Computational tools in pharmacology: Merging artificial and organic intelligence

VANDERBILT UNIVERSITY

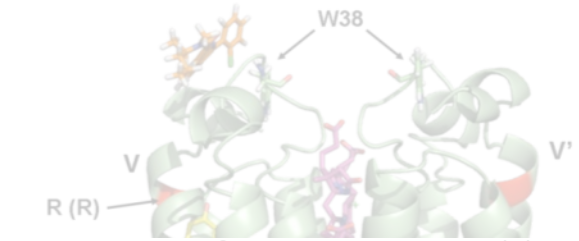
Shannon T. Smith
Ph.D. Candidate – Dr. Jens Meiler
6 May 2022

Disclaimers:

- Some slides were generously given by Dr. Jens Meiler
- Others were adapted from Dr. Carrie Jones's Modern Drug Discovery class
  - Drs. Dave Weaver, Craig Lindsley, Alex Waterson, Jeff Conn, Colleen Niswender

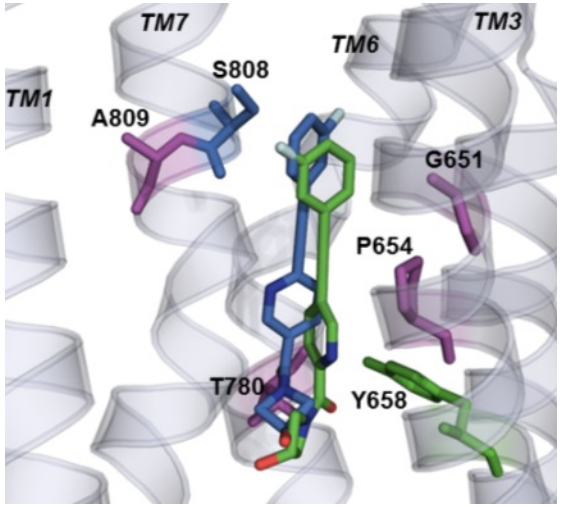
Olivia Marie Smith
Born May 3, 2022

Computational Structural and Chemical Biology in the Meiler Lab

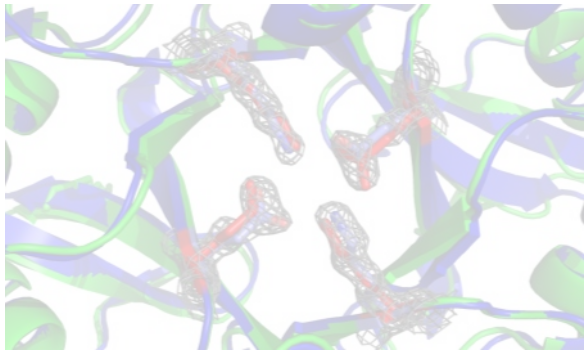


Weiner, et al., **Structure**, 2013, 21(7)

Protein structure prediction de novo and from limited experimental data

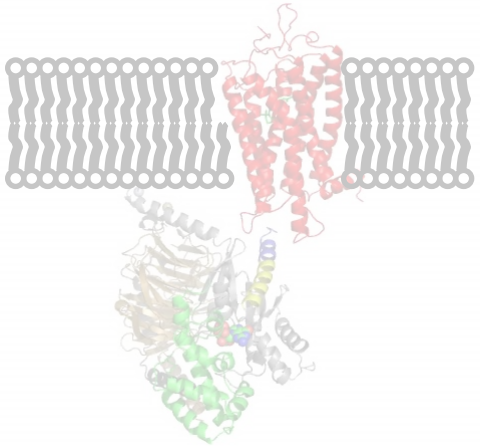


Dong, et al., **PLoS ONE**, 2013, 8(7)



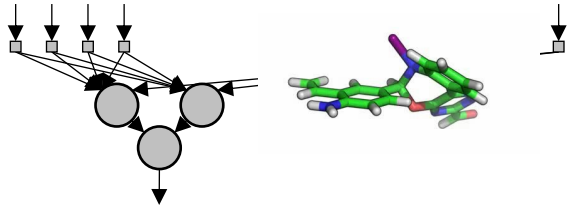
Fortenberry, et al., **JACS**, 2011, 133(45)

Design of large protein scaffolds, antibodies, and protein/ligand interfaces



Alexander et al., **NSMB**, 2014, 21(1)

Merging ligand- and structure-based computer-aided drug discovery

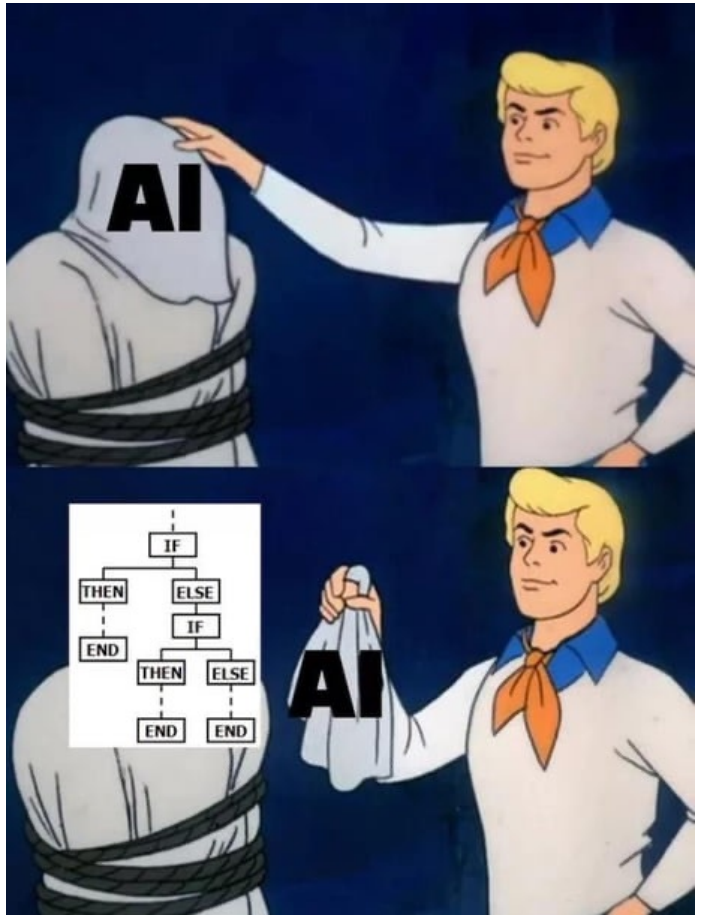


Muller, et al., **ChemMedChem**, 2012, 7(3)

Willis, et al., **PLoS Comp. Biol.**, 2013, 9(4)

Goals for this talk:

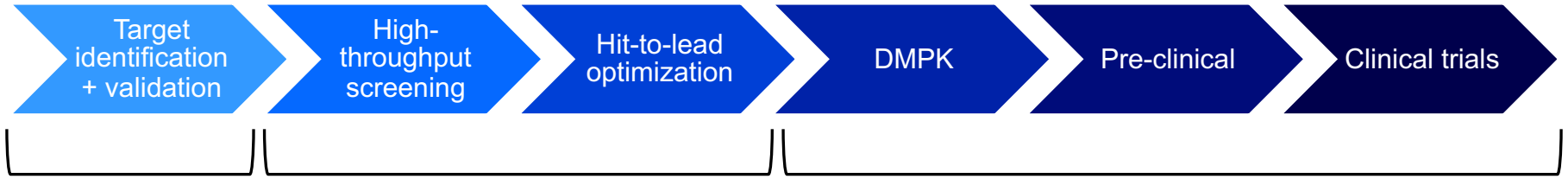
- Obtain a basic understanding for how computation, including AI, is currently applied in drug discovery at different stages and how you can apply this to your own work
- Demystifying AI in drug discovery
- Outline the advantages and disadvantages, provide realistic expectations



This talk is **NOT** about:

- Explaining the details of algorithms
- Me convincing you that AI will solve all your problems (though it can help)

Which stage are you working in?



Don't have a target

Have a target and in compound discovery phase

Have a target, a compound(s), and just want to make a drug now

I live here

Drug discovery: a crude history

pre-1800s | 1800s- early 1900s | 1900s | 2000s

Opium, ~3400 BCE

Coca leaves

Morphine **isolated** from opium (1805)

Chloral hydrate **synthesized** as synthetic sleeping drug (1832)

Arsphenamine/Compound 606 **synthesized for testing** for anti-syphilitic properties (1909)

Biological sciences:
molecular cloning, recombinant DNA technology, Human Genome Project, structural biology technologies

Chemical sciences:
Combinatorial chemistry

Computer science:
Uniprot, Protein Data Bank, AlphaFold, data storage/computational power

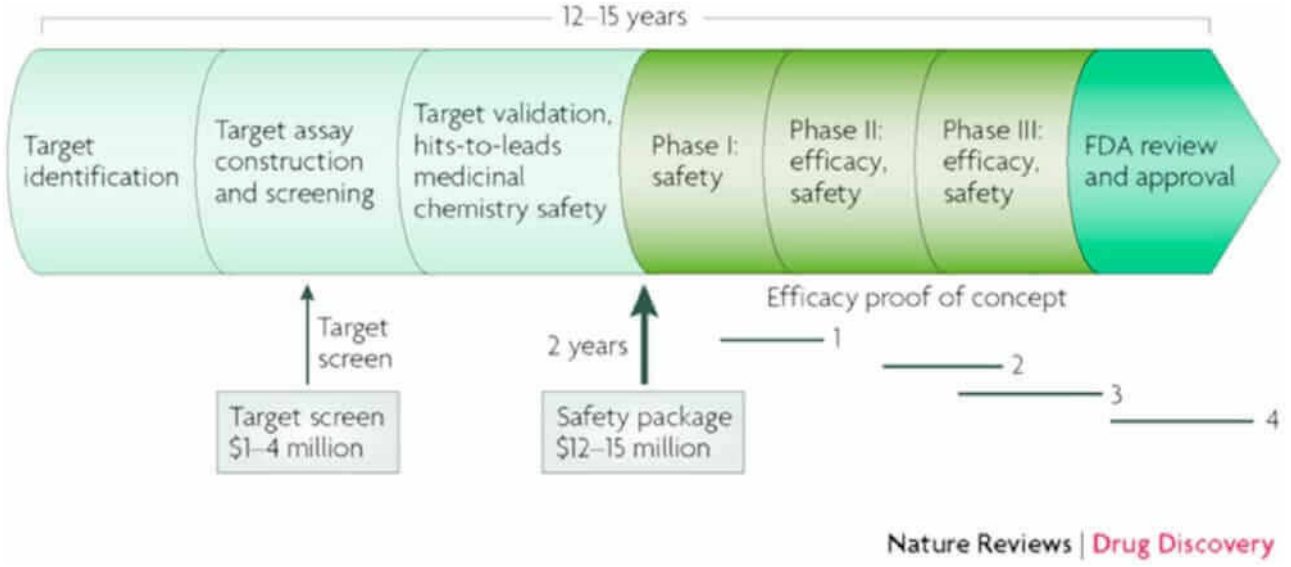
- MoA/target ID
- Screening technologies
- Rational-based design

- Computational tools in all of the above

Methods Mol Biol. 2009;572:3-12. doi: 10.1007/978-1-60761-244-5_1.

Why use AI in drug discovery?

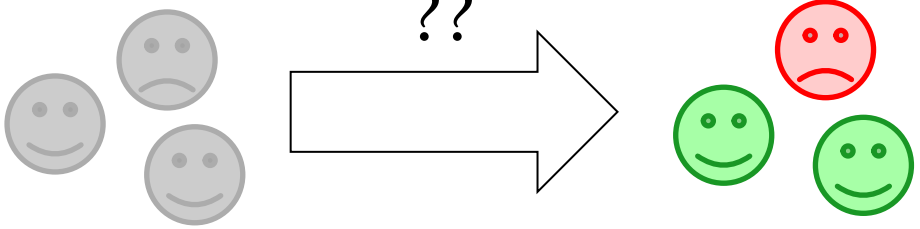
- Because we need to make drugs faster and cheaper



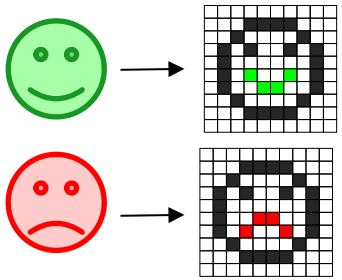
Roses, A. *Nat Rev Drug Discov* **7,** 807–817 (2008).

Machine Learning (ML) is great for pattern recognition

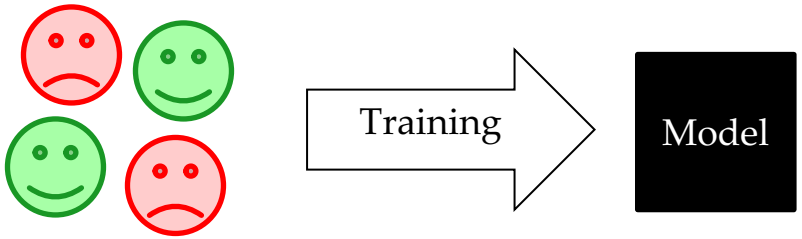
What determines if a face is a smile or frown?



