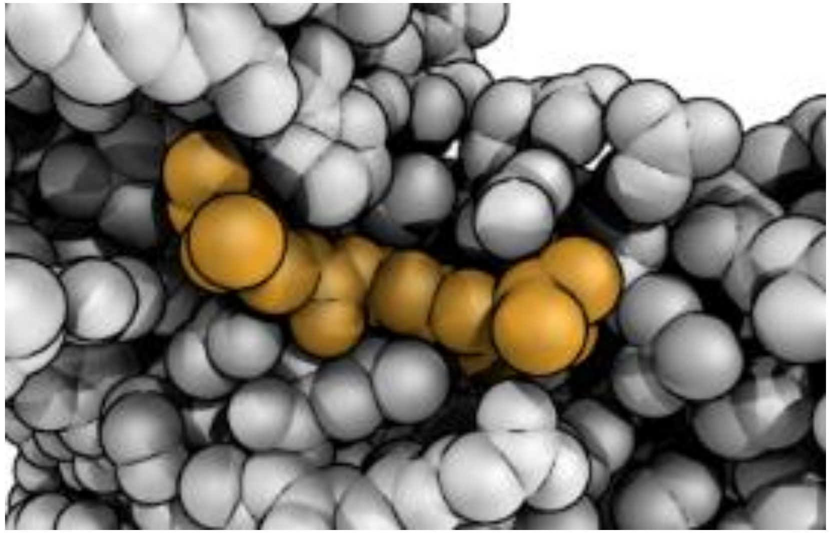
OSGUS, July 13, 2018

Exploring vHTS approaches with HTC

Spencer S. Ericksen

Associate Scientist

UW-CCC, Drug Development Core, Small Molecule Screening Facility



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

Open Science Grid

CENTER FOR HIGH THROUGHPUT COMPUTING

HTCondor
High Throughput Computing

Why am I here?

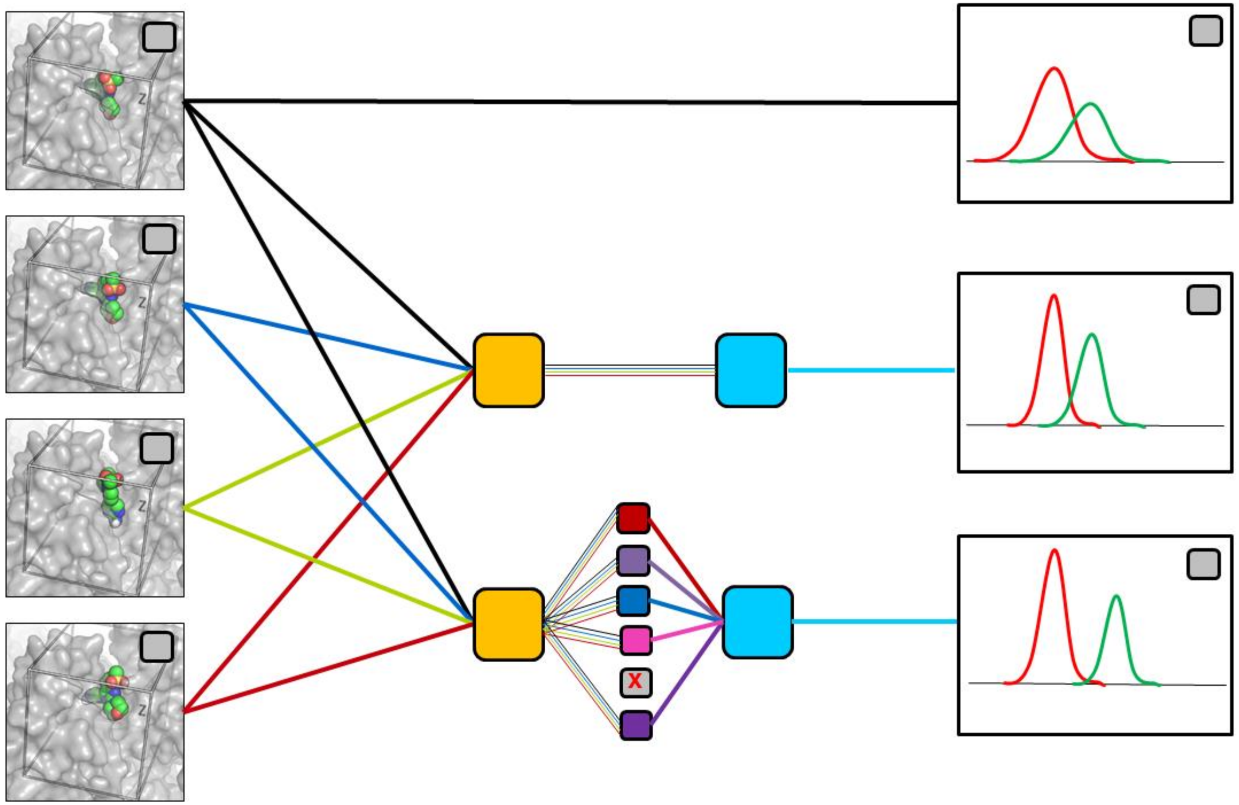
- We want to promote early stage drug discovery efforts on campus!
- Need to reduce costs to increase participation.
- Early Stage Drug Discovery: looking for needle in haystack.
- **HTS** assays of 10,000s to millions compounds.
- $1-100 per compound!

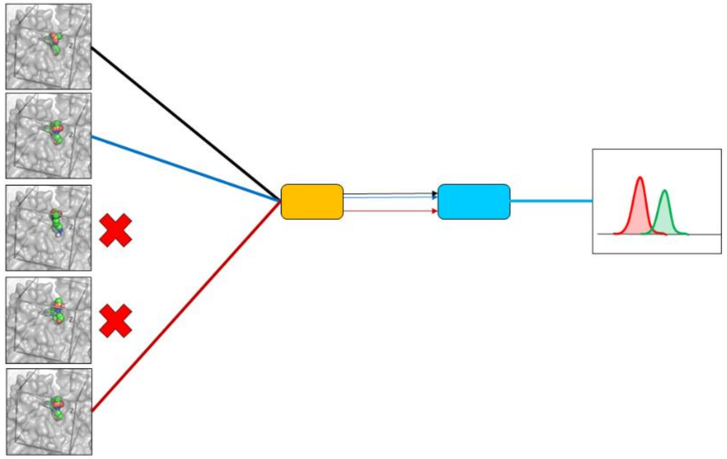
What is **v**HTS?

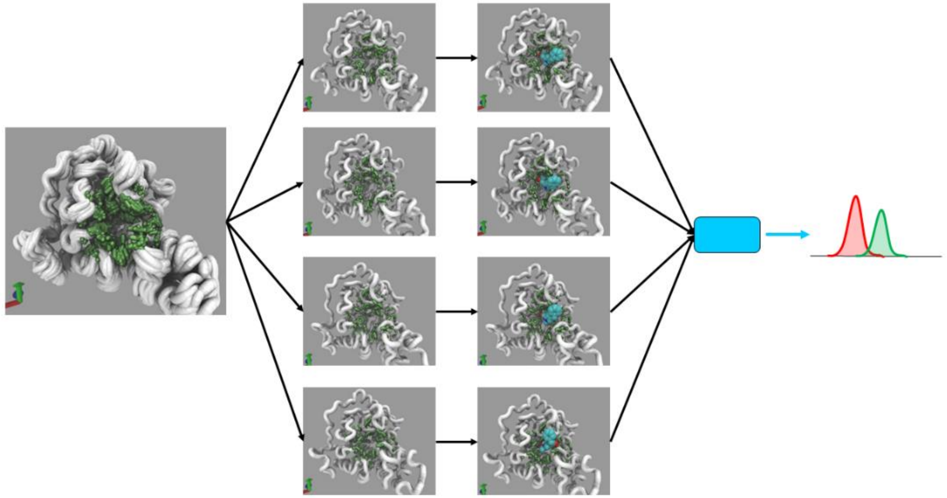
- Filter using **vHTS** first! Prioritize a much smaller subset testing.
- Enrichment for active compounds can greatly reduce costs.
- Use **docking** to predict potential for compound-**target** interaction.

Overview

- vHTS: the structure-based approach
- single docking programs
- consensus scoring
- advanced consensus scoring
- pose consensus scoring
- ensemble consensus scoring

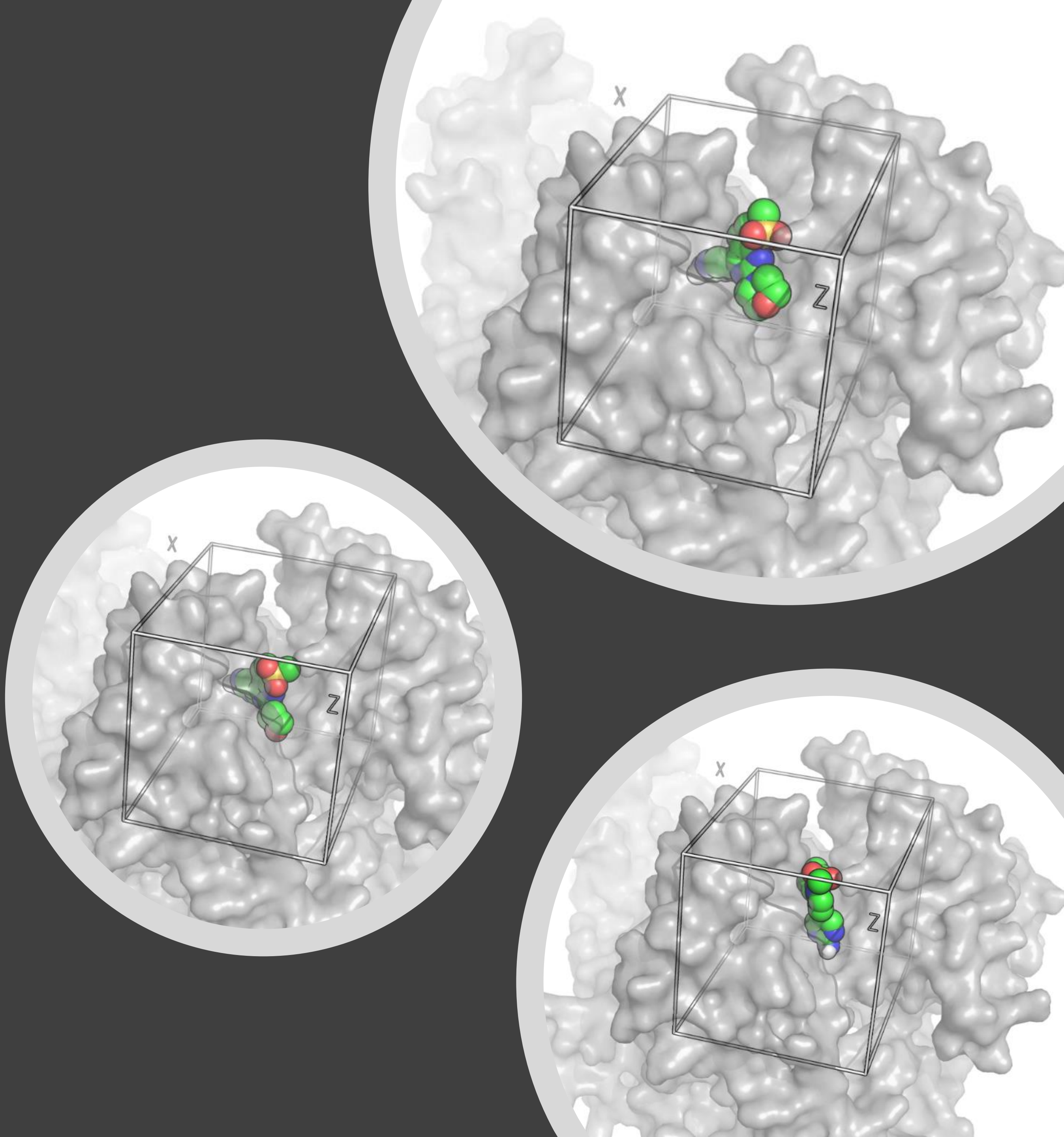






What is docking?

- Docking looks for best compound binding orientation on a 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.



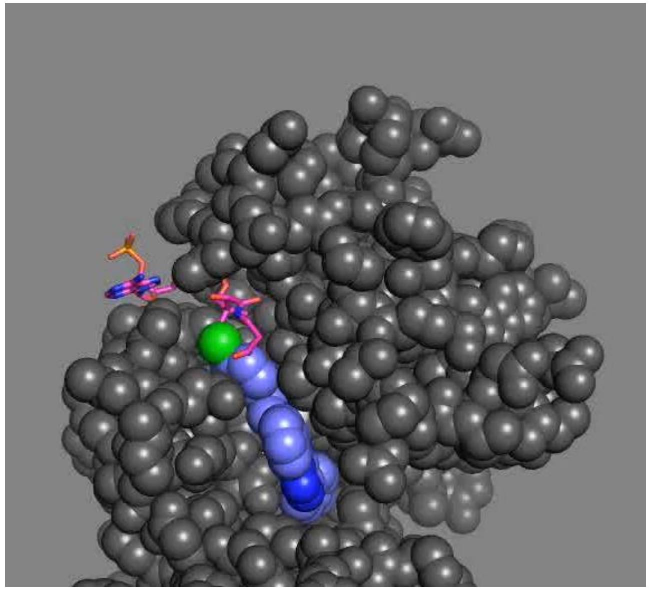
DUD-E: Benchmarking Data Set to Validate Docking-Based VS Methods

- *"A* ***D****atabase of* ***U****seful* ***D****ecoys:* ***E****nhanced"*
- 102 protein targets
- 22,886 active compounds with minimum potency 1 µM (or better)
- 100-600 ligands per target
- ~50 decoys for each active ligand (~2% actives)
- Decoys property-matched but dissimilar 2-D topology.
  - Properties: MW, LogP, HBA, HBD, rotatable bonds, net charge
  - ECFP4, keep 25% most dissimilar
- Actives are clustered. Diversity of actives is promoted by keeping max of 3 tightest binders in each cluster.

Mysinger et al., J. Med. Chem., 2012, Jul 5. doi 10.1021/jm300687e .

