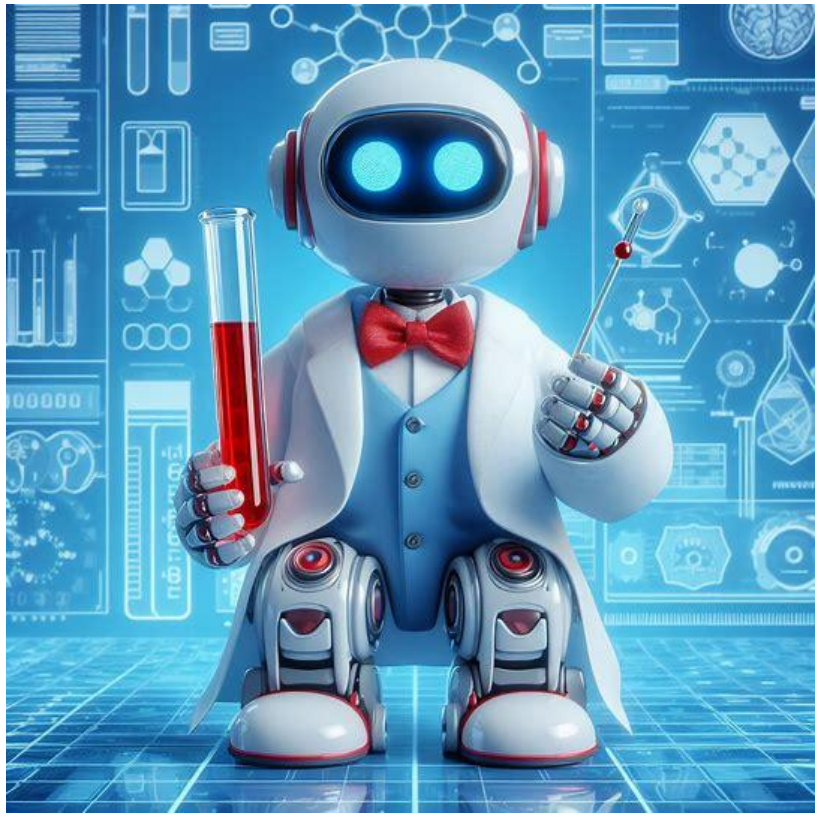


AI in Drug Discovery – An Overview
Session 1

September 16, 2024

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

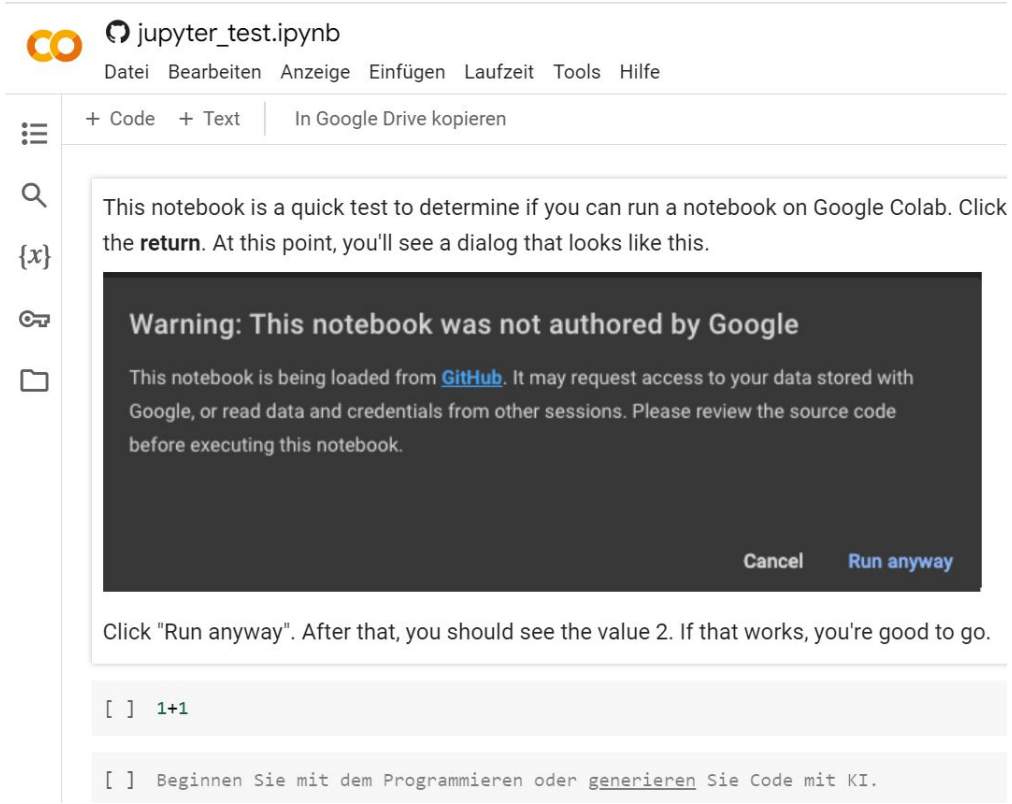
Session 2 - 3:00 - 4:00 pm

- 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 …



Tutorials and Slides Can Be Found On GitHub

https://github.com/volkamerlab/ai_in_chemistry_workshop

Definitions Can Be Tricky

artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.

Not a well-defined statement

https://www.britannica.com/technology/artificial-intelligence

AI and “The Rise of the Machines”

Andrew Chen Retweeted

Mat Velloso @matvelloso · Nov 22

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**

166 6.6K 19K

Show this thread

What Is Machine Learning?

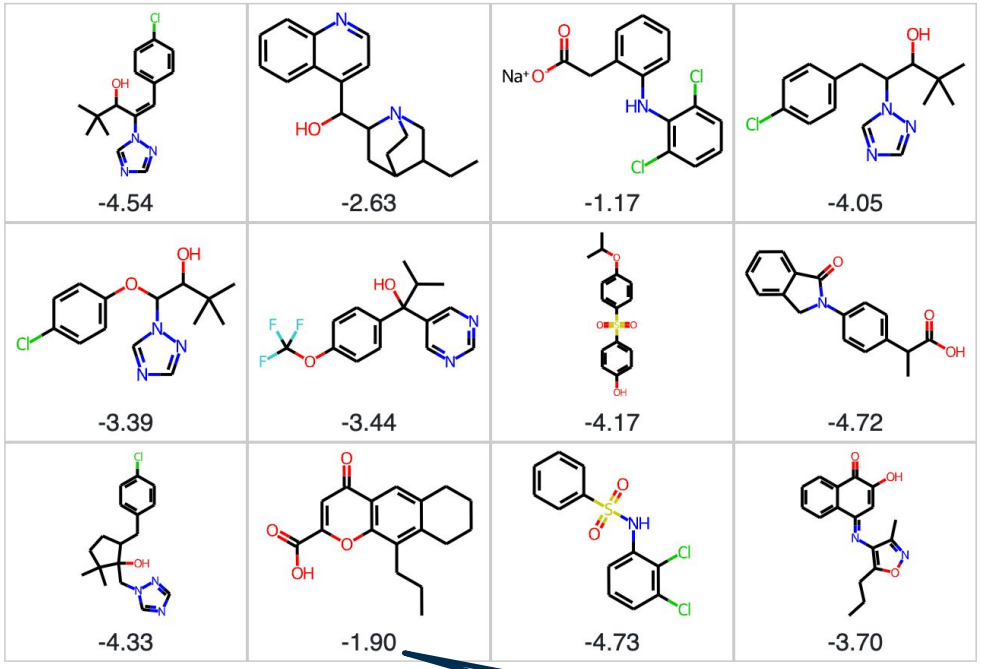
Machine learning is all about labeling things using examples

Cassie Kozyrkov, Google

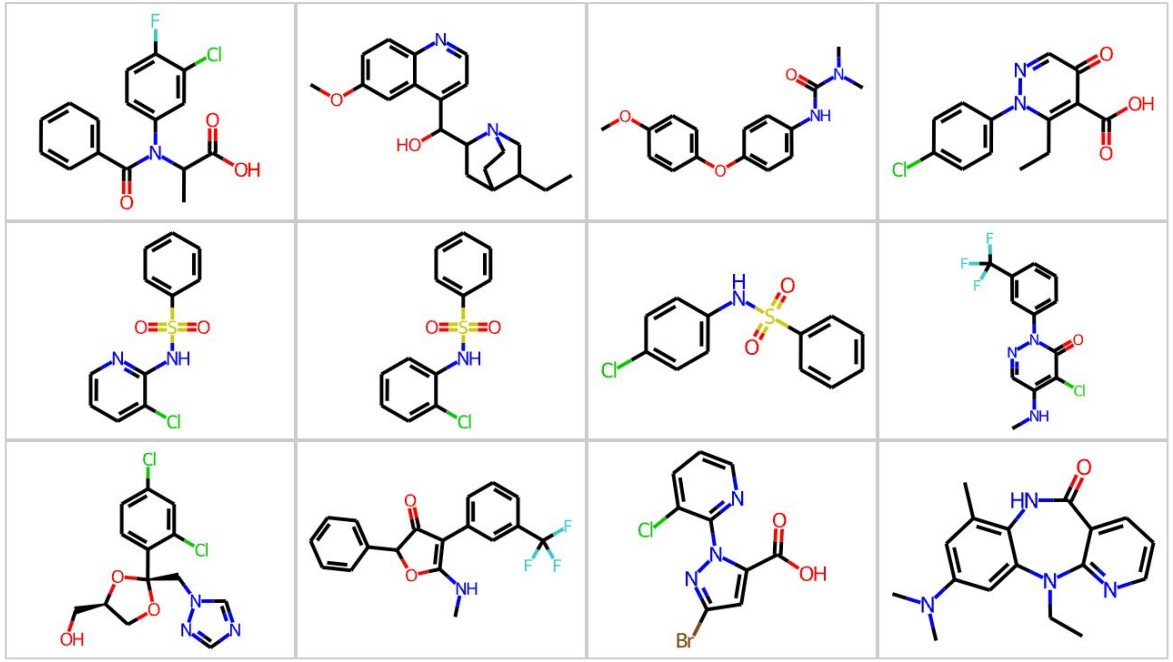
https://hackernoon.com/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c

