Where to begin with AI and ML?

(for chemistry)

Raquel Lopez-Rios de Castro
Clementi lab | Volkamer Lab |Chodera Lab
ChemSpider webinar series

AlphaFold

Claude

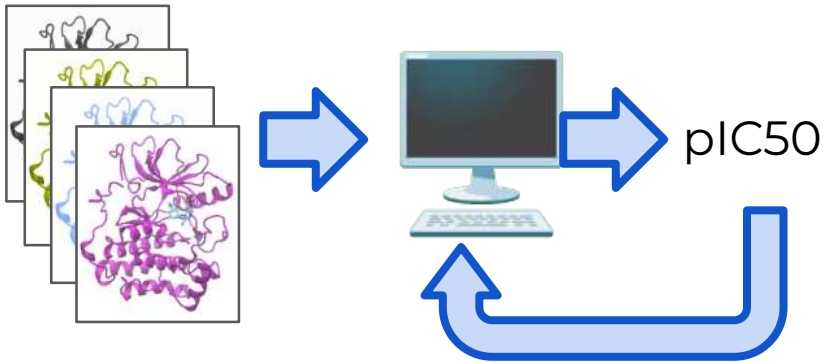
grammarly

deepseek

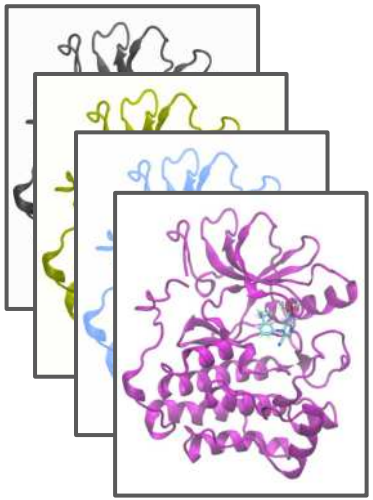
What is ML/AI?

"Machine Learning is all about labelling things using examples."

Pat Walters
OpenADMET
https://patwalters.github.io/



Drug-target complexes ⇔ pIC50 (activity)



```
[3.1,
4.3,
9.2,
6.8]
```

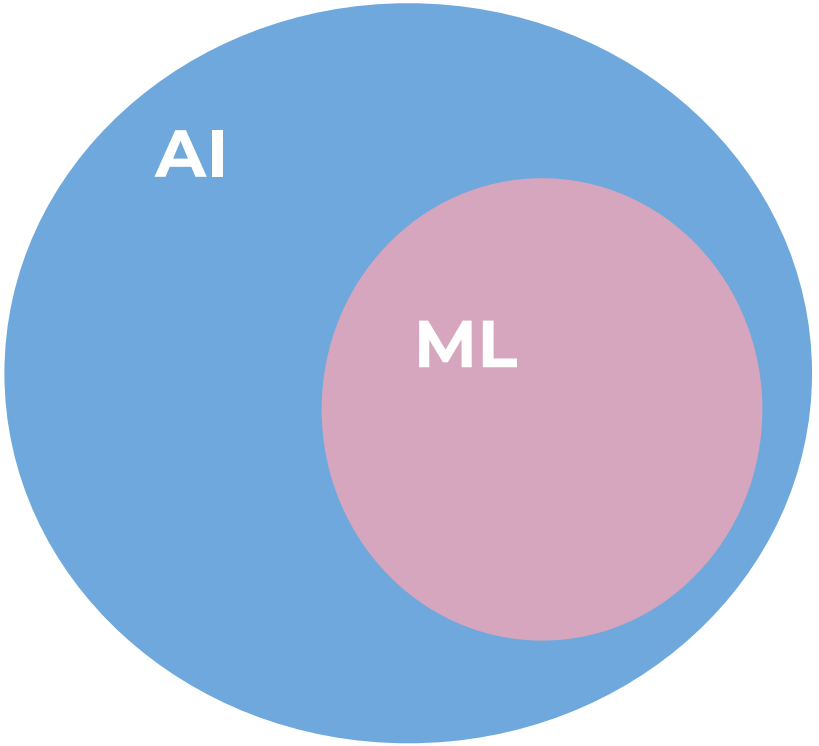
Training data

Mat Velloso
@matvelloso

Difference between machine learning and AI:

If it is written in Python, it's probably machine learning

If it is written in PowerPoint, it's probably AI



DATA

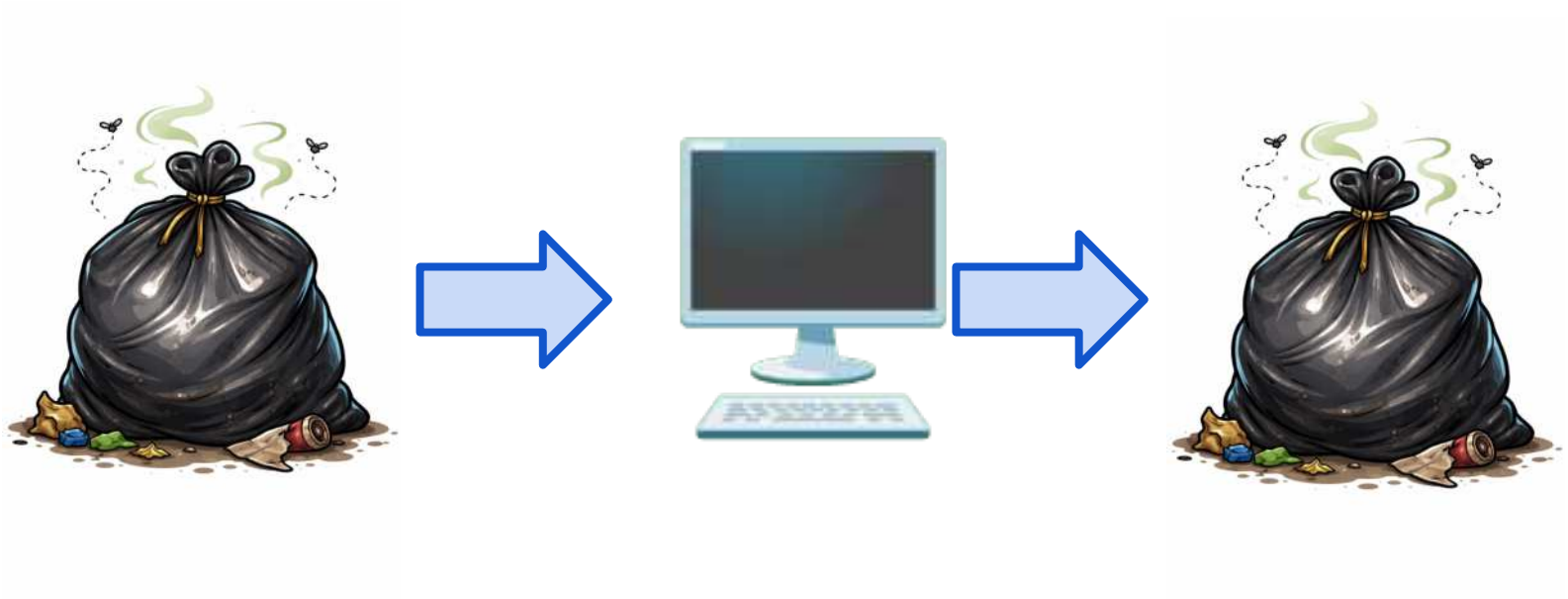
REPRESENTATIONS

ALGORITHMS

DATA REPRESENTATIONS ALGORITHMS

Why **Data** matters?

ML learns **from examples** so...



