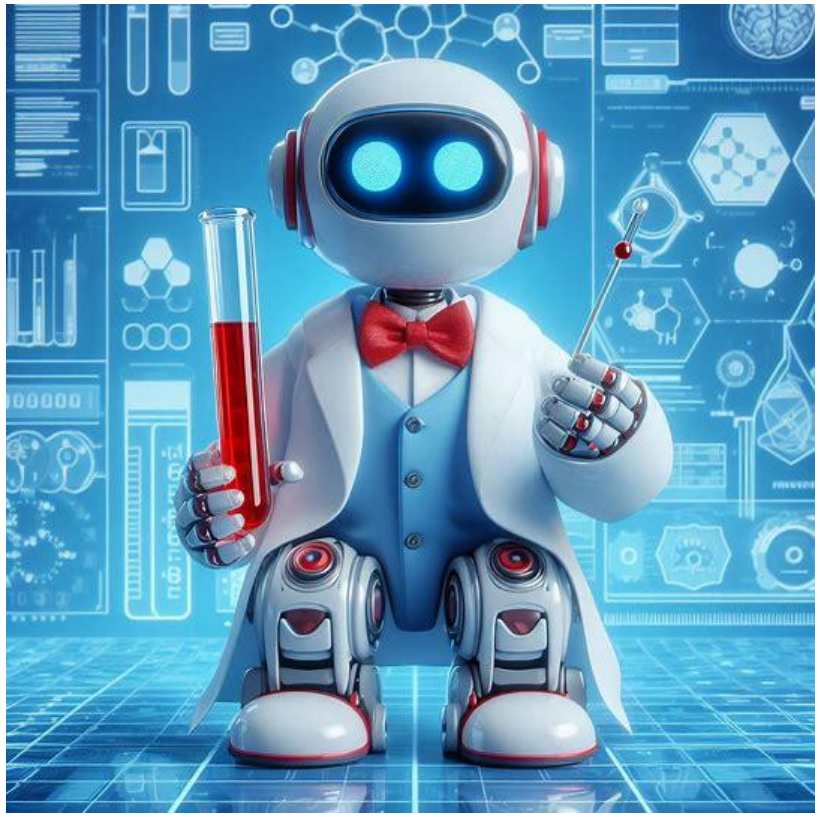


AI in Drug Discovery – An Overview

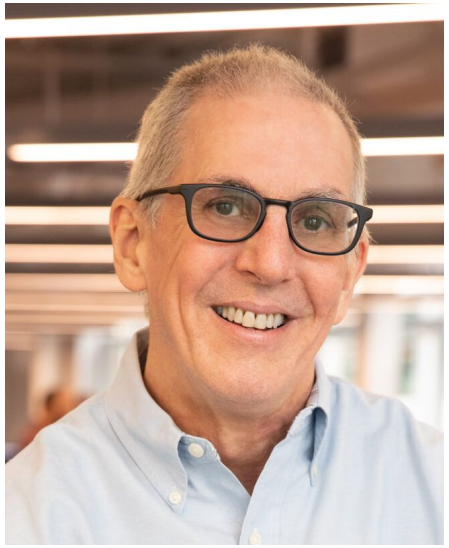
Andrea Volkamer and Pat Walters

September 16, 2024

Session 2 - Data is all you need!

Who we are!

Pat Walters
Relay Therapeutics



Raquel López-Ríos de Castro
Charité Berlin, MSKCC NYC

Michael Backenköhler
Saarland University

Andrea Volkamer
Saarland University

What we will do today

Session 1 - 1:30 - 2:30 pm

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

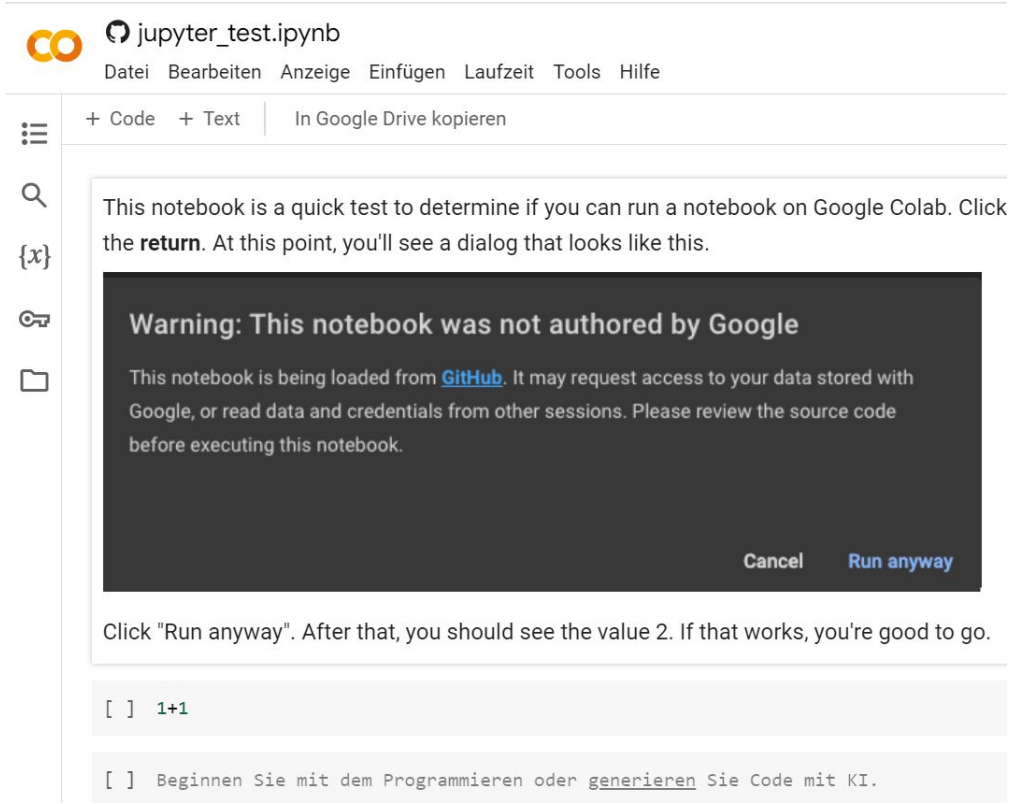
Session 2 - 3:00 - 4:00 pm

- Importance of data quality for AI and ML
- Data (pre)processing
- Exploratory data analysis
- Applicability domains

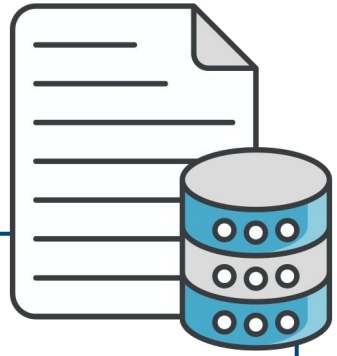
Session 3 - 4:30 - 5:30 pm

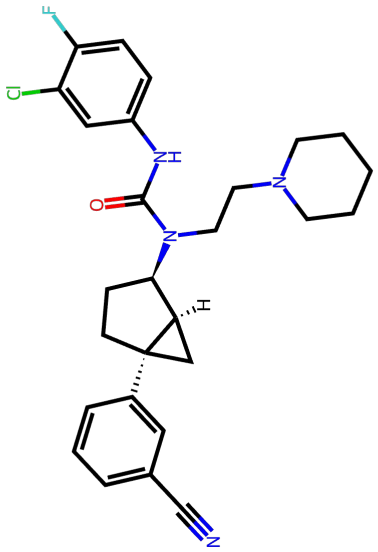
- AI in practice
- Molecule generation
- Active learning

Lectures supported by hands-on sessions …



The Path From Molecules to Properties

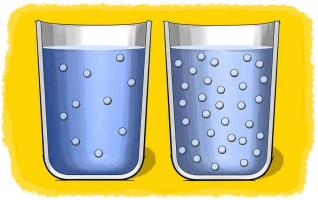


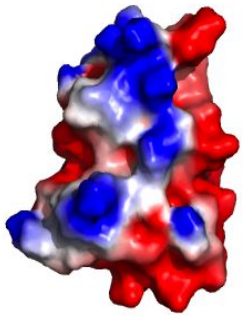


Supervised Learning (SL)

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

∞

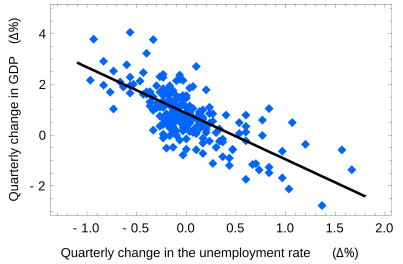
Limited

Representative

Just needs the right model!

