Machine Learning Meets (Quantum) Chemistry: Introduction and Short Overview

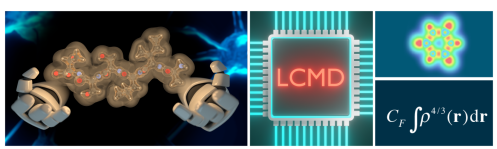
K. Briling, P. van Gerwen, A. Fabrizio, B. Meyer, R. Fabregat

École Polytechnique Fédérale de Lausanne
Laboratory for Computational Molecular Design (LCMD)
Prof. Clémence Corminboeuf

ksenia.briling@epfl.ch

12 May 2025

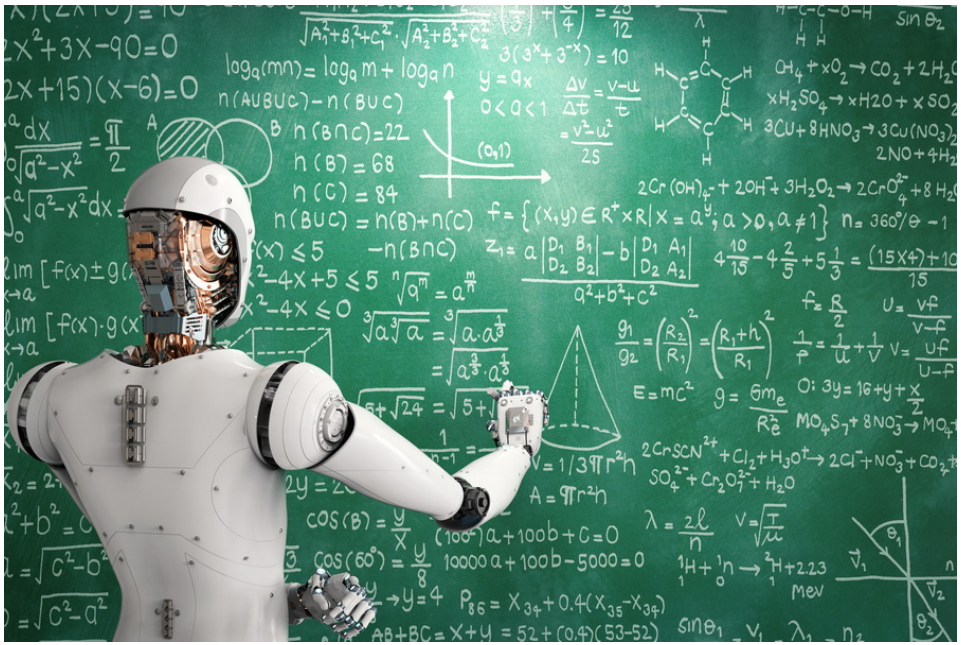
EPFL



Introduction

Introduction: why ML for chemistry?

Can a computer learn (quantum) chemistry?

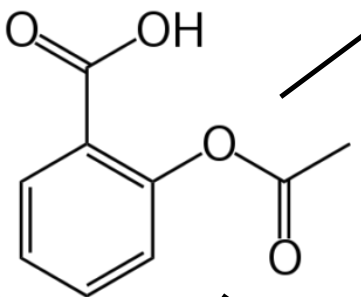


Learn what?

- For example, learn to predict the (quantum) chemical properties from molecular structural formulas

Pharmacological properties ?

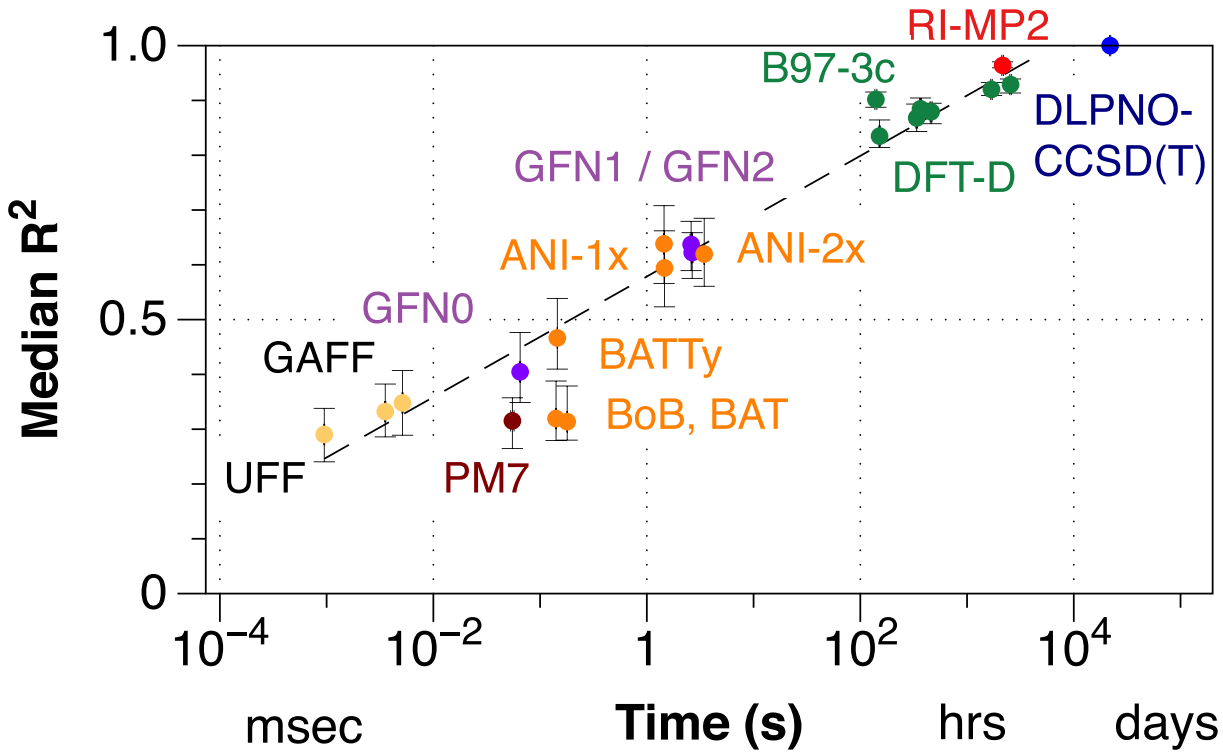
How does it react with another compound?



How are its spectra (Infrared, NMR,...)

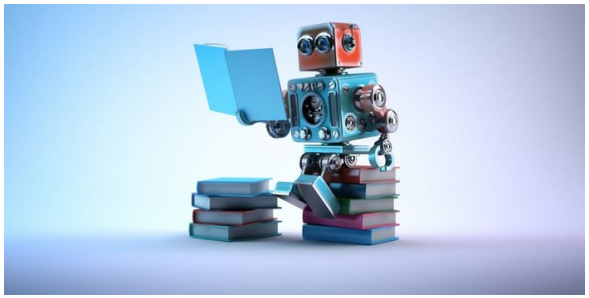
Is its solid form stable?

Learn why?



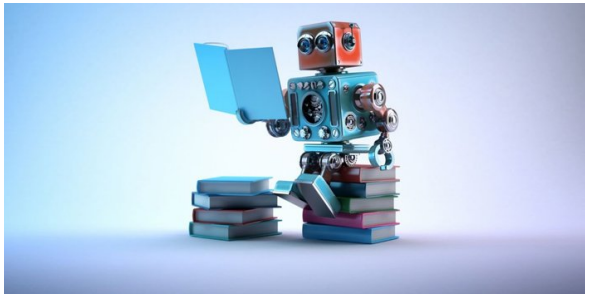
D. Folmsbee and G. Hutchison, "Assessing conformer energies using electronic structure and machine learning methods", Int. J. Quantum Chem. **121**, e26381 (2020)

Learn how?



