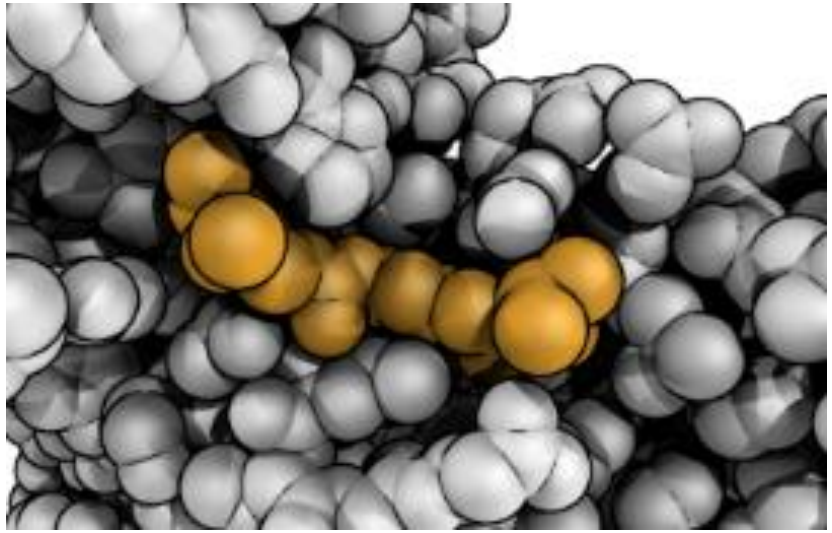
OSG Virtual School Showcase, August 11, 2021

Scaling Virtual Screening to Ultra-Large Virtual Chemical Libraries

Spencer S. Ericksen

UW Carbone Cancer Center
Drug Development Core
Small Molecule Screening Facility

sericksen@wisc.edu



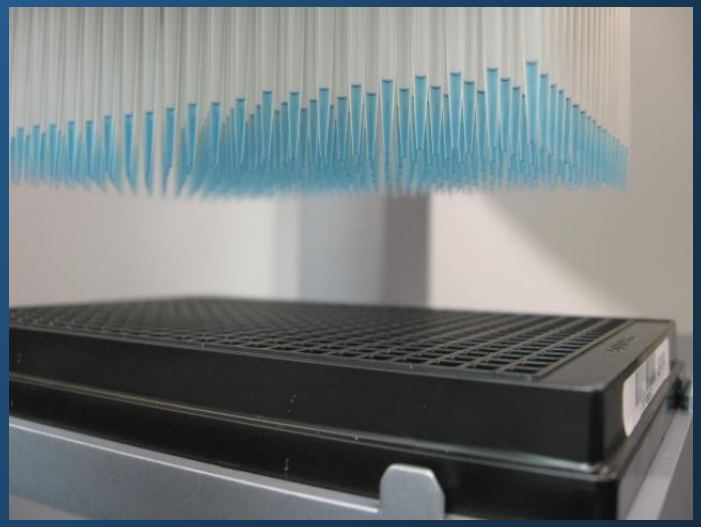
Carbone Cancer Center
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

