Scrutinizing Deep Learning: A Virtual Screening Case Study

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

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

MORGRIDGE INSTITUTE FOR RESEARCH

WISCONSIN UNIVERSITY OF WISCONSIN–MADISON

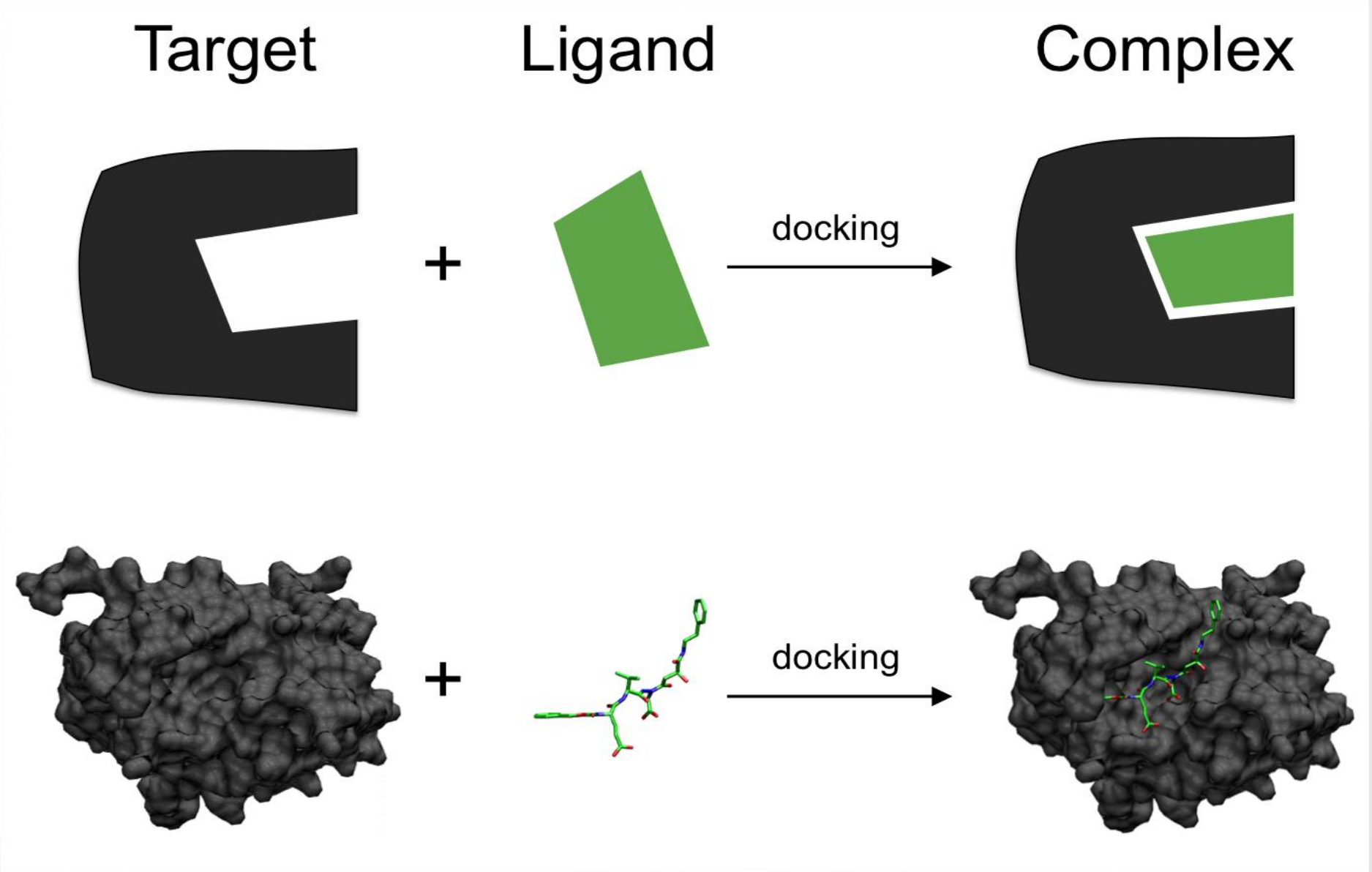
Introduction & Motivation

- **Problem:** Given a compound and target protein, determine whether the compound binds with the target. (Drug Discovery)
- Only way to be sure is physical tests (in vitro) in a molecule facility. Expensive and timely.

