Chimica Computazionale

Machine learning and computational chemistry

Emanuele Coccia

UNIVERSITÀ DEGLI STUDI DI TRIESTE

DSCF Dipartimento di Scienze Chimiche e Farmaceutiche

PhotoInduced Quantum Dynamics (PIQD) Group

PIQD

The Nobel Prize in Physics 2024

John J. Hopfield

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Geoffrey Hinton

"for foundational discoveries and inventions that enable machine learning with artificial neural networks"

Nobel prize in Chemistry 2024

The Nobel Prize in Chemistry 2024

David Baker

"for computational protein design"

Demis Hassabis

"for protein structure prediction"

John Jumper

"for protein structure prediction"

Artificial intelligence

- Computer systems able of mimicking decision-making and problem-solving tasks of a human mind
- Machine learning:
  - Pathway to AI that uses statistical models and training algorithms
  - Learns insights and patterns in the data
  - Makes new predictions without additional input/programming
- Large amount of data available

Machine learning in a nutshell

ML algorithms:

- Estimate relationships without any instruction of how to analyze or draw conclusions from the data
- Can recover mappings between a set of inputs/outputs or from the inputs alone
- Can discover structure in the data

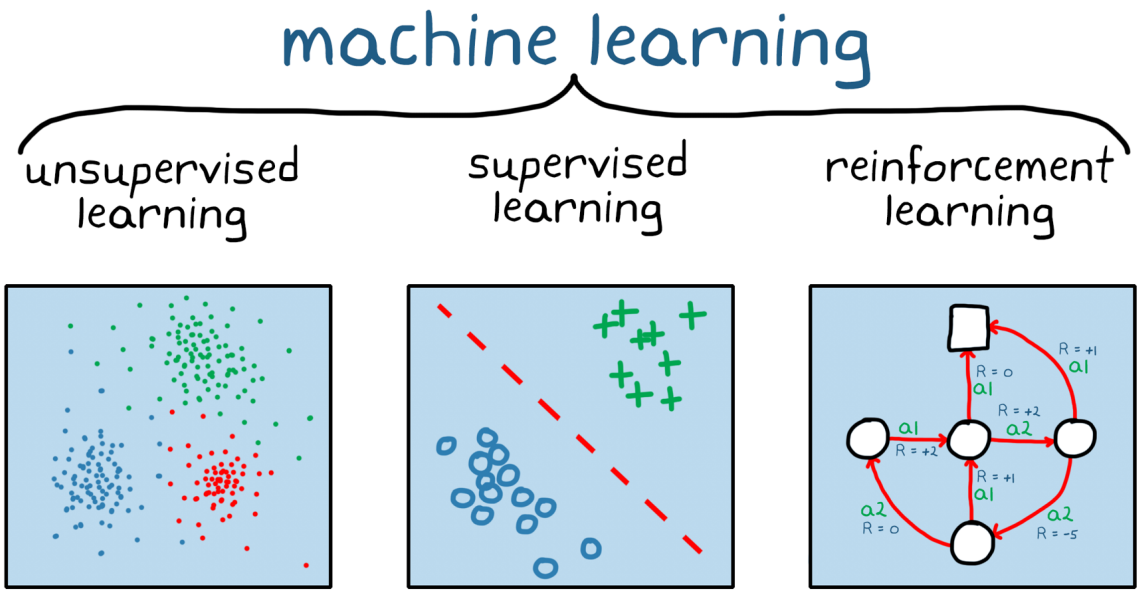
Machine learning in a nutshell

ML algorithms:

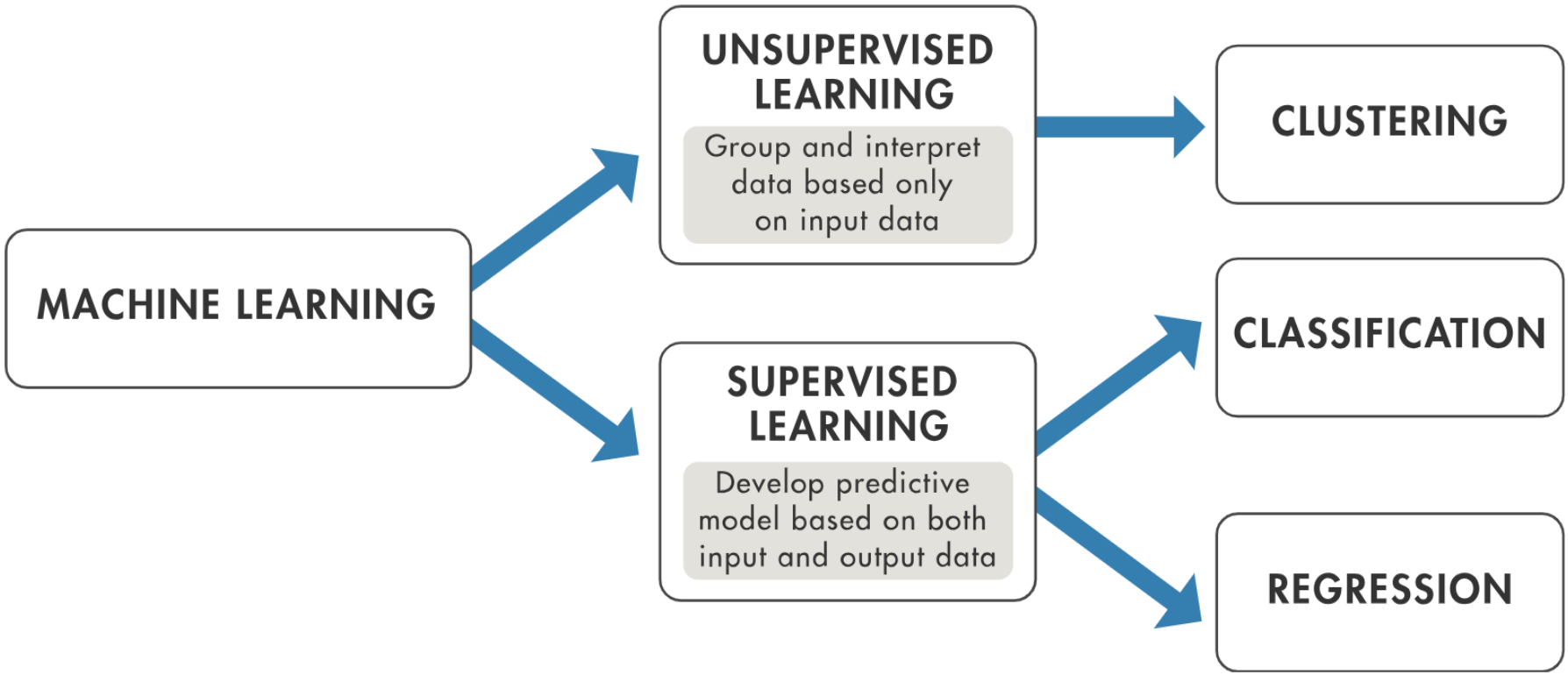
- Estimate relationships without any instruction of how to analyze or draw conclusions from the data
- Can recover mappings between a set of inputs/outputs or from the inputs alone
- Can discover structure in the data
- Use universal approximators

Types of Machine Learning

- **Supervised Learning**: Learn from labeled data (regression, classification)
- **Unsupervised Learning**: Find patterns in unlabeled data (e.g., clustering, dimensionality reduction)
- **Reinforcement Learning**: Learn through reward-based interaction with the environment

