Navigating the Molecular Maze: A Python-Powered Approach to Virtual Drug Screening

John Raicu [†] Valerie Hayot-Sasson (Advisor) [†] Kyle Chard (Advisor) [†‡] Ian Foster (Advisor) [†‡]

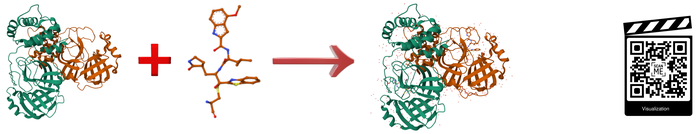
[†]University of Chicago [‡]Argonne National Laboratory

Abstract

The COVID-19 pandemic has highlighted the power of using computational methods for virtual drug screening. However, the molecular search space is enormous and the common protein docking methods are still computationally intractable without access to the world's largest supercomputers. AI methods provide a powerful new tool to help guide docking campaigns. In such approaches, a lightweight surrogate model is trained and then used to identify promising candidates for screening. We present ParslDock, a Python-based pipeline using the Parsl parallel programming library and the K-Nearest Neighbors machine learning model to screen a huge molecular space of molecules against arbitrary receptors. We achieved a 38X speedup with ParslDock compared to a brute-force docking approach.

Problem Statement

- **What is Protein Docking?** Predicting the optimal binding conformation of a protein receptor and ligand using a binding affinity scoring function



- **What are the challenges?** Machine learning model accuracy, sampling efficiency, and computational cost and complexity of docking workflow
- **What is the problem?** Identify the "best" ligands from a large dataset of potential molecules by efficiently combining simulation and machine learning algorithms on high performance computing resources

Background

Hydroxychloroquine SMILES String

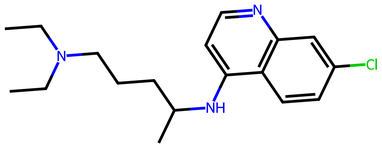
Example:
CCN(CCCC(C)NC1=C2C=CC(=CC2=NC=C1)Cl)CCO

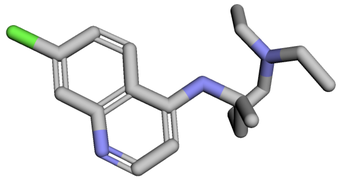
Explanation:
The simplified molecular-input line-entry system (SMILES) uses chemical notation to represent the structure of a molecule visualized in 2D below.

Hydroxychloroquine Fingerprint

Example:
11100100111101011111001111011011011111111001111111100000100110101

Explanation:
Molecular fingerprints are bit-vectors that help a machine learning model map a molecule description to a docking score.





Experimental Setup

Programming Tools

- Python 3.8.3 implements the computational pipeline
- Parsl 1.3.0.dev0 parallelizes various stages of the computational pipeline
- Jupyter Notebook 6.5.4 runs the Python code of the pipeline

Libraries

- AutoDock Vina 1.2.3 utilizes a scoring function and gradient-based optimization algorithm
- Visual Molecular Dynamics 1.9.3 visualizes and analyzes molecular simulations; Py3Dmol 2.0.3 enables interactive 3D molecular visualization directly in web browser; Matplotlib was used for general visualization
- Scikit-learn 1.3.0 was used for the machine learning KNN implementation
- NumPy 1.24.3 and Pandas 1.5.3 was used for general data processing and analysis

Hardware

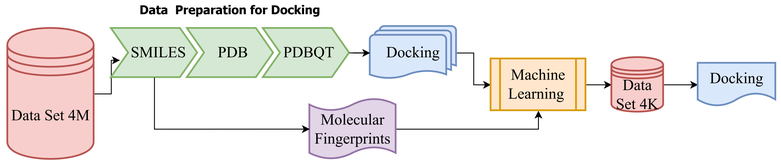
- 8c-laptop: 8-core Intel Core i9 CPU, 2.4GHz, 64GB DDR4, 8TB NVMe, MacOS 12.6.3
- 192c-server: 8x 24-core x86 Intel Xeon CPU, 2.1GHz, 786GB DDR4, 16TB SSD, Ubuntu Linux 22.04

Dataset

- 0.9 GB file containing four million ligands stored as SMILES strings

Proposed Solution

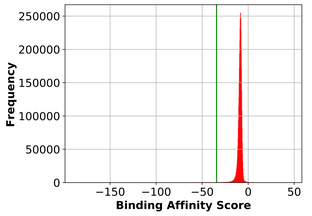
A python-powered automated pipeline that uses Parsl and machine learning to accelerate the docking process and improve resource utilization.