ML **excels** when **data** is

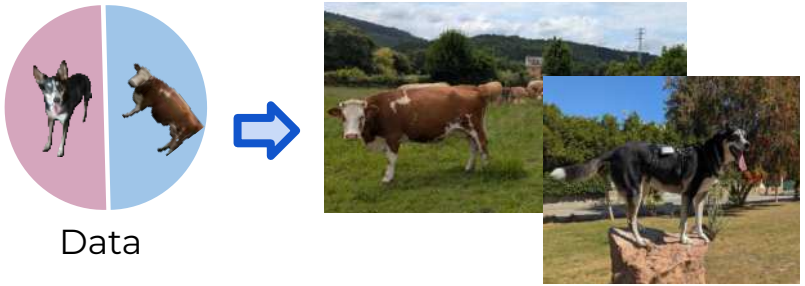
Large

~13 trillion tokens
~14 million images

Clean, consistent and self-defined

Dog Cow

Representative and balanced



Data

ML **excels** when **data** is

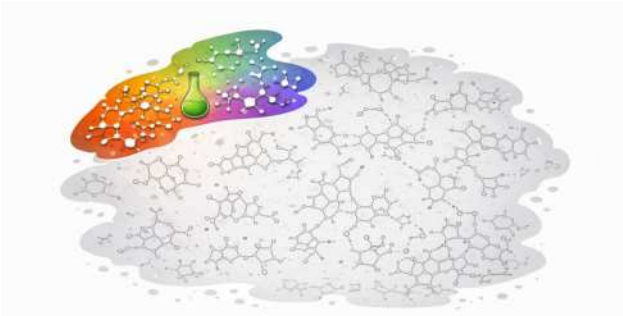
Large

~13 trillion tokens
~14 million images

ChEMBL	2.9 million compounds
Drug discovery campaign	500.000 compounds

Clean, consistent and self-defined

Truncated values, different experimental conditions, several values for the same molecule, small value range etc...

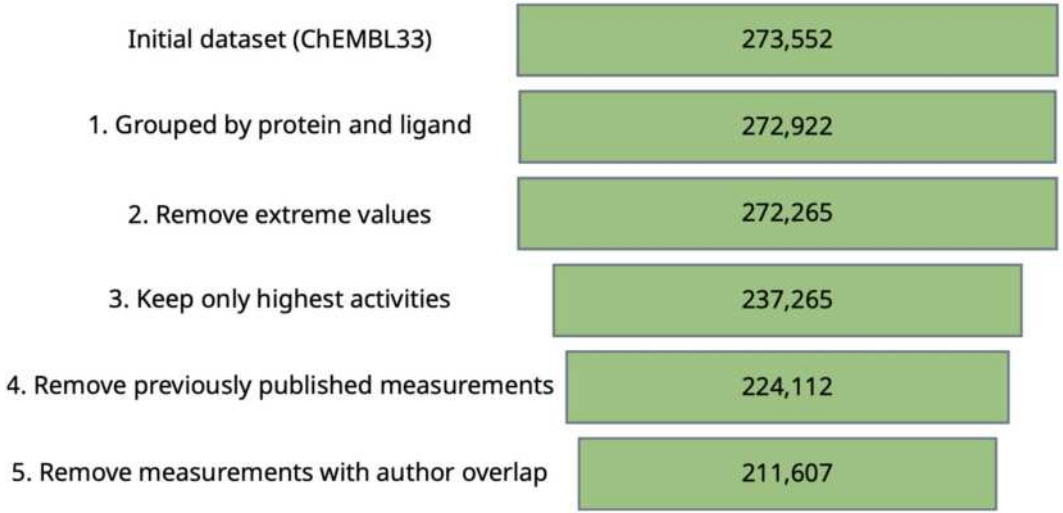


Representative and balanced

Uneven and small coverage of the chemical space

(Public) **databases**

Processing and **curating** chemical **data**



de Castro, Raquel López-Ríos, et al. *Living Journal of Computational Molecular Science* 6.1 (2025): 3875-3875.

