COMPUTATIONALLY ACCELERATING PROTEIN-LIGAND DOCKING FOR NEGLECTED TROPICAL DISEASES: A CASE STUDY ON DRUG REPURPOSING FOR LEISHMANIASIS

Anonymous authors
Paper under double-blind review

ABSTRACT

In this work, we propose a method blending representation learning and molecular docking to predict protein ligand interaction, a key building block of drug repurposing and discovery. Using Leishmaniasis as a case study, we analyze the speed-accuracy trade-off that representation learning methods provide when compared to more computationally intensive molecular docking methods. We find that while deep learning methods substantially reduce the screening burden for molecular docking by a factor of 600, they can not be trusted to find the top ligands binding to a given target. This suggests that current deep learning methods can be used to come up with a short list of most promising ligands but the final predictions should rely on molecular docking.

1 INTRODUCTION

Out of the 13,000 existing diseases known in the medical literature, roughly 5,000 have available treatments with the remaining 8,000 belonging to the rare disease category (Boycott et al., 2013). By definition, rare diseases affect a smaller or less affluent proportion of the global population and are not the major targets of pharmaceutical research programs (Feasey et al., 2010).

Many tropical diseases fall in this category. Furthermore, since the 1980s, drug discovery has become slower and more expensive over time, despite improvements in technology. This is known as Eroom's law (Scannell et al., 2012).

For neglected tropical diseases, drug repurposing seems like the best avenue to find treatments. Drug repositioning involves the investigation of existing drugs (1500 approved drugs) or drugs which have gone through advanced clinical trials for new therapeutic purposes (Pushpakom et al., 2019).

Depending on the approach, finding a new therapeutic use case for a drug involves identifying molecules (ligands) binding to proteins which are involved in pathway responsible for a symptoms or associated with the disease vector, be it a parasite or virus. Protein-ligand docking plays a crucial role in this process (Greenside et al., 2018).

In this work, we use Leishmaniasis as a concrete disease to showcase how deep learning methods can be paired with molecular simulation methods to accelerate protein-ligand docking.

We present a pipeline for drug repurposing which consists of applying deep learning methods (Huang et al., 2020) for Drug-Target interaction prediction, followed by the application of Protein-Ligand Docking with Autodock Vina (Morris et al., 2009).

We also quantify the agreement between deep learning methods and molecular simulation methods for protein-ligand prediction. Our main contributions are the quantification of the robustness of deep learning methods when compared against more established molecular docking methods and the presentation of a hybrid approach taking advantage of the speed of deep learning while not sacrificing accuracy.

2 RELATED WORK

2.1 LEISHMANIASIS

Leishmaniasis depicts a clinical spectrum of diseases caused by a parasitic flagellate protozoan that infests the blood of various vertebrates including humans by means of a vector usually the female phlebotomist (Torres-Guerrero et al., 2017). The signs and symptoms vary depending on the species of Leishmania. Two common forms include cutaneous Leishmaniasis which affects the skin and a more fetal form visceral leishmaniasis which affects internal organs such as the liver and spleen.

As a neglected tropical disease, Leishmaniasis affects the most vulnerable populations. These populations are exposed to costly, prolonged and even toxic treatment despite advances over the years (Murray et al., 2005). The use of existing (usually approved) drugs to cure these diseases, is a fast growing field called drug repurposing. Drug repurposing is made more feasible by the use of computational methods. These computational methods are either target or disease based (Sanseau & Koehler, 2011).

2.2 COMPUTATIONAL METHODS FOR PROTEIN-LIGAND PREDICTION

Molecular docking is a computational procedure that attempts to predict noncovalent binding of macromolecules or, more frequently, of a macromolecule (receptor) and a small molecule (ligand) efficiently, starting with their unbound structures, structures obtained from MD simulations, or homology modeling. Molecular docking is interested in reproducing chemical potentials, which determine the bound conformation preference and the free energy of binding. These methods take as input the 3D structure of ligands and proteins. The most common software used for molecular docking are AutoDock Vina, Rosetta Ligand, AutoDock 4 (Meiler & Baker, 2006; Morris et al., 2009; Trott & Olson, 2010). Molecular docking methods are widely trusted, used and played a role in the development of the first clinically approved HIV-1 integrase inhibitor (Goodsell et al., 2021; Schames et al., 2004). More recently, these methods have also been used to identify treatments for SARS-CoV-2 (Lokhande et al., 2020; Das et al., 2020).

