AI in Drug Discovery – An Overview

Andrea Volkamer and Pat Walters

September 22, 2025

Session 2 - Data is all you need!

Who we are!

Pat Walters
OpenADMET

Raquel López-Ríos de Castro
Chodera Lab, MSKCC NYC

Afnan Sultan
Saarland University

Lisa-Marie Rolli
Saarland University

Andrea Volkamer
Saarland University

What we will do today

Session 0 – 1:00– 1:30 pm

- Introduction to Jupyter notebooks

Session 1 - 1:30 - 2:30 pm

- An introduction to Artificial Intelligence (AI) and Machine Learning (ML)
- Molecular representations
- AI architectures

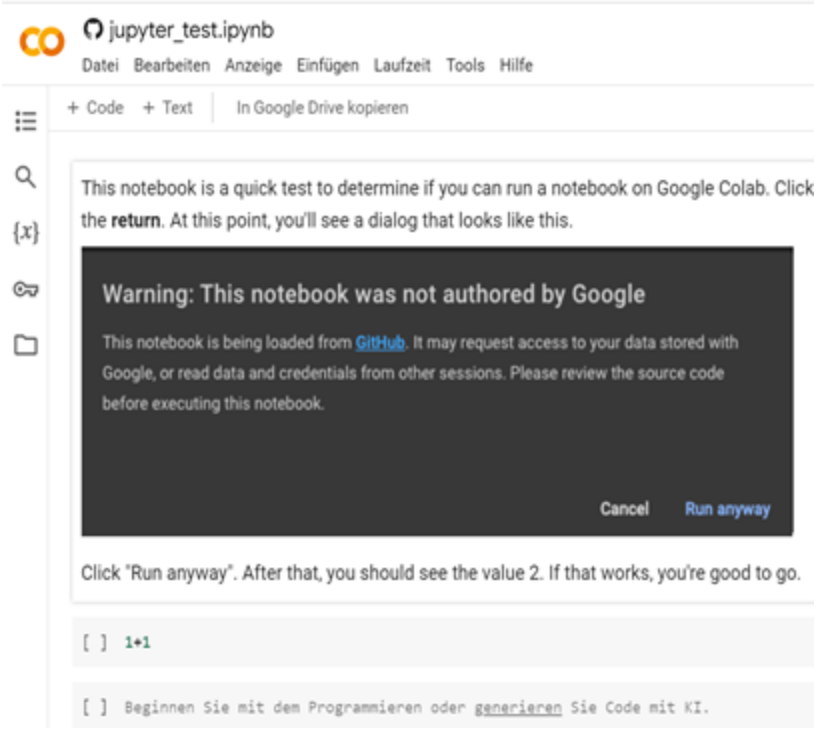
Session 2 - 3:00 - 4:00 pm

- The importance of data quality for AI/ML
- Exploratory data analysis
- Data preprocessing
- Applicability domains

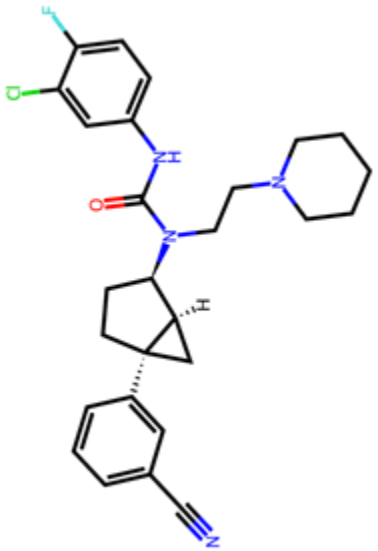
Session 3 - 4:30 - 5:30 pm

- AI in Practice
- Molecule generation
- Protein structure prediction
- Active learning

Lectures supported by hands-on sessions …



The Path From Molecules to Properties



Supervised Learning (SL)

Representation
- Physchem properties
- Morgan FPs
- Adjacency matrix

→

Model
- RF
- SVM
- GNN

→

Self-Supervised Learning (SSL)

Tokenization
- Sequences
- Graphs

→

Foundational Models
- Transformers
- Autoencoders

→

Fine-tuning

→

The Two Limiting Factors

Data

∞

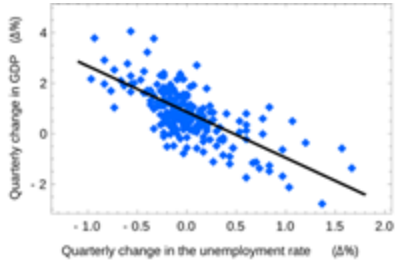
Limited

Representative

Just needs the right model!

Model

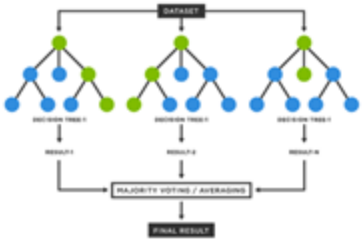
Simple



$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots + \epsilon$$

$$P = N_{features} + 2$$

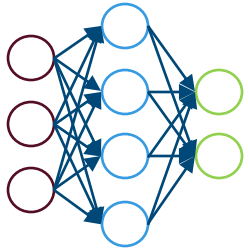
Moderate



$$Y = \frac{1}{T}\sum_{t=1}^{T} f_T(x)$$

$$P \propto N_{trees}, T_{depth}, \ldots$$

Complex



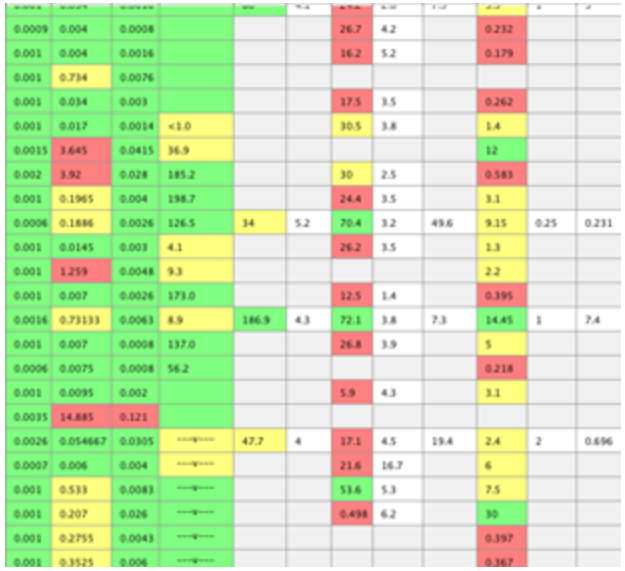
$$Y^l = W^l a^{l-1} + b^l$$

$$P \propto W, b$$

P = # parameter

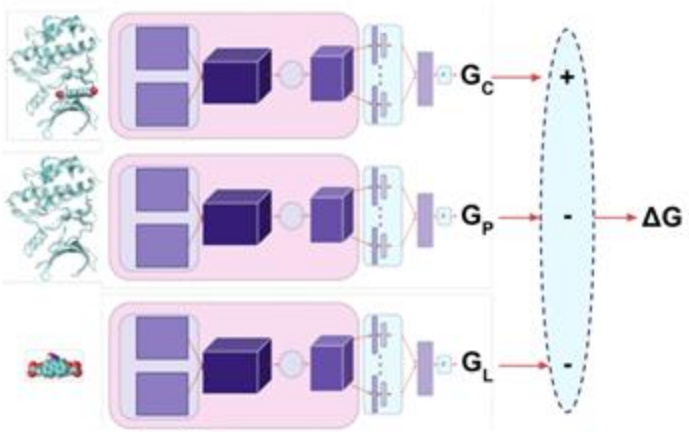
Key Components of Machine Learning in Drug Discovery (or Anything Else)

Data

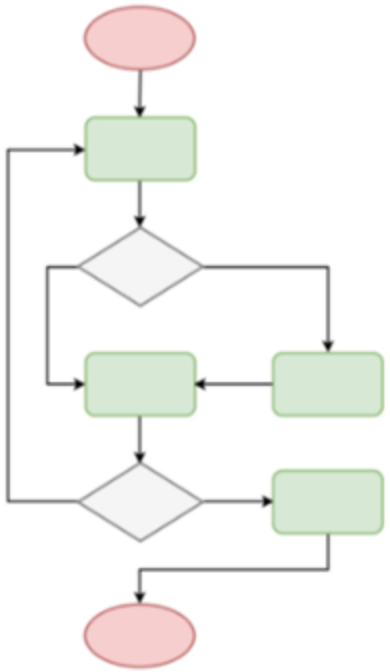


0.0009	0.004	0.0008				26.7	4.2		0.232		
0.001	0.004	0.0016				16.2	5.2		0.179		
0.001	0.734	0.0076									
0.001	0.034	0.003				17.5	3.5		0.262		
0.001	0.017	0.0014	<1.0			30.5	3.8		1.4		
0.0015	3.645	0.0415	36.9						12		
0.002	3.92	0.028	185.2			30	2.5		0.583		
0.001	0.1965	0.004	198.7			24.4	3.5		3.1		
0.0006	0.1886	0.0026	126.5	34	5.2	70.4	3.2	49.6	9.15	0.25	0.231
0.001	0.0145	0.003	4.1			26.2	3.5		1.3		
0.001	1.259	0.0048	9.3						2.2		
0.001	0.007	0.0026	173.0			12.5	1.4		0.395		
0.0016	0.73133	0.0063	8.9	186.9	4.3	72.1	3.8	7.3	14.45	1	7.4
0.001	0.007	0.0008	137.0			26.8	3.9		5		
0.0006	0.0075	0.0008	56.2						0.218		
0.001	0.0095	0.002				5.9	4.3		3.1		
0.0035	14.885	0.121									
0.0026	0.054667	0.0305	---v---	47.7	4	17.1	4.5	19.4	2.4	2	0.696
0.0007	0.006	0.004	---v---			21.6	16.7		6		
0.001	0.533	0.0083	---v---			53.6	5.3		7.5		
0.001	0.207	0.026	---v---			0.498	6.2		30		
0.001	0.2755	0.0043	---v---						0.397		
0.001	0.3525	0.006	---v---						0.367		

