

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 ...

The screenshot shows a Google Colab interface. At the top, there's a toolbar with icons for file operations, code, text, and Google Drive. The main area displays a notebook cell with the following content:

```
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 overlaid on the page, containing a warning message:

Warning: This notebook was not authored by Google

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 modal, there are "Cancel" and "Run anyway" buttons. Below the modal, a code cell shows the result of the execution:

```
[ ] 1+1
```

Another code cell below it contains German text:

```
[ ] Beginnen Sie mit dem Programmieren oder generieren Sie Code mit KI.
```

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



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?

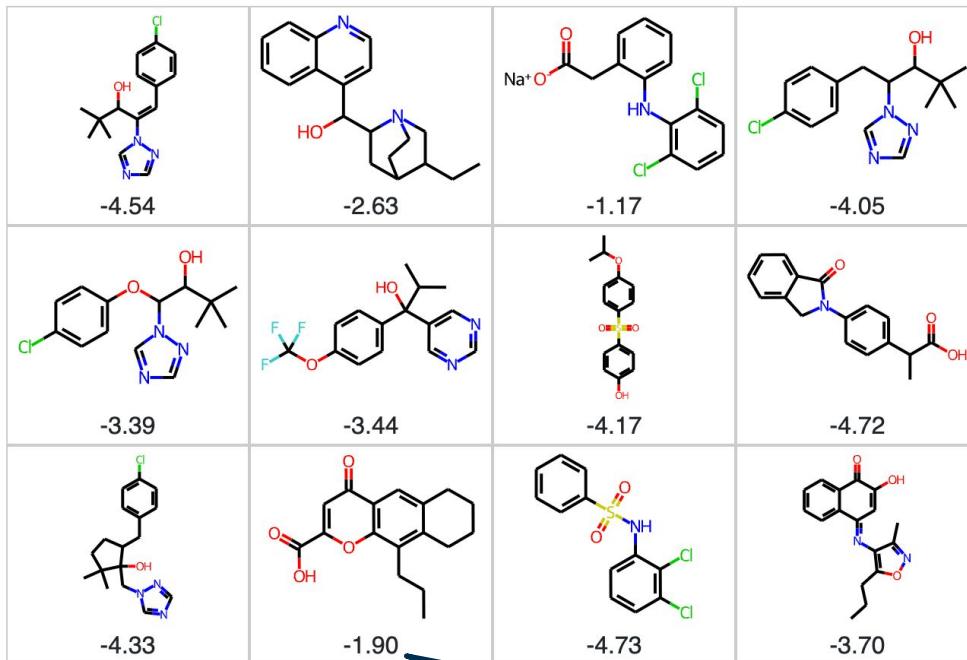
Machine learning is all about labeling things using examples



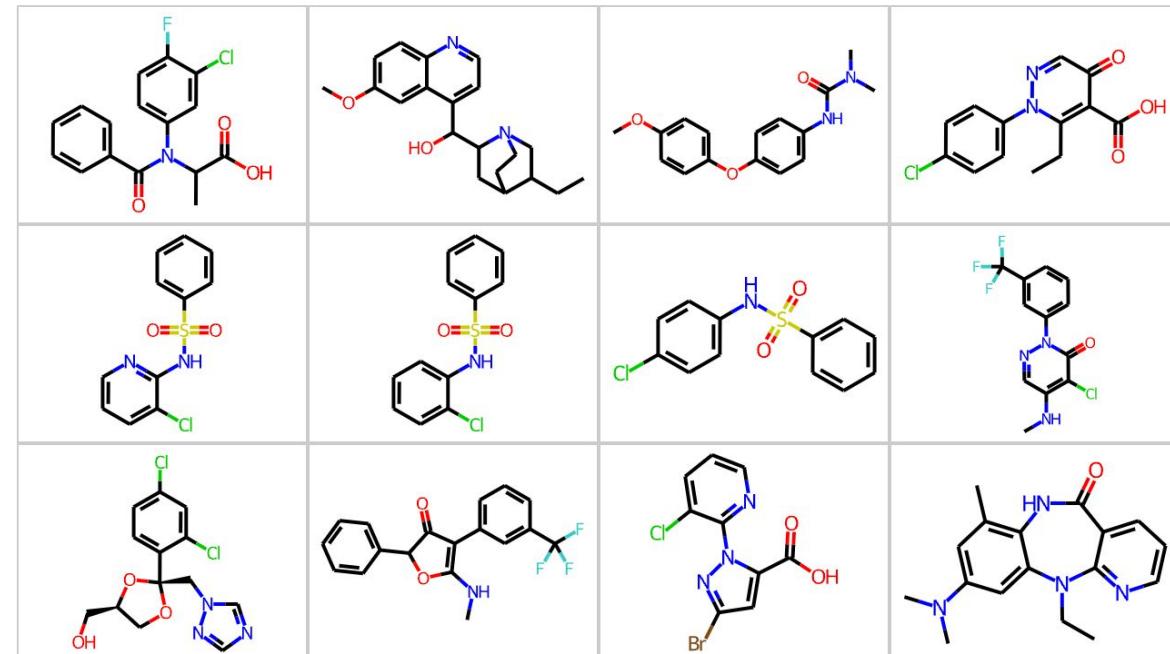
Cassie Kozyrkov, Google

Labeling Molecules Based on Examples

Molecules with measured data

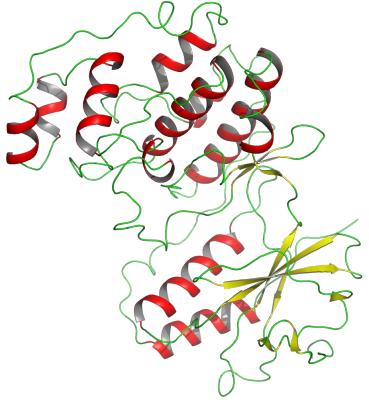


Molecules to be predicted

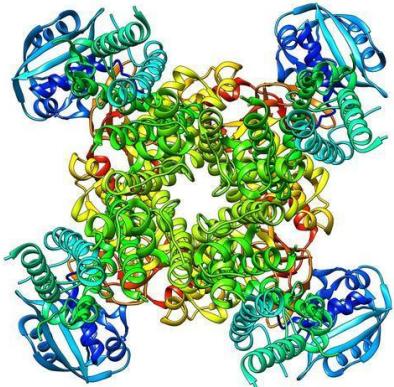


Log10(Molar Aqueous Solubility)

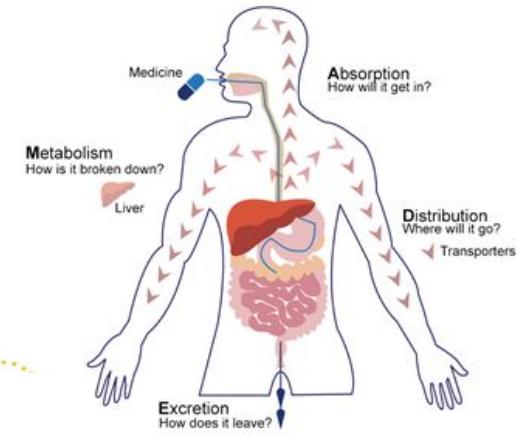
Using Predictive Models to Drive Drug Discovery



On-target Activity



Off-target Activity



Pharmacokinetics

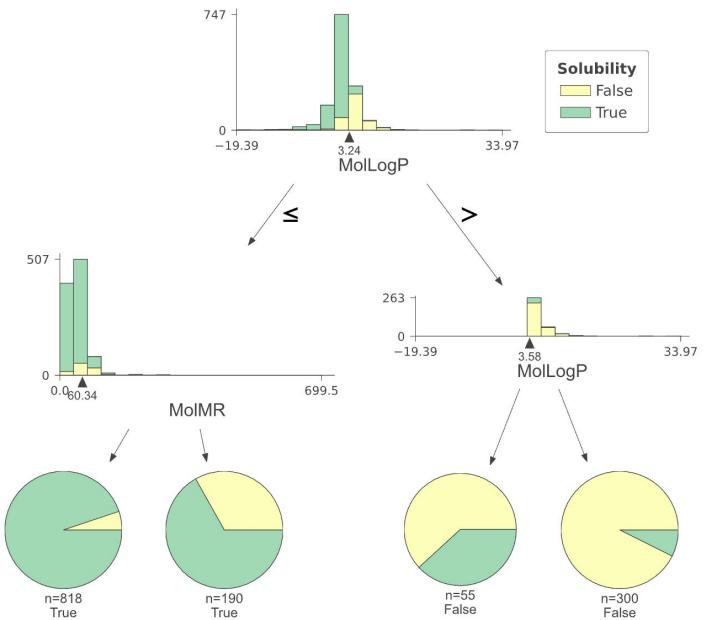