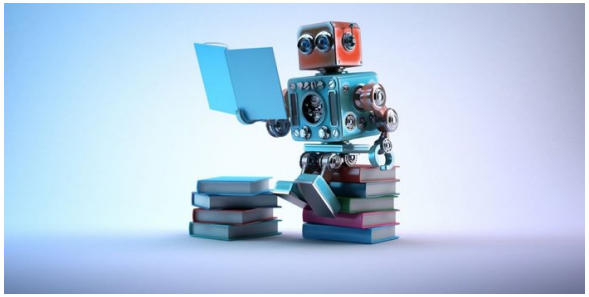
- A computer can learn, just like first experimental chemists have typically learned chemistry...

Learn how?



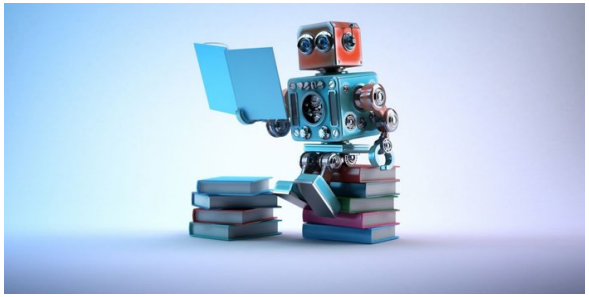
- A computer can learn, just like first experimental chemists have typically learned chemistry...
- Empirically: Based on their experience as opposed to *a priori* theoretical knowledge

Learn how?



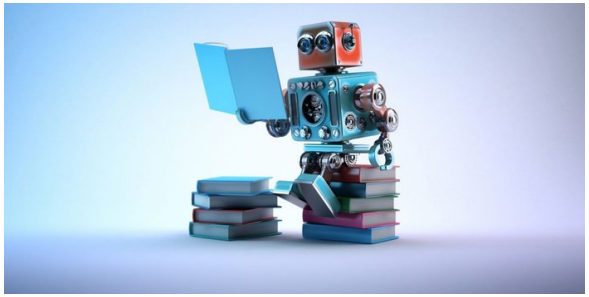
- A computer can learn, just like first experimental chemists have typically learned chemistry...
- Empirically: Based on their experience as opposed to *a priori* theoretical knowledge
- Starting from a dataset of theoretical or experimental data with molecular structures and the corresponding observable properties

Learn how?



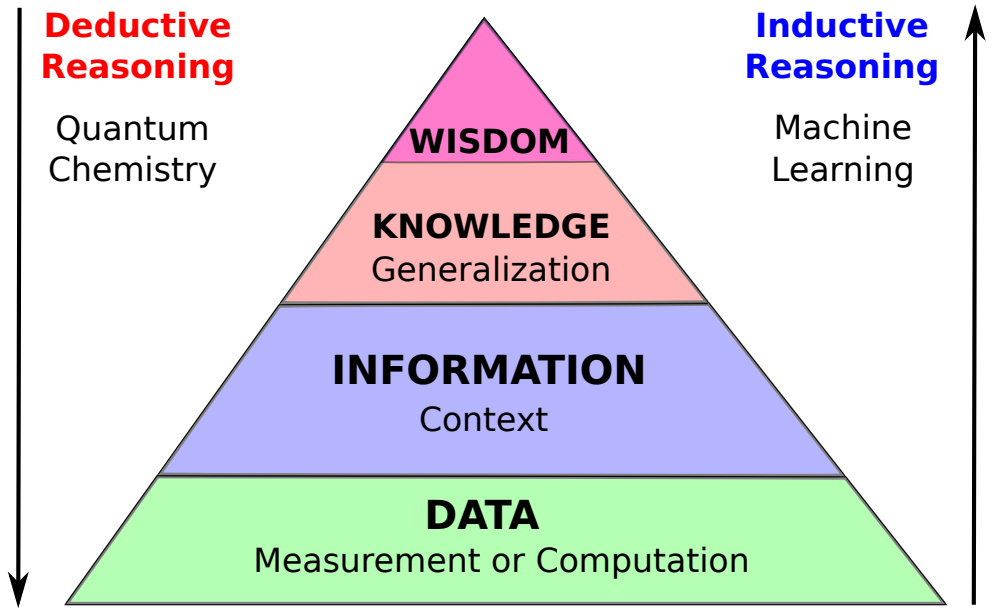
- A computer can learn, just like first experimental chemists have typically learned chemistry...
- Empirically: Based on their experience as opposed to *a priori* theoretical knowledge
- Starting from a dataset of theoretical or experimental data with molecular structures and the corresponding observable properties
- The computer can find relationships between molecular structures and the properties

Learn how?



- A computer can learn, just like first experimental chemists have typically learned chemistry...
- Empirically: Based on their experience as opposed to *a priori* theoretical knowledge
- Starting from a dataset of theoretical or experimental data with molecular structures and the corresponding observable properties
- The computer can find relationships between molecular structures and the properties
- It learns! And can apply the knowledge to new situations!

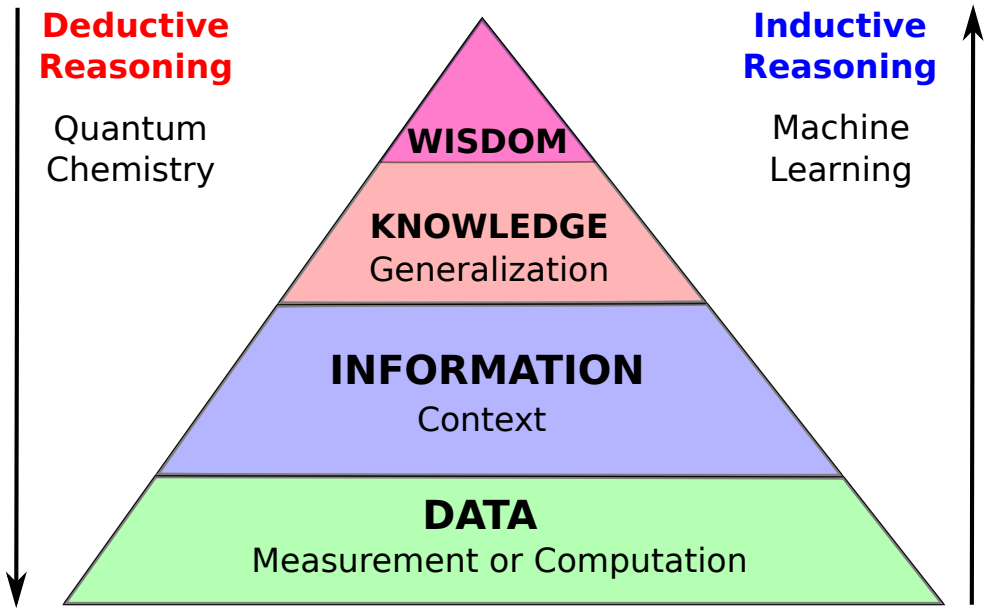
From data to knowledge



- ML is traditionally used in quantitative structure property (or activity) relationships (QSPR/QSAR)

O. A. von Lilienfeld, "Quantum machine learning in chemical compound space", Angew. Chem. Int. Ed. **57**, 4164–4169 (2018)

