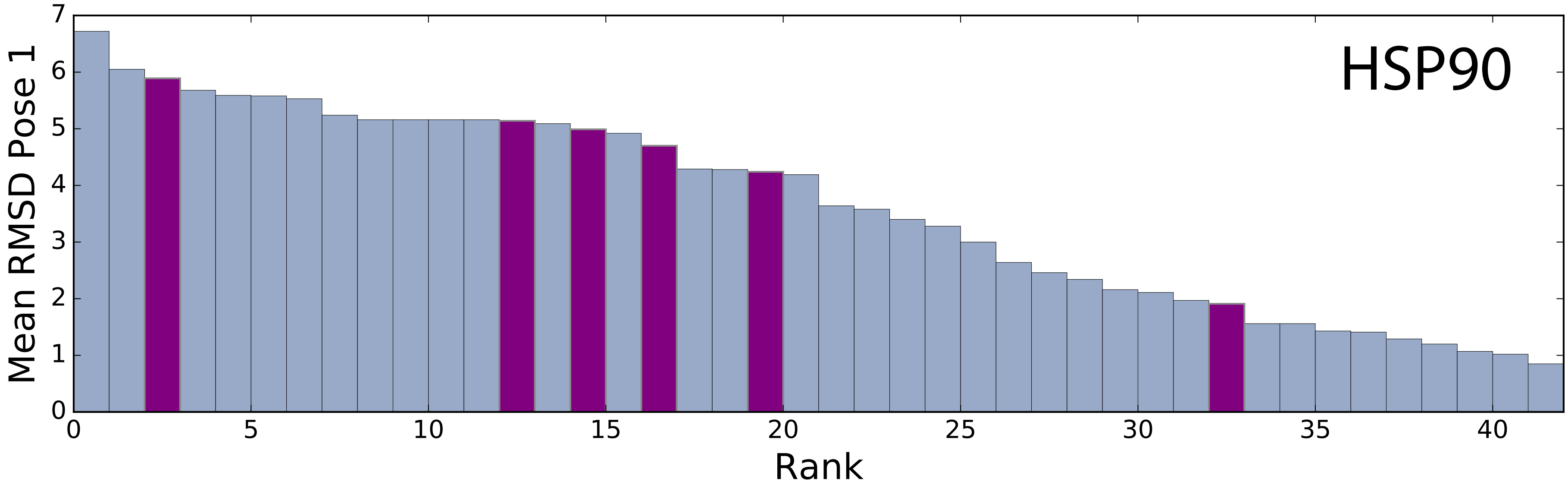
D3R Grand Challenge 2015

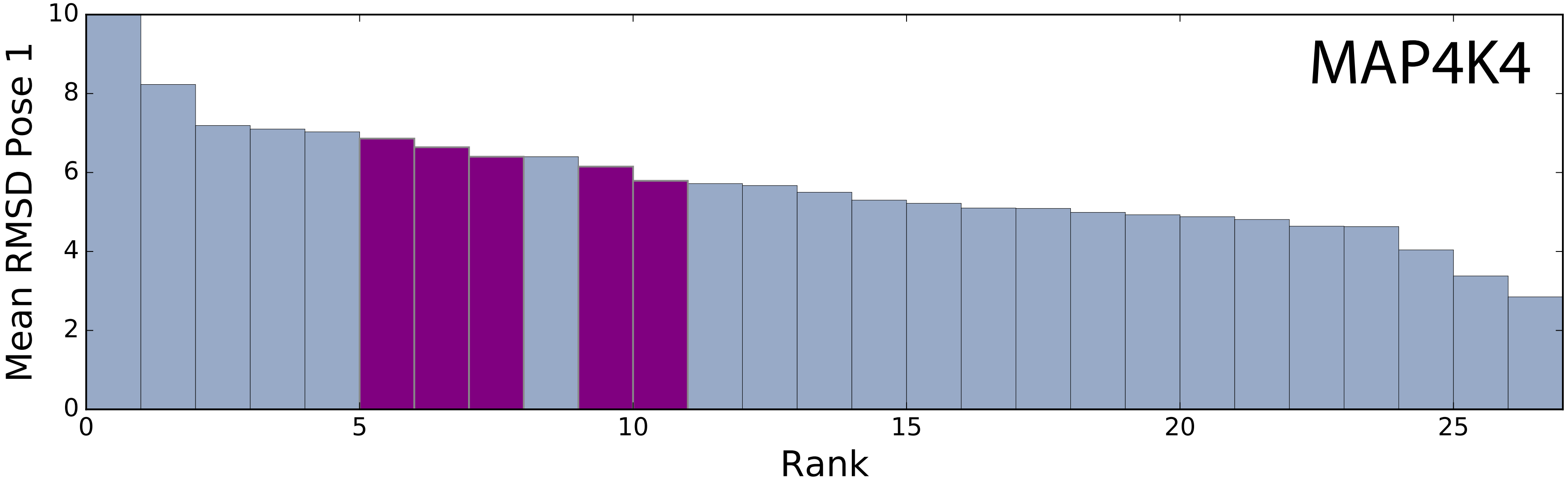
Machine Learning for Protein-Ligand Recognition

David Ryan Koes

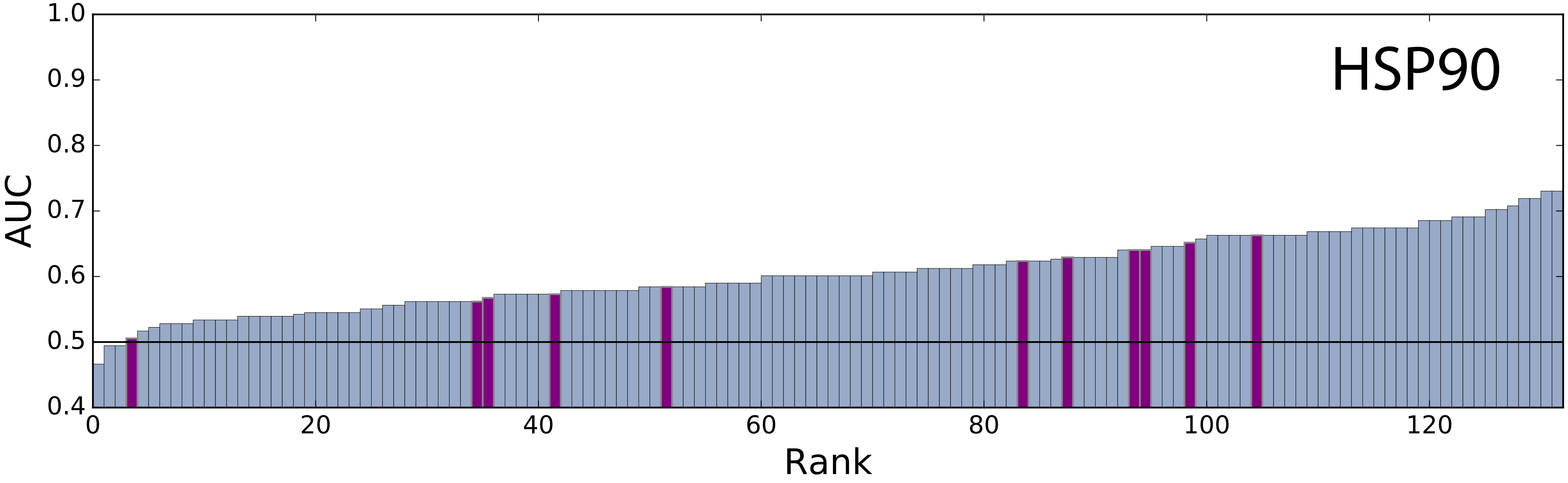
Computational and Systems Biology
University of Pittsburgh

