

# Machine Learning: A Starter Kit

Aditya Sonpal

CHE 596

# What the heck is this?

- AI: any technology which appears to do something smart, or mimics Human behavior.
- ML: a specific kind of AI but rather than a rule-based approach, the system learns how to do something by finding patterns in examples



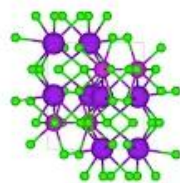
# AI AI Everywhere!

- Why? Abundance of data and computational power.
- image generated by AI

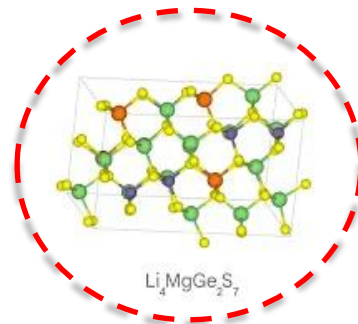


# AI AI Everywhere!

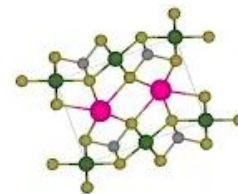
- Why? Abundance of data and computational power.
- Molecules generated by AI
  - [newly discovered crystals by Google's GNoME tool](#)



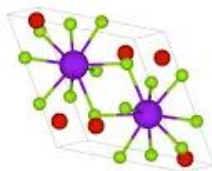
K<sub>2</sub>BiCl<sub>5</sub>



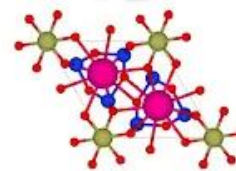
Li<sub>4</sub>MgGe<sub>2</sub>S<sub>7</sub>



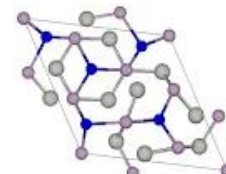
Mo<sub>5</sub>GeB<sub>2</sub>



KV<sub>3</sub>Se<sub>3</sub>



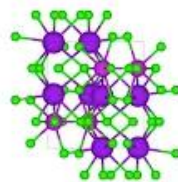
Rb<sub>2</sub>HfSi<sub>3</sub>O<sub>9</sub>



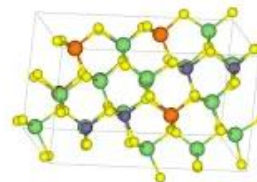
Tm<sub>5</sub>Pd<sub>9</sub>P<sub>7</sub>

# AI AI Everywhere!

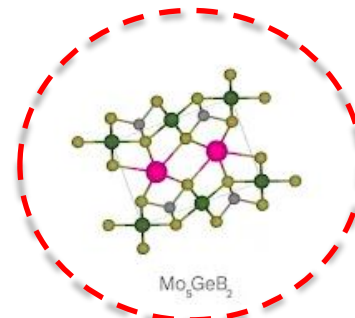
- Why? Abundance of data and computational power.
- Molecules generated by AI
  - [newly discovered crystals by Google's GNoME tool](#)



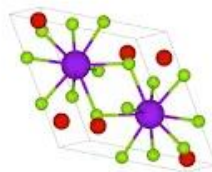
K<sub>2</sub>BiCl<sub>5</sub>



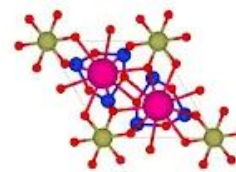
Li<sub>4</sub>MgGe<sub>2</sub>S<sub>7</sub>



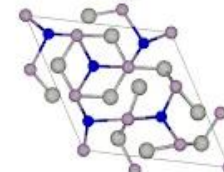
Mo<sub>5</sub>GeB<sub>2</sub>



KV<sub>3</sub>Se<sub>5</sub>



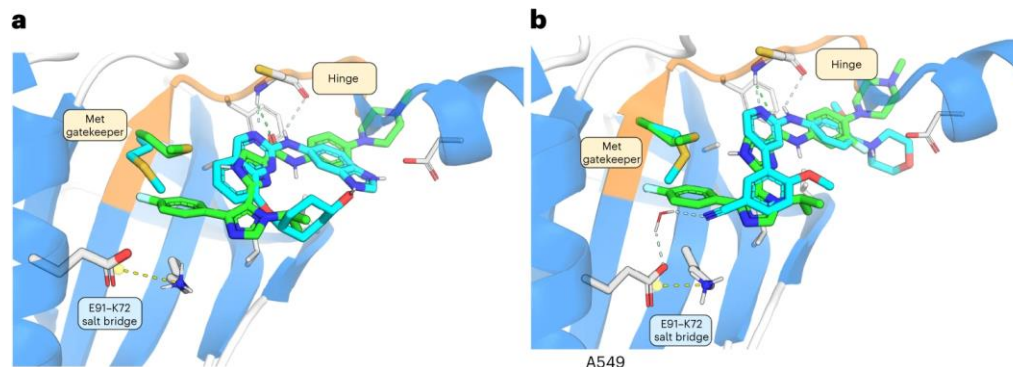
Rb<sub>2</sub>HfSi<sub>3</sub>O<sub>9</sub>



Tm<sub>5</sub>Pd<sub>9</sub>P<sub>7</sub>

# AI AI Everywhere!

- Why? Abundance of data and computational power.
- Molecules generated by AI
  - [newly discovered crystals by Google's GNoME tool](#)
  - [Insilico's new AI generated drug to treat Idiopathic pulmonary fibrosis](#)

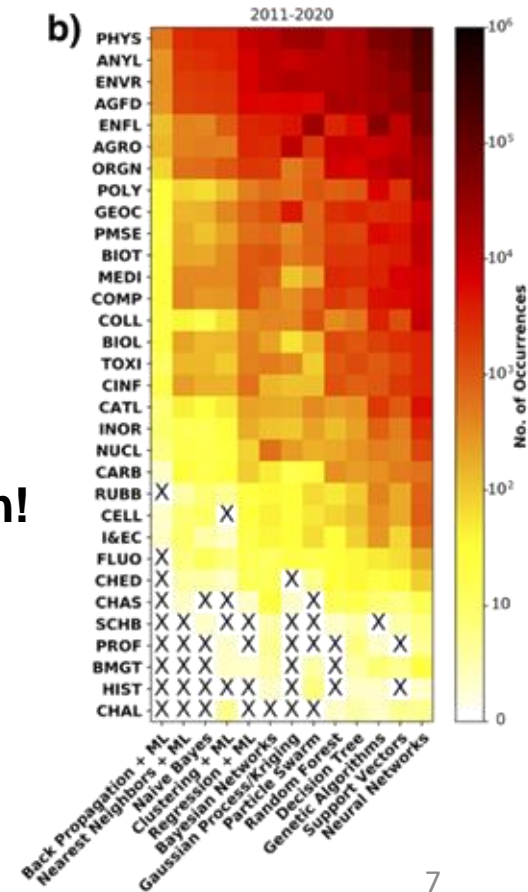


Source: <https://www.nature.com/articles/s41587-024-02143-0>

# Applications of ML in Chem

- Mapping molecular structure to target property
  - predict properties
  - categorize and cluster molecules
  - design new molecules

