2.1.3. Virtual Screening

- This module is a mini-lecture introducing virtual screening
- At the end of this module, you should be able to answer the following questions:
  - What is virtual screening and why is it used?
  - What types of chemical libraries are used in virtual screening?
  - How is a chemical library prepared for virtual screening?
  - What is a virtual screening hit and what are some desirable properties of a hit?
  - How are virtual screening programs assessed?

What is virtual screening and why is it used?

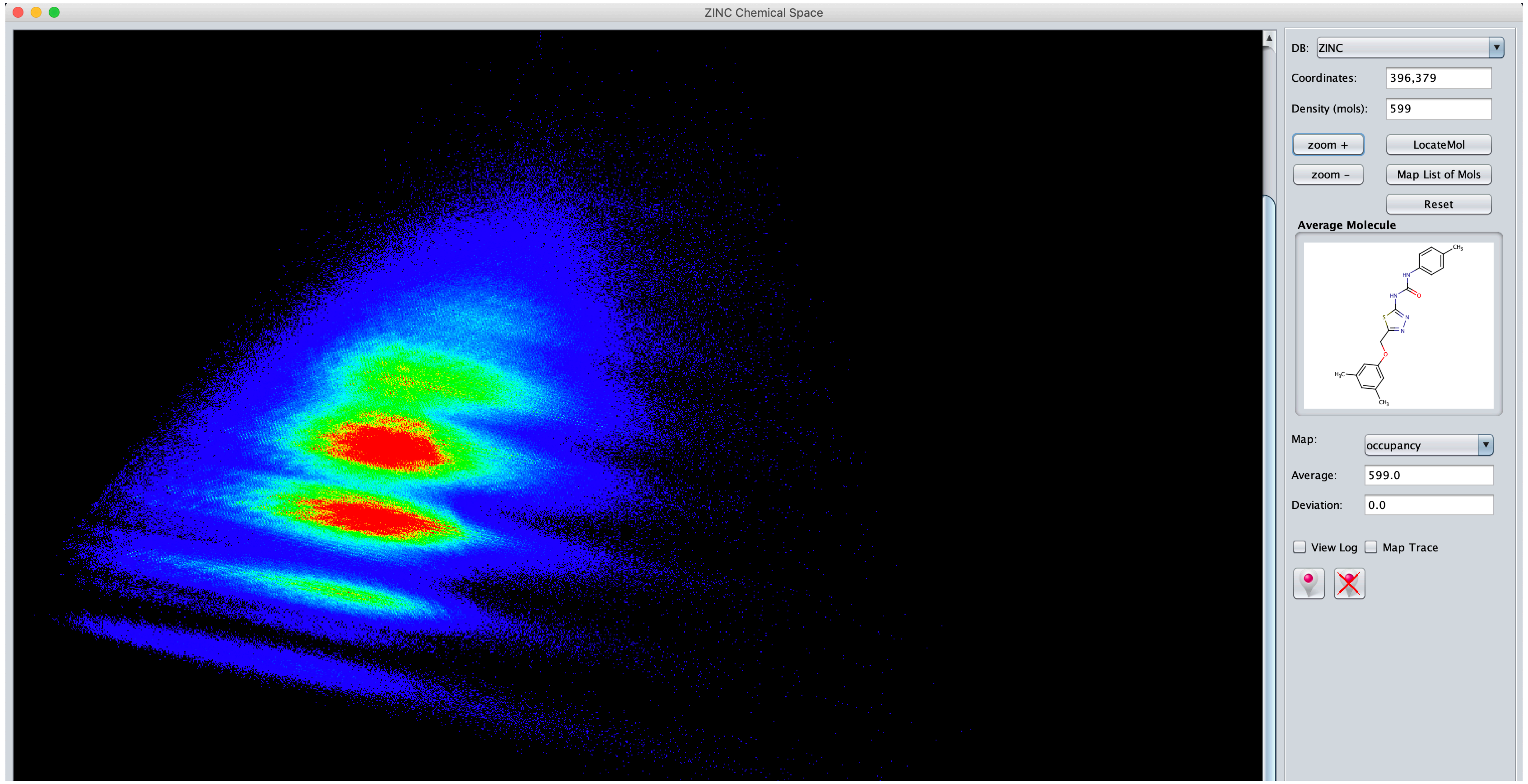
- Virtual screening
  - use of computation to estimate the activity (e.g. binding or inhibition) of a database of chemical compounds against a target
  - “virtual” in contrast to experimental high-throughput screening (HTS)
  - usually based on molecular docking, but machine learning in vogue
- Why is it used?
  - to help obtain leads for drugs and chemical probes
    - prioritizing compounds for experimental follow-up
    - faster predictions of specificity
  - compared to HTS
    - cheaper, accessible to academic laboratories
    - can screen larger libraries, increasing likelihood of good hit
    - HTS suffers from pan-assay interference compounds (PAINS)

