Atomwise

Chimeric molecules as adversarial training examples for machine learning

April 15th, 2021

Jon Sorenson, Ph.D.

Head of Technology Development

Ultra-large libraries + deep learning transform drug discovery

In a library of **138M** molecules Lyu, *et al* estimated **72,000** distinct scaffolds bind to D_4 receptor with **<10μM** affinity

The Enamine catalog now contains **17B** on-demand molecules

Lyu, et al, Ultra-large library docking for discovering new chemotypes, *Nature*, **566**, 224 (2019)

Ultra-large libraries + deep learning transform drug discovery

In a library of **138M** molecules Lyu, *et al* estimated **72,000** distinct scaffolds bind to D_4 receptor with **<10μM** affinity

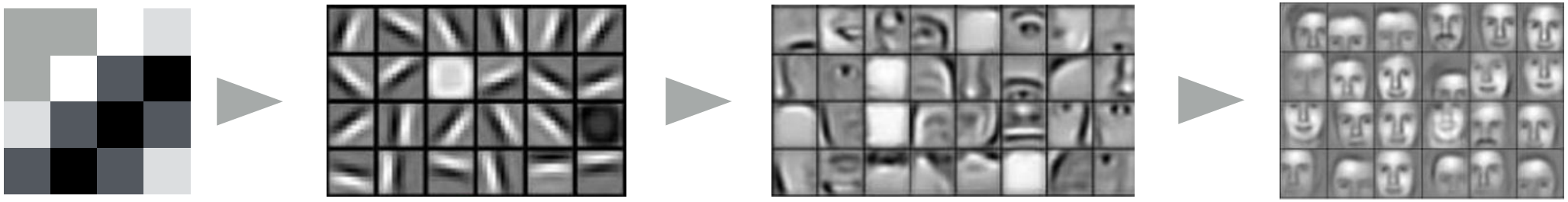
The Enamine catalog now contains **17B** on-demand molecules

How to find the stars in this galaxy that are potent, selective, and efficacious?

Lyu, et al, Ultra-large library docking for discovering new chemotypes, *Nature*, **566**, 224 (2019)

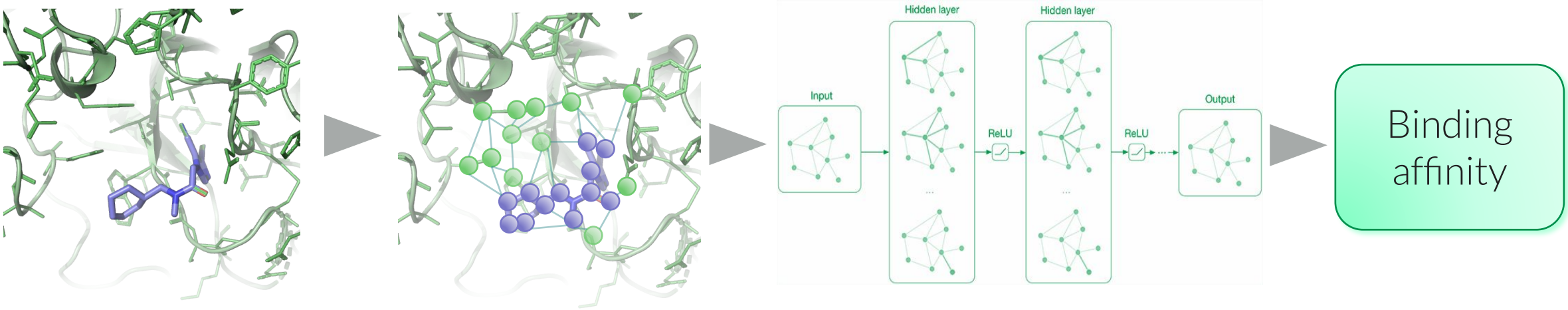
Structure-based drug design with deep learning

Convolutional neural networks for image recognition



Pixels | Edges | Eyes, Noses, Mouths | Faces

Convolutional neural networks for molecular recognition



Teaching AI to generalize wisely is hard

[Submitted on 6 Nov 2020 (v1), last revised 24 Nov 2020 (this version, v2)]

https://arxiv.org/abs/2011.03395

Underspecification Presents Challenges for Credibility in Modern Machine Learning

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, D. Sculley

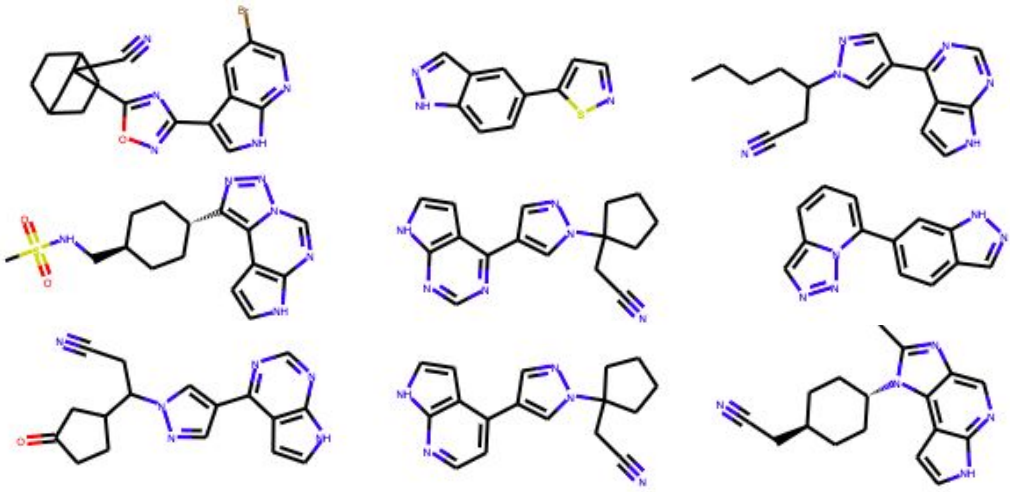
Underspecification

Models are optimized for performance on in-domain tasks and not sufficiently constrained to generalize well to novel inputs.

