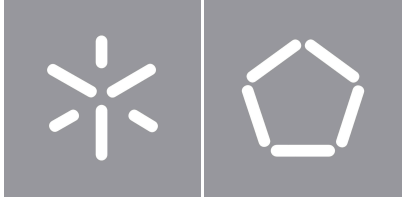
**University of Minho**
School of Engineering

Rodrigo Cordeiro Esperança

# Exploring Novel Targets for Mycobacterium tuberculosis through Machine Learning Methods

october 2025



**University of Minho**
School of Engineering

Rodrigo Cordeiro Esperança

# Exploring Novel Targets for Mycobacterium tuberculosis through Machine Learning Methods

Master's Dissertation

Master Degree in Bioinformatics
Dissertation supervised by
**Miguel Francisco Almeida Pereira Rocha**

october 2025

# Copyright and Terms of Use for Third Party Work



## License granted to users of this work:

# Acknowledgements

I would like to express my sincere gratitude to everyone who, in one way or another, contributed to the completion of this Master's journey.

First and foremost, I would like to thank my supervisor, Professor Miguel Rocha, for his guidance, availability, and valuable advice, which were essential throughout the development of this work.

A special thank you also goes to Nuno Alves for his constant support, patience, and dedication during this entire process, which made every challenge easier to overcome.

To my Master's colleagues, I extend my heartfelt thanks for the good moments shared, the teamwork, and the friendship that made this experience so much more enjoyable and enriching.

Finally, and most importantly, I would like to thank my family, my girlfriend and my girlfriend's family for their incredible support, understanding, and encouragement throughout the Master's degree. Without you, this achievement would not have been possible.

To all of you, my deepest thank you.

# Statement of Integrity

I hereby declare having conducted this academic work with integrity.

I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

University of Minho, Braga, october 2025

Assinado por: **Rodrigo Cordeiro Esperança**
Num. de Identificação: 15654375
Data: 2025.10.30 10:27:54+00'00'


Rodrigo Cordeiro Esperança

# Abstract

Tuberculosis, caused by *Mycobacterium tuberculosis*, remains one of the world's deadliest infectious diseases, a challenge further intensified by the rise of multidrug-resistant strains. Conventional drug discovery pipelines are often slow and costly, underscoring the urgent need for computational strategies capable of accelerating target identification and compound prioritisation.

This dissertation presents an integrated computational framework that leverages Machine Learning and Deep Learning techniques to predict drug–target interactions and identify novel therapeutic candidates against *Mycobacterium tuberculosis*. To enable accurate and disease relevant predictions, a tuberculosis specific bioactivity dataset was constructed by integrating multiple public databases and subsequently subjected to thorough manual curation to ensure biological relevance and data quality. This tailored dataset served as the foundation for all predictive modelling experiments.

A self-supervised learning approach based on the Barlow Twins architecture was employed to generate molecular and protein embeddings, which were then used as input features for supervised classifiers. In parallel, the BCM-DTI deep learning model provided a biologically constrained multimodal benchmark for comparative evaluation. To refine predictive outcomes, an automated post-processing pipeline was implemented, incorporating cheminformatics filters. Model interpretability was further enhanced through SHAP analysis, which enabled the identification of key molecular fragments driving activity predictions. The framework successfully identified several compounds with strong predicted affinities toward validated tuberculosis targets. These compounds were subsequently analysed through literature review to assess their therapeutic potential value.

Overall, this work establishes a comprehensive and reproducible computational pipeline that integrates semi-supervised learning, interpretability, and drug repurposing strategies to accelerate tuberculosis drug discovery.

**Keywords** Tuberculosis, Drug–Target Interaction, Machine Learning, Deep Learning, Drug Repurposing.

# Resumo

A tuberculose, causada por *Mycobacterium tuberculosis*, continua a ser uma das doenças infeciosas mais mortais do mundo, um desafio ainda mais agravado pelo surgimento de estirpes multirresistentes. As vias convencionais de descoberta de fármacos são frequentemente lentas e dispendiosas, o que sublinha a necessidade urgente de estratégias computacionais.

Esta dissertação apresenta uma estrutura computacional integrada que recorre a técnicas de Aprendizagem Automática e Aprendizagem Profunda para prever interações fármaco-alvo e identificar novos candidatos terapêuticos contra *Mycobacterium tuberculosis*. Para permitir previsões precisas e relevantes para a doença, foi construído um conjunto de dados de bioatividade específico para a tuberculose, através da integração de múltiplas bases de dados públicas, seguido de uma curadoria manual rigorosa para garantir a relevância biológica e a qualidade dos dados. Este conjunto de dados personalizado serviu de base a todas as experiências de modelação preditiva. Foi adotada uma abordagem de aprendizagem auto-supervisionada baseada na arquitetura Barlow Twins para gerar embeddings moleculares e proteicos, que foram posteriormente utilizados como variáveis de entrada em classificadores supervisionados. Em paralelo, o modelo profundo BCM-DTI foi utilizado como referência multimodal, para fins de comparação. Para refinar os resultados preditivos, foi implementado um processo automatizado de pós-processamento que integra filtros quimioinformáticos. A interpretabilidade dos modelos foi ainda aprimorada através da análise SHAP, que permitiu identificar fragmentos moleculares determinantes nas previsões de atividade. A estrutura desenvolvida identificou vários compostos com fortes afinidades preditas para alvos de tuberculose. Estes compostos foram posteriormente analisados através de revisão da literatura, de modo a avaliar o seu potencial terapêutico.

No seu conjunto, este trabalho estabelece um pipeline computacional abrangente e reprodutível que integra aprendizagem semi-supervisionada, interpretabilidade e estratégias de reposicionamento de fármacos para acelerar a descoberta de medicamentos contra a tuberculose.

**Palavras-chave** Tuberculose, Interações Fármaco–Alvo, Aprendizado de Máquina, Aprendizagem Profunda, Reaproveitamento de Fármacos.

# Contents

# List of Figures

# List of Tables

# Acronyms

**ACC** Accuracy.

**AUC** Area Under the Curve.

**BCM** Branch Chain Mining.

**BCM-DTI** Branch Chain Mining Drug–Target Interaction.

**CFM** Category Fragment Mapping.

**CNN** Convolutional Neural Network.

**CNNs** Convolutional Neural Networks.

**DL** Deep Learning.

**DNNs** Deep Neural Networks.

**DTI** Drug-target interactions.

**ECFP** Extended-Connectivity Fingerprints.

**ELM** Extreme Learning Machines.

**ELU** Exponential Linear Unit.

**ESM2** Evolutionary Scale Modeling 2.

**FN** False Negatives.

**FP** False Positives.

**FPR** False Positive Rate.

**GBM** Gradient boosting machine.

**GNNs** Graph Neural Networks.

**INH** Isoniazid.

**KNN** K-Nearest Neighbors.

**LSTM** Long Short-Term Memory.

**MAE** Mean Absolute Error.

**MCC** Matthews Correlation Coefficient.

**MDR-TB** Multidrug-resistant Tuberculosis.

**MHCII** Major Histocompatibility Complex class II.

**ML** Machine Learning.

**MSE** Mean Squared Error.

**MTB** *Mycobacterium tuberculosis*.

**NAD** Nicotinamide Adenine Dinucleotide.

**NN** Neural Network.

**PAMPs** Pathogen-associated molecular patterns.

**PCA** Principal Component Analysis.

**PR** Precision.

**PRRs** Pattern recognition receptors.

**QED** Quantitative estimate of drug-likeness.

**RECAP** Retrosynthetic Combinatorial Analysis Procedure.

**RF** Random forests.

**RIF** Rifampicin.

**RL** Reinforcement learning.

**RMSE** Root Mean Squared Error.

**RNNs** Recurrent Neural Networks.

**ROC** Receiver Operating Characteristic.

**SA** Synthetic accessibility.

**SHAP** SHapley Additive exPlanations.

**SL** Supervised learning.

**SSL** Semi-supervised learning.

**SVM** Support Vector Machines.

**t-SNE** t-Distributed Stochastic Neighbor Embedding.

**TB** Tuberculosis.

**TLR** Toll-like receptors.

**TN** True Negatives.

**TP** True Positives.

**TPR** True Positive Rate.

**TPSA** Topological polar surface area.

**UL** Unsupervised learning.

**UMAP** Uniform Manifold Approximation and Projection.

**XDR-TB** Extensively drug-resistant Tuberculosis.

# Chapter 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, it is the 13th leading cause of death, and since 2024, it has returned to being the leading cause of infectious death after being three years behind the disease caused by the coronavirus (COVID-19). Currently, there are still more than 10.6 million new cases of active **TB** worldwide and an estimated 1.25 million deaths [2]. The primary target of **MTB** are the lungs, which can be transmitted through the air, making it an easily transmitted disease. Therefore, finding new drugs that act quickly, with broader efficacy and minimal side effects is essential. Despite decades of research on immune responses that determine protection against **TB**, there is still no clear idea of the set of immune responses needed to prevent infection or disease progression [3].

Initially, **Rifampicin (RIF)** and **Isoniazid (INH)** were considered to be the most effective **TB** drugs. The increasing incidence of multidrug-resistant strains of **MTB** has further compounded the problem of **TB** control. In 2022, approximately 73% of people were confirmed to have developed resistance to **RIF** and **INH** after undergoing bacteriological tests [2], highlighting the need for new and effective treatments. For these reasons, computational methods, such as **Machine Learning (ML)**, can be potential alternative approaches to improve the accuracy and effectiveness of predicting drug-target interactions for **TB**, while simultaneously reducing the costs, time, and resources required for drug discovery and target identification [4].

**ML** is a subfield of Artificial intelligence that aims to find meaningful patterns in input data, within a predefined space of possibilities, driven by feedback mechanisms that iteratively refine and enhance the performance of the models [5]. In the context of **TB**, **ML** enables the analysis of complex interactions between drugs and **MTB** targets, which could facilitate the identification of new therapeutic targets and optimization of drug selection for treatment, especially given the challenge of drug resistance [6].

Among the many parts of the drug discovery process, the prediction of **Drug-target interactions (DTI)** is an essential part. **DTI** prediction is difficult and costly, as experimental trials are not only time-consuming but also expensive. *In silico* **DTI** predictions are therefore in high demand, as they can speed up the drug development process by systematically and promptly suggesting a new set of candidate molecules, which can save time and reduce the cost of the whole process [7].

In response to this demand, three types of *in silico* prediction methods have been proposed in the literature: molecular docking, similarity-based, and machine learning-based [7]. Methods for predicting interactions between drugs and proteins represent significant advancements in this field. Many approaches share the use of **ML** models capable of identifying patterns and predicting the affinity between molecules and target proteins. By leveraging raw sequence data, these methods reduce reliance on manual feature engineering and excel in learning meaningful representations directly from molecular and protein information.

## 1.2 Objectives

The objective of this thesis is to explore **ML** and **Deep Learning (DL)** models to predict **DTI**, using data from multiple sources. These models will be applied to the study of existing and identification of new therapeutic targets and drugs for **MTB**. The project aims to improve the accuracy of predictive models and potentially discover new or repurposed drugs for **TB**. To achieve this, several key objectives have been defined:

1. Review the state of the art on the relevant topics, including literature, methods, and computational tools;

2. Build and curate a large scale dataset of **DTI** by combining public repositories with a specific focus on **MTB** targets;

3. Exploring **ML** and deep **DL** for **DTI** prediction using the previous datasets;

4. Enhance models by augmenting available datasets and using advanced feature engineering and optimization techniques;

5. Perform drug repurposing experiments using alternative compounds predicted to interact with high priority **MTB** targets;

6. Apply an automated post processing and filtering pipeline combining physicochemical, structural, and drug likeness filters to identify viable compound candidates;

7. Conduct an explainability analysis using **SHapley Additive exPlanations (SHAP)** to identify the molecular features most influential in model predictions.

8. Validate predicted compounds using literature;

9. Propose an automated pipeline for novel drug and therapeutic target discovery.

## 1.3 Thesis Structure

The thesis is divided into five main chapters.

**Chapter 1 – Introduction** This chapter introduces the motivation, context, and objectives of the dissertation. It highlights the global impact of **TB**, the growing threat of multidrug-resistant strains, and the urgent need for computational approaches in drug discovery. The section outlines how **ML** and **DL** can accelerate the identification of novel therapeutic targets for **MTB**, concluding with the main objectives and structure of the work.

**Chapter 2 – State of the Art** This chapter provides an extensive review of the relevant literature. It begins with an overview of **TB**, its mechanisms of infection, and current antibiotics and resistance mechanisms. It then explores the fundamental principles of **ML**, including supervised, unsupervised, and semi-supervised learning, as well as molecular and protein representations. Finally, it discusses the role of **DL** architectures and their applications to **DTI** prediction, focusing on frameworks like Barlow Twins and BCM-DTI.

**Chapter 3 – Methodology** This chapter details the methodological framework followed in the study. It describes the Python tools and packages employed, the collection and preprocessing of data from public databases, and the construction of a **TB** focused dataset. It also explains the model selection process, the adaptation of the Barlow Twins and **Branch Chain Mining Drug–Target Interaction (BCM-DTI)** models, drug repurposing experiments, post processing pipelines, and the use of **SHAP** to interpret model predictions.

**Chapter 4 – Results and Discussion** In this chapter, the results obtained from the models and datasets are presented and discussed. It includes model performance metrics, comparison between different configurations, visualization of molecular and protein embeddings, and identification of potential drug candidates through filtering and explainability analyses. The chapter also reports the validation of

selected compounds through literature review and discusses their biological plausibility.

**Chapter 5 – Conclusions and Future Work** The final chapter summarizes the key findings and contributions of the dissertation, emphasizing the integration of semi-supervised learning, interpretability, and drug repurposing in **TB** research. It also proposes directions for future work, including improvements in dataset curation, expansion of molecular representations, and the incorporation of additional deep learning frameworks to enhance predictive accuracy and biological relevance.

# Chapter 2

# State of the art

## 2.1 Tuberculosis

**TB** is one of the leading infectious diseases in the world [8]. An estimated total of 10.6 million people worldwide fell ill with this deaseas in 2022, equivalent to 133 new cases per 100,000 individuals. These estimated increases in **TB** incidence in 2021 and 2022 are the consequence of disruptions in diagnosis and treatment during the COVID-19 pandemic when the reported number of people newly diagnosed with **TB** fell from 7.1 million in 2019 to 5.8 million in 2020 and 6.4 million in 2021. In 2022, an estimated total of 1,302,000 people died.[2].

This bacterial disease is caused by a bacterium of the phylum Actinobacteria known as **MTB**, the species of mycobacterium most frequently isolated in humans [9]. In most cases, exposure to a contagious individual leads to a subclinical infection localized in the lungs, where complete microbial eradication through innate and adaptive immunity is considered unlikely. **TB** is an airborne infectious disease, the most common mechanism of transmission of **MTB** is through small aerosolized particles that are transmitted from individual to individual through coughing, talking and sneezing [10].

In pulmonary **TB**, the main symptoms are cough, which occurs in 90% of the patients, hemoptysis, chest pain, hunger, night sweats, fever, weight loss, and extreme fatigue. The lungs are the main site affected, but there are cases in which the disease can spread to other parts of the body, which is called extrapulmonary **TB**. In this case, the lymph nodes become intertwined and develop abscesses. Additional complications include meningitis and miliary disease, the latter arising when bacteria spread through the bloodstream [11].

**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 [12].

## 2.2 Mechanism of infection

**TB** is transmitted by inhaling infectious droplets containing viable bacilli. After inhalation, the innate immune system is activated mobilizing various immune cells, including macrophages, dendritic cells, monocytes, and neutrophils to phagocytose and initiate an immune response aimed at eliminating the invading bacteria. If a macrophage infected with the bacillus fails to destroy it after phagocytosis, the intracellular replication of the pathogen may result in the host cell undergoing necrosis or apoptosis [13]. When cell necrosis occurs, the mycobacteria are released and can infect new macrophages or spread. In apoptosis, however, the integrity of the cell membrane is preserved, leading to the destruction of bacteria together with the macrophage [14].

The persistent influx of innate immune cells to the site of bacterial replication results in the formation of an early granuloma which is composed of Macrophages and other immune cells that recognize mycobacterial structures, specifically **Pathogen-associated molecular patterns (PAMPs)**, through membrane-associated **Pattern recognition receptors (PRRs)**, of which the most studied are **Toll-like receptors (TLR)** (**TLR**2, **TLR**4, and **TLR**9). When interacting with **TLR**s, signaling pathways are activated, leading to the production of predominantly pro-inflammatory cytokines [15].

Approximately 3 to 8 weeks after the initial infection, dendritic cells present **MTB** antigens to immature CD4+ T-cells via **Major Histocompatibility Complex class II (MHCII)** molecules. This process, coupled with IL-12 signaling and co-stimulatory interactions between B7 and CD28, drives the maturation of these T cells into Th1 lymphocytes [16].

In some individuals, the adaptive immune response and granuloma maturation are sufficient to sterilize the infection, leaving behind only a calcified and healed lesion. However, in most cases, the bacterial population persists, maintaining long term survival [17].

## 2.3 Antibiotics and Resistance

The most commonly used drugs for the treatment of **TB** are **RIF** and **INH**, but there are many other options. Typically, **TB** drugs are subdivided into first-line drugs: **INH**, **RIF**, rifabutin, pyrazinamide and ethambutol; and second-line drugs: cycloserine, ethionamide, streptomycin, kanamycin, levofloxacin, and moxifloxacin [18], as we show in Table 1. This division is based on effectiveness, cost and toxicity levels.

Normally, patients with non-resistant active **TB** are administered with a combination of some first-line antibiotics, over six to nine months [19]. To treat people with latent **TB**, the drug of choice is **INH**. Therapy

is typically continued for 6-9 months. Alternative regimens include a 4-month course of **RIF** [20].

However, several factors influence the choice of antibiotics such as the dynamic behaviour of **MTB** and the heterogeneity of its subpopulations. Bacteria can be in different states of activity, from active replication to dormant or persistent states. Dormant bacteria, with reduced metabolism, are less susceptible to antibiotics that affect actively growing cells. This, allied to the fact that latent **TB** has no symptoms and therefore people do not even know they have, are some of the reasons behind the difficulty in completely eradicating the infection since **MTB** populations in a persistent state can survive conventional treatments [21].

The choice of antibiotics involves considering these pharmacological characteristics and the conditions of the microenvironment where **MTB** is found. Pyrazinamide, for example, depends on the acidification of the endosomal compartment to be effective, but **MTB**'s ability to manipulate this environment can reduce its effectiveness. Bedaquiline, a lipophilic antibiotic, accumulates in lipid droplets in macrophages, acting as a reservoir to combat intracellular **MTB** populations during the phagocytosis process.

In addition, the activation of immune cells can induce tolerance to antibiotics. This is because the immune response increases the ability of macrophages to fight the infection, but it can also promote antibiotic resistance, which highlights the complexity of the dynamics between the pathogen and the host cell [22].

Table 1: First-line and second-line antibiotics for treatment of **TB**, as well as the type of treatment used, mode of action, the drug targets, and their mechanism of action. Data were retrieved from the DrugBank database [23].

| | **Antibiotics** | **Treatment** | **Mode of Action** | **Drug Targets** | **Mechanism of Action** |
|---|---|---|---|---|---|
| 1st line | Isoniazid | Active and latent **TB** | Bactericidal and bacteriostatic | Catalase-peroxidase | Inhibits the synthesis of mycobacterial cell wall lipids and nucleic acids |
| | Rifampin | Active and latent **TB** | Bactericidal | RNA Polymerase | Inhibits bacterial RNA polymerase, the enzyme responsible for DNA transcription |
| | Pyrazinamide | Active **TB** | Bactericidal | Fatty acid synthetase | Disrupts membrane energetics and inhibits membrane transport functions |
| | Ethambutol | Active and latent **TB** | Bacteriostatic | Arabinosyltransferase | Inhibits the transfer of arabinose into arabinogalactan, thereby disrupting cell wall integrity |
| 2nd line | Bedaquiline | Active **TB** | Bactericidal | ATP synthase subunit c | Inhibits mycobacterial ATP (adenosine 5'-triphosphate) synthase |
| | Cycloserine | Active **TB** | Bactericidal and bacteriostatic | D-alanine ligase A | Interferes with an early step in bacterial cell wall synthesis in the cytoplasm |
| | Kanamycin | Active **TB** | Bactericidal | 30S ribosomal protein S12 | Interferes with decoding site in the vicinity of nucleotide 1400 in 16S rRNA of 30S subunit |
| | Amikacin | Active **TB** | Bactericidal | 30S ribosomal protein S12 | Interferes with mRNA binding and tRNA acceptor sites |
| | Ciprofloxacin | Active **TB** | Bactericidal | DNA topoisomerase 4 | Prevents it from supercoiling the bacterial DNA which prevents DNA replication |

Antibiotic resistance is a huge problem that affects approximately 750,000 patients every year [2]. 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, i.e., spontaneous mutations appear that hinder the interaction between the drug and its target, impair the activation of prodrugs, or result in target overexpression. However, the resistance phenotypes of a significant portion of **MTB** clinical isolates cannot be fully explained by these mutations [24]. Drug-resistant **TB** can occasionally be transmitted directly from one person to another.

The two main categories of drug-resistant **TB** are: **Multidrug-resistant Tuberculosis (MDR-TB)**, defined as **MTB** resistant to at least the two main first-line drugs (e.g. **INH** and **RIF**) and **Extensively drug-resistant Tuberculosis (XDR-TB)**, defined as **MTB** resistant to drugs from both lines. [25] .

On average, around 450,000 new cases per year of **RIF** resistant **TB** are reported. Approximately 78% of patients with this resistant have **MDR-TB**, which considerably reduces the likelihood of successful treatment [26].

## 2.4 Machine Learning

Identifying targets remains a key challenge to the development of safe and effective drugs. **ML** can play a pivotal role in addressing this challenge by analysing vast and complex datasets more efficiently than traditional methods. For example, **ML** algorithms can identify patterns and correlations within data related to the chemical structure of antibiotics, mechanisms of action, and molecular targets. Furthermore, this models can predict potential resistance mechanisms by leveraging resistance profiles, enabling researchers to prioritize drug candidates that are less likely to encounter resistance. These capabilities make **ML** an invaluable resource for streamlining the drug discovery process. [27].

**ML** algorithms aim to find meaningful transformations taking into account the objective of the task. Therefore, **ML** can be defined 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 [28].

There are four main types of **ML**, **Supervised learning (SL)** (e.g. k-Nearest Neighbors, Naive Bayes, Decision Trees), **Semi-supervised learning (SSL)**,**Unsupervised learning (UL)** (e.g. Clustering, Dimensionality reduction), and **Reinforcement learning (RL)** [29].

Normally, a general workflow of an **ML** pipeline for drug discovery, starts with data gathering from relevant databases and proceeding through data preparation steps to ensure clean and usable input for

**ML** models. The data is then split into training, validation, and test sets, which are essential for building and evaluating predictive models. The training and validation steps involve fine-tuning hyperparameters to optimize model performance, while the test set is reserved for final model evaluation. The ultimate goal is to create a robust predictive model capable of identifying promising drug candidates efficiently, we can see this in the figure 1.

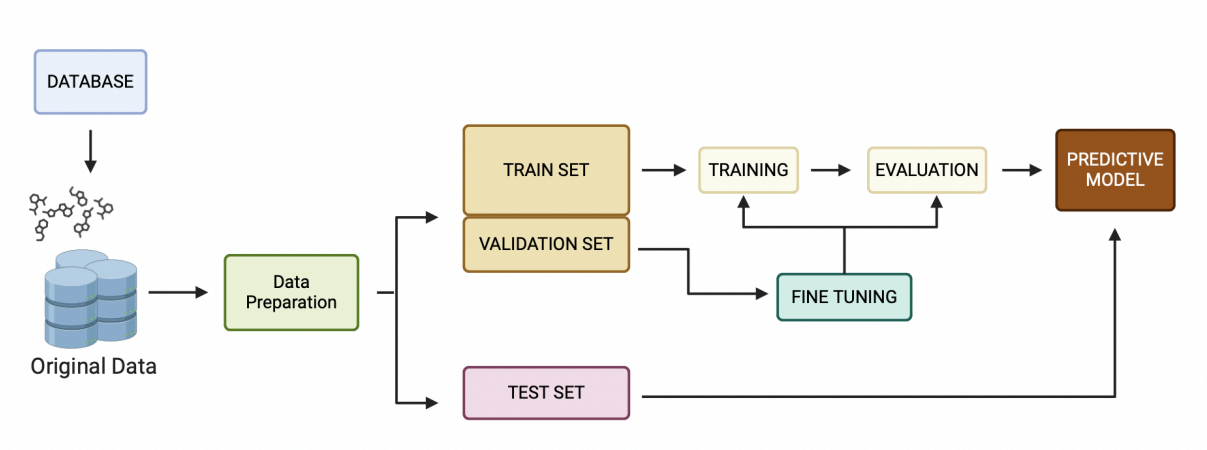



Figure 1: Schematic representation of common **ML** pipeline. The data is prepared before being split into training, validation, and test sets. The training set is used to train the model, while the validation set supports model evaluation and fine-tuning. Finally, the test set assesses the performance of the optimized predictive model. Adapted from [30] and [31]

## 2.5 Supervised Learning

**SL** algorithms are designed to deal with problems where the training data includes labelled samples. They are widely used and recognized due to their diverse applications [32]. After training, the model can classify or predict new cases. This type of learning falls into two main categories: classification and regression.

Classification consists of predicting specific categories or labels, such as predicting whether a certain chemical compound will be “effective” or “ineffective”. As mentioned earlier, the algorithm is trained with labelled data and learns to categorize new examples into one of the predefined classes. Regression, meanwhile, is used to predict continuous values, such as predicting the ideal dosage of a chemical compound to achieve therapeutic efficacy. The model learns to generate a function capable of estimating the numerical values of new cases.

This approach allows models to learn patterns from existing data and apply them to new situations efficiently and accurately [33].

### 2.5.1 Metrics

Performance metrics play a fundamental role in evaluating supervised learning models, enabling the measurement of their effectiveness and guiding improvements. These metrics vary according to the task at hand and are commonly classified into two main groups: classification metrics and regression metrics [34].

#### Classification metrics

Classification metrics provide a means to evaluate the performance of a classification algorithm. These metrics help to understand how well a model is distinguishing between different classes. Common classification metrics, specifically for binary cases, include **Accuracy (ACC)**, **Precision (PR)**, Sensitivity, Specificity, and F1-Score, ranging from 0 to 1 (Equations 2.1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.1a}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.1b}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{2.1c}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.1d}$$

$$\text{F1 Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.1e}$$

Where:

- **True Positives (TP)** represent the number of correctly predicted positive instances.
- **False Negatives (FN)** represent the number of positive instances incorrectly classified as negative.
- **True Negatives (TN)** represent the number of correctly predicted negative instances.
- **False Positives (FP)** represent the number of negative instances incorrectly classified as positive.
- **True Positive Rate (TPR)** measures the proportion of actual positives the model correctly identifies.

- **False Positive Rate (FPR)** measures the proportion of actual negatives incorrectly identified as positives by the model. It highlights the rate at which false alarms occur.

**Area Under the Curve (AUC)** is a performance metric used to evaluate the ability of a binary classification model to differentiate between positive and negative classes. Specifically, it is associated with the **Receiver Operating Characteristic (ROC)** curve. An **AUC** of 0.5 indicates no discriminatory power, equivalent to random guessing. An **AUC** of 1 signifies perfect discriminatory power. An **AUC** below 0.5 suggests the model is worse than random guessing, consistently misclassifying classes.

Performance metrics can also be applied to multi class classification problems. For these cases, metrics such as **PR**, recall, and F1-Score can be adapted using strategies like one vs one or one vs rest to evaluate performance across multiple classes, ensuring the model's ability to distinguish between more than two categories is adequately assessed [35].

### Regression Metrics

Regression metrics provide a way to evaluate the performance of a regression model, which predicts a continuous outcome. In this section, we introduce some of the common metrics used for regression analysis.