Open Science Grid

CENTER FOR HIGH THROUGHPUT COMPUTING

HTCondor
High Throughput Computing

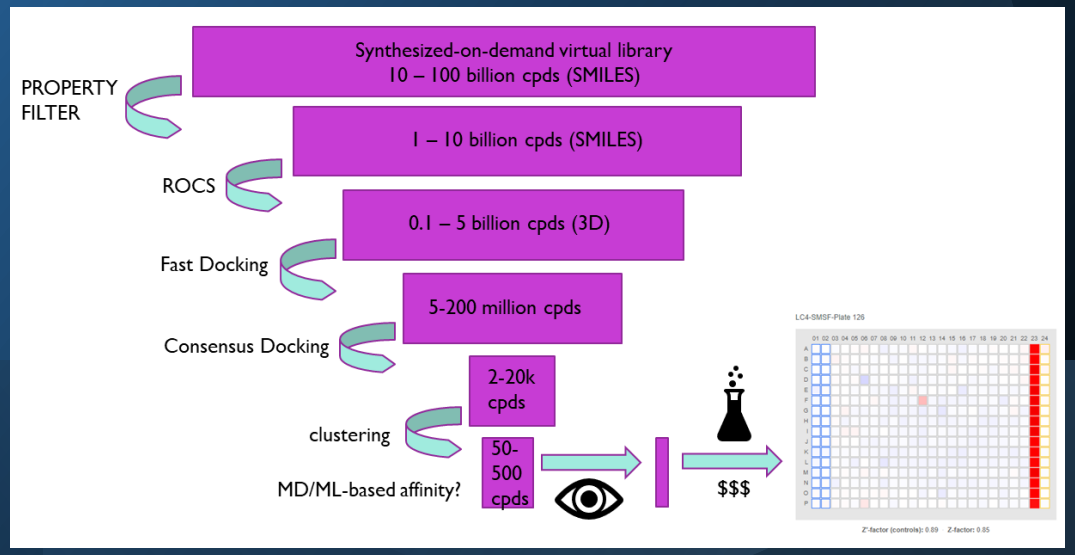
Early-stage drug discovery



- 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

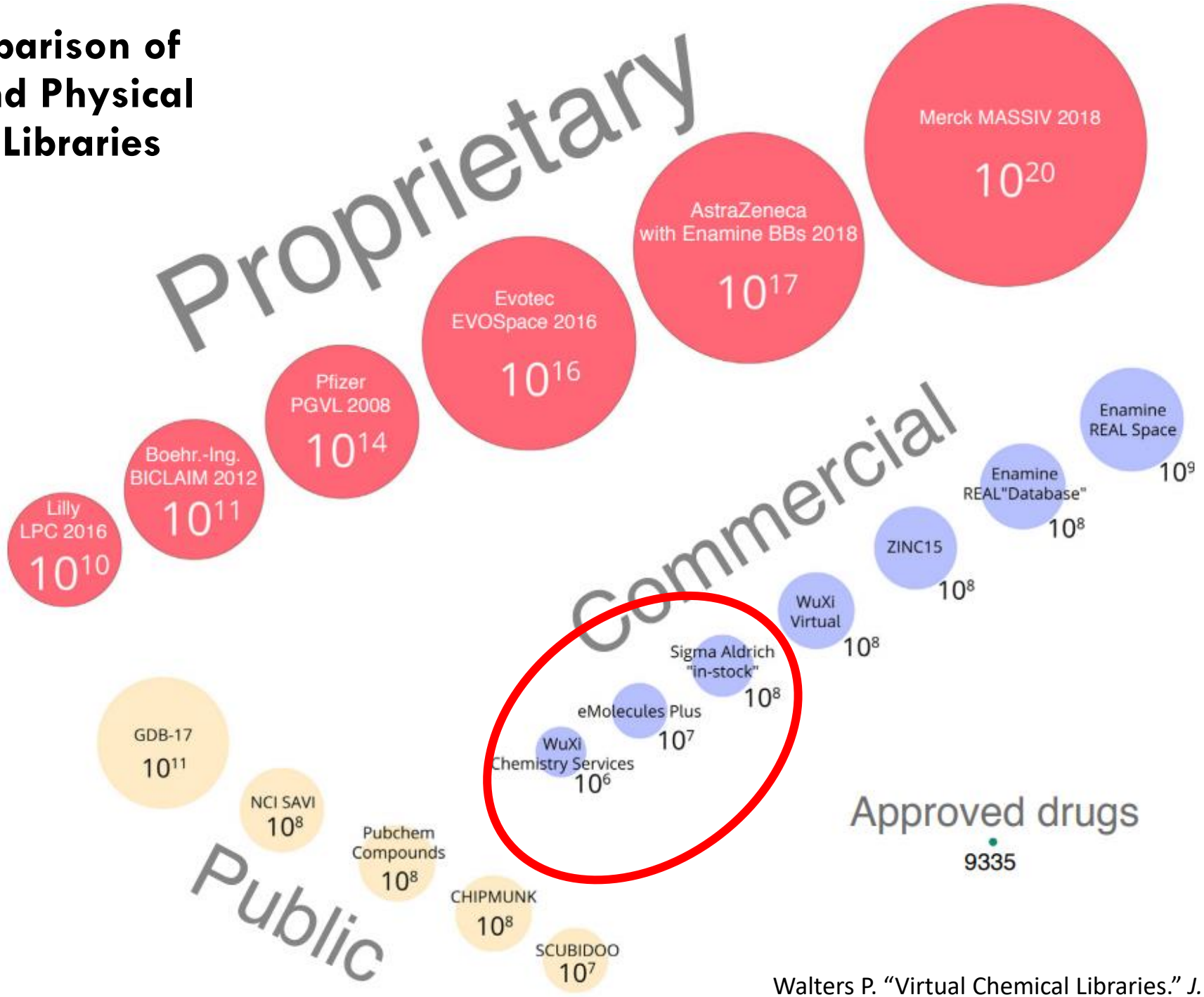
Real Screening (HTS)

- 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} → 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



Hoffmann & Gastreich "The next level in chemical space navigation: going far beyond enumerable compound libraries." *Drug Discovery Today*, 2019, 24, 5, 1148-1156.

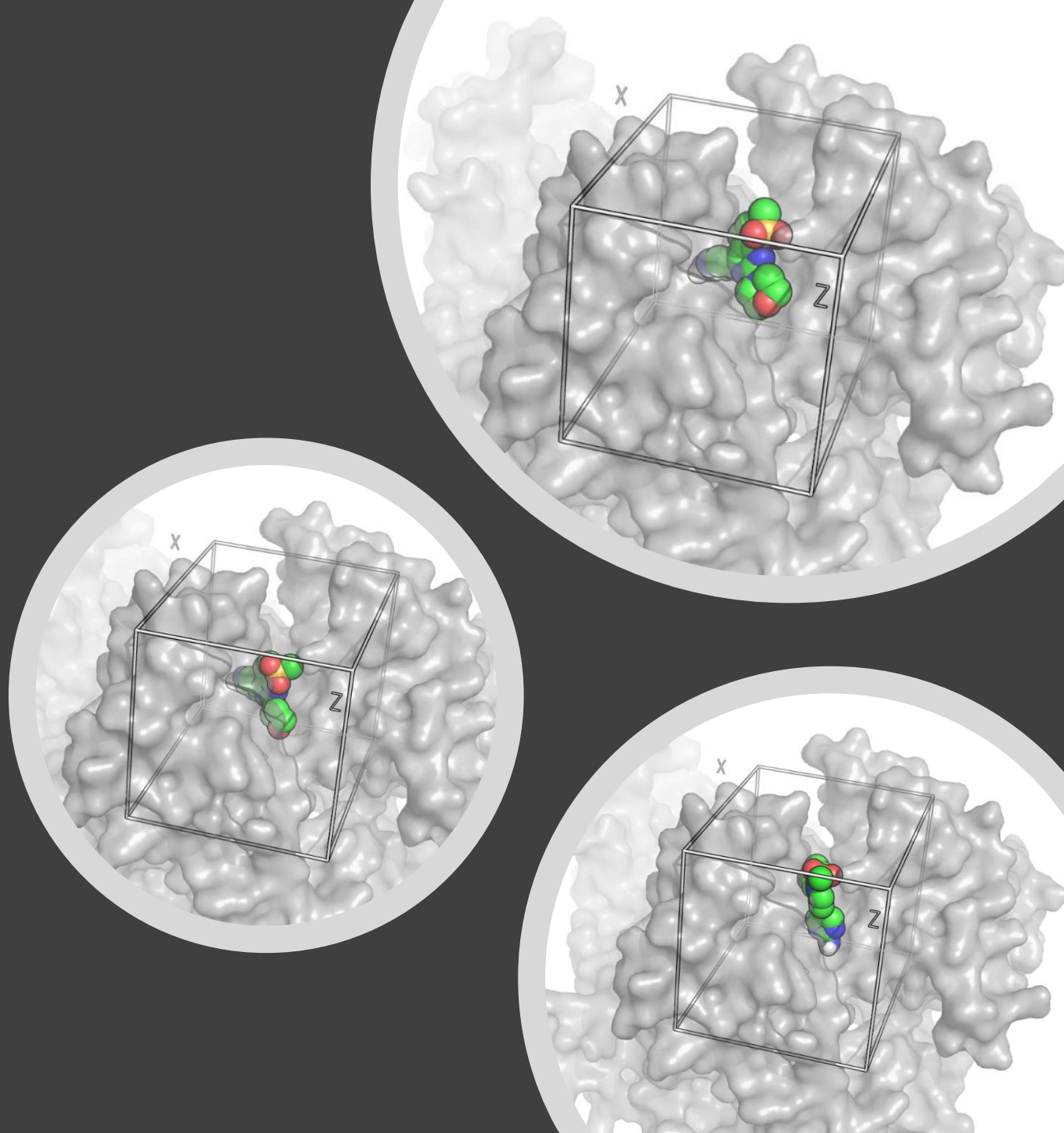
Walters P. "Virtual Chemical Libraries." *J. Med. Chem*. 2019, 62, 3, 1116-1124