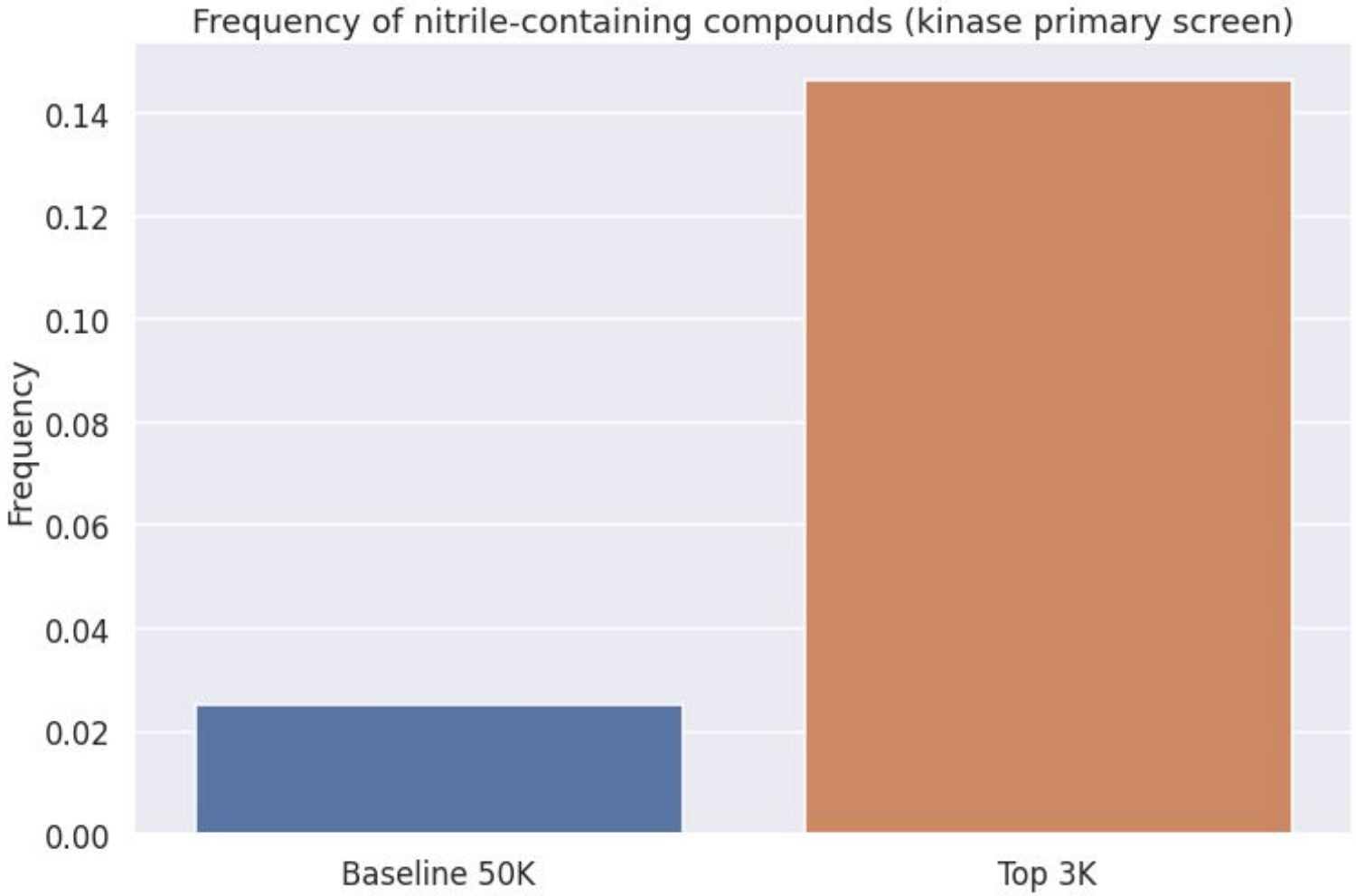
Solution

Supplement performance tests with stress tests to uncover cheating. Construct models and methods to beat these tests.

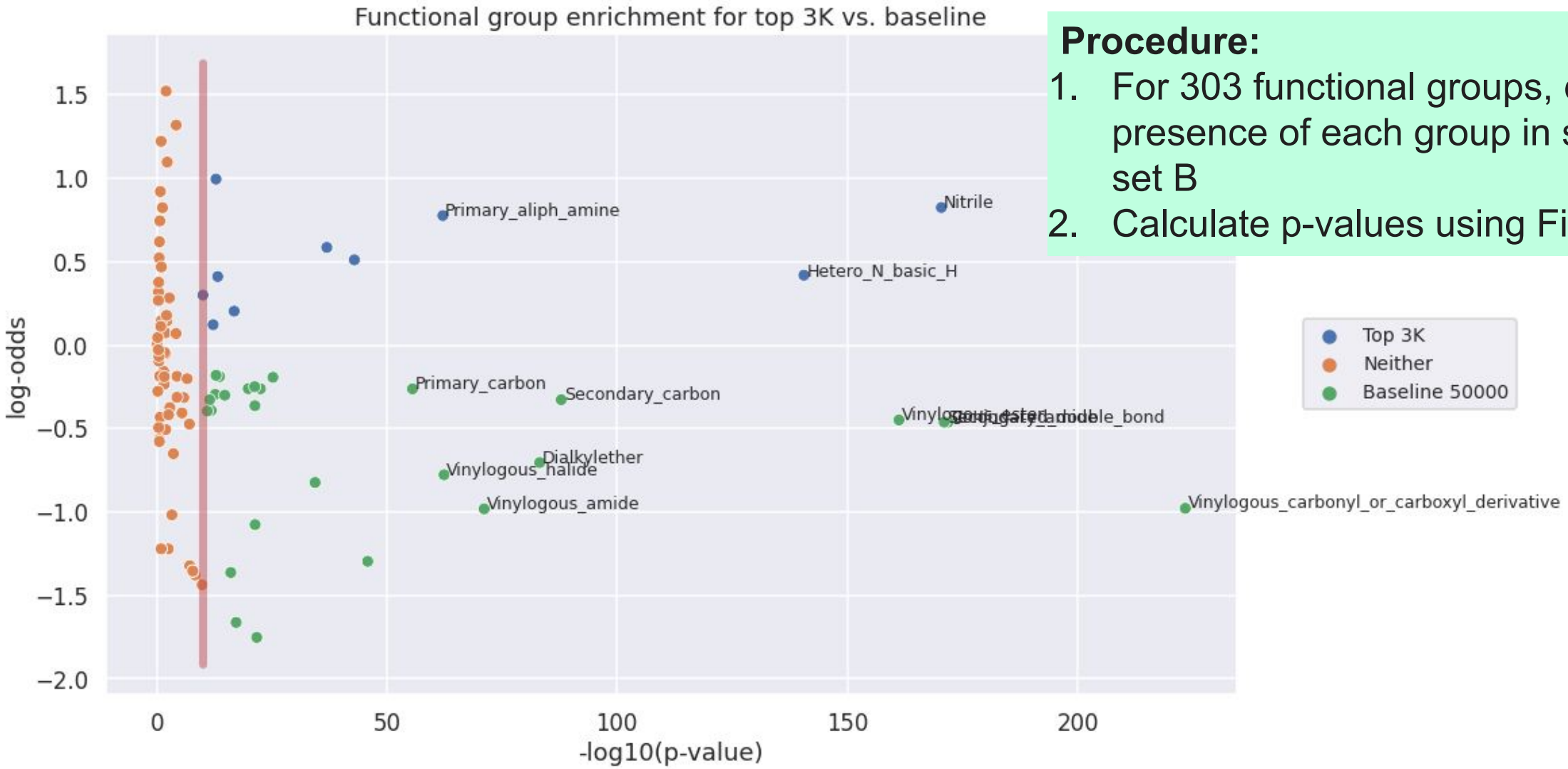
One way to cheat is to exploit functional group imbalances