Physical Properties

Two Types of ML Models – Classification and Regression

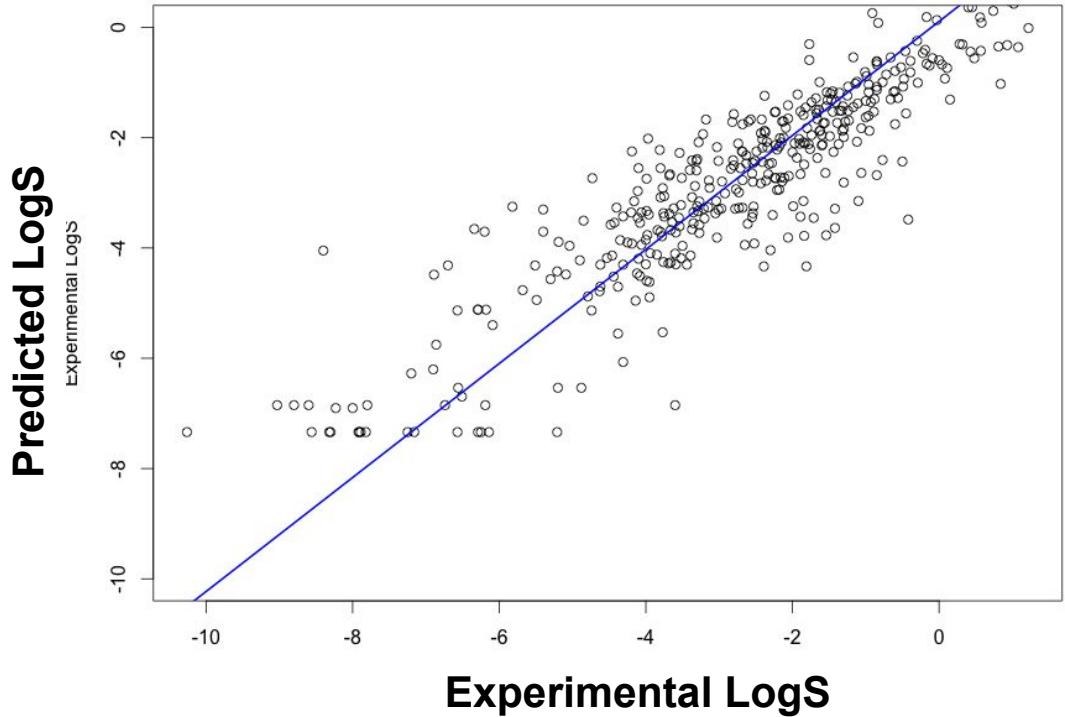
Classification

Out[32]:



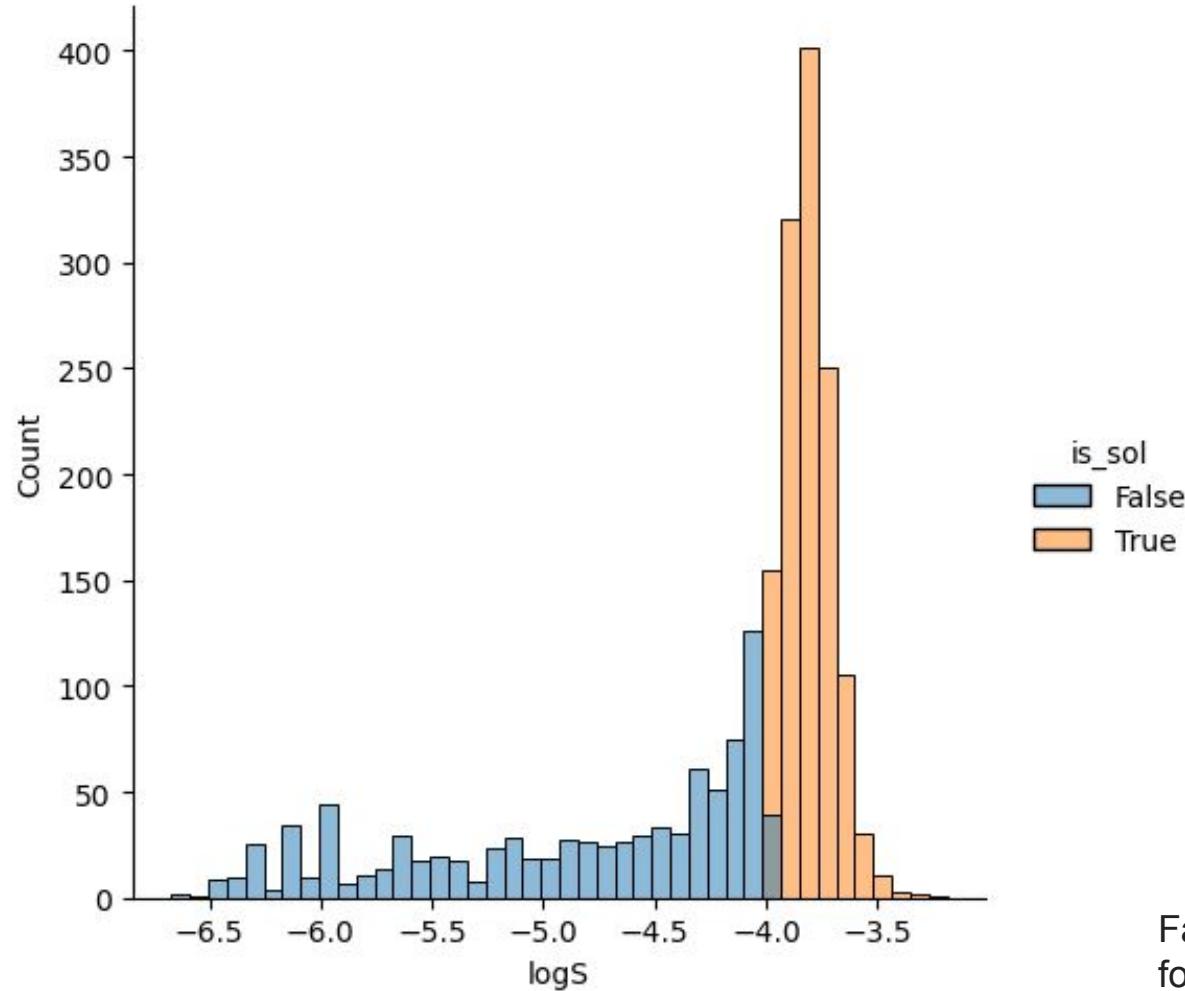
Predict a category, e.g.
soluble/insoluble

Regression



Predict a value, e.g. 4.2

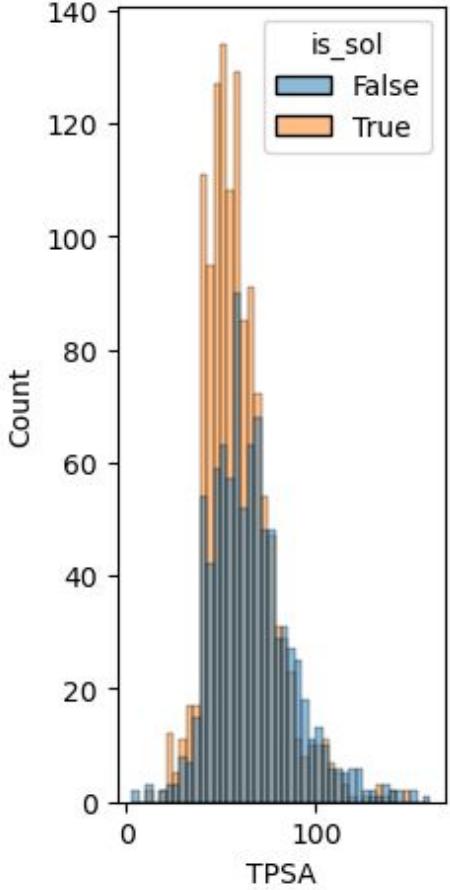
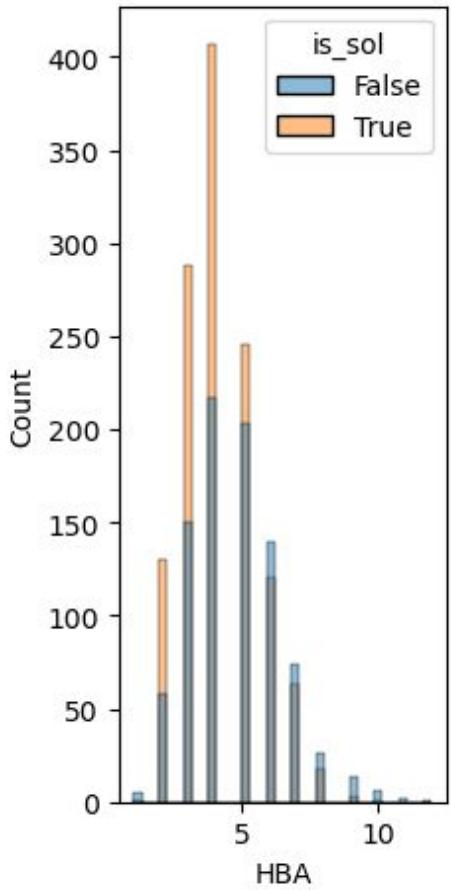
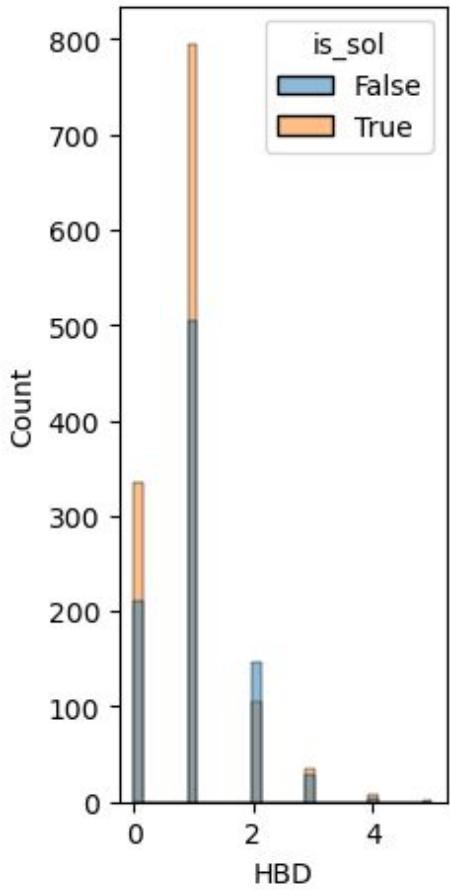
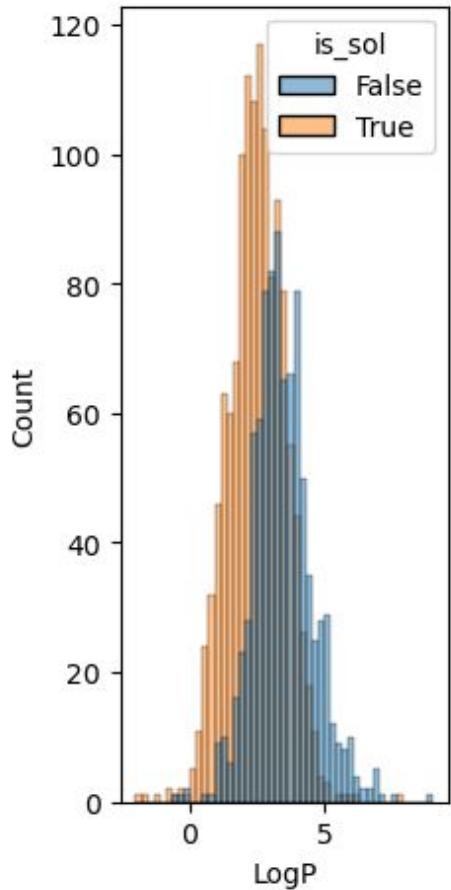
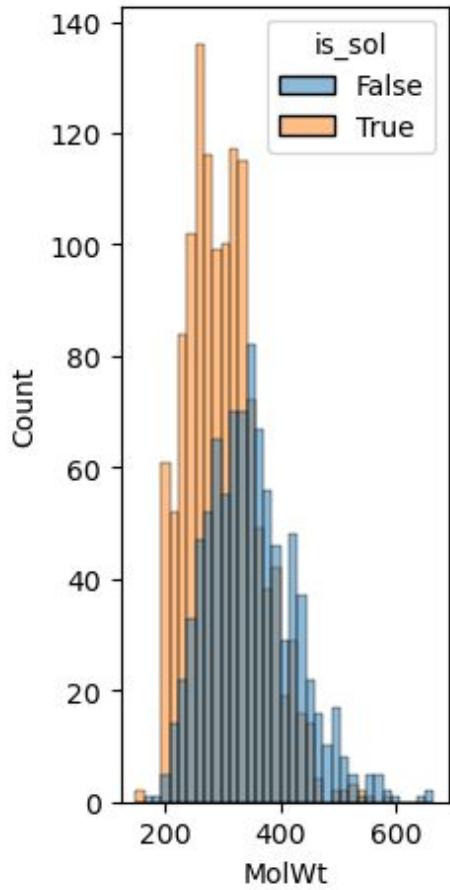
Let's Start With a Dataset



Aqueous solubility by CLND
2173 compounds
Min = 0.23 μ M
Max = 648 μ M

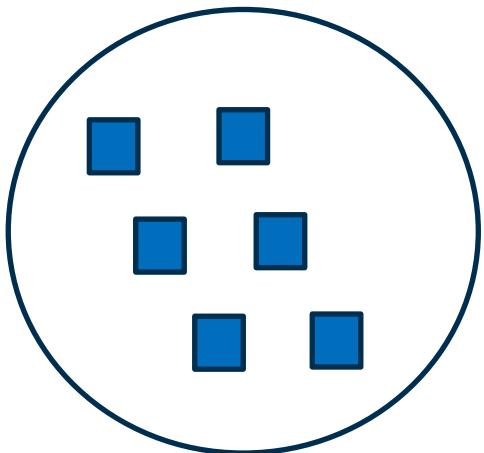
Fang, Cheng, et al. "Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective." *Journal of Chemical Information and Modeling* 63.11 (2023): 3263-3274.

