Scaling Virtual Screening to Ultra-Large Virtual Chemical Libraries

Spencer S. Ericksen
UW Carbone Cancer Center
Drug Development Core
Small Molecule Screening Facility

ssericksen@wisc.edu
Early-stage drug discovery

- Find rare molecules that affect a specific biological process. Develop as probes or drug candidates.

- Early-stage drug discovery is a needle-in-the-haystack problem—could be $10^{33}$ drug-like organic molecules.*

- High-Throughput Screening (HTS) is too expensive.

*Polishchuk PG, et al., JCAMD 2013 27(8):675-9
What is Virtual Screening?

- Virtual Screening: use a computer model to predict “active” molecules within large molecule sets.

- Structure-Based VS uses physics-based model to predict whether molecule will bind target protein

- Ligand-Based VS uses ML model to relate molecule structure to a property.

- Goal: reduce number of molecules that must be tested
HTS vs VS

Real Screening (HTS)

• test $10^4$-$10^6$ cpds
• generates valuable real data
• expensive
• noisy
• can’t scale to ultra-large libraries
• assay must scale to $10^4$-$10^6$

Virtual Screening + Real Focused Screening

• VS $10^8$-$10^{12} \rightarrow$ test $10^2$-$10^4$ cpds
• limited real data generation
• cheap
• **VERY** noisy
• scales to ultra-large libraries ($10^9$-$10^{12}$)
• VS models have data requirements
Size Comparison of Virtual and Physical Chemical Libraries


What is docking?

- Docking uses 3D molecular models to find best fit of molecule to active site of target.
- Search guided by a scoring function that evaluates favorability of each sampled configuration.
- Many docking programs are available.
- Docking score is crude estimate of binding favorability for a given compound.
Sort Compounds by Docking Scores

Dock Compound Library

<table>
<thead>
<tr>
<th>MOLID</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZINC42266748</td>
<td>58.72</td>
</tr>
<tr>
<td>ZINC6016217</td>
<td>56.21</td>
</tr>
<tr>
<td>ZINC6943749</td>
<td>56.14</td>
</tr>
<tr>
<td>ZINC27483165</td>
<td>49.65</td>
</tr>
<tr>
<td>ZINC13212603</td>
<td>49.39</td>
</tr>
<tr>
<td>ZINC5824695</td>
<td>48.79</td>
</tr>
<tr>
<td>ZINC20203103</td>
<td>48.72</td>
</tr>
<tr>
<td>ZINC21941263</td>
<td>48.46</td>
</tr>
<tr>
<td>ZINC11020165</td>
<td>48.35</td>
</tr>
<tr>
<td>ZINC03730730</td>
<td>48.29</td>
</tr>
<tr>
<td>ZINC15573870</td>
<td>48.29</td>
</tr>
<tr>
<td>ZINC23676750</td>
<td>48.23</td>
</tr>
<tr>
<td>ZINC28500076</td>
<td>48.16</td>
</tr>
<tr>
<td>ZINC8516771</td>
<td>48.13</td>
</tr>
<tr>
<td>ZINC21304840</td>
<td>48.11</td>
</tr>
<tr>
<td>ZINC20657708</td>
<td>48.10</td>
</tr>
<tr>
<td>ZINC20706070</td>
<td>48.07</td>
</tr>
<tr>
<td>ZINC04966068</td>
<td>48.05</td>
</tr>
<tr>
<td>ZINC20355597</td>
<td>48.03</td>
</tr>
<tr>
<td>ZINC21548393</td>
<td>48.02</td>
</tr>
<tr>
<td>ZINC25378232</td>
<td>47.98</td>
</tr>
<tr>
<td>ZINC27263206</td>
<td>47.96</td>
</tr>
<tr>
<td>ZINC26084921</td>
<td>47.92</td>
</tr>
</tbody>
</table>

Score Distributions

Actives

Inactives

Number of Compounds

Scores

Dock Compound Library

<table>
<thead>
<tr>
<th>MOLID</th>
<th>SCORE</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHEMBL323258</td>
<td>74.94</td>
</tr>
<tr>
<td>CHEMBL38532</td>
<td>74.19</td>
</tr>
<tr>
<td>ZINC36207525</td>
<td>69.07</td>
</tr>
<tr>
<td>ZINC14010625</td>
<td>68.48</td>
</tr>
<tr>
<td>ZINC21076300</td>
<td>68.36</td>
</tr>
<tr>
<td>ZINC61908006</td>
<td>66.40</td>
</tr>
<tr>
<td>ZINC64526095</td>
<td>65.96</td>
</tr>
<tr>
<td>CHEMBL419085</td>
<td>65.96</td>
</tr>
</tbody>
</table>

Structure-based virtual screening
Docking-based VS performance on 6 benchmark targets from DUD-E
Docking Compute Expense

- Compute time for docking depends on the search space, search quality, and complexity of the scoring function.
- To dock millions of compounds, we cut corners.
- Docking time varies between programs (~1 minute/compound).