Representative top predictions
Screening against the ATP-binding site of a protein kinase



Measuring enrichment using volcano plots



Procedure:

1. For 303 functional groups, count the presence of each group in set A and set B
2. Calculate p-values using Fisher's test

Constructing decoys to condition model learning

Functional group decoys

- For a particular active compound, create decoys that
  - Don’t share the overall structure of the compound

  — *but* —
  - Do have the same representation of functional groups

Require ECFP4/1024 similarity < 0.35

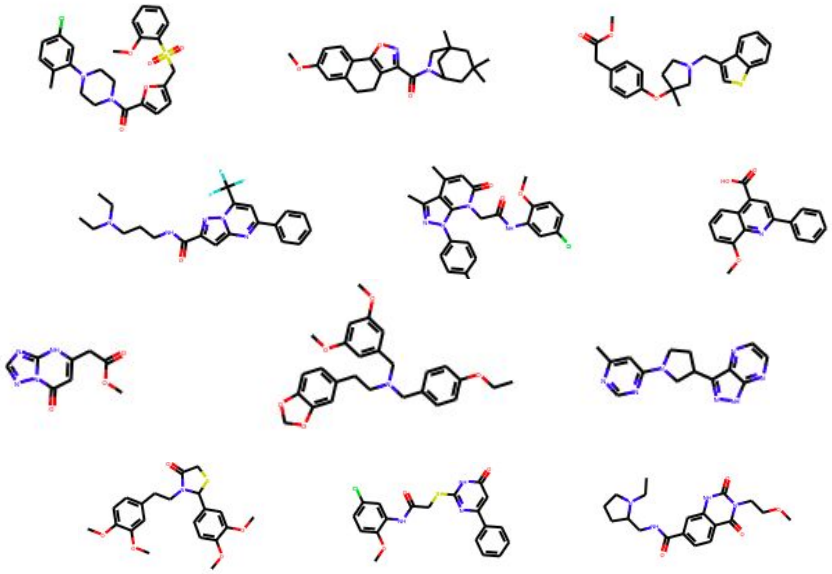
Keep similar molecular weight
Minimize MACCS/166 distance
Favor presence of same atom types

Minimize

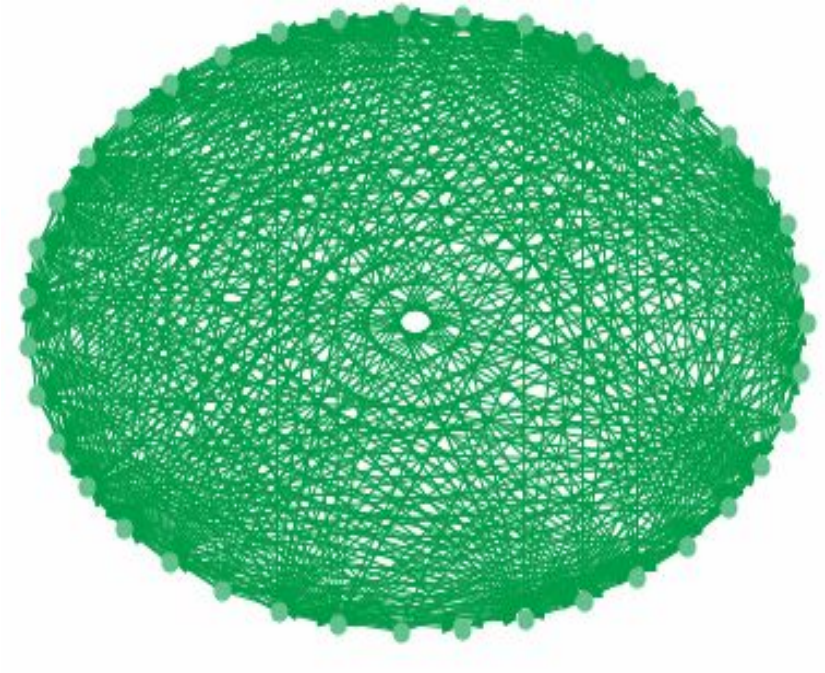
$\mathcal{L} = w_1 * (\%\ \text{difference in heavy atoms}) + w_2 * \text{MACCS distance} + \text{novel penalty} + \text{missing penalty}$

Setting up a genetic algorithm

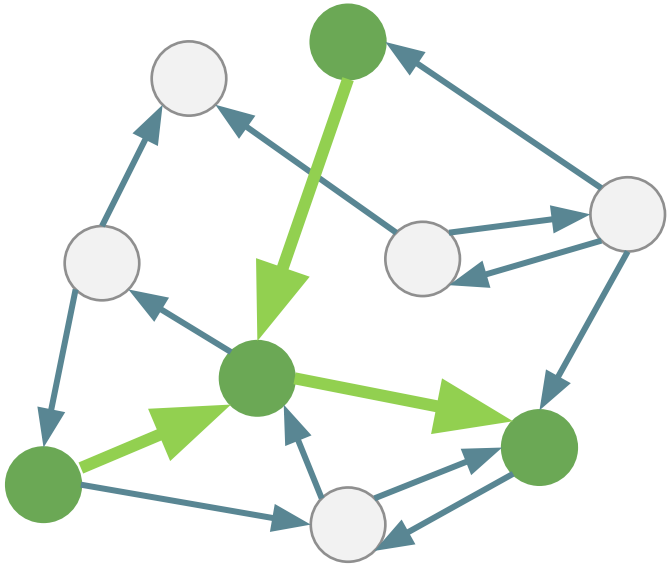
Using parallel fragment pools for molecular diversity



Create a fragment pool (BRICS) from a diverse library and the reference compound



Nodes are connected if a BRICS rule can link them



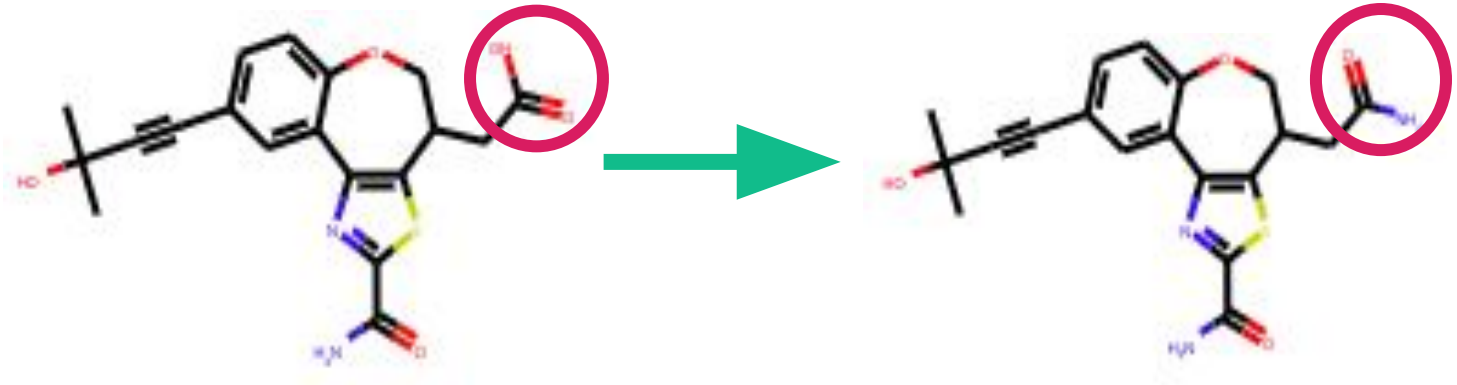
Candidate molecules are trees in this graph

Degen, *et al*. *ChemMedChem*. **3**, 1503 (2008).

