



**POLITECNICO**  
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE  
E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

## Combining Ligand-Based and Structure-Based Virtual Screening approaches for Efficient Drug Discovery experiments

LAUREA MAGISTRALE IN COMPUTER SCIENCE AND ENGINEERING - INGEGNERIA INFORMATICA

**Author:** SIMONE RIZZO

**Advisor:** PROF. GIANLUCA PALERMO

**Co-advisors:** DAVIDE GADIOLI, GIANMARCO ACCORDI

**Academic year:** 2022-2023

---

### 1. Introduction

Drug researchers have been enriching their pipelines with *in-silico* computer models. While *in-vitro* assays of candidate drugs, that test compounds on real cells and microorganisms, yield high-quality results, their throughput is not enough to screen the vast libraries of compounds available in the pharmaceutical industry. The solution to this problem is to start the drug research pipeline by employing *virtual screening* methods that can filter out compounds with low predicted activity. In particular, *molecular docking* is one of the most used computer models for the early discovery of drugs, representing a good compromise between accuracy and throughput. In docking simulations, a series of candidate compounds, named *ligands*, is positioned inside cavities of a protein, that needs to be inhibited to combat a target disease. Ligands that show good shape and chemical complementarities during these simulations are more likely to effectively interact and bind when dealing with real compounds and are thus selected for further analyses. HPC systems have reached exascale performance and are being used to aid this screening process. The recent pandemic demonstrated the need for urgent computing ca-

pabilities in drug discovery. However, with the decaying of Moore's law, simply allocating more resources is not enough to screen increasingly larger libraries of compounds when time is of the essence. Current drug research is focusing on ways to take advantage of computing resources more efficiently via software expedients. One of the possible approaches is to offload onerous computations to GPUs. This method was shown to allow significant speedups, in the order of 64×, in the EXSCALATE molecular docking platform [3], without any loss of precision. On the other hand, when a marginal loss in accuracy is allowed, approaches from the approximate computing field emerge. A possible solution was shown with GeoDock-MA [2], which can be tuned to approximate geometric conformations of candidate drugs, reaching a temporal improvement of over one order of magnitude with an accuracy degradation of only 30%. Finally, another common approach is *iterative screening* [4], where screening results from an initial selection of compounds are used to select subsequent batches of other molecules, guiding the exploration toward a promising chemical space.

The goal of this work is to reach even higher

virtual screening throughput to sustain drug discovery from massive libraries of compounds. A new virtual screening pipeline is proposed, applying optimizations along multiple directions. Previous works focused on single aspects of the screening pipeline, but experimental results suggest that combining optimization is beneficial for the whole process. In particular, the core optimizations explored by this thesis are applied to both ligands and proteins and are:

**Ligand filtering:** an initial set of ligands is regularly screened against the target protein and their results guide a selection of other ligands that are expected to maximize binding activity. This optimization can increase the throughput by discarding compounds that are unlikely to interact with the target.

**Anchor Point filtering:** each target protein may expose an arbitrary number of cavities where ligands can hook and interact. Molecular docking algorithms place the ligands in specific positions that are called *Anchor Points*, located in these cavities. The Anchor Points represent the initial position from where docking tools start the local optimization algorithm. By performing an initial screening, it is possible to measure how strongly an Anchor Point binds with specific classes of molecules. A speedup can be achieved by avoiding testing ligands on underperforming Anchor Points of the respective class.

In this thesis, the effects of both optimizations are studied, both independently and concurrently, reducing the computational load to achieve greater throughput for large libraries of ligands. The optimizations are automatically applied to satisfy time budget constraints.

This work is structured as follows: Section 2 gives a brief background on the fundamental concepts of computational chemistry; Section 3 explores previous attempts at optimizing virtual screening pipelines; Section 4 proposes the optimized screening solution; Section 5 examines the results of the experimental campaign; finally, Section 6 wraps up the work with the conclusions.

## 2. Background

This section will describe fundamental chemistry concepts for drug discovery, i.e. proteins and ligands, and will describe how they are used in

computer models.