Model

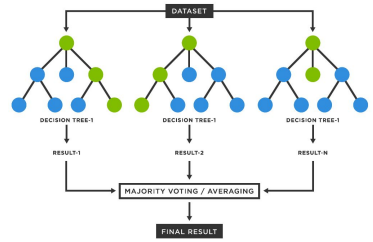
Simple



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

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

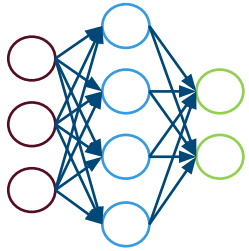
Moderate



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

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

Complex



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

$$P \propto W, b$$

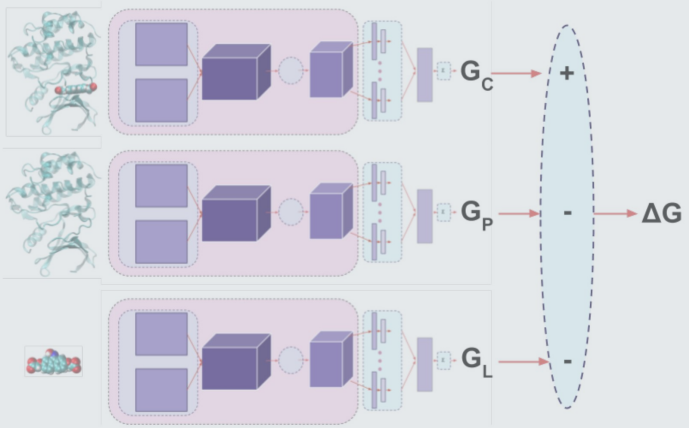
P = # parameter

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

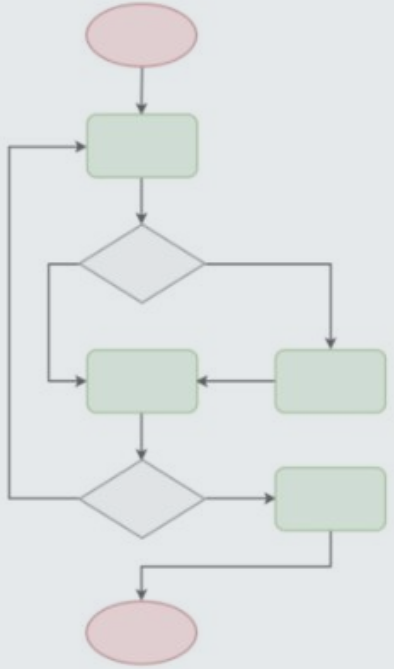
Data

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



Where Machine Learning Excels ⟷ Pharmaceutical Data Properties

(Ultra)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

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

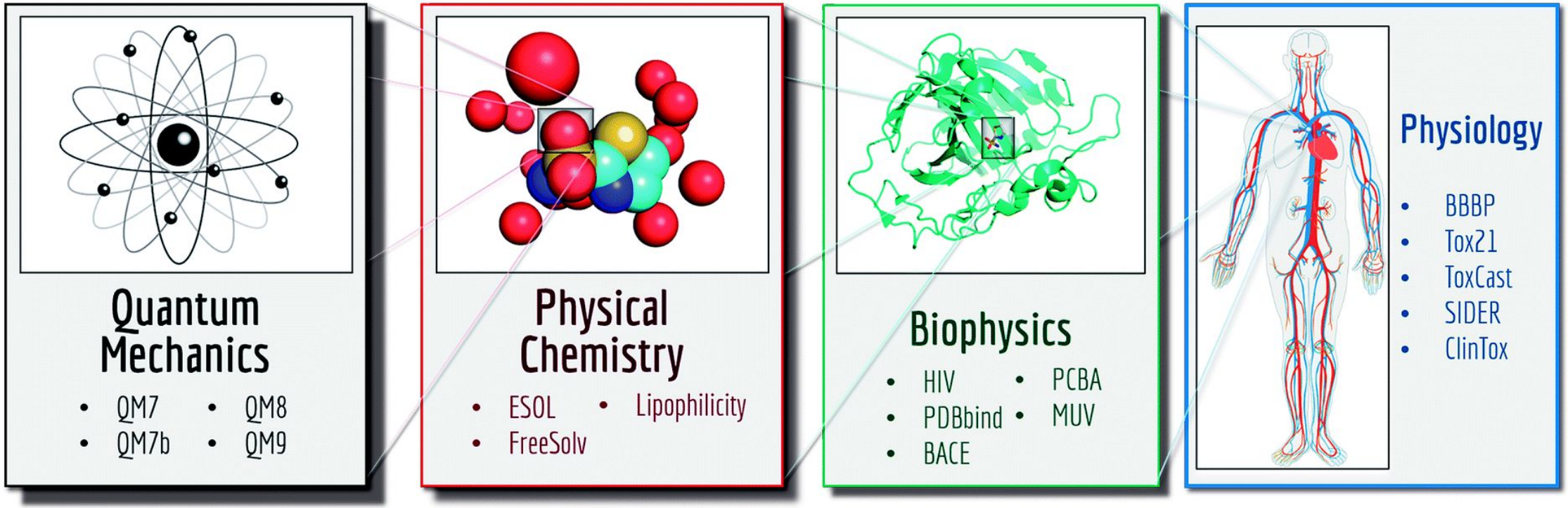
Data has a limited dynamic range
- Often spans only 2 or 3 logs
- Small dynamic range combined with experimental error makes significant correlations difficult

Even data from the “same” assay can be heterogeneous
- PK data measured with different doses, formulations
- Response can vary with operator, equipment, lab

Data covers a limited chemical space
- Even global models can be local

Small Molecule Benchmark Datasets: MoleculeNet

Common molecule benchmarks for molecular property prediction:
16 datasets divided into 4 categories.



https://moleculenet.org/

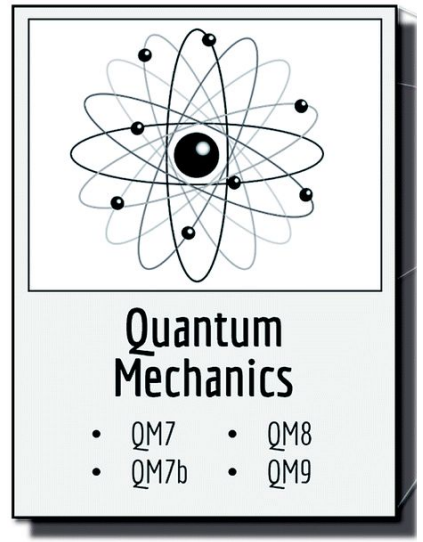
Small Molecule Benchmark Datasets: e.g., QM9 and Tox21

Quantum Mechanics data: QM9 → regression tasks

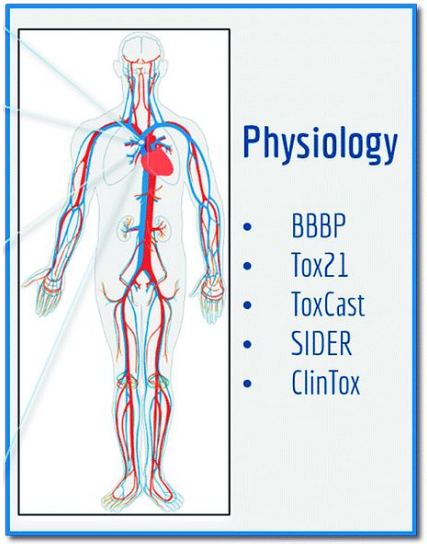
- ~134k molecules (SMILES, 3D) with <= 9 heavy atoms
- 12 properties: Geometric, energetic, electronic, and thermodynamic properties of DFT-modelled small molecules
- Commonly used due to its comparably large size

Tox21 → classification

- Toxicity data with molecules measured against toxicity-relevant endpoints
- ~8k molecules
- 12 tasks, here biological targets, including nuclear receptors and stress response pathways



https://moleculenet.org/



CAREFUL! See Pat's BlogPost
https://practicalcheminformatics.blogspot.com/2023/08/**we-need-better-benchmarks-for-machine**.html

Benchmark Data Sets: Therapeutic Data Commons (TDC)

Check for updates

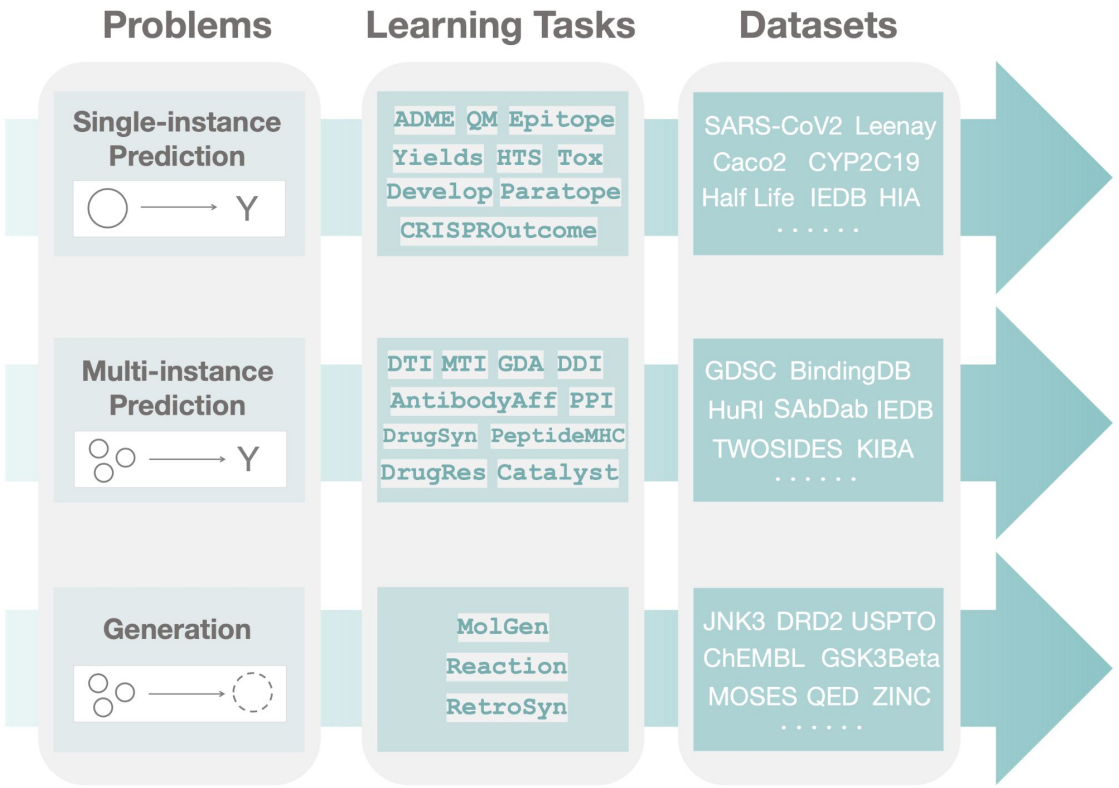
comment

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

Nat Chem Biol (2022). https://doi.org/10.1038/s41589-022-01131-2



CAREFUL! See Pat's BlogPost

https://practicalcheminformatics.blogspot.com/2023/08/**we-need-better-benchmarks-for-machine**.html