**Mean Absolute Error (MAE)** is the average of the absolute differences between predicted and actual values (Equation 2.2). It provides a simple measure of prediction accuracy where $N$ represents the total number of examples, $yi$ refers to the actual or observed value and $\hat{y}i$ corresponds the predicted value by the model.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i| \tag{2.2}$$

**Mean Squared Error (MSE)** is the average of the squares of the differences between predicted and actual values (Equation 2.3). Squaring the errors gives more weight to larger errors, making **MSE** sensitive to outliers.

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{2.3}$$

**Root Mean Squared Error (RMSE)** is the square root of **MSE**, translating the error into the same units as the original values (Equation 2.4). It also emphasizes larger errors more than **MAE**.

$$RMSE = \sqrt{MSE} \tag{2.4}$$

R-squared is the metric that quantifies the proportion of variance in the dependent variable that can be explained or predicted by the independent variables in the model (Equation 2.5). It indicates how well the model captures the variability of the target variable based on the given predictors.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \tag{2.5}$$

Adjusted R-squared, adjusts for the number of predictors in the model, providing a more accurate measure in the presence of multiple predictors (Equation 2.6).

$$\text{Adjusted } R^2 = 1 - \left(1 - R^2\right) \times \frac{N-1}{N-k-1} \tag{2.6}$$

Correlation is a statistical measure that quantifies the degree and direction of the relationship between two variables (Equation 2.7). It is commonly used in regression analysis to assess the strength and nature of their linear association.

$$Corr_R = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \tag{2.7}$$

where $x_i$ are individual values of the independent variable $x$, $y_i$ are the individual values of the dependent variable $y$, $\bar{x}$ is the mean of $x$ values, and $\bar{y}$ is the mean of $y$ values. This equation calculates the Pearson correlation coefficient $r$, which measures the linear relationship between two variables $x$ and $y$. It quantifies the strength and direction of their linear association, ranging from -1 (perfect negative correlation) to +1 (perfect positive correlation).

### 2.5.2 Algorithms

Several **SL** algorithms are available in the literature. This subsection reviews some of them.

#### Linear Regression

Linear regression is a linear method used to model and predict the relationship between variables. Its simplicity and effectiveness in identifying data patterns make it widely used across many fields. The coefficients of the linear equation are estimated by minimizing the sum of squared errors between predicted and actual values in the training set, typically using the ordinary least squares method. This technique forms the foundation of statistical **ML** models [36].

### Logistic Regression

Logistic Regression is a statistical linear approach used to predict binary outcomes. Unlike linear regression, logistic regression models the probability of an event occurring by transforming the linear combination of the independent variables through a sigmoid (logistic) function. The model parameters are adjusted iteratively, taking into account the difference between the predicted values and the actual class labels. The process also involves calculating the gradient of the cost function, which measures the classification error in the input data, followed by updating the parameters in the direction that minimizes this error. [33]

### SVM

**Support Vector Machines (SVM)**s are a collection of **SL** techniques used for classification and regression, grounded in statistical learning theory and convex optimization [37]. The goal of **SVM**s is to find the hyperplane that maximizes the margin between classes. This approach of maximizing the margin enables that algorithm to perform well in generalization and effectively handle data that may not be linearly separable. The optimization task can be found in equation (Equation 2.8) [33]:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i(w^T x_i + b) \geq 1, \forall i = 1, \ldots, n \tag{2.8}$$

where $w$ represents the weight vector, $b$ is the bias term, and $\|w\|$ is the Euclidean norm of $w$. Meanwhile, $w$ and $b$ can be computed using:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i \quad \text{and} \quad b = y_k - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_k \tag{2.9}$$

where $k$ is any support vector with $\alpha_k > 0$.

### Decision trees

Decision Trees are a type of predictive model used in **ML** for classification and regression tasks. They operate by representing decisions and their possible consequences in a hierarchical, tree-like structure. Each internal node in the tree corresponds to a test or decision on an attribute (e.g., "Is the age greater than 30?"), and each branch represents the outcome of that decision. The final output, found at the leaf nodes, represents the predicted value or class label. This intuitive approach allows Decision Trees to model complex decision making processes in a way that is both easily interpretable and highly flexible.

Decision Trees are constructed using a top-down, recursive process, where the dataset is partitioned based on specific criteria to create the tree-like structure. The initial node is referred to as the root node, while terminal nodes that do not split further are called leaf nodes. Often described as a divide and conquer

strategy, this method uncovers relationships within the data to make accurate predictions for new, unseen examples [38].

### RF

**Random forests (RF)** are a powerful ensemble learning algorithm for classification and regression tasks [39]. Ensemble learning refers to a machine learning approach that combines multiple models, often referred to as 'weak learners,' to produce a stronger, more robust model. By aggregating the predictions of these individual models, ensemble methods aim to improve accuracy and reduce overfitting [40].

In the case of **RF**, the ensemble consists of multiple decision trees, each trained on different subsets of data and features. The predictions of these trees are then aggregated, using the mode for classification tasks or the mean for regression, to enhance accuracy and robustness. During training, each tree in the forest is built from a random sample of the data, and at each split, a random subset of features is considered. This randomness helps to reduce overfitting and improves the generalization ability of the model.

Given an instance $x$ with $F$ classes, the final ensemble prediction from $N$ trees can be calculated using equation (Equation 2.10) [33]

$$H(N(x)) = \arg\max_{j} \sum_{k=1}^{N} \mathbb{1}(h_k(x) = j), \quad \text{for} \quad j = 1, \ldots, C \tag{2.10}$$

### KNN

**K-Nearest Neighbors (KNN)** is a foundational algorithm in **ML** and pattern recognition, valued for its straightforward approach and effectiveness in both classification and regression tasks. The algorithm functions by locating the $k$ nearest data points to a specified query point and making predictions based on the outputs of these neighbors. However, **KNN** has some limitations, including its sensitivity to the selection of $k$, inefficiency when applied to large datasets, and difficulty in managing outliers [41].

## 2.6 Unsupervised Learning

This technique is designed to uncover the input data's latent patterns or inherent structures. In contrast to **SL**, **UL** operates without the need for labelled datasets, offering a versatile and automated approach suitable for a wide range of applications [42].

A variety of **UL** algorithms are commonly applied in numerous fields. Methods like k-means and

hierarchical clustering are frequently used to detect hidden groupings in data. Another widely adopted approach is **Principal Component Analysis (PCA)**, which reduces data dimensionality while preserving its key characteristics.

### Principal Component Analysis

**PCA** is a widely recognized statistical technique known for its ability to simplify complex datasets. This simplification is achieved by transforming the data into a set of orthogonal variables called principal components. The main advantage of this method is its ability to reduce the dimensionality of the data while preserving its most relevant patterns and trends. As a multivariate technique, **PCA** examines data tables where observations are described by several intercorrelated quantitative variables, facilitating the identification and visualization of similarity patterns among observations and variables [43].

One of its most notable applications is dimensionality reduction, which transforms the data into a lower dimensional space while retaining its key feature's variance. The mathematical foundation of this technique includes the eigenvalue decomposition of positive semi definite matrices and the singular value decomposition of rectangular matrices. These principles enable the construction of principal components that maximize data variance while minimizing the distance between the original points and their projections in the new dimensional space. This approach makes **PCA** a powerful tool for analysing and interpreting high dimensional data [44].

### K-means clustering

K-means clustering is a widely used technique for partitioning data into $K$ distinct groups based on similarity. The algorithm assigns each data point to the cluster with the closest mean, effectively minimizing the variance within each cluster. It is especially popular for its simplicity and efficiency in handling large datasets. However, the traditional algorithm has some limitations, such as sensitivity to initial conditions and the need to specify the number of clusters in advance. Recent research has focused on addressing these challenges and improving the algorithm's efficiency and effectiveness [45].

Several extensions of the k-means algorithm have been developed to overcome its limitations. For instance, the U-k-means algorithm eliminates the need for initializations and parameter selection by automatically determining the optimal number of clusters. The global k-means algorithm, on the other hand, adds cluster centres incrementally through a deterministic global search, reducing the impact of initial conditions [46].

### t-SNE

**t-Distributed Stochastic Neighbor Embedding (t-SNE)** is a widely used technique to visualize high dimensional data by mapping it into a lower dimensional space, typically two or three dimensions. Its main advantage lies in its ability to reveal complex local structures within the data, making it an essential tool in areas such as single cell transcriptomics and large scale data visualization [47].

This technique was developed to create clear visual representations of complex data, minimizing the tendency to crowd points in the centre of the plot. This feature is particularly important for datasets that lie on low dimensional manifolds, helping to identify patterns across multiple scales.

### UMAP

**Uniform Manifold Approximation and Projection (UMAP)** is a nonlinear dimensionality reduction technique that aims to preserve both local and global structures of high dimensional data when mapping it into a lower dimensional space. It is grounded in manifold theory and topological data analysis, making it particularly effective in uncovering meaningful structures in complex datasets [48].

**UMAP** constructs a weighted graph representation of the data based on nearest neighbors and optimizes a low dimensional embedding that maintains the topological relationships of the original space. Compared to **t-SNE**, **UMAP** generally offers faster computation, better scalability to large datasets, and improved preservation of the global data structure.

## 2.7 Semi-supervised Machine Learning

Semi-supervised **ML** is a hybrid approach that combines labelled and unlabelled data to enhance learning tasks. This methodology bridges the gap between **SL**, which relies solely on labelled data, and **UL**, which uses only unlabelled data. This approach is particularly advantageous when acquiring labelled data is costly or time-consuming, while unlabelled data is readily available. The effectiveness of this approach is based on several key assumptions, such as smoothness, cluster or low density separation, manifold structure, and transduction. These assumptions guide the development of algorithms capable of effectively utilizing both labelled and unlabelled data to improve learning outcomes [49].

Among the techniques used in **SSL**, **SVM**s stand out. Semi-supervised **SVM**s extend the traditional **SVM** model to handle unlabelled data, incorporating models like transductive **SVM**s and Laplacian **SVM**s, to improve learning performance by incorporating the structure of the unlabelled data. Another important example is **Extreme Learning Machines (ELM)**, which have been adapted for semi-supervised tasks.

The application of manifold regularization to **ELM**s enhances their efficiency and applicability, allowing for optimized processing of both labelled and unlabelled data [50].

Additionally, advancements in semi-supervised **DL** have introduced methods, such as deep generative models, consistency regularization, and pseudo labelling. These methods are designed to improve model performance by effectively integrating unlabelled data into the training process, contributing significantly to better outcomes in learning tasks.

## 2.8 Representing Molecules

In drug discovery, it is common to use data consisting of molecules, including both small and large molecules of drugs. To characterize these structures, a comprehensive analysis of the main molecular representations is presented. Data representations are organized into two main categories: drug molecules and protein molecules, as illustrated in Figure 2, which provides a schematic overview of the different types of molecular representations.

Since the way data is represented directly influences the knowledge acquired by the learning model, selecting an appropriate representation is essential to optimize the predictive model's performance [34].

### 2.8.1 SMILES

SMILES (Simplified Molecular Input Line Entry System) is a widely adopted method for describing molecular structures in a textual format. This technique uses a linear, chain-based representation to encode the structure of a molecule in a concise and versatile manner [51].

An advantage feature is the ability to represent the same molecule in multiple ways, which can be leveraged for data augmentation in **ML** applications. This strategy has proven effective in improving predictive accuracy and reducing errors in tasks such as molecular property prediction.

### 2.8.2 Fingerprints

When dealing with feature vectors, each element of a molecular fingerprint corresponds to a specific molecular feature or descriptor. These elements can be used as input vectors to train models. Additionally, in traditional similarity-based methods, molecular fingerprints can be applied to calculate similarity measures between drug molecules [34].

The binary or sparse representation of molecular fingerprints simplifies the computation of molecular similarity. Among the most widely used molecular fingerprints are those based on topological structure,

physicochemical properties, and pharmacophores [52].

A particularly popular class of molecular fingerprints is **Extended-Connectivity Fingerprints (ECFP)**, which are circular fingerprints derived from the Morgan algorithm. **ECFP** encodes the local atomic environment of each atom within a molecule by iteratively expanding neighborhoods around atoms up to a predefined radius. This process captures detailed structural information and effectively represents molecular substructures as unique identifiers or bit strings [53].

### 2.8.3 Embeddings

In this approach, the structures and features of drug molecules are represented as points in a low dimensional vector space, referred to as embeddings. These embeddings are generated through the training of **DL** models, which effectively capture critical information about the structure and properties of molecules. This method exhibits high expressive power, enabling the extraction of complex features and the identification of nonlinear relationships, thereby providing more precise information on molecular properties. [54].

Recent studies highlight the growing use of **DL** to represent molecules, with autoencoders being one of the most promising techniques in this field [55].

### 2.8.4 Molecular graph

Molecular graphs are an essential representation in computational chemistry, where molecules are depicted as graphs with atoms represented as nodes and chemical bonds as edges. This form of characterization effectively preserves the structural features of molecules and provides detailed information crucial for more in-depth analysis and predictions. This approach is crucial for tasks such as molecular property prediction and the generation of new molecules [49].

### 2.8.5 Proteins Representation

Proteins play a central role in biological systems and are essential for understanding **DTI**. In computational approaches, protein molecules are typically represented in formats that capture their structural, sequential, or functional properties. These representations enables **ML** and deep **DL** to process and extract meaningful patterns for downstream tasks, such as predicting binding affinities or classifying biological functions.

One of the most common representations is aminoacid sequences, where proteins are expressed as

linear chains of amino acids using single letter codes. These sequences can be directly processed by **DL** models to capture sequential relationships and motifs that are critical for protein function prediction. Models like **Evolutionary Scale Modeling 2 (ESM2)** [56] demonstrate how large scale language models can directly infer three dimensional protein structures from primary sequences, achieving high resolution predictions while leveraging evolutionary patterns encoded in the sequences.

Furthermore, protein embeddings have gained popularity as a representation technique, leveraging pre-trained models to encode proteins into fixed length vectors. These embeddings encapsulate biophysical and functional properties derived from large scale datasets, enabling models to perform tasks like subcellular localization prediction or the classification of protein-protein interactions. These methods eliminate the need for multiple sequence alignments, simplifying the structure prediction pipeline while maintaining high accuracy [57].

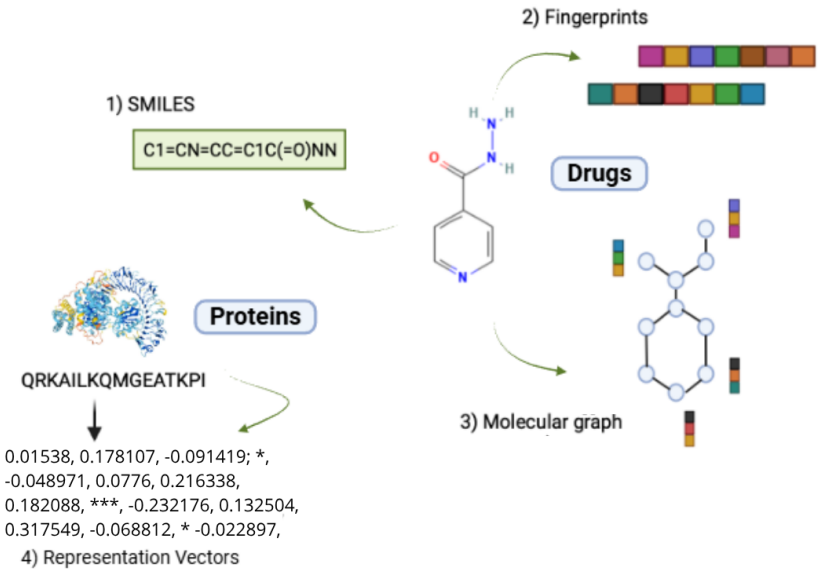



Figure 2: Drug and protein representation in computational analysis. (1) SMILES notation encodes molecular structures, serving as input for representation generation. (2) Fingerprints and (3) molecular graphs are constructed to capture structural and chemical properties of drugs. (4) Proteins are represented through sequence data and further converted into numerical vectors, enabling integration with drug data for predictive modeling. Adapted from [34]

## 2.9 Deep Learning

**DL** exhibits superior computational capacity and enhanced flexibility compared to traditional **ML** methodologies, which often rely on manual feature engineering and are less effective with high dimensional data. This is primarily attributable to the intricate architectures of **DL** models, characterized by multi-layered **Neural Network (NN)**s, which endow them with the capability to discern intricate patterns and relationships within vast datasets. Moreover, typical **DL** 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 refined and advanced learning representations [58].

### 2.9.1 Deep Neural Networks

**Deep Neural Networks (DNNs)**, inspired by the structure and functionality of the human brain, have demonstrated remarkable capabilities in function approximation. They achieve exponential precision for various functional classes, such as polynomials and sinusoidal functions. Recognized as Kolmogorov optimizers [59], **DNNs** are capable of approximating complex functions with fewer resources compared to alternative methods. Their expressive power is attributed to their depth and the nonlinearity of their activation functions, which allow them to effectively approximate functions, even those with low Besov smoothness [60].

Each **DNNs** instance is characterized by unique connectivity patterns and representational profiles due to differing initial conditions. These variations lead to distinct network representations, even when classification performance remains consistent [61].

Despite their impressive performance, this model are computationally intensive, requiring substantial resources. Enhancing energy efficiency and computational throughput without sacrificing accuracy is critical for their practical implementation in artificial intelligence systems. These advancements are essential to ensure their broader applicability and scalability across diverse domains.

**DL** models encompass a variety of architectures as we can see in Figure 3, each tailored to address specific types of problems and datasets. Among the most prominent are **Convolutional Neural Networks (CNNs)**, optimized for processing visual data [62]; **DNNs**, which form the foundation of many deep learning models, providing flexibility for general tasks [63]; **Recurrent Neural Networks (RNNs)**, well suited for sequential data [64]; and **Graph Neural Networks (GNNs)**, designed to handle structured data represented as graphs, such as molecular structures and complex networks [65]. Each of these approaches provides specialized solutions for complex challenges, highlighting the adaptability and broad

applicability of deep learning in diverse fields.

### 2.9.2 Convolutional Neural Networks

**CNNs** represent a class of **DL** models that are fundamental in various fields, designed to automatically and adaptively learn spatial hierarchies of features through multiple layers, including convolutional, pooling, and fully connected layers [66].

**CNNs** consist of several essential components:

- **Convolutional Layers**: Apply filters to input data to generate feature maps, capturing spatial hierarchies of attributes.
- **Pooling Layers**: Reduce the dimensionality of feature maps, often using techniques like max pooling or average pooling, which helps manage computational cost and control overfitting.
- **Activation Functions**: Nonlinear functions, such as ReLU, are crucial for introducing non-linearity into the model, enabling it to learn complex patterns.
- **Regularization Techniques**: Methods like dropout and weight regularization (L1, L2) are employed to prevent overfitting, improving the generalization of the model .

Although **CNNs** have achieved significant success, they still face challenges such as the need for large datasets and the risk of overfitting, especially in medical applications where data may be limited [67].

### 2.9.3 Recurrent Neural Networks

**RNNs** are a class of artificial neural networks designed to recognize patterns in data sequences, making them highly effective for tasks involving sequential information. These networks stand out for their ability to maintain a form of memory through internal loops, allowing them to process and analyse sequences of inputs efficiently. There are various types of **RNNs**, **Long Short-Term Memory (LSTM)** networks and Bidirectional **RNNs**, each with specific characteristics that enhance their ability to model sequential data [64].

#### Bidirectional RNNs

Bidirectional **RNNs** extend the capabilities of standard **RNNs** by processing data in both forward and backward directions. This approach enables this model to leverage information from both past and fu-

ture contexts, which can enhance performance in tasks such as phoneme classification and sequence prediction [68].

#### LSTM

**LSTM** networks are a type of **RNNs** designed to overcome the limitations of traditional **RNNs** in learning long-term dependencies. By incorporating gating functions into their architecture, this model is effectively address the vanishing gradient problem, which is a common challenge in standard **RNNs** when processing long sequences. **LSTM** have become the dominant architecture in many applications involving sequential data due to their superior performance in capturing long range dependencies [64].

### 2.9.4 Graph Neural Networks

**GNNs** have emerged as a powerful tool for analysing and learning from graph structured data. They extend **DL** techniques to non-Euclidean data, capturing complex relationships and dependencies within graph structures.

**GNNs** are specifically designed to process data with complex dependencies and relationships represented as graphs, where nodes are connected by edges. The goal is to learn representations for each node in the graph, integrating information from neighboring nodes and their relationships. The process of this model involves iteratively aggregating information from neighboring nodes, updating the node representations, and repeating this process across multiple layers to capture complex dependencies and information diffusion throughout the graph. This capability is particularly useful for predicting molecular activity, identifying potential therapeutic targets, and exploring interactions within biological pathways. The ability of **GNNs** to handle complex relational data makes them indispensable for the computational analysis of molecular and genetic data in drug discovery [33].

## 2.10 Deep Learning applied to Drug Target Interaction

### 2.10.1 Landscape of DL for DTI

In recent years, **DL** techniques have played a central role in computational biology, particularly excelling in **DTI** prediction and the processing of protein and chemical compound data. The application of these techniques enables robust multimodal approaches capable of exploring complex representations and identifying patterns that connect proteins and small molecules.

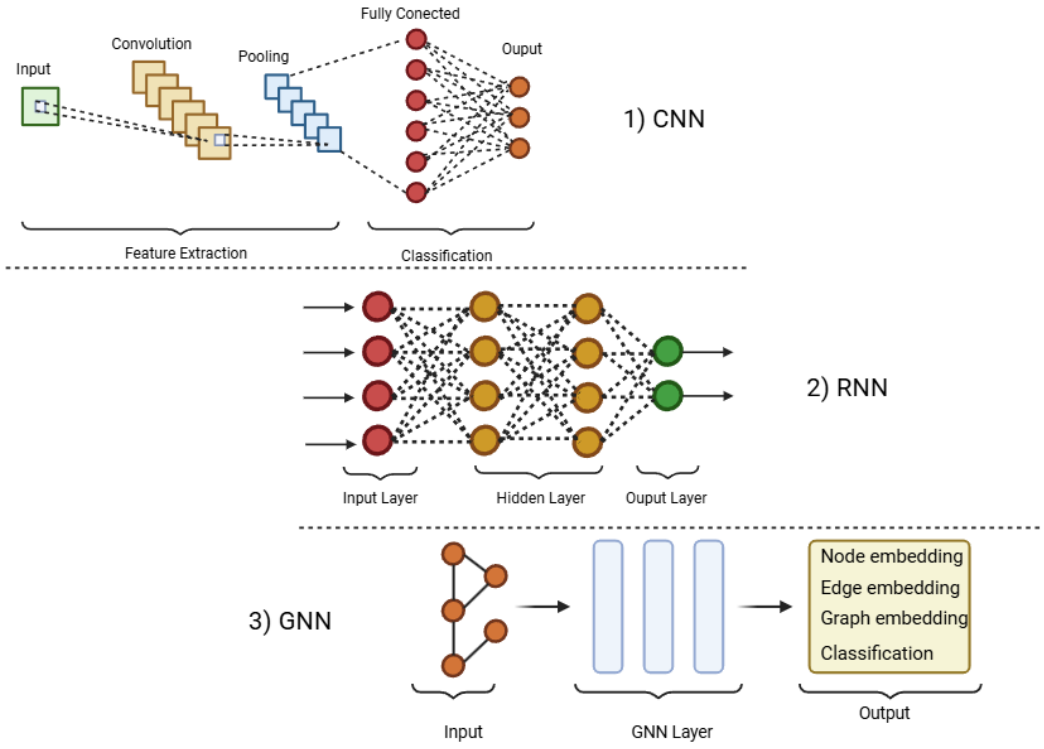



Figure 3: Three types of **DL**. (1) **CNNs** uses convolutional and pooling layers for feature extraction, followed by fully connected layers for output generation. (2) **RNNs** processes sequential data through input, hidden, and output layers, capturing temporal dependencies. (3) **GNNs** employs graph structures with node, edge, and graph embeddings for learning and classification tasks. Adapted from [33]

One of the most critical aspects in **DTI** prediction is the integration of high-quality and biologically relevant datasets. **ML** and **DL** models rely heavily on accurate experimental binding data to achieve meaningful generalisation. Consequently, databases that provide curated and standardised affinity information have become essential components of modern **DTI** pipelines. Among the most widely used resources are Open Targets [27], DrugBank [23], TTD [69], BindingDB [70], ChEMBL [71], PubChem [72], Davis [73], and KIBA [74], each contributing complementary types of interaction data. Together, these databases form the foundation for benchmark datasets that enable model comparison, reproducibility, and the development of new prediction strategies.

Various **DL** architectures have been developed to handle molecular and proteomic data, leveraging specialized representations such as SMILES, protein sequences, and molecular graphs. The creation of multimodal models relies on the fusion of these distinct representations, allowing for more comprehensive and precise analyses [75].

Another strategy involves combining structural and functional data through hybrid neural networks, which use embeddings to integrate semantic and structural information from proteins and chemical compounds, providing a richer perspective on their interactions as we can see in figure 4. Additionally, multi-target analysis, performed by advanced models, can predict multiple interactions between different targets and drugs, overcoming the limitations of conventional methods that focus on isolated interactions [76].

These models have proven highly effective in reducing the search space for candidate drugs, accelerating the process of discovering new compounds, and facilitating the repurposing of existing drugs. Furthermore, they provide efficient solutions for dealing with limited and imbalanced datasets, challenges often encountered in biological studies.

The application of **ML** techniques has become a cornerstone in **DTI** prediction due to their ability to handle diverse datasets and uncover hidden patterns in drug and target properties. As shown in Table 2, **ML** models such as Barlow Twins [77], utilize molecular descriptors, such as SMILES and fingerprints, to train classifiers. Algorithms like **SVM**, XGBoost have demonstrated considerable success in predicting interactions. However, these methods often struggle with challenges such as data imbalance and high dimensionality, requiring advanced preprocessing and feature selection strategies [6].

On the other hand, **DL** has transformed the field of **DTI** prediction by enabling automatic feature extraction and the integration of complex data types. Unlike traditional **ML** methods, which require manual feature engineering, **DL** models can process raw data to learn hierarchical representations. **CNNs**, for instance, are particularly effective in capturing the spatial and structural features of molecular data. Models like **BCM-DTI** [78] utilize **Convolutional Neural Network (CNN)** to predict **DTI** with high accuracy,

often integrating advanced techniques to handle data imbalance, **BCM-DTI** is a multimodal approach that combines drug and target features, using SMILES strings for molecular structures and protein sequences as inputs to independently extract features through **CNNs** layers, before merging them into a unified representation for interaction prediction. This allows the model to leverage complementary information from both drugs and targets, enhancing its predictive power.

**GNNs** have also gained prominence, with models such as NGCN [79] and Graphormerdti [80] leveraging topological and relational properties of molecules and biological targets to improve prediction accuracy. NGCN, in particular, employs graph based convolutional layers to capture interactions in heterogeneous networks by representing drugs and targets as nodes, with their connections encoded as edges. This enables the model to learn complex relationships across multiple biological entities, integrating diverse types of data such as molecular graphs, protein-protein interaction networks, and biological annotations. By incorporating these multimodal data sources, NGCN achieves high predictive accuracy and robustness in the context of **DTI**.

The prediction of **DTI**s relies heavily on diverse and comprehensive datasets that integrate chemical, genomic, and proteomic information. Commonly used databases include BindingDB [70], which provides extensive data on drug-target binding affinities, and ChEMBL [71], a database of bioactive molecules with drug-like properties. Other resources such as DrugBank [23] and PubChem [72] offer additional data on drug molecules, target proteins, and their associated pathways. The types of data utilized in these predictions typically include SMILES strings for encoding molecular structures, protein sequences for target characterization, and molecular fingerprints for structural analysis. Additionally, biological networks, such as protein-protein interaction networks, and gene expression data are often incorporated to enhance the predictive power of computational models. These heterogeneous datasets serve as the foundation for building **ML** and **DL** frameworks, enabling accurate and scalable **DTI** predictions.