Step 1: Data acquisition
Make faces machine-readable)



Step 2: Training
Given known smiles and frowns, computer figures out which features define each.

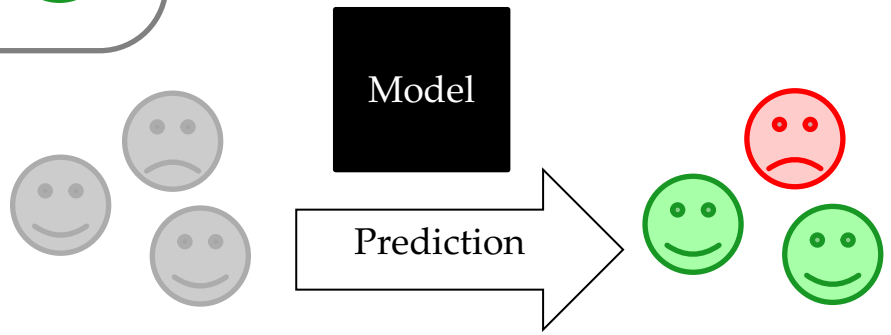


Step 3: Validation
How well do we perform on a known test set?

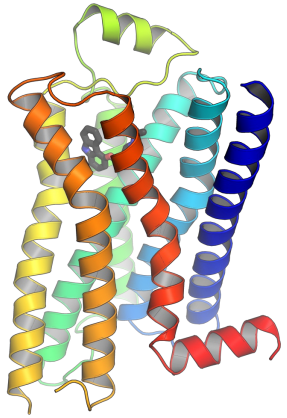
Predicted \ Actual	Smile	Frown
Smile	9	1
Frown	0	10

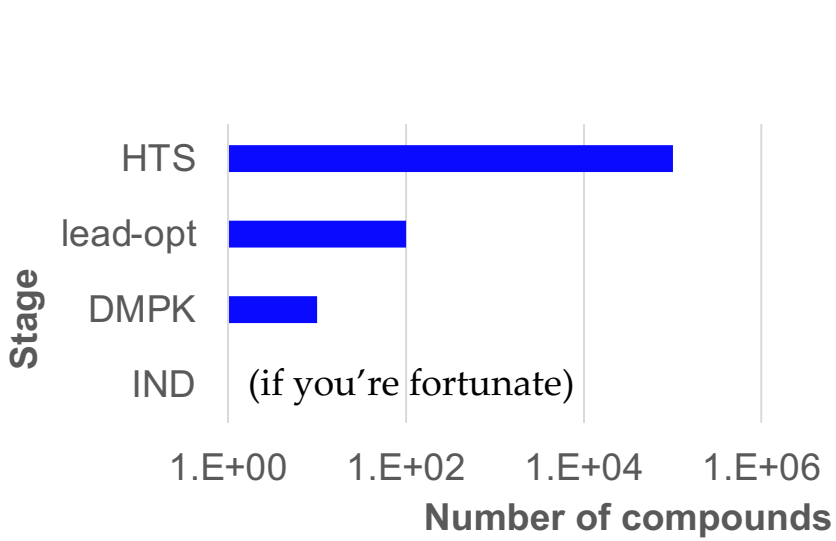
Accuracy = (9+10)/20 = 0.95

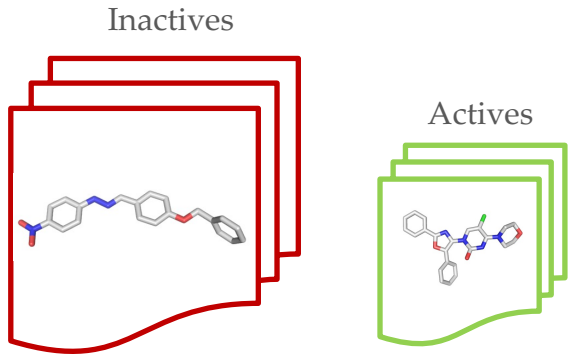
Step 4: Prediction
Given faces with unknown expressions, can we predict smile or frown?



Computer-aided drug discovery from a distance



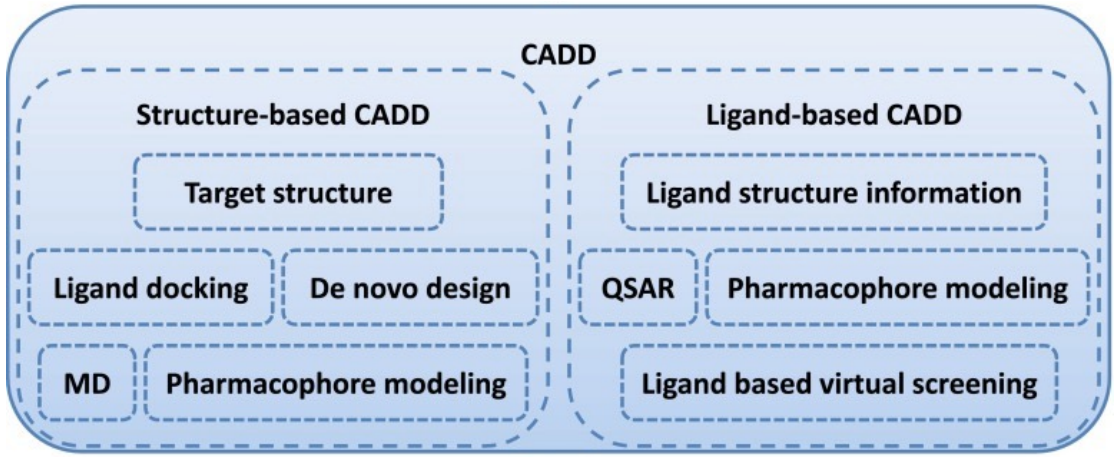




Structure-based Methods

Docking, Molecular Dynamics, SB-pharmacophore mapping

- Need high-resolution structure
- Understand mode of binding, key residues
- Rely on scoring functions
- Poorly ranks binding affinities of multiple compounds
- Slow/more expensive

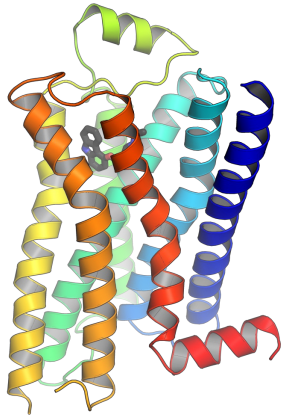


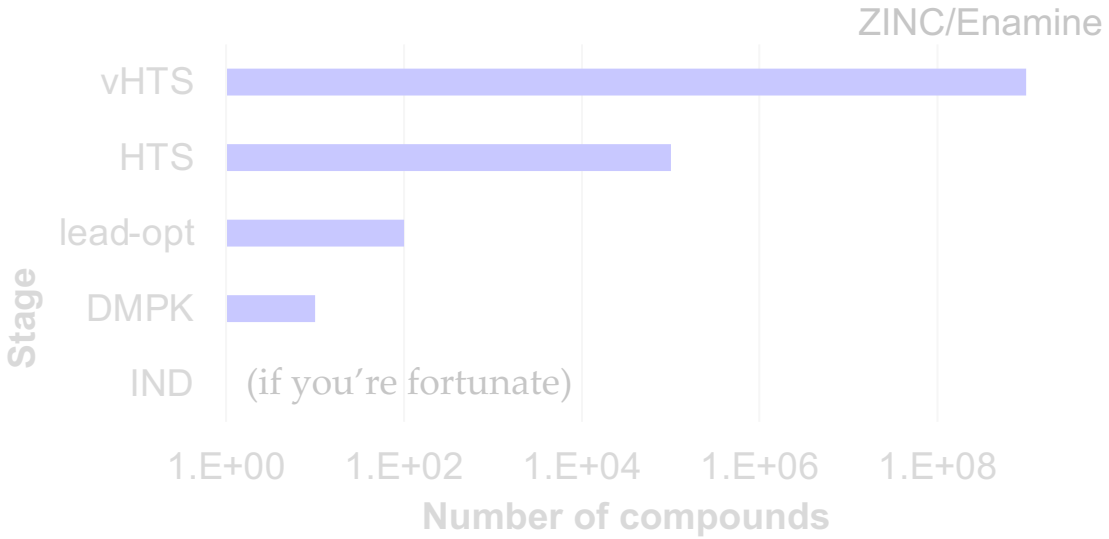
Ligand-based Methods

QSAR, LB-pharmacophore mapping, CoMFA

- No receptor pocket information
- Limited to chemical information provided by dataset (scaffold hopping)
- Consistency of experimental data used in reference set
- Ratio of descriptors to actives → over-training

Computer-aided drug discovery from a distance

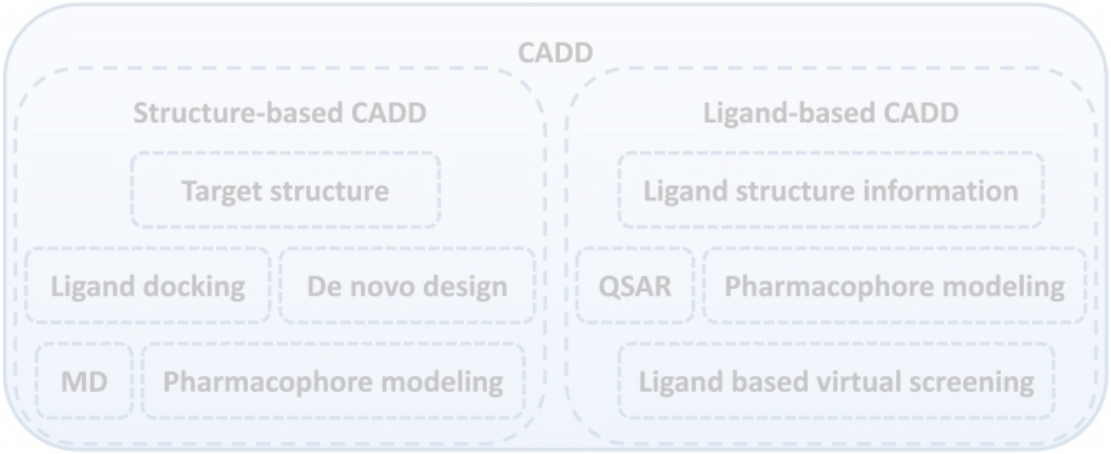




Structure-based Methods

Docking, Molecular Dynamics, SB-pharmacophore mapping

- Need high-resolution structure
- Understand mode of binding, key residues
- Rely on scoring functions
- Poorly ranks binding affinities of multiple compounds
- Slow/more expensive