Labeling Molecules Based on Examples

Molecules with measured data



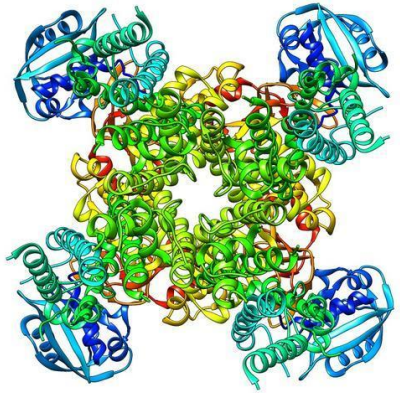
Molecules to be predicted



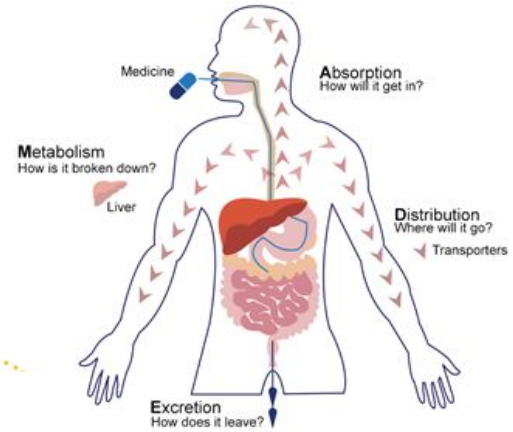
Log10(Molar Aqueous Solubility)

Using Predictive Models to Drive Drug Discovery

On-target Activity



Off-target Activity

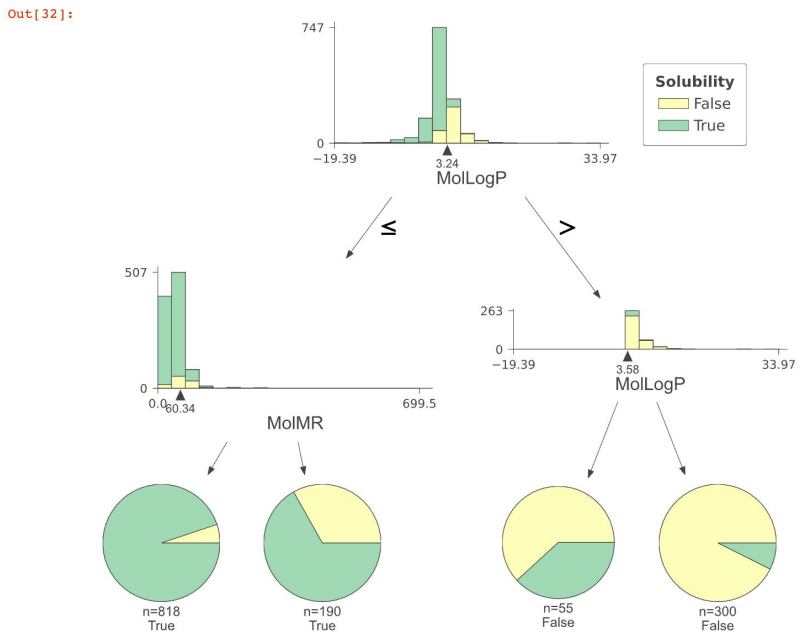


Pharmacokinetics

Physical Properties

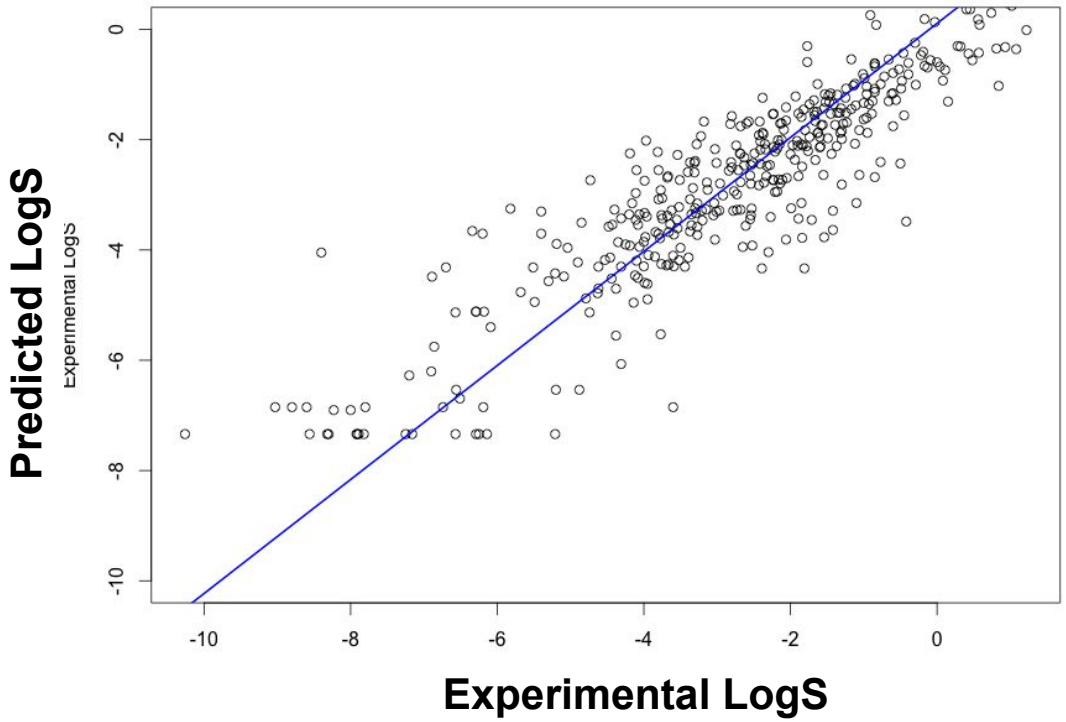
Two Types of ML Models – Classification and Regression

Classification



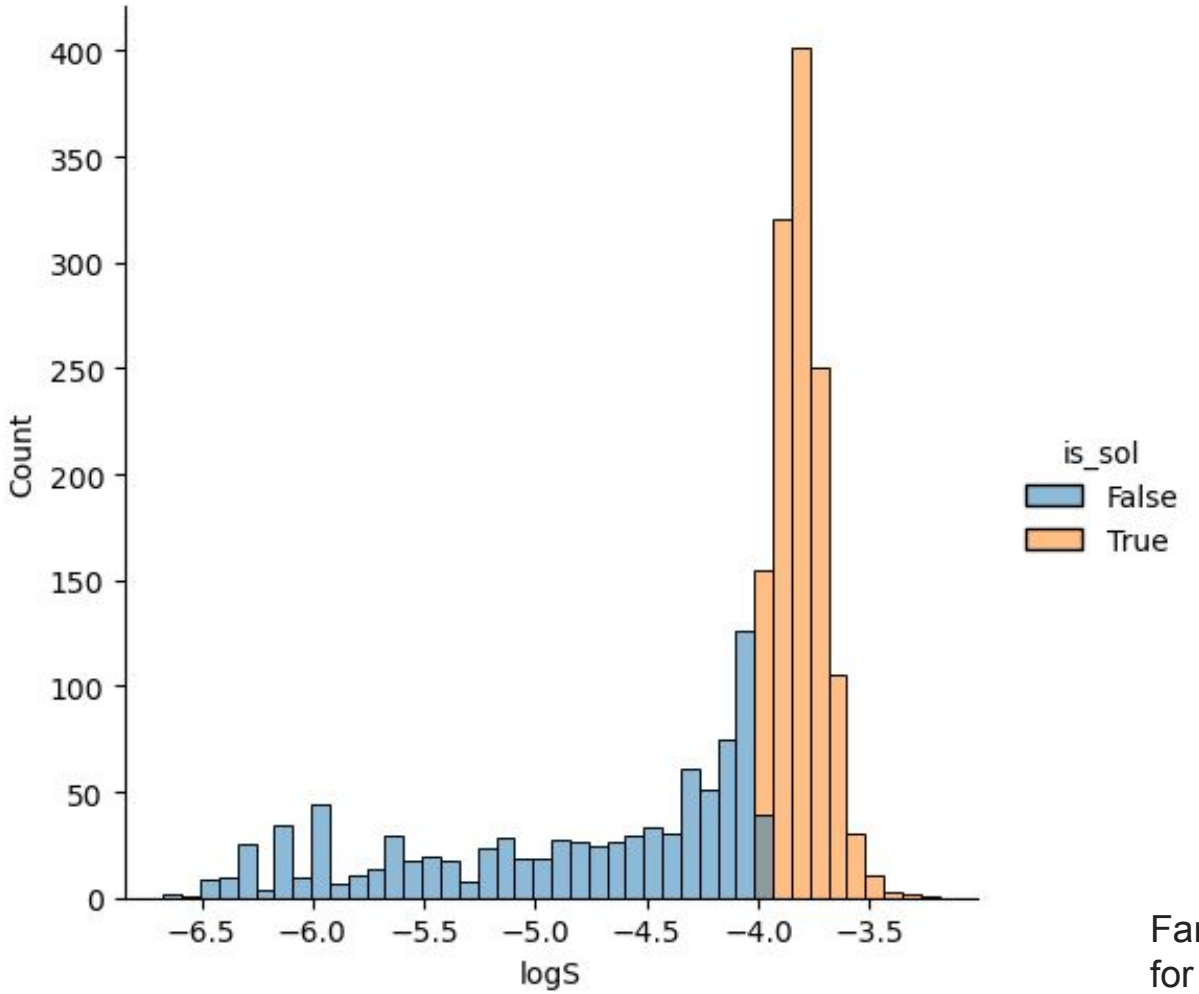
Predict a category, e.g. soluble/insoluble