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/

Small Molecule Datasets (unlabelled): ZINC and GuacaMol

ZINC (https://pubs.acs.org/doi/full/10.1021/acs.jcim.5b00559)

- Ultralarge collection of purchasable, "drug-like" molecules without target value
- ZINC15: ~120M molecules
- ZINC20: ~1.4 billion molecules,

GuacaMol (https://pubs.acs.org/doi/10.1021/acs.jcim.8b00839)

- Large collection of molecules from ChEMBL 24 without target value
- ~1.5M molecules
- Published training, test and validation set → for benchmarking

➔ **Commonly used benchmark for pretraining (LLMs) and generative models**

(Benchmark) Data Quality – Technical Issues

Inspired by Pat's BlogPost
we-need-better-benchmarks-for-machine

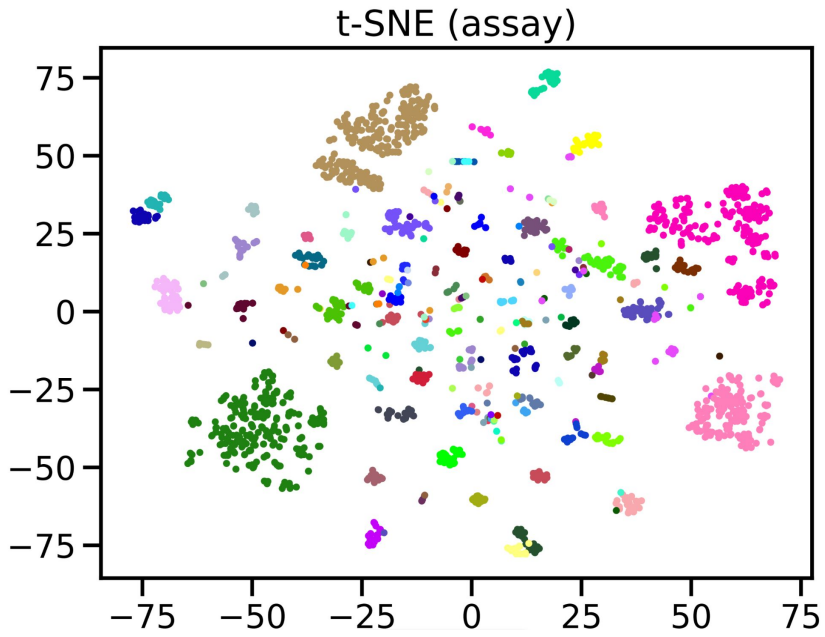
- **Valid structures**
  - Invalid SMILES, not be parsed by standard cheminformatics tools (e.g. rdkit)
- **Consistent chemical representations**
  - Standardized based on accepted conventions
- **Stereochemistry**
  - We need to know the correct structure
- **Consistent measurements**
  - Ideally experiments performed in same lab, but at least same standards
- **Realistic dynamic range and cutoffs**
  - Typical range: solubilities between 1 and 500 µM (cut at both ends '<' an '<')
  - Sensible cut-off choices for classification
- **Clear definitions of training, validation, and test sets**
  - Data leakage: random, scaffold and cluster splits
  - IDs need to be given in benchmark
- **Data curation errors**
  - Duplicate structures with differing labels

***Not so good* examples from MoleculeNet**

- 11 SMILES with uncharged tetravalent nitrogen atom in BBB
- 59 beta-lactam antibiotics in BBB, carboxylic acid in three different forms
- 71% of the molecules in BACE have at least one undefined stereocenter
- BACE data collected from 55 papers
- ESOL spans 13 logs, easy to separate

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

Applying ML to Drug Design with Trust

Do I have the right data?

Can I trust this prediction?

Is the model interpretable?

(Skipped due to time constraints)

- Model validation and data splits
- Model generalizability and applicability

Is My Training Data Relevant?

Machine learning is all about labeling things using examples

If I train on this?

Can I predict this?

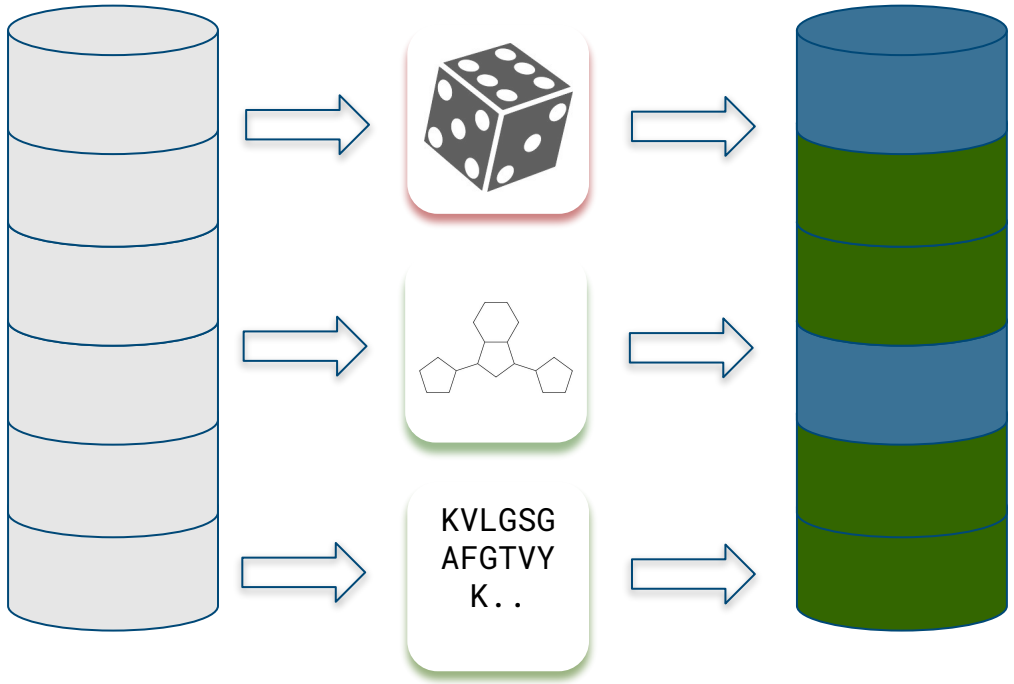
Example Scenarios

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

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

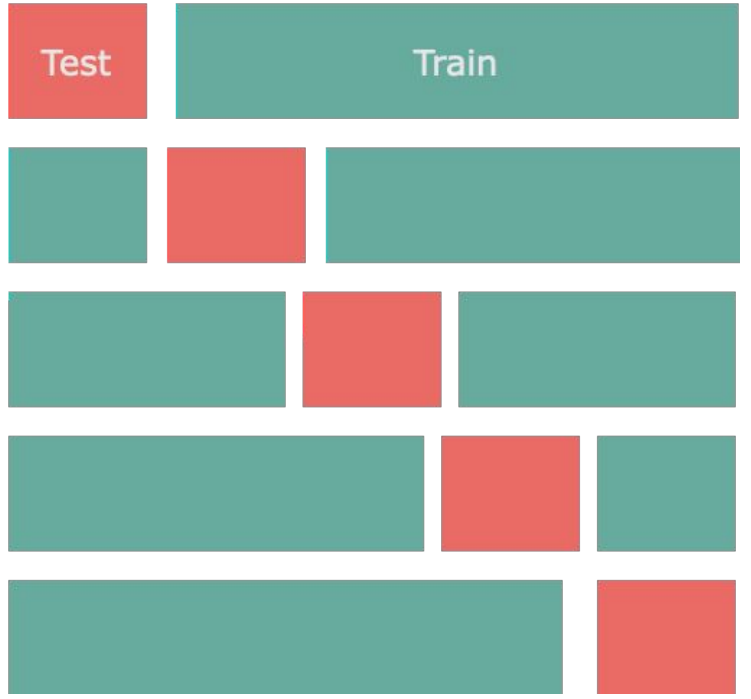
Data Split is Crucial to Assess Generalizability

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

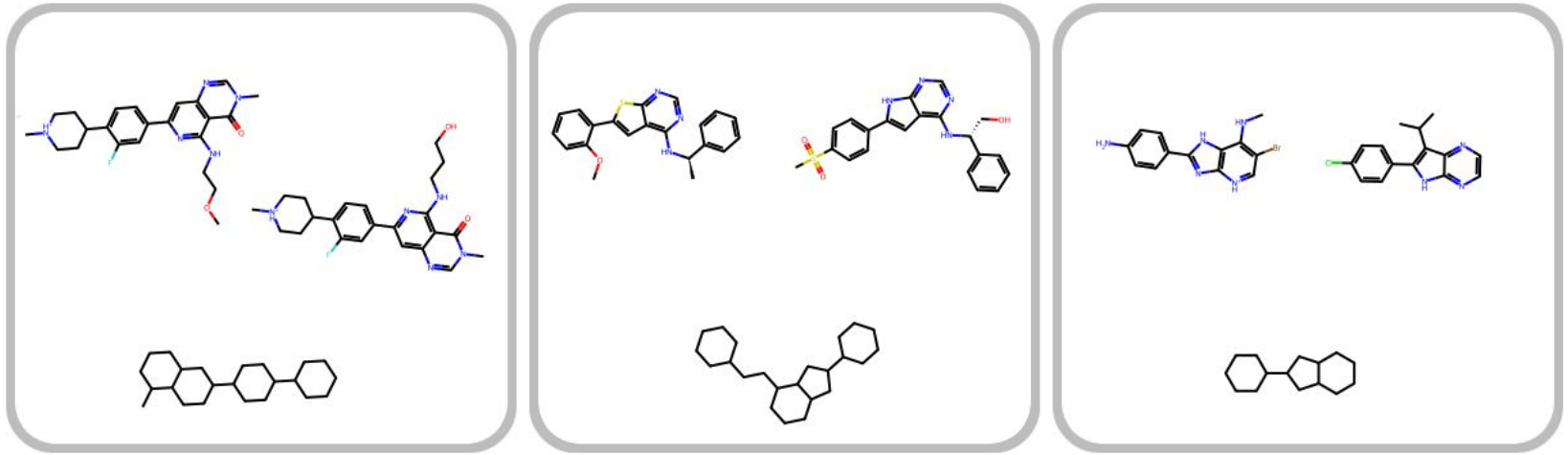


Grouped k-Fold Split

- 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)