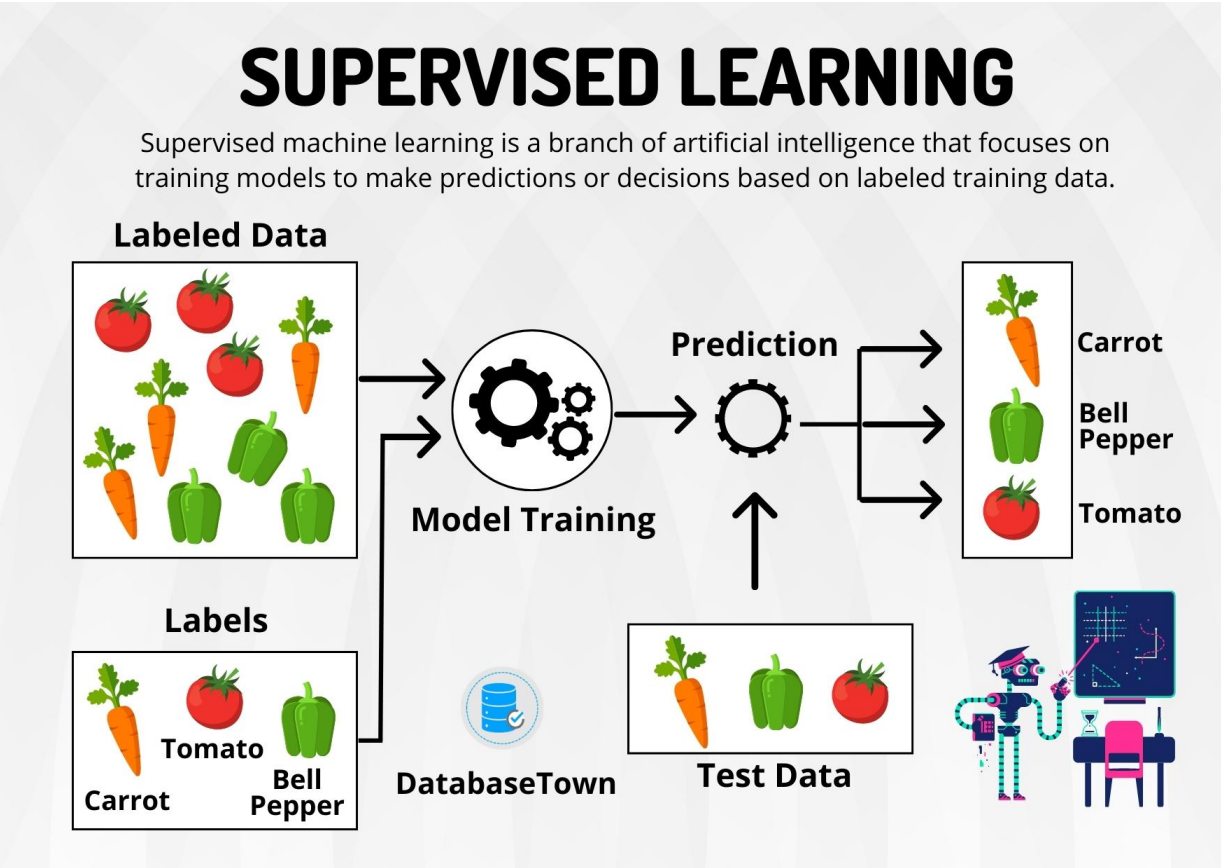


Types of Machine Learning



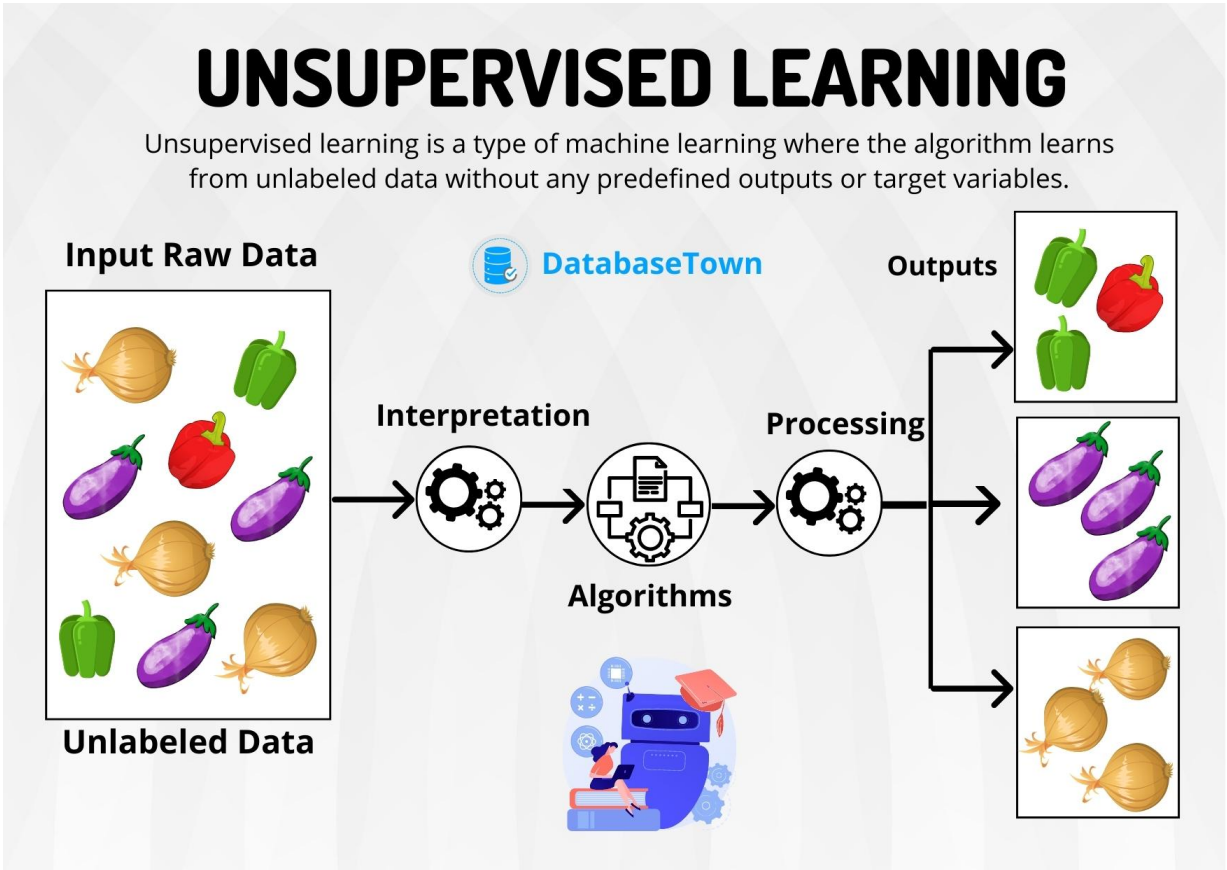
Supervised Learning

- Input and output pairs (labeled data)
- Train model to learn mapping: $f(x) = y$
- Examples: from structure to spectrum, acid or base



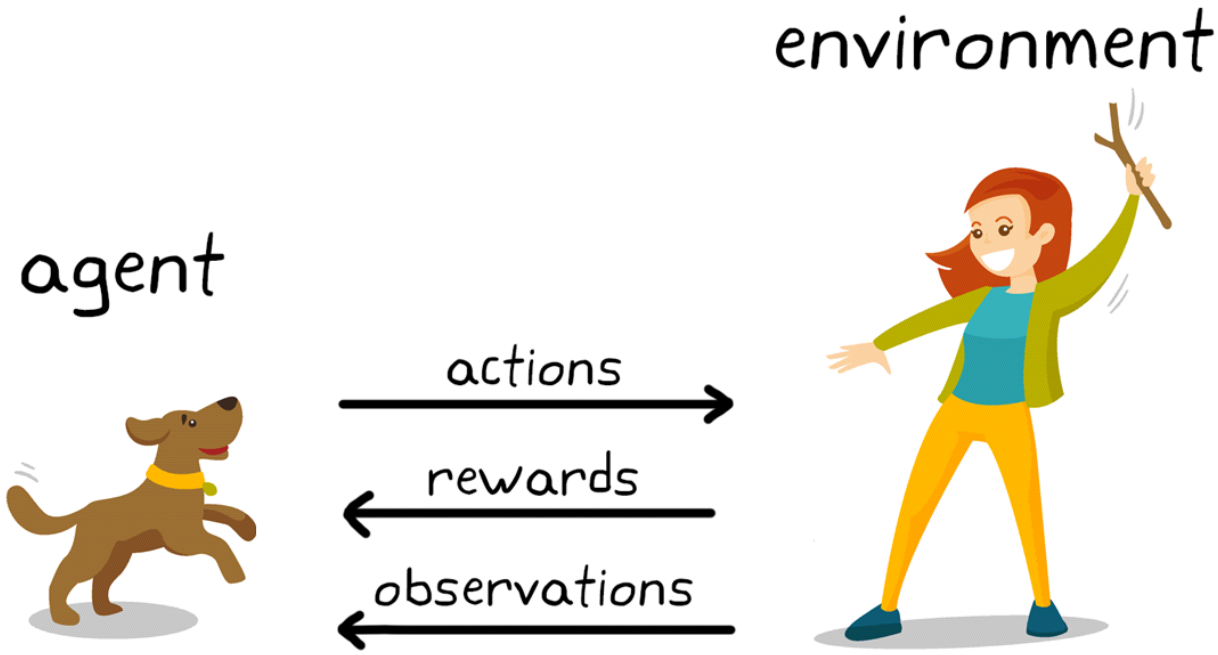
Unsupervised Learning

- No labeled output
- Aim: discover structure in data
- Examples: clustering from a MD trajectory



Reinforcement Learning

- Agent interacts with environment
- Learns to maximize cumulative reward
- Examples: game playing, robotics

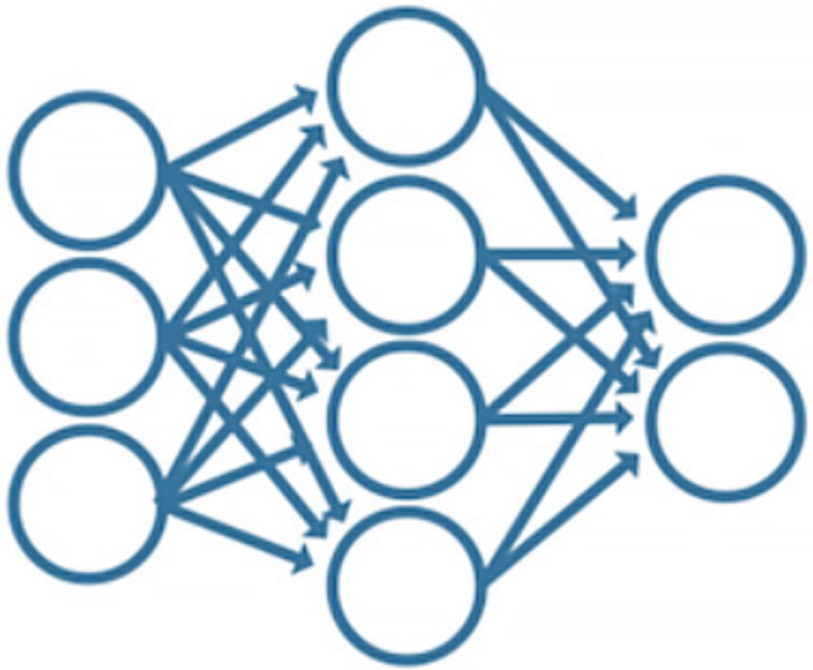


Common Algorithms

- Linear Regression
- Decision Trees
- k-Nearest Neighbors
- Support Vector Machines
- Neural Networks

What is a Neural Network?

