loss and extreme tiredness. The lungs are the main site affected by TB, but there are cases in which the disease can spread to other parts of the body, which is called extrapulmonary tuberculosis.

TB can be classified into two categories: latent infections, where common symptoms do not manifest themselves; and active disease, which occurs when the tubercle bacillus bypasses the immune system and multiplies [8].

2.2 Mechanism of infection

TB is transmitted by inhaling infectious droplets containing viable bacilli. After inhalation, the bacteria are phagocytosed by the alveolar macrophage, which rapidly activates the immune system and induces a response [9].

Macrophages, dendritic cells and other immune cells recognise mycobacterial structures, pathogen-associated molecular patterns (PAMPs) with membrane-associated pattern recognition receptors (PRRs), of which the most studied are Toll-like receptors (TLR2, TLR4, TLR9), when interacting with TLRs, signaling pathways are activated that lead to the production of predominantly pro-inflammatory cytokines, such as TNF, IL-1B, IL-12 and nitric oxide [10].

Ingestion of bacteria is then commonly destroyed through phagosome-lysosome fusion and acidification. However, Mtb can subvert this process and survive [11]. The innate immune response, led by macrophages, can result in three main scenarios: cell necrosis, apoptosis, or survival of the infected macrophages. When cell necrosis occurs, the mycobacteria are released and can infect new macrophages or spread. In apoptosis, on the other hand, the integrity of the cell membrane is not compromised, leading to the destruction of bacteria together with the macrophage.

The survival of the infected macrophages allows the mycobacteria to persist and even proliferate before the adaptive immune response is activated by specific T cells that have been selected in the regional lymph nodes.

2.3 Antibiotics and Resistance

TB can be cured with timely diagnosis and appropriate care. To treat drug-susceptible TB, several different antibiotics (isoniazid (INH), rifampin (RIF), ethambutol (EMB), and pyrazinamide (PZA)) are usually given in combination over six to nine months [12].

However, there are difficult-to-treat cases in which drug resistance to antibiotics is present. This happens when the bacterium that causes TB develops the ability to neutralise the effects of one or more of the drugs that are often used to treat the disease. Drug-resistant tuberculosis can occasionally be transmitted directly from one person to another.

The two main categories of drug-resistant TB are: Multidrug-resistant TB (MDR-TB), defined as Mtb resistant to at least the two main first-line drugs (isoniazid and rifampicin) and extensively drug-resistant TB (XDR-TB), defined as Mtb resistant to drugs from both lines. [13] .

Drug-resistant TB is a global health problem that is becoming more common. Furthermore, since the treatment and cure of DR-TB pose greater challenges due to the higher costs and increased difficulty compared to drug-susceptible TB, this problem is particularly pronounced in low- and middle-income countries where access to effective healthcare is limited [14].

2.4 Databases

Identifying targets remain key challenges to the development of safe and effective drugs. Databases containing information on the chemical structure of antibiotics, their mechanisms of action, molecular targets, and resistance profiles are valuable resources to help with this problem [15].

A database with potential binding applications should include various topics such as, analysis of ligands for a specific target to discover chemical characteristics that correlate with affinity, parameterization and validation of ligand detection methods, parameterization and validation of ligand detection methods, identifying candidate compounds for a new target by searching for ligands that bind to similar proteins, identifying drug candidates with a high risk of side effects, checking whether similar compounds bind to multiple receptors, among others [16].

Some of the databases that follow these criteria and will be used in this project are: Open Traget [15], Drugbank [17], TTD [18] and BindingDB [16].

2.5 Machine Learning

ML is directly related to statistics, but unlike statistics, ML has the capacity to handle huge and complex datasets for which statistical analysis (e.g. Bayesian analysis) would be impossible [4].

ML algorithms aim to find meaningful transformations taking into account the objective of the task. That is, we can define ML as the process of finding useful representations of some input data, within a predefined space of possibilities, using the orientation of a feedback signal. This simple idea makes it possible to solve a remarkably wide range of intellectual tasks [4].

There are two main types of ML, being unsupervised learning (Clustering, Dimensionality reduction, among others) and supervised learning (k-Nearest neighbors, Naive Bayes, Decision trees Kernel methods, among others). In the case of unsupervised learning, their respective models have the function of identifying the unknown patterns in the input data without any preexisting knowledge of their output. In the case of supervised learning, their models have the ability to "learn" and predict the expected values of similar data based on certain algorithms used for model training [19].

