AI in Drug Discovery – An Overview
Session 1

September 21, 2025

Who we are!

Pat Walters
OpenADMET

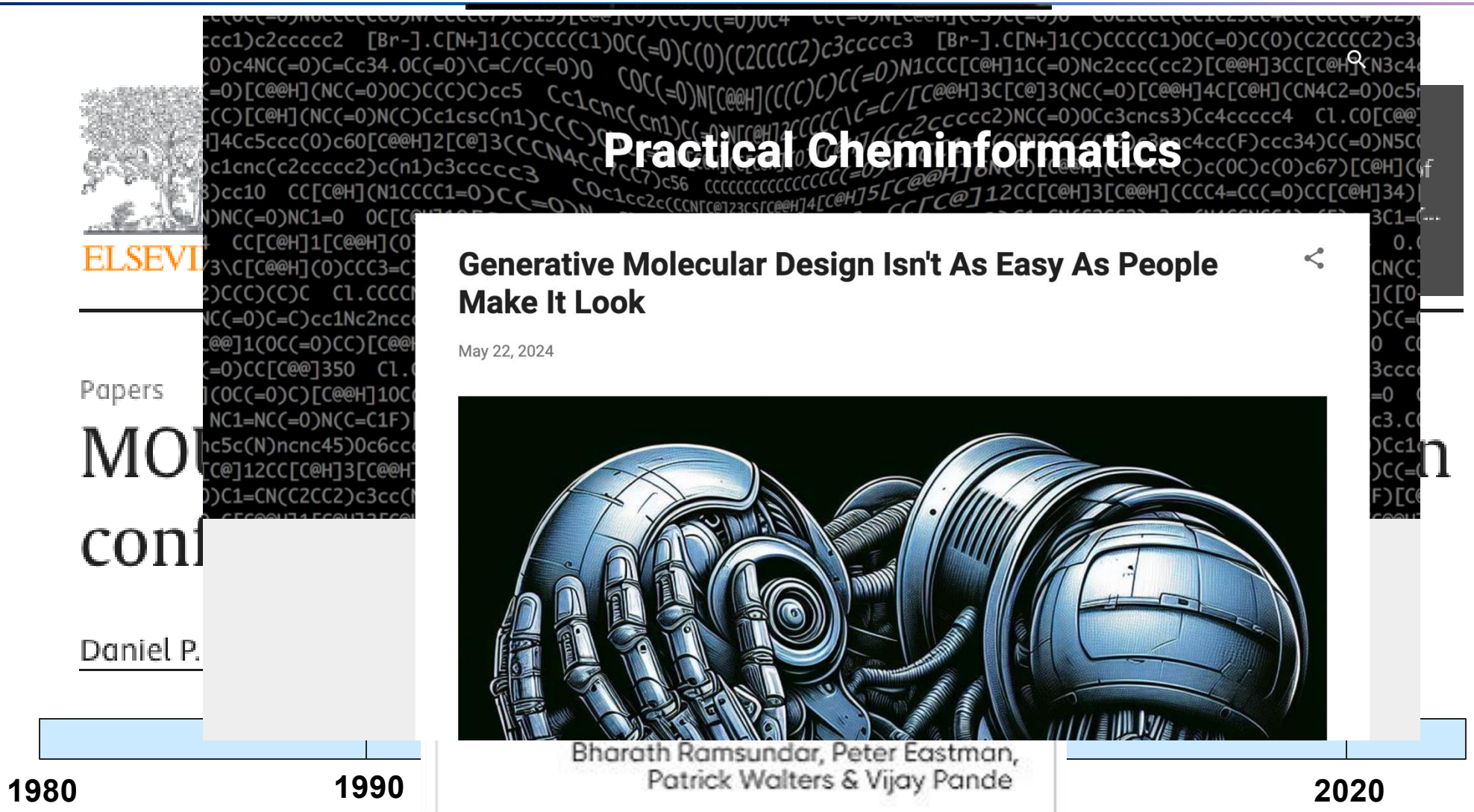
Raquel López-Ríos de Castro
Chodera Lab, MSKCC NYC

Afnan Sultan
Saarland University

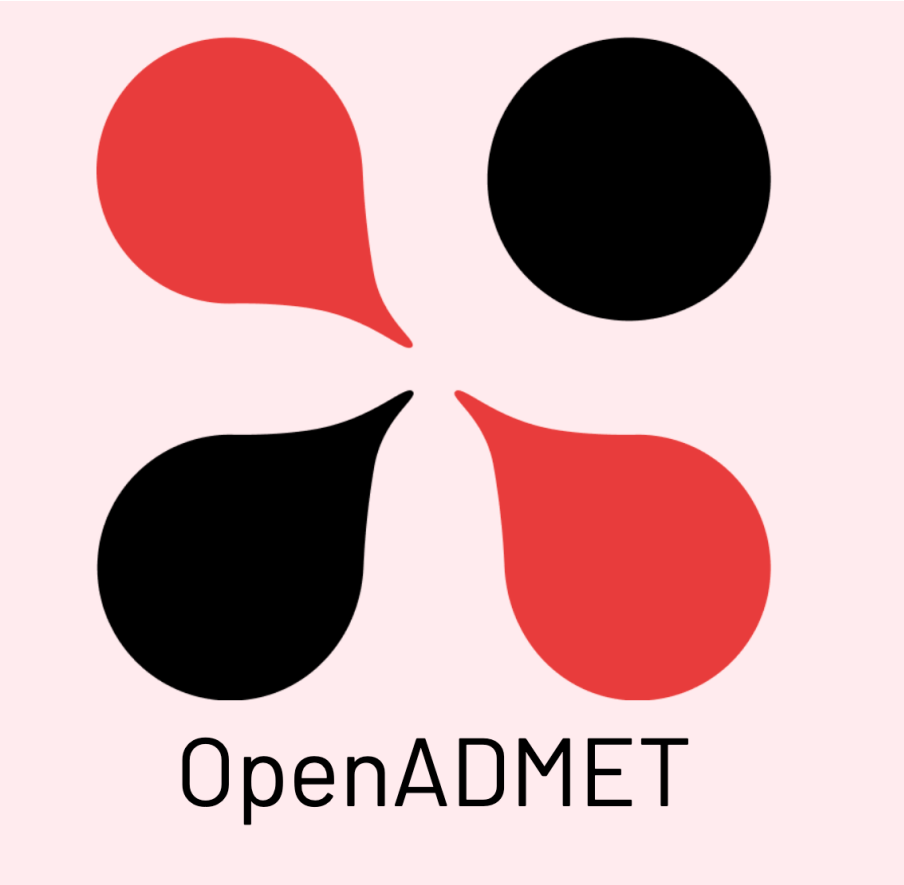
Lisa-Marie Rolli
Saarland University

Andrea Volkamer
Saarland University

My Timeline


Practical Cheminformatics
Generative Molecular Design Isn't As Easy As People Make It Look
May 22, 2024
Bharath Ramsundar, Peter Eastman, Patrick Walters & Vijay Pande
Papers
Daniel P.
1980
1990
2020

My New Adventure – Starting Week #2



High-throughput experimentation
- **Octant Bio – San Francisco**

Structural Biology
- **UCSF**

Machine Learning
- **Distributed Global Team**

Data Generation and Dissemination

Blind Challenges to Test Methods

Establishing Best Practices

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

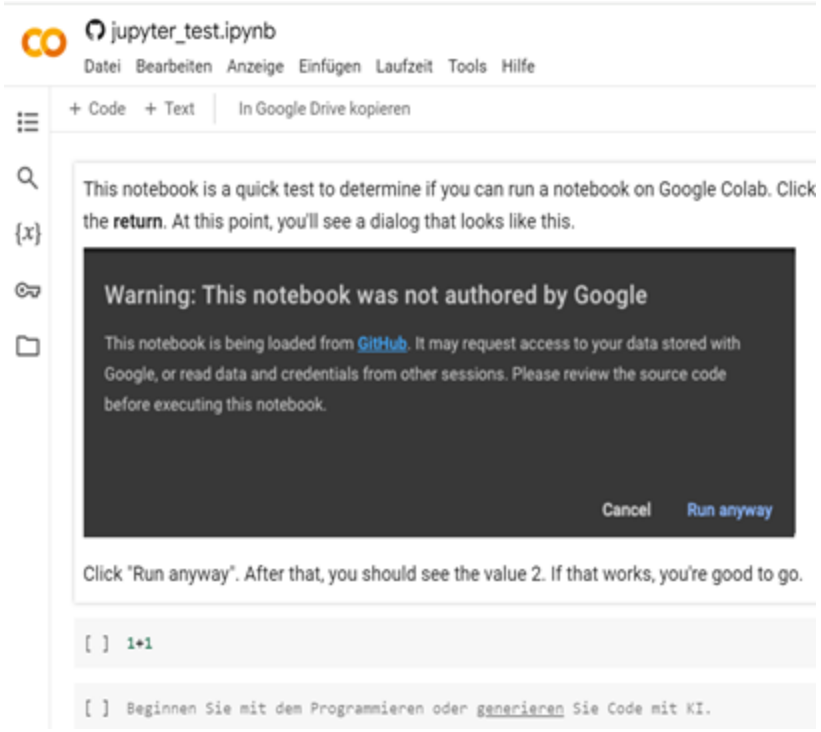
Session 2 - 3:00 - 4:00 pm

- 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 …



Tutorials and Slides Can Be Found On GitHub

https://github.com/volkamerlab/ai_in_chemistry_workshop_2025

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”

What Is Machine Learning?