Despite the significant progress in the field, there remain challenges such as data sparsity and the limited availability of labeled data, which make it difficult to train robust models, particularly in the context of **DL**. Additionally, class imbalance, characterized by the over representation of negative interactions in datasets, results in biased models that require the use of advanced balancing techniques.To overcome these challenges, future research should prioritize the collection of high quality data as well as the development of hybrid models that integrate traditional and deep learning approaches.

Table 2: Examples of **ML** and**DL** applied to **DTI**.

| | Author | Year | Data | Algorithms | Metrics | Representation | Outcome |
|---|---|---|---|---|---|---|---|
| **ML** | Olayan et al. [81] | 2018 | Yamanishi [82] | RF | AUC, AUPR | Fingerprints, Protein Descriptors, Graphs | Drug-Target binding (Binary) |
| | Nascimento et al. [83] | 2016 | Yamanishi [82] | SVM | AUPR | Protein Descriptors, SIMCOMP | Predicts Interaction (Regression) |
| | Schuh et al. [77] | 2025 | Davis [84], BindingDB [70] | Barlow (DL), GBM | ROC AUC, PR | SMILES, Sequence-Based Encoding | Drug-Target binding (Binary) |
| | El-Behery et al. [85] | 2021 | Benchmark and DrugBank [23] | RF, LightBoost, ExtraTree, ANN, SVM | AUC, PR, Accuracy | Sequence-Based Encoding, SMILES | Drug-Target binding (Regression) |
| **DL** | Yang et al. [86] | 2022 | Metz [87], KIBA [74], Davis [84] | CNN | MSE, CI, r2 | Sequence-Based Encoding, SMILES, Graphs | Drug-Target binding |
| | Cao et al. [79] | 2024 | Luo [88] | GCN | AUROC, AUPR | Graph | Drug-Target binding (Binary) |
| | Dou et al. [78] | 2023 | BindingDB[70], Davis [84] | CNN | AUC, PR, PRC | Amino Acid Sequences, SMILES | Drug-Target binding (Bunary) |
| | Gao et al. [80] | 2024 | DrugBank [23], KIBA [74], Davis [84] | GNN | AUC, AUPR, F1-score, MCC | Sequence-Based Encoding, Graph | Predicts interaction (Binary) |
| | Chen et al. [49] | 2021 | Tox21 [89] | GCN | ROC-AUC | Graph | Chemical toxicity prediction (Binary) |

### 2.10.2 Barlow Twins for DTI

The Barlow Twins framework [90] represents a milestone in **SSL** by proposing a conceptually simple, yet powerful objective based on the principle of redundancy reduction. The method aims to learn informative and non redundant feature representations without the need for labelled data, large batch sizes, or asymmetric network designs.

Barlow Twins employs a architecture consisting of two identical neural networks, referred to as the twin encoders. Each encoder processes a differently augmented view of the same input sample, denoted $y_A$ and $y_B$. These augmentations may involve random transformations (e.g., masking, dropout, or noise injection), ensuring that the model learns invariant features rather than memorising input-specific noise. Each encoder is composed of two main components:

- Encoder network $f_\theta$ — typically a deep neural network that maps the input to a latent feature space.
- Projection head $g_\phi$ — a smaller multilayer perceptron that projects the encoder's latent features into a representation space suitable for redundancy reduction.

Formally, given an input $x$, two distorted versions $x_A$ and $x_B$ are generated and passed through the twin networks as we can see on equation 2.11:

$$z_A = g_\phi(f_\theta(x_A)), \quad z_B = g_\phi(f_\theta(x_B)) \tag{2.11}$$

The two resulting embeddings, $z_A$ and $z_B$, are then batch normalised and used to compute a cross correlation matrix (equation 2.12:

$$C_{ij} = \frac{\sum_b z_{A,b,i} z_{B,b,j}}{\sqrt{\sum_b (z_{A,b,i})^2} \sqrt{\sum_b (z_{B,b,j})^2}}, \tag{2.12}$$

where $i$,$j$ index the embedding dimensions and b enumerates batch samples.

The learning objective, known as the Barlow Twins loss, is defined in equation 2.13:

$$\mathcal{L}_{BT} = \sum_i (1 - C_{ii})^2 + \lambda \sum_{i \neq j} C_{ij}^2, \tag{2.13}$$

where:

- Invariance term $(1 - C_{ii})^2$ penalises deviations of the diagonal elements from 1, encouraging both network branches to produce similar representations for different views of the same input.

- Redundancy reduction term $C_{ij}^2$ minimises the off-diagonal correlations, promoting independence between feature dimensions.

- The hyperparameter $\lambda$ controls the trade-off between the two objectives, typically set to a small value (e.g., $5 \times 10^{-3}$).

This formulation encourages the network to learn informative, decorrelated embeddings that are both robust to noise and compact in representation. Importantly, the Barlow Twins objective requires neither contrastive negative pairs nor large memory banks, leading to a simpler and more stable training process.

Building on this theoretical foundation, BarlowDTI [77] appeared, an adaptation of the Barlow Twins framework for **DTI** prediction. In this domain, the model learns joint molecular and protein representations using only one dimensional (1D) inputs, SMILES strings for compounds and amino acid sequences for proteins, thus eliminating the dependency on costly or uncertain 3D structural data as it can be seen in Figure 4 .

In this adaptation, Molecular structures are encoded using **ECFP**s derived from SMILES strings, which serve as fixed length binary descriptors capturing local atomic environments. Protein sequences are represented using transformer based embeddings, which encode biochemical and structural information directly from amino acid sequences. Two parallel encoders are trained using the Barlow Twins loss, one for the molecular branch and one for the protein branch. Each branch is followed by a projection head, and their outputs are concatenated to form a unified drug–target embedding space.

During pretraining, the model learns to align molecular and protein representations that are semantically related, promoting cross-modal invariance and redundancy reduction across the joint feature space. The resulting embeddings are biologically meaningful and can be transferred to downstream supervised tasks such as **DTI** classification or regression.

Following pretraining, the embeddings produced by the Barlow Twins encoders are used as inputs to traditional machine learning classifiers, such as XGBoost or linear SVM, which operate on the latent space to perform binary classification of drug–target pairs. This hybrid pipeline, self-supervised feature extraction followed by supervised classification, combines the interpretability and sample efficiency of gradient-boosted trees with the representational richness of deep self-supervised learning.

In the context of **DTI** prediction, BarlowDTI achieved competitive or superior performance compared to state-of-the-art supervised deep learning architectures, while maintaining lower computational cost and greater interpretability. The model successfully identified catalytically relevant residues and ligand-binding motifs purely from sequence-based inputs, underscoring its capacity to extract meaningful biochemical relationships from raw data.

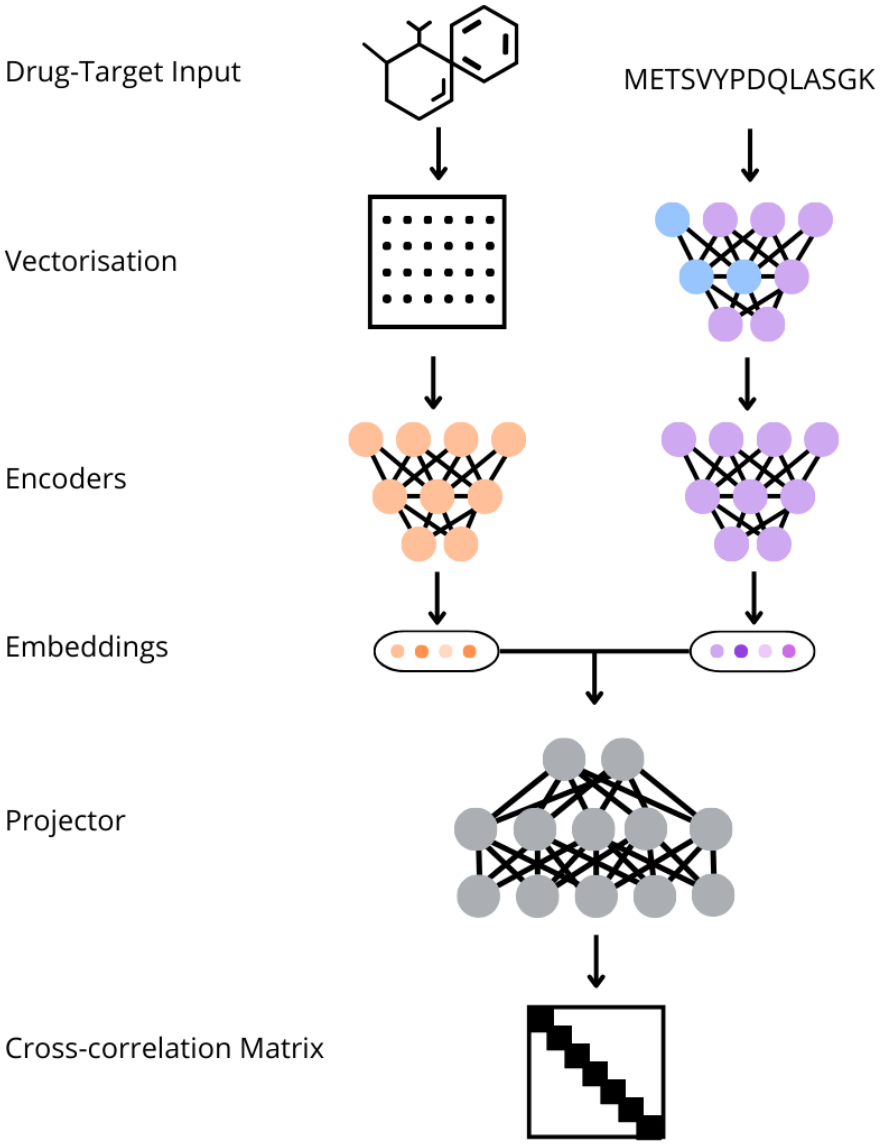



Figure 4: Schematic representation of the Barlow Twins architecture applied to **DTI**. The inputs (chemical structure of the compound and protein/sequence of the target) are first vectorized, processed by independent encoders, and transformed into embeddings. These embeddings are then passed through a projection layer, after which the cross correlation matrix is computed and used as the basis for the self-supervised loss function. Figure adapted from [77].

### 2.10.3 BCM-DTI framework

The **BCM-DTI** framework [78] represents a significant advance in the use of **DL** for **DTI** prediction. Traditional **DTI** models generally treat drug molecules and proteins as whole entities, often using global representations such as SMILES strings, molecular graphs, or amino acid sequences. However, such approaches overlook the fact that biological interactions typically occur between specific active fragments of a drug and localized regions of the target protein. To address this limitation, **BCM-DTI** introduces a fragment oriented learning paradigm, designed to explicitly model the substructural and functional elements that drive molecular recognition.

While previous models achieved strong results using semantic or topological embeddings, they remained limited by either high computational cost or insufficient interpretability. **BCM-DTI** overcomes these challenges by introducing a **Branch Chain Mining (BCM)** strategy for chemically informed fragment extraction and a **Category Fragment Mapping (CFM)** module for protein segmentation, combined in an end-to-end 1D **CNN** architecture.

The overall framework consists of three main components: Fragment Extraction, Feature Learning, and Prediction.

**Fragment Extraction**: **BCM-DTI** begins by decomposing both drugs and proteins into biologically meaningful fragments. For drugs, the **BCM** algorithm parses SMILES strings and identifies three complementary fragment types:

- Branch chains, representing substituents that influence molecular binding and flexibility.
- Common substructures, such as benzene, hydroxyl, carbonyl, or aldehyde groups.
- Motif fragments, derived using the **Retrosynthetic Combinatorial Analysis Procedure (RECAP)** rules to mimic synthetic building blocks based on reaction chemistry.

This approach contrasts with purely statistical segmentation method by integrating domain knowledge from medicinal chemistry, ensuring that the extracted fragments are chemically valid and functionally relevant. Each molecule is first canonicalised to its standard SMILES representation (via RDKit) before fragmentation, ensuring structural consistency across molecules.

For proteins, the **CFM** module groups amino acids according to physicochemical similarity (e.g., aliphatic, aromatic, acidic, basic, hydroxyl). Each amino acid is assigned to one of eight categories, forming a categorical sequence that is then segmented into non overlapping k-grams. This representation

captures both local physicochemical patterns and higher order dependencies across residues, producing a structured set of target fragments.

**Feature Learning**: Following fragmentation, **BCM-DTI** applies a **CNN** based feature extraction module to learn cooperative patterns between drug and protein fragments. Both drug and target fragment sequences are embedded via trainable embedding layers to produce dense numerical representations. For each modality, a stack of 1D convolutional blocks processes the embeddings, capturing local dependencies and fragment fragment interactions through sliding filters. Each CNN block consists of a 1D convolutional layer, batch normalization, **Exponential Linear Unit (ELU)** activation function, chosen for its ability to accelerate convergence and handle sparse activations as we can see on figure 5

Formally, the transformation can be expressed in equation 2.14:

$$X_i = \text{ELU}(X_{i-1} * W + b), \tag{2.14}$$

where $X_i$ is the output of the $i$-th layer, and $W$,$b$ are the convolutional parameters. The fragment level embeddings for drugs and targets are then concatenated to form a joint latent representation

**Prediction Layer**: The concatenated embedding is fed into a multi layer perceptron composed of three fully connected layers with ReLU activations and dropout regularization. The network outputs the final interaction probability through a sigmoid activation, optimised via binary cross-entropy loss:

$$\mathcal{L}_{BCE} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1-y_i)\log(1-p_i)] \tag{2.15}$$

where $y_i$ and $p_i$ denote the ground-truth label and predicted probability, respectively.

Overall, **BCM-DTI** represents a paradigm shift from molecule level to fragment level reasoning in **DTI** prediction. By modelling the synergistic interactions among chemically meaningful substructures, it provides a biologically interpretable and computationally efficient framework, positioning it as one of the leading approaches for large scale, data efficient drug discovery.

### 2.10.4 SHAP values for DTI

The **SHAP** framework [91], provides a unified and theoretically grounded approach for interpreting the predictions of complex **ML** models. Its core principle lies in attributing to each feature a quantitative contribution to the model's output, thereby offering an interpretable decomposition of predictions across both linear and non linear architectures. **SHAP** is particularly relevant in scientific domains where explainability is as crucial as accuracy. Understanding the factors driving predictions, for instance, molecular substructures influencing binding affinity, is essential for decision making.

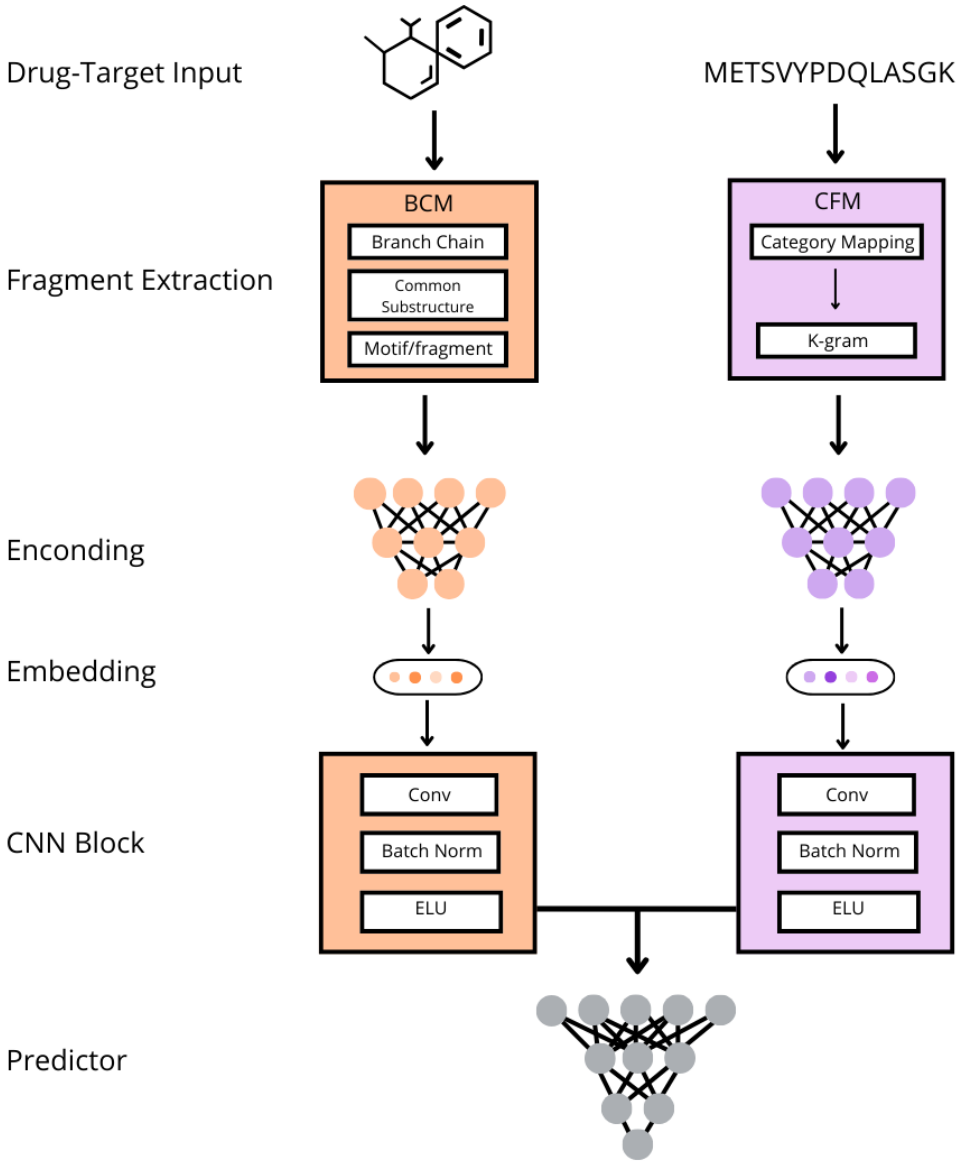



Figure 5: Schematic overview of the **BCM** and **CFM** framework for **DTI** prediction. A drug–target pair is processed through fragment extraction, where the drug is decomposed into branch chains, common substructures, and motifs/fragments (**BCM**), while the target protein sequence undergoes category mapping and k-gram processing (**CFM**). Both representations are then encoded, passed through **CNN** blocks for feature learning, and pooled into vector embeddings. Finally, the combined embeddings are fed into a fully connected layer to predict the interaction outcome. Figure adapted from [78].

The **SHAP** framework redefines model interpretability by expressing predictions as deviations from a baseline expectation $E[f(x)]$. For a given input $x$, the prediction can be represented on equation 2.16:

$$f(x) = \mathbb{E}[f(x)] + \sum_i \phi_i, \tag{2.16}$$

where $E[f(x)]$ represents the average model prediction, and $\phi_i$ represents the additive deviation introduced by feature $i$.

This formulation allows practitioners to decompose any prediction into a human interpretable contribution map, directly revealing which features increase or decrease the model's confidence in a particular output.

In the context of **DTI**, **SHAP** provides a powerful interpretability layer for complex predictive pipelines. When applied to models such as XGBoost, its possible to identify which molecular substructures (**ECFP** bits) most strongly influence the prediction of activity and how specific protein sequence features contribute to binding affinity.

By aggregating **SHAP** values across the test set, one can compute global feature importance metrics, such as the mean absolute **SHAP** value per bit or residue. These analyses help to reveal which fragments (e.g., hydroxyl groups, amides, or heterocycles), consistently drive predictions across molecules.

This interpretability enables mechanistic insight into model behaviour, facilitating hypothesis generation for medicinal chemistry and guiding compound prioritisation in virtual screening and drug repurposing.

# Chapter 3

# Methodology

This work followed a structured workflow to achieve its objectives, starting with the curation and preparation of **DTI** data and the construction of a **TB** specific dataset. Predictive models were then selected, trained, and evaluated using **ML** and **DL** approaches. The methodology included data preprocessing, model configuration, **SL** and **SSL**, evaluation using standard metrics, and interpretative analyses to extract biological insights from the predictions. All experiments were implemented in Python within Jupyter Notebooks.

## 3.1 Relevant Python packages and tools

As **ML** and **DL** have gained popularity, Python has emerged as a dominant language with a rich ecosystem of packages and tools. This thesis leveraged several key Python libraries, each playing a crucial role in its development and implementation. Below is a summary of the essential packages utilised:

**Pandas** [92]: is efficient for data manipulation and analysis. It provides high level structures for handling structured and labelled data, with functions to load datasets from formats such as CSV, Excel, and SQL. It supports fast operations, including filtering, sorting, grouping, merging, and reshaping.

**NumPy** [93]: provides a comprehensive collection of functions for mathematical and logical operations, array manipulation, sorting, selection, input/output, and other related tasks.

**Matplotlib** [94]: is a widely used library for data visualisation, providing functionalities such as bar plots, heatmaps, box plots, and many other types of graphical representations.

**chembl_webresource_client** [95]: is the official Python client for accessing the ChEMBL database via its REST API. It provides programmatic access to bioactivity data, compound information, and target annotations, allowing streamlined integration of ChEMBL resources into computational pipelines.

**Scikit-learn**[96]: it is an open-source library for **ML**, providing a flexible architecture to implement pipelines for supervised and unsupervised models.

**TensorFlow** [28]: developed by Google, is a platform for **ML** and **DL** research. It provides a comprehensive ecosystem of libraries, tools, and community support, excelling in the development and training of **NN**s. Its high performance, GPU scalability, and ability to export models across devices make it one of the most commonly adopted frameworks in the field.

**Keras** [28]: integrated with TensorFlow, Keras simplifies the construction of **NN**s by providing a high level interface for model design and training, with emphasis on simplicity, flexibility, and scalability.

**PyTorch** [97] stands out in tensor operations and automatic differentiation, enabling the development of complex **DL** models. As an open-source framework, it supports efficient manipulation of numerical data and is widely recognised for its performance and flexibility capabilities in scientific computing.

**Fair-esm** [56]: is a toolkit created by Facebook AI Research that provides streamlined access to advanced transformer based protein language models. It enables the extraction of detailed, per residue representations from protein sequences, utilising transformer architectures to support cutting edge research in protein analysis.

**UMAP** [48]: is a dimension reduction technique that preserves both local and global data structure when projecting high dimensional data into a lower dimensional space. It is particularly effective for visualising complex datasets, enabling the identification of patterns, clusters, and relationships that may not be apparent in the original feature space.

**RDKit** [98]: is an open source cheminformatics library widely used for molecular representation and analysis. It provides tools for working with chemical structures, including SMILES parsing, molecular fingerprinting, descriptor calculation, and substructure searching.

**SHAP** [91]: is designed to interpret **ML** models. Based on cooperative game theory, it assigns each feature an importance value that reflects its contribution to a given prediction. SHAP provides both global and local explanations, offering visualisations that make complex models more transparent and aiding in the extraction of meaningful insights from predictions.

## 3.2 Data Collection and Preprocessing

In the initial phase of this work, the focus was placed on the identification, collection, and analysis of relevant data for the study of **DTI**. To this end, a systematic exploration of scientific literature and specialised bioinformatics repositories was conducted to identify datasets related to **DTI** that could serve as a solid foundation for the development and training of predictive models. Several datasets were evaluated, including KIBA [74], Davis [84], BindingDB [70], Metz [87] and other specific subsets reported in

the literature. After gathering the relevant data, the preprocessing stage was carried out to ensure its consistency, reliability, and suitability for predictive modelling. To this end, each dataset was evaluated in terms of size, level of curation, and relevance. Duplicates, inconsistent records, and non standardised annotations were identified and removed. The dataset selected for model development was Papyrus [99], a comprehensive bioactivity resource. Although multiple datasets were considered during the data collection phase, Papyrus was ultimately chosen due to its scale, harmonisation strategy, high degree of manual and automated curation and versatility. The reasoning behind this selection will be thoroughly discussed in the results chapter 4, where its performance and contribution to the pipeline are evaluated in depth.

Papyrus is a large-scale dataset comprising millions of bioactivity measurements collected from reputable sources such ChEMBL [71], ExCAPE-DB [100], BindingDB [70] and others. A distinguishing feature is the classification of all data points into three quality levels, high, medium, and low based on specific criteria, including duplication flags, censored values, absence of quantitative measurements, or questionable activity records. For this work, only high quality interactions were retained, ensuring a trustworthy foundation for model training.

Another key advantage of Papyrus is its harmonisation of diverse bioactivity measurements, converting various bioactivity measurements (e.g., $IC_{50}$, $EC_{50}$, $K_i$, $K_d$) into standardised pChEMBL values. This metric represents the negative base-10 logarithm of the molar concentration of biological activity, as we can see in equation 3.1

$$pChEMBL = -\log_{10}\left(Activity\ [M]\right) \tag{3.1}$$

For instance, a compound with an $IC_{50}$ of 1 $\mu M$ would correspond to a pChEMBL value of 6, while an $IC50$ of 10 $nM$ corresponds to a pChEMBL value of 8. This transformation enables approximate comparisons across different compounds and experimental setups. Furthermore, targets are linked to well defined UniProt [101] identifiers, which facilitates the integration of protein sequence data and downstream analysis.

After selecting the dataset, the first step involved confirming the structural and biological representations of the entities: it was essential to verify that all compounds (drugs) were encoded in the SMILES format and that all targets were represented as amino acid sequences. This validation ensured that the data was compatible with the modelling pipelines and feature extraction techniques applied in subsequent stages. Following this, all entries containing missing values or duplicates were removed.

Protein sequences were then processed to ensure uniformity of input length. The distribution of sequence lengths was first analysed (Figure 6), revealing substantial variability across proteins. To balance

biological representativeness and computational efficiency, a truncation threshold was set at the 95th percentile of the observed distribution (≈ 1,390 amino acids). Sequences longer than this threshold were truncated, while shorter sequences were retained and zero padded to the same maximum length. This procedure preserved most of the biological variability while ensuring compatibility with batch training in **NN**.

For molecular representations, SMILES strings were canonicalised using RDKit [98] to standardise atom and bond ordering. The distribution of SMILES lengths was then analysed (Figure 6), showing that 95% of the molecules contained fewer than 76 characters. To ensure compatibility with batch processing, all SMILES were zero padded to a fixed maximum length of 76 characters. This strategy provided a uniform input representation while minimising unnecessary computational overhead and avoiding the truncation of structurally relevant information.

The following involved creating a binary classification label by applying a threshold to the pChEMBL values. In this work, compounds with $pChEMBL \geq 6.5$ were considered active, while those below this threshold were labelled as inactive. This cutoff was chosen to ensure biological relevance and to maintain consistency with the Papyrus dataset, where the same activity threshold was adopted for bioactivity classification [99]. This step allowed the original regression problem to be reformulated as a binary classification task. Such binarisation simplifies the prediction problem compared to modelling continuous activity values, and is a common strategy in cheminformatics to facilitate robust performance. It also makes the data directly suitable for training **ML** and **DL** models designed to distinguish active from inactive compounds.

To further enhance the robustness of model training and evaluation, a stratified data splitting strategy was employed. By using the class label as a stratification variable, the dataset was partitioned into 70% for training, 15% for validation, and 15% for testing, ensuring a balanced representation of both active and inactive instances across all subsets. This procedure was essential to prevent imbalanced class distributions, which could otherwise impair the model's generalisation ability and result in misleading performance metrics.

## 3.3 Tuberculosis dataset

A key objective of the project was to construct a curated dataset specifically focused on tuberculosis. This dataset was built by integrating and filtering data from three major public databases, DrugBank [23], BindingDB [70], and ChEMBL [71], while retaining only interactions relevant to **MTB**.