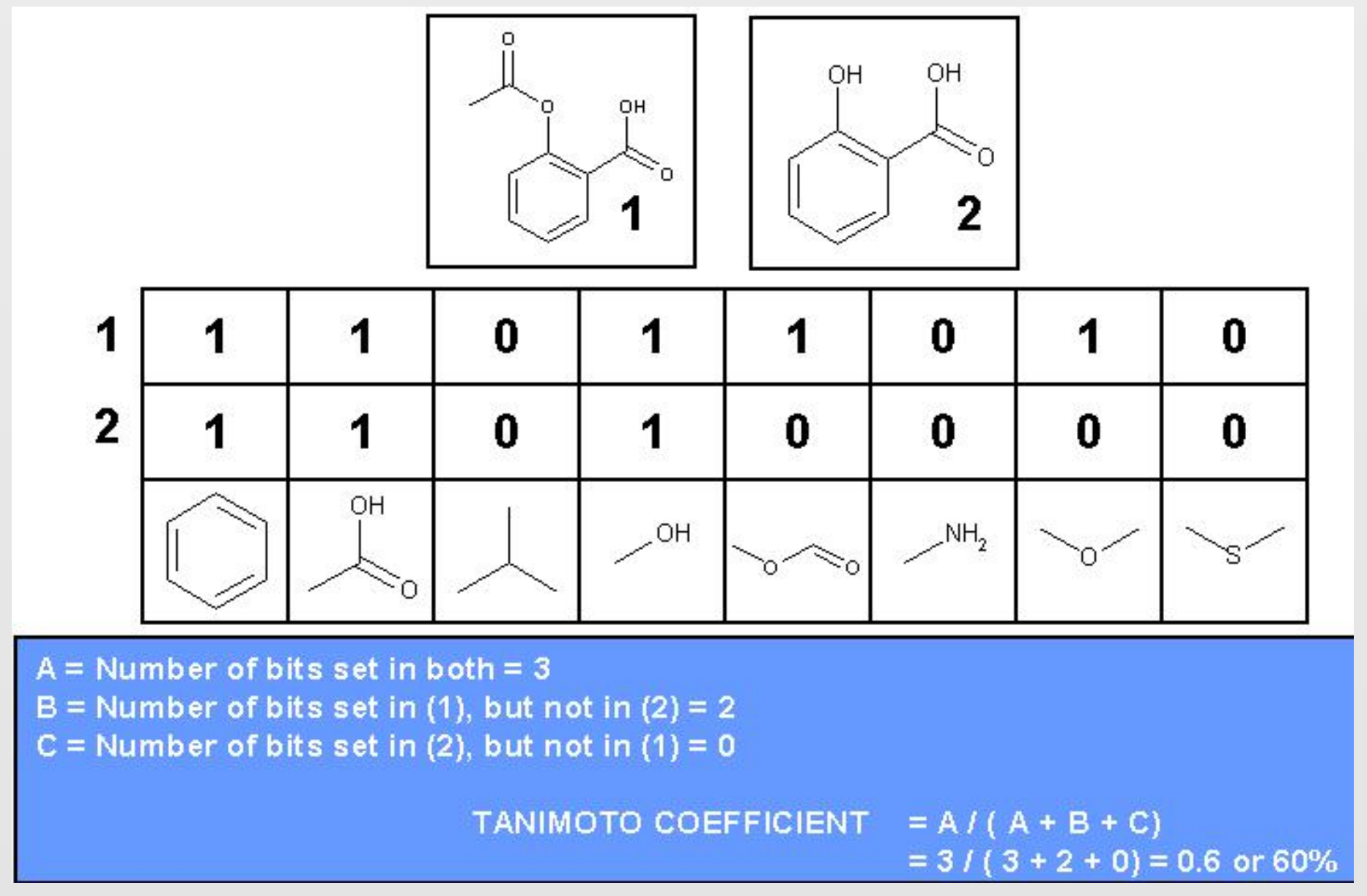
Virtual Screening can help accelerate drug discovery by **proposing most probable** compounds for testing. (in silico)