Structure-based virtual screening

SBVS

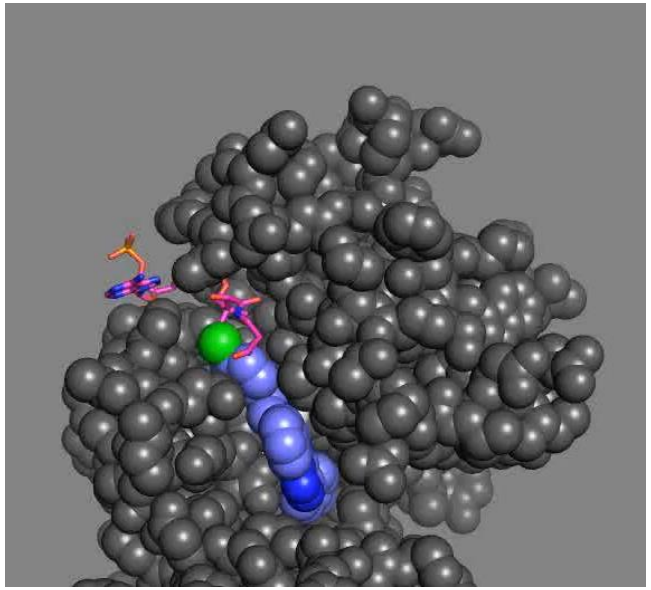
What is docking?

- 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.



Structure-based virtual screening

Dock Compound Library

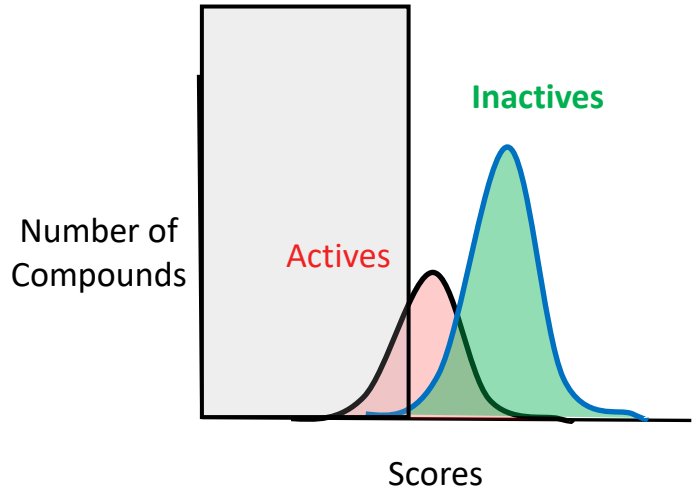


MOLID	SCORE
ZINC36206438	58.63
ZINC59310217	58.72
ZINC61596674	56.35
ZINC67458535	47.40
CHEMBL1221861	60.66
ZINC10123401	52.39
ZINC64526095	66.13
ZINC24002103	56.72
ZINC09612655	58.84
ZINC24002105	38.95
CHEMBL38532	74.19
ZINC40824467	50.10
ZINC59829723	58.29
ZINC37520295	44.78
ZINC49812309	38.01
ZINC14558020	53.31
CHEMBL472090	58.71
ZINC36207525	69.07
ZINC14010625	68.48
CHEMBL274782	63.97
ZINC63949457	55.35
ZINC39657146	48.74
ZINC23197109	58.72
ZINC25520953	63.14
ZINC09282496	43.71
ZINC60343267	62.18
ZINC58790750	62.53
CHEMBL400392	65.96
ZINC52096905	49.96
ZINC48922871	49.59
ZINC33058380	45.11
ZINC64684798	56.64
ZINC21076300	68.36
ZINC29461868	50.65
CHEMBL26183	58.56
ZINC61908006	66.40
ZINC15429053	54.10
CHEMBL323258	74.94
ZINC05091951	58.47
ZINC02759924	48.25
ZINC54596097	42.68
ZINC19899314	65.54
ZINC53113244	38.99
ZINC40947055	61.87
ZINC36611787	60.04
CHEMBL419085	65.96
ZINC35844701	58.57
ZINC01296699	39.07
ZINC39914438	49.68
ZINC00706129	48.34
ZINC34747432	52.55
ZINC43220997	47.45
ZINC37619890	54.49
ZINC15666896	55.50