The process began by parsing the DrugBank XML file, iterating through each drug entry. To identify

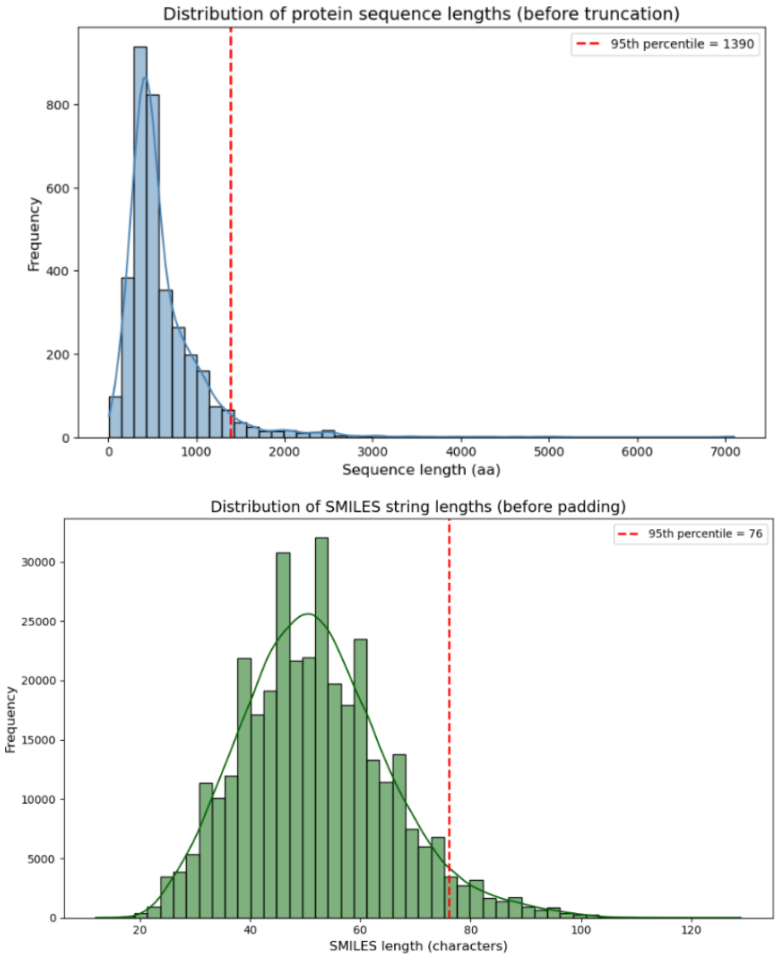



Figure 6: Distribution of input lengths before preprocessing. (Top) Protein sequence lengths (in amino acids) prior to truncation, with the 95th percentile corresponding to 1,390 residues (red dashed line). (Bottom) SMILES string lengths prior to padding, with the 95th percentile corresponding to 76 characters (red dashed line). These thresholds were used to standardize input sizes for model training.

compounds potentially relevant to **TB**, a semantic filtering strategy was employed using keywords such as "tuberculosis", "MTB", "TB", among others. For each drug matching these criteria, the following attributes were extracted: SMILES, mechanism of action, bioactivity measurement values (when available), and associated molecular targets. Only targets explicitly related to **TB**, either by name or by organism, were considered. For each valid target, both the UniProt ID and its corresponding amino acid sequence were retrieved when available. This process culminated in the creation of a drug $\times$ target matrix. Each pair was annotated with a binary label reflecting the evidence contained in DrugBank: active (1) when the UniProt identifier of a **TB** related target appeared in the list of targets reported for a given drug, indicating an experimentally validated interaction, and inactive (0) otherwise, when no such association was recorded in DrugBank. The labels were assigned automatically by cross-matching the list of **TB** drugs with the extracted targets, thereby ensuring a systematic and reproducible annotation procedure. This binarisation step standardised heterogeneous DrugBank evidence into a format suitable for supervised models.

The procedures for data extraction from BindingDB and ChEMBL followed an approach similar to the one used for DrugBank. BindingDB data were downloaded in TSV format and processed using the Python package pandas [92] for parsing, cleaning, and filtering entries related to **MTB**. For ChEMBL **TB** related targets were retrieved using the Python package chembl_webresource_client [95], querying the REST API for targets whose organism contained **MTB**. To ensure uniformity in activity representation, all values were converted into pChEMBL scores.

After extraction, data from three datasets were merged into a single unified dataset. Duplicate entries were removed based on shared SMILES, protein sequence, and pChEMBL values. In cases where the same drug–target pair had multiple activity entries, the median pChEMBL was computed and retained. A binary label was then assigned to each interaction: interactions with pChEMBL $\geq$ 6.5 were labelled as active (1), otherwise inactive (0). To ensure no data leakage between datasets, a strict filtering procedure was also applied to ensure that none of the drugs (based on SMILES) or protein targets (based on sequence or UniProt ID) present in this **TB** specific dataset overlapped with those used previously in the Papyrus dataset [99]. This guarantees the independence of the two datasets, which is essential for the evaluation of transfer learning and fine-tuning strategies.

The final dataset comprised approximately 9,089 unique drugs and 185 unique targets. To ensure biological relevance, all protein targets underwent manual curation. For each target, scientific literature was consulted such as [102], [103] or [104] to verify its association with **TB** and to clarify its biological function. This step led to a refined set of 155 validated targets, down from the initial 175.

To further assess the relationship between this curated **TB** dataset and the Papyrus dataset, a two

dimensional **UMAP** projection was generated using the Jaccard/Tanimoto similarity between molecular fingerprints.

Given two binary fingerprints $A$ and $B$, the Tanimoto similarity is defined as:

$$T(A, B) = \frac{c}{a + b - c}, \tag{3.2}$$

where $a = |A|$ is the number of bits set to 1 in $A$, $b = |B|$ is the number of bits set to 1 in $B$, and $c = |A \cap B|$ is the number of common bits set to 1 in both $A$ and $B$.

The Tanimoto index is especially attractive for large scale virtual screening and chemical space analysis due to its computational efficiency and interpretability. Numerous studies have demonstrated its robustness for fingerprint based similarity searching [105].

## 3.4 Model Selection

Based on the analysis presented in the State-of-the-Art section, and taking into account both the availability of code implementations and the suitability of the models to the **DTI** context, two predictive architectures were selected for experimental validation: Barlow Twins [77] and **BCM-DTI** [78]. The decision was supported by the comparative evidence summarised in Table 2, which reviews recent **ML** and **DL** approaches for **DTI** prediction across different datasets, algorithms, and molecular/protein representations. This analysis showed that contrastive and multimodal strategies consistently outperformed traditional methods, further motivating their selection for this work.

The Barlow Twins framework was chosen for its robust **SSL** capabilities. It enables the extraction of meaningful representations from molecular and protein sequences without requiring extensive labelled datasets, an advantageous property in the context of **TB**, where annotated data is often limited. The model has demonstrated strong generalisation in prior bioinformatics applications and was easily adapted to the molecular representation formats used in this work.

In parallel, the **BCM-DTI** model was selected as a **SL** baseline due to its ability to integrate multimodal biological information (e.g., SMILES strings, protein sequences) with biologically grounded constraints, improving interpretability and performance. Its prior success in generalising to unseen compounds and targets aligns well with the challenges of identifying novel **TB** targeted therapies.

The combination of these two models, one **SL** and one biologically constrained and supervised, offers a complementary methodological perspective for tackling **DTI** prediction. The **SSL** model (Barlow Twins) leverages large unlabelled datasets to learn generalisable molecular and protein representations, whereas

the SL (BCM-DTI) incorporates biological priors to guide prediction towards mechanistic relevance. By combining these approaches, the overall framework benefits both from data driven feature learning and from biologically informed constraints, thereby enhancing robustness and enabling evaluation across diverse molecular representations and learning paradigms.

## 3.5 Barlow Twins Model

The Barlow Twins model was adopted as a representative SSL architecture. In this work, its embeddings learned from perturbed molecular and protein sequences, were subsequently used as input to traditional ML classifiers (e.g., XGBoost and linear SVM).

In the initial phase of the work, the primary goal was to replicate the results originally reported by the authors of the Barlow Twins model, ensuring that the methodology and implementation were consistent with their study. This validation step was carried out using the same dataset employed in the original publication, allowing us to confirm the reproducibility of the approach before applying it to our specific use cases. Following this validation, experiments were conducted under three dataset configurations, corresponding to those introduced earlier in this chapter:

- **Papyrus dataset**: a large scale bioactivity dataset;
- **Tuberculosis dataset**: a manually curated dataset containing only drug–target pairs relevant to TB;
- **Papyrus + Tuberculosis dataset**: a merged version aimed at increasing variability and generalisation capacity.

Across these datasets, two feature extraction strategies were systematically compared. The first relied on descriptors used as a baseline, namely ECFP derived from SMILES strings for molecular structures and ESM2 based encodings for protein sequences, the second employed representations learned through the Barlow Twins model, which in this work was trained from scratch. In both approaches, molecular inputs were consistently represented as ECFP features generated from SMILES strings. The pretraining step was carried out using the original Barlow Twins loss function, where paired molecule–protein inputs were processed to learn joint representations. For the protein branch, ProtT5 embeddings were used only as initial input features to the encoder, which means that the Barlow Twins architecture was trained in its entirety, rather than relying on pre-trained incorporations.

The preprocessing of the data and the training configuration closely followed the procedure defined in the original Barlow Twins implementation. The dataset was split into training, validation, and test subsets based on a predefined "split" column to maintain consistency across partitions. All training parameters were kept at their default values as reported by the original authors, including a batch size of 512, the use of ReLU activation functions, and the AdamW optimiser with a learning rate of $1 \times 10^{-4}$ and a weight decay of $1.5 \times 10^{-3}$. The encoder and projector networks were composed of hidden layers with 4096 neurons, and the resulting embeddings had a dimensionality of 512.

The entire training process ran for 100 epochs, and model checkpoints correspond to the epoch with the lowest validation loss.

Following pretraining, the embeddings generated were passed to the XGBoost and Linear **SVM** classifiers. These models were evaluated using multiple metrics to assess classification performance: **ROC-AUC** and **PR-AUC** provided a global view of discriminative ability, while accuracy, precision, recall, F1-score, and **Matthews Correlation Coefficient (MCC)** offered insights into binary classification quality and class balance.

## 3.6 BCM-DTI model

Following the integration of Barlow Twins, the next step involved implementing the **BCM-DTI** model. This architecture is a supervised deep learning framework that combines molecular and protein information through a multimodal approach. Molecules are represented via a branch chain decomposition into chemically meaningful fragments, while proteins are encoded through categorical groupings of amino acids. Both representations are processed with 1D **CNNs**, and the resulting embeddings are concatenated into a joint representation to predict **DTI**.

The latent representations obtained from both the drug and protein branches are then concatenated and processed through a dense feedforward **NN** comprising three fully connected layers with 1024, 512, and 256 units, respectively. Each layer is followed by a dropout operation with a rate of 0.5 to prevent overfitting and improve generalisation. A final output layer with a sigmoid activation function produces a probability value between 0 and 1, indicating the likelihood of interaction between the drug and the target protein.

The **BCM-DTI** framework, similar to the procedure followed for the Barlow Twins model, was also subjected to a systematic evaluation. After replicating the original results reported by its authors, the model was tested on the three dataset configurations, under consistent training and evaluation settings.

Performance was measured using the same metrics which were used in the previous model.

Training was conducted using a batch size of 512, a learning rate of $1 \times 10^{-4}$, and the Adam optimiser. Early stopping was employed with a patience threshold of 15 epochs without improvement in the validation loss, ensuring that model selection was guided by generalisation performance rather than training fit. Gradient clipping was also applied during training to stabilise updates and prevent exploding gradients, especially in deeper layers of the model.

This configuration was implemented using the following parameters: the drug embedding size and protein embedding size were both set to 512. For the drug branch, the convolutional structure followed the default progression [25, 256, 512], where 25 corresponds to the input dimensionality of the molecular fragments and the subsequent layers (256, 512) represent successive convolutional channels. For the protein branch, the architecture followed the default progression [462, 128, 256, 512], where 462 corresponds to the input dimensionality of the protein feature space in the Papyrus **TB** datasets, and the following layers (128, 256, 512) capture increasingly abstract representations. Filters of size 32 were applied throughout the convolutional layers, and the model used three fully connected layers with sizes [1024, 512, 256]. All these values were kept as defined in the original **BCM-DTI** implementation. The optimiser learning rate was $1 \times 10^{-4}$ with weight clipping enabled.

Next, this model, originally trained on the Papyrus dataset, was fine-tuned to improve its ability to predict biologically relevant interactions specific to the disease.

Due to structural differences between the Papyrus and **TB** datasets, particularly in terms of average sequence length and molecular vocabulary, it was essential to ensure compatibility between them. To achieve this, padding and alignment operations were applied to the **TB** dataset so that the maximum fragment lengths of drugs and proteins matched those used during Papyrus training. Furthermore, the index mapping dictionaries generated from Papyrus were reused to guarantee a consistent numerical encoding of input tokens across datasets, ensuring that identical molecular fragments and amino acids were mapped to the same indices in both Papyrus and **TB**. The **BCM-DTI** model was then initialised with the weights obtained from pretraining on the Papyrus dataset and fine-tuned using only the data from the **TB** dataset. The model's performance was evaluated on the **TB** test set before and after fine-tuning, enabling a direct comparison of its predictive effectiveness.

## 3.7 Testing with Alternative Compounds (Drug Repurposing)

One of the central objectives of this work was to explore strategies for drug repurposing in the context of **TB**. The focus was on identifying compounds not currently employed in **TB** therapy but with a high potential to interact with molecular targets associated with the disease. To address this, a robust and interpretable prediction pipeline was developed, enabling the systematic screening of novel drug candidates using the models previously trained and fine-tuned throughout the project.

To improve computational efficiency and address time constraints, two additional dataset configurations were constructed by combining drug compounds from the Papyrus dataset with different subsets of tuberculosis targets.

The first included the 10 targets most strongly associated with the disease based on evidence from the scientific literature and curated biological databases, ensuring both biological relevance and methodological consistency. This subset included pantothenate synthetase, arabinosyltransferase A, the DNA-directed RNA polymerase beta chain, thymidylate kinase, and multiple isoforms of 3-oxoacyl-[acyl-carrier-protein] synthase, among others, while the second focused exclusively on InhA, the primary molecular target of Isoniazid, the most widely used drug in current **TB** treatment. The selection of these targets was guided by evidence from the scientific literature and refined with the assistance of AI based tools, ensuring both biological relevance and methodological consistency.

The drugs considered were all in SMILES format, while the protein targets were represented by their amino acid sequences. The resulting drug–target pairs formed a Cartesian product, representing all possible combinations between Papyrus compounds and **TB** specific targets. This newly generated dataset served as the input for the predictive models.

For the prediction stage, three previously trained models were employed: the **BCM-DTI** and the Barlow Twins framework with two different classifiers, XGBoost and linear **SVM**. These models, described in detail in earlier sections, were adapted and applied to **TB** specific data to evaluate their ability to identify potential drug–target interactions relevant to the disease. These approaches were selected as they had yielded the best performance in the comparative analyses conducted earlier, ensuring that the evaluation focused on the most promising strategies.

Each model produced, for every compound–target pair, a probability score representing the likelihood of interaction. This not only enabled binary classification but also provided a measure of the model's confidence in each prediction. A decision threshold of 0.5 was applied only for evaluation purposes, in order to compute standard binary classification metrics such as accuracy, precision, recall, F1-score, and

**MCC**. In contrast, for downstream candidate selection, predictions were ranked in descending order of probability, thereby prioritising the most promising compounds with the highest predicted likelihood of interaction.

After generating the predictions, two filtering strategies were applied to further increase the reliability of the results. In one case, only the compound–target pairs for which the model predicted an interaction with at least 95% confidence were retained; in the other, the threshold was increased to 99%. These filtered datasets were then used to perform the comparison across the three models. Finally, to ensure maximum robustness, only the pairs consistently classified as active by both approaches were kept for subsequent analysis.

## 3.8 Automated Filtering Post-Processing Pipeline

An automated computational pipeline was developed with the specific objective of filtering and prioritising candidate compounds predicted to be active against tuberculosis. This step followed the virtual screening task performed by the predictive models. It was designed to refine the initial list of predicted active molecules into a smaller subset that were not only computationally promising but also chemically viable, synthetically accessible, and compatible with essential medicinal chemistry criteria. In this way, the pipeline ensured that the final candidates retained for further analysis represented compounds with a realistic chance of progressing in a drug discovery setting. All scripts and resources developed for this pipeline are openly available at `https://github.com/esperancaa/TB_TESHIS`

The dataset used as input for this pipeline consisted of the filtered list of drug–target pairs predicted as active by the three models, Barlow Twins (XGBoost and SVM classifiers) and **BCM-DTI**, across both the Top 10 tuberculosis targets and the InhA, focused screening strategies described in the previous section. These predictions corresponded to the high confidence interactions (≥95% and ≥99% probability thresholds) retained after cross-model consensus filtering. Each compound, represented in the form of a SMILES string, was processed individually through a sequence of filters designed to replicate practical constraints encountered during pharmaceutical development. These filters were implemented in Python, making extensive use of the RDKit [98] library and associated cheminformatics utilities. The design of the pipeline was modular and sequential, meaning that each compound was tested against a series of ordered filters, and the point of rejection was explicitly recorded. This design provided full transparency and allowed for a detailed audit of where and why a molecule failed.

The first step of the pipeline was the verification of the validity of the SMILES representation, which

attempts to parse each string into a valid molecular graph; compounds that could not be parsed were immediately discarded. This was followed by an element check, in which molecules containing disallowed atoms such as silicon or tin were rejected, since these are rarely found in approved drugs and often lead to problematic chemistry [104]. In the subsequent stage, molecules carrying non-zero formal charges or radical electrons were eliminated, given that such features usually indicate unstable or reactive species.

Attention was then directed to the topology of the molecular structure. Molecules were discarded if they contained ring systems larger than eight members, as such macrocycles are frequently associated with synthetic difficulties and poor drug-like behaviour [106]. Similarly, molecules with more than two bridgehead atoms were excluded because of the structural strain and instability that these motifs can introduce. Phosphorus atoms were also carefully evaluated. Only compounds in which phosphorus appeared in chemically acceptable contexts, such as phosphate-like groups, were retained [107]. Structures in which phosphorus was found in unusual or reactive environments were removed from the dataset.

After these structural checks, the compounds were subjected to a medicinal chemistry filter. This filter consisted of two layers. The first was a basic list of forbidden functional groups, including well known problematic motifs such as aziridines, nitroso groups, acyl chlorides, reactive halogens, and strained triple bonds. These substructures are associated with instability, toxicity, or poor pharmacological performance. Any compound containing one or more of these groups was excluded. The second layer of the medicinal chemistry filter was an expanded list of undesirable motifs, designed to capture additional problematic chemotypes such as conjugated polyenes, quinones, isocyanates, epoxides, nitrosamines, or inappropriately halogenated functional groups [108]. The integration of these two levels ensured a broad coverage of chemically unstable or toxic functionalities.

Beyond structural alerts, physicochemical properties were also considered. Molecules were required to satisfy classical oral drug-likeness criteria, namely Lipinski's Rule of Five [109]. In practice, this meant that compounds had to fall within the following thresholds: molecular weight between 100 and 500 Daltons, calculated logP below 5, no more than 10 hydrogen bond acceptors, and no more than 5 hydrogen bond donors. In addition, more specific physicochemical thresholds were enforced, namely a **Topological polar surface area (TPSA)** of no more than $140\,\text{Å}^2$ and no more than 10 rotatable bonds. These criteria were included to improve the likelihood that the surviving molecules would display adequate solubility, permeability, and overall pharmacokinetic properties.

The final stage of the pipeline combined two widely used cheminformatics metrics: the **Synthetic accessibility (SA)** [110] score and the **Quantitative estimate of drug-likeness (QED)** [111]. The **SA** score provides an estimate of how easy or difficult it would be to synthesise a compound in the laboratory.

Lower values indicate higher feasibility, and in this work, a conservative threshold of SA $< 4.5$ was applied. The **QED** score, on the other hand, provides an aggregate measure of how closely a compound resembles approved drugs, taking into account multiple molecular properties simultaneously. Compounds with **QED** scores below 0.3 were considered unlikely to have sufficient drug-like character and were excluded. Only compounds that simultaneously satisfied both the **SA** and **QED** criteria were allowed to advance.

Altogether, this multistage design integrated three complementary dimensions of compound triage: structural integrity and stability, physicochemical appropriateness, and synthetic feasibility combined with drug-likeness. Every molecule processed was annotated with its outcome at each stage, including whether it passed or failed and, if rejected, the precise reason for its exclusion.

## 3.9 Explainability Analysis with SHAP

To enhance the interpretability of the predictions generated by the classification pipeline based on Barlow Twins embeddings and XGBoost, a **SHAP** analysis was conducted. **SHAP** assigns a value to each feature in every instance, quantifying how much that feature shifts the prediction away from a baseline. Positive values indicate that the feature increases the probability of the "active" class, while negative values indicate the opposite.

The main objective was to identify which molecular features, had the greatest influence on the model's output. By applying **SHAP** it was possible to trace the contribution of each input feature to the final prediction. In a final step, this analysis also aimed to verify whether any of the molecules that passed the previous filtering stages were among those most influential for model training, thus linking predictive outcomes with the internal importance assigned by the model.

Each test set instance was represented by an input vector composed of two main components: molecular **ECFP** descriptors, computed with radius 2 and fixed length of 1024 bits, and protein embeddings, previously extracted from the target sequences using the ProtT5 model. These representations were combined, resulting in the final input vectors used for XGBoost classification. Notably, the first 1024 elements of each vector exclusively corresponded to the **ECFP** bits, which enabled an isolated analysis of their individual contributions.

Based on this global importance ranking, the top 20 **ECFP** bits were selected. This list highlights the molecular substructures that most strongly influenced the classification of drug-target interactions toward the active class. It is worth noting that this is a global analysis, local variability across individual molecules is not captured in this ranking. Nevertheless, the **ECFP** bits can be directly mapped to their corresponding

substructures using the RDKit's bit info dictionary, allowing these features to be interpreted chemically.

Beyond global analysis, a method was implemented to identify which molecules in the test set activated the greatest number of relevant bits. For each molecule, it was determined which of the top ranked bits were present (i.e., set to 1). Each molecule was assigned a bit activation score corresponding to the total number of important bits present. In a subsequent step, we examined whether any of the previously identified and filtered compounds contained some of the most relevant bits for model learning.

## 3.10 Discussion

The methodological path followed throughout this work was designed to ensure scientific rigour, reproducibility, and alignment with the biological and pharmacological specificities of **TB**.

By integrating and curating bioactivity information from multiple sources it was possible to assemble a reliable benchmark that captured biologically relevant interactions while filtering out targets with uncertain or poorly evidenced associations with **TB**. This stage revealed a fundamental methodological insight: the predictive strength of **ML** systems is inherently limited by the quality, balance, and consistency of the input data. The manual verification of protein targets, together with the enforcement of biologically supported annotations, proved essential to ensure dataset validity and to mitigate the effects of data scarcity and heterogeneity.

Building on this foundation, the Barlow Twins framework introduced a **SSL** component that generated meaningful molecular and protein representations, enabling generalization beyond limited **TB** data. Combined with classical classifiers like XGBoost and **SVM**, it balanced **ML** representational strength with interpretability.

In parallel, the **BCM-DTI** model offered a complementary **DL** approach to capture complex molecular–protein interactions. Together, these models broadened methodological diversity and clarified how data quality and model design influence predictive performance.

Another methodological strength of this work lies in the systematic incorporation of interpretability through **SHAP** analysis, which identified molecular features driving predictions and linked computational outputs to chemical insights, while also revealing model biases and representation consistency.

Finally, the automated filtering pipeline ensured that the final set of predicted compounds adhered to real world medicinal chemistry constraints. Implemented using RDKit and custom Python scripts, this pipeline performed rigorous chemical validity checks as well as the exclusion of unstable or toxic substructures. This step transformed the purely computational predictions into chemically actionable

hypotheses suitable for drug discovery pipelines.

Taken together, the methodology developed in this work illustrates how a multi layered, biologically informed, and interpretable pipeline can address the intrinsic challenges of data scarcity, heterogeneity, and imbalance that characterise tuberculosis research. Each methodological decision, from dataset curation to representation learning, classifier design, interpretability, and chemical post-filtering, contributed to a coherent framework capable of balancing predictive performance with scientific rigour and biological plausibility.

Beyond its specific results, the value of this methodology lies in its transferability and reproducibility. The pipeline, designed with modularity in mind, can be readily adapted to other diseases or molecular systems with minimal modification.

In essence, the methodological journey of this work demonstrates that the strength of computational drug discovery does not reside solely in algorithmic sophistication, but in the synergy between biological insight, data integrity, and adaptive learning design. The proposed methodology thus offers a scientifically grounded and practically viable framework, capable of supporting the next generation of data-driven drug discovery efforts against tuberculosis and beyond.

# Chapter 4

# Results and Discussion

Following the implementation of the methodologies described previously, an empirical evaluation of the selected models was conducted to assess their predictive capabilities in the context of **DTI**, with particular emphasis on **TB**. The results presented in this chapter reflect both the quantitative performance of the Barlow Twins and **BCM-DTI** models across different experimental settings, as well as an interpretative analysis of the key factors influencing their predictions. The evaluation was performed on three datasets: the Papyrus dataset, the tuberculosis specific dataset curated for this work, and a combined dataset encompassing both sources.

The analysis includes not only standard classification metrics, such as **ROC-AUC**, **PR-AUC**, F1-Score, Precision, Recall, and **MCC**, but also the use of interpretability tools like **SHAP** to gain insight into the models decision making processes. The results support the methodological decisions made throughout the project and offer both quantitative and qualitative evidence of each model's potential to contribute to drug repurposing efforts and the identification of meaningful **DTI** in the fight against tuberculosis.

## 4.1 Data Collection and preprocessing

At the beginning of this work, several benchmark datasets commonly used in **DTI** studies were considered, including KIBA [74], Davis [84], Metz [87], BindingDB [70], Human[112], and Elangs[73]. These resources, although well established in the literature, present important limitations for the objectives of this project. Many of them are restricted in scope, either focusing exclusively on narrow subsets of targets, such as kinases, or providing relatively small numbers of interactions. Others rely on heterogeneous activity labels ($K_d$, $K_i$, $IC_{50}$, binary values), which complicates harmonisation and comparison across datasets, as one can see in Table 3.

In contrast, the Papyrus dataset was designed as a large scale, harmonised resource that integrates bioactivity data from multiple public repositories, including ChEMBL [95], PubChem [72], and ExCAPE-

DB [100]. One of its main distinguishing features is the systematic curation pipeline applied during its construction. All bioactivity measurements are normalised and mapped to a unified metric (pChEMBL), enabling consistent comparisons across different assays and sources.

The descriptive statistics obtained confirm the scale and robustness of this dataset: Papyrus contains a total of 522,475 drug–target interactions, spanning 349,554 unique compounds and 3,531 unique protein targets. After binarisation at the standard threshold of 6,5, the dataset exhibited a balanced distribution, with 276,002 active interactions (52.8%) and 246,415 inactive interactions (47.2%), which is particularly advantageous for model training since it mitigates the risk of bias caused by class imbalance.

The choice of Papyrus as the primary dataset was motivated by its extensive coverage of both chemical and biological aspects, and, most importantly, by the robustness of its curation process. In contrast to many other datasets, where noisy or inconsistent measurements can introduce bias and hinder generalisation, Papyrus provides a high level of data reliability through the harmonisation of activity scales, the filtering of problematic entries, and the systematic linking of all targets to well defined UniProt identifiers. This combination of scale, standardisation, and rigorous quality control makes Papyrus well suited for the development of **ML** and **DL** models for **DTI** prediction and, in the present context, for building a robust foundation for **TB** oriented repurposing experiments.

Table 3: Summary of benchmark datasets commonly used in **DTI** studies. The table reports the number of entries, unique drugs, unique targets, and the type of binding affinity label associated with each dataset. These data were compiled through a literature review to identify suitable datasets to train **DTI** predictive models.

| Dataset | Year | Nº of entries | Nº of Drugs | Nº Targets | Label Value | Target/Organism |
|---|---|---|---|---|---|---|
| KIBA [74] | 2014 | 117657 | 2068 | 229 | KIBA Score | Human kinases |
| Davis [84] | 2011 | 25772 | 68 | 379 | Kd | Human kinases |
| Metz [87] | 2011 | 35259 | 1423 | 170 | Ki | Human Proteins |
| Papyrus [99] | 2023 | 522475 | 349554 | 3531 | Pchembl | Multiple Species |
| BindingDB [70] | 2024 | 52274 | 10636 | 1413 | Kd | Multiple Organisms |
| BindingDB [70] | 2024 | 990630 | 538540 | 5077 | IC50 | Multiple Organisms |
| BindingDB [70] | 2024 | 374820 | 171831 | 3068 | Ki | Multiple Organisms |
| Human [112] | 2015 | 6728 | 2726 | 2001 | Binary affinity | Human Proteins |
| Elangs [73] | 2018 | 7786 | 1767 | 1876 | Binary affinity | Specific Organism |

## 4.2 Tuberculosis dataset

A major objective of this work was the construction of a curated dataset specifically focused on **MTB**. To this end, three complementary sources of bioactivity information were integrated: ChEMBL [95], BindingDB [70], and DrugBank [23]. Each dataset contributed a distinct set of compounds and targets relevant to tuberculosis. The ChEMBL subset initially comprised 9917 interactions, involving 7741 unique compounds and 99 targets. BindingDB contributed 1951 interactions, with 1452 unique compounds and 37 targets, while DrugBank provided 1596 interactions, corresponding to 18 compounds and 83 targets.