Structure-based virtual screening

Dock Compound Library

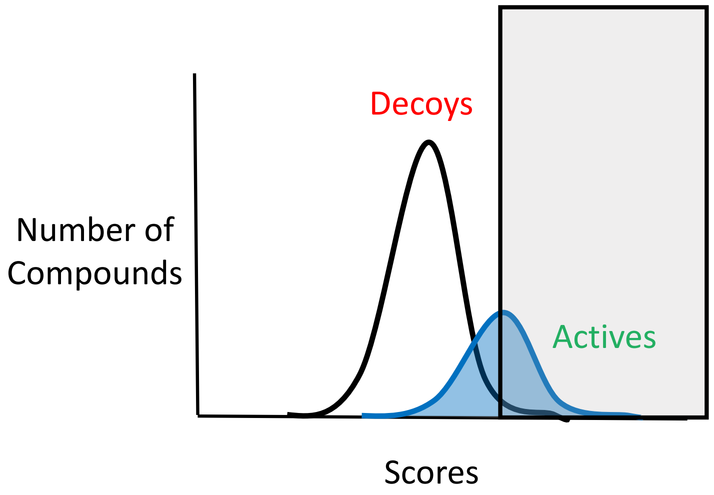


MOLID	SCORE
ZINC36206438	58.63
ZINC59310217	58.72
ZINC61596674	56.35
ZINC67458535	47.40
CHEMBL1221861	60.66
ZINC10123401	52.39
ZINC64526095	66.13
ZINC24002103	56.72
ZINC09612655	58.84
ZINC24002105	38.95
CHEMBL38532	74.19
ZINC40824467	50.10
ZINC59829723	58.29
ZINC37520295	44.78
ZINC49812309	38.01
ZINC14558020	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 programs have different search and scoring strategies

Docking Program	Search Algorithm	Scoring Function
AutoDock v4.2	Lamarkian Genetic Algorithm with Simulated Annealing	Forcefield
DOCK v6.7	Incremental Construction (Anchor-and-grow)	Forcefield
FRED v3.0.1	Exhaustive rigid docking search, discretized configuration space	Empirical
HYBRID v3.0.1	Exhaustive rigid docking search, discretized configuration space	Empirical + Knowledge-Based
PLANTS v1.2	Ant Colony Optimization	Empirical
rDock v2013.1	Genetic Algorithm, Monte Carlo, Minimization	Empirical
Smina (Vina) 1.1.2	Exhaustive flexible docking search, discretized configuration space	Knowledge-Based
Surflex v3.040	Incremental Construction with Matching Algorithm	Empirical

Docking

Scoring

No single program works for all targets

- No way to decide *a priori* which program is best for a new target

Individual Docking Algorithms

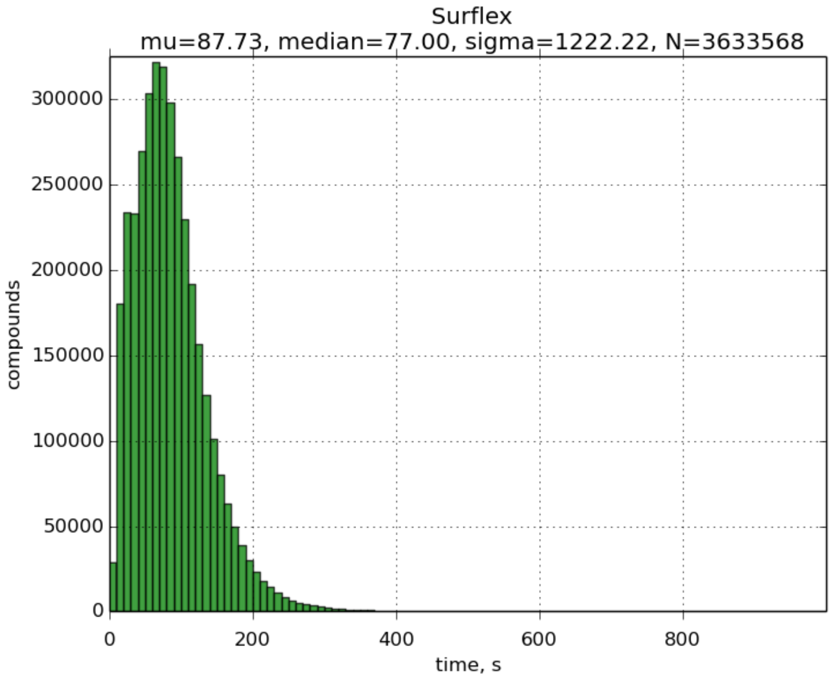
Target Class	Target	AD42	DOCK6	FRED	HYBRID	PLANTS	rDock	Smina	Surflex	Best
GPCR	ADRB1	0.68	0.78	0.77	0.65	0.86	0.81	0.79	0.80	0.86
GPCR	DRD3	0.69	0.59	0.79	0.81	0.69	0.66	0.68	0.71	0.81
Ion Channel	GRIA2	0.73	0.60	0.79	0.77	0.73	0.77	0.75	0.77	0.79
Kinase	BRAF	0.73	0.60	0.75	0.69	0.54	0.79	0.86	0.71	0.86
Kinase	CDK2	0.76	0.61	0.81	0.85	0.68	0.74	0.71	0.69	0.85
Kinase	PLK1	0.60	0.48	0.80	0.75	0.65	0.68	0.57	0.60	0.80
Kinase	SRC	0.65	0.64	0.65	0.66	0.52	0.68	0.67	0.66	0.68
Miscellaneous	FABP4	0.67	0.54	0.84	0.82	0.74	0.60	0.77	0.79	0.84
Receptor	ESR1	0.82	0.54	0.88	0.81	0.77	0.87	0.86	0.74	0.88
Receptor	ESR2	0.77	0.48	0.89	0.89	0.69	0.80	0.79	0.68	0.89
Other Enzymes	ACE	0.78	0.72	0.80	0.84	0.84	0.62	0.61	0.76	0.84
Other Enzymes	GLCM	0.55	0.60	0.70	0.81	0.64	0.77	0.51	0.79	0.81
Other Enzymes	HDAC8	0.70	0.90	0.87	0.76	0.82	0.71	0.86	0.83	0.90
Other Enzymes	HIVINT	0.54	0.65	0.74	0.60	0.76	0.67	0.81	0.66	0.81
Other Enzymes	PDE5A	0.68	0.65	0.84	0.82	0.79	0.78	0.74	0.66	0.84
Other Enzymes	PTN1	0.66	0.76	0.76	0.78	0.72	0.76	0.66	0.88	0.88
Protease	ADA17	0.51	0.40	0.59	0.69	0.58	0.58	0.54	0.70	0.70
Protease	FA10	0.86	0.81	0.79	0.82	0.80	0.90	0.84	0.76	0.90
Protease	HIVPR	0.63	0.66	0.74	0.78	0.79	0.64	0.74	0.81	0.81
Protease	MMP13	0.67	0.60	0.77	0.87	0.71	0.67	0.67	0.76	0.87
Protease	TRY1	0.79	0.82	0.80	0.83	0.81	0.74	0.75	0.93	0.93
	mean	0.69	0.64	0.78	0.78	0.72	0.73	0.72	0.75	0.84
	std. dev.	0.09	0.12	0.07	0.08	0.10	0.09	0.10	0.08	0.06

Docking Compute Expenses

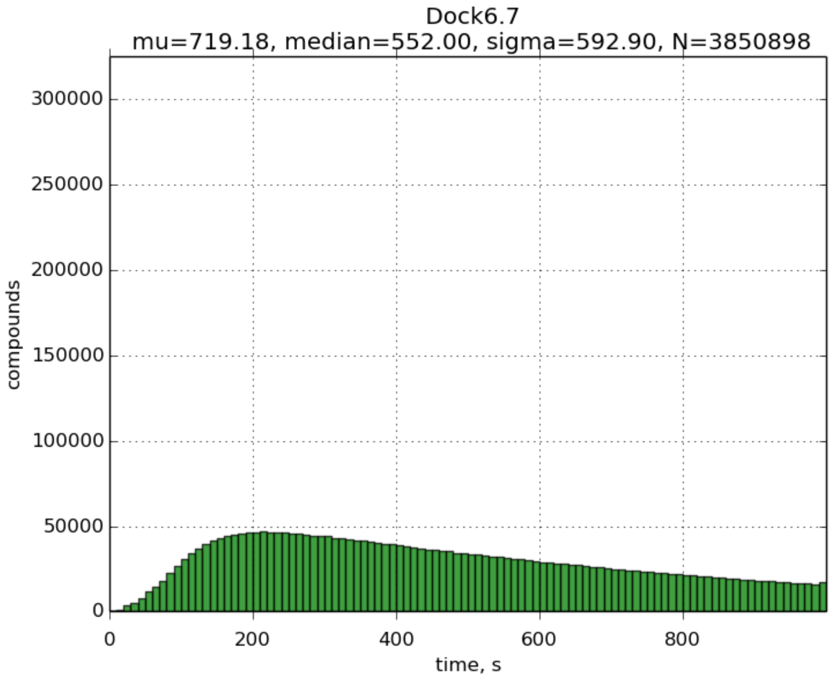
- Compute time for docking depends the search space, search quality, and complexity of the scoring function.
- To dock millions of compounds, we cut corners.
- Docking time varies between programs (~1 minute/compound).

(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



Surflex
mu=87.73, median=77.00, sigma=1222.22, N=3633568



Dock6.7
mu=719.18, median=552.00, sigma=592.90, N=3850898

How do we scale to HTC resources?

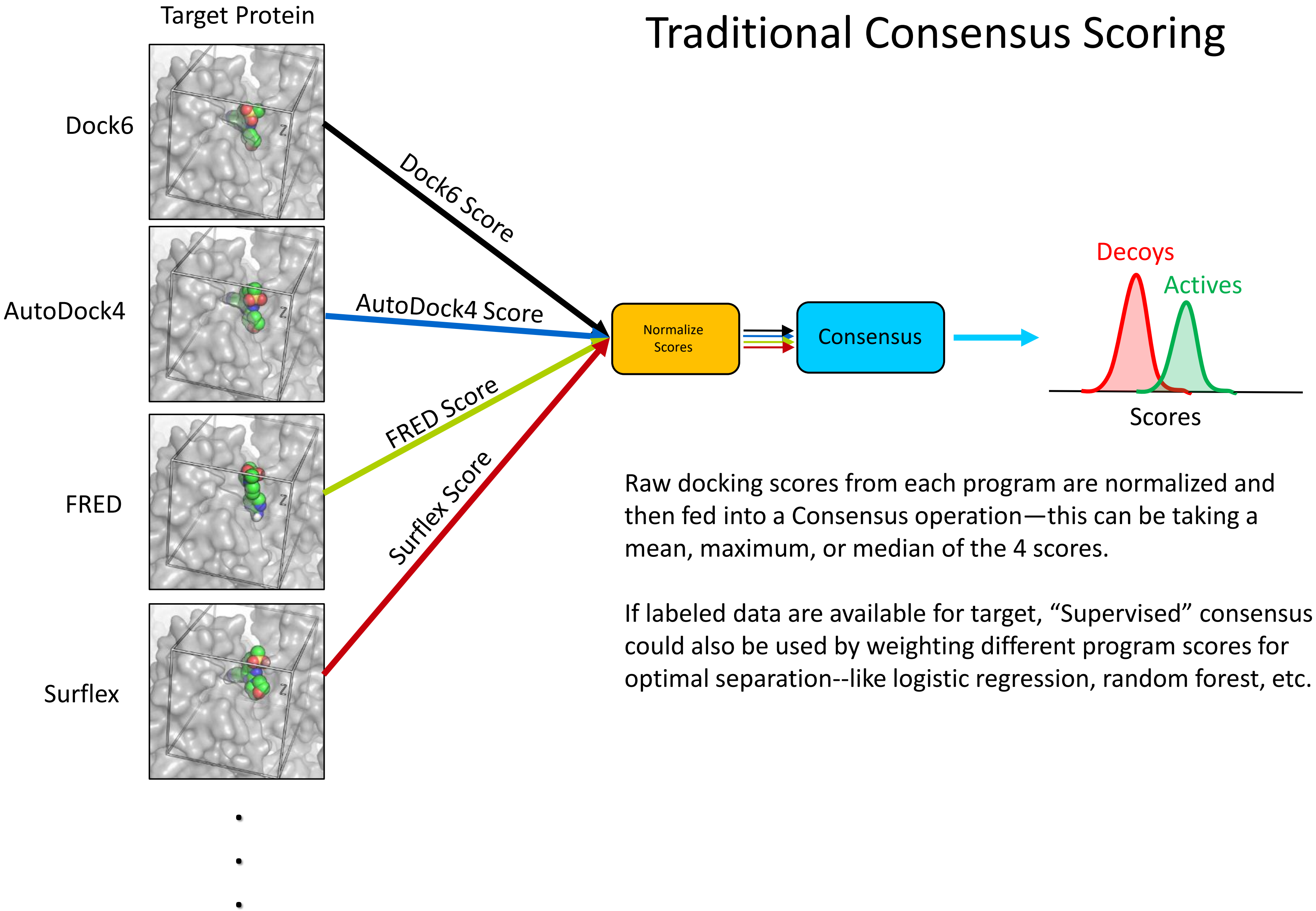
- Each docking run is independent--*pleasantly parallelizable*!
- Typical docking codes don’t benefit from specialized hardware or multiple cores.

- 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 within 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 do we benefit from HTC?

- Very large number of compounds
- Large numbers of targets
- Extensive docking parameter testing
- Benchmarking of different programs
- Hypothetical 100 node cluster = 3.5 million/day
- Local SMSF (3 nodes) = 35,000/day
- 100s of millions to billions of dockings!