Single Docking	Library Screen
Use GUI	Use scripts
Data in one directory	Data in tree structure
Single ligand PDBQT	Several ligand PDBQT
One interactive calculation	Submit jobs to cluster
Visually inspect all results	Visually inspect best results

based on Morris et al, 2008

How should we decide what to screen?

- Chemical space is vast and largely unexplored
  - Possible small organic molecules estimated > 10^{60}
  - Generated and collected in a database (GDB)
    - GDB-11: 26.4 million compounds with up to 11 atoms of C, N, O, and F
    - GDB-13: ~1 billion compounds with up to 13 atoms of C, N, O, S, and Cl
  - Exhaustive search not necessarily feasible or useful
- Different types of chemical libraries may be suitable



Generated by MQN-Mapplet (http://gdb.unibe.ch/tools/)
[Awale et al, 2013]

What types of chemical libraries are used?

Type of library	Analogy	Examples
Comprehensive	Search in the dark	ZINC15:~1 billion compounds in vendor catalogs. ~11 million in stock.
Combinatorial	Search in the dark	Enamine REAL:13 billion “readily accessible” molecules.
Diverse	Efficient search in the dark	Diverse REAL drug-like:15 million. NCI Diversity Set VI:1548 free.
“Focused” or “Targeted” for lead identification	Search with a flashlight	Filtered for a structural motif or pharmacophore
“Focused” or “Targeted” for lead optimization	Focusing the spotlights	Riboflavin analogues

Analogies from Morris et al, 2008

Molecular weight is an important factor

Class	Weight	Why do virtual screening?
Fragments	< 300 Da	to join together into leads
Lead-like	300-375 Da	low potency compounds that can be optimized
Drug-like	<500 Da	potential to be highly potent and suitable for preclinical testing

Zinc is not Commercial (ZINC15)

- a free database of commercially-available compounds for virtual screening
- divided into "tranches" by molecular weight and LogP, a measure of predicted hydrophobicity
  - predefined subsets of ZINC15 include fragments, lead-like, and drug-like
- organized into catalogs, notably
  - by vendor
  - BindingDB.org - binding affinity database
  - DrugBank-approved - approved drugs, helpful for repurposing
  - NCI HTS libraries: plated 2007 and diversity 3.
  - Sigma Aldrich building blocks
- can search for analogs of a particular compound (SAR by catalog)
- buyer beware!

ZINC15 Tranches

Rep. | 2D | 3D | React. | Standard | Purch. | Wait OK | pH | N/A | Charge | N/A

Predefined Subsets

- All
- None
- Shards
- Fragments
- Flagments
- Goldilocks
- Lead-Like
- Lugs
- Drug-Like
- Big-n-Greasy

Molecular Weight (up to, Daltons)