There are two main categories of supervised learning: classification, where the output values are categorical, and regression, where the output values are numeric and non-binary. In the classification category, we have the decision trees, SVM, neural networks methods as examples. In the regression category we have the Linear Regression, SVM Regression, Neural Network Regression methods as examples [19].

Decision trees are an essential building block for many ML algorithms. The idea behind decision trees is very intuitive and best represented in a visual form. The Support vector machine (SVM) for classification and support vector regression (SVR) for continuous outputs have found applications in computational biology for their ability to be robust against noise and to work with high-dimensional datasets found in genetics, transcriptomics, and proteomics. Neural networks constitute a collection of neurons and edges, where different weights can be applied to each edge connecting the neurons. At each neuron, an activation function is applied to the weighted input signal to generate an output signal. The number of hidden layers define whether the system is a shallow learning system (with one or a few hidden layer) or DL (with many hidden

layers) [19].

2.6 Deep Learning

Deep learning (DL) exhibits superior computational capacity and enhanced flexibility compared to traditional ML methodologies. This is primarily attributable to the intricate architectures of deep learning models, characterized by multilayered neural networks, which endow them with the capability to discern intricate patterns and relationships within vast datasets. Moreover, typical deep learning architectures boast millions of adaptable parameters, enabling them to encapsulate a broader spectrum of features and nuances present in complex real-world data, thus facilitating more nuanced and sophisticated learning representations. [20].

Like most neural network architectures, DL architectures are composed of layers (input, hidden and output), neurons and activation functions. The neurons act as feature detectors and are organised into lower and upper layers. The lower layers detect basic features and transmit them to the upper layers, which identify more complex features [21].

DL consists of several architectures, the most conventional of which are as follows: Deep neural network (DNN) trained to model complex non-linear relationships, extracting unique abstract features that help improve their performance [21], Convolutional neural network (CNN) used mainly for image processing applications [22], Recurrent neural network (RNN) better suited to dealing with sequential data. They are great for processing time-dependent information [23].

2.7 Machine learning and Deep learning applied to Drug-target interactions

The application of ML and DL techniques in the field of tuberculosis covers a variety of tasks, including predicting early diagnosis, identifying patterns in imaging scans, drug discovery, predicting antibiotic resistance, and personalising treatment regimens [24].

Among the many parts of the drug discovery process, the prediction of drug-target interactions (DTI) is an essential part. DTI is difficult and costly, as experimental trials are not only time-consuming but also expensive. *In silico* DTI predictions (performed on a computer) are therefore in high demand, as they can speed up the drug development process by systematically suggesting a new set of candidate molecules promptly, which can save time and reduce the cost of the whole process [25].

In response to this demand, three types of *in silico* DTI prediction methods have been proposed in the literature: molecular docking, similarity-based and machine learning/deep learning-based [25].

DTI-related tools represent significant advances in the prediction of drug-protein interactions. As a rule, these tools have in common the application of CNNs in their architectures, and some also use other models such as DNN

or BERT, with the aim of predicting the affinity between molecules (drugs) and target proteins. Using raw sequence data, such as SMILES and FASTA sequences, these tools eliminate the need for feature engineering and stand out for their ability to learn representations directly from molecular and protein data. We can see some of the tools in Table 1 that have the characteristics mentioned above.

Table 1: Examples of Machine Learning and Deep Learning applied to drug-target interaction.

Tool Name	**Algorithms**	**Database**	**Input**	**Output**
MT-DTI [25]	BERT, CNN, DNN	PubChem [26], Kiba [27], Davis [28]	SMILES (molecule), FASTA (protein)	Affinity Score (Regression)
DLM-DTI [29]	DNN	Davis[28], Bindingdb [30][17], IUPHAR	Protein sequence, SMILES	Drug-Target Interaction (Binary)
Deep DTA [31]	CNN, DNN	Davis [28], KIBA [27]	Protein sequences, SMILES	Drug-Target binding affinities (Regression)
MATTDTI [32]	CNN, FNN	Davis [28], KIBA [27]	SMILES (drugs) and FASTA (proteins)	Affinity Score of drug-target pairs

3 Methodology

The main objective of this work is to try to find new drugs for the treatment of tuberculosis through deep learning models, and to do this it is necessary to use data related to drug target interaction, so for this work we had two different approaches: Approach 1- Use of more general DTI-related datasets to train deep learning binary classification models to predict interactions between drugs and targets with the aim of later using a dataset with drugs related to the treatment of tuberculosis. After using the model to predict, see if there are new drugs with similar characteristics to tuberculosis through clustering. Approach 2 - similar to approach 1, but instead of using generic datasets, use a dataset with lots of data including drugs and targets related to Mtb and instead of using binary classification in the models, used regression. The general workflow is schematized in figure 1.