### 2.1. Proteins and ligands

Some of the most studied objects in molecular biology are *proteins*. They are conglomerates of other smaller objects, named *amino acids*, and play a fundamental role in many cellular functions e.g. defending against pathogens, coordinating biological processes, and providing a structure for the body. When an unwanted object, a *pathogen*, infects the body, it does so by employing proteins that interact with the host. For example, in coronaviruses, several *spike proteins* are scattered on the virus's surface and allow it to bind to cells and infect them. Drug researchers are thus interested in proteins because if a molecule can inhibit a pathogen's protein, the source of disease can be eradicated. In particular, the proteins that are targeted by drug researchers are named *receptors*.

Molecules inhibit a receptor by binding to them, altering the shape of the protein and its behavior. The molecules that can bind to a receptor are named *ligands*, and they are usually much smaller than target proteins. This dimension difference is convenient for the pharmaceutical industry since smaller compounds require fewer steps to be manufactured and they are more likely to be absorbed by the body.

When searching for a drug, it is useful to consider the *affinity* and the *specificity* of the possible ligand-receptor complex. A high affinity implies strong interactive forces between compounds, while a high specificity implies that the involved compounds are less likely to bind to other compounds. The two aspects are in some ways correlated, allowing us to focus on only one of them, usually affinity, at least in the first stages of ligand screening.

### 2.2. Virtual Screening

Since the molecular space is immense and performing experiments with all compounds is intractable, computer models can be used to approximate the affinity between ligands and receptors. A model of this kind is relatively fast and scalable compared to in-vitro methods and is referred to as *high-throughput virtual screening*. Computer models for drug screening are divided into two fundamental categories:

**Ligand-based virtual screening:** which can

be always applied and revolves around the application of the *similar property principle*. The latter states that similar compounds have similar properties, i.e. ligands that are highly similar to known actives on a receptor are more likely to exhibit biological activity on the same receptor. From an initial set of active molecules and, optionally, inactive ones other molecules are promoted for further analysis or discarded depending on *similarities*. The concept of similarity depends on the specific methods used. Notably, a common approach is to transform molecules into *fingerprints*, which encodes in a bit-vector the features of the molecules. The fingerprints are compared with appropriate distance metrics and the binding affinity of any molecule to a certain receptor can be inferred.

**Structure-based virtual screening:** which can be applied only when the 3D structure of the receptor is known, but produces high-quality results. The most used method of this type is *molecular docking*, where multiple binding poses are simulated and evaluated. Both the receptors and ligands have a certain degree of flexibility. However, ligands are much smaller than receptors and their configurations can be explored, whereas receptors are usually considered rigid to prevent excessive computational load. In such approaches, also known as *semi-flexible docking*, any ligand can assume different conformations, depending on its number of *rotatable bonds*. The two parts of the molecule joined by such a bond can freely rotate. An initial *search phase* finds plausible poses by rotating bonds and displacing the ligand on the cavities of the receptor, also known as *pockets*. A subsequent *scoring phase* evaluates poses by considering chemical and thermodynamics laws.

### 3. State of the art

This section will summarize the optimization directions of past works that aimed at increasing virtual screening throughput.

**GPU acceleration** techniques such as EXSCALATE [3] offload onerous parts of the screening pipeline to GPUs. These approaches typically focus on pure performance, but still dock and score the whole ligand dataset.

**Iterative screening** techniques perform accurate analysis, such as docking or in-vitro assays, on small sets of ligands and iteratively priori-

tize other ligands, maximizing similarity to active ones and dissimilarity to inactive ones. An increased throughput results from the possibility of stopping the full exploration until a fixed number of high-scoring molecules is found. However, these techniques do not typically tune computations to satisfy time constraints.

**Approximate computing** optimizations such as GeoDock-MA[2] inspect which parts of a screening pipeline marginally contribute to the screening quality and approximate them. These approaches typically focus on trade-offs between performance and quality, but still dock and score all the input ligands.

The hinted optimizations are independent and technically compatible with each other.

## 4. Proposed Methodology

This thesis proposes a pipeline to increase theoretical virtual screening throughput by intelligently filtering out ligands and binding sites while respecting user-defined time budgets dedicated to molecular docking. The increase in throughput is obtained by combining ligand-based and structure-based virtual screening approaches. Figure 1 depicts the main stages of the devised pipeline, highlighting in yellow stages that employ ligand-based methods and in green stages that employ structure-based ones.