- A neural network is a set of algorithms modeled after the human brain
- It is designed to recognize patterns from input data
- It consists of layers of interconnected neurons
  - Artificial neurons (nodes)
  - Weighted connections between neurons
  - Activation functions



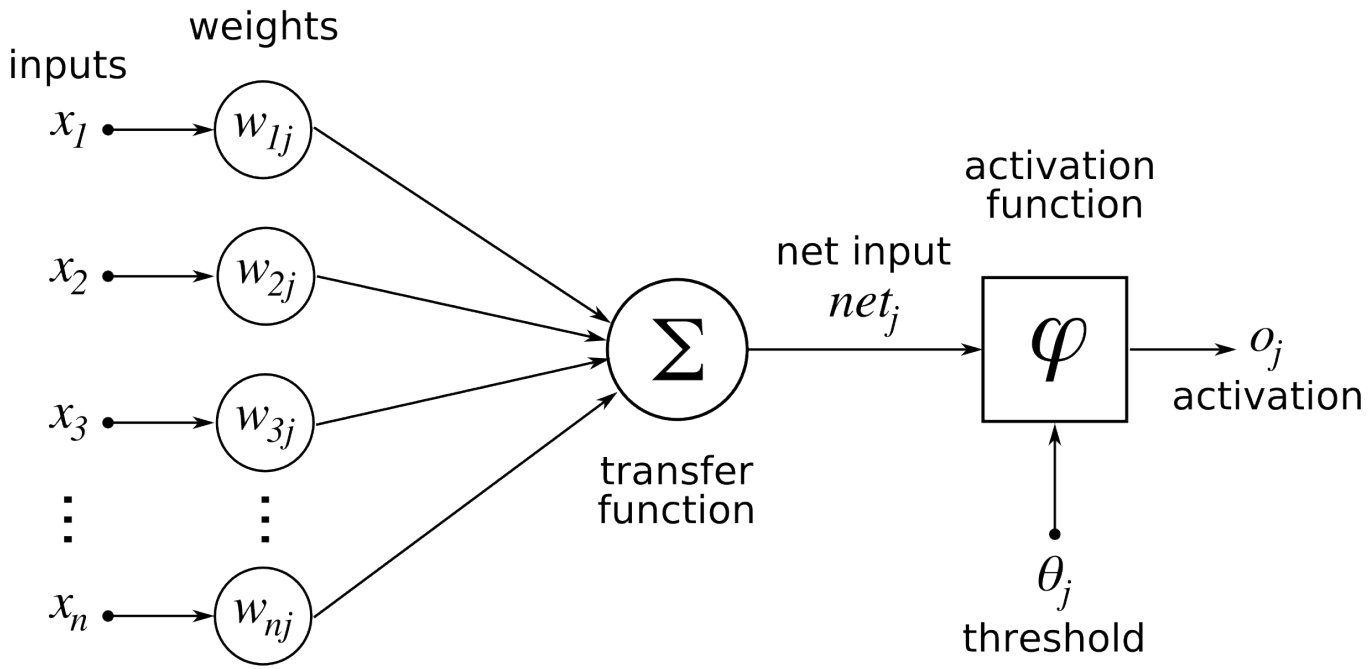
Basic Structure

- **Input layer**: Receives data
- **Hidden layers**: Process the data through weighted connections and activation functions
- **Output layer**: Produces final prediction or classification

Artificial Neuron

Each neuron performs:

- Inputs: $x_1, x_2, ..., x_n$
- Weights: $w_1, w_2, ..., w_n$
- A weighted sum of its inputs: $\sum w_i x_i + b$
- Output: $y = f(\sum w_i x_i + b)$



Activation Functions

- Add non-linearity to the model

Activation Functions

- Add non-linearity to the model
- Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$
- ReLU: $f(x) = \max(0, x)$
- Tanh: $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

Activation Functions

- Add non-linearity to the model
- Sigmoid: $\sigma(x) = \frac{1}{1+e^{-x}}$
- ReLU: $f(x) = \max(0, x)$
- Tanh: $\tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$

Training a Neural Network

1. **Forward Propagation**: Compute outputs layer by layer
2. **Loss Calculation**: Measure prediction error
3. **Backward Propagation**: Use gradients to adjust weights
4. **Optimization**: Apply updates using algorithms like gradient descent

- Common loss functions L:
  - Mean Squared Error
  - Cross-Entropy Loss
- Weight update rule:

$$w := w - \eta \cdot \frac{\partial L}{\partial w}$$

where η is the learning rate

Typical ML workflow

- Gathering and preparing data

Typical ML workflow

- Gathering and preparing data
- Choosing a representation

Typical ML workflow

- Gathering and preparing data
- Choosing a representation
- Training the model
  - Train model candidates
  - Evaluate model accuracy

Typical ML workflow

- Gathering and preparing data
- Choosing a representation
- Training the model
  - Train model candidates
  - Evaluate model accuracy
- Testing the model out of sample

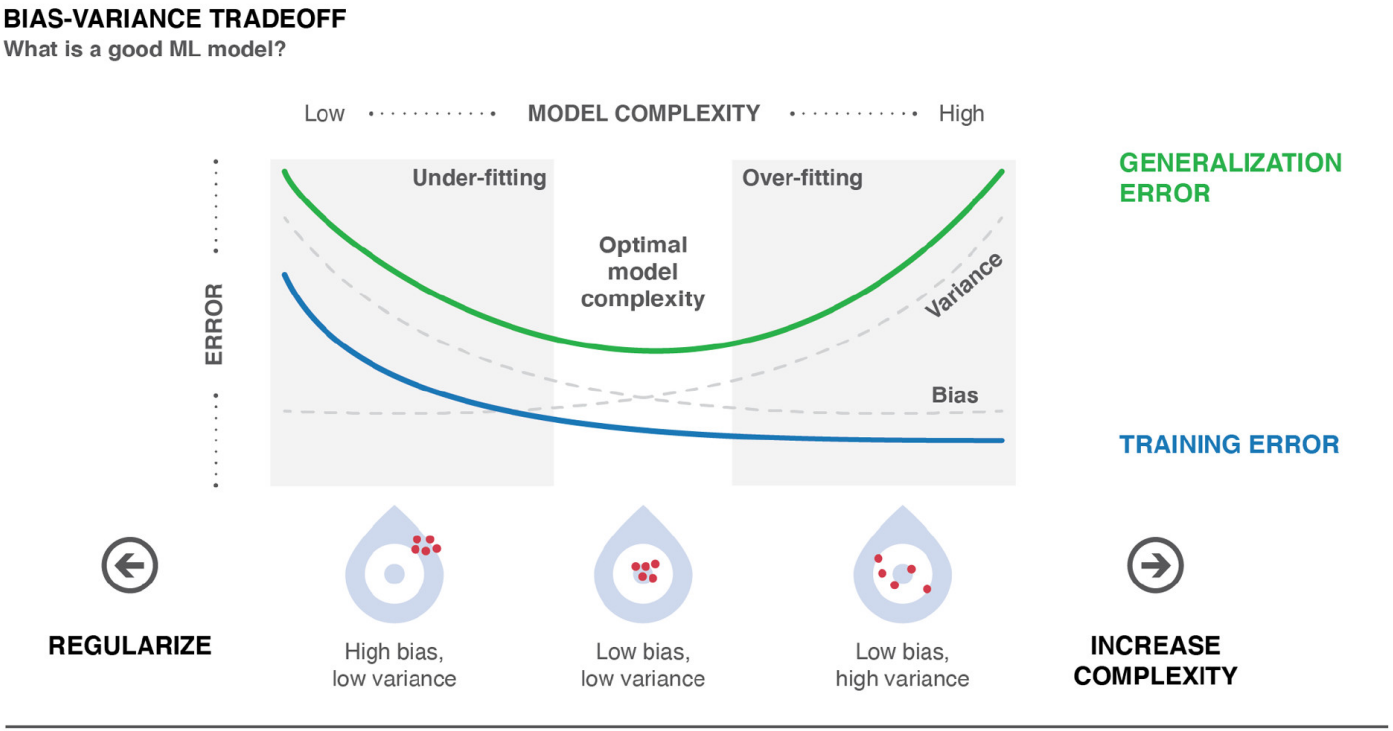
Typical ML workflow

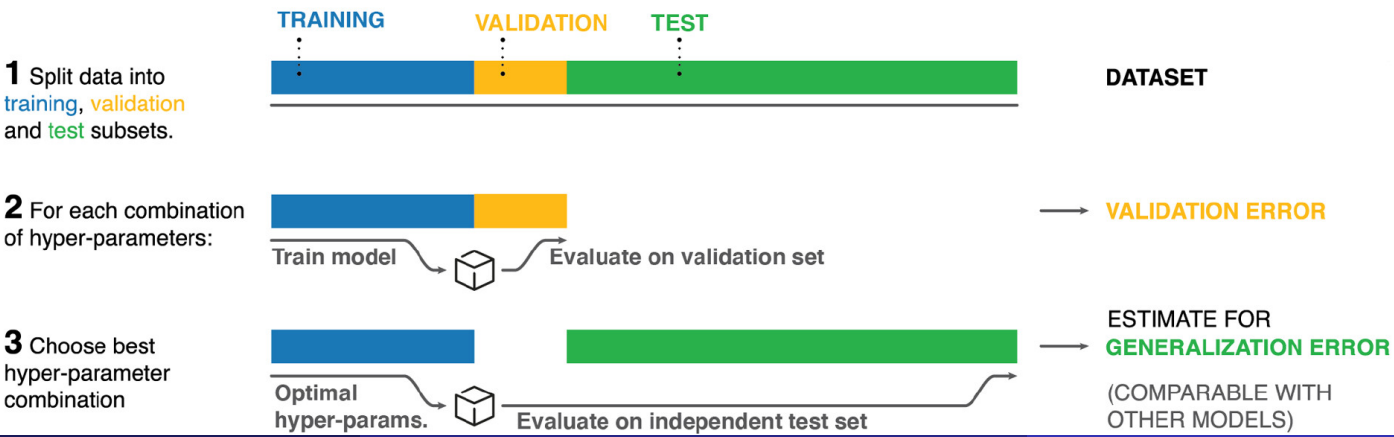
- Gathering and preparing data
- Choosing a representation
- Training the model
  - Train model candidates
  - Evaluate model accuracy
- Testing the model out of sample
- Deploy and monitor