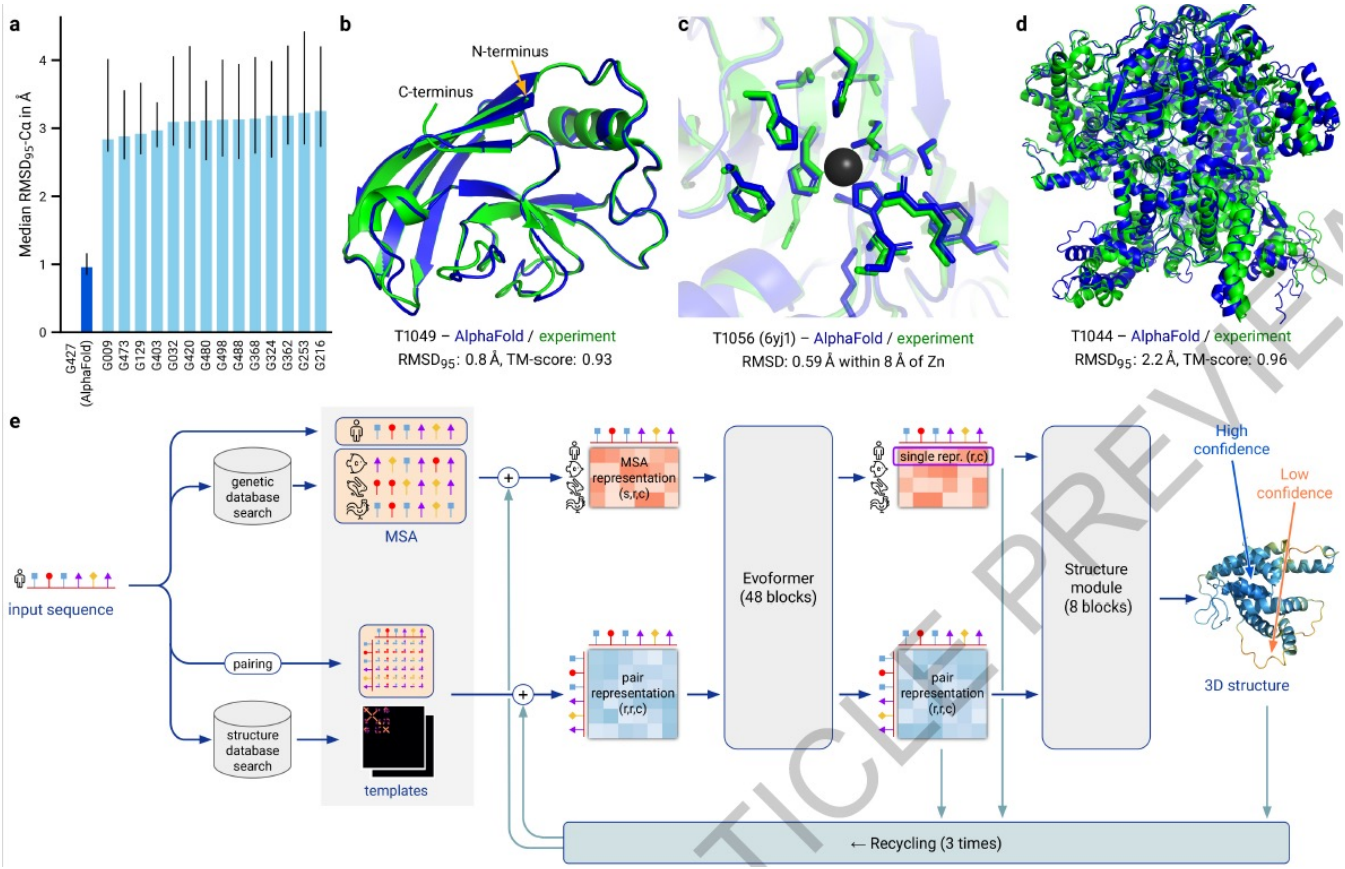
Ligand-based Methods

QSAR, LB-pharmacophore mapping, CoMFA

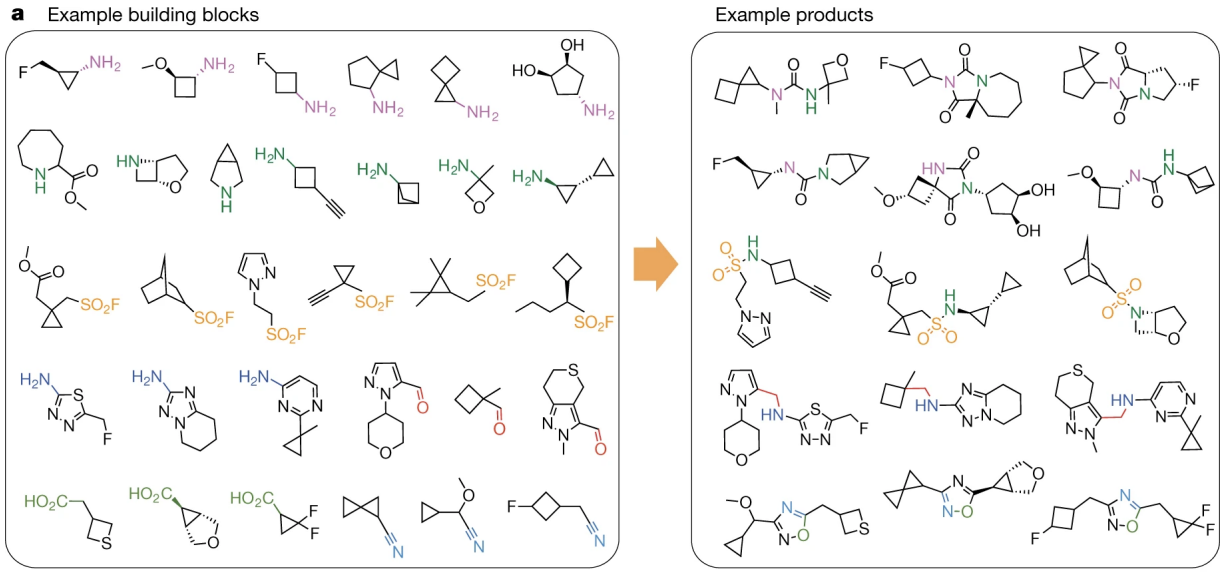
- No receptor pocket information
- Limited to chemical information provided by dataset (scaffold hopping)
- Consistency of experimental data used in reference set
- Ratio of descriptors to actives → over-training

AlphaFold: Highly accurate protein structure prediction

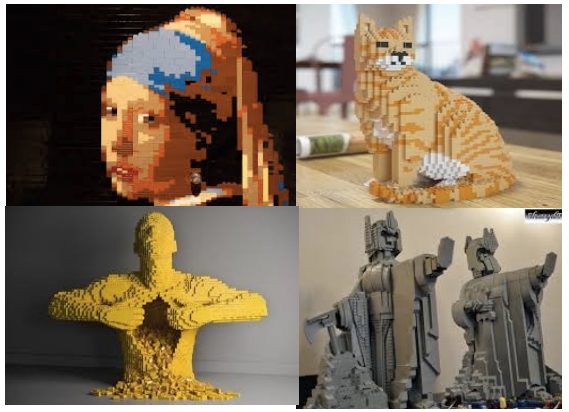
- DeepMind released AlphaFold algorithm and publicly available database of predicted structure of ”entire” human proteome and ~20 model organisms
- This was one big data problem “solved” with AI
- Accurate with membrane proteins also



Virtual synthesis library contains ~1 billion compounds to screen

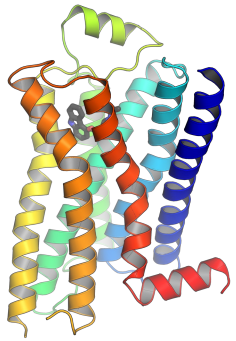


Enamine

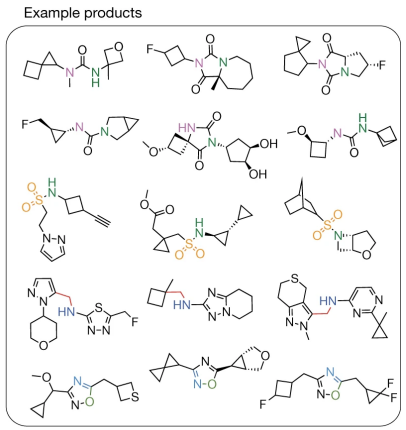


LEGO

Molecular docking of ultra-large libraries identifies binders of AmpC and D_4



+



AmpC

Dock 99 Million Compounds

↓

Cluster/select compounds to test

↓

44/55 synthesized

↓

Hits ranged from K_i = 1.3-400 uM

↓

Optimized to 77 nM

D_4 dopamine receptor

Dock 138 Million Compounds

↓

Cluster/select compounds to test

↓

549/589 synthesized

↓

81 hits
K_i < 10 uM

↓

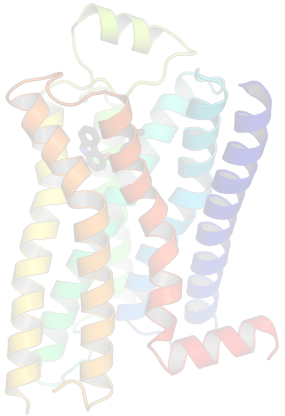
180 pM hit

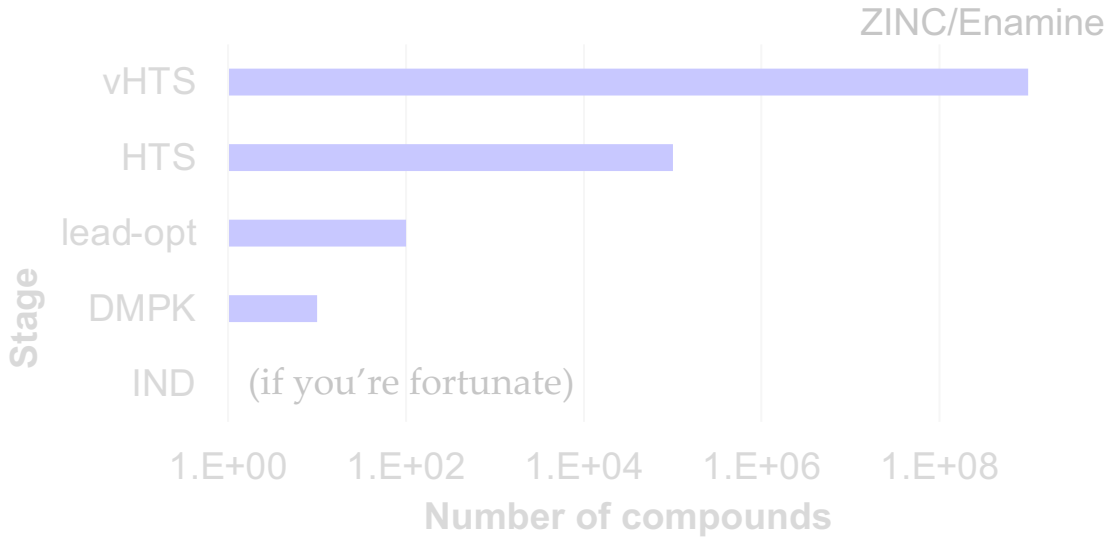
There are efforts to incorporate ML into these structure-based methods

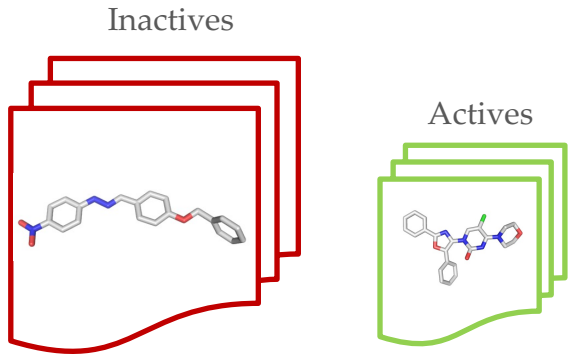
- Learning contact prediction
- de novo ligand design
- MD speed-up with ML

Shoichet (2019)

Computer-aided drug discovery from a distance



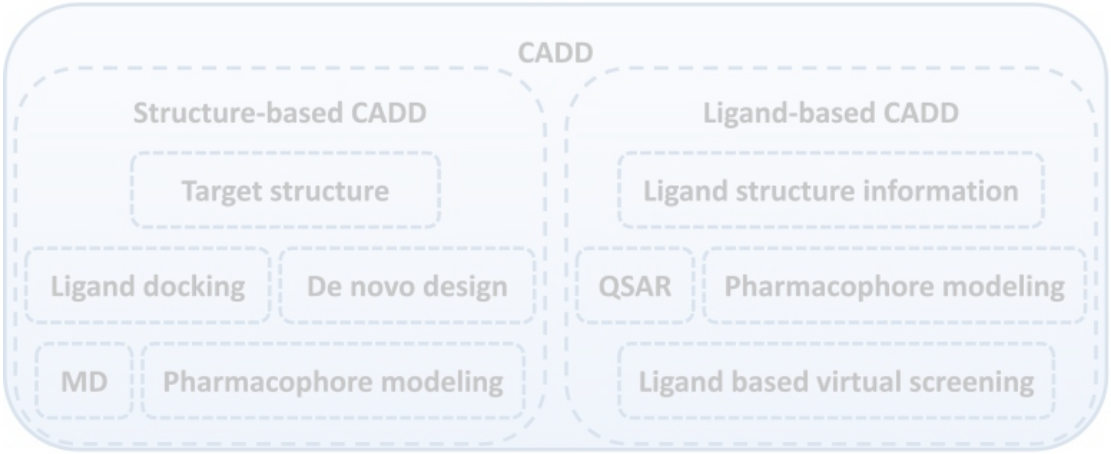




Structure-based Methods

Docking, Molecular Dynamics, SB-pharmacophore mapping

- Need high-resolution structure
- Understand mode of binding, key residues
- Rely on scoring functions
- Poorly ranks binding affinities of multiple compounds
- Slow/more expensive



Ligand-based Methods

QSAR, LB-pharmacophore mapping, CoMFA

- No receptor pocket information
- Limited to chemical information provided by dataset (scaffold hopping)
- Consistency of experimental data used in reference set
- Ratio of descriptors to actives → over-training

Journal of Medicinal Chemistry

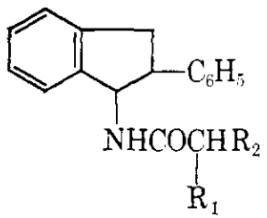
VOLUME 7, NUMBER 4 JULY 6, 1964

A Mathematical Contribution to Structure–Activity Studies

SPENCER M. FREE, JR., AND JAMES W. WILSON

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pennsylvania

Received February 4, 1964



	R_1	
R_2	H	CH_3
$N(CH_3)_2$	2.13	1.64
$N(C_2H_5)_2$	1.28	0.85

LD_{50} values (mg/10g)

What is the SAR (qualitatively)?