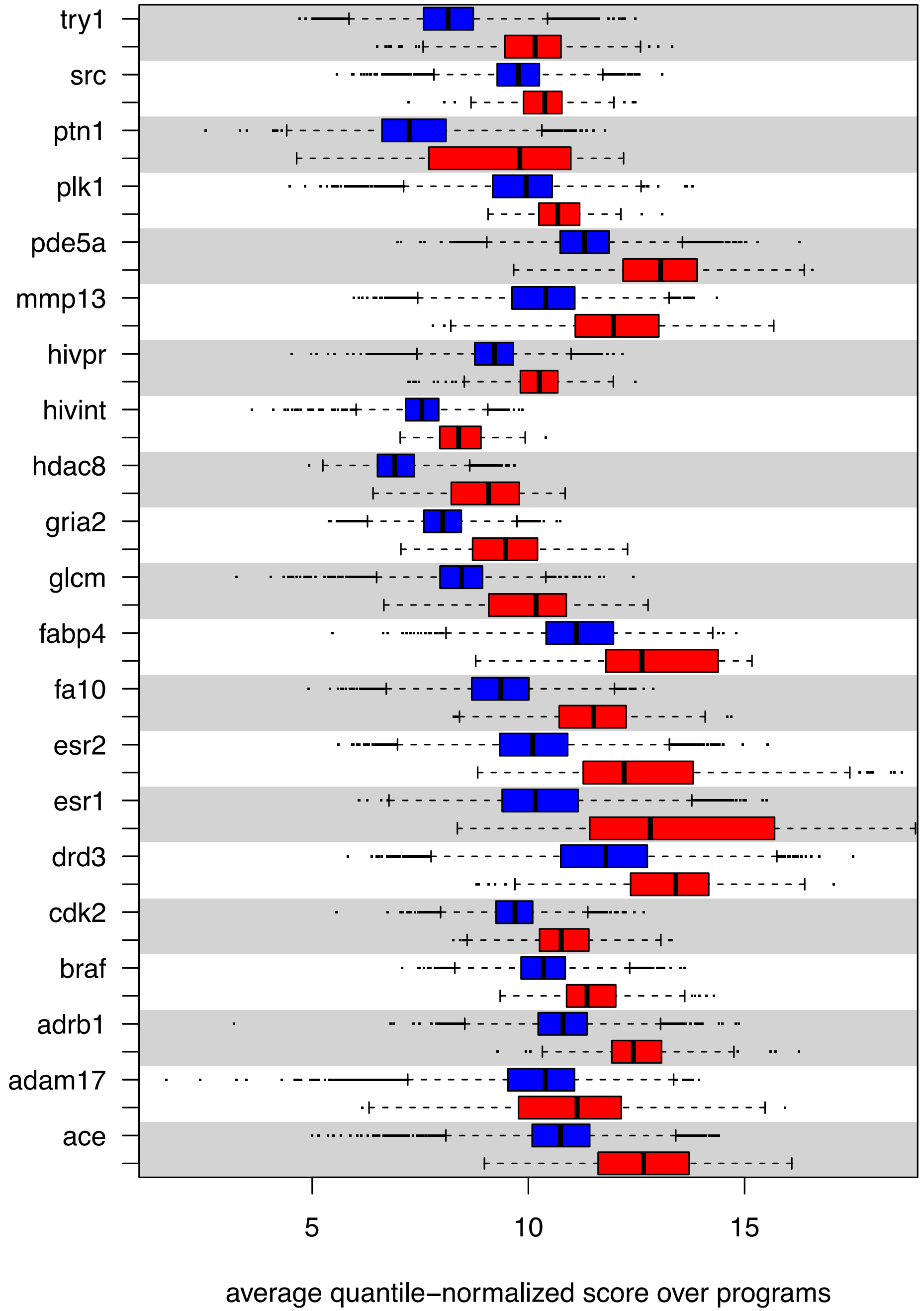
Traditional Consensus Scoring



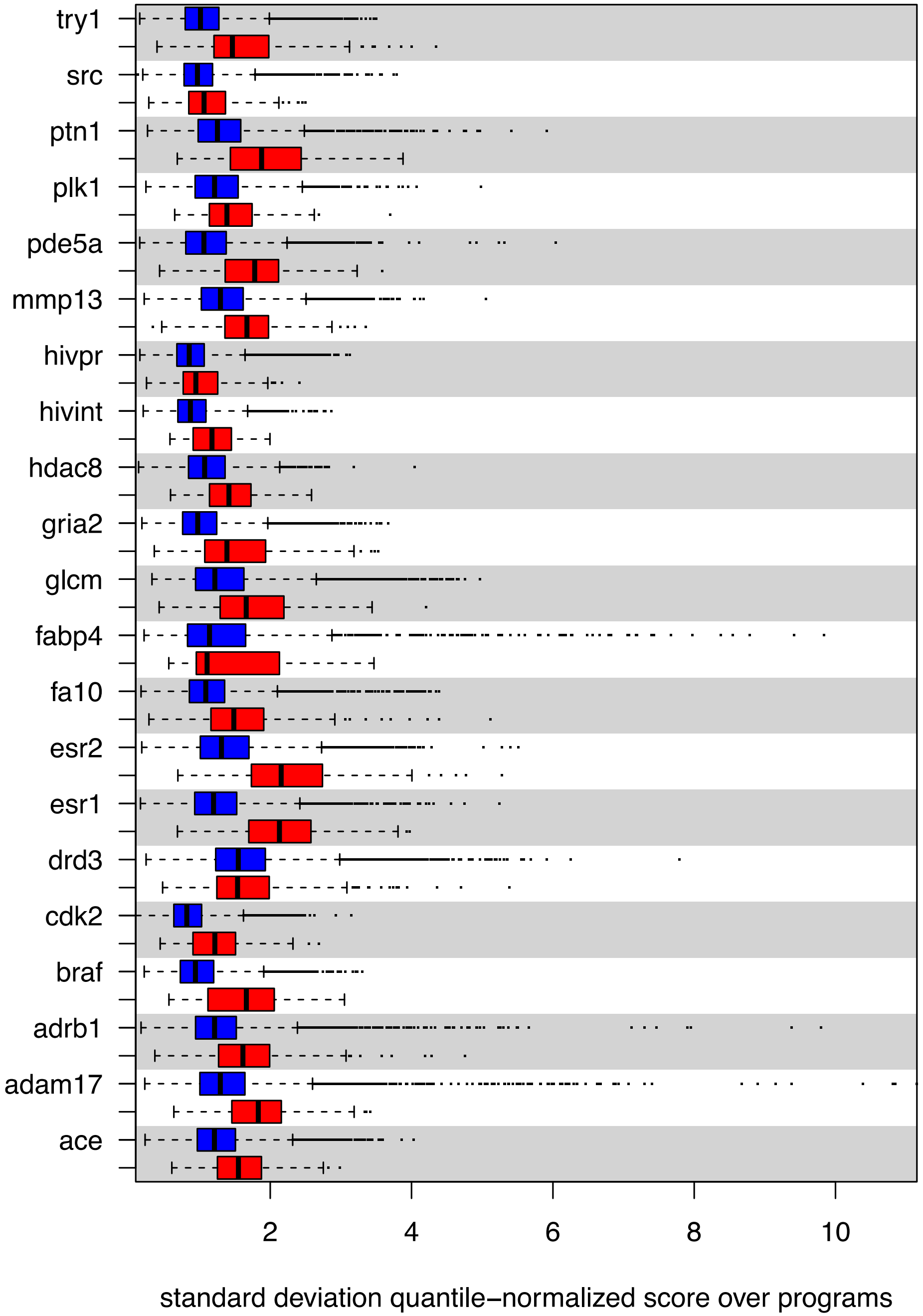
Raw docking scores from each program are normalized and then fed into a Consensus operation—this can be taking a mean, maximum, or median of the 4 scores.

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.

Actives (red) have higher average−over−program scores than decoys (blue)

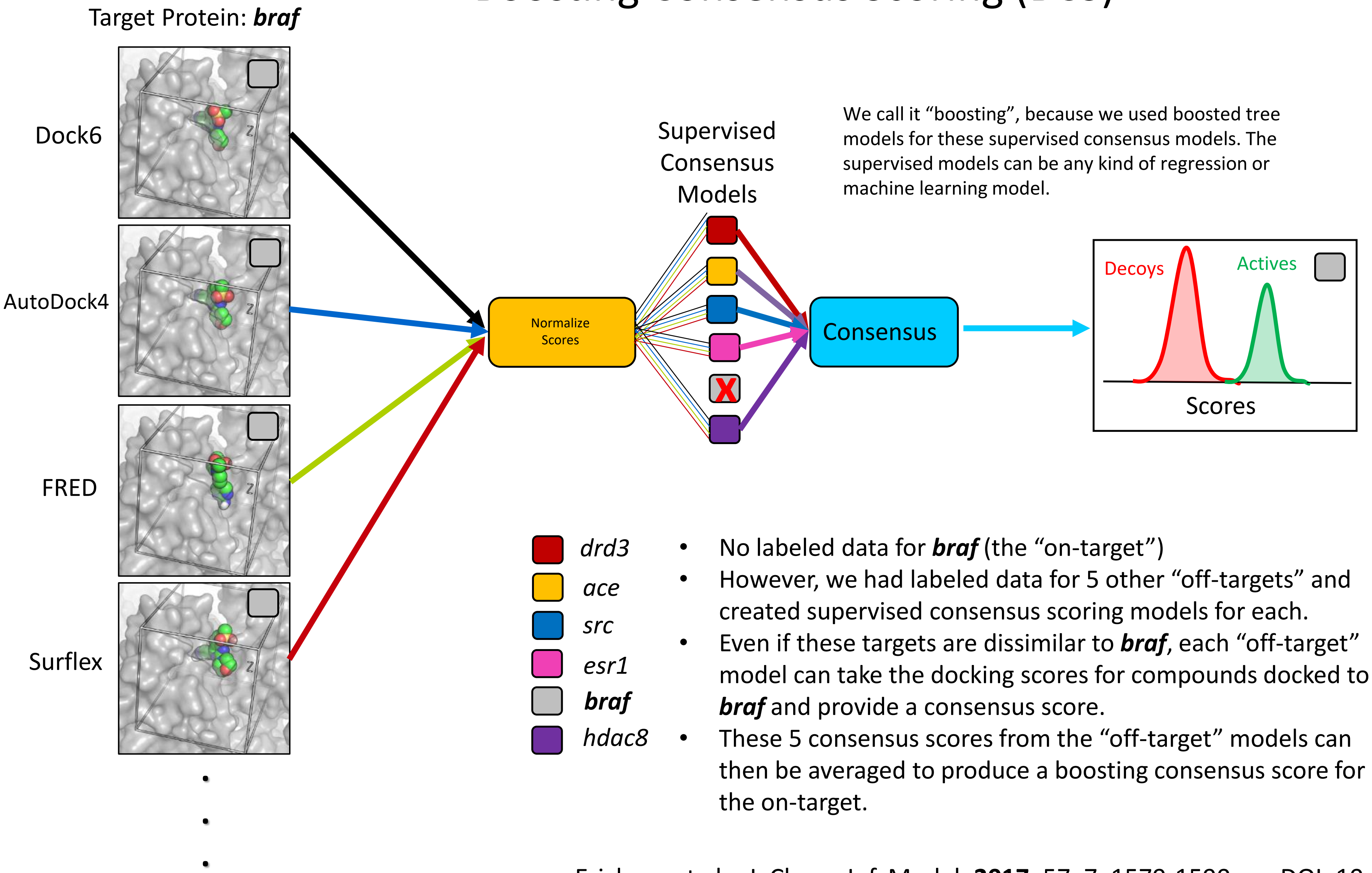


Actives (red) have higher SD−over−program scores than decoys (blue)



- As expected the mean score for actives (red) was higher than for decoys (blue).
- Interestingly, the standard deviation in scores was also higher for actives than for decoys

Boosting Consensus Scoring (BCS)

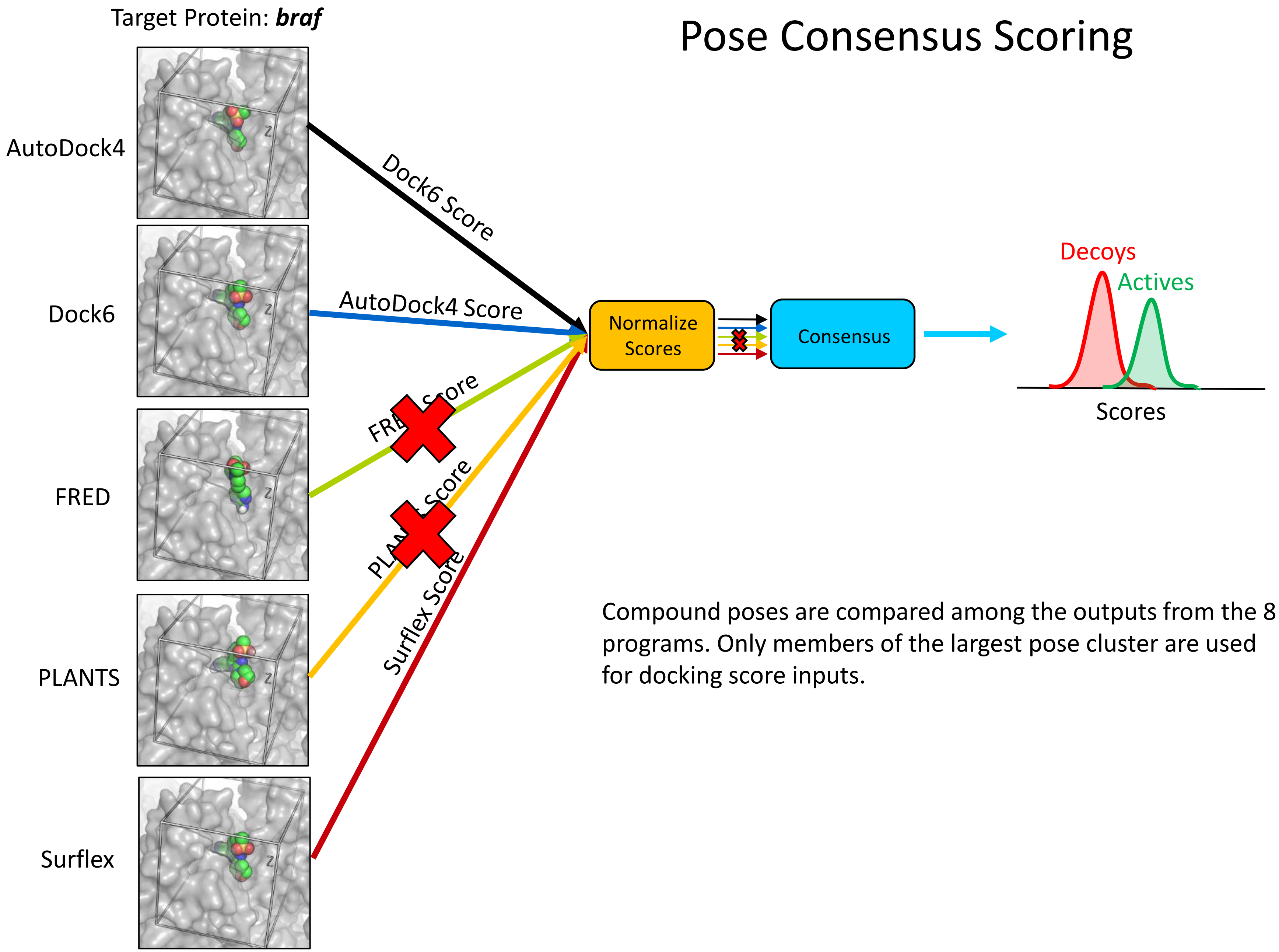


We call it "boosting", because we used boosted tree models for these supervised consensus models. The supervised models can be any kind of regression or machine learning model.

- No labeled data for ***braf*** (the "on-target")
- However, we had labeled data for 5 other "off-targets" and created supervised consensus scoring models for each.
- Even if these targets are dissimilar to ***braf***, each "off-target" model can take the docking scores for compounds docked to ***braf*** and provide a consensus score.
- These 5 consensus scores from the "off-target" models can then be averaged to produce a boosting consensus score for the on-target.