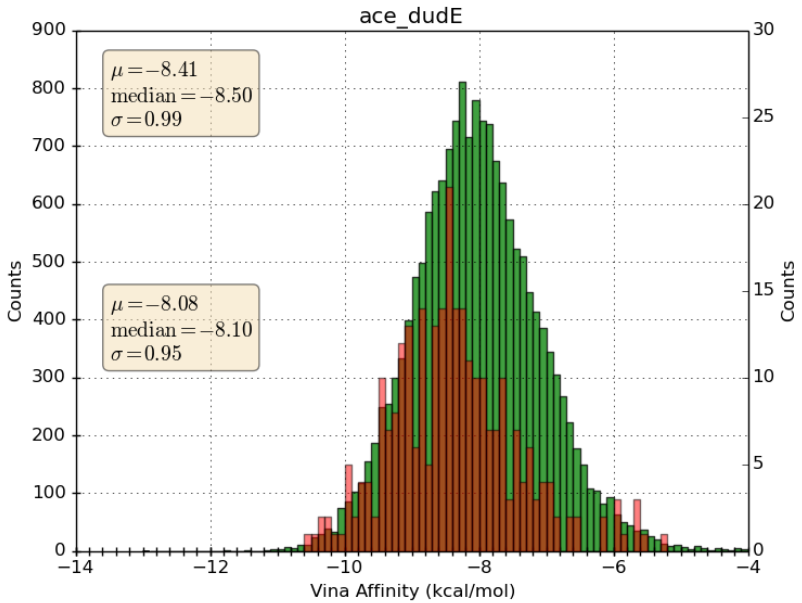
Sort Compounds by Docking Scores

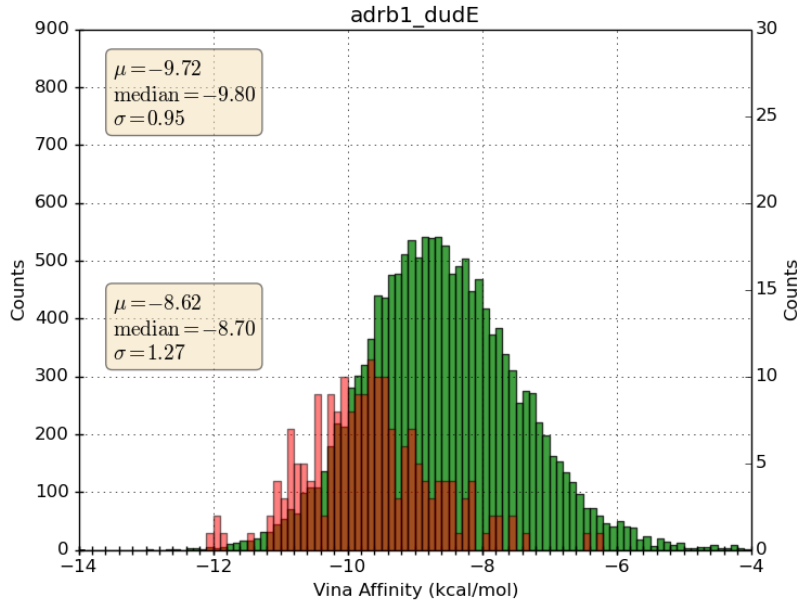
MOLID	SCORE
CHEMBL323258	74.94
CHEMBL38532	74.19
ZINC36207525	69.07
ZINC14010625	68.48
ZINC21076300	68.36
ZINC61908006	66.40
ZINC64526095	66.13
CHEMBL419085	65.96
CHEMBL400392	65.96
ZINC19899314	65.54
CHEMBL274782	63.97
ZINC25520953	63.14
ZINC58790750	62.53
ZINC60343267	62.18
ZINC40947055	61.87
CHEMBL1221861	60.66
ZINC36611787	60.04
ZINC09612655	58.84
ZINC59310217	58.72
ZINC23197109	58.72
CHEMBL472090	58.71
ZINC36206438	58.63
ZINC35844701	58.57
CHEMBL26183	58.56
ZINC05091951	58.47
ZINC59829723	58.29
ZINC24002103	56.72
ZINC64684798	56.64
ZINC61596674	56.35
ZINC15666896	55.50
ZINC63949457	55.35
ZINC37619890	54.49
ZINC15429053	54.10
ZINC14558020	53.31
ZINC34747432	52.55
ZINC10123401	52.39
ZINC29461868	50.65
ZINC40824467	50.10
ZINC52096905	49.96
ZINC39914438	49.68
ZINC48922871	49.59
ZINC39657146	48.74
ZINC00706129	48.34
ZINC02759924	48.25
ZINC43220997	47.45
ZINC67458535	47.40
ZINC33058380	45.11
ZINC37520295	44.78
ZINC09282496	43.71
ZINC54596097	42.68
ZINC01296699	39.07
ZINC53113244	38.99
ZINC24002105	38.95
ZINC49812309	38.01

