

## **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**  
OpenADMET



**Raquel López-Ríos de Castro**  
Saarland University



**Afnan Sultan**  
Saarland University



**Lisa-Marie Rolli**  
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

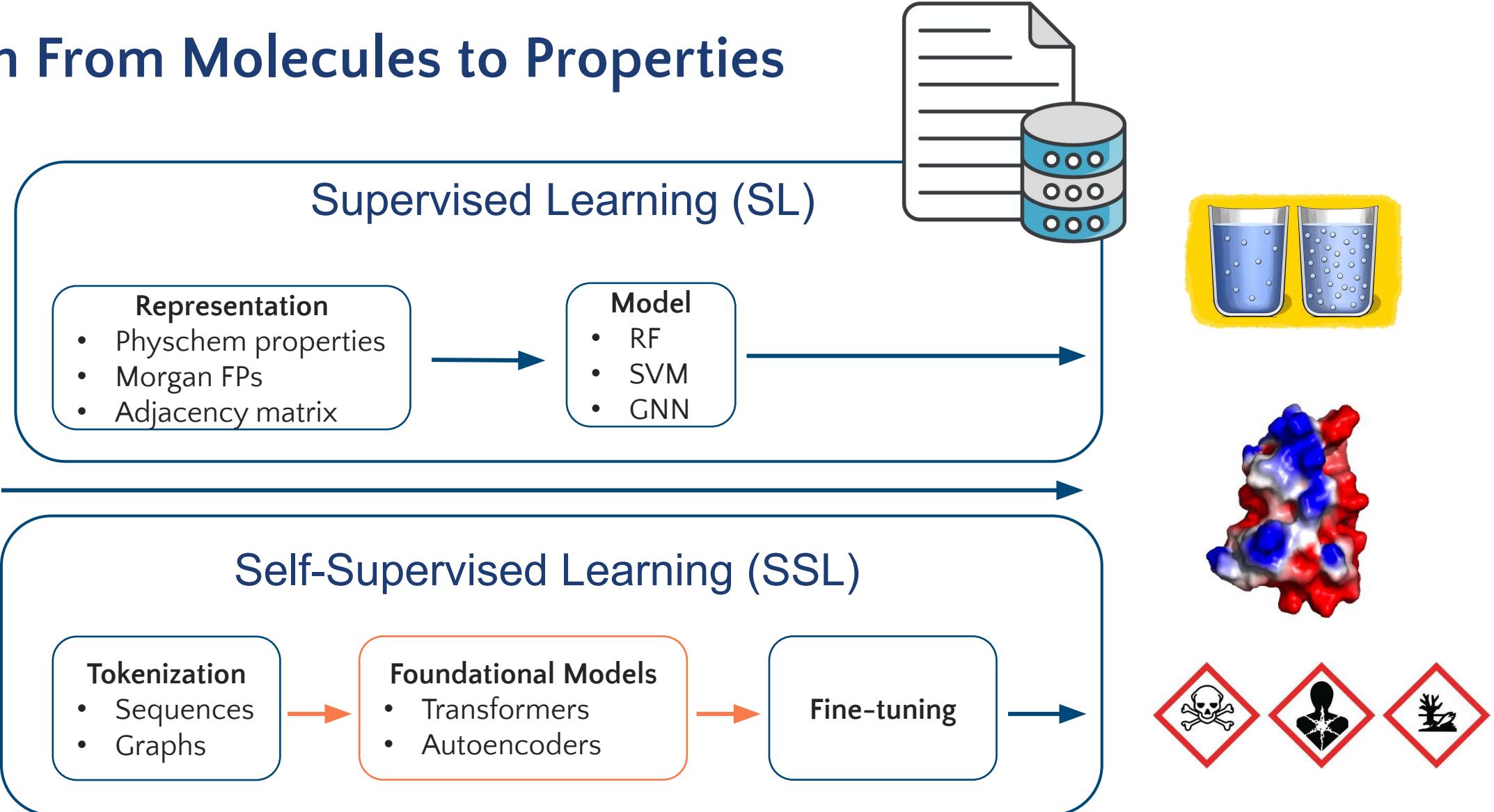
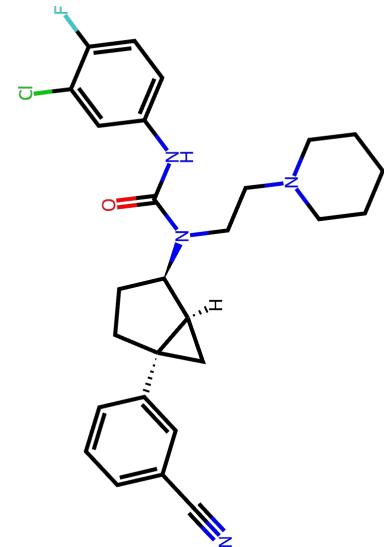
## Session 3 - 4:30 - 5:30 pm

- AI in practice
- Molecule generation
- Active learning

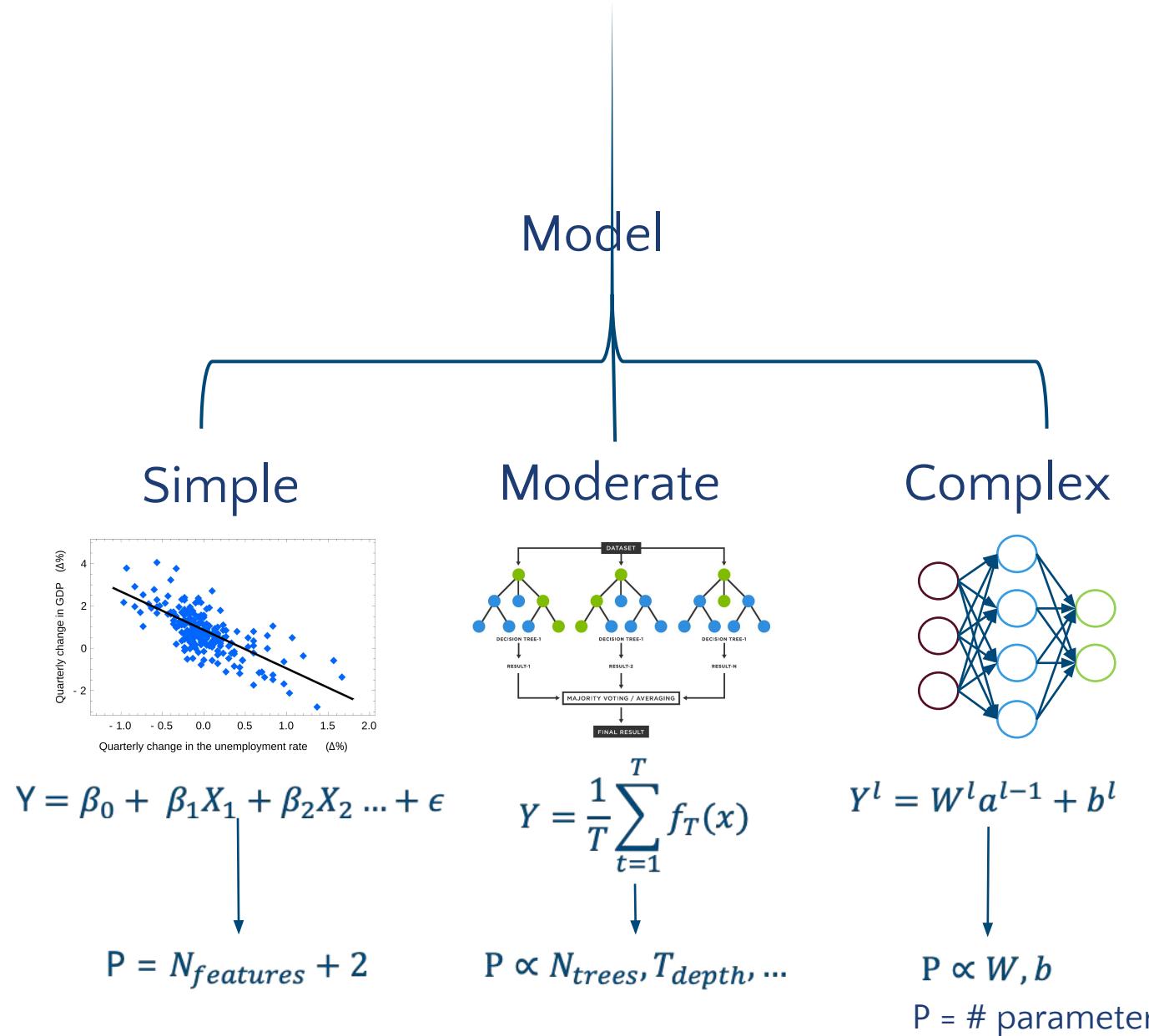
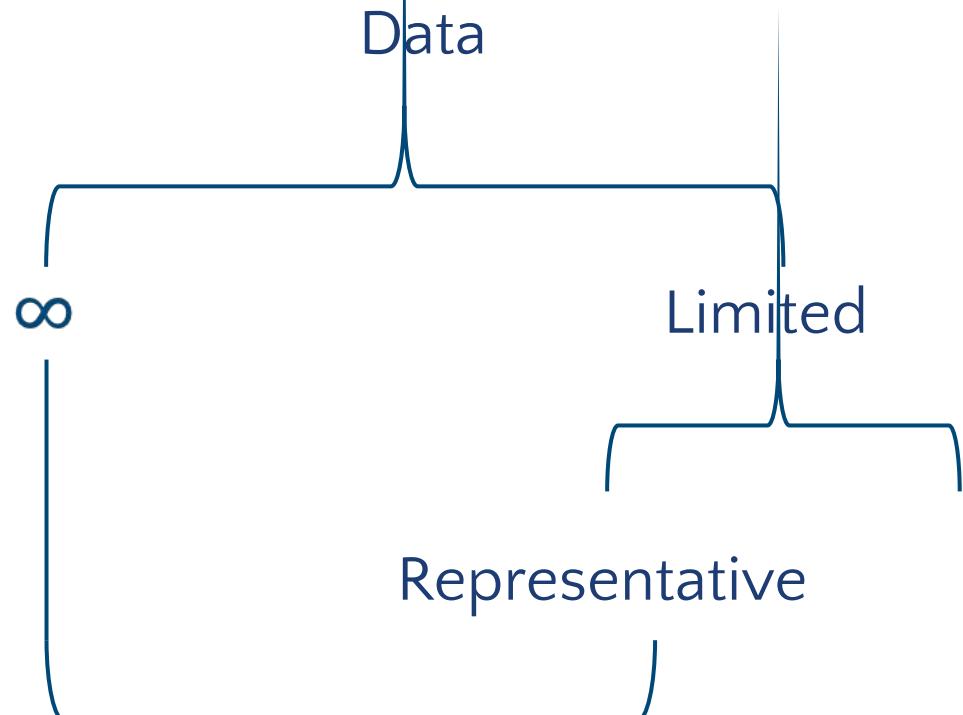
Lectures supported by hands-on sessions ...

