Docking Billions of Molecules with Open-Source Software

Oliv Eidam[1] and Barbara Dravec[1,2]

1 CADD Consulting GmbH, Switzerland
2 Technit Discovery Engine, Switzerland

Email: oeidam@cadd-consulting.com
Website: cadd-consulting.com

CADD Consulting

1 Background

- **Molecular docking** is a simple and powerful approach to screen molecules in silico. In parallel, the **available chemical space for screening is growing fast**. Screening this space on in-house clusters becomes increasingly challenging. Not surprisingly, the use of docking in the cloud is popular. [1]
- In 2020, **an open-source drug discovery platform enabling large-scale docking screens (> 1 billion molecules)** was published. [2]
- In 2023, **VirtualFlow 2.0 – a platform enabling adaptive screens of 69 billion molecules** was published. [3]
- We were privileged to get early access to the "ready-to-dock" 69 billion Enamine library (to be released) and report first tests.

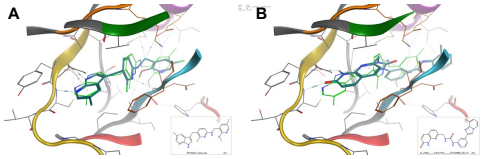
We report on the setup, speed, and cost of VirtualFlow 2.0, open-source software for large-scale docking.

3 Results

- We set up VirtualFlow 1.0 earlier in 2023 on our in-house cluster with 128 CPUs. That is where the numbers in Table 1 of the first primary screen come from. The cost for that screen is difficult to estimate. It is our opinion that buying hardware is more expensive than renting. VirtualFlow 2.0 setup and runs were done in May 2024 on AWS.
- In our tests, we docked 1.4 and 1.1 M molecules (11-16%) of the Sparse10 REAL library containing 8.7-87 M molecules. One molecule per tranche was selected. The primary screen against CSF1R was limited to 1 million molecules.

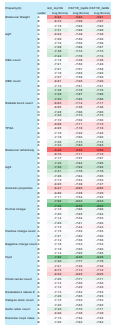
Target	Type of screen	# of mols	Speed (sec/mol)	vCPU hours	Cost ($)
GK	primary	1259523	4.9	1728	-
GK	ATG-VS	1392683	6.55	2534	51
CSF1R	ATG-VS	1114274	6.52	2020	40
CSF1R	primary	1000242	6.26	1738	35

Table 1: Docking speed and cost.
First screen was run on in-house cluster, screens 2-4 on AWS.
- On AWS, the **speed per molecule was 6.3±0.7sec/mol/vCPU**
- Based on this, the **compute cost to dock 1M molecules is $35±4**
- **Projected compute costs:**
  - Prescreen of 8.7 (or 87) M molecules: $303±33 ($3000)
  - ATG-derived primary screen of 50M: $1750±190.
  - Docking of 5 billion molecules: $175k±19k



Qvina2 docking poses (teal) in CSF1R.
A. Docking pose of PLX-5622 (positive control, teal color) overlaps well with crystallographic ligand (green) from pdb_id 6n33.
B. Docking pose of high-ranking diazanaphthalen-one forming important H-bonds to hinge of kinase. This scaffold could be an interesting variation of the well-known quinazolinones and an interesting molecule for testing.

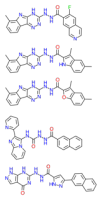
Visualization of average docking scores computed for each tranche of each property.
Color scheme: red-white-green **per** run.
Sparse matrices are the basis to select the most active tranches from the ATG Prescreen for the primary screen (step 3).