Genetic algorithm: evolution and selection

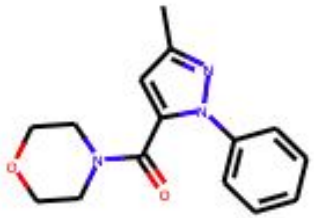
2 children

Cross-over. Break parents at an edge that has the same edge type (LHS & RHS). Form two new children.



Mutation. Replace a terminal node (head or tail) with another eligible terminal node.

What do the chimeric decoys look like?



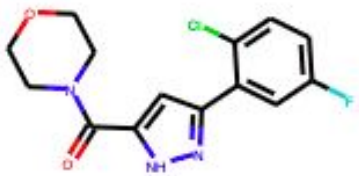
reference

0

1

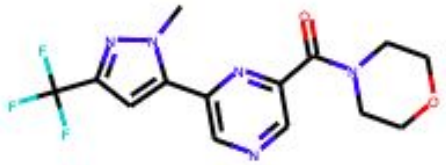
2

3

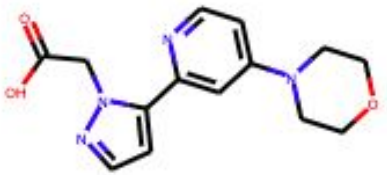


4

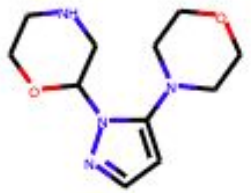
5



6



7



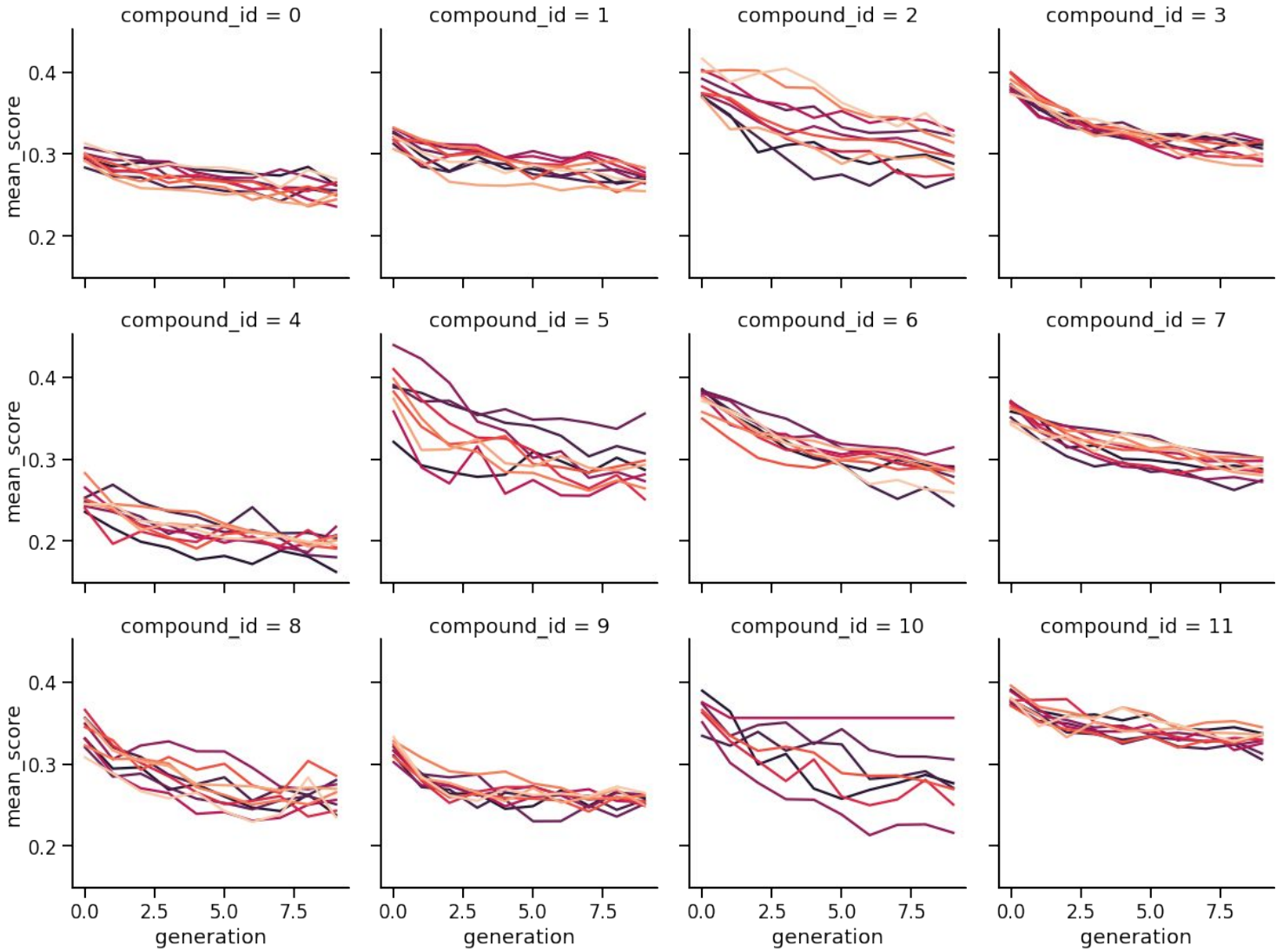
8

Varying combinations of *morpholine, carboxyl, diazole, methyl,* and *phenyl*

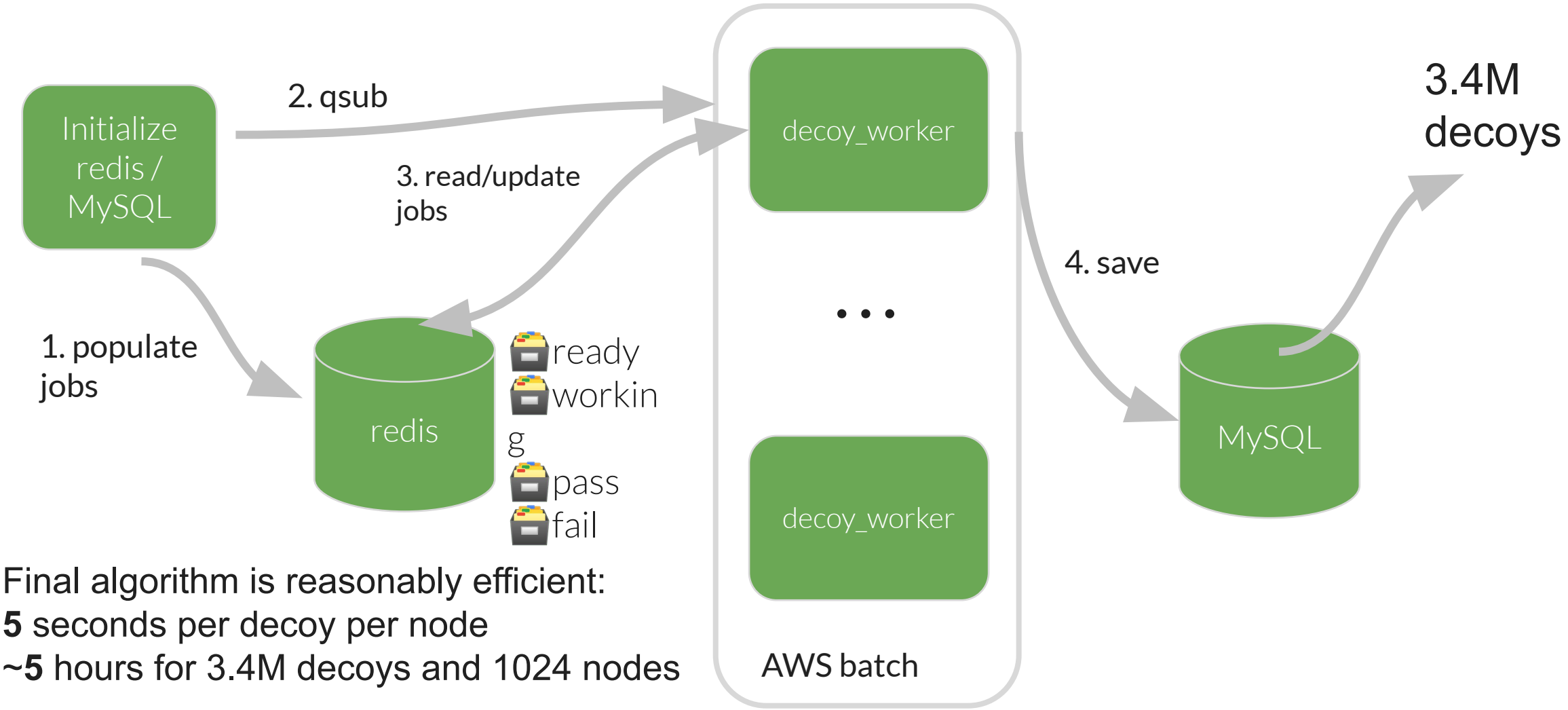
