AI in Drug Discovery – An Overview
Session 1

September 16, 2024

Who we are!

Pat Walters
Relay Therapeutics

Raquel López-Ríos de Castro
Charité Berlin, MSKCC NYC

Michael Backenköhler
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

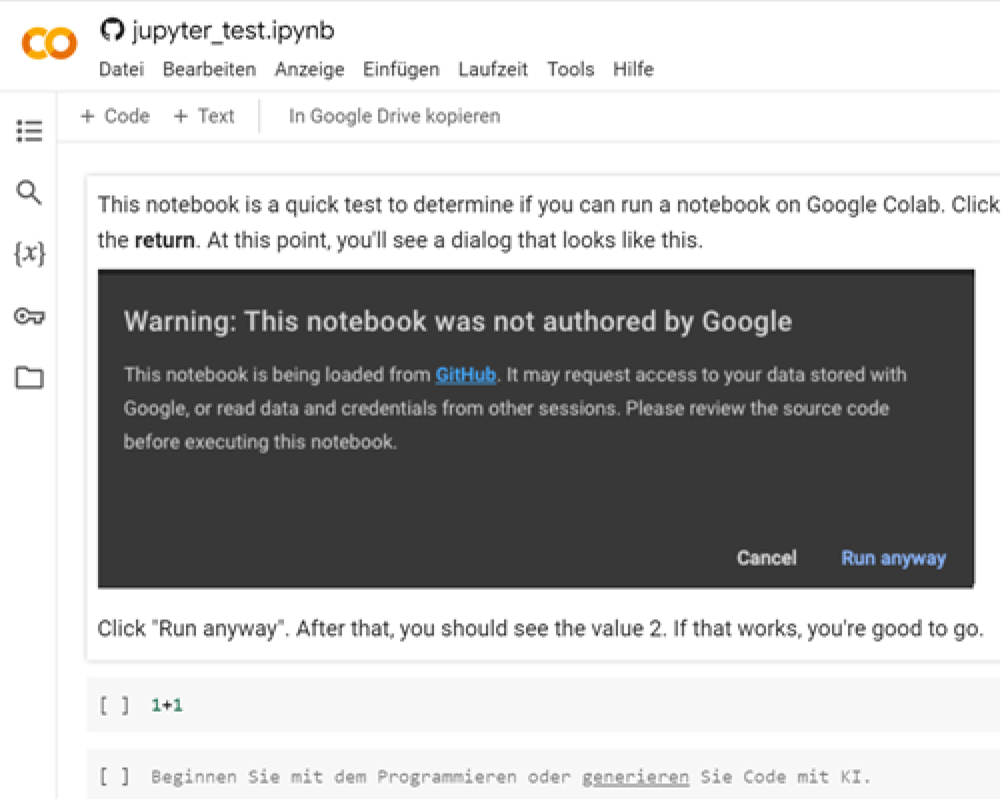
Session 2 - 3:00 - 4:00 pm

- 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
- Active learning

Lectures supported by hands-on sessions …



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.

Warning: This notebook was not authored by Google

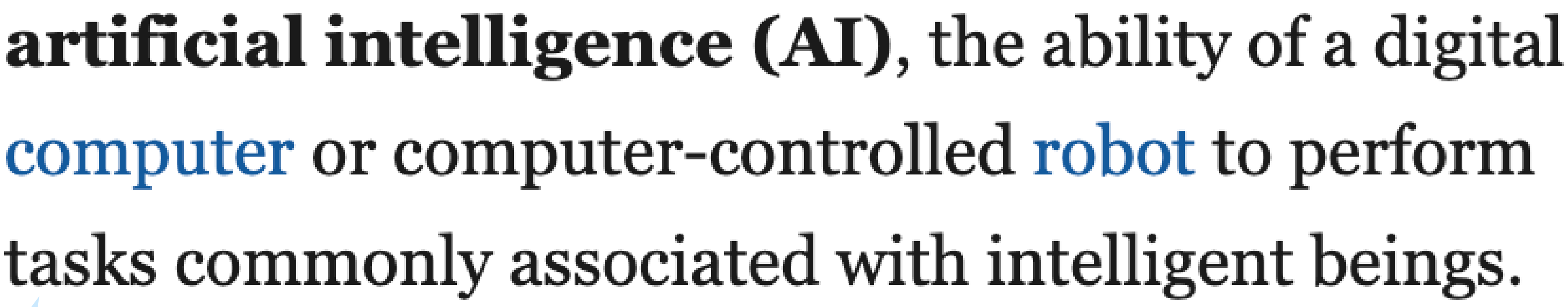
This notebook is being loaded from GitHub. It may request access to your data stored with Google, or read data and credentials from other sessions. Please review the source code before executing this notebook.

Click "Run anyway". After that, you should see the value 2. If that works, you're good to go.

```
[ ] 1+1
```

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

Definitions Can Be Tricky

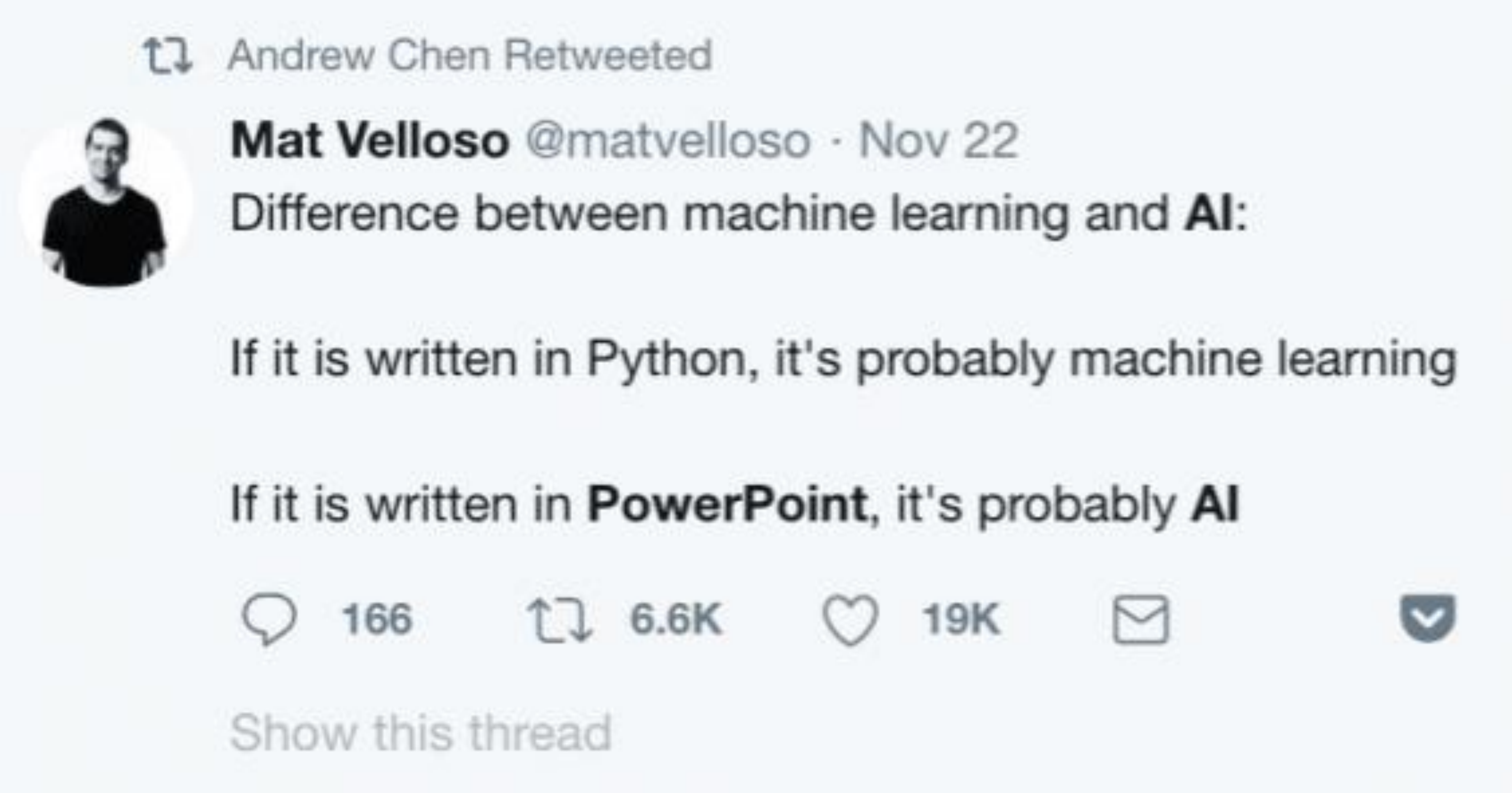


artificial intelligence (AI), the ability of a digital computer or computer-controlled robot to perform tasks commonly associated with intelligent beings.

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**

What Is Machine Learning?

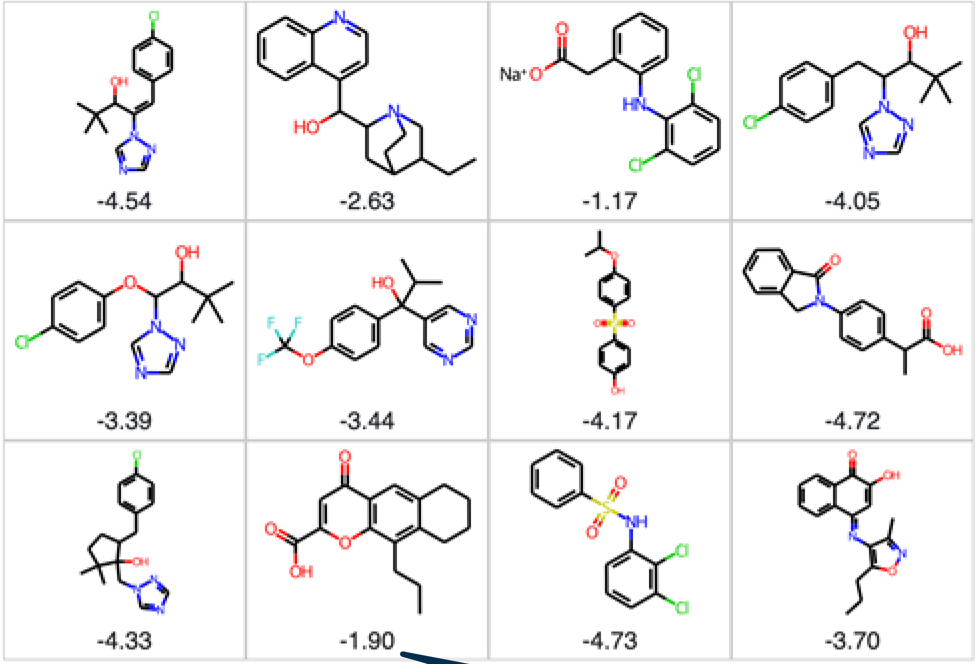
Machine learning is all about labeling things using examples