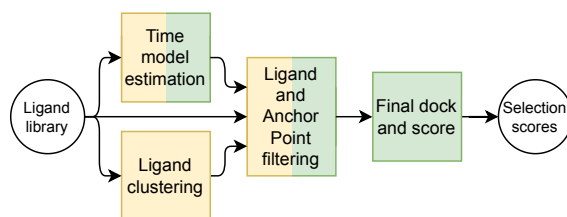


Figure 1: Stages of the proposed pipeline

The first stages can be run once for multiple pockets, and are thus *target-independent*. Their objectives are to estimate the time to dock and score any ligand on a target pocket and to elect a selection of ligands as representatives of the whole ligands library. The last stages are instead *target-dependent* and the core idea behind them is to *pre-screen* the representative ligands to infer the quality of other ligands on specific Anchor Points of the target pocket. The user can define how much time to allocate for molecular docking and how many Anchor Points should be kept, approximating results but achieving higher

screening throughput. The quantity of ligands to dock and to score is automatically computed to respect the time budget.

The logic is written in Python 3.10 and the selected suite of virtual screening software is LiGen [1]. Still, nothing prevents us from using the same pipeline with any other program able to expose information on Anchor Points.

The rest of this section will explain each stage.

#### 4.1. Time model estimation

The goal of this stage is to provide a dock and score time model for the filtering stage so that the user-defined time budgets are respected. The time to dock and score each molecule spans over an order of magnitude and it is useful to associate classes of molecules with their computational costs instead of considering them as equal. This stage is target-independent, but it is possible to train the time model on the target pocket to improve accuracy.

The first operation performed by the stage is to select different types of molecules. The chemical properties used to differentiate ligands are the count of rotatable bonds and heavy atoms. The number of rotatable bonds defines the flexibility of the ligand, allowing more torsional conformations. The number of heavy atoms, i.e. atoms that are not hydrogen, is instead indicative of the size of the molecule. Experimental data suggest a linear dependency between the complexity of a molecule and its time to score it. In this context, the complexity corresponds to the number of rotatable bonds times the number of heavy atoms.

The second operation of this stage is to dock and score each type of ligand on a sample pocket, keeping track of the processing time. A series of *buckets* with a sufficient number of ligands with shared properties is created and each of them is submitted to be docked and scored.

Finally, the time to dock and score a specific type of ligand is computed as the mean of the tested ligands of the same type. A linear regression is performed to obtain the linear coefficients of the model that predicts the time to dock and score, which is  $t_{lig} = \alpha \cdot H_{lig} \cdot R_{lig} + \beta$ , where  $t_{lig}$ ,  $H_{lig}$  and  $R_{lig}$  are respectively the estimated time, number of heavy atoms and number of rotatable bonds of the ligand.

The computed time model is employed by the

following filtering stage to determine the cost of docking and scoring input ligands.

#### 4.2. Ligand clustering

This stage aims to select a reduced set of ligands that can represent the whole library in the pre-screen. To accomplish this selection, ligands are clustered and a representative is chosen for each cluster. With this approach, the dataset can be partitioned into different groups, and diversity in the representatives helps cover the whole chemical space available. Moreover, molecules in a certain cluster are similar and thus likely to exhibit similar biological activity. Clustering does not depend on the target pocket but should be repeated each time the ligands library changes. Ligands are transformed into fingerprints, in particular 2048 bits Morgan ones generated by Rdkit, since common clustering algorithms work on vectors and not molecular graphs. An initial run of *k-means* clustering is performed on fingerprints and is accelerated by offloading the computation to an NVIDIA GPU, using a library named cuML. For each cluster, the ligand whose fingerprint is nearest to the centroid of the cluster is elected as its *leader*. Finally, all the fingerprints of the ligands library are compared to the leaders using Tanimoto distance, keeping track of how many ligands resulted as the most similar to each leader. This is done for coherency since the filtering stage uses Tanimoto similarities to associate molecules to leaders and does not predict ligand positions inside k-means clusters. Leaders and their relative cluster sizes are propagated to the filtering stage, which can estimate the computation load due to docking and scoring all the molecules from any cluster.