1. **Dataset 4M**: four million ligands represented by SMILES strings
2. **SMILES→PDB→PDBQT**: To prepare the data for docking, the SMILES strings are converted into PDB files and then into PDBQT files
3. **Docking**: Docking runs Monte Carlo simulations on the 1iep protein receptor PDBQT file with a ligand PDBQT file and outputs a binding-affinity score
4. **Molecular Fingerprints**: Morgan fingerprints are generated as a 128-bit vector with a depth of 8 from a SMILES string
5. **Machine Learning**: Morgan Fingerprints and docking scores are paired as the input to the machine learning model K-Nearest Neighbor (KNN)
6. **Dataset 4K**: four thousand ligands with the best docking scores (lowest binding-affinity scores)
7. **Docking**: Runs docking simulations on a smaller optimal subset of data containing four thousand ligands instead of four million

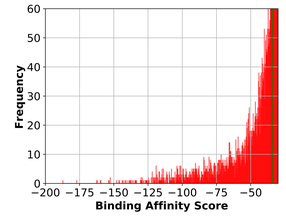
Results

Total Distribution of Docking Scores



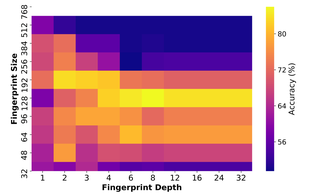
- Binding affinity scores have a normal distribution

Top 0.1% of Docking Scores



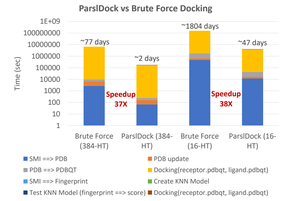
- Top-4k samples based on binding affinity score

Machine Learning Parameter Optimization



- KNN performance is sensitive to Morgan Fingerprint parameters (size and depth)
- Significantly better performance is achieved at a bit vector size of 128 and depth of 8.

ParslDock Performance Evaluation



- Up to 38X speedup on ParslDock vs. Brute Force Docking
- Linear scalability from 8-core laptop to 192-core server

Conclusions

- ParslDock: A Python-powered automated pipeline that uses Parsl and machine learning to accelerate the docking process, efficiently utilize compute resources, and reduce the time to discovery
- ParslDock achieves 38X speedup in performance that makes it possible to execute the virtual drug screening pipeline on a personal computer

References

[1] Yadu Babuji, Anna Woodard, Zhuozhao Li, Daniel S. Katz, Ben Clifford, Rohan Kumar, Lukasz Lacinski, Ryan Chard, Justin M. Wozniak, Ian Foster, Michael Wilde, and Kyle Chard. Parsl: Pervasive parallel programming in python. In *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, HPDC '19, page 25–36, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366700. doi:10.1145/3307681.3325400. URL `https://doi.org/10.1145/3307681.3325400`.

[2] Austin Clyde, Thomas Brettin, Alexander Partin, Hyunseung Yoo, Yadu Babuji, Ben Blaiszik, Andre Merzky, Matteo Turilli, Shantenu Jha, Arvind Ramanathan, et al. Protein-ligand docking surrogate models: A sars-cov-2 benchmark for deep learning accelerated virtual screening. *arXiv preprint arXiv:2106.07036*, 2021.

[3] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967. doi:10.1109/TIT.1967.1053964.

[4] Jerome Eberhardt, Diogo Santos-Martins, Andreas F Tillack, and Stefano Forli. Autodock vina 1.2. 0: New docking methods, expanded force field, and python bindings. *Journal of chemical information and modeling*, 61(8):3891–3898, 2021.

johnny.raicu@gmail.com