LogP (up to)	200	250	300	325	350	375	400	425	[illegible]	500	>500	Totals, by LogP
-1	31,520	210,198	784,279	1,124,026	2,310,970	854,208	300,607	56,[illegible]	[illegible]	23,759	5,598	5,801,749
0	154,788	1,095,532	3,992,760	5,390,653	10,938,346	3,784,188	1,767,726	467,[illegible]	[illegible]	209,867	3,862	28,410,922
1	414,273	3,374,769	13,196,175	17,182,223	34,851,540	12,860,730	7,279,946	2,568[illegible]	[illegible]	1,206,882	8,249	95,133,476
2	555,197	5,533,895	25,622,912	33,383,218	88,330,050	28,983,131	19,267,814	8,741[illegible]	[illegible]	8,570,868	22,307	226,760,093
2.5	214,735	2,723,802	14,831,118	19,890,876	40,686,668	20,475,678	15,126,147	8,256,[illegible]	[illegible]	7,879,918	23,424	137,854,889
3	125,492	2,145,037	13,281,388	18,488,586	37,136,913	22,333,418	17,838,728	11,103,726	10,986,050	6,968,806	38,186	140,446,330
3.5	57,244	1,386,997	10,135,959	14,739,692	29,791,008	21,458,732	18,737,428	13,319,769	13,704,019	9,066,863	61,984	132,459,695
4	18,215	642,659	6,131,454	8,393,140	12,628,179	15,472,307	16,892,846	13,888,572	12,864,529	10,473,721	91,903	97,497,525
4.5	2,275	180,303	2,873,064	4,781,472	7,950,935	10,959,424	12,773,295	12,430,134	11,562,431	10,654,985	127,265	74,295,583
5	94	22,897	852,691	1,984,202	4,012,544	6,416,455	8,397,678	9,229,001	8,818,548	9,234,778	155,546	49,124,434
>5	28	895	44,519	178,728	557,180	1,226,923	2,066,211	2,641,238	3,062,143	2,471,283	799,799	13,048,947
Totals, by Weight	1,573,861	17,316,984	91,746,319	125,536,816	269,194,333	144,825,194	120,448,426	82,703,885	79,387,972	66,761,730	1,338,123	**1,000,833,643** Substances; **1.7K** Tranches

How is a chemical library prepared for virtual screening?

- Chemicals in libraries are usually specified by a string of characters
  - Simplified Molecular-Input Line-Entry system (SMILES)
    - Element abbreviation, possibly in square brackets
    - Bonds between atoms
  - IUPAC International Chemical Identifier (InChI)
- Docking requires 3D structures and molecular mechanics parameters
  - Conformer generation programs such as OpenEye Omega or Balloon can be used to create 3D structures. ZINC provides 3D structures based on Omega.
  - In AutoDock, parameters for each ligand atom are the partial charge and atom type

What is a virtual screening hit and what are some desirable properties of a hit?

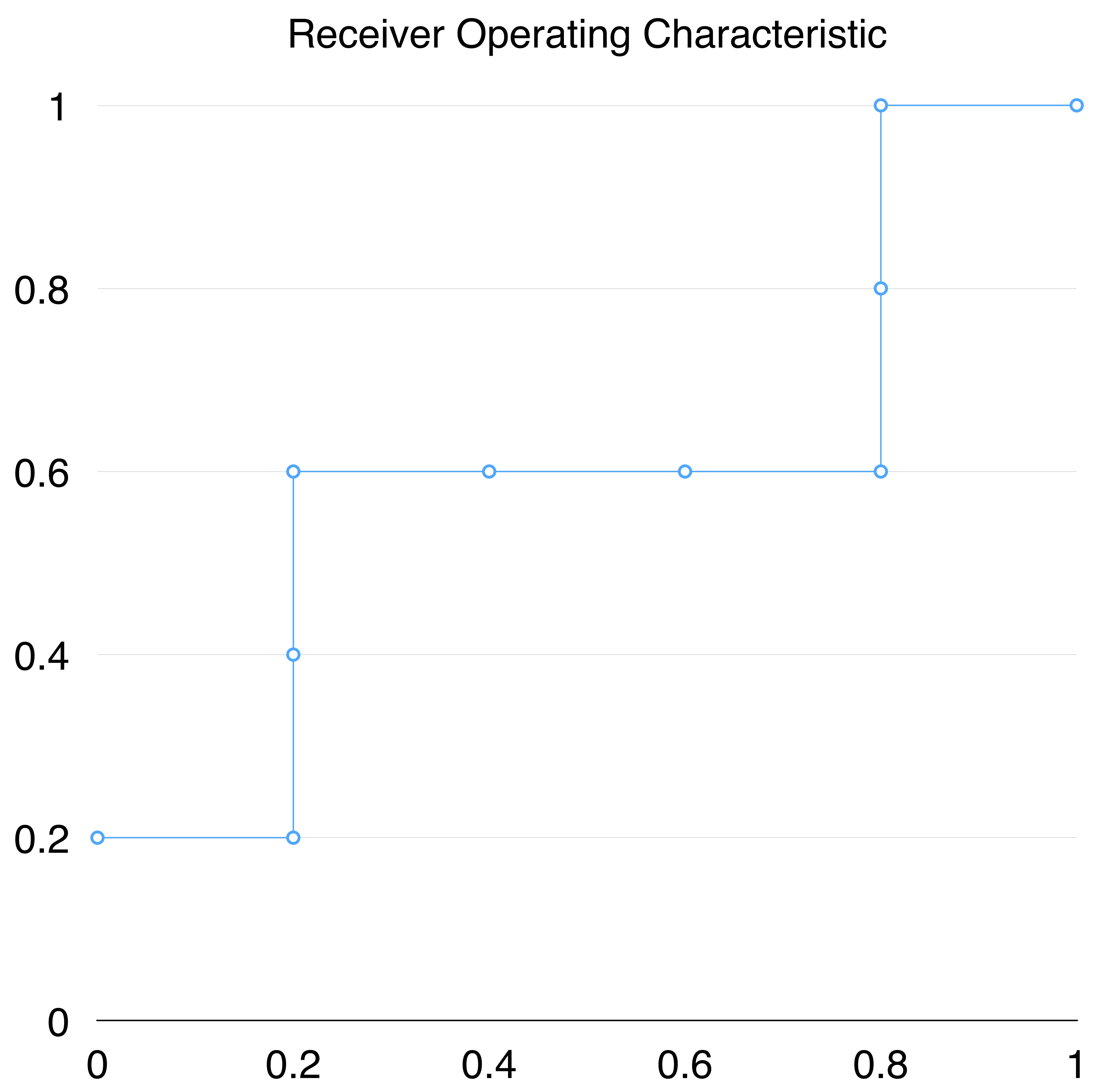
- A virtual screening hit has a low docking score
- If there is a positive control, the docking score is comparable or lower than the positive control
- Ideally, a virtual screening hit should
  - have a number of clear contacts with the receptor
  - have most atoms in contact with the receptor, opposed to in solvent