Regression



Predict a value, e.g. 4.2

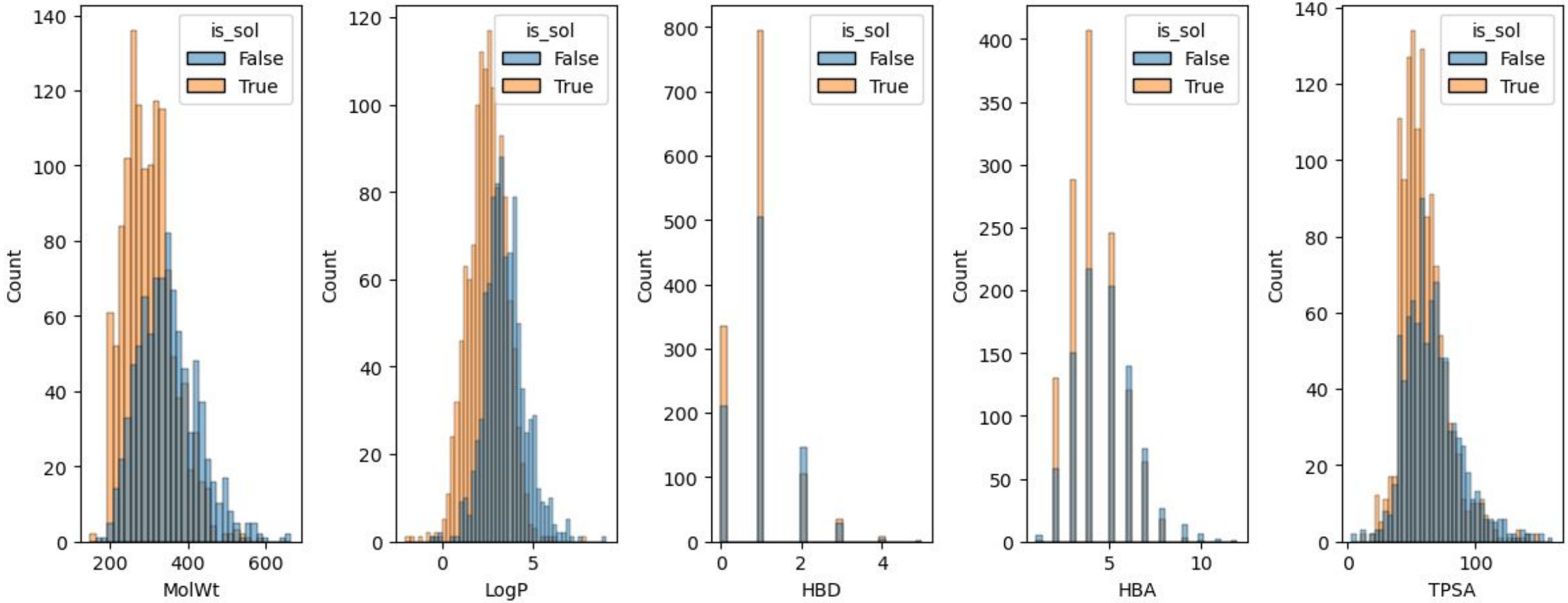
Let’s Start With a Dataset



Aqueous solubility by CLND
2173 compounds
Min = 0.23μM
Max = 648μM

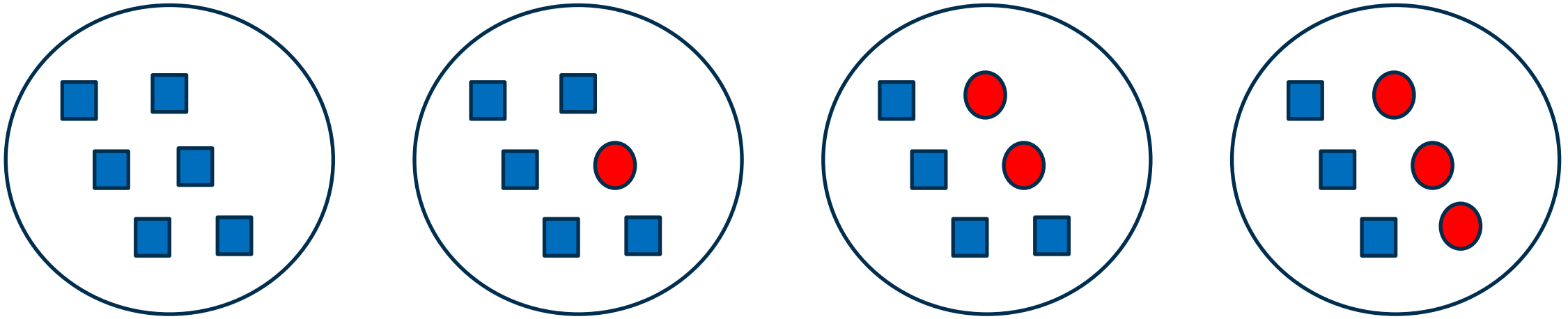
Fang, Cheng, et al. "Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective." *Journal of Chemical Information and Modeling* 63.11 (2023): 3263-3274.

Which Property Best Separates Soluble vs Insoluble Molecules?



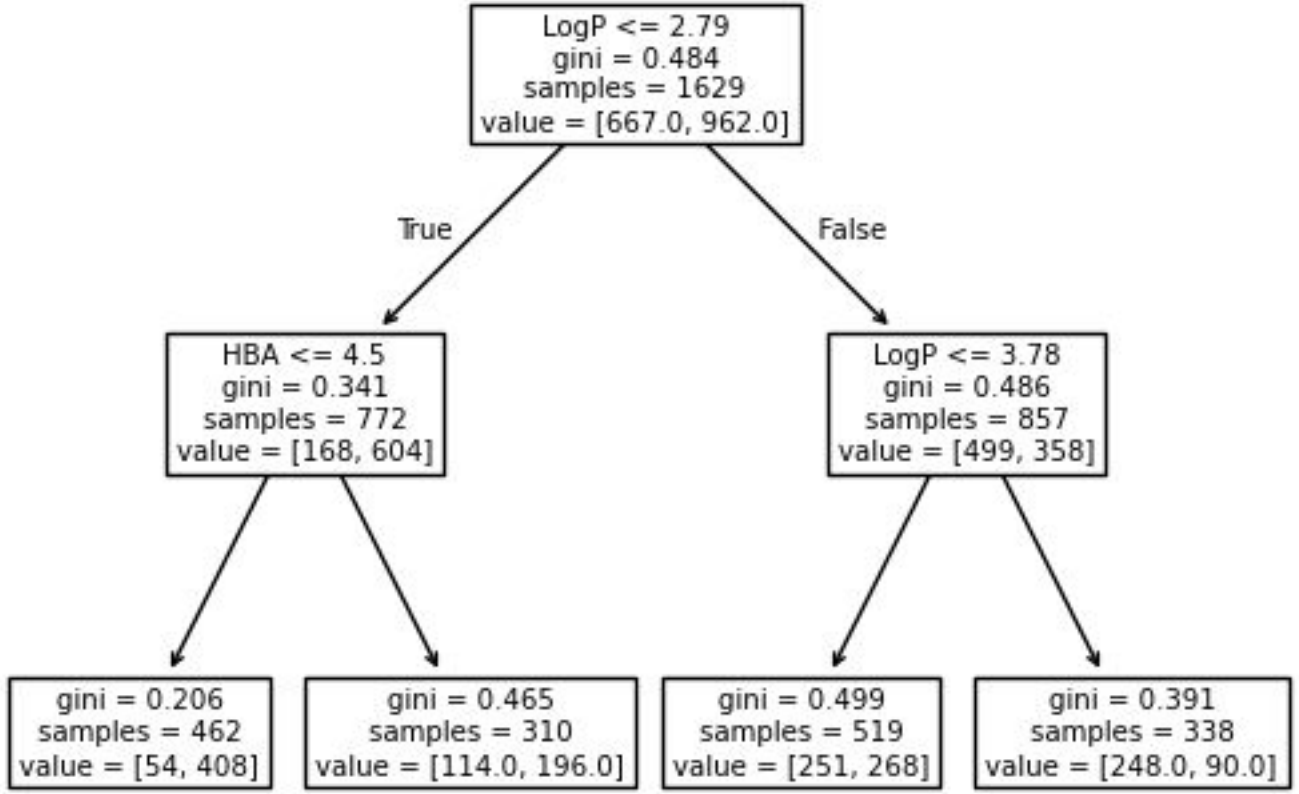
The Gini Index Quantifies the “Purity” of a Split