<table>
<thead>
<tr>
<th>Program</th>
<th>Time (seconds)</th>
<th>Std. Dev. (seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AD4</td>
<td>435.6</td>
<td>197.1</td>
</tr>
<tr>
<td>Dock</td>
<td>719.2</td>
<td>592.9</td>
</tr>
<tr>
<td>Fred</td>
<td>15.6</td>
<td>5.7</td>
</tr>
<tr>
<td>Hybrid</td>
<td>9.3</td>
<td>2.9</td>
</tr>
<tr>
<td>Plants</td>
<td>43.4</td>
<td>20.5</td>
</tr>
<tr>
<td>rDock</td>
<td>49.3</td>
<td>26.7</td>
</tr>
<tr>
<td>Smina</td>
<td>250.1</td>
<td>172.8</td>
</tr>
<tr>
<td>Surflex</td>
<td>78.9</td>
<td>1159.6</td>
</tr>
</tbody>
</table>
Consensus Scoring

- No single program is reliable

- Use multiple docking programs

- Consensus scores are more reliable than those from any individual docking program.

Virtual screening performance on 21 benchmark targets

Target Class | Target
---|---
GPCR | ADRB1
GPCR | DRD3
Ion Channel | GRIA2
Kinase | BRAF
Kinase | CDK2
Kinase | PLK1
Kinase | SRC
Miscellaneous | FABP4
Receptor | ESR1
Receptor | ESR2
Other Enzymes | ACE
Other Enzymes | GLCM
Other Enzymes | HDAC8
Other Enzymes | HIVINT
Other Enzymes | PDE5A
Other Enzymes | PTN1
Protease | ADA17
Protease | FA10
Protease | HIVPR
Protease | MMP13
Protease | TRY1

P > 0.05

Pairwise t-test

PI Mitchell (Gitter/Hoffmann co-Pis)
https://research.wisc.edu/funding/uw2020/round-3-projects/an-adaptive-computational-pipeline-drug-discovery/

DOI: 10.1021/acs.jcim.7b00153
How do we scale with HTC resources?

- Each docking run is independent—*pleasently parallelizable*!
- Typical docking codes don’t benefit from specialized hardware or multiple cores.

- To maximize throughput:
  - Enable “Flock” and “Glide” to access more nodes.
  - Split compound library up into small chunks.
    - Number of compounds should run in ~2hr for a given docking program.
    - Chunk size varies from 5—500 compounds!
  - Dock each chunk on a single slot to scavenge ANY open slots. Dock compounds in chunk serially.
  - Checkpointing is enabled and a wrapper script is used to track the compounds completed in case job is evicted and migrates to another node.
How does SBVS benefit from HTC?

• Couldn’t really see how docking-based VS works without proper testing/validation!

• Examine performance over many targets

• Benchmarking of different docking programs

• Extensive docking parameter testing/validation

• Dock large compound sets
  • Routinely perform SBVS on libraries of 10-40 million cpds

• Hypothetical 100 node cluster = 3.5 million/day
  —100s of millions to BILLIONS of dockings!
ligand-based virtual screening

LBVS
Ligand-Based Virtual Screening—a ML hit-finding model

VS on Ultra-Large Virtual Chemical Library

Train RF model on prior screening data (PriA-SSB interaction)
- LifeChem Diversity Sets 1-3: 75,000 cpds (primary and retest)
- LifeChem Diversity Set 4: 25,000 cpds (primary only)
- MLPCN (NIH probe set): 337,000 cpds (primary and retest)

Training Data: 427,000 cpds, number of actives: 554 (hit rate = 0.13%)

**VS Procedure**
- Download Enamine REAL database 1.1 billion molecules (Oct 11, 2019)
- Split library up into 18 batches (each 60.3 million)
  - Average compute time of **3.24 ms per compound**
  - Mean run time per 60 million cpd batch = 53.2 hrs

https://enamine.net/compound-collections/real-compounds/real-database

Gitter Lab: Alnami M. et al., “Scalable supervised learning for synthesize-on-demand chemical libraries.” manuscript in prep
HTC is a fabulous resource for VS.

Effective VS requires rapid cycles of development, testing, validation. HTS enables this!

HTC allows VS to scale to new ultra-large virtual chemical libraries.
Acknowledgments

- CHTC Facilitators:
  Lauren Michael & Christina Koch

- Tony Gitter & Michael Newton
  - “A Machine Learning Platform for Adaptive Chemical Screening.” 1R01GM135631-01A1

- UWCCC-Drug Development Core
  Tim Bugni, Mike Hoffmann, Weiping Tang

- Computational Chemists
  Scott Wildman, Moayad Alnammi, Ken Satyshur