Two main VS methods

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%

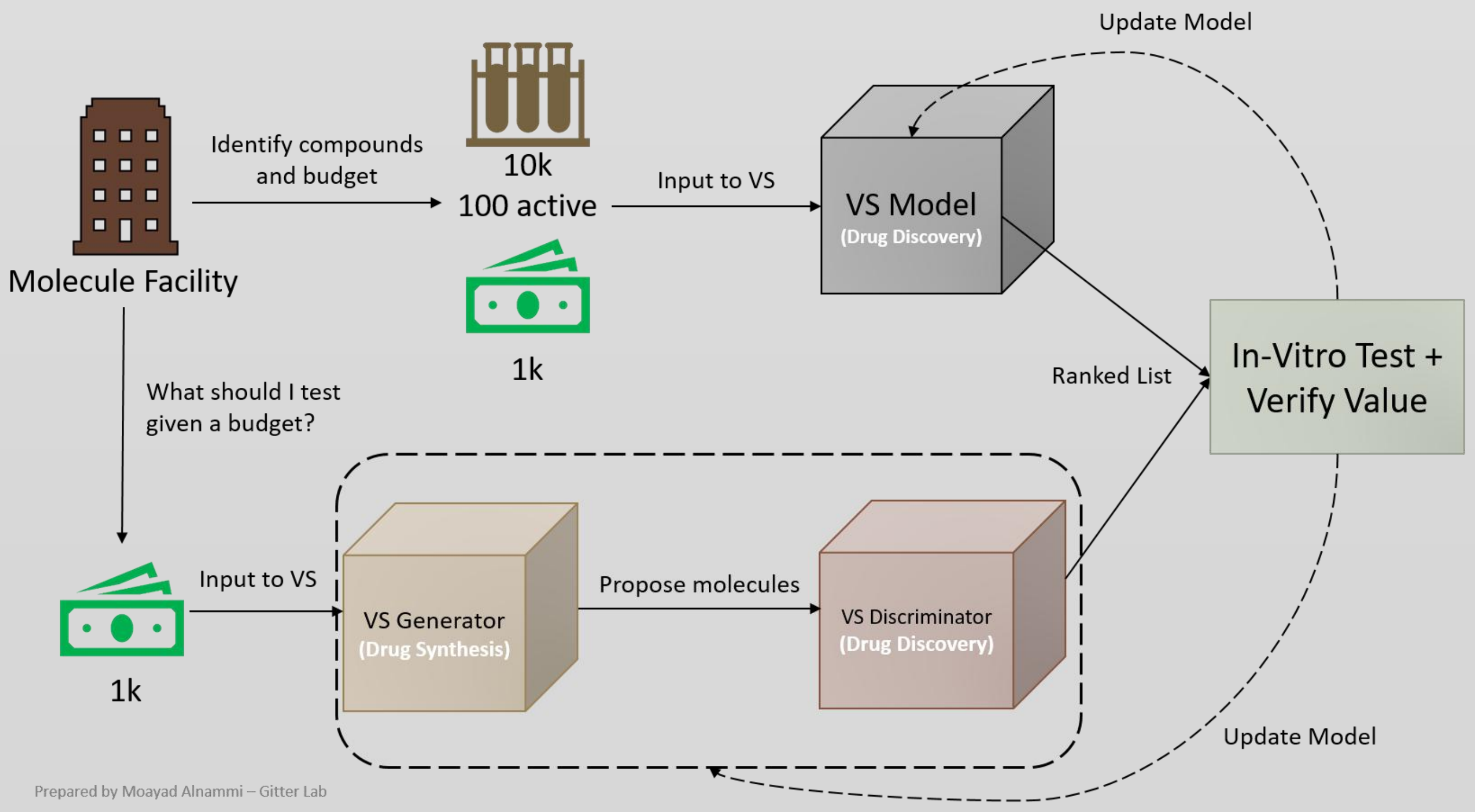
LB Fingerprint Concept. Figure from [2]

Case Study: SSB-PriA

- Keck lab has given 75k ligand-protein interaction data for 3 targets. **(known)**
- Later another 25k ligand interaction for these 3 targets. **(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 metrics** as it translates to real world value.