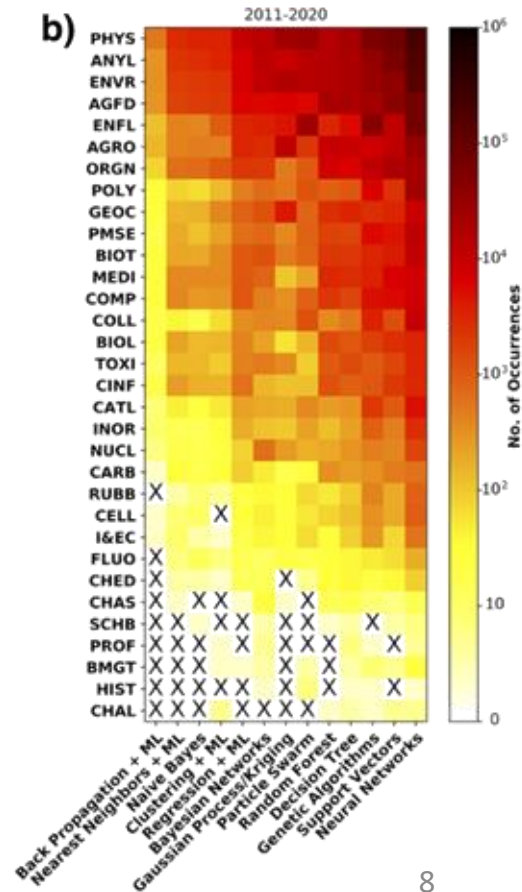
**Invaluable in drug and materials discovery and design!**



# Applications of ML in Chem

- Mapping molecular structure to target property
- Accelerating computational chemistry
  - predicting forcefield parameters
  - predicting DFT level energies

<https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107>

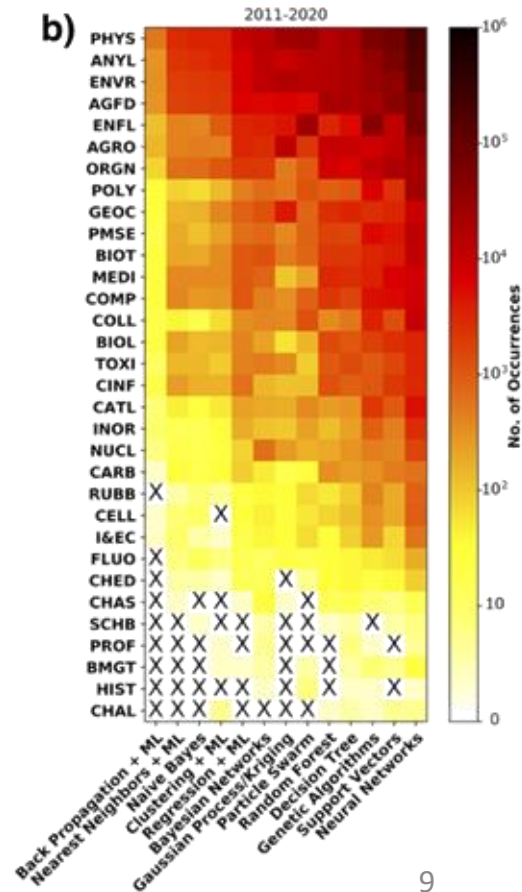




# Applications of ML in Chem

- Mapping molecular structure to target property
- Accelerating computational chemistry
- With experiments:
  - analyze spectroscopic data
  - predict phase maps
  - predict reaction pathways
  - select experimental candidates
  - retrosynthesis: viable synthetic routes for organic compounds

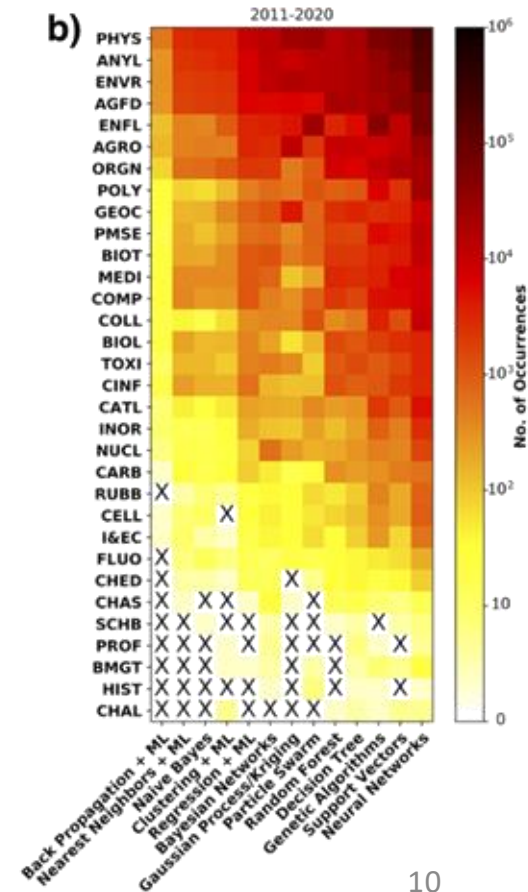
<https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107>



# Applications of ML in Chem

- Mapping molecular structure to target property
- Accelerating computational chemistry
- With experiments:
- Crazy ones: 🦇
  - language models to search literature for experimental parameters
  - robots to perform experiments
  - robots to enhance decision making in experiments