Which Property Best Separates Soluble vs Insoluble Molecules?

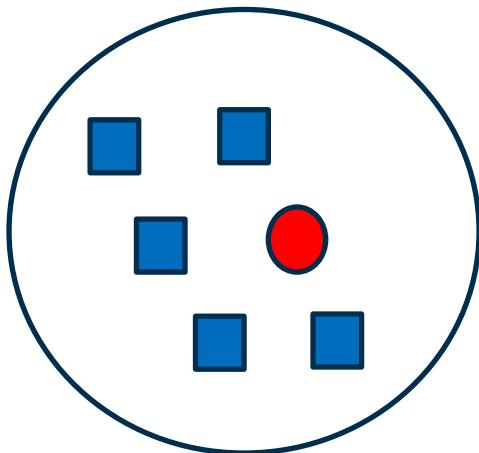


The Gini Index Quantifies the “Purity” of a Split

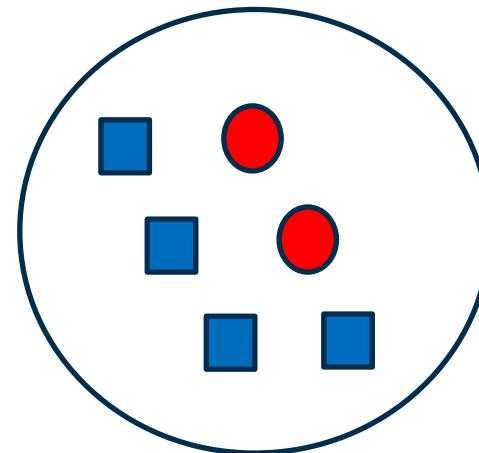
$$\text{Gini Index} = 1 - \sum_{i=1}^n (P_i)^2$$



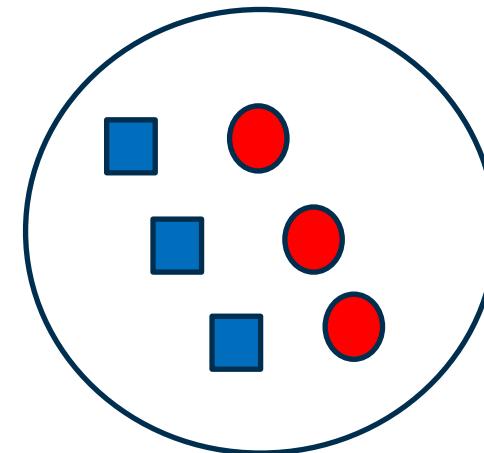
1.00



0.83

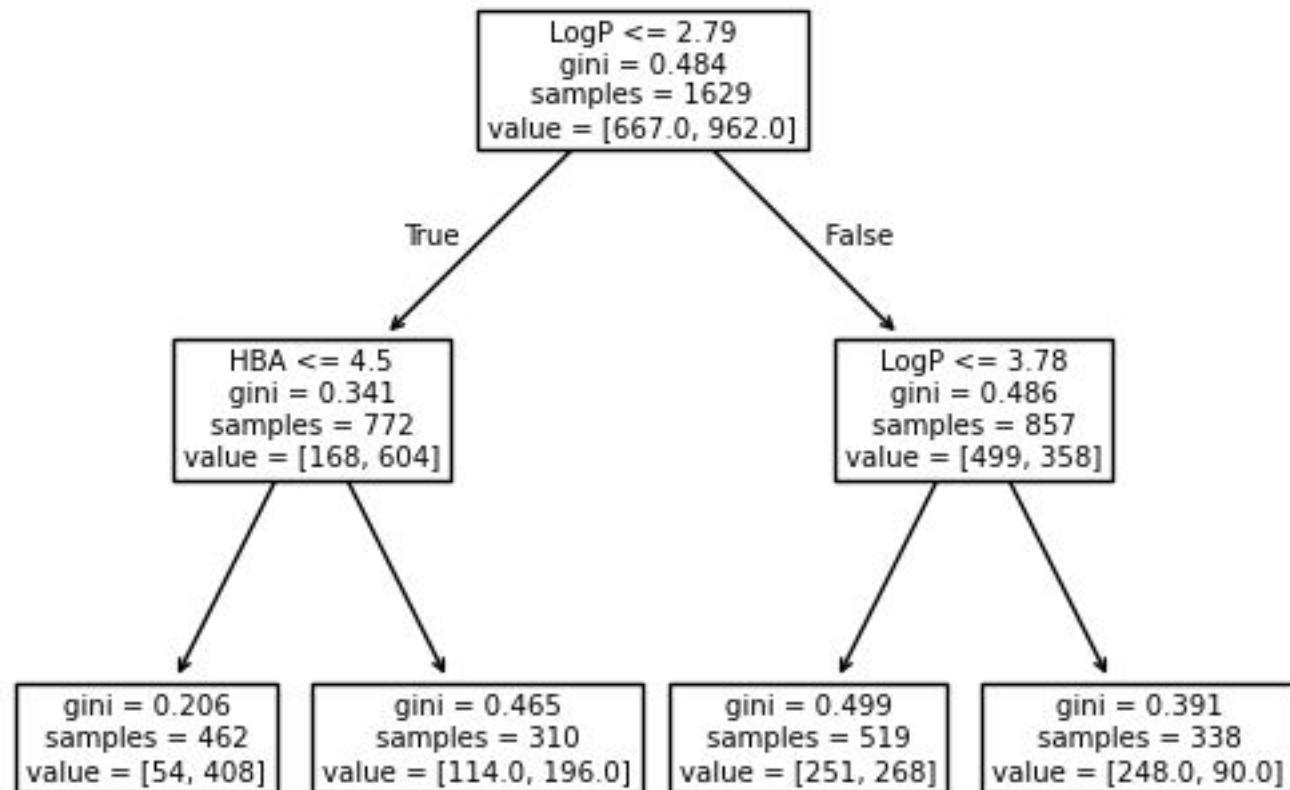


0.66

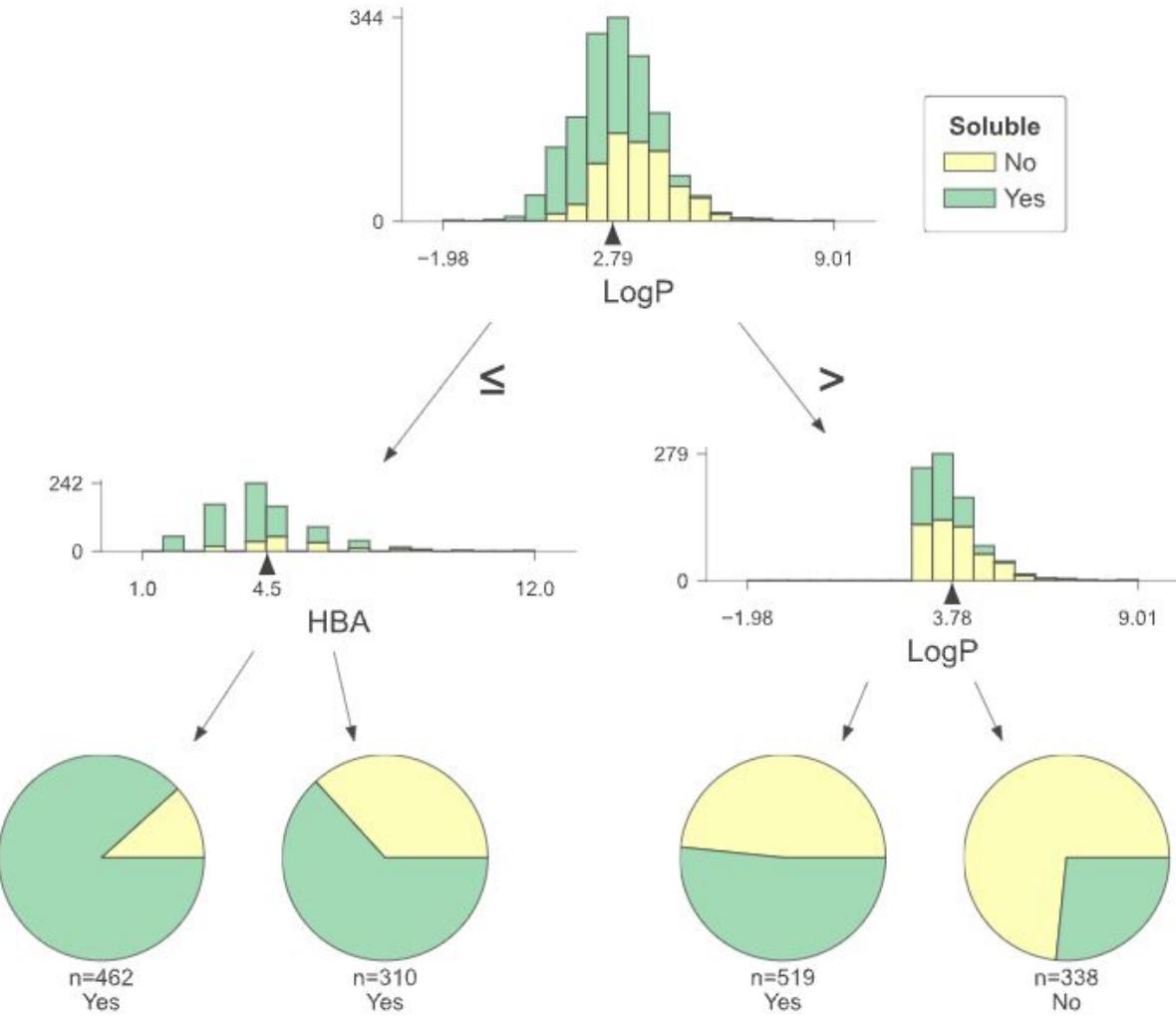


0.50

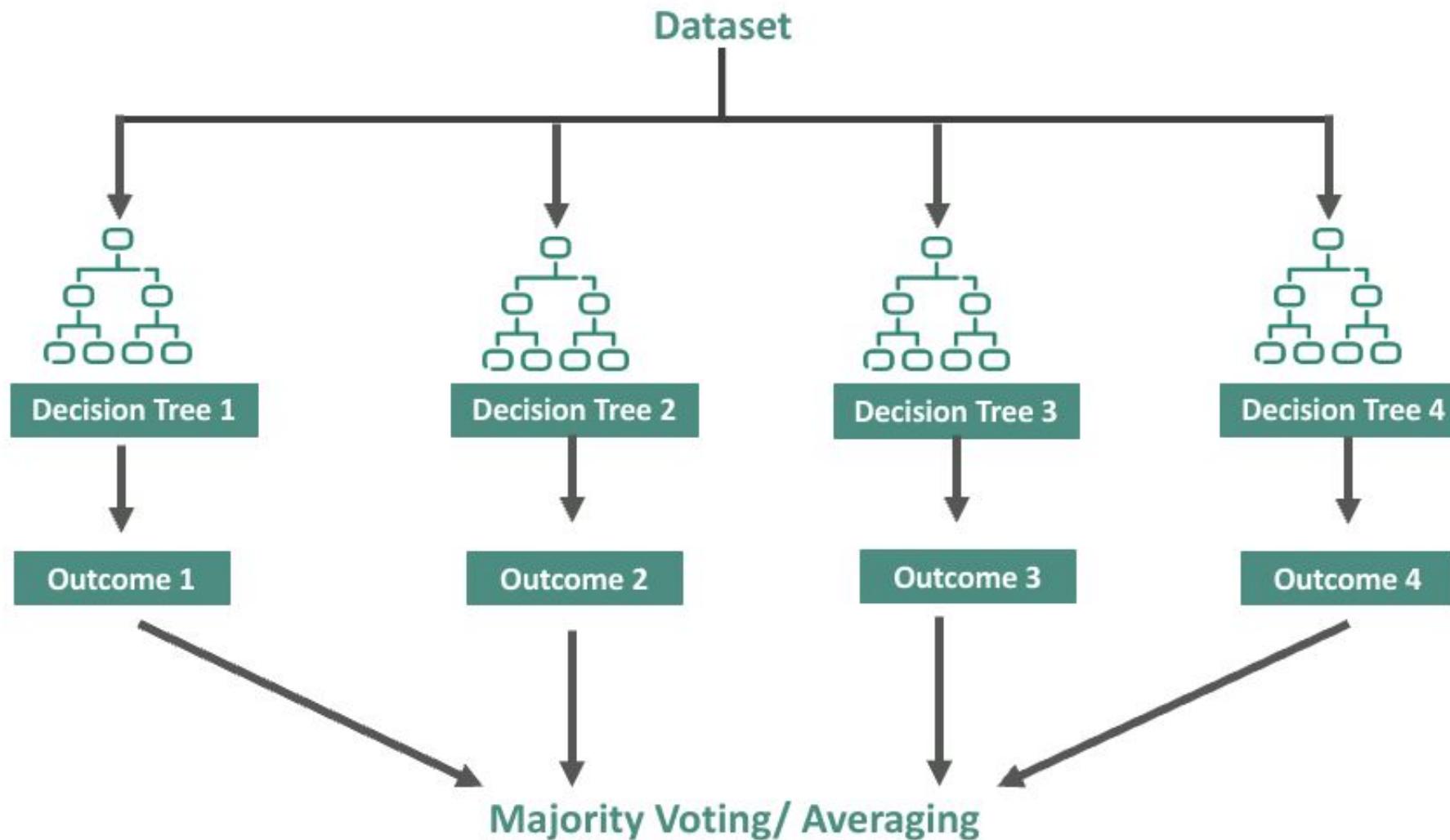
Build a Decision Tree



A Better (IMO) Decision Tree Visualization



Random Forest Uses an Ensemble of Decision Trees



There Are Many Tree Ensemble Methods



XGBoost

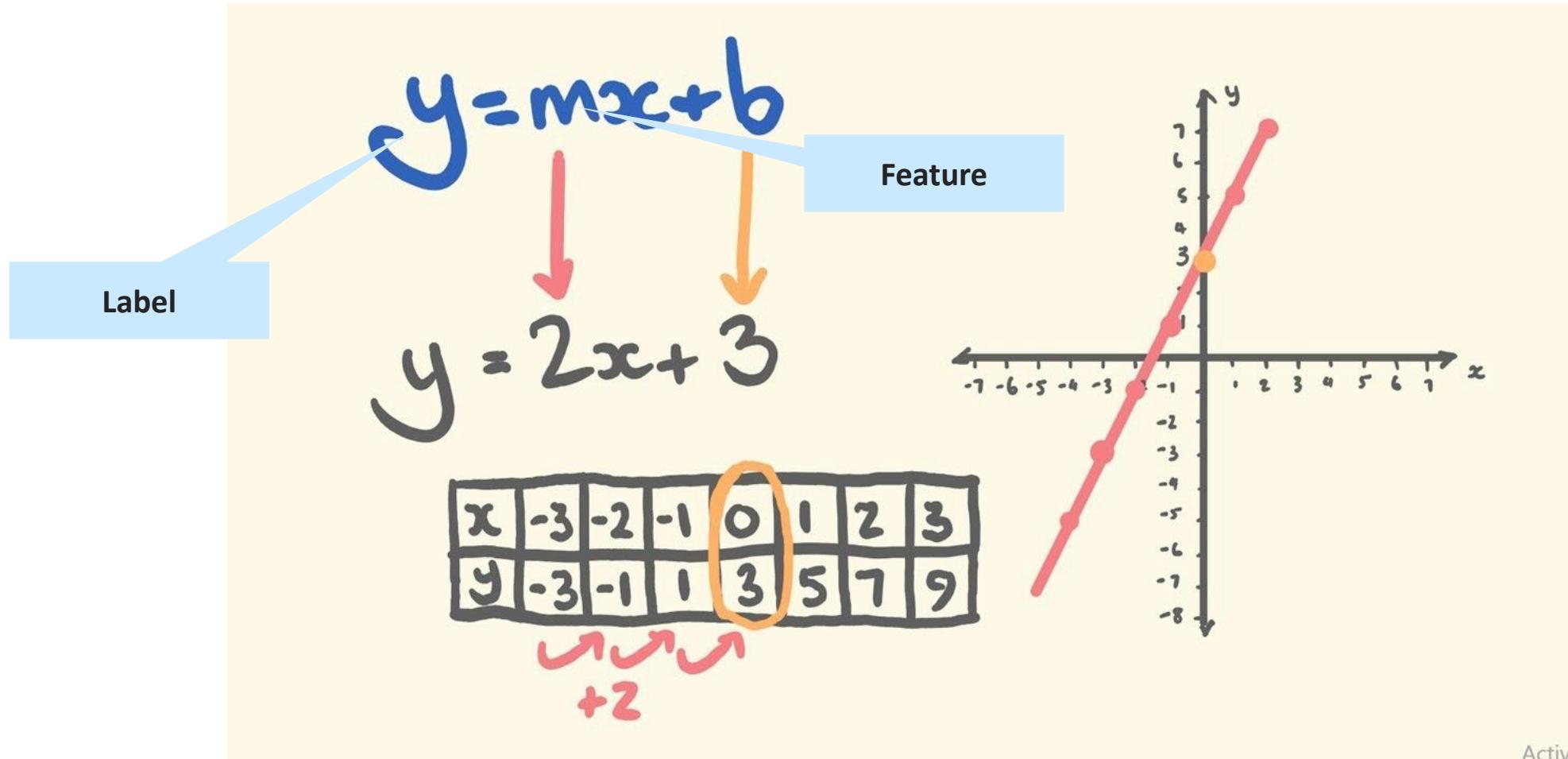


