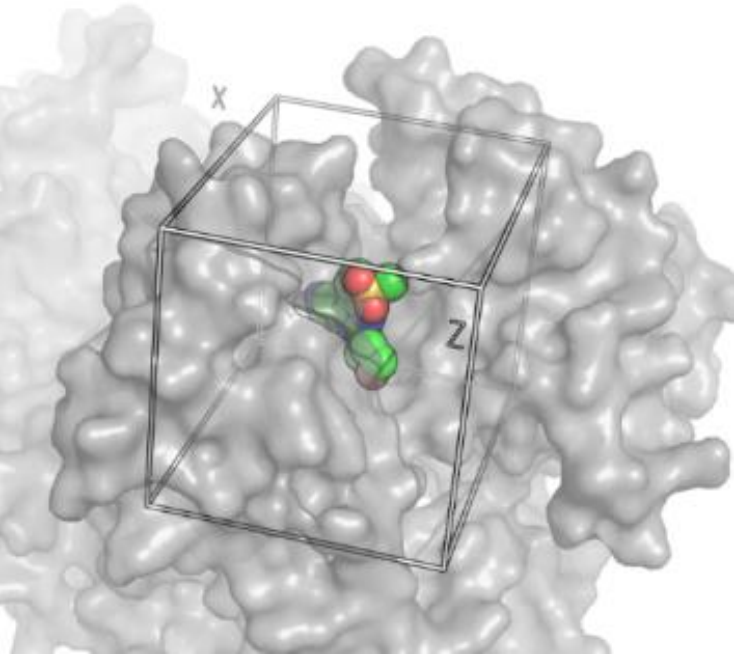
AI Applications

Exploring a billion chemicals for drug discovery

Anthony Gitter

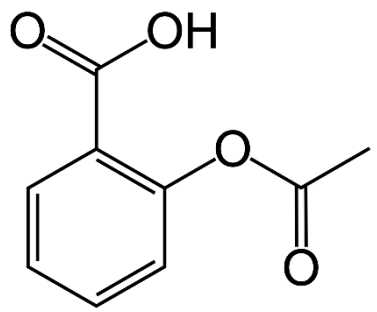
gitter@biostat.wisc.edu

University of Wisconsin-Madison



Chemicals as drugs (medicines)

