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

Rodrigo Esperança[1], Nuno Alves[1], and Miguel Rocha[1]

Universidade do Minho, Campus de Gualtar, 4710-057 Braga, Portugal
rodrigoce9@gmail.com, id10075@alunos.uminho.pt,
mrocha@di.uminho.pt

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**	**Author**
MT-DTI [25]	BERT, CNN, DNN	PubChem [26], Kiba [27], Davis [28]	SMILES (molecule), FASTA (protein)	Affinity Score (Regression)	Bonggun Shin
Deep ConV-DTI [29]	CNN, DNN	DrugBank [17], IUPHAR [30], KEGG [31]	Raw protein sequence, Morgan/Circula Digital Printing	Drug-Target Interaction (Binary)	Lee
Deep DTA [32]	CNN, DNN	Davis [28], KIBA [27]	Protein sequences, SMILES	Drug-Target binding affinities (Regression)	Hakime
MATTDTI [33]	CNN, FNN	Davis [28], KIBA [27]	SMILES (drugs) and FASTA (proteins)	Affinity Score of drug-target pairs	Zeng
MDeedPred [34]	(CNN), Regressor	Davis [28], PDBBind [35]	Aminoacid (proteins), SMILES (compounds)	Predicted affinity score the input compound-target (XC50)	Rifaioglu

Bibliography

[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] I. Lee, J. Keum, and H. Nam, "Deepconv-dti: Prediction of drug-target interactions via deep learning with convolution on protein sequences," *PLoS Computational Biology*, vol. 15, 6 2019.

[30] S. D. Harding, J. F. Armstrong, E. Faccenda, C. Southan, S. H. Alexander, A. P. Davenport, M. Spedding, and J. A. Davies, "The iuphar/bps guide to pharmacology in 2024," *Nucleic Acids Research*, vol. 52, pp. D1438–D1449, 1 2024.

[31] M. Kanehisa, "Kegg: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, pp. 27–30, 1 2000.

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

[33] 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.

[34] A. S. Rifaioglu, R. C. Atalay, D. C. Kahraman, T. Doğan, M. Martin, and V. Atalay, "Mdeepred: novel multi-channel protein featurization for deep learning-based binding affinity prediction in drug discovery," *Bioinformatics*, vol. 37, pp. 693–704, 5 2021.

[35] M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li, and R. Wang, "Comparative assessment of scoring functions: The casf-2016 update," *Journal of Chemical Information and Modeling*, vol. 59, pp. 895–913, 2 2019.