$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$

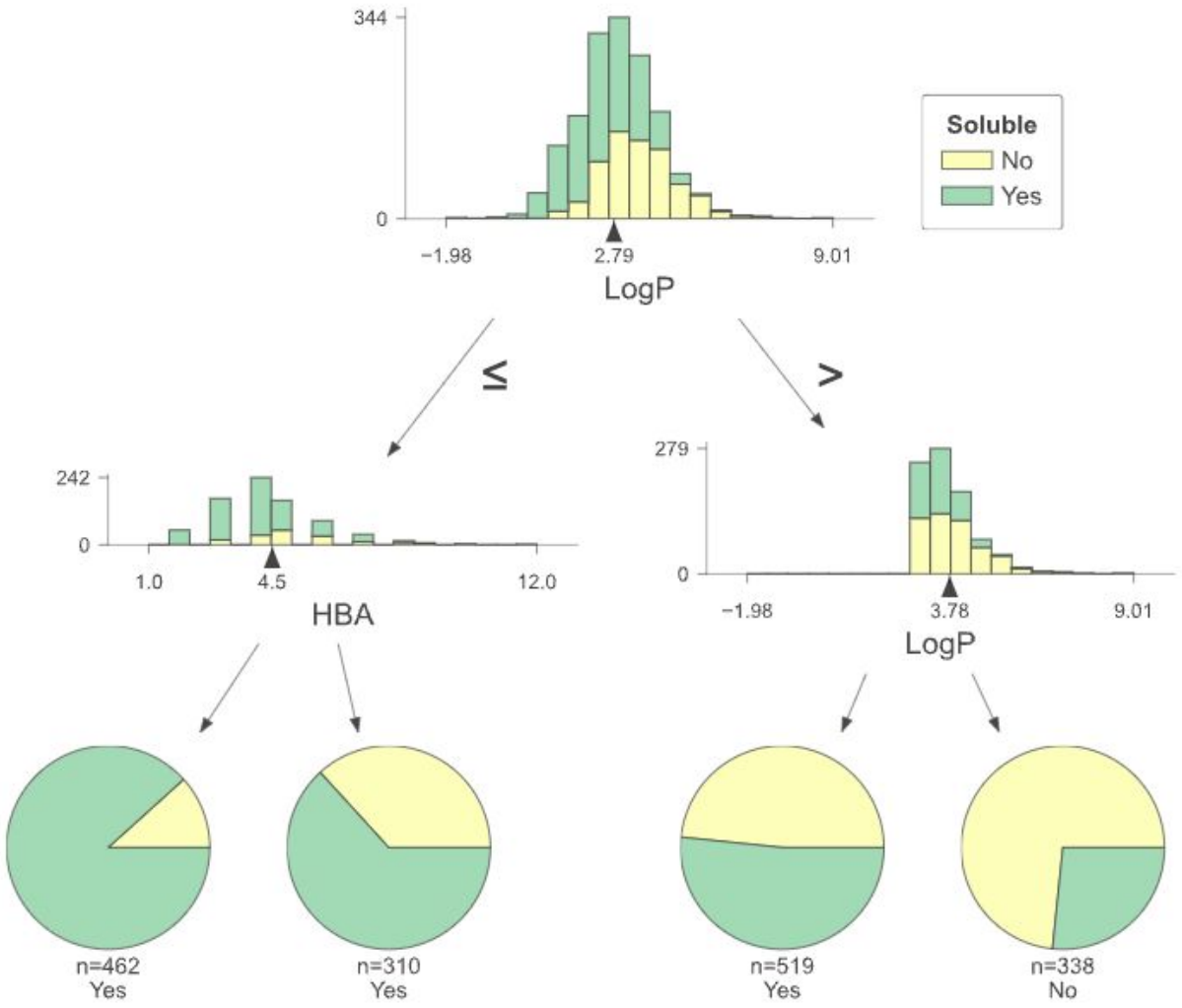


1.00 0.83 0.66 0.50

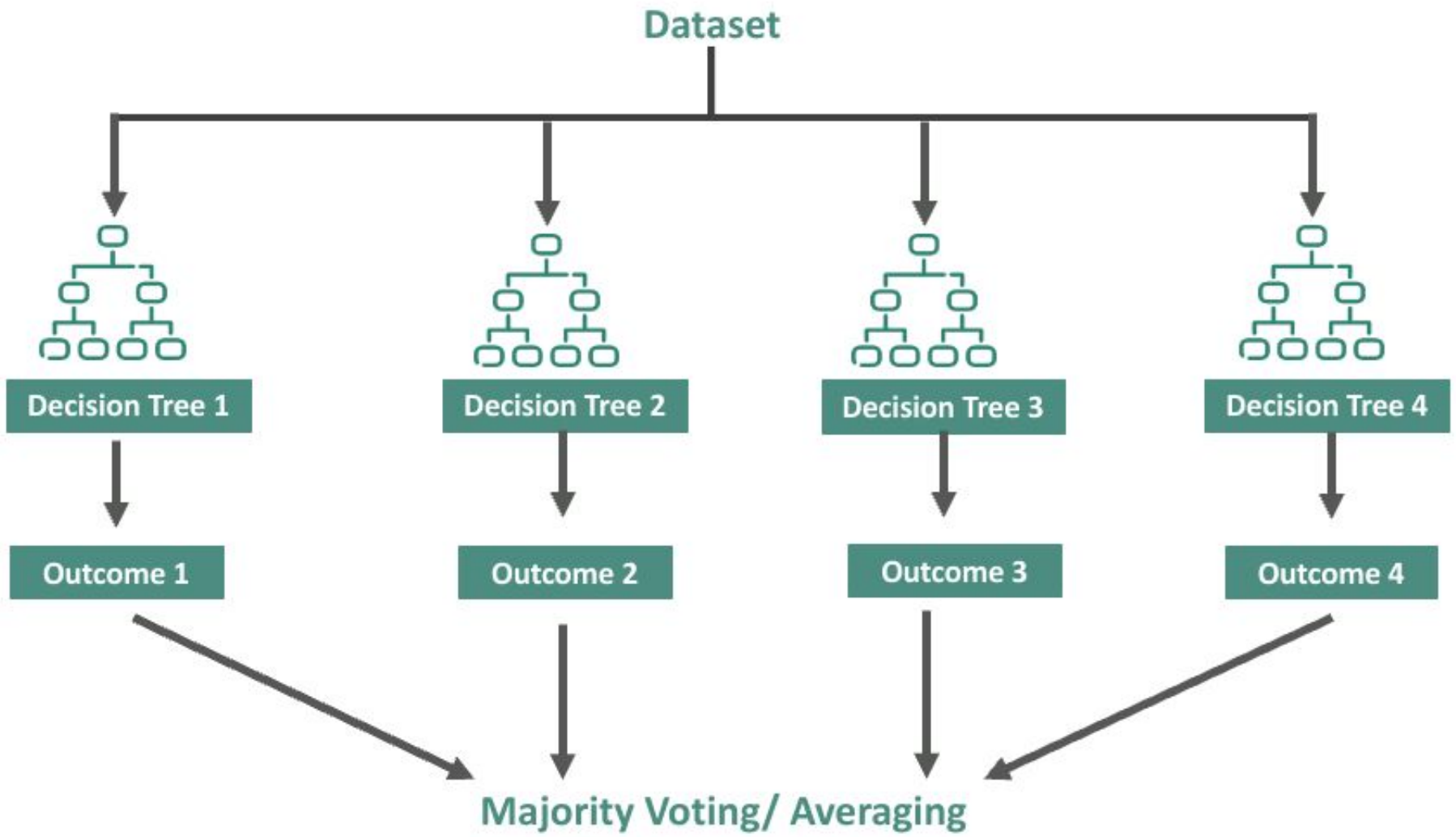
Build a Decision Tree



A Better (IMO) Decision Tree Visualization

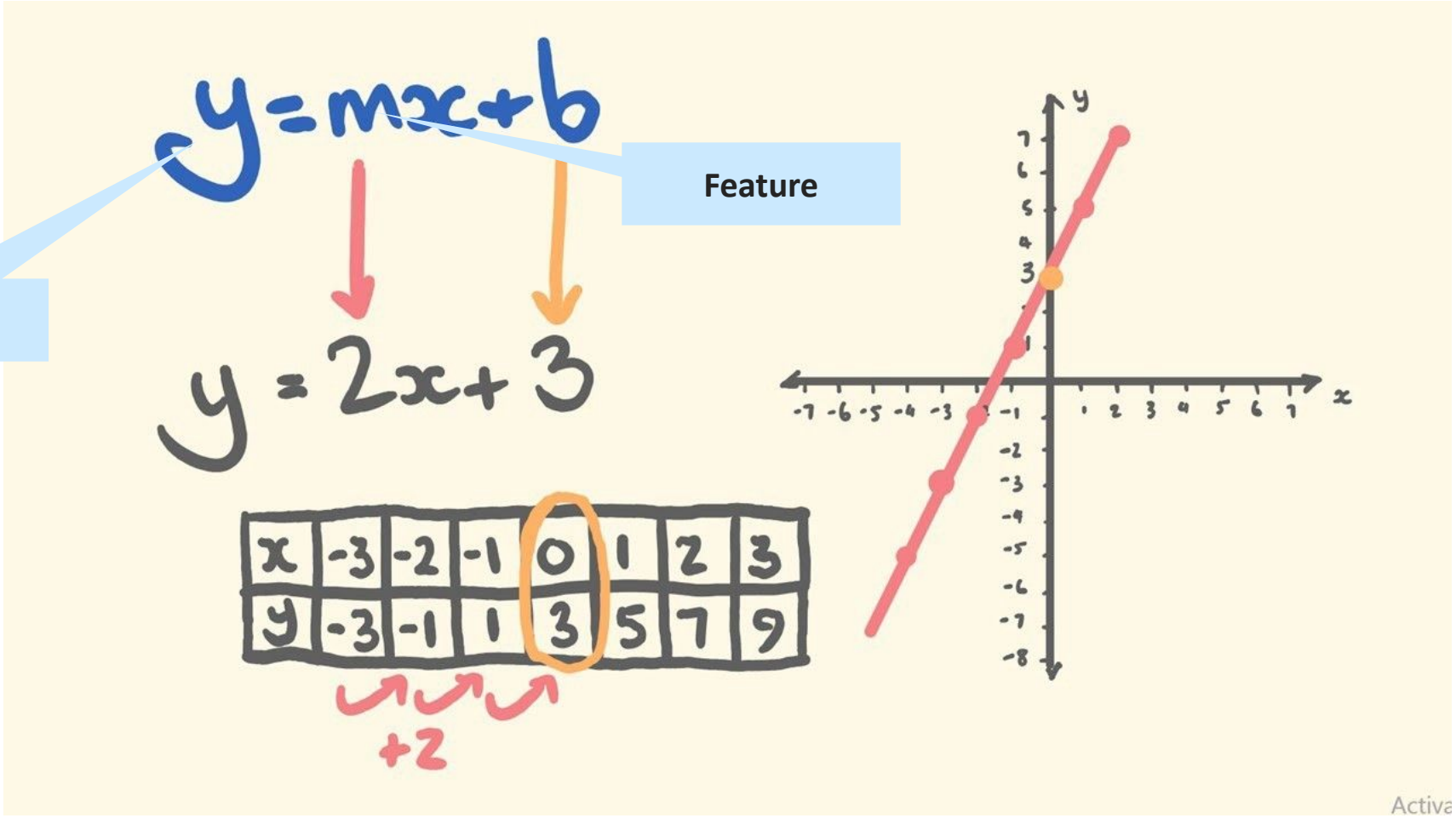


Random Forest Uses an Ensemble of Decision Trees

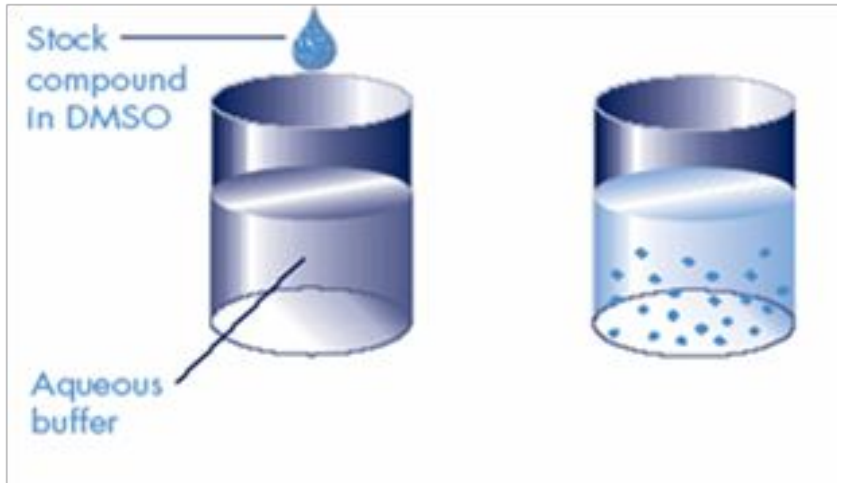


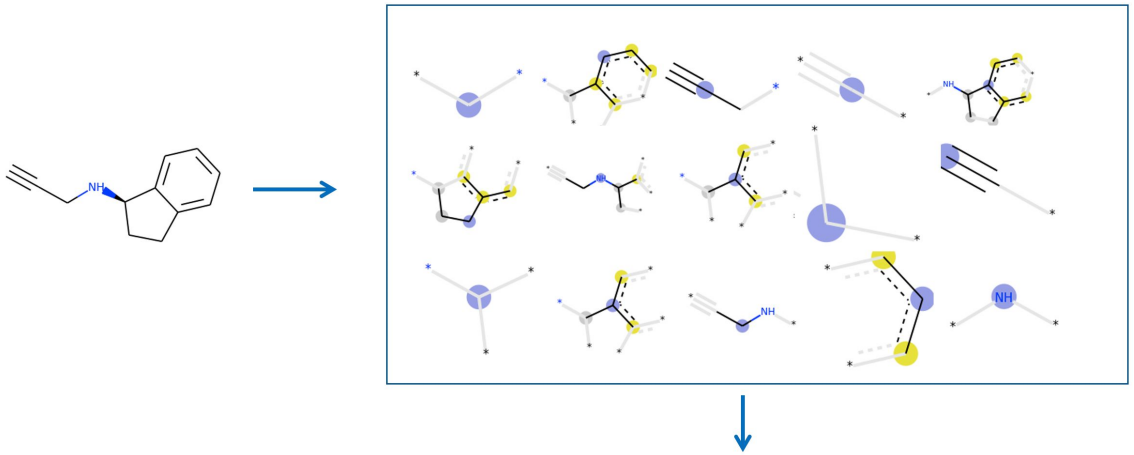
There Are Many Tree Ensemble Methods

ML Predicts a Set of Labels (y) Based on Features (X)



Defining Labels (y) and Features (X) - An Example





01011001000100011100100011001010101010111

y = Log Aqueous Solubility

X = A Vector Representing a Molecule

Label

Features

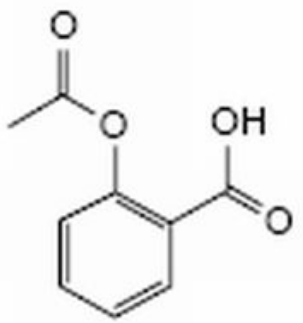
Define Features Based on Books People Have Read

Fred	1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0
Sally	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	0
Jane	0	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0

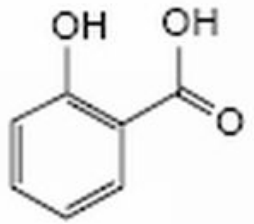
Create vectors representing books purchased by individuals
1 = bought book
0 = did not buy book