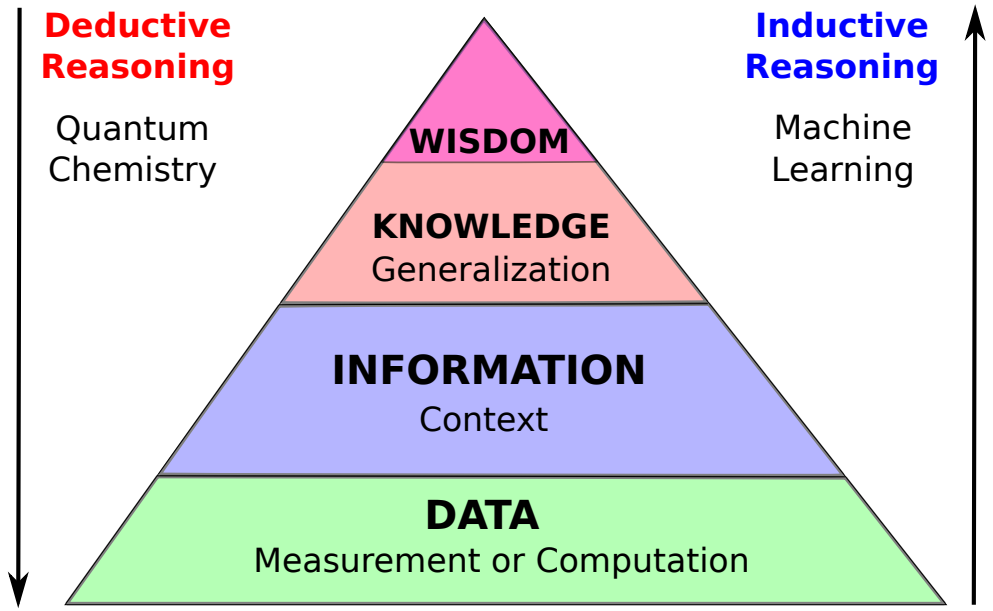
From data to knowledge



- ML is traditionally used in quantitative structure property (or activity) relationships (QSPR/QSAR)
- Can we unify the two worlds to extend the current quantum chemistry toolbox by "quantum machine learning" (QML)?

O. A. von Lilienfeld, "Quantum machine learning in chemical compound space", Angew. Chem. Int. Ed. **57**, 4164–4169 (2018)

From data to knowledge



- ML is traditionally used in quantitative structure property (or activity) relationships (QSPR/QSAR)
- Can we unify the two worlds to extend the current quantum chemistry toolbox by "quantum machine learning" (QML)?
- QML: classical machine learning applied to quantum-chemical properties

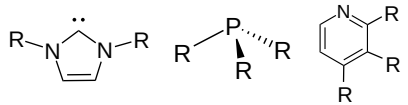
---

O. A. von Lilienfeld, "Quantum machine learning in chemical compound space", Angew. Chem. Int. Ed. **57**, 4164–4169 (2018)

ML meets quantum chemistry: guideline

Catalysts Library

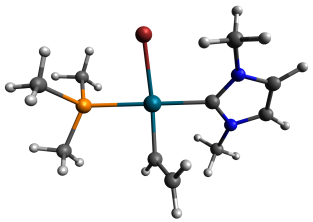
$L_1—M—L_2$

L_1 / L_2 = 

M = Ni Pd Pt Cu Ag Au

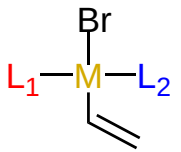
P(C)(C)(C)[Pd@SP1](*)(*)C1(=N(C)CCN(C)1)

25,116 catalysts



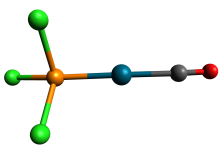
3D Geometry

ab initio Geometries and Energies for

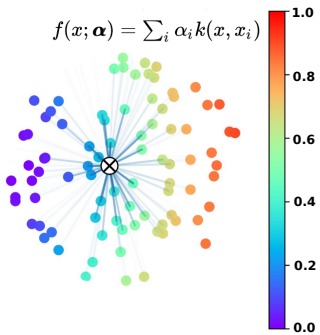
$L_1—M—L_2$ and 

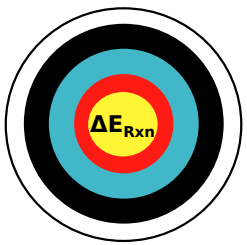
E_{el}(**1**) E_{el}(**2**)

Machine Learning Models



Descriptors





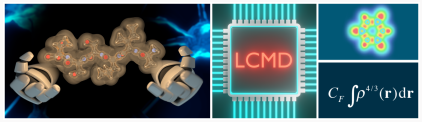
Targeted Properties

$\Delta E_{RxnA} = E_{el}(\mathbf{2}) - E_{el}(\mathbf{1})$

A. Fabrizio et al., "Quantum chemistry meets machine learning", CHIMIA **73**, 983 (2019)

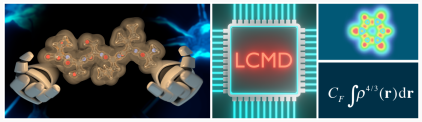
Machine learning methods: a short overview

Three paradigms of ML



- Supervised learning (*Classification*, *Regression*)
  Find relations between your data (molecules) and a target variable (chemical properties) that you want to be able to predict

Three paradigms of ML



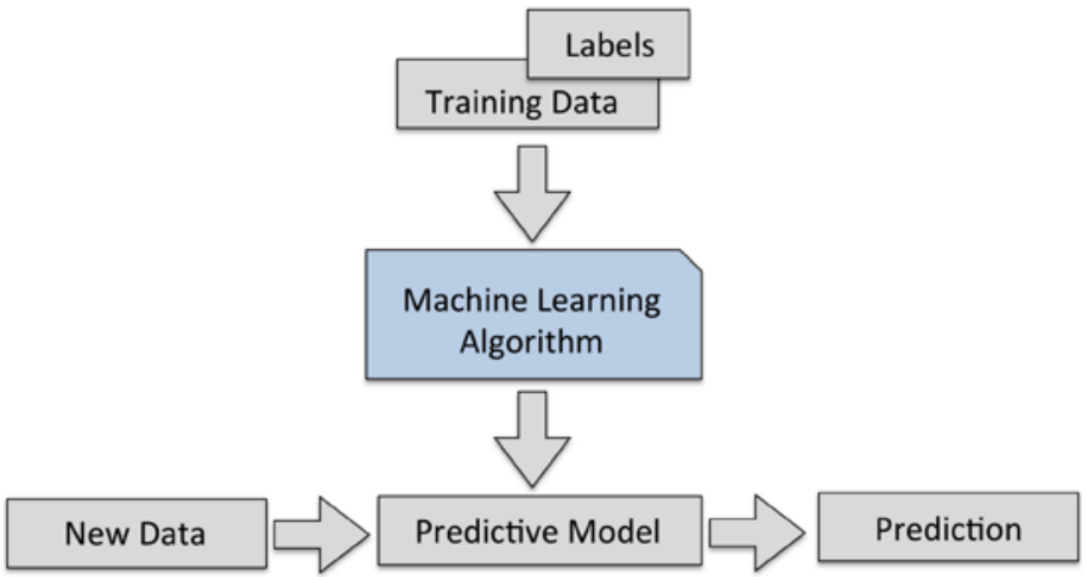
- Supervised learning (*Classification, Regression*)
  Find relations between your data (molecules) and a target variable (chemical properties) that you want to be able to predict
- Unsupervised learning (*Clustering, Dimensionality Reduction*)
  Searching for indirect hidden structures, clusters, patterns or features in the raw (molecular) data

Three paradigms of ML

- Supervised learning (*Classification*, *Regression*)
  Find relations between your data (molecules) and a target variable (chemical properties) that you want to be able to predict
- Unsupervised learning (*Clustering*, *Dimensionality Reduction*)
  Searching for indirect hidden structures, clusters, patterns or features in the raw (molecular) data
- Reinforcement learning
  Solving interactive problems with an environment (*e.g.*, a chess engine)
  Close to supervised learning but reward (feedback) instead of labels