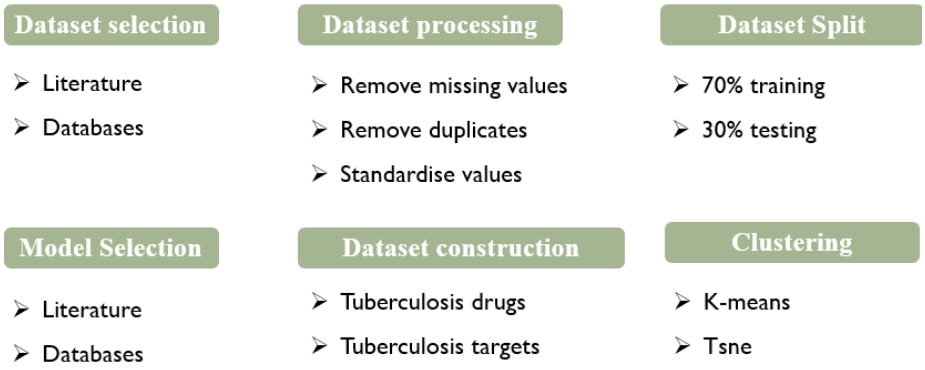

Figure 1: A schematic representation of the workflow used in the two approaches of this work.

3.1 Data loading and preparation

3.1.1 dataset selection:

For the first approach, we chose two of the most widely used DTI datasets in the literature to train DL models: the KIBA dataset [27] and the Davis dataset [28].

The KIBA dataset [27] covers a wide range of information, providing details on both inhibitors (or ligands) and target proteins (in this case, kinases). Each entry in the dataset consists of the protein sequence, the structure of the ligand (usually represented as SMILES) and, crucially, the associated KIBA score.

The Davis [28] dataset is a structured compilation of protein and ligand pairs, accompanied by their respective inhibition constant (Ki) and dissociation constant (Kd) values. Each entry in this dataset offers a view of the target protein, the ligand in question and the associated affinity measure.

In the second approach, we chose to use Bindingdb [30] to extract a bigger dataset with data related to Mtb in order to train a model. The extracted dataset contains all the Bindingdb data.

Each entry in the dataset consists of the protein sequence, the ligand structure and the associated Kd, Ki and IC50 values. For this approach, we also used the Metz [33] dataset to evaluate the model. This dataset consists of protein kinases and their respective inhibitors with the associated Ki values.

3.1.2 Pre-processing:

This stage was characterised by the "cleaning" of the data set, i.e. removing rows containing "NaN" values, incorrect characters (ex. "¡", "¿") in the sequences, duplicates, non-relevant columns (only with zero and repeated values), and applying a normalisation process to the data results after this filtering. For the first approach, it was necessary to convert the Kd and KIBA values into 0 and 1. To do this, a threshold was set where above this value the data was

transformed into 1 and below it into 0. For approach 2, we only selected the data that contained Ki values because the dataset that was going to be used to evaluate the model only contained proteins and their respective inhibitors with associated Ki values. It was also necessary to apply the negative logarithm in base 10 to these values in order to scale the data.

3.2 Model Selection

Two deep learning models linked to the prediction of drug-target interaction were selected from the literature. To train these models, we decided to only use 70 per cent of the datasets, reserving 15 percent of the data for validation and 15 percent for test sets. The models ran for 50 epochs, with a batch size of 32 and we used model checkpoint to save the model with the best validation loss and early stopping to stop training if the validation loss does not improve after 10 epochs. These two methods were imported from the Keras package [34].

3.2.1 DL models:

In the first approach, two DL models were used, DeepDTA [31] and DLM-DTI [29], to evaluate their performance on both datasets. In the second approach only the DeepDTA model was used [31].

