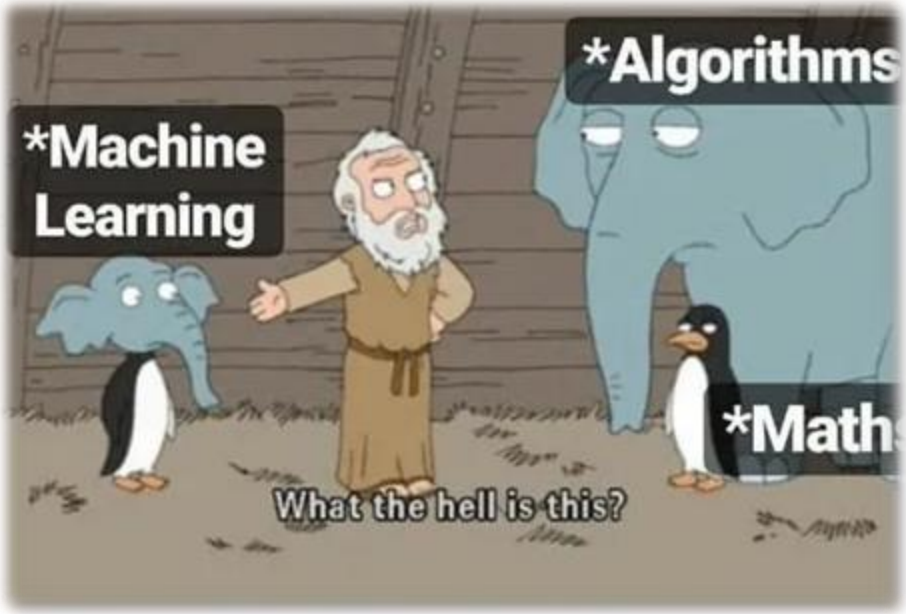
Machine Learning: A Starter Kit

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

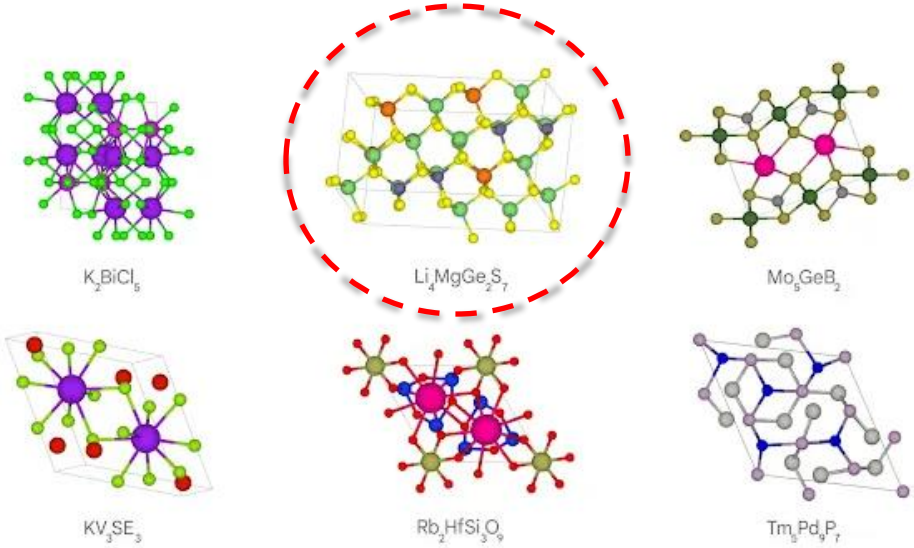


AI AI Everywhere!

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

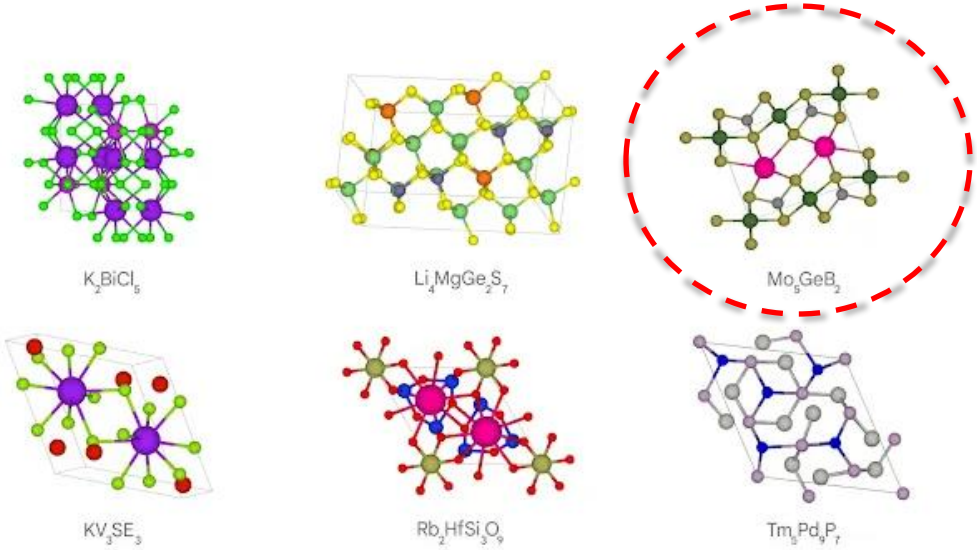
AI AI Everywhere!

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



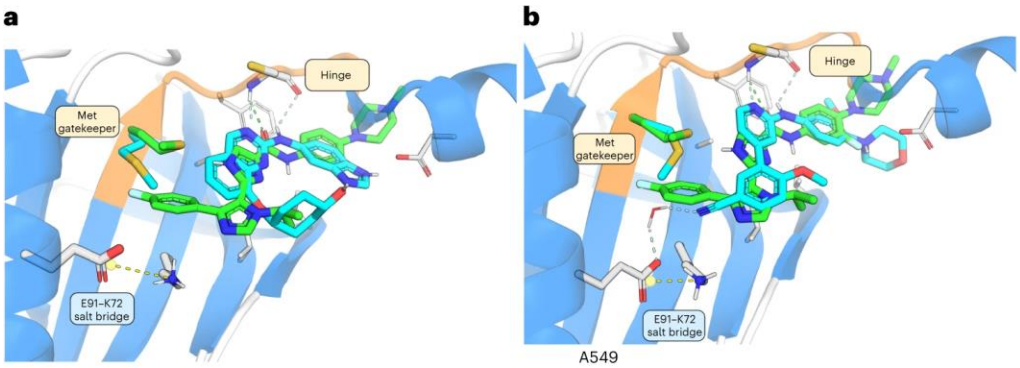
AI AI Everywhere!