Docking Results

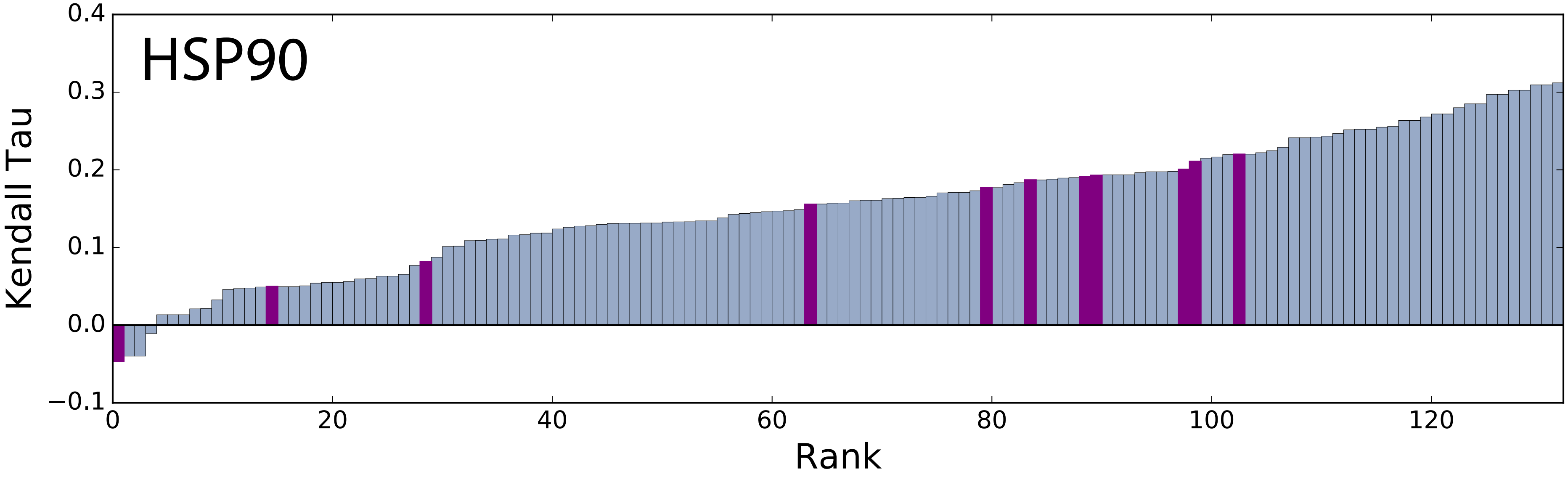


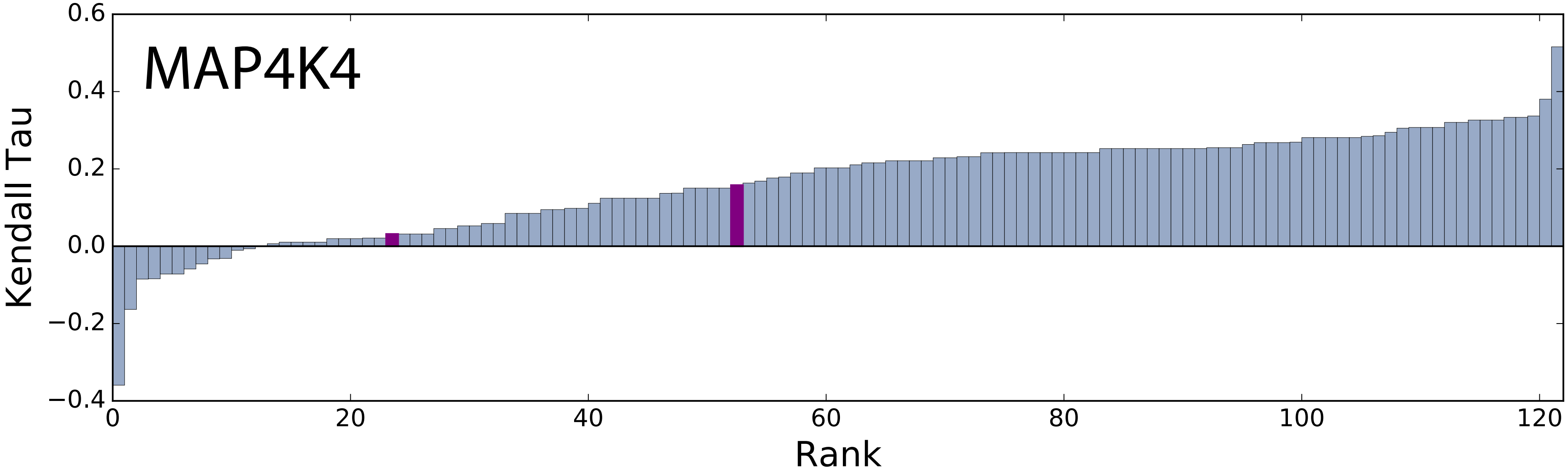


Screening Results

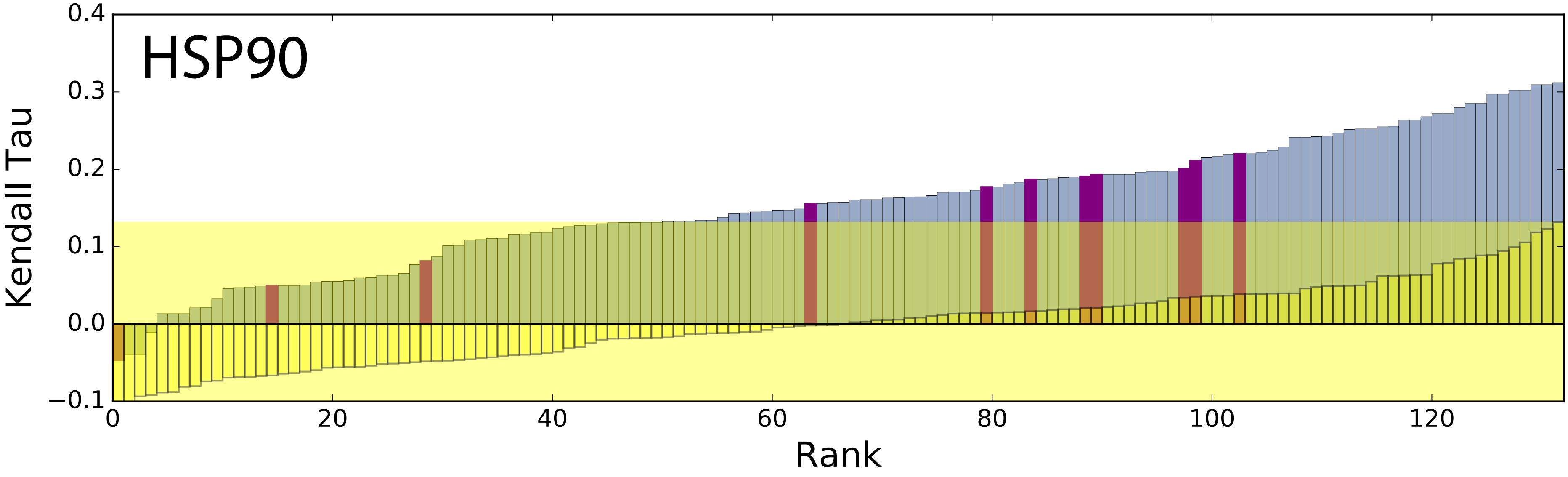


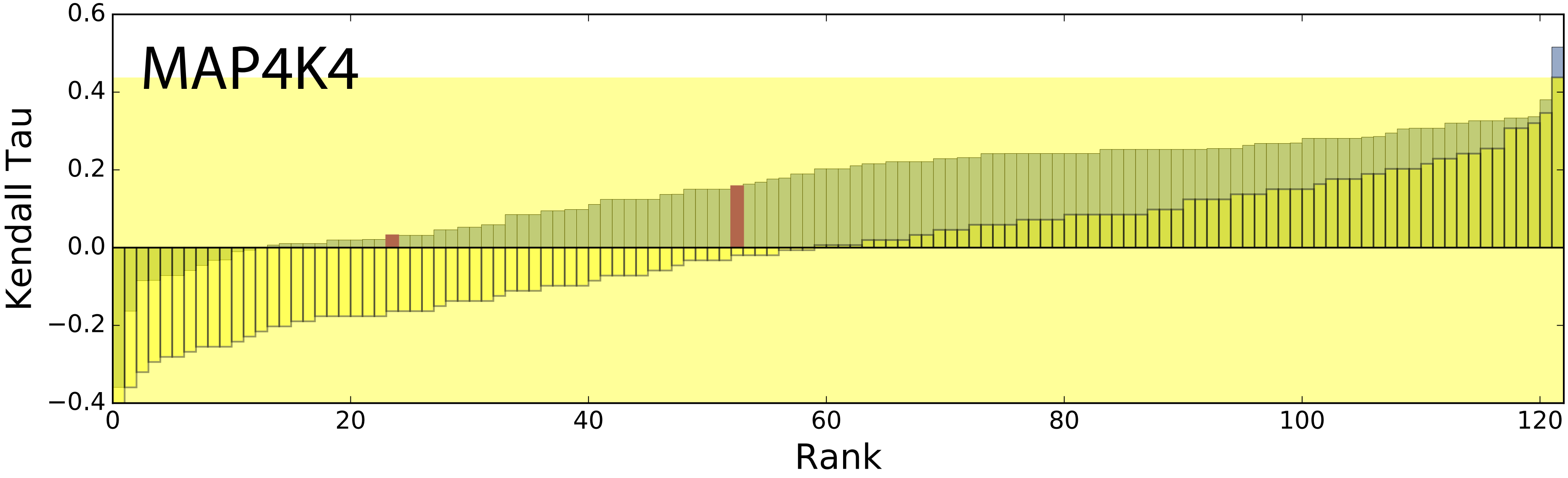
Affinity Results



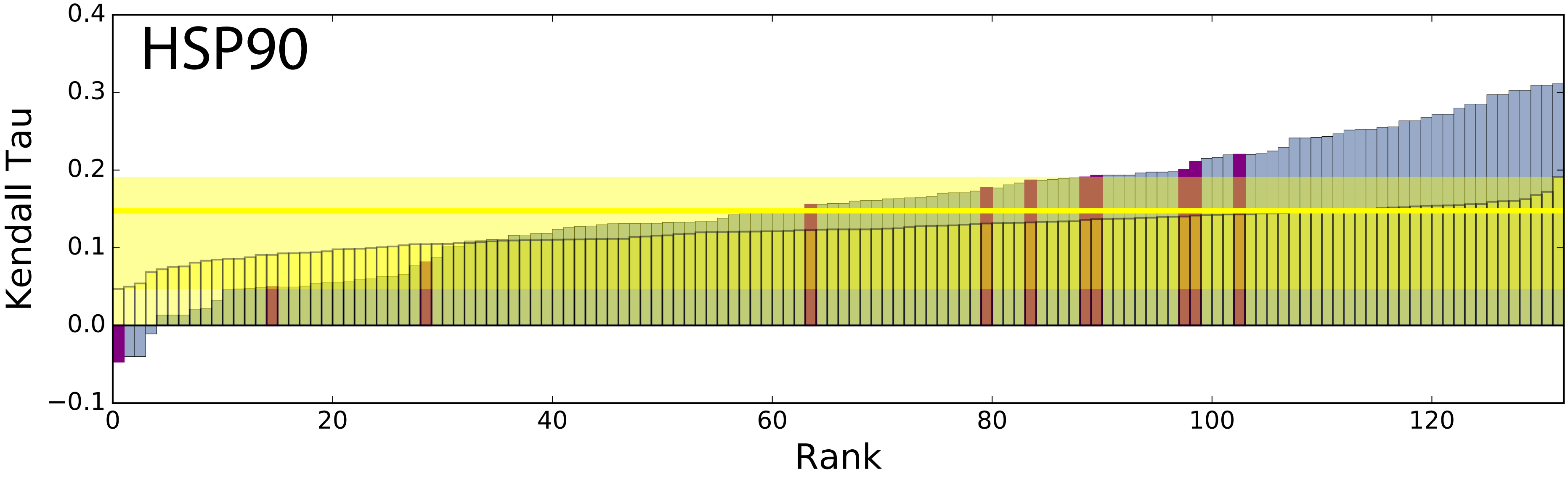


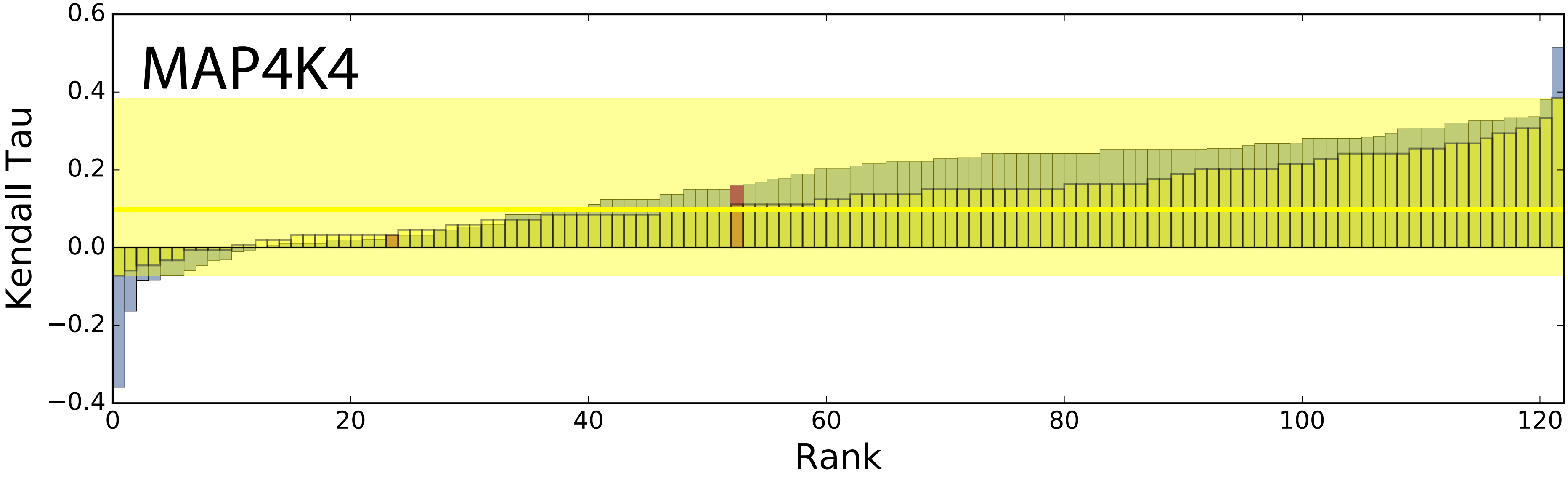
Visualizing Significance



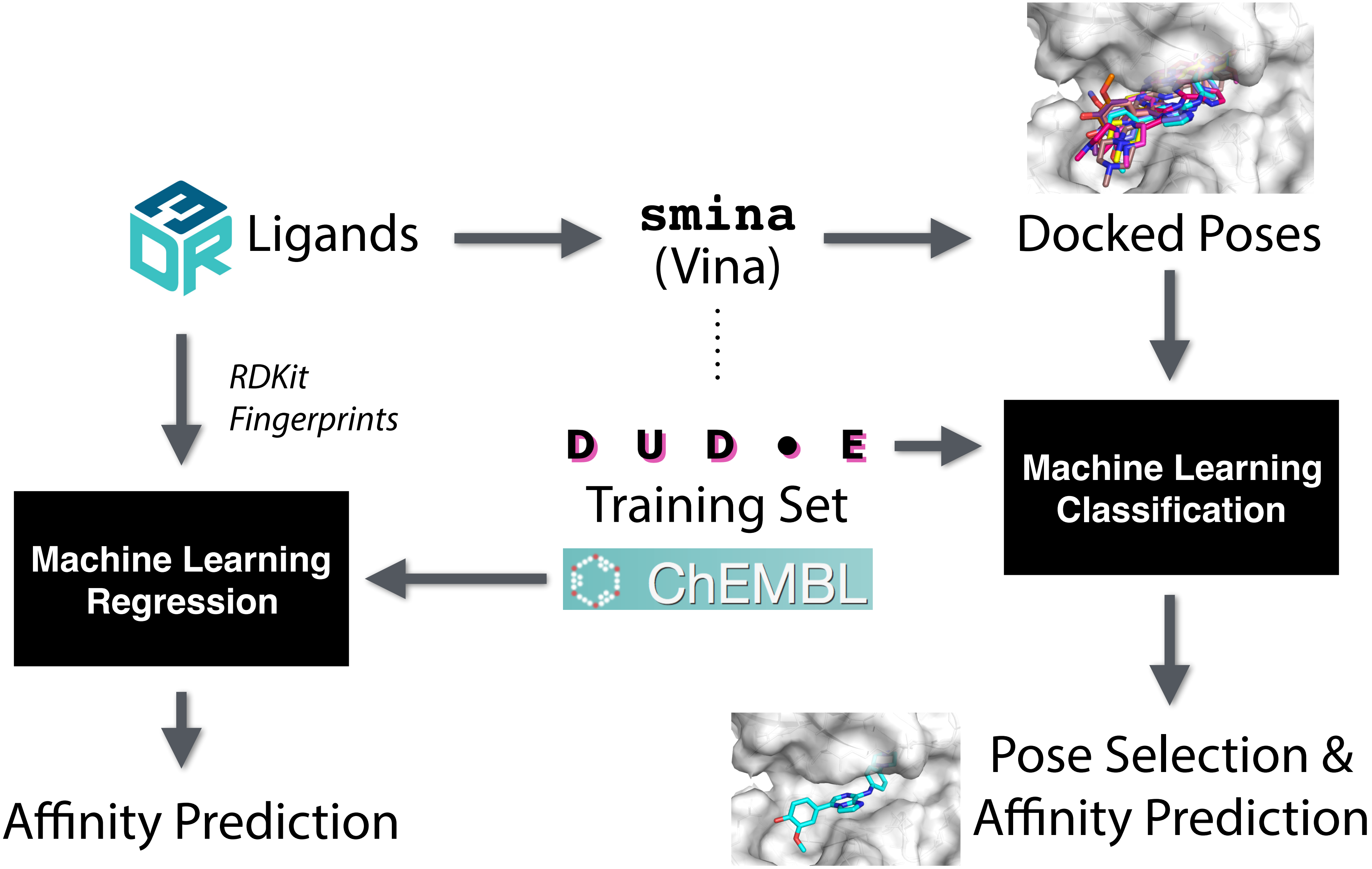


Visualizing Significance





Overall Approach



Ligands

smina
(Vina)

Docked Poses

RDKit Fingerprints

DUD•E
Training Set
ChEMBL

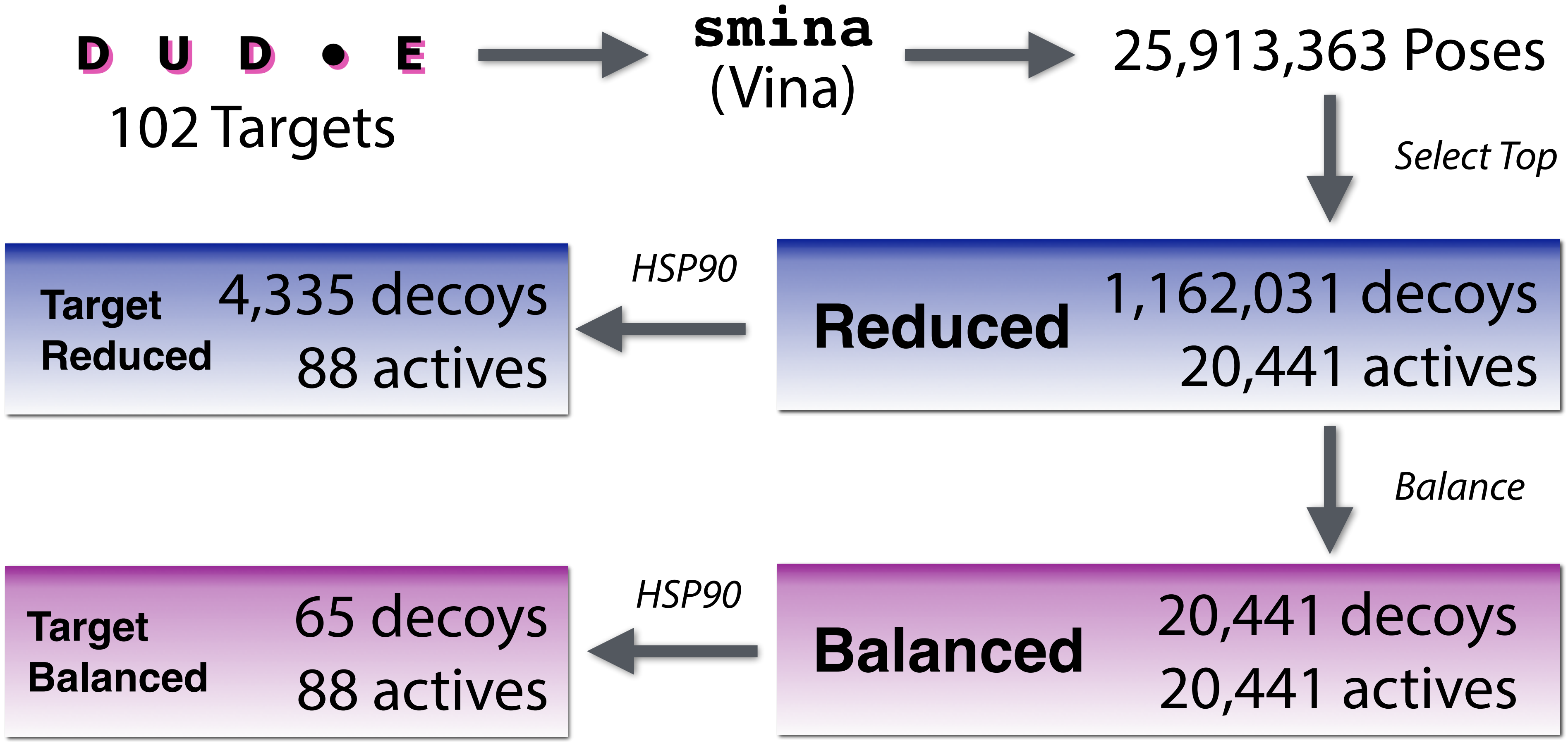
Machine Learning Classification

Machine Learning Regression

Affinity Prediction

Pose Selection & Affinity Prediction

Training Set – Classification



DUD•E

102 Targets

smina

(Vina)

25,913,363 Poses

Select Top

Reduced

1,162,031 decoys

20,441 actives

HSP90

Target Reduced

4,335 decoys

88 actives

Balance

Balanced

20,441 decoys

20,441 actives

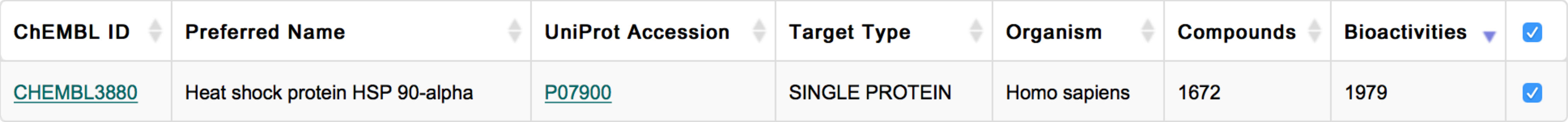
HSP90

Target Balanced

65 decoys

88 actives

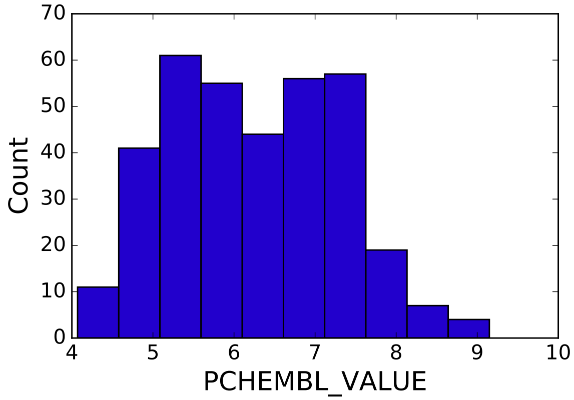
Training Set – Regression



ChEMBL ID	Preferred Name	UniProt Accession	Target Type	Organism	Compounds	Bioactivities	☑
CHEMBL3880	Heat shock protein HSP 90-alpha	P07900	SINGLE PROTEIN	Homo sapiens	1672	1979	☑