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

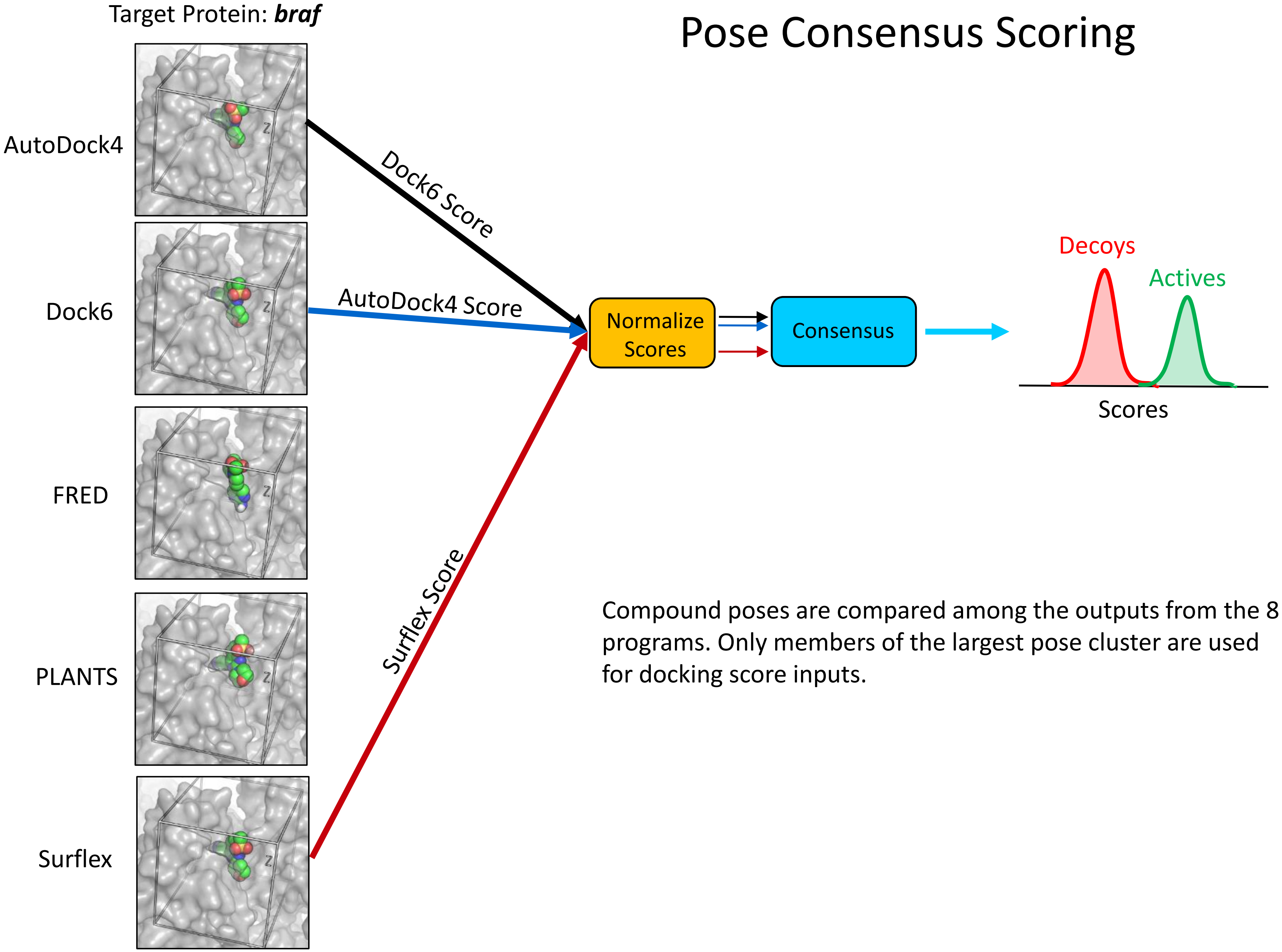
Pose Consensus Scoring



Compound poses are compared among the outputs from the 8 programs. Only members of the largest pose cluster are used for docking score inputs.

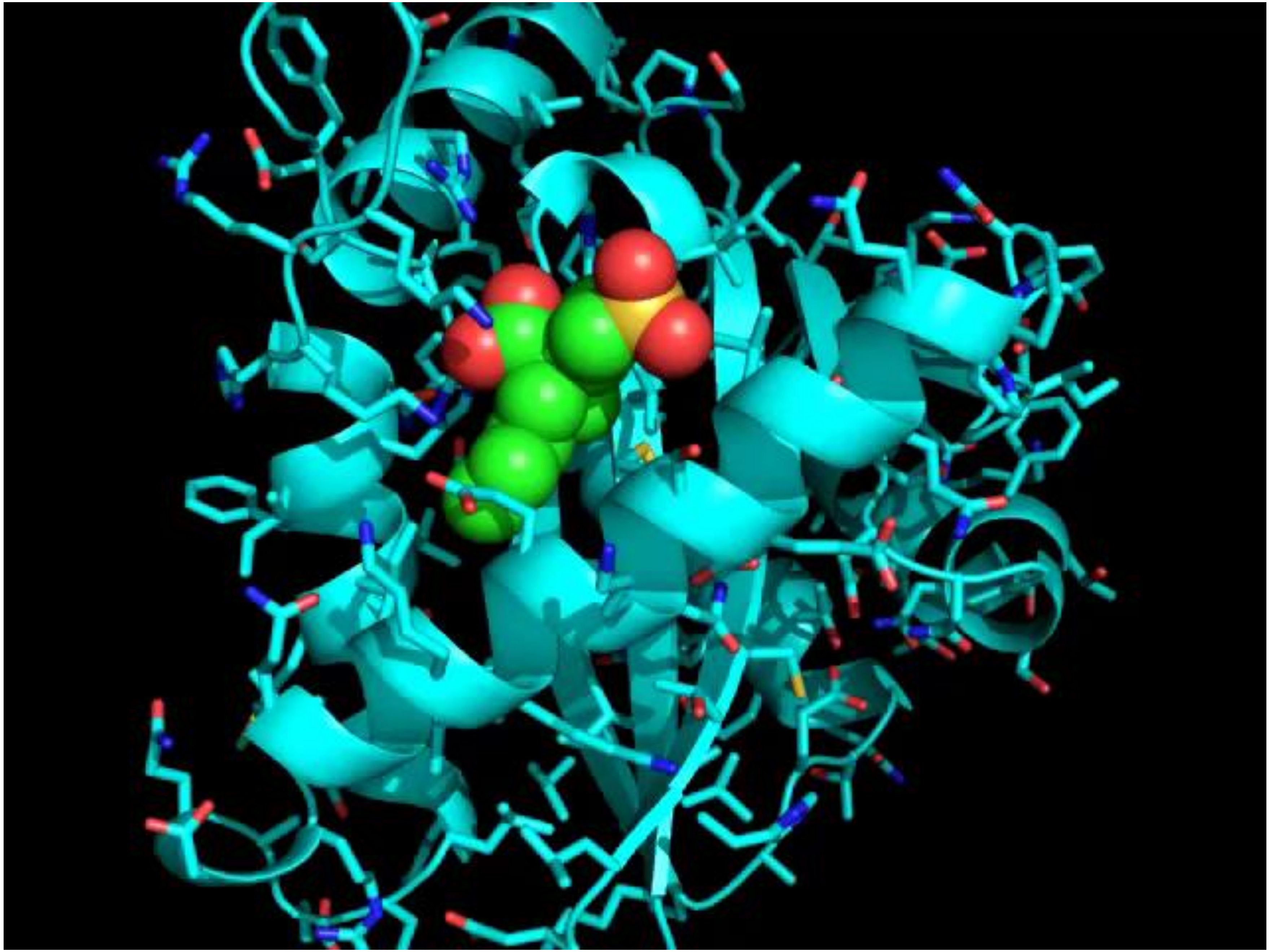
x

Pose Consensus Scoring

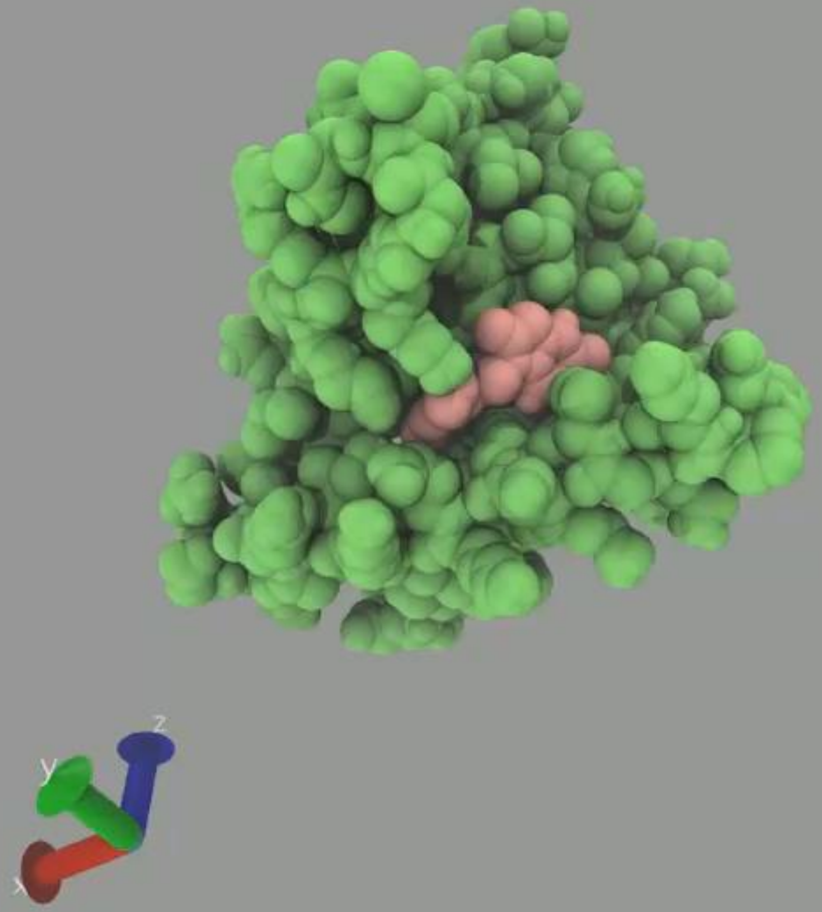


Compound poses are compared among the outputs from the 8 programs. Only members of the largest pose cluster are used for docking score inputs.

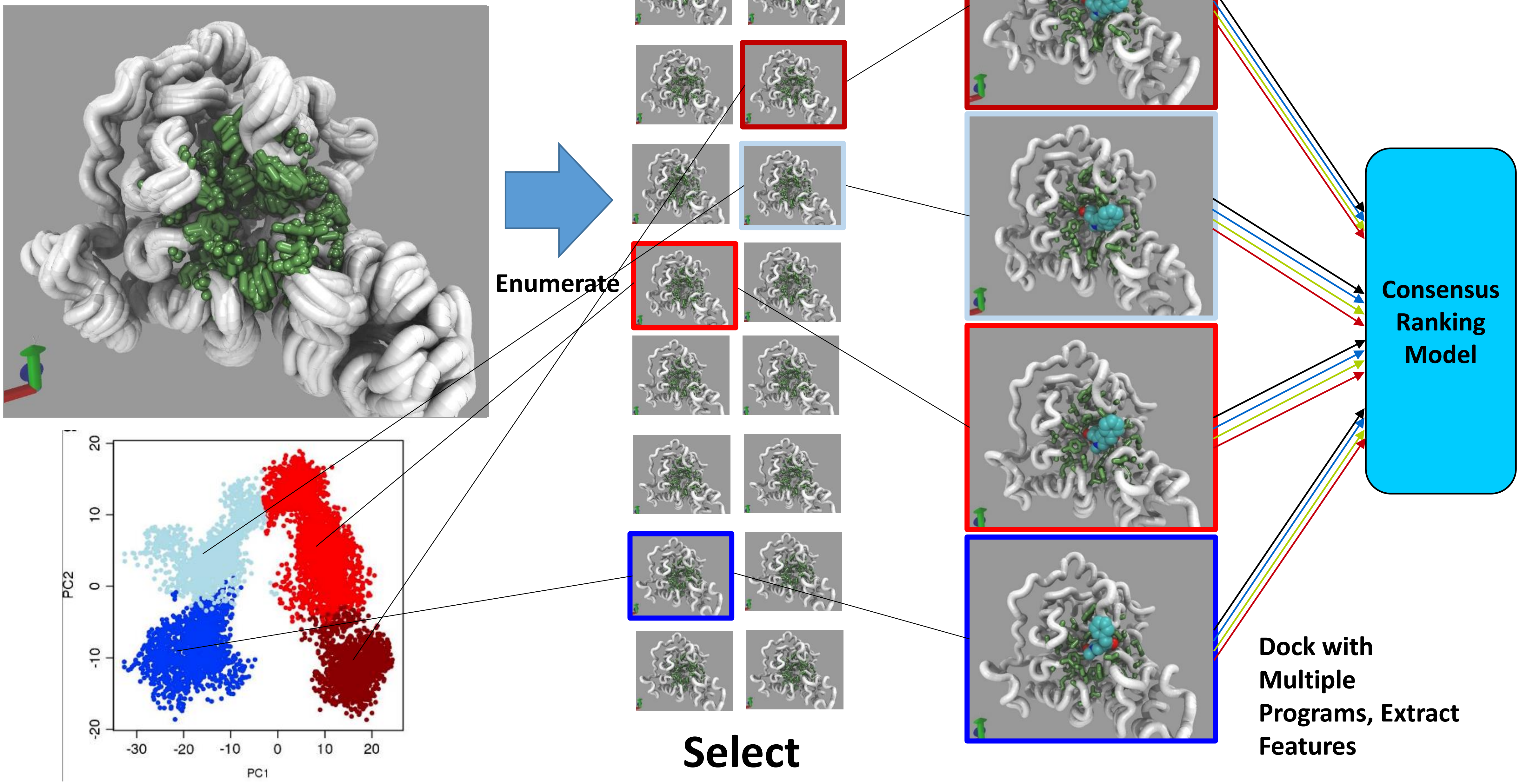
x



In docking, the static approximation of target protein is severe!