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



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



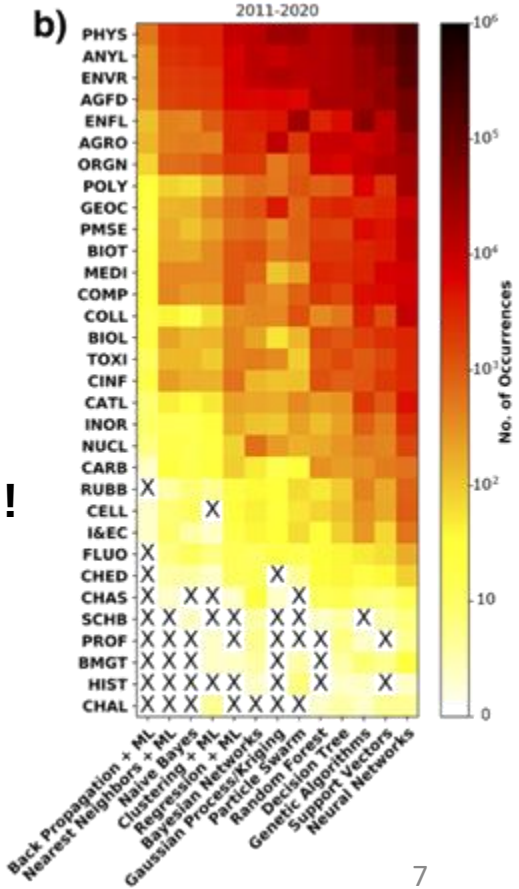
Source: https://www.nature.com/articles/s41587-024-02143-0

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

Applications of ML in Chem

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

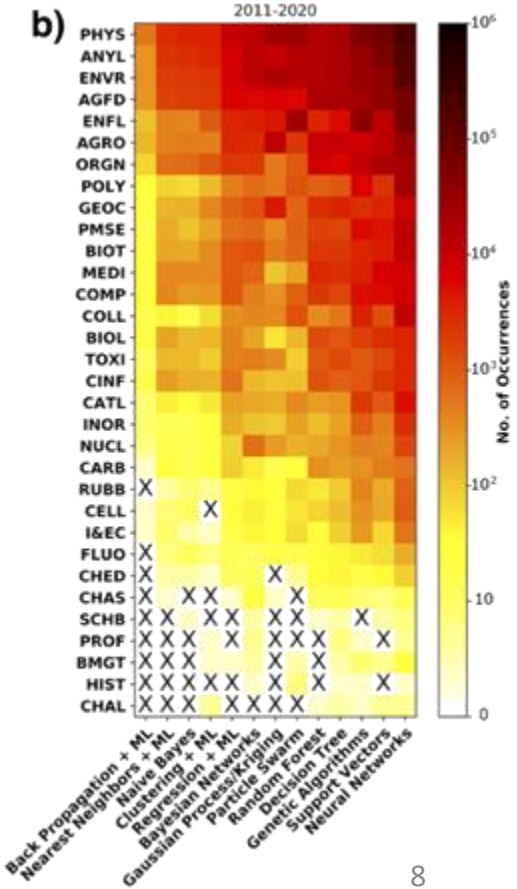
Invaluable in drug and materials discovery and design!



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

Applications of ML in Chem

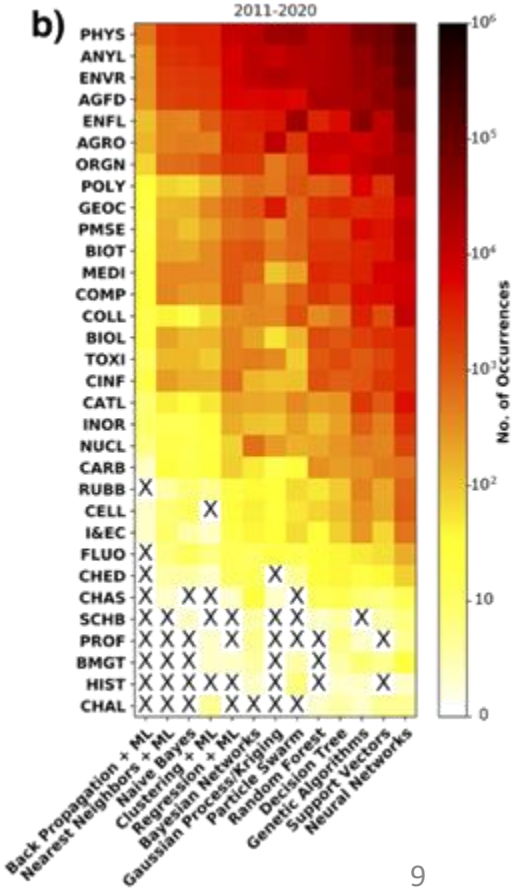
- Mapping molecular structure to target property
- Accelerating computational chemistry
  - predicting forcefield parameters
  - predicting DFT level energies



