Practical Model Selection for Virtual Chemical Screening

Shengchao Liu[1,7], Moayad Alnammi[1,7], Spencer Ericksen[2,3,8], Andrew Voter[4], James Keck[4], Michael Hoffmann[2,5], Scott Wildman[2], Anthony Gitter[1,3,6,7,8]

[1]Department of Computer Sciences; [2]Small Molecule Screening Facility;[3]Center for Predictive Computational Phenotyping; [4]Department of Biomolecular Chemistry;
[5]McArdle Laboratory for Cancer Research;[6]Department of Biostatistics and Medical Informatics; University of Wisconsin-Madison, Madison, WI
[7]Morgridge Institute for Research, Madison, WI

CPCP

MORGRIDGE INSTITUTE FOR RESEARCH

WISCONSIN UNIVERSITY OF WISCONSIN-MADISON

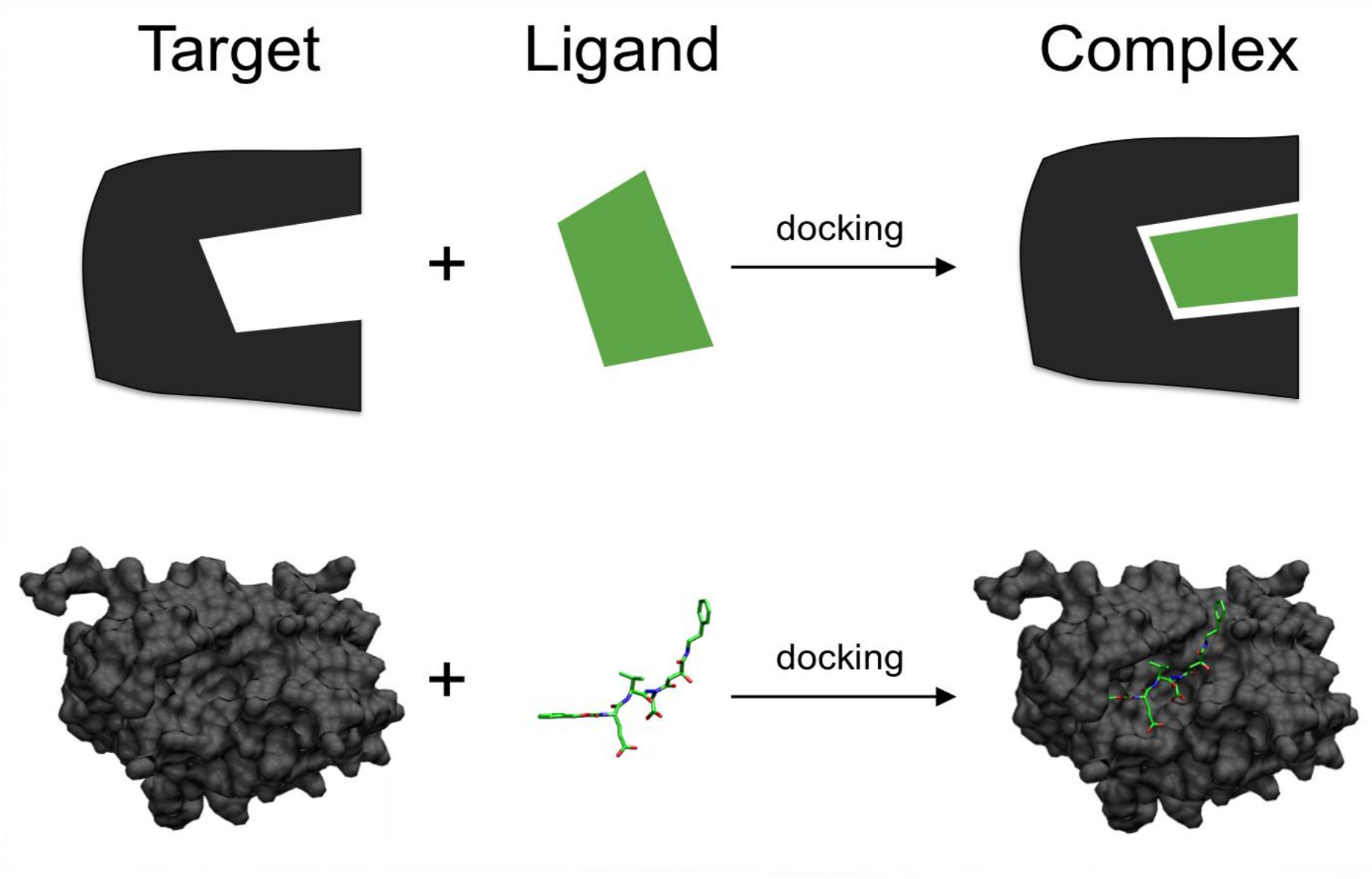
Introduction & Motivation

- **Problem:** Given a chemical compound and target protein, determine whether the compound binds with the target.
- Experimental tests in a small molecule screening facility are expensive.

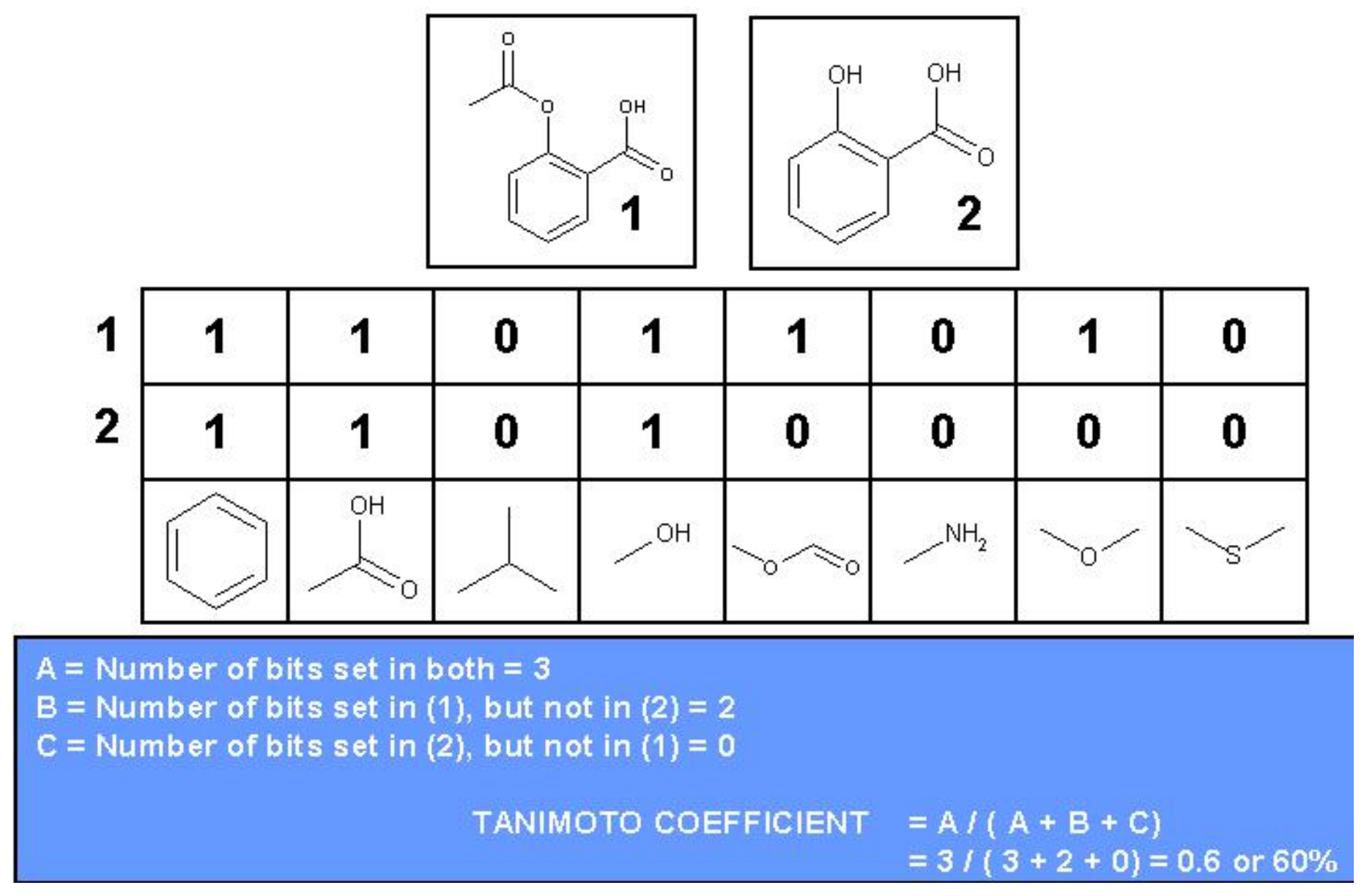
Virtual Screening (VS) can help accelerate drug discovery by **proposing the most probable** compounds for experimental testing.