Optimizing hyperparameters

Grid search across # of building blocks, mutation rate, crossover rate, # of generations **vs**. quality and speed

Parameter	Best value
% crossover	**40%**
% mutation	**30%**
% new compounds	**30%**
# of generations	**5**



Generating millions of chimeric decoys

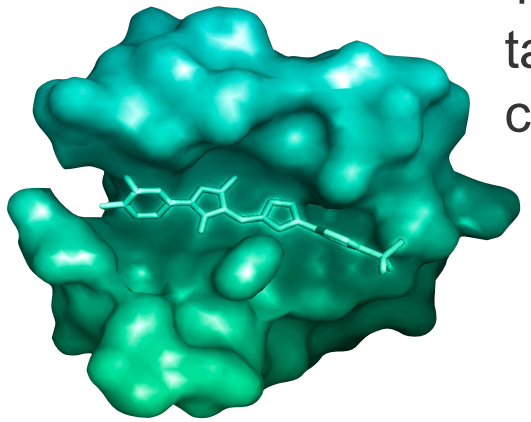


Final algorithm is reasonably efficient:
5 seconds per decoy per node
~5 hours for 3.4M decoys and 1024 nodes

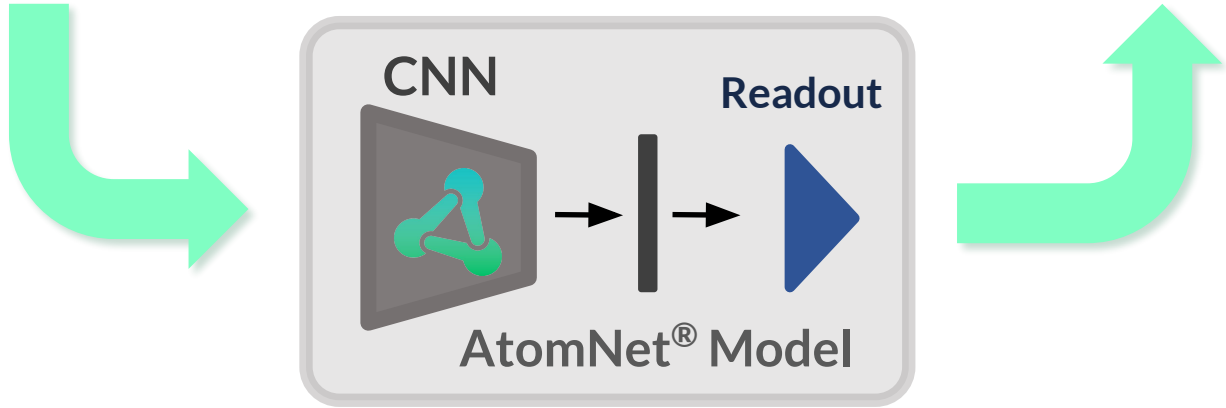
Results

Evaluation

- New activity classification models
  - Varying fraction of chimeric decoys and *actives-as-decoys**
- Baseline
  - 100% *actives-as-decoys* and measured negative compounds