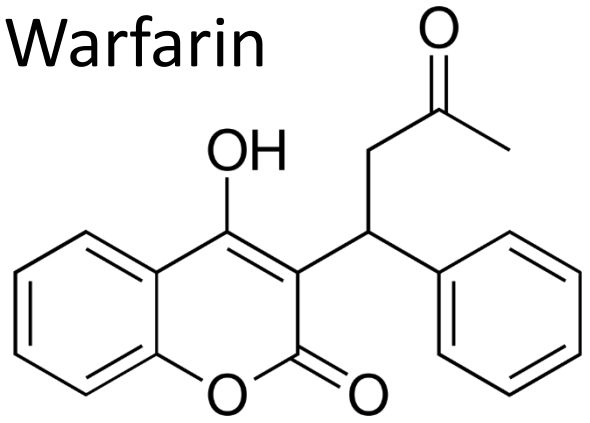
Aspirin



Wikipedia

➡ Reduce pain, inflammation

Warfarin



Wikipedia

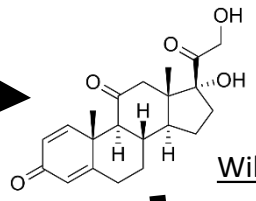
➡ Treat blood clots

??? ➡ New disease

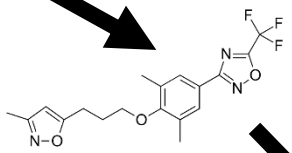
Standard chemical screening

Drug discovery

Amazon



Wikipedia



Wikipedia

Ineffective

Effective

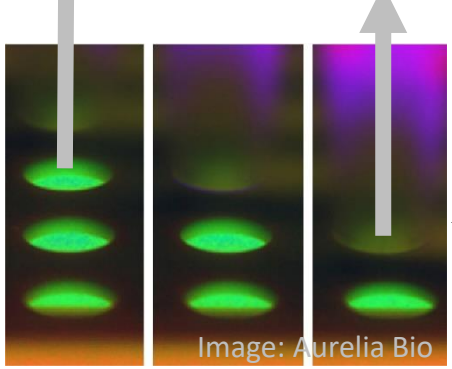


Image: Aurelia Bio

Image: Norah Trent

Culinary

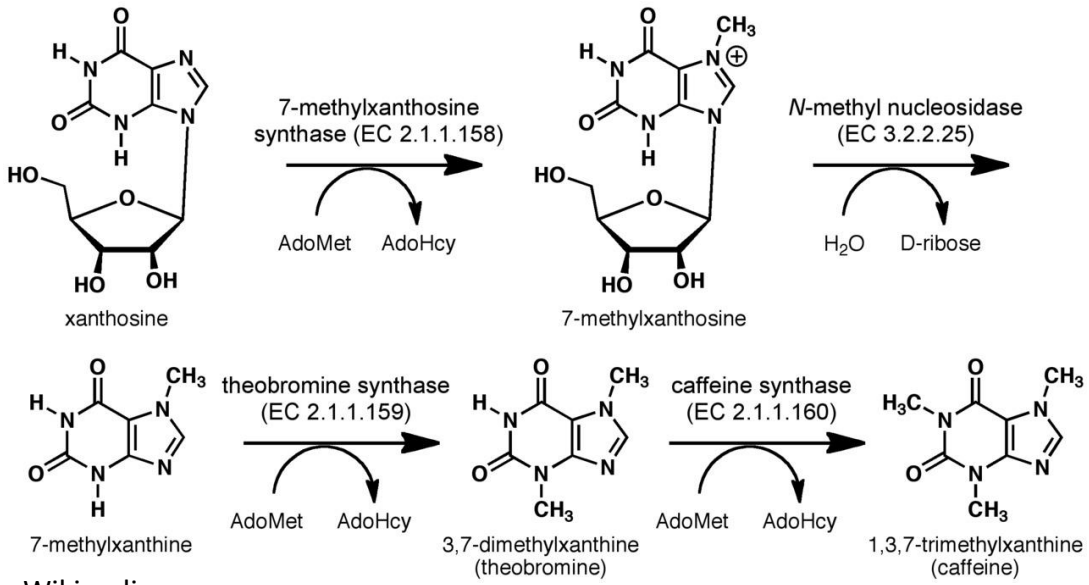
Wikipedia

Wikipedia

Instagram

Exploring chemical recipes

Drug discovery



Wikipedia

Culinary

Wikipedia

Serious Eats: Stephanie Cameron

Recipes for > 1 billion chemicals

Chemical synthesis

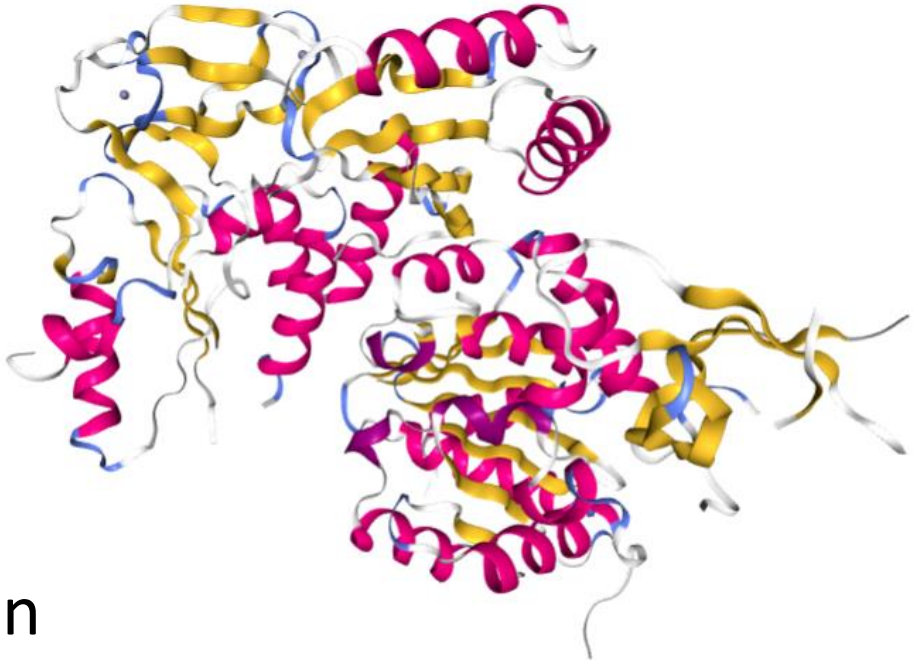
Machine learning to guide

French	-	Oats	-
Ethiopian	+	Chicken	+
Vietnamese	?	Saffron	?