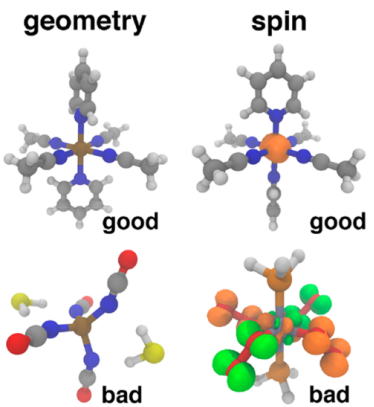
Supervised ML:
Making predictions about the future

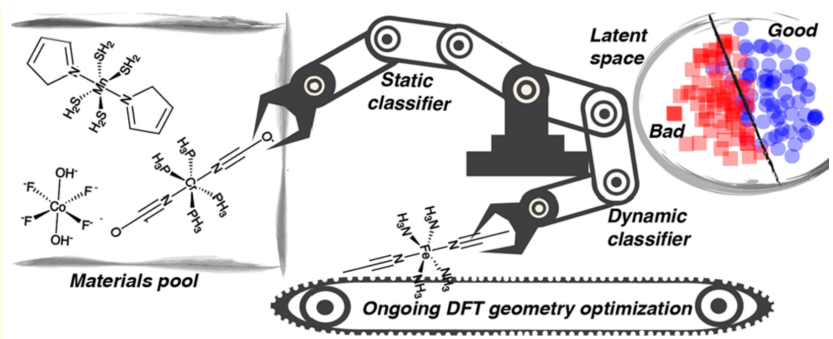
- In supervised learning the model is constructed from *training* molecules with known properties that allows us to make predictions about unseen molecules
- *Classification* task: assign discrete class labels
- *Regression* task: the chemical property is a continuous value

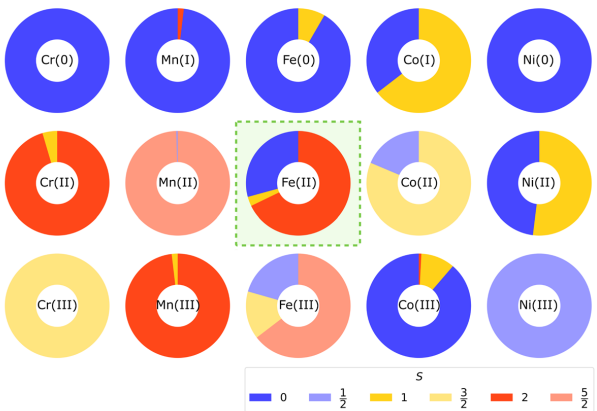


Supervised ML: Classification

- Goal: predict the categorical labels of new molecules based on past observations
- Applications: electronic structure computations outcomes (good/bad), ground state spins, ...

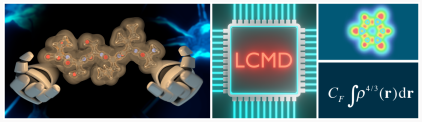






C. Duan et al., "Learning from failure: Predicting electronic structure calculation outcomes with machine learning models", J. Chem. Theory Comput. **15**, 2331–2345 (2019); Y. Cho et al., "Automated prediction of ground state spin for transition metal complexes", Digital Discovery **3**, 1638–1647 (2024)

Supervised ML: Regression



- Task: prediction of continuous properties
- Applications: heat capacity at room temperature, HOMO–LUMO gap, receptor–ligand binding, stability of molecular conformers, . . .

M. Rupp, O. A. von Lilienfeld, and K. Burke, "Guest editorial: Special topic on data-enabled theoretical chemistry", J. Chem. Phys. **148**, 241401 (2018)

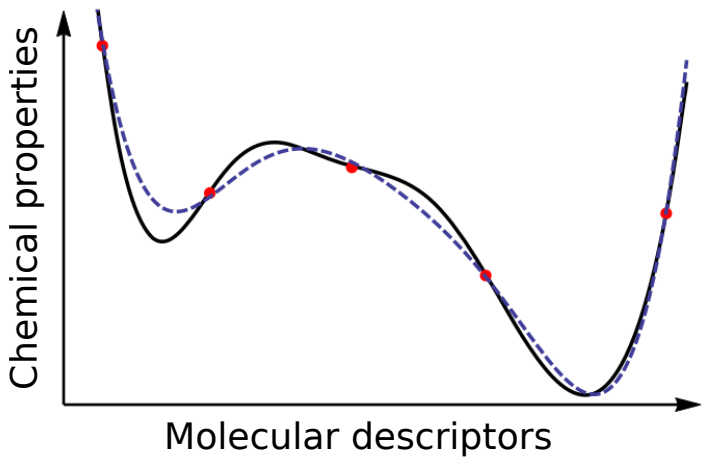
Supervised ML: Regression

- Task: prediction of continuous properties
- Applications: heat capacity at room temperature, HOMO–LUMO gap, receptor–ligand binding, stability of molecular conformers, ...

$$\mathbf{Y} = \mathbf{f}(descriptors)$$

Can be applied successfully if:

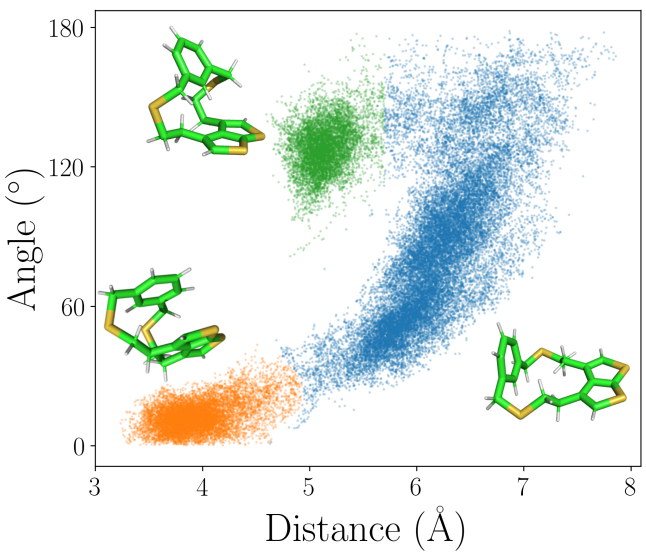
- Cause and effect relationship connecting system to property
- Query scenario is interpolative in nature
- Sufficient training data available

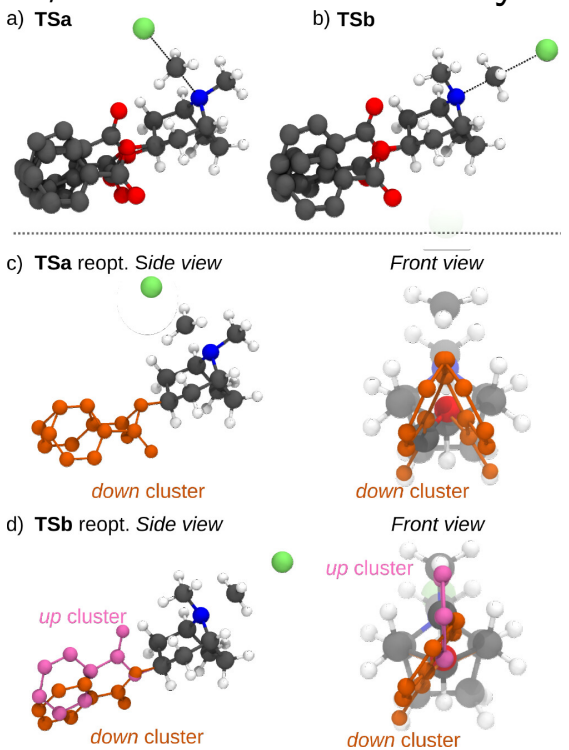


M. Rupp, O. A. von Lilienfeld, and K. Burke, "Guest editorial: Special topic on data-enabled theoretical chemistry", J. Chem. Phys. **148**, 241401 (2018)