Challenges in Machine Learning

- Quality and quantity of data
- Overfitting and underfitting
- Model interpretability
- Ethical and societal impacts

Machine learning in a nutshell





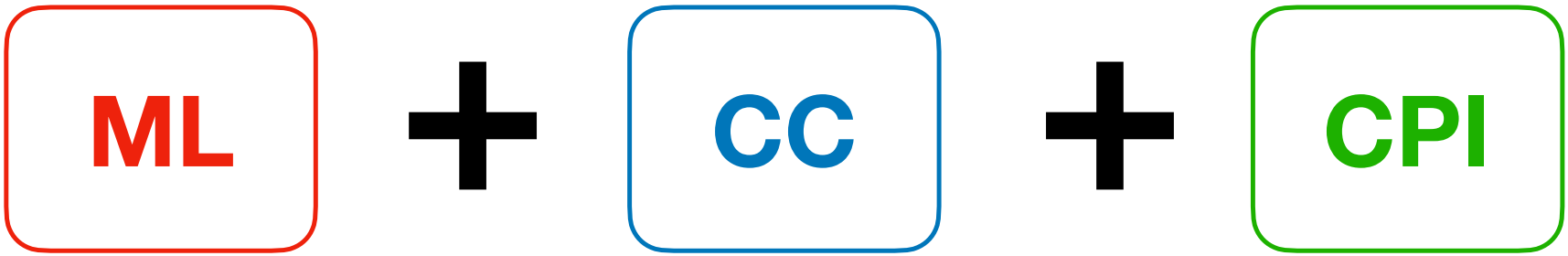
Machine learning in chemistry: overview

Machine learning models

- **Transformative impact on chemical sciences**
- **Dramatic acceleration of computations**
- **Amplifying insights available from chemistry methods**
- **Coaction of expertise in computer and physical/chemical sciences**

Machine learning in chemistry: overview

- ML = machine learning
- CC = computational chemistry
- CPI = chemical and physical intuition



- Well suited for nonlinear relationships
- Need robust data sets

- High-quality data
- Robust data sets

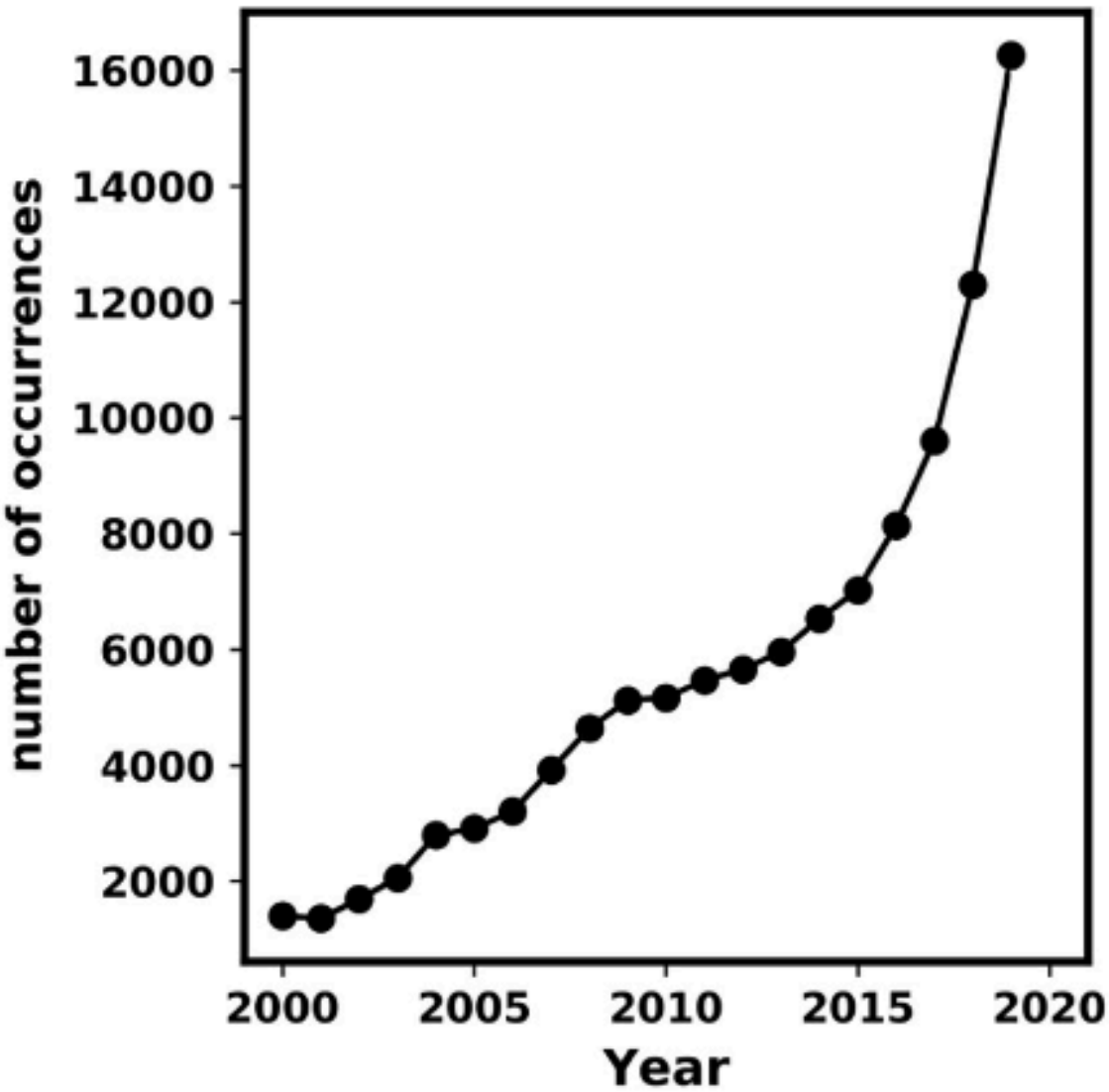
- Selecting appropriate methods
- Limited understanding

=

Catalyst accelerating data-driven hypotheses generation

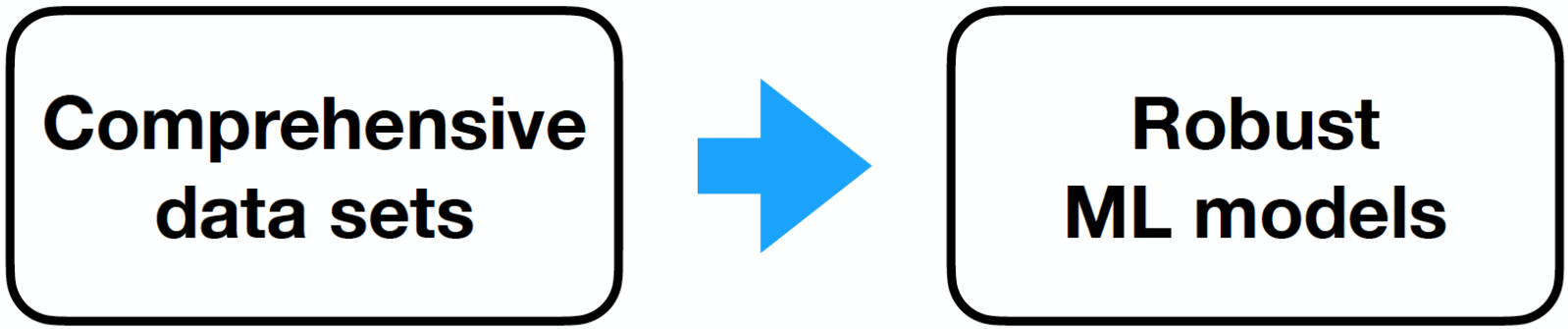
Machine learning in chemistry: overview

Occurrence of any ML term in American Chemical Society journals



Data sets

- ML models as sophisticated parametrizations of data sets
- Data set must be representative
- Quality of data set determines the model effectiveness
- Avoid/reduce biases or artifacts
- CPI to reduce the function space
- A priori removing of unphysical solutions

