Scaling Virtual Screening to Ultra-Large Virtual Chemical Libraries

Spencer S. Ericksen
UW Carbone Cancer Center
Drug Development Core
Small Molecule Screening Facility
Why am I here?

- To promote early-stage drug discovery efforts on campus!

- Find active molecules that modulate therapeutically relevant mechanism. Develop as probes or leads.

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

- Conventional HTS approach too expensive.

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

- Virtual Screen: use a computer model to evaluate a chemical library. Prioritize some subset for testing.

- VS models predict potential for compound-target interaction or assay read-out.

- Goal is in enrichment for actives. Highly enriched subset reduces costs—enables focused screening.
## Searching chemical space for hits

<table>
<thead>
<tr>
<th>High Throughput Screening</th>
<th>Virtual Screening + focused testing</th>
</tr>
</thead>
<tbody>
<tr>
<td>• test $10^4$-$10^6$ cpds</td>
<td>• VS $10^8$-$10^{12}$ $\rightarrow$ test $10^2$-$10^4$ cpds</td>
</tr>
<tr>
<td>• generates valuable real data</td>
<td>• limited real data generation</td>
</tr>
<tr>
<td>• expensive</td>
<td>• cheap</td>
</tr>
<tr>
<td>• noisy</td>
<td>• VERY noisy</td>
</tr>
<tr>
<td>• can’t scale to ultra-large libraries</td>
<td>• scales to ultra-large libraries ($10^9$-$10^{12}$)</td>
</tr>
<tr>
<td>• Assay must be developed!</td>
<td>• VS models have data requirements</td>
</tr>
</tbody>
</table>
Virtual and Physical Chemical Libraries


NPMI = normalized ratios of principle moments of inertia

Structure-based virtual screening

SBVS
What is docking?

- Docking uses 3D molecular models to determine the optimal compound binding orientation on a given target.
- Search is guided by a scoring function that evaluates favorability of each sampled configuration.
- Many docking programs exist with different search strategies and scoring functions.
- Docking score is crude estimate of binding favorability for a given compound.
Structure-based virtual screening

Dock Compound Library

Sort Compounds by Docking Scores

Number of Compounds

Score Distributions

Inactives

Actives
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>
Raw docking scores from each program are normalized and then fed into a Consensus operation—this can be taking a mean, maximum, or median of the 4 scores.

If labeled data are available for target, “Supervised” consensus could also be used by weighting different program scores for optimal separation—like logistic regression, random forest, etc.
Virtual screening performance on N=21 benchmark targets

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

- Each docking run is independent—*pleasantly 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?

• Can’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
  • Recently performed SBVS on 8 million and 40 million cpd in-stock libraries
• Hypothetical 100 node cluster = 3.5 million/day
• 100s of millions to BILLIONS of dockings!
ligand-based virtual screening

LBVS
Ligand-Based VS—a ML hit-finding model

LBVS on Ultra-Large Virtual Chemical Library

Train RF model on prior screening data (PriA-SSB interaction)

- LifeChem Diversity Sets 1-3: 74,763 cpds (primary and retest)
- LifeChem Diversity Set 4: 25,278 cpds (primary only)
- MLPCN (NIH probe set): 337,104 cpds (primary and retest)

Total: 427,300 cpds, number of actives: 554 (hit rate = 0.13%)

VS Procedure

- Download Enamine REAL database 1.077 billion cpds (Oct 11, 2019)—SMILES format.
- Split library up into 18 batches (each 60.3 million)
- Run each batch as a single job on generic CPU compute node on HTCondor
  - Single core, ~5GB RAM, 32 GB disk
  - SMILES canonicalized/de-salted and converted to ECFP4 fingerprints
  - ECFP4 fingerprints scored by pre-trained random forest classifier.
  - Average compute time of 3.24 ms per compound
  - Mean run time per 60 million cpd batch = 53.2 hours (standard deviation=6.4 hr)

Gitter Lab: Alnammi M. et al., “Scalable supervised learning for synthesize-on-demand chemical libraries.” manuscript in prep
Dose-response testing of 68 compounds ordered from Enamine

Gitter Lab: Alnamm M. et al., “Scalable supervised learning for synthesize-on-demand chemical libraries.” manuscript in prep
VS for ultra-large virtual libraries

• LBVS with RF and fingerprints easily scales to 1.0 billion cpds

• SBVS study required 5 million CPU hours to evaluate 1.3 billion cpds.

• Another SBVS used 27,612 GPUs to score 1.37 billion in < 24 hours.

• We have applied SBVS in consensus docking screens with 6 programs on libraries up to ~40 million cpds.
TIERED APPROACH FOR SBVS

Synthesized-on-demand virtual library
10 – 100 billion cpds (SMILES)

1 – 10 billion cpds (SMILES)

0.1 – 5 billion cpds (3D)

5-200 million cpds

2-20k cpds

50-500 cpds

MD/ML-based affinity?

*$*$
Conclusions

HTC is a fabulous resource for VS.

Rapid cycles of development, testing, validation of VS

Scaling to ultra-large virtual chemical libraries.

Access to large numbers of GPU nodes might enable CNN-based scoring in docking or rigorous MD-based approaches for absolute ligand binding free energy.
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
- Score distribution of actives (red) is shifted relative to inactives (blue).

- Interestingly, the standard deviation in scores was also higher for actives than for decoys.