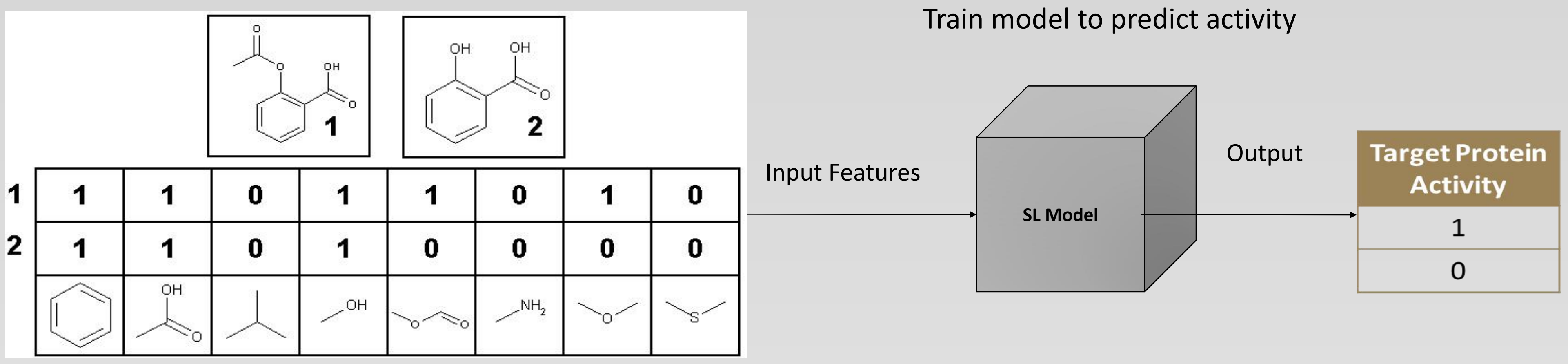
Real-World Impact: Help molecule facilities by proposing top 1000 most likely compounds. Perfect ranking not important.



Prepared by Moayad Alnammi – Gitter Lab

Supervised Learning Setup

Train model to predict activity



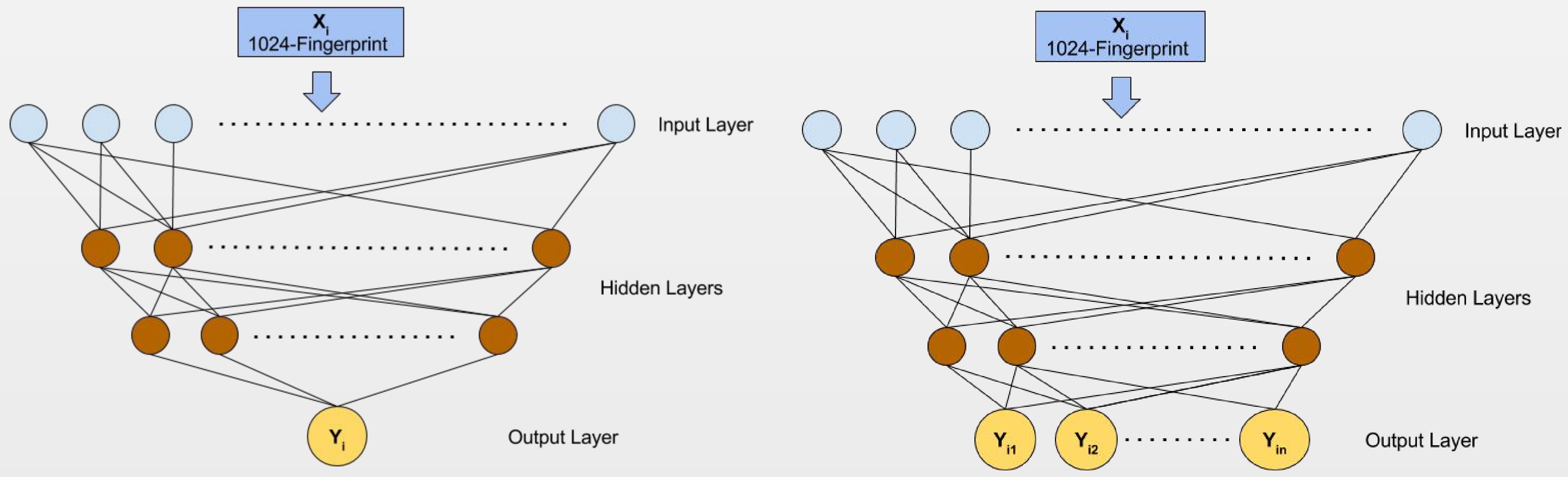
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