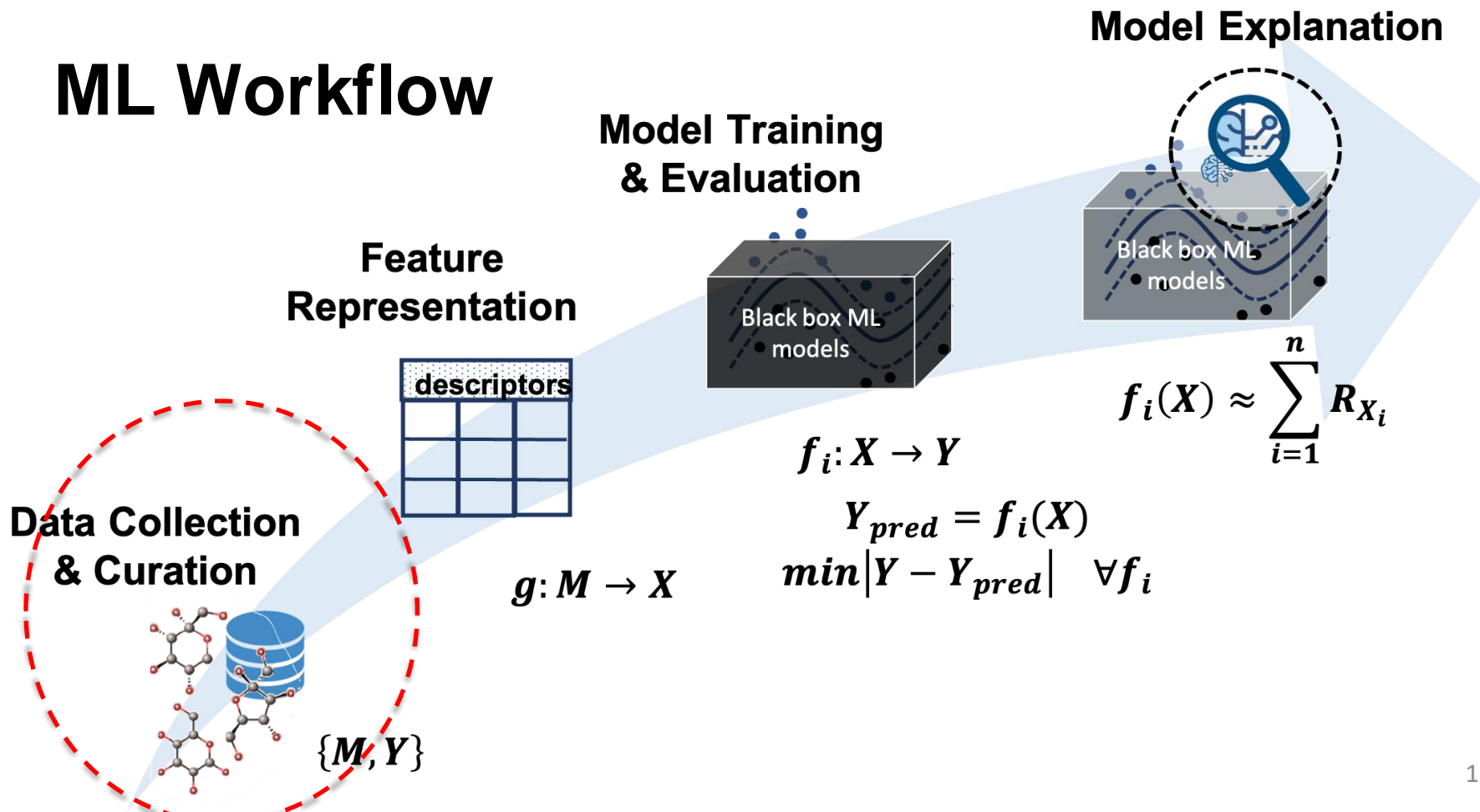
**This is not an exhaustive list!**



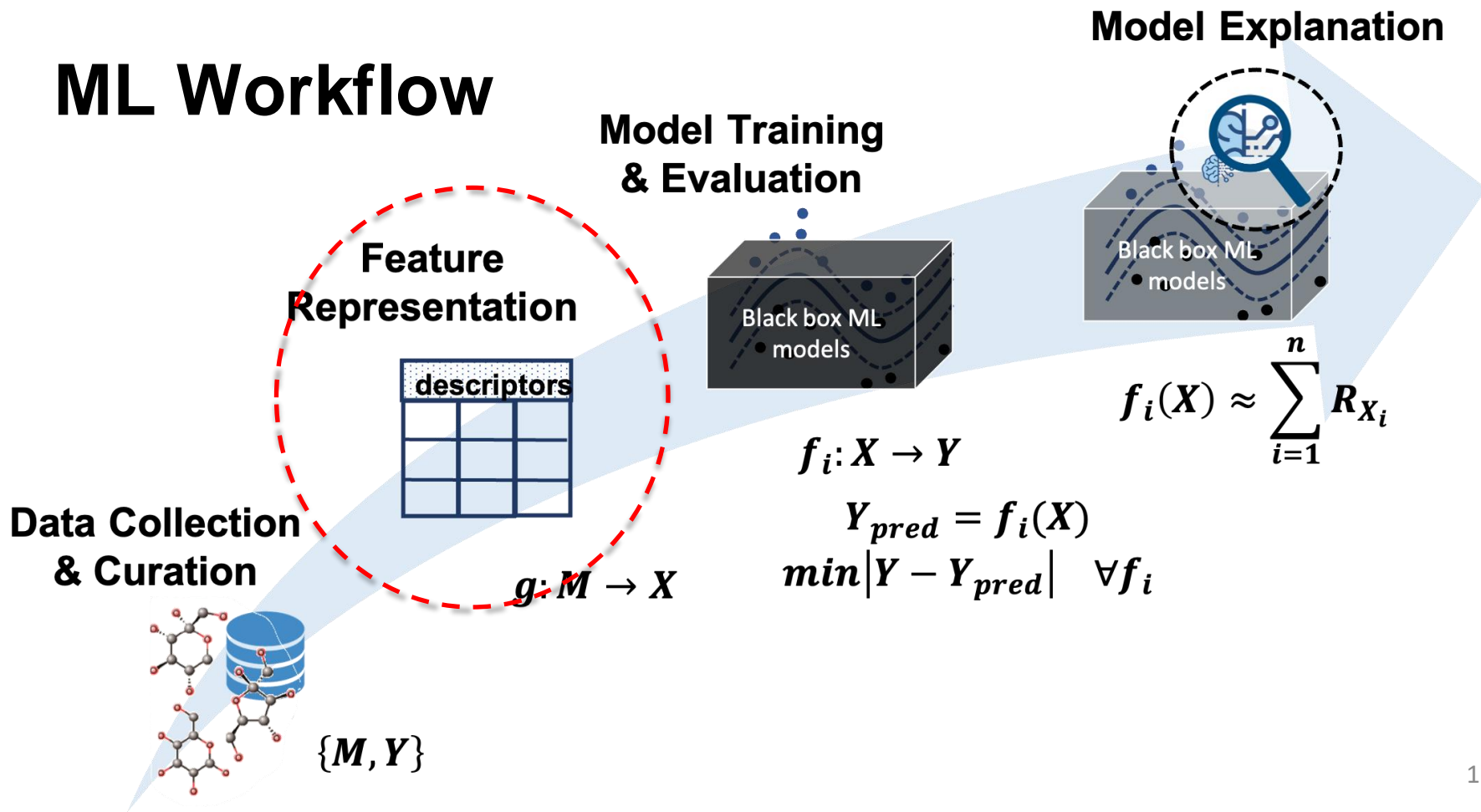
Breathe?