KinoML and Kinodata:
https://github.com/openkinome/kinodata/blob/master/kinase-bioactivities-in-chembl/kinase-bioactivities-in-chembl.ipynb

Kramer, et al. *Journal of medicinal chemistry*55.11 (2012): 5165-5173.

Processing and **curating** chemical **data**

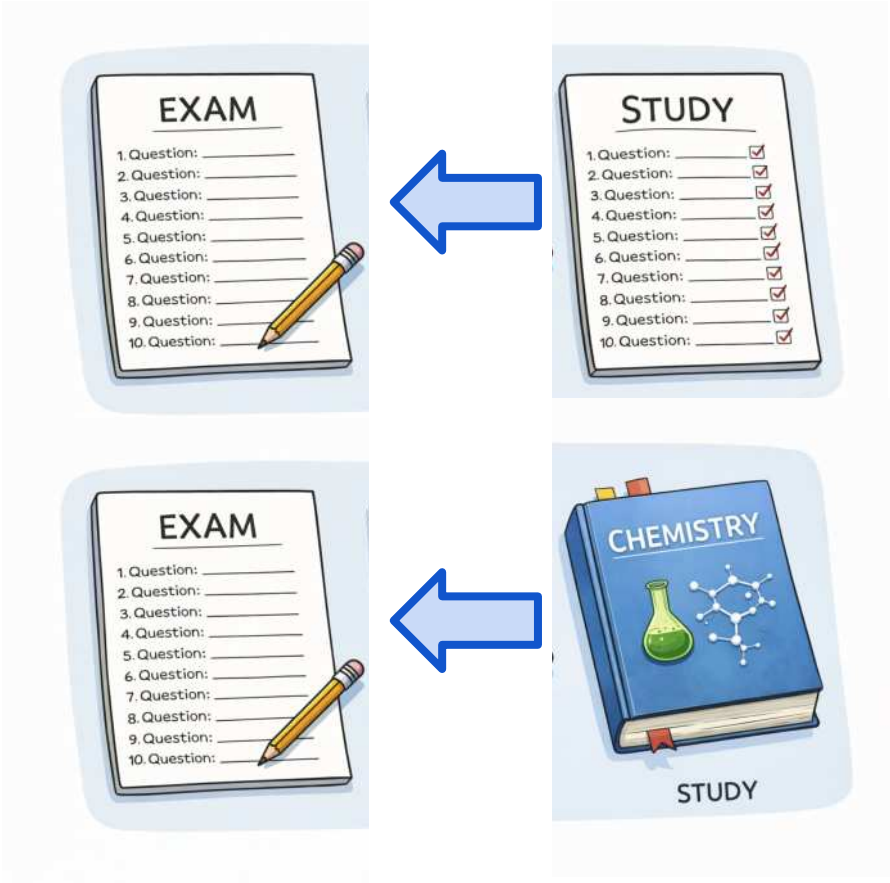
Initial dataset (ChEMBL33)	273,552
1. Grouped by protein and ligand	272,922
2. Remove extreme values	272,265
3. Keep only highest activities	237,265
4. Remove previously published measurements	224,112
5. Remove measurements with author overlap	211,607

de Castro, Raquel López-Ríos, et al. *Living Journal of Computational Molecular Science* 6.1 (2025): 3875-3875.

KinoML and Kinodata:
https://github.com/openkinome/kinodata/blob/master/kinase-bioactivities-in-chembl/kinase-bioactivities-in-chembl.ipynb

Kramer, et al. *Journal of medicinal chemistry*55.11 (2012): 5165-5173.