- R_1: larger group contributes to lower LD_{50} values
- R_2: larger group contributes to lower LD_{50} values
- Assuming R_1 and R_2 are independent

What is the Quantitative SAR?

response = average + effect of R_1 substituent + effect of R_2 substituent

$$LD_{50} = \mu + a[H] + a[CH_3] + b[N(CH_3)_2] + b[N(C_2H_5)_2]$$

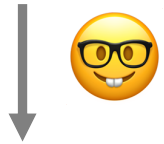
where μ = over-all average

$a[H]$ = contribution of H substituent at position R_1

$a[CH_3]$ = contribution of CH_3 substituent at position R_1

$b[N(CH_3)_2]$ = contribution of $N(CH_3)_2$ substituent at position R_2

$b[N(C_2H_5)_2]$ = contribution of $N(C_2H_5)_2$ substituent at position R_2



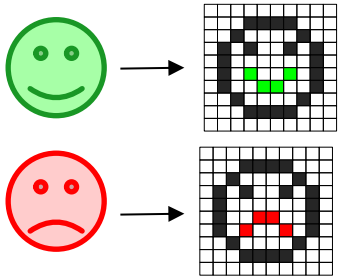
Assuming
1. R_1 and R_2 are independent
2. Symmetric contributions to R group total

Multiple Linear Regression problem

- R_1 = H contributes +0.245 to LD_{50}
- R_1 = CH_3 contributes -0.245 to LD_{50}
- R_2 = $N(CH_3)_2$ contributes +0.425 to LD_{50}
- R_2 = $N(C_2H_5)_2$ contributes -0.425 to LD_{50}

"The suggested mathematical models do not compensate for the three dimensionality of compounds, pH, pK_a, or other similar physical properties. Perhaps, in time, these can be built into the models for better estimation."

Chemical “descriptors” allow us to describe models with numeric values



1-D (scalar) descriptors
- MW
- # of C/N/O/H atoms
- Formal charge

2-D descriptors
- TPSA
- # of HBA/HBD
- # of rotatable bonds
- logP
- # of rings (by ring size)

- 2-D autocorrelations: Property distribution by bond distance

3-D descriptors
- Volume
- Density
- Radius of gyration
- Principal moment of inertia

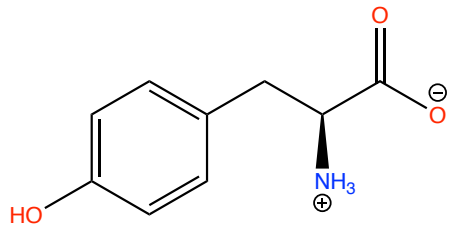
- 3-D autocorrelations: Property distribution by Euclidian distance

4-D descriptors
- Determined by conformational ensemble of molecule

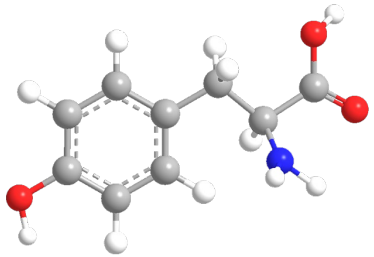
- Grid occupancy of lower-dimensional descriptors

$C_9H_8O_3N$

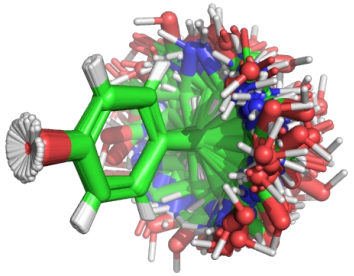
Number of atom types



Connectivity (+ stereochemistry)



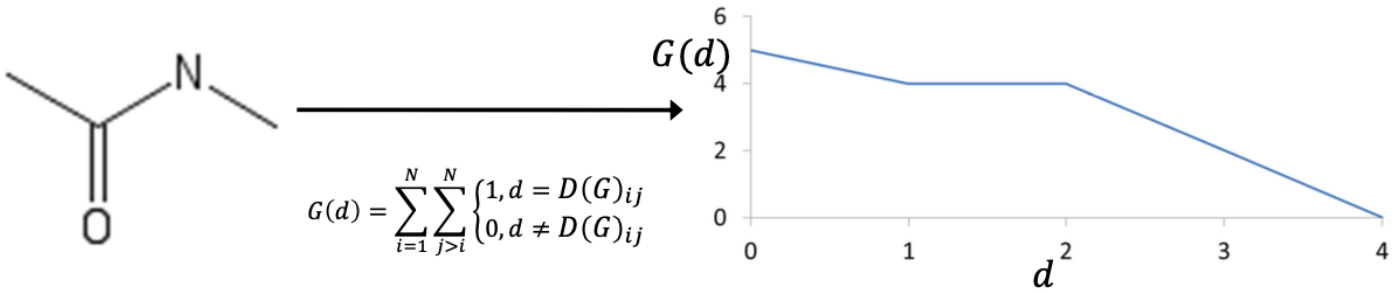
Connectivity with spatial info



Dynamic information

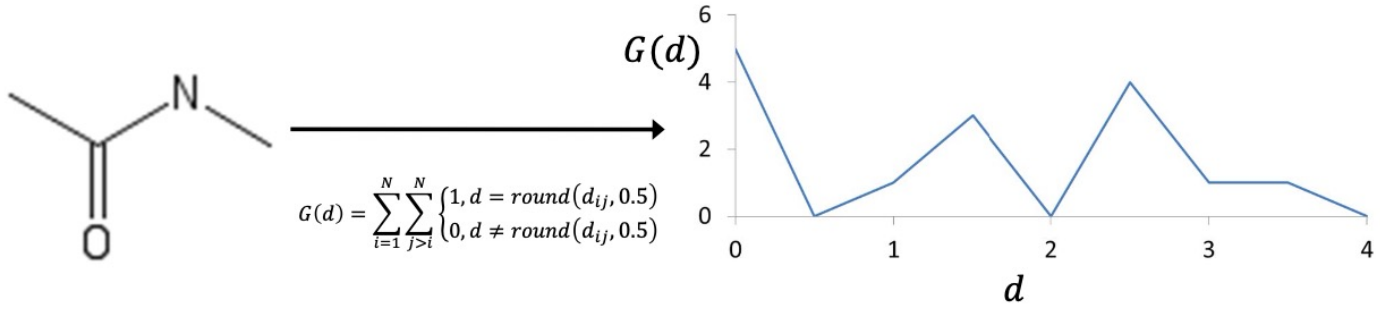
2D and 3D autocorrelation descriptors

Distribution of number of bonds



	C	C	N	C	O
C	0	1	2	3	2
C	1	0	1	2	1
N	2	1	0	1	2
C	3	2	1	0	3
O	2	1	2	3	0

Distribution of atom pair distances

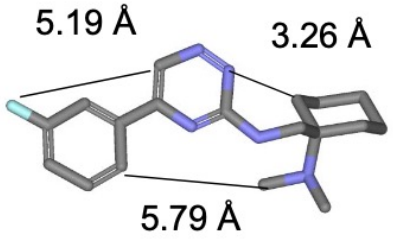


	C	C	N	C	O
C	0.00	1.54	2.5*	3.5*	2.5*
C	1.54	0.00	1.47	2.5*	1.23
N	2.5*	1.47	0.00	1.47	2.5*
C	3.5*	2.5*	1.47	0.00	3.0*
O	2.5*	1.23	2.5*	3.0*	0.00
					* - estimates

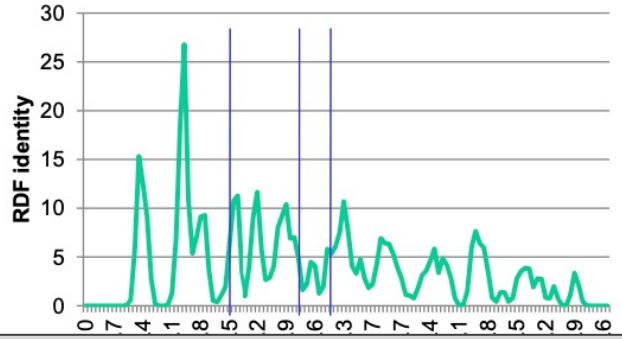
Radial distribution functions describe 3D distribution of descriptors

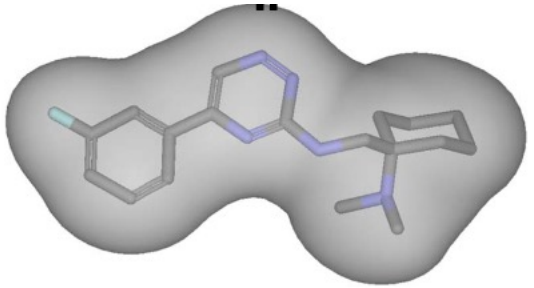
What is the 3D distribution of…

Atoms?



$$g(d) = \sum_{\substack{atom \\ pairs}} e^{-B(d-d_{ij})^2}$$

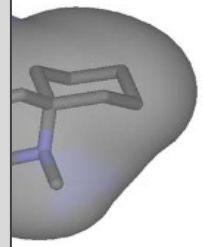




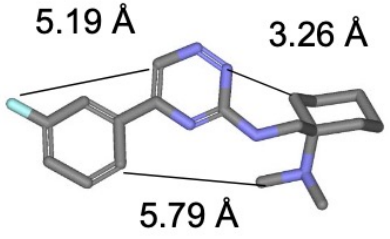
Lone pair electrons?

Bottom line:

1. **These descriptors can get complicated**
2. **These are <u>really</u> useful descriptors…why do you think that is?**

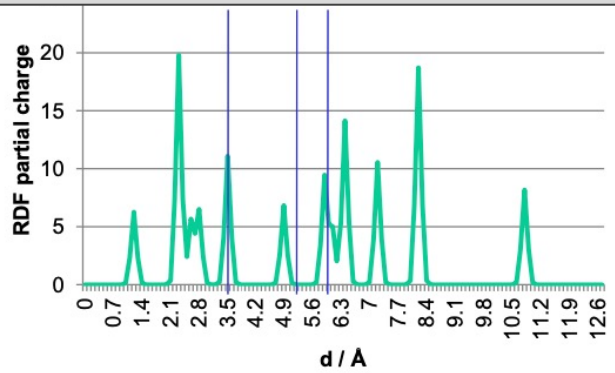


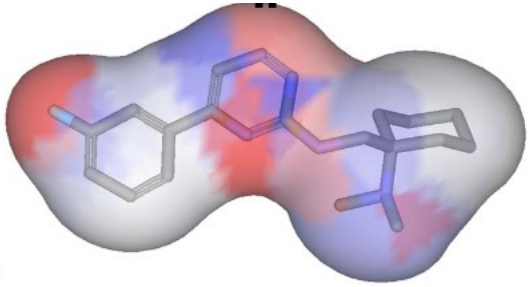
Partial charges?



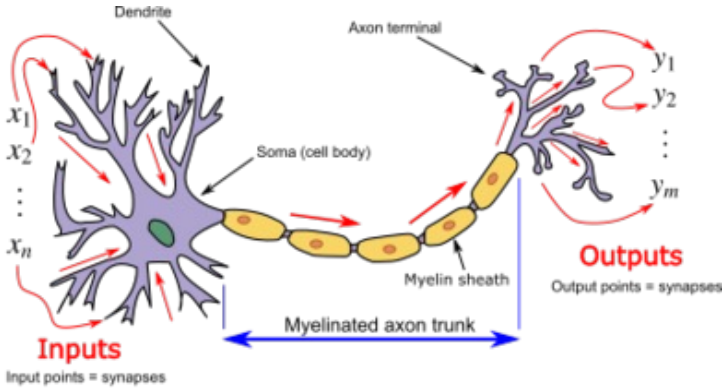
Gilson (2003)

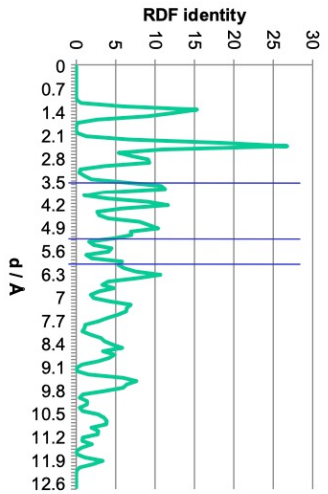
$$g(d) = \sum_{\substack{atom \\ pairs \\ i,j}} A_i A_j \, e^{-B(d-d_{ij})^2}$$



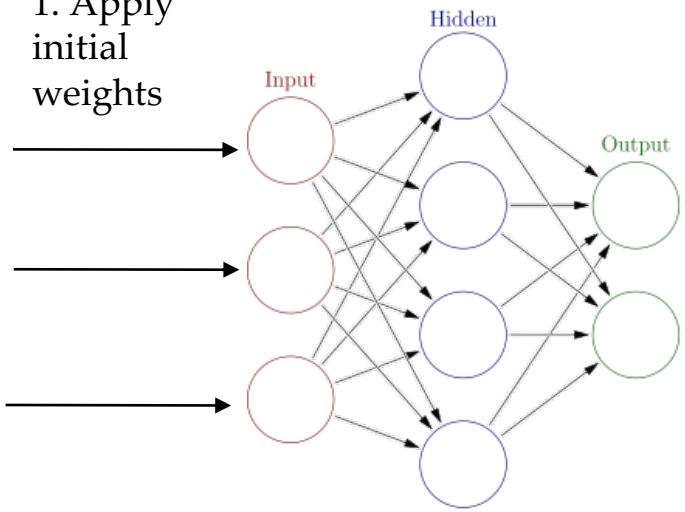


Artificial Neural Networks (ANNs)