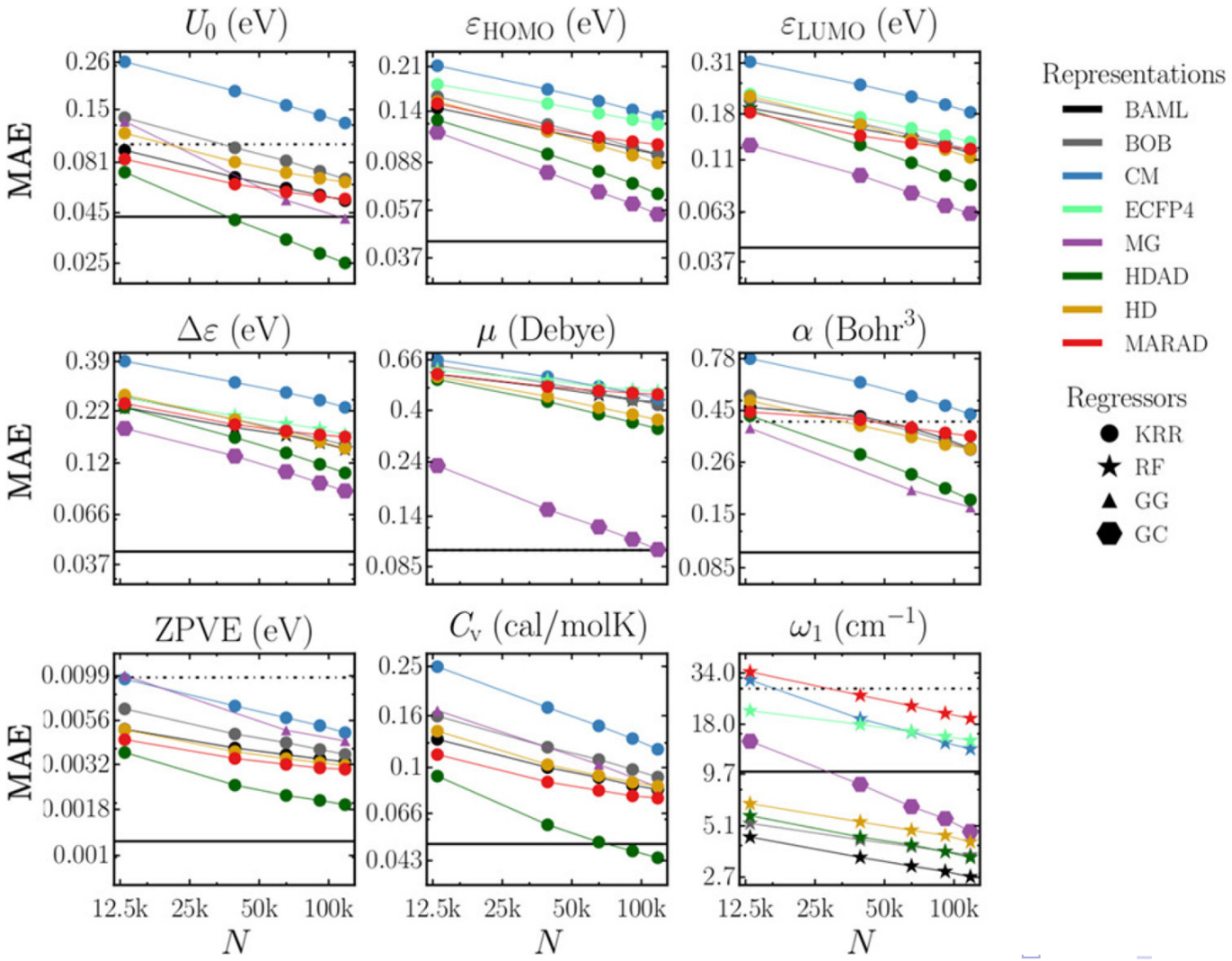


CC databases for ML

database	description	location
AFLOWLIB	databases containing calculated properties of over 625k materials[510]	http://www.aflowlib.org
ANI-1	large computational DFT database, which consists of more than 20 M off equilibrium conformations for 57.5k small organic molecules[511,512]	https://github.com/isayev/ANI1_dataset
ANI-1x/ANI-1ccx	ANI-1x contains multiple QM properties from 5 M DFT calculations, while ANI-1ccx contains 500k data points obtained with an accurate CCSD(T)/CBS extrapolation[513]	https://github.com/aiqm/ANI1x_datasets
BindingDB	measured binding affinities focusing on interactions of proteins considered to be candidates as drug-targets; 1 200 000 binding data for 5500 proteins and over 520 000 drug-like molecules[514]	http://www.bindingdb.org
Clean Energy Project	contains ~10 000 000 molecular motifs of potential interest which cover small molecule organic photovoltaics and oligomer sequences for polymeric materials[515]	http://cepdb.molecularspace.org
CoRE MOF	database containing over 4700 porous structures of metal−organic frameworks with publicly available atomic coordinates; includes important physical and chemical properties[516]	10.11578/1118280
FreeSolv	experimental and calculated hydration free energies for neutral molecules in water[517]	http://www.escholarship.org/uc/item/6sd403pz
GDB	GDB-11, GDB-13, and GDB-17; together these databases contain billions of small organic molecules following simple chemical stability and synthetic feasibility rules[518]	http://gdb.unibe.ch/downloads/
Hypothetical Zeolites	contains approximately 1 M zeolite structures[519]	http://www.hypotheticalzeolites.net/
Materials Project	contains computed structural, electronic, and energetic data for over 500k compounds[520]	https://www.materialsproject.org
MD17	data sets in this package range in size from 150k to nearly 1 M conformational geometries; all trajectories are calculated at a temperature of 500 K and a resolution of 0.5 fs[372]	http://www.sgdml.org
MoleculeNet	contains data on the properties of over 700k compounds[521]	http://moleculenet.ai
Open Catalyst Project	1.2 M molecular relaxations with results from over 250 M DFT calculations relevant for renewable energy storage[522]	https://opencatalystproject.org/index.html
OQMD	consists of DFT predicted crystallographic parameters and formation energies for over 200k experimentally observed crystal structures[523]	http://oqmd.org
PubChemQC PM6	provides 221 million molecular structures optimized with the PM6 method and several electronic properties computed at the same level of theory[524]	http://pubchemqc.riken.jp/pm6_datasets.html
PubChemQC	provides ~3 million molecular structures optimized by DFT and excited states for over 2 million molecules using TD-DFT[525]	http://pubchemqc.riken.jp/
QM7-X	comprehensive data set of 42 physicochemical properties for ~4.2 M equilibrium and nonequilibrium structures of small organic molecules with up to seven non-hydrogen (C, N, O, S, Cl) atoms[526]	https://zenodo.org/record/4288677#.X9jHNC2ZNTY
QM9	geometric, energetic, electronic, and thermodynamic properties for 134k stable small organic molecules out of GDB-17[527]	https://figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules/978904
Synthesis Project	collection of aggregated synthesis parameters computed using the text contained within over 640 000 journal articles[528]	www.synthesisproject.org
quantum-machine.org	a repository of diverse data sets, including valence electron densities, chemical reactions, solvated protein fragments, and molecular Hamiltonians	http://quantum-machine.org/datasets/

Benchmarking data sets

- Learning curves (QM9 database, 134K organic molecules)
- Target and DFT/B3LYP accuracy



Applications

- Searching stationary points of a PES
- Generating force fields for MM and MD
- Use in metadynamics (collective variables)
- Chemometrics
- Text mining for extracting scientific information
- Structure/property relationship in spectroscopies
- Retrosynthesis
- Materials
- Drug design
- ...

Applications: vibrational spectroscopy

- One-to-one spectrum-structure relationships
- Conventionally with CC

Applications: vibrational spectroscopy

- One-to-one spectrum-structure relationships
- Conventionally with CC
- Machine-learning protocol to correlate spectral fingerprints with local molecular structures
  - Quick and accurate prediction of infrared (IR) and Raman spectra
  - Structure recognition of functional groups from vibrational spectral features

IR and Raman with quantum chemistry

- Vibrational modes
  - Diagonalization of the mass-weighted Hessian matrix
  - Eigenvectors: normal modes **q**
  - Eigenvalues: frequencies
  - Harmonic approximation

IR and Raman with quantum chemistry

- Vibrational modes
  - Diagonalization of the mass-weighted Hessian matrix
  - Eigenvectors: normal modes $\mathbf{q}$
  - Eigenvalues: frequencies
  - Harmonic approximation
- IR
  - Change in the dipole moment $\boldsymbol{\mu}$

$$\text{IR intensity} \propto \left(\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{q}} \right)^2$$

IR and Raman with quantum chemistry

- Vibrational modes
  - Diagonalization of the mass-weighted Hessian matrix
  - Eigenvectors: normal modes $\mathbf{q}$
  - Eigenvalues: frequencies
  - Harmonic approximation
- IR
  - Change in the dipole moment $\boldsymbol{\mu}$