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

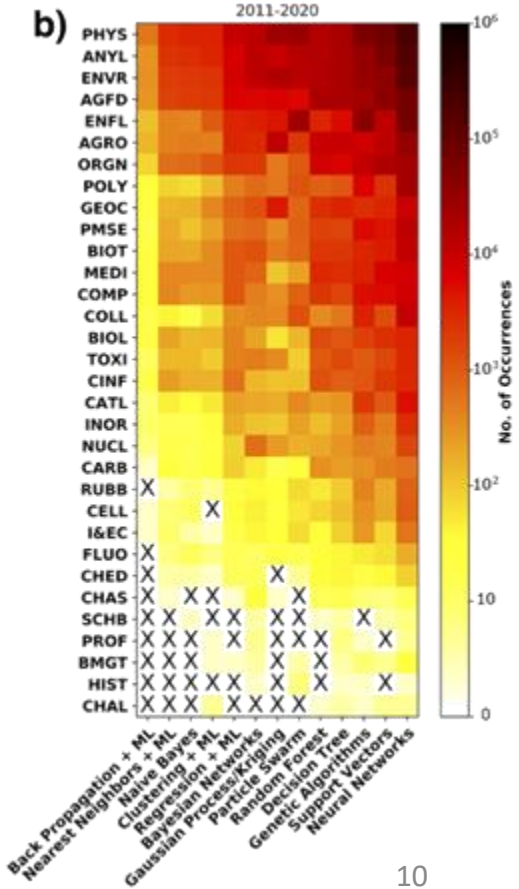


https://pubs.acs.org/doi/10.1021/acs.chemrev.1c00107

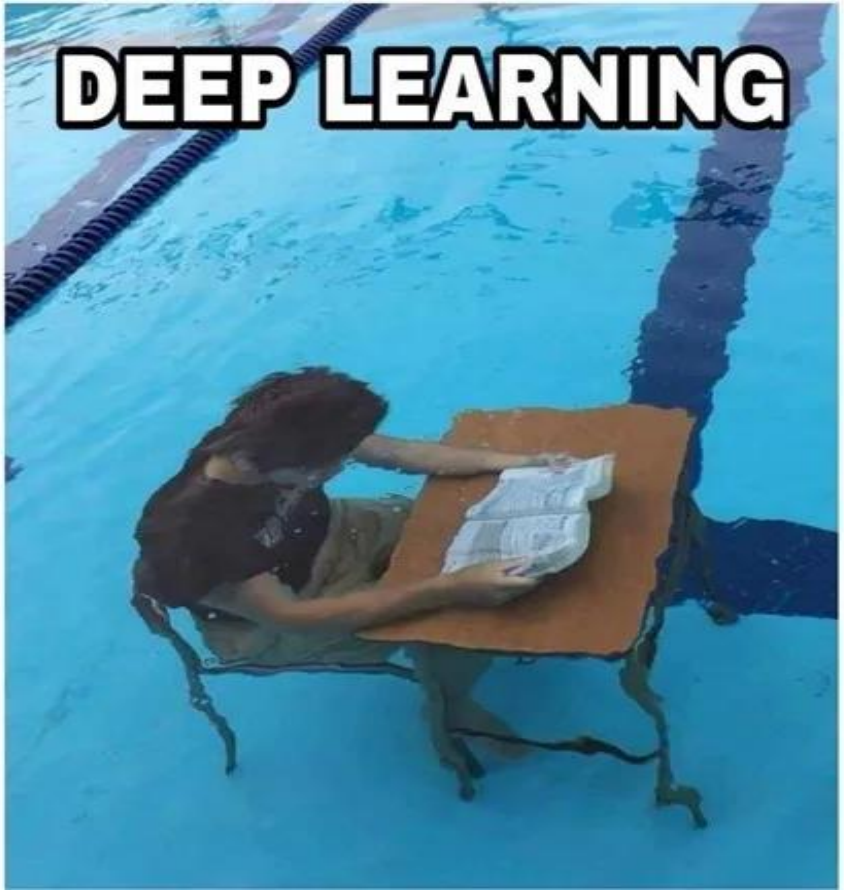
Applications of ML in Chem

- 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

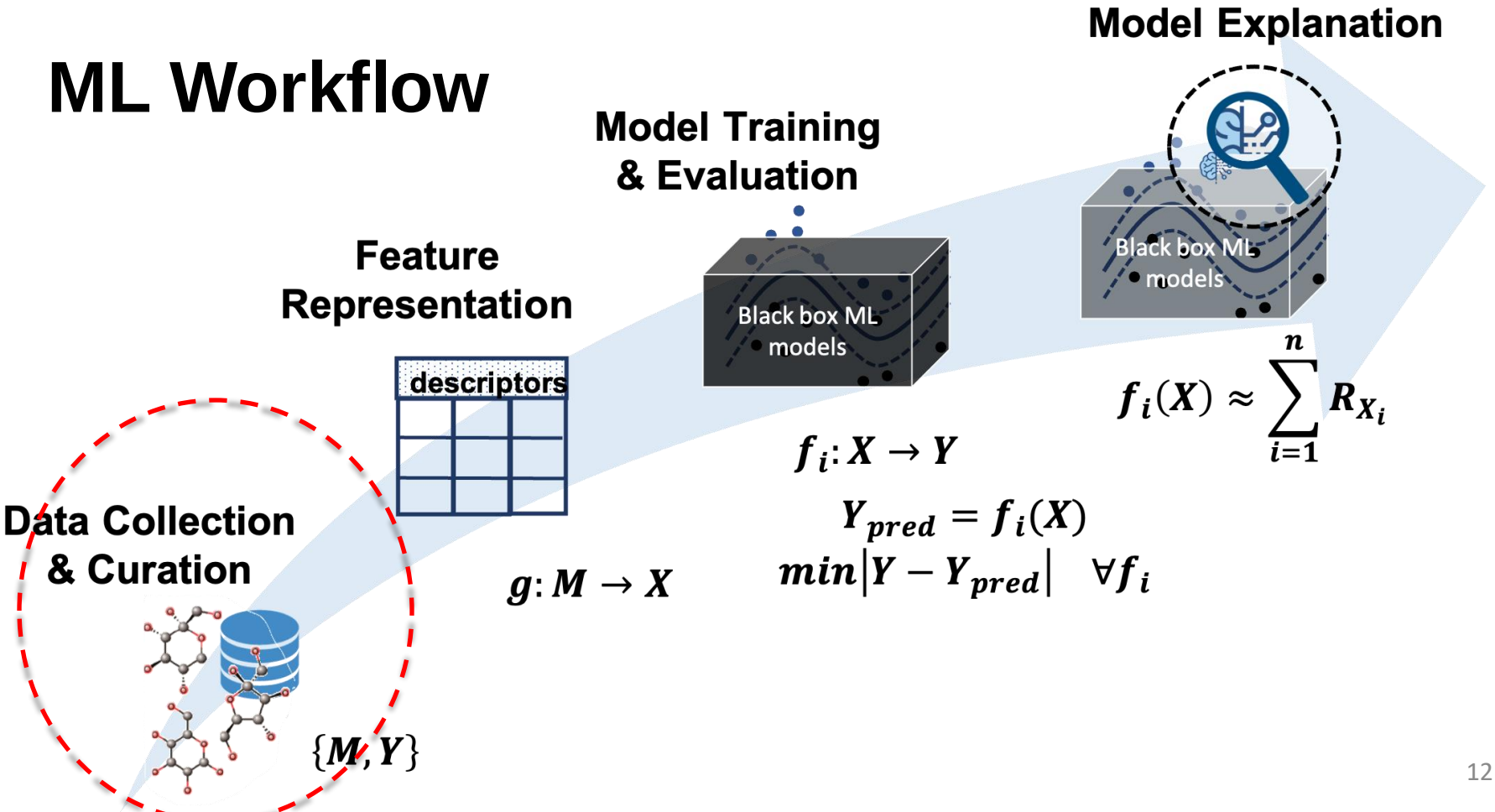
This is not an exhaustive list!