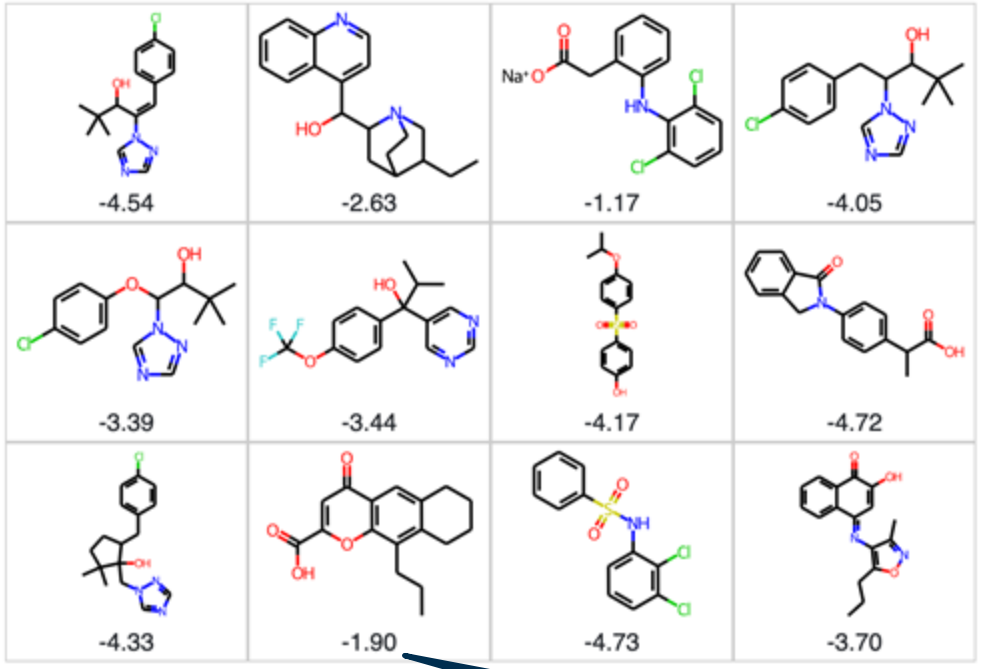
Machine learning is all about labeling things using examples

Cassie Kozyrkov, Google

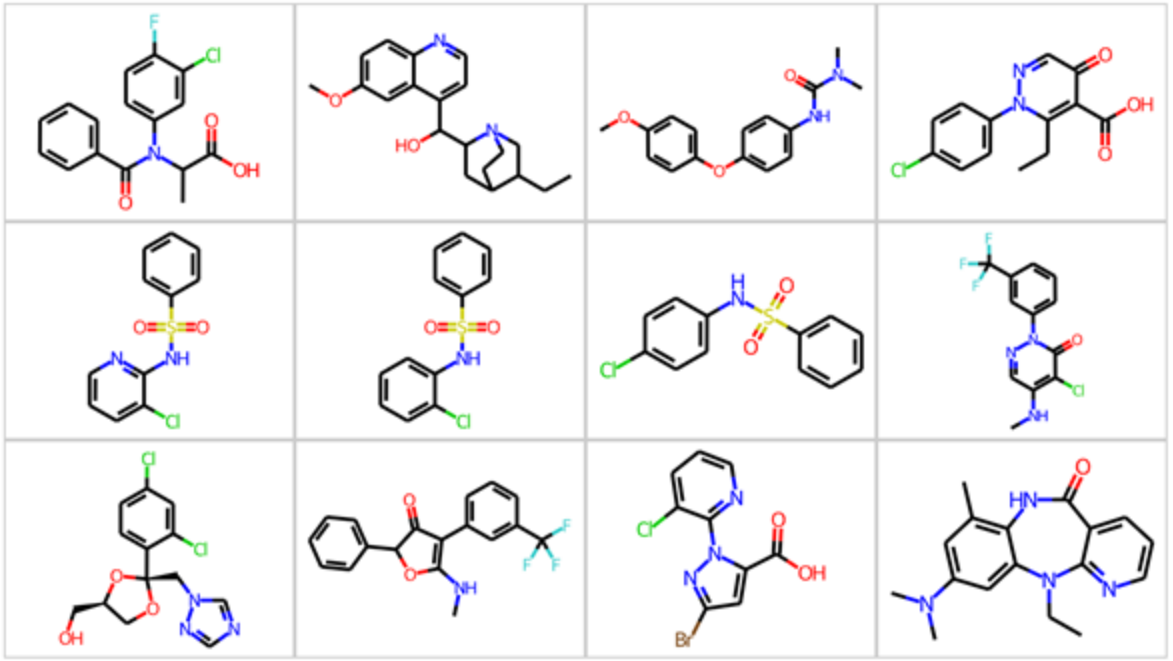
https://hackernoon.com/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c

Labeling Molecules Based on Examples

Molecules with measured data

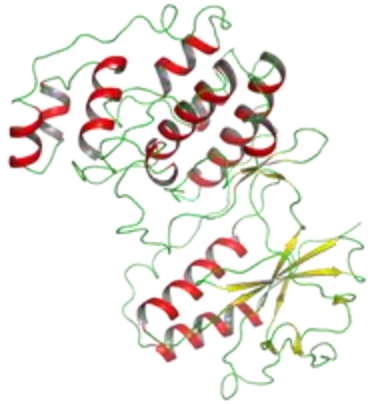


Molecules to be predicted

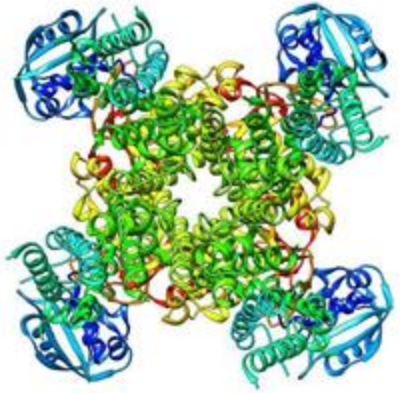


Log10(Molar Aqueous Solubility)

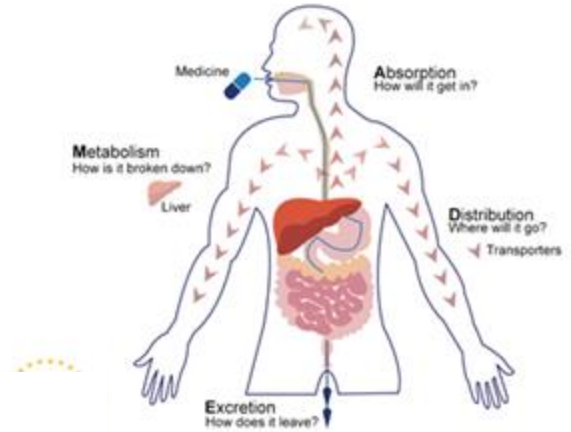
Using Predictive Models to Drive Drug Discovery



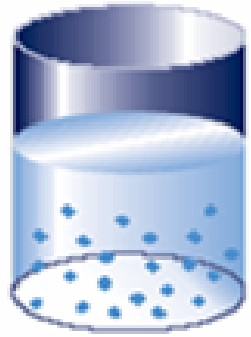
On-target Activity



Off-target Activity



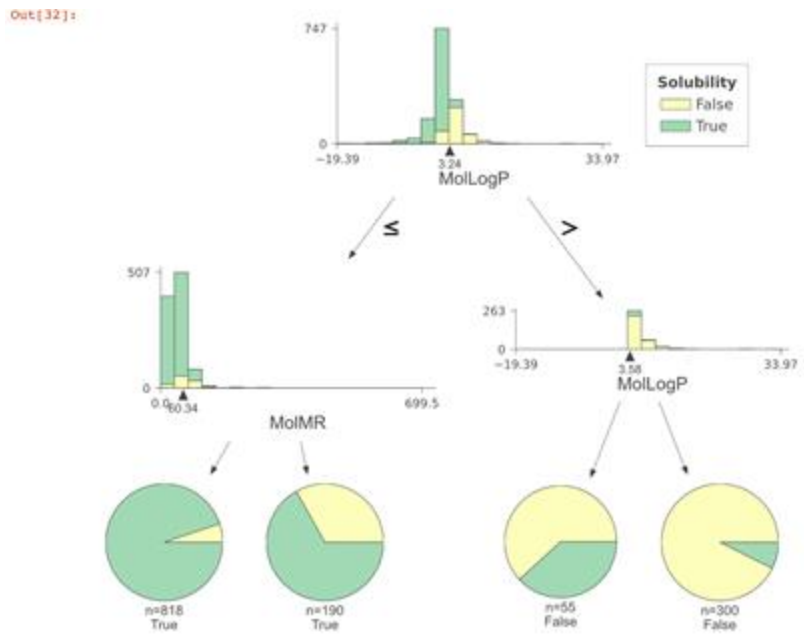
Pharmacokinetics



Physical Properties

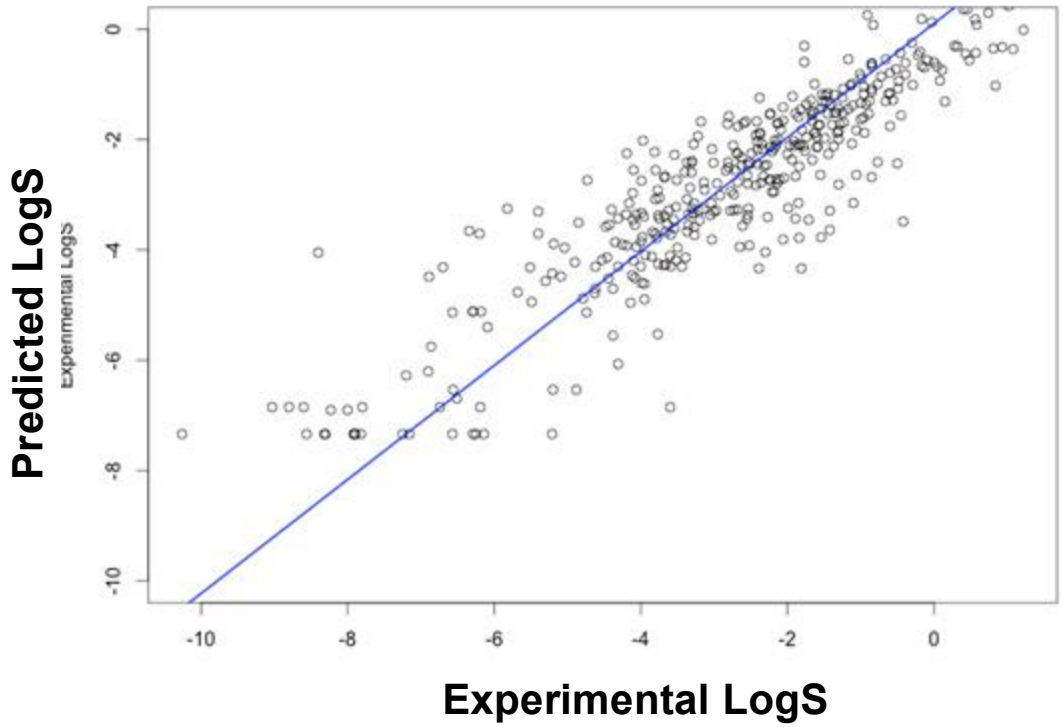
Two Types of ML Models – Classification and Regression