1.9M target/compound complexes

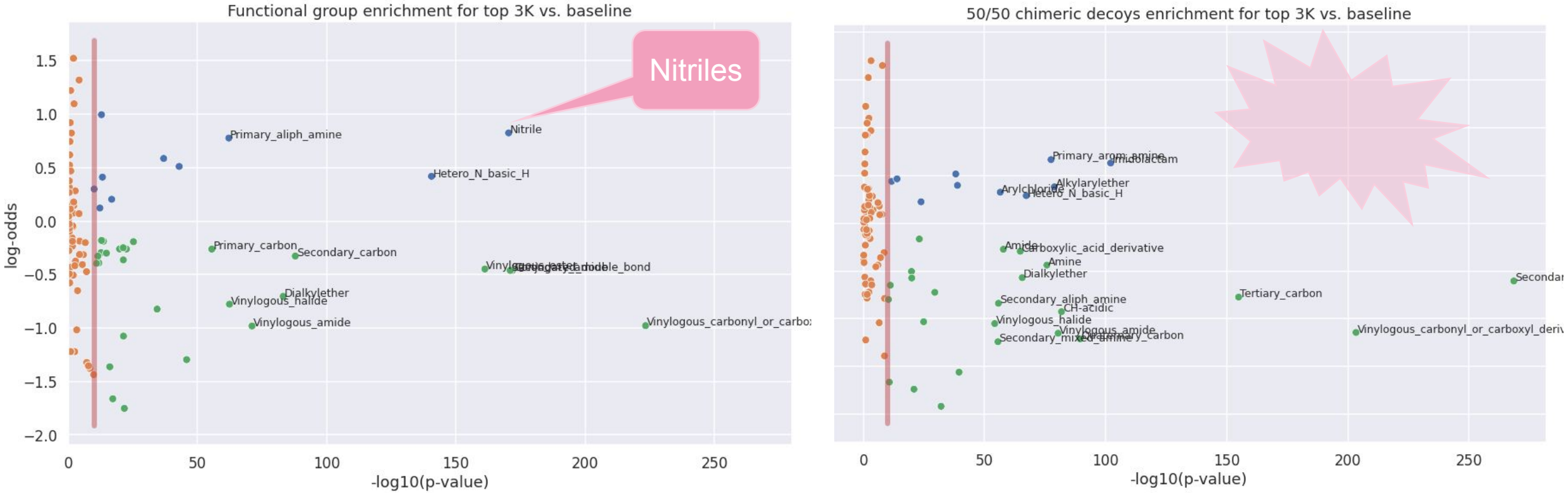


Binding classification

***actives-as-decoys:** presenting active compounds as decoys against non-active targets

Training with chimeric decoys reduces bias

Revisiting original project results



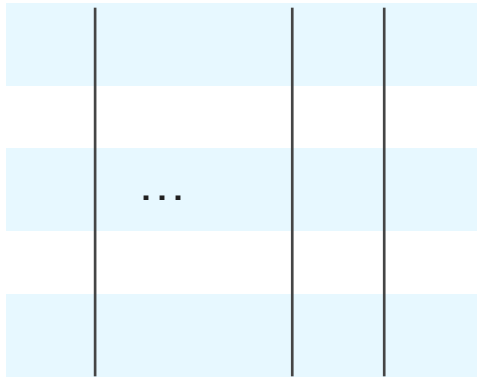
Top 3K Neither Baseline 50K

Evaluating bias comprehensively

Visualizing good bias and undesirable bias

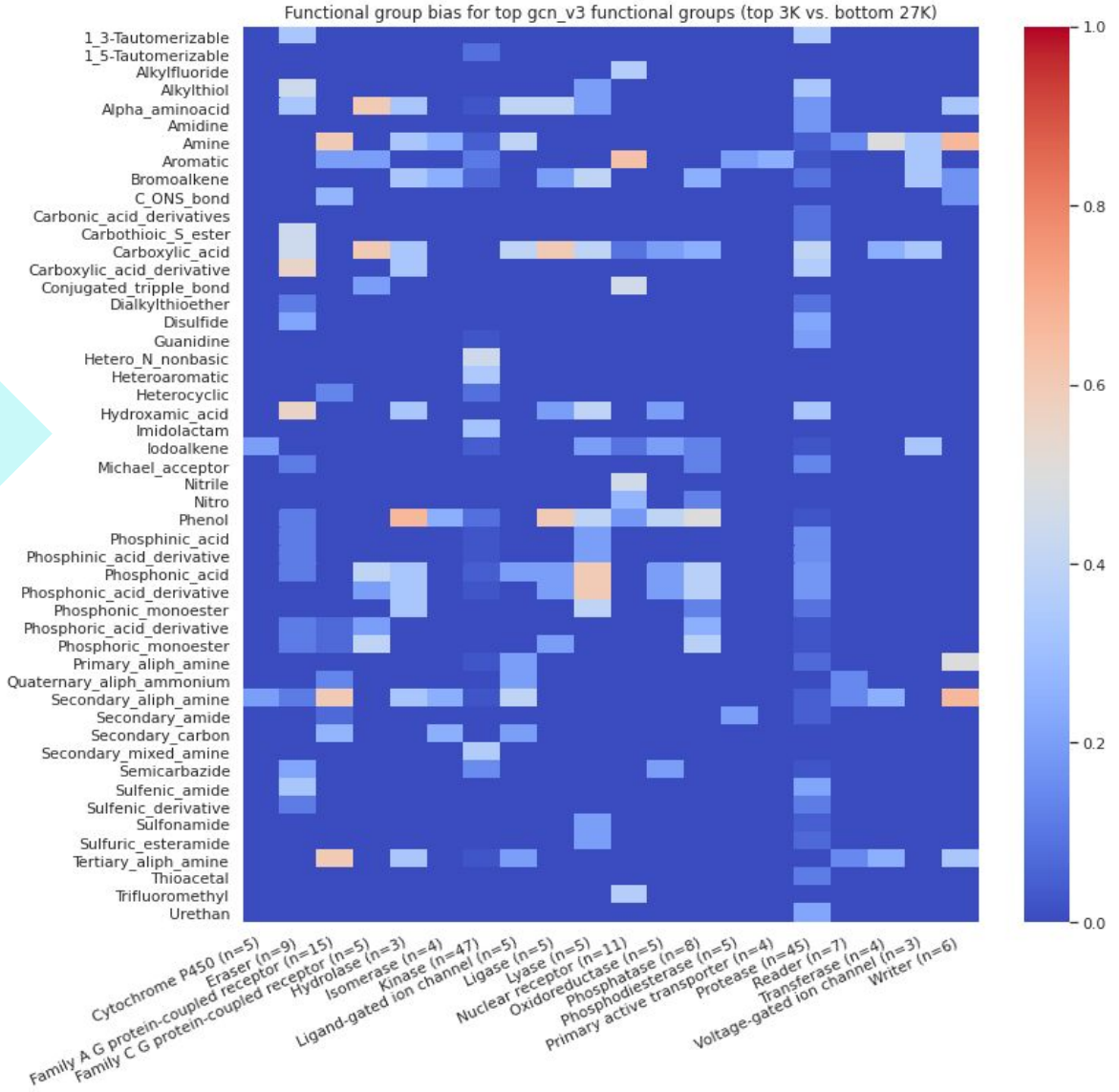
200 targets →

30,000 fragments ↓



Enrichment of functional groups in top predictions (Fisher's test)