When combined, the three sources resulted in a dataset of 9021 interactions, covering 8976 compounds and 175 unique protein targets. Of these interactions, 1199 were classified as active and 7822 as inactive after applying the standard pChEMBL threshold (6.5). To ensure biological reliability, all targets were subsequently subjected to manual curation, where 20 proteins were excluded due to insufficient or unclear association with **TB** reported in the literature.

The final curated **TB** dataset consisted of 6850 drug–target pairs, involving 6693 unique compounds and 155 validated targets, with a class distribution of 923 active interactions and 5927 inactive interactions (Table 4). Building this dataset proved to be particularly challenging, as available bioactivity data directly related to Tb is scarce. This scarcity not only limited the scale of the dataset but also contributed to a pronounced class imbalance.

Despite these limitations, the dataset provided a disease-specific benchmark for evaluating predictive models and for exploring drug repurposing strategies. Its manual refinement, particularly the elimination of targets lacking clear biological evidence, ensured that only molecular interactions of interest were retained.

Table 4: Overview of the **TB** dataset considered in this study. The table details the number of entries, unique drugs, unique targets, and the distribution of active versus inactive interactions. The Combined dataset aggregates information from 3 sources, while the Cured dataset corresponds to the final curated version used for model training.

| Source | Nº of entries | Nº of Drugs | Nº Targets | Active int. | Inactive int. |
|---|---|---|---|---|---|
| DrugBank | 1596 | 18 | 83 | 26 (1,6%) | 1570 (98,4%) |
| Chembl | 9917 | 99 | 379 | 1113 (11,2%) | 8804 (88,8%) |
| Bindingdb | 1951 | 1452 | 37 | 96 (4,9%) | 1855 (95,1%) |
| Combined | 9021 | 9089 | 175 | 1199 (13,3%) | 7822 (86,7%) |
| Curated | 8585 | 6692 | 155 | 1081 (12,6%) | 7504 (87,4%) |

To better characterise the curated **TB** dataset, we analysed the distribution of interactions across its protein targets. As shown in Figure 7, the dataset is markedly imbalanced, with a small subset of targets concentrating the majority of annotated interactions. For instance, the three most represented proteins (P9WPC5, P9WFT3, and P9WGR1) together account for over 20% of all drug–target pairs, while the majority of other targets are supported by only a limited number of interactions. This uneven distribution reflects the experimental focus historically placed on a few well studied proteins, leaving many potential targets relatively unexplored. Such skewness emphasises the dual challenges of training robust predictive models under class imbalance and the need for computational approaches to prioritise less studied targets that might hold therapeutic potential.

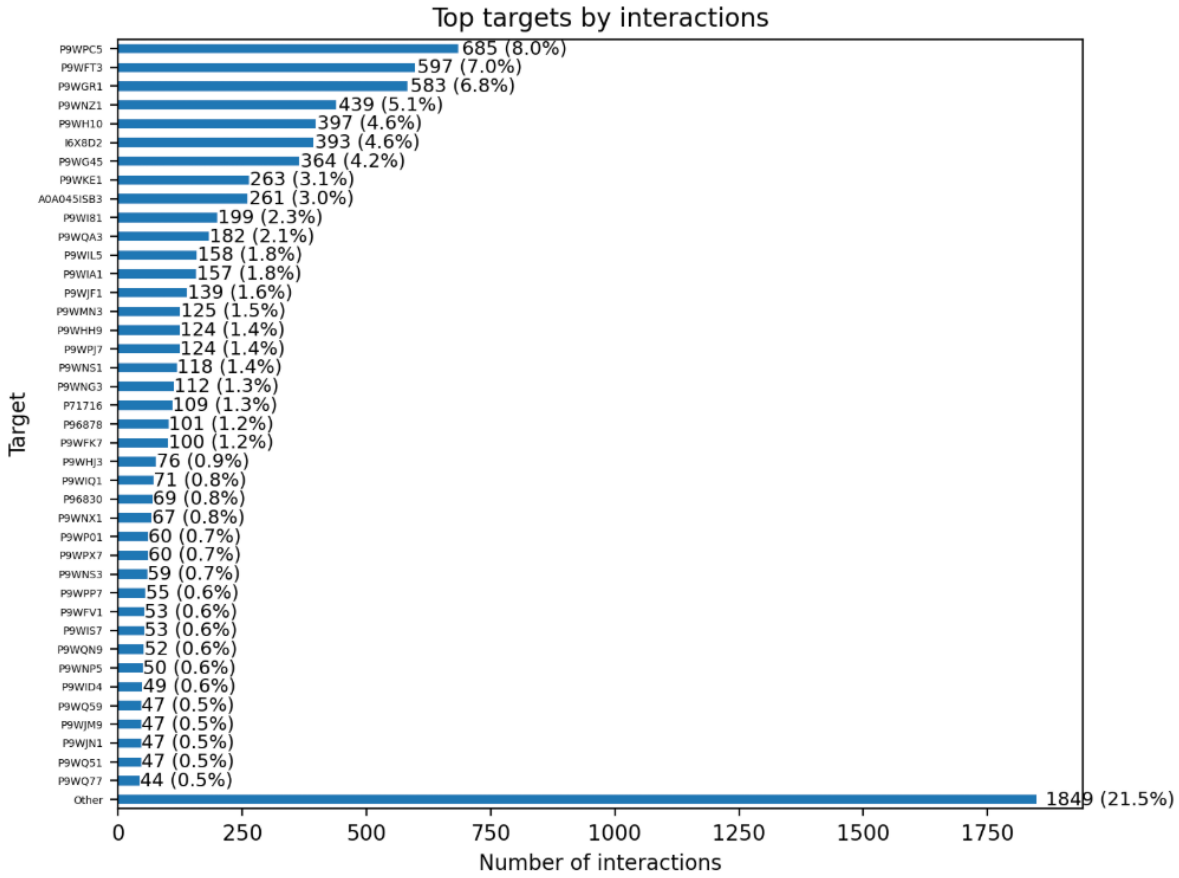



Figure 7: Targets ranked by number of interactions. The target P9WPC5 shows the highest number of interactions (685; 8.0%), followed by P9WFT3 (597; 7.0%) and P9WGR1 (583; 6.8%). All remaining targets were grouped under “Other” (1,849; 21.5%).

To further explore the structural diversity of the curated **TB** dataset, a **UMAP** projection was generated in which molecules were coloured according to their associated targets (Figure 8). The visualisation reveals that, despite the presence of multiple distinct protein classes, the majority of compounds occupy a highly

overlapping chemical space, with only a few scattered clusters corresponding to target specific chemotypes. This overlap suggests that different **TB** targets are frequently addressed by structurally related scaffolds, reflecting both the reuse of privileged chemical motifs in drug discovery and the inherent difficulty of designing selective inhibitors for enzymes with similar binding environments. The limited formation of well separated clusters highlights the chemical redundancy of the dataset and further illustrates why predictive models face challenges in discriminating activity across multiple targets within such a constrained chemical landscape.

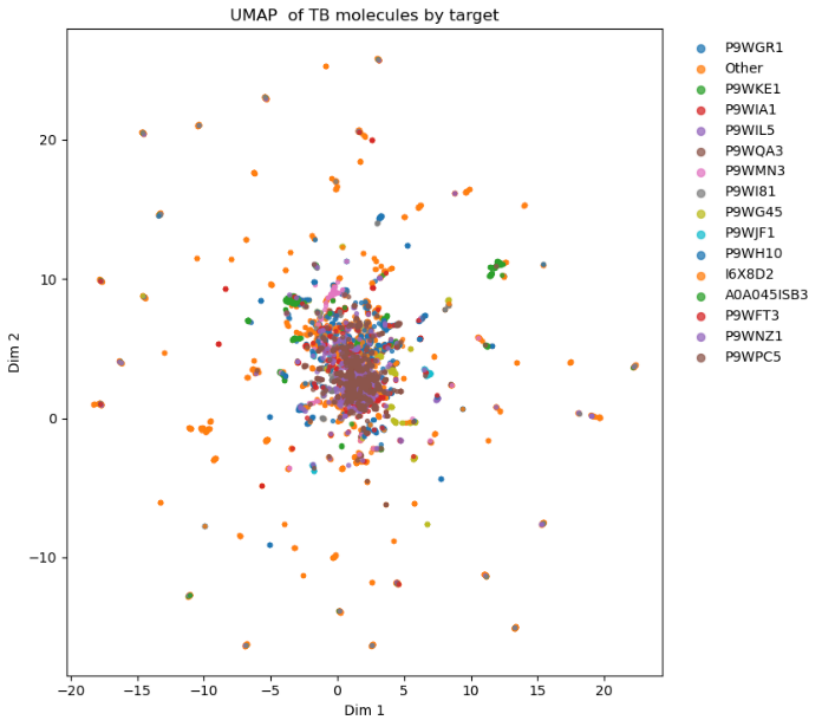



Figure 8: **UMAP** projection of **TB** related molecules colored by their associated protein targets. Each point represents a compound, with different colors corresponding to distinct targets. The plot illustrates the overlap and clustering of molecules across targets, highlighting the lack of strong separation in chemical space between compounds associated with different TB targets.

Finally, to assess the broader representativeness of the curated **TB** dataset, its chemical space was compared against that of the large-scale Papyrus collection using a **UMAP** projection (Figure 9). The analysis, based on Jaccard/Tanimoto similarity of molecular fingerprints, showed that **TB** compounds are not isolated from the Papyrus space but instead embedded within it. Rather than forming distinct clusters, the **TB** molecules appear as a compact and narrowly distributed region, highlighting their reduced chemical diversity compared to the full Papyrus set. This observation confirms that the curated **TB** dataset constitutes a focused subset of Papyrus: while it shares many structural features with the broader collection, it

also provides a more specific and disease-relevant chemical profile, tailored to tuberculosis-related targets.

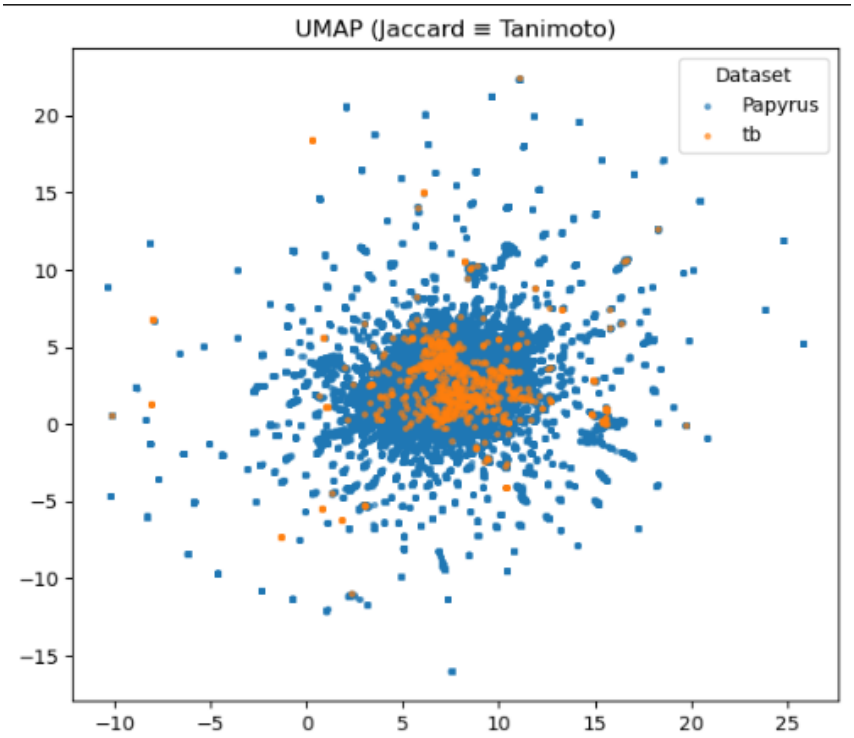



Figure 9: **UMAP** projection based on Jaccard/Tanimoto similarity of molecular fingerprints, comparing compounds from the **TB** dataset (orange) and the Papyrus dataset (blue).

## 4.3 Barlow Twins Model

Before applying the selected architectures to the selected data, an initial step consisted of reproducing the results reported by the original authors of the models. This validation step was critical to ensure methodological consistency and to confirm that our implementations accurately reflected the original studies.

For this purpose, the BindingDB dataset was used, replicating the same experimental conditions described in the reference publications. Table 5 summarises the performance achieved in this work compared to the theoretical values reported. The original results indicated a **ROC-AUC** of 0.9364 and a **PR-AUC** of 0.7344, while our experimental reproduction achieved a **ROC-AUC** of 0.9302 and a **PR-AUC** of 0.7119.

The close agreement between the reported and reproduced metrics demonstrates that the model implementation was consistent with the original description and that no significant deviations occurred during replication. This reproducibility provided a solid foundation for subsequent experiments, ensuring that any performance differences observed on **TB** related data could be attributed to dataset specific

characteristics rather than implementation inconsistencies.

Table 5: Replication of results obtained with the Barlow Twins model on the BindingDB dataset. Performance metrics are reported as ROC-AUC and PR-AUC for both the theoretical and experimental settings, confirming the robustness and reproducibility of the model.

| | **Dataset** | **ROC-AUC** | **PR-AUC** |
|---|---|---|---|
| Theorical | Bindingdb | 0.9364 | 0.7344 |
| Experimental | Bindingdb | 0.9302 | 0.7119 |

To further validate the implementation, the Barlow Twins model was trained under three different dataset configurations: the large scale Papyrus dataset, the combined Papyrus+**TB** dataset, and the **TB** dataset. The corresponding training and validation loss curves are shown in Figure 10. These experiments were designed to generate high quality feature embeddings that could be subsequently used by supervised classifiers. Moreover, they allowed us to investigate how dataset size, balance, and biological specificity influenced representation learning and, ultimately, the downstream classification performance in predicting **DTI**.

For the Papyrus dataset, the training loss decreased sharply during the initial epochs and quickly stabilised at low values. The validation loss initially followed this downward trend but diverged after approximately the $10^{th}$ epoch, plateauing at higher values. This behaviour is characteristic of overfitting: while the model continues to optimise its performance on the training set, it fails to achieve equivalent improvements on unseen data (validation set).

In the Papyrus+**TB** configuration, the general pattern remained similar, with steep reductions in both training and validation losses followed by stabilisation. However, the gap between the curves was narrower compared to Papyrus alone. The integration of **TB** specific interactions increased data diversity, effectively acting as a form of regularisation and partially mitigating overfitting. This demonstrates that the addition of domain specific information can balance the advantages of large scale data with improved generalisation.

The **TB** dataset, in contrast, presented the most challenging learning scenario. Both training and validation losses remained substantially higher than in the previous two cases, and their reduction was slower and less pronounced. The validation curve plateaued at elevated values, reflecting the reduced size and pronounced class imbalance of the dataset. Nevertheless, the relatively small gap between training and validation losses indicates that, although the model was less optimized overall, it also avoided the severe overfitting observed with Papyrus. Instead, the model followed a more conservative learning trajectory, constrained by the scarcity of data.

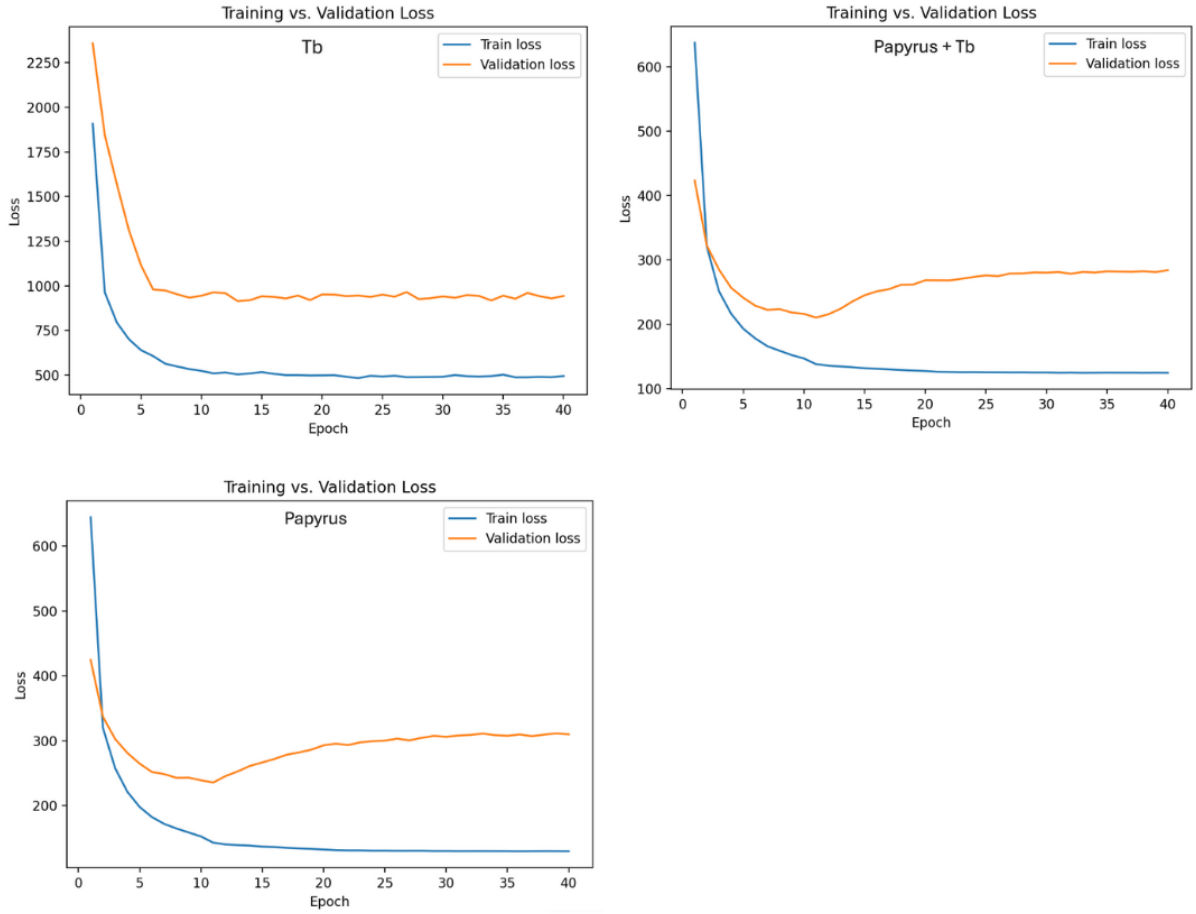



Figure 10: Training and validation loss curves for the Barlow Twins model under different dataset configurations. (Top left) Training using only the **TB** dataset; (Top right) training with the combined Papyrus + **TB** dataset; (Bottom) training using only the Papyrus dataset. The results show distinct convergence behaviours, with Papyrus and Papyrus + **TB** presenting lower and more stable validation losses compared to **TB** alone.

Building on the previous analysis, table 6 provides a comparative assessment of the different classifiers (XGBoost and linear **SVM**) and feature extraction strategies (**ECFP**s/**ESM2** vs. Barlow Twins) across the three dataset configurations: Papyrus, **TB**, and Papyrus+**TB**. Performance was evaluated using a previously defined independent test set, ensuring an unbiased assessment of generalisation capability. Results are reported across multiple evaluation metrics, including **ROC-AUC**, Accuracy, Precision, Recall, F1-score, and **MCC**.

For the Papyrus dataset, XGBoost consistently outperformed **SVM** across all metrics, regardless of the feature type. Both **ECFP**s/**ESM2** and Barlow Twins embeddings achieved strong results with XGBoost, yielding **ROC-AUC** values above 0.86 and F1-scores above 0.80. These results are consistent with the training and validation curves, where Papyrus provided a rich signal for rapid optimisation, albeit at the cost of overfitting. The weaker performance of **SVM**, particularly with **ECFP**s/**ESM2**, suggests that this method faces challenges in very high dimensional and sparse feature spaces, where capturing complex interaction patterns becomes more challenging.

In the **TB** dataset, the differences between classifiers and feature types were more pronounced. XGBoost with **ECFP**s/**ESM2** achieved the highest **ROC-AUC** (0.8718) and accuracy (0.8885), demonstrating the utility of handcrafted descriptors in low data regimes. By contrast, **SVM** performance dropped substantially, with **ROC-AUC** values below 0.80 in most cases and low F1-scores, reflecting the difficulty of learning from a imbalanced dataset. The overall lower scores compared to Papyrus align with the higher validation losses observed in the **TB** learning curves, which reflect the limited size and imbalance of this dataset.

The Papyrus+**TB** configuration provided the best compromise between scale and domain specific diversity. Here, Barlow Twins embeddings combined with XGBoost achieved the strongest overall performance, with a **ROC-AUC** of 0.8720, accuracy of 0.7867, and balanced precision (0.7909) and recall (0.8038). The **MCC** value of 0.5723 further supports the robustness of this configuration. These results align with the loss curves, where the integration of **TB** data acted as a natural regularizer, narrowing the gap between training and validation losses and producing more balanced generalization.

Overall, we can conclude that dataset scale and quality directly influence performance, Papyrus enables strong results but risks overfitting, **TB** alone is limited by scarcity and imbalance, while Papyrus+**TB** strikes a balance between the two. The choice of classifier plays a critical role. In our experiments, XGBoost generally outperformed **SVM**, particularly under low data conditions, likely due to its greater robustness in handling high dimensional and heterogeneous representations. Representation type has dataset dependent effects, Barlow Twins embeddings perform competitively and reach their best potential

when combined with Papyrus+**TB**.

Together with the loss analyses, these results demonstrate that large scale curated resources like Papyrus provide a strong foundation for training, but in our case, that **TB** knowledge is essential to achieve biologically relevant predictions. The combination of both datasets, coupled with powerful classifiers such as XGBoost, provides the most promising strategy for tuberculosis **DTI** prediction and repurposing.

Table 6: Comparative performance of different classifiers (XGBoost and **SVM**) and feature extraction strategies (**ECFP**s/**ESM2** and Barlow Twins) across three dataset configurations: Papyrus, Tuberculosis, and the combined Papyrus + **TB** dataset. Results are reported using multiple evaluation metrics, including **ROC-AUC**, Accuracy, Precision, Recall, F1-Score, and **MCC**, highlighting the impact of dataset choice and representation method on predictive performance.

| Dataset | classifier | Feature Type | ROC-AUC | Accuracy | Precision | Recall | F1-Score | MCC |
|---|---|---|---|---|---|---|---|---|
| **Papyrus** | XGBoost | ECFPs/ESM | 0.8659 | 0.7847 | 0.7790 | 0.8279 | 0.8024 | 0.5677 |
| | | Barlow Twins | 0.8709 | 0.7875 | 0.7941 | 0.8071 | 0.8005 | 0.5733 |
| | SVM | ECFPs/ESM | 0.7360 | 0.6260 | 0.5922 | 0.9383 | 0.7261 | 0.2901 |
| | | Barlow Twins | 0.7239 | 0.6702 | 0.6729 | 0.7313 | 0.7009 | 0.3363 |
| **Tuberculosis** | XGBoost | ECFPs/ESM | 0.8718 | 0.8885 | 0.6086 | 0.3456 | 0.4409 | 0.4032 |
| | | Barlow Twins | 0.8642 | 0.8893 | 0.6329 | 0.3086 | 0.4149 | 0.3903 |
| | SVM | ECFPs/ESM | 0.7977 | 0.2166 | 0.1396 | 1.0000 | 0.2450 | 0.1196 |
| | | Barlow Twins | 0.7960 | 0.8548 | 0.4109 | 0.3272 | 0.3643 | 0.2858 |
| **Papyrus + Tb** | XGBoost | ECFPs/ESM | 0.8665 | 0.7851 | 0.7780 | 0.8232 | 0.8000 | 0.5694 |
| | | **Barlow Twins** | **0.8720** | **0.7867** | **0.7909** | **0.8038** | **0.7973** | **0.5723** |
| | SVM | ECFPs/ESM | 0.7399 | 0.6304 | 0.5926 | 0.9334 | 0.7250 | 0.3045 |
| | | Barlow Twins | 0.7355 | 0.6805 | 0.6749 | 0.7482 | 0.7097 | 0.3589 |

In addition to the comparative results summarised in Table 6, Figure 11 presents the **ROC-AUC** and Precision Recall curves for the XGBoost classifier using Barlow Twins embeddings across the three dataset configurations: Papyrus, **TB**, and Papyrus+**TB**. These visualisations provide a more detailed perspective on the trade-offs between sensitivity and specificity, and on classifier performance under class imbalance.

For the Papyrus dataset, the **ROC** curve shows strong separability between active and inactive interactions, with an **AUC** of 0.871. The **PR** curve further supports this result, yielding an average precision of 0.877. These values indicate robust predictive power, consistent with the balanced class distribution and large scale of Papyrus.

In the **TB** dataset, performance is substantially lower. The **ROC-AUC** remains competitive (0.864), suggesting that the model is still able to distinguish between classes. However, the **PR** curve reveals an average precision of only 0.499, reflecting the strong class imbalance and reduced number of active samples. The discrepancy between **ROC** and **PR** performance highlights the limitations of **ROC** based metrics in imbalanced settings, reinforcing the relevance of **PR** curves as a more sensitive indicator of performance in **TB** data.

The Papyrus+**TB** dataset exhibits the most favourable balance. The **ROC-AUC** (0.872) is comparable to Papyrus, and the **PR** cruve (0.875) remains very high. This indicates that the integration of **TB** data did not compromise the separability achieved by Papyrus but, in fact, enhanced the relevance of predictions for imbalanced cases. The consistency between **ROC** and **PR** metrics confirms that the combined dataset successfully preserved generalisation while mitigating overfitting tendencies.

Together, these curves reinforce three central observations: Papyrus enables robust predictive performance, yet the models tend to overfit when not carefully regularised, highlighting the interplay between dataset complexity and modelling choices.

The **TB** data illustrate the dual challenges of scarcity and class imbalance: while **ROC** metrics may suggest acceptable performance, precision–recall curves more accurately expose the difficulty of reliably identifying active compounds.

Papyrus+**TB** offers the most balanced and reliable performance, combining the scale and balance of Papyrus with the disease specificity of the tuberculosis dataset.

These findings, consistent with both the training/validation loss curves and the comparative metrics in Table 6, confirm that the hybrid Papyrus+**TB** strategy provides the most promising foundation.

Complementing the results observed for XGBoost, Figure 11 shows the **ROC** and Precision Recall curves obtained with the **SVM** classifier across the Papyrus, tuberculosis, and Papyrus+**TB** datasets. In contrast, the results demonstrate that **SVM** struggles to achieve consistent predictive performance,

particularly under conditions of data scarcity and imbalance.

For the Papyrus dataset, the **ROC** curve yields an **AUC** of 0.724, while the **PR** curve shows an average Precision of 0.719. These values are considerably lower than those achieved with XGBoost, reflecting the difficulty when dealing with high-dimensional molecular embeddings and the intricate non linear relationships inherent to drug–target interaction data. This behaviour underscores the increased sensitivity of such classifiers to sparse and complex feature spaces, where optimisation becomes challenging without sufficient regularisation.