3.2.2 Hyperparameters:

In the first approach, we opted to use binary classification.The labels used to train the model contained the Kd values in the case of the Davis dataset, and the KIBA values in the case of the KIBA dataset. The DeepDTA model consists of three convolutional layers with different kernel sizes and progressively increasing numbers of filters, a GlobalMaxPooling1D layer and a dense output layer with 3 fully connected layers (1024 ,1024, 512) and sigmoid activation and dropout, to predict the probability of interaction between the protein and the small molecule. The model is compiled using the Adam optimiser and the cross-entropy binary loss function.

The DLM-DTI model is similar using three sequential blocks, each composed of a fully connected layer (2048, 1024, 512), GeLU activation and dropout. The model is compiled using the AdamW optimiser

In the second approach, we opted to use regression. The labels in this case, contained the Ki values. With regard to the hyperparameters, the activation was removed and the loss function was changed to Mean Squared Error.

3.3 Testing trained model on TB drugs

In this step, we used the literature and databases to collect the most commonly used drugs in the treatment of tuberculosis, such as isoniazid, rifampin and ethambutol, and their respective targets, including dihydrofolate reductase, the beta subunit of DNA-directed RNA polymerase and cytochrome P450 2C19. We

then added this data to the Davis [28] and KIBA [27] sets for the first approach, and to the Metz [33] dataset for the second. We processed the data again, as before, and used the model with the best results to make the predictions.

3.4 Clustering

In this final phase, we used tools such as k-means and t-SNE to create clusters with the model's predictions for the dataset built earlier to identify drugs or targets similar to those used to treat tuberculosis.

4 Results and discussion

4.1 Models' Performance

The two different approaches used throughout this work had as their main objectives the exploration of targets directly associated with drugs used in the treatment of tuberculosis.

In Approach 1, after the datasets were processed, Davis dataset [28] had 229 proteins, 2111 compounds and 119,000 interactions. the KIBA [27] dataset had 442 proteins, 68 compounds, 25,000 interactions.

The table below 2, shows the performance results of the DeepDTA and DLM-DTI models on different data sets, measured by the ROC-AUC and PR-AUC metrics. The ROC-AUC metric assesses the model's ability to distinguish between classes. Higher values indicate better model performance in terms of sensitivity and specificity, while PR-AUC metric is especially useful for unbalanced data sets. It assesses the relationship between precision and recall. Higher values indicate better performance in identifying positive examples.

On the KIBA dataset, the DeepDTA model achieved an ROC-AUC of 0.91 and a PR-AUC of 0.79, while the DLM-DTI model achieved an ROC-AUC of 0.91 and a PR-AUC of 0.77. These results show that both models have a reasonable ability to distinguish between classes in the KIBA dataset, with DeepDTA showing a slight advantage in the PR-AUC metric.

On the Davis dataset, the DeepDTA model achieved an ROC-AUC of 0.90 and a PR-AUC of 0.48, while the DLM-DTI model achieved an ROC-AUC of 0.90 and a PR-AUC of 0.42. Although both models maintain a good performance on the ROC-AUC metric, the PR-AUC values are significantly lower compared to the results on the KIBA dataset, indicating greater challenges in identifying positive examples on this dataset.

We can conclude, that the combination between model and dataset that shows better performance is the DeepDTA [31] model with the KIBA [27] dataset.

The next step was to analyse the confusion matrix present in figure 2 to understand the balance of the dataset using the DeepDTA model. We can then conclude that we have a very good performance regarding the negative class,

Table 2: Results of deep learning models in approach 1.

Dataset	Model	ROC-AUC	PR-AUC
KIBA [27]	DeepDTA [31]	0.91	0.79
Davis [28]	DeepDTA [31]	0.90	0.48
KIBA [27]	DLM-DTI [29]	0.91	0.77
Davis[28]	DLM-DTI [29]	0.90	0.42

only 515 false positives, and regarding the positive class we have a reasonable performance with 1393 false negatives.