Two Main VS Strategies

1. **Structure-Based:** docking methods that requires target structure info.
2. **Ligand-Based:** *similar* compounds bind similarly. No structure knowledge of target required.



SB Docking Concept. Figure from [1]



1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

A = Number of bits set in both = 3
B = Number of bits set in (1), but not in (2) = 2
C = Number of bits set in (2), but not in (1) = 0

TANIMOTO COEFFICIENT = A / (A + B + C)
= 3 / (3 + 2 + 0) = 0.6 or 60%

LB Fingerprint Concept. Figure from [2]

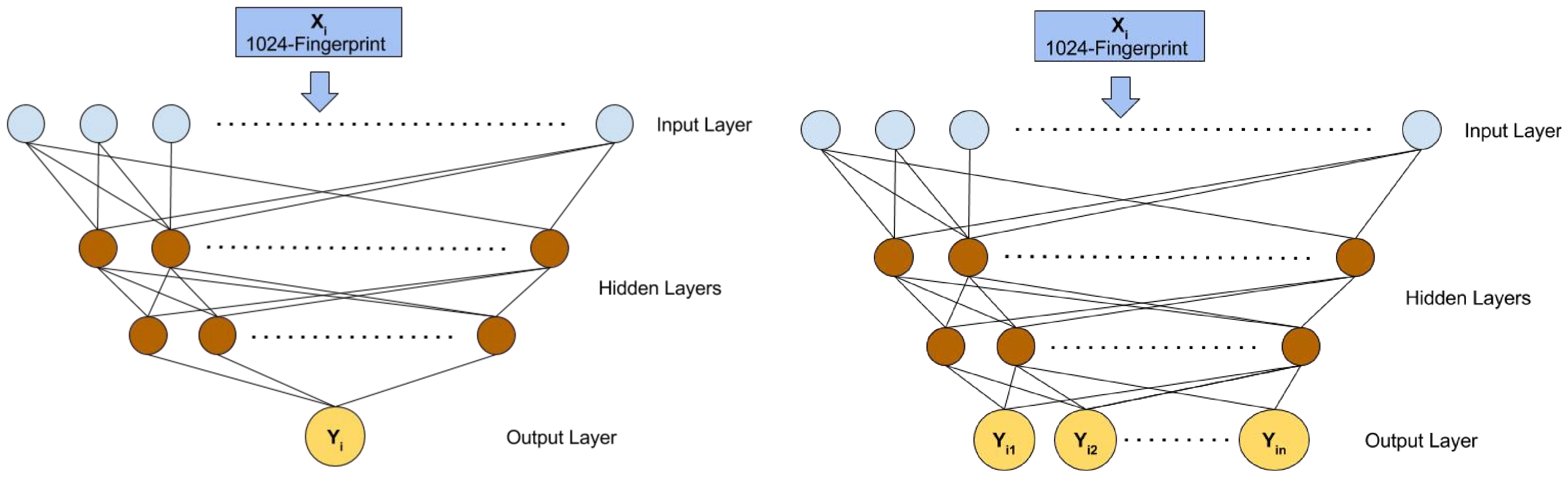
Case Study: SSB-PriA

- Keck lab screened 75000 compounds to see which disrupt the SSB-PriA interaction. **(known)**
- Untested library of 25000 new compounds. **(unknown)**