The screenshot shows a Google Colab interface. At the top, there's a toolbar with icons for file operations (Datei, Bearbeiten, Anzeige, Einfügen, Laufzeit, Tools, Hilfe), a code editor tab (+ Code), a text editor tab (+ Text), and a link to copy the notebook to Google Drive. Below the toolbar, a message says: "This notebook is a quick test to determine if you can run a notebook on Google Colab. Click the return. At this point, you'll see a dialog that looks like this." A dark modal dialog box is displayed, containing a warning message: "Warning: This notebook was not authored by Google". It continues: "This notebook is being loaded from GitHub. It may request access to your data stored with Google, or read data and credentials from other sessions. Please review the source code before executing this notebook." At the bottom right of the dialog are "Cancel" and "Run anyway" buttons. Below the dialog, a text input field shows the command "[ ] 1+1" and a status bar at the bottom says "[ ] Beginnen Sie mit dem Programmieren oder generieren Sie Code mit KI."

# The Path From Molecules to Properties



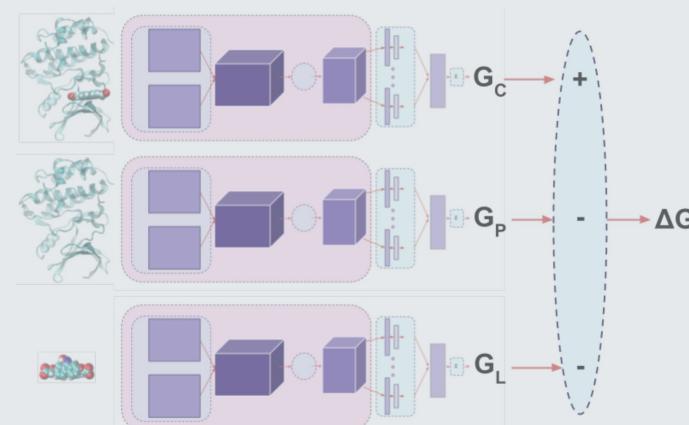
# The Two Limiting Factors



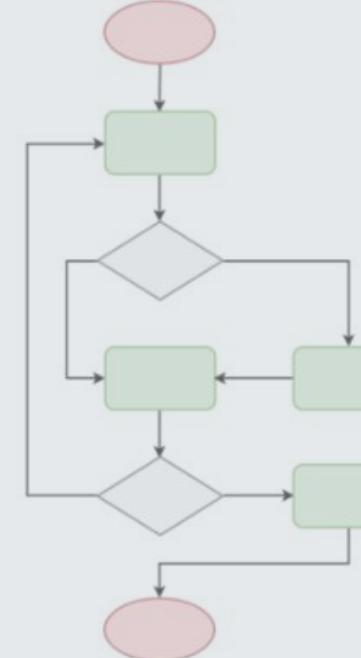
# Key Components of Machine Learning in Drug Discovery (or Anything Else)

# Data

# Representation



# Algorithms



# Where Machine Learning Excels $\longleftrightarrow$ Pharmaceutical Data Properties

## (Ultra)Large amounts of data

- Pharmaceutical data is minuscule 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

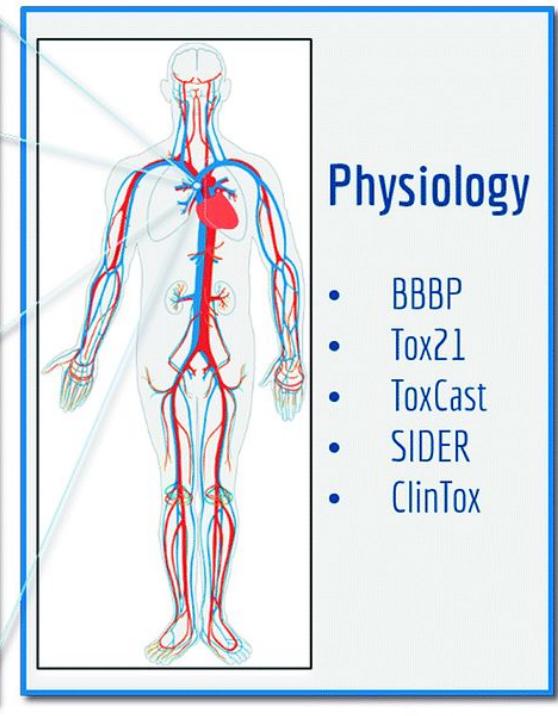
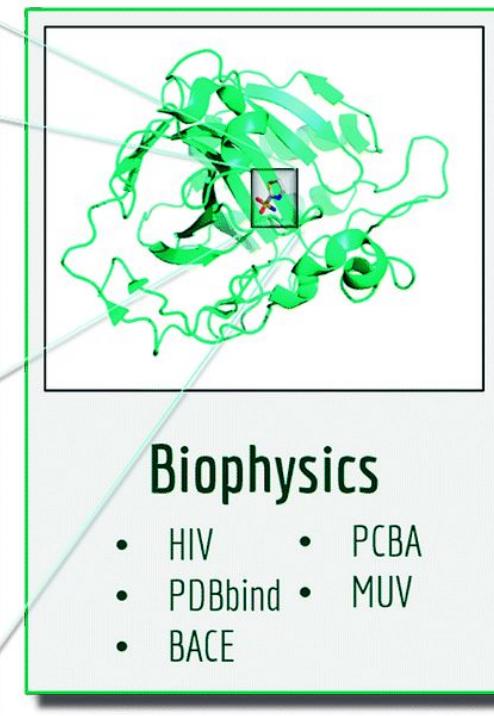
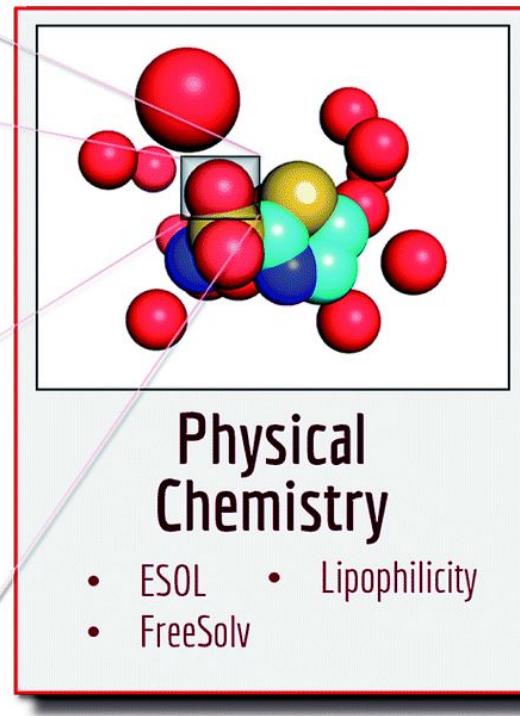
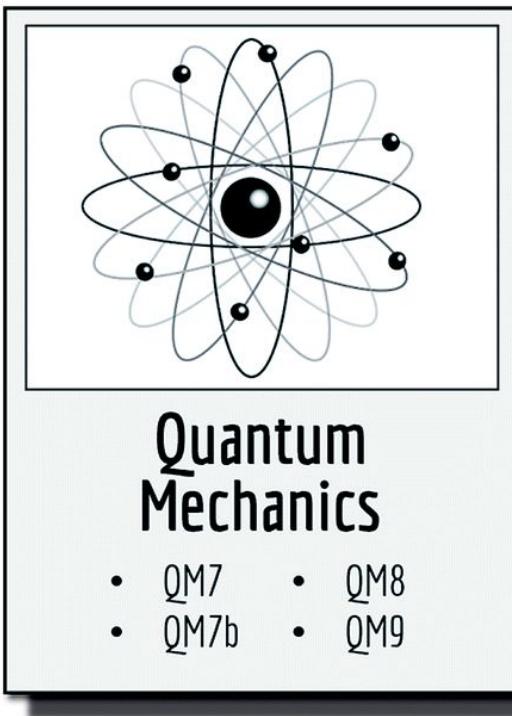
## Data covers a limited chemical space

- 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/>

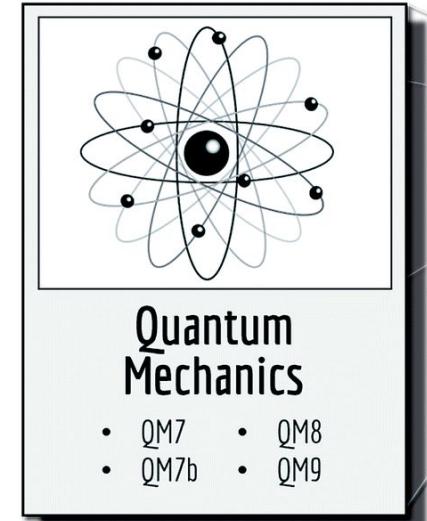
# Small Molecule Benchmark Datasets: e.g., QM9 and Tox21