Cassie Kozyrkov, Google

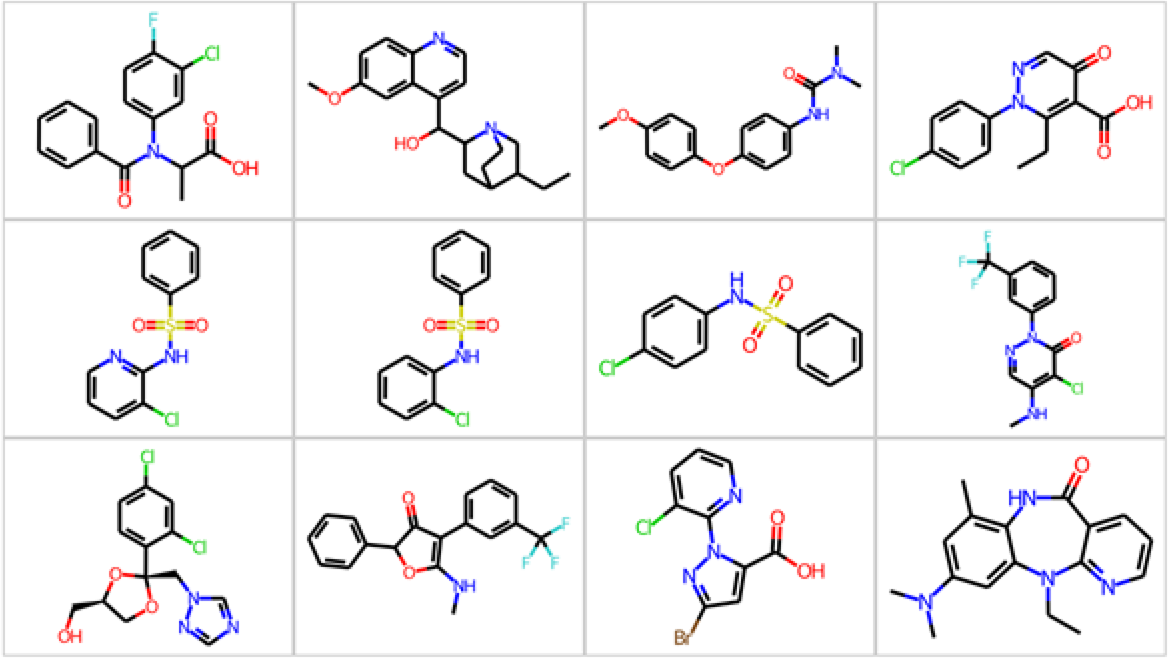
https://hackernoon.com/the-simplest-explanation-of-machine-learning-youll-ever-read-bebc0700047c

Labeling Molecules Based on Examples

Molecules with measured data

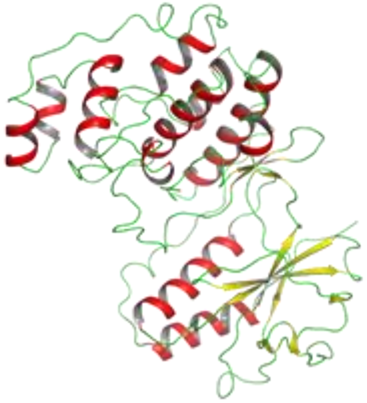


Molecules to be predicted



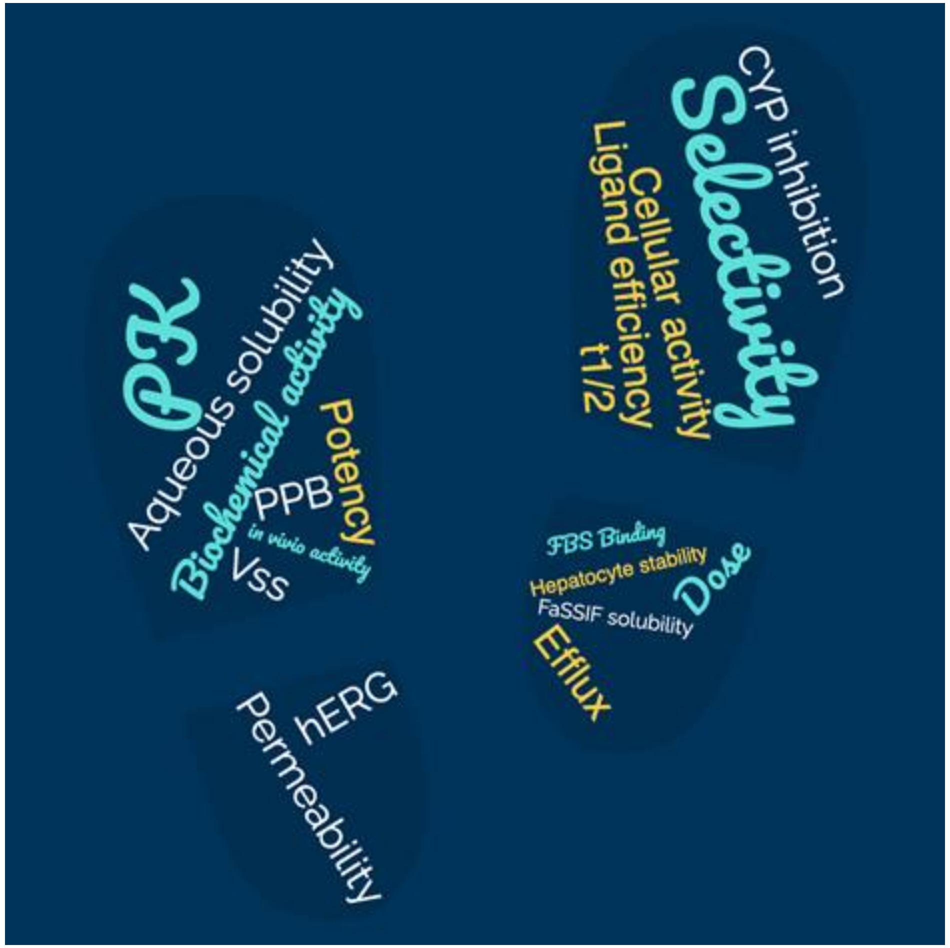
Log10(Molar Aqueous Solubility)

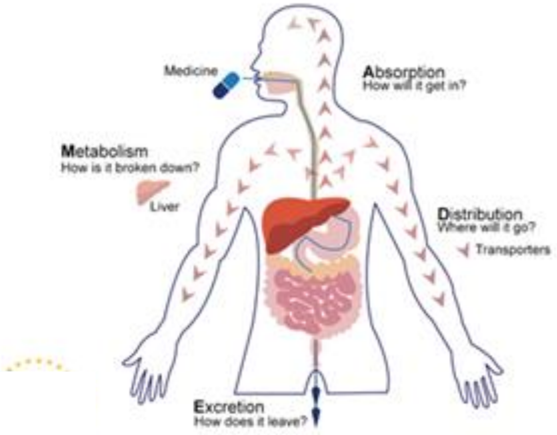
Using Predictive Models to Drive Drug Discovery



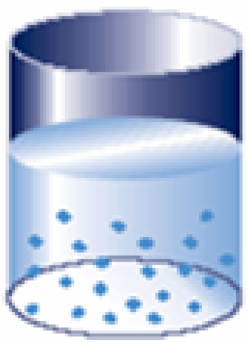
On-target Activity

Off-target Activity





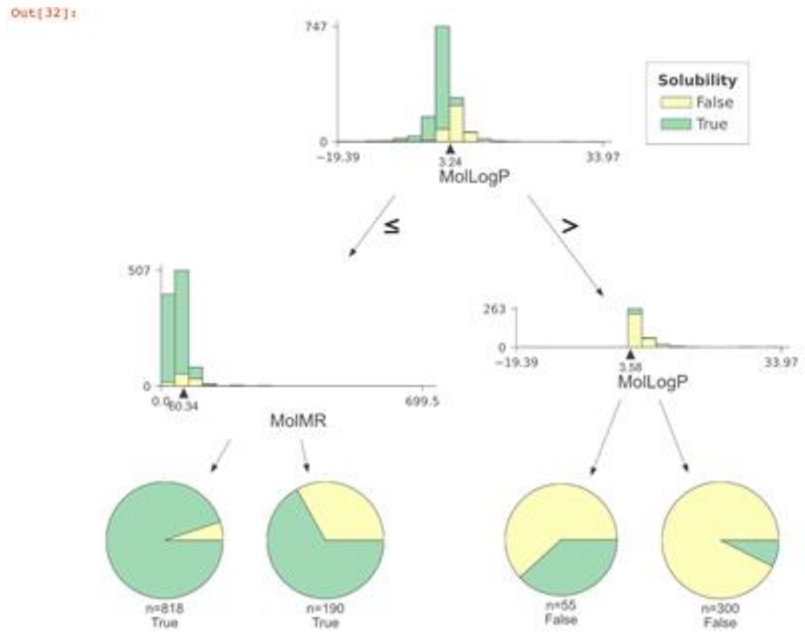
Pharmacokinetics



Physical Properties

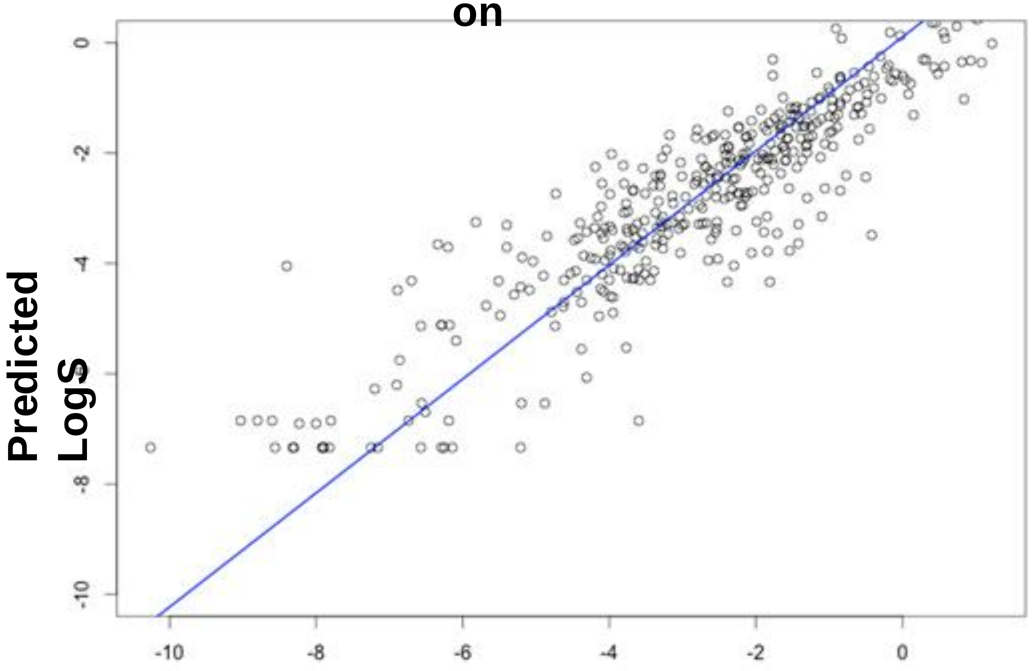
Two Types of ML Models – Classification and Regression