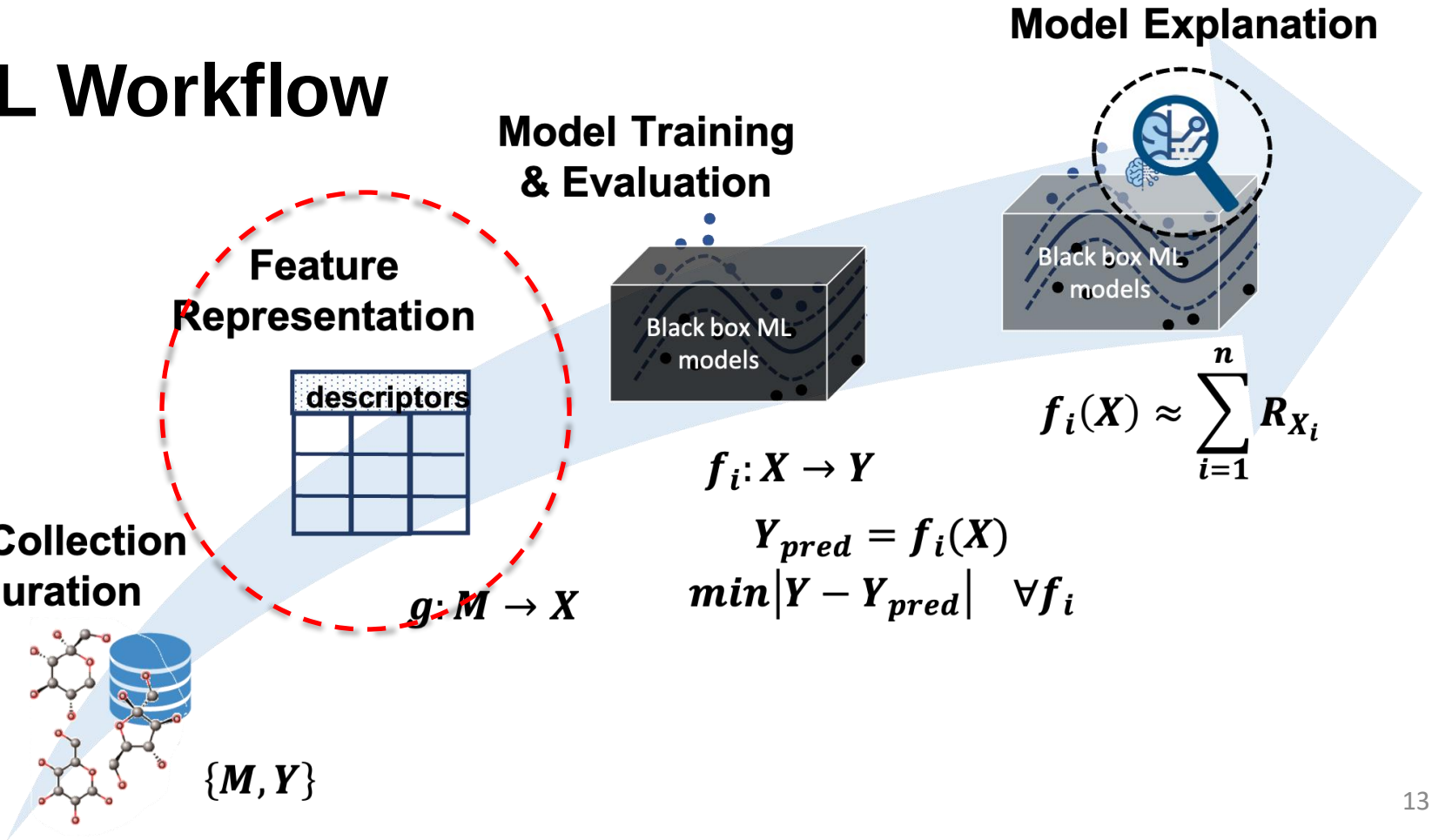
Breathe?



ML Workflow

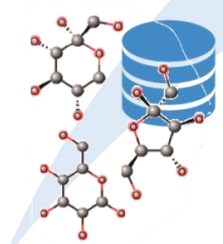

Model Explanation
Model Training
& Evaluation
Feature
Representation
descriptors
Black box ML
models
Black box ML
models
Data Collection
& Curation
$\{M, Y\}$
$g: M \rightarrow X$
$f_i: X \rightarrow Y$
$Y_{pred} = f_i(X)$
$min|Y - Y_{pred}| \quad \forall f_i$
$f_i(X) \approx \sum_{i=1}^{n} R_{X_i}$

ML Workflow


Model Explanation
Model Training
& Evaluation
Feature
Representation
descriptors
Black box ML
models
Black box ML
models
Data Collection
& Curation
$f_i(X) \approx \sum_{i=1}^{n} R_{X_i}$
$f_i: X \rightarrow Y$
$Y_{pred} = f_i(X)$
$min|Y - Y_{pred}| \quad \forall f_i$
$g: M \rightarrow X$
$\{M, Y\}$