We display results of three runs screening different targets and tranches of Sparse10:
1. GK_ATG-VS: top 10k tranches
2. CSF1R_ATG-VS-1: top 8k tranches
3. CSF1R_ATG-VS-2: tail 8k tranches

Take-aways:
✓ GK has different scores than CSF1R
❖ Heatmaps are similar despite different targets and tranche selection in prescreen

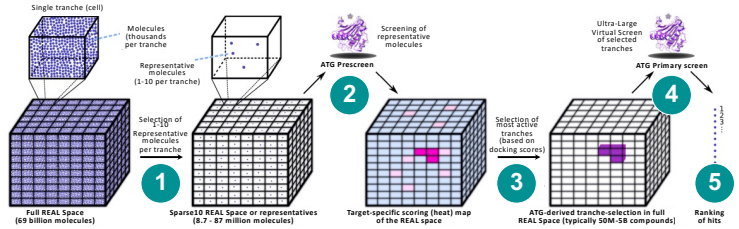
Analysis of top 500 molecules from CSF1R primary screen.
- 295 out of the best scoring 500 molecules are acetohydrazides (see chemical structures of examples on right) which may be problematic
- Best score is -15.1
- Xtal-lig scores -11.0 and would be ranked outside of the top 10'000 (top 1%) of the primary screen
- Top scoring azaindole ranks 469 (S -13.2)
- Good mols: cluster of diazanaphthalen-ones forms important H-bonds to hinge



Example structures of acetohydrazides

2 Methods

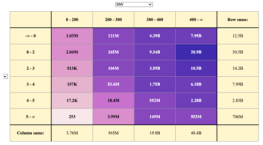
- We tested VirtualFlow 2.0 (VF2) using AWS Batch and cost-effective Spot instances (~$0.02/vCPU hour) and AWS Linux. In 2023, we tested VirtualFlow 1.0 (VF1) on an in-house computer with 128 vCPUs (AMD ThreadRipper), Ubuntu and SLURM.
- The docking method was always QuickVina 2.0 with exhaustiveness=1. Targets were glucokinase (GK from tutorial) and CSF1R.
- The screening libraries were Enamine REAL Space 2022q12 for VF2, and 2018q12 in VF1 tests, with 69 and 1.4 billion mols, respectively.
- Adaptive Target-Guided Virtual Screening (ATG-VS) is a new method in VF2 which reduces compute time/cost up to 5000 times. [3]



- Workflow adapted from Figure 3 in reference [3]
- Interactive 2D projection of 69 B molecule chemical space available at ref [5]

4 Discussion

- **Adaptive Target-Guided Virtual Screening (ATG-VS) enables cost efficient screening of 69 billion molecule space.** The effectiveness of the ATG-VS method to find validated hits remains to be shown experimentally.
- Note that **computational costs do not reflect the full costs of a virtual screen**. The time for the setup of VirtualFlow, target preparation, control calculations, custom library preparation (optional), result storage and analysis is not included.
- **ATG-VS merits to be investigated further**, especially with smaller, more lead-like, subsets of the Enamine REAL space, and perhaps more exhaustive sampling, or alternative engine.



VirtualFlow implementation and documentation are great. Setup requires good Linux knowledge. Docking speed per molecule was moderate. New ATG-VS method enables cost efficient screening of 69 billion molecule space.

References

[1] Tingle BI and Irwin JJ (2023). *JCIM*. Large-Scale Docking in the Cloud.
[2] Gorgulla C. et al. (2020). *Nature*. An open-source drug discovery platform enables ultra-large virtual screens.
[3] Gorgulla C. et al. (2023). *bioRxiv* [Preprint]. VirtualFlow 2.0 - The Next Generation Drug Discovery Platform Enabling Adaptive Screens of 69 Billion Molecules
[4] Github repository for VirtualFlow: https://github.com/VirtualFlow/
[5] VirtualFlow website: https://virtual-flow.org/

About CADD Consulting

We have 16 years of experience in virtual screening using docking. We have a track record of numerous validated hits from docking against various target, and we are experts in preparing focused or diverse screening libraries.

Contact us today
for your Large-Scale Docking screen on your target

CADD Consulting

Download this poster from our website at cadd-consulting.com