CatBoost

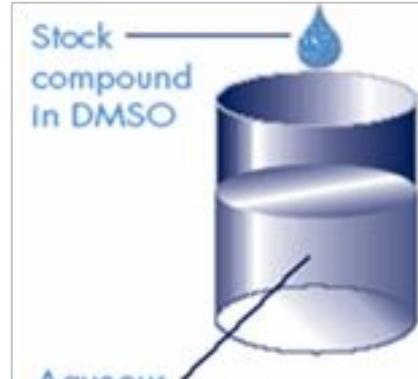


LightGBM

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



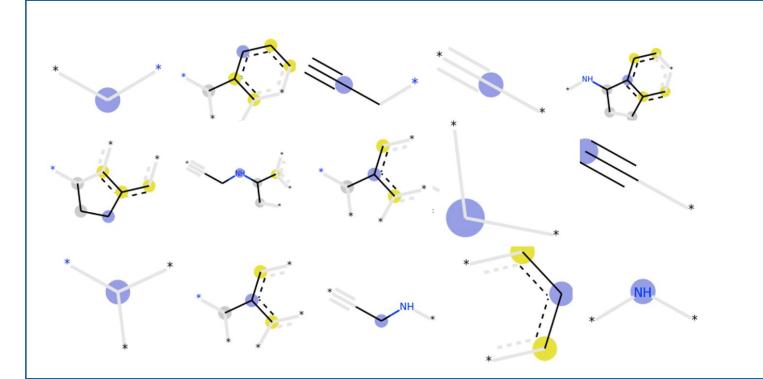
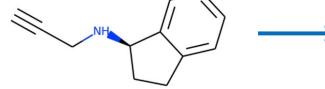
Defining Labels (y) and Features (X) - An Example



$y = \text{Log Aqueous Solubility}$

Label

Features



010110010001000111001000110010010101010111

$X = \text{A Vector Representing a Molecule}$

Define Features Based on Books People Have Read



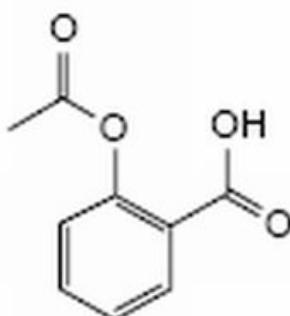
Fred	1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0
------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

Sally	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	0
-------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