Representation Learning methods on the other hand are more recent and bypass the need for 3D structures. Through unsupervised and supervised learning, these methods leverage protein and molecule vectors to predict binding affinity. Protein vectors are represented by either directly extracting features from the amino acid sequence as shown by Reczko & Bohr (1994); Chou (2005); Shen et al. (2007); Chou (2000); Kim et al. (2019) or using protein language models trained on the task of next amino acid prediction Alley et al. (2019); Elnaggar et al. (2020); Rives et al. (2019); Heinzinger et al. (2019). Ligand vectors are represented as either hand-crafted features from SMILES strings such as in Toropov et al. (2005); Rogers & Hahn (2010); Kim et al. (2019) or graph neural networks taking as input atoms and their bonds (Gilmer et al., 2017). The protein and ligand representations are fed to a decoder trained to predict binding affinity.

3 METHODS

3.1 PIPELINE OVERVIEW

In the first part of our pipeline, we computed the affinity scores between targets which characterized the Leishmaniasis disease and a large list of drugs using DeepPurpose. DeepPurpose introduced by Huang et al. (2020) is the current state-of-the-art deep learning based method for virtual screening.

Based on the affinity scores obtained from DeepPurpose we shortened the list of drug candidates by keeping the first 50 drugs for each target. We then select those pairs and compute interface energies using molecular docking.

We started our experiments with data from the Indaba Grand Challenge [1]. It consists of a list of approved drugs (4000+ drugs) and all the known targets (500+ targets) associated with Leishmaniasis.

[1] https://deeplearningindaba.com/grand-challenges/leishmaniasis/

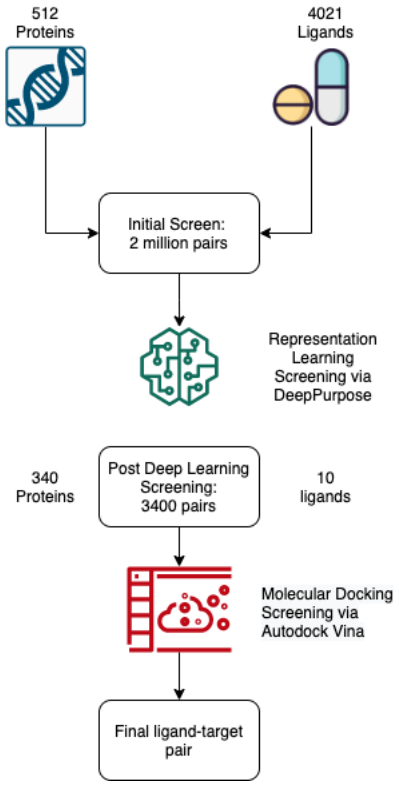


Figure 1: Overview of hybrid deep learning and molecular docking approach for drug repurposing

3.2 Measuring the speed-accuracy trade-off of deep learning methods

A central question of this work is understanding the speed-accuracy trade-off of deep learning methods. To understand by how much deep learning methods accelerate protein-ligand matching, we compare the number of CPU hours that would have been required to exhaustively perform protein ligand docking solely relying on molecular docking software compared to the number of CPU hours required using our approach.

To understand to what extent deep learning methods agree with molecular docking methods, we analyzed the correlation between the results of deep learning based methods and computational molecular docking tool.

Both methods output an affinity score for a given ligand-target pair but have different rules for ligand appreciation. For DeepPurpose, a higher binding score indicates the ligand is better, while for Autocking, a lower free energy implies the ligand is better. To compute the correlation of the scores, for each target, we picked a shortlist of the top 50 drugs according to the deep learning methods.

For these 50 drugs, we ranked them using scores from Autodock Vina and compared the ranking correlation between them using the Kendall rank coefficient (Abdi, 2007). The closer the rank correlation is to 1, the better.

For each target and the top 50 drugs identified by Deep Learning, we also measured the Pearson correlation between the binding score computed by DeepPurpose and the interface energies from molecular docking (Benesty et al., 2009). The close the correlation is to -1, the better.

Furthermore, given the nature of this problem, most protein-ligand pairs are irrelevant and the most important part is coming up with a handful of strong candidate drugs. To take this into account, we measured the average rank that the top drug according to molecular docking is assigned when relying on deep learning methods.