z
y
x

Ensemble Docking

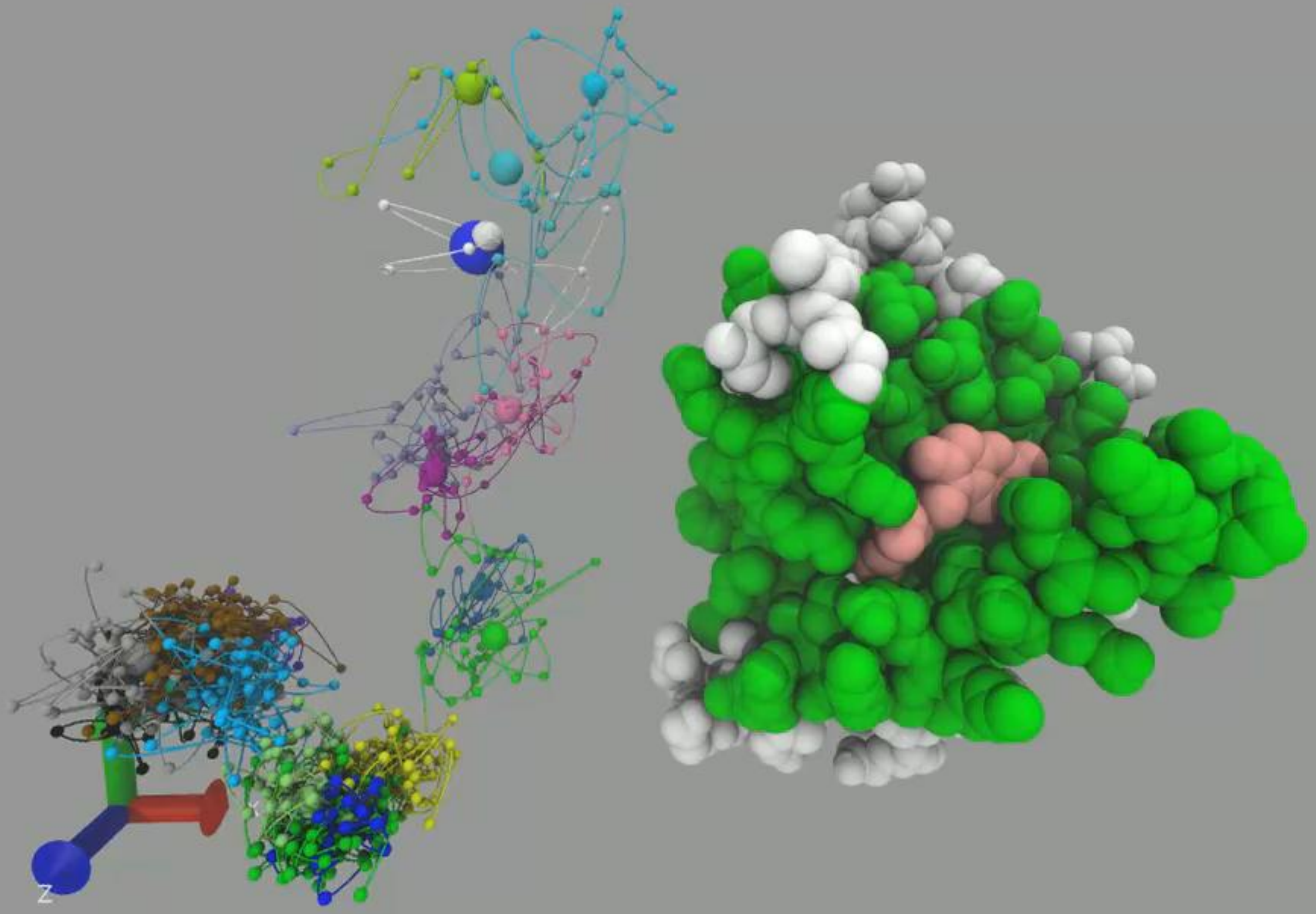


Enumerate

Select

Dock with Multiple Programs, Extract Features

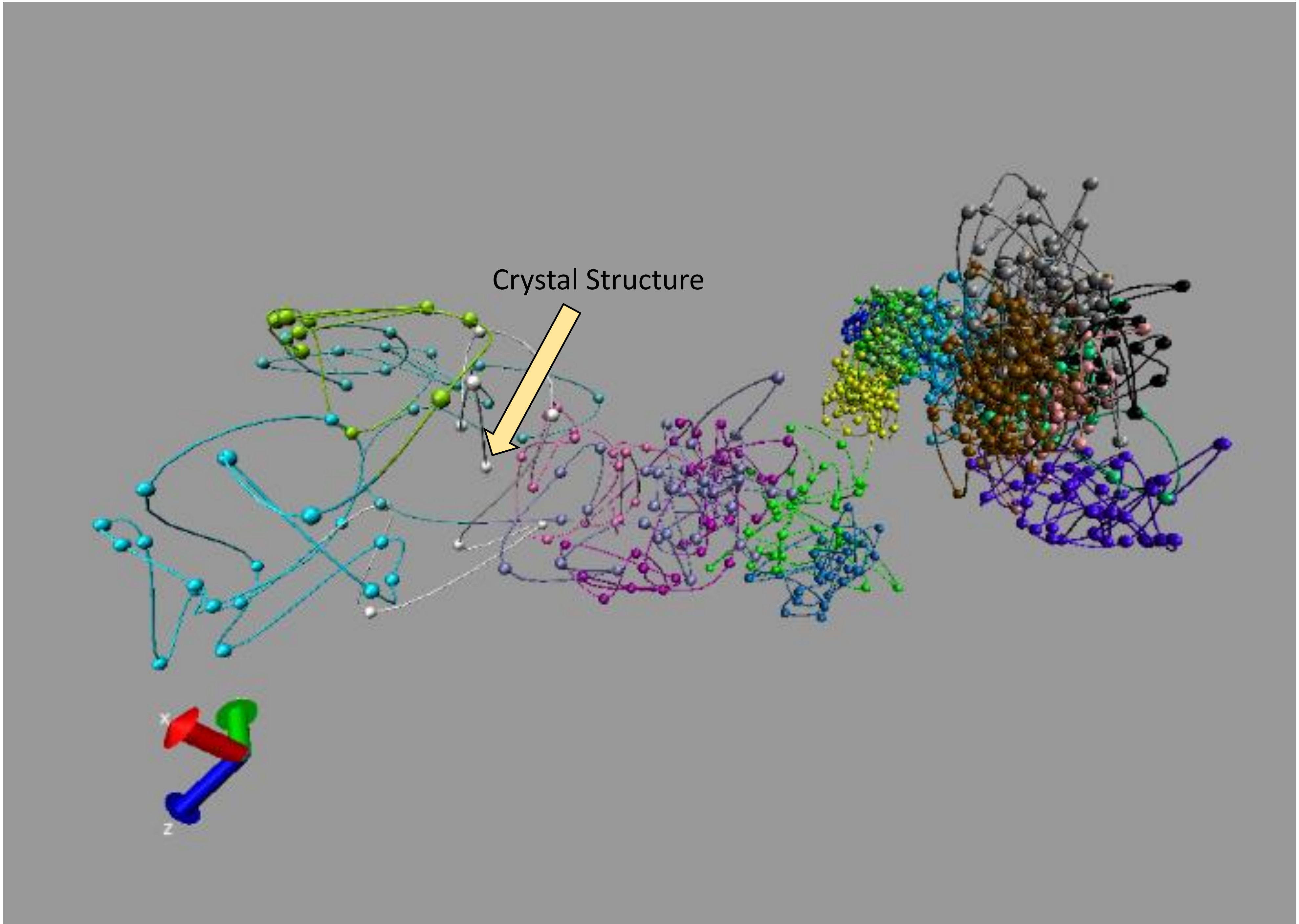
Consensus Ranking Model


Z

Colored string shows progress of MD trajectory of HIV integrase projected onto 3 principal component dimensions based on binding pocket geometry.

The trajectory begins near the crystal structure conformation (labeled). Individual snapshots from trajectory are shown as small spheres.

Protein conformations were clustered based on binding pocket geometry.



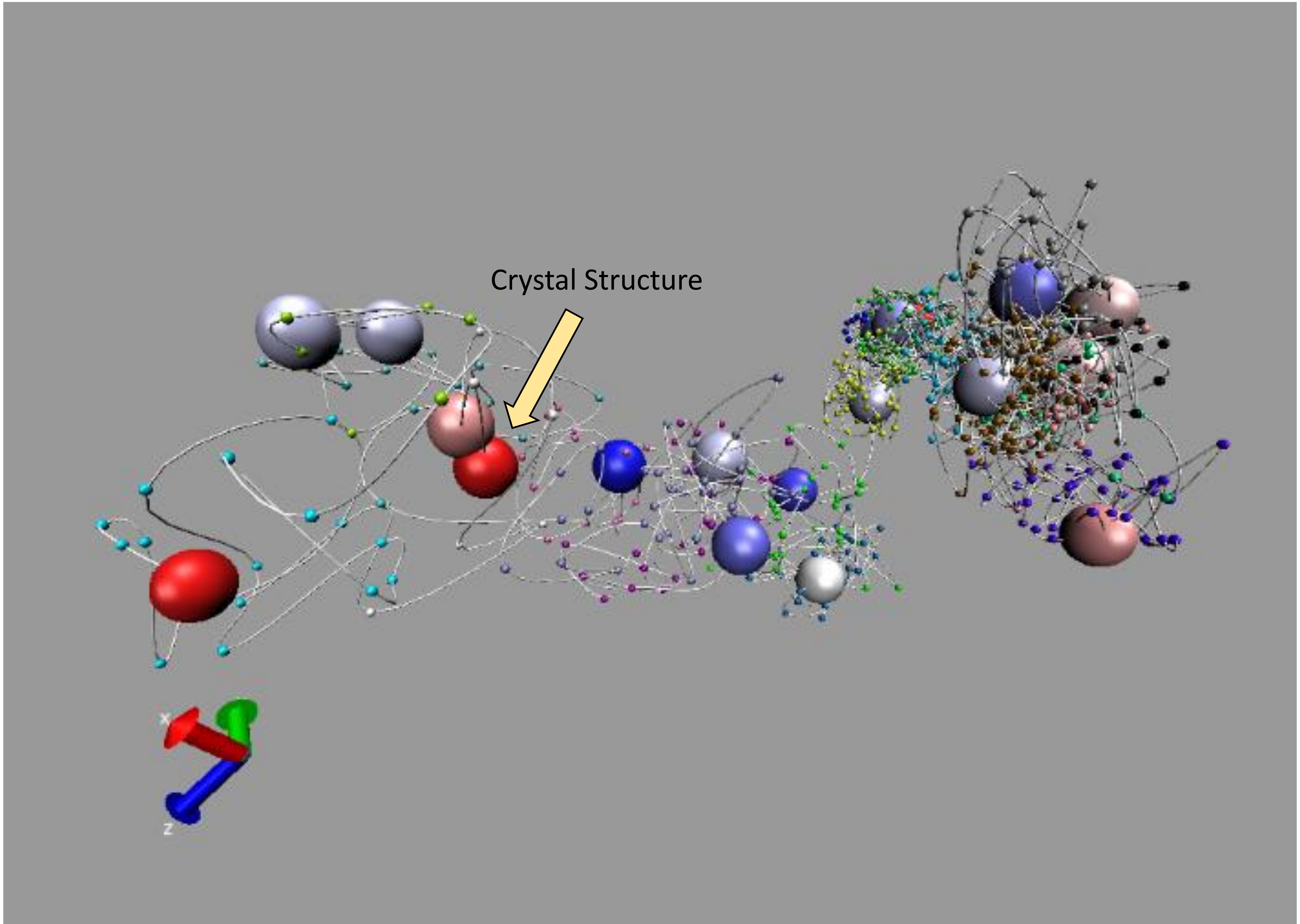
White string shows progress of MD trajectory of HIV integrase projected onto 3 principal component dimensions based on binding pocket geometry.

The trajectory begins near the crystal structure conformation (labeled). Individual snapshots from trajectory are shown as small spheres.

Protein conformations were clustered based on binding pocket geometry. Conformers are colored by their cluster ID (small colored spheres).

The 20 other large spheres indicate conformers selected as cluster reps (most central conformation in each cluster).

The crystal structure and the other 20 representatives were docked using the program smina. These were colored by their virtual screening performance based on the ROCAUC metric (red=0.74 to blue=0.860).