Goal: Assess **quality of MTNN and other common methods** on this unknown set. We are only given one chance. Also gives us a chance to assess **quality of evaluation metrics** as they translate to real world value.

Real-World Impact: Help screening facilities by proposing top 250 most likely compounds. Perfect ranking not important.

Single Task vs. Multi-Task Neural Networks



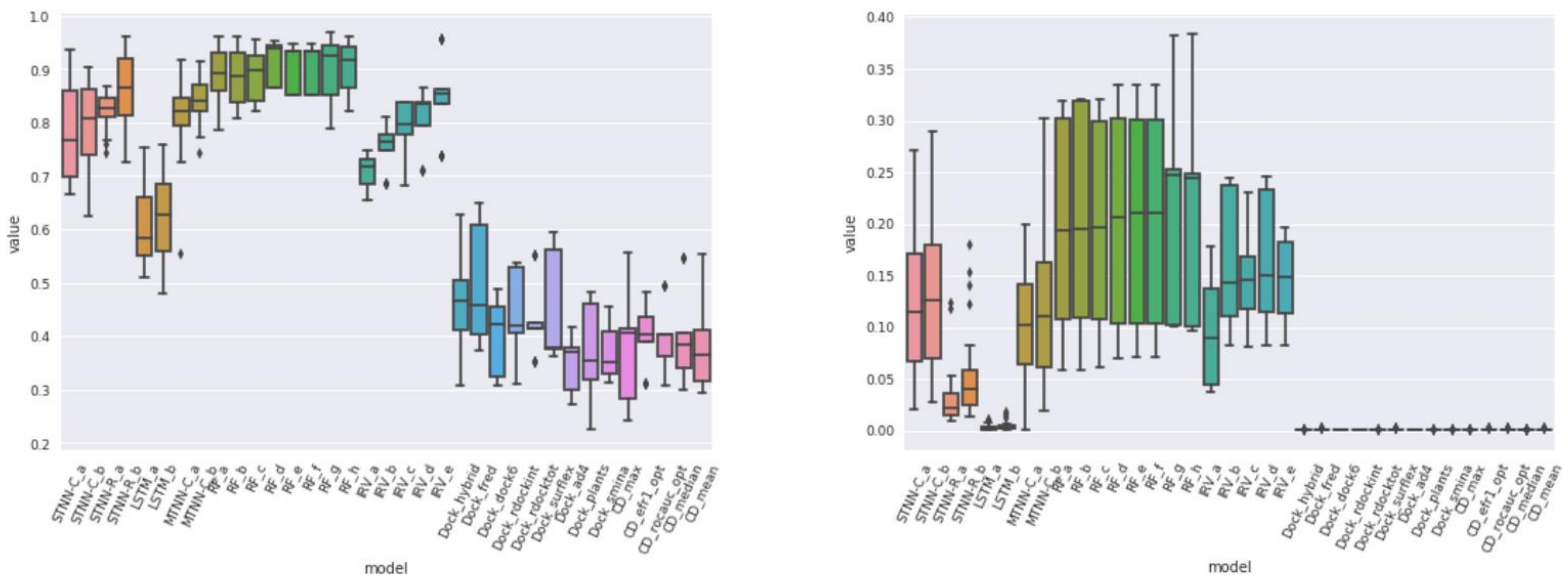
SSB-PriA

SSB-PriA Additional targets

Project Pipeline

- Stage 1: Hyperparameter Selection Stage, prune hyperparameter space
- Stage 2: Cross Validation Stage, select best model based on early enrichment
- Stage 3: Prospective Screening Stage, evaluate best models with new experiments