In the tuberculosis dataset, the **ROC-AUC** improves to 0.796, suggesting that the **SVM** can separate active from inactive cases to some degree. However, the **PR** curve reveals an **AUC** of only 0.341, underscoring the impact of class imbalance. The low precision across most recall levels indicates that the classifier produces a high number of false positives when applied to tuberculosis specific data, limiting its practical utility in this setting.

The Papyrus+**TB** configuration shows slightly better balance, with **ROC-AUC** 0.736 and AP 0.721. While still inferior to XGBoost, these results confirm that integrating tuberculosis specific interactions improves generalisation compared to Papyrus alone. Nevertheless, the gains remain modest, and the classifier remains less robust overall.

Overall, the results reveal some insights. First, **SVM** underperforms relative to XGBoost across all configurations, with the discrepancy being most pronounced in **PR** based evaluations, which are more sensitive to class imbalance. Second, the **TB** dataset accentuates these shortcomings, showing acceptable **ROC** scores but substantially weaker **PR** performance. Third, although the Papyrus+**TB** configuration yields the most balanced setting, the improvements remain insufficient to render **SVM** competitive.

This comparison confirms that while **SVM** provides a useful baseline, it lacks the capacity to model complex non-linearities in high dimensional feature spaces, making XGBoost the more reliable choice for tuberculosis focused **DTI** prediction.

Given that the overarching goal of this work was to predict novel drug–target interactions specifically for **TB**, the subsequent analyses focused on the tuberculosis subset of the combined Papyrus+**TB** configuration, the model that had previously demonstrated the most balanced and robust performance across datasets.

Table 7 summarises the performance metrics obtained when evaluating only the tuberculosis related interactions within the combined dataset. This targeted evaluation offers a more realistic measure of the model's applicability for antitubercular drug discovery, where data scarcity and class imbalance present significant challenges.

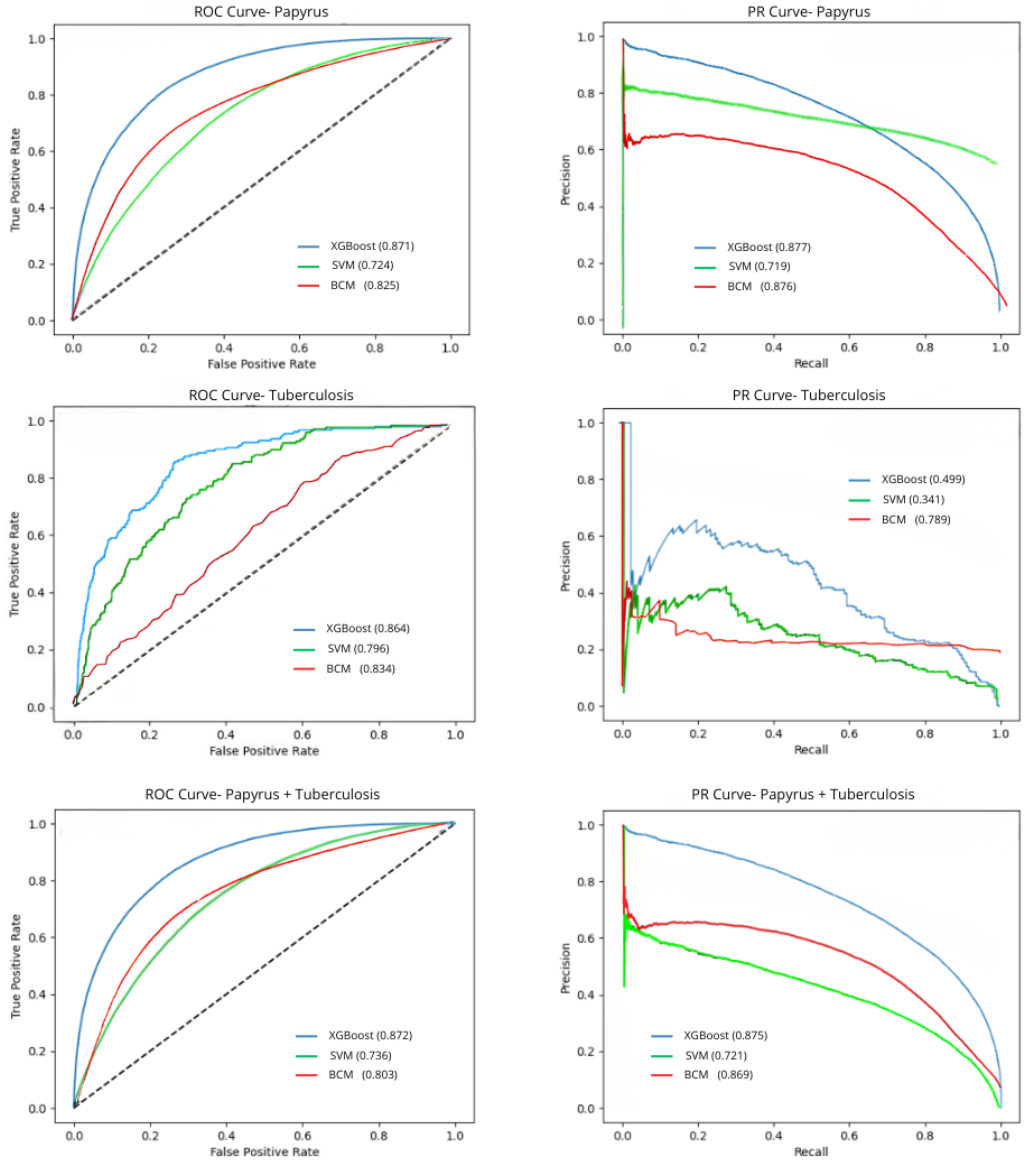



Figure 11: **ROC** and **PR** curves comparing different classifiers and datasets for d**DTI** prediction. Results are shown for XGBoost, **SVM**, and **BCM-DTI** models, evaluated on Papyrus, tuberculosis, and the combined Papyrus + **TB** datasets. The curves illustrate that XGBoost achieves the highest overall performance across datasets, while **SVM** displays lower discriminative ability. **BCM-DTI** shows intermediate performance, with Papyrus and Papyrus **TB** outperforming **TB** alone.

The results confirm that the XGBoost classifier, trained using Barlow Twins embeddings, retained a reasonable ability to discriminate between active and inactive interactions, achieving a **ROC-AUC** of 0.83 and an overall accuracy of 0.86. However, both precision (0.45) and recall (0.23) values indicate that, in this domain, the model tends to overpredict the positive class, generating a high number of false positives. This behaviour reflects the inherent difficulty of identifying truly bioactive compounds when active examples are under represented in the data.

The **SVM** model, by contrast, exhibited a more pronounced performance drop under the same conditions, with a **ROC-AUC** of 0.63 and extremely low recall (0.08), failing to capture most active interactions. This discrepancy reinforces the superior adaptability of XGBoost in data limited and imbalanced environments, where its ensemble based nature allows for more flexible decision boundaries.

Table 7: Performance metrics computed exclusively on the **TB** subset within the combined Papyrus + **TB** dataset. Results are reported for two classifiers: XGBoost with Barlow Twins features and **SVM** with Barlow features. The evaluation includes **ROC-AUC**, Accuracy, Precision, Recall, F1-Score, and **MCC**, focusing specifically on **TB** related interactions.

| Dataset | classifier | Feature Type | ROC-AUC | ACC | PR | Recall | F1-Score | MCC |
|---|---|---|---|---|---|---|---|---|
| **Papyrus + TB** | XGBoost | Barlow Twins | 0.8302 | 0.8654 | 0.4458 | 0.2270 | 0.3008 | 0.2514 |
| | SVM | Barlow | 0.6340 | 0.8592 | 0.3023 | 0.0798 | 0.1262 | 0.0978 |

## 4.4 BCM-DTI

The **BCM-DTI** model was also subjected to a reproducibility assessment to confirm the validity of its implementation. As in the original publication, the BindingDB dataset was used to replicate the reported results under comparable experimental conditions. Table 8 presents the performance values for both the theoretical (originally reported) and experimental (reproduced) settings, evaluated with **ROC-AUC**, Precision, and Recall.

The original study reported a **ROC-AUC** of 0.933, a Precision of 0.893, and a Recall of 0.684. Our experimental reproduction achieved a **ROC-AUC** of 0.908, a Precision of 0.880, and a Recall of 0.629. These results show a high degree of agreement with the theoretical values, with only minor reductions observed across metrics. The slight variations can be attributed to differences in dataset splitting, preprocessing procedures, or stochasticity inherent to model training.

Overall, the reproducibility experiments demonstrate that the **BCM-DTI** implementation is consistent with the original work and maintains robust performance on BindingDB. This validation provided confidence

that the model could be reliably applied and adapted to **TB** specific data in the subsequent stages of this study.

Table 8: Replication of results obtained with the **BCM-DTI** model on the BindingDB dataset. Performance metrics are reported as **ROC-AUC**, Precision and Recall for both the theoretical and experimental settings, confirming the robustness and reproducibility of the model.

| | **Dataset** | **ROC-AUC** | **PR** | **Recall** |
|---|---|---|---|---|
| Original | Bindingdb | 0.933 | 0.893 | 0.684 |
| Experimental | Bindingdb | 0.908 | 0.880 | 0.629 |

Figure 12 shows the training and validation loss trajectories of the **BCM-DTI** model across the three dataset configurations: tuberculosis, Papyrus, and Papyrus+**TB**. These curves provide important insights into the model's learning behaviour and highlight the influence of dataset scale and diversity on convergence patterns.

For the Tb dataset, the model converges extremely rapidly, with both training and validation losses stabilising within the first few epochs and remaining closely aligned throughout. This pattern suggests that the model can fit the limited dataset without signs of severe overfitting. However, the fast convergence and relatively high plateau of the validation loss indicate that the model's ability to capture deeper representations is constrained by the scarcity and imbalance of **TB** interactions.

The Papyrus dataset displays a more gradual and stable decrease in both training and validation losses. While the training curve continues to decline steadily across epochs, the validation curve flattens at slightly higher values, with modest fluctuations. This divergence reflects the richer variability of Papyrus, which provides sufficient signal for continued optimisation but also increases the likelihood of overfitting as the number of epochs grows.

In the Papyrus+**TB** configuration, the behaviour lies between the two extremes. Training and validation losses decrease in parallel during the early epochs, with only a modest gap maintained throughout the training process. The inclusion of **TB** interactions appears to add biological relevance without disrupting the optimisation stability afforded by Papyrus. This results in a balanced convergence, mitigating the underfitting observed in the **TB** only setting and the stronger overfitting characteristic of Papyrus alone.

In summary, the **BCM-DTI** loss curves highlight how dataset characteristics shape model optimisation. The **TB** dataset leads to rapid but shallow convergence, Papyrus enables deeper learning but carries overfitting risks, and Papyrus+**TB** offers the most balanced trajectory, combining large scale information with disease specific diversity.

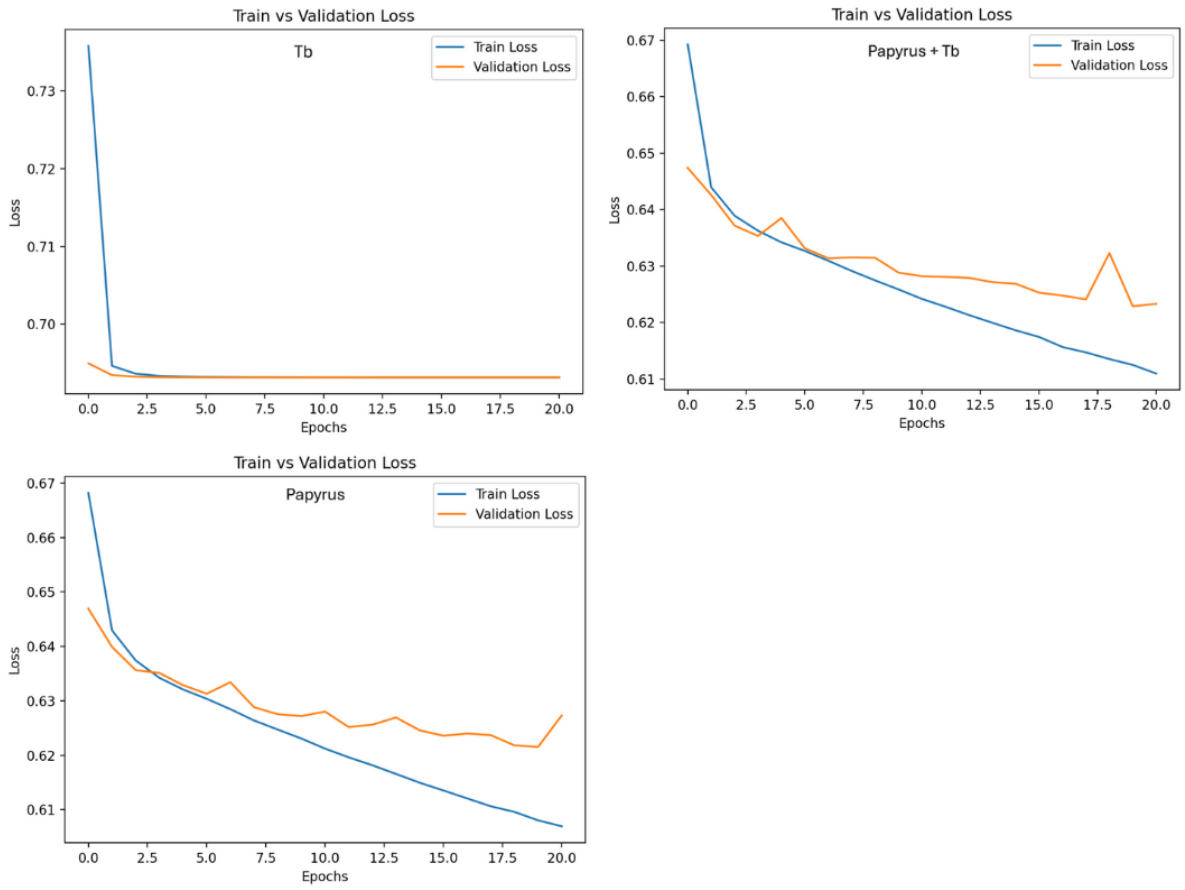



Figure 12: Training and validation loss curves for the **BCM-DTI** model under different dataset configurations. (Top left) Training with the Tuberculosis dataset; (Top right) training with the combined Papyrus+**TB** dataset; (Bottom) training with the Papyrus dataset alone. The loss trajectories indicate that Papyrus and Papyrus + **TB** provide smoother convergence and lower validation losses compared to the **TB** only setup.

To complement the loss curve analysis, the **BCM-DTI** model was systematically evaluated across four dataset configurations: Papyrus, Tuberculosis, Papyrus+**TB**, and Fine-Tuning. These experiments were designed to assess how dataset scale, diversity, and transfer learning strategies influence supervised model performance. The Fine-Tune approach initialised the model with weights trained on Papyrus and adapted it exclusively to **TB** data, testing the effectiveness of transfer learning.

Table 9 summarises the results using six evaluation metrics: **ROC-AUC**, Accuracy, Precision, Recall, F1-score, and **MCC**. The Papyrus+**TB** configuration achieved the strongest overall performance, with **ROC-AUC** of 0.8309, Accuracy of 0.8257, and **MCC** of 0.5058, confirming the advantage of combining scale with disease specific information. The Papyrus baseline also yielded competitive results, while the **TB** setting performed poorly due to limited size and imbalance. Interestingly, the fine-tuning strategy led to a significant increase in recall, but at the expense of substantial reductions in precision and overall balance. This indicates that the model, when exposed exclusively to the **TB** dataset, shifted to predict a higher number of positive interactions, reducing false negatives but increasing false positives. Such behaviour is in line with the mathematical relationship between Precision and Recall and highlights the model's difficulty in maintaining discriminatory power under data constrained conditions, demonstrating the challenges of effective knowledge transfer when the target domain is small and imbalanced.

Table 9: Performance metrics of the **BCM-DTI** model across different dataset configurations. Results are reported for Papyrus, Tuberculosis, Papyrus+**TB** and Fine-Tune experiments. Metrics include **ROC-AUC**, Accuracy, Precision, Recall, F1-Score, and **MCC**, providing a comprehensive evaluation of predictive performance.

| **Dataset** | **ROC-AUC** | **ACC** | **PR** | **Recall** | **F1-Score** | **MCC** |
|---|---|---|---|---|---|---|
| **Papyrus** | 0.8248 | 0.8213 | 0.8765 | 0.7087 | 0.7701 | 0.4962 |
| **Tuberculosis** | 0.8348 | 0.4163 | 0.7898 | 0.3271 | 0.3004 | 0.3975 |
| **Papyrus + Tb** | **0.8309** | **0.8257** | **0.8686** | **0.7131** | **0.7524** | **0.5058** |
| **Fine Tune** | 0.8168 | 0.6554 | 0.2505 | 0.8580 | 0.3504 | 0.3256 |

The following figure 11 illustrates the **ROC** and **PR** curves for the **BCM-DTI** model across the same dataset configurations presented in Table 9. These visualisations provide a detailed perspective on the classifier's discriminative ability and highlight the impact of dataset scale and domain specificity on predictive reliability.

For the Papyrus dataset, the **ROC** curve shows a clear separation between classes, while the **PR** curve indicates consistently high precision across a wide range of recall values. These results confirm that

**BCM-DTI** benefited from the scale and balanced distribution of Papyrus, achieving robust performance in line with the quantitative metrics.

In the **TB** configuration, performance degraded substantially. The **ROC** curve approached the diagonal, suggesting near random discrimination, and the PR curve revealed very low precision across all recall levels. This outcome aligns with the quantitative metrics and the loss curve analysis, emphasising that the scarcity and imbalance of tuberculosis-specific interactions severely limit the model's ability to generalise.

The Papyrus+**TB** configuration produced the most balanced outcome. The **ROC** curve achieved a high **AUC** comparable to Papyrus, while the **PR** curve maintained strong precision even as recall increased. This confirms that the hybrid dataset successfully combined the generalizability of Papyrus with the disease relevance of **TB**, mitigating overfitting and improving prediction stability.

Together, these curves reinforce the conclusions drawn from the quantitative results: Papyrus enables strong baseline performance, Tuberculosis is severely constrained by data scarcity, Papyrus+**TB** offers the most reliable and biologically relevant predictions, and fine-tuning introduces challenges that must be addressed in future work.

As was done for previous models, a final evaluation was performed focusing exclusively on the tuberculosis subset of the Papyrus+**TB** configuration. Table 10 reports the performance metrics obtained for this subset. The results show a marked reduction in predictive ability compared to the global Papyrus+**TB** evaluation, with **ROC-AUC** and accuracy values around 0.48, and a low precision of 0.12 despite a moderate recall of 0.49. This indicates that, although the model still identifies some true positives, it produces a substantial number of false positives, an expected outcome given the imbalance of tuberculosis data.

These findings highlight the challenges of applying broad, data rich models to highly specific biomedical contexts. Despite the benefits of large-scale pretraining, the embeddings derived from Papyrus do not fully capture the structural and biochemical nuances characteristic of tuberculosis targets. Nonetheless, the model's relatively high recall remains useful for early-stage virtual screening, where the priority is to recover potentially active compounds rather than exclude all inactives.

Table 10: Performance metrics computed exclusively on the tuberculosis subset within the combined Papyrus+**TB** dataset. Results are reported for **BCM-DTI** model. The evaluation includes **ROC-AUC**, Accuracy, Precision, Recall, F1-Score, and **MCC**, focusing specifically on **TB** related interactions.

| **Dataset** | **ROC-AUC** | **ACC** | **PR** | **Recall** | **F1-Score** | **MCC** |
|---|---|---|---|---|---|---|
| **Papyrus + Tb** | 0.4765 | 0.4781 | 0.1205 | 0.4908 | 0.1935 | 0.0220 |

The results obtained in this study demonstrate that the prediction of **TB** drug–target interactions re-

mains a challenging task. Both the Barlow Twins and **BCM-DTI** models achieved competitive performance when trained on large scale or hybrid datasets, but their predictive ability decreased substantially when restricted to **TB** data. This outcome is not unexpected: the pronounced class imbalance and heterogeneous experimental annotations, rising from the use of multiple assay types, measurement scales (e.g., $IC_{\text{e.g., } IC_{50},\, EC_{50},\, K_i,\, K_d}$), and activity thresholds across different sources and targets places strong limitations on the ability of **ML** and **DL** methods to generalise. In this sense, the imperfect performance of the models should be understood as a direct reflection of the constraints imposed by the available data rather than as a failure of the modelling approaches themselves.

Across all experiments, we can concluded that models for **TB DTI** prediction are far from perfect. Strategic use of complementary modelling approaches and hybrid datasets can partly overcome these challenges, producing results that, while not definitive, represent a significant step forward in the application of computational methods to tuberculosis drug repurposing. In our case, both the Barlow Twins and **BCM-DTI**, trained in Papyrus+**TB**, emerge as the most promising models.

## 4.5 Testing with Alternative Compounds (Drug Repurposing)

To initiate the drug repurposing analysis, two complementary evaluation strategies were implemented. The first focused on the ten **TB** related molecular targets most frequently reported in the literature. The second strategy concentrated exclusively on InhA, the primary target of **INH**. For both approaches, all possible drug–target pairs were generated using the Papyrus compound space and evaluated through three predictive models, Barlow Twins with XGBoost, Barlow Twins with **SVM**, and **BCM-DTI**, each assigning a probability score representing the likelihood of interaction. To identify the most reliable candidates, probability thresholds of 95% and 99% were applied, distinguishing broader predictions from high-confidence interaction sets.

### Filtering for 10 targets

When applying the 95% probability cut-off across the ten most important **TB** targets, the predictive pipeline retained a substantial number of candidate compounds. In total, 5,786 unique drug–target pairs were identified, reflecting a broad chemical space of potential interactions.

To ensure robustness, only the interactions that simultaneously exceeded the probability threshold in at least two out of the three predictive models (Barlow Twins + XGBoost, Barlow Twins + **SVM**, and **BCM-DTI**) were retained for further consideration. This cross model consensus filtering reduced the likelihood

of spurious predictions and strengthened the confidence in the retained candidates.

This large pool highlighted the diversity of compounds with strong predicted activity according to the models, providing a wide foundation for subsequent filtering and prioritisation.

Increasing the cut-off to 99% drastically reduced the number of surviving candidates, narrowing the results to 317 drug–target pairs. This reduction represents the shift from a sensitive to a highly specific strategy, where only the interactions with the strongest computational support are preserved. Despite the smaller scale, the retained molecules represent highly reliable predictions.

### Filtering for InhA

For the second approach, focusing exclusively on InhA, the application of the 95% probability threshold resulted in the identification of 974 candidate molecules. Importantly, this outcome suggests that the models were capable of capturing a rich variety of chemotypes potentially capable of modulating InhA activity.

When the analysis was restricted to the 99% probability threshold for InhA, the number of candidates decreased sharply to 30 molecules. This result highlights the selectivity of the predictive models when only the highest-confidence interactions are considered. Although the chemical diversity is reduced, the reliability of these predictions is greatly enhanced. The 30 surviving compounds thus represent a highly curated list of promising candidates for further computational or experimental validation, with potential to complement or even inspire alternatives to **INH**.

Together, these analyses demonstrate the power of combining **ML** and **DL** models for drug repurposing in **TB**. While the broader thresholds (95%) enable the exploration of a large and chemically diverse set of candidates, the stricter thresholds (99%) ensure a focused selection of highly reliable compounds. Importantly, both strategies revealed molecules with strong predicted interactions against clinically validated targets, such as InhA. The dual approach, therefore, provides complementary perspectives: wide exploration for hypothesis generation and narrow prioritisation for downstream validation.

## 4.6 Automated filtering Pipeline

The introduction of this automated pipeline thus transformed the large and heterogeneous list of predicted interactions into a much smaller, interpretable, and practically useful collection of compounds. By integrating structural rules, physicochemical thresholds, and medicinal chemistry expertise, the framework ensured that the candidates selected were not only computational predictions but also aligned with the

practical requirements of real world drug discovery.

After obtaining the high confidence predictions for drug–target interactions, it was necessary to subject these results to a rigorous filtering process to ensure that the final candidates were not only computationally promising but also chemically viable and compatible with established principles of medicinal chemistry as we can see on figure 13.

### Filtering for 10 targets

When this pipeline was applied to the predictions obtained for the ten most relevant **TB** targets, the results revealed a marked reduction in the number of candidates as the filtering process advanced. At a probability threshold of 95%, the process started with 5,786 drug–target pairs for the 10 selected targets, which were then sequentially filtered until only 969 survived all stages. The filter that contributed most to this reduction was the basic medicinal chemistry filter, which eliminated 1,559 molecules, primarily due to the presence of reactive halogens or other chemically undesirable groups. This indicates that even when computational models predict high probabilities of interaction, a large proportion of the candidate molecules contain chemical moieties incompatible with safe or stable drug design. Another stage with a strong impact was the physicochemical filter, which removed 1,366 molecules, most of them due to excessive rotatable bonds or polar surface area, properties that compromise oral bioavailability. At the stricter threshold of 99%, the pipeline began with 317 molecules and concluded with only 14 survivors. Again, the most significant reduction occurred in the basic medicinal chemistry filter, which eliminated 157 molecules, followed by the physicochemical filter, which discarded 87 additional compounds. These results highlight how chemical reactivity and poor physicochemical balance are the most common issues in high-probability candidates, strongly limiting their progression to the final stages.

### Filtering for InhA

The same analysis was carried out for the isoniazid target, InhA. At 95%, 974 predicted interactions were initially obtained, but after passing through all filters, only 133 molecules remained. In this case, the most restrictive filter was again the physicochemical filter, which eliminated 266 compounds, followed by the basic medicinal chemistry filter, which excluded 225 molecules. This reinforces the idea that many molecules predicted for InhA present unfavourable polarity or flexibility, making them less compatible with the requirements of drug-likeness. Under the 99% threshold, the process started with 30 molecules and was reduced to a single final candidate. At this level, the largest reduction occurred during the physicochemical filter, which eliminated 7 molecules, and the Lipinski filter, which removed 5 additional

candidates. Despite the reduced scale, the reasons for elimination were consistent with those observed in the broader analyses: excessive rotatable bonds, poor polarity balance, or violations of established medicinal chemistry rules.

The overall results of this filtering step reveal a clear trade-off between inclusiveness and selectivity. Broader thresholds allow for the retention of a larger number of molecules, preserving chemical diversity and increasing the chances of identifying unexpected repurposing opportunities. On the other hand, stricter thresholds drastically reduce the number of candidates, focusing on a very limited set of compounds but ensuring higher confidence in their viability. In all cases, the filters that eliminated the largest number of molecules were those associated with medicinal chemistry rules and physicochemical constraints, underlining the importance of these criteria in distinguishing theoretically active molecules from practically usable drug candidates.

In conclusion, the filtering process was decisive in transforming the large number of computational predictions into a much smaller and more reliable set of candidates. By integrating structural validation, medicinal chemistry principles, physicochemical balance, and synthetic feasibility, the pipeline ensured that only the most promising molecules were retained. This dual perspective, with results analysed at both 95% and 99% thresholds, highlights not only the robustness of the strategy but also the inherent challenges of drug repurposing in tuberculosis, where the scarcity and imbalance of available data make the search for viable candidates particularly demanding.

## 4.7 Explainability Analysis with SHAP

To interpret the XGBoost classifier, SHAP values were computed on the concatenated representations produced by the pipeline, for global importance we used the mean absolute SHAP.

The analysis revealed the 20 most influential **ECFP** bits. The corresponding mean absolute **SHAP** values are reported in figure 14, where higher scores reflect stronger contributions to the model's predictions.

The interpretation of the highlighted bits, rich in hydrogen bond donors and acceptors (hydroxyls, Bit 49, 106, 223, amides, Bit 218 and carbonyls, Bit 100, 123, 39, 218, 71), nitrogen containing heterocycles (Bits 71, 58, 6, 61, 160), and hydrophobic or halogenated substituents (Bits 137, 91, 72, 156), is consistent with known patterns of molecular recognition in **MTB** drug targets [113].