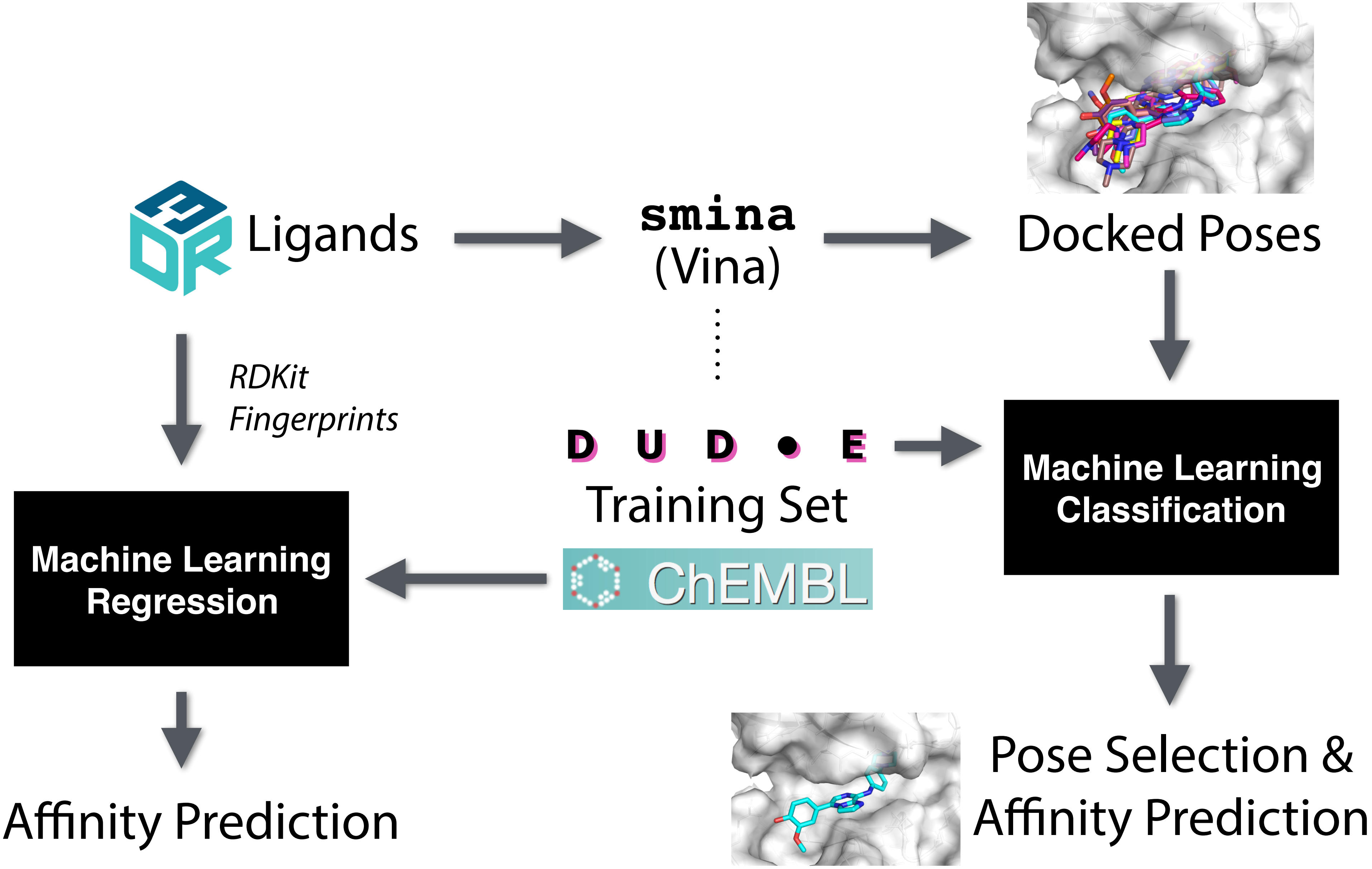
```
import pandas as pd
hsp = pd.read_csv('bioactivity-15_21_16_49.txt',sep='\t')
smi = hsp[(hsp.STANDARD_TYPE == 'IC50') & (hsp.RELATION == '=') &
(hsp.STANDARD_UNITS == 'nM') & (hsp.PCHEMBL_VALUE > 0)].loc[:,
['CANONICAL_SMILES','PCHEMBL_VALUE']]
smi.to_csv('hsp90.smi',sep='\t',index=False,header=False)
```

Remove salts



355 Active Compounds

Overall Approach



Ligands

smina
(Vina)

Docked Poses

RDKit Fingerprints

DUD•E
Training Set
ChEMBL

Machine Learning Classification

Machine Learning Regression

Affinity Prediction

Pose Selection &
Affinity Prediction

Features - Classification

```
gauss(o=0,_w=0.5,_c=8)
gauss(o=3,_w=2,_c=8)
gauss(o=1.5,_w=0.3,_c=8)
gauss(o=2,_w=0.9,_c=8)
gauss(o=1,_w=0.9,_c=8)
gauss(o=1,_w=0.5,_c=8)
gauss(o=1,_w=0.3,_c=8)
gauss(o=1,_w=0.7,_c=8)
gauss(o=2,_w=0.5,_c=8)
gauss(o=2,_w=0.7,_c=8)
gauss(o=3,_w=0.9,_c=8)
repulsion(o=0,_c=8)
hydrophobic(g=0.5,_b=1.5,_c=8)
hydrophobic(g=0.5,_b=1,_c=8)
hydrophobic(g=0.5,_b=2,_c=8)
hydrophobic(g=0.5,_b=3,_c=8)
non_hydrophobic(g=0.5,_b=1.5,_c=8)
vdw(i=4,_j=8,_s=0,_^=100,_c=8)
vdw(i=6,_j=12,_s=1,_^=100,_c=8)
e_vdw
non_dir_h_bond(g=-0.7,_b=0,_c=8)
non_dir_h_bond(g=-0.7,_b=0.2,_c=8)
non_dir_h_bond(g=-0.7,_b=0.5,_c=8)
non_dir_h_bond(g=-1,_b=0,_c=8)
non_dir_h_bond(g=-1,_b=0.2,_c=8)
non_dir_h_bond(g=-1,_b=0.5,_c=8)
non_dir_h_bond(g=-1.3,_b=0,_c=8)
non_dir_h_bond(g=-1.3,_b=0.2,_c=8)
non_dir_h_bond(g=-1.3,_b=0.5,_c=8)
```

```
non_dir_anti_h_bond_quadratic(o=0,_c=8)
non_dir_anti_h_bond_quadratic(o=0.5,_c=8)
non_dir_anti_h_bond_quadratic(o=1,_c=8)
non_dir_h_bond_lj(o=-0.7,_^=100,_c=8)
non_dir_h_bond_lj(o=-1,_^=100,_c=8)
non_dir_h_bond_lj(o=-1.3,_^=100,_c=8)
e_hb
e_ligPen
ad4_solvation(d-sigma=3.6,_s/q=0.01097,_c=8)
ad4_solvation(d-sigma=3.6,_s/q=0.01097,_c=8)
e_s1
e_s2
e_s3
e_s4
e_s5
electrostatic(i=1,_^=100,_c=8)
electrostatic(i=2,_^=100,_c=8)
e_E0
e_E1
num_tors_div
num_heavy_atoms_div
num_heavy_atoms
num_tors_add
num_tors_sqr
num_tors_sqrt
num_hydrophobic_atoms
ligand_length
numBonds
bfO
bfN
myRotors
```