Target Protein Activity
1
0

Fingerprint Concept. Figure from [2]

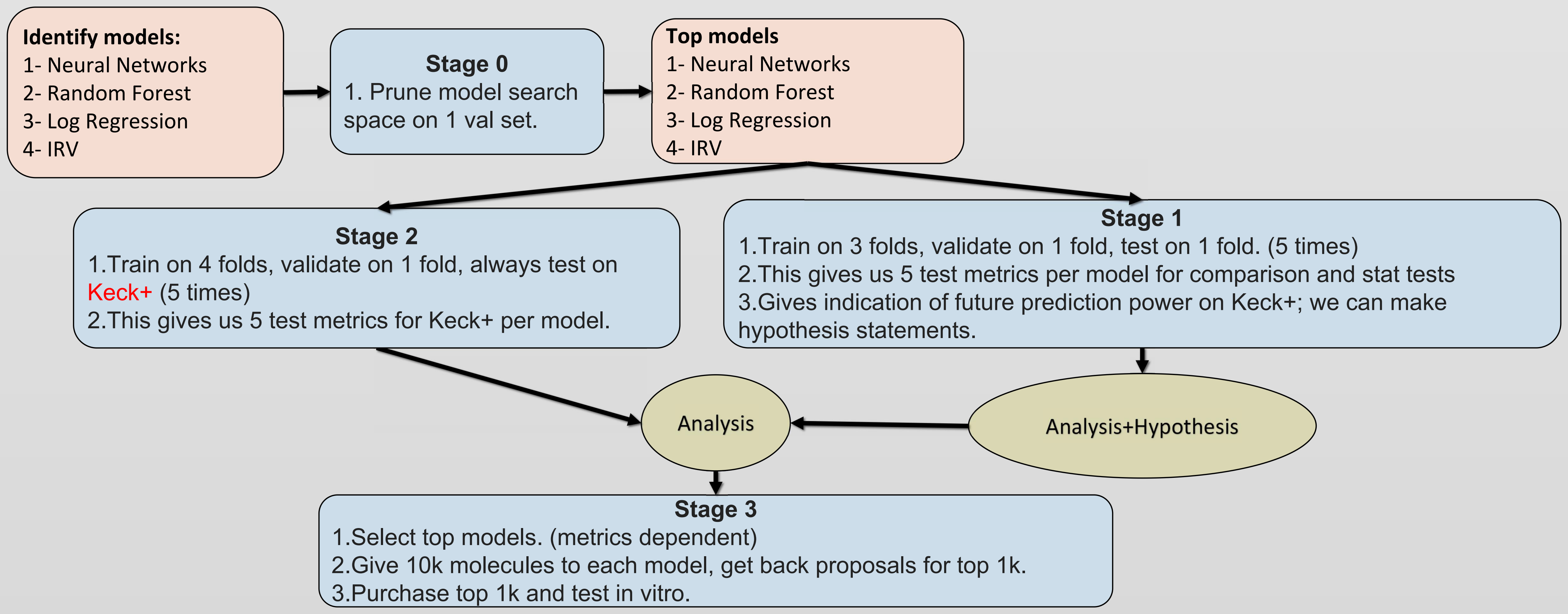
Goal: Given a new molecule, use trained model to predict its activity.

Single Task NN vs Multi-Task NN



Issue	Multiple STNN	Single MTNN
Imbalanced classes	Easy weight adjustment	- Careful weight adjustment - Target error can dominate others
Merging Datasets	No need	Missing labels
Stratified train/val/test	Easy 1-column split	- Complicated multi-col - Greedy col-by-col splits
Shared Weights	None	- Captures semantic structural info - local minima/regularizers
NN Hyperparameters	- Activation functions: relu, **elu**, etc. - Optimizer: **adam**, sgd, etc. - **Dropout,** BatchNorms, weight initializers, architetcture.	
Evaluation Metrics	- ROC, PR, EF, BEDROC. - Translation to real-world value for molecule facilities.	

Project Pipeline



Identify models:
1- Neural Networks
2- Random Forest
3- Log Regression
4- IRV

Stage 0
1. Prune model search space on 1 val set.