## Quantum Mechanics data: QM9 → regression tasks

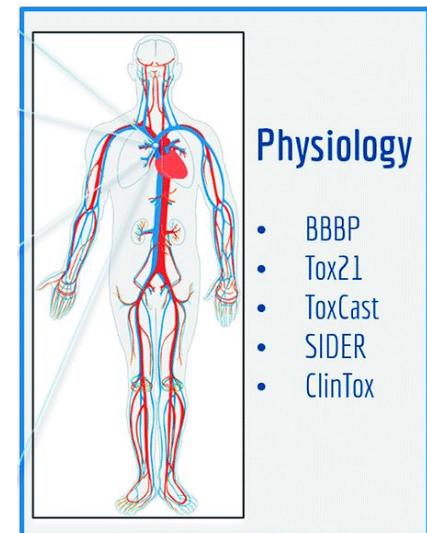
- ~134k molecules (SMILES, 3D) with <= 9 heavy atoms
- 12 properties: Geometric, energetic, electronic, and thermodynamic properties of DFT-modelled small molecules
- Commonly used due to its comparably large size

## 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

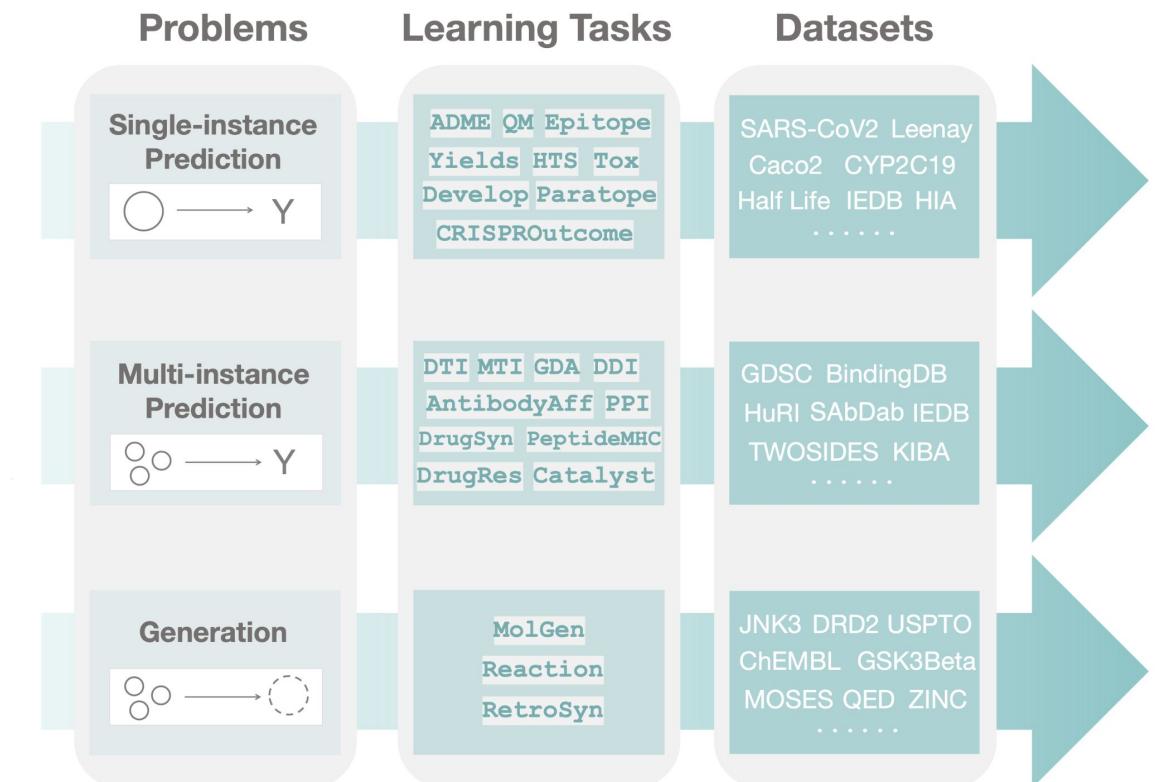
comment

## 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		
CHEMBL4650319		MOBOCERTINIB	AP32788, AP-32788, Mobocertinib, Tak-788, TAK-788	Small molecule	4	585.71		
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		



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

# Recommended Datasets Supplied By the Polaris Initiative

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

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

asap-discovery/antiviral-potency-2025-unblinded V2 Molecule Potency ✓  
Size: 1,328 Date: 2025-03-28

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

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

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

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

adaptyv-bio/egfr-binders-v1 protein-design ✓  
Size: 213 Download: 66 KB Benchmark: 1 Date: 2024-10-25

adaptyv-bio/egfr-binders-v0 protein-design ✓  
Size: 202 Download: 61 KB Benchmark: 1 Date: 2024-09-26

polaris/drewry2017-pkis2-subset-v2 Kinase HitDiscovery ✓  
Size: 640 Download: 54 KB Benchmark: 14 Date: 2024-07-10

biogen/adme-fang-v1 adme ✓  
Size: 3,521 Download: 384 KB Benchmark: 7 Date: 2024-07-10

Polaris certified datasets have been evaluated and approved by industry experts

# 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

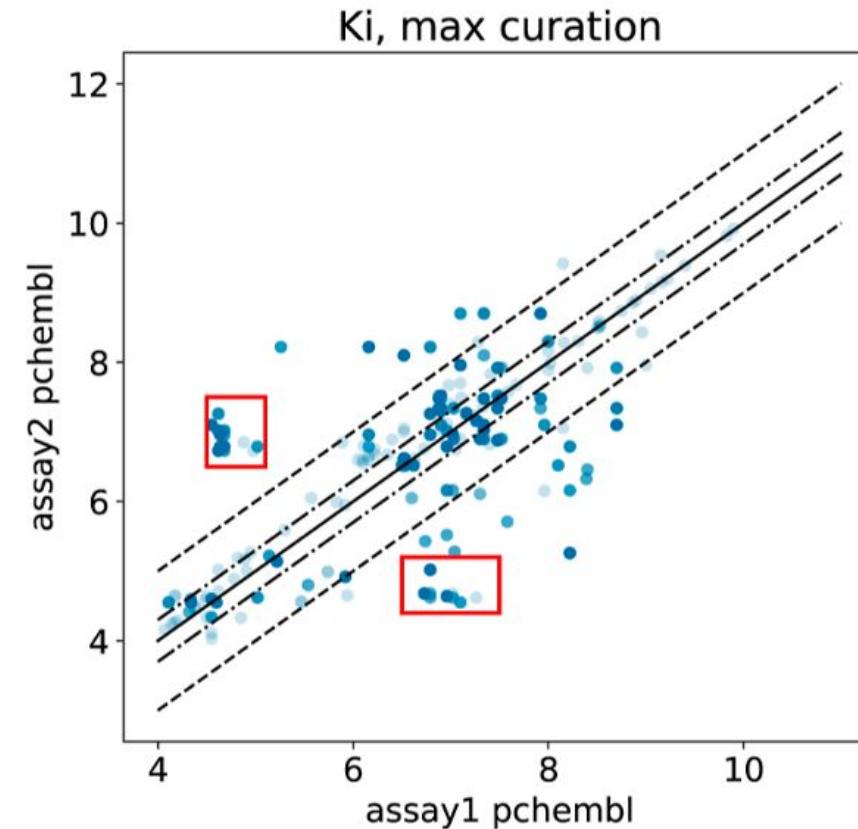
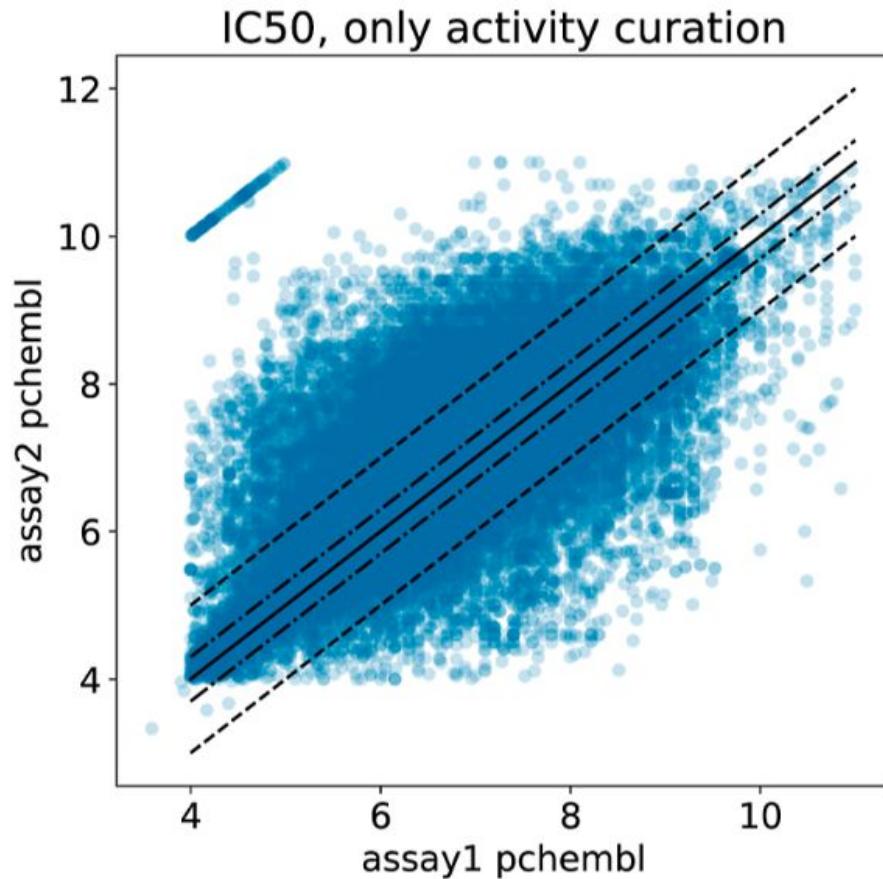
Inspired by Pat's BlogPost  
[we-need-better-benchmarks-for-machine-learning/](https://patcc.org/we-need-better-benchmarks-for-machine-learning/)

- **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  $\mu\text{M}$  (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