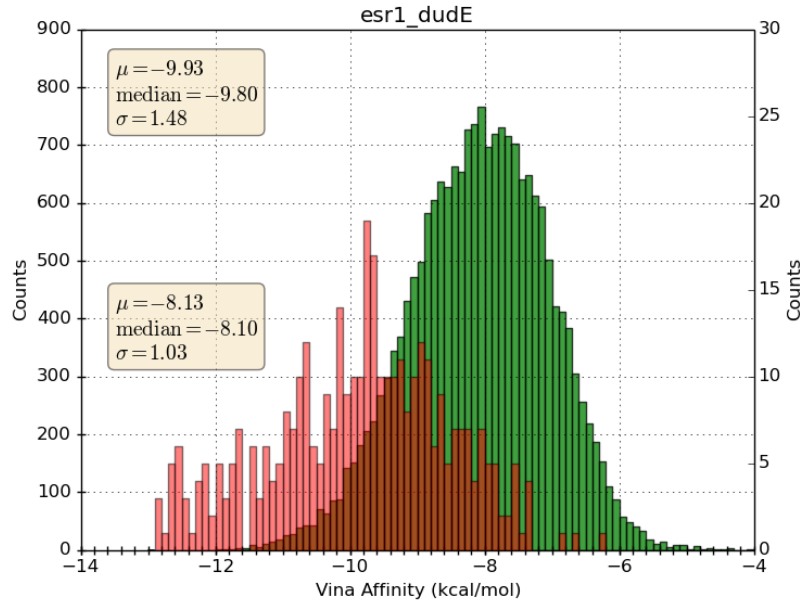
Score Distributions

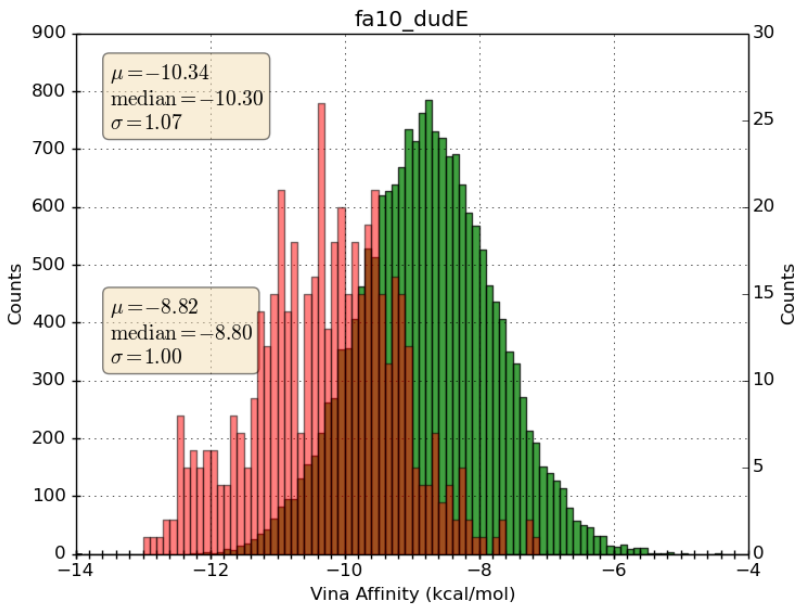


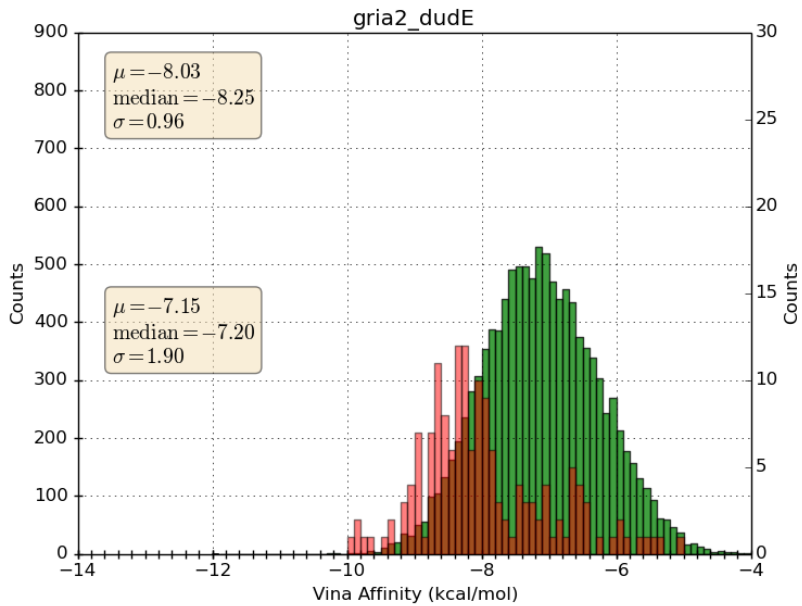
Docking-based VS performance on 6 benchmark targets from DUD-E

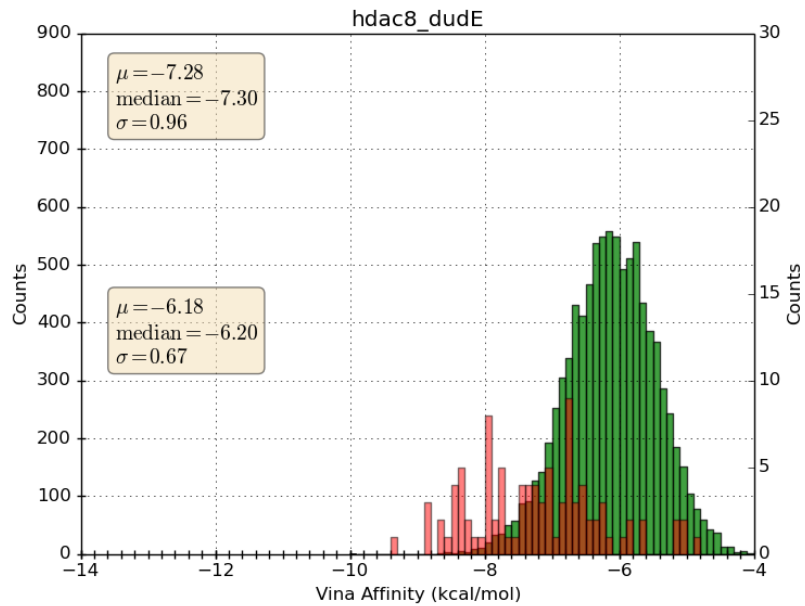






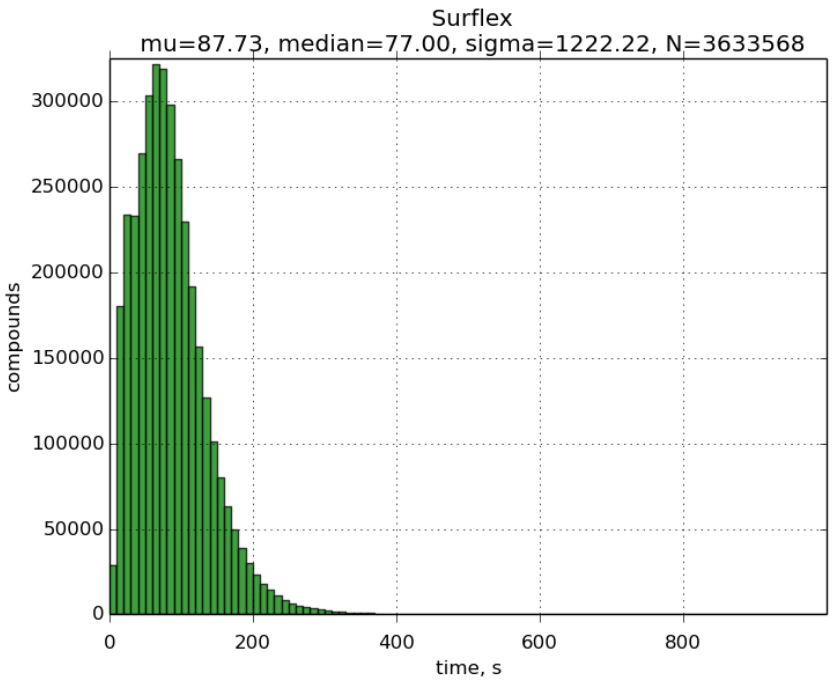


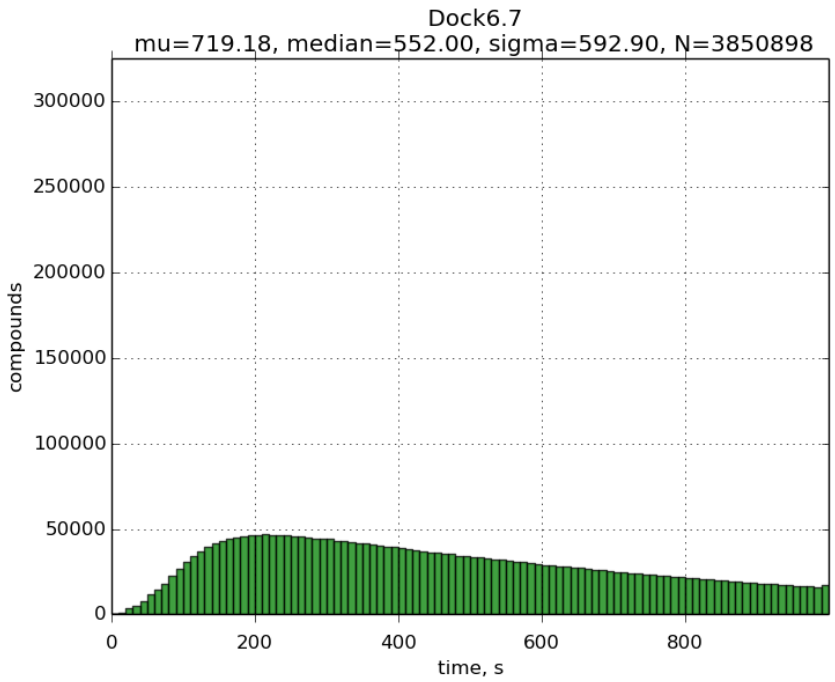




Docking Compute Expense

- Compute time for docking depends 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).