In general, hydrogen bond networks and polar contacts underpin ligand-protein affinity and selectivity, while donor and acceptor distribution strongly influences permeability and solubility, which are key

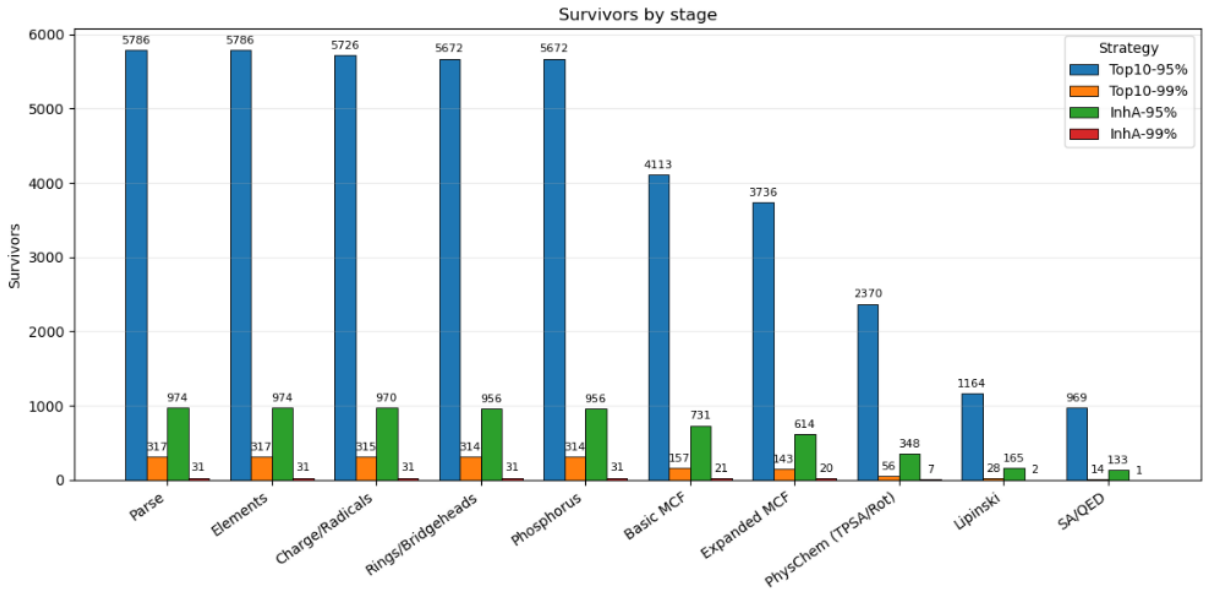



Figure 13: Number of compounds retained ("survivors") at each filtering stage for different selection strategies. The figure compares the evolution of compound counts across physicochemical and structural filters (e.g., rings, radicals, phosphorus, Lipinski, and **SA**/**QED**) for the Top10–95%, Top10–99%, InhA–95%, and InhA–99% strategies. The progressive reduction highlights the impact of increasingly stringent filtering criteria on the candidate space.

determinants in the optimisation of anti-tuberculosis drugs [114].

On the other hand, the recurrence of nitrogenous heterocycles in the model bits is consistent with the enriched landscape of anti-**TB** chemotypes based on heteroaromatic rings capable of mimicking cofactors (e.g., **Nicotinamide Adenine Dinucleotide (NAD)**$^+$/**NAD**H) or anchoring in polar pockets of active sites. [115].

Sulfur and fluorine containing compounds play important roles in tuberculosis therapy, with hydrophobic substitutions and halogenation patterns contributing to effective drug binding and activity [116].

Finally, the alkyl/aliphatic fragments and halogenation suggested by the model make sense in view of **MTB** extremely lipid rich cell wall, which imposes a permeability barrier and favours ligands with a hydrophobic surface sufficient to cross/root themselves in lipid microenvironments. Thus, the combination of polar groups (for specific anchoring) with well distributed hydrophobic mass (for permeation and accommodation in cavities) is a recurring strategy in the design of anti **TB** drugs [117].

After identifying the top 20 most influential **ECFP** bits, we next examined the molecules that passed all filtering stages to determine whether any of them contained these high importance fragments. The objective was to assess whether the prioritised candidates shared substructural features consistently recognised by the model as predictive of activity.

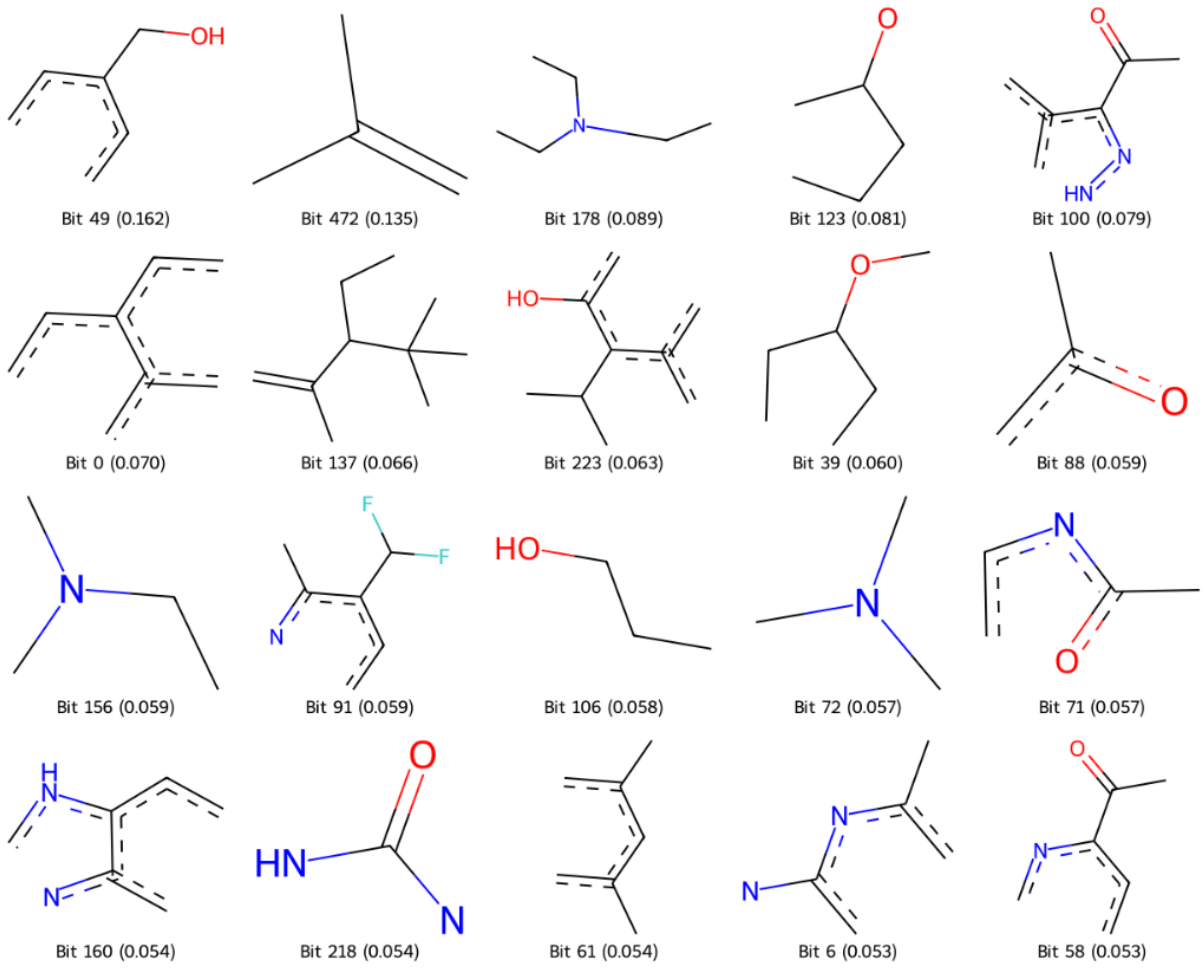



Figure 14: Structural fragments corresponding to the 20 most influential **ECFP** bits identified through **SHAP** analysis of the XGBoost classifier. The importance of each bit was quantified using mean absolute **SHAP** values, where higher values reflect a stronger contribution to the model's predictions.

From the 14 compounds retained in the strategy based on the top 10 **TB** related targets, 3 molecules were found to contain at least one of the top 20 **SHAP** bits. In addition, it was found that the retained compound from the **INH** strategy also contained bits from the top 20 as we can see in figure 15.

**Compound 1 (CHEMBL3646837)**: contained bit 223, a nitrogen heterocycle fragment;

**Compound 2 (CHEMBL1778675)**: triggered bit 160, a sulfur containing fragment;

**Compound 3 (CHEMBL4519484)**: triggered bit 160, a sulfur containing fragment;

**Compound 4 (CHEMBL4566916)**: activated bits 49 and 61, oxygenated sulfur fragments.

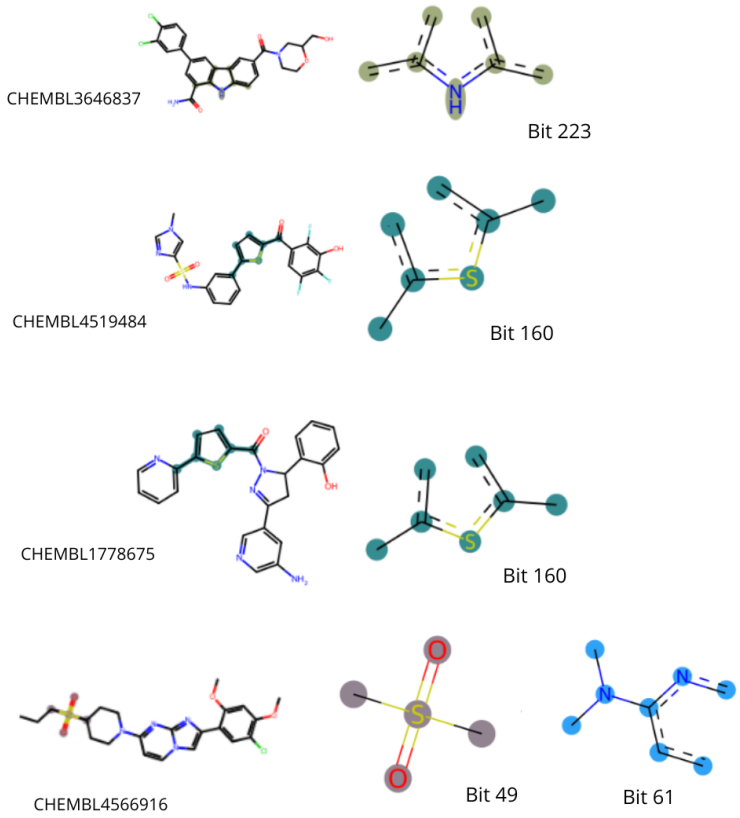



Figure 15: Retained compounds containing top 20 **SHAP** bits. CHEMBL3646837 (bit 223), CHEMBL1778675 (bit 160), CHEMBL4519484 (bit 160), and CHEMBL4566916 (bits 49 and 61).

These observations suggest that, although the majority of filtered candidates did not overlap with the most discriminative **SHAP** features, a subset of molecules presented substructures already prioritised by the model as relevant for activity prediction. This overlap provides an additional layer of confidence in their selection, highlighting potential structural motifs that may underpin the model's decision making process.

## 4.8 Validation of compounds through literature

To probe the biological plausibility of the candidates prioritised by the model, we examined in detail the top four molecules that passed all downstream filters (physicochemical, RDKit validity, fragment sanity checks, and **SHAP** based bit presence). For each compound we contrasted the predicted **MTB** target with literature and ChEMBL annotations.

For compound 1 (CHEMBL3646837), the predicted target is P9WNG3 (FabH ($\beta$-ketoacyl-ACP synthase III), obtained a prediction of 0.990 by our model. Primary literature associates this scaffold with JAK2 inhibition in oncology (carbazole/carboline "hinge binder" profile) [118], with no record of antimycobacterial activity nor evidence for FabH inhibition.

FabH accepts acyl-ACP/CoA-like substrates in a polar/catalytic cavity (Cys–His–Asn triad), effective inhibitors typically mimic acyl intermediates or carry acidic/anion stabilising groups [119]. CHEMBL3646837 is planar, lipophilic and bulky, optimised for an ATP pocket rather than an acyl tunnel, and lacks the polar handles expected for FabH [118]. While some flat aromatics have been reported against FabH, the overall pocket chemotype mismatch suggests a low likelihood of true FabH inhibition absent docking [120].

For Compound 2 (CHEMBL4519484), predicted target is P9WNS3 (DXS, 1-deoxy-D-xylose-5-phosphate synthase), with a model score of 0.992. Literature reports this scaffold as a potent 17$\beta$-HSD2 inhibitor (endocrine/osteoporosis context), with no evidence of antimycobacterial activity [121].

DXS requires a $Mg^{2+}$ dependent active site, favouring highly polar phosphorylated intermediates, effective inhibitors are typically TPP mimetics or phosphate analogues [122]. CHEMBL4519484 is a pyrimidinone sulfonamide bearing hydrophobic and aromatic substituents but lacking phosphate or acidic headgroups required for substrate mimicry [121]. This chemotype mismatch strongly limits the likelihood of direct DXS inhibition, unless binding occurs at a yet undescribed allosteric pocket.

Compound 3 (CHEMBL1778675), predicted target is P9WIL5 (PanC, pantothenate synthetase), with a model score of 0.994. The associated series is documented as a B-Raf (V600E) kinase inhibitor in oncology, with no antimycobacterial link [123].

PanC catalyses adenylation chemistry, typically inhibited by acyl-adenylate mimics (e.g., acyl-AMS, sulfamoyl-adenylates) that reproduce the AMP like scaffold and polar contacts [124]. CHEMBL1778675 is a $\pi$-rich (rich in aromatic or unsaturated regions), lipophilic kinase-like hinge-binder lacking AMP-mimetic motifs and polar functionality [123]. Given this discrepancy, the probability of true PanC inhibition is very low, despite the superficial presence of heteroaromatic motifs.

Compound 4 (CHEMBL4566916) is predicted to target P9WGR1 (InhA, enoyl-ACP reductase), with a

model score of 0.995. Public data associate this structure with inhibition of the histone methyltransferase SUV39H2, although the imidazo[1,2-a]pyridines present in CHEMBL4566916 are a chemotype with precedents in drug discovery **TB** (notably against QcrB) [125], there are no reports of anti-**TB** activity of this compound [126].

The InhA cofactor coupled pocket favours ligands that mimic NAD(H) interactions or form strong hydrogen bonds near the nicotinamide subsite [127]. CHEMBL4566916 is predominantly hydrophobic, with only a sulfone as a polar group, and lacks resemblance to canonical InhA pharmacophores [126]. Consequently, although the chemotype family is relevant in **TB** medicinal chemistry, this specific molecule shows poor fit for InhA and would require structural validation to support inhibition.

All four molecules achieved very high scores in the model, and two featured fragments highlighted by **SHAP** (sulphone/sulfoxide and N-rich heterocycles) that the classifier associates with "active" results. However, cross-checking with primary literature and ChEMBL annotations shows that these are human-targeted oncological/epigenetic structures with no documented activity on the predicted **MTB** enzymes; Furthermore, their chemotypes poorly match the substrate/cofactor characteristics and chemistry of the FabH, DXS, PanC, and InhA cavities, respectively. This tension between statistical signals and mechanistic plausibility is consistent with our broader conclusion that the **TB** dataset is highly unbalanced and chemically specialised (**UMAP** analysis), which increases the risk of overconfident extrapolations from patterns learned in a broader chemical space.

In summary, although the pipeline successfully identified high-scoring candidates, a mechanistic review suggests that three compounds may have limited likelihood of acting on their predicted **MTB** targets, while the fourth would benefit from structural validation. Nevertheless, these findings highlight promising chemotypes that merit further investigation, and additional docking and experimental validation studies will be essential to confirm their potential inhibitory activity.

# Chapter 5

# Conclusions and future work

## 5.1 Conclusions

This work demonstrated the potential of integrating **ML** and **DL** strategies to accelerate the discovery of novel therapeutic targets and compounds for **MTB**. Through the combination of large-scale data integration, **SSL**, and predictive modelling, the study proposed a pipeline capable of identifying candidate drugs and providing interpretable insights into **DTI**.

The Barlow Twins model proved effective in learning molecular and protein embeddings, while the **BCM-DTI** architecture achieved consistent predictive performance across multiple datasets. Furthermore, the integration of traditional classifiers such as XGBoost allowed a direct comparison between descriptor-based and learned molecular features. When evaluated on the combined Papyrus+**TB** dataset, the models achieved strong results, confirming that data augmentation and heterogeneous information contribute to improved generalisationn.

However, when the models were applied exclusively to the **TB** specific dataset, the predictive metrics, particularly **ROC-AUC**, F1-score, and **MCC**, were considerably lower. This decline in performance highlights important limitations associated with the available **TB** data. This dataset contained fewer samples, imbalanced activity labels, and high redundancy among molecular structures, leading to limited variability and poor model discrimination capacity. n addition, many **TB** entries lacked quantitative binding affinities and were instead annotated as binary activity labels (active/inactive) derived from screening data, typically based on standard potency thresholds. This reduction in data resolution further limited the granularity of learning. These factors collectively explain the reduced reliability and statistical robustness of **TB** specific predictions.

Despite these constraints, the automated filtering and post-processing pipeline enabled the identification of chemically viable and drug compounds, while the SHAP analysis provided a transparent view of the molecular features driving predictions. The validation of selected compounds through literature evidence

reinforced the practical relevance of the proposed models. While several candidates were unlikely to act on their predicted targets, others showed reported anti-tubercular activity, underscoring the models' potential to highlight chemically relevant scaffolds even when target-level correspondence remains uncertain.

Overall, this dissertation contributes to the computational exploration of **MTB** drug discovery by integrating explainability, data-driven modelling, and automated filtering into a cohesive and reproducible framework. It also emphasises the importance of data quality and representation balance in the performance of deep learning models applied to **DTI** prediction.

## 5.2 Future Work

Future research should aim to improve the reliability and biological relevance of specific predictions for tuberculosis. Based on the limitations identified in this study, several directions are proposed:

**Data enrichment and curation** — It will be crucial to expand the **TB** dataset by incorporating additional experimentally validated binding affinities and well-characterised negative interactions, to complement the currently available experimental data and improve statistical balance. Incorporating additional public resources, such as affinity data or activity values, could reduce label imbalance and increase statistical representativeness.

**Toxicity prediction and filtering**— While the current post-processing filters address chemical feasibility and binding affinity, future iterations should incorporate predictive toxicity and pharmacokinetic models. This integration will help ensure that prioritised compounds not only exhibit strong target engagement but also display favourable safety and bioavailability profiles, ultimately increasing the translational relevance of the predictions.

**Experimental and in silico validation** — Molecular docking simulations on the highest-scoring compounds against **MTB** targets will provide structural confirmation of predicted interactions, helping to prioritise candidates for in vitro validation. This combined approach will validate binding poses, estimate interaction energies, and support the experimental design needed to confirm inhibitory activity and refine future model iterations.

By addressing these aspects, future work will enhance the reliability, biological interpretability, and translational potential of the proposed approach. Incorporating toxicity filters and molecular docking validation will move the workflow closer to real world drug discovery, providing a crucial link between computational modeling and experimental pharmacology for combating multidrug resistant **TB**.

# Attachments

Table 11: List of the ten high priority **MTB** targets selected for this study, along with their biological roles. The selected proteins include well established therapeutic targets involved in essential biosynthetic and metabolic pathways such as cell wall formation, nucleotide synthesis, and protein translation.

| **Target Name** | **Role** |
|---|---|
| Pantothenate synthetase [128] | Catalyses the ATP dependent condensation of pantoate and -alanine to form pantothenate |
| Arabinosyltransferase A [129] | Critical role in the biosynthesis of the mycobacterial cell wall |
| DNA-directed RNA polymerase beta chain [104] | Essential enzyme for DNA transcription in RNA |
| Thymidylate kinase [130] | Biosynthesis of thymidine nucleotides |
| 3-oxoacyl-[acyl-carrier-protein] synthase 3[131] | Catalyst for the condensation reaction between long-chain acyl-CoA and malonyl-ACPshipping container |
| 3-oxoacyl-[acyl-carrier-protein] synthase 2 [132] | Biosynthesis of mycolic acids, essential components of the cell wall |
| 3-oxoacyl-[acyl-carrier-protein] synthase 1 [133] | Biosynthesis of mycolic acids, essential components of the cell wall |
| Cell division protein FtsZ [134] | Crucial for the dynamics of the Z ring |
| 1-deoxy-D-xylulose-5-phosphate synthase [122] | Biosynthesis of isoprenoids, essential components for the survival of the bacteria |
| Peptide deformylase[135] | Critical role in bacterial protein synthesis |

# Bibliography


[1] Muhammad M. Ibrahim, Tom M. Isyaka, Umoru M. Askira, Jidda B. Umar, Mustafa A. Isa, Adam Mustapha, and Akbar Salihu. Trends in the incidence of rifampicin resistant mycobacterium tuberculosis infection in northeastern nigeria. *Scientific African*, 17:e01341, 9 2022. ISSN 24682276. doi: 10.1016/j.sciaf.2022.e01341.

[2] *Global Tuberculosis Report 2024*. World Health Organization, 2024. ISBN 9789240101531.

[3] JoAnne L. Flynn and John Chan. Immune cell interactions in tuberculosis. *Cell*, 185:4682–4702, 12 2022. ISSN 00928674. doi: 10.1016/j.cell.2022.10.025.

[4] Zixuan E, Guanyu Qiao, Guohua Wang, and Yang Li. Gsl-dti: Graph structure learning network for drug-target interaction prediction. *Methods*, 223:136–145, 3 2024. ISSN 10462023. doi: 10.1016/j.ymeth.2024.01.018.

[5] Lijun Cai, Jiaxin Chu, Junlin Xu, Yajie Meng, Changcheng Lu, Xianfang Tang, Guanfang Wang, Geng Tian, and Jialiang Yang. Machine learning for drug repositioning: Recent advances and challenges. *Current Research in Chemical Biology*, 3:100042, 2023. ISSN 26662469. doi: 10.1016/j.crchbi.2023.100042.

[6] Wouter Deelder, Sofia Christakoudi, Jody Phelan, Ernest Diez Benavente, Susana Campino, Ruth McNerney, Luigi Palla, and Taane G. Clark. Machine learning predicts accurately mycobacterium tuberculosis drug resistance from whole genome sequencing data. *Frontiers in Genetics*, 10, 9 2019. ISSN 1664-8021. doi: 10.3389/fgene.2019.00922.

[7] Keunsoo Kang, Joyce Ho, Bonggun Shin, Sungsoo Park, and Joyce C Ho. Self-attention based molecule representation for predicting drug-target interaction. *Proceedings of Machine Learning Research*, 106:1–18, 2019. doi: 10.48550/arXiv.1908.06760. URL https://www.researchgate.net/publication/335258141.

[8] Giovanni Sotgiu, Philippe Glaziou, Charalambos Sismanidis, and Mario Raviglione. *Tuberculosis Epidemiology*, pages 229–240. Elsevier, 2017. doi: 10.1016/B978-0-12-803678-5.00507-5.

[9] Iñaki Comas, Mireia Coscolla, Tao Luo, Sonia Borrell, Kathryn E Holt, et al. Out-of-africa migration and neolithic coexpansion of mycobacterium tuberculosis with modern humans. *Nature Genetics*, 45:1176–1182, 10 2013. ISSN 1061-4036. doi: 10.1038/ng.2744.

[10] Maxime Barbier and Thierry Wirth. The evolutionary history, demography, and spread of the mycobacterium tuberculosis complex. *Microbiology Spectrum*, 4, 8 2016. ISSN 21650497. doi: 10.1128/microbiolspec.tbtb2-0008-2016.

[11] Caroline Weekes and Lakshmi P. Kotra. *Mycobacterium Tuberculosis Infections*, pages 1–7. Elsevier, 2007. doi: 10.1016/B978-008055232-3.60889-X.

[12] Ahmed Abduljabbar Jaloob Aljanaby, Qassim Muhsin Hashim Al-Faham, Israa Abduljabbar Jaloob Aljanaby, and Thualfakar Hayder Hasan. Epidemiological study of mycobacterium tuberculosis in baghdad governorate, iraq. *Gene Reports*, 26, 3 2022. ISSN 24520144. doi: 10.1016/j.genrep.2021.101467.

[13] Deepak Vats, Geeta Rani, Alisha Arora, Vidushi Sharma, Isha Rathore, Shaikh Abdul Mubeen, and Archana Singh. Tuberculosis and t cells: Impact of t cell diversity in tuberculosis infection. *Tuberculosis*, 149:102567, 12 2024. ISSN 14729792. doi: 10.1016/j.tube.2024.102567.

[14] Brenda Sáenz, Rogelio Hernandez-Pando, Gladis Fragoso, Oscar Bottasso, and Graciela Cárdenas. The dual face of central nervous system tuberculosis: A new janus bifrons?, 3 2013. ISSN 14729792.

[15] Reinout Van Crevel, Tom H.M. Ottenhoff, and Jos W.M. Van der Meer. Innate immunity to mycobacterium tuberculosis, 2002. ISSN 08938512.

[16] Avinash Khadela, Vivek P. Chavda, Humzah Postwala, Yesha Shah, Priya Mistry, and Vasso Apostolopoulos. Epigenetics in tuberculosis: Immunomodulation of host immune response. *Vaccines*, 10:1740, 10 2022. ISSN 2076-393X. doi: 10.3390/vaccines10101740.

[17] Phillip. P. Salvatore and Ying. Zhang. *Tuberculosis: Molecular Basis of Pathogenesis*. Elsevier, 2017. doi: 10.1016/B978-0-12-801238-3.95697-6.

[18] Payam Nahid and Philip C. Hopewell. Tuberculosis treatment, 2017.

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

[20] E. Phelan, A. El-Gammal, and T.M. O'Connor. Tuberculosis, 2011.

[21] Peter D Craggs and Luiz Pedro S de Carvalho. Bottlenecks and opportunities in antibiotic discovery against mycobacterium tuberculosis. *Current Opinion in Microbiology*, 69:102191, 10 2022. ISSN 13695274. doi: 10.1016/j.mib.2022.102191.

[22] Nathan J. Day, Pierre Santucci, and Maximiliano G. Gutierrez. Host cell environments and antibiotic efficacy in tuberculosis. *Trends in Microbiology*, 32:270–279, 3 2024. ISSN 0966842X. doi: 10.1016/j.tim.2023.08.009.

[23] Craig Knox, Mike Wilson, Christen. M Klinger, Mark Franklin, Eponine Oler, et al. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad976. URL https://doi.org/10.1093/nar/gkad976.

[24] Ana María García-Marín, Irving Cancino-Muñoz, Manuela Torres-Puente, Luis M Villamayor, Rafael Borrás, María Borrás-Máñez, Montserrat Bosque, Juan J Camarena, Ester Colomer-Roig, Javier Colomina, Isabel Escribano, Oscar Esparcia-Rodríguez, Ana Gil-Brusola, Concepción Gimeno, Adelina Gimeno-Gascón, Bárbara Gomila-Sard, Damiana González-Granda, Nieves Gonzalo-Jiménez, María Remedio Guna-Serrano, José Luis López-Hontangas, Coral Martín-González, Rosario Moreno-Muñoz, David Navarro, María Navarro, Nieves Orta, Elvira Pérez, Josep Prat, Juan Carlos Rodríguez, María Montserrat Ruiz-García, Hermelinda Vanaclocha, Fernando González-Candelas, Victoria Furió, and Iñaki Comas. Role of the first who mutation catalogue in the diagnosis of antibiotic resistance in mycobacterium tuberculosis in the valencia region, spain: a retrospective genomic analysis. *The Lancet Microbe*, 5, 1 2024. ISSN 26665247. doi: 10.1016/S2666-5247(23)00252-5.

[25] Mamoru Fujiwara, Masanori Kawasaki, Norimitsu Hariguchi, Yongge Liu, and Makoto Matsumoto. Mechanisms of resistance to delamanid, a drug for mycobacterium tuberculosis. *Tuberculosis*, 108: 186–194, 1 2018. ISSN 14729792. doi: 10.1016/j.tube.2017.12.006.