20

Chemical Fingerprints as Molecular Descriptors

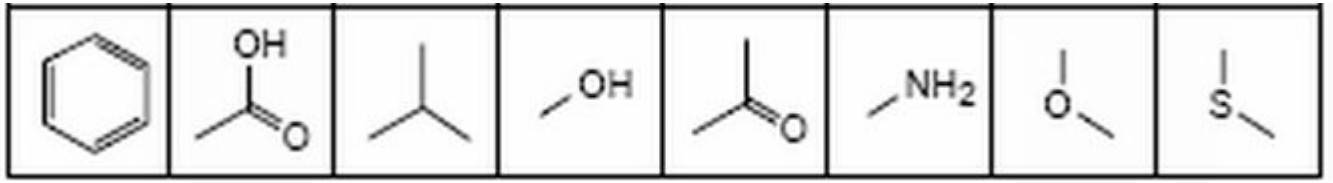


1

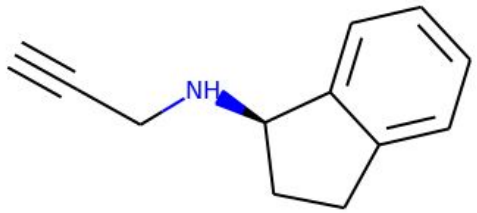


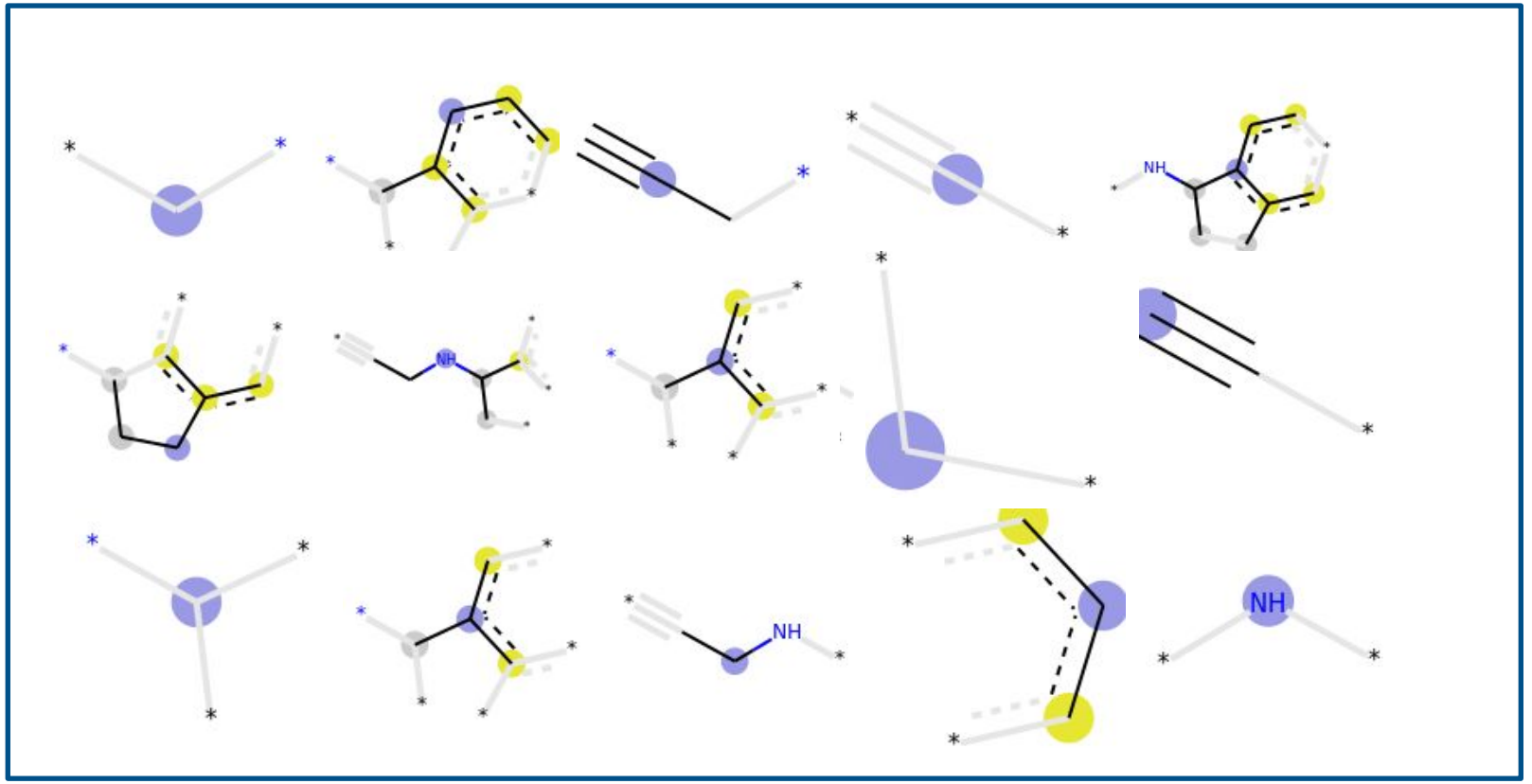
2

1	1	1	0	1	1	0	1	0
2	1	1	0	1	0	0	0	0

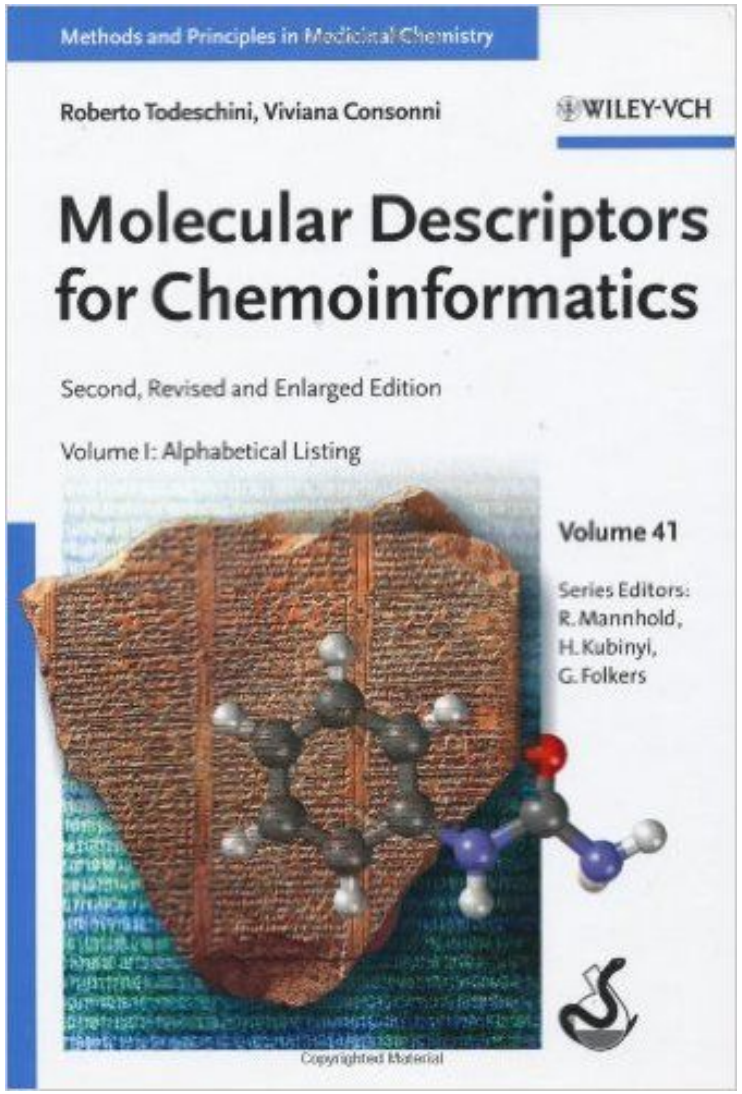


Chemical Fingerprints as Vector Representations of Molecules





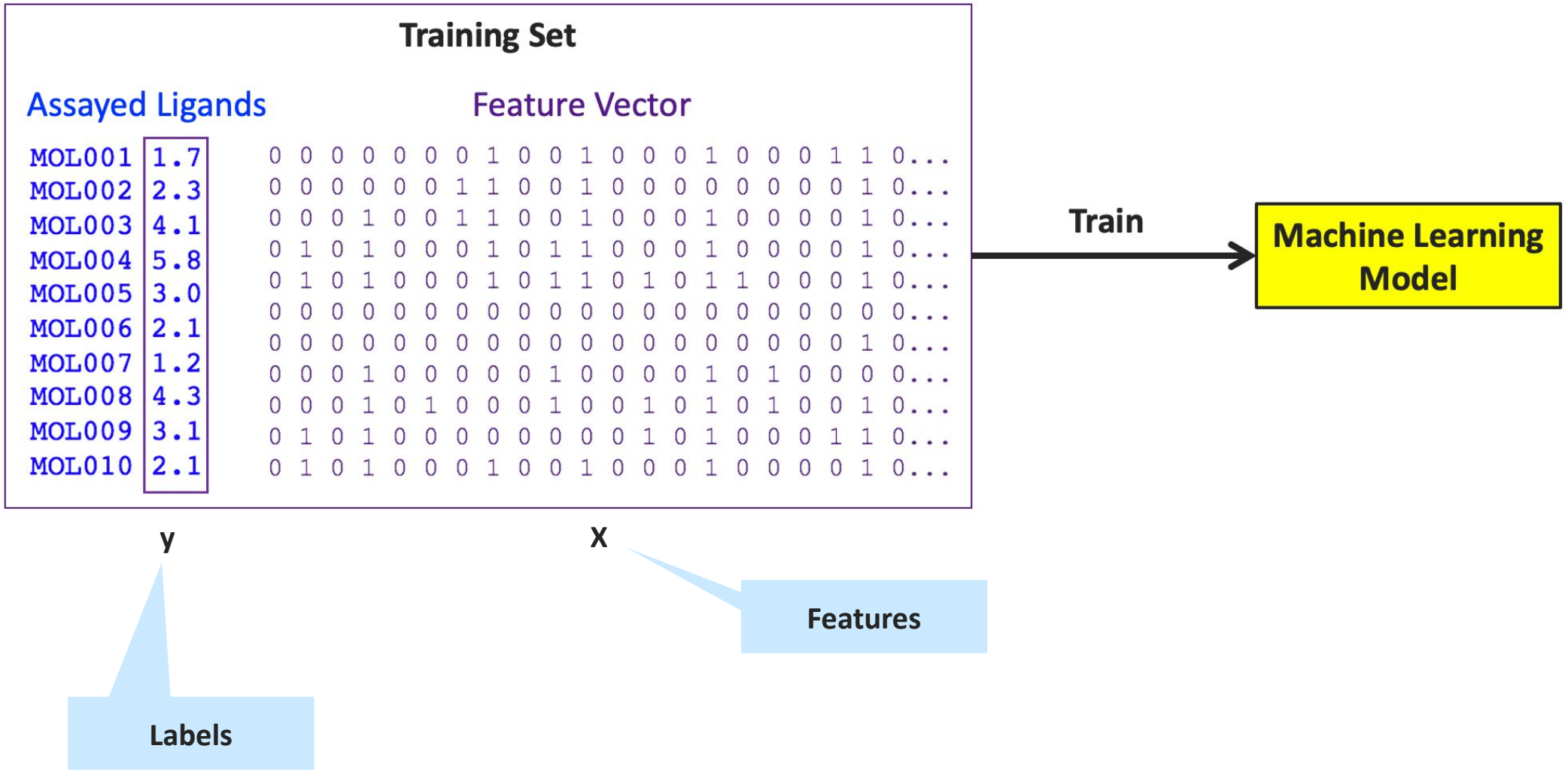
There are A LOT of ways to do this



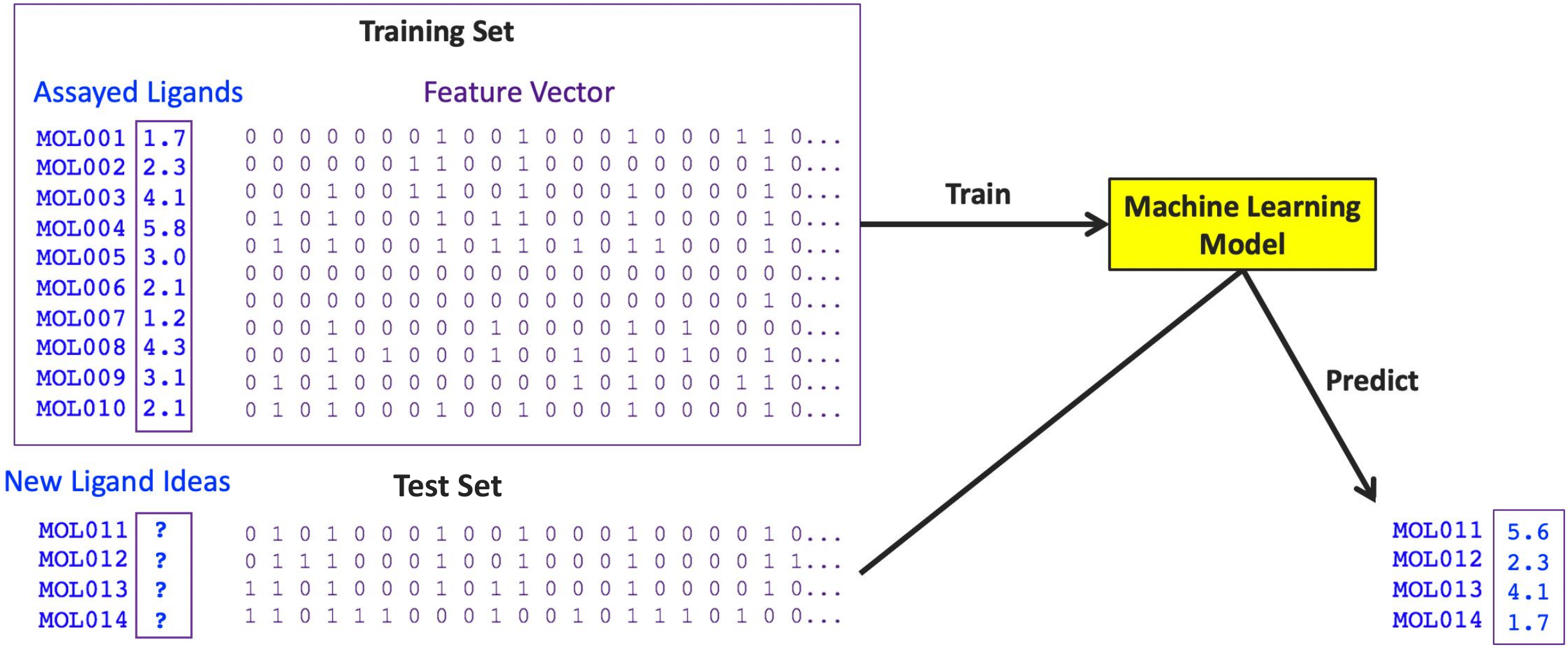
2 Volumes
6,000 references from 450 journals