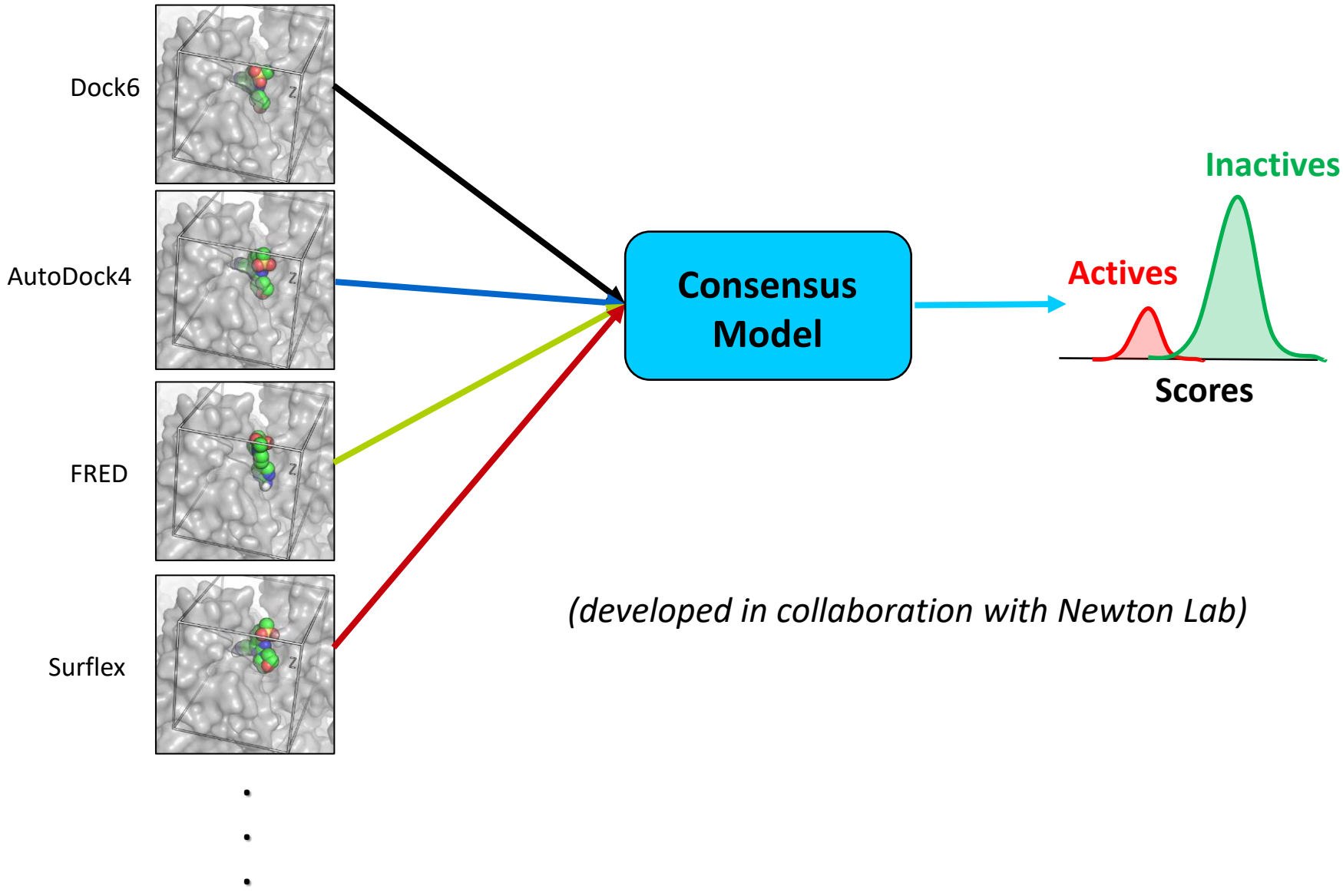


(seconds)

Program	Time	Std. Dev.
AD4	435.6	197.1
Dock	719.2	592.9
Fred	15.6	5.7
Hybrid	9.3	2.9
Plants	43.4	20.5
rDock	49.3	26.7
Smina	250.1	172.8
Surflex	78.9	1159.6