Drug discovery case study: PriA-SSB

Image: Kateryna Kon

PriA bound to SSB (PDB:4NL8)



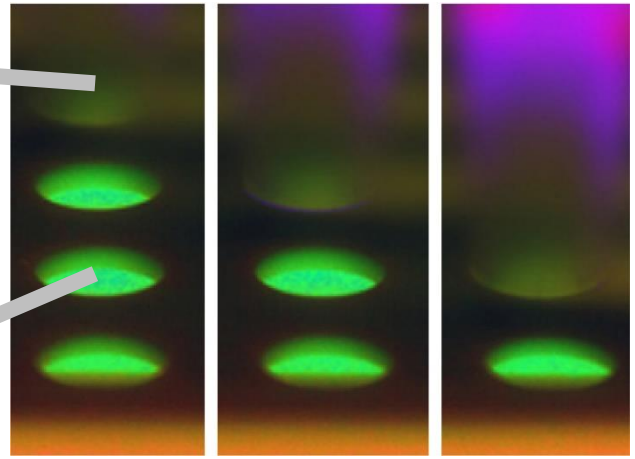
- *Klebsiella pneumonia*
- Bacterial protein-protein interaction
- DNA repair, recombination, and replication

Goal: block interaction, kill bacteria

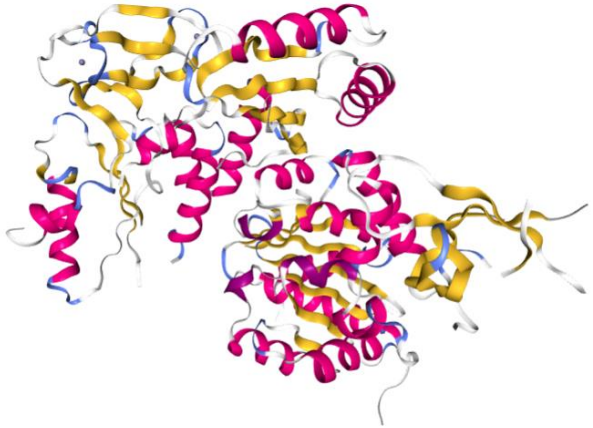
Computational models prioritize chemicals

PriA-SSB binding assay

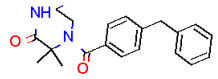
Effective
→ blocked interaction
→ active or “hit”

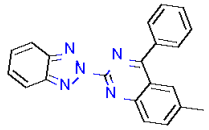


Ineffective
→ proteins bind
→ inactive

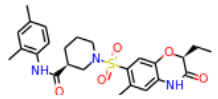


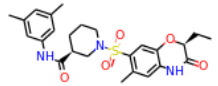
Test some initial chemicals

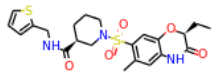


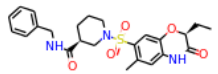


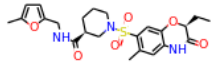
Prioritize remaining chemicals

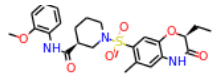
1 

2 

3 

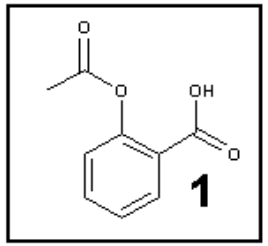
4 

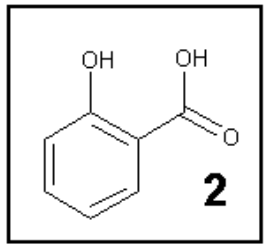
5 

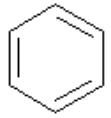
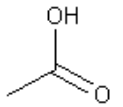
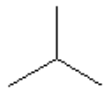
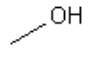
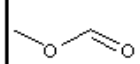
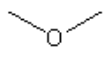
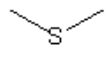
6 

Formulating a classification problem

Active Inactive





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

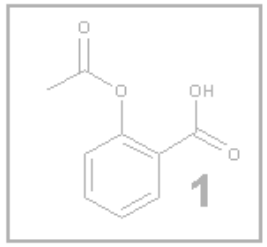
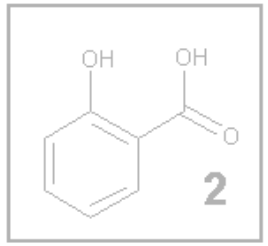
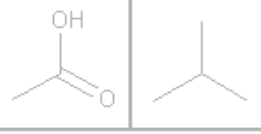
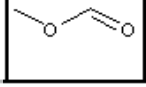
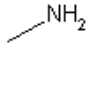
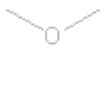
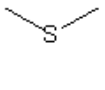
Chemical fingerprints

2-D searching tutorial

What distinguishes the active and inactive compounds?