Classification



Predict a category, e.g. soluble/insoluble

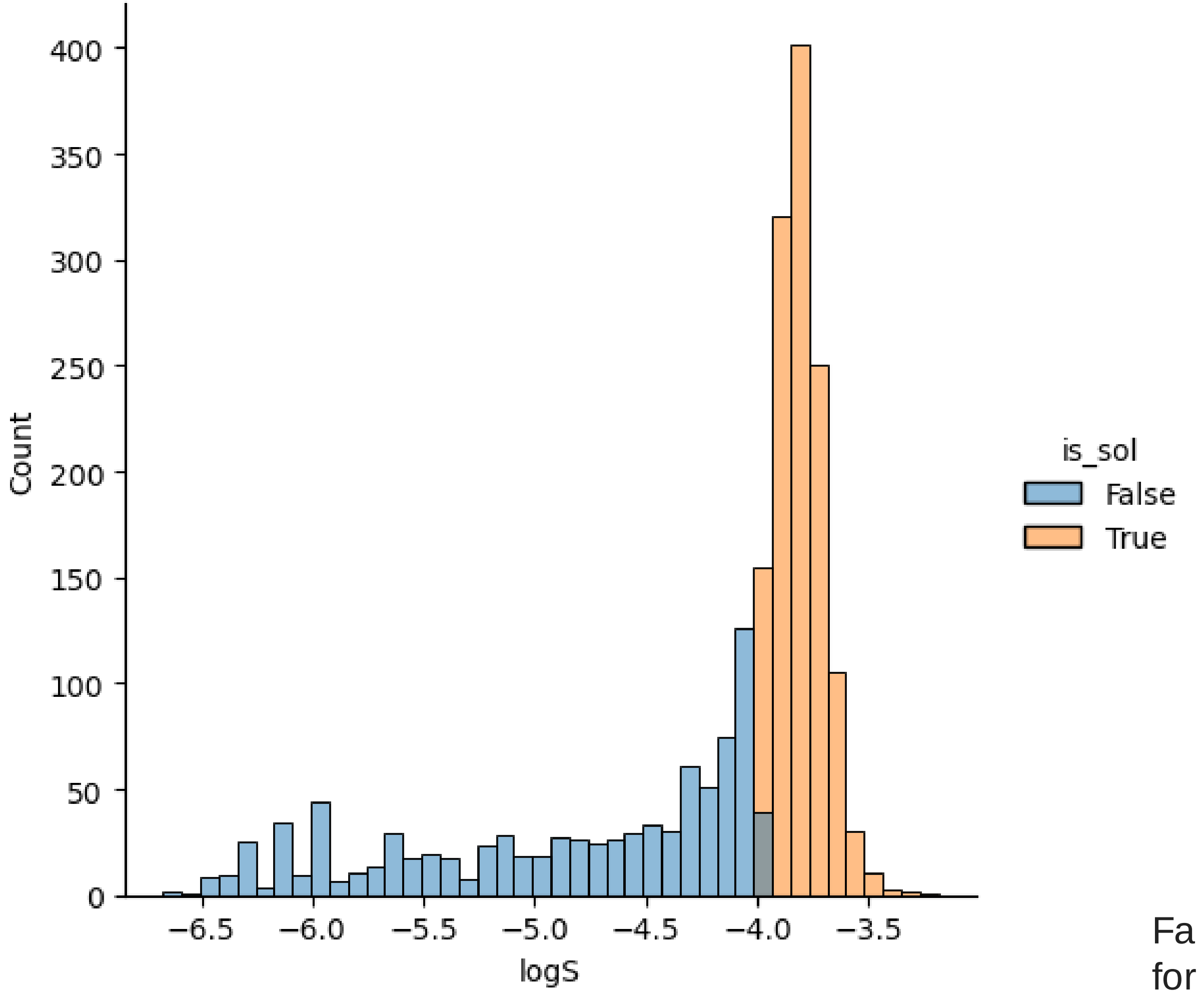
Regression



Experimental LogS

Predict a value, e.g. 4.2

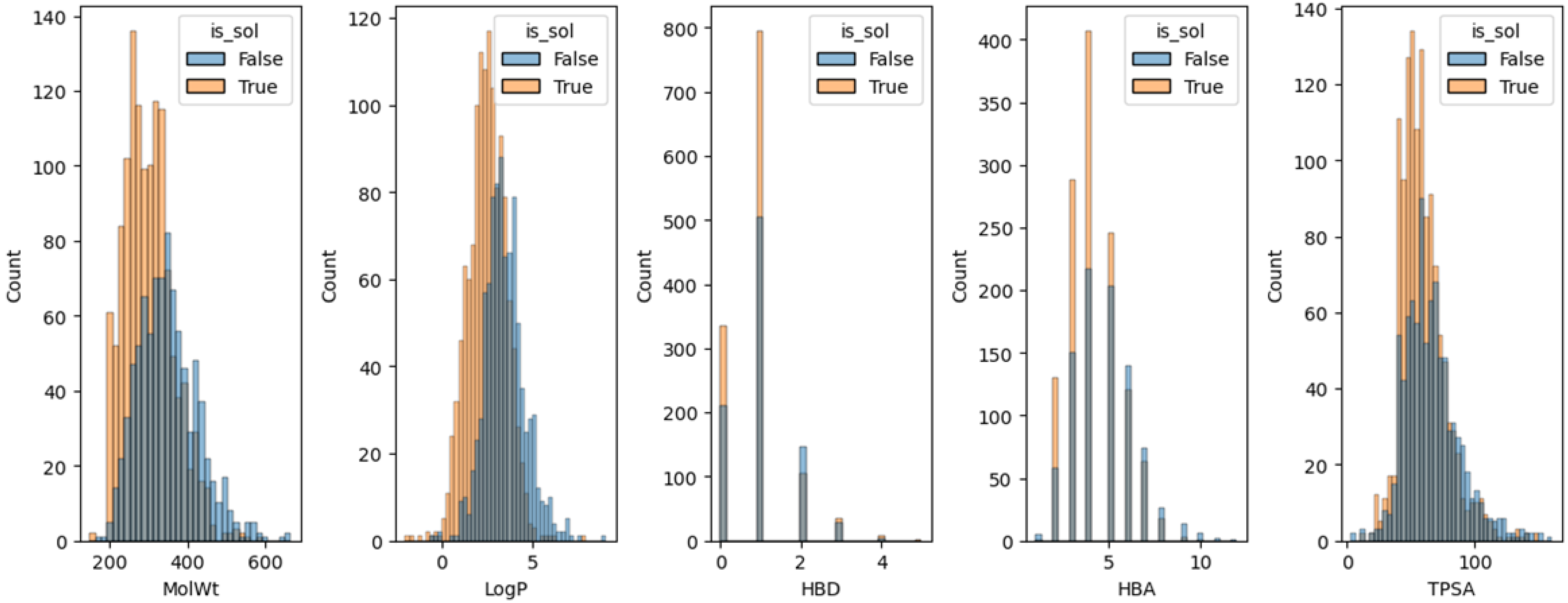
Let’s Start With a Dataset



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

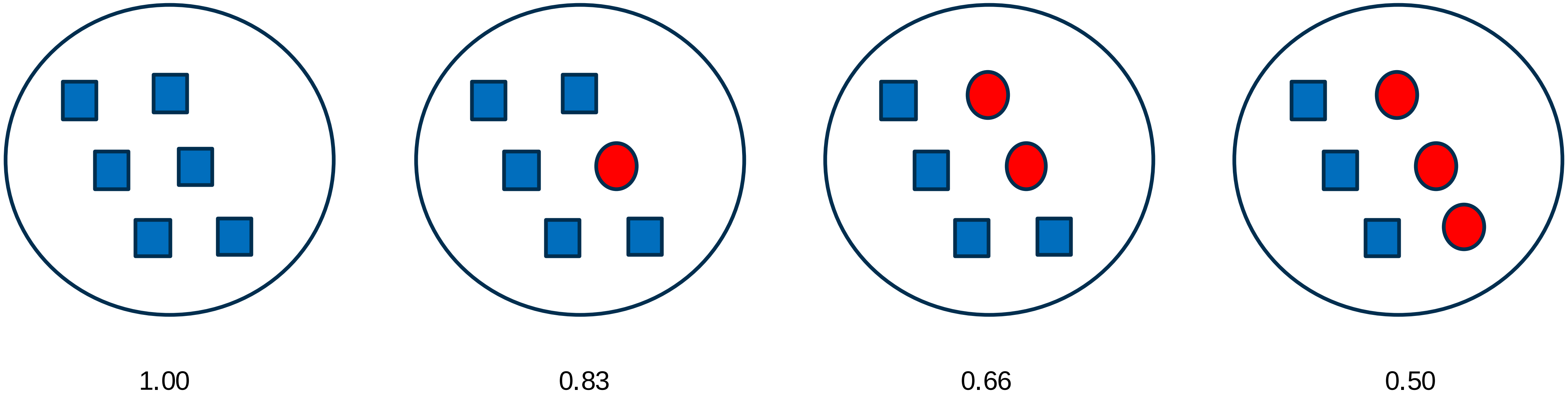
Fang, Cheng, et al. "Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective." *Journal of Chemical Information and Modeling* 63.11 (2023): 3263-3274.

Which Property Best Separates Soluble vs Insoluble Molecules?