60 Terms

Steric

Hydrophobic

van der Waals

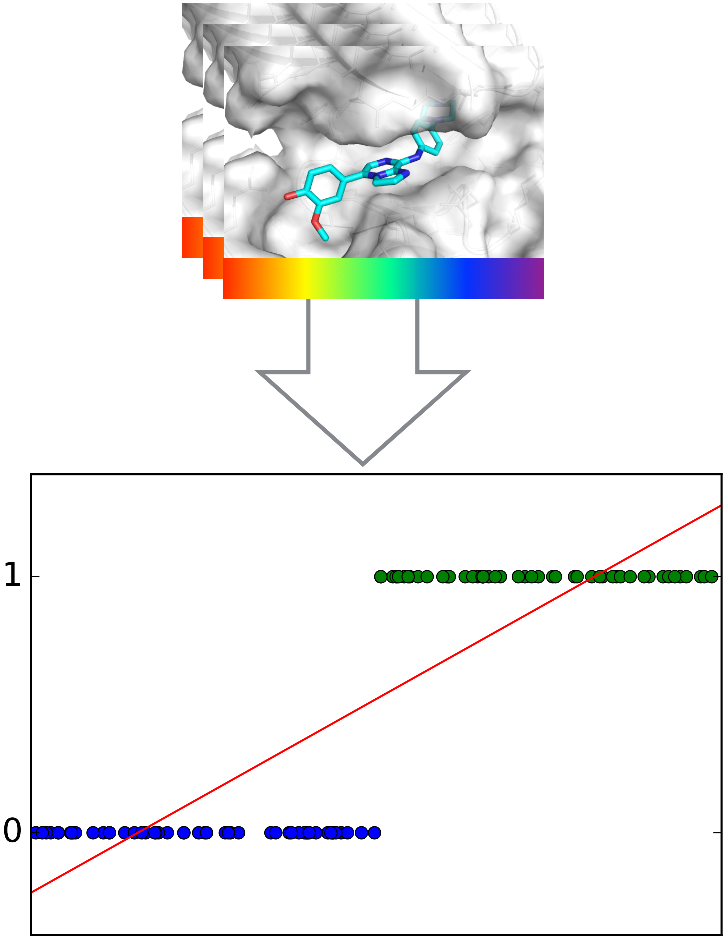
Hydrogen Bond

Solvation

Electrostatic

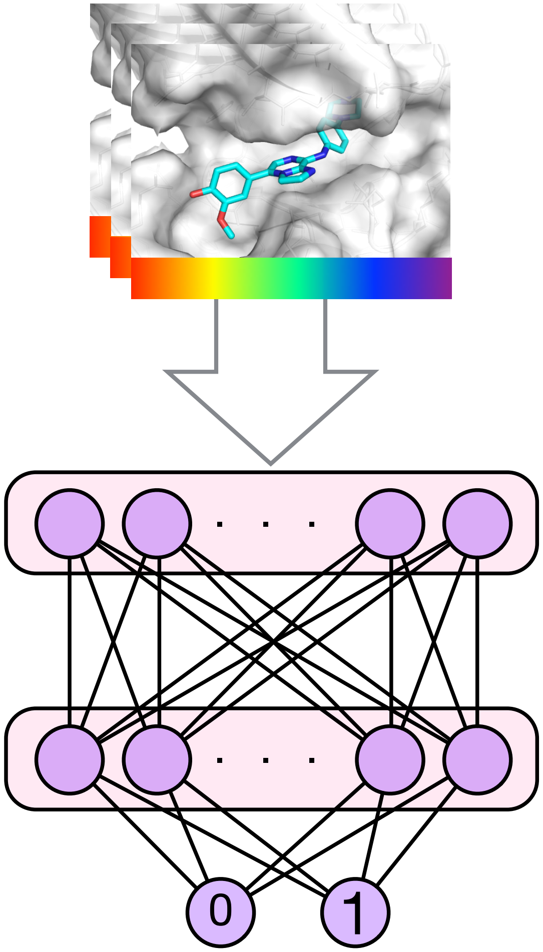
Counts

Models - Classification



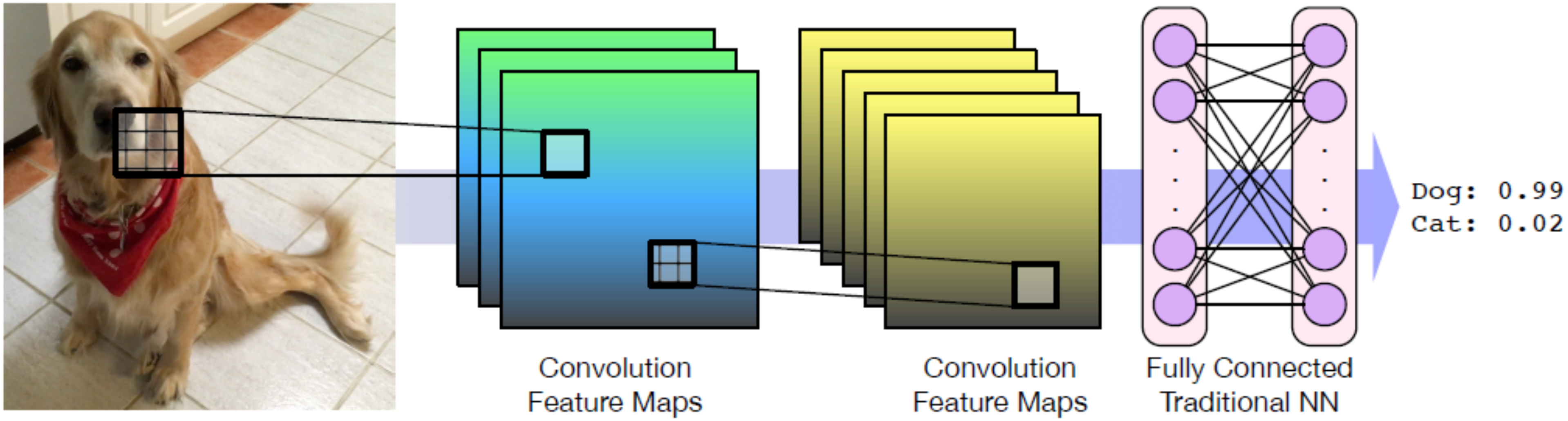
Linear Regression

also LASSO

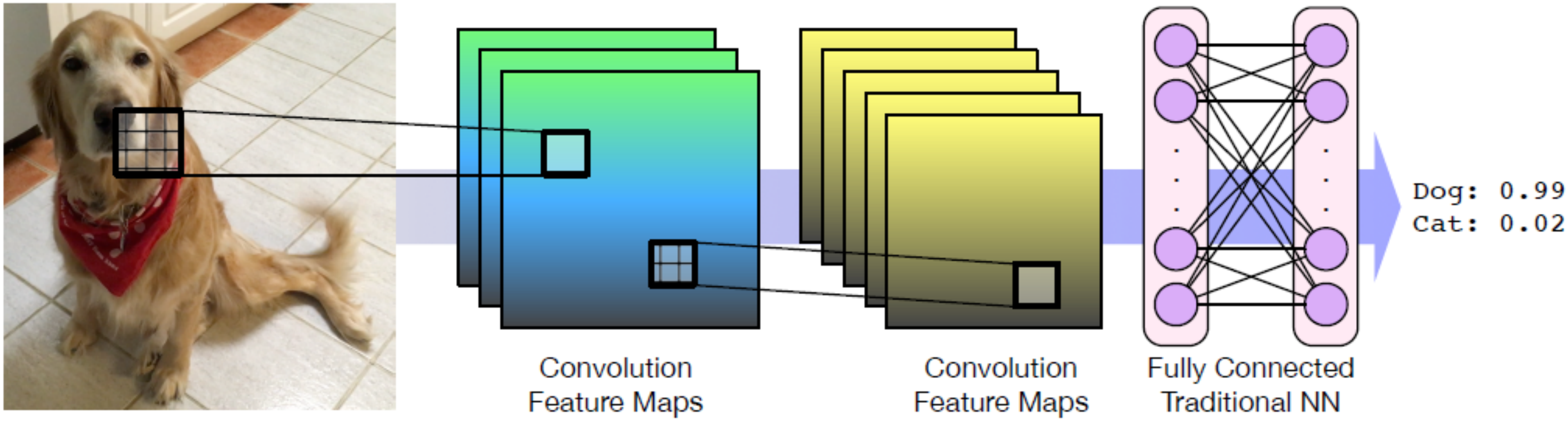


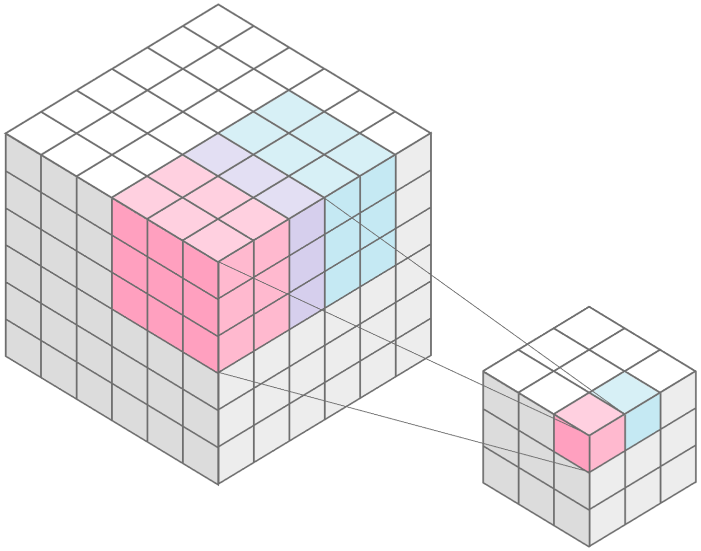
Artificial Neural Network

Convolutional Neural Net

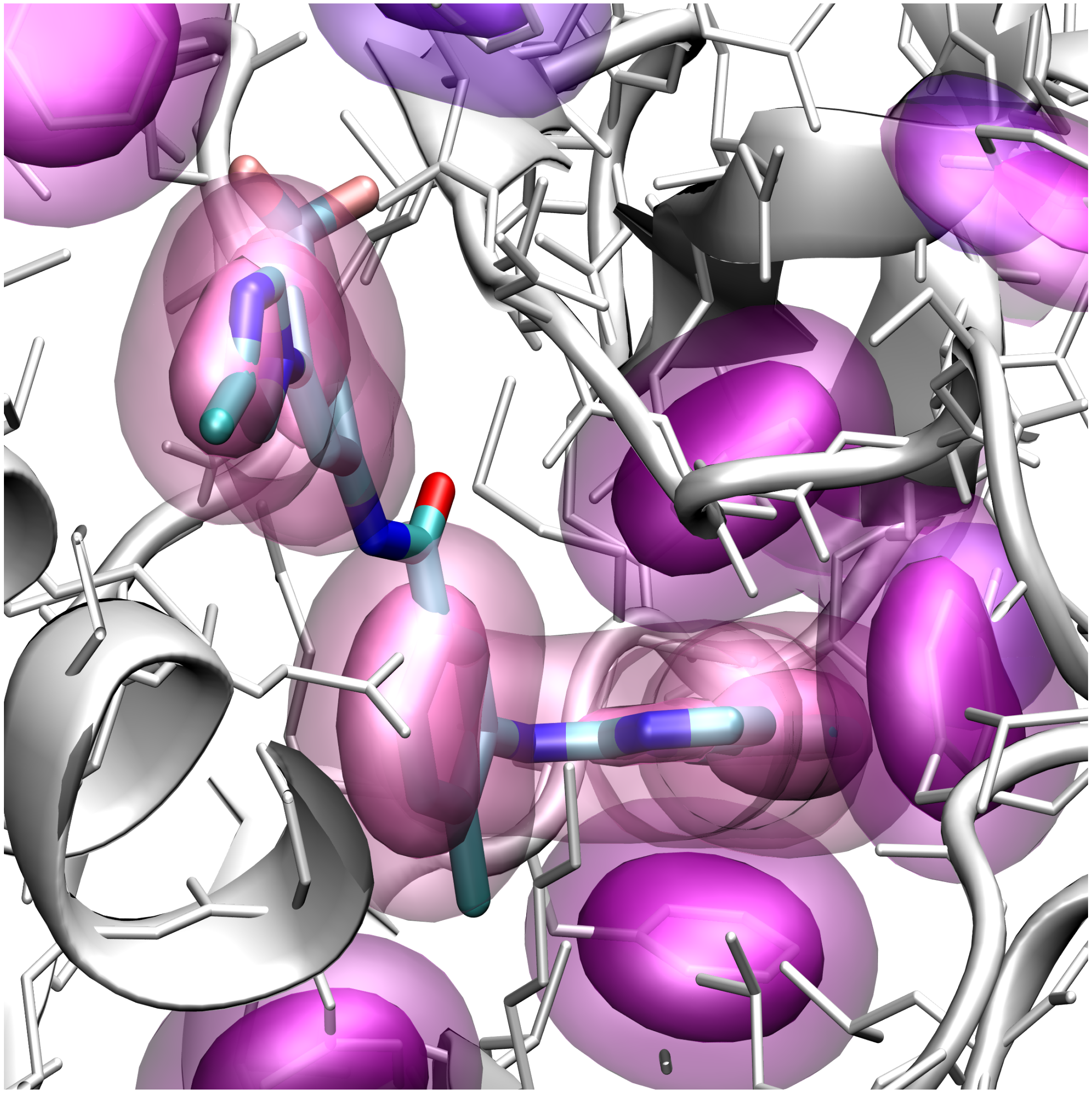


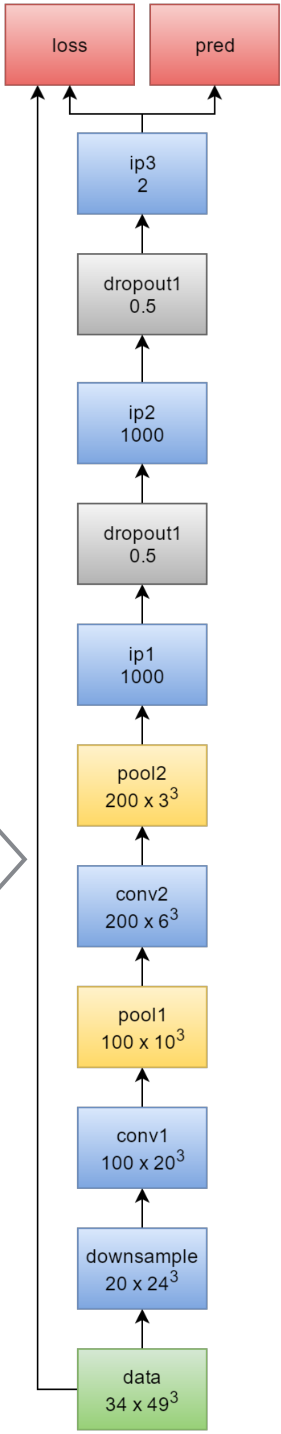
Convolutional Neural Net





Convolutional Neural Net





loss

pred

ip3
2

dropout1
0.5

ip2
1000

dropout1
0.5

ip1
1000

pool2
200 x 3^3

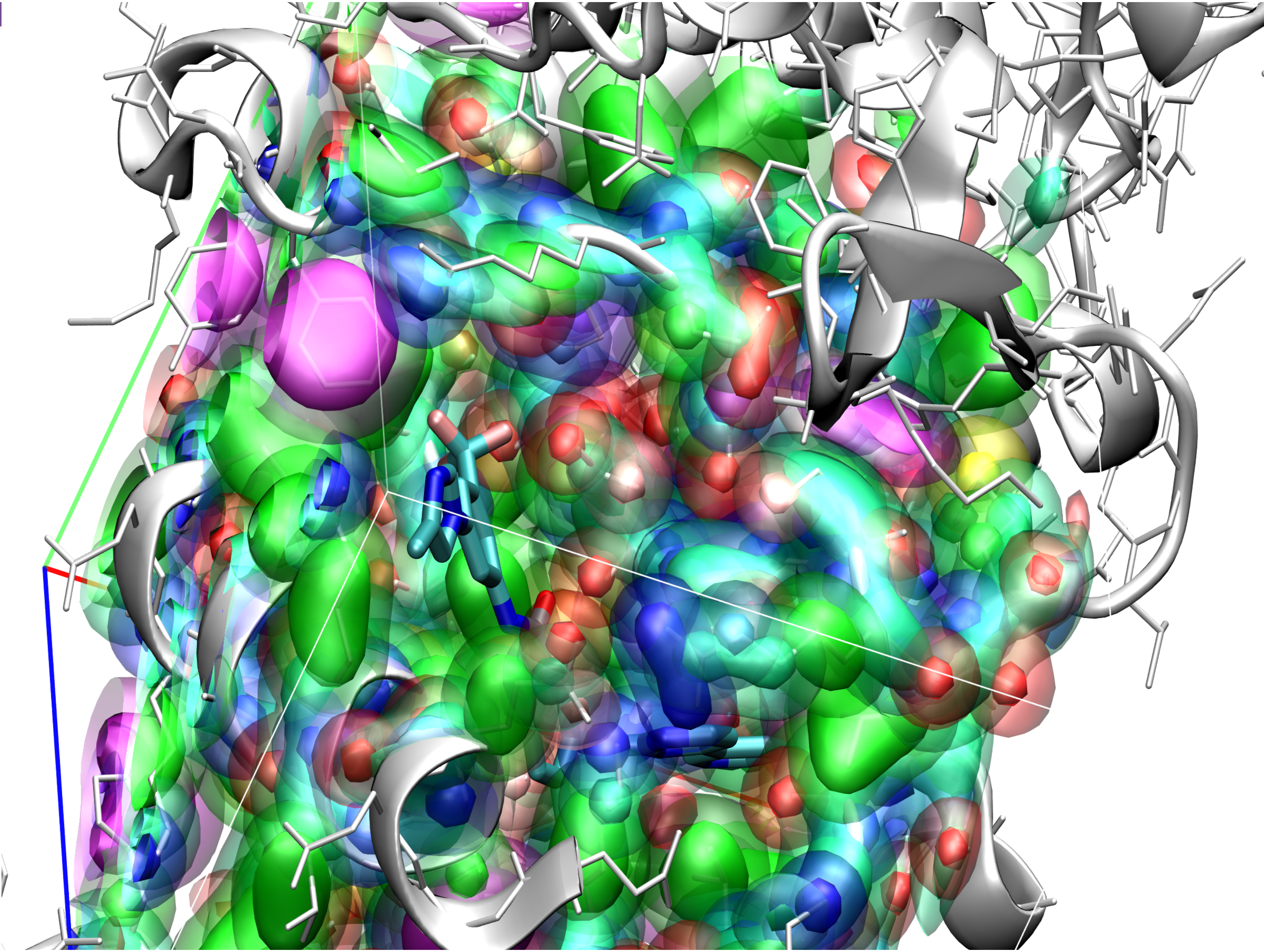
conv2
200 x 6^3

pool1
100 x 10^3

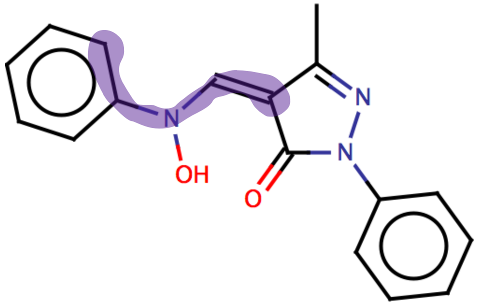
conv1
100 x 20^3

downsample
20 x 24^3

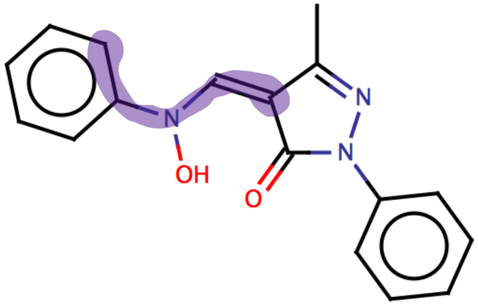
data
34 x 49^3



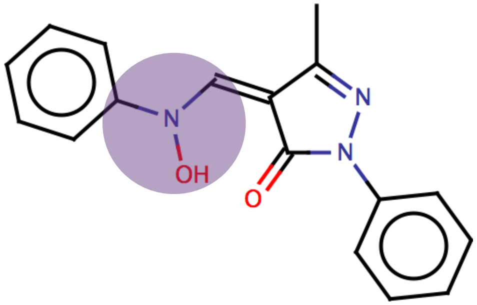
Regression



RDKit
path
2048 bits



SMARTS
path
n bits



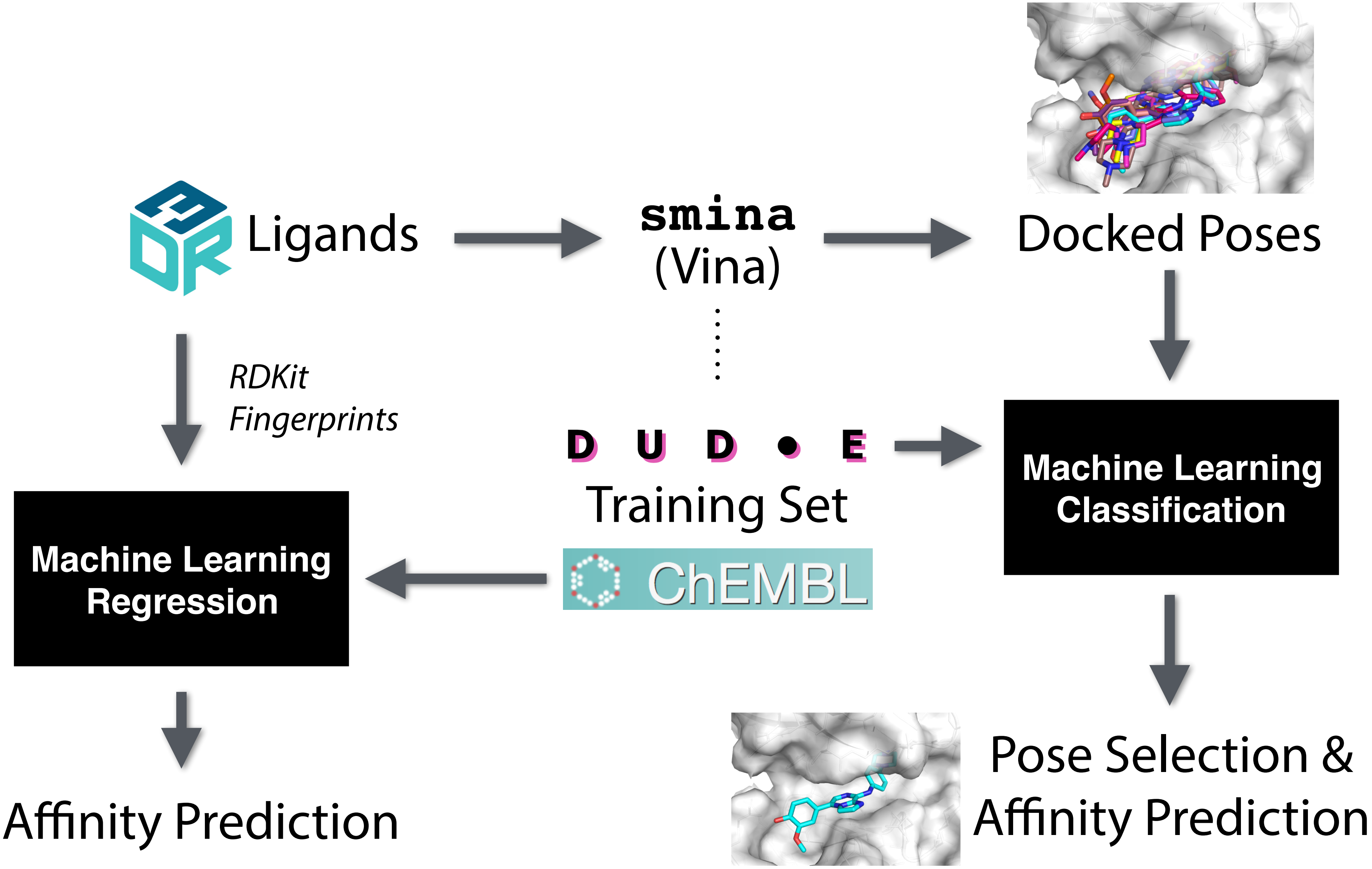
ECFP6
circular
2048 bits

ElasticNetCV

$$\underset{w}{min} \frac{1}{2n_{samples}}||Xw - y||_2^2 + \alpha\rho||w||_1 + \frac{\alpha(1-\rho)}{2}||w||_2^2$$

`https://github.com/dkoes/qsar-tools`

Overall Approach



Ligands

smina
(Vina)