Data **splitting**



❌

✅

Data **splitting**

K-fold **splitting**

E.g. 5-fold:

Fold 1: Train on Folds 2, 3, 4, 5 | Test on Fold 1

Fold 2: Train on Folds 1, 3, 4, 5 | Test on Fold 2

Fold 3: Train on Folds 1, 2, 4, 5 | Test on Fold 3

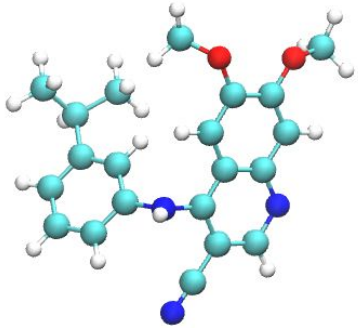
Fold 4: Train on Folds 1, 2, 3, 5 | Test on Fold 4

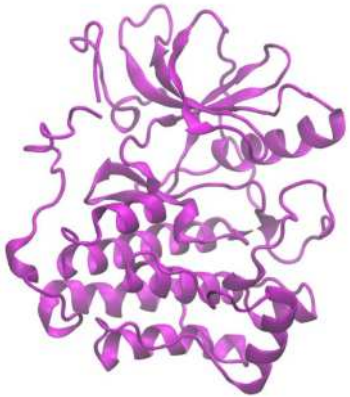
Fold 5: Train on Folds 1, 2, 3, 4 | Test on Fold 5

Specially useful when **limited data** is available!

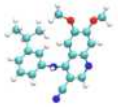
Data **splitting**

You can split by:

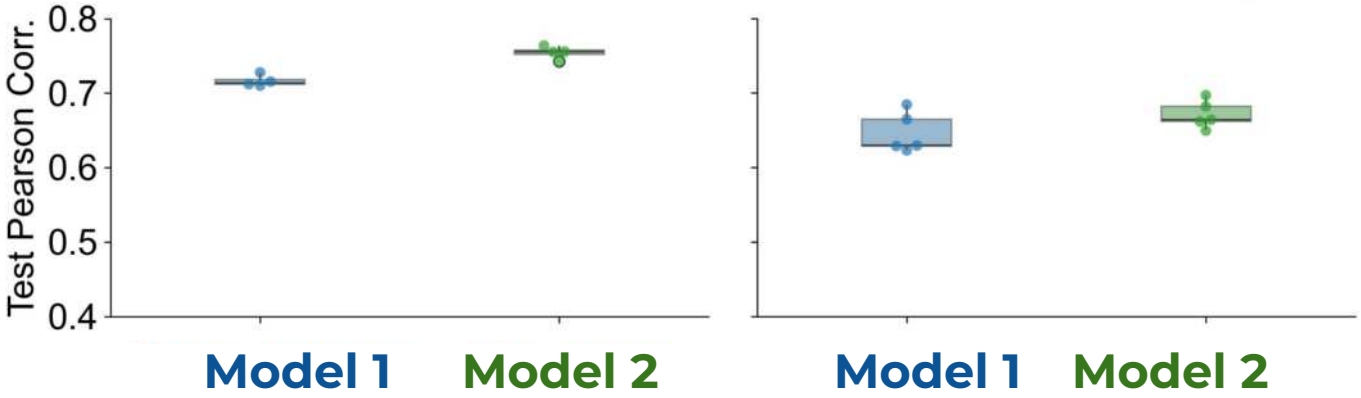




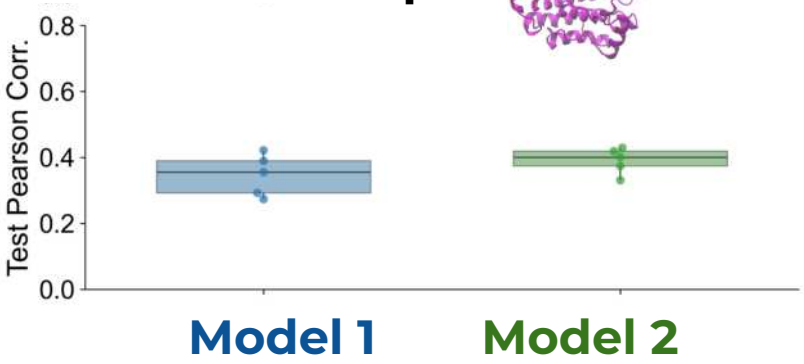
Random split



Scaffold split



Pocket split



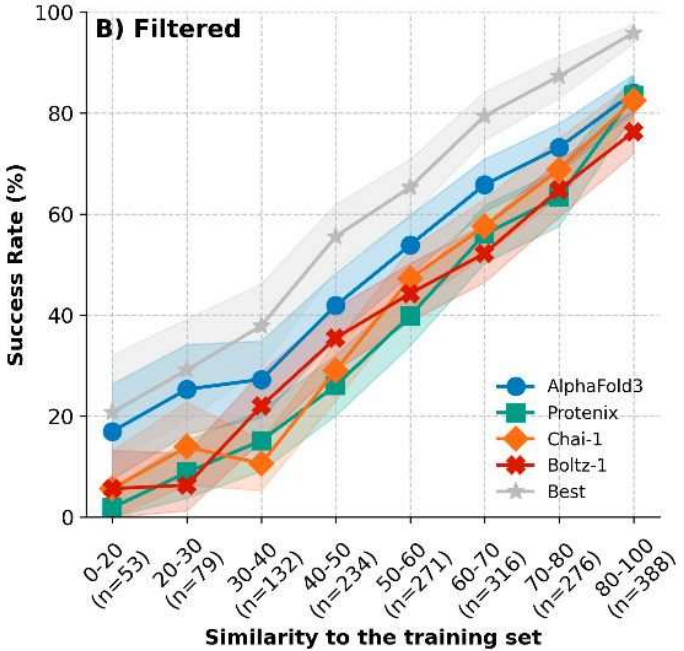
HAVE PROTEIN-LIGAND CO-FOLDING METHODS MOVED BEYOND MEMORISATION?

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
peter.skrinjar@unibas.ch

Jérôme Eberhardt
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
jerome.eberhardt@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
janani.durairaj@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
torsten.schwede@unibas.ch



ABSTRACT

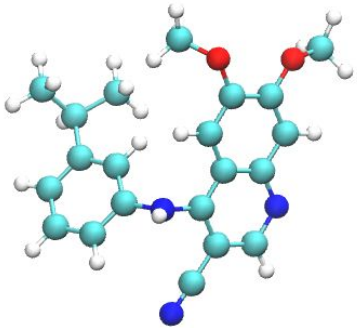
Deep learning has driven major breakthroughs in protein structure prediction, however the next critical advance is accurately predicting how proteins interact with other molecules, especially small molecule ligands, to enable real-world applications such as drug discovery and design. Recent deep learning all-atom methods have been built to address this challenge, but evaluating their performance on the prediction of protein-ligand complexes has been inconclusive due to the lack of relevant benchmarking datasets. Here we present a comprehensive evaluation of four leading all-atom cofolding deep learning methods using our newly introduced benchmark dataset Runs N' Poses, which comprises 2,600 high-resolution protein-ligand systems released after the training cutoff used by these methods. We demonstrate that current co-folding approaches largely memorise ligand poses from their training data, hindering their use for *de novo* drug design. This limitation is especially pronounced for ligands that have only been seen binding in one pocket, whereas more promiscuous ligands such as cofactors show moderately improved performance. With this work and benchmark dataset, we aim to accelerate progress in the field by allowing for a more realistic assessment of the current state-of-the-art deep learning methods for predicting protein-ligand interactions.

DATA

REPRESENTATIONS

ALGORITHMS

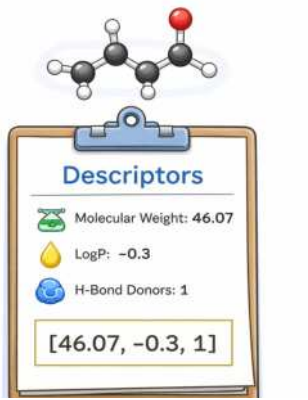
What is a **representation**?