Consensus Scoring

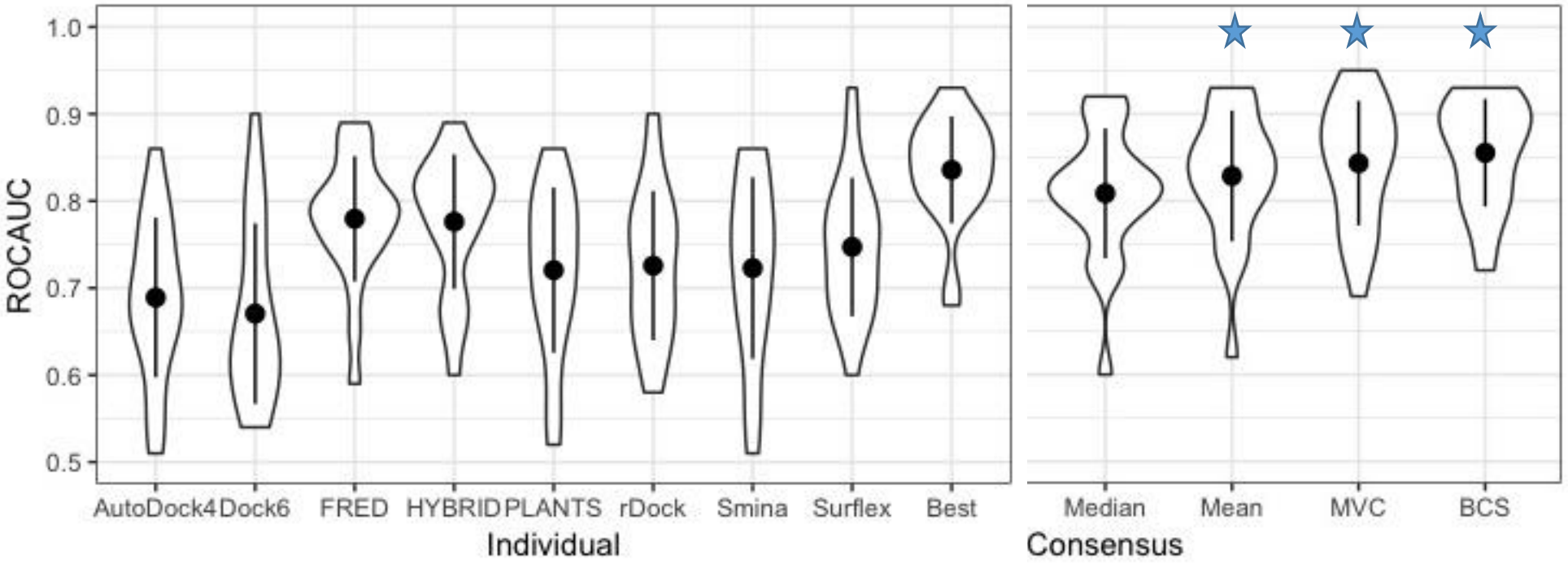
- No single program is reliable
- Use multiple docking programs
- Consensus scores are more reliable than those from any individual docking program.



(developed in collaboration with Newton Lab)

Ericksen et al., *J. Chem. Inf. Model*. **2017**, 57, 7, 1579-1590 DOI: 10.1021/acs.jcim.7b00153

Virtual screening performance on 21 benchmark targets



★ *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/

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

Ericksen et al., *J. Chem. Inf. Model*. **2017**, 57, 7, 1579-1590 DOI: 10.1021/acs.jcim.7b00153

How do we scale with 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.

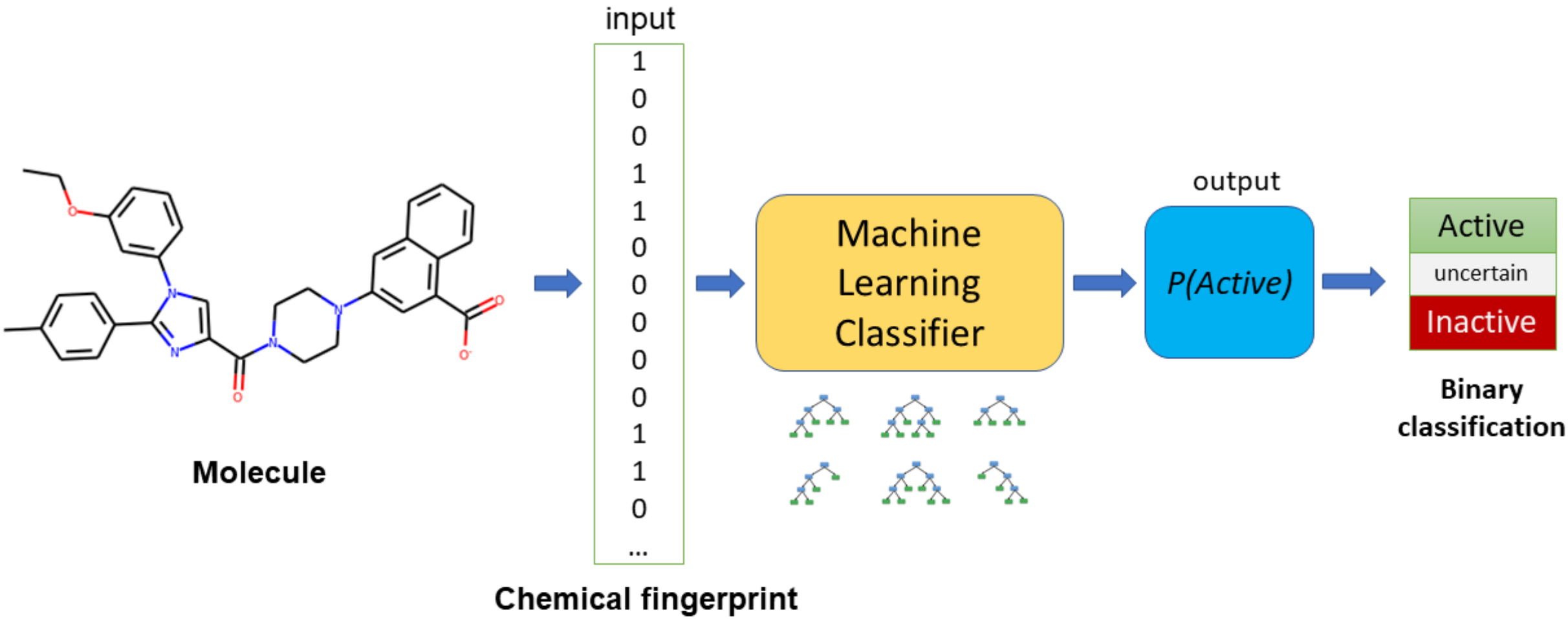
How does SBVS benefit from HTC?