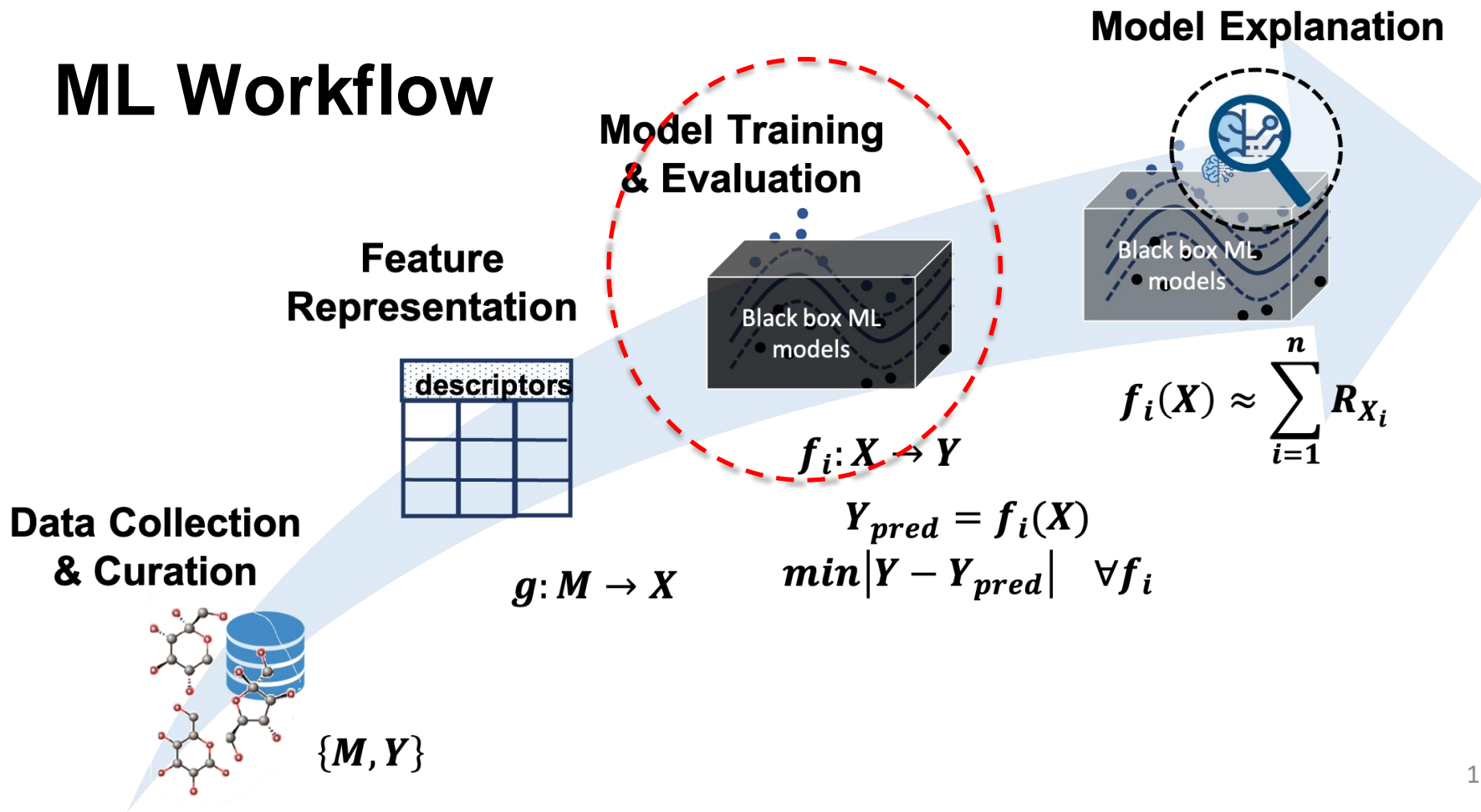
# ML Workflow



# ML Workflow

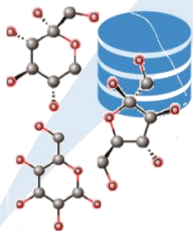


# ML Workflow



# ML Workflow

Data Collection  
& Curation



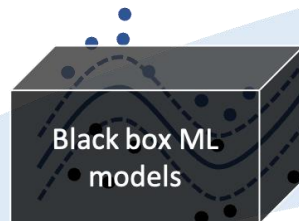
$\{M, Y\}$

Feature  
Representation

descriptors		

$$g: M \rightarrow X$$

Model Training  
& Evaluation

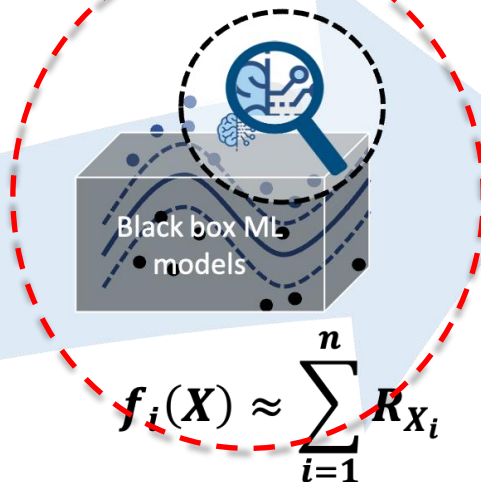


$$f_i: X \rightarrow Y$$

$$Y_{pred} = f_i(X)$$

$$\min |Y - Y_{pred}| \quad \forall f_i$$

Model Explanation



$$f_i(X) \approx \sum_{i=1}^n R_{X_i}$$

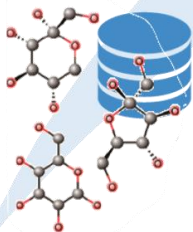


## Data

Model Training  
& Evaluation

## Model Explanation

## Feature Representation

Data Collection  
& Curation $\{M, Y\}$ 

$$g: M \rightarrow X$$

**Table 1**  
A list of popular chemical databases commonly used in ML.

Classification	Name	Content	URL
Chemical reaction databases	SciFinder	Information on chemical compounds, bibliographic data, and chemical reactions (commercial database)	<a href="https://scifinder.cas.org/">https://scifinder.cas.org/</a>
	Reaxys	Chemical reaction and bibliographic information (commercial database)	<a href="https://www.reaxys.com/">https://www.reaxys.com/</a>
	USPTO	Chemical structure and reaction	<a href="https://www.repository.cam.ac.uk/handle/1810/244727">https://www.repository.cam.ac.uk/handle/1810/244727</a>
	ORD	Organic chemical reaction data	<a href="https://github.com/open-reaction-database">https://github.com/open-reaction-database</a>
	NextMove	Chemical reaction data	<a href="https://www.nextmovesoftware.com/about.html">https://www.nextmovesoftware.com/about.html</a>
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	<a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>
	NIST	Standard physicochemical properties of compounds	<a href="https://webbook.nist.gov/chemistry/">https://webbook.nist.gov/chemistry/</a>
	ChemSpider	Structure and property of compounds	<a href="https://www.chemspider.com">https://www.chemspider.com</a>
	ChEMBL	Drug-like properties of bioactive molecules	<a href="https://www.ebi.ac.uk/chembl">https://www.ebi.ac.uk/chembl</a>
	DrugBank	Properties of drug molecules	<a href="https://go.drugbank.com/releases/latest">https://go.drugbank.com/releases/latest</a>
Material databases	Tox21	Toxic effects of substances	<a href="https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html">https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html</a>
	ESOL	Water solubility of compounds	<a href="https://doi.org/10.1021/ci034243x">https://doi.org/10.1021/ci034243x</a>
	FreeSolv	Water solubility of small neutral molecules	<a href="https://github.com/MobleyLab/FreeSolv">https://github.com/MobleyLab/FreeSolv</a>
	Lipophilicity	Lipid solubility of organic compounds	<a href="https://doi.org/10.1002/cem.2718">https://doi.org/10.1002/cem.2718</a>
	CSD	Organic and metal-organic crystal structures	<a href="https://www.ccdc.cam.ac.uk/">https://www.ccdc.cam.ac.uk/</a>
Computational chemistry databases	ICSD	Inorganic and metal-organic crystal structures	<a href="https://icdd.products.fiz-karlsruhe.de/">https://icdd.products.fiz-karlsruhe.de/</a>
	PDF	Diffraction data of inorganic and organic compounds	<a href="https://www.icdd.com/pdfssearch/">https://www.icdd.com/pdfssearch/</a>
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	<a href="https://matweb.com/">https://matweb.com/</a>
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	<a href="https://data.nasa.gov/dataset/Li-Ion-Battery-Aging-Datasets/u5r-zjdb">https://data.nasa.gov/dataset/Li-Ion-Battery-Aging-Datasets/u5r-zjdb</a>
	HTeM	Experimental information of inorganic thin-film materials	<a href="https://nrel.gov/">https://nrel.gov/</a>
Computational chemistry databases	CDB-17	Structures of organic molecules up to 17 atoms	<a href="https://www.gdb.unibe.ch/downloads/">https://www.gdb.unibe.ch/downloads/</a>
	QM9	Quantum chemical properties of organic molecules	<a href="https://quantum-machine.org/datasets/">https://quantum-machine.org/datasets/</a>
	ANI-1	Energy and force of non-equilibrium molecules	<a href="https://github.com/isaeye/ANI1_dataset">https://github.com/isaeye/ANI1_dataset</a>
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://next-gen.materialsproject.org/">https://next-gen.materialsproject.org/</a>
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://oqmd.org/">https://oqmd.org/</a>
Computational chemistry databases	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://aflowlib.org/">https://aflowlib.org/</a>
	MD17/ISO-17	Energy and force of non-equilibrium molecules	<a href="https://quantum-machine.org/datasets/">https://quantum-machine.org/datasets/</a>
	LASP	Global PES dataset of molecules/materials	<a href="https://www.lasphub.com">https://www.lasphub.com</a>
	OC20	Adsorption energy of molecules in catalysts	<a href="https://open-catalyst-project.org/">https://open-catalyst-project.org/</a>
	Atom3D	3D structure of molecules, RNA, and proteins	<a href="https://www.atom3d.ai/">https://www.atom3d.ai/</a>