Representation



Algorithms



Key Components of Machine Learning in Drug Discovery (or Anything Else)

Data

0.0009	0.004	0.0008				26.7	4.2		0.232		
0.001	0.004	0.0016				16.2	5.2		0.179		
0.001	0.734	0.0076									
0.001	0.034	0.003				17.5	3.5		0.262		
0.001	0.017	0.0014	<1.0			30.5	3.8		1.4		
0.0015	3.645	0.0415	36.9						12		
0.002	3.92	0.028	185.2			30	2.5		0.583		
0.001	0.1965	0.004	198.7			24.4	3.5		3.1		
0.0006	0.1886	0.0026	126.5	34	5.2	70.4	3.2	49.6	9.15	0.25	0.231
0.001	0.0145	0.003	4.1			26.2	3.5		1.3		
0.001	1.259	0.0048	9.3						2.2		
0.001	0.007	0.0026	173.0			12.5	1.4		0.395		
0.0016	0.73133	0.0063	8.9	186.9	4.3	72.1	3.8	7.3	14.45	1	7.4
0.001	0.007	0.0008	137.0			26.8	3.9		5		
0.0006	0.0075	0.0008	56.2						0.218		
0.001	0.0095	0.002				5.9	4.3		3.1		
0.0035	14.885	0.121									
0.0026	0.054667	0.0305	---v---	47.7	4	17.1	4.5	19.4	2.4	2	0.696
0.0007	0.006	0.004	---v---			21.6	16.7		6		
0.001	0.533	0.0083	---v---			53.6	5.3		7.5		
0.001	0.207	0.026	---v---			0.498	6.2		30		
0.001	0.2755	0.0043	---v---						0.397		
0.001	0.3525	0.006	---v---						0.367		

Representation

Algorithms

Where Does Machine Learning Excel?

Large amounts of data

- **Pharmaceutical data is miniscule compared to many other fields**

Responses are definitive

Samples are independently distributed

- **No cases where the same example falls into 2 different categories**

Samples are identically distributed

- **Equal distributions of positive and negative examples**

Training data is representative of what is being predicted

Pharmaceutical Data is Not Ideal for Machine Learning

Data is sparse

- **Rarely have a complete data matrix**

Data is truncated

- **Many assay values report as “<1” or “>30”**
- **Difficult to know the true value**

Data has a limited dynamic range

- **Often spans only 2 or 3 logs**
- **Small dynamic range combined with experimental error makes significant correlations difficult**

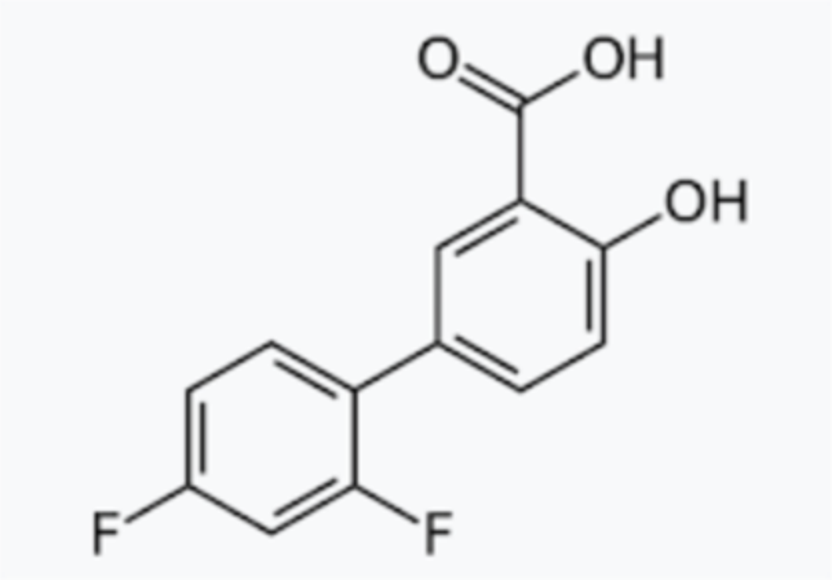
Even data from the “same” assay can be heterogeneous

- **PK data measured with different doses, vehicles and formulations**
- **Response can vary with operator, equipment, lab**

Data covers a limited chemical space

- **Even global models can be local**

We Can Have Multiple Experimental Values For the Same Compound



Diflunisal

Form	Solubility µg/ml	LogS mol/L
1	26	-3.9
2	7.6	-4.5
3	0.93	-5.4
4	0.29	-5.9

Different crystal polymorphs of Diflunisal have different aqueous solubilities

Small Molecule Bioactivity Datasets: ChEMBL or PubChem

Database for collecting binding affinities

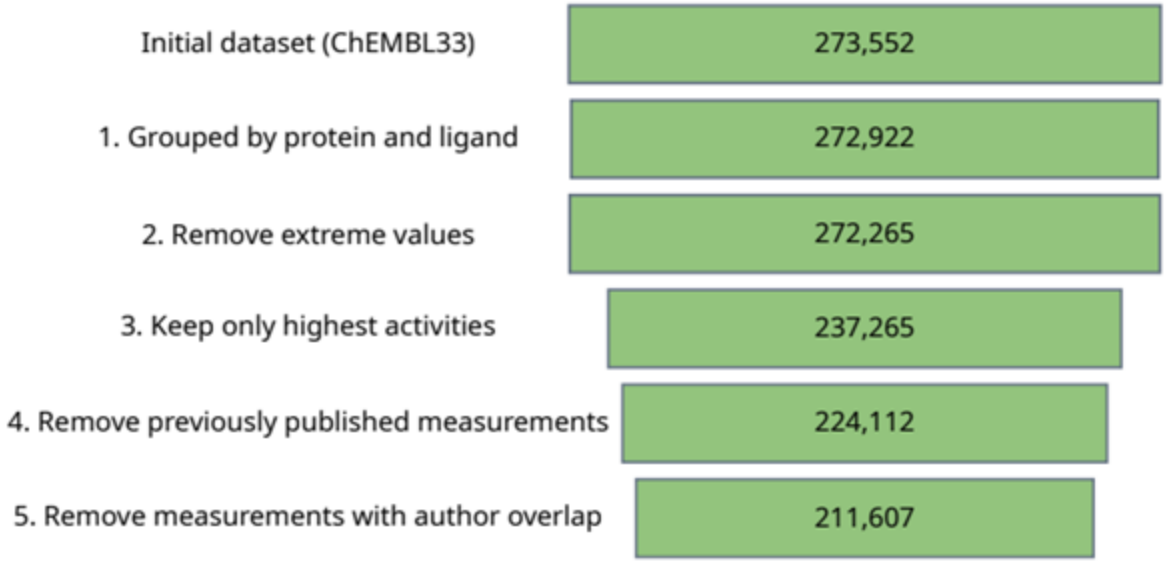
ChEMBL ID	Search Hit	Name	Synonyms	Type	Max Phase	Molecular Weight	Targets	Bioactivities
CHEMBL4558324		LAZERTINIB	C-18112003-G, GNS1480, GNS-1480, JNJ-73841937-AAA, Lazertinib, Yh25448, YH25448, YH-25448	Small molecule	3	554.66	12 By Type:	44 By Std. Type:
CHEMBL4650319		MOBOCERTINIB	AP32788, AP-32788, Mobocertinib, Tak-788, TAK-788	Small molecule	4	585.71	14 By Type:	55 By Std. Type:
CHEMBL4761468		AUMOLERTINIB	Almonertinib, Ameile, Amerol, Aumolertinib, Egfr t790m inhibitor hs-10296, EQ143, EQ-143, HS-10206, Hs 10296, Hs-10296, HS-10296	Small molecule	3	525.66	5 By Type:	23 By Std. Type:

https://www.ebi.ac.uk/chembl/

How to Preprocess such Data Properly?