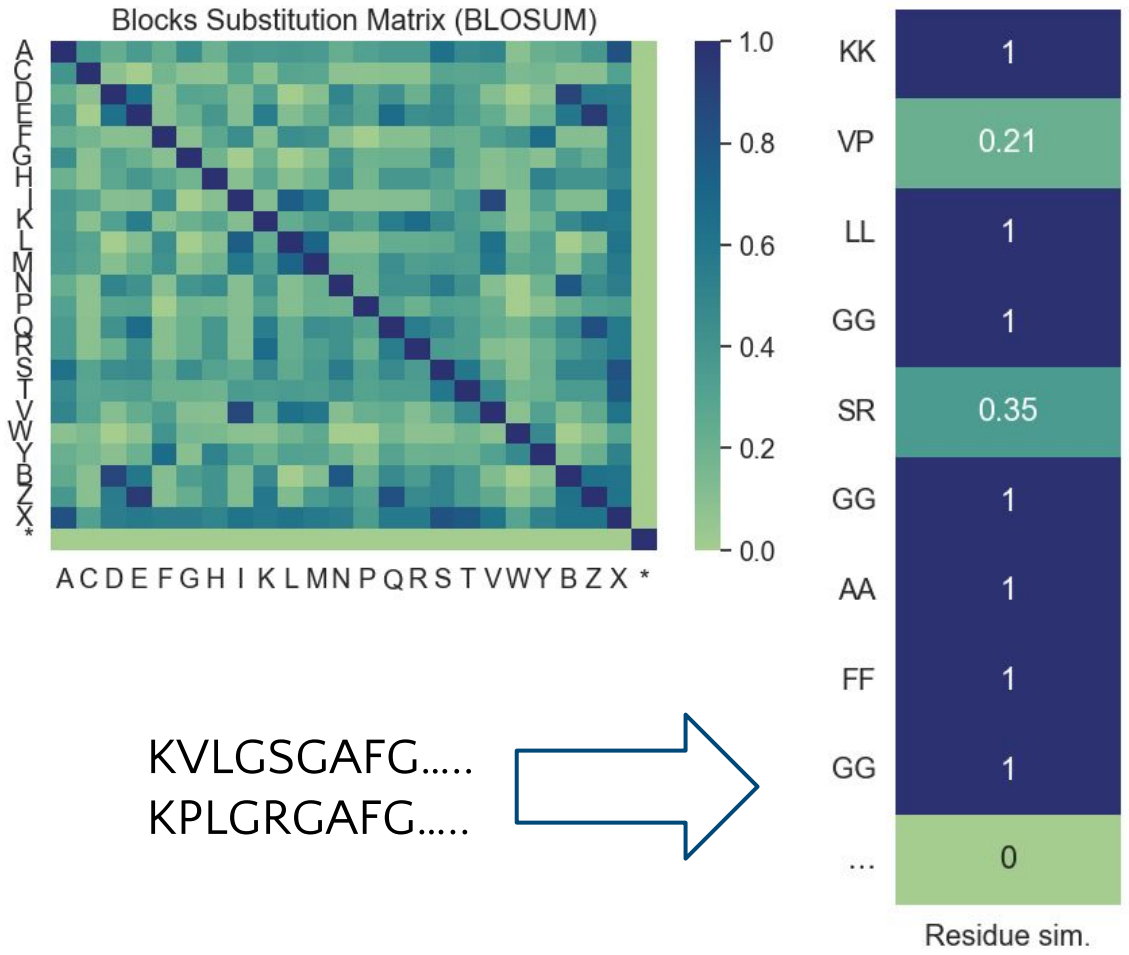
Data Splits – Scaffold Split



- Avoid data leakage over structurally similar ligands
- Group by generic scaffold *(or similarity)*
- Ensure each group only in one of {train, test, val}

Data Splits – Target-Based Split

- Cluster pocket sequences based on BLOSUM-similarity
- Generate splits by randomly assigning clusters to {train, val, test}



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

Applicability and Uncertainty of Machine Learning Models

Do I have the right data?

Can I trust this prediction?

Is the model interpretable?

Conformal Prediction

Bootstrap

Gaussian Process Regression

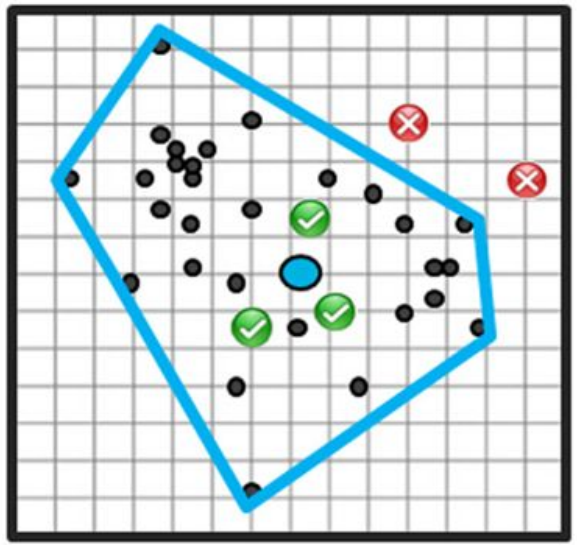
Jackknife

Monte Carlo Dropout

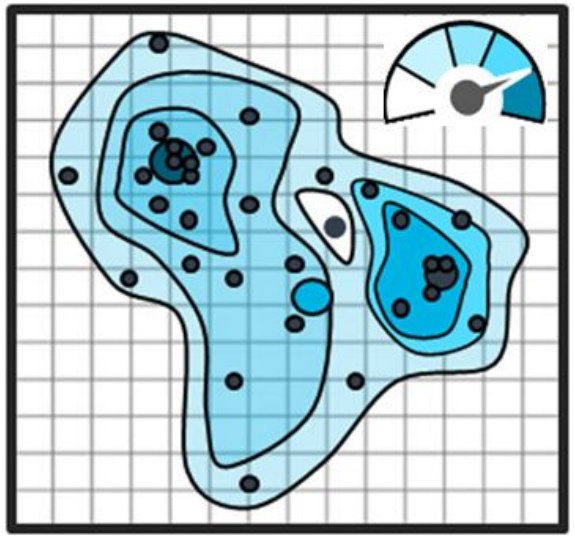
Applicability Domain

Definition: The applicability domain of a (Q)SAR/ML model is the response and chemical structure space in which the model makes predictions with a given reliability.

1. Applicability:
Interpolation within the chemical structure space



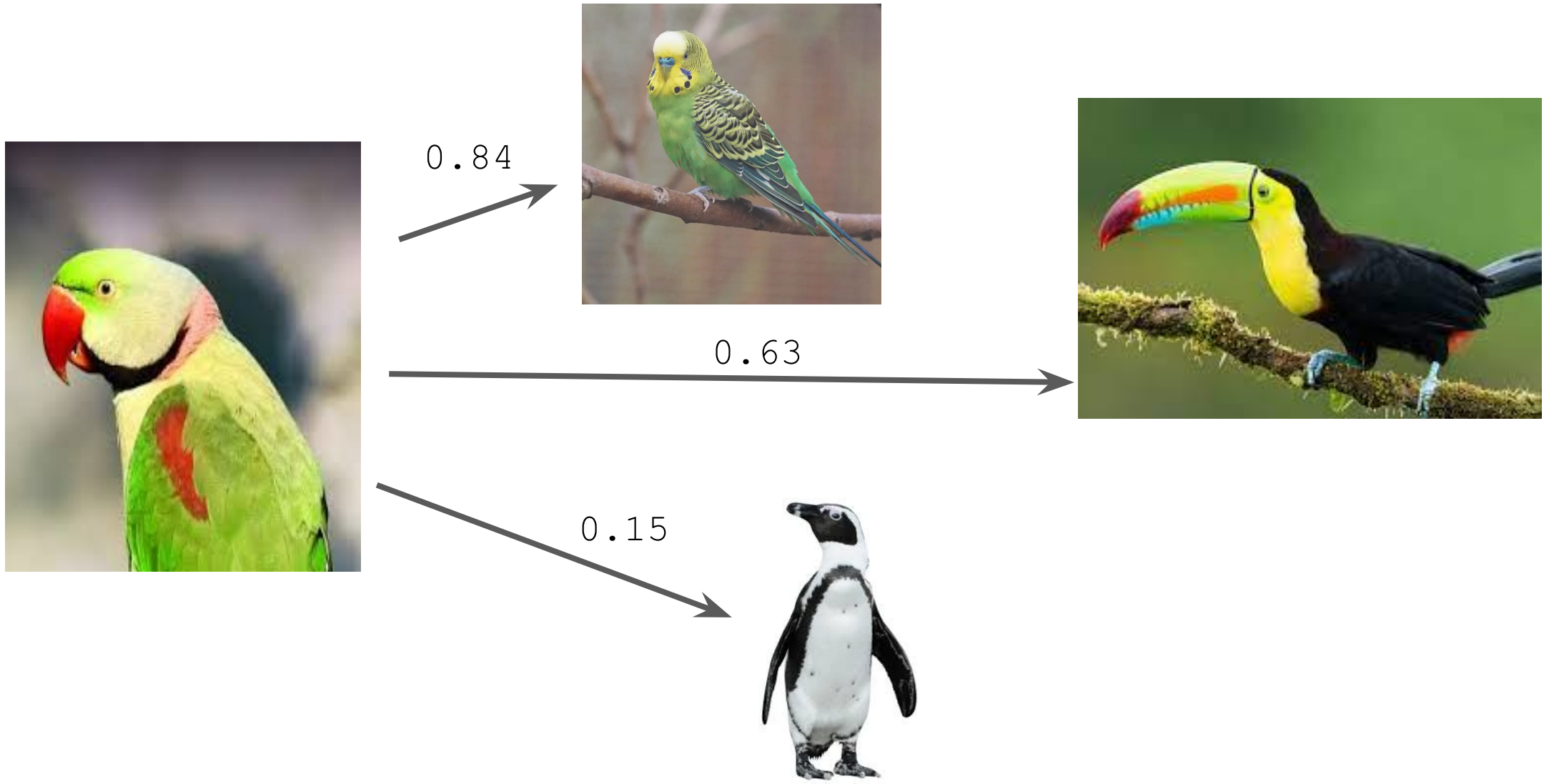
2. Reliability:
Density of knowledge around the query chemical structure