Docked Poses

RDKit Fingerprints

DUD•E
Training Set
ChEMBL

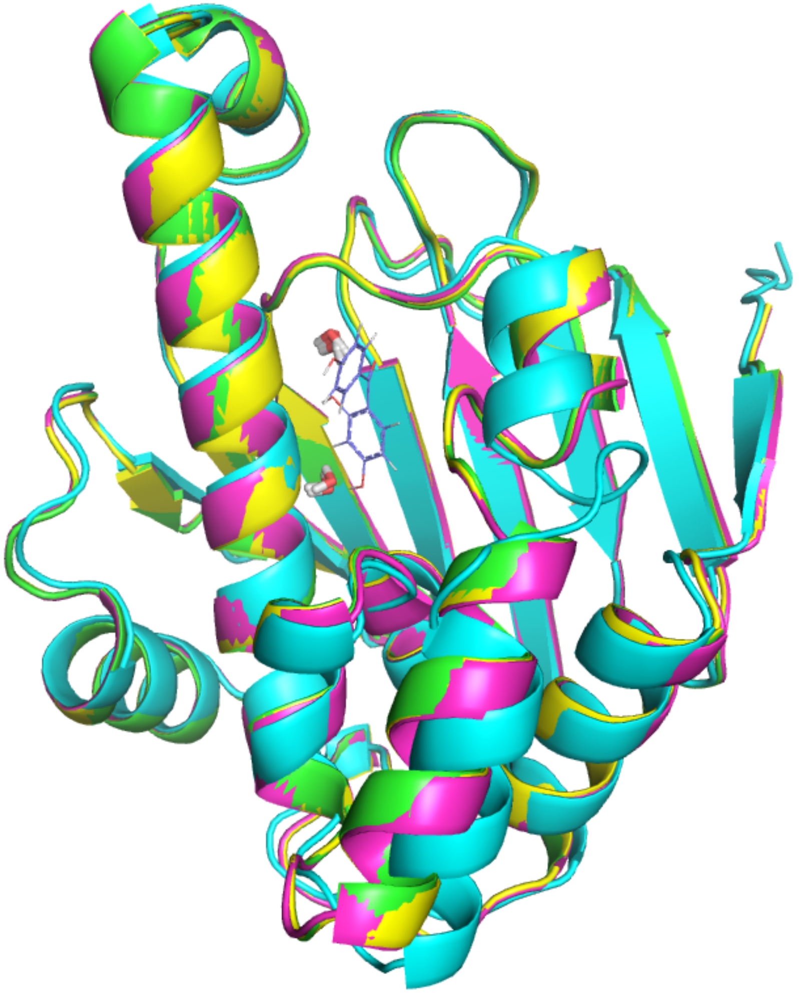
Machine Learning Classification

Machine Learning Regression

Affinity Prediction

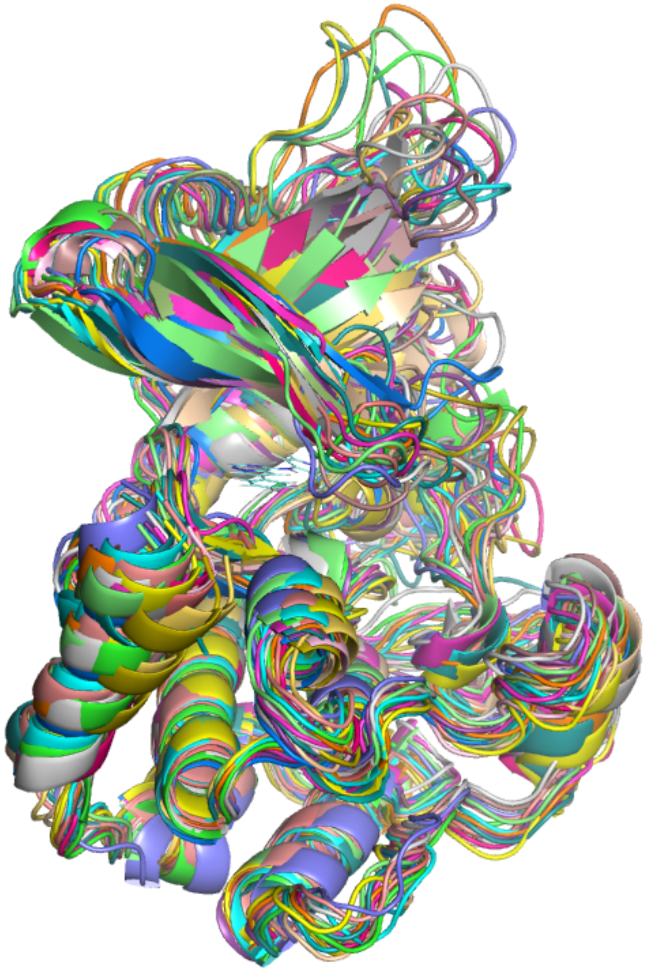
Pose Selection & Affinity Prediction

Docked Poses – Receptors



HSP90

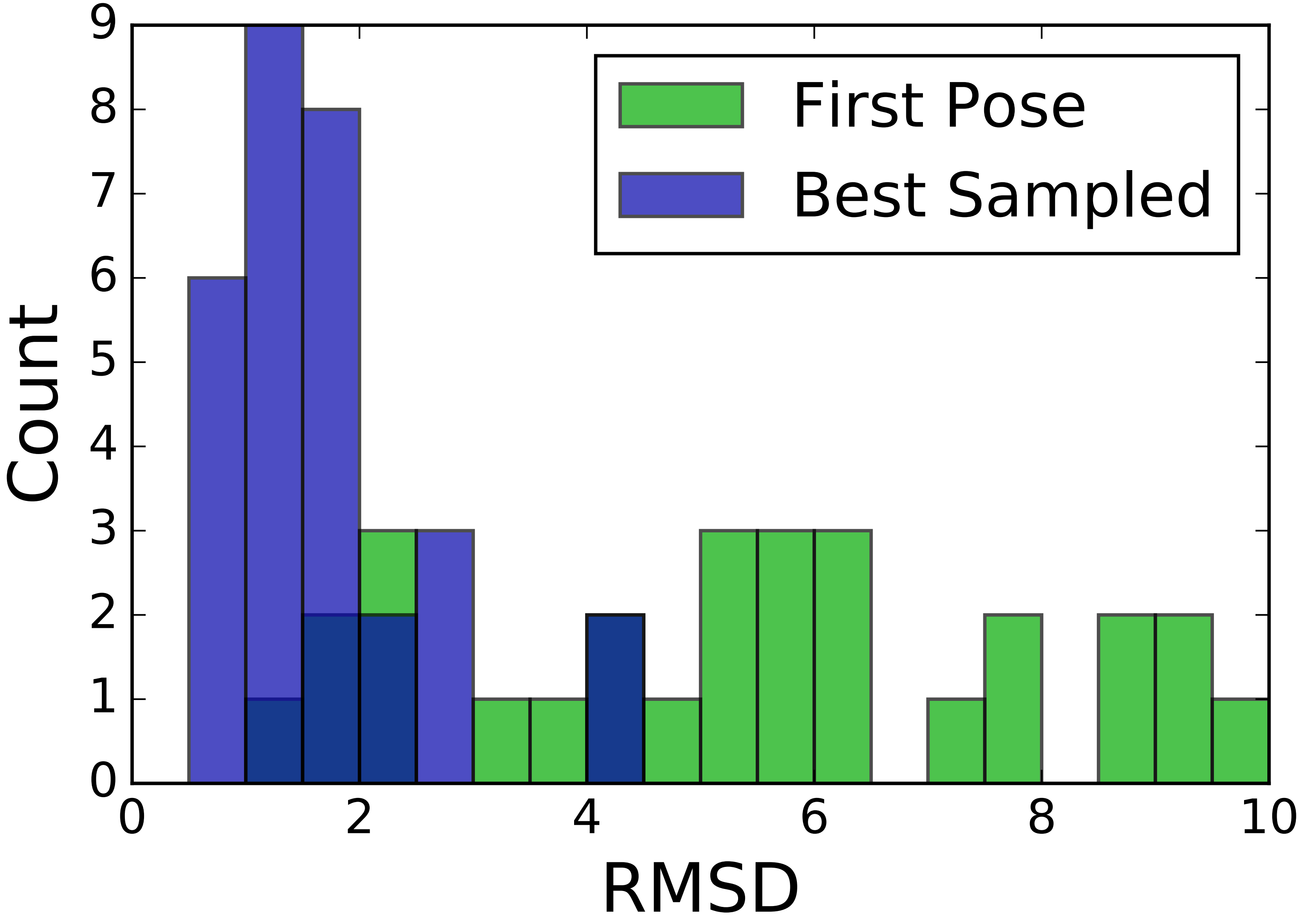
2JJC,2XDX,4YKQ,4YKR
with and without binding site waters



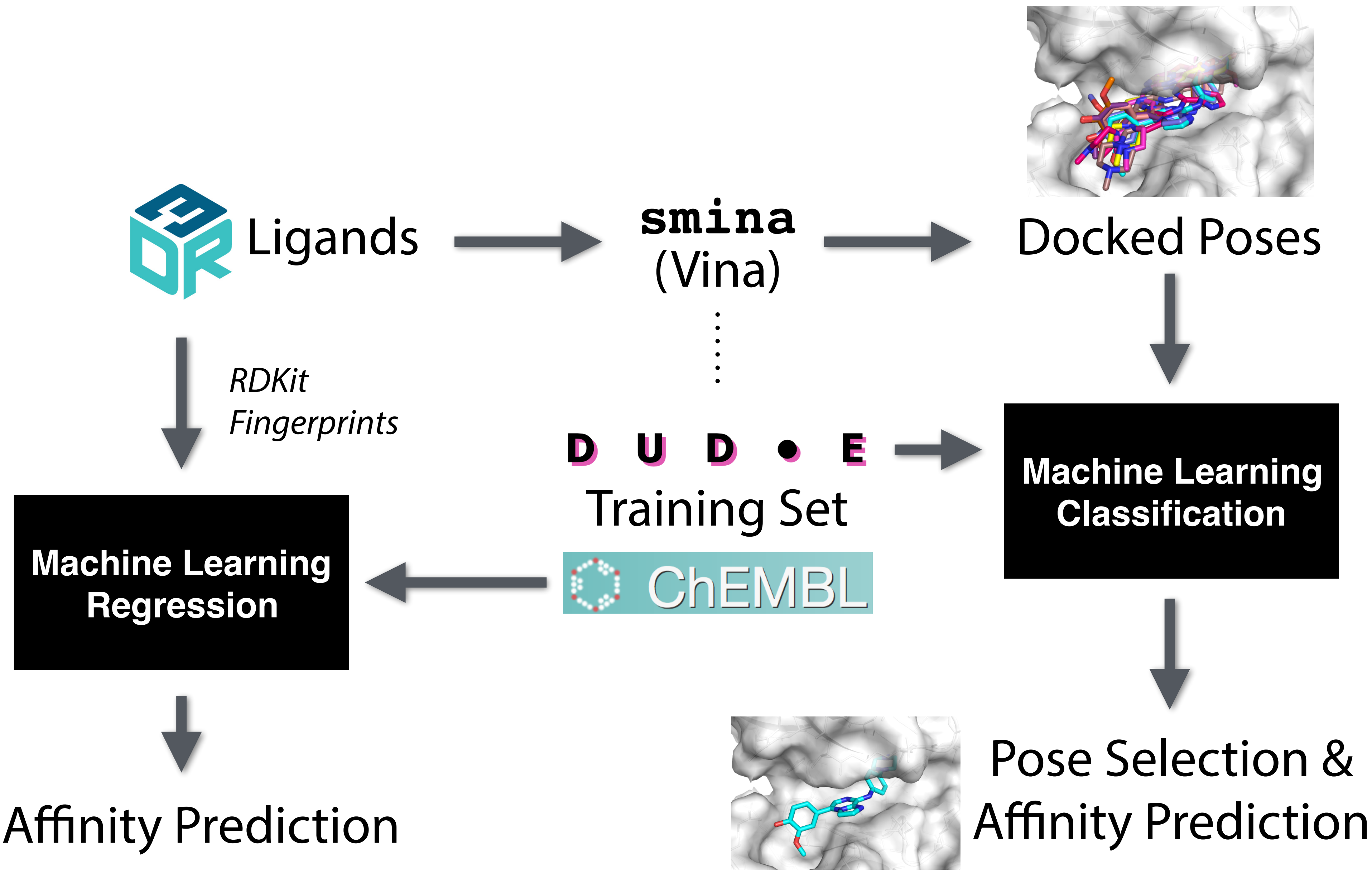
MAP4K4

4OBO,4U44
plus 10 representative snapshots
from 100ns MD

Sampled vs. Selected



Overall Approach



Ligands

smina (Vina)