Classification



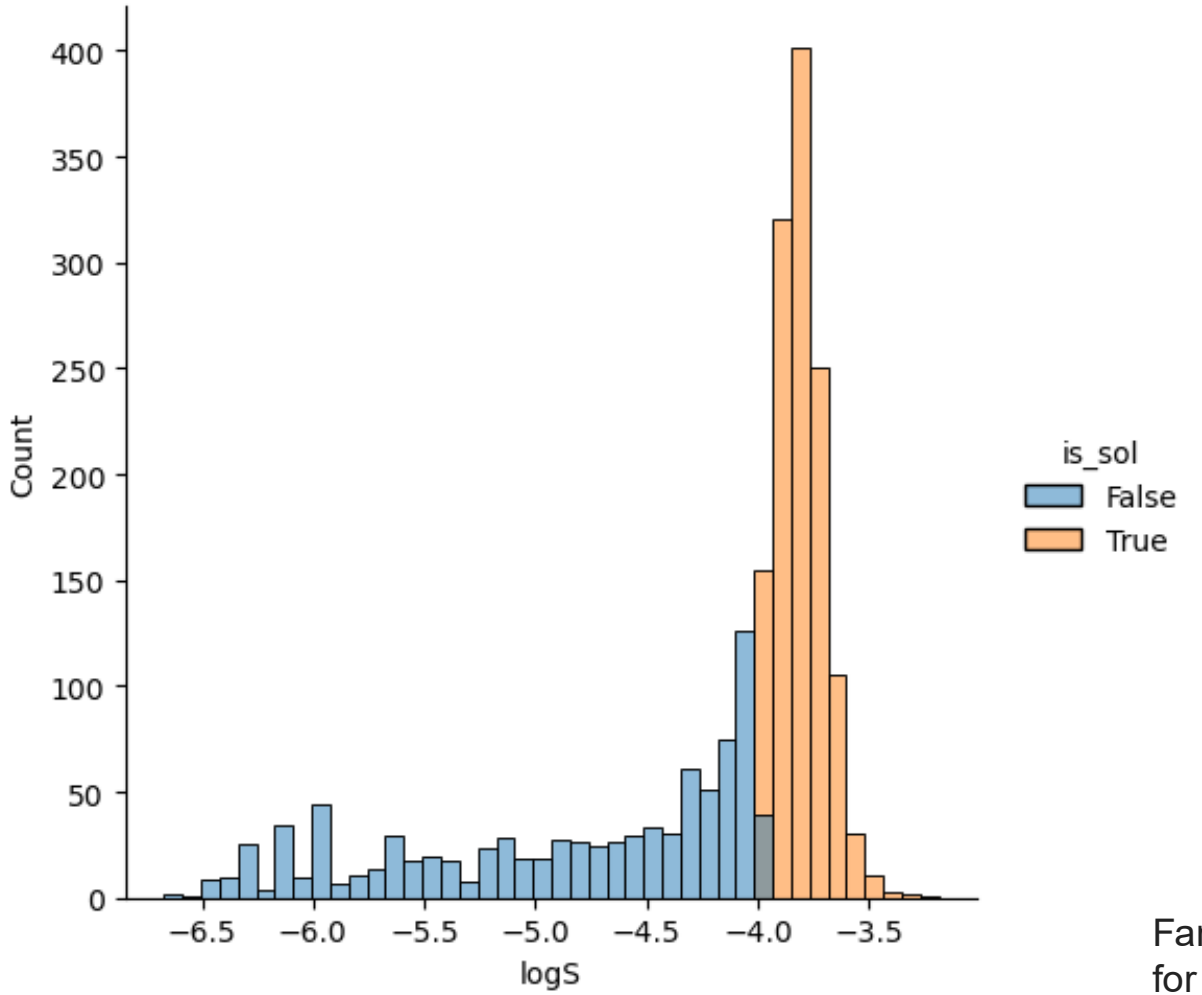
Predict a category, e.g. soluble/insoluble

Regression



Predict a value, e.g. 4.2

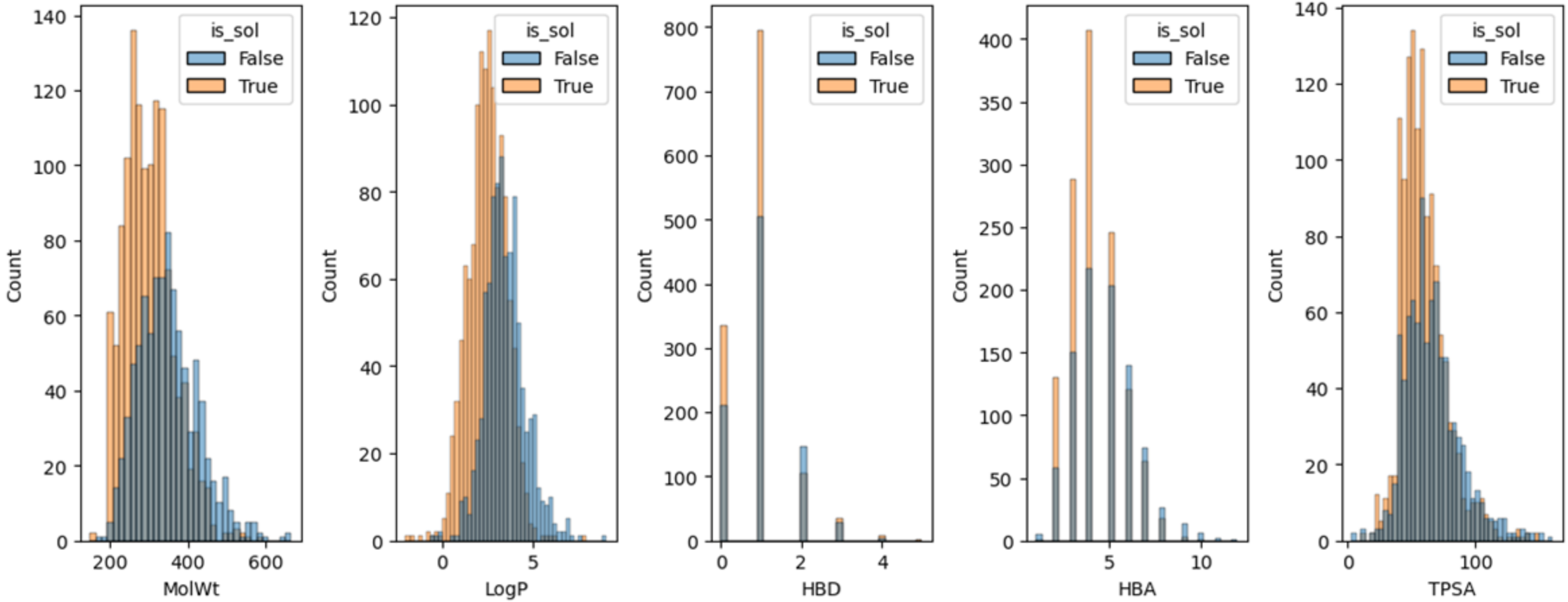
Let’s Start With a Dataset



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

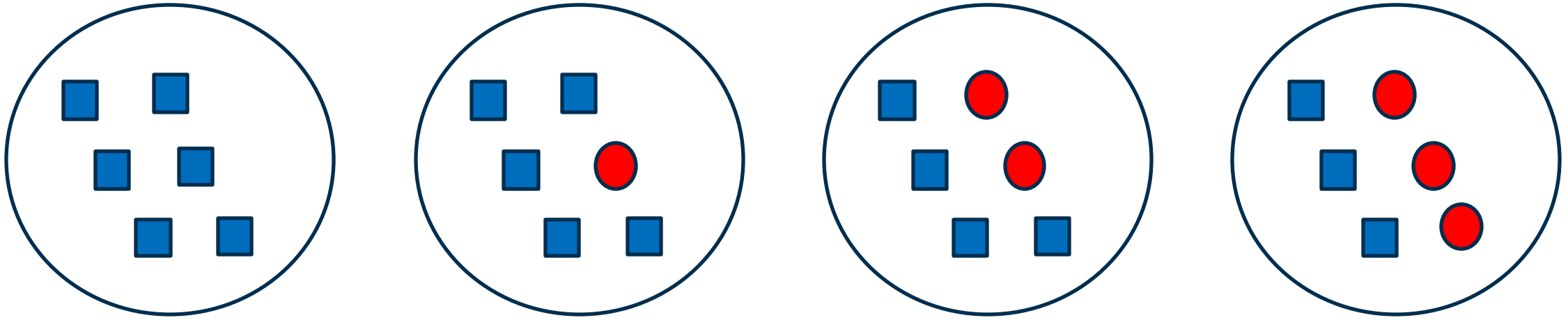
Fang, Cheng, et al. "Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective." *Journal of Chemical Information and Modeling* 63.11 (2023): 3263-3274.