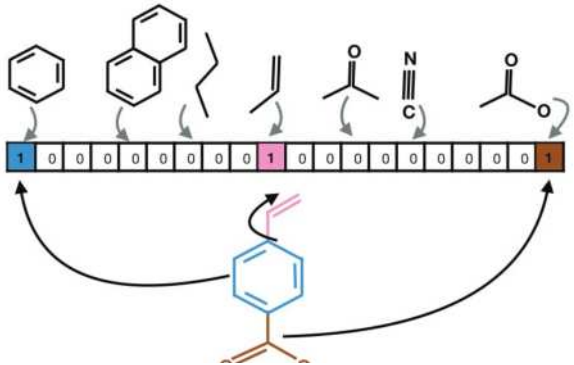
Featurization

Chemical structure **Vector**

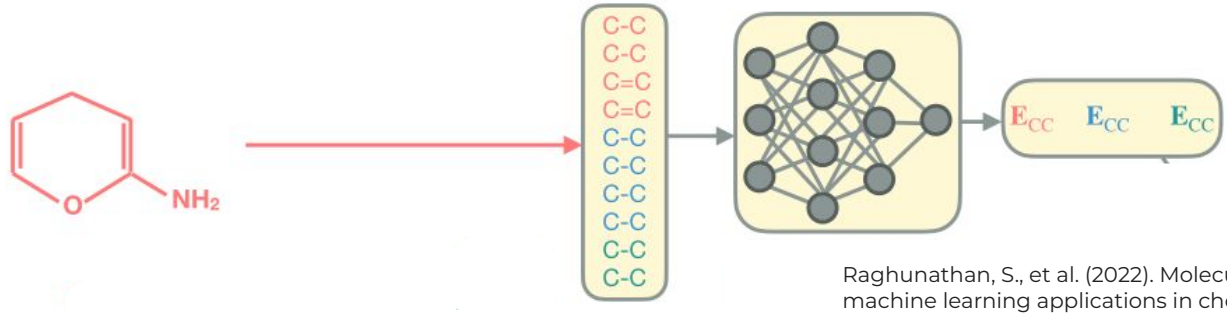
Descriptors (molecular properties)



Chemical fingerprints



Graphs

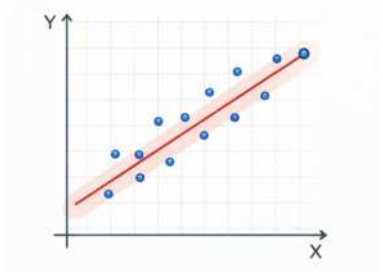


Raghunathan, S., et al. (2022). Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chemistry*, *122*(7), e26870.

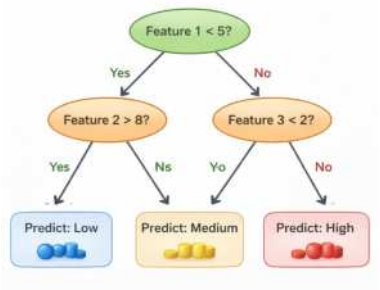
DATA REPRESENTATIONS **ALGORITHMS**

Algorithms are the **mathematical methods** that **identify**/learn the **patterns** in data

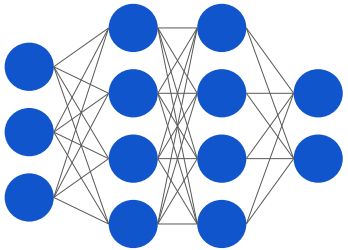
Linear models



Tree-based models



Neural networks



Complexity

Interpretability

Conclusions

- Data curation!
- Never trust a result until you see their data strategy
- Towards better curated **open-source molecular databases**

AI in chemistry workshop

https://github.com/volkamerlab/ai_in_chemistry_workshop_2025/tree/main

TeachopenCADD

https://github.com/volkamerlab/teachopencadd

KinoData

https://github.com/volkamerlab/kinodata-3D

KinoML

https://github.com/openkinome/kinoml