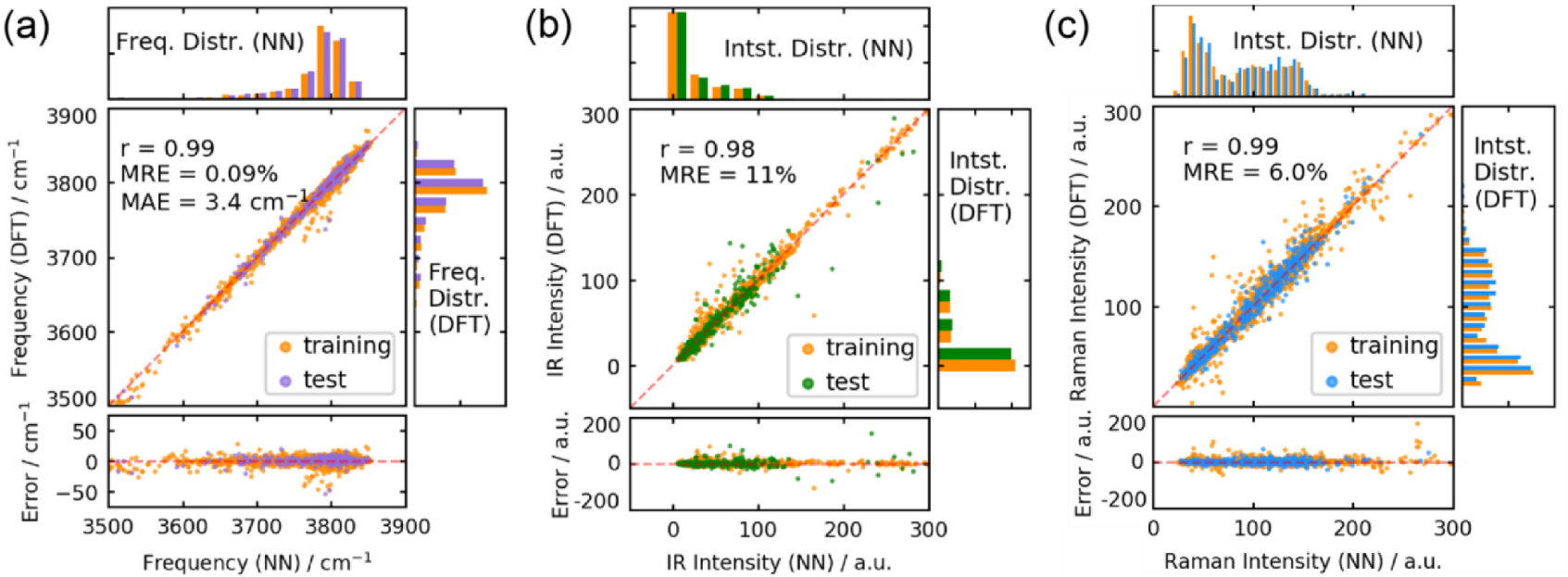
$$\text{IR intensity} \propto \left(\frac{\partial \boldsymbol{\mu}}{\partial \mathbf{q}}\right)^2$$

- Raman
  - Change in the polarizability $\boldsymbol{\alpha}$

$$\text{Raman intensity} \propto \left(\frac{\partial \boldsymbol{\alpha}}{\partial \mathbf{q}}\right)^2$$

Applications: vibrational spectroscopy

- Hydroxyl (OH, 3000-4000 cm^{-1}) and carbonyl (C=O, 1400-2000 cm^{-1}) groups
- Dataset with around 21,000 molecules
- Spectra with DFT/ B3LYP/6-31G(2df,p)

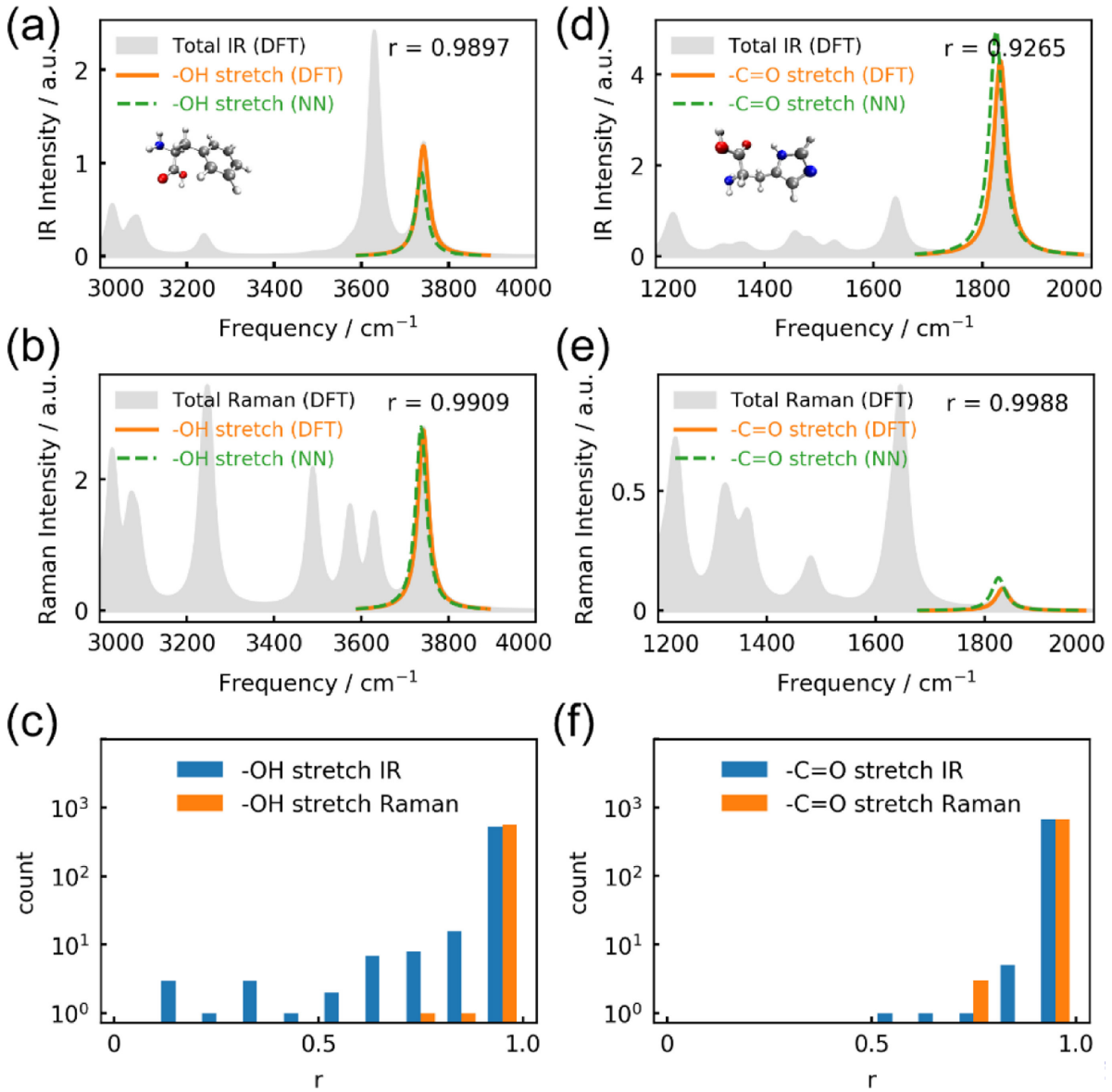


$$MAE = \frac{1}{n}\sum_i |x_i - x_{ref,i}|$$

$$MRE = \frac{1}{n}\sum_i |x_i - x_{ref,i}|/|x_{ref}|$$

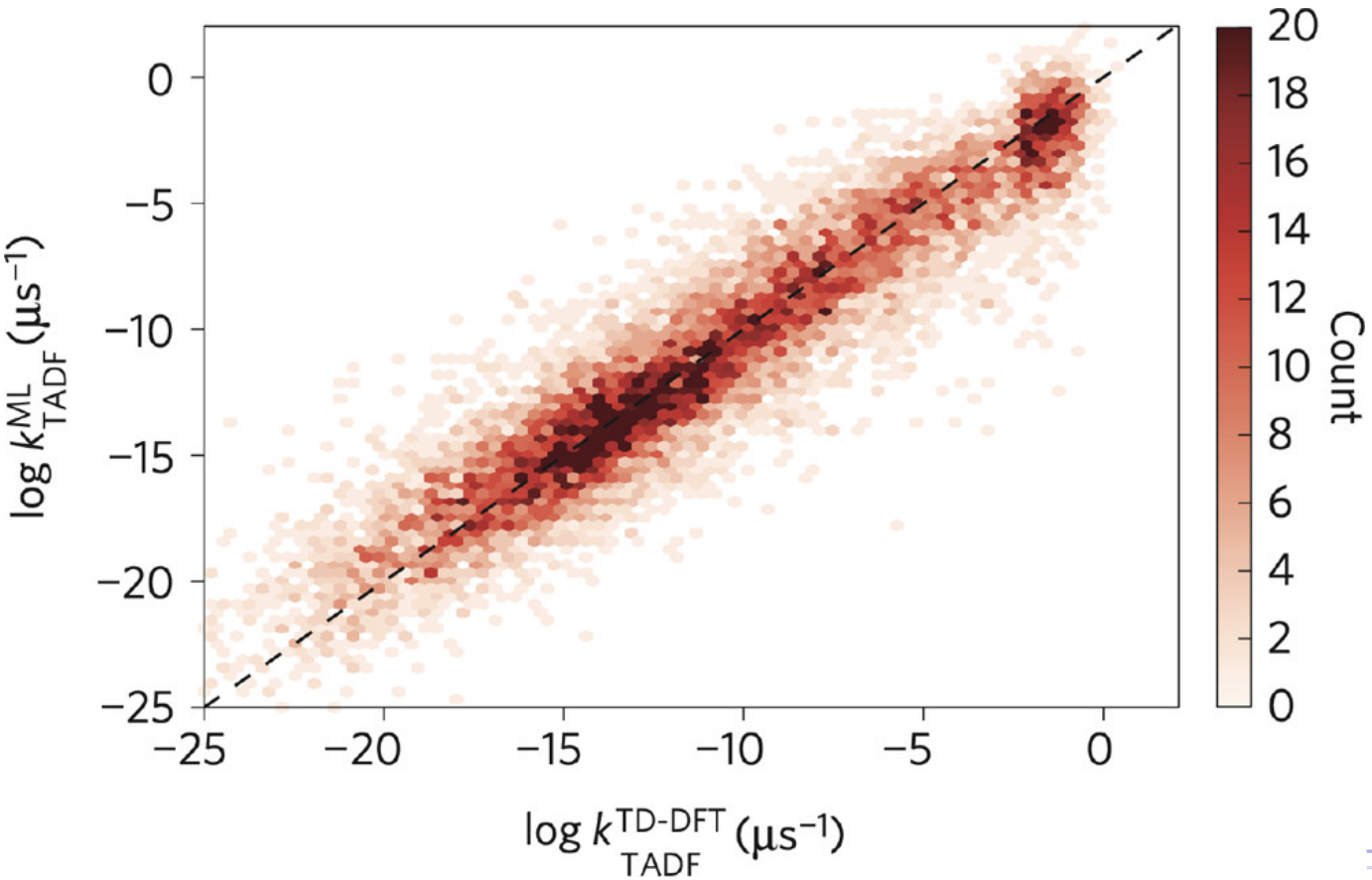
Applications: vibrational spectroscopy

Histidine (left) and phenylalanine (right)