Which Property Best Separates Soluble vs Insoluble Molecules?



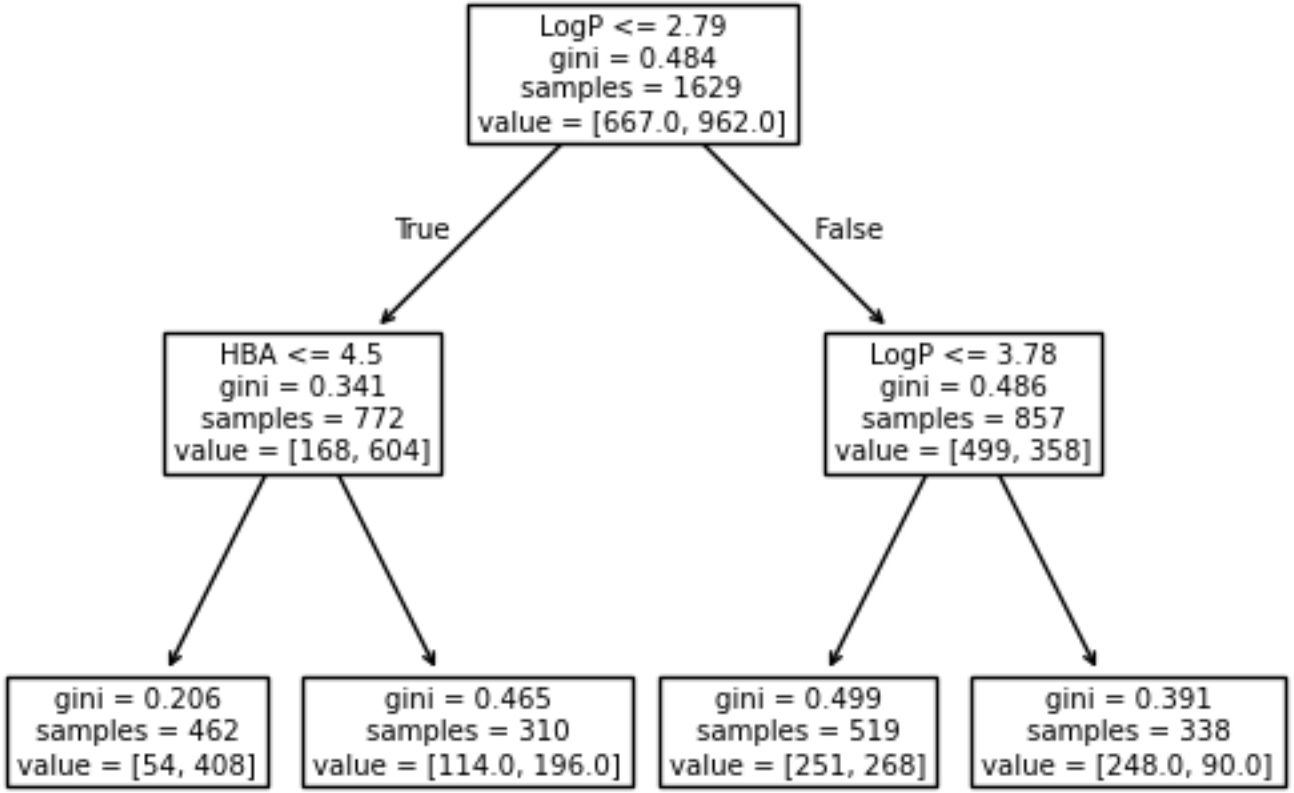
The Gini Index Quantifies the “Purity” of a Split