Unsupervised ML:
Finding subgroups with clustering

- Exploratory data analysis technique that allows to organize information into meaningful subgroups (clusters) without any prior knowledge
- Each cluster defines a group of objects that share a certain degree of similarity but are more dissimilar to objects in other clusters
- Applications: representative sample, subsets selection, conformational analysis

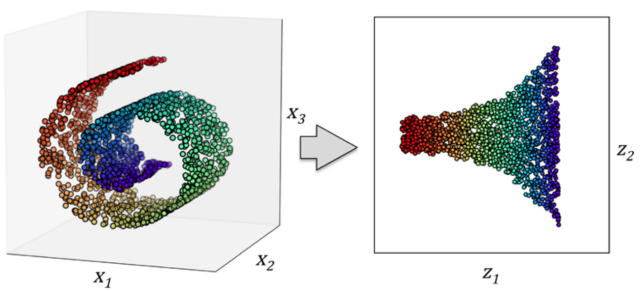


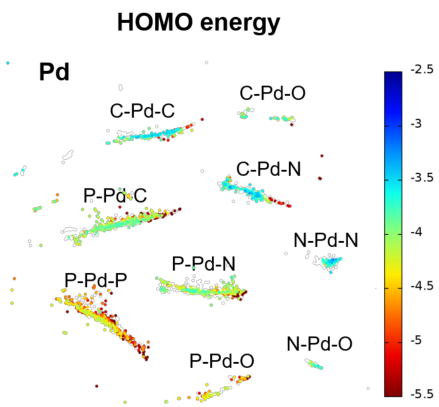


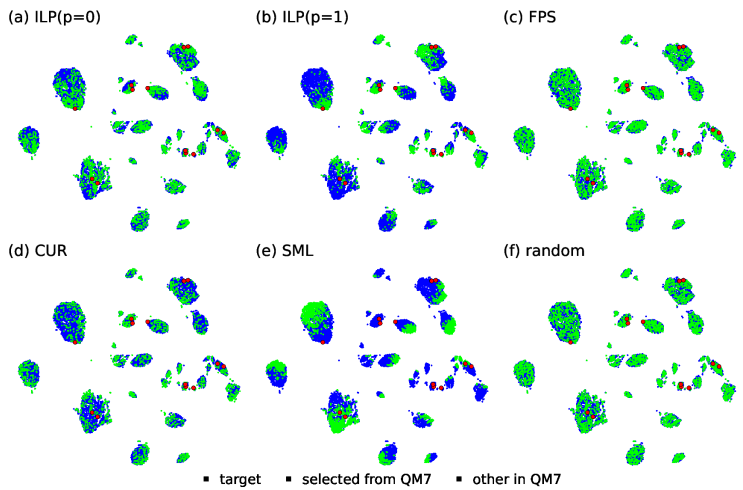
R. Laplaza, M. D. Wodrich, and C. Corminboeuf, "Overcoming the pitfalls of computing reaction selectivity from ensembles of transition states", J. Phys. Chem. Lett. **15**, 7363–7370 (2024)

Unsupervised ML: Dimensionality reduction

- Clustering identifies agglomeration of data but do not offer an overall picture of the relations between different structures
- Dimensionality reduction compresses the data onto a smaller dimensional subspace while retaining most of the relevant information
- Data visualization: a high-dimensional feature set can be projected onto 1-, 2-, or 3D spaces
- Applications: representation of chemical space, pattern recognition



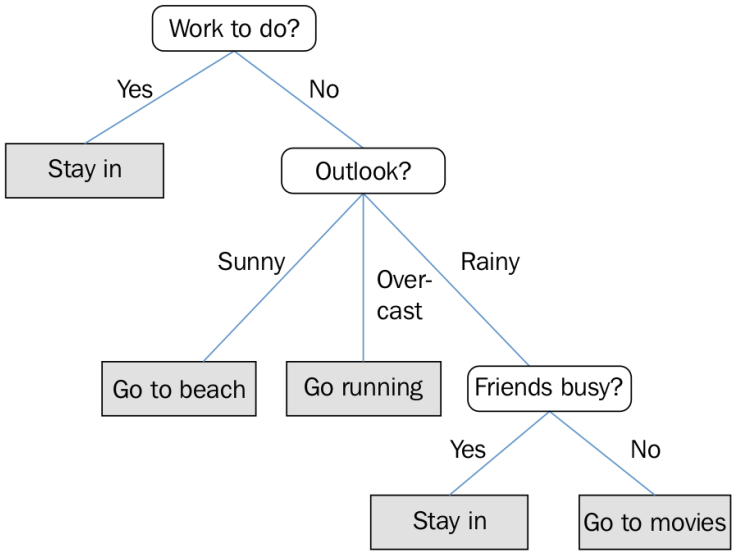




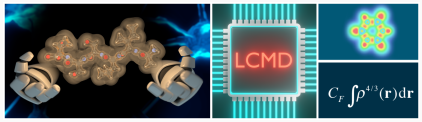
B. Sawatlon et al., "Data mining the C–C cross-coupling genome", ChemCatChem **11**, 4096–4107 (2019); M. Haeberle et al., "Integer linear programming for unsupervised training set selection in molecular machine learning", Mach. Learn.: Sci. Technol. **6**, 025030 (2025)

ML method example: Decision Trees

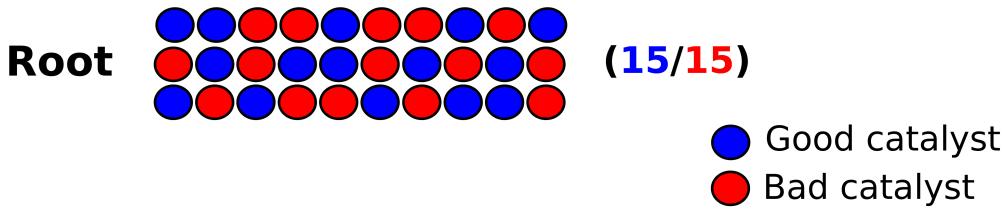
- Decision Trees (DTs) are a non-parametric supervised learning method used for classification and regression
- Learns simple decision rules inferred from the data features
- DT split a set of objects into subsets (usually 2 in binary trees) that are purer in composition
- Rules are selected based on how well splits can differentiate



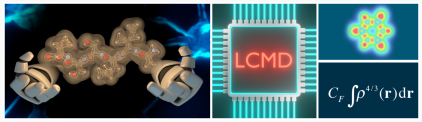
Decision Trees: chemical example



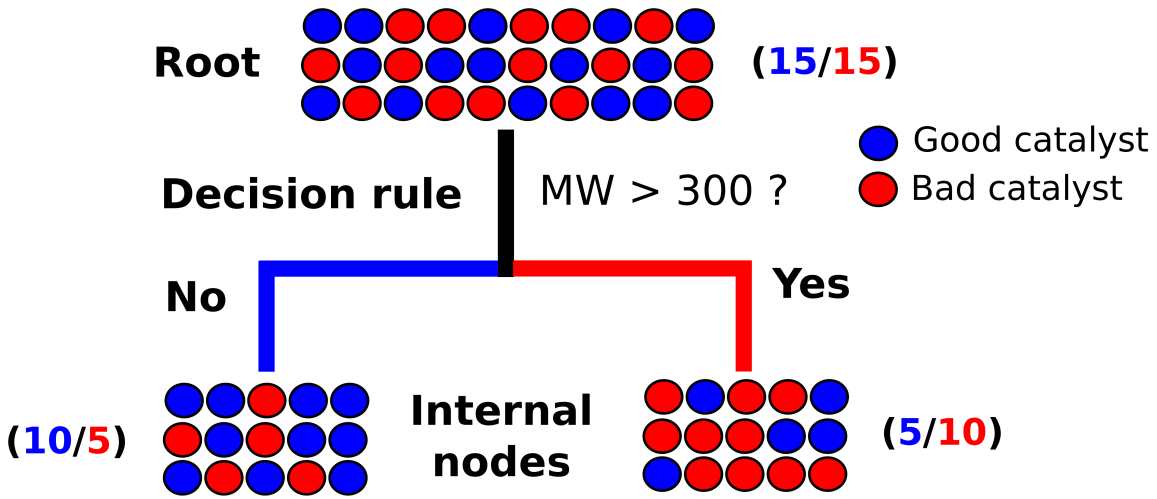
- Starting with a root composed by a set of 30 catalysts and knowing the ones that are "good" or "bad" for a given reaction
- Each catalyst is described by several simple molecular descriptors (*e.g.*, molecular weight, shape index, molecular volume, # of C atoms, # of rotatable bonds...