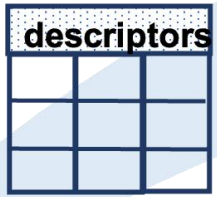
ML Workflow

Data Collection & Curation



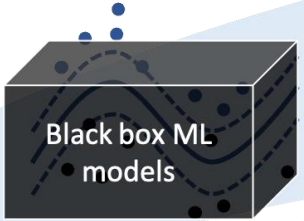
$\{M, Y\}$

Feature Representation



$g: M \rightarrow X$

Model Training & Evaluation

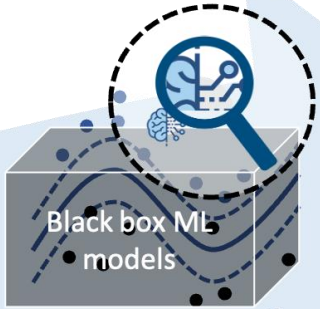


$f_i: X \rightarrow Y$

$Y_{pred} = f_i(X)$

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

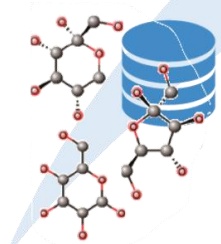
Model Explanation



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

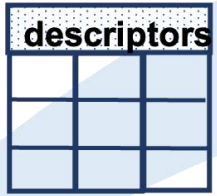
ML Workflow

Data Collection & Curation



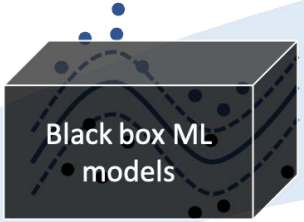
$\{M, Y\}$

Feature Representation



$g: M \rightarrow X$

Model Training & Evaluation

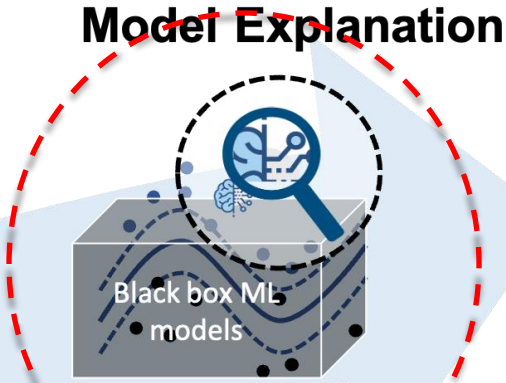


$f_i: X \rightarrow Y$

$Y_{pred} = f_i(X)$

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

Model Explanation



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

Data

Model Explanation

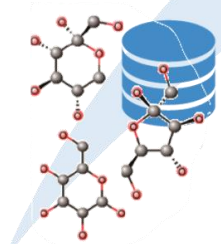
Model Training & Ev

Feature Representation

descriptors

$g: M \rightarrow X$

Data Collection & Curation



$\{M, Y\}$

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)	https://scifinder.cas.org/
	Reaxys	Chemical reaction and bibliographic information (commercial database)	https://www.reaxys.com/
	USPTO	Chemical structure and reaction	https://www.repository.cam.ac.uk/handle/1810/244727
	ORD	Organic chemical reaction data	https://github.com/open-reaction-database
	NextMove	Chemical reaction data	https://www.nextmovesoftware.com/about.html
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	https://pubchem.ncbi.nlm.nih.gov/
	NIST	Standard physicochemical properties of compounds	https://webbook.nist.gov/chemistry/
	ChemSpider	Structure and property of compounds	https://www.chemspider.com
	ChemBL	Drug-like properties of bioactive molecules	https://www.ebi.ac.uk/chembl/
	DrugBank	Properties of drug molecules	https://go.drugbank.com/releases/latest
	Tox21	Toxic effects of substances	https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html
	ESOL	Water solubility of compounds	https://doi.org/10.1021/ci034243x
	FreeSolv	Water solubility of small neutral molecules	https://github.com/MobleyLab/FreeSolv
	Lipophilicity	Lipid solubility of organic compounds	https://doi.org/10.1002/cem.2718
Material databases	CSD	Organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk/
	ICSD	Inorganic and metal–organic crystal structures	https://icsd.products.fiz-karlsruhe.de/
	PDF	Diffraction data of inorganic and organic compounds	https://www.icdd.com/pdfsearch/
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	https://matweb.com/
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/uj5r-zjdb
	HTEM	Experimental information of inorganic thin-film materials	https://htem.nrel.gov/
Computational chemistry databases	GDB-17	Structures of organic molecules up to 17 atoms	https://www.gdb.unibe.ch/downloads/
	QM9	Quantum chemical properties of organic molecules	https://quantum-machine.org/datasets/
	ANI-1	Energy and force of non-equilibrium molecules	https://github.com/isayev/ANI1_dataset
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://next-gen.materialsproject.org/
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://oqmd.org/
	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://aflowlib.org/
	MD17/ISO-17	Energy and force of non-equilibrium molecules	https://quantum-machine.org/datasets/
	LASP	Global PES dataset of molecules/materials	https://www.lasphub.com
	OC20	Adsorption energy of molecules in catalysts	https://opencatalystproject.org/
	Atom3D	3D structure of molecules, RNA, and proteins	https://www.atom3d.ai/

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

notebooks: get_data.ipynb, clean_data.ipynb

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

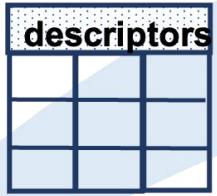
Data is King! Your model is only as good as the data you provide it!

Features

Model Explanation

Model Training & Evaluation

Feature Representation



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

$$f_i: X \rightarrow Y$$

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

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

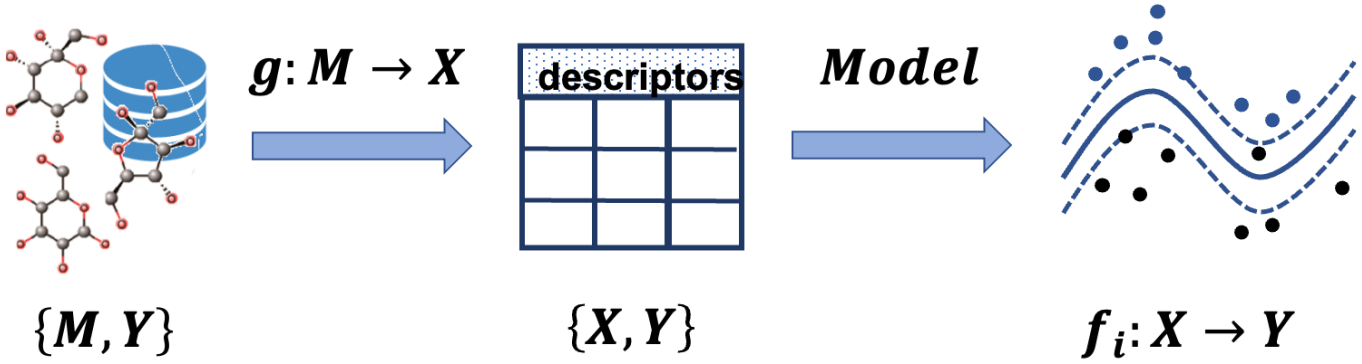
Data Collection & Curation

$$g: M \rightarrow X$$

$$\{M, Y\}$$

Feature Representation

- function g: M → X converts a chemical representation m ∈ M to a feature input x ∈ 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.)