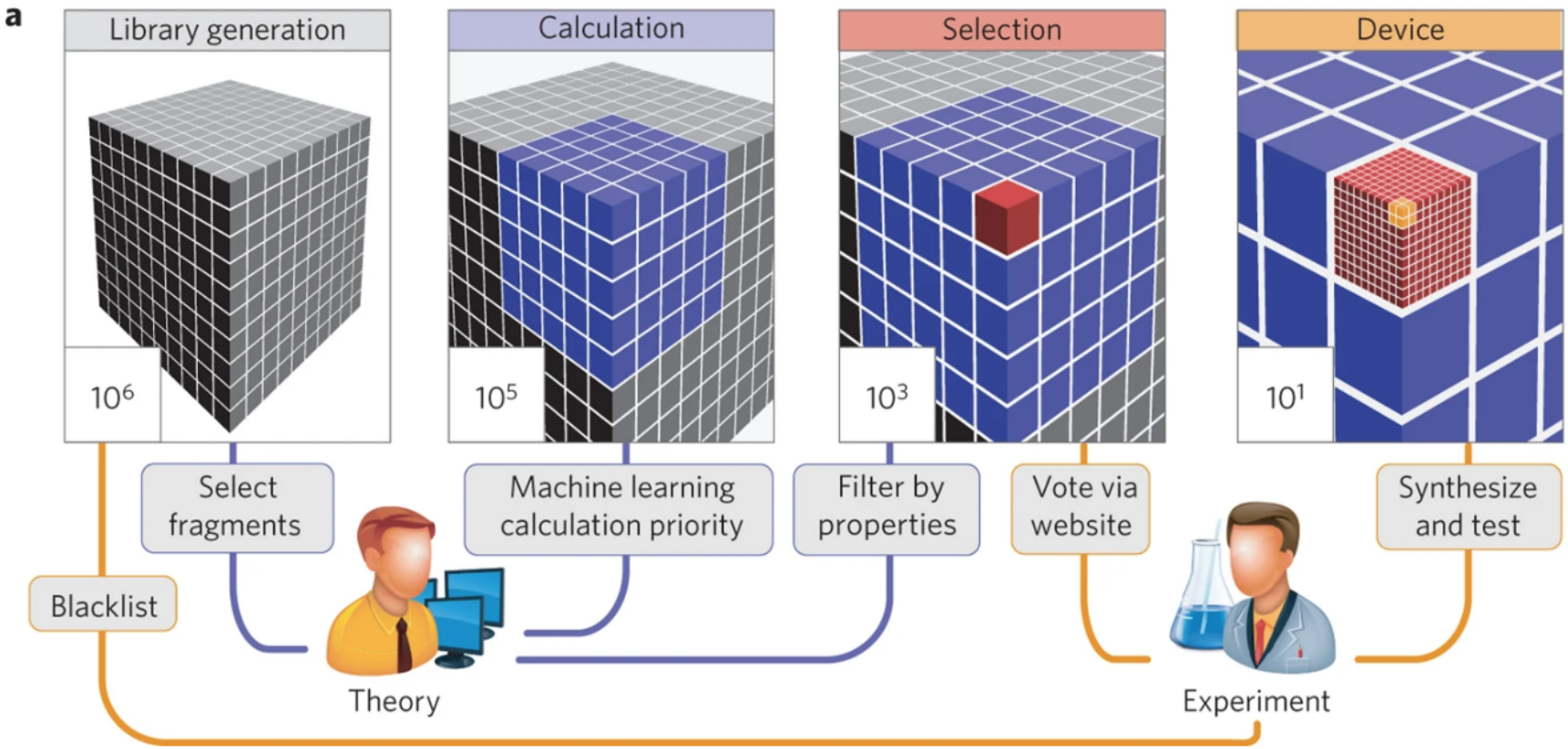


Applications: molecular and material design

- Identify compounds with desired properties (high-throughput screening)
- OLED emitters (k_{TADF} delayed fluorescence rate constant)
- ML comparable to CC calculations, at a fraction of the computational cost

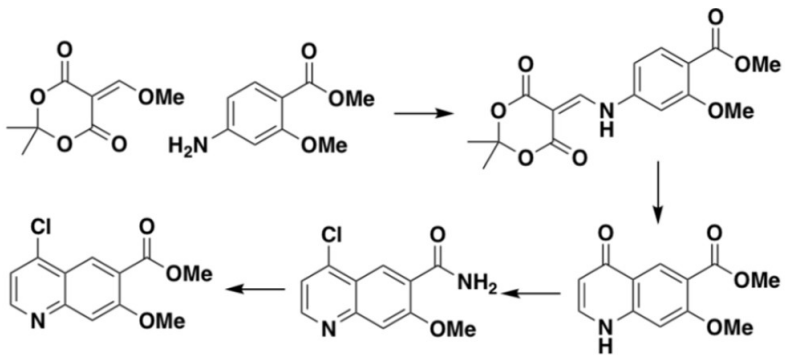


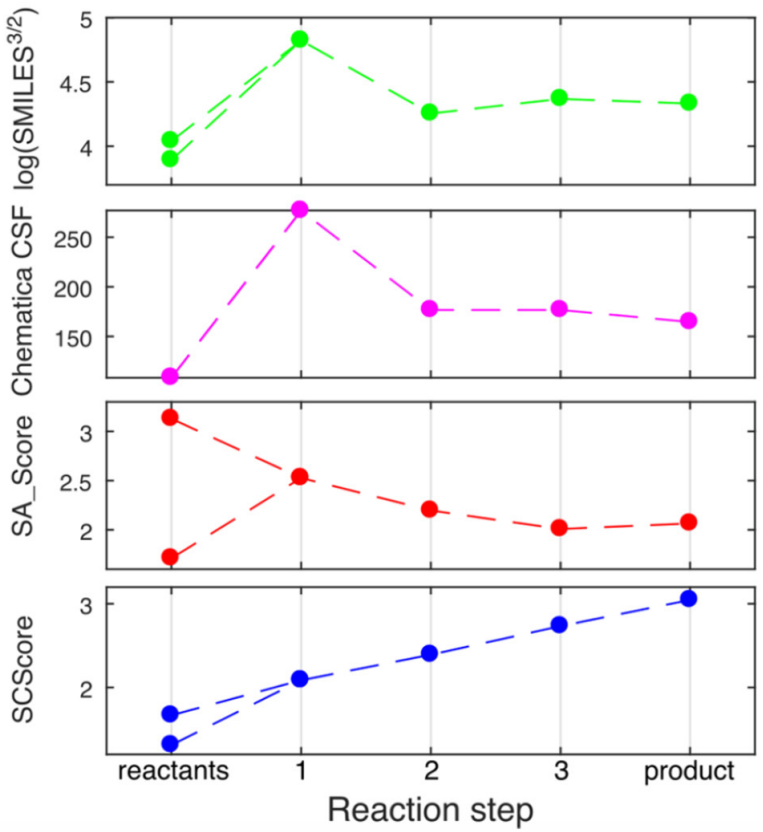
Applications: molecular and material design