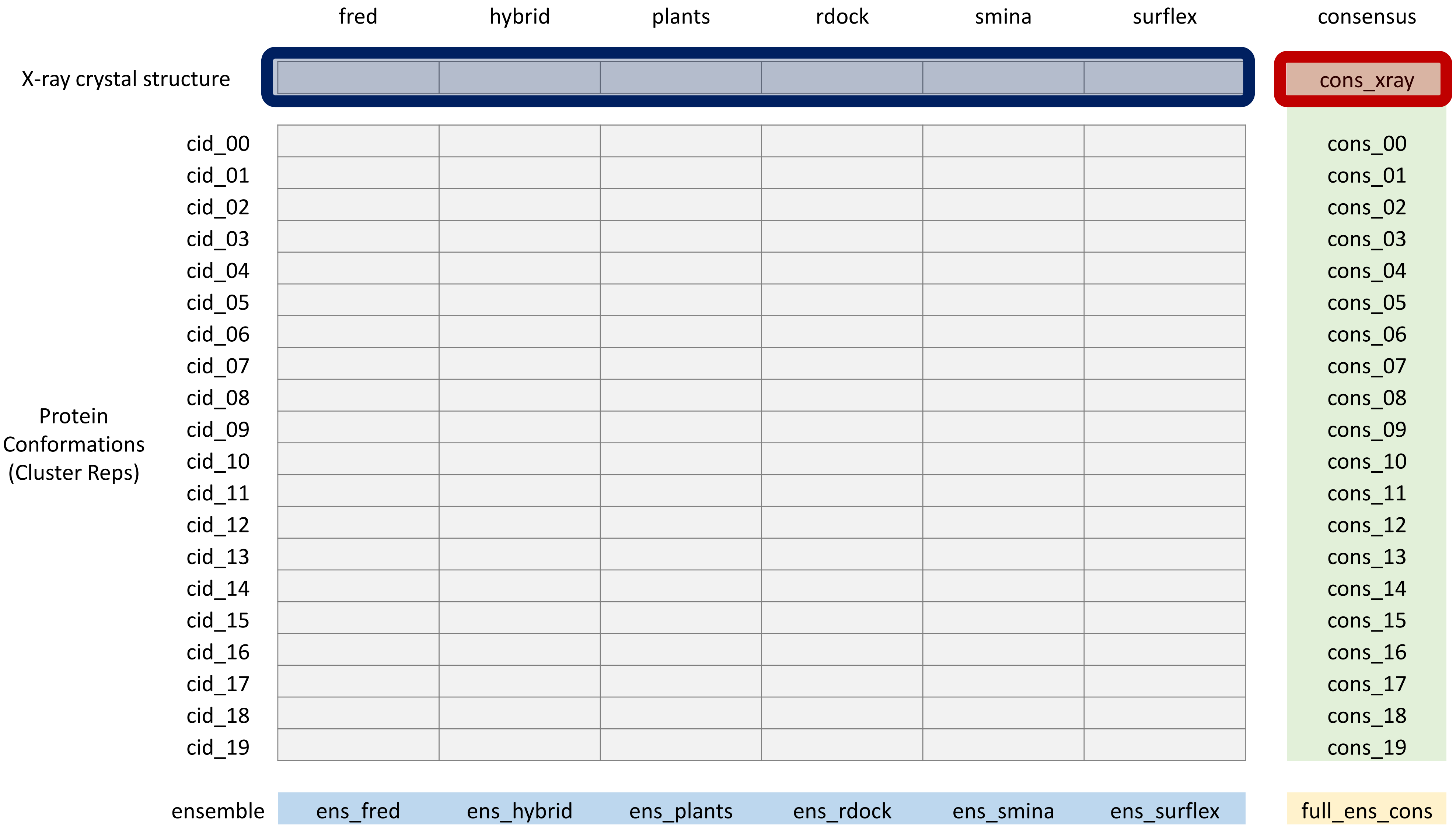


Single Program

Docking Score Matrix for a Single Compound

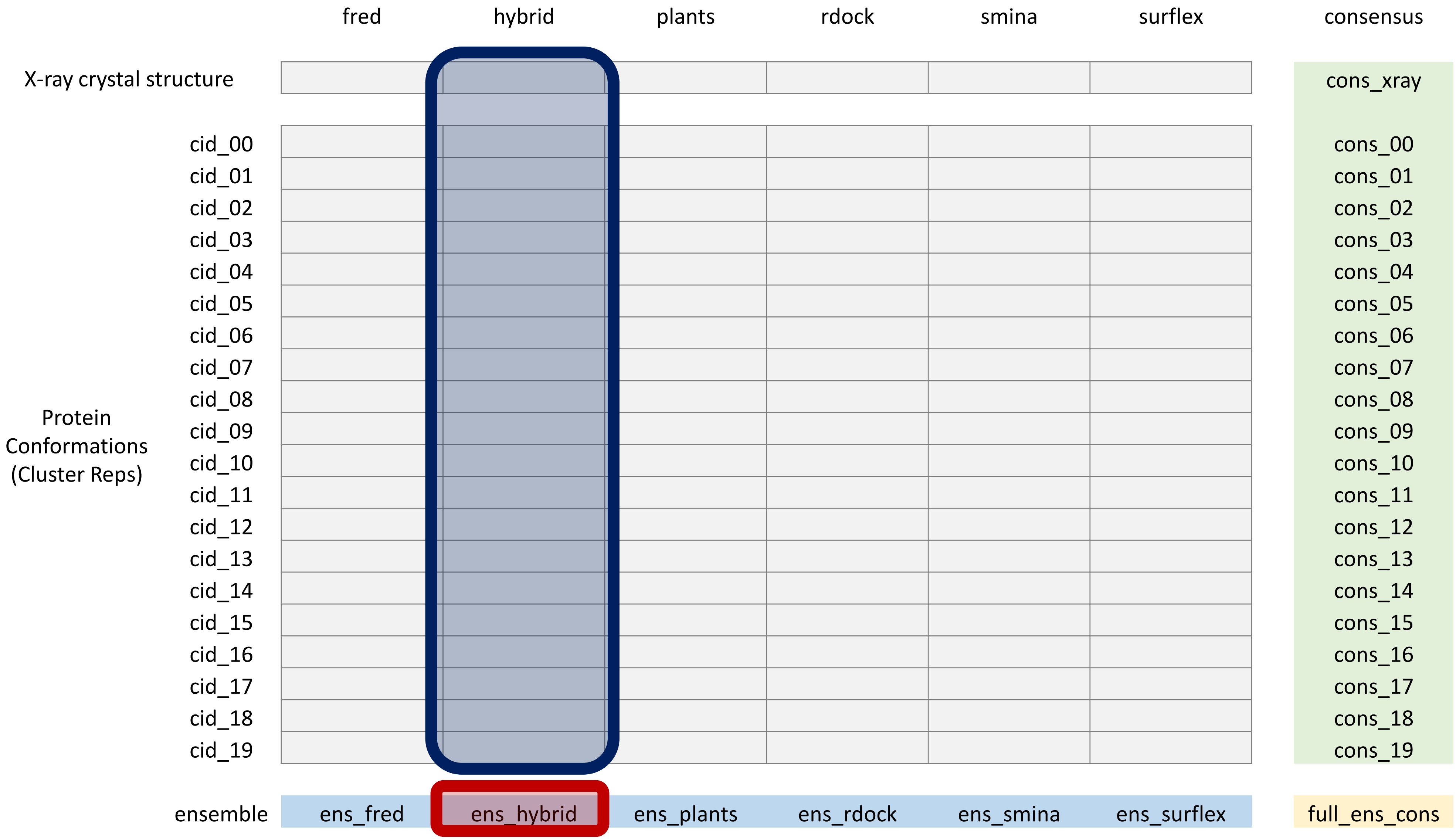
		fred	hybrid	plants	rdock	smina	surflex	consensus
X-ray crystal structure			▨					cons_xray
Protein Conformations (Cluster Reps)	cid_00							cons_00
	cid_01							cons_01
	cid_02							cons_02
	cid_03							cons_03
	cid_04							cons_04
	cid_05							cons_05
	cid_06							cons_06
	cid_07							cons_07
	cid_08							cons_08
	cid_09							cons_09
	cid_10							cons_10
	cid_11							cons_11
	cid_12							cons_12
	cid_13							cons_13
	cid_14							cons_14
	cid_15							cons_15
	cid_16							cons_16
	cid_17							cons_17
	cid_18							cons_18
	cid_19							cons_19
	ensemble	ens_fred	ens_hybrid	ens_plants	ens_rdock	ens_smina	ens_surflex	full_ens_cons

Consensus Scoring



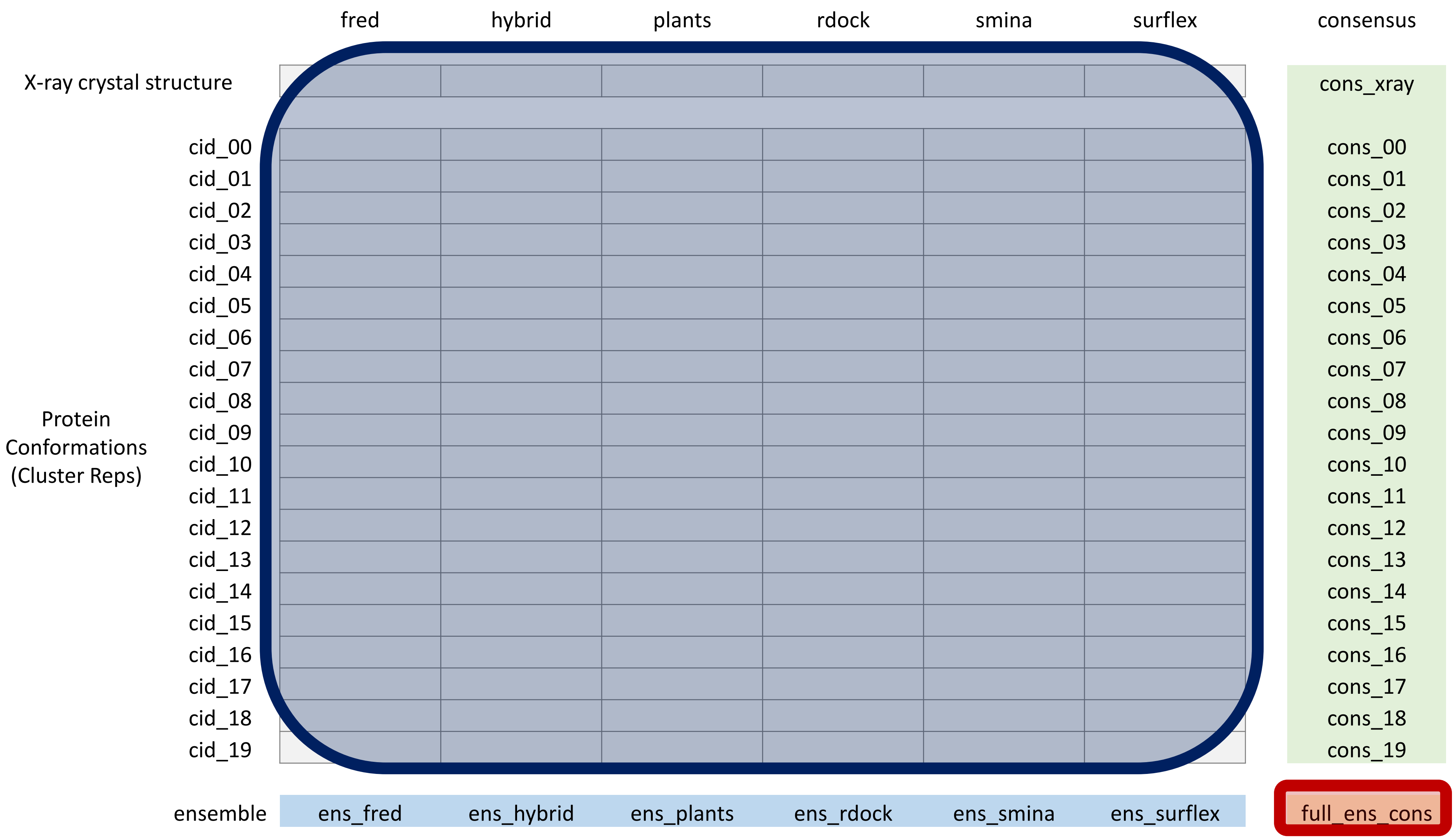
		fred	hybrid	plants	rdock	smina	surflex	consensus
X-ray crystal structure								cons_xray
Protein Conformations (Cluster Reps)	cid_00							cons_00
	cid_01							cons_01
	cid_02							cons_02
	cid_03							cons_03
	cid_04							cons_04
	cid_05							cons_05
	cid_06							cons_06
	cid_07							cons_07
	cid_08							cons_08
	cid_09							cons_09
	cid_10							cons_10
	cid_11							cons_11
	cid_12							cons_12
	cid_13							cons_13
	cid_14							cons_14
	cid_15							cons_15
	cid_16							cons_16
	cid_17							cons_17
	cid_18							cons_18
	cid_19							cons_19
	ensemble	ens_fred	ens_hybrid	ens_plants	ens_rdock	ens_smina	ens_surflex	full_ens_cons

Ensemble Scoring



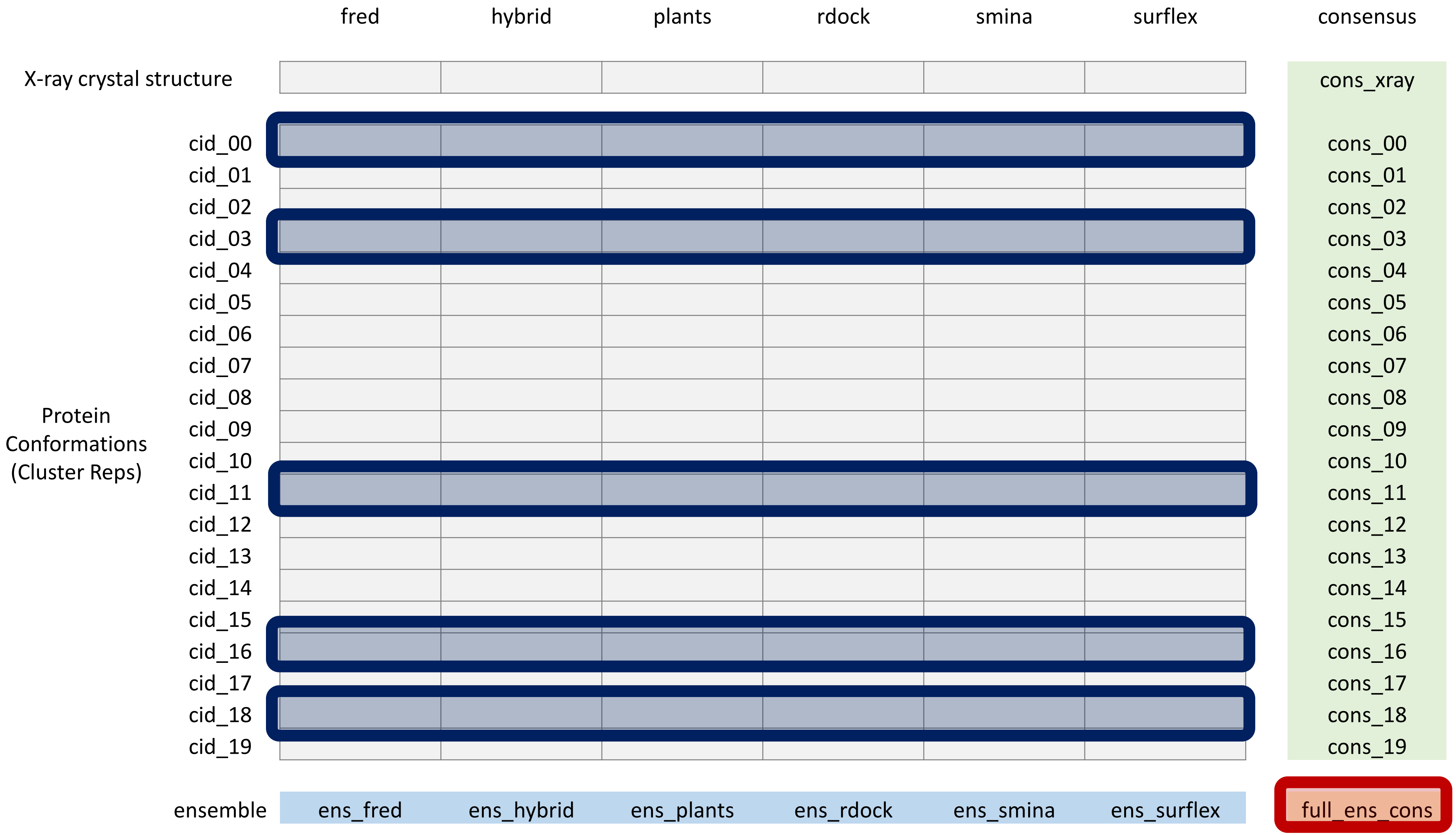
		fred	hybrid	plants	rdock	smina	surflex	consensus
	X-ray crystal structure							cons_xray
Protein Conformations (Cluster Reps)	cid_00							cons_00
	cid_01							cons_01
	cid_02							cons_02
	cid_03							cons_03
	cid_04							cons_04
	cid_05							cons_05
	cid_06							cons_06
	cid_07							cons_07
	cid_08							cons_08
	cid_09							cons_09
	cid_10							cons_10
	cid_11							cons_11
	cid_12							cons_12
	cid_13							cons_13
	cid_14							cons_14
	cid_15							cons_15
	cid_16							cons_16
	cid_17							cons_17
	cid_18							cons_18
	cid_19							cons_19
	ensemble	ens_fred	ens_hybrid	ens_plants	ens_rdock	ens_smina	ens_surflex	full_ens_cons