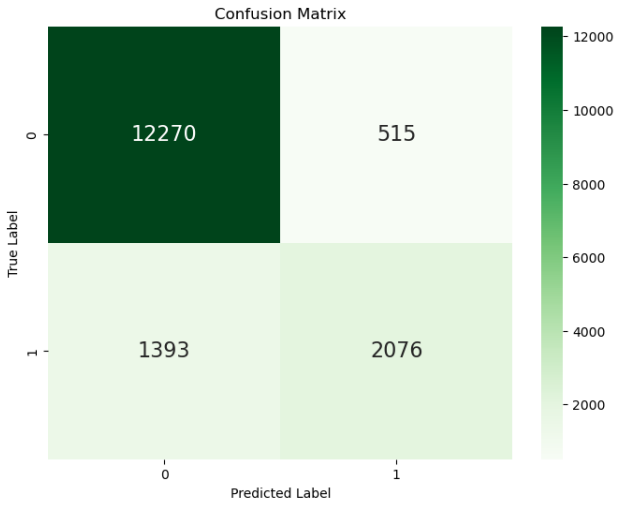

Figure 2: Confusion Matrix of DeepDTA with KIBA dataset, where true negatives (TN) are in the top left cell, false positives (FP) in the top right cell, false negatives (FN) in the bottom left cell and true positives (TP) in the bottom right cell..

After evaluating the desired dataset with the trained model, we concluded that the model was not working as expected because it could not predict the interaction between the drugs and the targets that we knew existed for TB. This may have been due to the dataset not being balanced, or the model not being trained with data related to TB targets. Because of these results, we decided to carry out a second approach in which we used a much larger dataset containing *Mtb* data to refine the model.

For Approach 2, the data set used, after being processed, had 119000 compounds, 2350 proteins, and 513820 interactions.

3 shows the results for the trained model. The errors (MSE, MAE, RMSE) are relatively low, indicating that the model is making reasonably accurate predictions. The R^2 shows that the model explains a good deal of the variance

in the data, but there is still room for improvement.

In summary, the model has a decent performance at predicting the actual values, but there may be opportunities for adjustments and improvements to further increase the accuracy of the predictions.

Table 3: Results of deep learning model in approach 2.

Dataset	**Model**	**MSE**	**MAE**	**RMSE**	**R^2**
Bindingdb [30]	DeepDTA [31]	0.77	0.66	0.88	0.67

In Figure 3, we can see that the scatter plot shows a generally linear relationship between actual and predicted values. We can see that there are points very close to the red line at the extremes, suggesting that the model performs better in this range.

Overall, the graph shows that the model has reasonably good predictive performance, with an R^2 of 0.67 and an acceptable amount of points around the perfect prediction line, although there are some deviations and outliers.

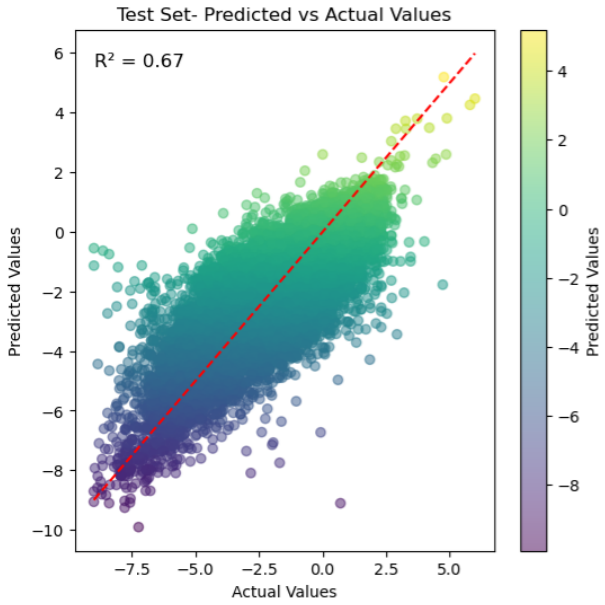

Figure 3: Scatter plot of the test set showing the relationship between the predicted values and the actual values. The dashed red line indicates the line of identity (predicted values equal actual values), and the colouring of the dots represents the density of the data.

After evaluating the desired dataset with the trained model, we concluded that this model was already working better than the other one because it could

already predict the interactions of TB drugs and targets properly, as we can see in table 4.

Table 4: Model predictions on approach 2 for the interactions between antibiotics used in TB and their respective targets.

Drugs	**Targets**	**-log10(Ki)**
Isoniazid	Dihydrofolate reductase (P9WNX1)	-2.63026
Rifampin	DNA-directed RNA polymerase subunit beta (P9WGY9)	-3.085087
Pyrazinamide	Probable fatty acid synthase Fas (P95029)	-5.143974
Ethambutol	Cytochrome P450 2C19 (P33261)	-5.316166
Kanamycin	30S ribosomal protein S12 (P0A7S3)	-1.946355

Once the results of the model had been obtained, we proceeded to clustering to try to find potential new candidate drugs for TB. There are 5 different types of clusters and most of the drugs used in the treatment of tuberculosis are found in cluster 2. Although the model does not predict any other drugs with a direct interaction with tuberculosis targets, through figure 4 one can see that there are drugs with similar characteristics such as Kinome3582 [OC(=O)c1csc2C(=O)NCCNc12] and Kinome1171[OCC(Cc1ccc(Cl)cc1)NC(=O) c2ccc(cc2)c3ccncc3].