Decision Trees: chemical example

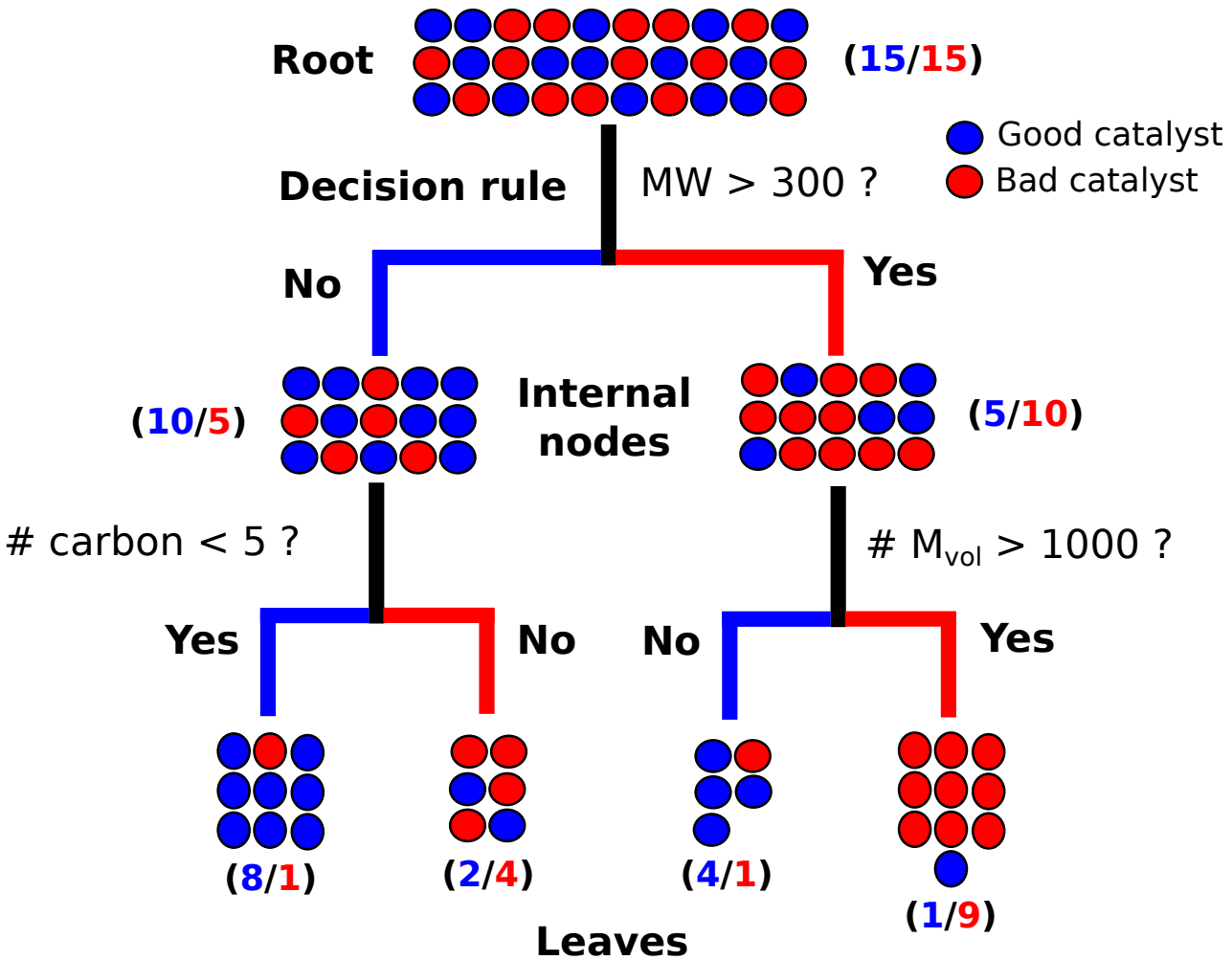


- DT algorithm searches a rule that yields the highest decrease in impurity
- Once a rule is selected and splits a node into two

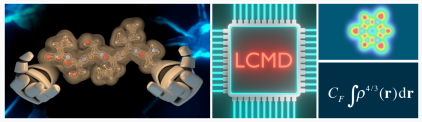


Decision Trees: chemical example

- The same logic is applied to each "child" node
- Stops when no further gain / stopping rules met

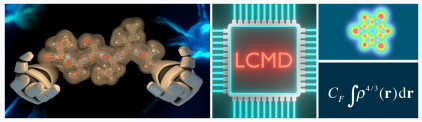


Decision Trees



Why classification and regression trees (CART) is a successful tool?

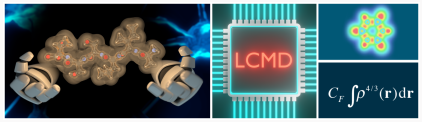
Decision Trees



Why classification and regression trees (CART) is a successful tool?

- Universally applicable to classification and regression problems with no assumptions on the data structure
- The picture of the tree structure gives valuable insights into which variables are important
- Terminal nodes give a natural clustering of the data into homogenous groups
- Can handle large data sets: $O(DM \log M)$
  (M is # of molecules, D is # of descriptors)

Decision Trees



Why classification and regression trees (CART) is a successful tool?

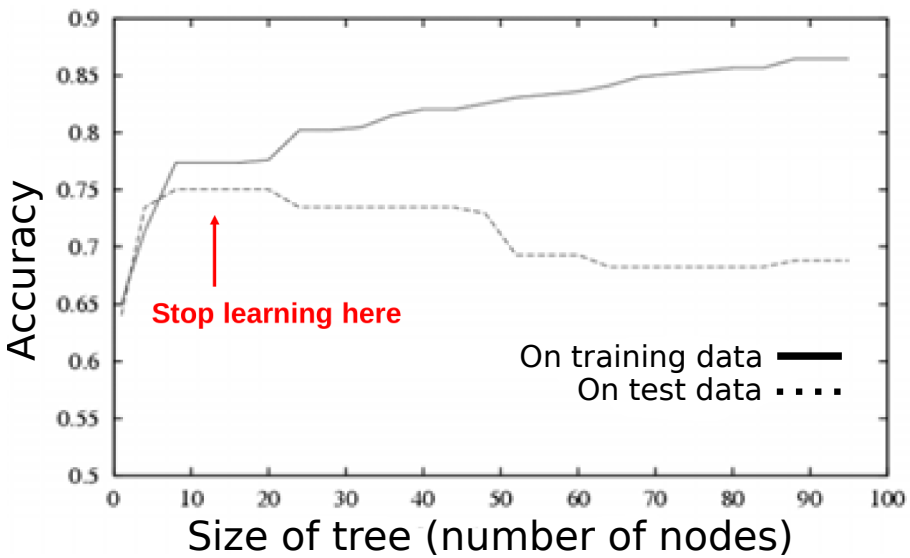
- Universally applicable to classification and regression problems with no assumptions on the data structure
- The picture of the tree structure gives valuable insights into which variables are important
- Terminal nodes give a natural clustering of the data into homogenous groups
- Can handle large data sets: $O(DM \log M)$
  (M is # of molecules, D is # of descriptors)

But...