Consensus + Ensemble Scoring



		fred	hybrid	plants	rdock	smina	surflex	consensus
X-ray crystal structure								cons_xray
Protein Conformations (Cluster Reps)	cid_00							cons_00
	cid_01							cons_01
	cid_02							cons_02
	cid_03							cons_03
	cid_04							cons_04
	cid_05							cons_05
	cid_06							cons_06
	cid_07							cons_07
	cid_08							cons_08
	cid_09							cons_09
	cid_10							cons_10
	cid_11							cons_11
	cid_12							cons_12
	cid_13							cons_13
	cid_14							cons_14
	cid_15							cons_15
	cid_16							cons_16
	cid_17							cons_17
	cid_18							cons_18
	cid_19							cons_19
	ensemble	ens_fred	ens_hybrid	ens_plants	ens_rdock	ens_smina	ens_surflex	full_ens_cons

“Smart” Consensus + Ensemble Scoring



Protein Conformations (Cluster Reps)

	fred	hybrid	plants	rdock	smina	surflex	consensus
X-ray crystal structure							cons_xray
cid_00							cons_00
cid_01							cons_01
cid_02							cons_02
cid_03							cons_03
cid_04							cons_04
cid_05							cons_05
cid_06							cons_06
cid_07							cons_07
cid_08							cons_08
cid_09							cons_09
cid_10							cons_10
cid_11							cons_11
cid_12							cons_12
cid_13							cons_13
cid_14							cons_14
cid_15							cons_15
cid_16							cons_16
cid_17							cons_17
cid_18							cons_18
cid_19							cons_19
ensemble	ens_fred	ens_hybrid	ens_plants	ens_rdock	ens_smina	ens_surflex	full_ens_cons

cdk2

	fred	hybrid	plants	rdock	smina	surflex	Cons Scoring	frame	cpop	RMSD
xtal structure	0.53	0.75	0.66	0.79	0.64	0.68	0.81			
cluster reps	0.79	0.79	0.69	0.79	0.75	0.64	0.82	3	5	1.8
	0.73	0.76	0.65	0.79	0.74	0.67	0.81	20	26	1.9
	0.63	0.70	0.58	0.76	0.65	0.60	0.72	65	44	2.8
	0.66	0.77	0.62	0.78	0.68	0.67	0.76	88	29	2.9
	0.65	0.74	0.57	0.73	0.66	0.61	0.73	320	93	2.9
	0.70	0.76	0.63	0.76	0.69	0.64	0.76	422	20	2.9
	0.67	0.77	0.60	0.76	0.64	0.69	0.75	150	53	3.0
	0.70	0.74	0.62	0.76	0.70	0.67	0.76	115	28	3.0
	0.58	0.64	0.47	0.62	0.60	0.60	0.63	482	65	3.0
	0.56	0.63	0.49	0.68	0.60	0.62	0.65	536	115	3.1
	0.58	0.68	0.55	0.70	0.64	0.62	0.68	223	50	3.1
	0.63	0.73	0.59	0.73	0.65	0.63	0.71	342	44	3.2
	0.66	0.74	0.59	0.72	0.66	0.62	0.71	857	46	3.3
	0.68	0.75	0.66	0.77	0.73	0.69	0.79	624	41	3.3
	0.62	0.69	0.55	0.70	0.66	0.61	0.70	942	78	3.3
	0.67	0.76	0.61	0.74	0.67	0.61	0.73	672	55	3.3
	0.62	0.72	0.54	0.67	0.62	0.61	0.69	834	78	3.4
	0.58	0.69	0.58	0.75	0.69	0.69	0.74	722	68	3.4
	0.61	0.70	0.54	0.69	0.65	0.67	0.70	380	41	3.5
	0.61	0.71	0.53	0.73	0.61	0.65	0.70	917	22	3.6
Ensemble Scoring	0.68	0.77	0.59	0.78	0.70	0.67	0.75			

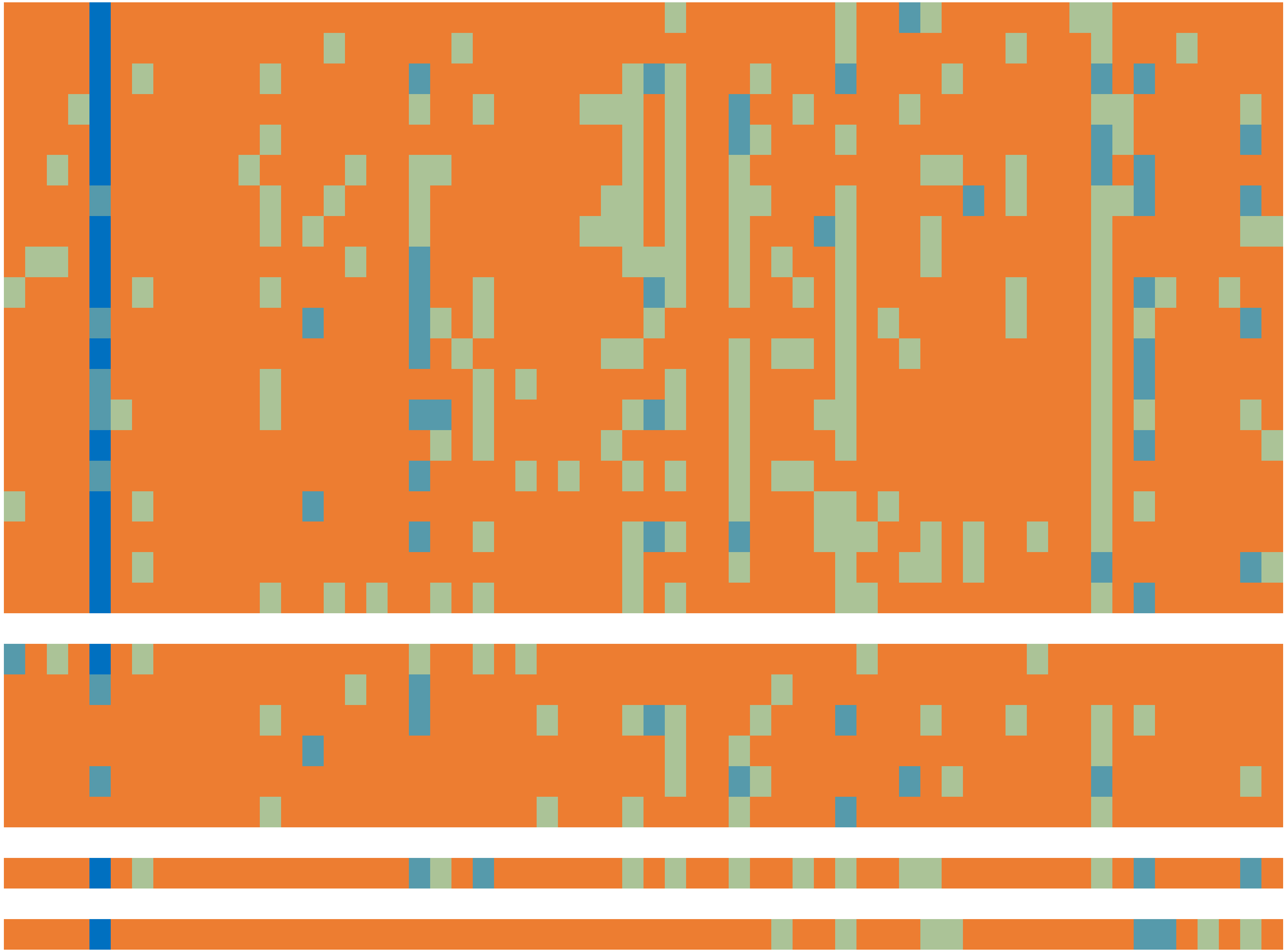
hivpr

	fred	hybrid	plants	rdock	smina	surflex	Cons Scoring	frame	cpop	RMSD
xtal structure	0.67	0.68	0.77	0.66	0.77	0.64	0.82			
cluster reps	0.75	0.70	0.77	0.67	0.78	0.68	0.83	2	9	1.6
	0.57	0.57	0.70	0.63	0.75	0.67	0.76	15	18	2.1
	0.69	0.66	0.78	0.66	0.79	0.62	0.83	153	28	2.4
	0.66	0.64	0.76	0.73	0.82	0.62	0.85	102	30	2.6
	0.66	0.66	0.78	0.72	0.84	0.66	0.89	74	38	2.6
	0.68	0.67	0.75	0.70	0.84	0.61	0.84	126	16	2.7
	0.65	0.62	0.75	0.74	0.80	0.62	0.83	211	87	2.9
	0.61	0.62	0.75	0.71	0.84	0.60	0.82	182	38	2.9
	0.57	0.67	0.73	0.66	0.81	0.62	0.80	31	9	2.9
	0.62	0.59	0.74	0.77	0.83	0.66	0.81	39	15	3.0
	0.59	0.56	0.75	0.66	0.83	0.67	0.82	328	43	3.4
	0.61	0.53	0.71	0.68	0.76	0.63	0.77	394	118	3.6
	0.61	0.60	0.72	0.69	0.79	0.58	0.77	409	71	3.6
	0.61	0.63	0.76	0.70	0.79	0.69	0.84	962	46	3.9
	0.56	0.61	0.78	0.74	0.80	0.58	0.82	736	154	3.9
	0.62	0.67	0.73	0.68	0.82	0.59	0.80	552	75	4.0
	0.62	0.65	0.74	0.61	0.80	0.57	0.78	837	35	4.0
	0.60	0.66	0.75	0.70	0.81	0.67	0.83	906	41	4.1
	0.56	0.62	0.73	0.61	0.80	0.56	0.75	803	39	4.1
	0.54	0.57	0.76	0.70	0.83	0.58	0.79	647	91	4.2
Ensemble Scoring	0.64	0.65	0.76	0.73	0.85	0.65	0.87			

Target: HIV integrase

protein conf	scaffolds retrieved	total
cons_mean_14	7	60
cons_mean_00	7	60
cons_mean_15	12	60
cons_mean_11	14	60
cons_mean_17	10	60
cons_mean_12	14	60
cons_mean_05	16	60
cons_mean_01	15	60
cons_mean_04	13	60
cons_mean_09	16	60
cons_mean_10	12	60
cons_mean_18	12	60
cons_mean_16	9	60
cons_mean_02	15	60
cons_mean_19	9	60
cons_mean_06	10	60
cons_mean_03	10	60
cons_mean_07	14	60
cons_mean_13	11	60
cons_mean_08	12	60
ens_mean_fred	9	60
ens_mean_hybrid	4	60
ens_mean_plants	12	60
ens_mean_rdock	4	60
ens_mean_smina	8	60
ens_mean_surflex	6	60
full_mean	15	60
cons_mean_xtal	9	60

Active Bemis-Murcko Scaffolds



Conclusions

HTC is a fabulous resource for massive structure-based vHTS

HTC enables rapid cycles of development, testing, validation of docking-based VS

HTC will enable more sophisticated MD-based approaches to SBVS

Thank You!

- UWCCC-Drug Development Core
- Mike Hoffmann (PI)
- Scott Wildman & Ken Satyshur
- Michael Newton & Tony Gitter
- Open Science Grid & CHTC
- Facilitators: Lauren Michael & Christina Koch

CENTER FOR HIGH THROUGHPUT COMPUTING

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

OpenEye SCIENTIFIC

WARF
Wisconsin Alumni Research Foundation

Open Science Grid

NSF

Extra Slides on Ensemble Docking

Ensemble Docking general procedure

- Choose some reference structure of target protein with bound compound from theory or experiment.
- ***Enumeration***: perform MD simulation to examine possible conformational states of the bound-state of target protein
- ***Selection***: select subset of snapshots from simulation to serve as representative target conformers.
- ***Docking***: Dock large library of decoy/active compounds to each target conformer. Apply consensus? How many programs?
- ***Scoring***: *Models, Features, Predictions*: Besides docking scores, what other types of features may be derived from docking outputs.

Ensemble Docking technical procedure

- Reference structure based on crystal structure used in DUD-E data set.
  - Standard protein preparation: add missing atoms, strip water molecules, detergents, non-essential ions, etc.
  - structure is energy minimized—this is the reference structure.
- ***Enumeration***:
  - NPT MD simulation of holoform for100 ns at 1 atm, 300K, explicit water, neutralizing Na^+ or Cl^- ions. PBC
  - Frames dumped every 0.100 ns (1000 total frames), energy minimized.
- ***Selection***:
  - Target protein conformers are aligned on C_a reference atoms.
  - Conformers clustered using pocket atom coordinates as features.
  - Pocket atoms defined as heavy atoms from residues within 10Å of original bound compound position. HAC, n=20 clusters, Ward's linkage, select most central representatives.

Ensemble Docking technical procedure

- ***Docking***:
  - Dock DUD-E decoy/active compounds to each target conformer
  - Used 6 of our best programs with default docking prototcols.
  - Tried both static and dynamic search space: use initial ligand position and dynamic ligand position.
- ***Scoring***
  - Each program produces 20 scores for each compound—one for each target conformation.
  - Scores were normalized (z-scores) for each conformer/program
  - 120 z-scores for each compound (6 progs *20 target conformers)
  - Take mean, median, max of these scores for final compound ranking.
  - Apply standard ROCAUC and EF1 metrics

single program: “smina”