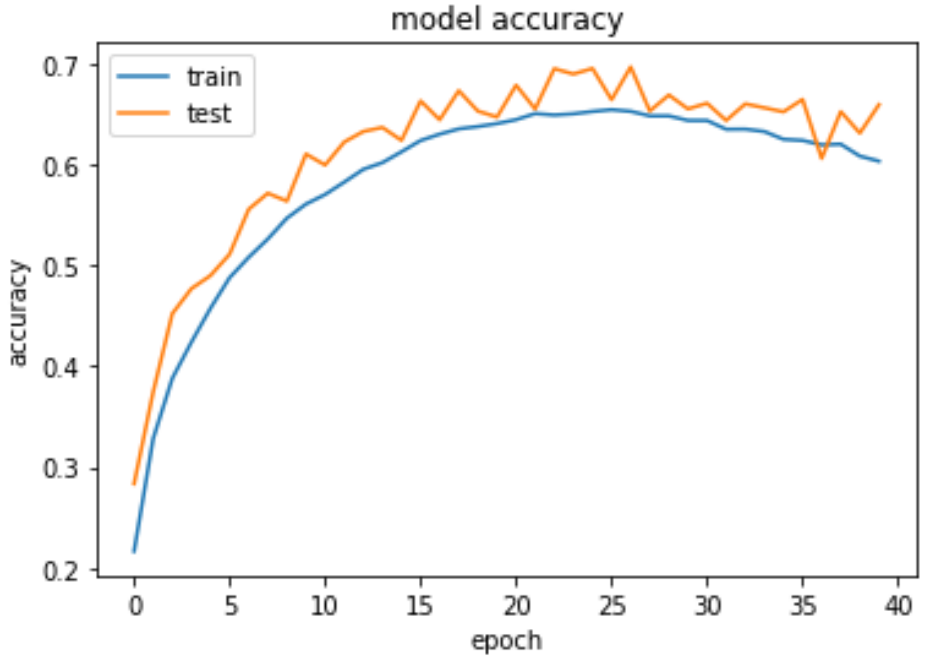




1. Apply initial weights

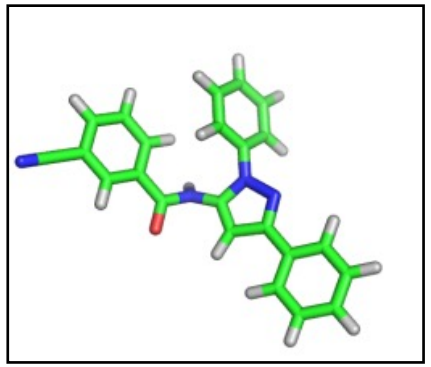


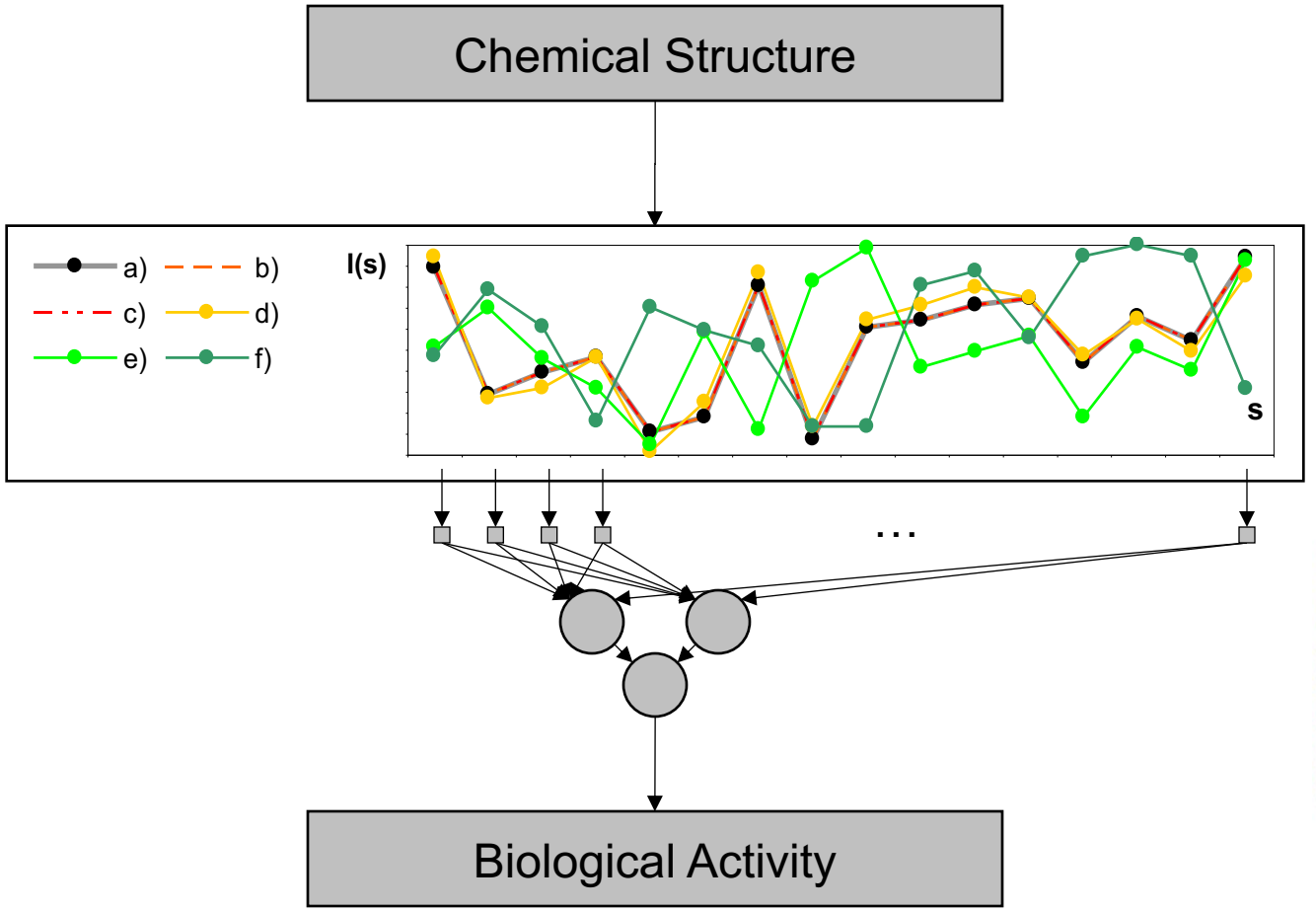
2. Apply weights
3. Neuron activated?
4. Complete task
5. Calculate error
5. Adjust accordingly and try again

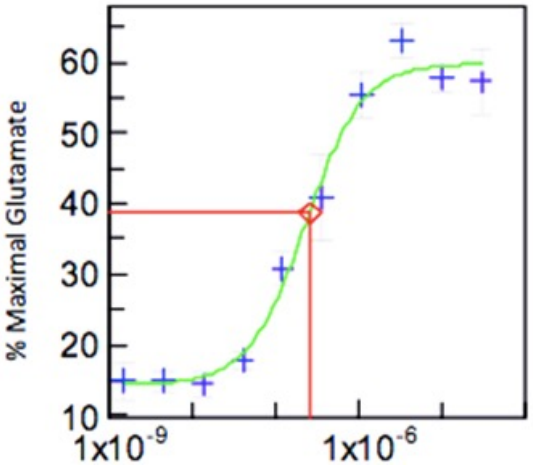


https://en.wikipedia.org/wiki/Artificial_neural_network

Artificial Neural Networks (ANNs) predicted biological activity from compound descriptors



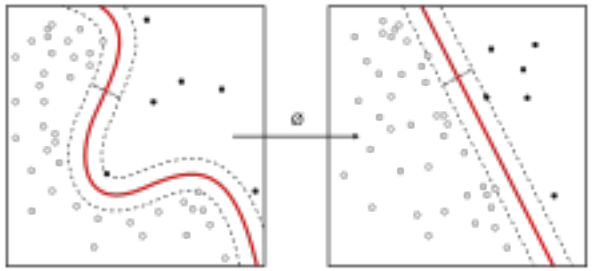


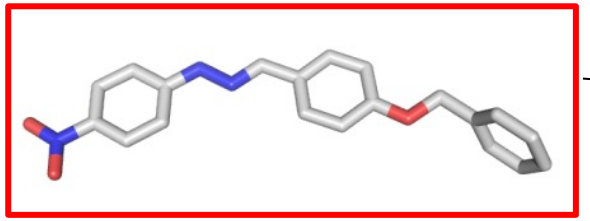


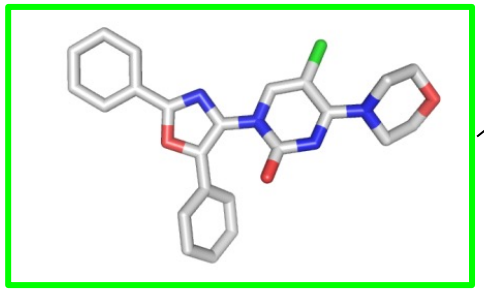
Mapping descriptor space into hyperspace

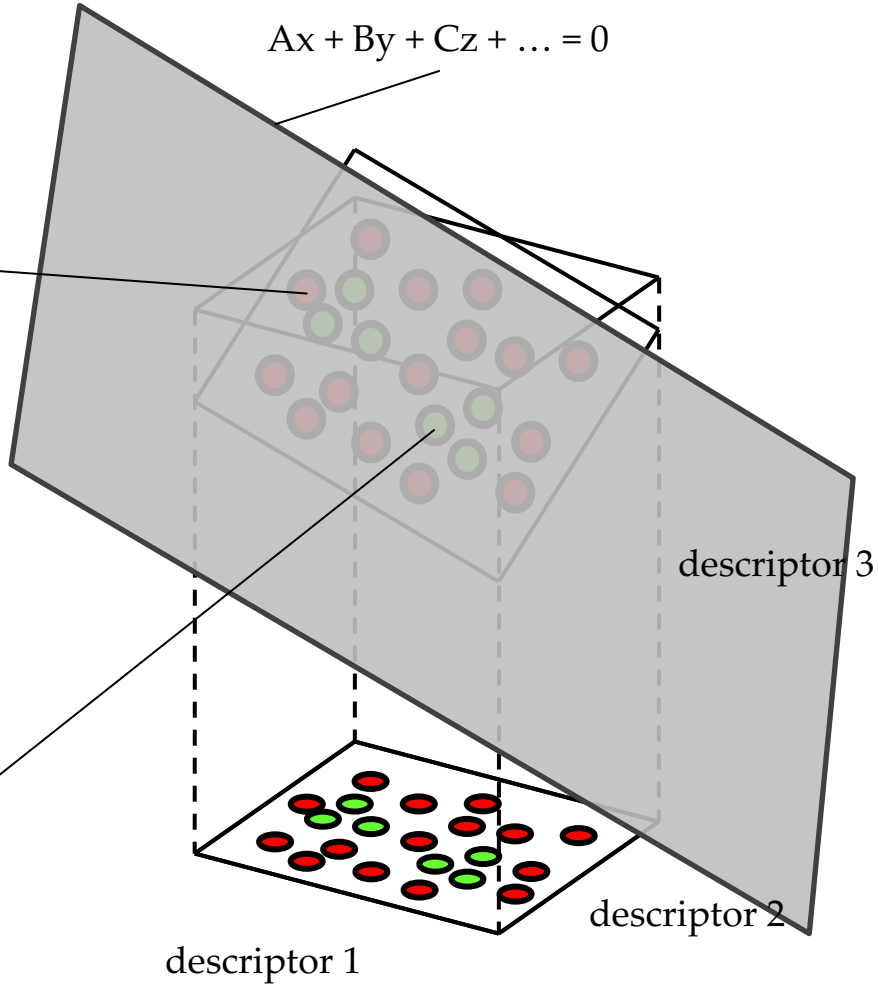
Which descriptors, and their weights, are important to separate actives and inactives?