URL: uniform resource locator; USPTO: United States Patent and Trademark Office; ORD: Open Reaction Database; NIST: National Institute of Standards and Technology; CSD: Cambridge Structural Database; ICSD: Inorganic Crystal Structure Database; PDF: Powder Diffraction File; HTeM: High-Throughput Experimental Materials; OQMD: Open Quantum Materials Database; OC20: Open Catalyst 2020; DFT: density functional theory; PES: potential energy surface.



# Demo

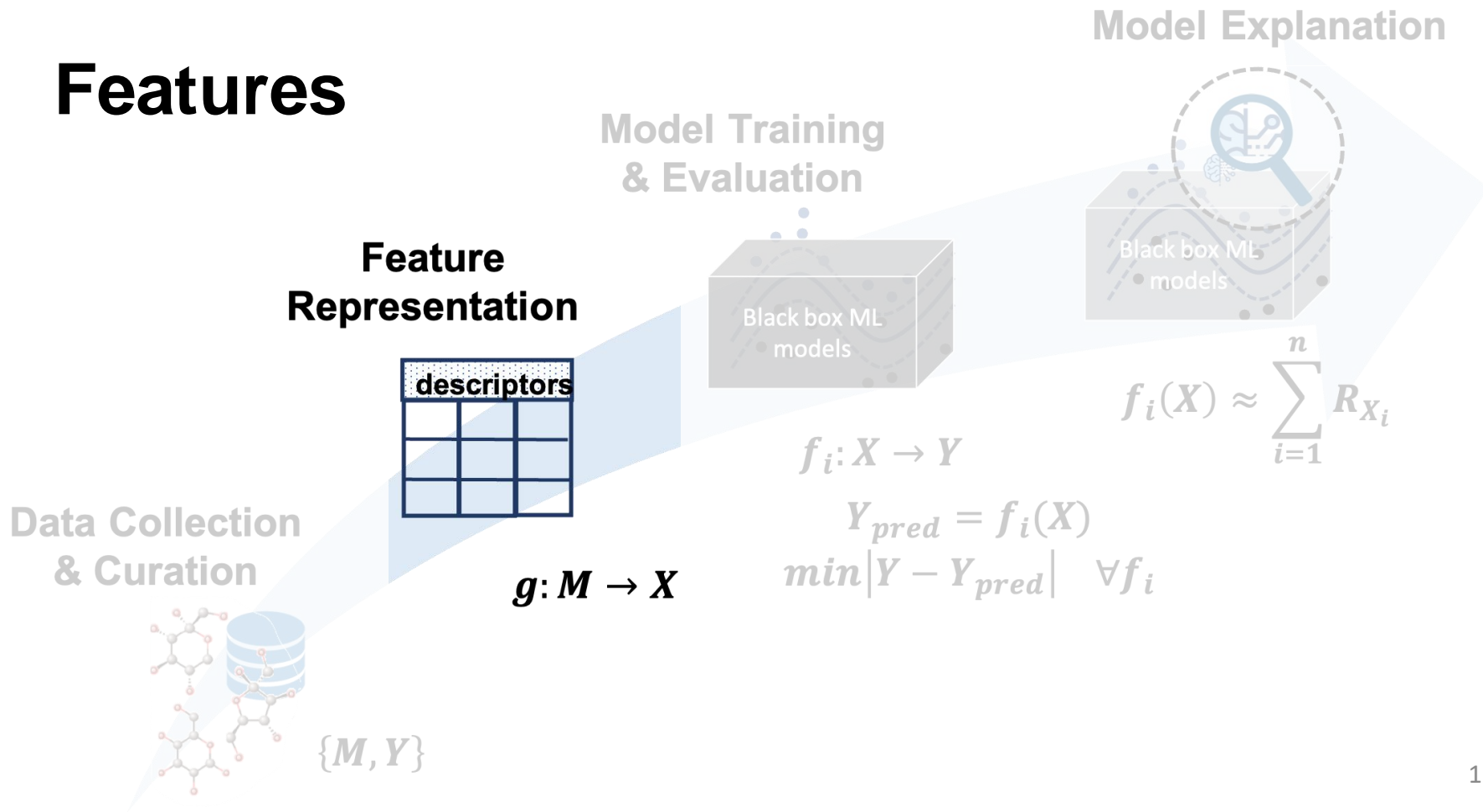
# Data Collection & Curation

- Essential practices:
  - identify & document source, context, and limitations of data (biases, domain of applicability)
  - data cleaning – duplicates, missing values, identifying unphysical values, unit conversion, homogenization, normalization
  - document data cleaning steps comprehensively
  - adhering to **F**indable **A**ccessible **I**nteroperable and **R**eusable (**FAIR**) principles of scientific data management



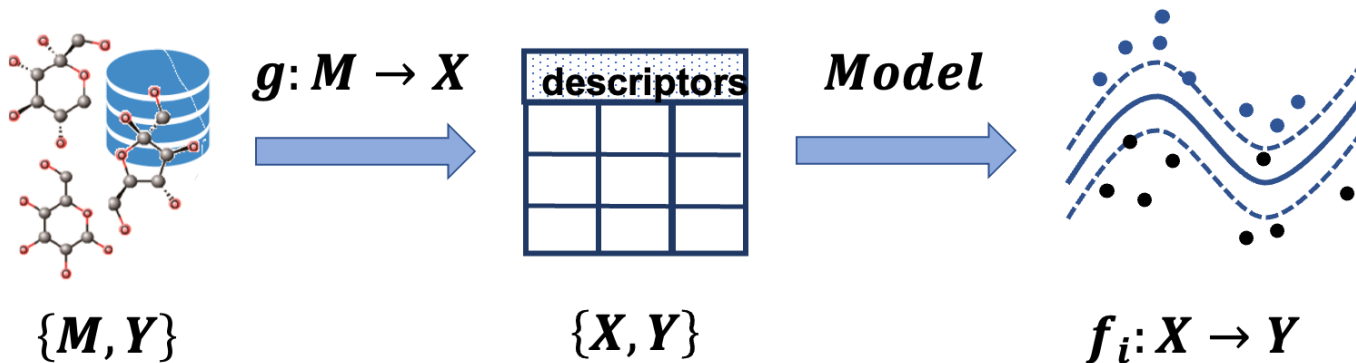
**Data is King! Your model is only as good as the data you provide it!**

# Features



# Feature Representation

- function  $g: M \rightarrow X$  converts a chemical representation  $m \in M$  to a feature input  $x \in X$
- $m$  can be SMILES, InChI, XYZ, CIF, PDB, etc.
- $x$  is a numeric or vector representation of  $m$  (descriptors, fingerprints, coulomb matrices, learned features, text-based, etc.)