[26] Wenli Wang, Hongjuan Zhou, Long Cai, and Tingting Yang. Association between the rifampicin resistance mutations and rifabutin susceptibility in mycobacterium tuberculosis: a meta-analysis.

*Journal of Global Antimicrobial Resistance*, 11 2024. ISSN 22137165. doi: 10.1016/j.jgar.2024.11.014.

[27] David Ochoa, Andrew Hercules, Miguel Carmona, Daniel Suveges, Jarrod Baker, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. *Nucleic Acids Research*, 51 (D1):D1353–D1359, 11 2022. ISSN 0305-1048. doi: 10.1093/nar/gkac1046. URL https://doi.org/10.1093/nar/gkac1046.

[28] François Chollet. Deep learning with python.

[29] Solveig Badillo, Balazs Banfai, Fabian Birzele, Iakov I. Davydov, Lucy Hutchinson, et al. An introduction to machine learning. *Clinical Pharmacology and Therapeutics*, 107:871–885, 4 2020. ISSN 15326535. doi: 10.1002/cpt.1796.

[30] A. Suruliandi, T. Idhaya, and S. P. Raja. Drug target interaction prediction using machine learning techniques – a review. *International Journal of Interactive Multimedia and Artificial Intelligence*, 8: 86–100, 2024. ISSN 19891660. doi: 10.9781/ijimai.2022.11.002.

[31] Han Shi, Simin Liu, Junqi Chen, Xuan Li, Qin Ma, and Bin Yu. Predicting drug-target interactions using lasso with random forest based on evolutionary information and chemical structure. *Genomics*, 111:1839–1852, 12 2019. ISSN 08887543. doi: 10.1016/j.ygeno.2018.12.007.

[32] Yalin Baştanlar and Mustafa Özuysal. *Introduction to Machine Learning*, pages 105–128. 2014. doi: 10.1007/978-1-62703-748-8_7.

[33] George Obaido, Ibomoiye Domor Mienye, Oluwaseun F. Egbelowo, Ikiomoye Douglas Emmanuel, Adeola Ogunleye, Blessing Ogbuokiri, Pere Mienye, and Kehinde Aruleba. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges, and prospects. *Machine Learning with Applications*, 17:100576, 9 2024. ISSN 26668270. doi: 10.1016/j.mlwa.2024.100576.

[34] Wen Shi, Hong Yang, Linhai Xie, Xiao-Xia Yin, and Yanchun Zhang. A review of machine learning-based methods for predicting drug–target interactions. *Health Information Science and Systems*, 12:30, 4 2024. ISSN 2047-2501. doi: 10.1007/s13755-024-00287-6.

[35] Zhongchen Ma and Songcan Chen. Multi-dimensional classification via a metric approach. *Neurocomputing*, 275:1121–1131, 1 2018. ISSN 09252312. doi: 10.1016/j.neucom.2017.09.057.

[36] Gilbert Berdine MD and Shengping Yang. Linear regression. *The Southwest Respiratory and Critical Care Chronicles*, 2, 2014. ISSN 23259205. doi: 10.12746/swrccc2014.0206.077.

[37] Alessia Mammone, Marco Turchi, and Nello Cristianini. Support vector machines. *WIREs Computational Statistics*, 1:283–289, 11 2009. ISSN 1939-5108. doi: 10.1002/wics.49.

[38] Lior Rokach. Decision forest: Twenty years of research. *Information Fusion*, 27:111–125, 1 2016. ISSN 15662535. doi: 10.1016/j.inffus.2015.06.005.

[39] Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001. ISSN 08856125. doi: 10.1023/A:1010933404324.

[40] Bertrand S. Clarke and Jennifer L. Clarke. Ensemble methods. *Predictive Statistics*, pages 449–523. doi: 10.1017/9781139236003.012.

[41] Oliver Kramer. *K-Nearest Neighbors*, pages 13–23. 2013. doi: 10.1007/978-3-642-38652-7_2.

[42] Umesh R. Hodeghatta and Umesh Nayak. *Unsupervised Machine Learning*, pages 161–186. Apress, 2017. doi: 10.1007/978-1-4842-2514-1_7.

[43] P. Gemperline. Principal component analysis. *Technometrics*, 45:276 – 276, 2003. doi: 10.1007/978-0-387-30164-8_665.

[44] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis. *Nature Methods*, 14:641–642, 2017. doi: 10.1038/nmeth.4346.

[45] Kristina P. Sinaga and Miin-Shen Yang. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727, 2020. doi: 10.1109/ACCESS.2020.2988796.

[46] A. Likas, N. Vlassis, and J. Verbeek. The global k-means clustering algorithm. *Pattern Recognit.*, 36:451–461, 2003. doi: 10.1016/S0031-3203(02)00060-2.

[47] L. Maaten and Geoffrey E. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[48] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

[49] Jiarui Chen, Yain-Whar Si, Chon-Wai Un, and Shirley W. I. Siu. Chemical toxicity prediction based on semi-supervised learning and graph convolutional neural network. *Journal of Cheminformatics*, 13:93, 12 2021. ISSN 1758-2946. doi: 10.1186/s13321-021-00570-8.

[50] Jesper E. van Engelen and Holger H. Hoos. A survey on semi-supervised learning. *Machine Learning*, 109:373–440, 2 2020. ISSN 0885-6125. doi: 10.1007/s10994-019-05855-6.

[51] Daniel S. Wigh, J. Goodman, and A. Lapkin. A review of molecular representation in the age of machine learning. *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 12, 2022. doi: 10.1002/wcms.1603.

[52] L. Pattanaik and Connor W. Coley. Molecular representation: Going long on fingerprints. *Chem*, 2020. doi: 10.1016/j.chempr.2020.05.002.

[53] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50:742–754, 5 2010. ISSN 1549-9596. doi: 10.1021/ci100050t.

[54] Yoshua Bengio, Aaron C. Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2012. doi: 10.1109/TPAMI.2013.50.

[55] Dor Bank, Noam Koenigstein, and R. Giryes. Autoencoders. *ArXiv*, abs/2003.05991, 2020.

[56] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic level protein structure with a language model, 7 2022.

[57] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. Learned protein embeddings for machine learning. *Bioinformatics*, 34:2642–2648, 8 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty178.

[58] James Zou, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, and Amalio Telenti. A primer on deep learning in genomics. *Nature Genetics*, 51:12–18, 1 2019. ISSN 1061-4036. doi: 10.1038/s41588-018-0295-5.

[59] Dennis Elbrächter, Dmytro Perekrestenko, P. Grohs, and H. Bölcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67:2581–2623, 2019. doi: 10.1109/TIT.2021.3062161.

[60] R. Gribonval, Gitta Kutyniok, M. Nielsen, and Felix Voigtländer. Approximation spaces of deep neural networks. *Constructive Approximation*, 55:259–367, 2019. doi: 10.1007/S00365-021-09543-4.

[61] J. Mehrer, Courtney J. Spoerer, N. Kriegeskorte, and Tim C Kietzmann. Individual differences among deep neural network models. *Nature Communications*, 11, 2020. doi: 10.1038/s41467-020-19632-w.

[62] Zewen Li, Fan Liu, Wenjie Yang, Shouheng Peng, and Jun Zhou. A survey of convolutional neural networks: Analysis, applications, and prospects. *IEEE Transactions on Neural Networks and Learning Systems*, 33:6999–7019, 2020. doi: 10.1109/TNNLS.2021.3084827.

[63] Radoslaw Martin Cichy and Daniel Kaiser. Deep neural networks as scientific models. *Trends in Cognitive Sciences*, 23:305–317, 2019. doi: 10.1016/j.tics.2019.01.009.

[64] Yong Yu, Xiaosheng Si, Changhua Hu, and Jian xun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural Computation*, 31:1235–1270, 2019. doi: 10.1162/neco_a_01199.

[65] Vijay Prakash Dwivedi, Chaitanya K. Joshi, T. Laurent, Yoshua Bengio, and X. Bresson. Benchmarking graph neural networks. *ArXiv*, abs/2003.00982, 2023.

[66] James L. Crowley. Convolutional neural networks. *Nature Methods*, 20:1269–1270, 2020. doi: 10.1038/s41592-023-01973-1.

[67] R. Yamashita, M. Nishio, R. Do, and K. Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9:611 – 629, 2018. doi: 10.1007/s13244-018-0639-9.

[68] M. Schuster and K. Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45:2673–2681, 1997. doi: 10.1109/78.650093.

[69] Ying Zhou, Yintao Zhang, Donghai Zhao, Xinyuan Yu, Xinyi Shen, et al. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Research*, 52(D1):D1465–D1477, 09 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad751. URL https://doi.org/10.1093/nar/gkad751.

[70] Tiqing Liu, Yuhmei Lin, Xin Wen, Robert N. Jorissen, and Michael K. Gilson. BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic Acids Research*, 35(suppl_1):D198–D201, 12 2006. ISSN 0305-1048. doi: 10.1093/nar/gkl999.

[71] Anna Gaulton, Anne Hersey, Michał Nowotka, A. Patrícia Bento, Jon Chambers, David Mendez, Prudence Mutowo, Francis Atkinson, Louisa J. Bellis, Elena Cibrián-Uhalte, Mark Davies, Nathan Dedman, Anneli Karlsson, María Paula Magariños, John P. Overington, George Papadatos, Ines Smit, and Andrew R. Leach. The chembl database in 2017. *Nucleic Acids Research*, 45:D945–D954, 1 2017. ISSN 0305-1048. doi: 10.1093/nar/gkw1074.

[72] Sunghwan Kim, Paul A. Thiessen, Evan E. Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A. Shoemaker, Jiyao Wang, Bo Yu, Jian Zhang, and Stephen H. Bryant. Pubchem substance and compound databases. *Nucleic Acids Research*, 44:D1202–D1213, 1 2016. ISSN 0305-1048. doi: 10.1093/nar/gkv951.

[73] Masashi Tsubaki, Kentaro Tomii, and Jun Sese. Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics*, 35:309–318, 1 2019. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty535.

[74] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. *Journal of Chemical Information and Modeling*, 54:735–743, 3 2014. ISSN 1549-9596. doi: 10.1021/ci400709d.

[75] Ying Wang, Yangguang Su, Kairui Zhao, Diwei Huo, Zhenshun Du, Zhiju Wang, Hongbo Xie, Lei Liu, Qing Jin, Xuekun Ren, Xiujie Chen, and Denan Zhang. A deep learning drug screening framework for integrating local-global characteristics: A novel attempt for limited data. *Heliyon*, 10, 7 2024. ISSN 24058440. doi: 10.1016/j.heliyon.2024.e34244.

[76] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34:i821–i829, 9 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/bty593.

[77] Maximilian G. Schuh, Davide Boldini, Annkathrin I. Bohne, and Stephan A. Sieber. Barlow twins deep neural network for advanced 1d drug–target interaction prediction. *Journal of Cheminformatics*, 17:18, 2 2025. ISSN 1758-2946. doi: 10.1186/s13321-025-00952-2.

[78] Liang Dou, Zhen Zhang, Dan liu, Ying Qian, and Qian Zhang. Bcm-dti: A fragment-oriented method for drug–target interaction prediction using deep learning. *Computational Biology and Chemistry*, 104:107844, 6 2023. ISSN 14769271. doi: 10.1016/j.compbiolchem.2023.107844.

[79] Junyue Cao, Qingfeng Chen, Junlai Qiu, Yiming Wang, Wei Lan, Xiaojing Du, and Kai Tan. Ngcn: Drug-target interaction prediction by integrating information and feature learning from heterogeneous network. *Journal of Cellular and Molecular Medicine*, 28, 4 2024. ISSN 1582-1838. doi: 10.1111/jcmm.18224.

[80] Mengmeng Gao, Daokun Zhang, Yi Chen, Yiwen Zhang, Zhikang Wang, Xiaoyu Wang, Shanshan Li, Yuming Guo, Geoffrey I. Webb, Anh T.N. Nguyen, Lauren May, and Jiangning Song. Graphormerdti: A graph transformer-based approach for drug-target interaction prediction. *Computers in Biology and Medicine*, 173:108339, 5 2024. ISSN 00104825. doi: 10.1016/j.compbiomed.2024.108339.

[81] Rawan S Olayan, Haitham Ashoor, and Vladimir B Bajic. Ddr: efficient computational method to predict drug–target interactions using graph mining and machine learning approaches. *Bioinformatics*, 34:1164–1173, 4 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx731.

[82] Yoshihiro Yamanishi, Michihiro Araki, Alex Gutteridge, Wataru Honda, and Minoru Kanehisa. Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, 24:i232–i240, 7 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn162.

[83] André C. A. Nascimento, Ricardo B. C. Prudêncio, and Ivan G. Costa. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC Bioinformatics*, 17:46, 1 2016. ISSN 1471-2105. doi: 10.1186/s12859-016-0890-3.

[84] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29:1046–1051, 11 2011. ISSN 1087-0156. doi: 10.1038/nbt.1990.

[85] Heba El-Behery, Abdel-Fattah Attia, Nawal El-Fishawy, and Hanaa Torkey. Efficient machine learning model for predicting drug-target interactions with case study for covid-19. *Computational Biology and Chemistry*, 93:107536, 8 2021. ISSN 14769271. doi: 10.1016/j.compbiolchem.2021.107536.

[86] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. Mgraphdta: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, 13:816–833, 2022. ISSN 2041-6520. doi: 10.1039/D1SC05180F.

[87] James T Metz, Eric F Johnson, Niru B Soni, Philip J Merta, Lemma Kifle, and Philip J Hajduk. Navigating the kinome. *Nature Chemical Biology*, 7:200–202, 4 2011. ISSN 1552-4450. doi: 10.1038/nchembio.530.

[88] Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications*, 8:573, 9 2017. ISSN 2041-1723. doi: 10.1038/s41467-017-00680-8.

[89] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9:513–530, 2018. ISSN 2041-6520. doi: 10.1039/C7SC02664A.

[90] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. 6 2021. URL http://arxiv.org/abs/2103.03230.

[91] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[92] McKinney W. and Team P. pandas: powerful python data analysis toolkit. Technical report, 2015.

[93] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, and Pauli Virtanen. Array programming with numpy. *Nature*, 585:357–362, 9 2020. ISSN 0028-0836. doi: 10.1038/s41586-020-2649-2.

[94] John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9: 90–95, 2007. ISSN 1521-9615. doi: 10.1109/MCSE.2007.55.

[95] Mark Davies, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. Chembl web services: streamlining access to drug discovery data and utilities. *Nucleic acids research*, 43:W612–20, 7 2015. ISSN 1362-4962. doi: 10.1093/nar/gkv352.

[96] Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, Edouard Duchesnay, and Gilles

Louppe. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 01 2012.

[97] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

[98] Landrum Greg. Rdkit: Open-source cheminformatics. https://www.rdkit.org.

[99] O. J. M. Béquignon, B. J. Bongers, W. Jespers, A. P. IJzerman, B. van der Water, and G. J. P. van Westen. Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *Journal of Cheminformatics*, 15:3, 1 2023. ISSN 1758-2946. doi: 10.1186/s13321-022-00672-x.

[100] Jiangming Sun, Nina Jeliazkova, Vladimir Chupakhin, Jose-Felipe Golib-Dzib, Ola Engkvist, Lars Carlsson, Jörg Wegner, Hugo Ceulemans, Ivan Georgiev, Vedrin Jeliazkov, Nikolay Kochev, Thomas J. Ashby, and Hongming Chen. Excape-db: an integrated large scale dataset facilitating big data analysis in chemogenomics. *Journal of Cheminformatics*, 9:17, 12 2017. ISSN 1758-2946. doi: 10.1186/s13321-017-0203-5.

[101] Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, and Aduragbemi Adesina. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53:D609–D617, 1 2025. ISSN 0305-1048. doi: 10.1093/nar/gkae1010.

[102] Shazia Khan, Sathya Narayanan Nagarajan, Amit Parikh, Sharmishtha Samantaray, Albel Singh, Devanand Kumar, Rajendra P. Roy, Apoorva Bhatt, and Vinay Kumar Nandicoori. Phosphorylation of enoyl-acyl carrier protein reductase inha impacts mycobacterial growth and survival. *Journal of Biological Chemistry*, 285:37860–37871, 11 2010. ISSN 00219258. doi: 10.1074/jbc.M110.143131.

[103] Anita G. Amin, Renan Goude, Libin Shi, Jian Zhang, Delphi Chatterjee, and Tanya Parish. Emba is an essential arabinosyltransferase in mycobacterium tuberculosis. *Microbiology*, 154:240–248, 1 2008. ISSN 1350-0872. doi: 10.1099/mic.0.2007/012153-0.

[104] Wei Lin, Soma Mandal, David Degen, Yu Liu, Yon W. Ebright, Shengjian Li, Yu Feng, Yu Zhang, Sukhendu Mandal, Yi Jiang, Shuang Liu, Matthew Gigliotti, Meliza Talaue, Nancy Connell, Kalyan Das, Eddy Arnold, and Richard H. Ebright. Structural basis of mycobacterium tuberculosis transcription and transcription inhibition. *Molecular Cell*, 66:169–179.e8, 4 2017. ISSN 10972765. doi: 10.1016/j.molcel.2017.03.001.

[105] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7:20, 12 2015. ISSN 1758-2946. doi: 10.1186/s13321-015-0069-3.

[106] Jennifer Alisa Amrhein, Stefan Knapp, and Thomas Hanke. Synthetic opportunities and challenges for macrocyclic kinase inhibitors. *Journal of Medicinal Chemistry*, 64:7991–8009, 6 2021. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.1c00217.

[107] Hanxiao Yu, He Yang, Enxue Shi, and Wenjun Tang. Development and clinical application of phosphorus-containing drugs. *Medicine in Drug Discovery*, 8:100063, 12 2020. ISSN 25900986. doi: 10.1016/j.medidd.2020.100063.

[108] AkshatKumar Nigam, Robert Pollice, Gary Tom, Kjell Jorner, John Willes, Luca A. Thiede, Anshul Kundaje, and Alan Aspuru-Guzik. Tartarus: A benchmarking platform for realistic and practical inverse molecular design, 2023. URL https://arxiv.org/abs/2209.12487.

[109] Christopher A. Lipinski, Franco Lombardo, Beryl W. Dominy, and Paul J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews*, 23:3–25, 1 1997. ISSN 0169409X. doi: 10.1016/S0169-409X(96)00423-1.

[110] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1:8, 12 2009. ISSN 1758-2946. doi: 10.1186/1758-2946-1-8.

[111] G. Richard Bickerton, Gaia V. Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L. Hopkins. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4:90–98, 2 2012. ISSN 1755-4330. doi: 10.1038/nchem.1243.

[112] Hui Liu, Jianjiang Sun, Jihong Guan, Jie Zheng, and Shuigeng Zhou. Improving compound–protein

interaction prediction by building up highly credible negative samples. *Bioinformatics*, 31:i221–i229, 6 2015. ISSN 1367-4811. doi: 10.1093/bioinformatics/btv256.

[113] Hemant Joshi, Divya Kandari, and Rakesh Bhatnagar. Insights into the molecular determinants involved in mycobacterium tuberculosis persistence and their therapeutic implications, 2021. ISSN 21505608.

[114] Xiaopan Gao, Xia Yu, Kaixiang Zhu, Bo Qin, Wei Wang, Pu Han, Justyna Aleksandra Wojdyla, Meitian Wang, and Sheng Cui. Crystal structure of mycobacterium tuberculosis elongation factor g1. *Frontiers in Molecular Biosciences*, 8, 9 2021. ISSN 2296-889X. doi: 10.3389/fmolb.2021.667638.

[115] Yoanna Teneva, Rumyana Simeonova, Violeta Valcheva, and Violina T. Angelova. Recent advances in anti-tuberculosis drug discovery based on hydrazide–hydrazone and thiadiazole derivatives targeting inha. *Pharmaceuticals*, 16:484, 3 2023. ISSN 1424-8247. doi: 10.3390/ph16040484.

[116] Samreen Fatima, Ashima Bhaskar, and Ved Prakash Dwivedi. Repurposing immunomodulatory drugs to combat tuberculosis. *Frontiers in Immunology*, 12, 4 2021. ISSN 1664-3224. doi: 10.3389/fimmu.2021.645485.

[117] Alice Italia, Mohammed Monsoor Shaik, and Francesco Peri. Emerging extracellular molecular targets for innovative pharmacological approaches to resistant mtb infection. *Biomolecules*, 13: 999, 6 2023. ISSN 2218-273X. doi: 10.3390/biom13060999.

[118] Ashok Vinayak Purandare, Douglas G. Batt, Qingjie Liu, Harold Mastalerz, and Kurt Zimmermann. Carbazole and carboline kinase inhibitors, August 26 2014. URL https://patents.google.com/patent/US8815840B2/en. United States Patent.

[119] Karthik Raman, Kalidas Yeturu, and Nagasuma Chandra. targettb: A target identification pipeline for mycobacterium tuberculosis through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology*, 2:109, 12 2008. ISSN 1752-0509. doi: 10.1186/1752-0509-2-109.

[120] Bin Jia, Yang min Ma, Bin Liu, Pu Chen, Yan Hu, and Rui Zhang. Synthesis, antimicrobial activity, structure-activity relationship, and molecular docking studies of indole diketopiperazine alkaloids. *Frontiers in Chemistry*, 7, 11 2019. ISSN 2296-2646. doi: 10.3389/fchem.2019.00837.

[121] Ahmed S. Abdelsamie, Mohamed Salah, Lorenz Siebenbürger, Ahmed Merabet, Claudia Scheuer, Martin Frotscher, Sebastian T. Müller, Oliver Zierau, Günter Vollmer, Michael D. Menger, Matthias W.

Laschke, Chris J. van Koppen, Sandrine Marchais-Oberwinkler, and Rolf W. Hartmann. Design, synthesis, and biological characterization of orally active 17-hydroxysteroid dehydrogenase type 2 inhibitors targeting the prevention of osteoporosis. *Journal of Medicinal Chemistry*, 62:7289–7301, 8 2019. ISSN 0022-2623. doi: 10.1021/acs.jmedchem.9b00932.

[122] Victor O. Gawriljuk, Alaa Alhayek, Anna K.H. Hirsch, and Matthew R. Groves. Apo structure of mycobacterium tuberculosis 1-deoxy-d-xylulose 5-phosphate synthase dxps: Dynamics and implications for inhibitor design. *Biochemical and Biophysical Research Communications*, 747:151246, 2 2025. ISSN 0006291X. doi: 10.1016/j.bbrc.2024.151246.

[123] Matthew O. Duffey, Ruth Adams, Christopher Blackburn, Ryan W. Chau, Susan Chen, Katherine M. Galvin, Khristofer Garcia, Alexandra E. Gould, Paul D. Greenspan, Sean Harrison, Shih-Chung Huang, Mi-Sook Kim, Bheemashankar Kulkarni, Steven Langston, Jane X. Liu, Li-Ting Ma, Saurabh Menon, Masayuki Nagayoshi, R. Scott Rowland, Tricia J. Vos, Tianlin Xu, Johnny J. Yang, Shaoxia Yu, and Qin Zhang. Discovery and optimization of pyrazoline compounds as b-raf inhibitors. *Bioorganic & Medicinal Chemistry Letters*, 20:4800–4804, 8 2010. ISSN 0960894X. doi: 10.1016/j.bmcl.2010.06.113.

[124] Zi-Wei Chen, Karoline Fuchs, Werner Sieghart, R. Reid Townsend, and Alex S. Evers. Deep amino acid sequencing of native brain gabaa receptors using high-resolution mass spectrometry. *Molecular & Cellular Proteomics*, 11:M111.011445, 1 2012. ISSN 15359476. doi: 10.1074/mcp.M111.011445.

[125] Sauvik Samanta, Sumit Kumar, Eswar K. Aratikatla, Sandeep R. Ghorpade, and Vinayak Singh. Recent developments of imidazo[1,2- a ]pyridine analogues as antituberculosis agents. *RSC Medicinal Chemistry*, 14:644–657, 2023. ISSN 2632-8682. doi: 10.1039/D3MD00019B.

[126] Jr. Morley, Robert E., Edward J. Richter, and George L. Engel. Method and apparatus for authenticating a magnetic fingerprint signal using an adaptive analog to digital converter, May 1 2007. URL https://patents.google.com/patent/US7210627B2/en. United States Patent.

[127] Asesh Banerjee, Eugenie Dubnau, Annaik Quemard, V. Balasubramanian, Kyung Sun Um, Theresa Wilson, Des Collins, Geoffrey de Lisle, and William R. Jacobs. inha , a gene encoding a target for isoniazid and ethionamide in mycobacterium tuberculosis. *Science*, 263:227–230, 1 1994. ISSN 0036-8075. doi: 10.1126/science.8284673.

[128] Yanhui Yang, Peng Gao, Yishuang Liu, Xinyue Ji, Maoluo Gan, Yan Guan, Xueqin Hao, Zhuorong Li, and Chunling Xiao. A discovery of novel mycobacterium tuberculosis pantothenate synthetase inhibitors based on the molecular mechanism of actinomycin d inhibition. *Bioorganic & Medicinal Chemistry Letters*, 21:3943–3946, 7 2011. ISSN 0960894X. doi: 10.1016/j.bmcl.2011.05.021.

[129] Yicheng Gong, Chuancun Wei, Jun Wang, Nengjiang Mu, Qinhong Lu, Chengyao Wu, Ning Yan, Huifang Yang, Yao Zhao, Xiuna Yang, Sudagar S. Gurcha, Natacha Veerapen, Sarah M. Batt, Zhiqiang Hao, Lintai Da, Gurdyal S. Besra, Zihe Rao, and Lu Zhang. Structure of the priming arabinosyltransferase afta required for ag biosynthesis of mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 120, 6 2023. ISSN 0027-8424. doi: 10.1073/pnas.2302858120.

[130] Souleymane Konate, Koffi N'Guessan Placide Gabin Allangba, Issouf Fofana, Raymond Kre N'Guessan, Eugene Megnassan, Stanislav Miertus, and Vladimir Frecer. Improved inhibitors targeting the thymidylate kinase of multidrug-resistant mycobacterium tuberculosis with favorable pharmacokinetics. *Life*, 15:173, 1 2025. ISSN 2075-1729. doi: 10.3390/life15020173.

[131] Emily M. Cross, Felise G. Adams, Jack K. Waters, David Aragão, Bart A. Eijkelkamp, and Jade K. Forwood. Insights into acinetobacter baumannii fatty acid synthesis 3-oxoacyl-acp reductases. *Scientific Reports*, 11:7050, 3 2021. ISSN 2045-2322. doi: 10.1038/s41598-021-86400-1.

[132] Vijai Singh and Pallavi Somvanshi. Homology modelling of 3-oxoacyl-acyl carrier protein synthase ii from mycobacterium tuberculosis h37rv and molecular docking for exploration of drugs. *Journal of Molecular Modeling*, 15:453–460, 5 2009. ISSN 1610-2940. doi: 10.1007/s00894-008-0426-5.

[133] Hiten J. Gutka, Jasper Marc G. Bondoc, Ryan Patwell, Shahebraj Khan, Edyta M. Grzelak, Rajendra Goswami, Martin I. Voskuil, and Farahnaz Movahedzadeh. Rv0100: An essential acyl carrier protein from m. tuberculosis important in dormancy. *PLoS ONE*, 19, 6 2024. ISSN 19326203. doi: 10.1371/journal.pone.0304876.

[134] Hongjuan Zhang, Ying Chen, Yu Zhang, Luyao Qiao, Xiangyin Chi, Yanxing Han, Yuan Lin, Shuyi Si, and Jiandong Jiang. Identification of anti-mycobacterium tuberculosis agents targeting the interaction of bacterial division proteins ftsz and sepfe. *Acta Pharmaceutica Sinica B*, 13:2056–2070, 5 2023. ISSN 22113835. doi: 10.1016/j.apsb.2023.01.022.

[135] Anshika Sharma, Gopal K Khuller, and Sadhna Sharma. Peptide deformylase – a promising therapeutic target for tuberculosis and antibacterial drug discovery. *Expert Opinion on Therapeutic Targets*, 13:753–765, 7 2009. ISSN 1472-8222. doi: 10.1517/14728220903005590.