Prepare 6M complexes spanning representative groups and target classes

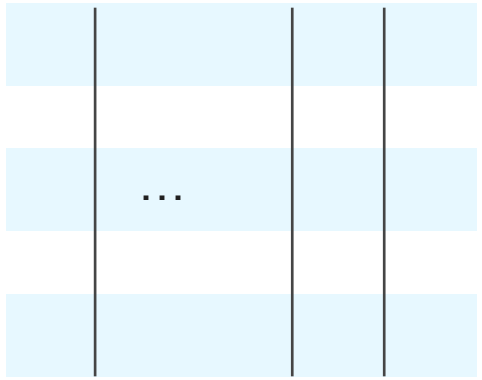


Evaluating bias comprehensively

Visualizing good bias and undesirable bias

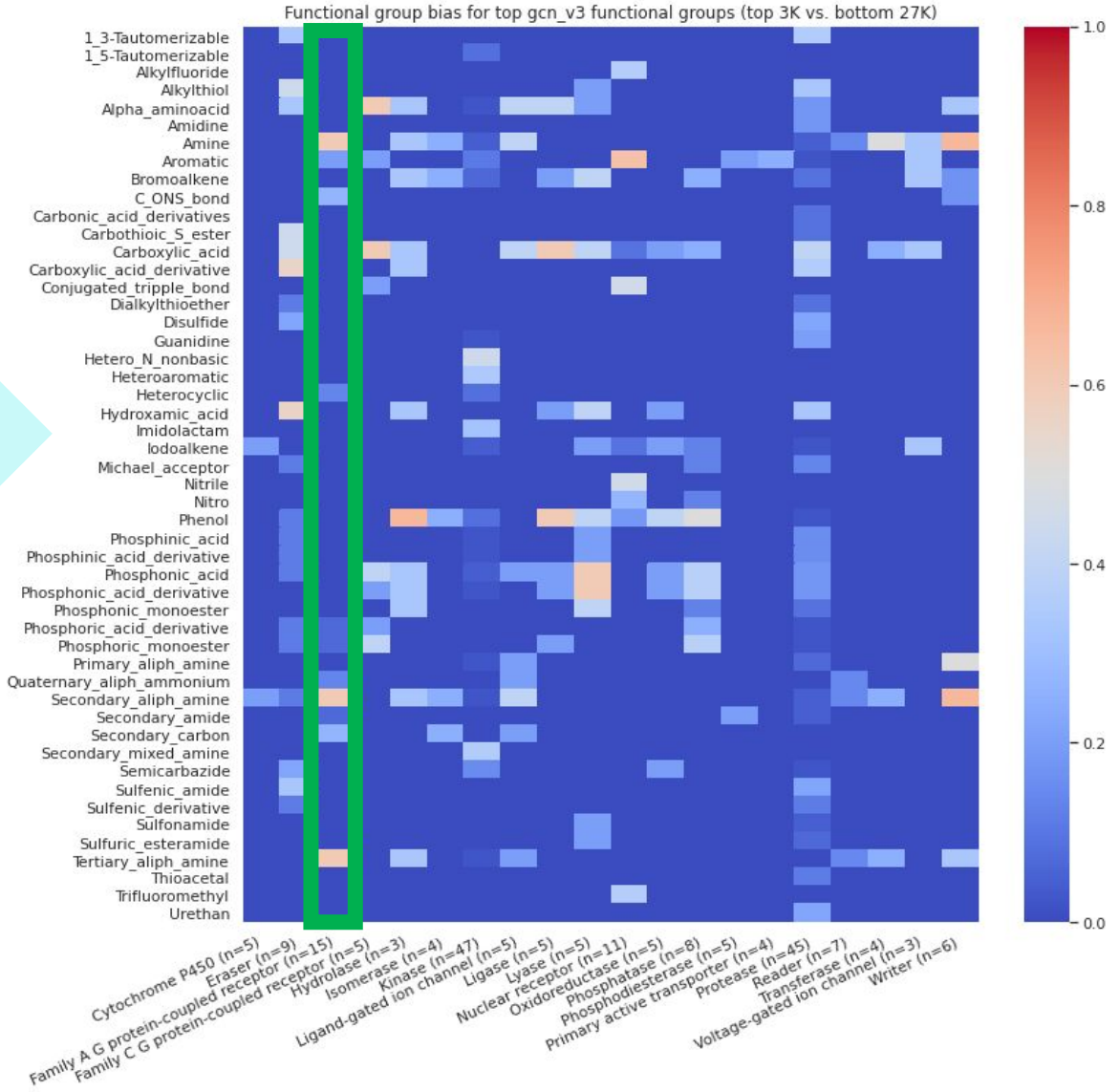
200 targets →

30,000 fragments ↓



Prepare 6M complexes spanning representative groups and target classes

Enrichment of functional groups in top predictions (Fisher's test)