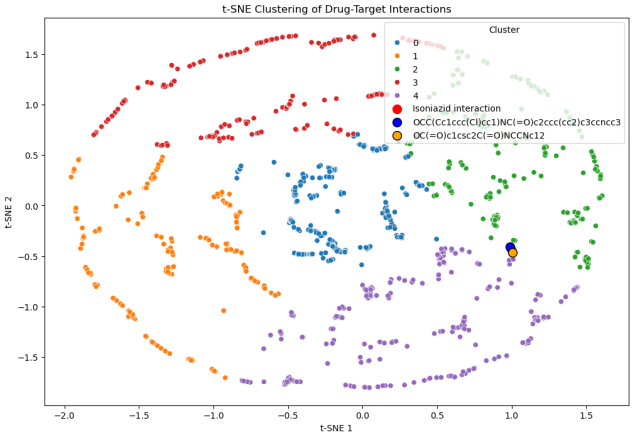

Figure 4: Clustering of drug-target interactions using t-distributed Stochastic Neighbor Embedding (t-SNE). Each point represents a drug, and its position is determined by the similarity of its target interactions to those of other drugs.

5 Conclusion and future work

With this work we can conclude that the DeepDTA model [31] obtains slightly better results than the DLM-DTI model [29]. However, it was necessary to use a more specific Mtb dataset for the model to be able to predict interactions correctly, since when using an unbalanced and generic dataset such as KIBA [27] the model was unable to predict interactions between TB drugs and their respective targets.

This work provides valuable information that can serve as a basis for future research. However, the next step would be to explore drugs with similar characteristics obtained from clustering in the literature.

In addition, it would be interesting to optimise the model's architecture. Another interesting approach for future work might be to use the dataset extracted from Bindingdb [30] to train the DeepDTA model [31] with binary classification and then compare the results obtained between the three datasets.

6 Code availability

The pipelines implemented in this work were developed in Python, and are available at `https://github.com/esperancaa/Projeto_PG50923`

References

[1] M. M. Ibrahim, T. M. Isyaka, U. M. Askira, J. B. Umar, M. A. Isa, *et al.*, "Trends in the incidence of rifampicin resistant mycobacterium tuberculosis infection in northeastern nigeria," *Scientific African*, vol. 17, 9 2022.

[2] *Global tuberculosis report 2023*, 2023.

[3] J. A. L. Flynn and J. Chan, "Immune cell interactions in tuberculosis," pp. 4682–4702, 12 2022.

[4] F. Chollet, "Deep learning with python."

[5] G. Sotgiu, P. Glaziou, C. Sismanidis, and M. Raviglione, *Tuberculosis Epidemiology*. Elsevier Inc., 10 2016, pp. 229–240.

[6] I. Comas, M. Coscolla, T. Luo, S. Borrell, K. E. Holt *et al.*, "Out-of-africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans," *Nature Genetics*, vol. 45, pp. 1176–1182, 10 2013.

[7] M. Barbier and T. Wirth, "The evolutionary history, demography, and spread of the mycobacterium tuberculosis complex," *Microbiology Spectrum*, vol. 4, 8 2016.

[8] A. A. J. Aljanaby, Q. M. H. Al-Faham, I. A. J. Aljanaby, and T. H. Hasan, "Epidemiological study of mycobacterium tuberculosis in baghdad governorate, iraq," *Gene Reports*, vol. 26, 3 2022.

[9] C. C. Weekes, "Mycobacterium tuberculosis infections."

[10] R. V. Crevel, T. H. Ottenhoff, and J. W. V. der Meer, "Innate immunity to mycobacterium tuberculosis," pp. 294–309, 2002.

[11] B. Sáenz, R. Hernandez-Pando, G. Fragoso, O. Bottasso, and G. Cárdenas, "The dual face of central nervous system tuberculosis: A new janus bifrons?" pp. 130–135, 3 2013.