Docked Poses

RDKit Fingerprints

DUD•E Training Set

ChEMBL

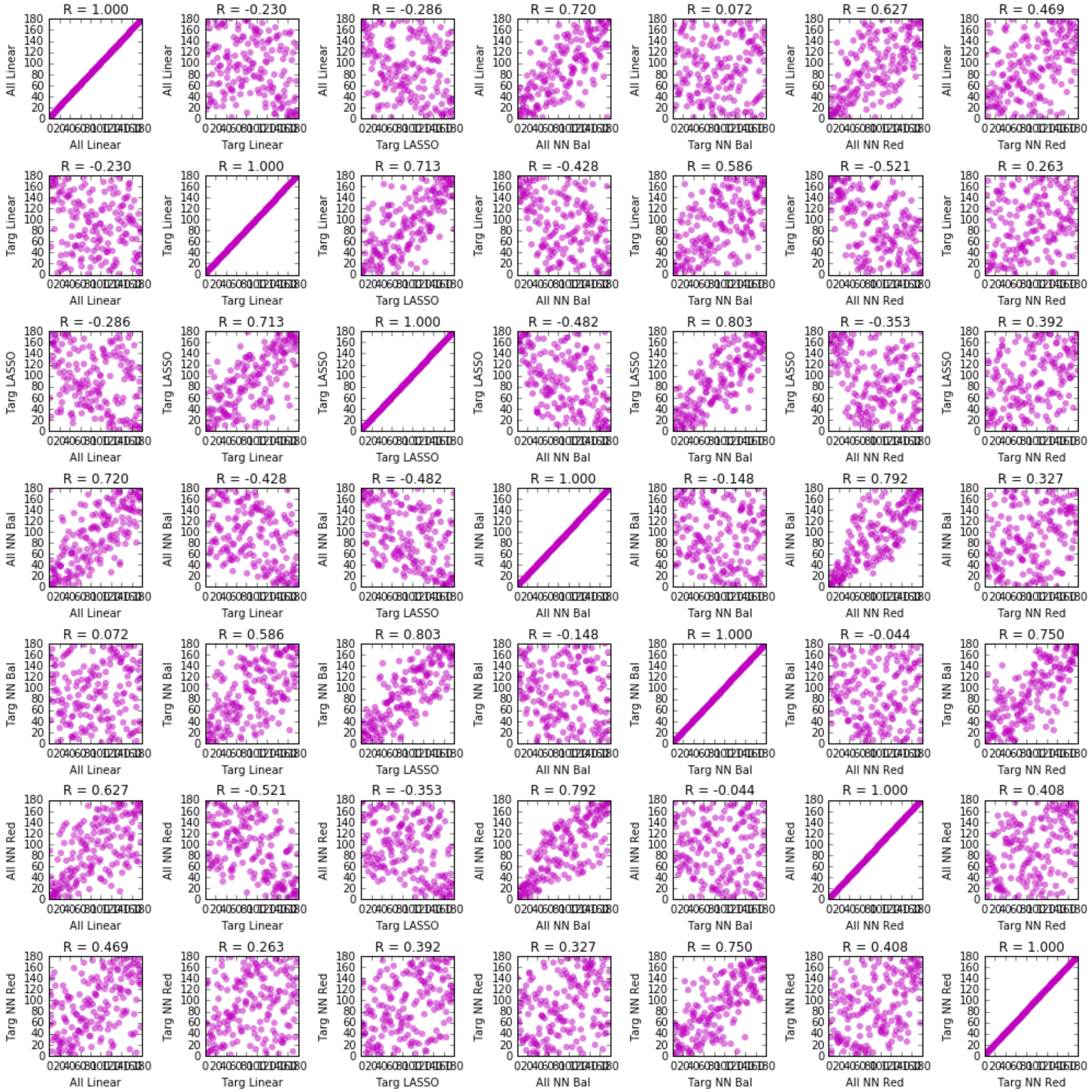
Machine Learning Classification

Machine Learning Regression

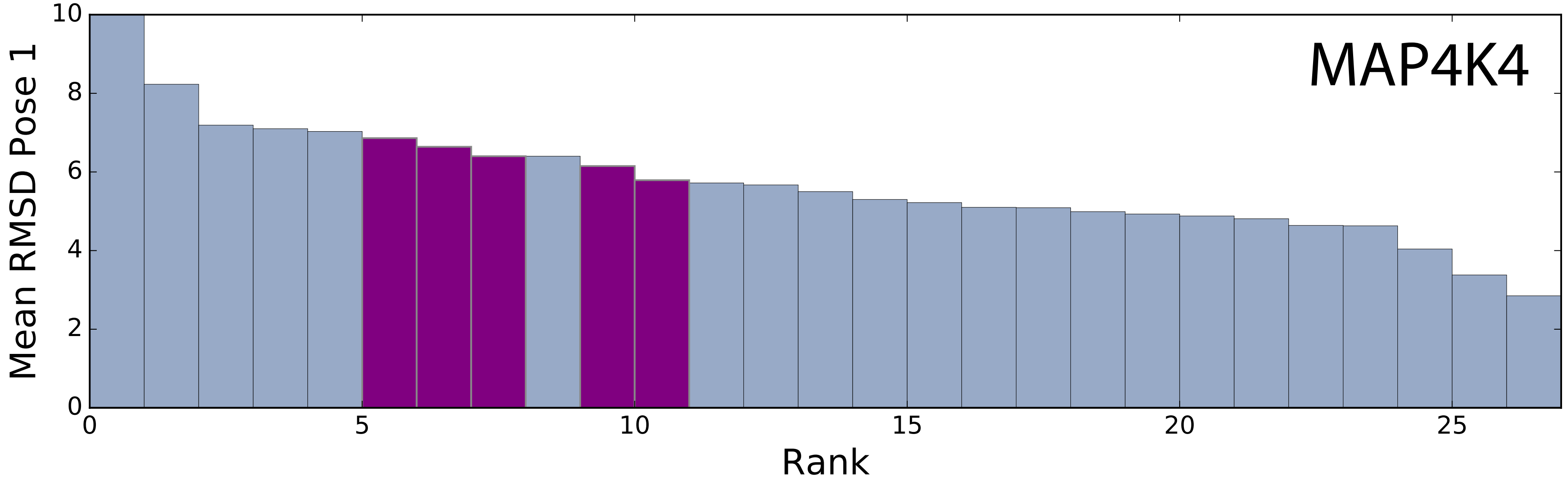
Affinity Prediction

Pose Selection & Affinity Prediction

Results


R = 1.000
R = -0.230
R = -0.286
R = 0.720
R = 0.072
R = 0.627
R = 0.469
R = -0.230
R = 1.000
R = 0.713
R = -0.428
R = 0.586
R = -0.521
R = 0.263
R = -0.286
R = 0.713
R = 1.000
R = -0.482
R = 0.803
R = -0.353
R = 0.392
R = 0.720
R = -0.428
R = -0.482
R = 1.000
R = -0.148
R = 0.792
R = 0.327
R = 0.072
R = 0.586
R = 0.803
R = -0.148
R = 1.000
R = -0.044
R = 0.750
R = 0.627
R = -0.521
R = -0.353
R = 0.792
R = -0.044
R = 1.000
R = 0.408
R = 0.469
R = 0.263
R = 0.392
R = 0.327
R = 0.750
R = 0.408
R = 1.000
All Linear
Targ Linear
Targ LASSO
All NN Bal
Targ NN Bal
All NN Red
Targ NN Red
0
20
40
60
80
100
120
140
160
180
All Linear
Targ Linear
Targ LASSO
All NN Bal
Targ NN Bal
All NN Red
Targ NN Red