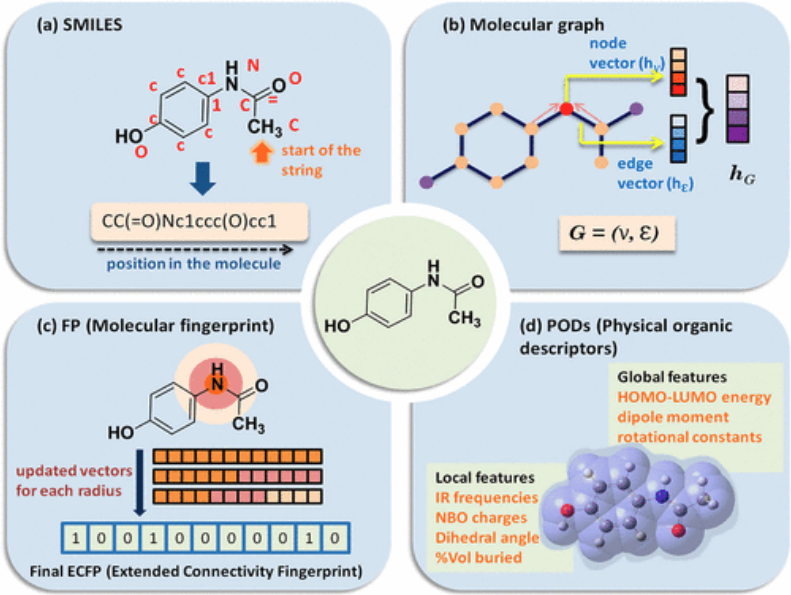
Demo

notebook: generate_features.ipynb

Feature Representation

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

- 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



Model

Model Explanation

Model Training
& Evaluation

Feature
Representation

descriptors

Black box ML
models

Black box ML
models

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

$$f_i: X \rightarrow Y$$

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

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

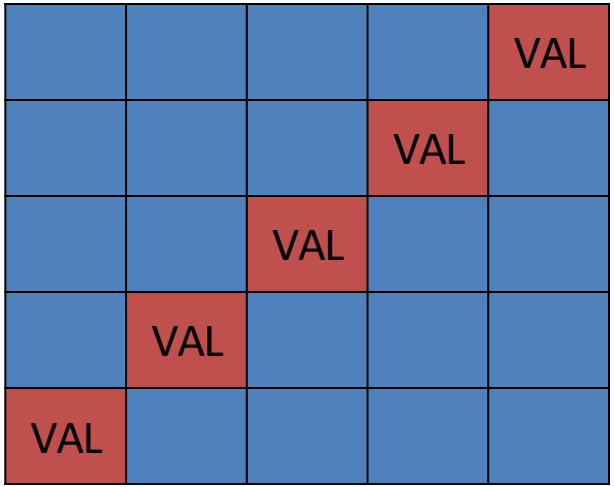
Data Collection
& Curation

$$g: M \rightarrow X$$

$$\{M, Y\}$$

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



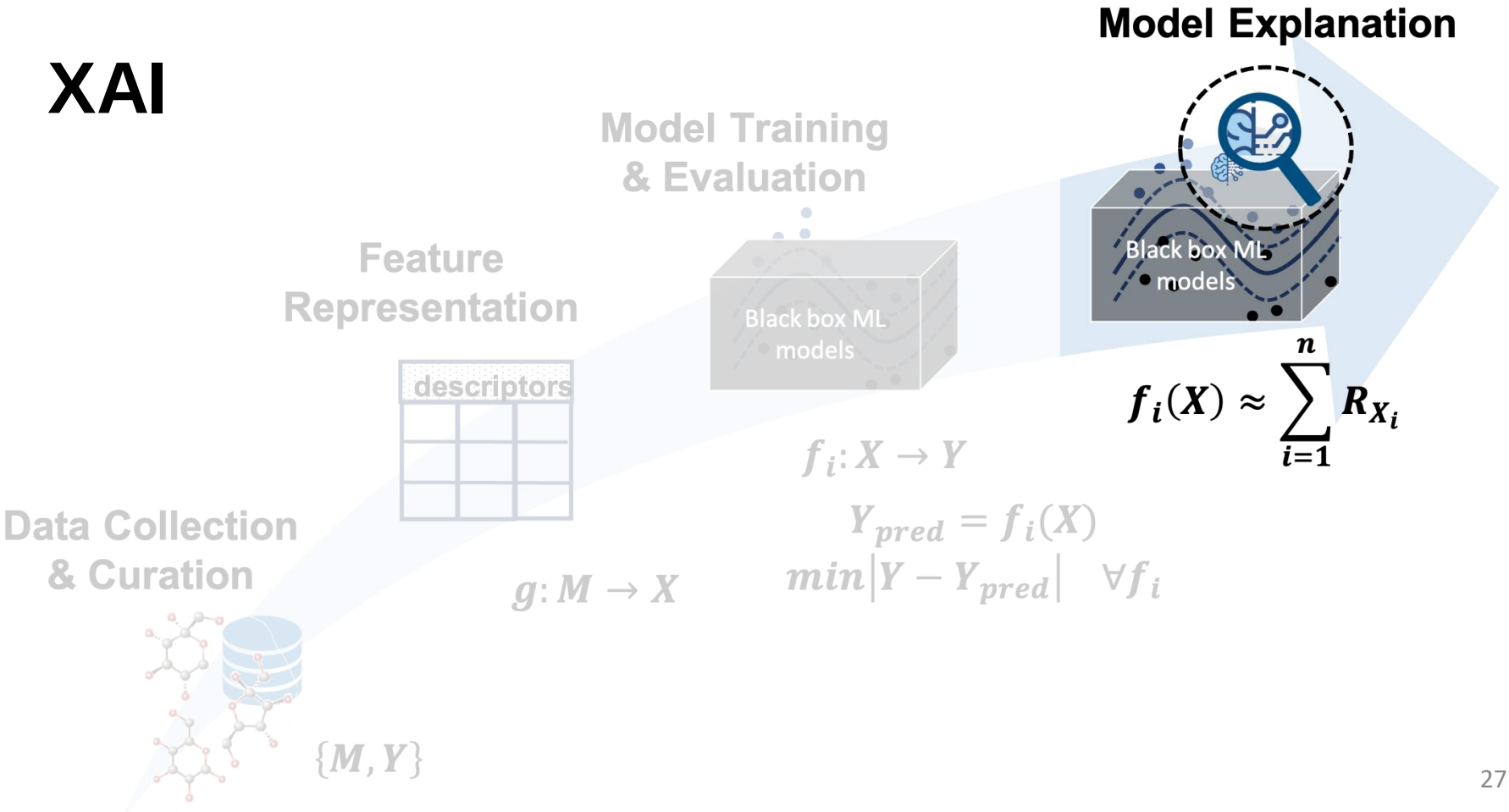
Demo

notebook: train_model.ipynb

Model Training & Evaluation

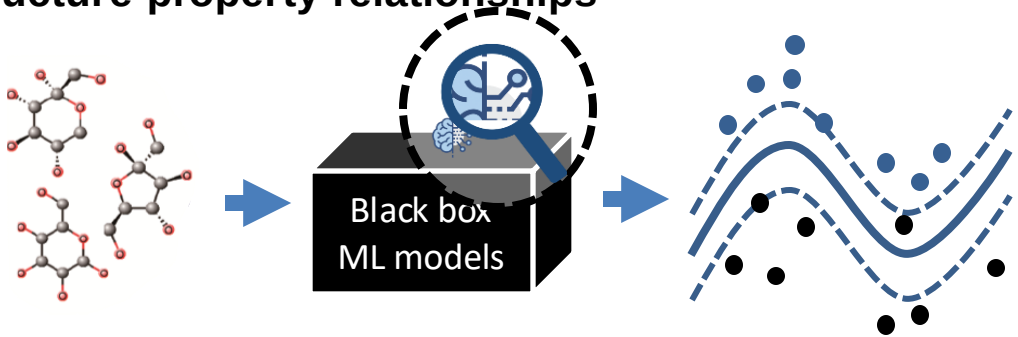
- multiple ML algorithms may work, comprehensive selection process based on performance, cost & interpretability
- balanced splits for training, validation, & test sets, reporting & optimizing hyperparameters, learning curves, etc., common yet essential steps
- comprehensive set of error metrics, error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

XAI


Model Explanation
Model Training & Evaluation
Feature Representation
descriptors
Black box ML models
Black box ML models
Data Collection & Curation
$f_i: X \rightarrow Y$
$Y_{pred} = f_i(X)$
$min|Y - Y_{pred}| \quad \forall f_i$
$g: M \rightarrow X$
$\{M, Y\}$
$f_i(X) \approx \sum_{i=1}^{n} R_{X_i}$

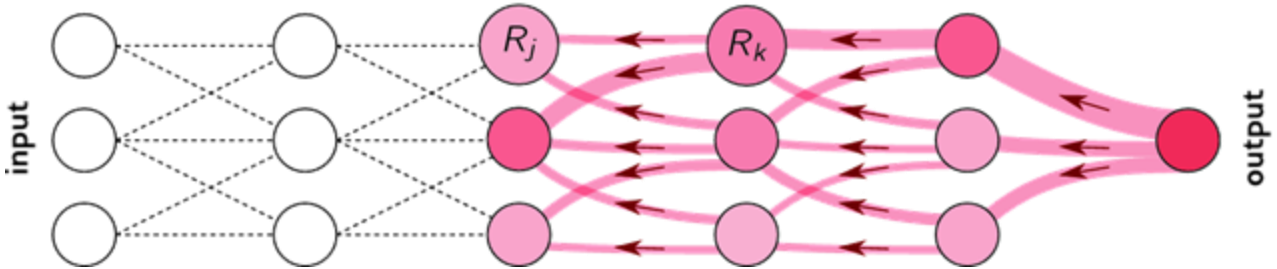
Model Explanation

- ML approaches (esp. powerful deep learning techniques) are **black-box models, ∴ 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

notebook: with_chemml.ipynb

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

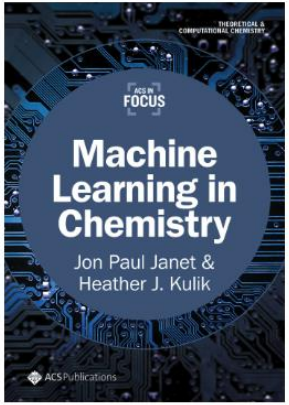
Table 1
A list of popular chemical databases commonly used in ML.

Classification	Name	Content	URL
Chemical reaction databases	SciFinder	Information on chemical compounds, bibliographic data, and chemical reactions (commercial database)	https://scifinder.cas.org/
	Reaxys	Chemical reaction and bibliographic information (commercial database)	https://www.reaxys.com/
	USPTO	Chemical structure and reaction	https://www.repository.cam.ac.uk/handle/1810/244727
	ORD	Organic chemical reaction data	https://github.com/open-reaction-database
	NextMove	Chemical reaction data	https://www.nextmovesoftware.com/about.html