Cross Validation



AUC[ROC]

AUC[Precision Recall]

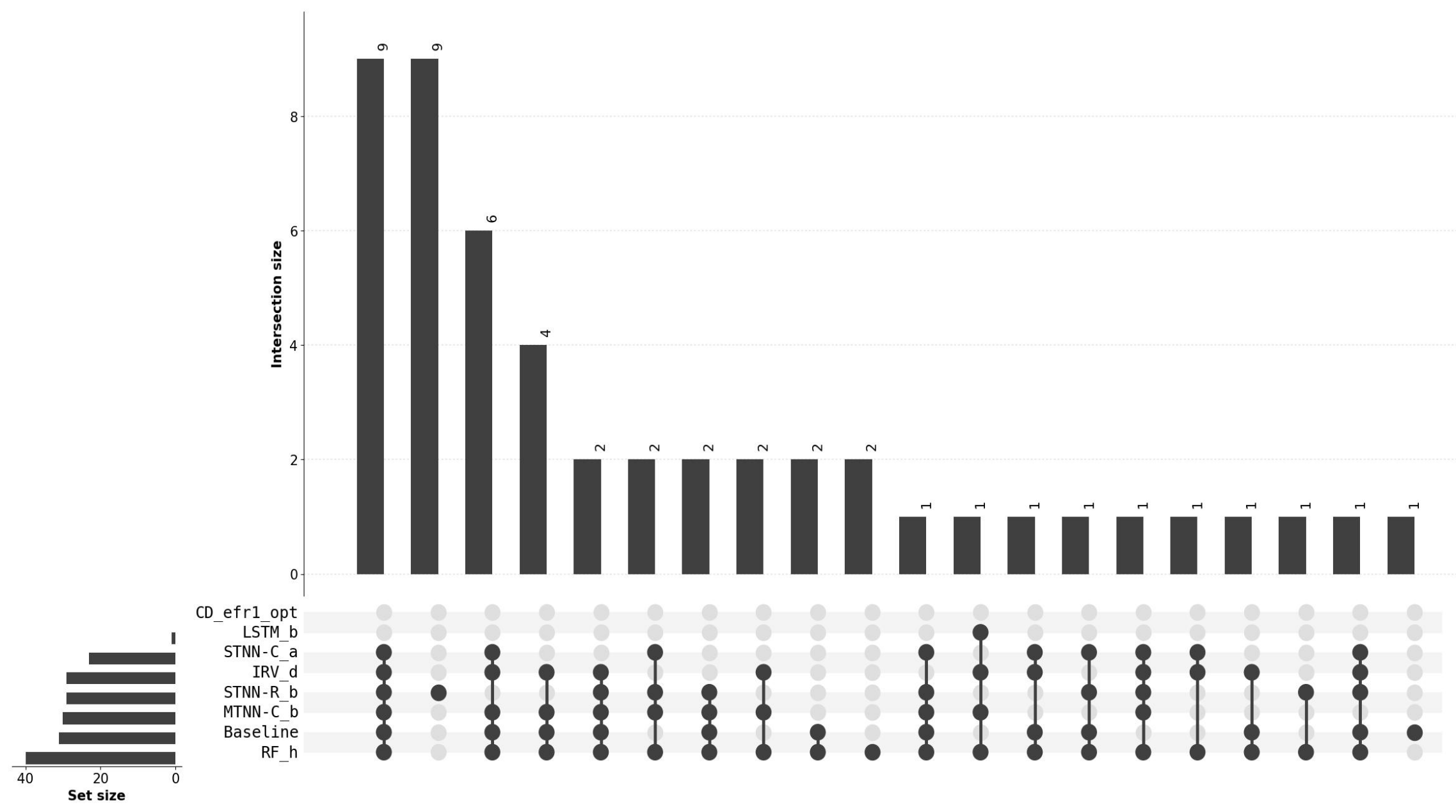
Evaluation metrics on Pria-SSB AS for all models.