[12] A. R. Rees, *A New History of Vaccines for Infectious Diseases: Immunization - Chance and Necessity*, 2022.

[13] M. Fujiwara, M. Kawasaki, N. Hariguchi, Y. Liu, and M. Matsumoto, "Mechanisms of resistance to delamanid, a drug for mycobacterium tuberculosis," *Tuberculosis*, vol. 108, pp. 186–194, 1 2018.

[14] A. Kashyap and P. K. N. Sharma, *Drug Resistant TB Demystified: Doctor's Secret Guide*, 2023.

[15] D. Ochoa, A. Hercules, M. Carmona, D. Suveges, J. Baker *et al.*, "The next-generation Open Targets Platform: reimagined, redesigned, rebuilt," *Nucleic Acids Research*, vol. 51, no. D1, pp. D1353–D1359, 11 2022. [Online]. Available: https://doi.org/10.1093/nar/gkac1046

[16] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities," *Nucleic Acids Research*, vol. 35, no. $\text{suppl}_1, pp.D198 - -D201, 122006$.

[17] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler *et al.*, "DrugBank 6.0: the DrugBank Knowledgebase for 2024," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1265–D1275, 11 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad976

[18] Y. Zhou, Y. Zhang, D. Zhao, X. Yu, X. Shen *et al.*, "TTD: Therapeutic Target Database describing target druggability information," *Nucleic Acids Research*, vol. 52, no. D1, pp. D1465–D1477, 09 2023. [Online]. Available: https://doi.org/10.1093/nar/gkad751

[19] S. Badillo, B. Banfai, F. Birzele, I. I. Davydov, L. Hutchinson *et al.*, "An introduction to machine learning," *Clinical Pharmacology and Therapeutics*, vol. 107, pp. 871–885, 4 2020.

[20] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature Genetics*, vol. 51, pp. 12–18, 1 2019.

[21] T. D. Akinosho, L. O. Oyedele, M. Bilal, A. O. Ajayi, M. D. Delgado *et al.*, "Deep learning in the construction industry: A review of present status and future innovations," *Journal of Building Engineering*, vol. 32, p. 101827, 11 2020.

[22] S. Wang and T. Huang, *Applications of Deep Learning in Biomedicine*. Elsevier, 2021, pp. 29–39.

[23] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model." ISCA, 9 2010, pp. 1045–1048.

[24] W. Deelder, S. Christakoudi, J. Phelan, E. D. Benavente, S. Campino, R. McNerney, L. Palla, and T. G. Clark, "Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data," *Frontiers in Genetics*, vol. 10, 9 2019.

[25] B. Shin, S. Park, K. Kang, and J. C. Ho, "Self-attention based molecule representation for predicting drug-target interaction," pp. 1–18, 2019. [Online]. Available: https://mt-dti.deargendev.me/

[26] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "Pubchem 2023 update," *Nucleic Acids Research*, vol. 51, pp. D1373–D1380, 1 2023.

[27] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen *et al.*, "Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis," *Journal of Chemical Information and Modeling*, vol. 54, pp. 735–743, 3 2014.

[28] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar, "Comprehensive analysis of kinase inhibitor selectivity," *Nature Biotechnology*, vol. 29, pp. 1046–1051, 11 2011.

[29] J. Lee, D. W. Jun, I. Song, and Y. Kim, "Dlm-dti: a dual language model for the prediction of drug-target interaction with hint-based learning," *Journal of Cheminformatics*, vol. 16, p. 14, 2 2024.

[30] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "Bindingdb: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Research*, vol. 35, pp. D198–D201, 1 2007.

[31] H. Öztürk, A. Özgür, and E. Ozkirimli, "Deepdta: deep drug–target binding affinity prediction," *Bioinformatics*, vol. 34, pp. i821–i829, 9 2018.

[32] Y. Zeng, X. Chen, Y. Luo, X. Li, and D. Peng, "Deep drug-target binding affinity prediction with multiple attention blocks," *Briefings in Bioinformatics*, vol. 22, 9 2021.

[33] J. T. Metz, E. F. Johnson, N. B. Soni, P. J. Merta, L. Kifle, and P. J. Hajduk, "Navigating the kinome," *Nature Chemical Biology*, vol. 7, pp. 200–202, 4 2011.

[34] F. Chollet *et al.*, "Keras," https://keras.io, 2015.