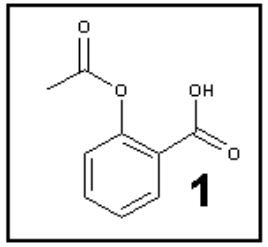
Example classifier: decision tree

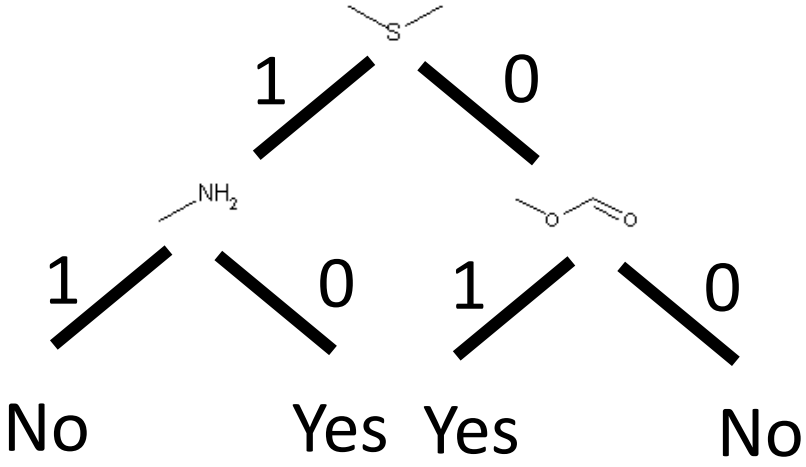
Are these features present?

	Active	Inactive						
								
1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0
								

2-D searching tutorial

Is compound active?

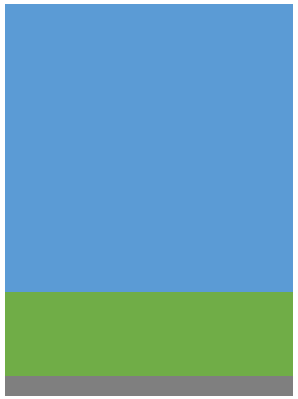




Activity prediction

Combine many decision trees into random forest

Evaluating prospective performance



Train on 72k chemicals, PriA-SSB inhibition labels

Only 79 active chemicals

Choose among many possible models, cross-validation

Best models select 250 of 22k new chemicals

Compare many types of computational models

Protein structure-based

- 9 docking scores
- Consensus docking

(Ericksen et al. JCIM 2017 DOI:10.1021/acs.jcim.7b00153)

Chemical-based

- Single-task supervised learning
- Multi-task supervised learning
- Chemical similarity baseline

Random forest performs best in prospective screen

Model	Actives	Actives not in baseline	SIM clusters	MCS clusters
Experimental	**54**	--	27	35
Similarity baseline	31	--	14	17
Consensus docking	0	0	0	0
STNN-C	21	2	11	13
STNN-R	28	8	14	18
MTNN-C	27	7	12	17
LSTM	1	1	1	1
Random forest	**37**	**7**	**17**	**22**
IRV	29	4	15	18

Only 54 experimental actives

Random forest is best overall

Models select 250 of 22k new chemicals

Available chemical libraries

Life Chemicals

10^5 chemicals
On campus
Previous study

NIH's MLPCN

3×10^5 chemicals
On campus

Aldrich Market Select

8×10^6 chemicals
Commercial

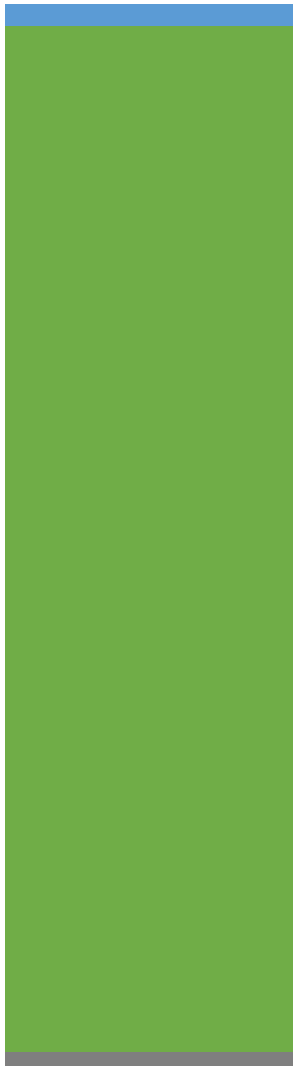
Can we train a model on these…

…to select chemicals from these

Enamine REAL database

10^9 synthesizable chemicals
Commercial

Prioritizing 1 billion chemicals: Enamine REAL



Random forest model
Trained on 427k chemicals, PriA-SSB inhibition labels

Predictions are fast: 18 jobs,
1 CPU and 6 GB RAM each,
mean runtime 53h

Random forest selects 100 of 1 billion new chemicals
Only 68 of the 100 can be synthesized successfully

Enamine REAL is a major success

31 of the **68** chemicals are hits!

Chemical selection	Library	Library size	Chemicals tested	Hits	Hit rate
Entire library	Life Chemicals + MLPCN	427,000	427,000	554	0.13%
Random forest	Enamine REAL	1 billion	68	31	46%

351 times better hit rate