- Classic support vector machine (SVM) problem
- Works better with small sample sizes
- Easier interpretability of results: you can figure out most important features









https://en.wikipedia.org/wiki/Support-vector_machine

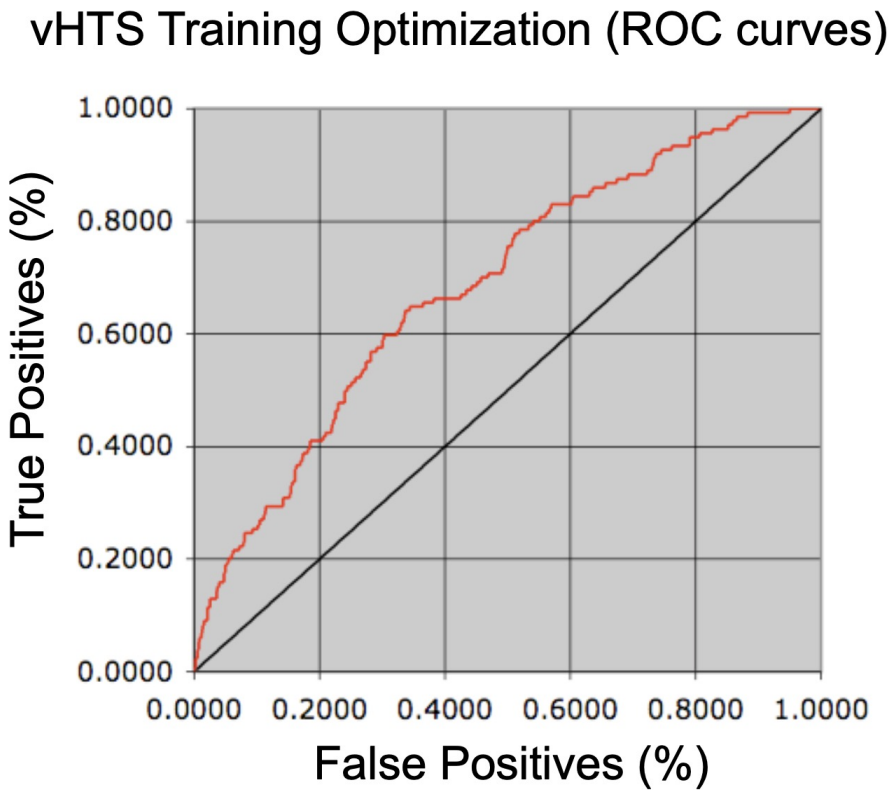
Optimizing descriptors for given test set

The more generalizable, the better! Less over-fitting.

Number descriptors

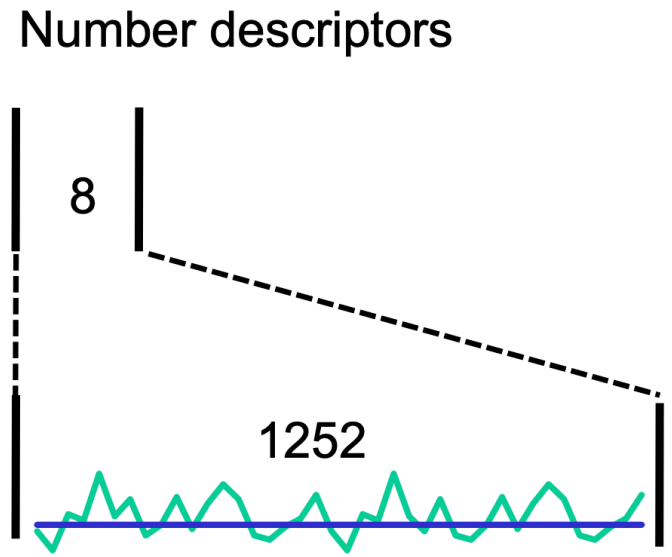
| 8 |

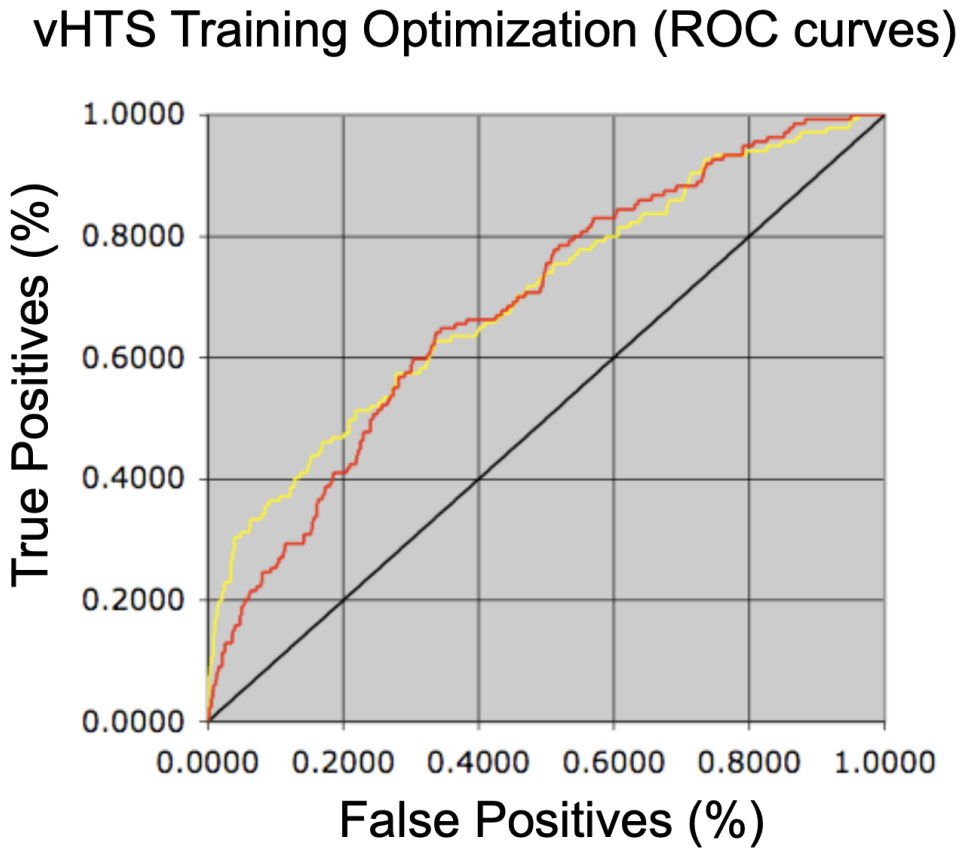
- Molecular Weight
- Number H bond donors
- Number H bond acceptors
- XlogP
- Polar surface area
- Mean molecular polarizability
- Molecular dipole moment
- Aqueous solubility



Optimizing descriptors for given test set

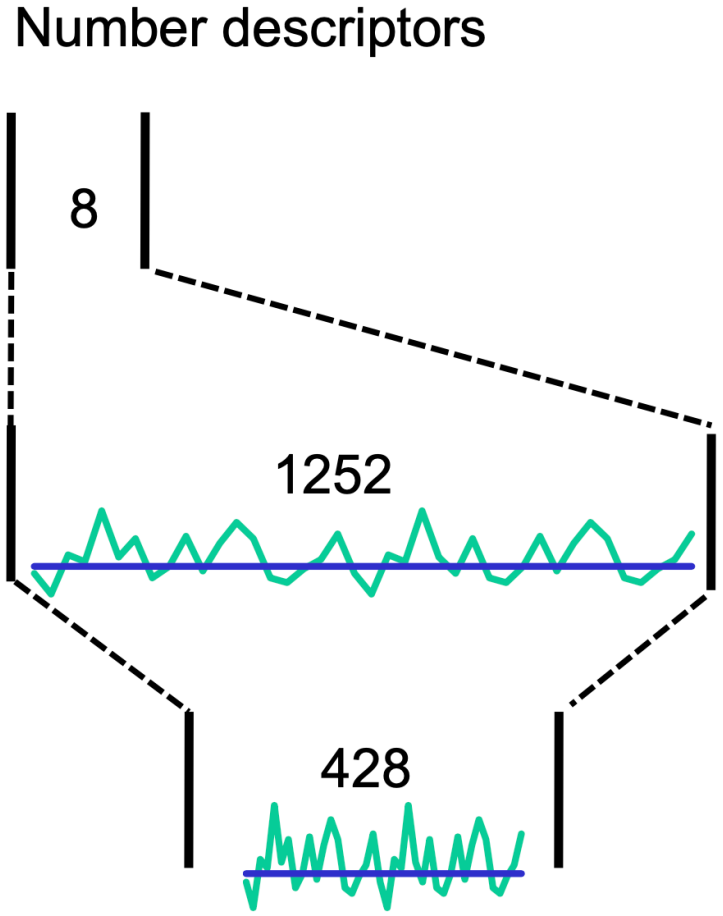
The more generalizable, the better! Less over-fitting.

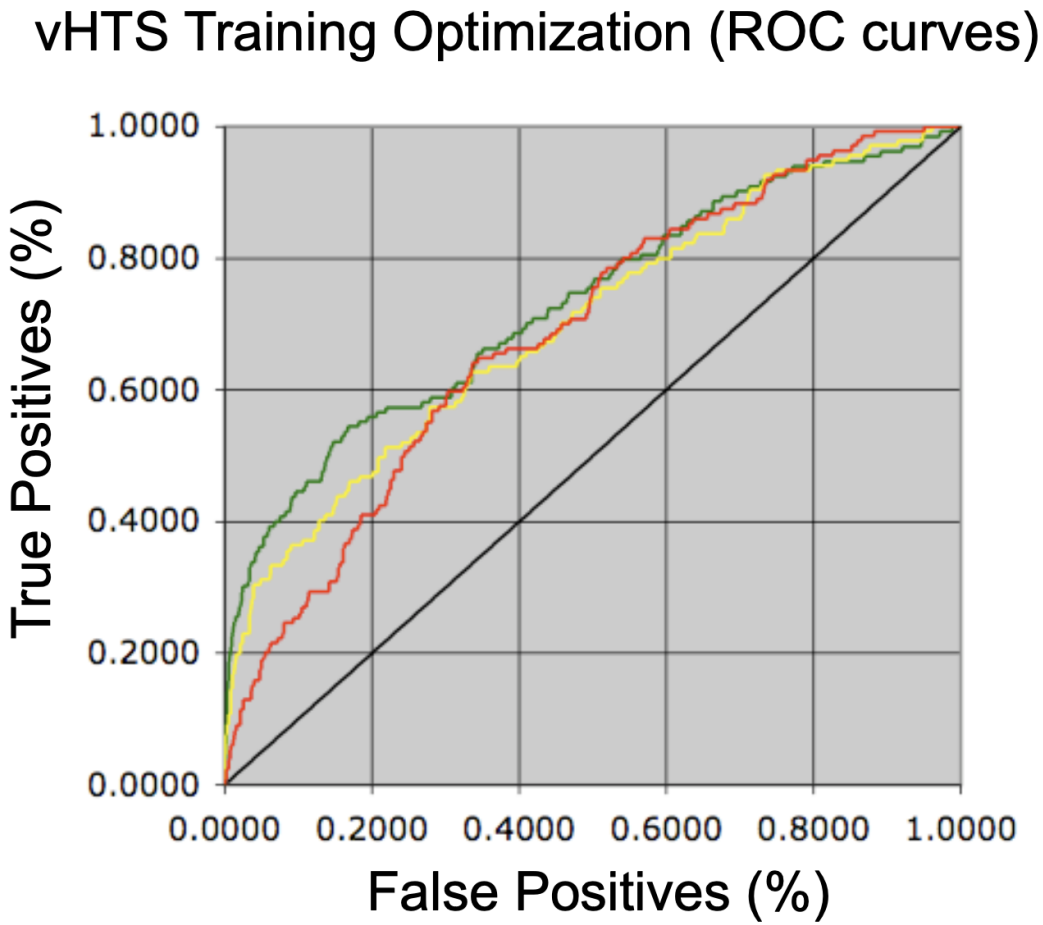




Optimizing descriptors for given test set

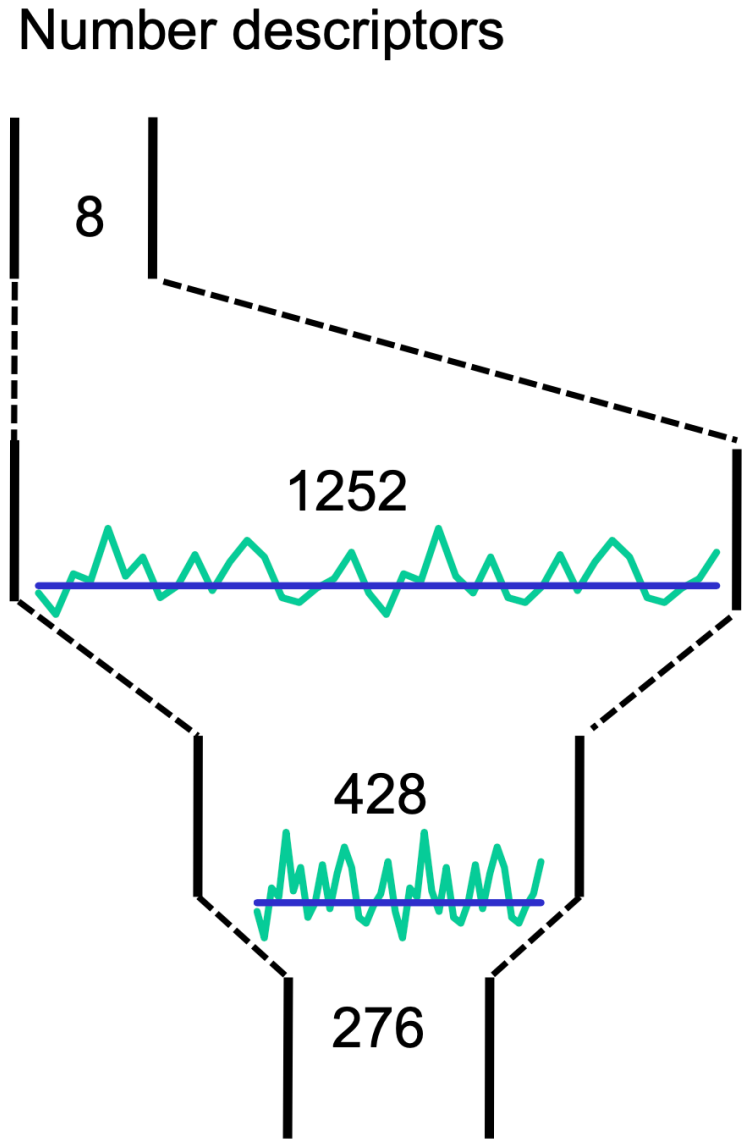
The more generalizable, the better! Less over-fitting.