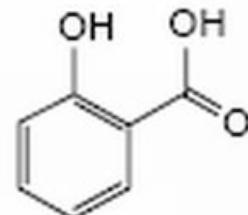
Jane	0	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0
------	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---

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

Chemical Fingerprints as Molecular Descriptors



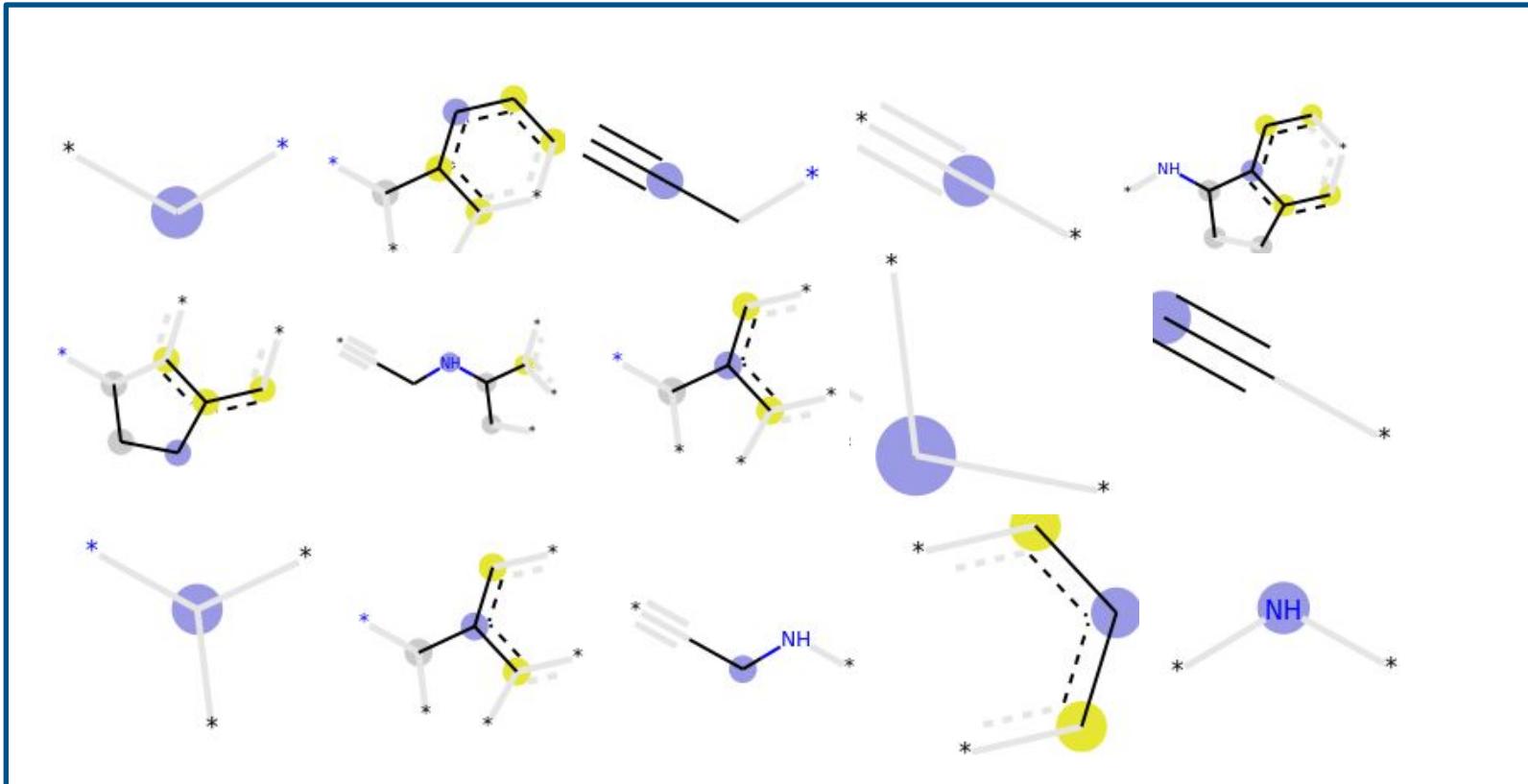
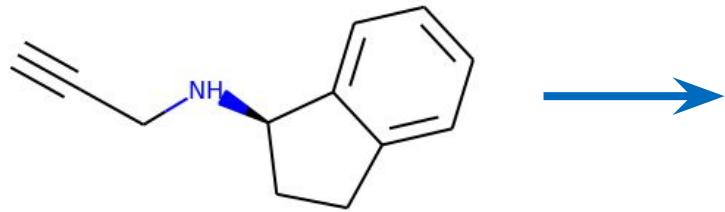
1



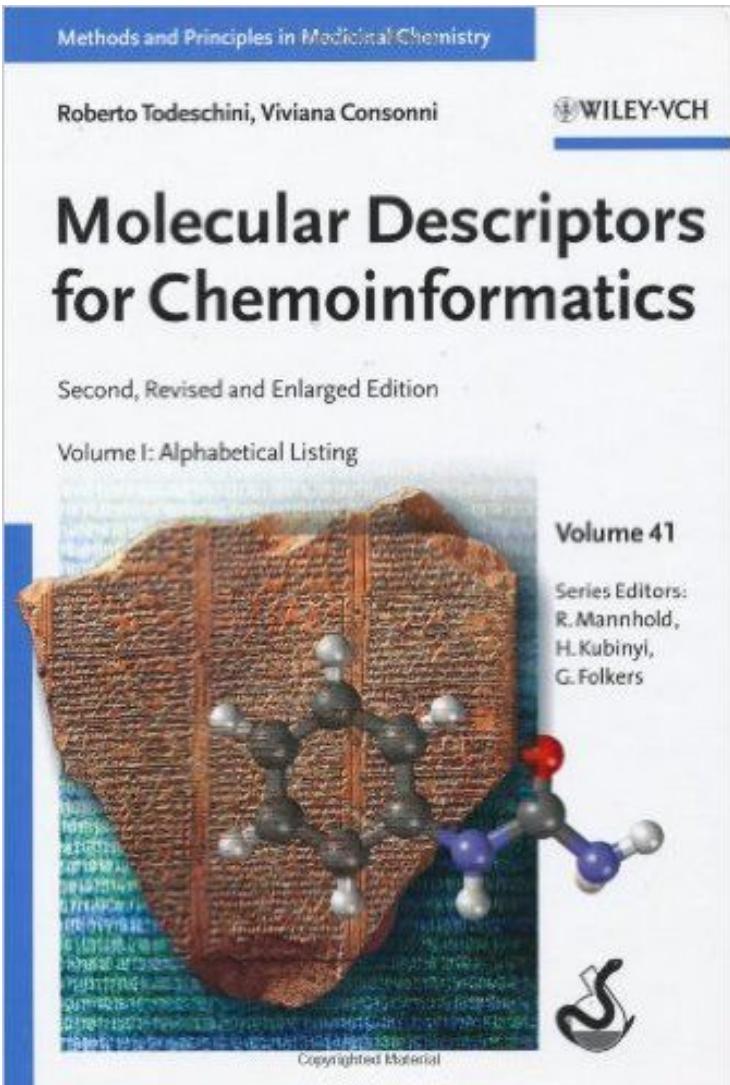
2

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



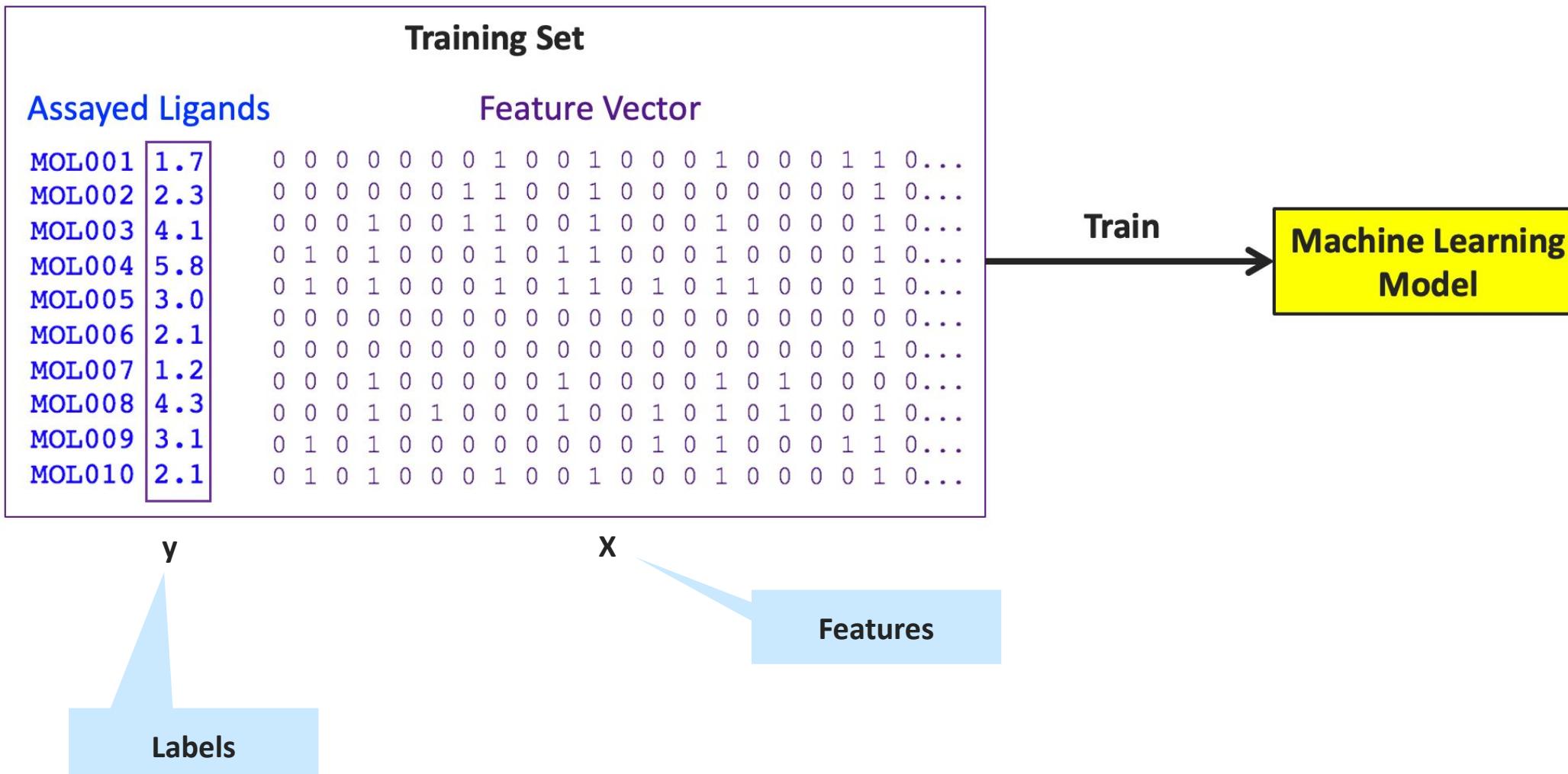
There are A LOT of ways to do this



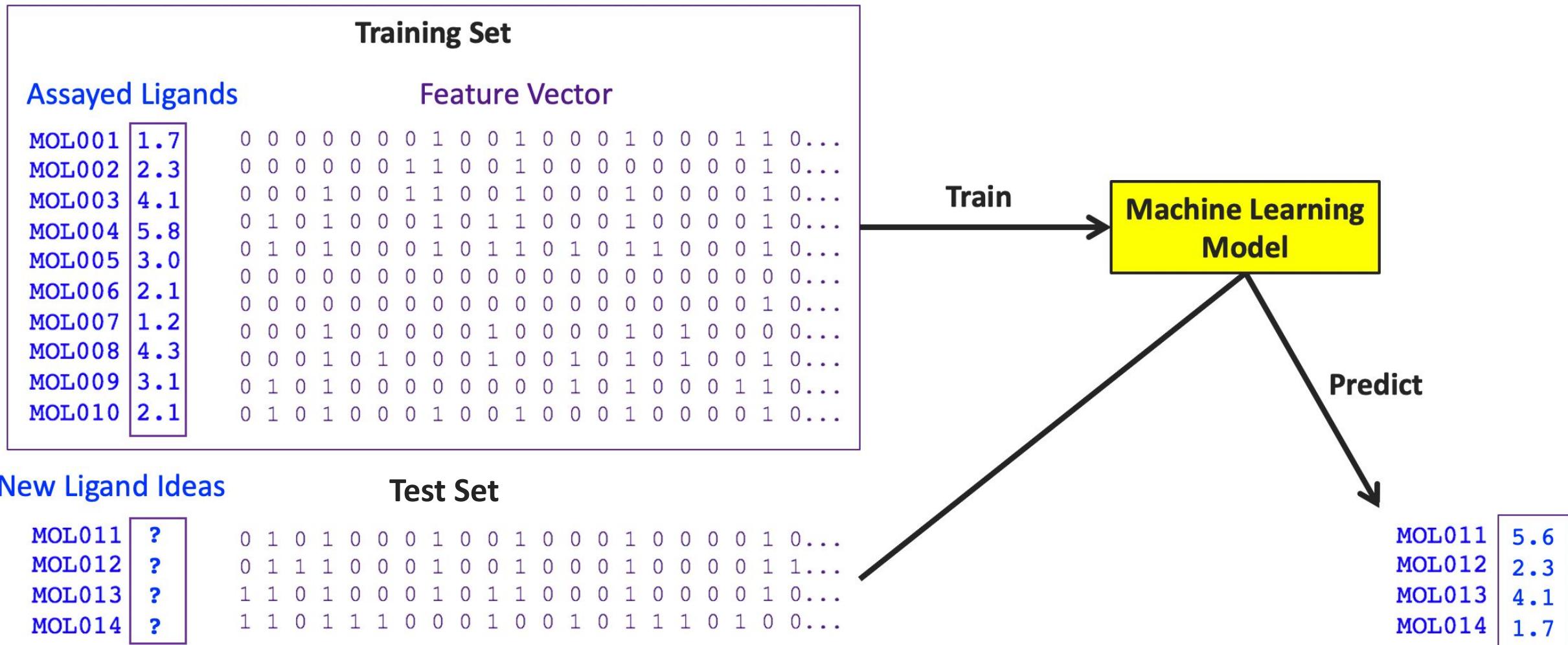
2 Volumes
6,000 references from 450 journals

DRAGON 7 has 5,270 descriptors z

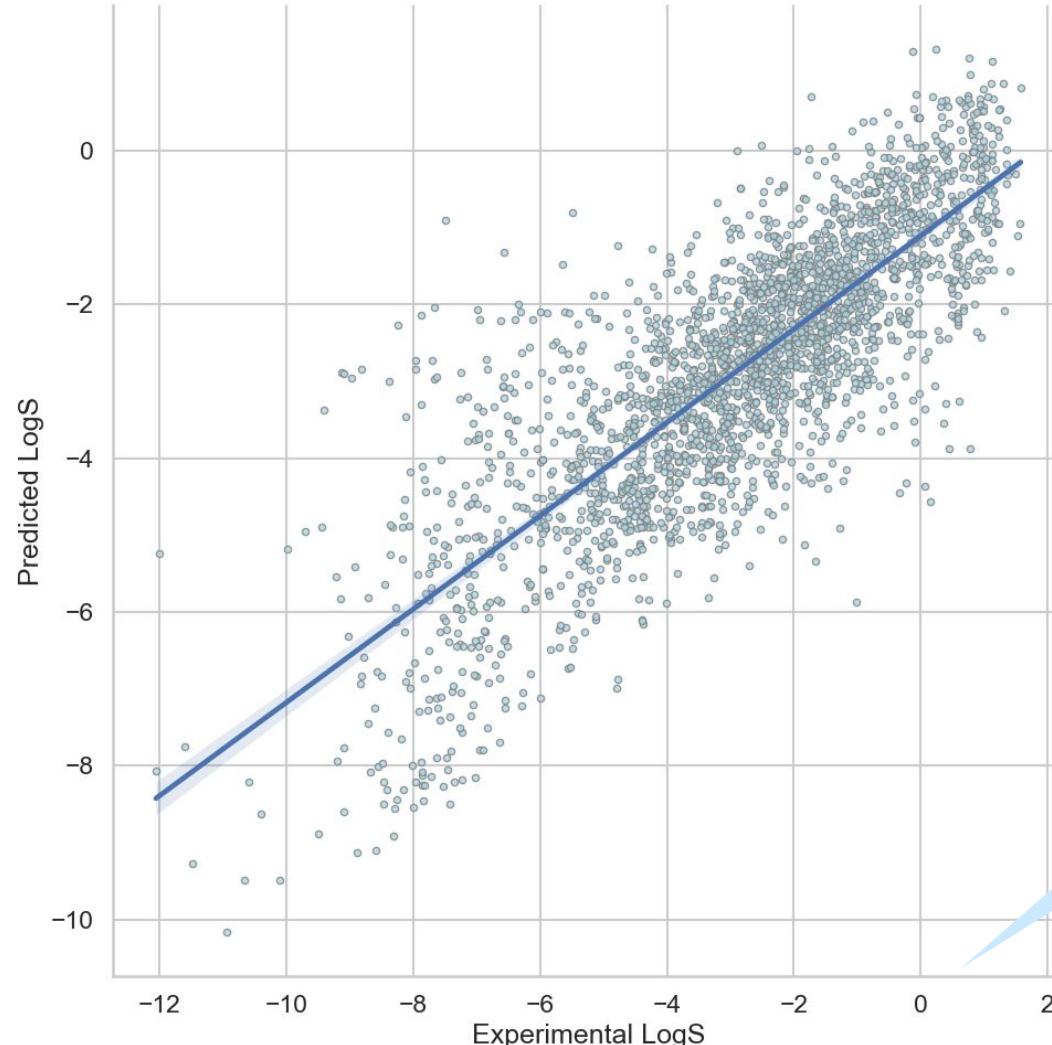