	Min	Max	Mean
Kendall τ Rank Correlation	-0.33	0.10	-0.09
Pearson Correlation	-0.30	0.07	-0.12

Table 1: Results of agreements between deep learning and molecular docking methods

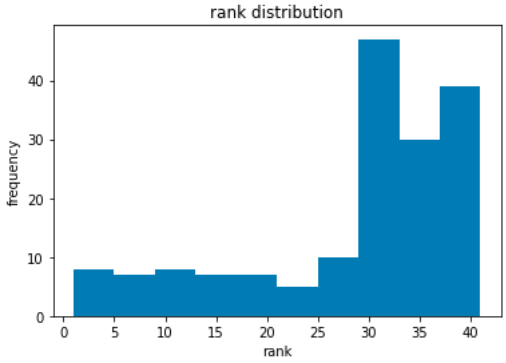


Figure 2: Rank assigned by deep learning methods to top drug according to molecular docking

4 RESULTS

4.1 DEEP LEARNING IS A POWERFUL TOOL FOR SCREENING BUT A WEAK TOOL FOR RANKING

The first insight that this work reveals is that deep learning methods are currently weak tool to perform top drug selection. As can be seen in Table 1, when compared with molecular docking methods, across targets, the rank and correlation scores are low.

However, when looking at the rank assigned by deep learning methods to the top drug found by Autodock Vina, we find that the median rank is 32 with a worst rank case of 41. This means that for all the targets we have tried, the top drug according to molecular docking always will be found most of the time between position 30 and 40 of the sorted drugs according to deep learning methods. Exceptionally, it will be in the top 30. Given that the initial ligand library has 4021 ligands, deep learning methods are able to significantly reduce the space of drug to look at by a factor of 100. That said, after this initial reduction, they can't be trusted to exactly identify the top drug.

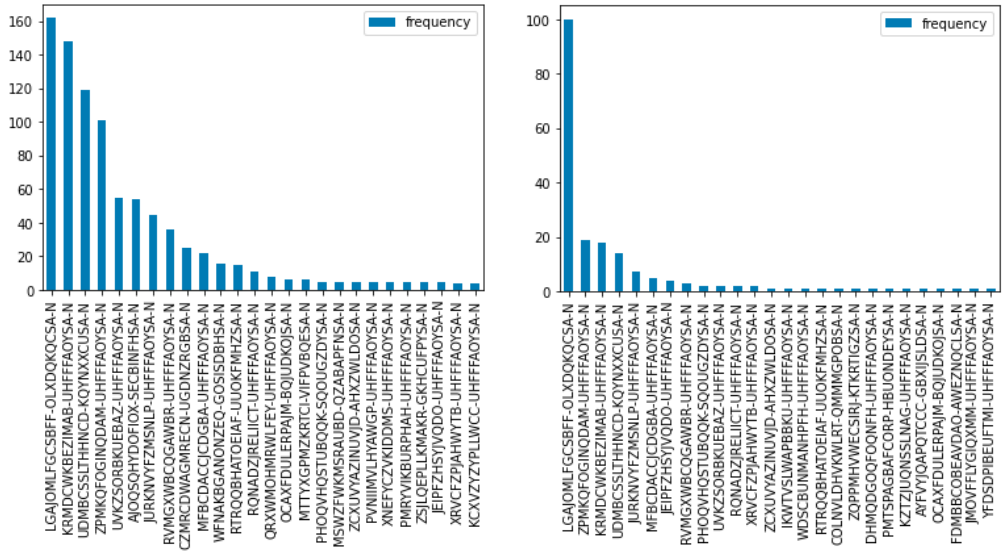
4.2 A FEW DRUGS CONSISTENTLY COME AT THE TOP

After shrinking the list of drug candidates obtained and predicting ligand target interaction, we measured how often various drugs were the best to bind to each target. As the disease is characterized by many targets it is important to find the ligands that will be well bounded with much receptors as possible. We noticed that top hits were dominated by a few drugs.

As we hypothesized, there are ligands that can be bounded effectively with multiply receptors. These ligands serve as potential candidate for treatment. As it is shown in the frequency histogram, Figure 3, the ligand *LGAJOMLFGCSBFF-OLXDQKQCSA-N* most frequently appears as the top drug candidate for several targets, hence it can be considered as the most susceptible drug to treat Leishmaniasis.