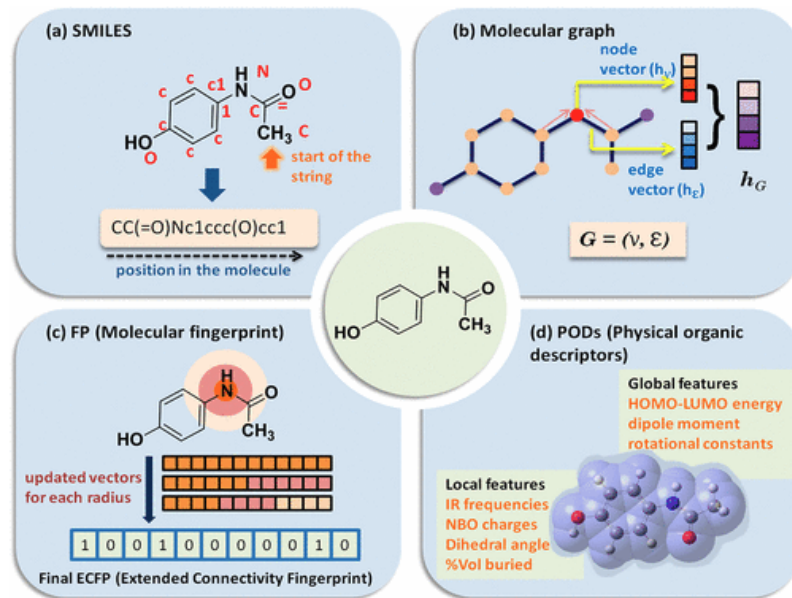


# Demo

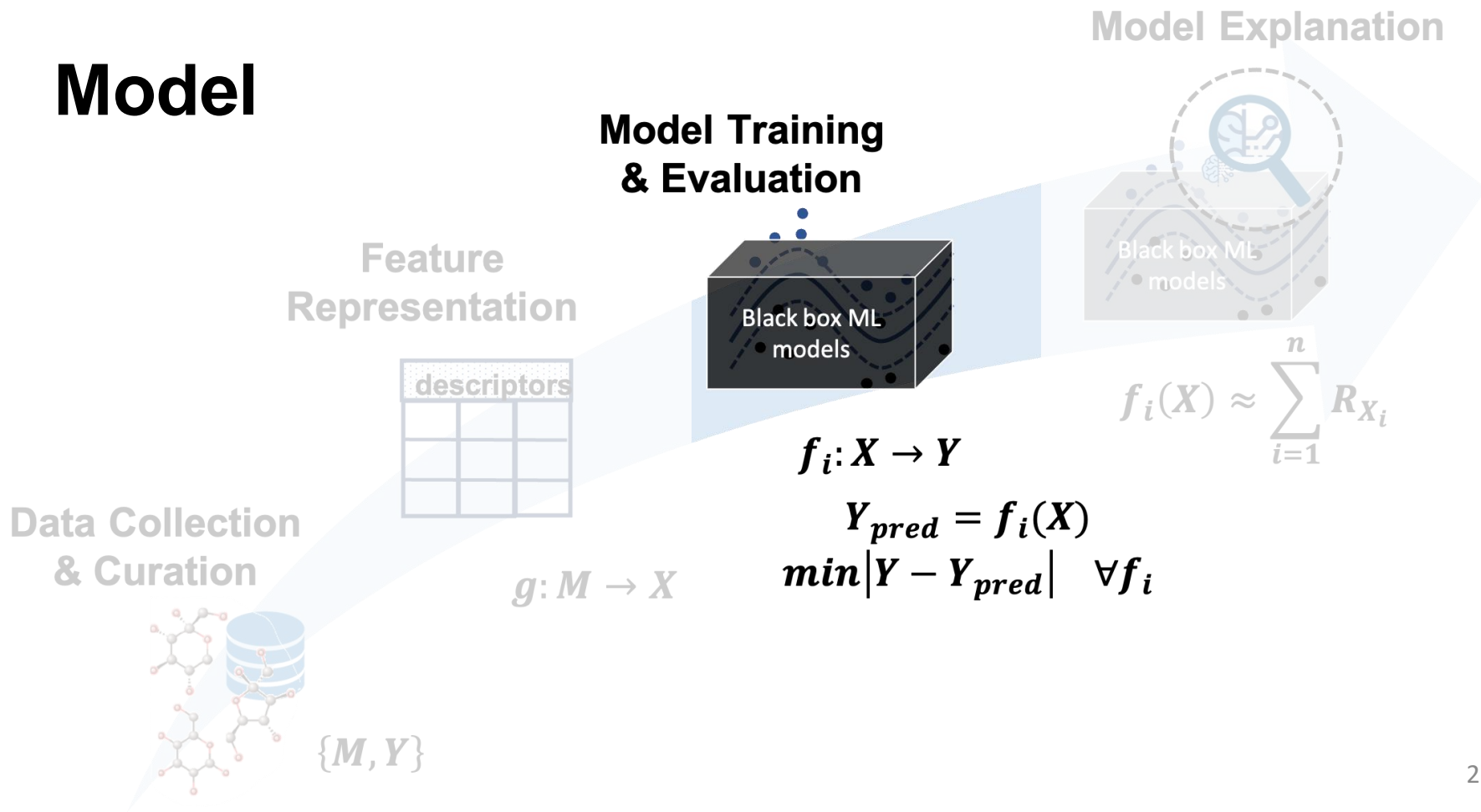
# Feature Representation

- qualities of a good feature representation: **comprehensiveness, feasibility, cost-effectiveness, performance, interpretability**
- multiple correct ways to represent different molecules
- systematic comparison & benchmarking strategies to choose the right featurization method → essential practice
  - cost vs benefit
  - feasibility analysis
  - comprehensiveness vs interpretability

<https://pubs.acs.org/doi/10.1021/acs.jpca.3c04779>



# Model



# Model Training & Evaluation

- Mapping feature space to target space, algorithm learns patterns from training examples, calculates loss, updates its parameters, and repeats until loss is minimized
- Model training
  - model selection (no free lunch theory, various algorithms exist, more than 1 may work for a problem, usually try a bunch of them, see which one works)
  - hyperparameter tuning, this determines starting point of the optimization problems, the steps it takes towards the optimized solution, etc.
- The above steps are done during cross validation

				VAL
			VAL	
		VAL		
	VAL			
VAL				



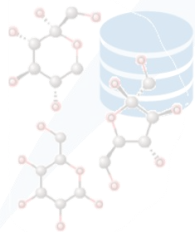
# Demo

# Model Training & Evaluation

- multiple ML algorithms may work, comprehensive selection process based on performance, cost & interpretability
- balanced splits for training, validation, & test sets, reporting & optimizing hyperparameters, learning curves, etc., common yet essential steps
- comprehensive set of error metrics, error bars & uncertainty
- documentation