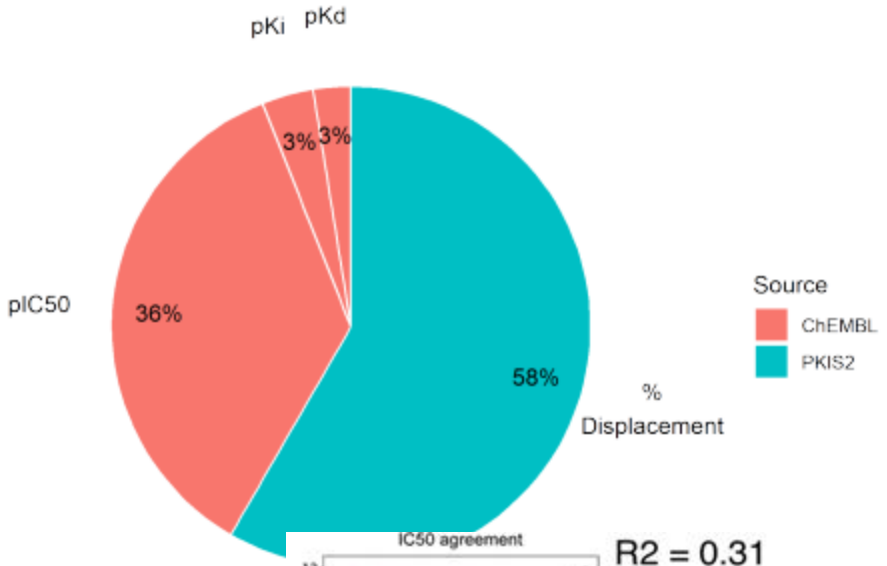
- Automated curation pipeline to reduce errors and to increase reliability

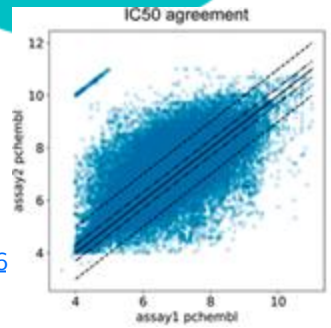
see Kramer, et al. *JMedChem* **2012**; 55(11):5165–5173



López-Ríos de Castro, *et al.*, *biorxiv*, **2024**
https://doi.org/10.1101/2024.09.10.612176

- Bioactivity assay measurement classes vary by and within data set

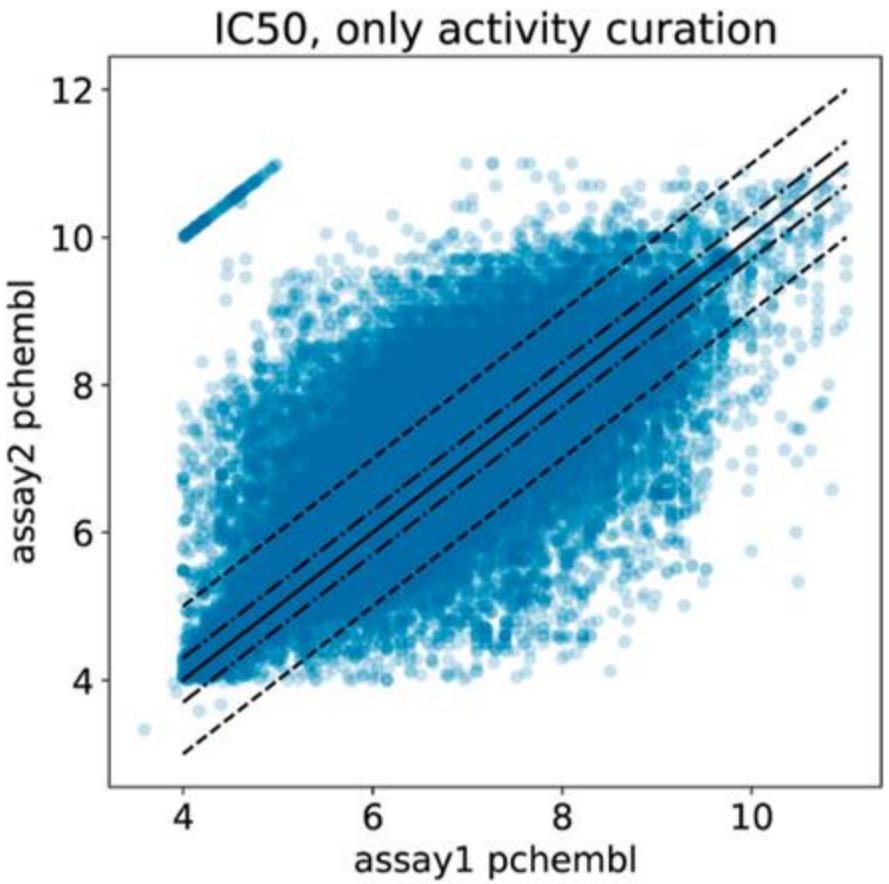


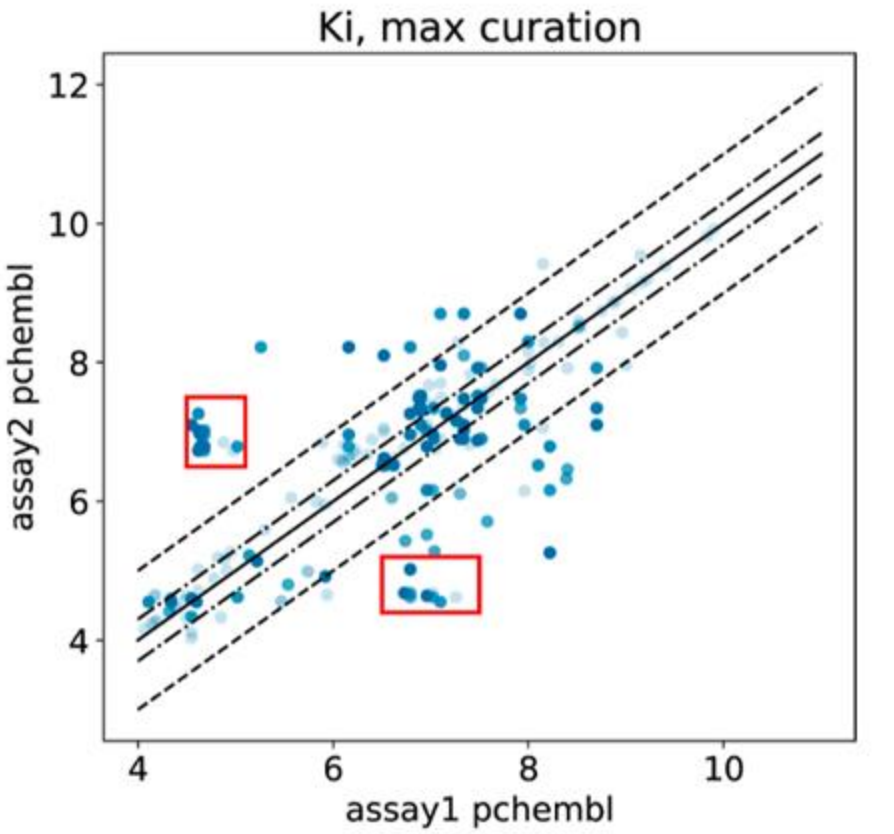


R2 = 0.31
MAE = 0.50
Kendall's tau = 0.51

Landrum, Riniker., JCIM, **2024**, 64,5
https://doi.org/10.1101/2024.09.10.612176

Inconsistent Data Can Make ML Modeling Difficult

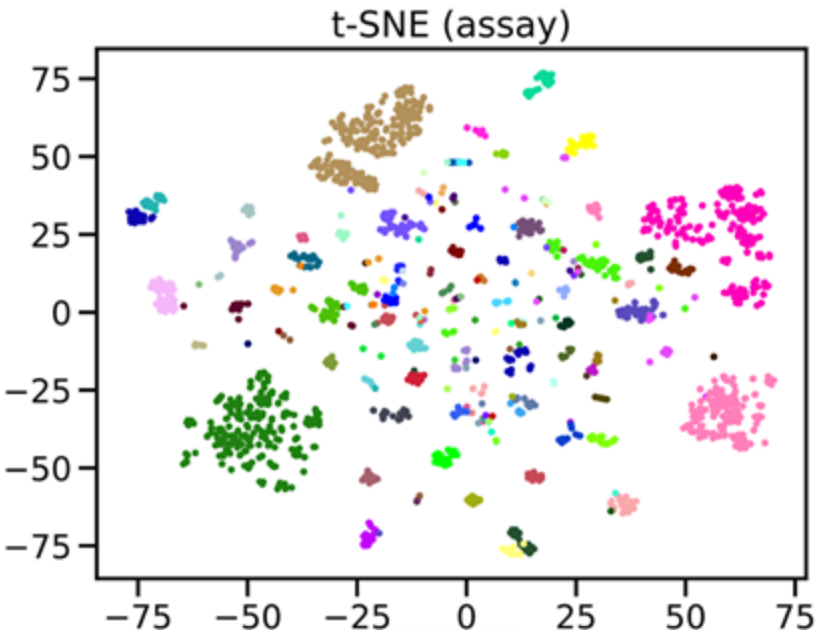




Landrum, Gregory A., and Sereina Riniker. "Combining IC50 or K i values from different sources is a source of significant noise." *Journal of Chemical Information and Modeling* 64.5 (2024): 1560-1567.

Typical Challenges in (Public) Chemical Datasets

- Different experimental protocols
- Inconsistent values
- Missing documentation
- Data tends to be heavily clustered



- kinodata-EGFR ligand activities
- t-SNE on 2048 bit Morgan fingerprints
- colors = assays

Literature Datasets are Problematic


Check for updates

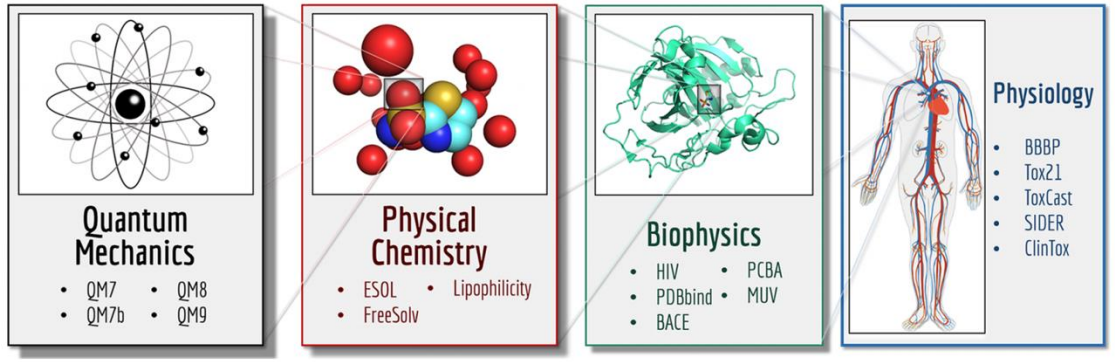
comment

Artificial intelligence foundation for therapeutic science

Artificial intelligence (AI) is poised to transform therapeutic science. Therapeutics Data Commons is an initiative to access and evaluate AI capability across therapeutic modalities and stages of discovery, establishing a foundation for understanding which AI methods are most suitable and why.