4.3 BIOLOGICAL ANALYSIS FOR MOST PROMISING DRUG

Our model predicts the best compound to use as a drug is alpha-glutamicin. It is a non-steroidal anti-inflammatory drug that adheres to Lipinski's rule. In addition to its anti-inflammatory actions, it exhibits analgesic, antipyretic, and platelet-inhibitory actions. Its mechanism of action includes blocking the synthesis of prostaglandins by inhibiting cyclooxygenase, which converts arachidonic acid to



(a) Top 5 (b) Top 1

Figure 3: Frequency of drug's appearance at the top of the list for each target

cyclic endoperoxides, precursors of prostaglandins. Inhibition of prostaglandin synthesis accounts for their analgesic, antipyretic, and platelet-inhibitory actions; other mechanisms may contribute to their anti-inflammatory effects (PubChem). Furthermore, previous studies have hypotheses that prostaglandins are parasite virulence factors and play a role in host interactions (Alves-Ferreira et al., 2020). While the underlying mechanism is yet to be understood, one can hypothesize that blocking the formation of the substrate, arachidonic acid, required by the parasite to make prostaglandin can help prevent parasitic action. Based on our model, the best drug-target pair is LGAJOMLFGCSBFF-OLXDQKQCSA-N and A0A5K1VKU9 with binding energy of -10.7Kcal/mol. Nevertheless, the chosen compound has negative and similar binding energies with various Leishmania targets making multi-target attack possible.

4.4 Computational Speedup: From 2 months to 14 hours

Molecular docking is a traditional method used to compute the configuration of a molecule when bounded to another. It was de facto methodology and the ground truth for drug discovery.

The dataset used to conduct the screening has 2,058,752 protein-ligand pairs. Given that the majority of drug-target pairs are irrelevant, screening them is important to focus on the most promising hits. Using Deep Learning to screen all protein-ligands pairs, we are able to come up with a shortlist of 3400 pairs in which for each target, the top drug matching to it is consistently found. We bring a computation that would have taken 2 months to 14 hours.

5 Conclusion

To conclude, in this work we use Leishmaniasis to understand the role that deep learning methods can play to accelerate protein ligand prediction. We find that deep learning methods can provide a reliable shortlist of promising ligands but they can't be trusted to find the best ligand. Based on this insight, we present an approach taking advantage of the speedup provided by deep learning methods while relying on the accuracy of molecular docking. Future work can extend this analysis to other diseases and move from molecular docking to experimental characterization of most promising protein ligand pairs.

REFERENCES

Hervé Abdi. The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pp. 508–510, 2007.

Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.

Eliza Vanessa Carneiro Alves-Ferreira, Tiago Rodrigues Ferreira, Pegine Walrad, Paul M. Kaye, and Angela Kaysel Cruz. Leishmania braziliensis prostaglandin f(2α) synthase impacts host infection. *Parasites & vectors*, 13(1):9–9, Jan 2020. ISSN 1756-3305. doi: 10.1186/s13071-020-3883-z. URL `https://pubmed.ncbi.nlm.nih.gov/31915065`. 31915065[pmid].

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer, 2009.

Kym M Boycott, Megan R Vanstone, Dennis E Bulman, and Alex E MacKenzie. Rare-disease genetics in the era of next-generation sequencing: discovery to translation. *Nature Reviews Genetics*, 14(10):681–691, 2013.

Kuo-Chen Chou. Prediction of protein subcellular locations by incorporating quasi-sequence-order effect. *Biochemical and biophysical research communications*, 278(2):477–483, 2000.

Kuo-Chen Chou. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. *Bioinformatics*, 21(1):10–19, 2005.

Sourav Das, Sharat Sarmah, Sona Lyndem, and Atanu Singha Roy. An investigation into the identification of potential inhibitors of sars-cov-2 main protease using molecular docking study. *Journal of Biomolecular Structure and Dynamics*, pp. 1–11, 2020.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Debsindhu Bhowmik, et al. Prottrans: Towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.

Nick Feasey, Mark Wansbrough-Jones, David CW Mabey, and Anthony W Solomon. Neglected tropical diseases. *British medical bulletin*, 93(1):179–200, 2010.