#### 4.3. Ligand and Anchor Point filtering

This stage selects ligands and associated Anchor Points that are worth docking and scoring. The user can define the maximum number of Anchor Points  $m_{thresh}$  to consider for each ligand, set by default to 3.

The first step is a pre-screening of the cluster leaders, which consists of an initial dock and score on the target pocket. The stage is thus target-dependent. The obtained scores are sorted to obtain a ranking of leaders. The ranking is traversed from top to bottom, computing

the time that would be necessary to dock and score the cluster associated with each leader, using the time model,  $m_{thresh}$ , and cluster dimensions. The computing complexity of all molecules in a cluster is assumed to be comparable to the one of the leader. The ranking traversal is stopped when the time budget is reached, keeping only *worthy leaders*. Tanimoto similarities are computed between each ligand and all leaders. If the most similar leader is worthy, the ligand is assigned to be docked and scored on the best  $m_{thresh}$  Anchor Points found during the pre-screening of the specific leader. The output of the stage is a *key-value mapping* where keys are Anchor Points and values are lists of assigned ligands.

#### 4.4. Final dock and score

This stage concludes the pipeline by performing a final dock and score of the molecules promoted by previous filtering. The stage is thus target-dependent, being strictly associated with the pocket used before. The assignments between Anchor Points and ligands are considered: for each single Anchor Point, all its assigned ligands are docked and scored on the specific Anchor Point. The results of this final molecular docking phase can be merged and analyzed for further investigation.

## 5. Experimental Evaluation

This section describes the results obtained from experimental data. The experiments were run using a selection of over 7 million ligands and 27 different pockets with a varying number of Anchor Points between 1 and 14. The experiments were on a machine with 2× AMD EPYC 7282 CPUs, 64 GiB of RAM, and 2× NVIDIA A100 GPUS with 40 GiB of VRAM each.

The proposed pipeline was tested in all its stages, when applicable, and as a whole. The time model estimation stage demonstrated a good fit between experimental data and the time model with its linear coefficients, resulting in an adjusted  $R^2$  value for the regression of 0.972. The proposed leader selection and clustering showed better results than random selection and clustering in terms of leader diversity, intra-cluster similarity, and low presence of excessively small clusters.

The filtering results are the most relevant and

can be considered indicative of the effectiveness of the whole approach since they allow the theoretical throughput to increase by lowering the computation load for docking and scoring. For each filtering test, the scores of the selected combinations of Anchor Points and ligands were compared to the scores that would have been obtained by perfectly selecting an identical amount of ligands and Anchor Points. The evaluation metrics deriving from these two sets of combinations were:

**Best score** which is a percentage representing the number of perfect combinations that were also found in the filtering selection.

**Score ratio** which is computed as the mean score of selected combinations over the mean score of perfect combinations.

Of the two, *best score* is the most relevant, since it focuses on the selection quality, while *score ratio* can vary depending on the distribution of score values of the dataset and is less indicative. Here follow *best score* results for each filter type, compared to a random downsample baseline.

### 5.1. Anchor Point filtering

Figure 3 depicts the differences in *best score* when purely filtering Anchor Points, with varying parameters for the number of leaders and Anchor Points for each ligand. The proposed filtering outperforms the random baseline, which is represented with dashed lines. Notably, using 2 selected Anchor Points results in a similar performance to using 4 random ones, but halving the computational load.

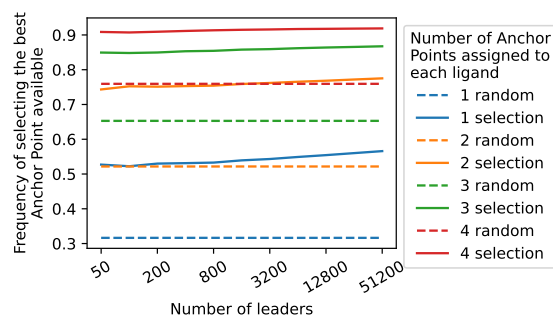


Figure 2: Mean results for Anchor Point filtering

### 5.2. Ligand filtering

Figure 3 depicts the differences in *best score* when filtering only ligands, with varying parameters for time budget allocated and count of leaders. The proposed filtering outperforms

a random selection of ligands, represented with a dashed line. Notably, in the case of a 20% relative time budget allocated (compared to exhaustive docking), the proposed solution selects ligands 2.3× better than the baseline.