How are virtual screening program assessed?

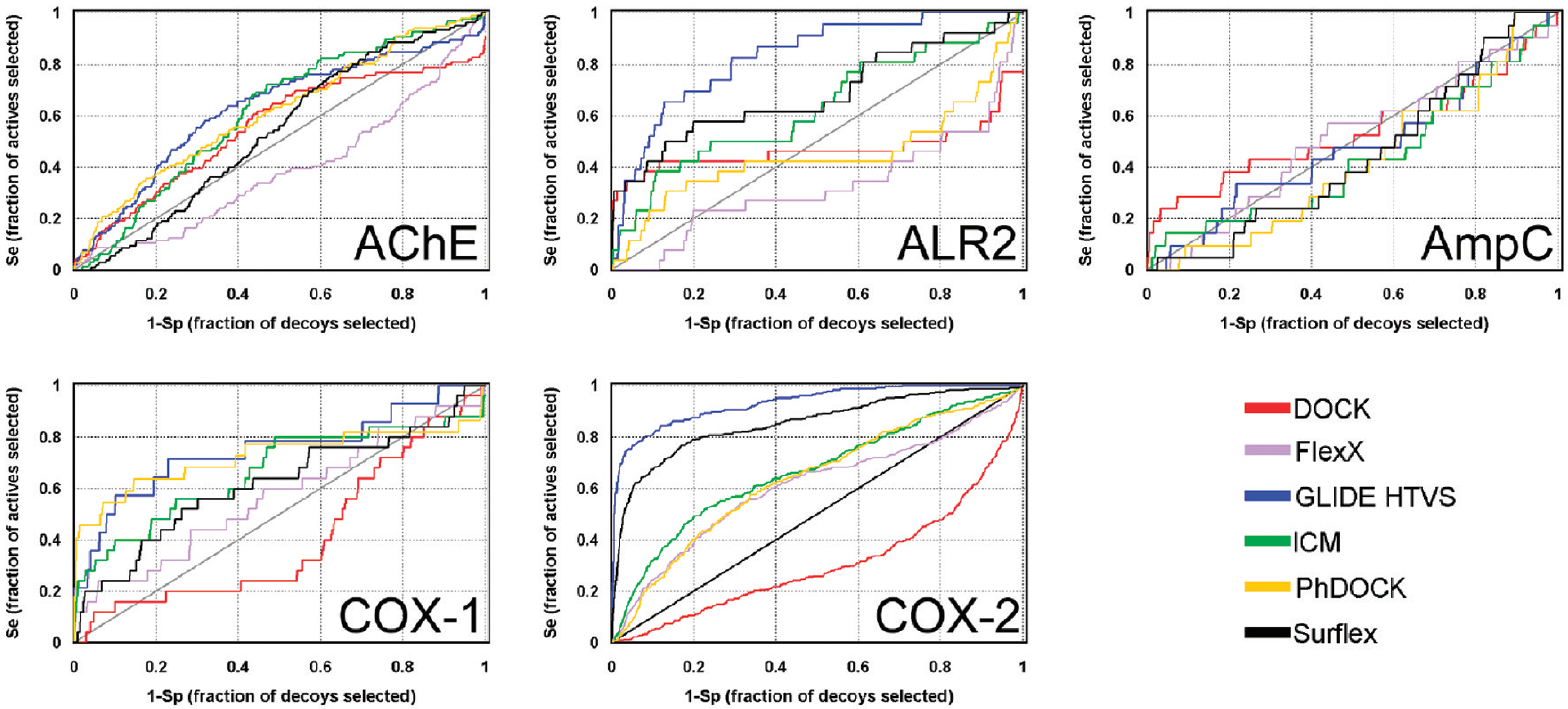
- Binding poses
  - Similarity of predicted poses to solved structures
  - Root mean square deviation (RMSD) < 2 A considered good
- Docking scores
  - Correlation between scores and measured affinities
  - Active molecules should have lower scores than
    - inactive compounds
    - decoys, which are have similar chemical properties but different connectivity than active molecules
  - Quantified by receiver operating characteristic (ROC) curve, area under the ROC curve (AUC), enrichment factor
- Reported values can be biased
  - Computational chemists, especially methods developers, tempted to tweak approach (e.g. parameters) until experiments are reproduced.
  - Blinded challenges (e.g. D3R Grand Challenge) reduce bias

