Project Description

AI for discovering affordable therapies against neglected tropical diseases

Supervisors: Dr Diego A. Oyarzún and Dr Shay Cohen

Abstract

This project aims to address the urgent need for affordable therapies against neglected tropical diseases (NTDs), focusing specifically on Chagas disease. Despite its significant global health impact, Chagas disease lacks affordable effective treatment options. We will build a machine learning pipeline to design compound libraries for experimental screening, in collaboration with the DNDi. The approach involves training binary classifiers of drug action using ensemble models and graph neural networks, followed chemical large language models (LLM) to screen compound libraries. The research will utilize phenotypic screening data from DNDi, incorporating nearly 900,000 chemical structures with various readouts of effect. The project has the potential to significantly contribute to improve health outcomes in low-income countries affected by Chagas disease.

Introduction

Neglected Tropical Diseases comprise a diverse group of infectious diseases that primarily affect populations in low-income regions of the world, particularly in tropical and subtropical climates. These diseases disproportionately affect poor and marginalized communities with limited access to adequate healthcare, sanitation, and clean water. Despite their significant impact on global health, these diseases often receive little attention from the pharmaceutical industry due to their unprofitability and lack of financial incentives for R&D aimed at new therapies.

Chagas disease is a potentially life-threatening illness caused by the protozoan parasite *Trypanosoma cruzi*. It primarily affects people in impoverished areas of Latin America and increased population migration have carried Chagas disease to new regions such as the the United States and European countries. The acute phase typically occurs shortly after infection and may exhibit mild symptoms or go unnoticed. However, if left untreated, the infection progresses to the chronic phase, which can last for years or even decades. During the chronic phase, the parasite can cause severe damage to the heart, digestive system, and other organs, leading to complications such as heart failure, arrhythmias, and other serious abnormalities [1].

The impact of Chagas disease on global health is significant, with an estimated 6-7 million people infected worldwide and approximately 10,000 deaths annually. The available treatment options are limited; there are only two drugs, benznidazole and nifurtimox, currently approved for use, but these have significant limitations, including high cost, lengthy treatment durations, and potential side effects.

Research Challenge

Our general aim is to develop a machine learning pipeline to design libraries of compounds for further experimental screening by our partner DNDi. The specific aims are:

1. Train binary classifiers of drug action, using a mix of techniques, including classic ensemble models, which the supervisor has successfully employed in a recent study for senolytic therapy [2], as well as graph neural networks that were recently employed for discovering novel antibiotics against resistant pathogens [3].
2. Build a chemical large language model (LLM) from the screened compounds against *T. cruzi*. The supervisor has recently trialled a publicly available chemical LLM [4]

pretrained on 10M chemical structures from PubChem for both antibiotics and senolytics discovery. The results suggest a strong potential of these models for early hit discovery.

3. Computationally screen unlabelled compound libraries designed in collaboration with DNDi, and chosen based on chemical diversity, physicochemical properties, availability, and cost.

Data & Methodology

We will employ the phenotypic screening data from DNDi collected on their in-house *T. cruzi* assay. The dataset includes 900,000 chemical structures alongside five readouts of effect derived from dose-response curves: IC50, IC90, CC50, % Max inhibition, and SI. The *T. cruzi* data covers almost 100-fold more compounds than those employed in recent successful applications of AI for hit discovery [2], [3], and thus may lead to models with better success rate.

In Objective 1, we will first employ classic statistical learning available in scikit-learn and related Python packages, including tree-based methods, feed-forward neural networks and ensemble models. The models will be trained on combinations of physicochemical descriptors and molecular fingerprints computable with chemoinformatics software such as RDKit [5]. In a second stage we proceed with deep learning models based on message-passing neural networks (ChemProp [6]), currently the gold standard for molecular property prediction from large data. We will extensively compare both approaches through cross-validation and benchmarking.

In Objective 2, we will employ the BERT architecture for masked language models. We will explore various strategies for tokenization and training starting from the ideas from recent chemical LLMs, ChemBerta [4] and ChemGPT [7]. The pretrained LLM will be employed for compound classification using both the learned embeddings for supervised classification, as well as fine-tuning with labelled compounds. After trying with off-the-shelf pretrained models, we will explore modifications to the model and/or pre-training it on the labelled *T. cruzi* data itself through task-specific pre-training.

In Objective 3, we will query the trained models on various combinations of libraries with varying chemical properties, including toxicity or the presence of beneficial fragments based on DNDi's and Sandexis' expertise. The computational screen will require extensive unsupervised analysis of compound libraries using a combination of clustering, dimensionality reduction and multivariate statistics.

RRI/Ethical considerations

This project addresses an urgent need for affordable new therapies against Chagas disease. The burden of Chagas is particularly heavy in low-resource settings, where access to healthcare services and diagnostic tools is limited. Moreover, the chronic nature of the disease places a considerable economic burden on affected individuals and healthcare systems, as it often requires long-term medical care and monitoring. This project will help early-stage identification of chemical scaffolds with potential for downstream optimization, leveraging the knowhow of DNDi and their experience in public-private-academic partnerships to address some of the most pressing health challenges in low-income countries.

Expected outcome & Impact

Outcomes:

1) Suite of trained supervised machine learning predictors of drug action on *T. cruzi*.

2) Pre-trained chemical LLM suitable for prediction, dimensionality reduction, and clustering analyses of chemical structures against *T. cruzi*.

3) A library of chemical structures prioritized for experimental screening in *T. cruzi*, informed by the predicted hits from the machine learning models.

Impact:

The project promises to substantially contribute to the discovery of new therapies against Chagas disease. There is significant potential for impact delivery and amplification through the close link with DNDi, a leading non-profit with a successful track record in end-to-end development of low-cost therapies for tropical diseases, and its global network of public-private partnerships that make a real-world impact improving the lives of some of the most marginalized communities in the planet.

References

[1] A. Rassi and J. A. Marin-Neto, ‘Chagas disease’, The Lancet, vol. 375, no. 9723, pp. 1388–1402, Apr. 2010, doi: 10.1016/S0140-6736(10)60061-X.

[2] V. Smer-Barreto et al., ‘Discovery of senolytics using machine learning’, Nat Commun, vol. 14, no. 1, Art. no. 1, Jun. 2023, doi: 10.1038/s41467-023-39120-1.

[3] G. Liu et al., ‘Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii’, Nat Chem Biol, vol. 19, no. 11, Art. no. 11, Nov. 2023, doi: 10.1038/s41589-023-01349-8.

[4] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar, ‘ChemBERTa-2: Towards Chemical Foundation Models’. arXiv, Sep. 04, 2022. doi: 10.48550/arXiv.2209.01712.

[5] ‘RDKit’. Accessed: Mar. 16, 2023. [Online]. Available: https://www.rdkit.org/

[6] E. Heid et al., ‘Chemprop: Machine Learning Package for Chemical Property Prediction’. ChemRxiv, Jul. 24, 2023. doi: 10.26434/chemrxiv-2023-3zcfl.

[7] N. C. Frey et al., ‘Neural scaling of deep chemical models’, Nat Mach Intell, vol. 5, no. 11, pp. 1297–1305, Nov. 2023, doi: 10.1038/s42256-023-00740-3.