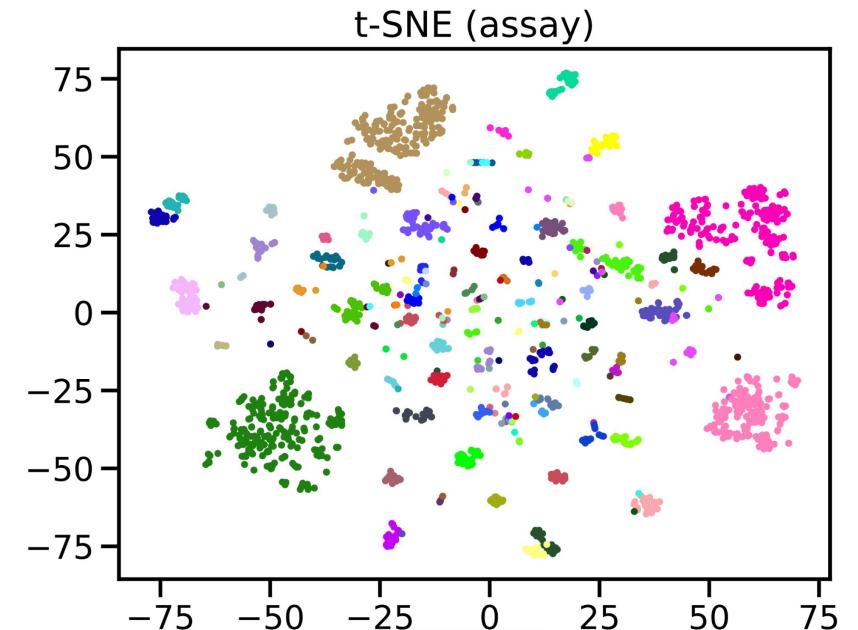
## Inconsistent Data Can Make ML Modeling Difficult



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

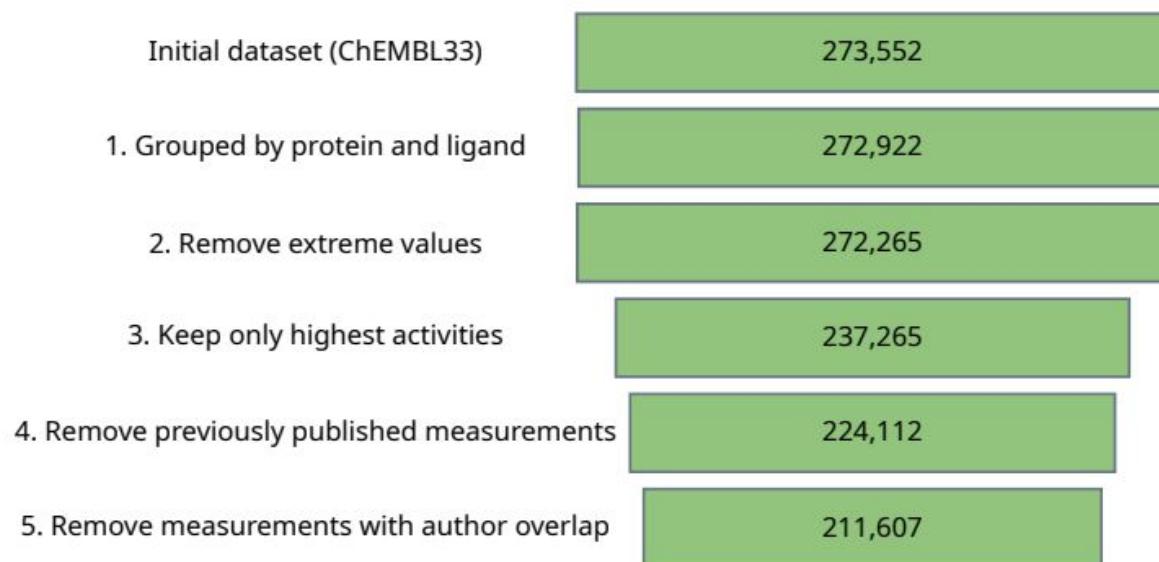


- kinodata-EGFR ligand activities
- t-SNE on 2048 bit Morgan fingerprints
- colors = assays

# How to Preprocess such Data Properly?

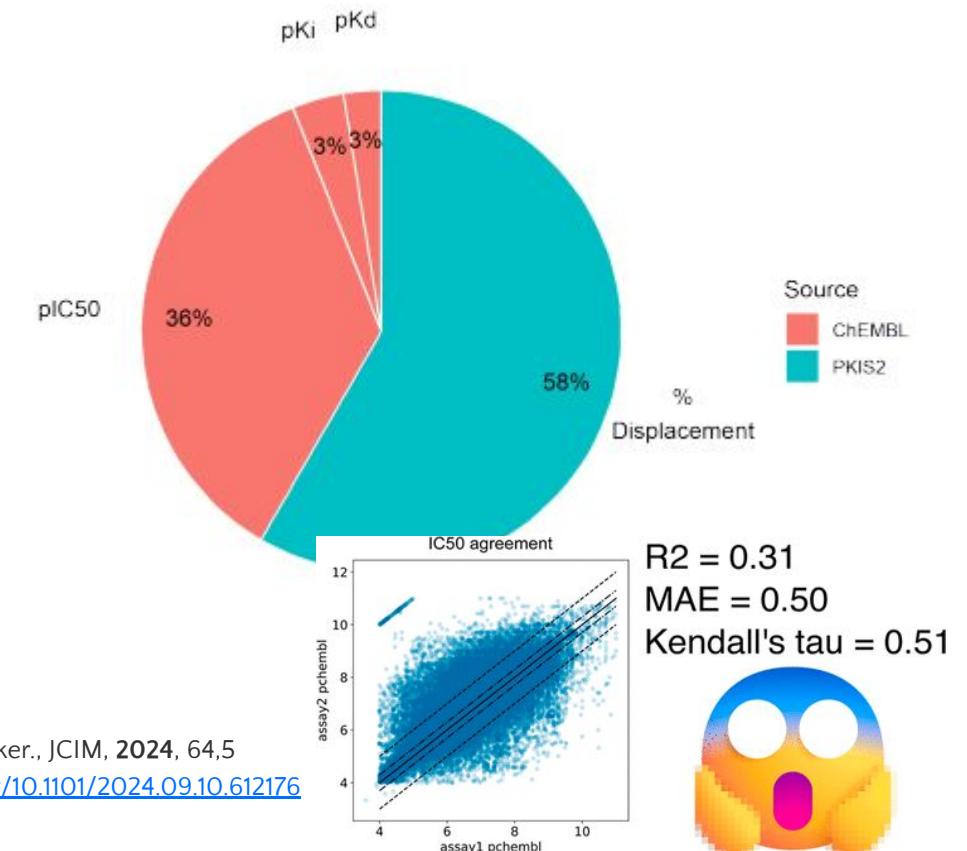
- Automated curation pipeline to reduce errors and to increase reliability

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



# 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



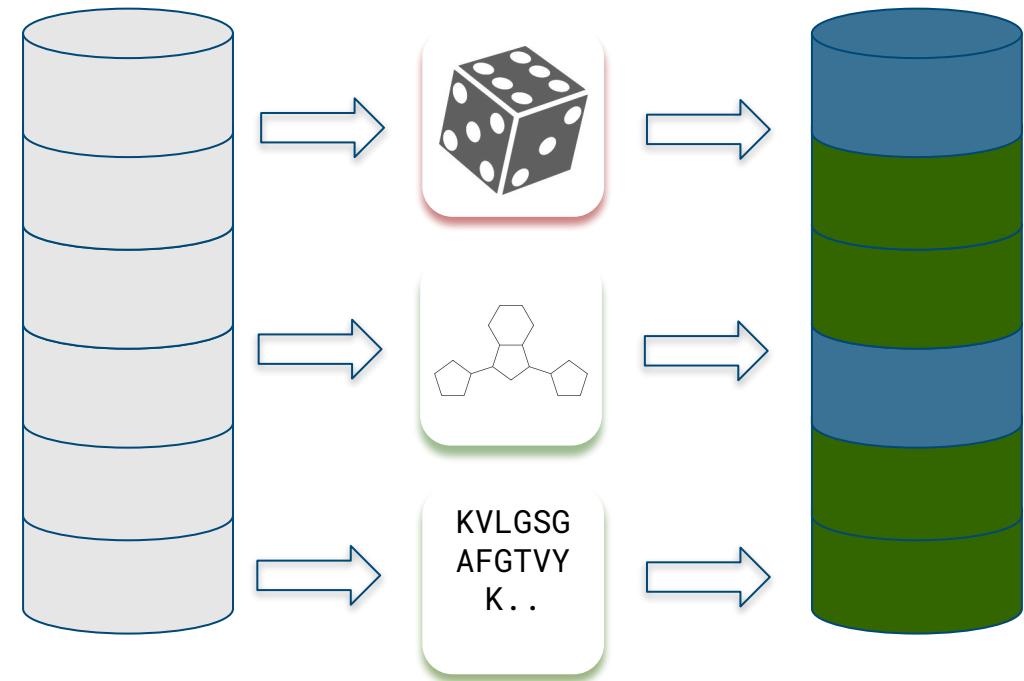
Can I predict properties for my molecules based on literature data?