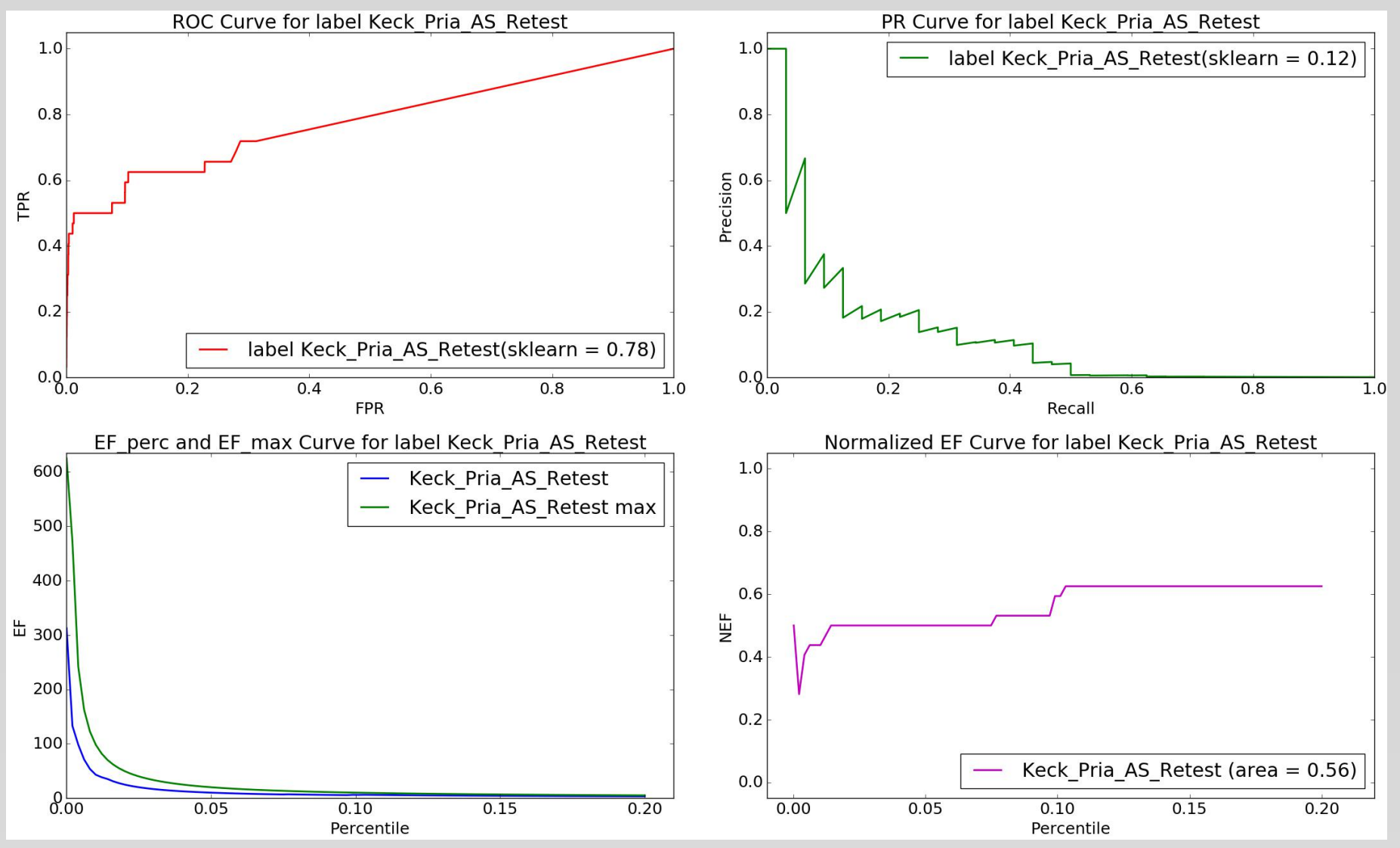
Top models
1- Neural Networks
2- Random Forest
3- Log Regression
4- IRV

Stage 2
1.Train on 4 folds, validate on 1 fold, always test on Keck+ (5 times)
2.This gives us 5 test metrics for Keck+ per model.