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

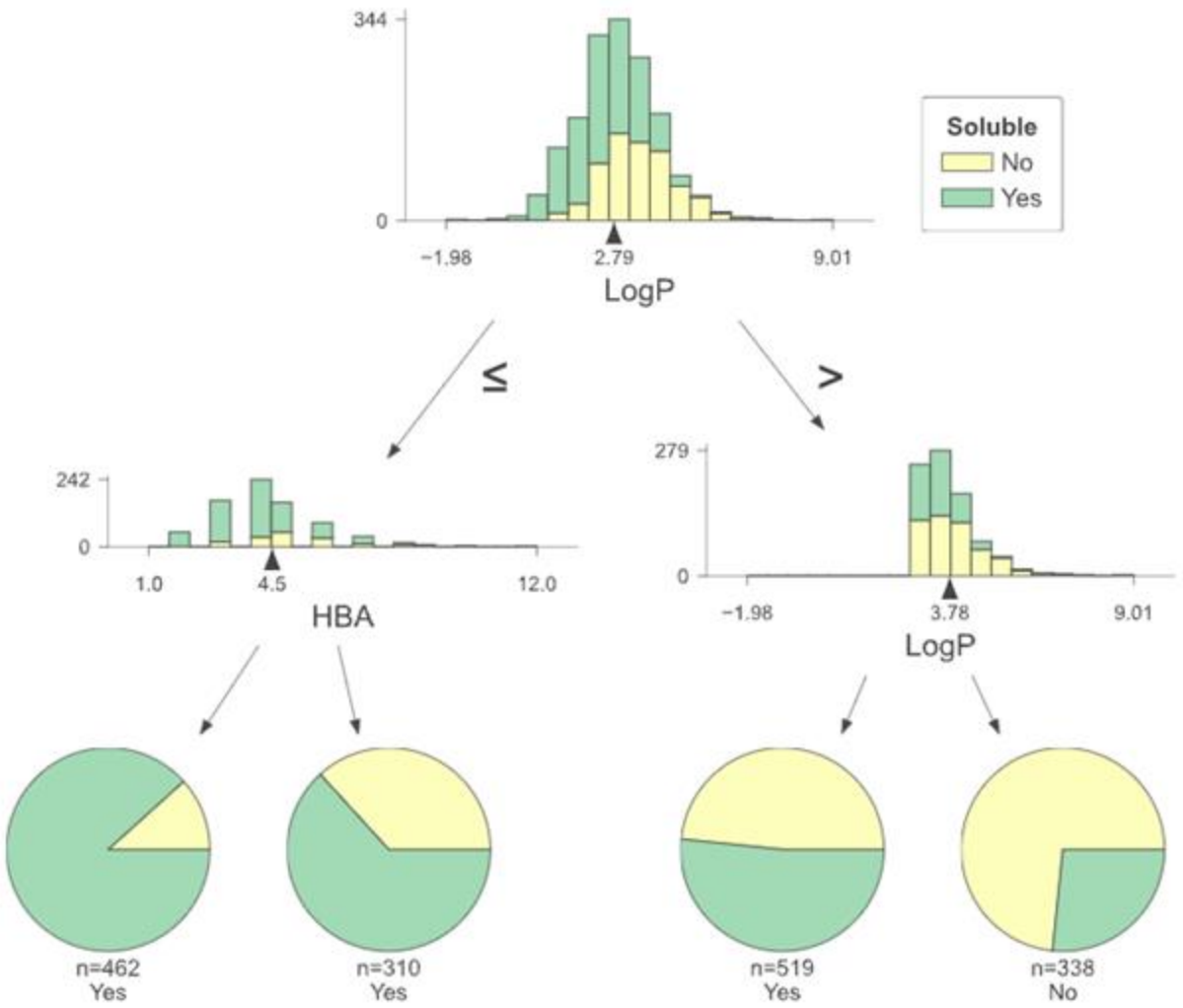


1.00 0.83 0.66 0.50

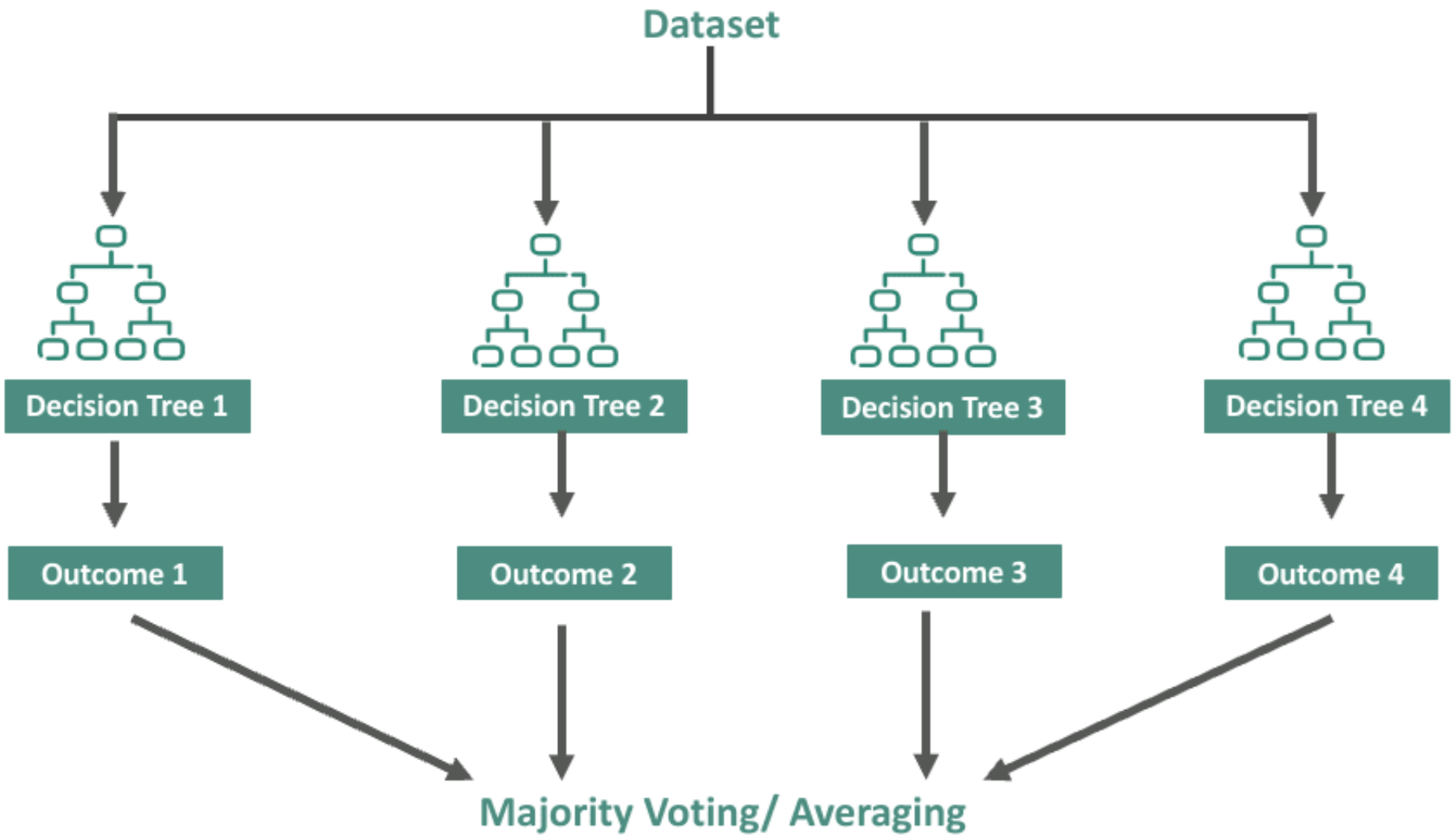
Build a Decision Tree



A Better (IMO) Decision Tree Visualization

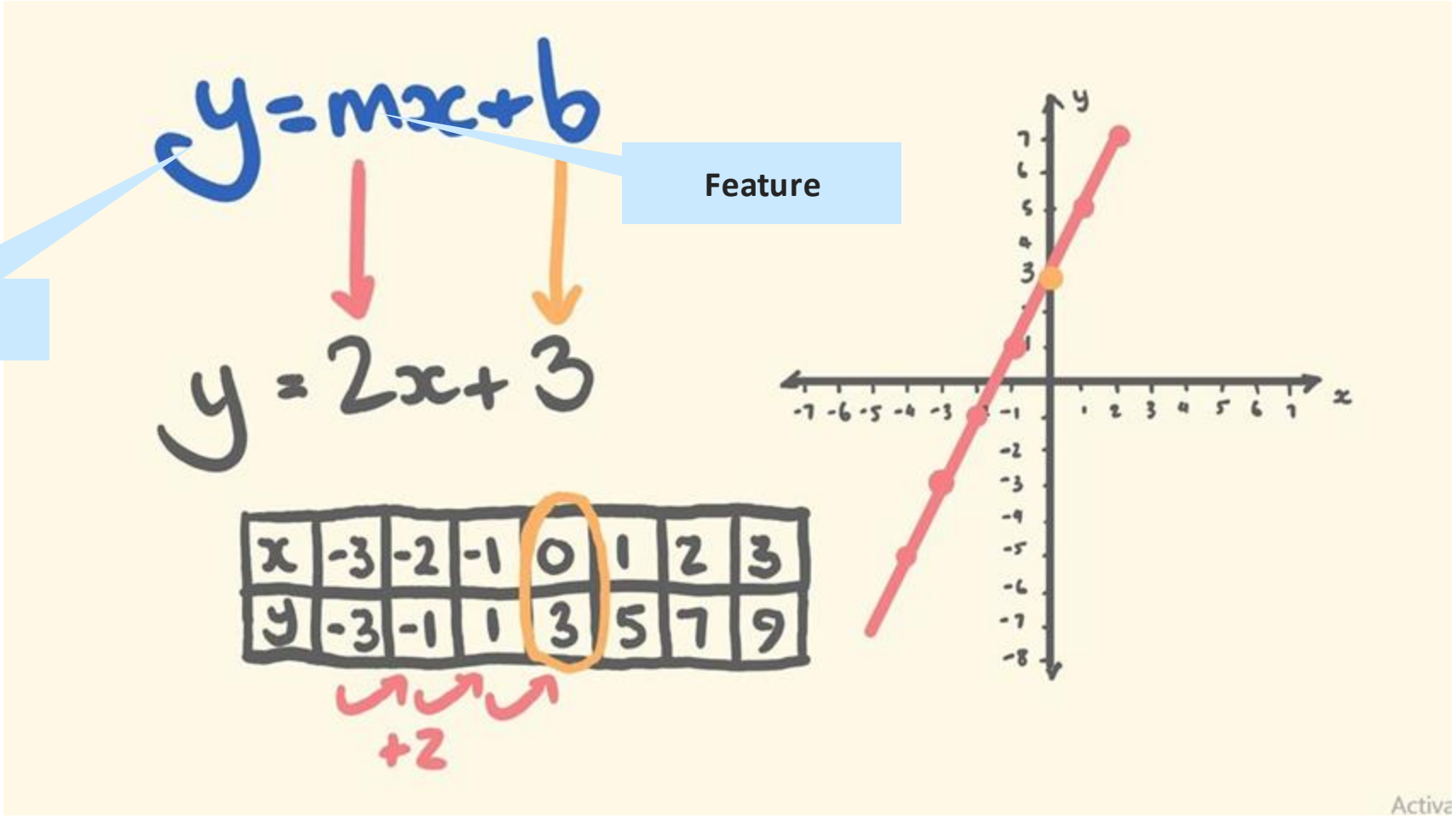


Random Forest Uses an Ensemble of Decision Trees

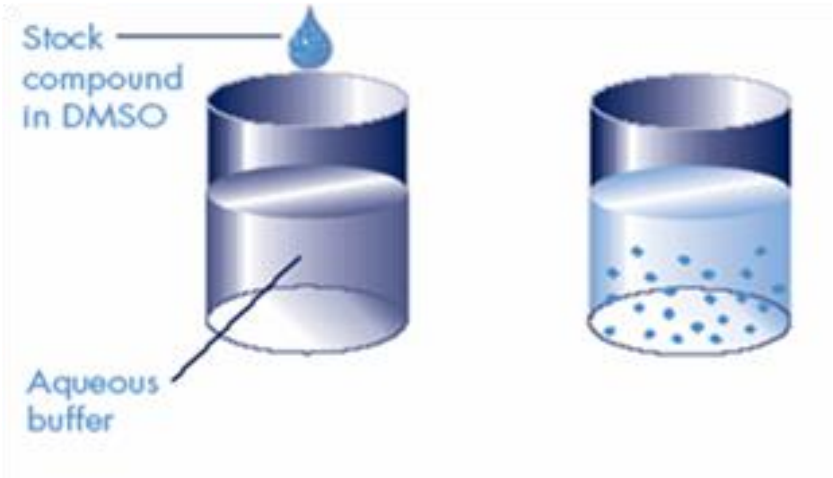


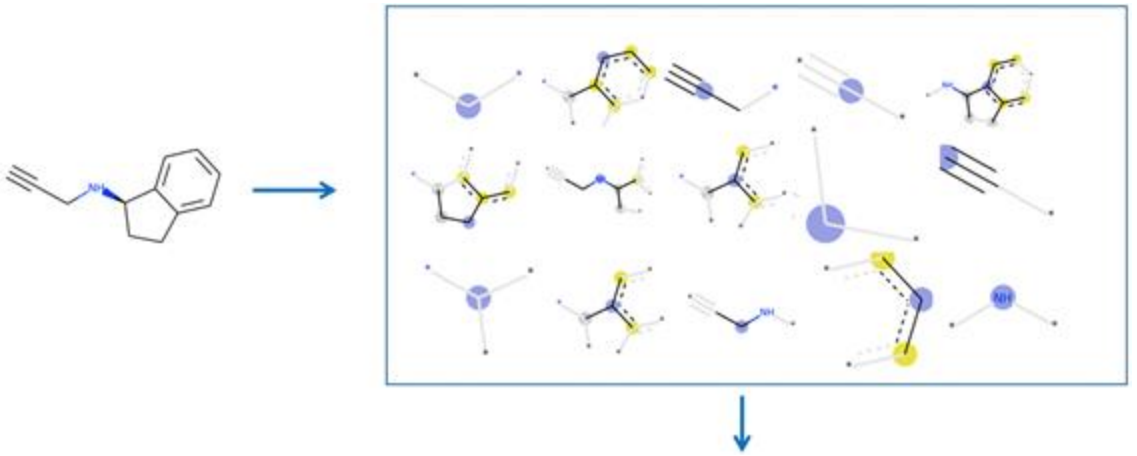
There Are Many Tree Ensemble Methods

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



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





0101100100010001110010001100101010101111

y = Log Aqueous Solubility

X = A Vector Representing a Molecule

Label

Features

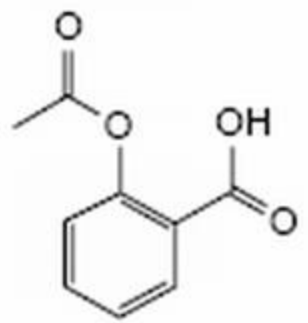
Define Features Based on Books People Have Read

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

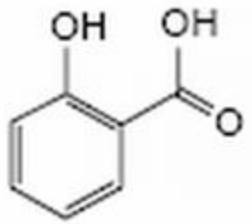
Create vectors representing books purchased by individuals
1 = bought book
0 = did not buy book