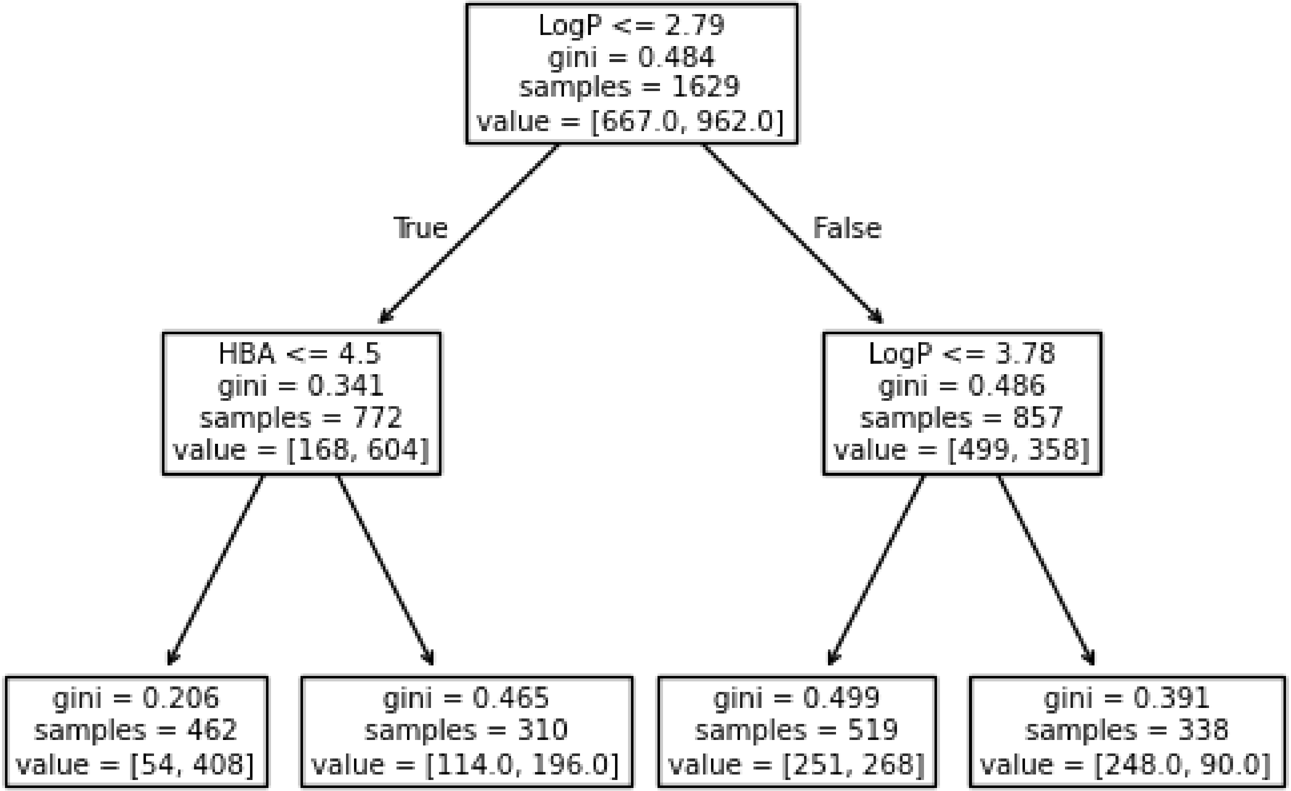


The Gini Index Quantifies the “Purity” of a Split

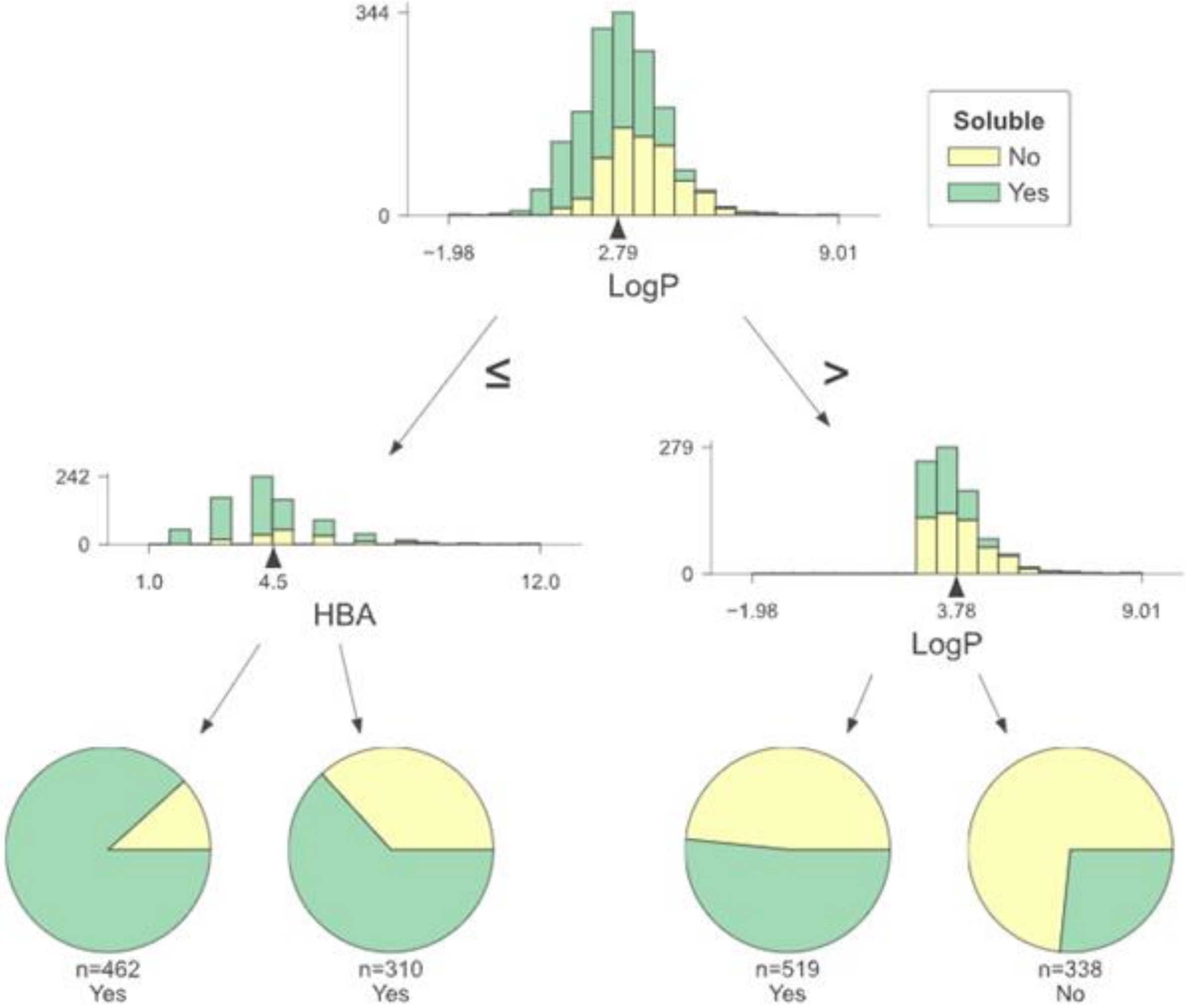
$$\text{Gini Index} = 1 - \sum_{i=1}^{n} (P_i)^2$$