## XAI

Data Collection  
& Curation


 $\{M, Y\}$ 

Feature  
Representation

descriptors			

$$g: M \rightarrow X$$

Model Training  
& Evaluation

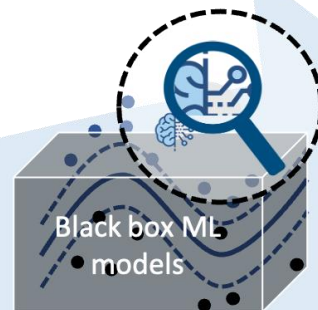


$$f_i: X \rightarrow Y$$

$$Y_{pred} = f_i(X)$$

$$\min |Y - Y_{pred}| \quad \forall f_i$$

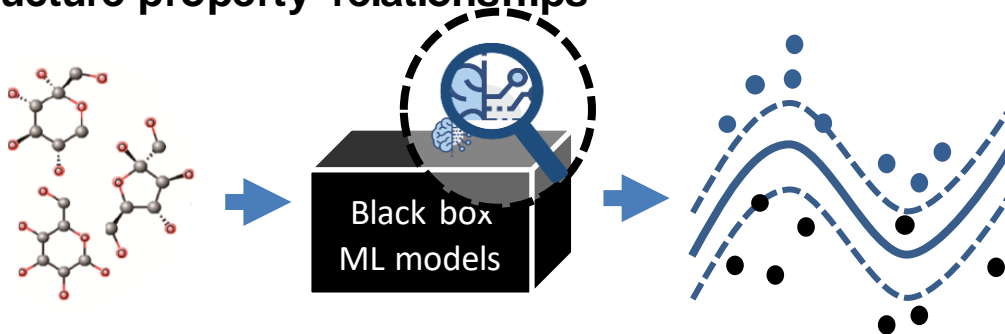
Model Explanation



$$f_i(X) \approx \sum_{i=1}^n R_{X_i}$$

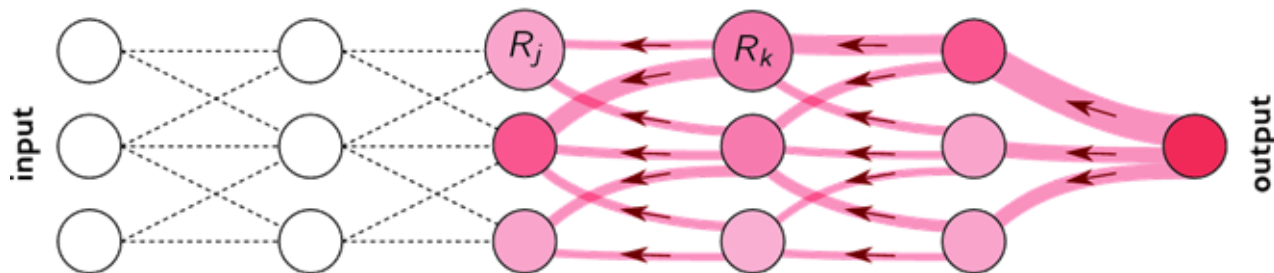
# Model Explanation

- ML approaches (esp. powerful deep learning techniques) are **black-box models**,  $\therefore$  **low trust**
- new field of **eXplainable Artificial Intelligence (XAI)** creates visibility in AI/ML models
- explains **rationale** behind model's predictions, helps **ensure scientific basis**
- helps understand intricate **structure-property relationships** – identifies important features



# Model Explanation

- Layer-wise Relevance Propagation: propagates the values of the outputs back to the input layer. The portion of the value that reaches each input feature is the relevance of that feature for the particular prediction. This relevance is then aggregated in various ways to assign a 'global' relevance score to each feature.



# Demo

# Summary

- Data is King!
- Features are what your model actually sees, make sure that they hold sufficient information and do not prove to be a computational bottleneck
- A problem may have multiple solutions, so try different models, do cross validation, never compares apples to oranges, when comparing 2 models, do not change anything else but the ML algorithm
- Different metrics are like pieces of a puzzle, individually they may not convey everything, but together they can tell the entire story.
- To convince an experimentalist... explain your model using XAI, remember they may still not be convinced.

# Resources

- Data sources

**Table 1**  
A list of popular chemical databases commonly used in ML.

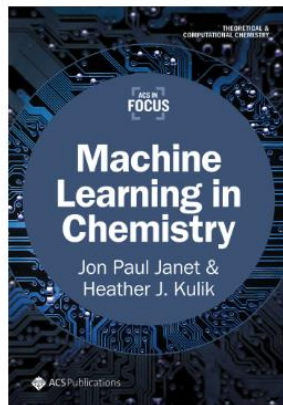
Classification	Name	Content	URL
Chemical reaction databases	SciFinder	Information on chemical compounds, bibliographic data, and chemical reactions (commercial database)	<a href="https://scifinder.cas.org/">https://scifinder.cas.org/</a>
	Reaxys	Chemical reaction and bibliographic information (commercial database)	<a href="https://www.reaxys.com/">https://www.reaxys.com/</a>
	USPTO	Chemical structure and reaction	<a href="https://www.repository.cam.ac.uk/handle/1810/244727">https://www.repository.cam.ac.uk/handle/1810/244727</a>
	ORD	Organic chemical reaction data	<a href="https://github.com/open-reaction-database">https://github.com/open-reaction-database</a>
	NextMove	Chemical reaction data	<a href="https://www.nextmovesoftware.com/about.html">https://www.nextmovesoftware.com/about.html</a>
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	<a href="https://pubchem.ncbi.nlm.nih.gov/">https://pubchem.ncbi.nlm.nih.gov/</a>
	NIST	Standard physicochemical properties of compounds	<a href="https://webbook.nist.gov/chemistry/">https://webbook.nist.gov/chemistry/</a>
	ChemSpider	Structure and property of compounds	<a href="https://www.chemspider.com">https://www.chemspider.com</a>
	ChEMBL	Drug-like properties of bioactive molecules	<a href="https://www.ebi.ac.uk/chembl/">https://www.ebi.ac.uk/chembl/</a>
	DrugBank	Properties of drug molecules	<a href="https://go.drugbank.com/releases/latest">https://go.drugbank.com/releases/latest</a>
Material databases	Tox21	Toxic effects of substances	<a href="https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html">https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html</a>
	ESOL	Water solubility of compounds	<a href="https://doi.org/10.1021/ci034243x">https://doi.org/10.1021/ci034243x</a>
	FreeSolv	Water solubility of small neutral molecules	<a href="https://github.com/MobleyLab/FreeSolv">https://github.com/MobleyLab/FreeSolv</a>
	Lipophilicity	Lipid solubility of organic compounds	<a href="https://doi.org/10.1002/jcem.2718">https://doi.org/10.1002/jcem.2718</a>
	CSD	Organic and metal-organic crystal structures	<a href="https://www.ccdc.cam.ac.uk/">https://www.ccdc.cam.ac.uk/</a>
	ICSD	Inorganic and metal-organic crystal structures	<a href="https://icsd.products.fiz-karlsruhe.de/">https://icsd.products.fiz-karlsruhe.de/</a>
	PDF	Diffraction data of inorganic and organic compounds	<a href="https://www.icdd.com/pdfssearch/">https://www.icdd.com/pdfssearch/</a>
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	<a href="https://matweb.com/">https://matweb.com/</a>
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	<a href="https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/u55r-gjdb">https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/u55r-gjdb</a>
	HTeM	Experimental information of inorganic thin-film materials	<a href="https://item.nrel.gov/">https://item.nrel.gov/</a>
Computational chemistry databases	GDB-17	Structures of organic molecules up to 17 atoms	<a href="https://www.gdb.unibe.ch/downloads/">https://www.gdb.unibe.ch/downloads/</a>
	QM9	Quantum chemical properties of organic molecules	<a href="https://quantum-machine.org/datasets/">https://quantum-machine.org/datasets/</a>
	ANI-1	Energy and force of non-equilibrium molecules	<a href="https://github.com/isaev/ANI1_dataset">https://github.com/isaev/ANI1_dataset</a>
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://next-gen.materialsproject.org/">https://next-gen.materialsproject.org/</a>
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://oqmd.org/">https://oqmd.org/</a>
	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	<a href="https://aflowlib.org/">https://aflowlib.org/</a>
	MD17/ISO-17	Energy and force of non-equilibrium molecules	<a href="https://quantum-machine.org/datasets/">https://quantum-machine.org/datasets/</a>
	LASP	Global PES dataset of molecules/materials	<a href="https://www.lasphub.com">https://www.lasphub.com</a>
	OC20	Adsorption energy of molecules in catalysts	<a href="https://opencatalystproject.org/">https://opencatalystproject.org/</a>
	Atom3D	3D structure of molecules, RNA, and proteins	<a href="https://www.atom3d.ai/">https://www.atom3d.ai/</a>