3. Decidability:
Distance of the query chemical compound to the decision boundary of model

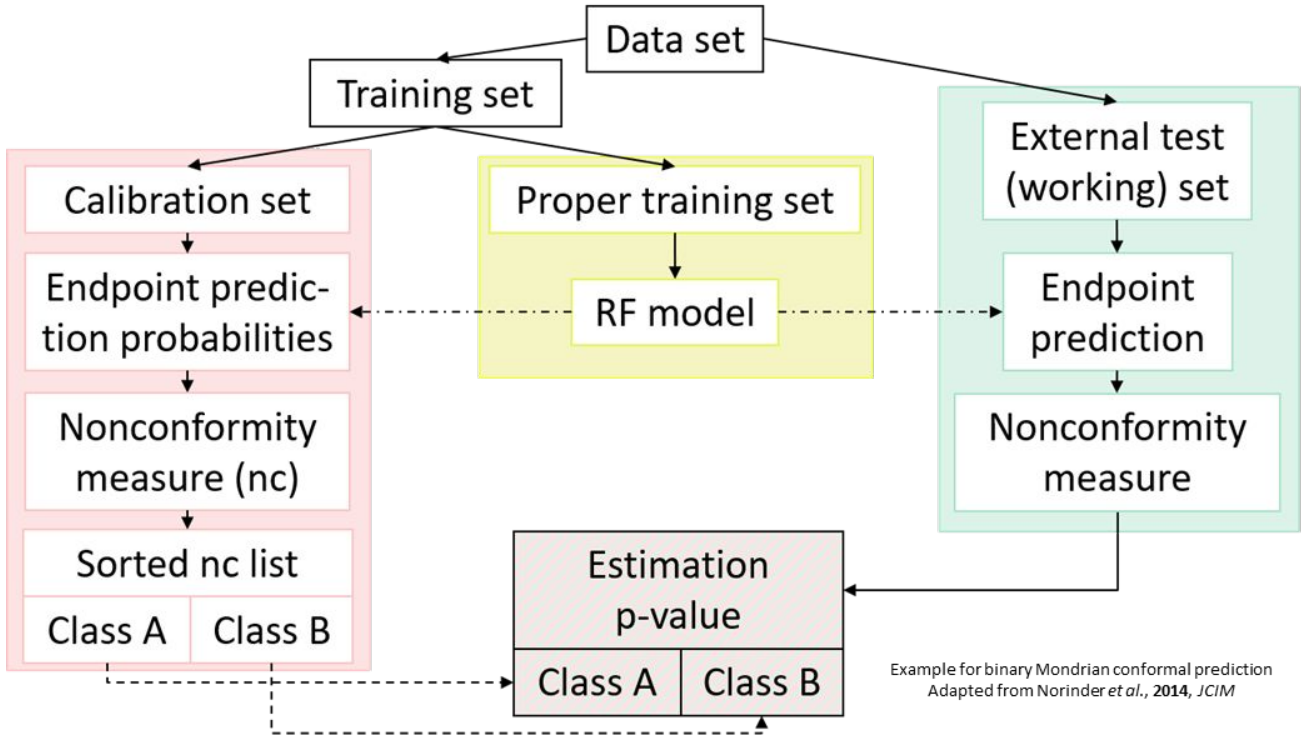
Figures taken from: Hanser, *et al.*, *SAR and QSAR in Environmental research*, **2016**, 27:11, 865-881

Using Similarity to Assess Applicability

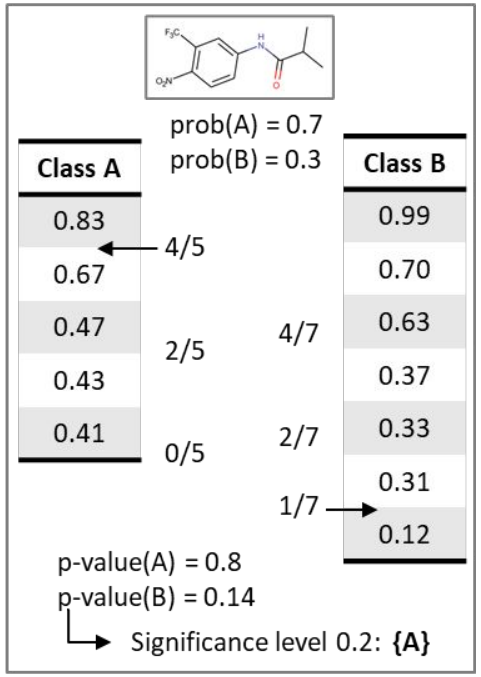


Conformal Prediction Framework Based on ML

- Statistically valid on a given confidence level (if exchangeability is given)
- Additional calibration step: Compare predictions to those previously seen (p-value)
- Output prediction sets in binary classification: **{A}, {B}, {A,B}** or **{ }**
- Validity = % of correct classifications
- Efficiency = % of single class predictions



Example for binary Mondrian conformal prediction
Adapted from Norinder *et al.*, **2014**, *JCIM*



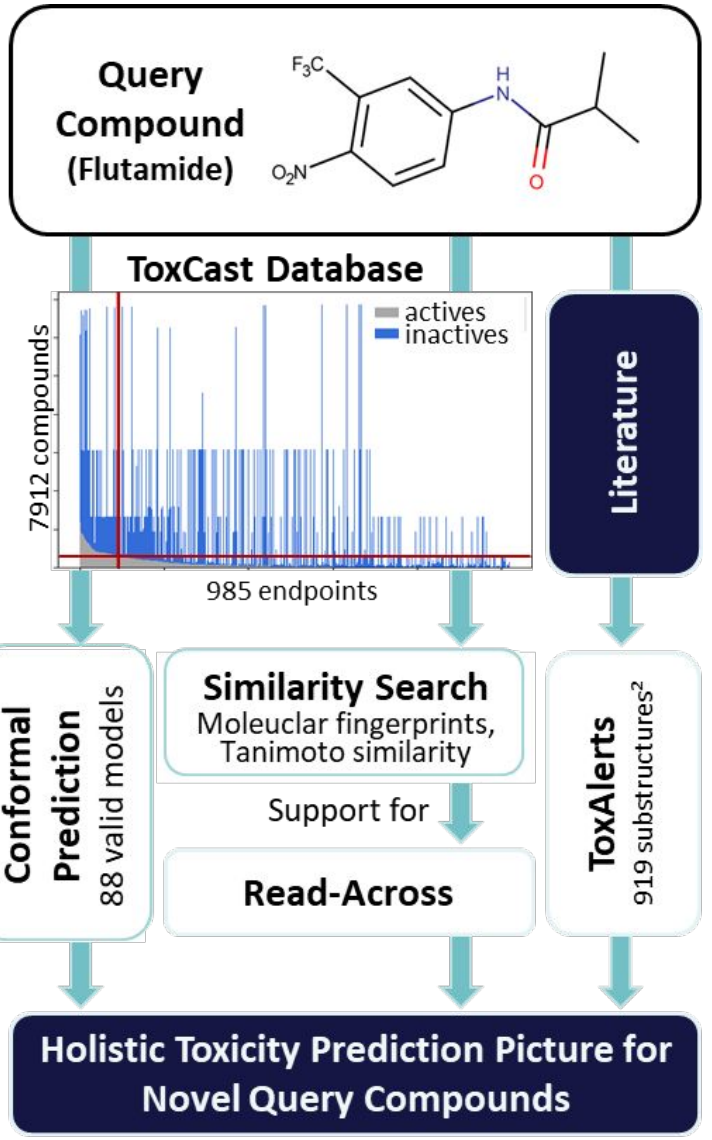
Know-Tox – Predictions with Confidence

- ToxCast data set
  - ~ 8000 compounds * 1000 endpoints (sparse)
  - Pharmaceutical, pesticides, environmental chemicals
  - Cell cycle, steroid receptors, cytotoxicity, …

Can we apply the model to new (BASF) data?