Chemical property databases	PubChem	Chemical and physical properties, biological activities, and toxicity of substances	https://pubchem.ncbi.nlm.nih.gov/
	NIST	Standard physicochemical properties of compounds	https://webbook.nist.gov/chemistry/
	ChemSpider	Structure and property of compounds	https://www.chemspider.com
	ChemBL	Drug-like properties of bioactive molecules	https://www.ebi.ac.uk/chembl/
	DrugBank	Properties of drug molecules	https://go.drugbank.com/releases/latest
	Tox21	Toxic effects of substances	https://ntp.niehs.nih.gov/whatwestudy/tox21/index.html
	ESOL	Water solubility of compounds	https://doi.org/10.1021/ci034243x
	FreeSolv	Water solubility of small neutral molecules	https://github.com/MobleyLab/FreeSolv
	Lipophilicity	Lipid solubility of organic compounds	https://doi.org/10.1002/cem.2718
Material databases	CSD	Organic and metal–organic crystal structures	https://www.ccdc.cam.ac.uk/
	ICSD	Inorganic and metal–organic crystal structures	https://icsd.products.fiz-karlsruhe.de/
	PDF	Diffraction data of inorganic and organic compounds	https://www.icdd.com/pdfsearch/
	MatWeb	The thermoplastic and thermoset of polymers, metals, and other engineering materials	https://matweb.com/
	Li-ion Battery Aging Datasets	Charge and discharge curves of lithium batteries	https://data.nasa.gov/dataset/Li-ion-Battery-Aging-Datasets/uj5r-zjdb
	HTEM	Experimental information of inorganic thin-film materials	https://htem.nrel.gov/
Computational chemistry databases	GDB-17	Structures of organic molecules up to 17 atoms	https://www.gdb.unibe.ch/downloads/
	QM9	Quantum chemical properties of organic molecules	https://quantum-machine.org/datasets/
	ANI-1	Energy and force of non-equilibrium molecules	https://github.com/isayev/ANI1_dataset
	Materials Project	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://next-gen.materialsproject.org/
	OQMD	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://oqmd.org/
	Aflowlib	DFT relaxed material structures and their thermal, electronic, and elastic properties	https://aflowlib.org/
	MD17/ISO-17	Energy and force of non-equilibrium molecules	https://quantum-machine.org/datasets/
	LASP	Global PES dataset of molecules/materials	https://www.lasphub.com
	OC20	Adsorption energy of molecules in catalysts	https://opencatalystproject.org/
	Atom3D	3D structure of molecules, RNA, and proteins	https://www.atom3d.ai/