22

Chemical Fingerprints as Molecular Descriptors

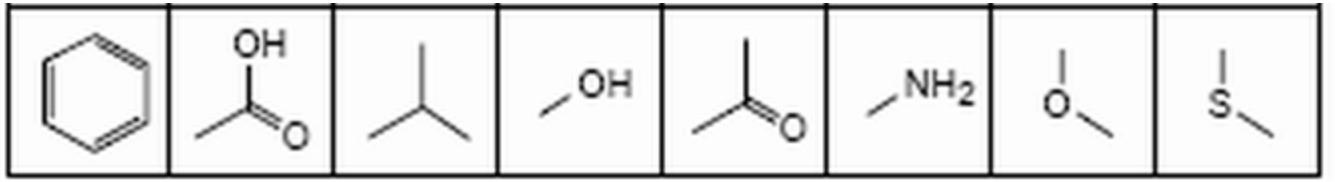


1

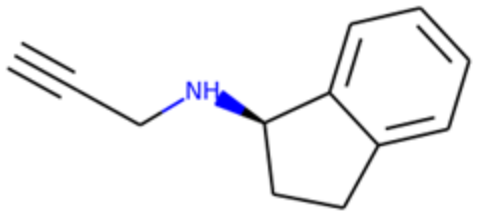


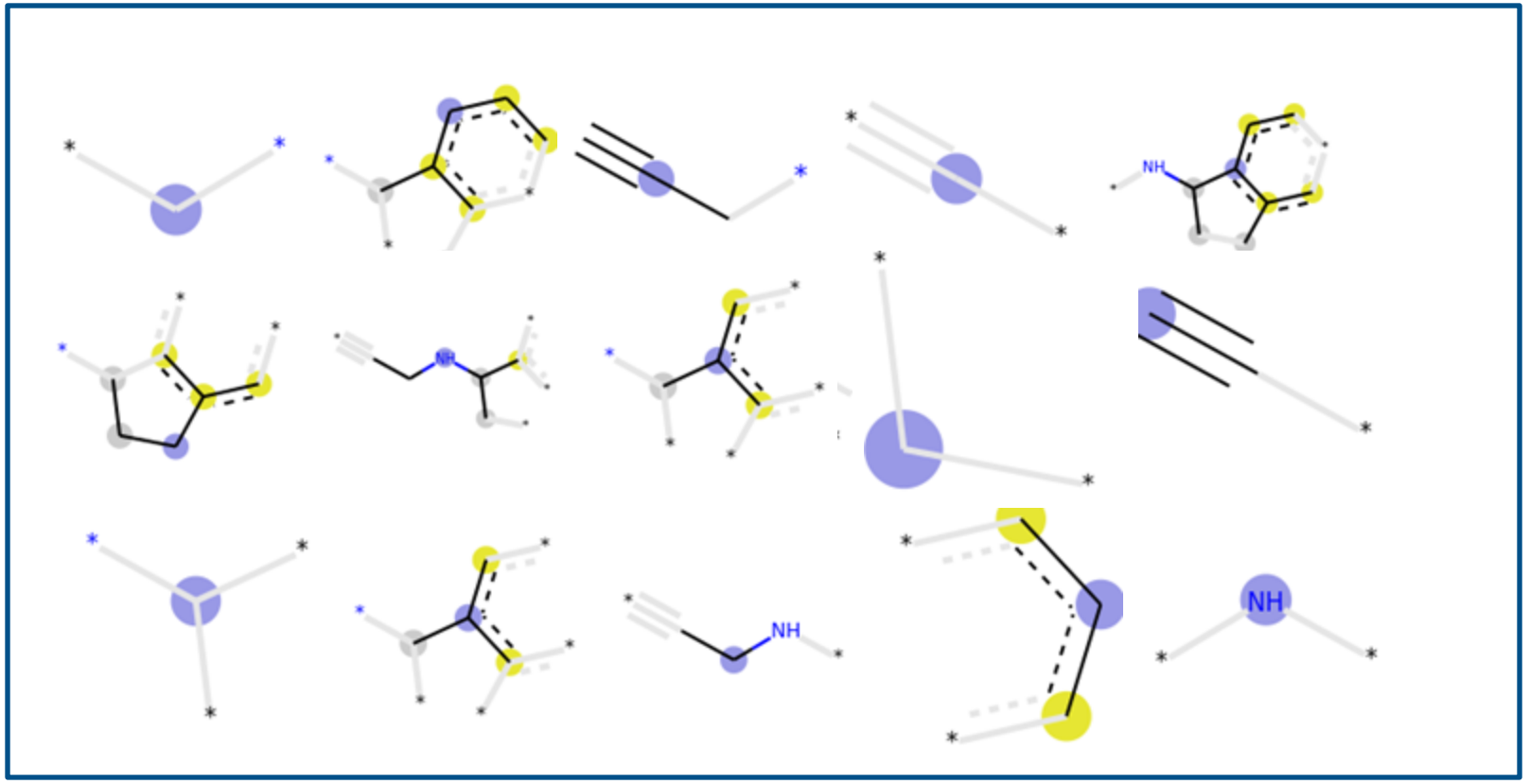
2

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

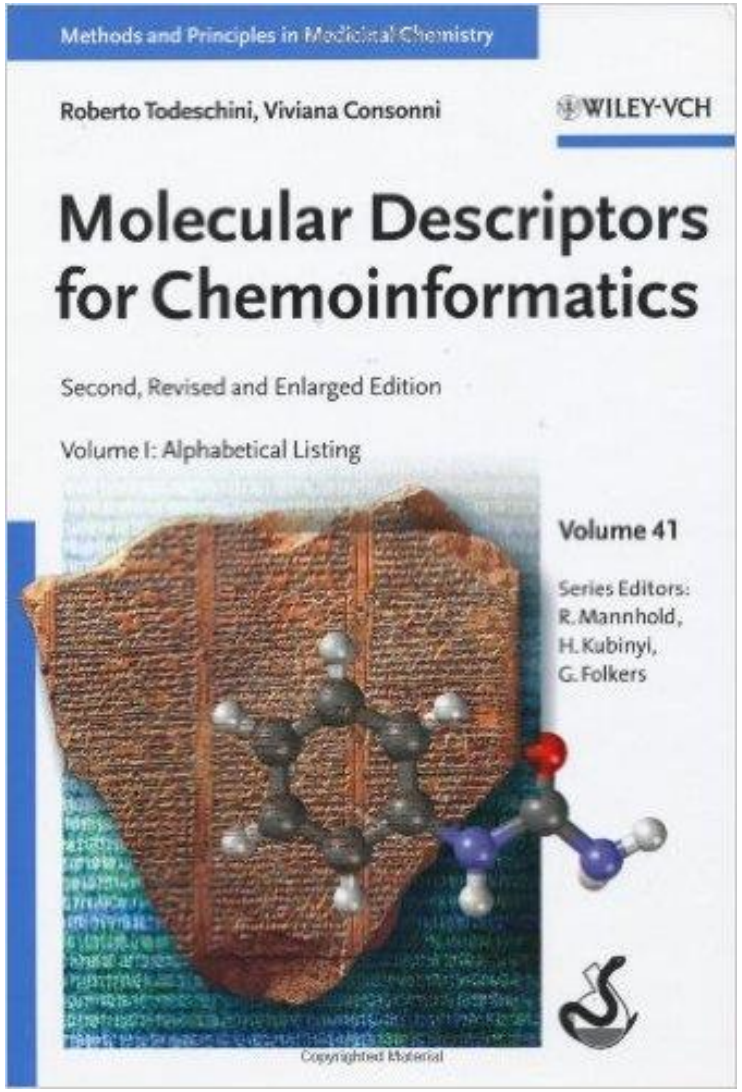


Chemical Fingerprints as Vector Representations of Molecules





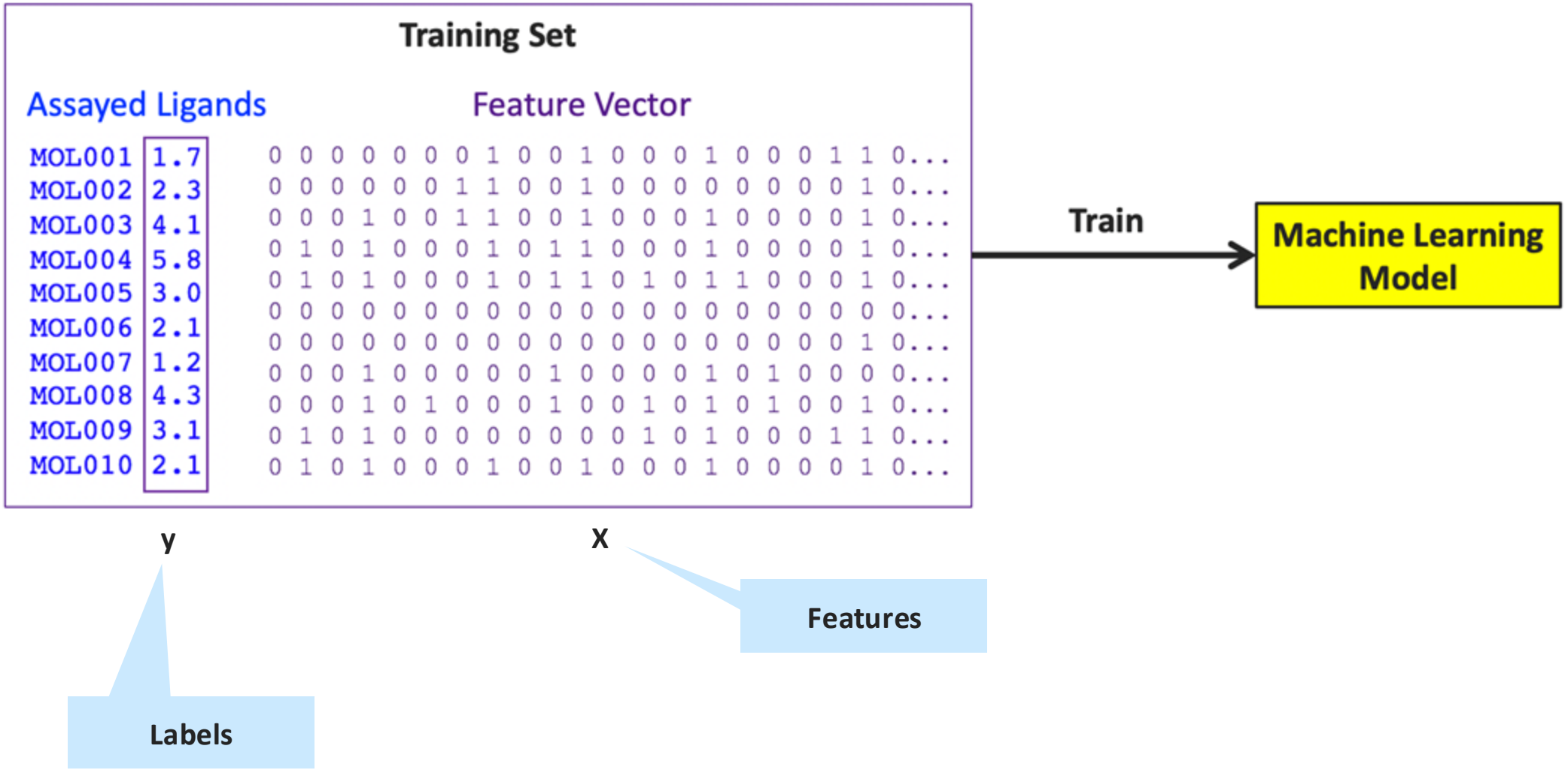
There are A LOT of ways to do this



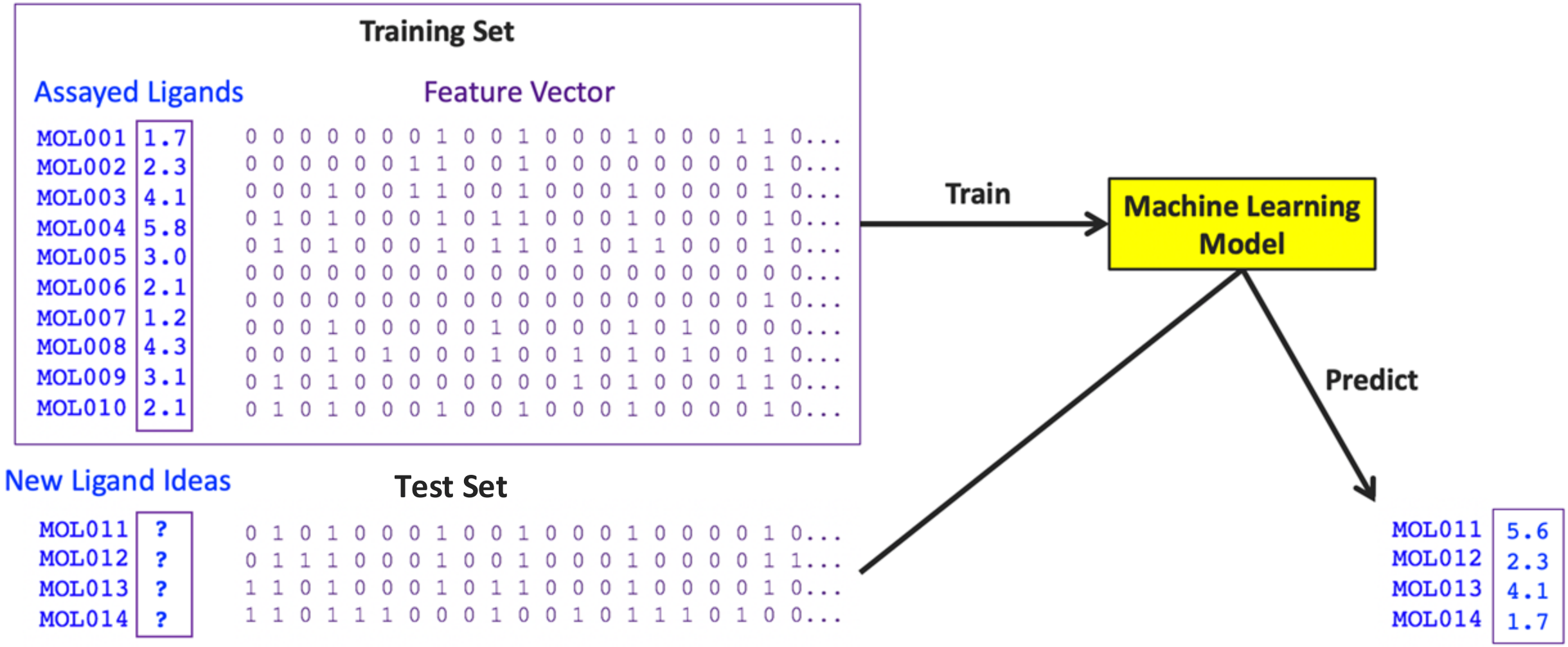
2 Volumes
6,000 references from 450 journals