Applications: retrosynthesis

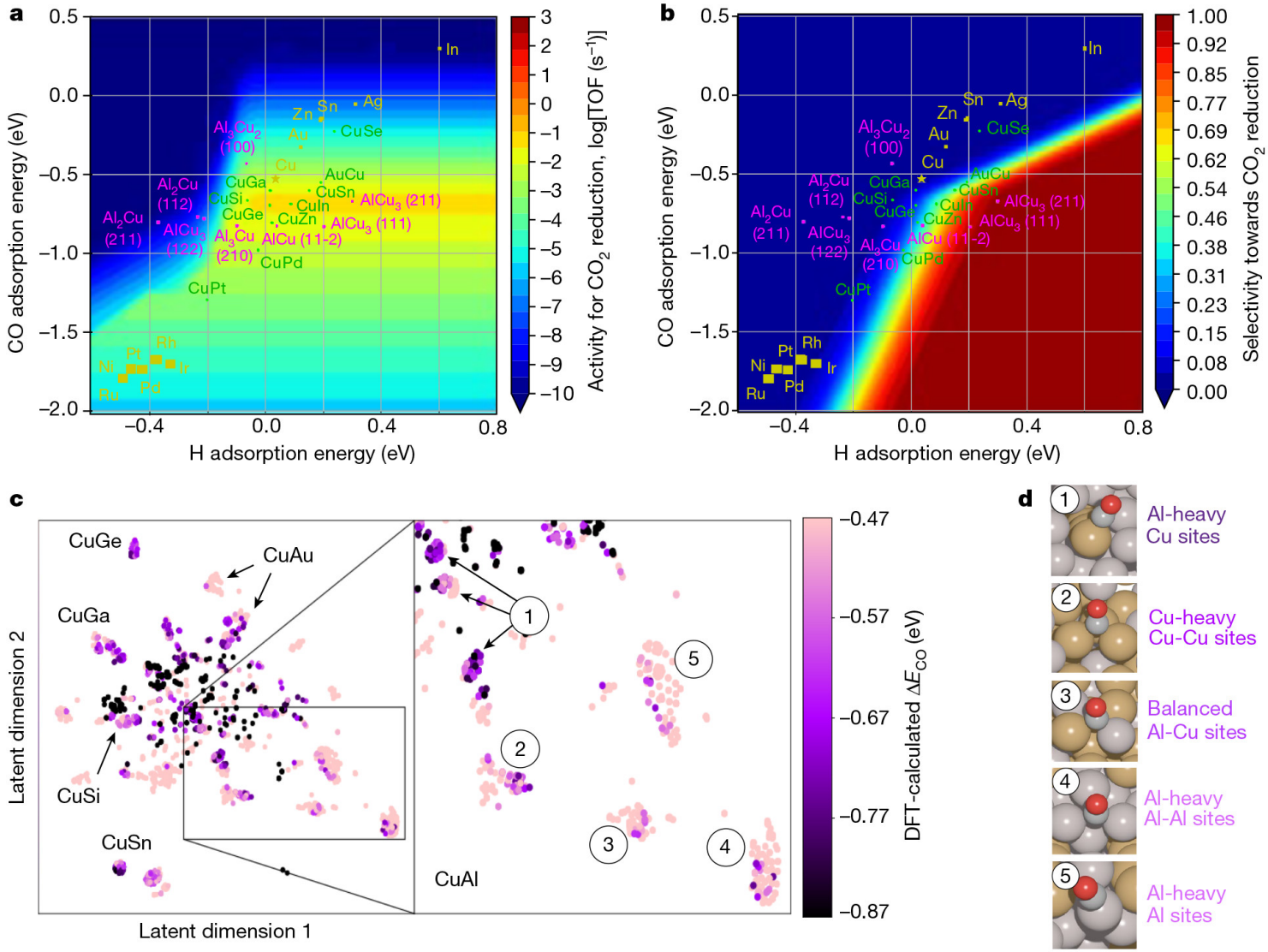
- Design of chemical steps
- SCScore: data-driven metric specific for reactions
- Monotonic increase in complexity with SCScore
- ML to overcome the generalization issues of rule-based algorithms
- Synthesis of a precursor to lenvatinib





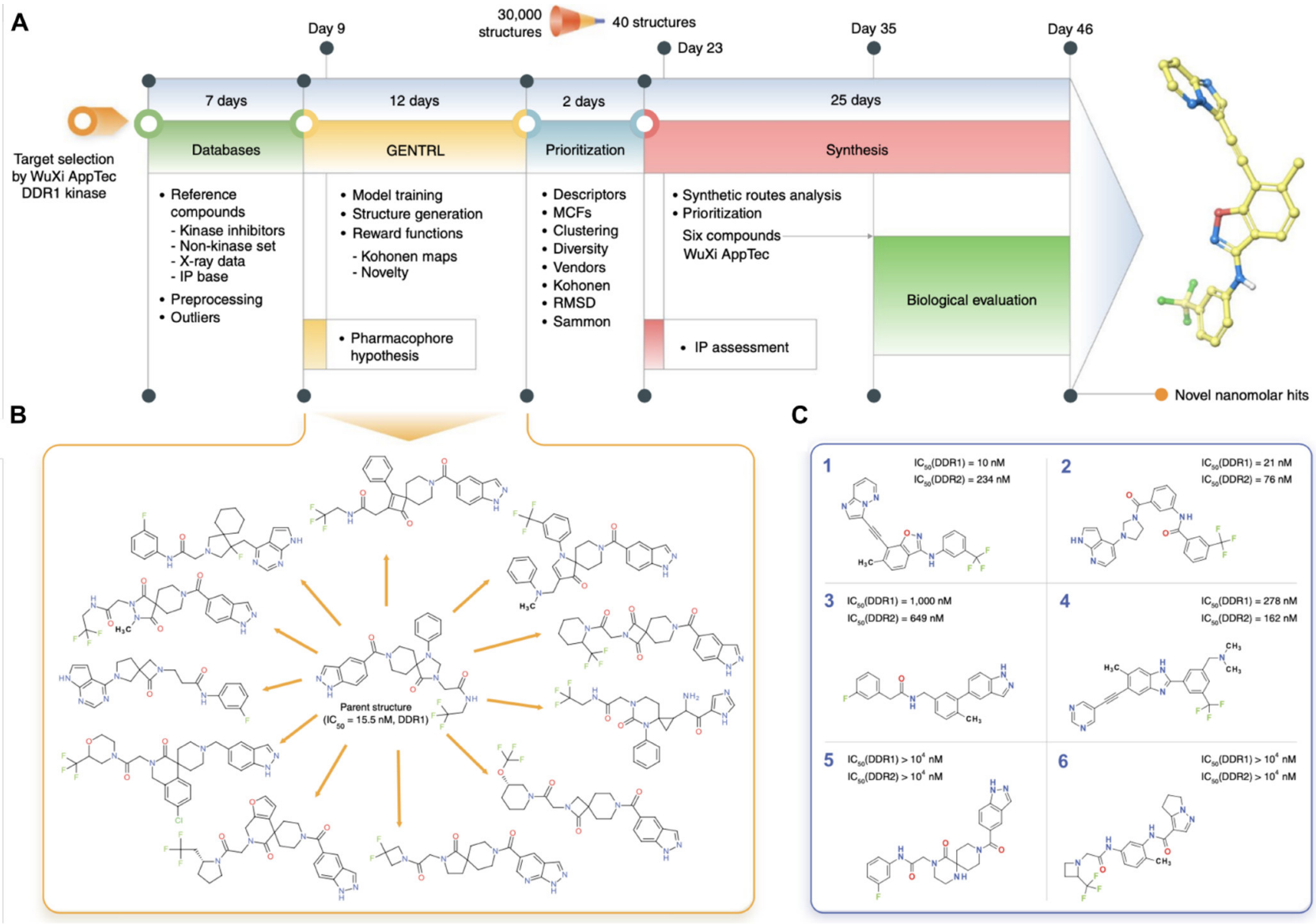
Applications: catalysis

Accelerated discovery of CO_2 electrocatalysts using ML



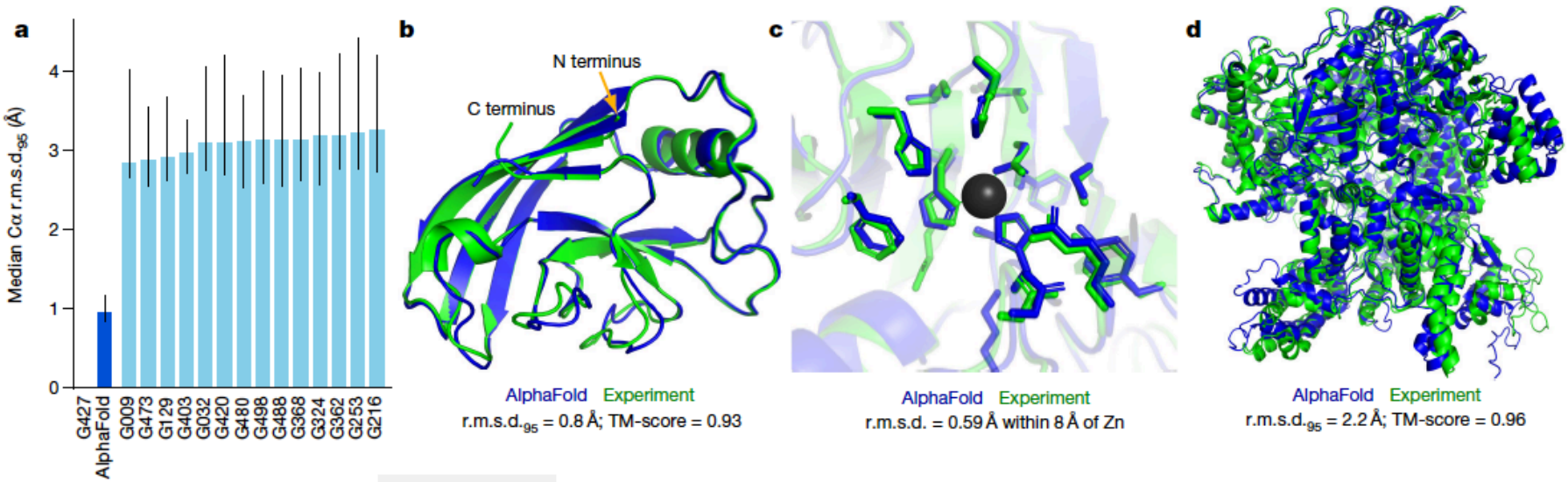
Applications: drug design

Discoidin domain receptor 1 (DDR1)



Applications: protein structures (AlphaFold 2)

- Experimental efforts have yielded $\approx 100,000$ unique protein structures
- Tiny fraction of the billions of known protein sequences
- The experimental determination of a single structure often requires months to years
- Long-standing goal of computational biology: predict the three-dimensional structure based solely on the amino acid sequence



Applications: protein structures (AlphaFold 2)

Key Accuracy Metrics

- AlphaFold 2 was validated in the 14th Critical Assessment of protein Structure Prediction
- It demonstrated accuracy competitive with experimental structures in a majority of cases, vastly outperforming other competing methods
- AlphaFold 2 achieved a median backbone accuracy of **0.96 r.m.s.d.$_{95}$** (C_α root-mean-square deviation at 95% residue coverage)
- The next best method achieved 2.8 r.m.s.d.$_{95}$
- AlphaFold all-atom accuracy was 1.5 r.m.s.d.$_{95}$ compared to 3.5 r.m.s.d.$_{95}$ for the best alternative method
- The method is scalable, successfully predicting a 2180-residue protein with accurate domain packing