DRAGON 7 has 5,270 descriptors z

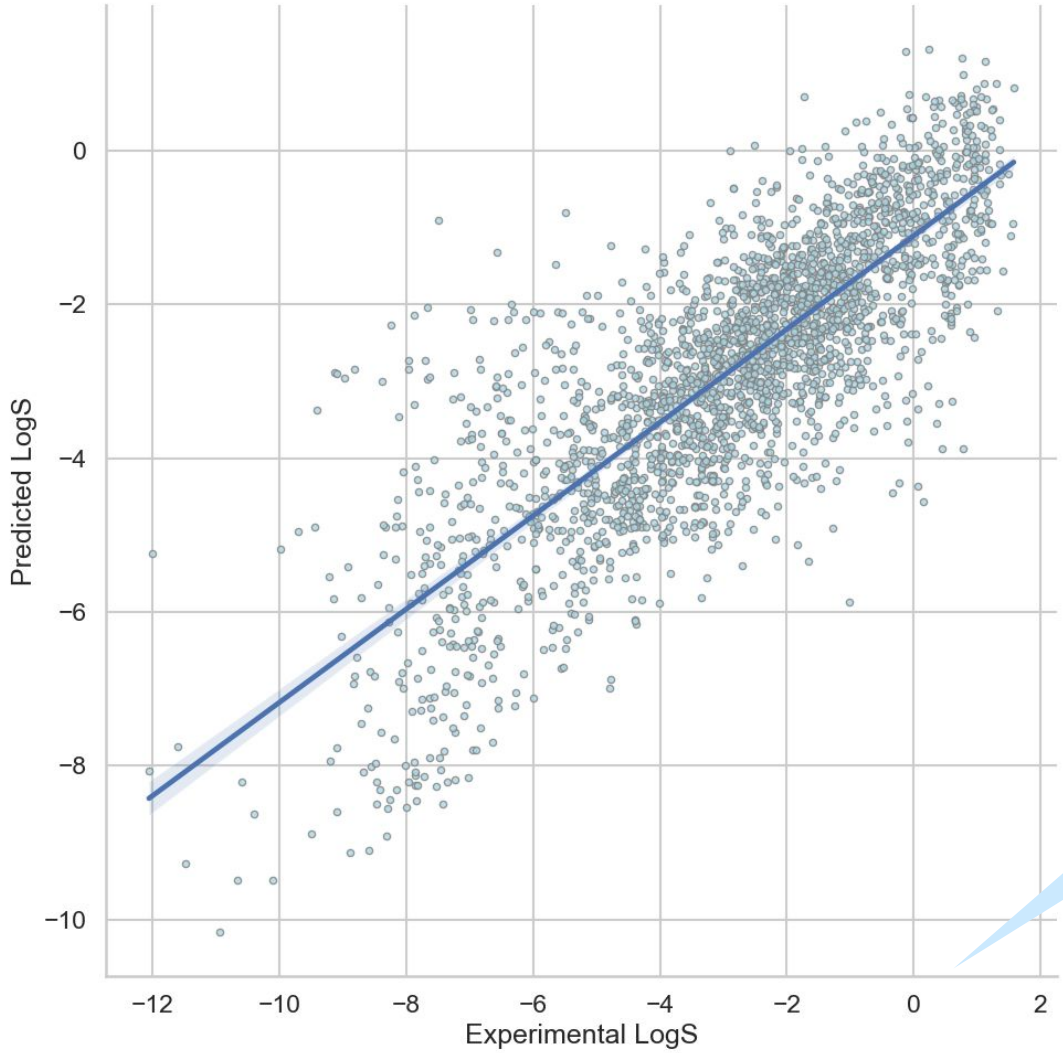
Training a Machine Learning Model



Making Predictions With a Machine Learning Model

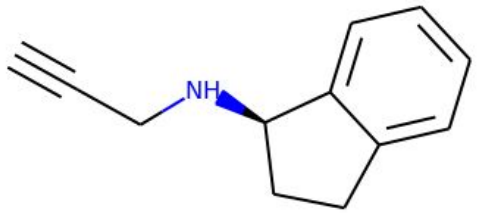


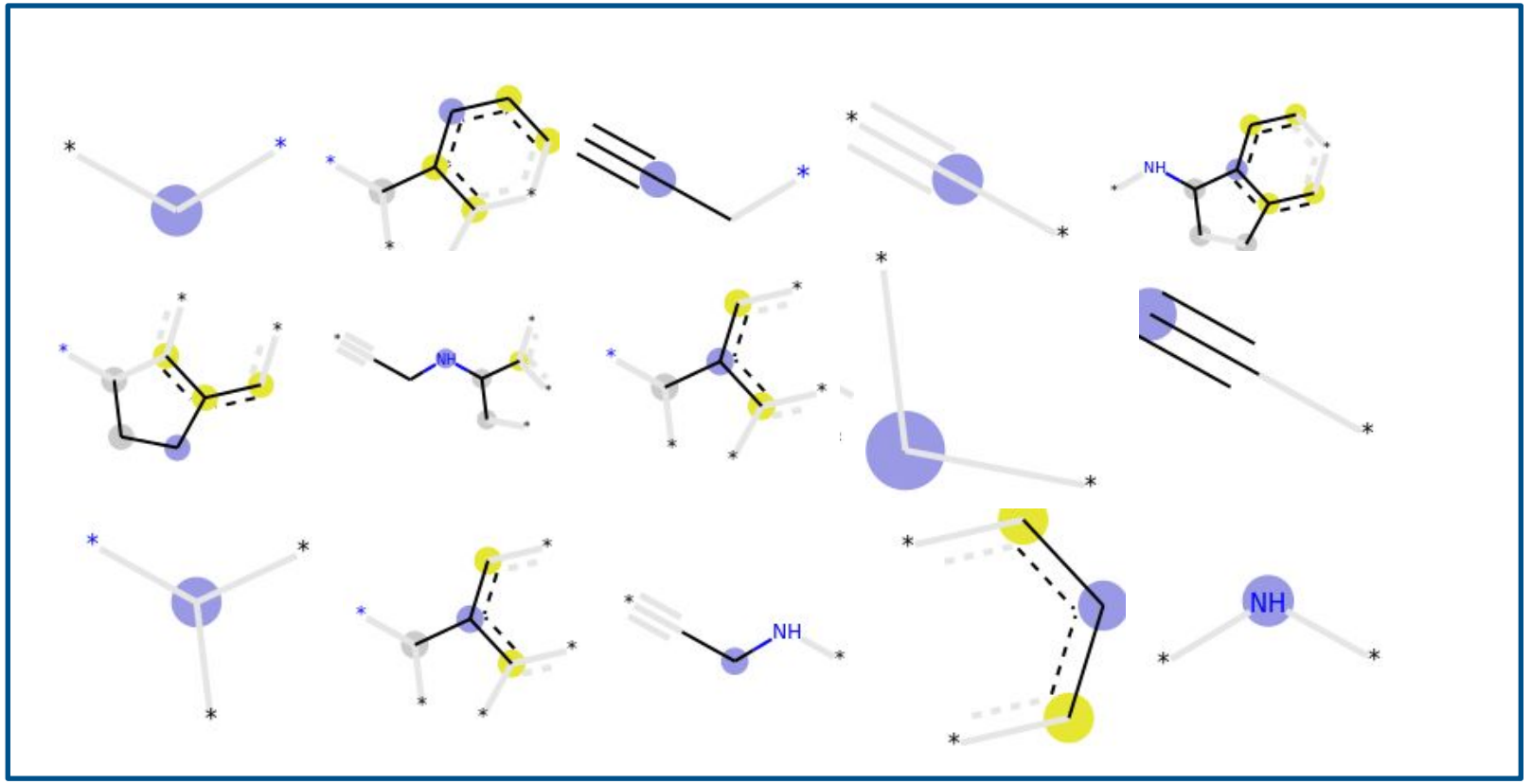
Prediction Performance of an Aqueous Solubility Model



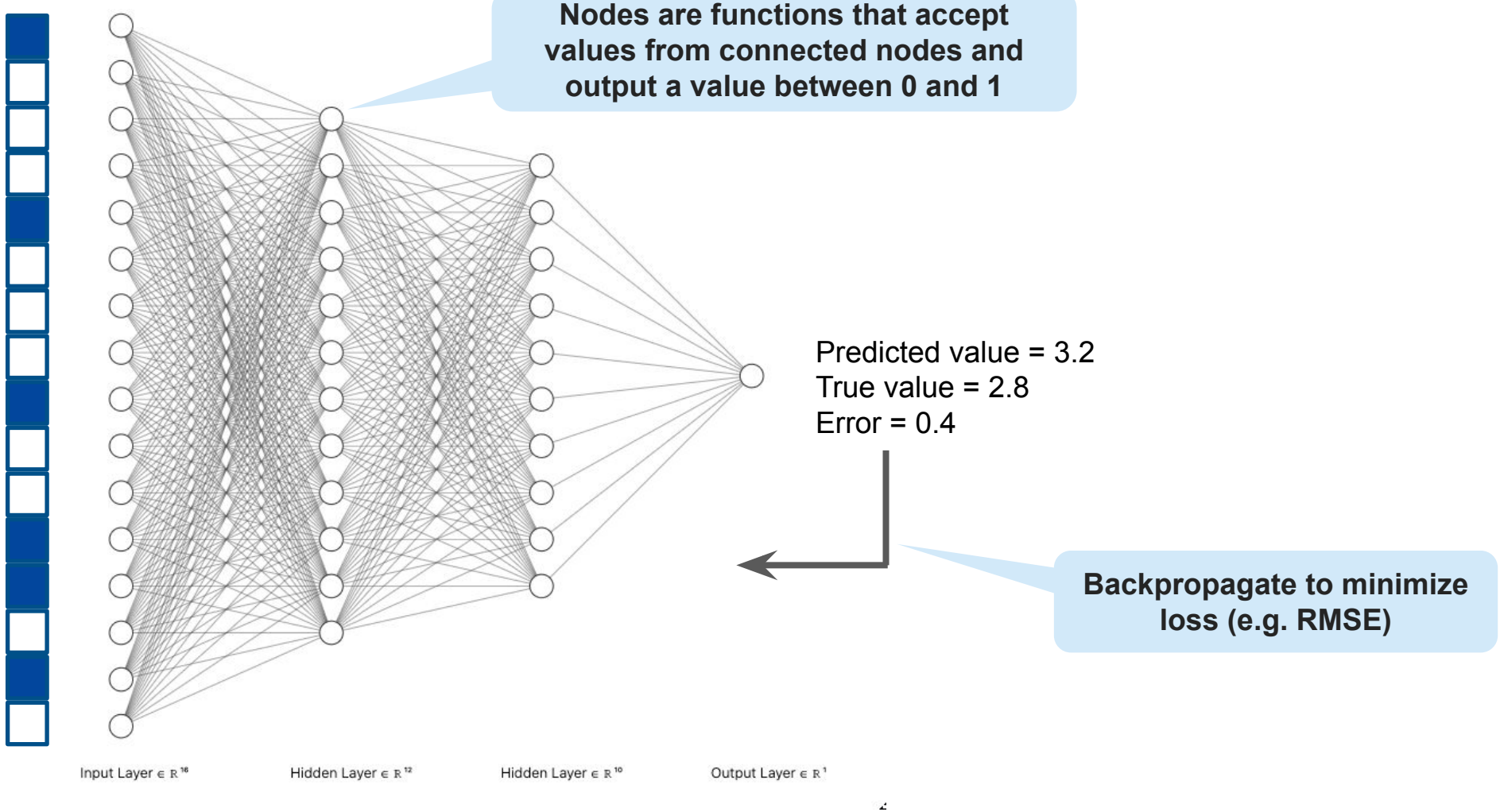
Dynamic range is very large, does not represent a typical use case