DRAGON 7 has 5,270 descriptors z

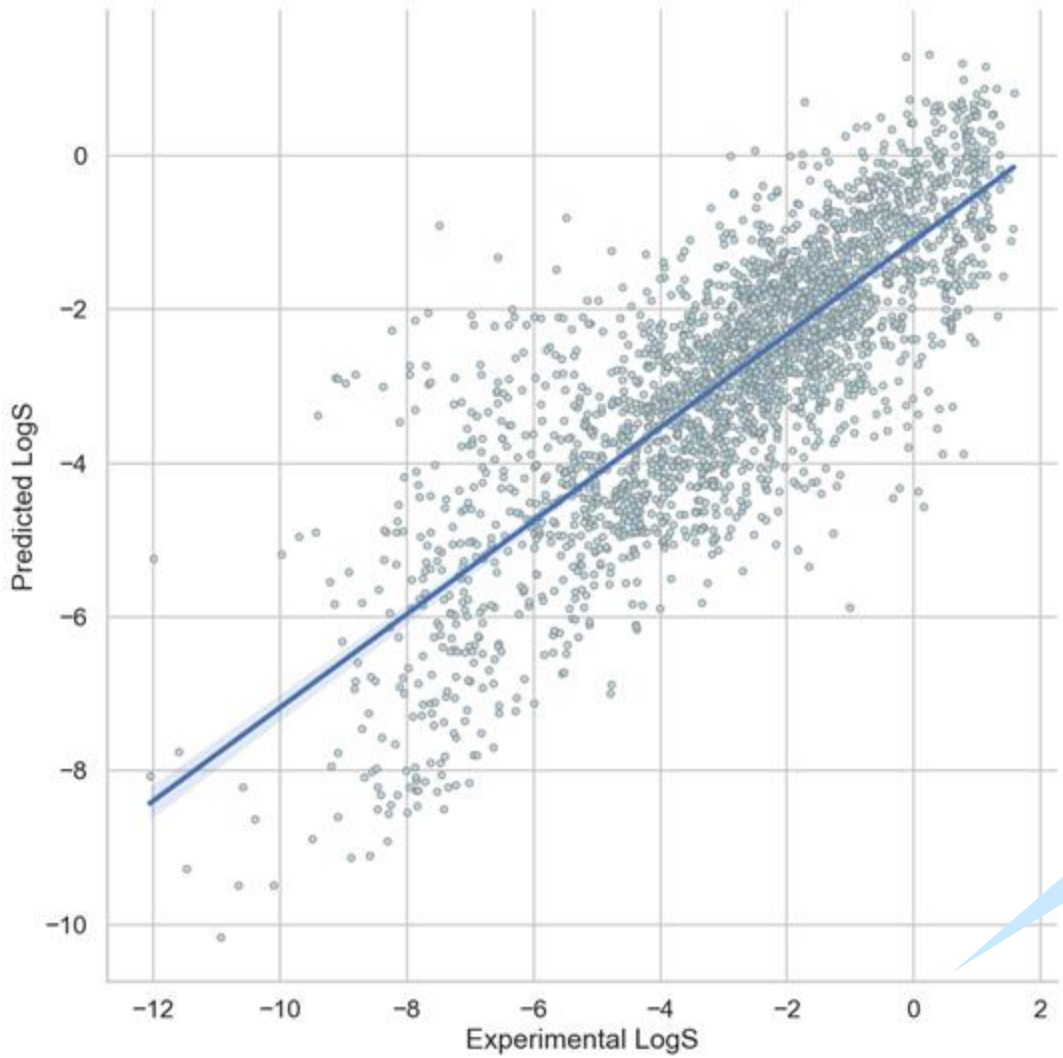
Training a Machine Learning Model



Making Predictions With a Machine Learning Model

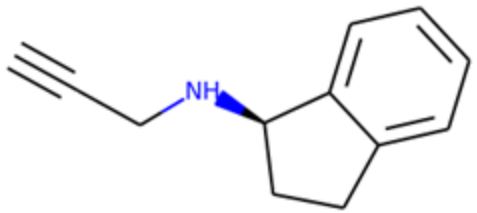


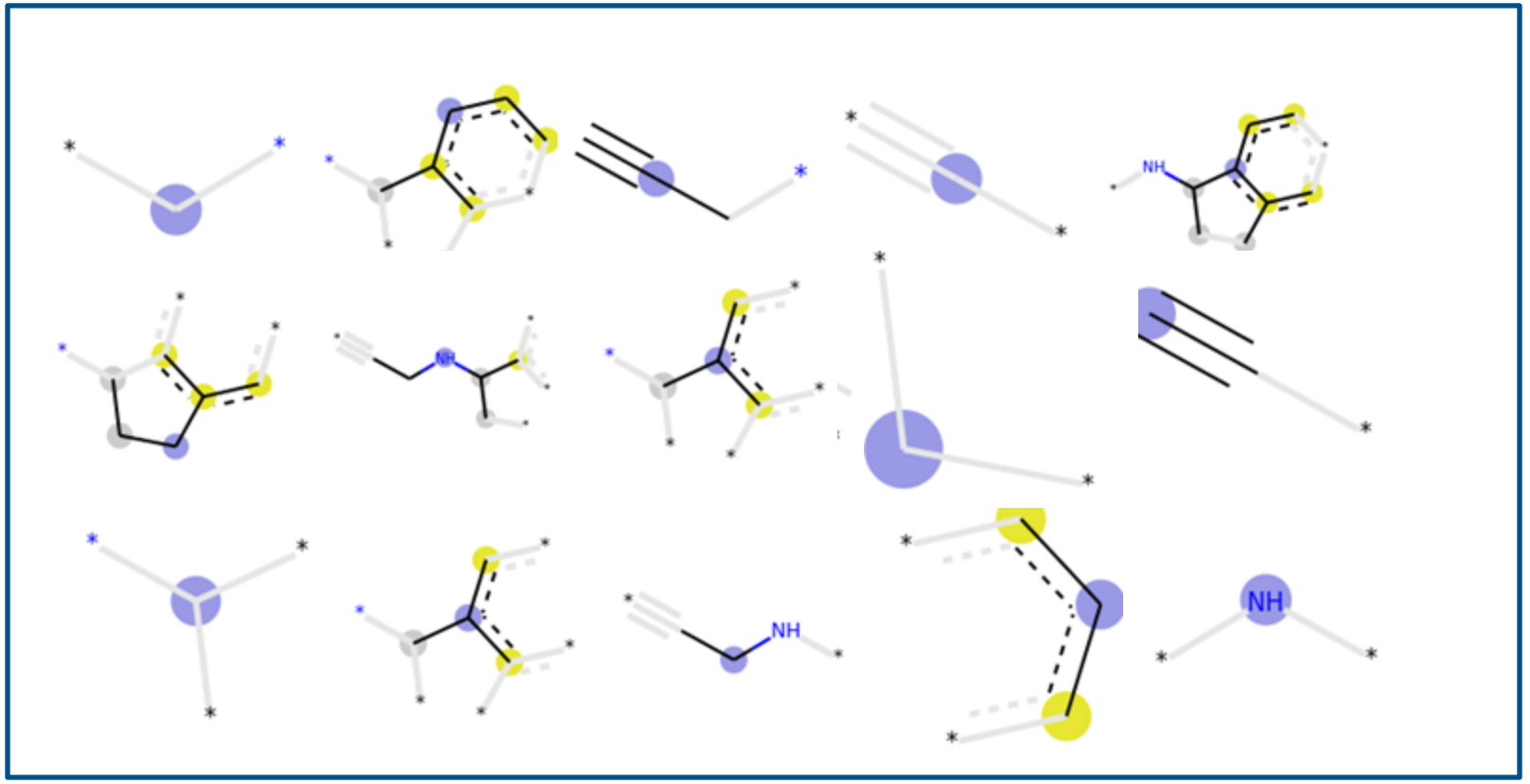
Prediction Performance of an Aqueous Solubility Model



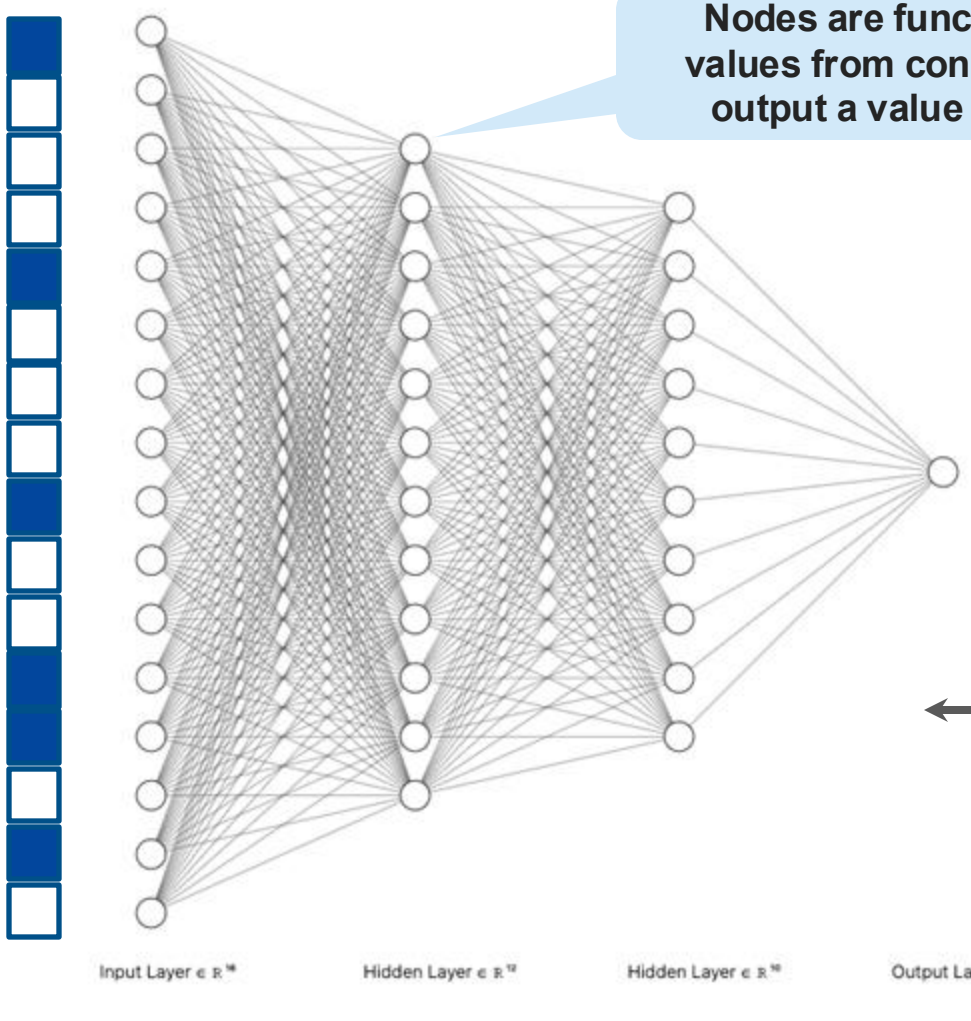
Dynamic range is very large, does not represent a typical use case

Chemical Fingerprints as Vector Representations of Molecules





Neural Networks Adjust Weights to Minimize a Loss Function



Nodes are functions that accept values from connected nodes and output a value between 0 and 1

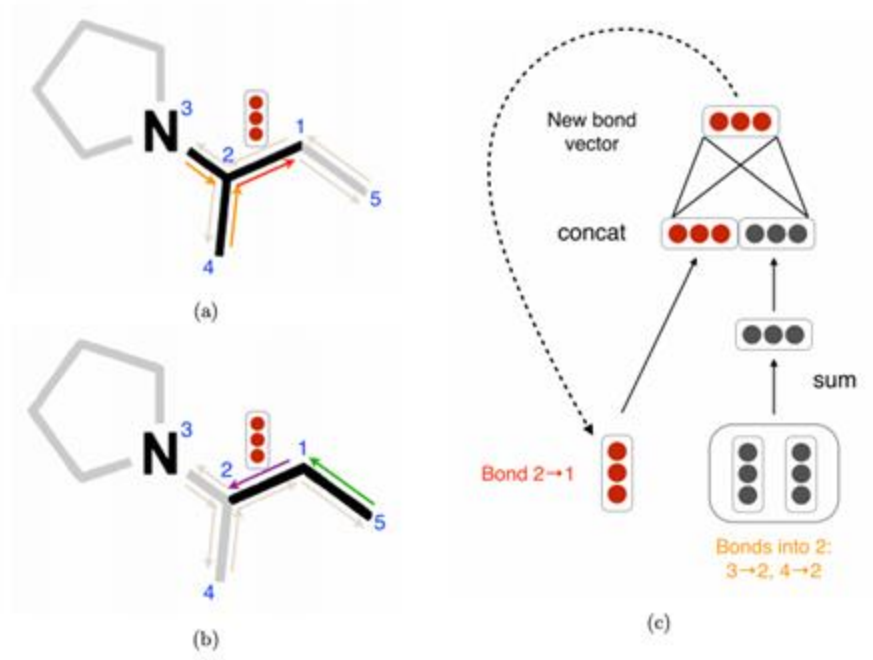
Predicted value = 3.2
True value = 2.8
Error = 0.4

Backpropagate to minimize loss (e.g. RMSE)

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

How to Find Me

https://pwalters.github.io

Practical Cheminformatics Publications Tutorials Blog Videos Resources

Pat Walters

Cheminformatics, ML

Cambridge, MA

Email

Google Scholar

Github

Bluesky

LinkedIn

X (formerly Twitter)

Pat Walters is Chief Scientist at OpenADMET, an open science initiative that combines insights from high-throughput experimentation, structural biology, and machine learning to improve the prediction of drug absorption, metabolism, excretion, and toxicology. Before his current role, Pat spent thirty years in leadership positions at Relay Therapeutics and Vertex Pharmaceuticals. Pat is the 2023 recipient of the Herman Skolnik Award for Chemical Information Science from the American Chemical Society. He is a member of the editorial advisory boards for the Journal of Chemical Information and Modeling and Artificial Intelligence in the Life Sciences, and previously held a similar role with the Journal of Medicinal Chemistry. Pat is co-author of the book "Deep Learning for the Life Sciences", published in 2019 by O'Reilly and Associates. He received his Ph.D. in Organic Chemistry from the University of Arizona where he studied the application of artificial intelligence in conformational analysis. Prior to obtaining his Ph.D., Pat worked at Varian Instruments as both a chemist and a software developer. He received his B.S. in Chemistry from the University of California, Santa Barbara.