The Receiver Operating Characteristic (ROC) Curve

score	fraction of actives	fraction of inactives
-53.4	0.2	0
-50.2	0.2	0.2
-49.2	0.4	0.2
-45.7	0.6	0.2
-42.1	0.6	0.4
-35.2	0.6	0.6
-30.0	0.6	0.8
-21.3	0.8	0.8
-20.7	1.0	0.8
-4.2	1.0	1.0



Molecular docking is often an unreliable binary classifier



Cross et al, 2009

Review

- What is virtual screening and why is it used?
- What types of chemical libraries are used in virtual screening?
- How is a chemical library prepared for virtual screening?
- What is a virtual screening hit and what are some desirable properties of a hit?
- How are virtual screening programs assessed?

References

- Awale, M.; van Deursen, R.; Reymond, J.-L. MQN-Mapplet: Visualization of Chemical Space with Interactive Maps of DrugBank, ChEMBL, PubChem, GDB-11, and GDB-13. J. Chem. Inf. Model. 2013, 53 (2), 509–518. https://doi.org/10.1021/ci300513m.
- Cross, J. B.; Thompson, D. C.; Rai, B. K.; Baber, J. C.; Fan, K. Y.; Hu, Y.; Humblet, C. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. Journal of Chemical Information and Modeling 2009, 49 (6), 1455–1474. https://doi.org/10.1021/ci900056c.
- [Morris et al, 2008a] Presentation: Using AutoDock 4 for Virtual Screening (Handouts, PDF document, 1.1 MB)
  - http://autodock.scripps.edu/faqs-help/tutorial/using-autodock4-for-virtual-screening/VSTutorial2.2008.pdf
  - This presentation also describes some virtual screening success stories
- [Morris et al, 2008b] Instructions: Using AutoDock 4 for Virtual Screening (PDF document, 464 KB)
  - http://autodock.scripps.edu/faqs-help/tutorial/using-autodock4-for-virtual-screening/UsingAutoDock4forVirtualScreening_v4.pdf
- Sterling, T.; Irwin, J. J. ZINC 15 – Ligand Discovery for Everyone. J. Chem. Inf. Model. 2015, 55 (11), 2324–2337. https://doi.org/10.1021/acs.jcim.5b00559.