Chemical Fingerprints as Vector Representations of Molecules

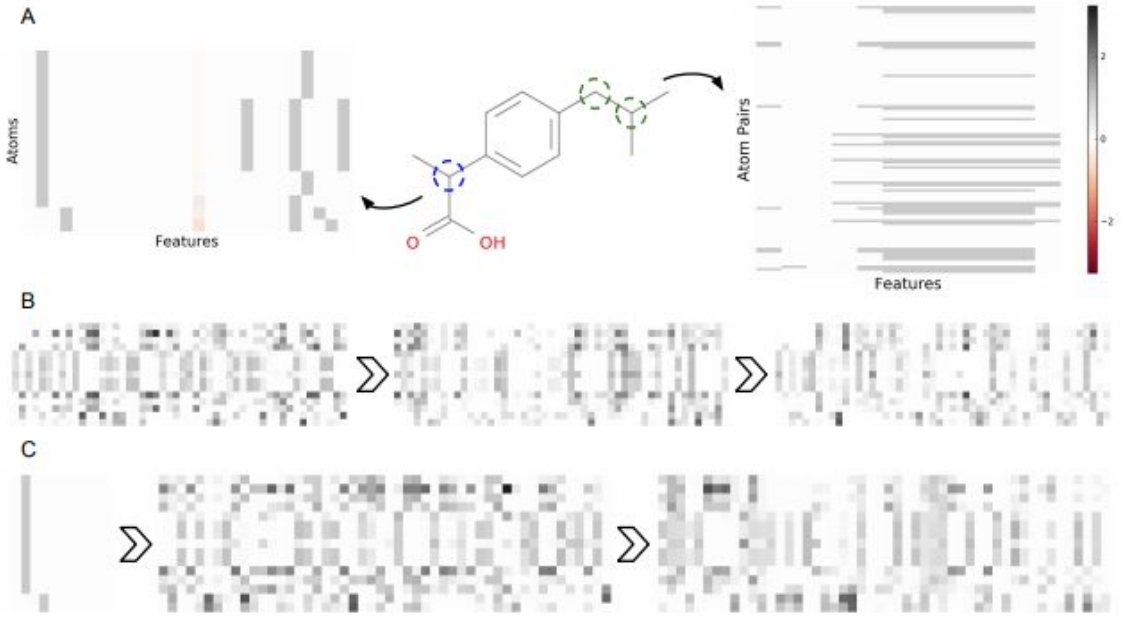




Neural Networks Adjust Weights to Minimize a Loss Function

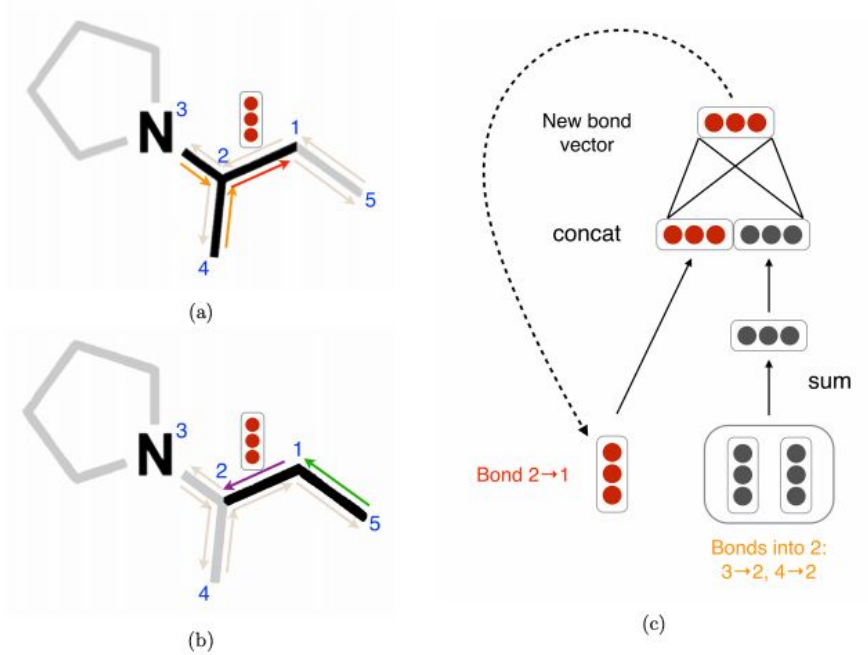


Using Neural Networks to Create New Molecular Representations



Graph Convolutions

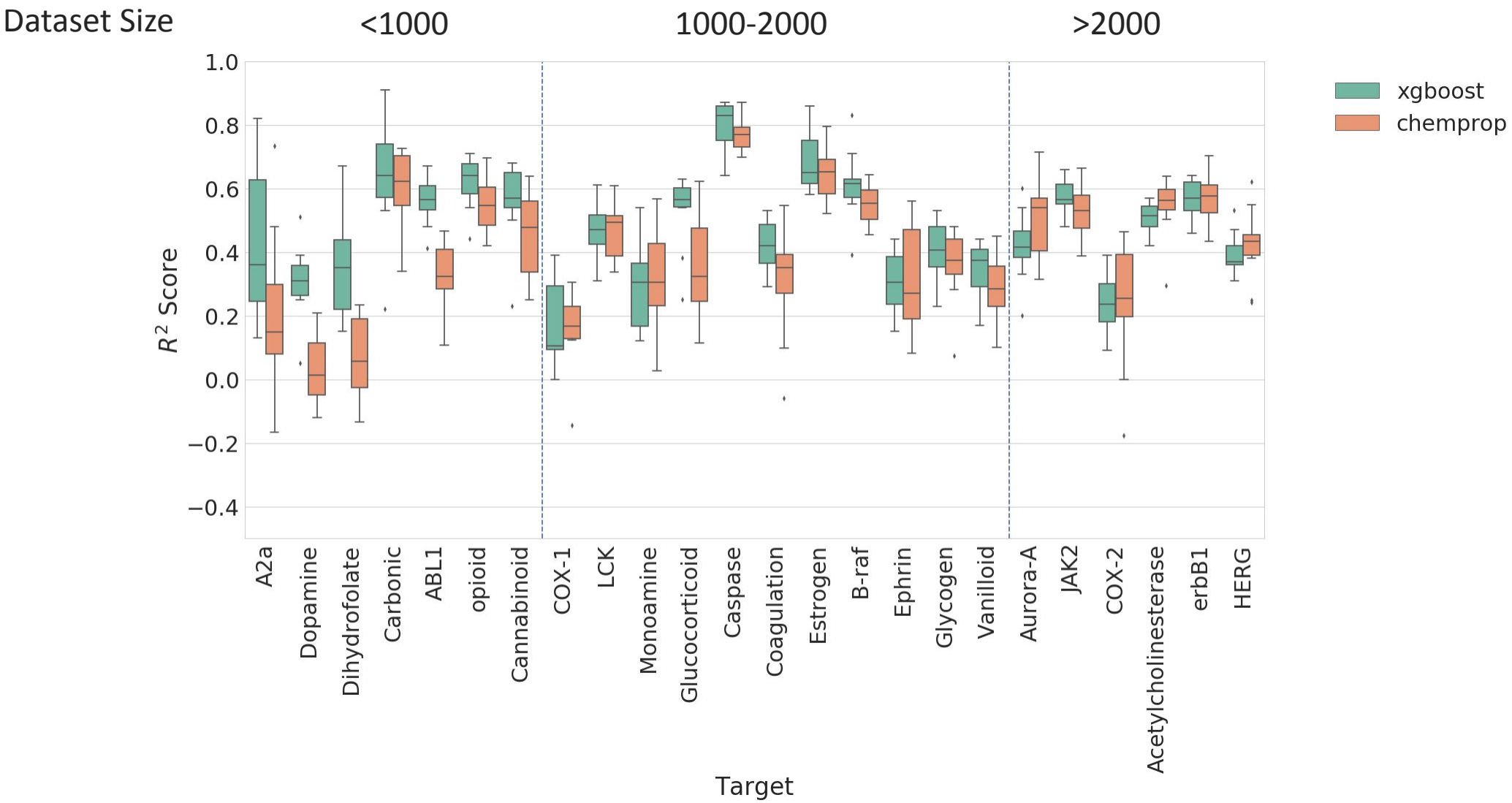
J. Comput. Aided Mol. Des. 2016, 595–608



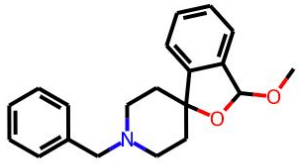
Message Passing Neural Network

J. Chem. Inf. Model. 2019, 59, 3370−3388

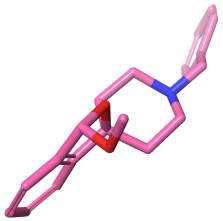
Are Neural Network Representations Better?



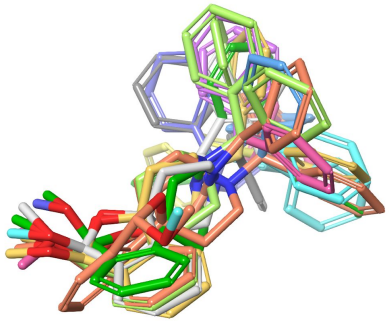
Incorporating 3D into Molecular Machine Learning is an Unsolved Problem



2D single-instance



3D single-instance



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 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...
```

```
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

5.21

5.21

How to Find Me

https://pwalters.github.io