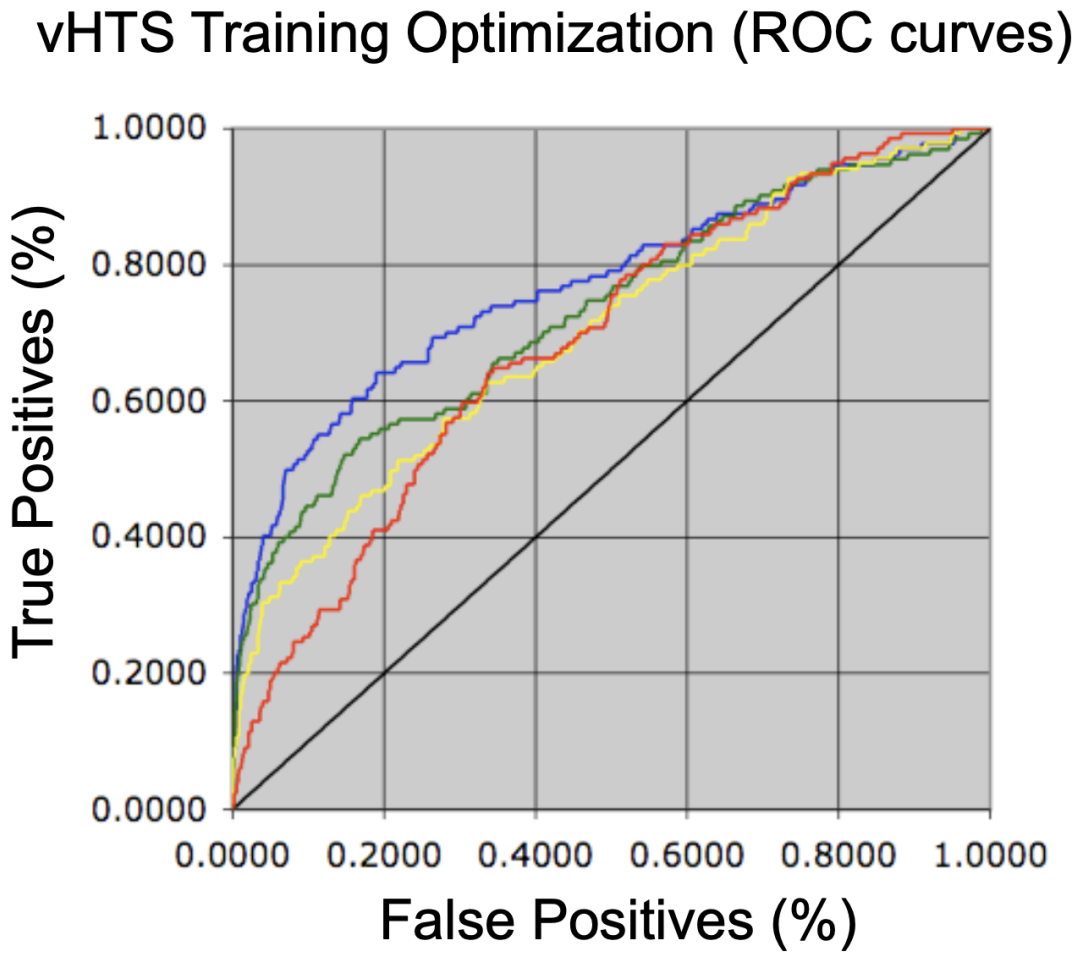




Optimizing descriptors for given test set

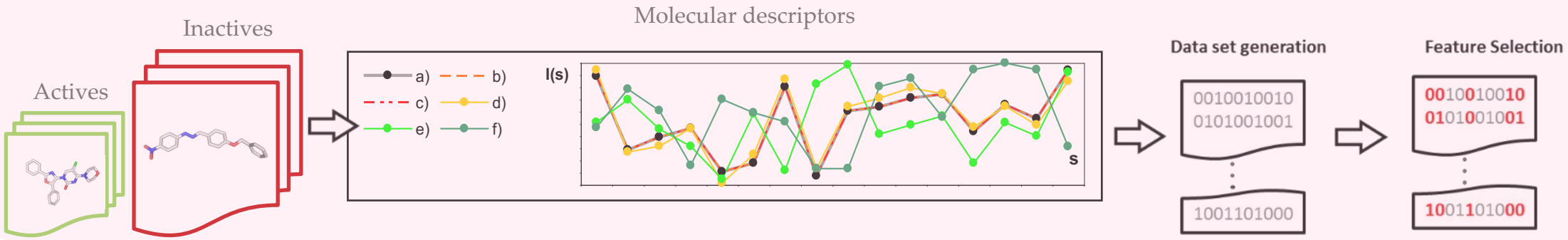
The more generalizable, the better! Less over-fitting.



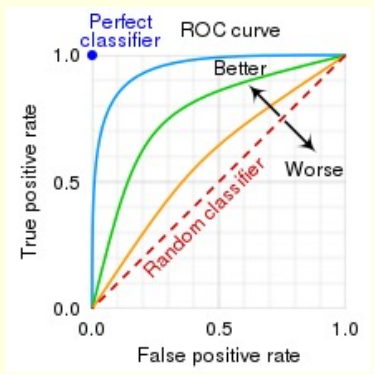


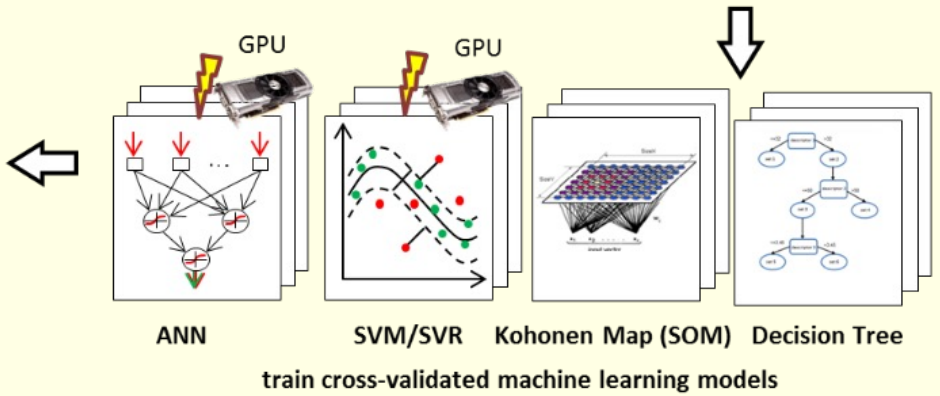
Putting all this together → LB-vHTS pipeline

Dataset acquisition

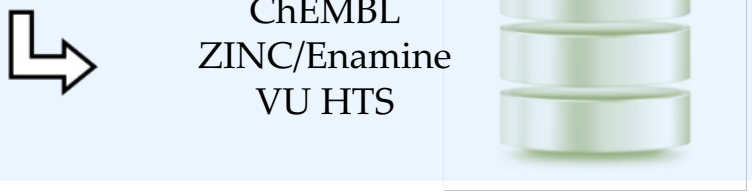


Training and validation





Prediction



ChEMBL
ZINC/Enamine
VU HTS

Rank compounds based on model score for testing

J. Mendenhall and J. Meiler; "Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout"; *J Comput Aided Mol Des*; **2016**.

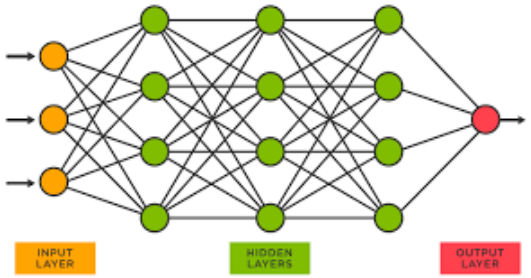
ANN QSAR identifies $mGluR_5$ allosteric modulators

150,000 compounds were experimentally tested for allosteric modulation of $mGluR_5$ measuring receptor-induced intracellular release of calcium

- 1,387 (0.94%) compounds were verified as $mGluR_5$ PAMs
- 345 (0.23%) compounds were verified as $mGluR_5$ NAMs

Niswender, C. M.; Johnson, K. A.; Luo, Q.; Ayala, J. E.; Kim, C.; Conn, P. J.; Weaver, C. D. *Mol Pharmacol* ***2008, 73, 1213-24.***

~750,000 compounds (ChemBridge) virtually screened with ANN QSAR for allosteric modulation of $mGluR_5$



- 749 compounds predicted EC50 < 10uM
- 12 novel NAM scaffolds

Enrichment = 3.6% / 0.23% = 16

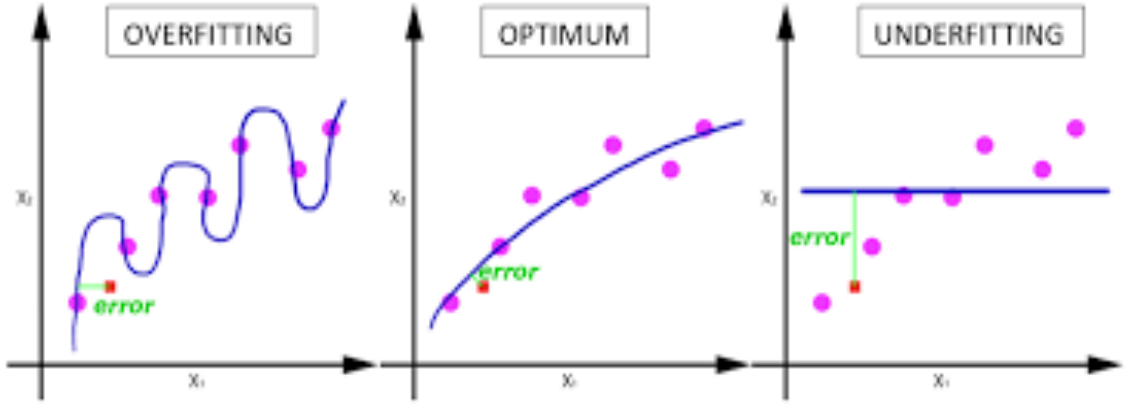
R. Mueller, E. S. Dawson, J. Meiler, A. L. Rodriguez, B. A. Chauder, B. S. Bates, A. S. Felts, J. P. Lamb, U. N. Menon, S. B. Jadhav, A. S. Kane, C. K. Jones, K. J. Gregory, C. M. Niswender, P. J. Conn, C. M. Olsen, D. G. Winder, K. A. Emmitte and C. W. Lindsley. *ChemMedChem*; **2012**; Vol. 7 (3): p. 406-14.

A couple things to be weary of with QSAR

- Less confidence in predictions for molecules further in chemical space from your test set
  - We have information of only a tiny fraction of available compound space
  - Nobody likes to report inactive molecules
  - And the actives are usually chemically similar

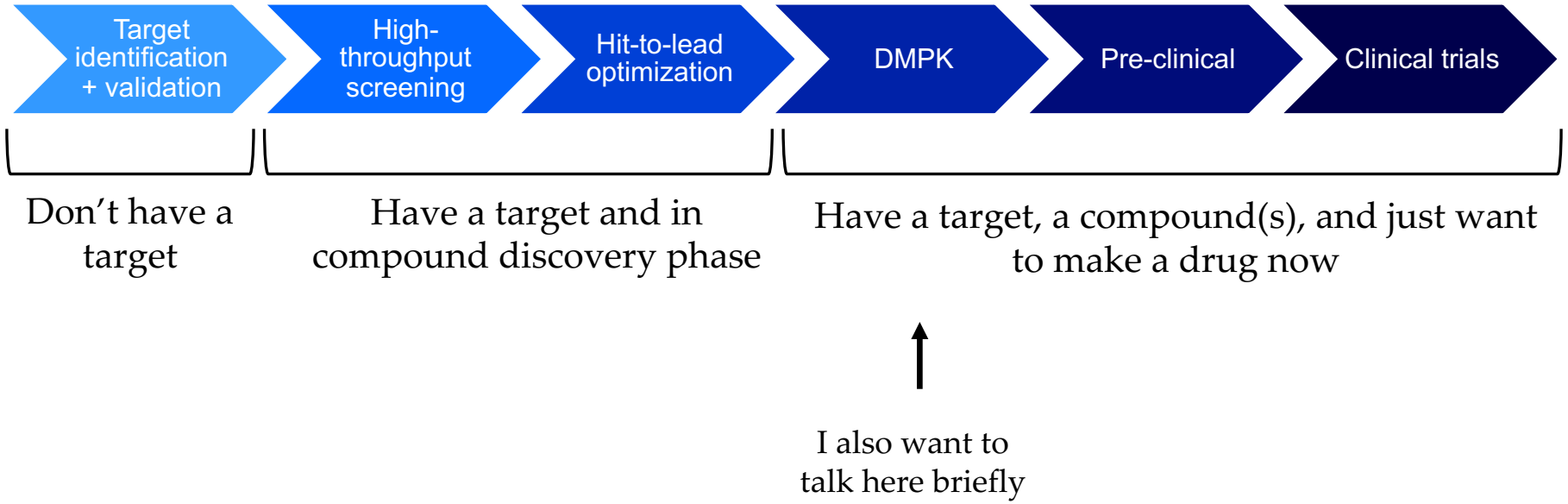
- Heavily biased dataset composition
  - Many descriptors and only a few outcomes
  - Over-fitting if not careful