Kexin Huang, Tianfan Fu, Wenhao Gao, Yue Zhao, Yusuf Roohani, Jure Leskovec, Connor W. Coley, Cao Xiao, Jimeng Sun and Marinka Zitnik

Therapeutic Data Commons

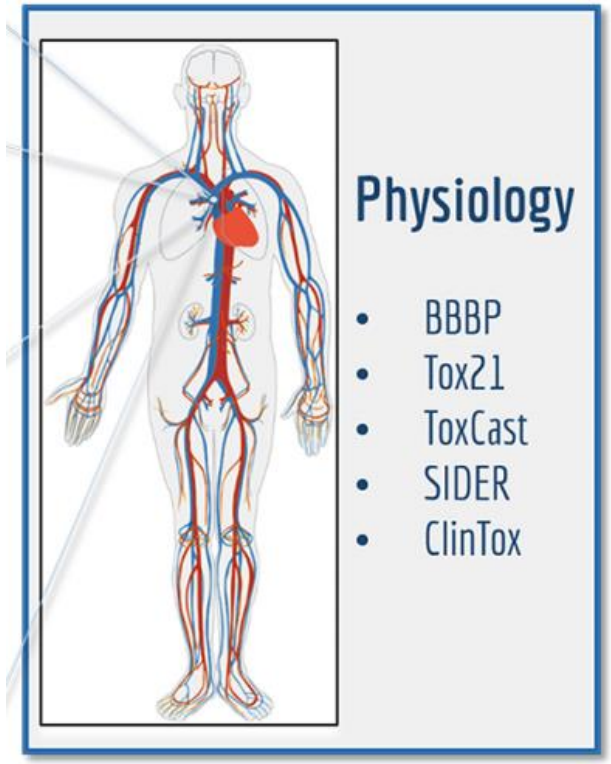


MoleculeNet

Chemical structure errors
Unspecified stereochemistry
Inconsistent experiments
Curation errors
Unrealistic dynamic range
Irrelevant experiments
Poorly defined endpoints
Assay artifacts

Please Don't Use These Datasets!

For Now, Focus on Simple, Consistent, Well-Defined Endpoints



SIDER Categories

Hepatobiliary disorders
Metabolism and nutrition disorders
Product issues
Eye disorders
Investigations
Musculoskeletal and connective tissue disorders
Gastrointestinal disorders
Social circumstances
Immune system disorders
Reproductive system and breast disorders
Neoplasms benign, malignant and unspecified (incl cysts and polyps)
General disorders and administration site conditions
Endocrine disorders
Surgical and medical procedures
Vascular disorders
Blood and lymphatic system disorders
Skin and subcutaneous tissue disorders
Congenital, familial and genetic disorders
Infections and infestations
Respiratory, thoracic and mediastinal disorders
Psychiatric disorders
Renal and urinary disorders
Pregnancy, puerperium and perinatal conditions
Ear and labyrinth disorders
Cardiac disorders
Nervous system disorders
Injury, poisoning and procedural complications

Recommended Datasets Supplied By the Polaris Initiative

https://polarishub.io/datasets?certifiedOnly=true

asap-discovery/antiviral-admet-2025-unblinded V2 Molecule ADMET
Size: 560 2025-03-28

asap-discovery/antiviral-potency-2025-unblin... V2 Molecule Potency
Size: 1,328 2025-03-28

asap-discovery/antiviral-ligand-poses-202... V2 Molecule Molecule 3 D
Size: 965 2025-03-28

recursion/rxrx3-core V2 Molecule Image
Size: 222,601 2024-12-11

roman-bushuiev/massspecgym V2 Molecule Small molecule discovery
Size: 231,104 2024-11-26

leash-bio/belka-v1 V2 Molecule Screen
Size: ~99.3M 2024-10-30

adaptyv-bio/egfr-binders-v1 protein-design
Size: 213 66 KB 1 benchmark 2024-10-25

adaptyv-bio/egfr-binders-v0 protein-design
Size: 202 61 KB 1 benchmark 2024-09-26

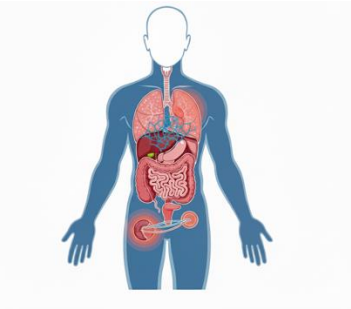
polaris/drewry2017-pkis2-subset-v2 Kinase HitDiscovery
Size: 640 54 KB 14 benchmarks 2024-07-10

biogen/adme-fang-v1 adme
Size: 3,521 384 KB 7 benchmarks 2024-07-10

Polaris certified datasets have been evaluated and approved by industry experts

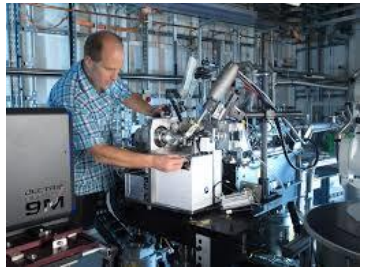
Two Complimentary Efforts That Can Address Some of the Gaps

OpenADMET



How drugs interact with the body
- Absorption
- Metabolism
- Excretion
- Toxicology

OpenBind



How drugs bind to proteins
- Structure
- Binding affinity
- 500K structures / 5yrs

Data Generation and Dissemination

Blind Challenges to Test Methods

Establishing Best Practices

Applicability Domains

Is the training data relevant to what is being predicted?

Is My Training Data Relevant?

Machine learning is all about labeling things using examples

If I train on this?

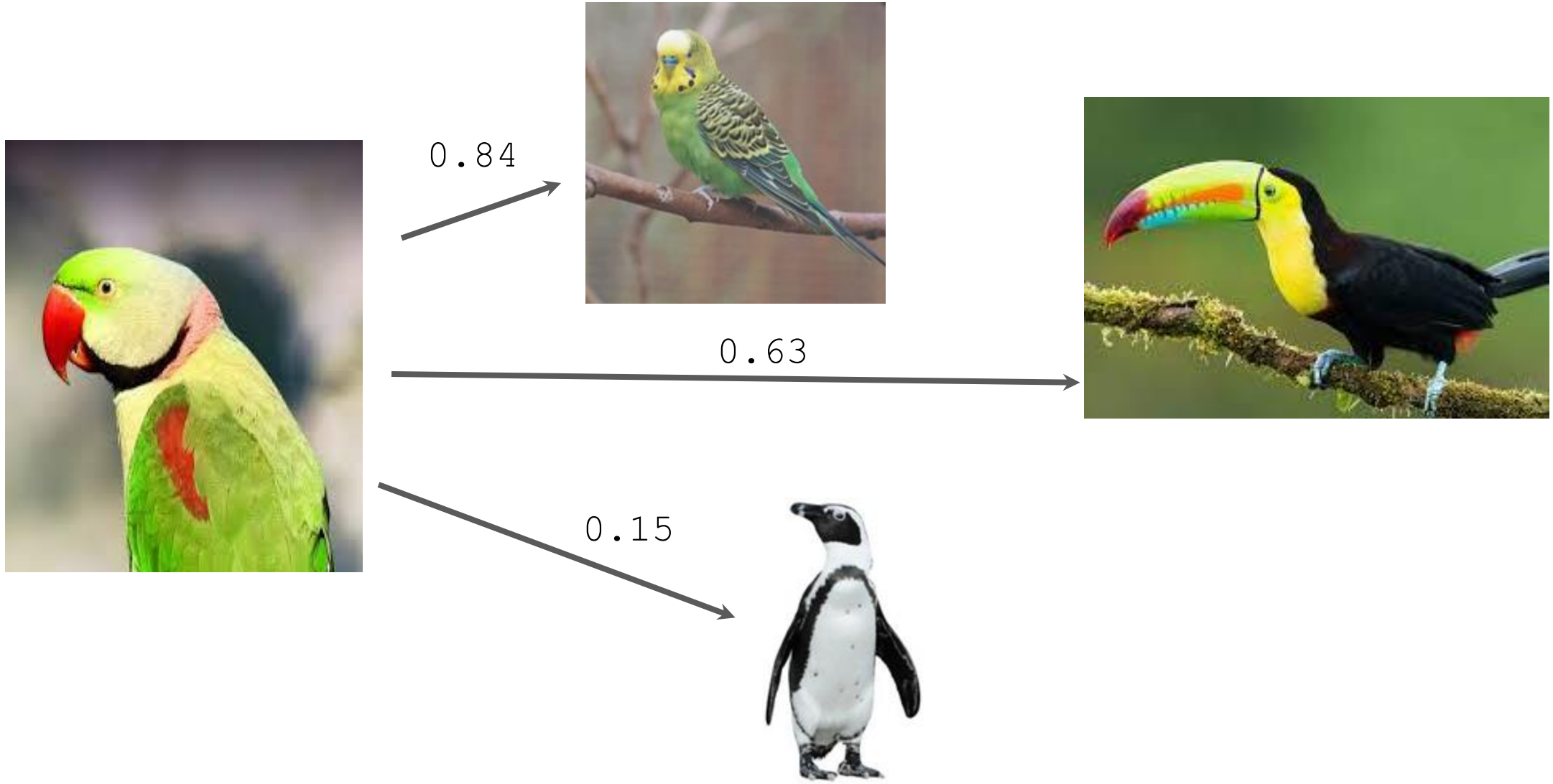
Can I predict this?