	Protein Conformer	cdk2 ROCAUC	fa10 ROCAUC	glcm ROCAUC	hivint ROCAUC	hivpr ROCAUC	Protein Conformer	All 5 targets ROCAUC
xtal Rep	cid_gro1	0.724	0.762	0.493	0.769	0.740	cid_gro1	0.698
Cluster Reps	cid_00	0.654	0.747	0.529	0.749	0.745	cid_00	
	cid_01	0.598	0.754	0.535	0.841	0.764	cid_01	
	cid_02	0.608	0.711	0.512	0.836	0.761	cid_02	
	cid_03	0.608	0.751	0.523	0.783	0.721	cid_03	
	cid_04	0.664	0.619	0.558	0.811	0.762	cid_04	
	cid_05	0.729	0.756	0.600	0.810	0.750	cid_05	
	cid_06	0.640	0.704	0.559	0.746	0.742	cid_06	
	cid_07	0.653	0.628	0.558	0.826	0.728	cid_07	
	cid_08	0.733	0.707	0.516	0.787	0.714	cid_08	
	cid_09	0.656	0.786	0.482	0.812	0.738	cid_09	
	cid_10	0.702	0.756	0.586	0.810	0.730	cid_10	
	cid_11	0.631	0.654	0.579	0.800	0.726	cid_11	
	cid_12	0.624	0.662	0.537	0.830	0.745	cid_12	
	cid_13	0.652	0.583	0.519	0.774	0.708	cid_13	
	cid_14	0.655	0.712	0.563	0.775	0.732	cid_14	
	cid_15	0.682	0.676	0.616	0.812	0.701	cid_15	
	cid_16	0.642	0.762	0.570	0.789	0.740	cid_16	
	cid_17	0.752	0.709	0.594	0.862	0.748	cid_17	
	cid_18	0.688	0.768	0.593	0.823	0.732	cid_18	
	cid_19	0.676	0.687	0.573	0.832	0.733	cid_19	
		0.662	0.707	0.555	0.805	0.736	cid_00-19	0.693
Consensus	mean_norm	0.699	0.758	0.589	0.872	0.760	mean_norm	0.736
	mean_raw	0.702	0.759	0.609	0.871	0.760	mean_raw	0.740

cdk2		ROCAUC							EF1					
cid	fred	hybrid	plants	rdock	smina	surflex	cons	fred	hybrid	plants	rdock	smina	surflex	cons
gro1	0.532	0.749	0.664	0.787	0.643	0.679	0.808	1.7	12.4	4.0	14.6	4.2	3.2	12.9
0	0.653	0.740	0.572	0.731	0.662	0.613	0.731	3.6	6.1	1.7	7.2	1.9	1.5	5.7
1	0.564	0.630	0.486	0.676	0.603	0.621	0.651	1.9	2.5	1.1	1.3	1.1	1.3	0.2
2	0.581	0.644	0.472	0.621	0.603	0.597	0.631	1.5	1.9	0.4	1.5	0.8	1.5	1.9
3	0.625	0.725	0.535	0.670	0.619	0.615	0.687	1.5	4.9	1.9	3.0	1.7	0.2	1.9
4	0.668	0.764	0.609	0.737	0.671	0.609	0.733	3.6	4.9	1.3	7.4	3.6	0.8	3.8
5	0.683	0.754	0.656	0.766	0.730	0.685	0.792	3.8	9.1	2.5	7.8	5.1	1.7	7.8
6	0.582	0.676	0.551	0.696	0.639	0.622	0.680	1.9	3.0	1.1	4.4	1.7	0.4	3.0
7	0.615	0.693	0.554	0.702	0.661	0.610	0.702	3.0	2.7	1.3	6.6	3.2	0.6	3.0
8	0.732	0.756	0.645	0.789	0.741	0.668	0.806	8.2	5.5	3.4	10.4	7.8	1.9	10.3
9	0.627	0.705	0.578	0.761	0.652	0.605	0.717	2.3	4.2	0.2	8.7	3.6	1.3	2.5
10	0.699	0.737	0.617	0.760	0.703	0.670	0.764	3.4	7.6	1.9	9.1	4.0	2.7	7.4
11	0.673	0.769	0.597	0.764	0.637	0.694	0.751	3.8	7.8	4.0	10.8	3.2	2.3	5.7
12	0.611	0.711	0.530	0.730	0.609	0.650	0.702	3.2	2.5	2.3	7.8	0.4	1.3	1.1
13	0.611	0.702	0.537	0.693	0.651	0.670	0.702	1.7	3.0	0.8	3.8	1.9	2.5	2.3
14	0.656	0.737	0.595	0.717	0.657	0.619	0.715	3.4	4.4	1.3	7.8	3.6	0.4	3.8
15	0.577	0.692	0.579	0.749	0.689	0.695	0.737	0.8	2.3	0.8	5.1	0.8	2.3	2.1
16	0.635	0.732	0.585	0.733	0.648	0.633	0.710	4.0	5.9	1.1	7.6	1.5	0.6	4.2
17	0.788	0.787	0.694	0.795	0.752	0.641	0.816	15.6	14.6	4.0	11.4	8.0	2.5	17.5
18	0.699	0.764	0.630	0.758	0.685	0.636	0.762	3.8	8.2	4.2	5.5	4.2	1.1	6.3
19	0.655	0.766	0.618	0.776	0.679	0.667	0.764	3.8	9.5	3.6	11.7	4.4	1.5	8.9
ens	0.682	0.773	0.594	0.777	0.701	0.673	0.754	4.0	8.4	2.7	11.0	2.5	0.6	3.8
								28278	28278	28294	28303	28304	28304	28305

hivpr	ROCAUC							EF1						
cid	fred	hybrid	plants	rdock	smina	surflex	cons	fred	hybrid	plants	rdock	smina	surflex	cons
gro1	0.654	0.685	0.803	0.598	0.740	0.806	0.824	1.5	5.5	13.8	4.7	4.7	11.6	20.0
0	0.616	0.628	0.805	0.669	0.740	0.837	0.812	3.8	4.3	19.2	8.0	9.3	14.4	19.4
1	0.578	0.635	0.821	0.715	0.761	0.826	0.839	3.0	6.8	22.2	7.3	8.8	17.9	21.1
2	0.588	0.605	0.832	0.714	0.758	0.823	0.835	3.0	5.5	21.8	9.1	5.6	17.2	20.5
3	0.561	0.491	0.800	0.662	0.720	0.791	0.759	2.7	2.1	16.6	5.4	3.9	13.4	11.8
4	0.543	0.602	0.845	0.694	0.747	0.818	0.821	1.3	4.5	19.2	4.5	7.5	9.3	19.6
5	0.596	0.576	0.827	0.720	0.743	0.825	0.839	4.6	4.7	18.3	6.0	5.8	19.8	19.4
6	0.601	0.602	0.810	0.667	0.733	0.823	0.810	5.3	6.0	17.7	7.5	6.2	16.0	19.6
7	0.673	0.629	0.830	0.672	0.723	0.818	0.839	6.5	6.0	20.9	6.9	6.2	13.2	20.1
8	0.606	0.610	0.816	0.712	0.737	0.796	0.817	4.2	4.3	19.4	9.3	4.5	12.1	18.8
9	0.604	0.610	0.795	0.668	0.734	0.824	0.811	4.4	4.2	17.7	4.5	5.8	17.0	18.3
10	0.590	0.615	0.845	0.666	0.731	0.825	0.824	2.5	5.5	18.5	4.9	7.5	9.1	22.2
11	0.480	0.531	0.827	0.700	0.717	0.789	0.777	2.1	4.7	15.3	7.8	2.8	7.8	17.2
12	0.555	0.596	0.824	0.740	0.748	0.819	0.817	2.7	7.8	20.1	6.2	5.4	13.8	17.5
13	0.531	0.547	0.781	0.637	0.696	0.770	0.745	1.5	3.8	11.8	6.3	5.6	11.0	14.9
14	0.540	0.577	0.814	0.739	0.728	0.804	0.812	2.7	6.2	15.7	8.2	5.0	11.2	17.4
15	0.511	0.593	0.812	0.700	0.697	0.800	0.783	1.1	3.4	14.6	6.3	4.1	12.3	17.5
16	0.689	0.665	0.821	0.654	0.741	0.803	0.839	6.6	3.8	17.7	3.7	6.2	12.3	20.7
17	0.601	0.584	0.834	0.679	0.740	0.797	0.809	4.4	6.2	15.9	5.8	7.3	8.8	19.2
18	0.612	0.629	0.840	0.750	0.727	0.821	0.845	3.8	6.0	20.5	7.5	3.9	12.5	23.1
19	0.549	0.601	0.823	0.737	0.736	0.811	0.827	2.3	5.9	15.7	7.3	4.3	14.2	18.8
ens	0.604	0.617	0.834	0.726	0.754	0.843	0.836	5.5	12.5	21.8	11.8	6.0	25.4	24.4
								35199	35786	36174	36188	36184	36193	36224

ROCAUC

using our best 6 docking programs

	cdk2	fa10	glcm	hivint	hivpr
Protein Conf	ROCAUC	ROCAUC	ROCAUC	ROCAUC	ROCAUC
cons_mean_gro1 (xtal)	0.789	0.834	0.691	0.808	0.823
cons_mean_00	0.714	0.650	0.668	0.757	0.815
cons_mean_01	0.647	0.781	0.584	0.810	0.844
cons_mean_02	0.623	0.732	0.584	0.780	0.839
cons_mean_03	0.675	0.724	0.602	0.760	0.770
cons_mean_04	0.719	0.597	0.669	0.815	0.825
cons_mean_05	0.783	0.794	0.596	0.831	0.843
cons_mean_06	0.675	0.644	0.578	0.751	0.816
cons_mean_07	0.693	0.594	0.645	0.796	0.842
cons_mean_08	0.791	0.653	0.634	0.810	0.822
cons_mean_09	0.711	0.849	0.611	0.815	0.810
cons_mean_10	0.752	0.777	0.679	0.815	0.829
cons_mean_11	0.744	0.640	0.692	0.812	0.789
cons_mean_12	0.694	0.667	0.597	0.871	0.825
cons_mean_13	0.695	0.625	0.573	0.835	0.751
cons_mean_14	0.700	0.728	0.668	0.812	0.819
cons_mean_15	0.736	0.626	0.627	0.792	0.793
cons_mean_16	0.703	0.805	0.620	0.730	0.840
cons_mean_17	0.798	0.760	0.665	0.825	0.817
cons_mean_18	0.751	0.822	0.658	0.763	0.849
cons_mean_19	0.756	0.673	0.653	0.806	0.832
ens_mean_fred	0.655	0.538	0.480	0.634	0.614
ens_mean_hybrid	0.764	0.630	0.722	0.632	0.632
ens_mean_plants	0.594	0.542	0.673	0.764	0.834
ens_mean_rdock	0.777	0.854	0.668	0.729	0.726
ens_mean_smina	0.701	0.764	0.576	0.851	0.754
ens_mean_surflex	0.673	0.654	0.594	0.645	0.843
full_mean	0.744	0.745	0.658	0.855	0.841

EF1

using our best 6 docking programs

Protein Conf	cdk2 EF1	fa10 EF1	glcm EF1	hivint EF1	hivpr EF1
cons_mean_gro1 (xtal)	13.3	10.6	29.6	11.0	18.5
cons_mean_00	5.3	1.9	22.2	9.0	19.0
cons_mean_01	0.6	7.1	9.3	12.0	20.1
cons_mean_02	1.3	3.9	16.7	14.0	19.8
cons_mean_03	1.1	4.7	20.4	11.0	11.9
cons_mean_04	3.2	1.3	20.4	13.0	19.2
cons_mean_05	6.3	4.7	7.4	12.0	17.9
cons_mean_06	2.1	2.4	18.5	5.0	17.7
cons_mean_07	2.5	8.0	13.0	12.0	18.3
cons_mean_08	10.8	2.6	16.7	8.0	18.8
cons_mean_09	3.0	8.4	22.2	19.0	17.9
cons_mean_10	6.8	4.8	24.1	8.0	20.7
cons_mean_11	5.7	7.6	18.5	11.0	15.1
cons_mean_12	0.6	1.9	16.7	11.0	16.4
cons_mean_13	1.5	7.1	22.2	9.0	13.8
cons_mean_14	2.5	6.1	20.4	10.0	15.7
cons_mean_15	1.7	0.9	20.4	16.0	15.3
cons_mean_16	4.9	5.0	11.1	9.0	19.8
cons_mean_17	15.2	5.8	20.4	10.0	17.9
cons_mean_18	5.9	9.3	18.5	10.0	20.3
cons_mean_19	7.2	7.3	22.2	11.0	17.7
ens_mean_fred	3.2	1.0	8.7	11.4	4.8
ens_mean_hybrid	7.8	2.3	15.2	7.6	11.3
ens_mean_plants	2.7	2.4	31.4	15.0	21.8
ens_mean_rdock	11.0	13.4	24.1	5.0	11.8
ens_mean_smina	2.5	3.2	5.6	12.0	6.0
ens_mean_surflex	0.6	4.7	18.5	7.0	25.4
full_mean	3.2	5.0	24.1	13.0	22.6

Is there hope for “Smart” Ensemble Scoring?

	mean				max			
ROCAUC	xtal	ens	xtal+ens	ens5	xtal	ens	xtal+ens	ens5
cdk2	0.808	0.754	0.760	0.807	0.784	0.778	0.787	0.799
fa10	0.844	0.746	0.756	0.844	0.835	0.822	0.827	0.838
glcm	0.706	0.666	0.670	0.695	0.738	0.692	0.698	0.746
hivint	0.824	0.875	0.877	0.910	0.781	0.790	0.792	0.815
hivpr	0.824	0.836	0.837	0.857	0.798	0.818	0.815	0.838

EF1	xtal	ens	xtal+ens	ens5	xtal	ens	xtal+ens	ens5	
cdk2	12.9	3.8	4.2	12.0	9.3	10.8	11.4	14.6	
fa10	12.3	6.0	6.3	8.0	9.9	6.5	8.2	11.4	
glcm	27.8	22.2	24.1	29.6	20.4	14.8	14.8	13.0	18.5
hivint	13.0	21.0	21.0	24.0	4.0	6.0	6.0	9.0	
hivpr	20.0	24.4	25.0	26.9	6.2	10.3	10.1	13.4	

rocauc	best_cids	ef1	best_cids	rocauc	best_cids	ef1	best_cids
cdk2	17,8,5,19,10	cdk2	17,8,19,5,10	cdk2	17,8,5,19,10	cdk2	17,19,8,10,5
fa10	9,18,16,5,1	fa10	18,9,7,1,13	fa10	9,18,6,5,1	fa10	9,7,1,13,16
glcm	11,10,17,4,14	glcm	19,0,10,9,6	glcm	11,8,4,7,10	glcm	9,0,18,4,14
hivint	12,5,13,17,14	hivint	9,5,2,7,15	hivint	12,10,11,4,19	hivint	14,15,17,9,3
hivpr	18,1,5,7,16	hivpr	18,10,1,16,2	hivpr	1,18,2,12,5	hivpr	1,19,6,5,9

Next Steps for Ensemble Docking

- get missing docking data from key programs (Fred/Hybrid)
- evaluate diversity of active compound retrieval
- explore longer trajectory or enhanced sampling techniques
  - apo vs. holo?
- Boolean masking of docking scores based on pose consensus?
- consider more protein conformers?
- consider fewer protein conformers (smart selection):
  - transfer learning
  - supervised learning
  - active learning?
  - unsupervised selection?
- consider using ensemble of available experimental structure?

ML Approaches

- Transfer Learning: train ML model using data from all off-targets:

```
label   model                   features

 y   =   f   ( dp1 { min, 25%, 50%, 75%, max },
               dp2 { min, 25%, 50%, 75%, max },
               dp3 { min, 25%, 50%, 75%, max },
                 …
               dp6 { min, 25%, 50%, 75%, max } )
```

distribution from n=20
scores produced for
compound against 20
target conformations using
docking program 1

- apply model to predict labels for compounds docked to on-target

ML Approaches

- Supervised, pick subset of protein conformers:
  - “all-stars”: take top 5 best performing conformers (5 out of 20) ☺
  - “the draft” (heuristic): start with best conformer than add conformers incrementally with maximal gain 😐
  - “champions”: (brute-force) take optimal set of 5 from all combinations: 20-choose-5 = 15504 ☹