- Conformal prediction
  - Case study: Antiandrogen activity (AA)
  - Validity, efficiency and accuracy like other studies

	Dataset	Efficiency	Accuracy (SCPs) all	cl.1	cl.0	# toxic/ non-toxic
cross-validation	ToxCast AA	0.87	0.78	0.80	0.78	868/5842
external	Norinder[3]	0.79	0.68	0.70	0.67	160/201
In-house	BASF (I)	0.94	**0.56**	**0.97**	**0.07**	280/254



https://github.com/volkamerlab/knowtox_manuscript_SI

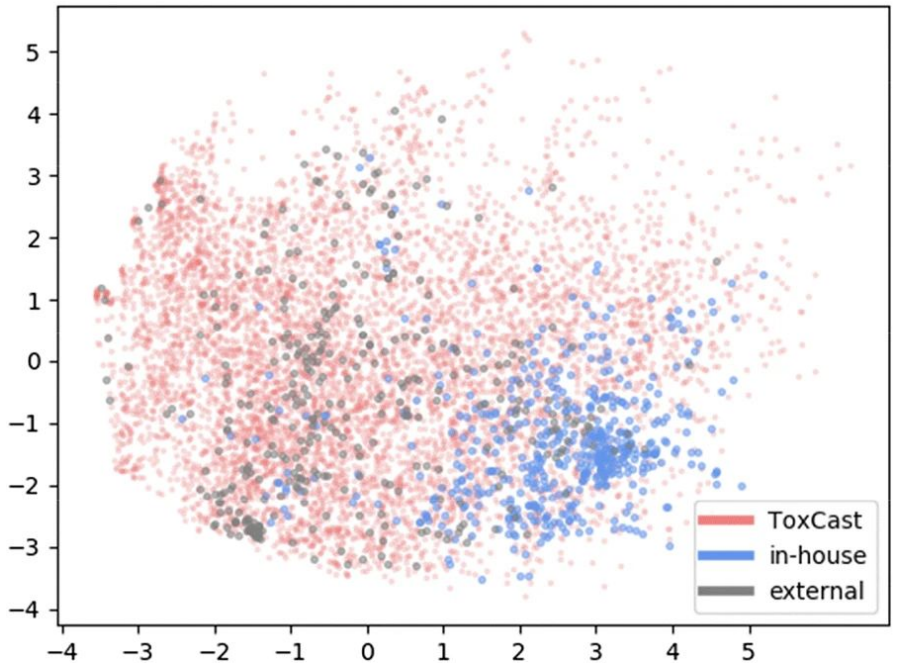
Know-Tox – Predictions with Confidence

- ToxCast data set
  - ~ 8000 compounds * 1000 endpoints (sparse)
  - Pharmaceutical, pesticides, environmental chemicals
  - Cell cycle, steroid receptors, cytotoxicity, …

Can we apply the model to new (BASF) data?

- Conformal prediction
  - Case study: Antiandrogen activity (AA)
  - Validity, efficiency and accuracy like other studies
  - kNN normalization & balancing
  - Less efficient, but higher accuracy

	Dataset	Efficiency	Accuracy (SCPs) all	cl.1	cl.0	# toxic/ non-toxic
cross-validation	ToxCast AA	0.87	0.78	0.80	0.78	868/5842
external	Norinder[3]	0.79	0.68	0.70	0.67	160/201
In-house	BASF (I)	0.94	**0.56**	**0.97**	**0.07**	280/254
	BASF (II)	**0.20**	**0.75**	**0.80**	**0.71**	280/254

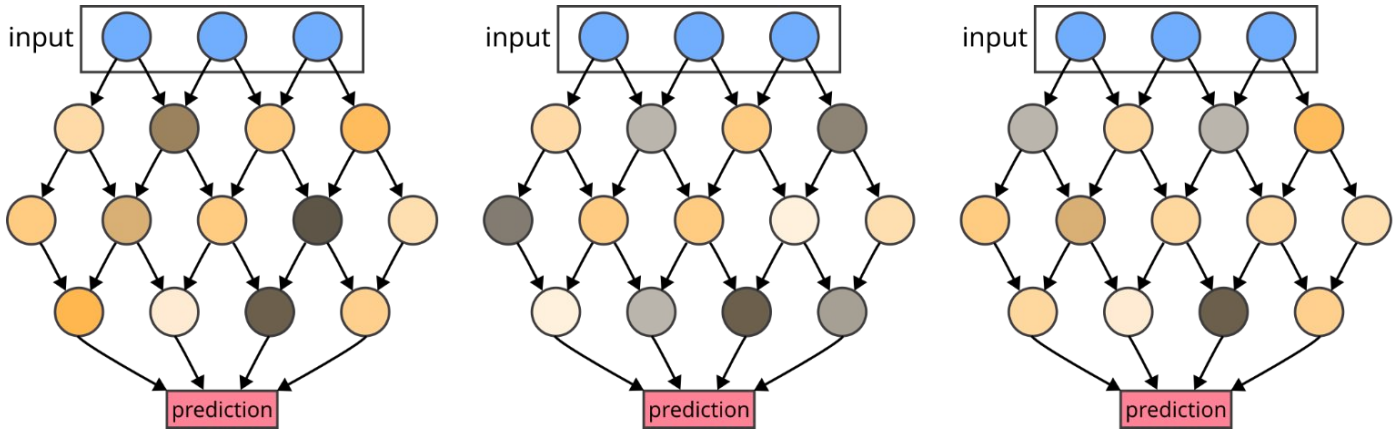


Descriptor space of a 2-component PCA trained on ToxCast-AA data

Morger, A., *et al. J Cheminform* 12, 24 (**2020**)
https://github.com/volkamerlab/knowtox_manuscript_SI

Using Uncertainty to Assess Model Confidence

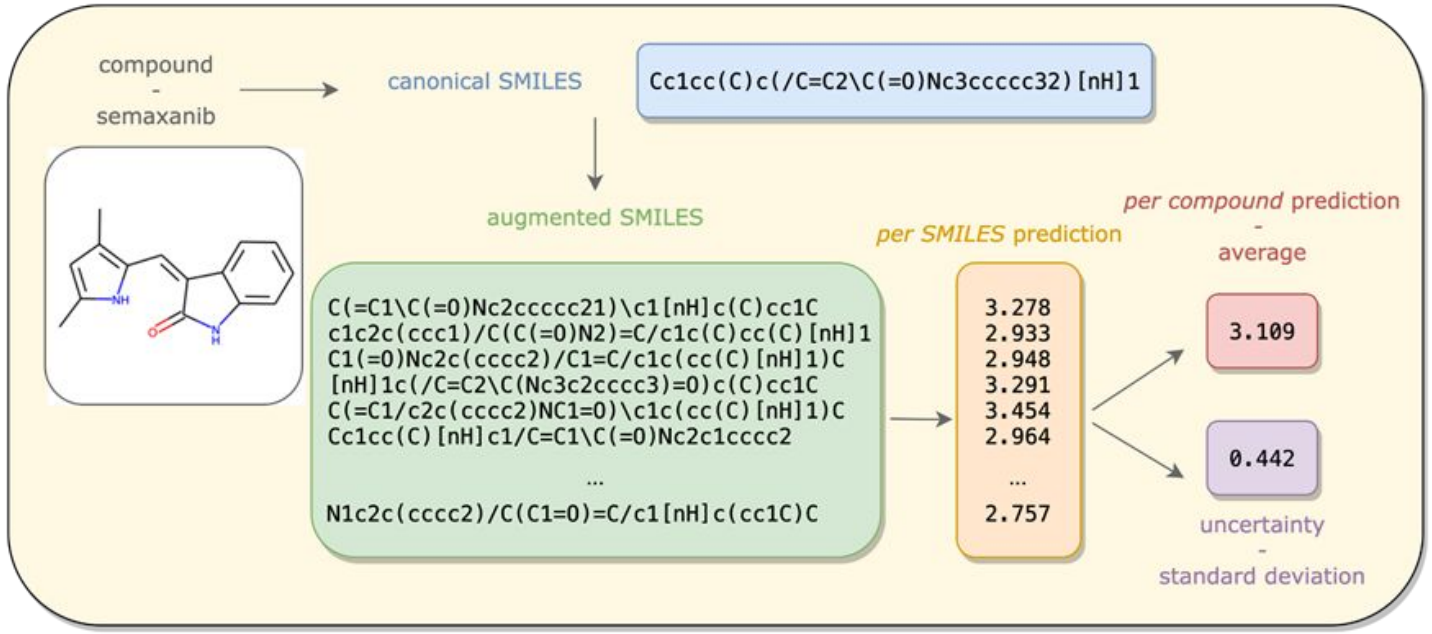
- **Single deterministic methods**
  - Classification setting: predicted class probabilities
  - Secondary model trained to predict uncertainty for an already trained model
- **Ensemble methods**
  - Ensemble of similar but different models -> Variance as uncertainty
  - Varying random seeds, resulting in different weights
  - Varying the training data (*bagging*)
- **Training- and test-time data augmentation**
  - Augmeted set of points using stochastic noise or domain knowledge