- 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



Gitter Lab: Liu, et al., "**Practical Model Selection for Prospective Virtual Screening.**" *J. Chem. Inf. Model.* 2019, 59, 1, 282–293. https://doi.org/10.1021/acs.jcim.8b00363

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: 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

IC50 (uM)	C 0.5 (% ne	C 1.0	C 2.1	C 4.1	C 8.2	C 16	C 33	C 66	Active
0.5	93.2	40.8	21.0	11.6	8.1	6.6	5.9	8.6	1
2.1	78.2	66.2	45.9	25.2	10.3	7.8	7.1	6.0	1
0.5	116.4	69.3	50.5	24.2	11.4	8.8	7.1	10.3	1
3.4	79.7	74.1	63.1	33.9	9.9	6.0	6.2	5.3	1
4.4	85.3	78.5	72.0	47.9	11.8	4.5	3.5	3.4	1
4.8	72.8	63.2	54.6	43.7	33.6	19.4	10.5	8.3	1
4.9	85.7	77.9	70.4	54.4	19.6	8.0	5.6	5.7	1
6.1	80.1	67.9	67.8	53.2	33.5	15.1	9.0	8.6	1
6.8	88.1	83.0	81.0	68.8	35.0	10.4	5.5	4.5	1
9.1	85.0	81.5	81.0	73.0	50.9	14.3	6.5	6.2	1
3.9	119.8	86.5	82.0	68.2	47.7	15.7	7.3	9.8	1
2.4	126.3	91.8	81.4	72.5	47.3	19.4	7.6	8.5	1
10.8	86.8	84.3	79.5	76.0	58.9	24.2	8.8	6.2	1
0.5	132.0	91.3	82.4	73.1	56.0	24.0	9.9	9.5	1
12.4	85.5	85.9	85.4	79.8	63.8	30.9	8.0	5.0	1
48.2	81.6	75.7	67.9	60.5	55.0	45.8	30.8	25.8	1
66.0	121.5	88.3	82.6	75.7	63.1	37.5	9.1	7.2	1
12.6	87.7	87.0	82.8	83.9	71.1	28.1	8.0	6.1	1
12.1	89.2	89.7	86.1	80.1	66.8	30.2	10.3	6.7	1
11.5	91.4	87.2	87.5	84.5	67.7	26.5	10.9	6.3	1
15.5	83.3	88.0	77.3	74.0	62.2	41.1	24.7	12.3	1
35.4	123.9	94.4	86.1	81.9	68.7	40.8	9.4	8.4	1
13.9	87.7	86.1	87.4	85.7	72.9	38.0	11.4	9.3	1
21.0	86.6	83.3	82.5	75.9	67.8	47.8	24.9	9.2	1
15.7	87.8	89.6	89.8	83.9	70.5	43.8	15.8	5.7	1
7.0	80.0	78.9	78.9	71.7	53.7	44.8	39.8	36.9	1
66.0	125.3	92.1	86.0	86.3	74.4	49.5	15.4	10.8	1
22.2	90.0	90.6	92.1	86.8	76.4	62.5	22.7	7.2	1
37.8	88.5	87.6	87.0	80.1	77.6	62.3	36.5	11.0	1
24.9	84.6	87.5	88.0	89.1	80.0	65.2	29.9	8.9	1
66.0	140.0	94.7	97.5	91.0	83.1	62.3	25.1	10.1	1
20.2	86.5	85.7	83.3	80.3	80.3	68.4	50.3	43.6	0
66.0	136.0	100.0	89.3	86.0	86.7	74.1	49.7	28.3	1
66.0	83.7	86.8	87.7	82.8	76.5	63.1	60.3	65.1	0
66.0	86.9	87.8	85.9	86.7	78.5	78.8	71.7	69.4	0
7.3	131.5	93.7	90.9	90.4	90.5	82.4	70.5	49.2	0
66.0	122.0	95.1	90.0	84.9	88.9	84.5	74.2	63.6	0
66.0	83.3	83.6	80.6	82.6	82.0	86.1	84.0	86.0	0
66.0	84.2	82.1	84.5	83.4	83.3	81.6	84.1	87.9	0
66.0	134.3	95.6	87.6	85.4	85.2	84.7	77.4	72.8	0
66.0	134.5	97.4	87.2	83.1	84.6	86.1	77.9	74.2	0
66.0	136.5	93.6	88.8	89.0	87.8	89.4	78.2	63.8	0
66.0	85.7	86.8	83.5	80.2	87.7	82.1	84.7	87.2	0
66.0	89.4	88.1	89.9	89.3	81.7	82.2	80.6	80.9	0
66.0	87.3	84.5	88.6	83.2	86.9	81.6	85.0	88.7	0
66.0	86.8	85.6	84.8	86.7	86.2	83.9	88.4	85.0	0
66.0	134.3	100.4	87.4	88.0	89.0	87.8	77.2	71.0	0
66.0	89.2	88.1	87.4	89.4	88.2	84.2	82.2	82.8	0
66.0	95.1	87.1	84.9	87.5	87.3	88.0	86.8	81.4	0
66.0	83.7	85.4	88.7	88.1	86.9	86.8	85.6	84.4	0
66.0	131.0	93.0	91.7	84.5	87.9	87.0	82.7	79.7	0
66.0	129.3	95.0	89.2	86.5	90.2	87.5	80.7	78.1	0
66.0	88.1	88.0	83.6	85.4	83.6	89.5	86.7	90.7	0
66.0	136.3	93.7	92.4	86.6	91.3	81.6	85.5	78.3	0
66.0	138.0	96.4	92.7	85.1	89.7	82.9	84.5	78.9	0
66.0	85.0	88.3	91.9	88.0	87.6	88.7	84.8	86.7	0
66.0	87.8	88.8	90.8	88.7	88.9	84.7	89.5	84.7	0
66.0	88.2	88.0	88.1	89.4	88.9	91.0	90.2	87.0	0
66.0	130.5	94.1	93.9	93.5	93.2	86.5	81.2	81.4	0
66.0	119.8	91.6	93.1	90.0	88.0	87.7	90.8	84.4	0
66.0	126.3	96.0	92.6	91.1	92.1	87.8	85.4	81.5	0
66.0	89.3	90.8	91.0	89.7	86.5	88.6	88.2	91.7	0
66.0	141.3	99.5	92.2	89.1	91.5	86.7	87.8	82.3	0
66.0	131.5	95.4	94.9	91.6	95.0	88.0	87.6	85.4	0
66.0	125.3	94.6	94.9	89.9	90.1	93.7	88.5	87.4	0
66.0	133.0	96.7	92.9	91.7	93.3	91.3	91.6	97.2	0
66.0	85.7	90.2	92.9	92.2	89.1	92.9	96.6	101.8	0
66.0	140.8	101.8	93.2	93.1	97.6	108.8	125.5	170.5	0

Conclusions

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

CENTER FOR HIGH THROUGHPUT COMPUTING

NIH National Institutes of Health
Turning Discovery Into Health

Carbone Cancer Center
UNIVERSITY OF WISCONSIN
SCHOOL OF MEDICINE AND PUBLIC HEALTH

WARF
Wisconsin Alumni Research Foundation

OpenEye
SCIENTIFIC

Open Science Grid

NSF