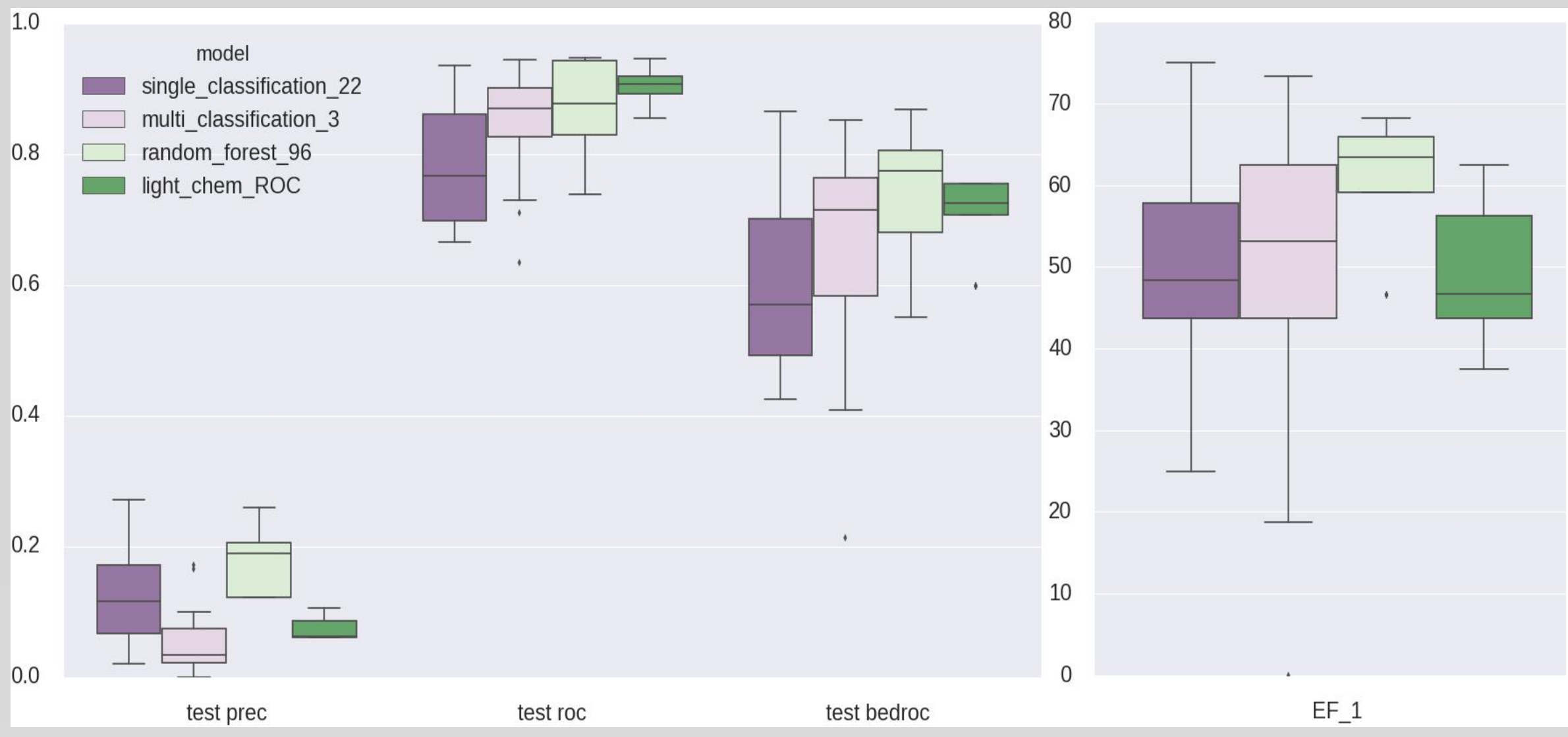
Stage 1
1.Train on 3 folds, validate on 1 fold, test on 1 fold. (5 times)
2.This gives us 5 test metrics per model for comparison and stat tests
3.Gives indication of future prediction power on Keck+; we can make hypothesis statements.

Stage 3
1.Select top models. (metrics dependent)
2.Give 10k molecules to each model, get back proposals for top 1k.
3.Purchase top 1k and test in vitro.

Preliminary Results



Sample MTNN evaluation results using different metrics. How do we relate these metrics to actual value?



Preliminary results among four different classes of models: STNN, MTNN, Random Forest, and LightChem. The results are on four metrics on the test set.

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