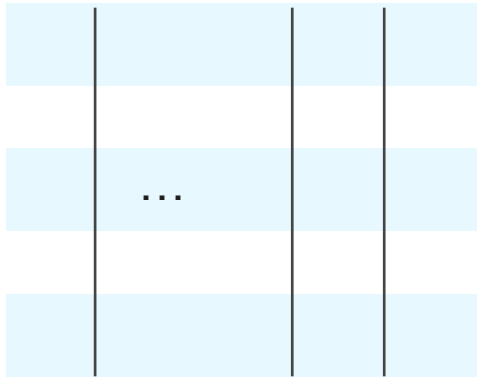
Good bias: Family A GPCRs favor amines

Evaluating bias comprehensively

Visualizing good bias and undesirable bias

200 targets →

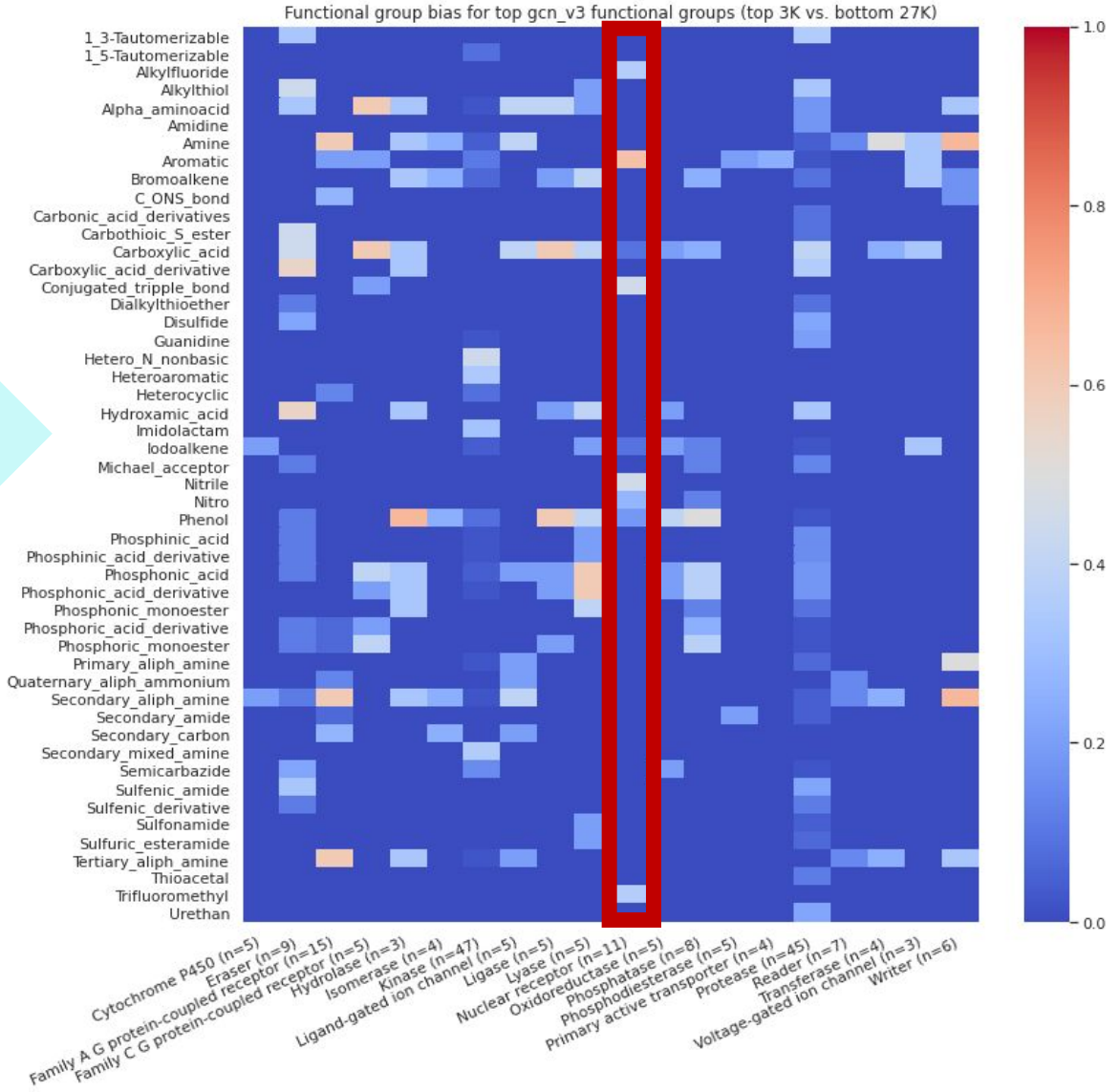
30,000 fragments ↓



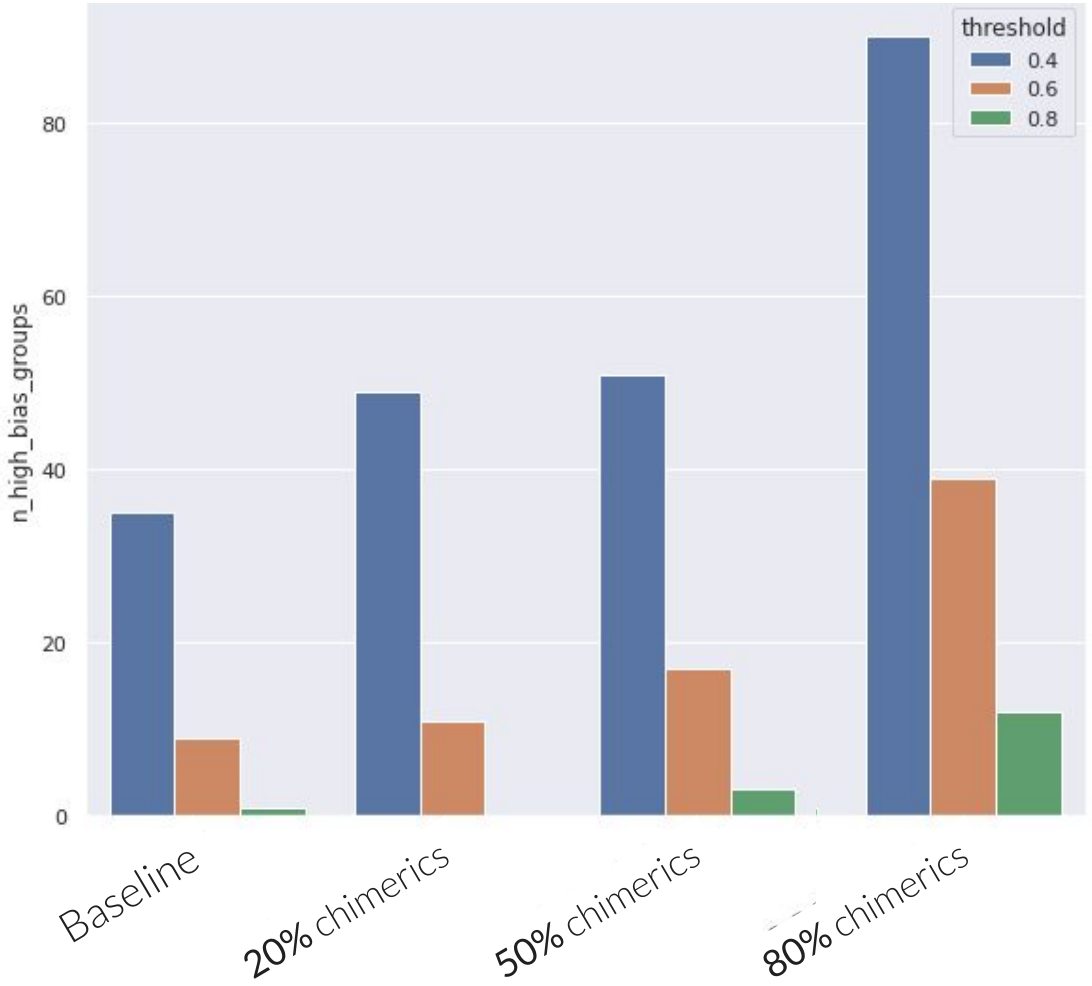
Enrichment of functional groups in top predictions (Fisher's test)

Prepare 6M complexes spanning representative groups and target classes

Bad bias: fluorinated alkyls for nuclear receptors



…comprehensive examination shows bias was not reduced



Increasing bias across target classes

What happened?

Did the models exploit new trivial hyperplanes?

Procedure

1. Train random forests to discriminate **decoy versus real** using MACCS+properties (5 trials × 465 compounds)
2. Train random forests with 50% reference vs. 50% reference for null model.

Input	AUC
reference/chimerics	**0.77** ± 0.02
reference/reference	0.497

What happened?

Did the models exploit new trivial hyperplanes?

Procedure

1. Train random forests to discriminate **decoy versus real** using MACCS+properties (5 trials × 465 compounds)
2. Train random forests with 50% reference vs. 50% reference for null model.

Input	AUC
reference/reference	**0.77** ± 0.02
reference/chimerics	0.497

Decoys are different by a MACCS bit and common global properties

RF feature importance

feature	
maccs_62	0.051317
tpsa	0.042159
CrippenClogP	0.036687
exactmw	0.035783
labuteASA	0.032326
CrippenMR	0.031435
FractionCSP3	0.029717
lipinskiHBD	0.026461
NumHBA	0.023386
NumAmideBonds	0.023269

Future directions

- Characterize hyperplanes
- Revisit fitness function and/or fragment pool diversity
- Online adversarial decoys
- Other generative approaches
  - GANs
  - Focus decoys on the decision boundary

Thank you!

Acknowledgments

Thanks to Abe, Izzy, and the cheminformatics, ML and software teams at Atomwise

Other Talks & Posters

https://info.atomwise.com/acs_spring2021

Join Us!

https://www.atomwise.com/careers/