Prospective Screening

Hits in Top 250 Predictions

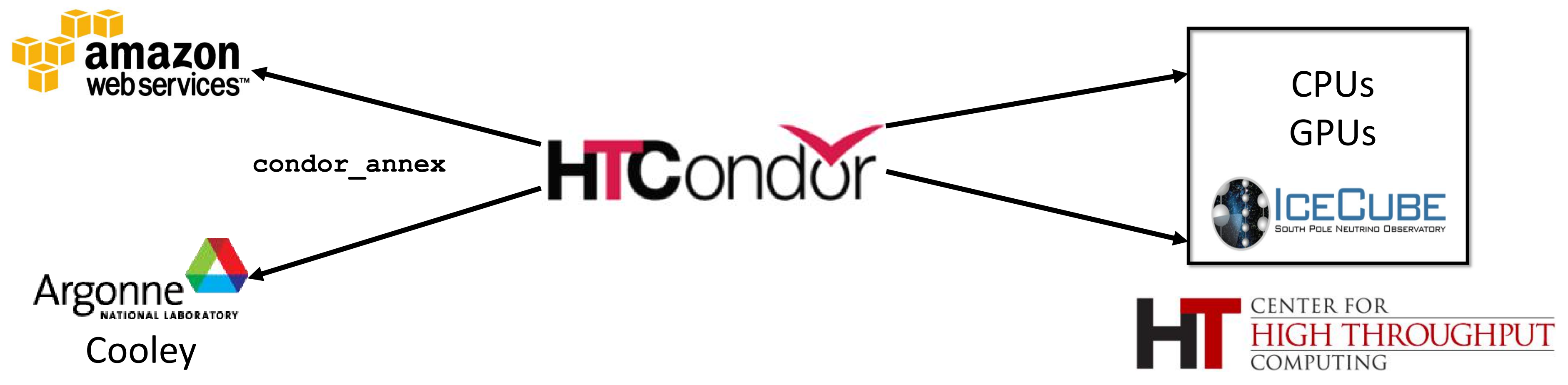
Number of active compounds in top 250 predictions from seven selected models and a chemical similarity baseline compared to the number of experimentally-identified actives.

Model	Actives	Actives not in baseline	SIM clusters	MCS clusters
Experimental	62	--	32	37
Similarity Baseline	31	--	14	8
Consensus Docking	0	0	0	0
STNN-C	23	4	12	7
STNN-R	29	13	16	11
MTNN-C	30	6	15	9
LSTM	1	1	1	1
Random Forest	40	10	16	9
IRV	29	5	13	7



An UpSet plot showing the overlap between the selected models and the chemical similarity baseline on PriA-SSB prospective. The plot generalizes a Venn diagram by indicating the overlapping sets with dots on the bottom and the size of the overlaps with the bar graph.

High-throughput Computing


amazon webservices
condor_annex
HTCondor
Argonne NATIONAL LABORATORY
Cooley
CPUs
GPUs
ICECUBE
CENTER FOR HIGH THROUGHPUT COMPUTING

Future Work

- Test ensembles that combine classification and regression models
- Scale to more diverse chemical libraries with millions of untested chemicals
- Assess alternative chemical feature representations

References

1. Scigenis. "Schematic illustration of docking a small molecule ligand (green) to a protein target (black) forming a protein-ligand complex." en.wikipedia.org/wiki/Docking_(molecular)
2. S. Lusher and G. Schaftenaar. "2-D searching Tutorial" http://www.cmbi.ru.nl/edu/bioinf4/2D-Prac/2d.shtml
3. GitHub repository https://github.com/gitter-lab/pria_lifechem