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

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

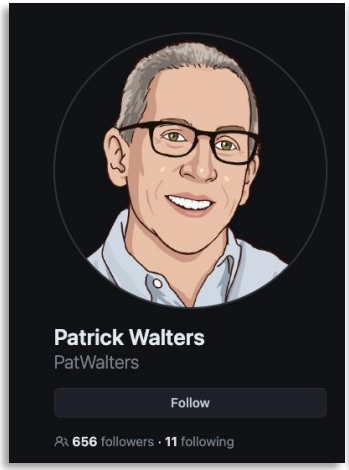
Publisher: American Chemical Society

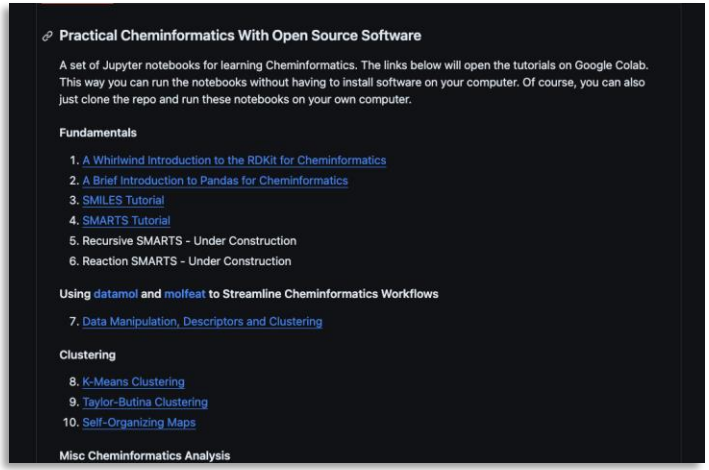
Get it on Google Play

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

Resources

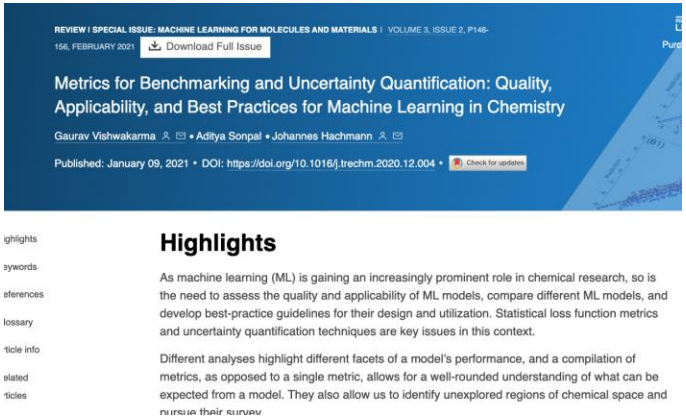
- Data sources
- A good introductory book written by Chemists (only 18 pages)
- A great github repository for practical cheminformatics





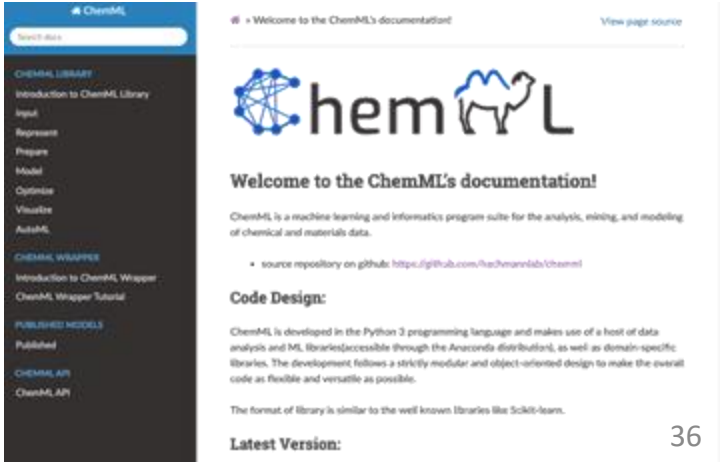
Resources

- 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



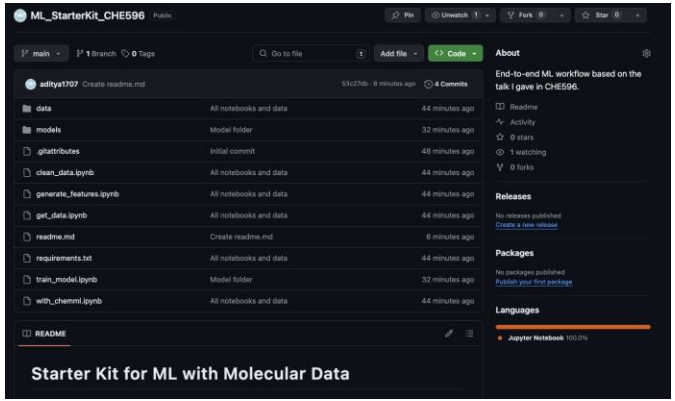
Resources

- 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