Training a Machine Learning Model



Making Predictions With a Machine Learning Model

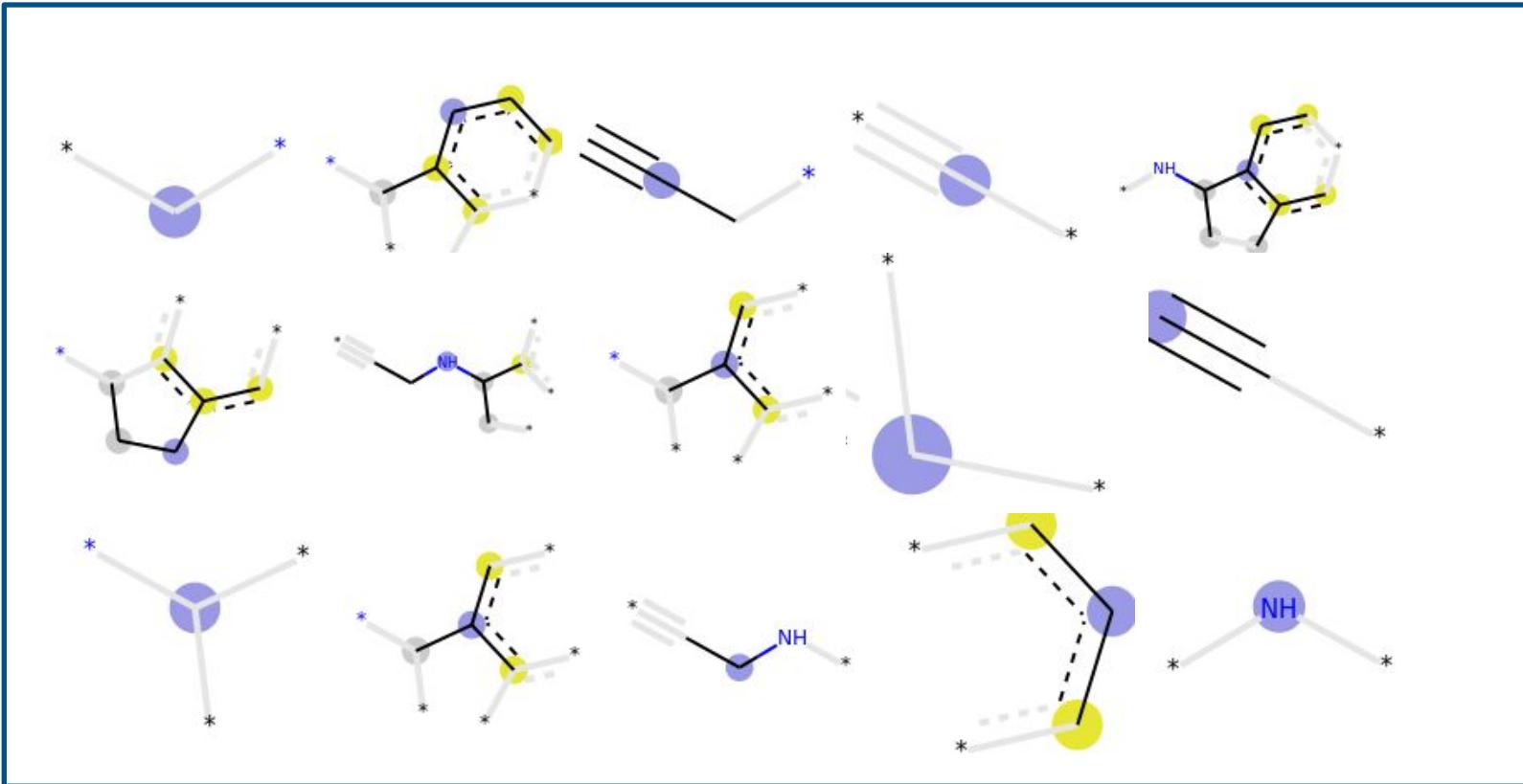
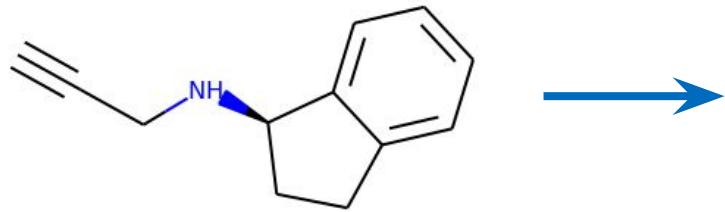


Prediction Performance of an Aqueous Solubility Model

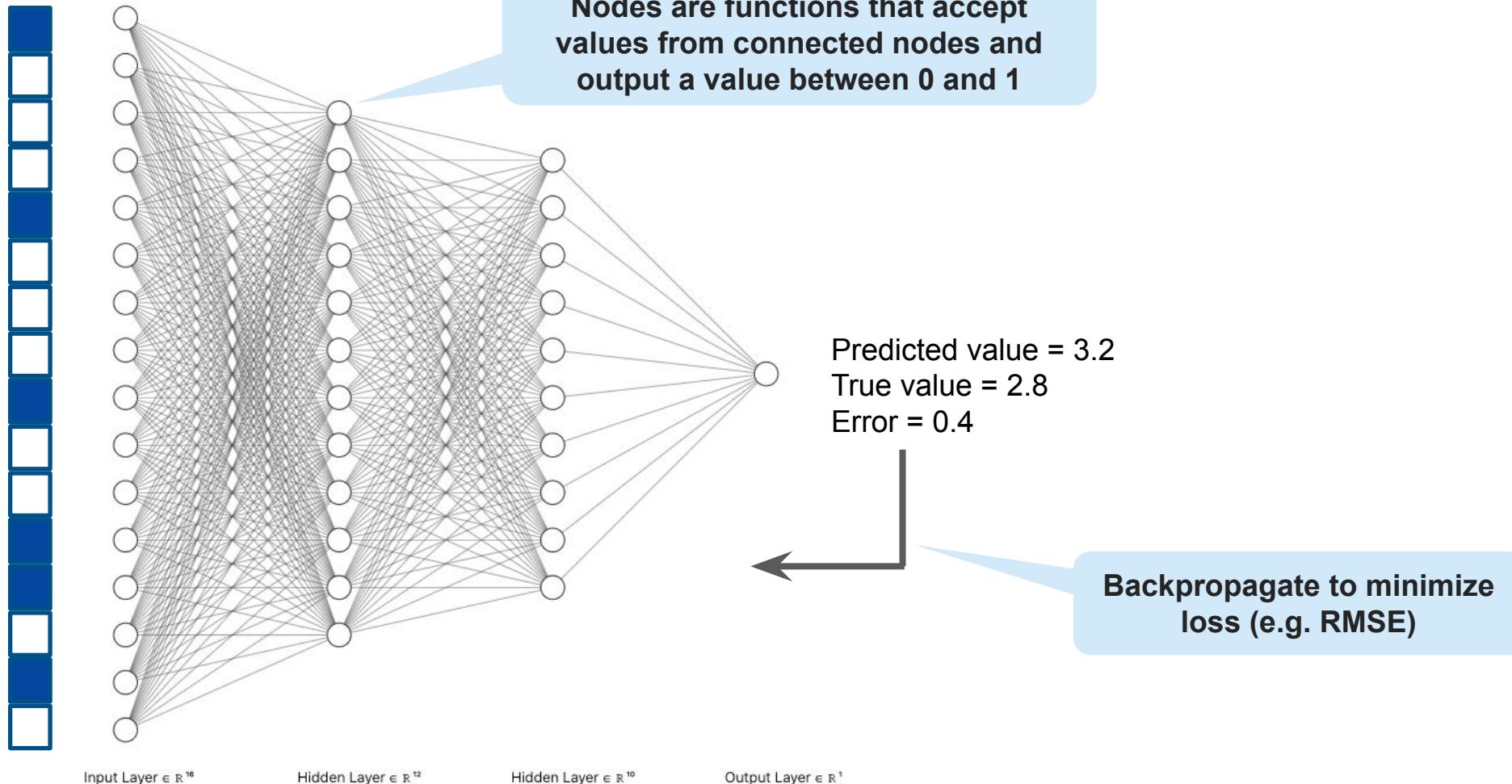


Dynamic range is very large,
does not represent a typical use
case

Chemical Fingerprints as Vector Representations of Molecules



Neural Networks Adjust Weights to Minimize a Loss Function

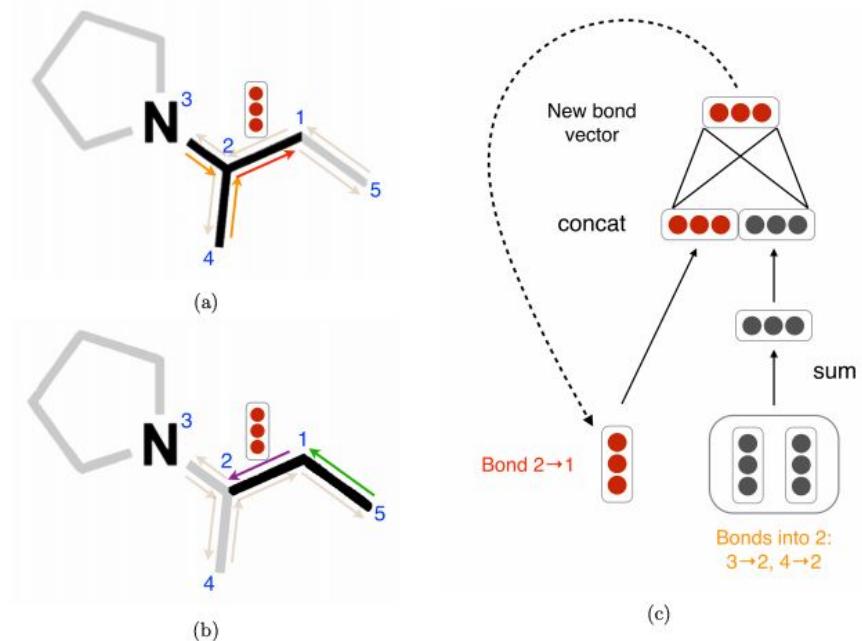


Using Neural Networks to Create New Molecular Representations



Graph Convolutions

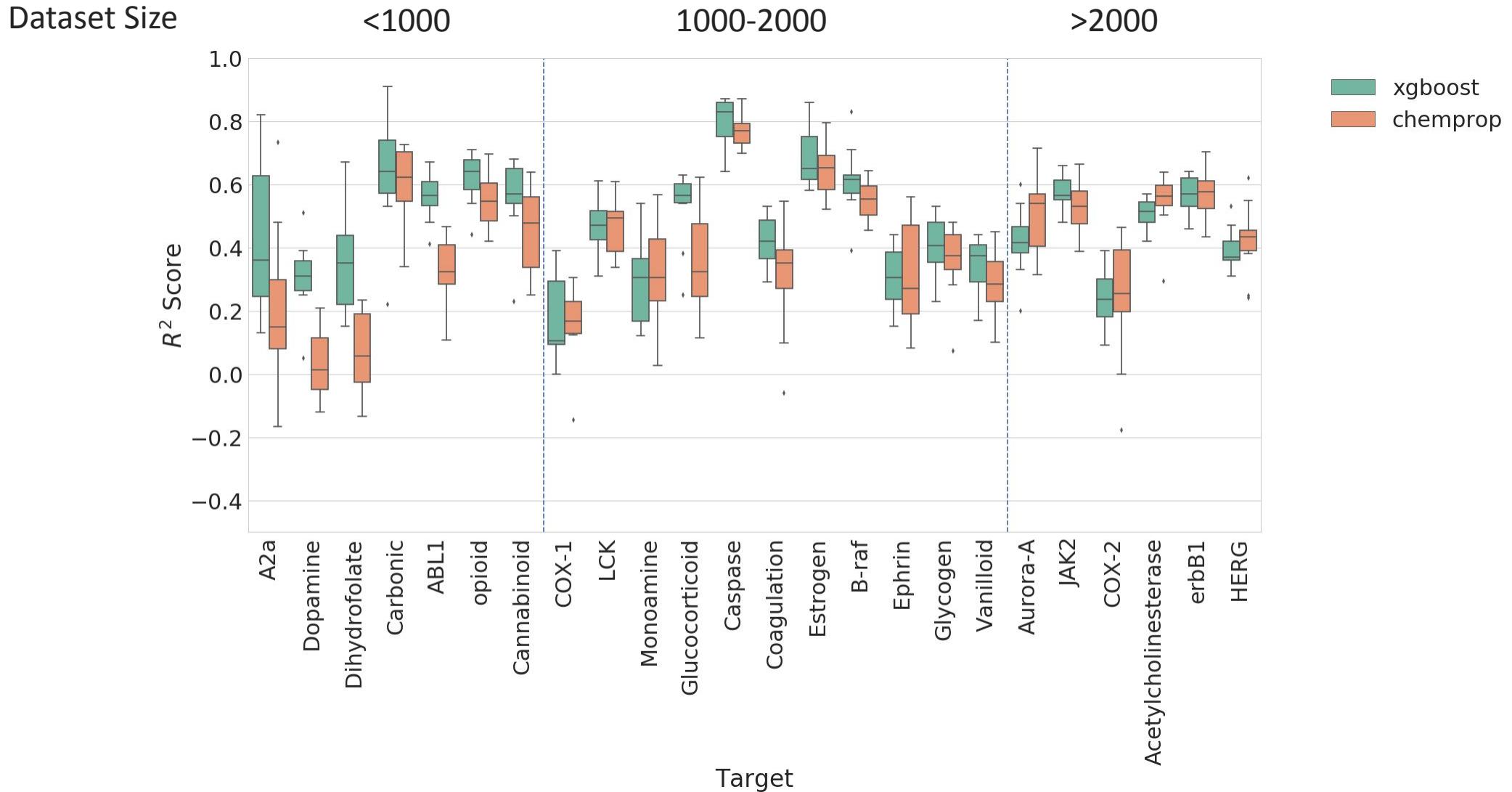
J. Comput. Aided Mol. Des. 2016, 595–608



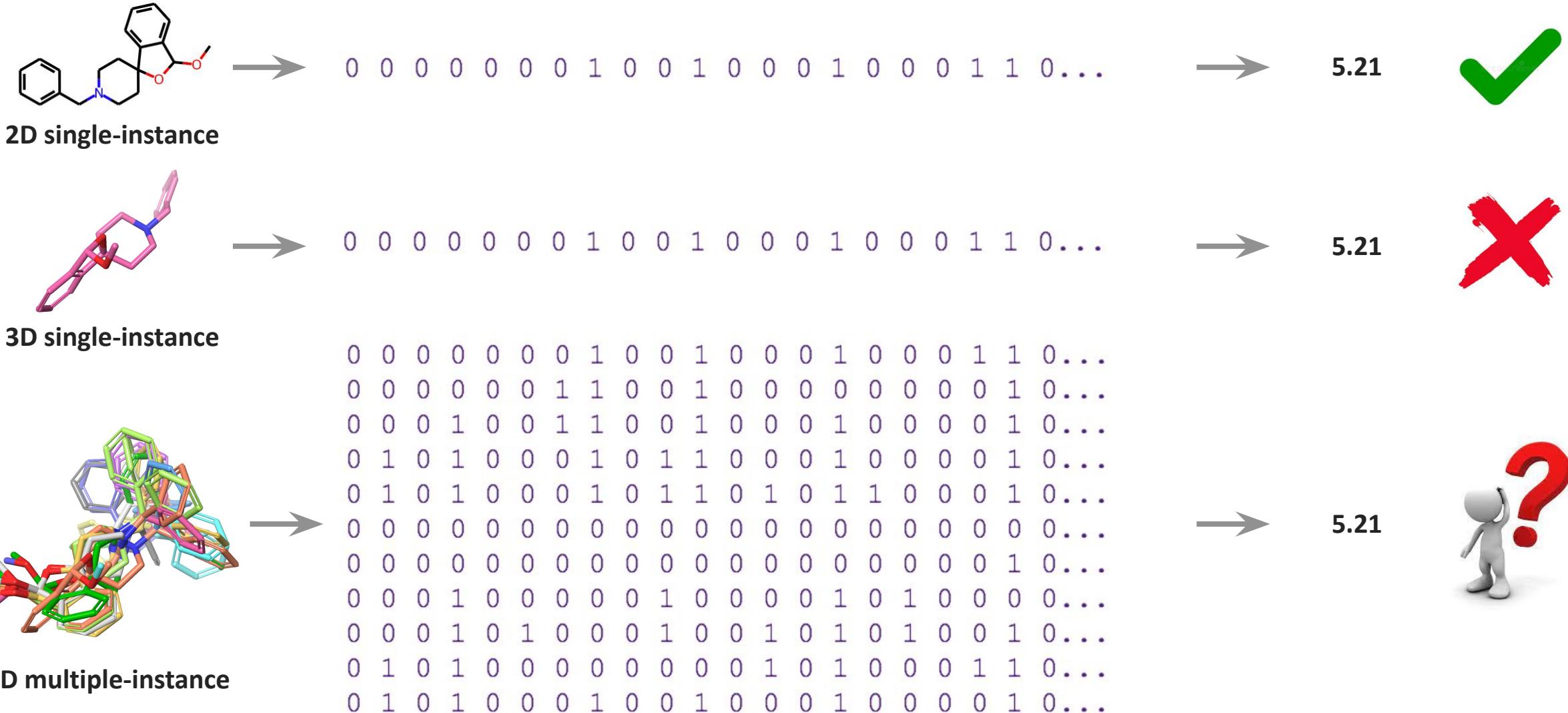
Message Passing Neural Network

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

Are Neural Network Representations Better?



Incorporating 3D into Molecular Machine Learning is an Unsolved Problem



How to Find Me

<https://pwalters.github.io>

My Boring Website

Publications

Tutorials

Blog

Videos



Pat Walters

Cheminformatics, ML

📍 Cambridge, MA

✉️ Email

✉️ Google Scholar

⌚ Github

.linkedin LinkedIn

𝕏 X (formerly Twitter)

Pat Walters is Chief Data Officer at Relay Therapeutics in Cambridge, MA. Prior to joining Relay, he spent more than 20 years at Vertex Pharmaceuticals where he was Global Head of Modeling & Informatics. 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.