Example Scenarios

Can I predict properties for my molecules based on literature data?

Can I make predictions for a new project from my existing data?

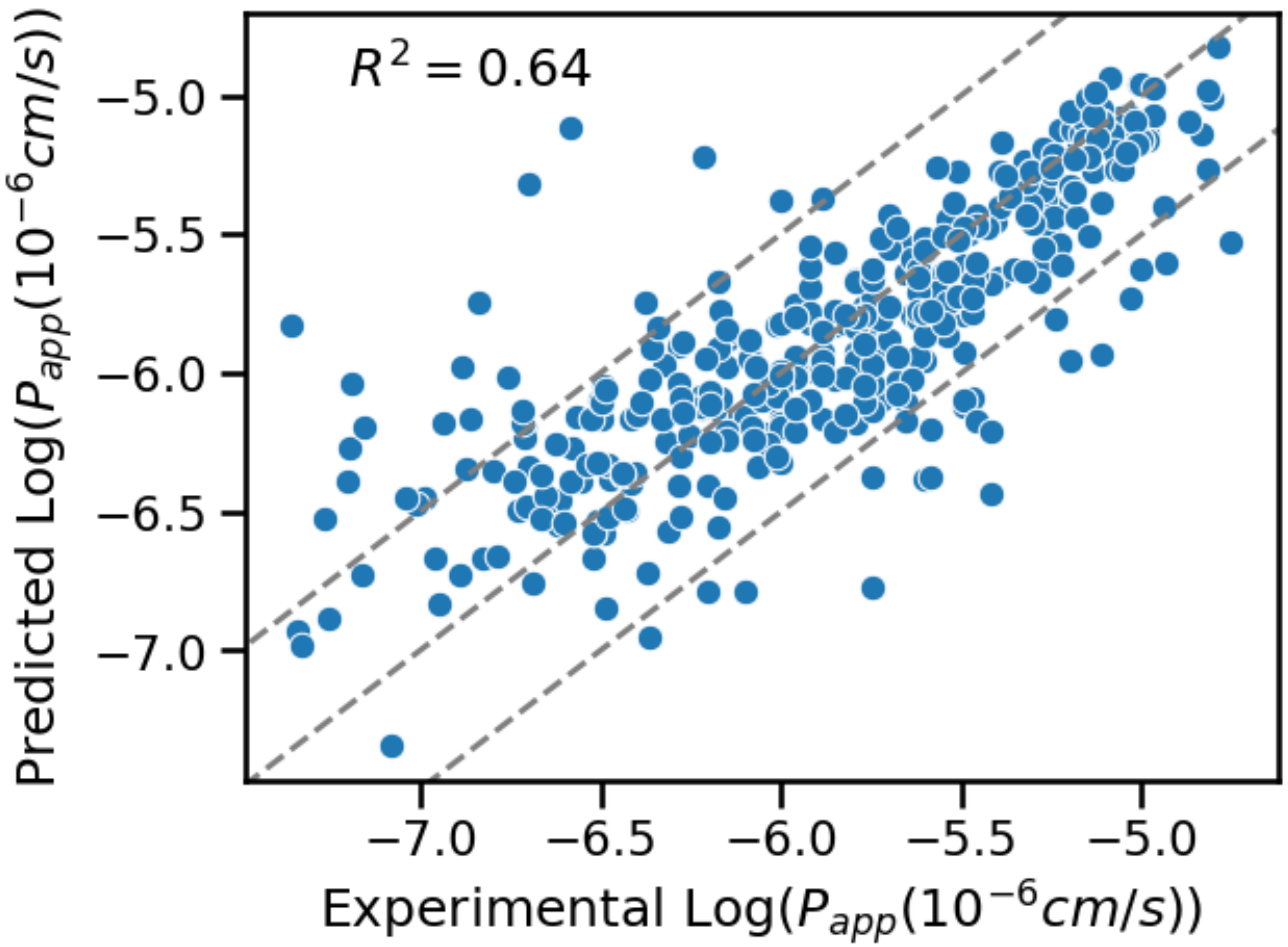
Using Similarity to Assess Applicability



Caco2 Permeability Data From the Scientific Literature

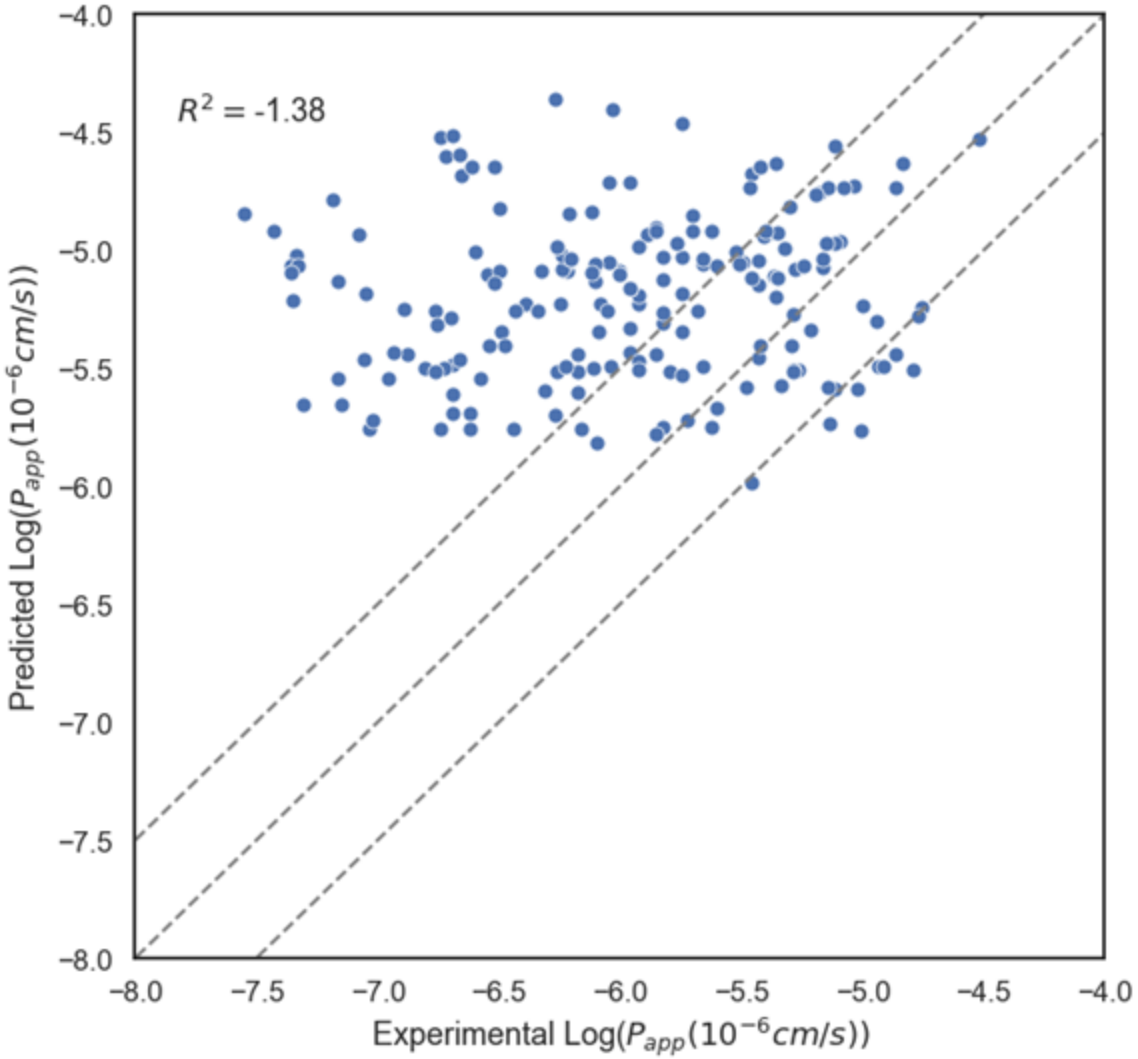
Therapeutic Data Commons - https://tdcommons.ai