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

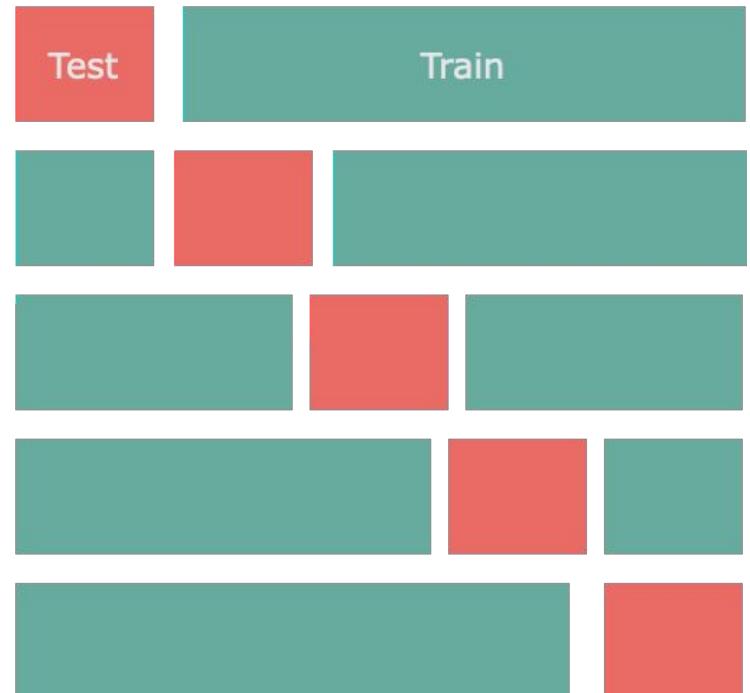
# Data Split is Crucial to Assess Generalizability

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

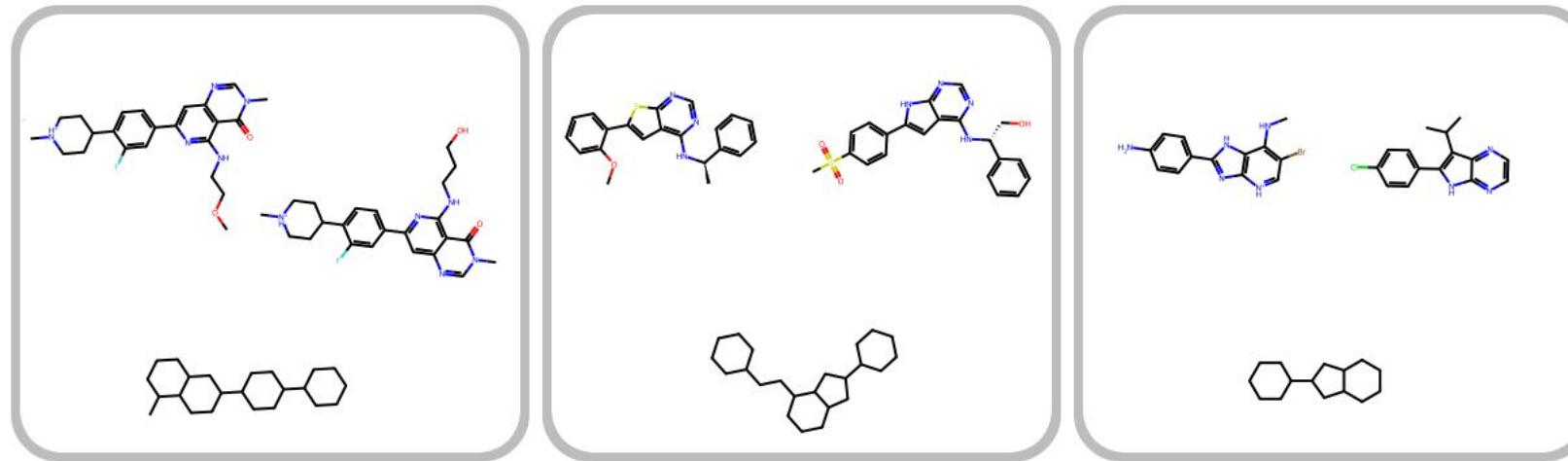


# Grouped k-Fold Split

- 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)



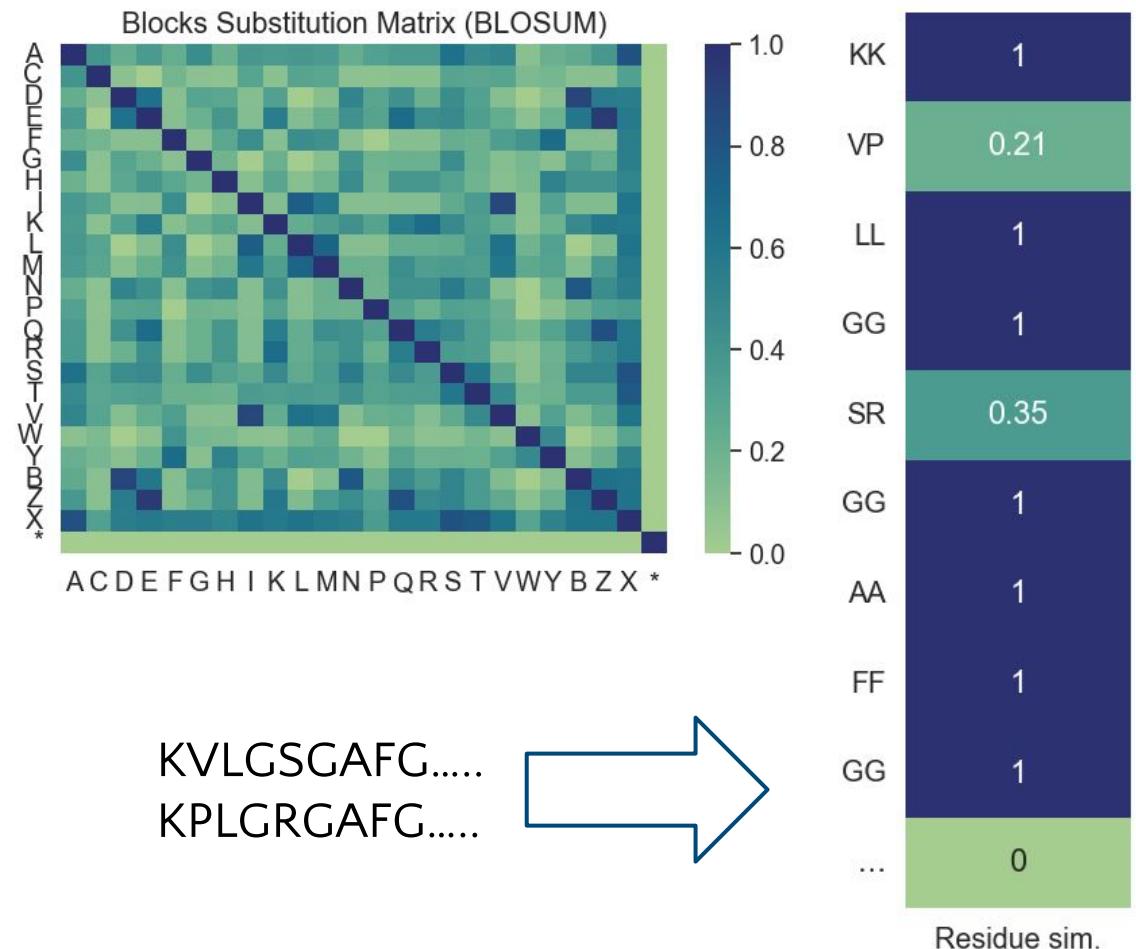
# Data Splits – Scaffold Split



- 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

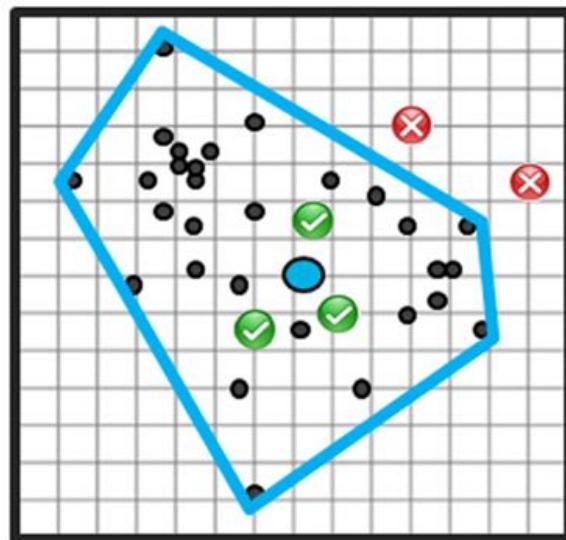
Monte Carlo Dropout

# 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.

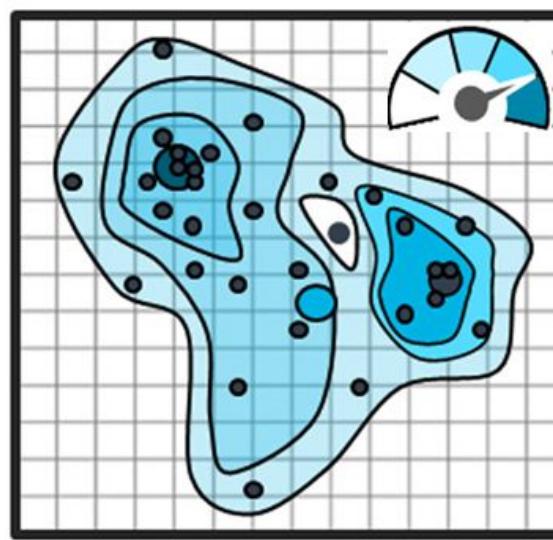
## 1. Applicability:

Interpolation within the chemical structure space



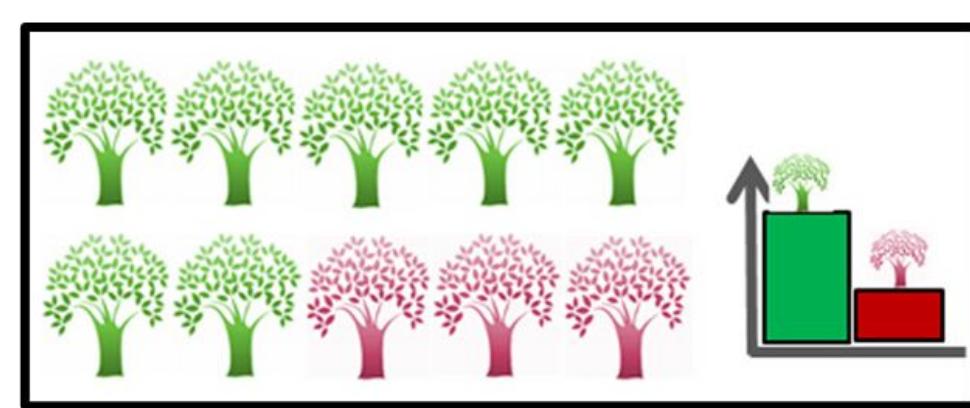
## 2. Reliability:

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



0.84



0.63

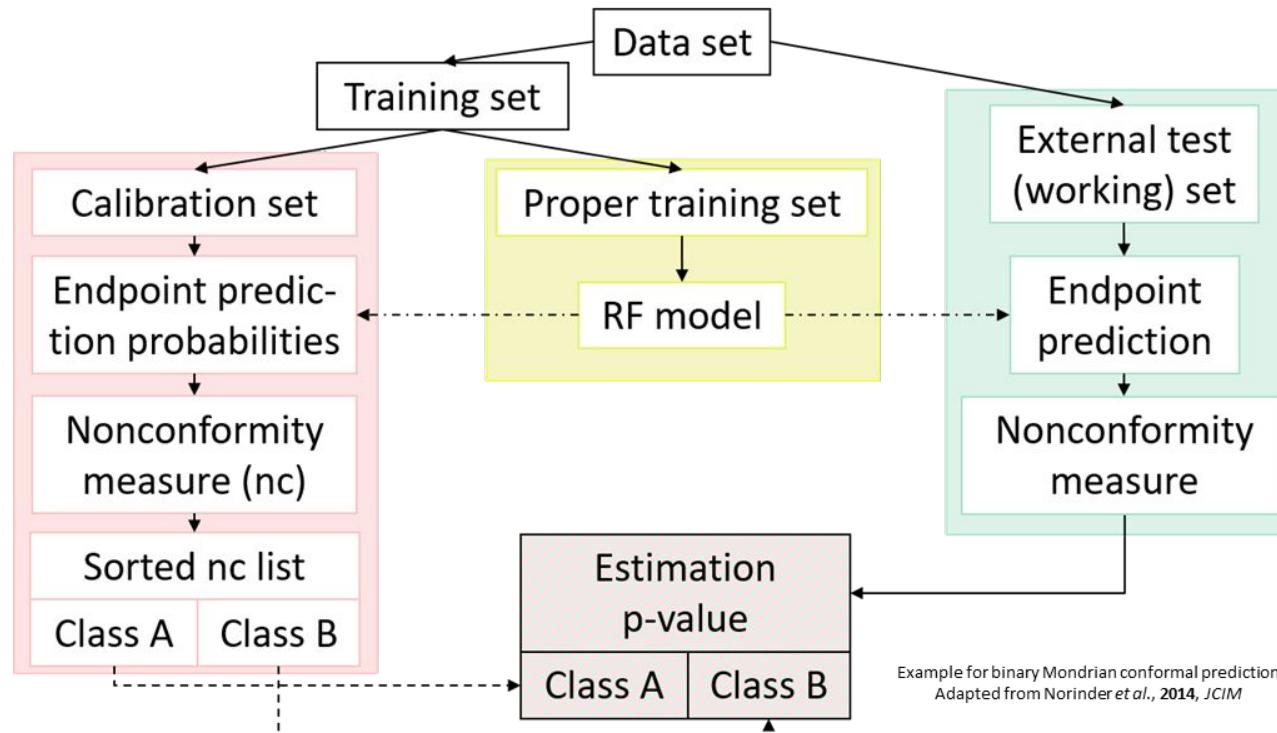


0.15



# 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



		<chem>O=[N+]([O-])c1ccc(F)c(F)c1</chem>	prob(A) = 0.7	prob(B) = 0.3		
			Class A	Class B	0.99	
0.83					4/5	
0.67						0.70
0.47					2/5	0.63
0.43						0.37
0.41					0/5	0.33
					2/7	0.31
						1/7 → 0.12

p-value(A) = 0.8  
p-value(B) = 0.14  
→ Significance level 0.2: {A}

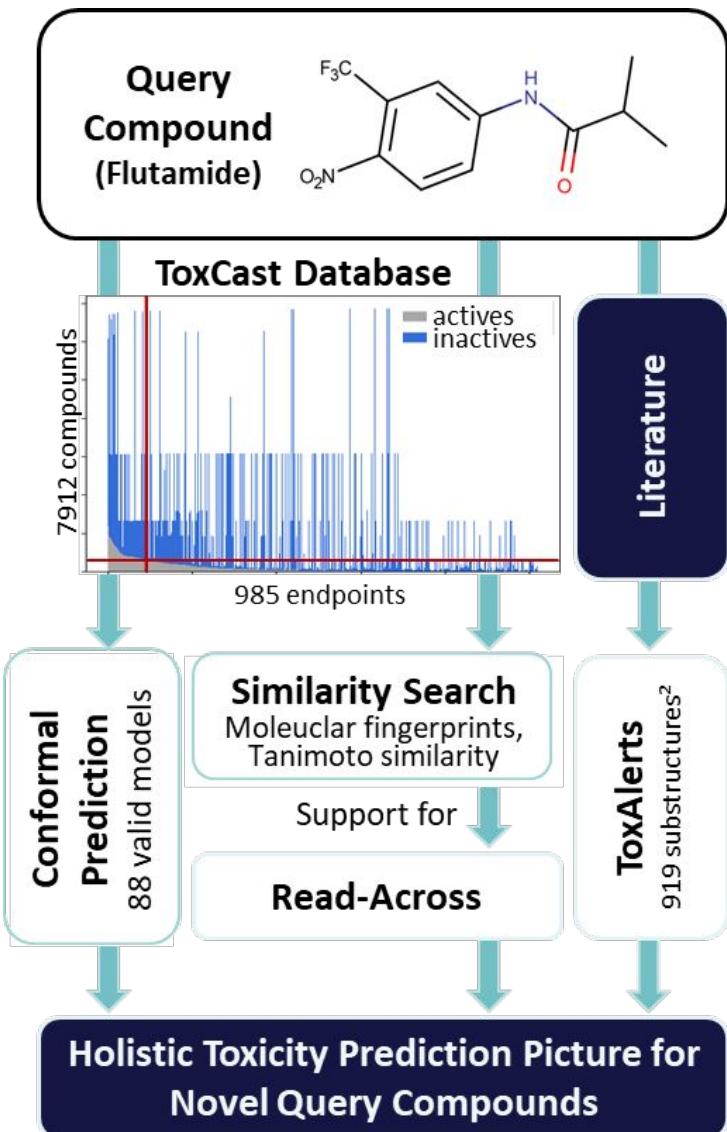
# Know-Tox – Predictions with Confidence

- 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)			# toxic/non-toxic
			all	cl.1	cl.0	
cross-validation	ToxCast AA	0.87	0.78	0.80	0.78	868/5842
external	Norinder <sup>3</sup>	0.79	0.68	0.70	0.67	160/201
In-house	BASF (I)	0.94	<b>0.56</b>	<b>0.97</b>	<b>0.07</b>	280/254



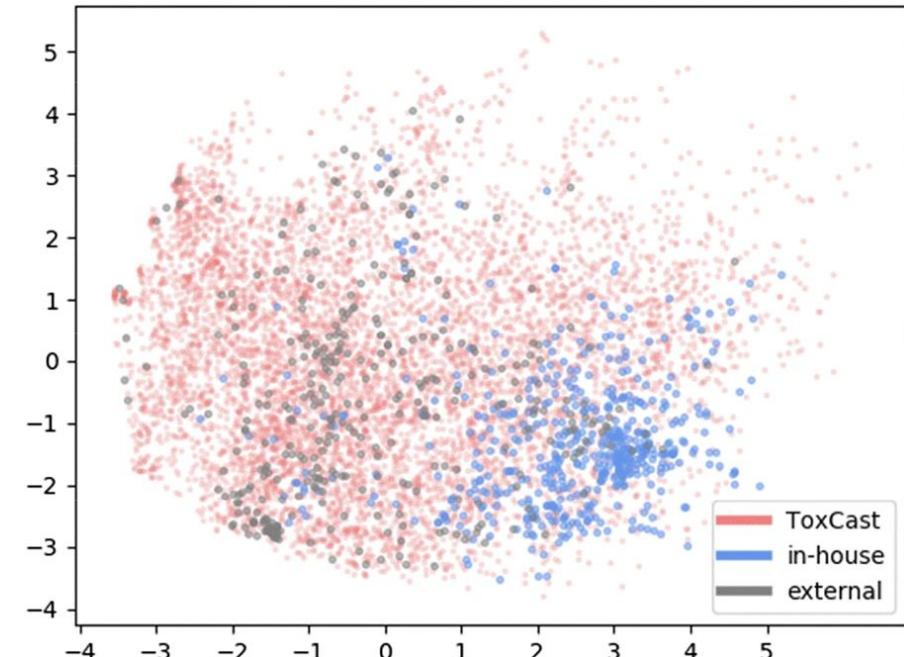
# Know-Tox – Predictions with Confidence

- 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

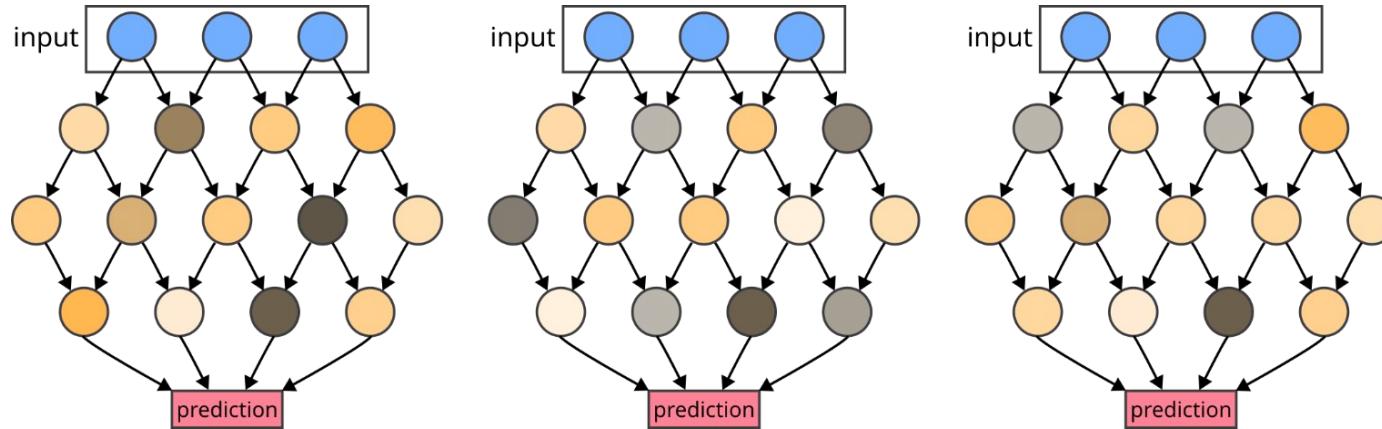
	<b>Dataset</b>	<b>Efficiency</b>	<b>all</b>	<b>Accuracy (SCPs)</b>	<b># toxic/non-toxic</b>	
				cl.1	cl.0	
cross-validation	ToxCast AA	0.87	0.78	0.80	0.78	
	Norinder <sup>3</sup>	0.79	0.68	0.70	0.67	
	BASF (I)	0.94	<b>0.56</b>	<b>0.97</b>	<b>0.07</b>	
In-house	BASF (II)	<b>0.20</b>	<b>0.75</b>	<b>0.80</b>	<b>0.71</b>	280/254