- Decision-tree learners can create over-complex trees that do not generalize the data well
- This is called overfitting

Overfitting and early stopping

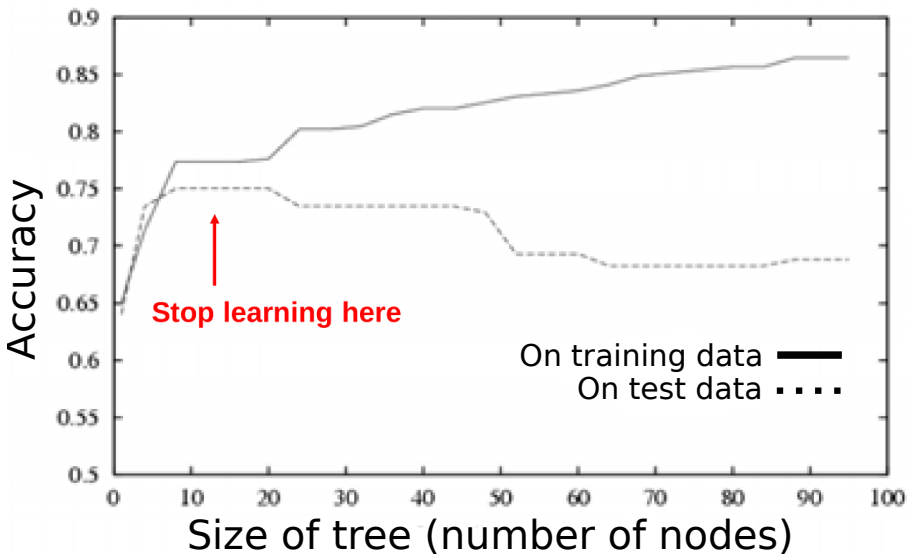
- Decision trees can overfit data
- So, it is necessary to use a validation set in order to prune the tree at an optimal size

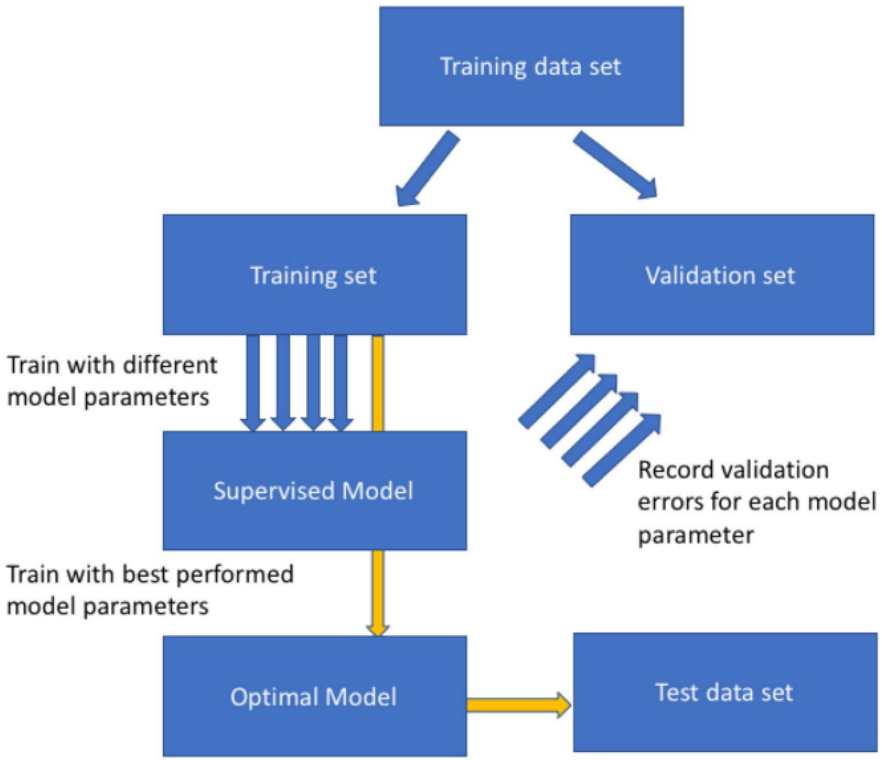


Y. Xu and R. Goodacre, "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning", J. Anal. Test. **2**, 249–262 (2018)

Overfitting and early stopping

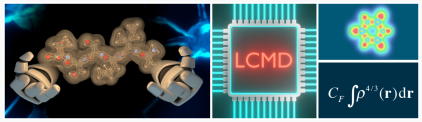
- Decision trees can overfit data
- So, it is necessary to use a validation set in order to prune the tree at an optimal size



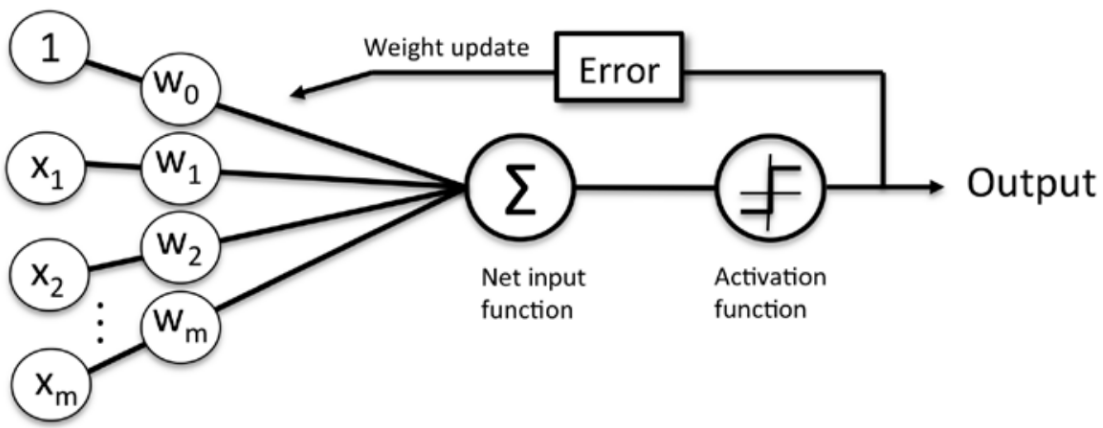


Y. Xu and R. Goodacre, "On splitting training and validation set: a comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning", J. Anal. Test. **2**, 249–262 (2018)