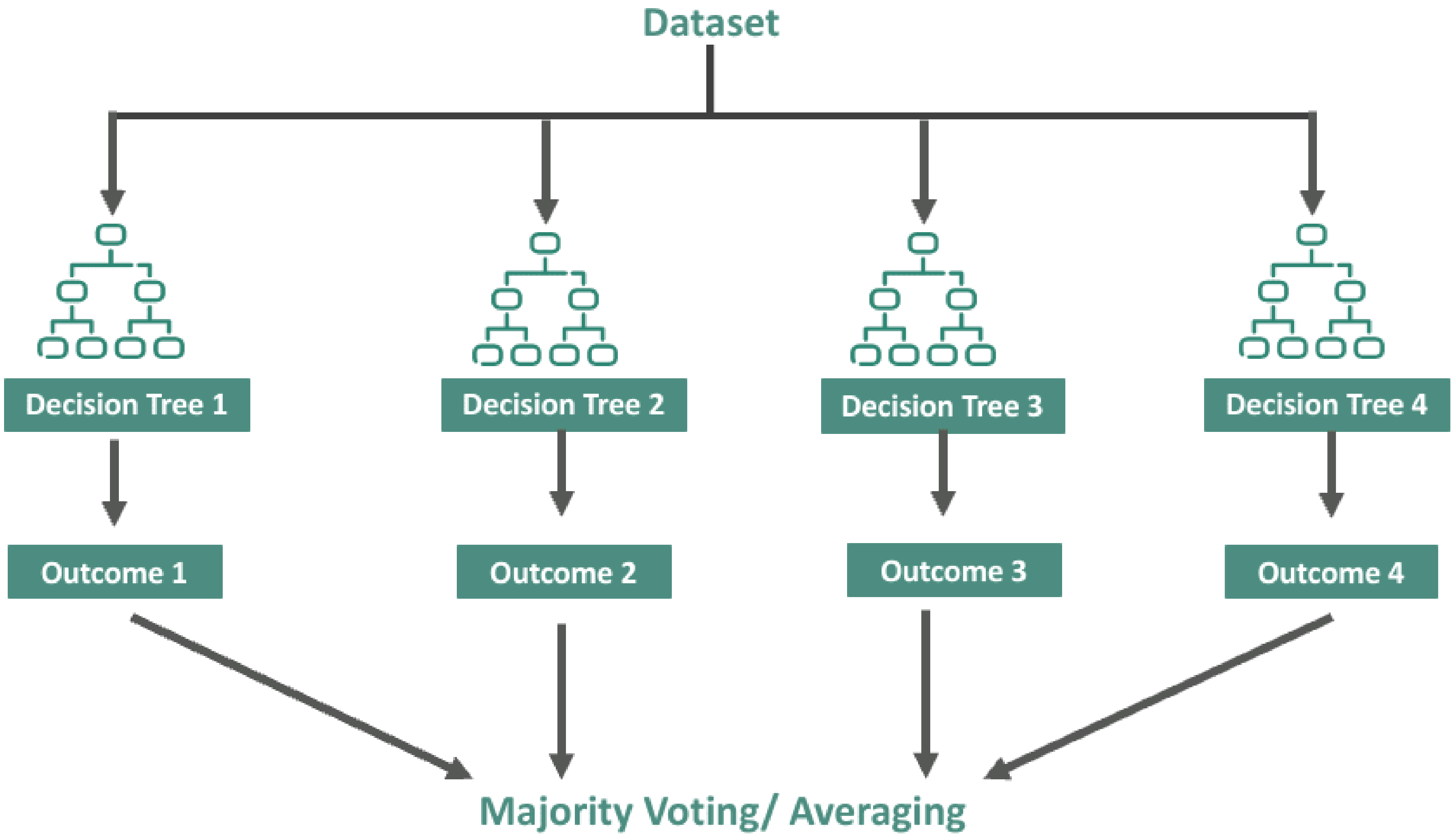
Build a Decision Tree



A Better (IMO) Decision Tree Visualization



Random Forest Uses an Ensemble of Decision Trees



There Are Many Tree Ensemble Methods

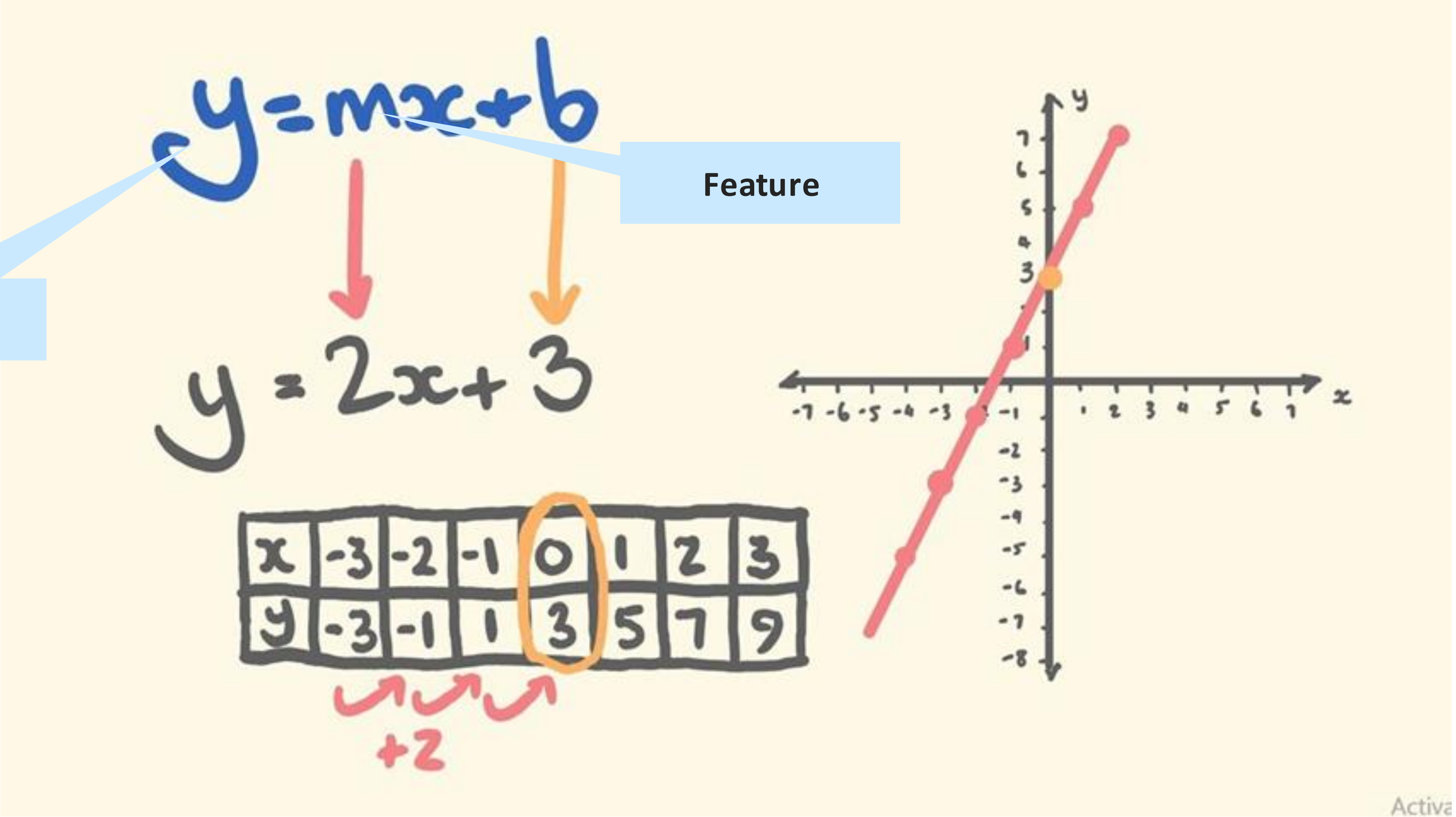
RANDOM FOREST

CatBoost

XGBoost

LightGBM

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



$$y = mx + b$$

$$y = 2x + 3$$

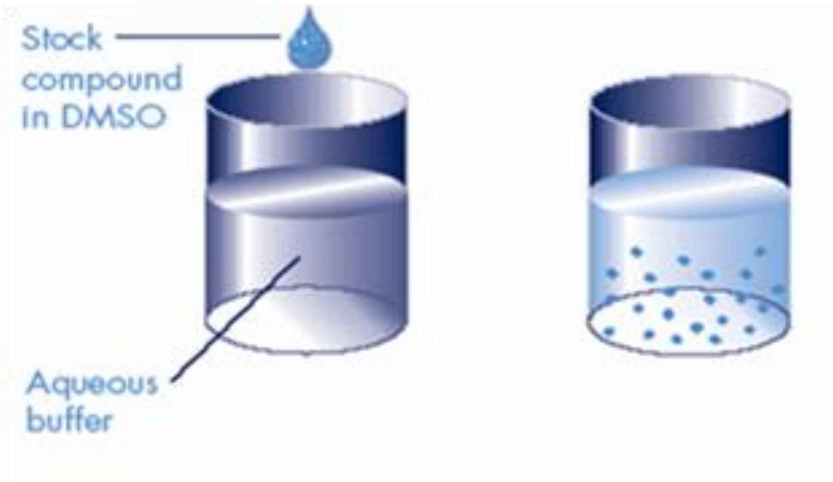
x	-3	-2	-1	0	1	2	3
y	-3	-1	1	3	5	7	9

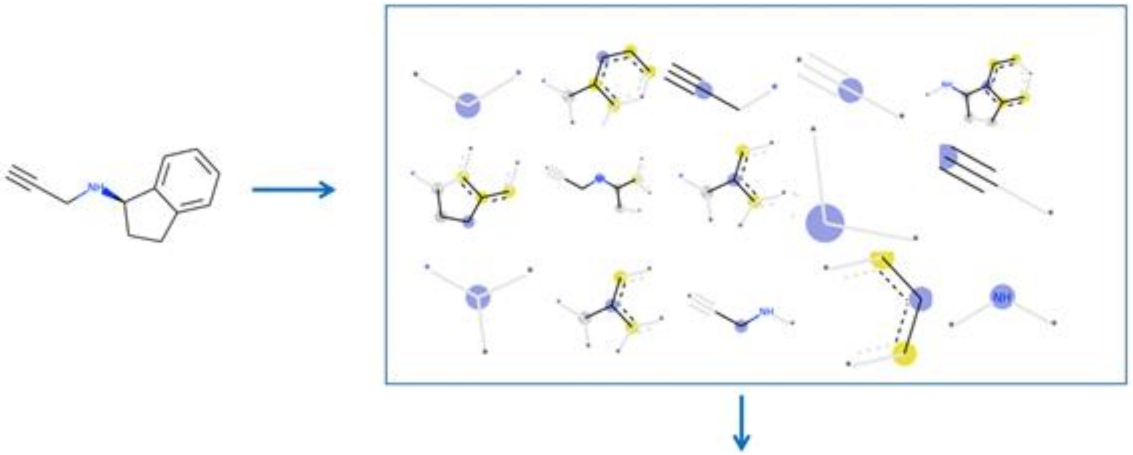
Feature

Label

Activa

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





010110010001000111001000110010101010111

y = Log Aqueous Solubility

X = A Vector Representing a Molecule

Features

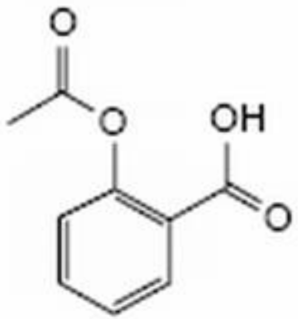
Label

Define Features Based on Books People Have Read

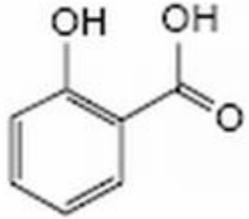
Fred	1	0	0	1	1	1	0	0	1	0	0	1	1	1	0	0
Sally	1	0	1	0	1	1	0	0	0	0	1	1	0	1	0	0
Jane	0	0	1	0	0	0	1	1	0	0	1	1	1	0	0	0

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

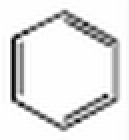
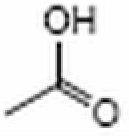
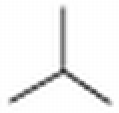
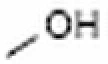
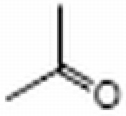
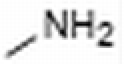
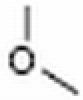
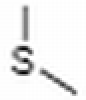
Chemical Fingerprints as Molecular Descriptors