- **960 Compounds**
- **Data collected from 23 papers**

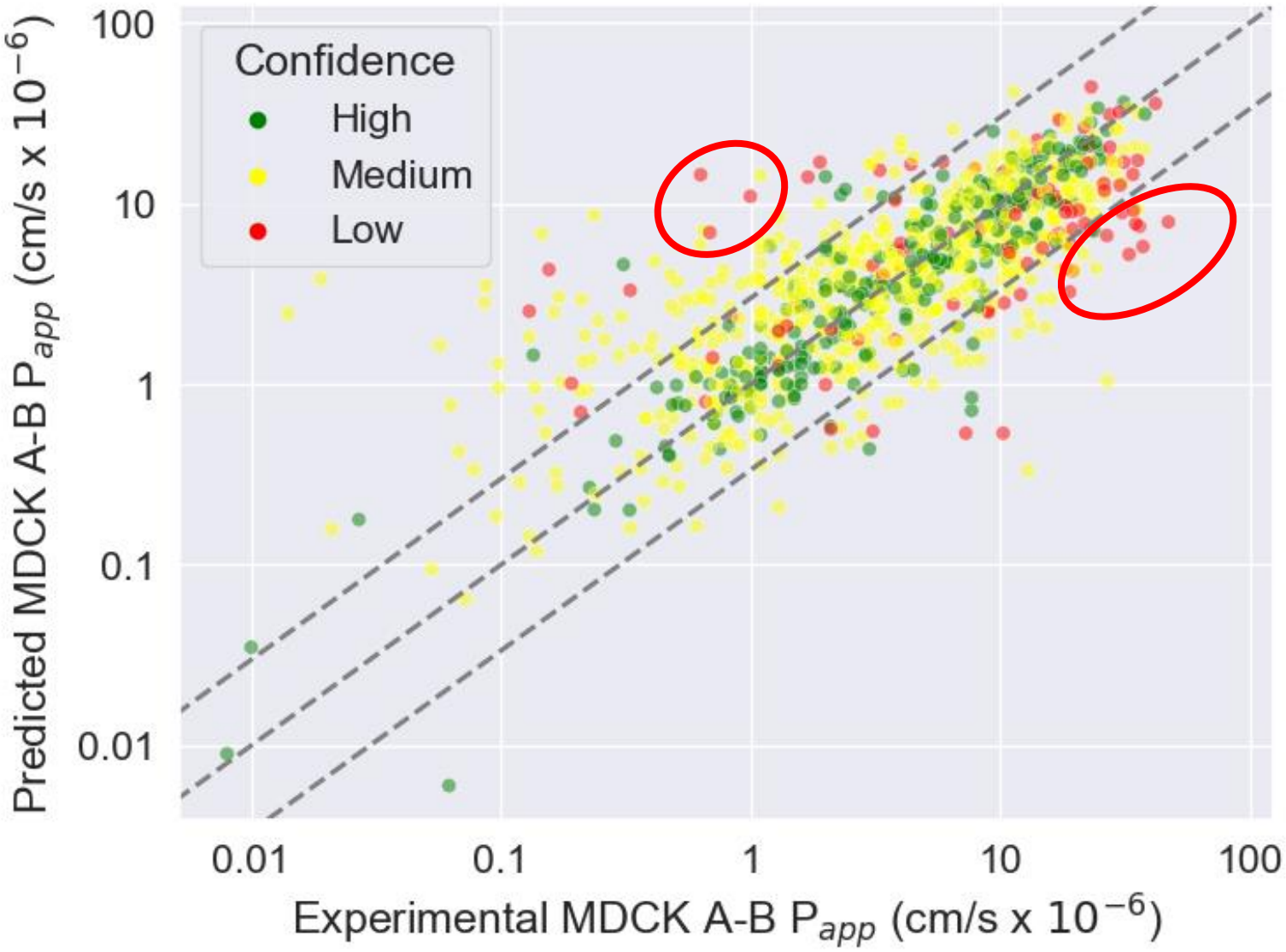


Wang, *J.Chem.Inf.Model* (2016) 763

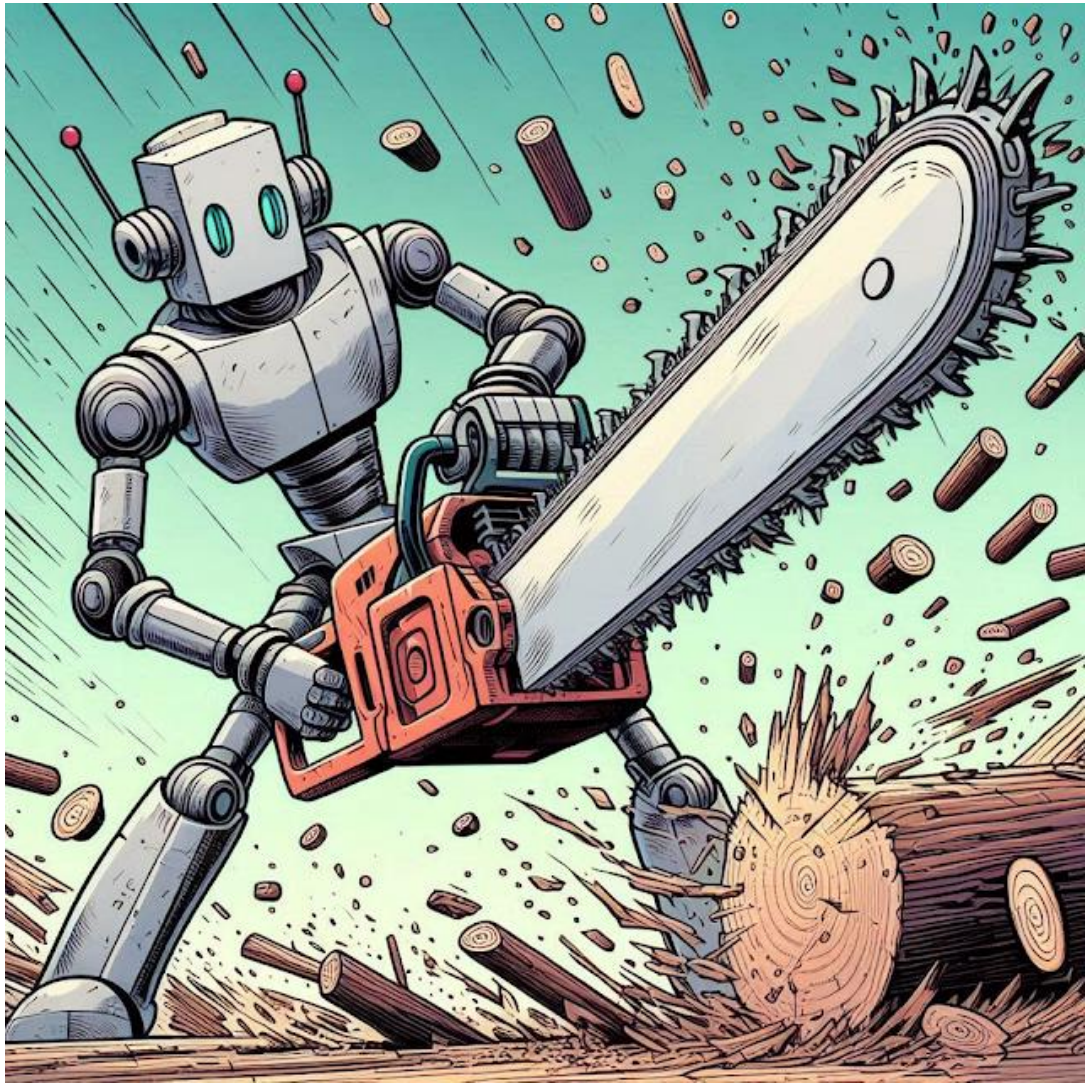
Poor Performance When Predicting Based on Literature Training Data



Using Uncertainty to Guide Experiments

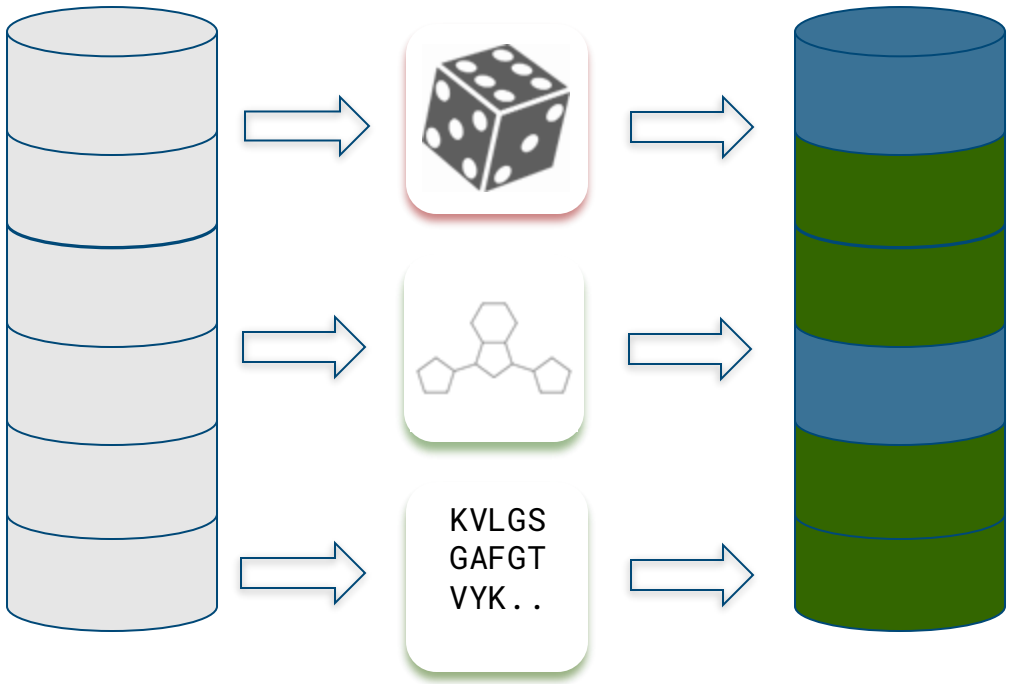


Splitting Datasets

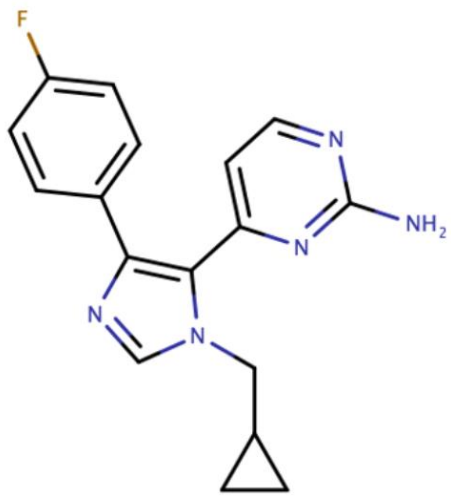


Data Split is Crucial to Assess Generalizability

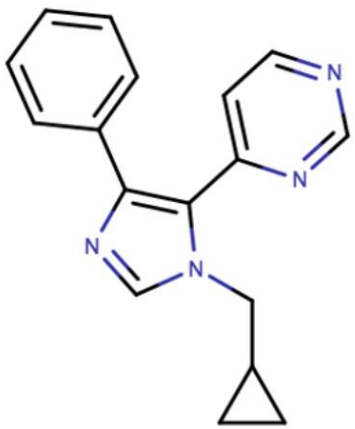
- Train/val/test split
- Random splits overestimate out-of-domain generalization
- Better: Splits based on
  - Ligand scaffold
  - Protein sequence