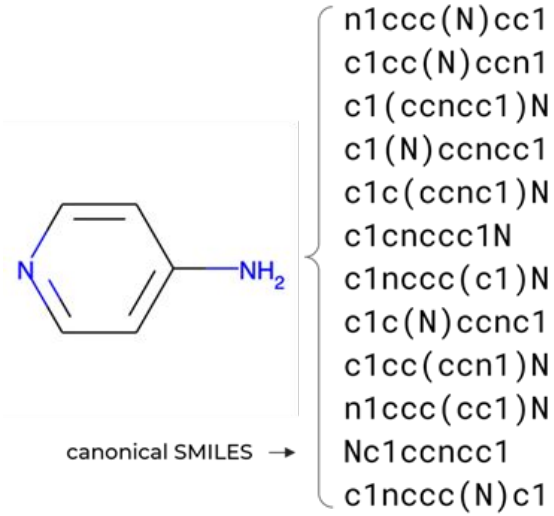


https://github.com/volkamerlab/teachopencadd

SMILES Augmentation Improves Model Performance

Maximizing molecular property prediction performance with confidence estimation using SMILES augmentation and DL



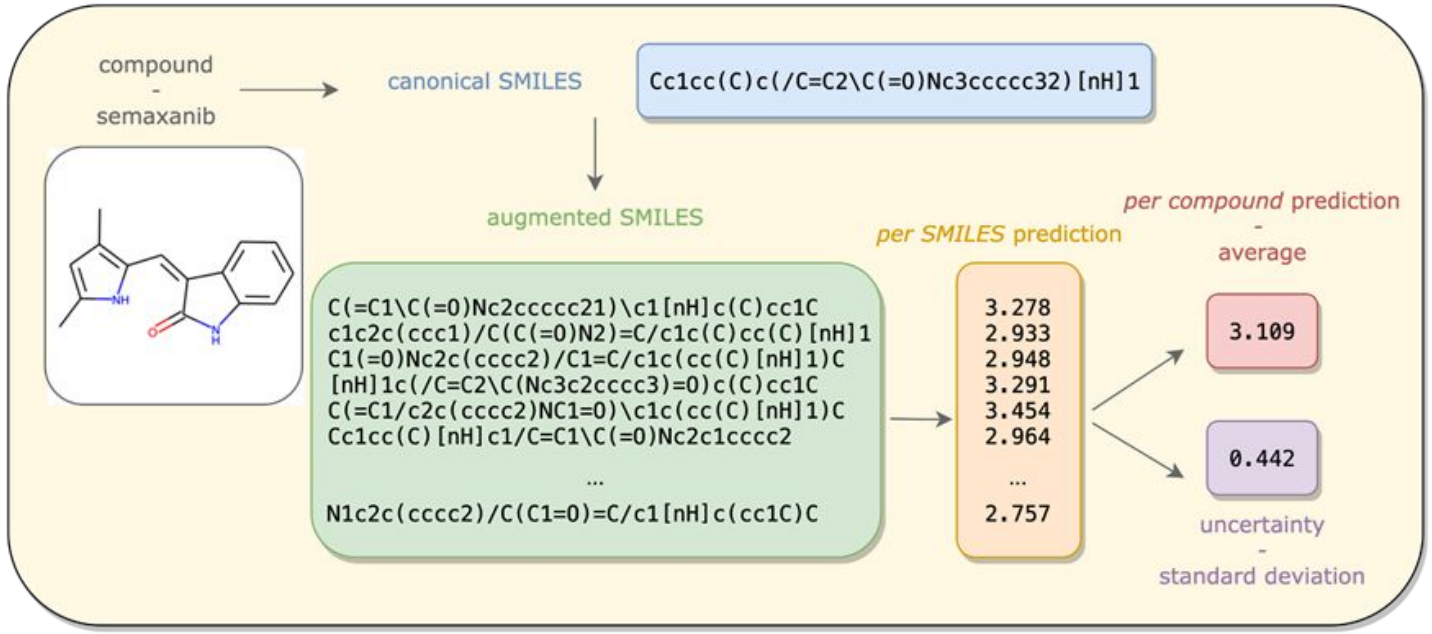


Kimber *et al., AI in Life Sciences,* **2021**, 1, 100014
https://github.com/volkamerlab/maxsmi

CBI CENTER FOR BIOINFORMATICS

SMILES Augmentation Improves Model Performance

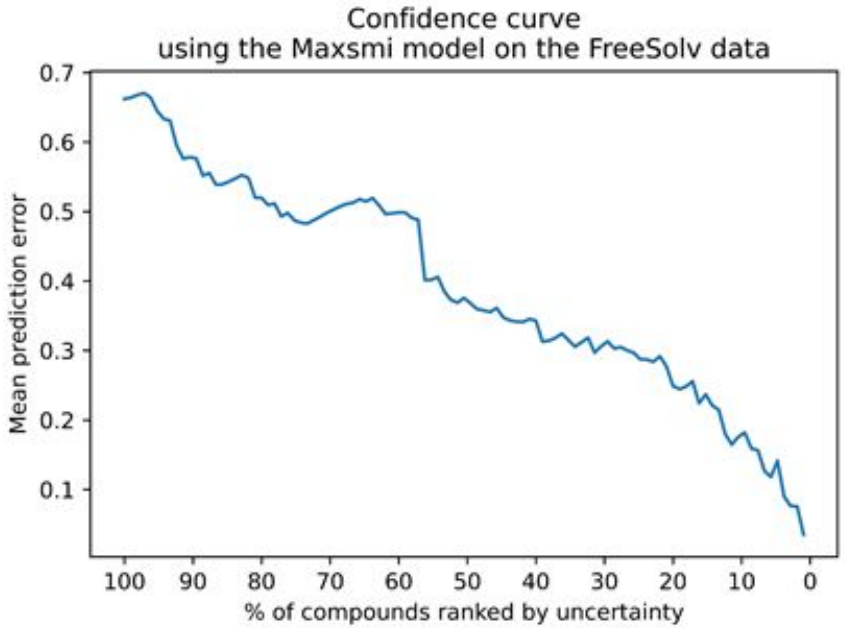
Maximizing molecular property prediction performance with confidence estimation using SMILES augmentation and DL



Kimber *et al., AI in Life Sciences,* **2021**, 1, 100014
https://github.com/volkamerlab/maxsmi

Uncertainty prediction

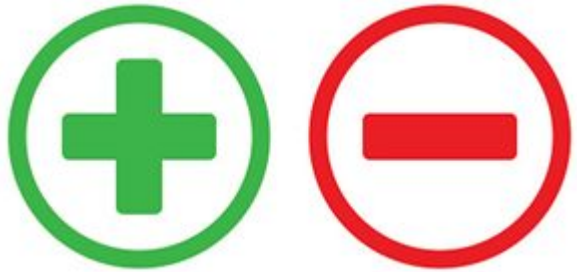
More confident model implies smaller prediction error



CBI CENTER FOR BIOINFORMATICS

Take Home Messages

- More data available
- Increased computational power
- Advances in neural network algorithms
- Open-source libraries



- True predictive power
- Applicability and reliability
- Interpretability

- **Data quality, relevance and proper processing are key**
- **Proper model evaluation, always consider baselines when applying ML/DL**
- **Check applicability of your model, ideally prospective studies**
- **Open source and reproduibility**

TeachOpenCADD in a Nutshell

https://github.com/volkamerlab/teachopencadd

What? Pipelines for common CADD* tasks using open resources

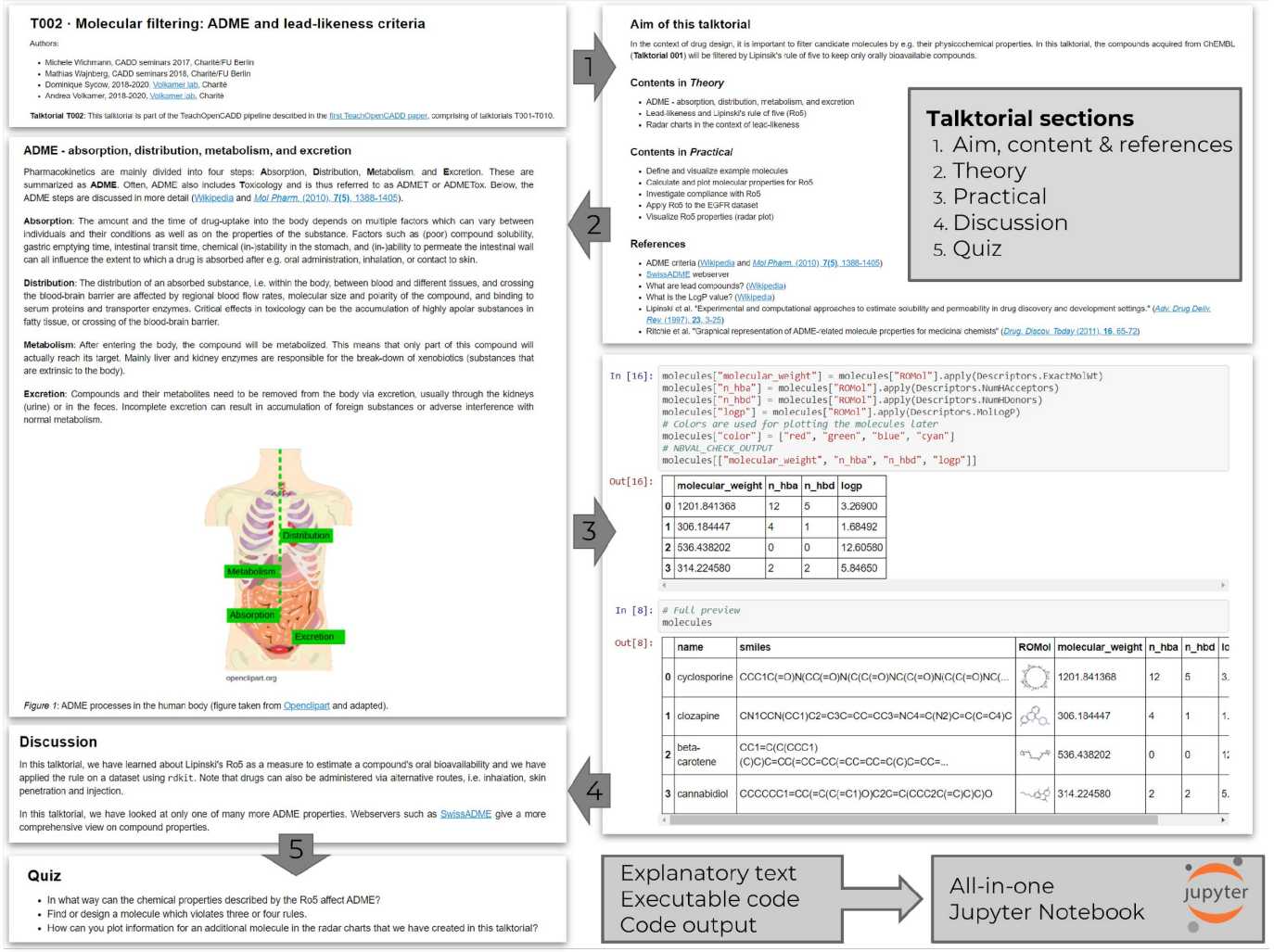
How? Coding-based Jupyter notebooks[1] (Python) & GUI-based KNIME workflows[2]