- Can be a black-box and tough to interpret



You simply memorized the actives

You really didn't learn anything


Target identification + validation
High-throughput screening
Hit-to-lead optimization
DMPK
Pre-clinical
Clinical trials
Don't have a target
Have a target and in compound discovery phase
Have a target, a compound(s), and just want to make a drug now
I also want to talk here briefly

“It’s relatively easy to discovery a potent ligand—it’s damn tough to discover a drug” – E.H. Cordes

90% of drug candidates that get IND fail in CTs

Causes for attrition	
Animal tox/safety	20%
ADME	41%
Lack of efficacy	27%
Marketing consideration	7%
Formulation	4%
Miscellaneous	20%

	Antibiotic	CNS	Cardiovascular	Respiratory
Failure due to **safety** (Phase 1)	85-100%	20-30%	30-40%	50%
Failure due to **efficacy** (Phase 2/3)	5-15%	70%	60%	50%

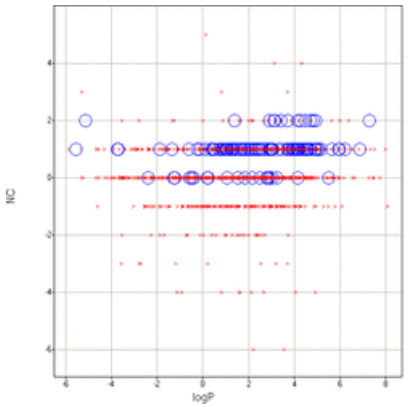
Failure is industry-wide
Failure is indication-wide
Reasons for the failures vary between indications, but NOT between companies

What do we use now?
Lipinski’s Rule of 5 (sort of)

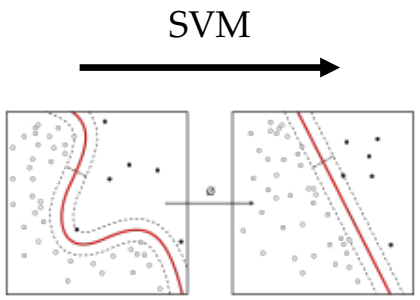
Can we figure out which molecular features don’t pass tox?

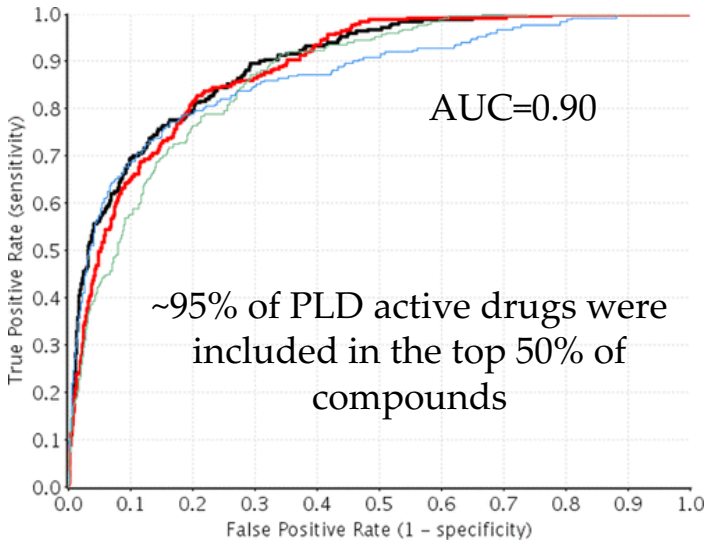
Drug-induced phospholipidosis (PLD) prediction using SVMs

- PLD is a phospholipid storage disorder characterized by excessive accumulation of intracellular phospholipids is a variety of tissue types
- Many drugs can accumulate in tissues by forming complexes with the polar phospholipids → "drug-induced PLD"
- Potentially toxic and can halt drug development in later stages
- **Can we identify these compounds beforehand? Let's fail earlier!**



Commonly-used descriptors for this task do not separate PLD inducing drugs





Most important features:

1. N16 (a positively charged nitrogen atom in a saturated ring
2. C3 (an aromatic carbon atom with no substitution and adjacent to two aromatic carbon atoms)
3. H2 (a hydrogen bonded to an aromatic carbon)
4. M12 (a number of aromatic rings)
5. S5 (sulfur in a ring bonded to two aromatic atoms)

J. Chem. Inf. Model. 2012, 52, 7, 1798–1805

Drug discovery: a crude history

pre-1800s | 1800s- early 1900s | 1900s | 2000s

Opium, ~3400 BCE

Coca leaves

Morphine **isolated** from opium (1805)

Chloral hydrate **synthesized** as synthetic sleeping drug (1832)

Arsphenamine/Compound 606 **synthesized for testing** for anti-syphilitic properties (1909)

Biological sciences:
molecular cloning, recombinant DNA technology, Human Genome Project, structural biology technologies

Chemical sciences:
Combinatorial chemistry

Computer science:
Uniprot, Protein Data Bank, AlphaFold, data storage/computational power

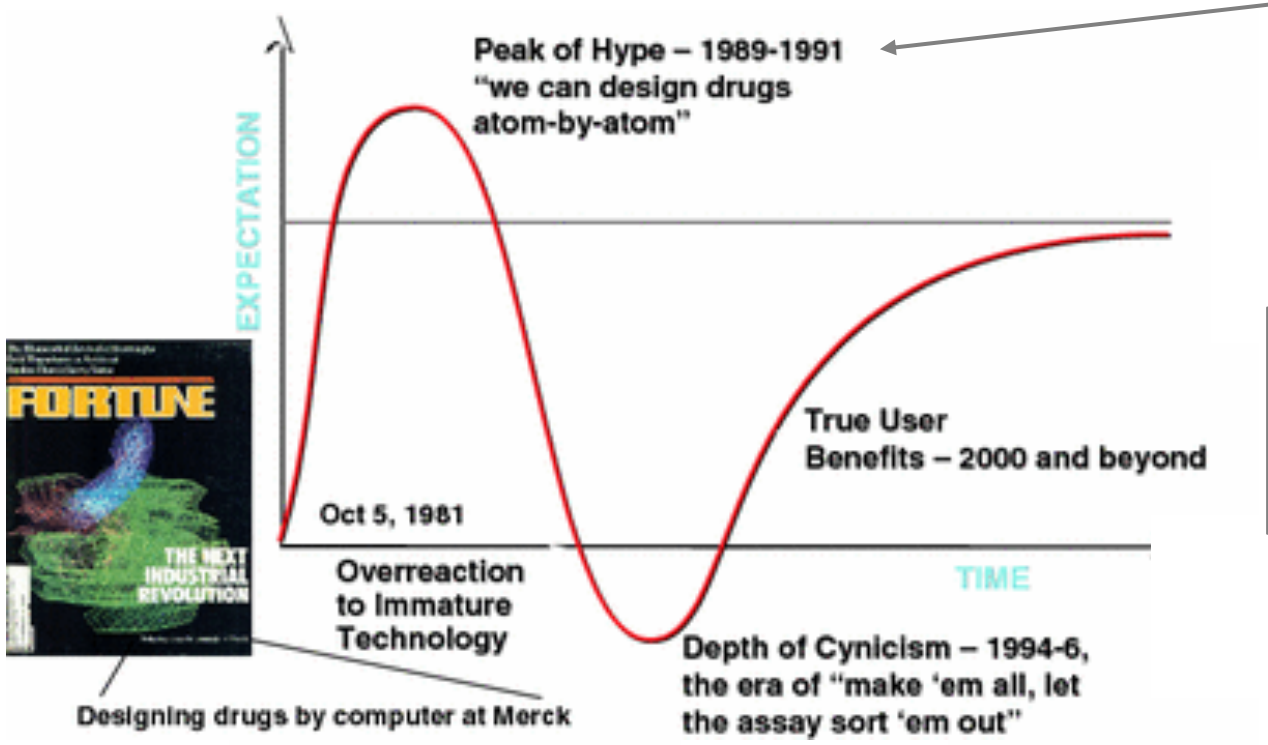
- MoA/target ID
- Screening technologies
- Rational-based design

- Computational tools in all of the above

Methods Mol Biol. 2009;572:3-12. doi: 10.1007/978-1-60761-244-5_1.

”Computer-aided drug design: the next 20 years” (2007)

Paper based on talk given at 2007 ACS meeting:
“The aim in this Perspectives is to provoke some thinking about what may lie ahead.”



Peter Goodford (1989)
“We must…

F 1. deal properly with water
C 2. remember that conformation depends on structure and environment
B 3. combine theory and experiment
D 4. predict solubility
A 5. improve homology modeling”

- (mid-1990s) HIV protease inhibitors via SB-CADD
- (1995): de novo ligand design of binders to dihydrofolate reductase and thymidylate synthase
- (1999) tirofiban by pharmacophore vHTS

https://link.springer.com/article/10.1007/s10822-007-9142-y

This is becoming more accessible

- Python sklearn is very well-documented and freely accessible—go try it!

Dr. Fiona Marshall Visit

Merck/MSD
Senior VP Head of Discovery, Preclinical and Translational Medicine

2018-2021:
- Senior VP Head of Discovery, Preclinical & Translational Medicine
- VP Head of Neuroscience and Head of MSD UK Discovery Research
- VP Head of Discovery Research MRL UK

Sosei Group
2015-2018: Executive VP and Chief Scientific Officer

Heptares Therapeutics
2006-2018: Chief Scientific Officer and Founder

Tuesday, May 17th from 2:00-5:00pm

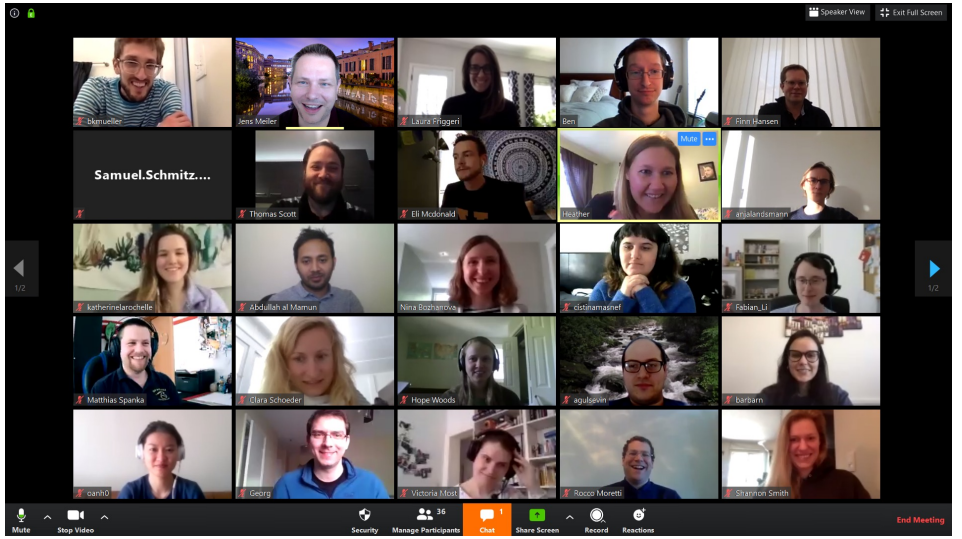
CPB Retreat Speaker on May 18th
9:00 in first floor class ESB

Please talk to me or email for details
(shannon.t.smith.1@vanderbilt.edu)

Acknowledgements

Meiler Lab
Jens Meiler
Ben Brown
Rocco Moretti
Oanh Vu
Jeffrey Mendenhall

Funding



You can throw the kitchen sink at your problem

- Python sklearn is very well-documented and freely accessible—go try it!

```
from sklearn.neural_network import MLPClassifier
from sklearn.neighbors import KNeighborsClassifier
from sklearn.svm import SVC
from sklearn.gaussian_process import GaussianProcessClassifier
from sklearn.gaussian_process.kernels import RBF
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier, AdaBoostClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.discriminant_analysis import QuadraticDiscriminantAnalysis

    Classifiers = [
    KNeighborsClassifier(3),
    SVC(kernel="linear", C=0.025),
    SVC(gamma=2, C=1),
    GaussianProcessClassifier(1.0 * RBF(1.0)),
    DecisionTreeClassifier(max_depth=5),
    RandomForestClassifier(max_depth=5, n_estimators=20, max_features='auto'),
    MLPClassifier(alpha=1, max_iter=1000),
    AdaBoostClassifier(),
    GaussianNB(),
    QuadraticDiscriminantAnalysis()]
```