Pose Prediction



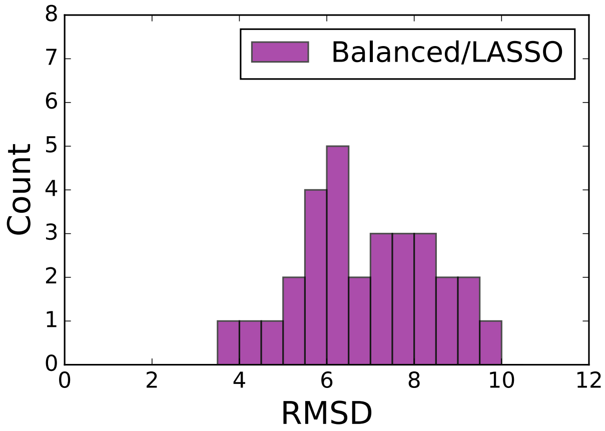
Balanced
LASSO

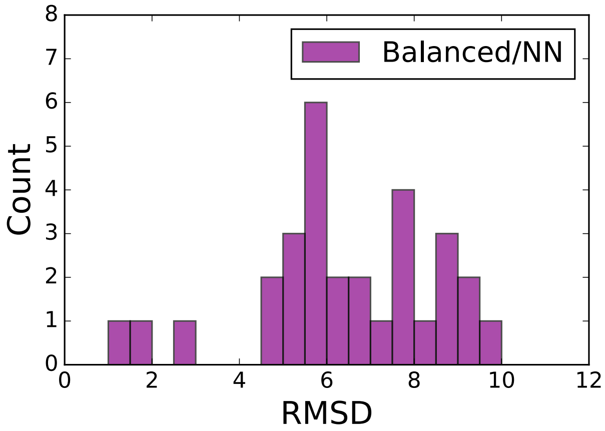
Reduced
NN

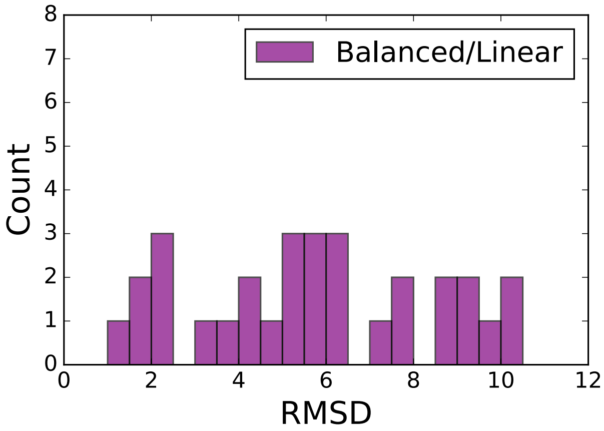
Balanced
NN

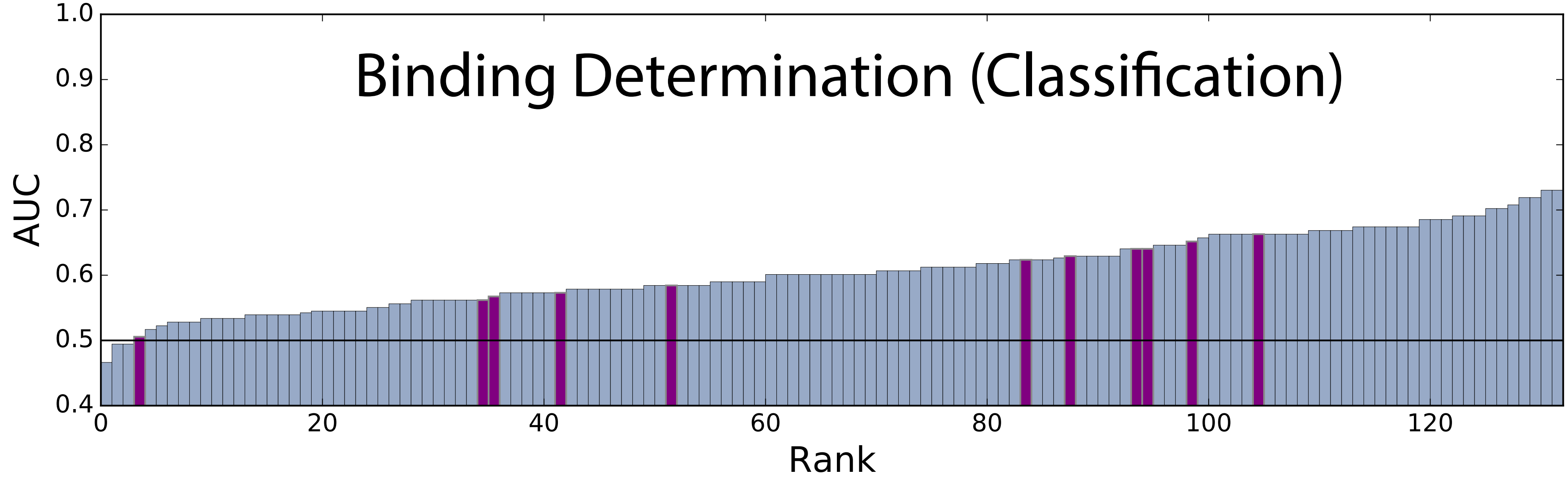
Balanced
Linear

Reduced
Linear









Target LASSO

Target Reduced NN

QSAR SMARTS

QSAR RDKit

Balanced CNN2

Balanced CNN1

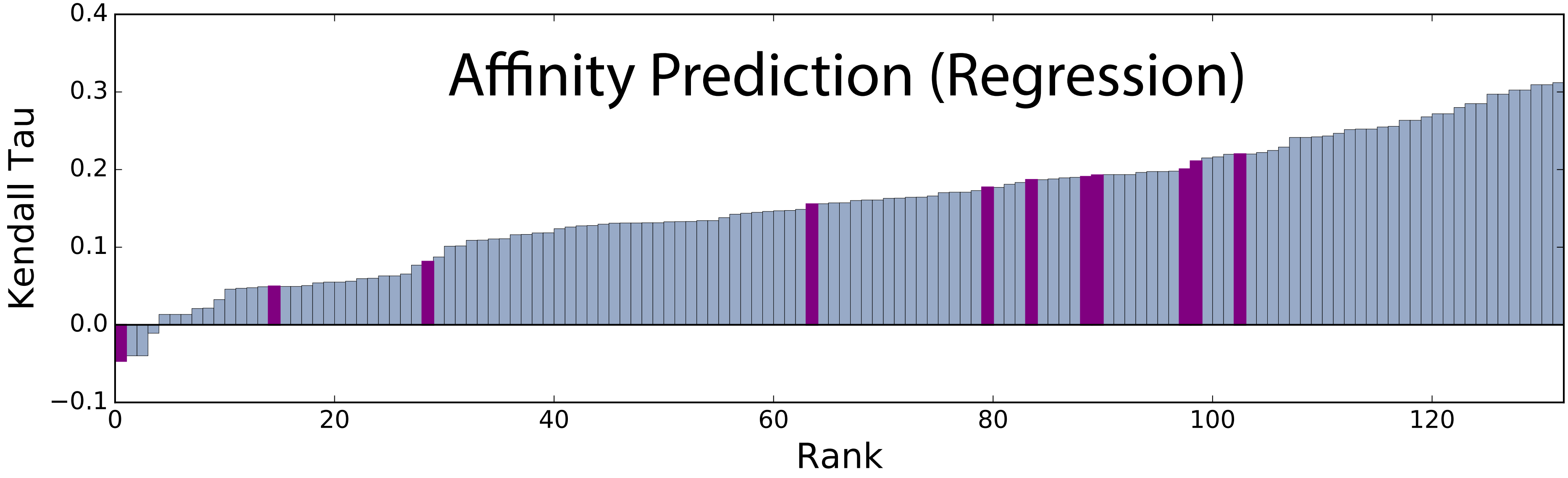
Target Balanced NN

Reduced NN

QSAR ECFP6

Balanced NN

Balanced Linear



Target LASSO

QSAR SMARTS

Target Balanced NN

QSAR RDKit

Balanced CNN2

Reduced NN

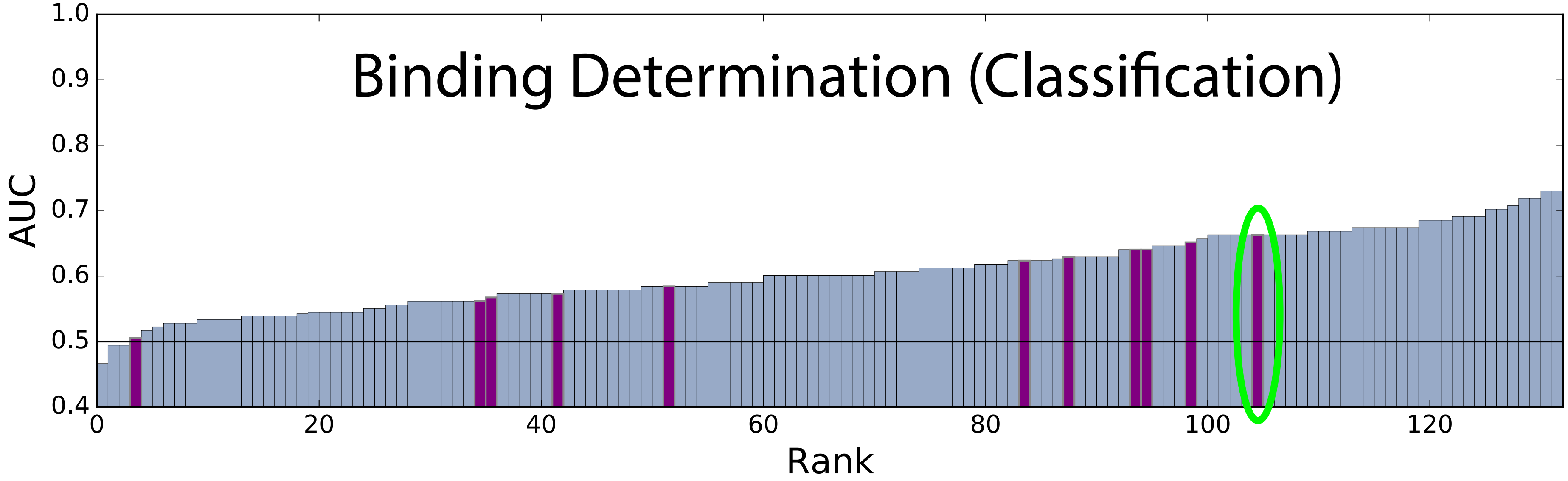
Target Reduced NN

Balanced CNN1

QSAR ECFP6

Balanced NN

Balanced Linear



Target LASSO

Target Reduced NN

QSAR SMARTS

QSAR RDKit

Balanced CNN2

Balanced CNN1

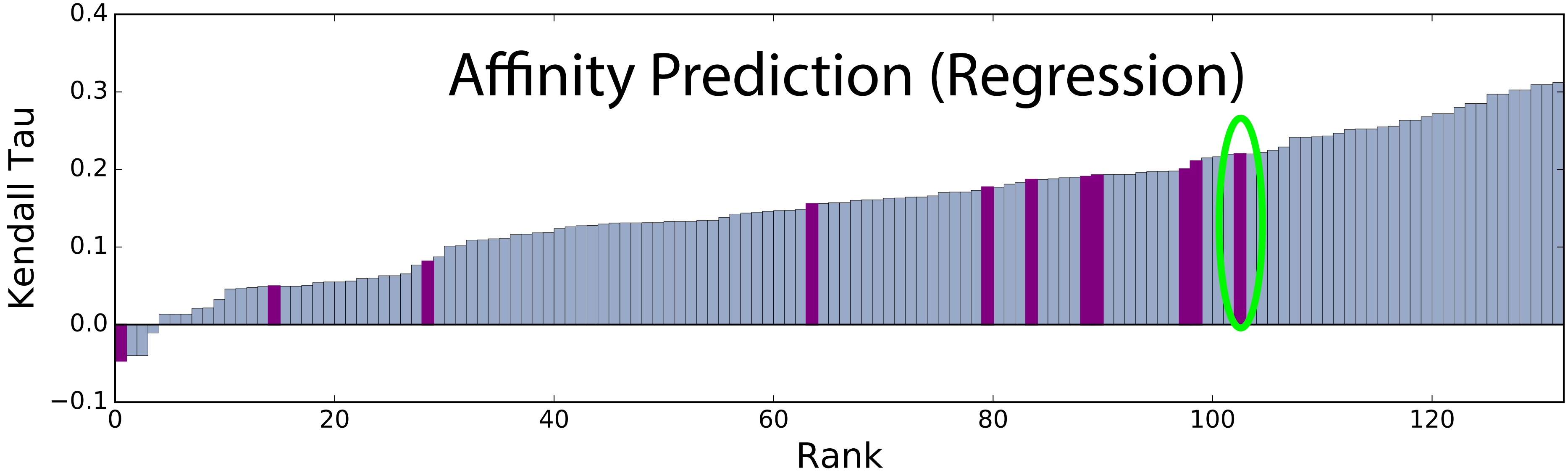
Target Balanced NN

Reduced NN

QSAR ECFP6

Balanced NN

Balanced Linear



Target LASSO

QSAR SMARTS

Target Balanced NN

QSAR RDKit

Balanced CNN2

Reduced NN

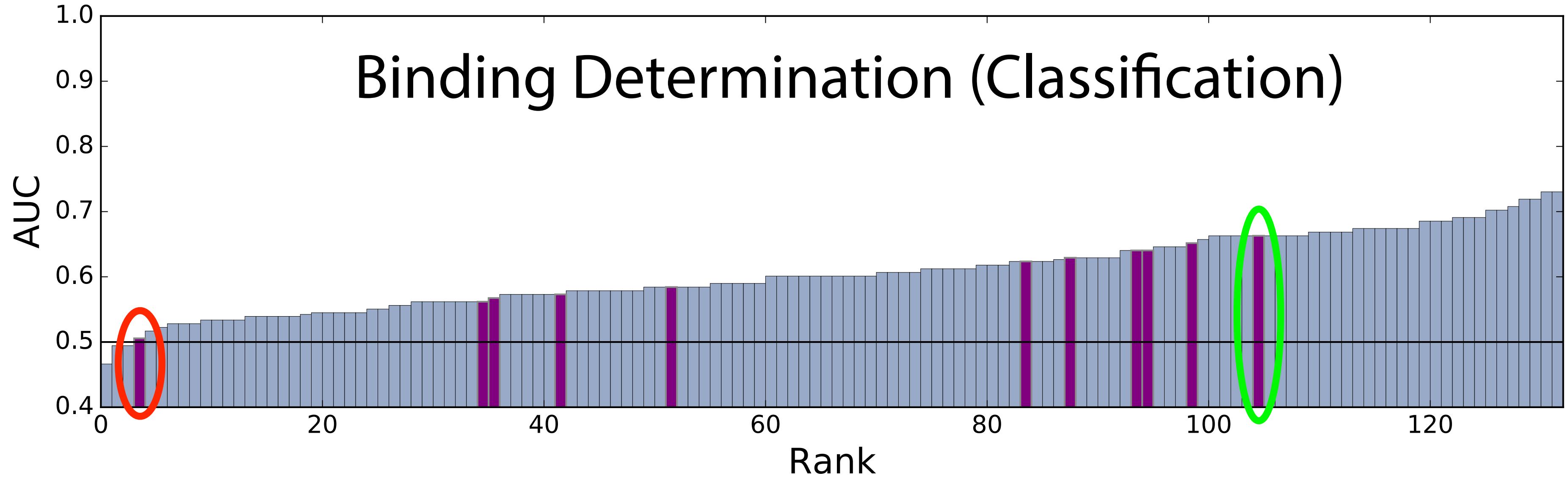
Target Reduced NN

Balanced CNN1

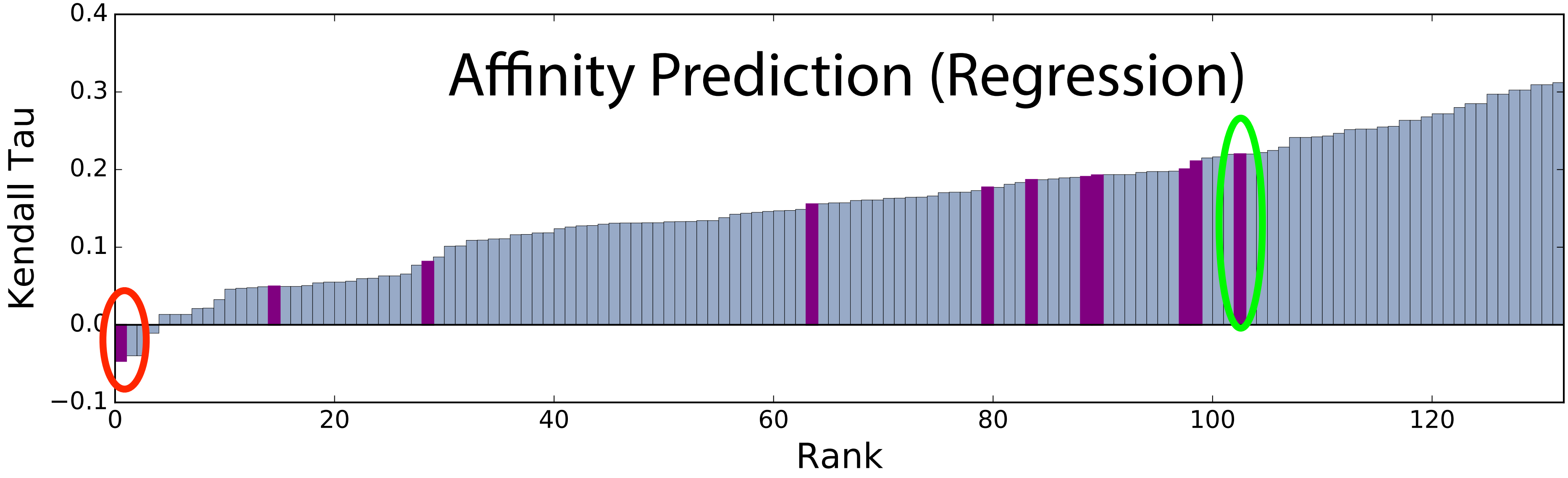
QSAR ECFP6

Balanced NN

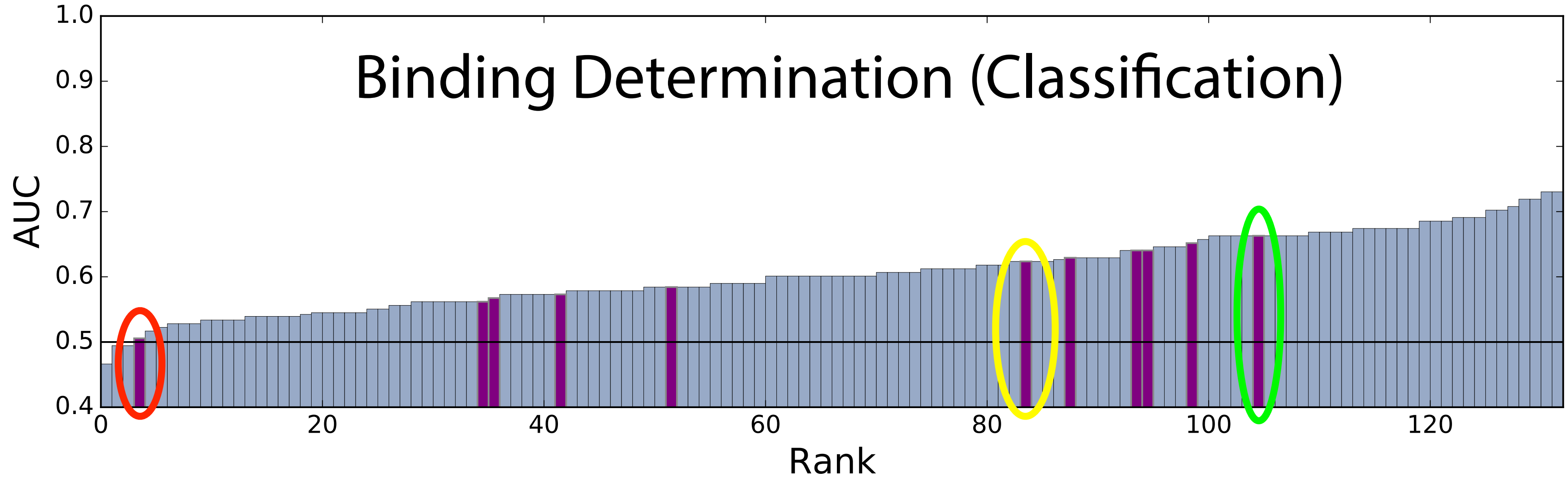
Balanced Linear



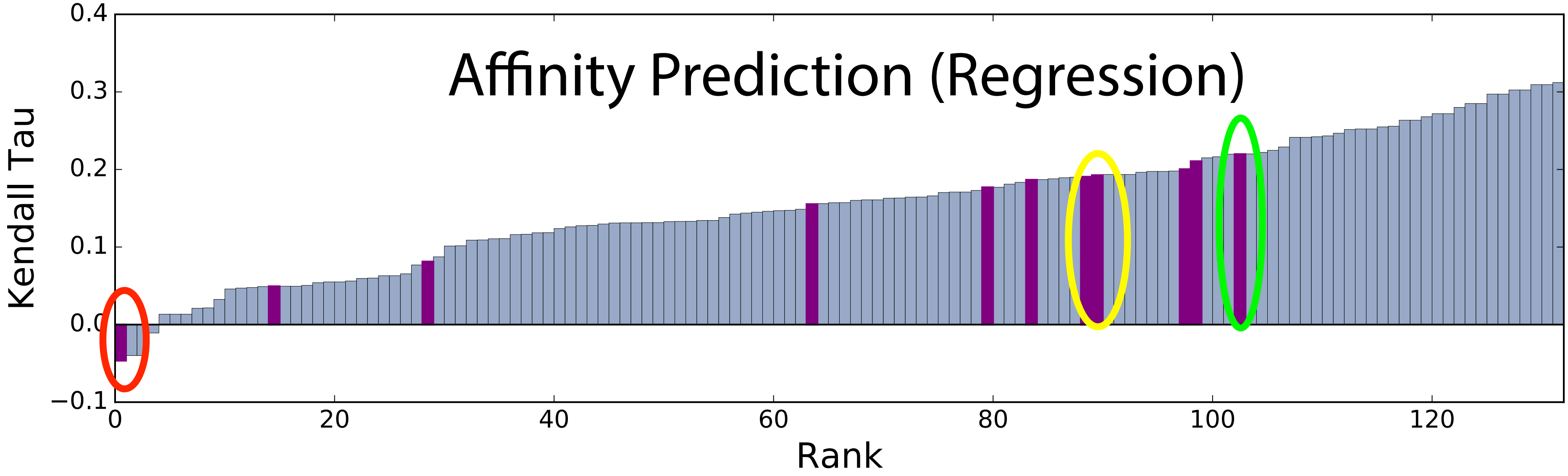
Target LASSO | Target Reduced NN | QSAR SMARTS | QSAR RDKit | Balanced CNN2 | Balanced CNN1 | Target Balanced NN | Reduced NN | QSAR ECFP6 | Balanced NN | Balanced Linear



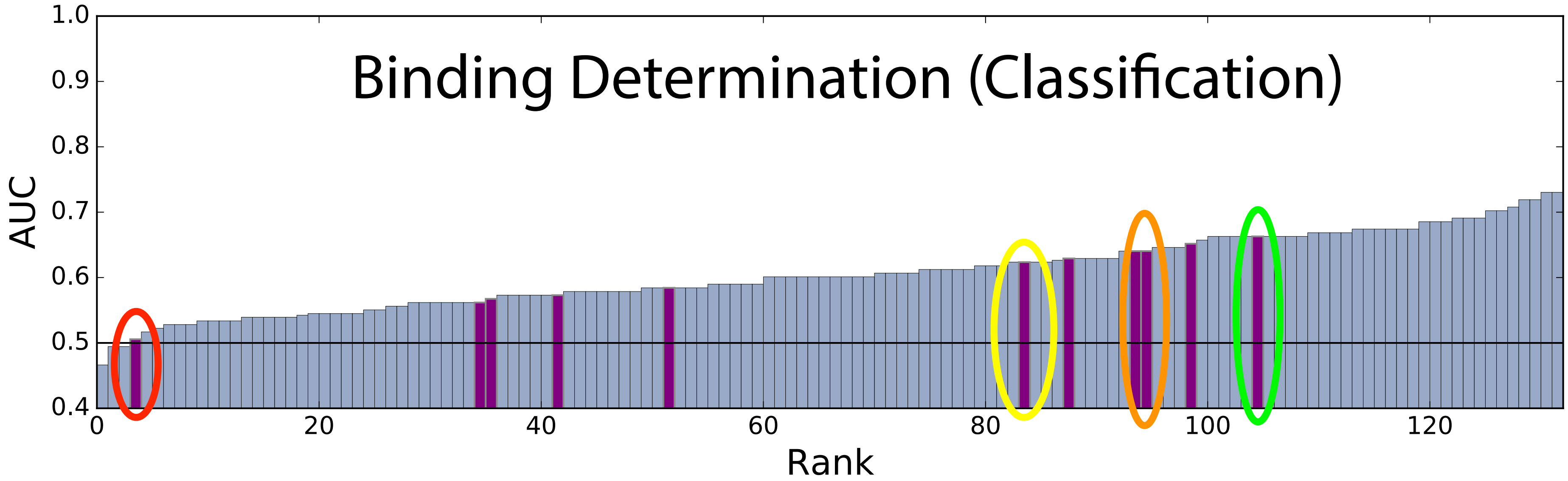
Target LASSO | QSAR SMARTS | Target Balanced NN | QSAR RDKit | Balanced CNN2 | Reduced NN | Target Reduced NN | Balanced CNN1 | QSAR ECFP6 | Balanced NN | Balanced Linear



Target LASSO | Target Reduced NN | QSAR SMARTS | QSAR RDKit | Balanced CNN2 | Balanced CNN1 | Target Balanced NN | Reduced NN | QSAR ECFP6 | Balanced NN | Balanced Linear



Target LASSO | QSAR SMARTS | Target Balanced NN | QSAR RDKit | Balanced CNN2 | Reduced NN | Target Reduced NN | Balanced CNN1 | QSAR ECFP6 | Balanced NN | Balanced Linear