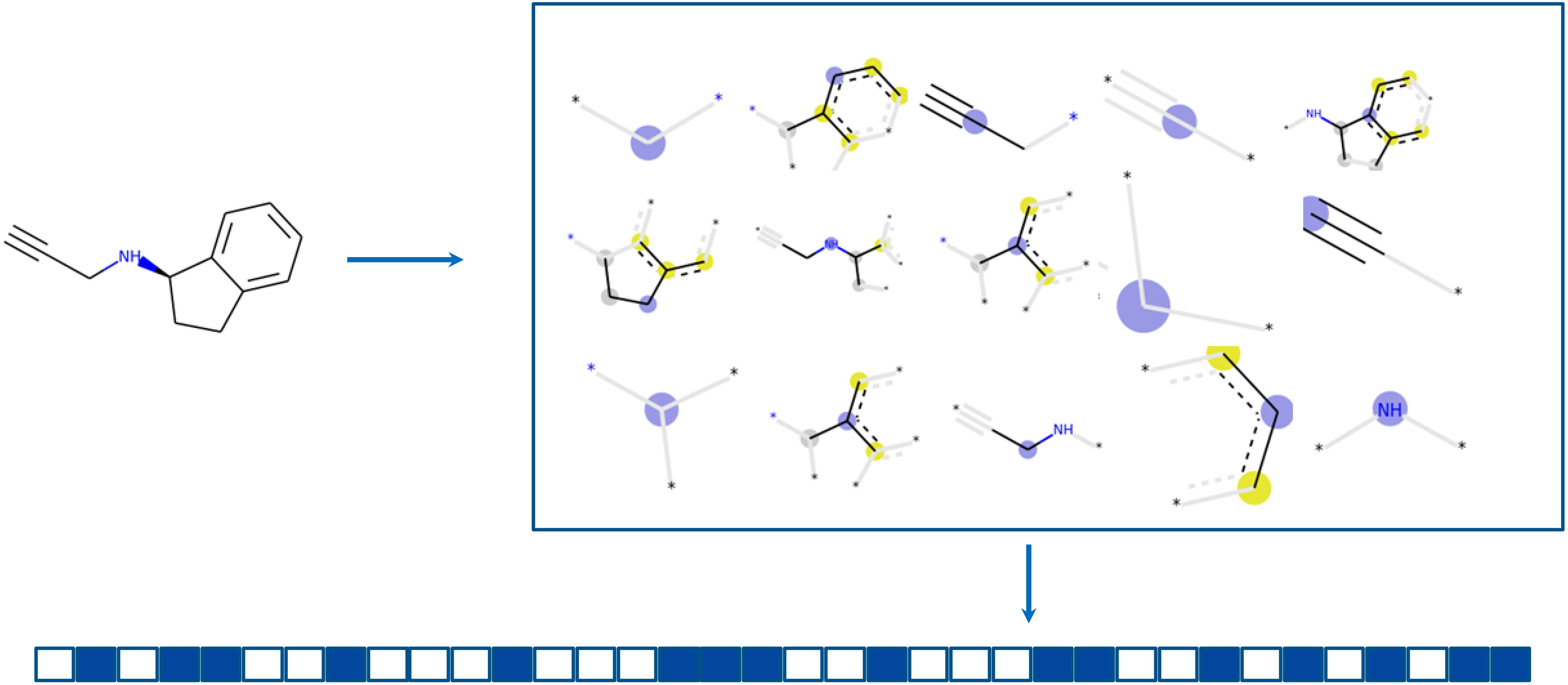
1



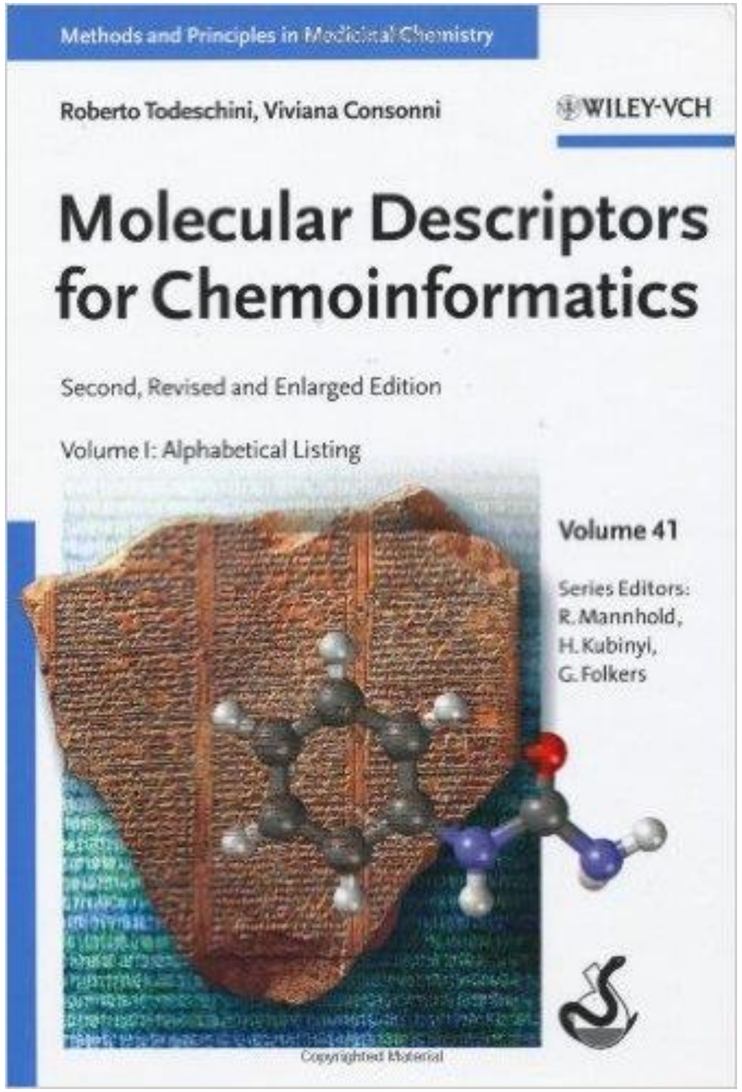
2

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



Methods and Principles in Medicinal Chemistry

Roberto Todeschini, Viviana Consonni

WILEY-VCH

Molecular Descriptors for Chemoinformatics

Second, Revised and Enlarged Edition

Volume I: Alphabetical Listing

Volume 41

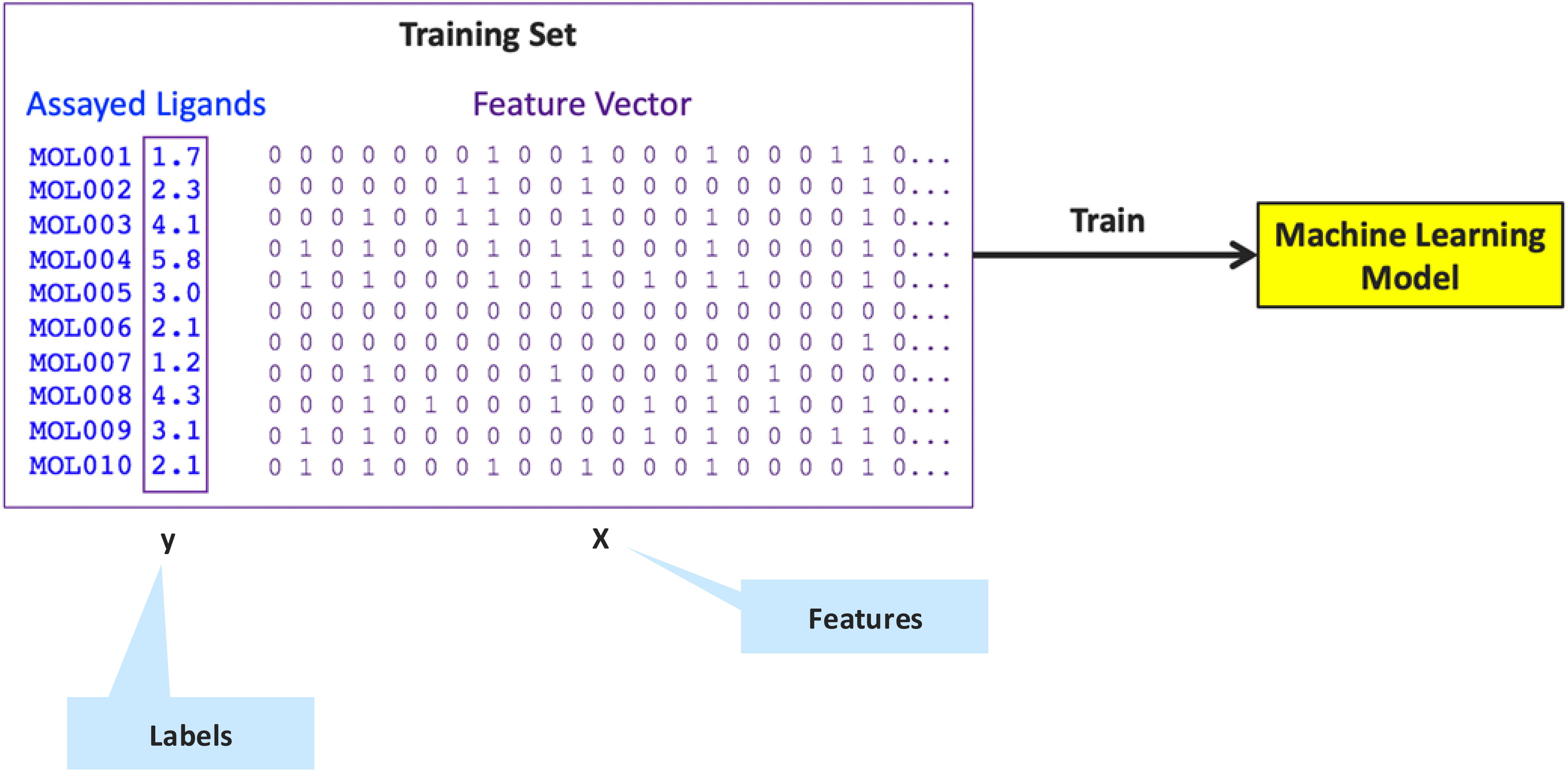
Series Editors:
R. Mannhold,
H. Kubinyi,
G. Folkers

Copyrighted Material

2 Volumes
6,000 references from 450 journals

DRAGON 7 has 5,270 descriptors z

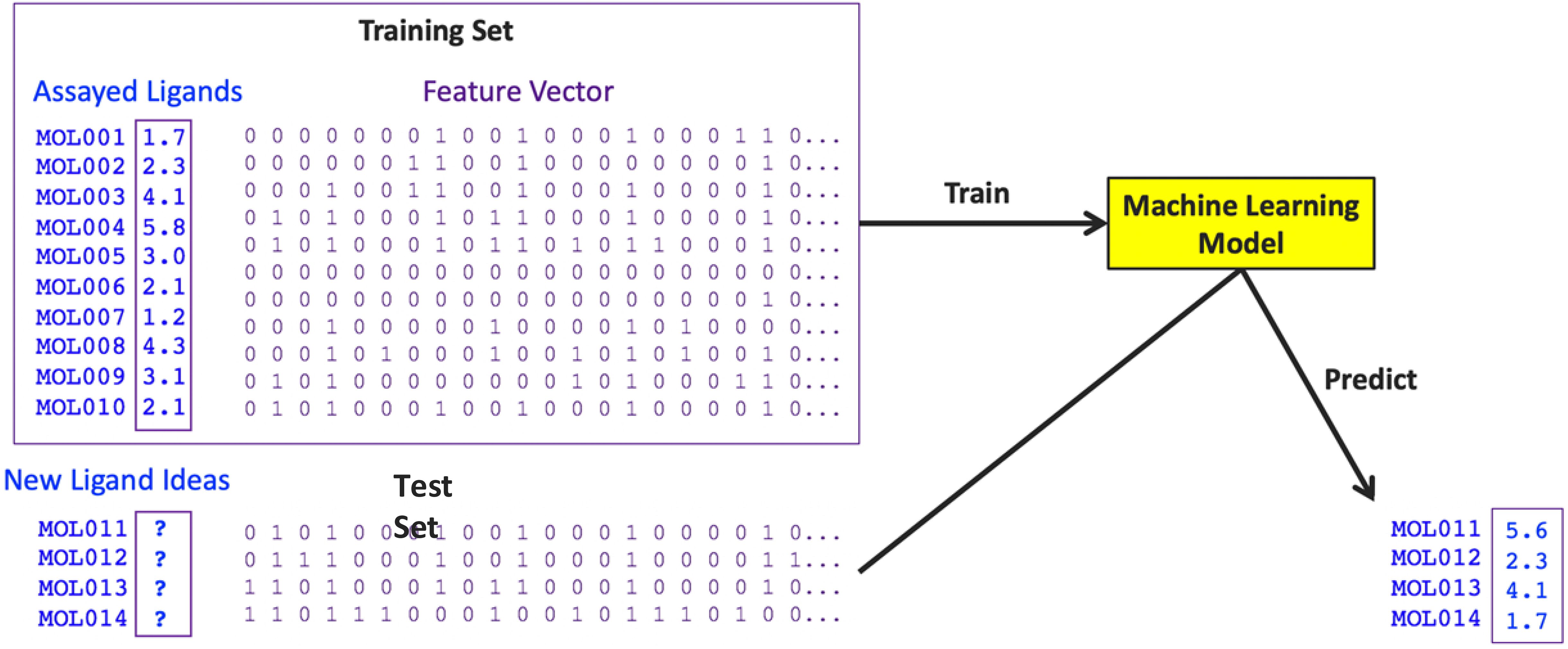
Training a Machine Learning Model



Training Set

Assayed Ligands		Feature Vector
MOL001	1.7	0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...
MOL002	2.3	0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0...
MOL003	4.1	0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 1 0...
MOL004	5.8	0 1 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 1 0...
MOL005	3.0	0 1 0 1 0 0 0 1 0 1 1 0 1 0 1 1 0 0 0 1 0...
MOL006	2.1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0...
MOL007	1.2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0...
MOL008	4.3	0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0...
MOL009	3.1	0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 1 0...
MOL010	2.1	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...

Making Predictions With a Machine Learning Model



Training Set

Assayed Ligands		Feature Vector
MOL001	1.7	0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0...
MOL002	2.3	0 0 0 0 0 0 1 1 0 0 1 0 0 0 0 0 0 0 0 1 0...
MOL003	4.1	0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 0 0 0 0 1 0...
MOL004	5.8	0 1 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 1 0...
MOL005	3.0	0 1 0 1 0 0 0 1 0 1 1 0 1 0 1 1 0 0 0 1 0...
MOL006	2.1	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0...
MOL007	1.2	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0...
MOL008	4.3	0 0 0 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 0 0 0...
MOL009	3.1	0 0 0 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 1 0...
MOL010	2.1	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...