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

# Using Uncertainty to Assess Model Confidence

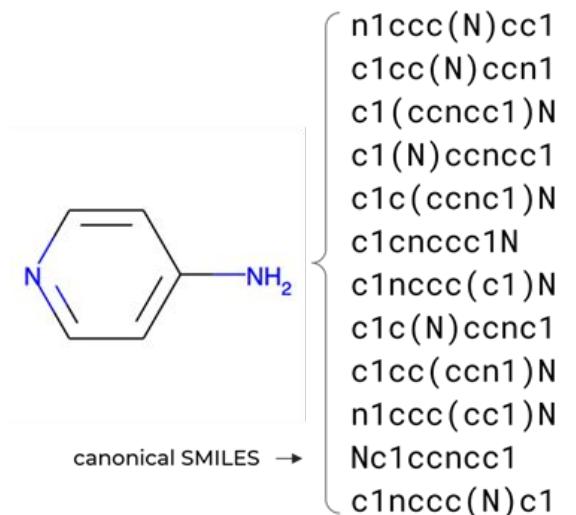
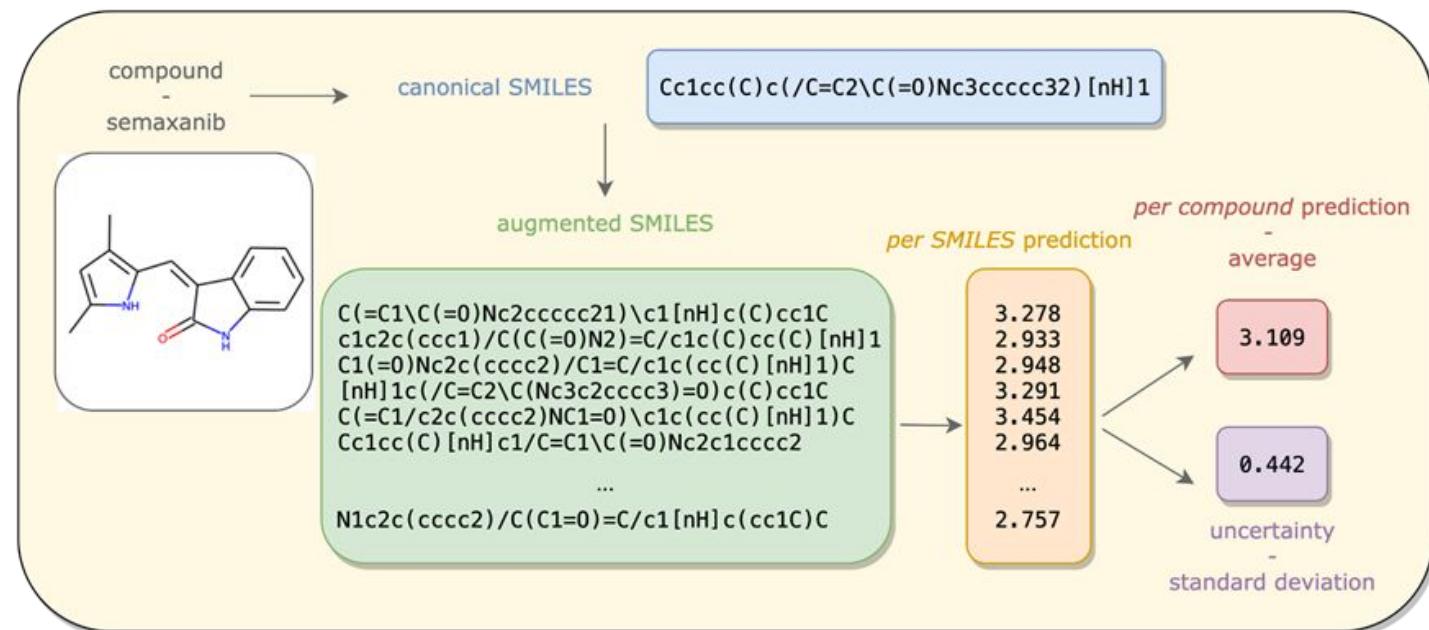
- Single deterministic methods
  - Classification setting: predicted class probabilities
  - Secondary model trained to predict uncertainty for an already trained model
- Ensemble methods
  - Ensemble of similar but different models -> Variance as uncertainty
  - Varying random seeds, resulting in different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

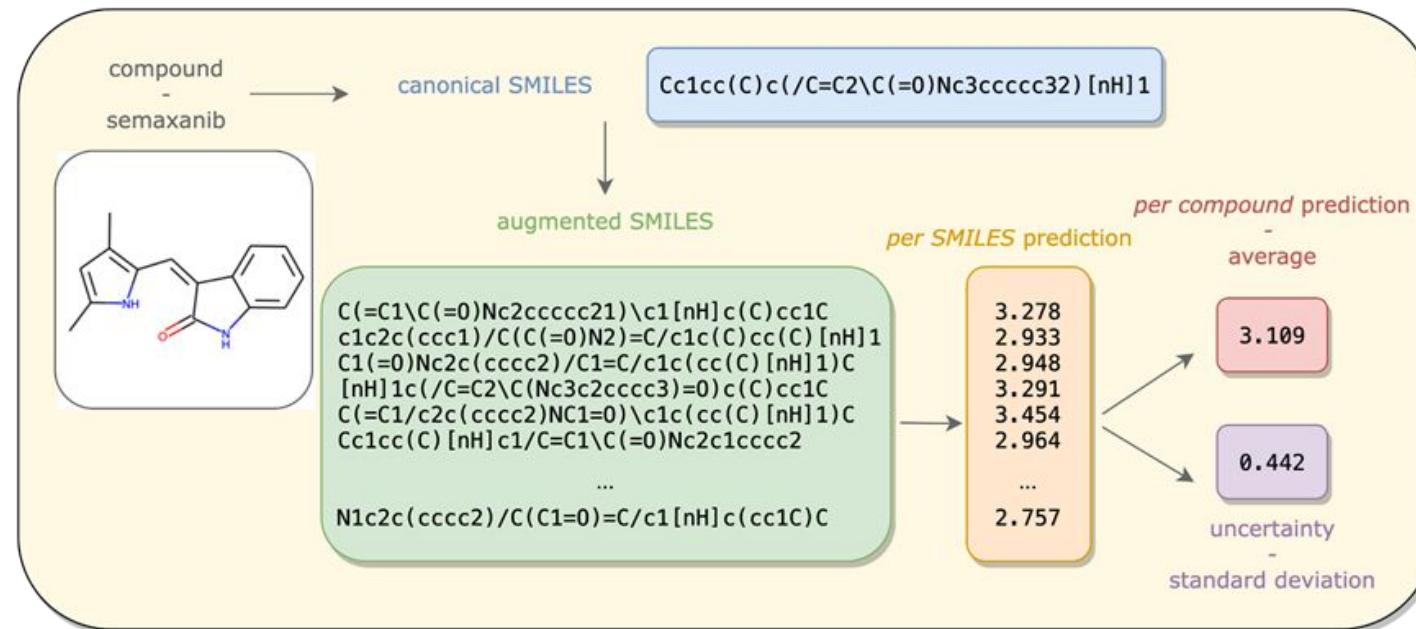


Kimber et al., AI in Life Sciences, 2021, 1, 100014

<https://github.com/volkamerlab/maxsmi>

# SMILES Augmentation Improves Model Performance

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

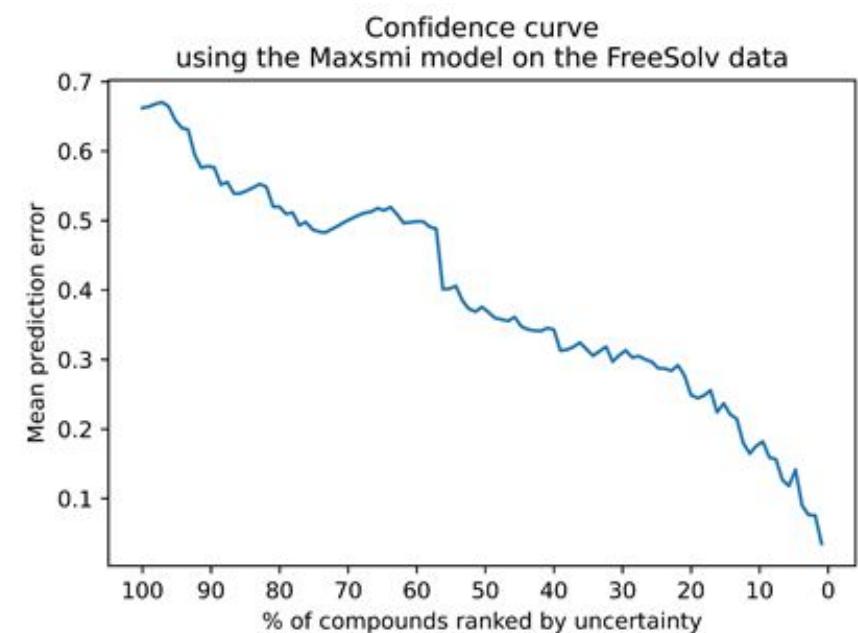


Kimber et al., AI in Life Sciences, 2021, 1, 100014

<https://github.com/volkamerlab/maxsmi>

## Uncertainty prediction

More confident model implies smaller prediction error



# TeachOpenCADD in a Nutshell

<https://github.com/volkamerlab/teachopencadd>

**What?** Pipelines for common CADD\* tasks using open resources

**How?** Coding-based Jupyter notebooks<sup>1</sup> (Python) & GUI-based KNIME workflows<sup>2</sup>



**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

**T002 - Molecular filtering: ADME and lead-likeness criteria**

Autors:

- Michele Wichmann, CADD seminars 2017, Charité/FU Berlin
- Mathias Wajnberg, CADD seminars 2018, Charité/FU Berlin
- Dominique Sycow, 2018-2020, Volkamer lab, Charité
- Andrea Volkamer, 2018-2020, Volkamer lab, Charité

Talktorial T002: This talktorial is part of the TeachOpenCADD pipeline described in the [first TeachOpenCADD paper](#), comprising of talktoria T001-T10.