Justin Gilmer, Samuel S Schoenholz, Patrick F Riley, Oriol Vinyals, and George E Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, pp. 1263–1272. PMLR, 2017.

David S Goodsell, Michel F Sanner, Arthur J Olson, and Stefano Forli. The autodock suite at 30. *Protein Science*, 30(1):31–43, 2021.

Peyton Greenside, Maureen Hillenmeyer, and Anshul Kundaje. Prediction of protein-ligand interactions from paired protein sequence motifs and ligand substructures. In *PSB*, pp. 20–31. World Scientific, 2018.

Michael Heinzinger, Ahmed Elnaggar, Yu Wang, Christian Dallago, Dmitrii Nechaev, Florian Matthes, and Burkhard Rost. Modeling aspects of the language of life through transfer-learning protein sequences. *BMC bioinformatics*, 20(1):1–17, 2019.

Kexin Huang, Tianfan Fu, Lucas M Glass, Marinka Zitnik, Cao Xiao, and Jimeng Sun. Deeppurpose: a deep learning library for drug–target interaction prediction. *Bioinformatics*, 2020.

Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, et al. Pubchem 2019 update: improved access to chemical data. *Nucleic acids research*, 47(D1):D1102–D1109, 2019.

Kiran Bharat Lokhande, Sayali Doiphode, Renu Vyas, and K Venkateswara Swamy. Molecular docking and simulation studies on sars-cov-2 mpro reveals mitoxantrone, leucovorin, birinapant, and dynasore as potent drugs against covid-19. *Journal of Biomolecular Structure and Dynamics*, pp. 1–12, 2020.

Jens Meiler and David Baker. Rosettaligand: Protein–small molecule docking with full side-chain flexibility. *Proteins: Structure, Function, and Bioinformatics*, 65(3):538–548, 2006.

Garrett M Morris, Ruth Huey, William Lindstrom, Michel F Sanner, Richard K Belew, David S Goodsell, and Arthur J Olson. Autodock4 and autodocktools4: Automated docking with selective receptor flexibility. *Journal of computational chemistry*, 30(16):2785–2791, 2009.

Henry W Murray, Jonathan D Berman, Clive R Davies, and Nancy G Saravia. Advances in leishmaniasis. *The Lancet*, 366(9496):1561–1577, 2005. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(05)67629-5. URL https://www.sciencedirect.com/science/article/pii/S0140673605676295.

PubChem. Glucametacin. URL https://pubchem.ncbi.nlm.nih.gov/compound/3033980.

Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

M Reczko and H Bohr. The def data base of sequence based protein fold class predictions. *Nucleic acids research*, 22(17):3616, 1994.

Alexander Rives, Siddharth Goyal, Joshua Meier, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *bioRxiv*, pp. 622803, 2019.

David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

Philippe Sanseau and Jacob Koehler. Editorial: Computational methods for drug repurposing. *Briefings in Bioinformatics*, 12(4):301–302, 07 2011. ISSN 1467-5463. doi: 10.1093/bib/bbr047. URL https://doi.org/10.1093/bib/bbr047.

Jack W Scannell, Alex Blanckley, Helen Boldon, and Brian Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature reviews Drug discovery*, 11(3):191–200, 2012.

Julie R Schames, Richard H Henchman, Jay S Siegel, Christoph A Sotriffer, Haihong Ni, and J Andrew McCammon. Discovery of a novel binding trench in hiv integrase. *Journal of medicinal chemistry*, 47(8):1879–1881, 2004.

Juwen Shen, Jian Zhang, Xiaomin Luo, Weiliang Zhu, Kunqian Yu, Kaixian Chen, Yixue Li, and Hualiang Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.

Andrey A Toropov, Alla P Toropova, Dilya V Mukhamedzhanoval, and Ivan Gutman. Simplified molecular input line entry system (smiles) as an alternative for constructing quantitative structure-property relationships (qspr). 2005.

Edoardo Torres-Guerrero, Marco Romano Quintanilla-Cedillo, Julieta Ruiz-Esmenjaud, and Roberto Arenas. Leishmaniasis: a review. *F1000Research*, 6:750–750, May 2017. ISSN 2046-1402. URL https://pubmed.ncbi.nlm.nih.gov/28649370. 28649370[pmid].

Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.