Train

Machine Learning Model

Predict

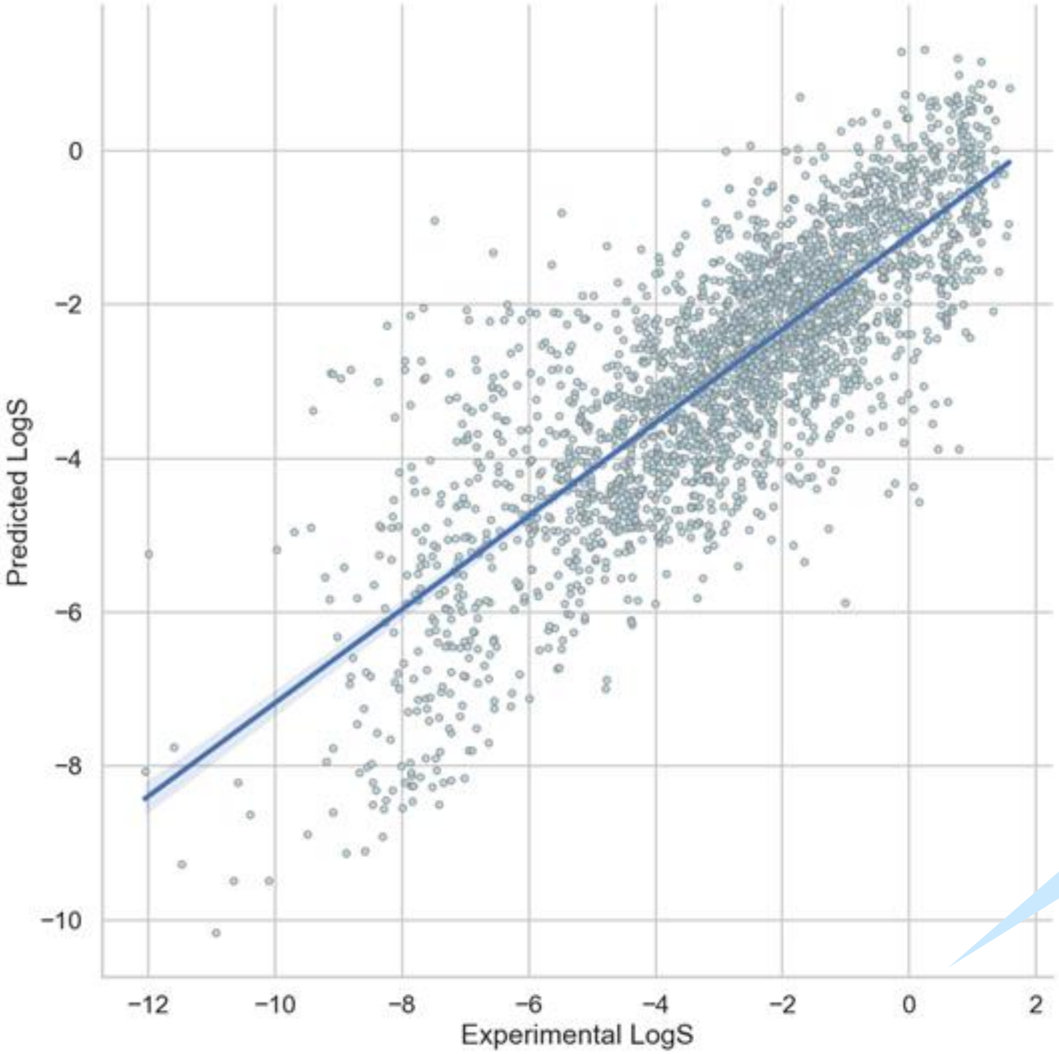
New Ligand Ideas

Test Set

MOL011	?	0 1 0 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 0...
MOL012	?	0 1 1 1 0 0 0 1 0 0 1 0 0 0 1 0 0 0 0 1 1...
MOL013	?	1 1 0 1 0 0 0 1 0 1 1 0 0 0 1 0 0 0 0 1 0...
MOL014	?	1 1 0 1 1 1 0 0 0 1 0 0 1 0 1 1 1 0 1 0 0...

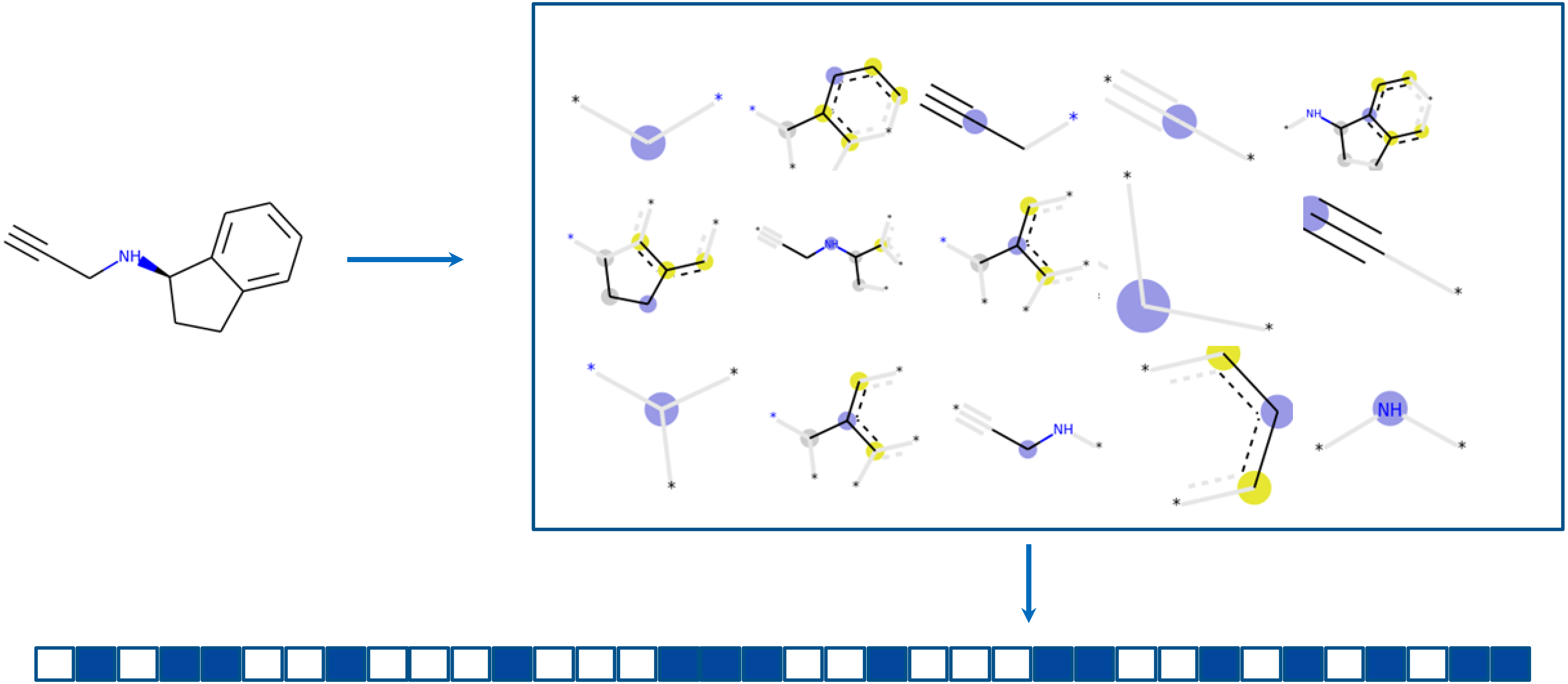
MOL011	5.6
MOL012	2.3
MOL013	4.1
MOL014	1.7

Prediction Performance of an Aqueous Solubility Model

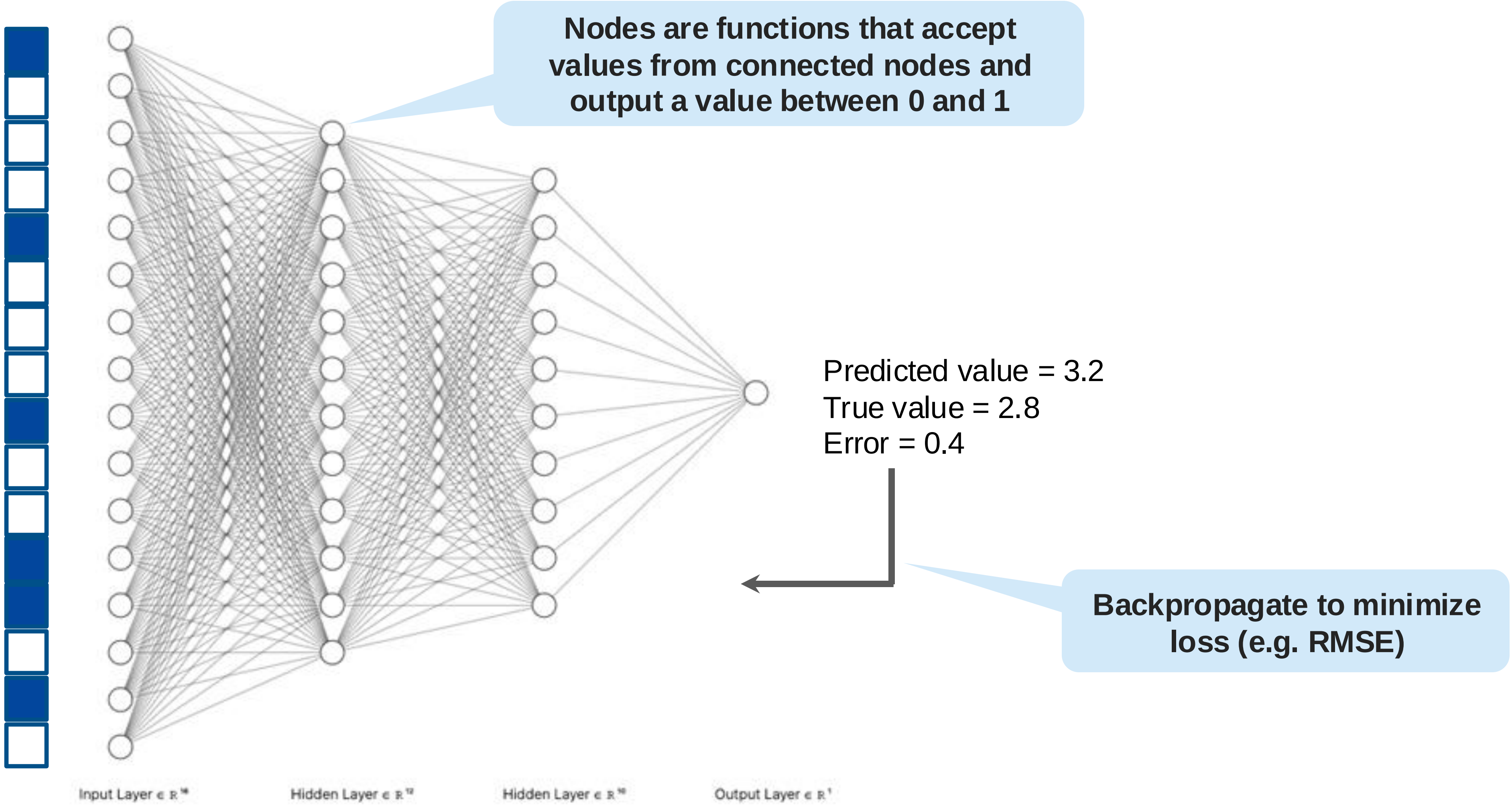


Dynamic range is very large, does not represent a typical use case

Chemical Fingerprints as Vector Representations of Molecules



Neural Networks Adjust Weights to Minimize a Loss Function



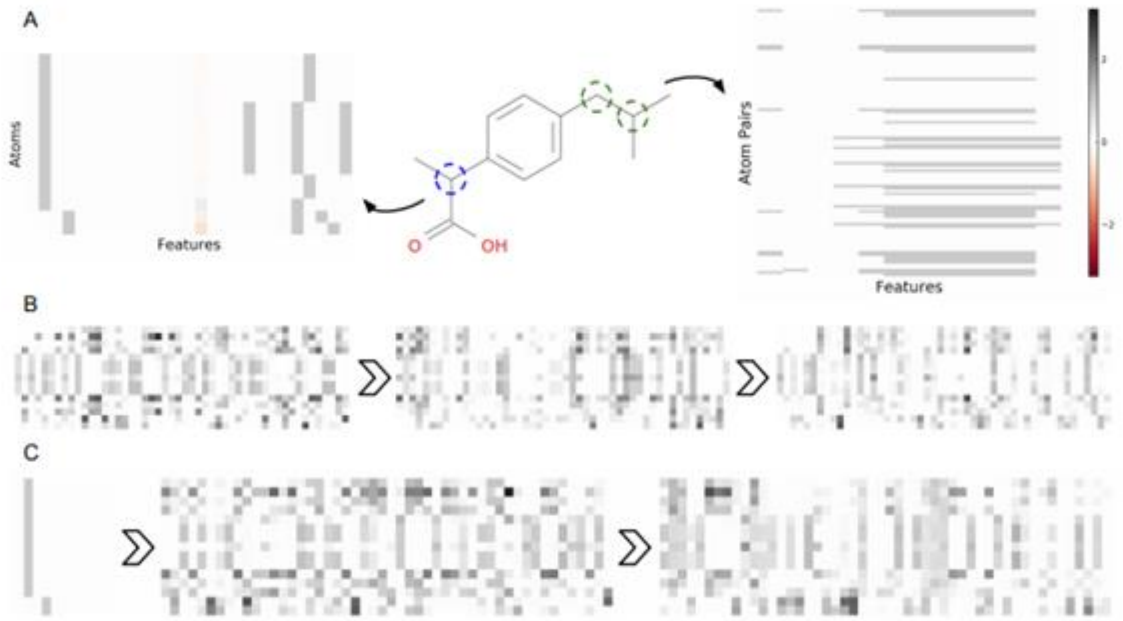
Input Layer ∈ $\mathbb{R}^{16}$
Hidden Layer ∈ $\mathbb{R}^{12}$
Hidden Layer ∈ $\mathbb{R}^{10}$
Output Layer ∈ $\mathbb{R}^{1}$