URL: uniform resource locator; USPTO: United States Patent and Trademark Office; ORD: Open Reaction Database; NIST: National Institute of Standards and Technology; CSD: Cambridge Structural Database; ICSD: Inorganic Crystal Structure Database; PDF: Powder Diffraction File; HTeM: High-Throughput Experimental Materials; OQMD: Open Quantum Materials Database; OC20: Open Catalyst 2020; DFT: density functional theory; PES: potential energy surface.



# Resources

- [Data sources](#)
- [A good introductory book written by Chemists \(only 18 pages\)](#)

[Read Now](#)

## Machine Learning in Chemistry

**Author(s):** Jon Paul Janet, Heather J. Kulik

**Publication Date:** May 29, 2020

Copyright © 2020 American Chemical Society

[Subscribed](#)

**Cite This:** *Machine Learning in Chemistry*, American Chemical Society, 2020. DOI: 10.1021/acs.infocus.7e4001

**eISBN:** 9780841299009

**DOI:** 10.1021/acs.infocus.7e4001

**Subjects:** Algorithms, Chemical engineering and industrial chemistry, Computational modeling, Machine learning, Neural networks, Theoretical and computational chemistry

**Read Time:** five to six hours

**Collection:** Inaugural


**Publisher:** American Chemical Society

[Get it on Google Play](#)

Recent advances in machine learning or artificial intelligence for vision and natural language processing

# Resources

- [Data sources](#)
- [A good introductory book written by Chemists \(only 18 pages\)](#)
- [A great github repository for practical cheminformatics](#)



Patrick Walters  
PatWalters

Follow

656 followers · 11 following

Practical Cheminformatics With Open Source Software

A set of Jupyter notebooks for learning Cheminformatics. The links below will open the tutorials on Google Colab. This way you can run the notebooks without having to install software on your computer. Of course, you can also just clone the repo and run these notebooks on your own computer.

Fundamentals

1. [A Whirlwind Introduction to the RDKit for Cheminformatics](#)
2. [A Brief Introduction to Pandas for Cheminformatics](#)
3. [SMILES Tutorial](#)
4. [SMARTS Tutorial](#)
5. Recursive SMARTS - Under Construction
6. Reaction SMARTS - Under Construction

Using [datamol](#) and [molfeat](#) to Streamline Cheminformatics Workflows

7. [Data Manipulation, Descriptors and Clustering](#)

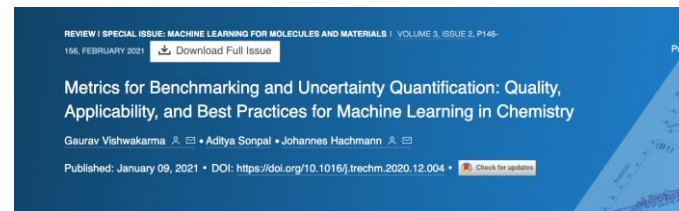
Clustering

8. [K-Means Clustering](#)
9. [Taylor-Butina Clustering](#)
10. [Self-Organizing Maps](#)

Misc Cheminformatics Analysis

# Resources

- [Data sources](#)
- [A good introductory book written by Chemists \(only 18 pages\)](#)
- [A great github repository for practical cheminformatics](#)
- [Plugging a review article I co-authored about metrics](#)



Highlights

Keywords

References

Glossary

Title Info

Related

Citations

## Highlights

As machine learning (ML) is gaining an increasingly prominent role in chemical research, so is the need to assess the quality and applicability of ML models, compare different ML models, and develop best-practice guidelines for their design and utilization. Statistical loss function metrics and uncertainty quantification techniques are key issues in this context.

Different analyses highlight different facets of a model's performance, and a compilation of metrics, as opposed to a single metric, allows for a well-rounded understanding of what can be expected from a model. They also allow us to identify unexplored regions of chemical space and pursue their survey.

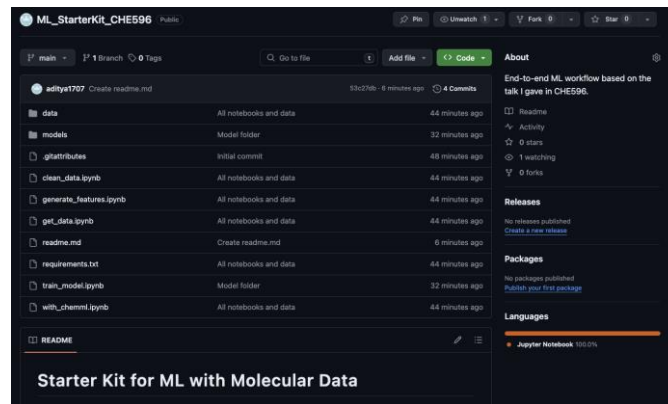
# Resources

- [Data sources](#)
- [A good introductory book written by Chemists \(only 18 pages\)](#)
- [A great github repository for practical cheminformatics](#)
- [Plugging a review article I co-authored about metrics](#)
- [Plugging a software I co-wrote for ML, esp. XAI](#)



# Resources

- [Data sources](#)
- [A good introductory book written by Chemists \(only 18 pages\)](#)
- [A great github repository for practical cheminformatics](#)
- [Plugging a review article I co-authored about metrics](#)
- [Plugging a software I co-wrote for ML, esp. XAI](#)
- [GitHub repository to all the code I demonstrated](#)



# List of Tools & Libraries

- Data Collection & curation: Pandas, numpy, scipy, and a gazillion others
- Feature representation: rdkit, openbabel, etc.
- ML algorithm: Scikit-learn, pytorch, tensorflow, LightGBM
- Plotting and visualization: Matplotlib, seaborn, pandas
- XAI: ChemML