Scaffold Splits Can Be Problematic

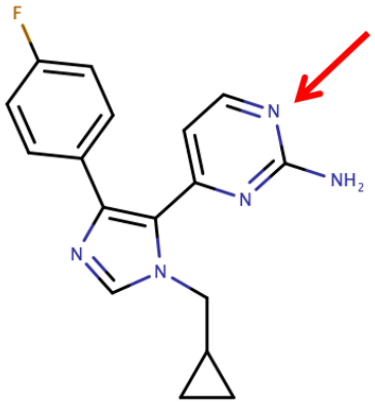


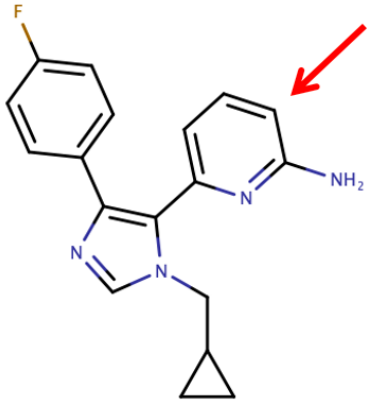
molecule



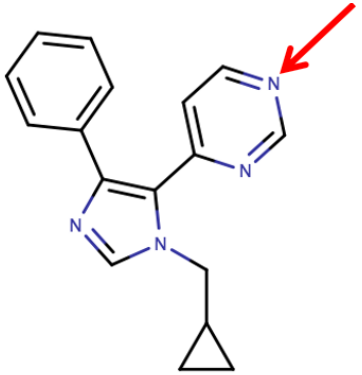
scaffold

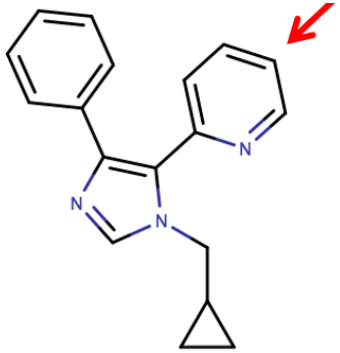
Remove acyclic atoms and leave linkers





Molecules differ by one atom



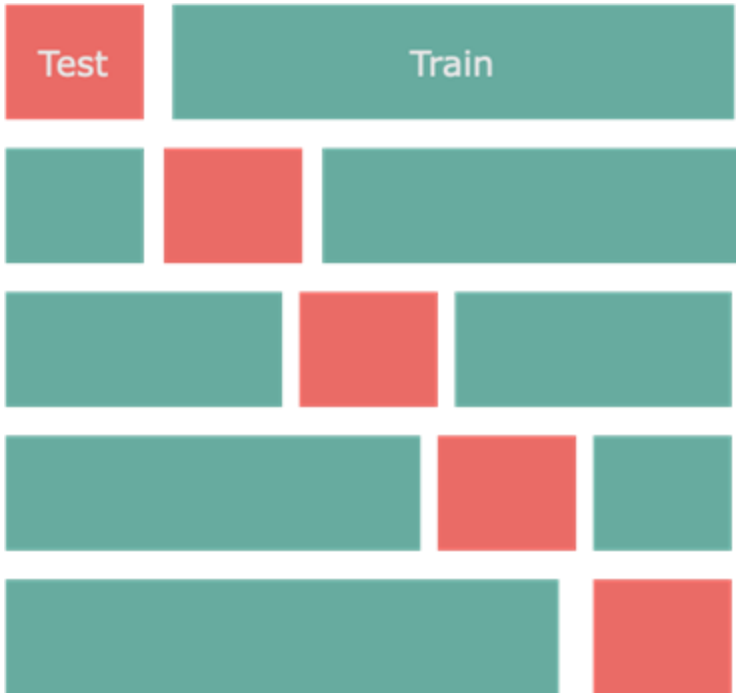


Different scaffolds

Bemis, Guy W., and Mark A. Murcko. "The properties of known drugs. 1. Molecular frameworks." *Journal of medicinal chemistry* 39.15 (1996): 2887-2893.

Grouped k-Fold Split

- Folds are in accordance with groups (scaffold clusters)
- Every data point is part of the test set *exactly once*
- Folds may not be of exact equal size
- Consider labels (stratified split)



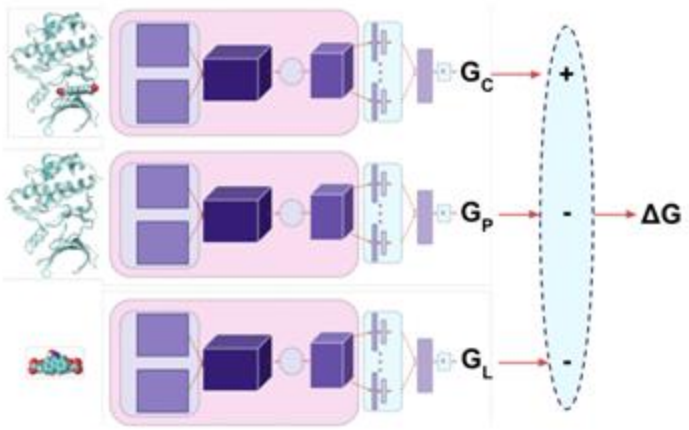
DataSAIL – Data Splitting Against Information Leakage

Developed by Roman Joeres at Prof. Kalinina's lab, Saarland University
uses mathematical optimization to identify the most difficult split
https://github.com/kalininalab/DataSAIL

Key Factors for Success with AI in Drug Discovery (or Anywhere Else)

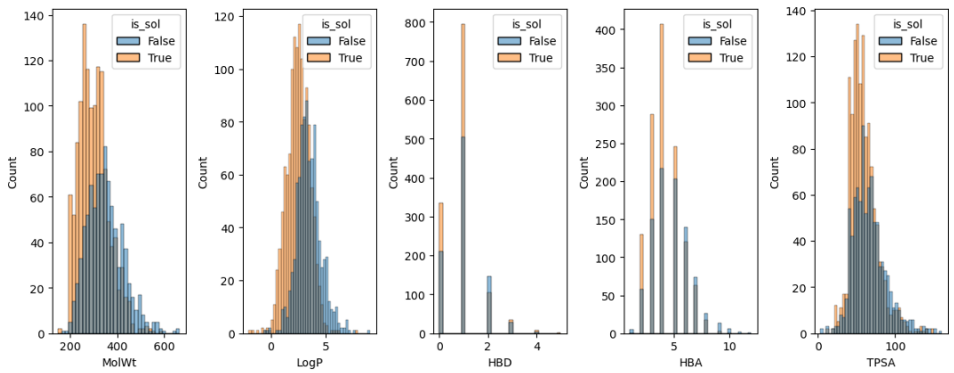
Data

Representation

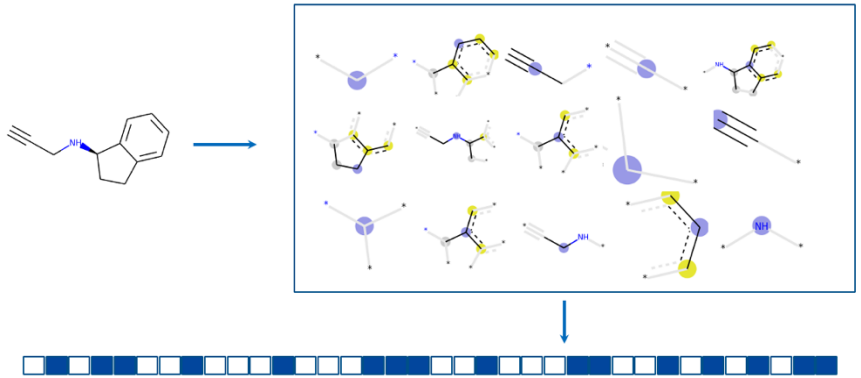


Algorithms

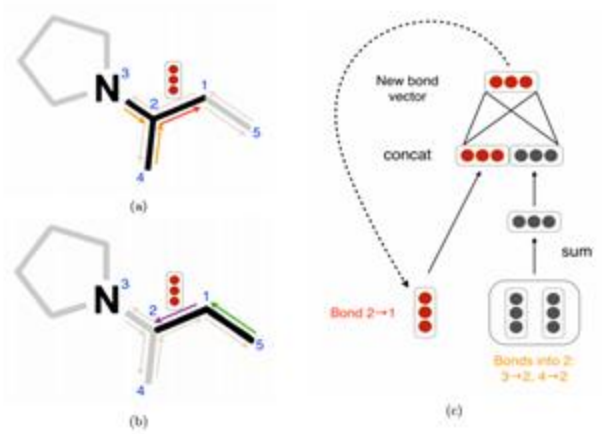
Representation Transforms a Chemical Structure to a Vector