## ADME - absorption, distribution, metabolism, and excretion

Pharmacokinetics are mainly divided into four steps: Absorption, Distribution, Metabolism, and Excretion. These are summarized as **ADME**. Often, ADME also includes Toxicology and is thus referred to as ADMET or ADMETox. Below, the ADME steps are discussed in more detail ([Wikipedia](#) and [Mol.Pharm.\(2010\), 7\(5\), 1388-1405](#)).

**Absorption:** The amount and the time of drug uptake into the body depends on multiple factors which can vary between individuals and their conditions as well as on the properties of the substance. Factors such as (poor) compound solubility, gastric emptying time, intestinal transit time, chemical (in-)stability in the stomach, and (in-)ability to permeate the intestinal wall can all influence the extent to which a drug is absorbed after e.g. oral administration, inhalation, or contact to skin.

**Distribution:** The distribution of an absorbed substance, i.e. within the body, between blood and different tissues, and crossing the blood-brain barrier are affected by regional blood flow rates, molecular size and polarity of the compound, and binding to serum proteins and transporter enzymes. Critical effects in toxicology can be the accumulation of highly polar substances in fatty tissue, or crossing of the blood-brain barrier.

**Metabolism:** After entering the body, the compound will be metabolized. This means that only part of this compound will actually reach its target. Mainly liver and kidney enzymes are responsible for the break-down of xenobiotics (substances that are extrinsic to the body).

**Excretion:** Compounds and their metabolites need to be removed from the body via excretion, usually through the kidneys (urine) or in the feces. Incomplete excretion can result in accumulation of foreign substances or adverse interference with normal metabolism.



Figure 1: ADME processes in the human body (figure taken from [Openclipart](#) and adapted).

## Discussion

In this talktorial, we have learned about Lipinski's  $\text{Ro}_5$  as a measure to estimate a compound's oral bioavailability and we have applied the rule on a dataset using rdkit. Note that drugs can also be administered via alternative routes, i.e. inhalation, skin penetration and injection.

In this talktorial, we have looked at only one of many more ADME properties. Webservers such as [SwissADME](#) give a more comprehensive view on compound properties.

## Quiz

- In what way can the chemical properties described by the  $\text{Ro}_5$  affect ADME?
- Find or design a molecule which violates three or four rules.
- How can you plot information for an additional molecule in the radar charts that we have created in this talktorial?

## Aim of this talktorial

In the context of drug design, it is important to filter candidate molecules by e.g. their physicochemical properties. In this talktorial, the compounds acquired from ChEMBL (Talktorial 001) will be filtered by Lipinski's rule of five to keep only orally bioavailable compounds.

## Contents in Theory

- ADME - absorption, distribution, metabolism, and excretion
- Lead-likeness and Lipinski's rule of five ( $\text{Ro}_5$ )
- Radar charts in the context of lead-likeness

## Contents in Practical

- Define and visualize example molecules
- Calculate and print molecular properties for  $\text{Ro}_5$
- Investigate compliance with  $\text{Ro}_5$
- Apply  $\text{Ro}_5$  to the EGF-R dataset
- Visualize  $\text{Ro}_5$  properties (radar plot)

## References

- ADME criteria ([Wikipedia](#) and [Mol.Pharm.\(2010\), 7\(5\), 1388-1405](#))
- [SwissADME](#) webserver
- What are lead compounds? ([Wikipedia](#))
- What is the  $\text{LogP}$  value? ([Wikipedia](#))
- Lipinski et al., "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings." ([Adv.Drug.Deliv.Res.\(1997\), 23, 3-25](#))
- Ritchie et al., "Graphical representation of ADME-related molecule properties for medicinal chemists" ([Drug.Discovery.Today.\(2011\), 16, 65-72](#))

In [16]:

```
molecules["molecular_weight"] = molecules["ROMol"].apply(Descriptors.ExactMolWt)
molecules["n_hba"] = molecules["ROMol"].apply(Descriptors.NumAcceptors)
molecules["n_hbd"] = molecules["ROMol"].apply(Descriptors.NumHDonors)
molecules["logP"] = molecules["ROMol"].apply(Descriptors.MolLogP)
# Colors are used for plotting the molecules later
molecules["color"] = ["red", "green", "blue", "cyan"]
# NDVAL_CHECK_OUTPUT
molecules[["molecular_weight", "n_hba", "n_hbd", "logP"]]
```

Out[16]:

	molecular_weight	n_hba	n_hbd	logP
0	1201.841368	12	5	3.26900
1	306.184447	4	1	1.68492
2	536.438202	0	0	12.60580
3	314.224580	2	2	5.84650

In [8]: # Full preview  
molecules

Out[8]:

	name	smiles	ROMo	molecular_weight	n_hba	n_hbd	ic
0	cyclosporine	CCC1C(=O)N(CC(=O)N(C(C(=O)NC(C(=O)N(C(C(=O)NC(=O)C1)C)C)C)C)C)C		1201.841368	12	5	3.
1	clozapine	CN1CCN(CC1)C2=C3C=CC=C3C3=N4C4=C(N2)C=C(C=C4)C		306.184447	4	1	1.
2	beta-carotene	CC1=CC(CCC1)(C)C=CC=CC=C(C=C(C=C1)C)C		536.438202	0	0	1.
3	cannabidiol	CCCCCC1=CC=C(C=C1O)C2=C(CCC2C(=C)C)O		314.224580	2	2	5.

Explanatory text  
Executable code  
Code output

## Talktorial sections

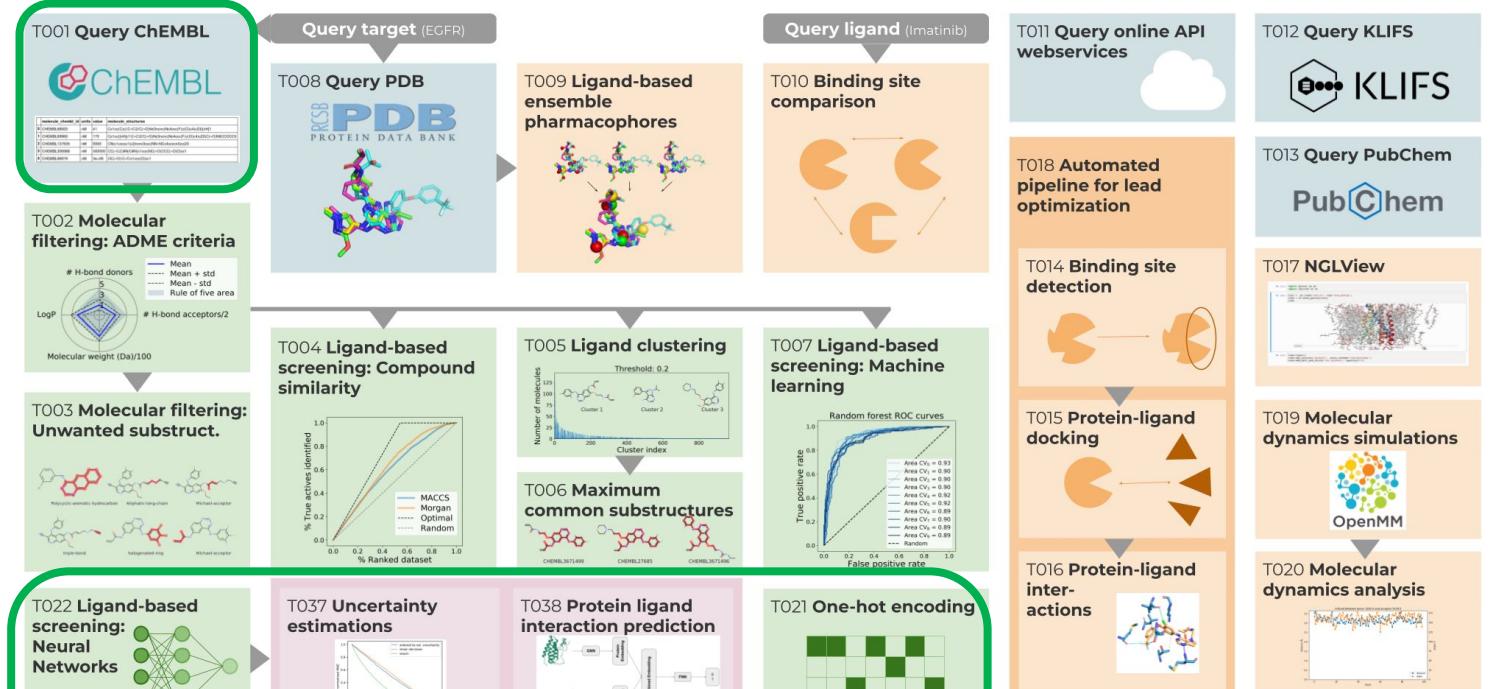
- Aim, content & references
- Theory
- Practical
- Discussion
- Quiz



# TeachOpenCADD – Open-Source Python Pipelines

Relevant for today's session:

API queries



DL examples

Kinase-centric structure-based ML  
<https://github.com/openkinome>

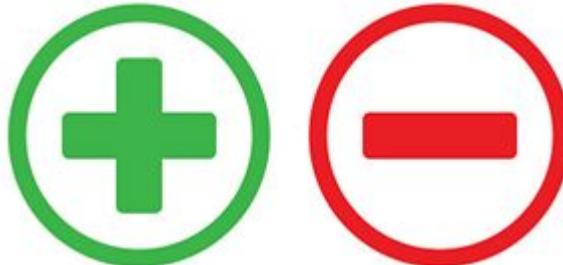
Lessons learned during the journey of data: from experiment to model for predicting kinase affinity, selectivity, polypharmacology, and resistance,

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

<https://github.com/volkamerlab/teachopencadd>

# 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 reproducibility



UNIVERSITÄT  
DES  
SAARLANDES

Thank you for your attention!