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

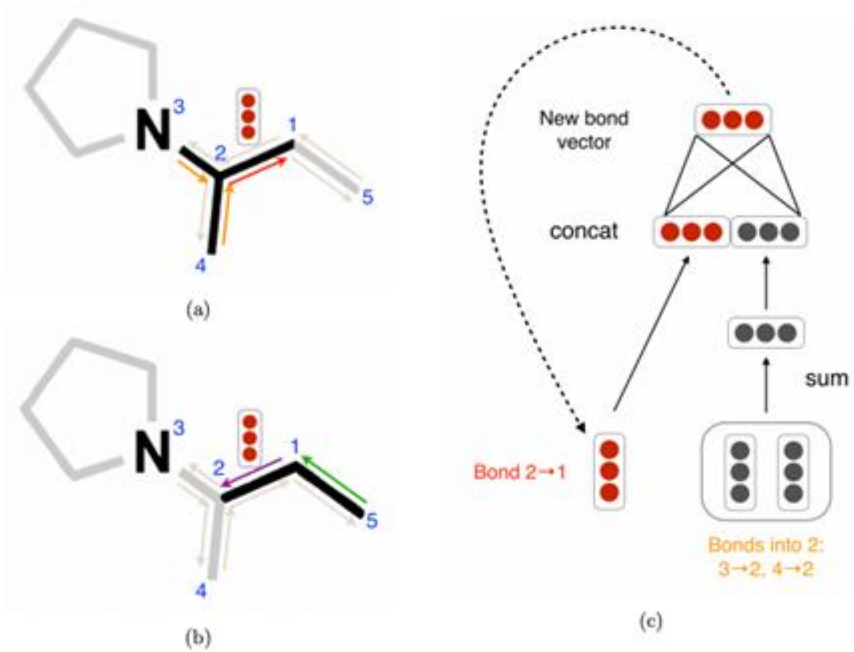
Backpropagate to minimize loss (e.g. RMSE)

Using Neural Networks to Create New Molecular Representations



Graph Convolutions

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

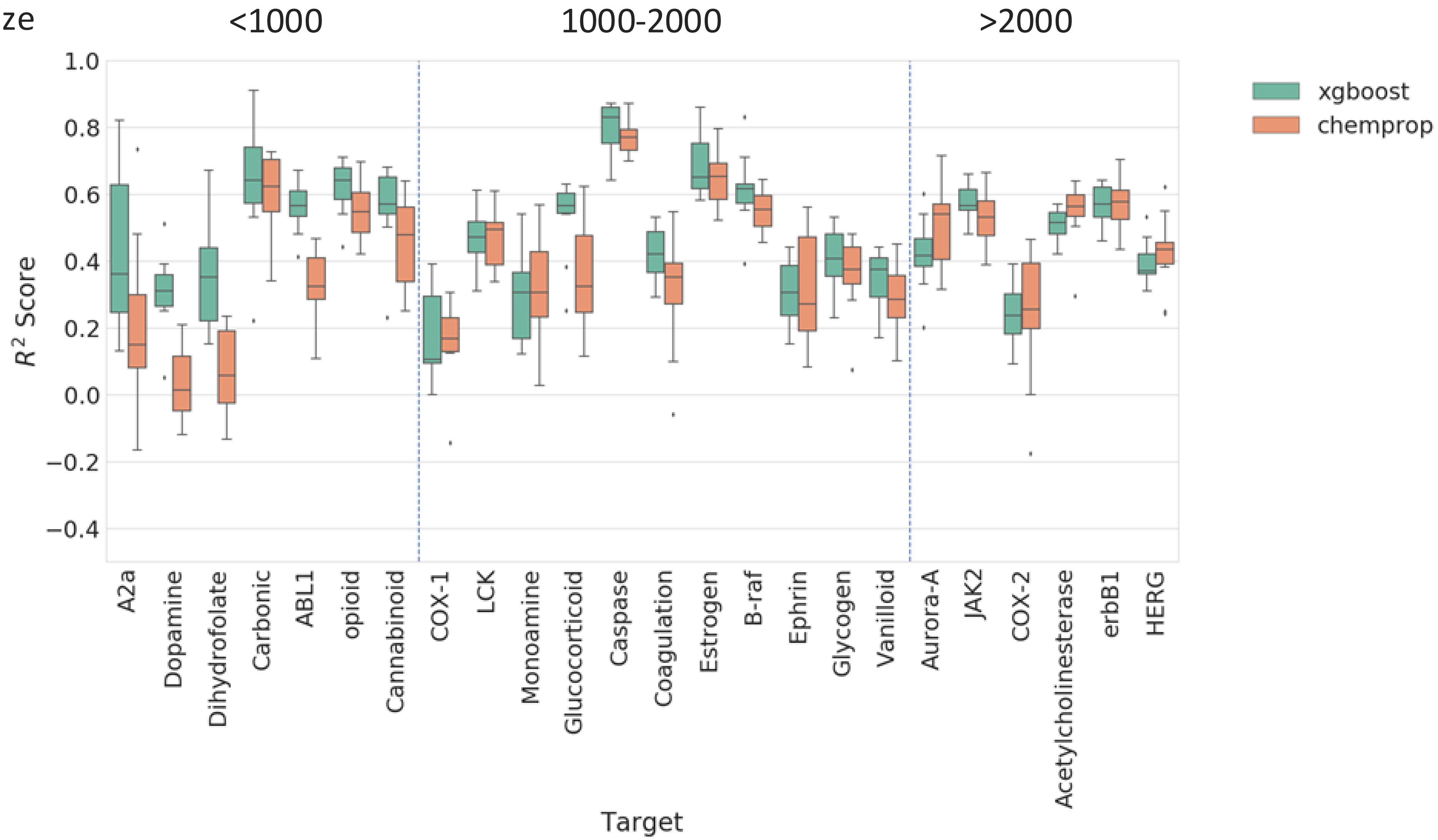


Message Passing Neural Network

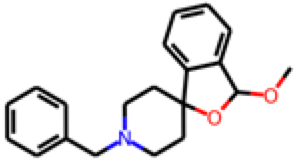
J. Chem. Inf. Model. 2019, 59, 3370−3388

Are Neural Network Representations Better?

Dataset Size



Incorporating 3D into Molecular Machine Learning is an Unsolved Problem

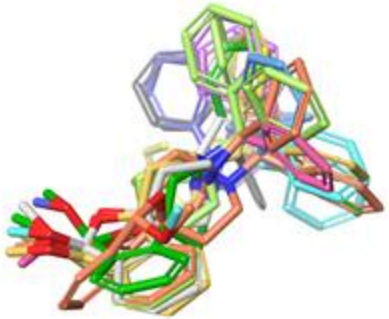


2D single-instance

→ 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0... → 5.21 ✓

3D single-instance

→ 0 0 0 0 0 0 0 1 0 0 1 0 0 0 1 0 0 0 1 1 0... → 5.21 ✗



3D multiple-instance

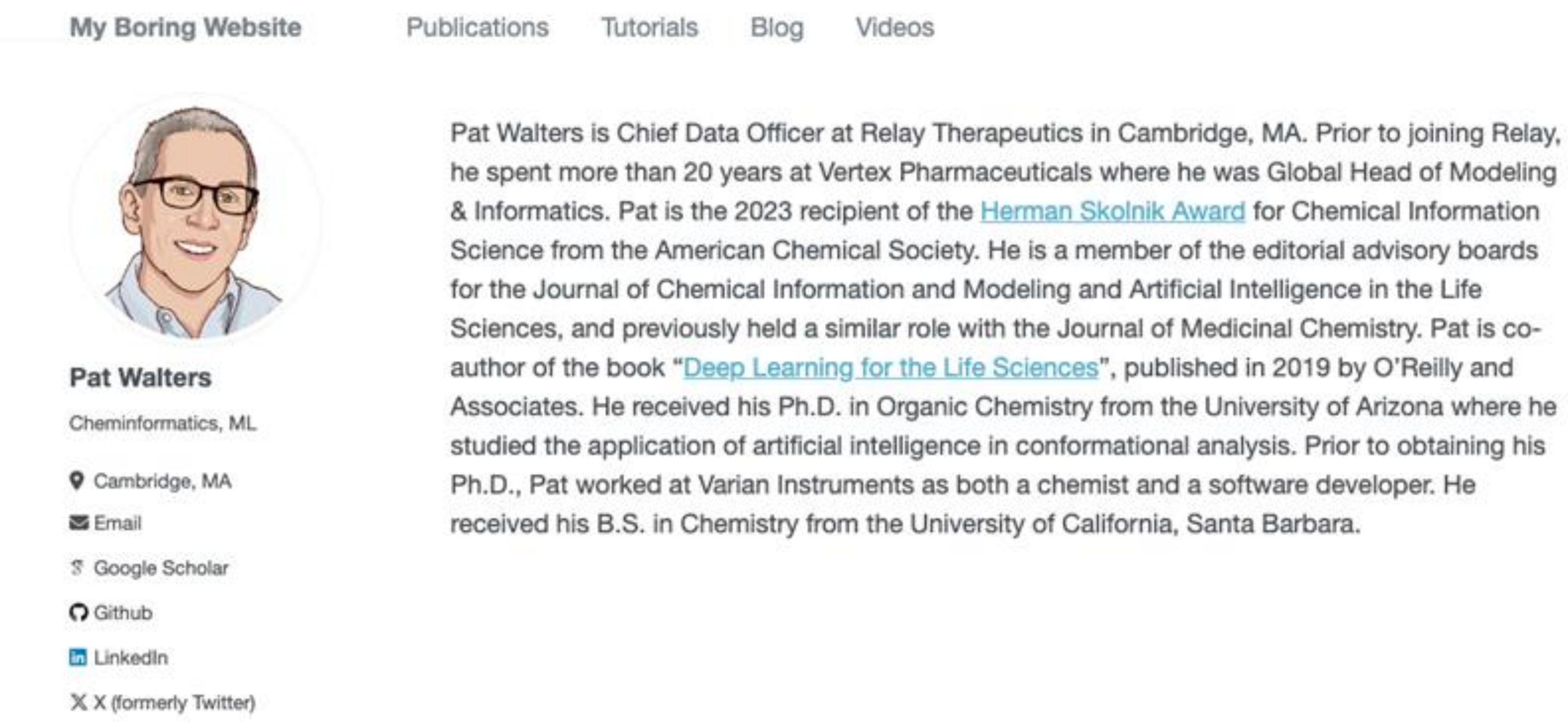
→

```
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 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

How to Find Me

https://pwalters.github.io



Pat Walters

Cheminformatics, ML

- Cambridge, MA
- Email
- Google Scholar
- Github
- LinkedIn
- 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.