Molecular Properties



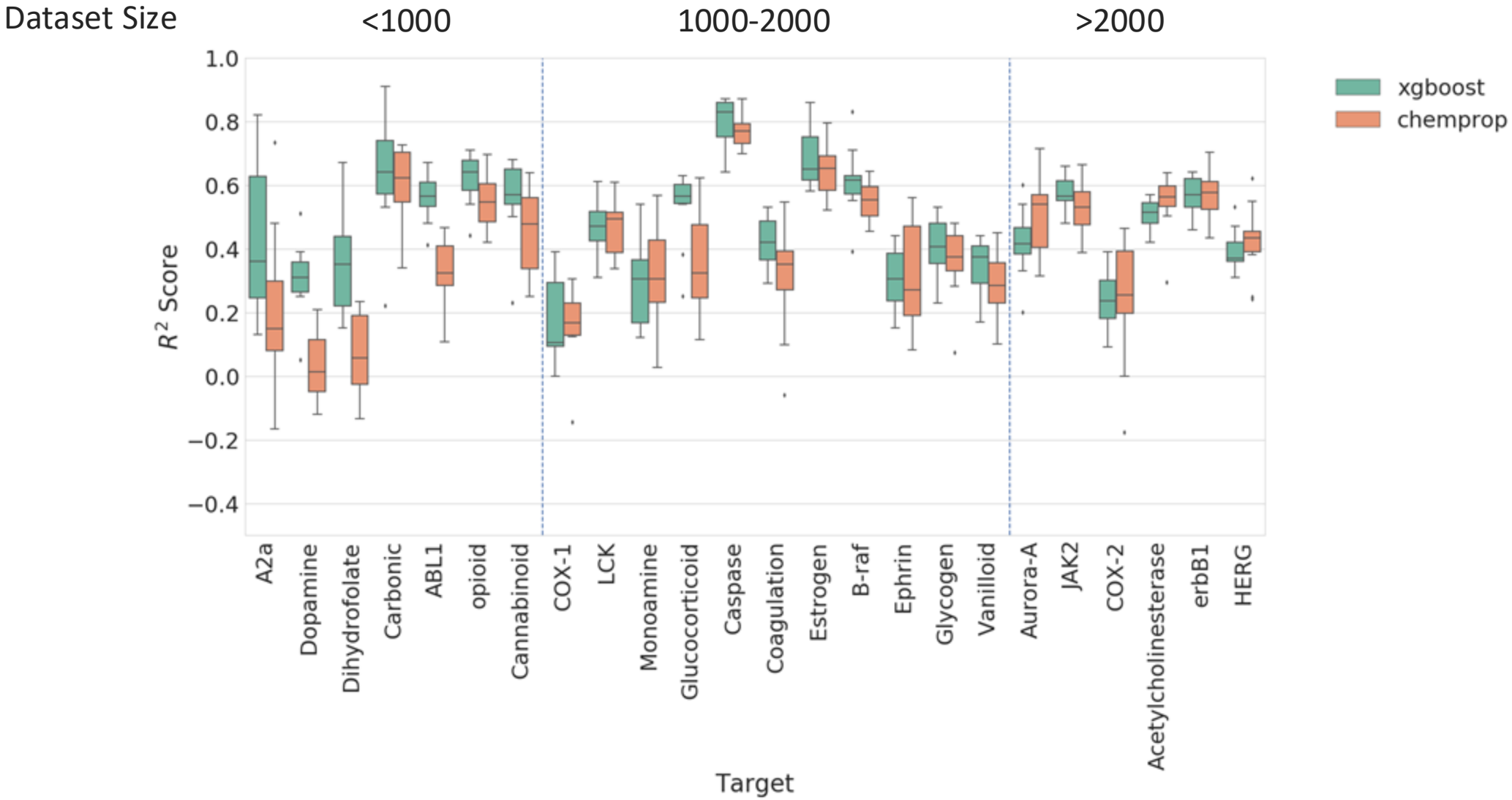
Chemical Fingerprints



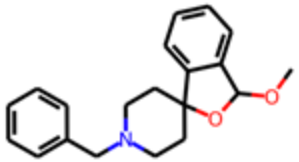
Message Passing Neural Network (MPNN)

Graph Convolutional Neural Network (GCNN)

Are Neural Network Representations Better?



Incorporating 3D into Molecular Machine Learning is an Unsolved Problem



2D single-instance

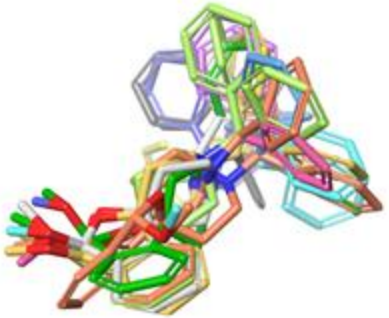
0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...

5.21

3D single-instance

0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...

5.21



3D multiple-instance

0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...
0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0...
0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 1 0...
0 1 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 1 0...
0 1 0 1 0 0 0 1 0 1 1 0 1 0 1 1 0 0 0 1 0...
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0...
0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0...
0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0...
0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 1 0...
0 1 0 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0...
0 1 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0...

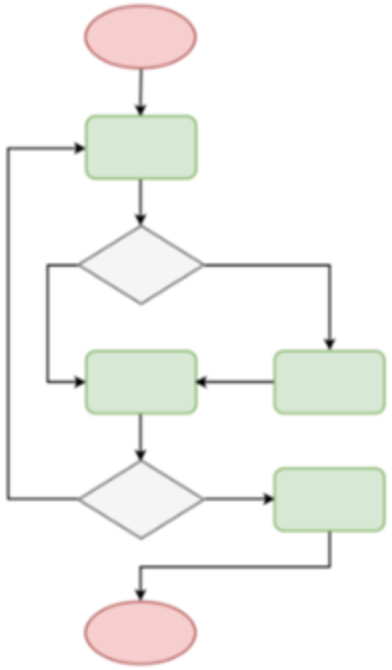
5.21

Key Factors for Success with AI in Drug Discovery (or Anywhere Else)

Data

Representation

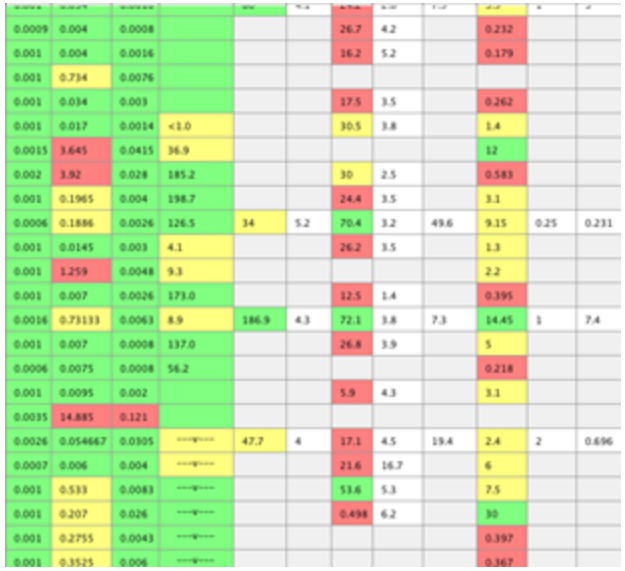
Algorithms



Taking Advantage of a Rapidly Evolving Machine Learning Ecosystem

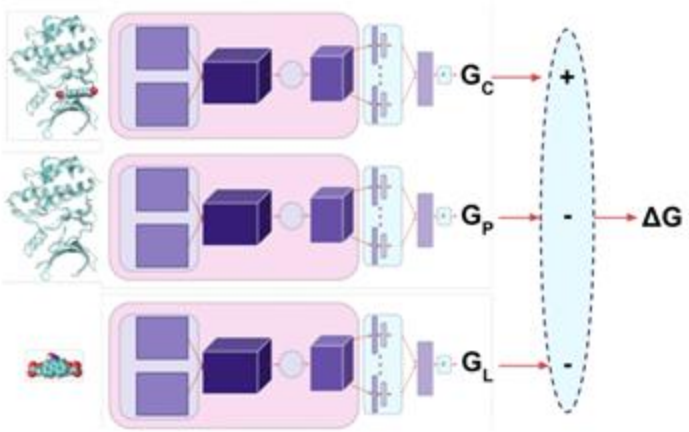
Key Components of Machine Learning in Drug Discovery (or Anything Else)

Data



0.0009	0.004	0.0008				26.7	4.2		0.232		
0.001	0.004	0.0016				16.2	5.2		0.179		
0.001	0.734	0.0076									
0.001	0.034	0.003				17.5	3.5		0.262		
0.001	0.017	0.0014	<1.0			30.5	3.8		1.4		
0.0015	3.645	0.0415	36.9						12		
0.002	3.92	0.028	185.2			30	2.5		0.583		
0.001	0.1965	0.004	198.7			24.4	3.5		3.1		
0.0006	0.1886	0.0026	126.5	34	5.2	70.4	3.2	49.6	9.15	0.25	0.231
0.001	0.0145	0.003	4.1			26.2	3.5		1.3		
0.001	1.259	0.0048	9.3						2.2		
0.001	0.007	0.0026	173.0			12.5	1.4		0.395		
0.0016	0.73133	0.0063	8.9	186.9	4.3	72.1	3.8	7.3	14.45	1	7.4
0.001	0.007	0.0008	137.0			26.8	3.9		5		
0.0006	0.0075	0.0008	56.2						0.218		
0.001	0.0095	0.002				5.9	4.3		3.1		
0.0035	14.885	0.121									
0.0026	0.054667	0.0305	---v---	47.7	4	17.1	4.5	19.4	2.4	2	0.696
0.0007	0.006	0.004	---v---			21.6	16.7		6		
0.001	0.533	0.0083	---v---			53.6	5.3		7.5		
0.001	0.207	0.026	---v---			0.498	6.2		30		
0.001	0.2755	0.0043	---v---						0.397		
0.001	0.3525	0.006	---v---						0.367		

Representation



Algorithms