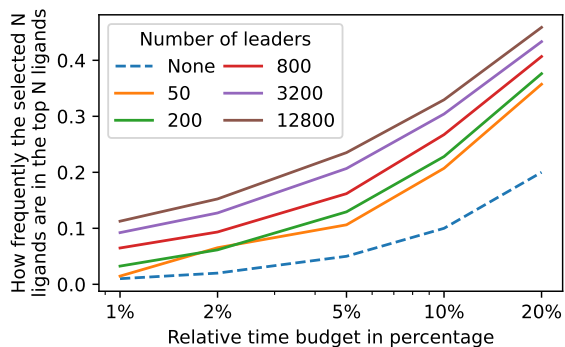


Figure 3: Mean results for ligand filtering

### 5.3. Combined filtering

Figure 4 depicts the differences in *best score* using full filtering, with varying parameters for time budget allocated and count of Anchor Points for each ligand, using 12800 leaders. The random baseline is represented with dashed lines. Combining filtering methods is beneficial, as visible by comparing results with the same time budgets and 12800 leaders in Figure 3. The approach is again better than random. For example, using 3 Anchor Points and 20% relative time budget allocated, full filtering selects ligands 2.1× better than the baseline. Alternatively, again with 3 Anchor Points, the quality of scores randomly selected by using 20% relative time budget is matched by the devised filtering, but using less than 5% of the time budget.

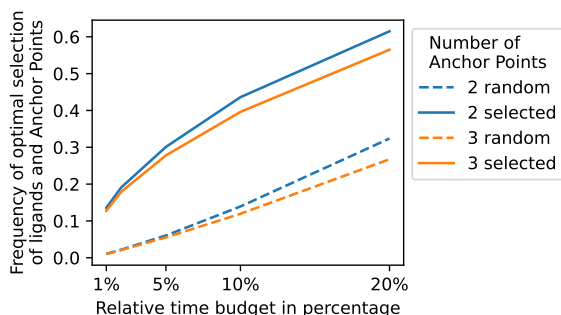


Figure 4: Mean results for full filtering

## 6. Conclusions

This thesis presents an optimized virtual screening pipeline, intending to increase processing

throughput by discarding ligands and binding sites without excessively sacrificing results quality. The pipeline combines ligand-based and structure-based virtual screening methods. Most notably, it performs an initial molecular docking step and uses its result to guide the filtering of other ligands using molecular similarity, automatically respecting time constraints. At the same time, it also reduces the Anchor Points to be evaluated for each ligand. The experimental evaluations demonstrate that, from an absolute point of view, reducing the time budget for docking by 80% results in a decrease in quality of only 30%. Compared to a random baseline, the pipeline achieves 2.1× better scores at a fixed time budget and requires 4× less computational resources to achieve the same quality.

## References

- [1] Claudia Beato, Andrea R. Beccari, Carlo Cavazzoni, Simone Lorenzi, and Gabriele Costantino. Use of experimental design to optimize docking performance: The case of ligendock, the docking module of ligen, a new de novo design program. *Journal of Chemical Information and Modeling*, 53(6):1503–1517, April 2013.
- [2] Davide Gadioli, Gianluca Palermo, Stefano Cherubin, Emanuele Vitali, Giovanni Agosta, Candida Manelfi, Andrea R. Beccari, Carlo Cavazzoni, Nico Sanna, and Cristina Silvano. Tunable approximations to control time-to-solution in an hpc molecular docking mini-app. *The Journal of Supercomputing*, 77(1):841–869, April 2020.
- [3] Davide Gadioli, Emanuele Vitali, Federico Ficarelli, Chiara Latini, Candida Manelfi, Carmine Talarico, Cristina Silvano, Carlo Cavazzoni, Gianluca Palermo, and Andrea Rosario Beccari. Exscalate: An extreme-scale virtual screening platform for drug discovery targeting polypharmacology to fight sars-cov-2. *IEEE Transactions on Emerging Topics in Computing*, 11(1):170–181, January 2023.
- [4] David E. Graff, Eugene I. Shakhnovich, and Connor W. Coley. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chemical Science*, 12(22):7866–7881, 2021.