jupyter Open for Innovation KNIME

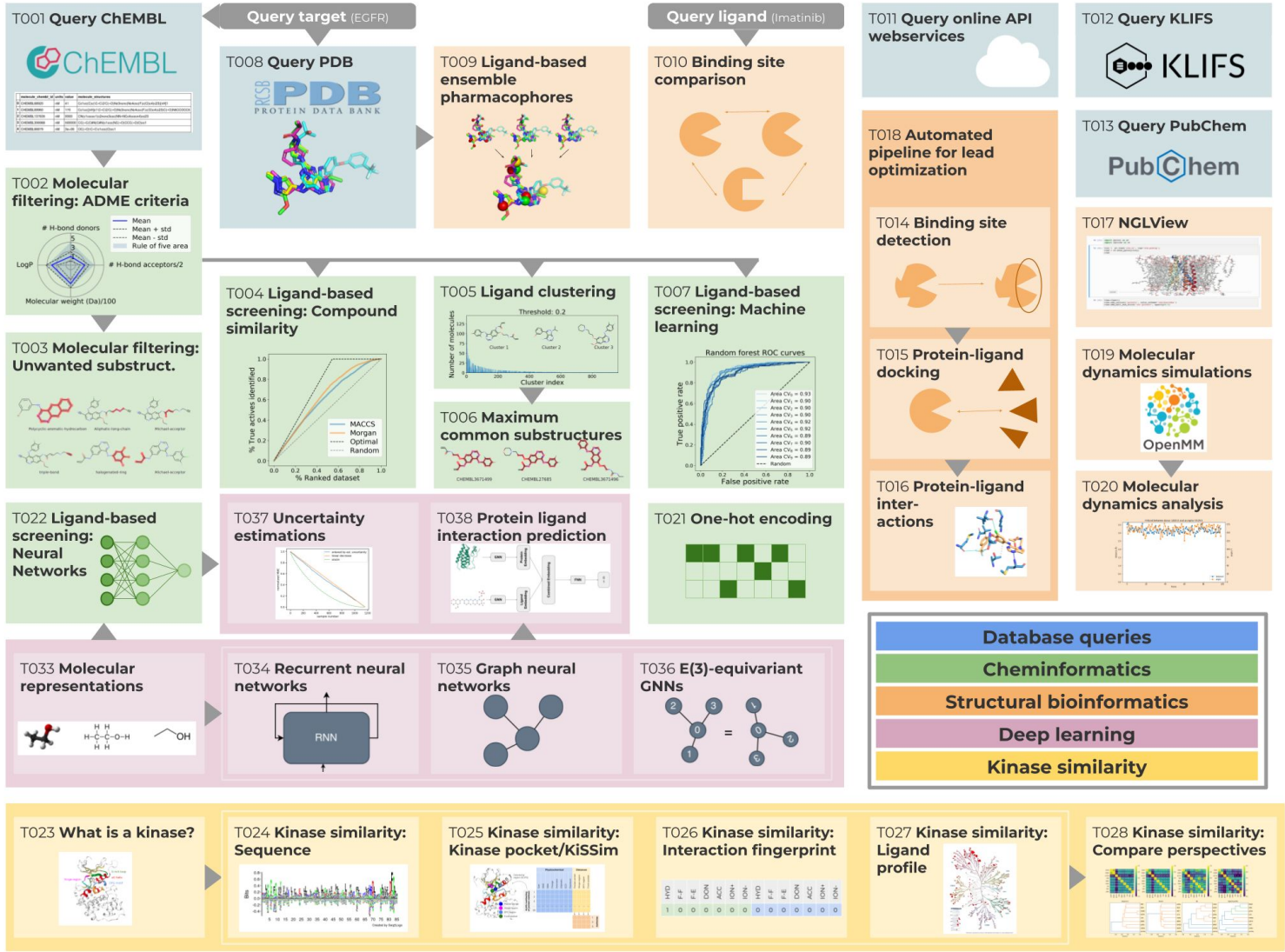
For what? Teaching, self studies & starting point for research projects

From and for whom? Students, group, community -> Beginners & advanced users

*CADD = computer-aided drug design



TeachOpenCADD – Open-Source Python Pipelines



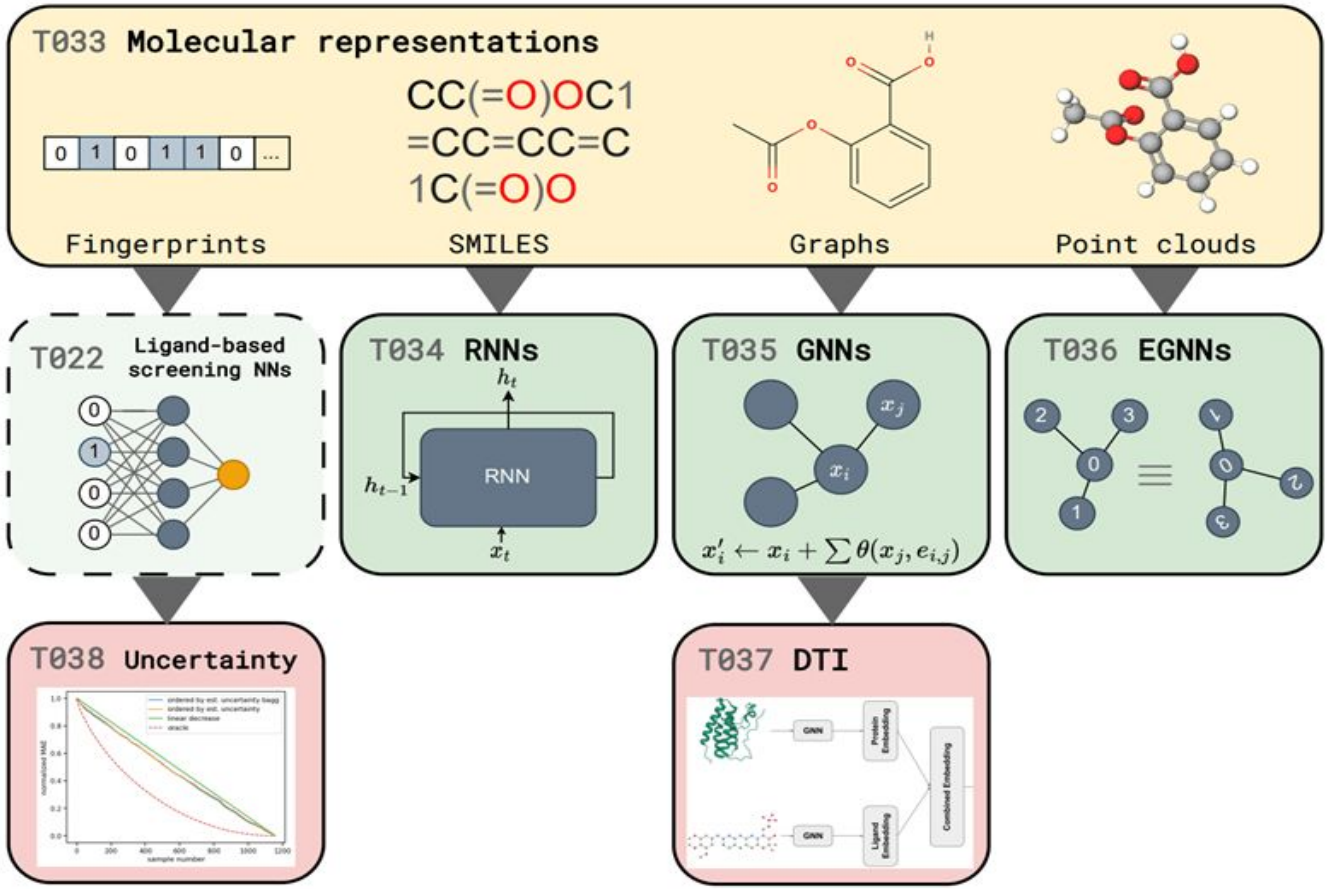
https://github.com/volkamerlab/teachopencadd

TeachOpenCADD goes Deep Learning (DL)

Introduction to molecular representations and architectures

- Molecular representation
  - Fingerprints
  - Text (SMILES)
  - Molecular graphs
  - Point clouds
  - Embeddings (learned)

- Different architectures
  - Multi-layer neural network
  - Recurrent neural network (RNN)
  - Graph neural network (GNN)
  - E(3)-invariant GNN



https://github.com/volkamerlab/teachopencadd

UNIVERSITÄT
DES
SAARLANDES

Thank you for your attention!