Target LASSO

Target Reduced NN

QSAR SMARTS

QSAR RDKit

Balanced CNN2

Balanced CNN1

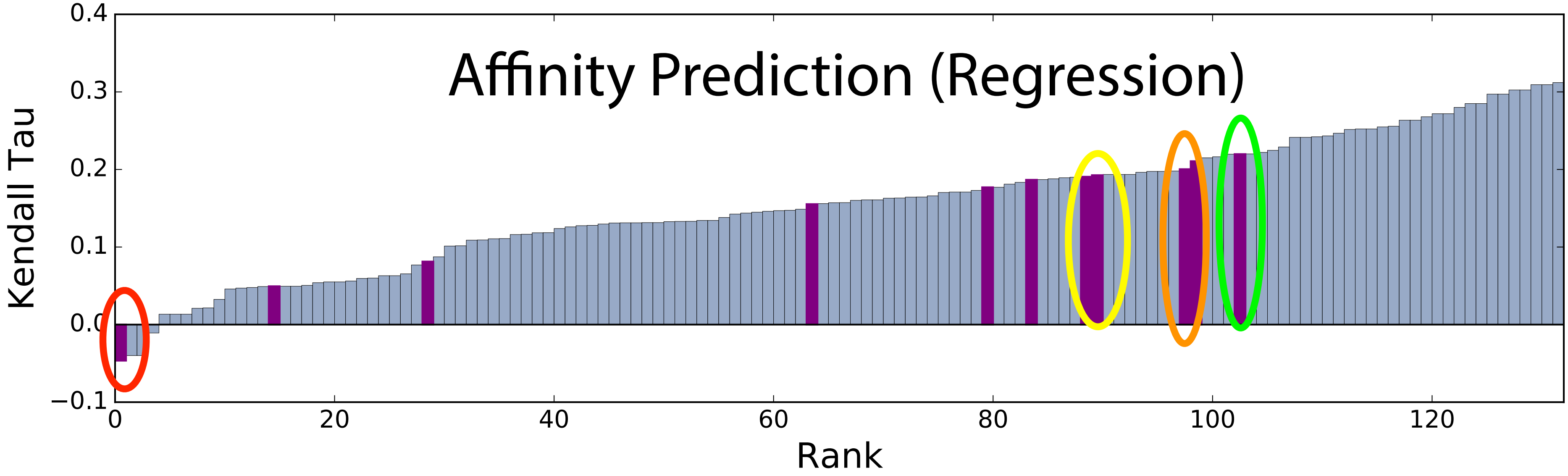
Target Balanced NN

Reduced NN

QSAR ECFP6

Balanced NN

Balanced Linear



Target LASSO

QSAR SMARTS

Target Balanced NN

QSAR RDKit

Balanced CNN2

Reduced NN

Target Reduced NN

Balanced CNN1

QSAR ECFP6

Balanced NN

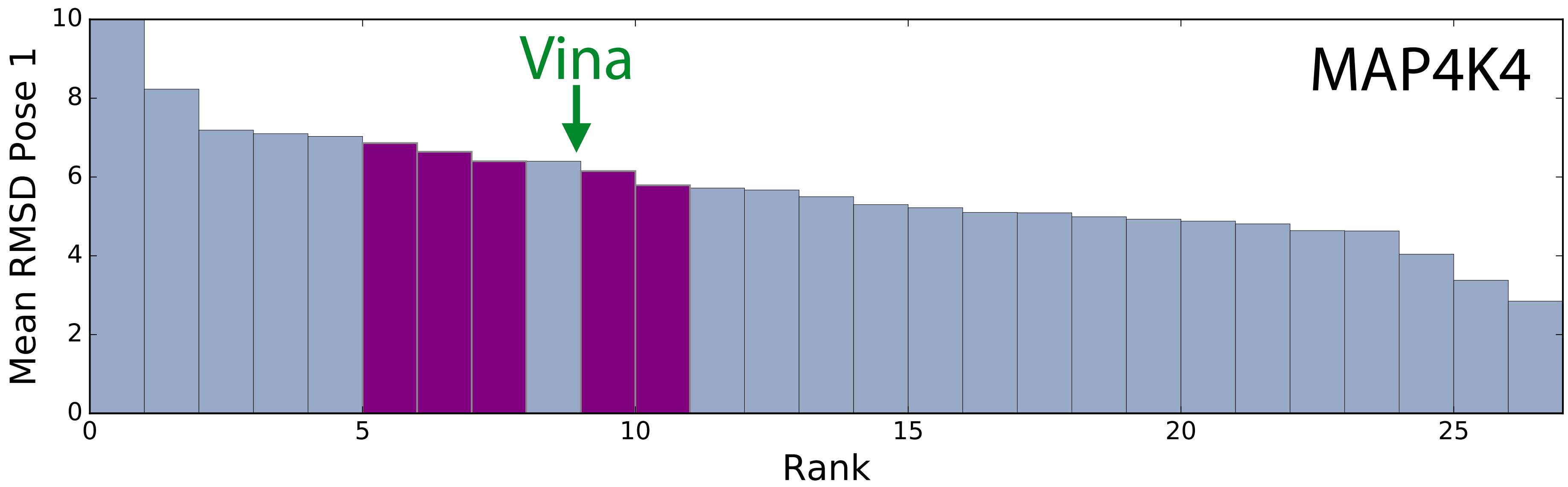
Balanced Linear

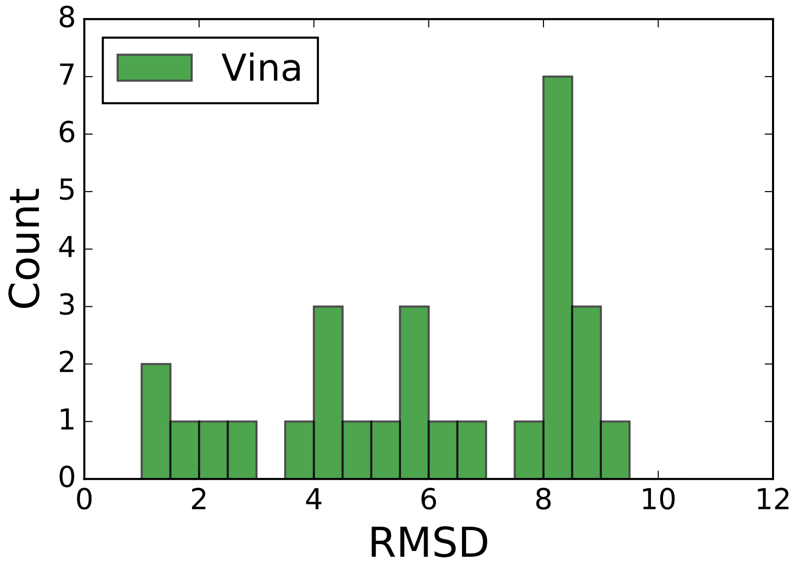
Did we improve Vina?

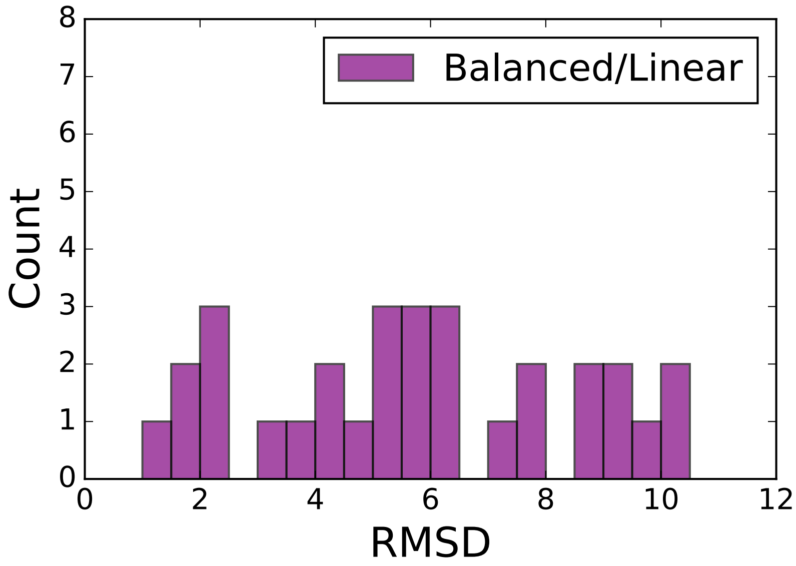
Did we improve Vina?

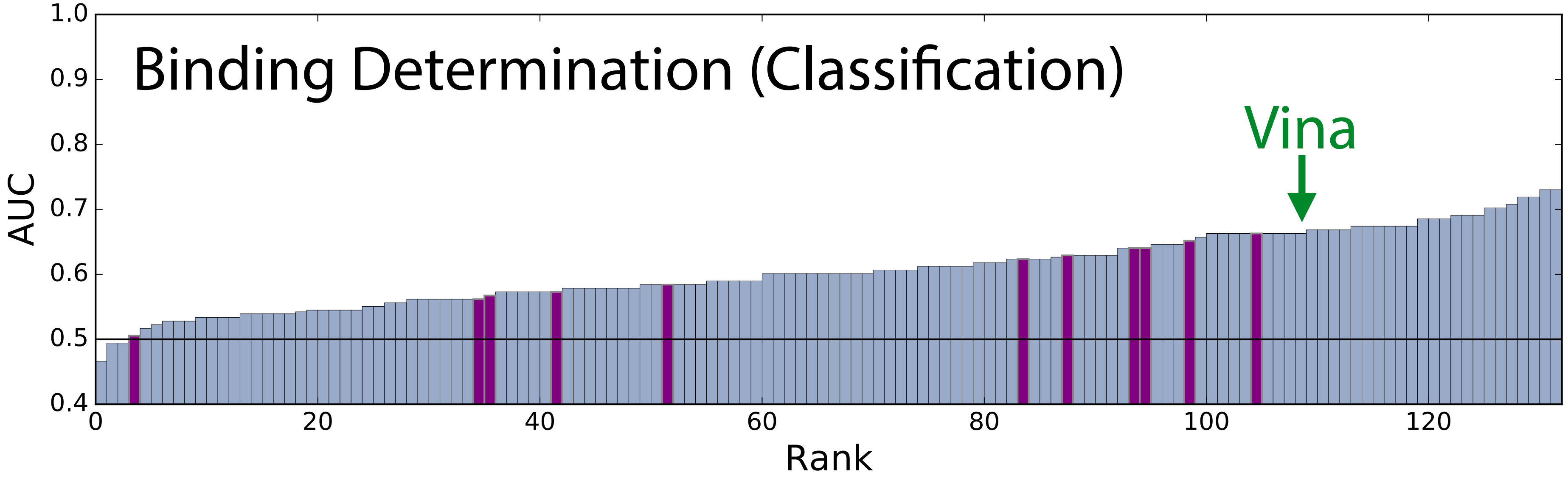
Not really.

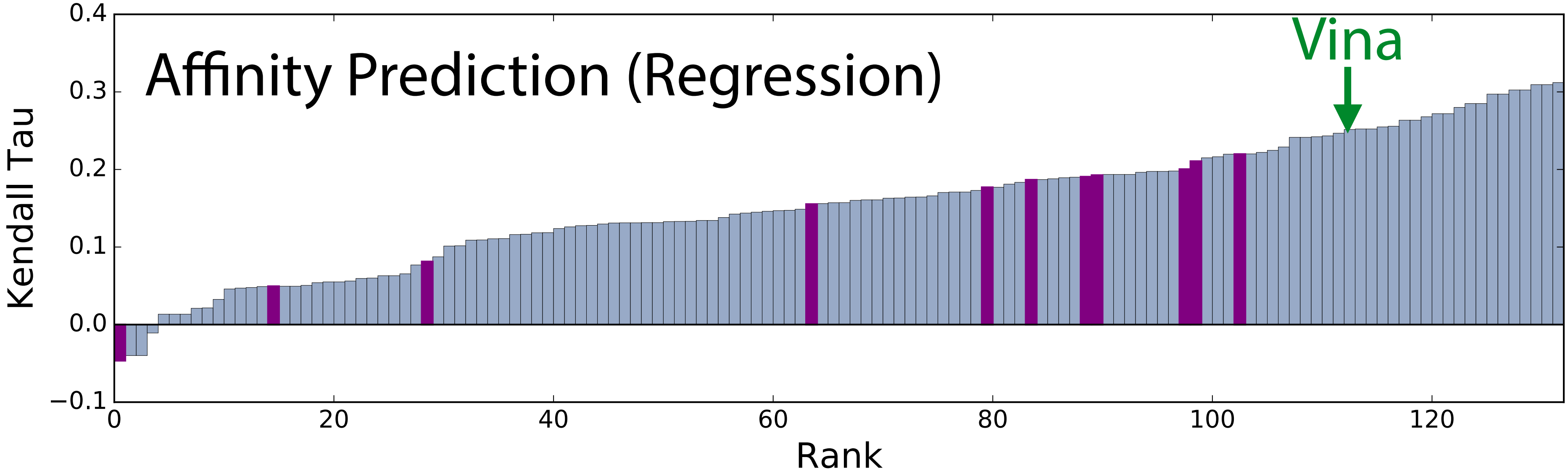
Did we improve Vina?








Binding Determination (Classification)
Vina
AUC
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0
20
40
60
80
100
120
Rank


Affinity Prediction (Regression)
Vina
Kendall Tau
0.4
0.3
0.2
0.1
0.0
−0.1
0
20
40
60
80
100
120
Rank

Conclusions

- Minimal improvement in pose selection
- No improvement with affinity prediction

- Training set is key
- Cross-purpose validation is possible
- CNN scoring has promise
- 2D did not beat 3D

Acknowledgements

Drug Design Data Resource

Students

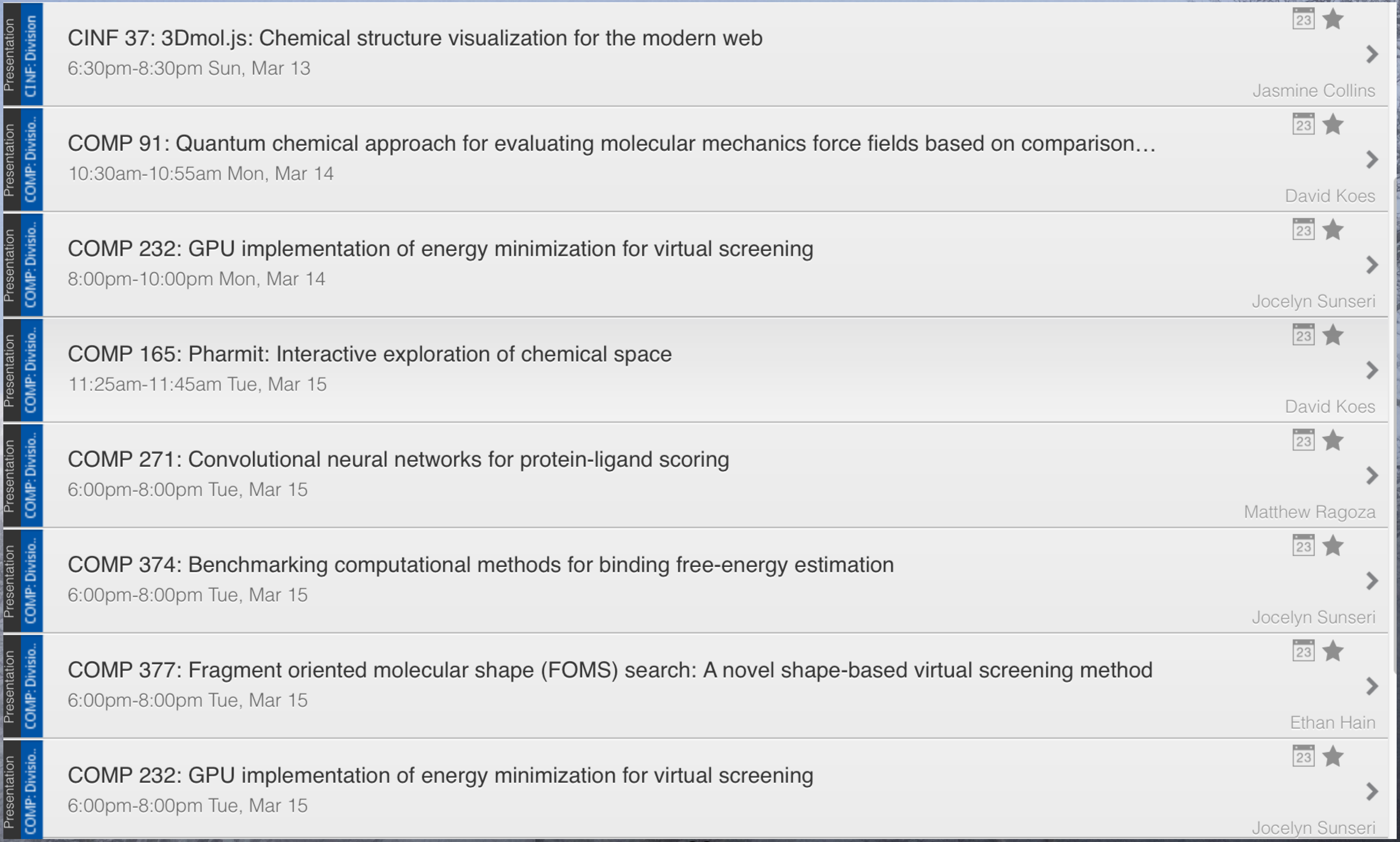
Jasmine Collins
Matthew Ragoza

Noah Bastola
Jesse Bracho
Jocelyn Sunseri

Funding

R01GM108340

Questions?



CINF 37: 3Dmol.js: Chemical structure visualization for the modern web
6:30pm-8:30pm Sun, Mar 13
Jasmine Collins

COMP 91: Quantum chemical approach for evaluating molecular mechanics force fields based on comparison…
10:30am-10:55am Mon, Mar 14
David Koes

COMP 232: GPU implementation of energy minimization for virtual screening
8:00pm-10:00pm Mon, Mar 14
Jocelyn Sunseri

COMP 165: Pharmit: Interactive exploration of chemical space
11:25am-11:45am Tue, Mar 15
David Koes

COMP 271: Convolutional neural networks for protein-ligand scoring
6:00pm-8:00pm Tue, Mar 15
Matthew Ragoza

COMP 374: Benchmarking computational methods for binding free-energy estimation
6:00pm-8:00pm Tue, Mar 15
Jocelyn Sunseri

COMP 377: Fragment oriented molecular shape (FOMS) search: A novel shape-based virtual screening method
6:00pm-8:00pm Tue, Mar 15
Ethan Hain

COMP 232: GPU implementation of energy minimization for virtual screening
6:00pm-8:00pm Tue, Mar 15
Jocelyn Sunseri