Perceptron

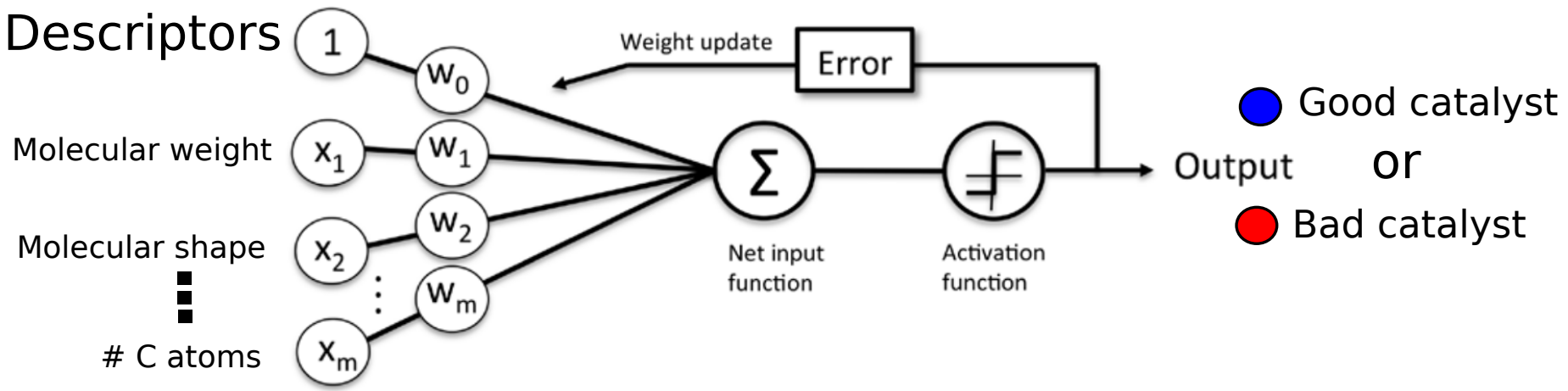


- Artificial neurons are inspired from biological neurons
- An artificial neuron has one or many inputs, each associated to a weight
- If the weighted sum of inputs is lower than a threshold, the neuron remains inactive
- If the weighted sum bypass the threshold, the neuron is activated and produce an output signal
- During the learning phase, this output is used to calculate the error of the prediction and update the weights



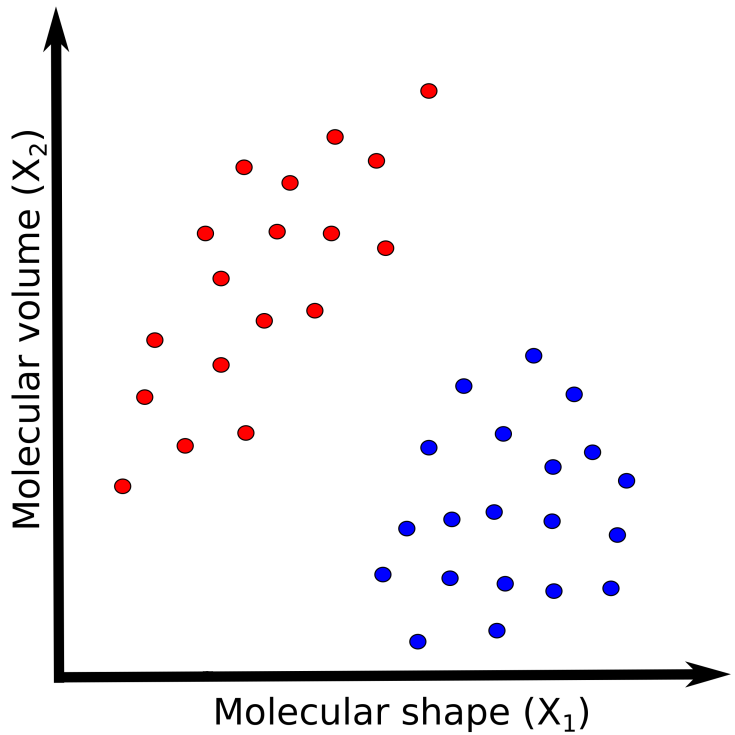
Perceptron: Chemical example

Starting with the same set of 30 catalysts and knowing the ones that are "good" or "bad" for a given reaction

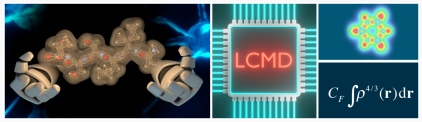


Perceptron: Chemical example

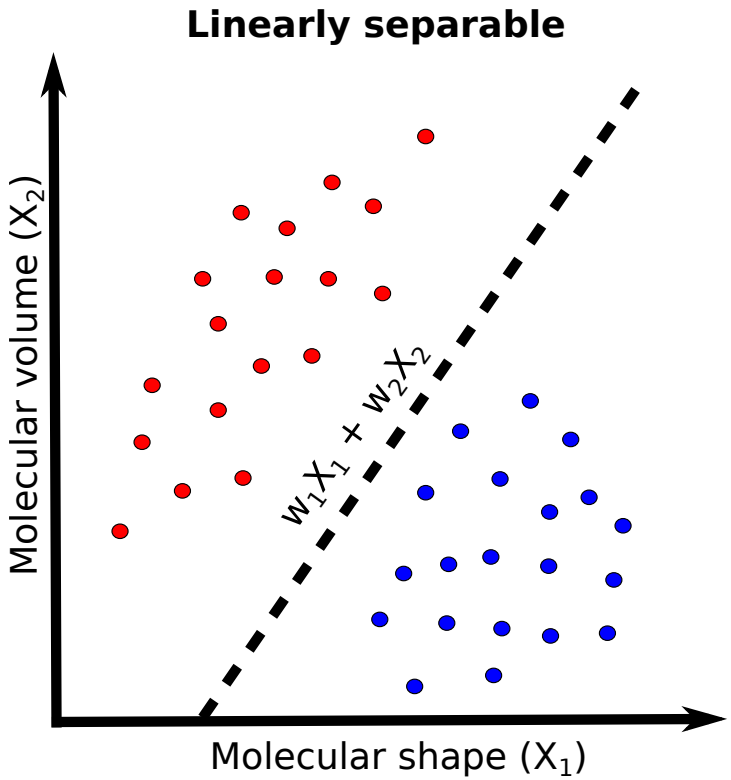
During the learning phase, the outputs are used to update the weights



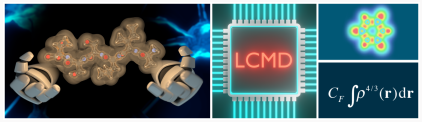
Perceptron: Chemical example



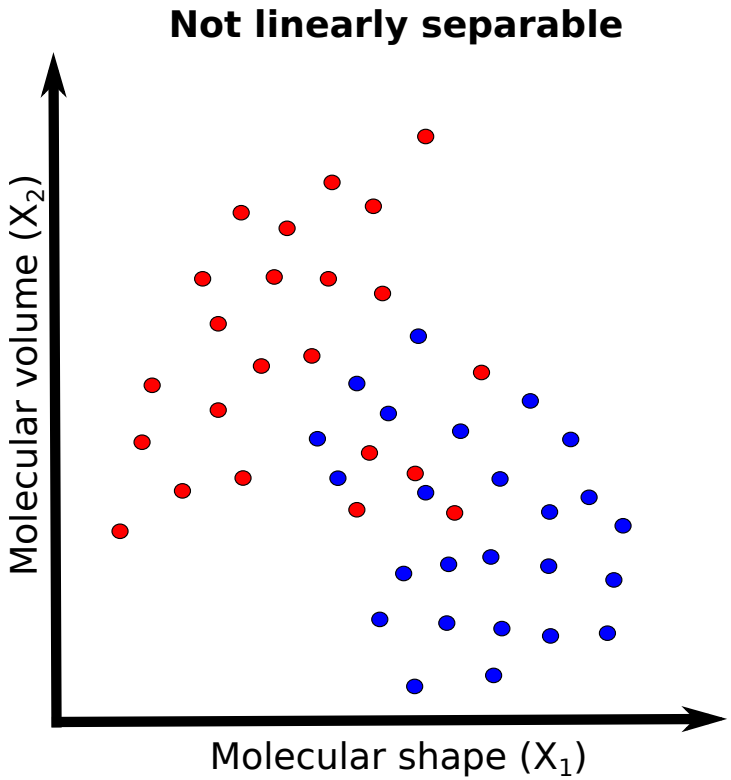
If the two classes are linearly separable, the perceptron will converge to a solution



Perceptron: Chemical example

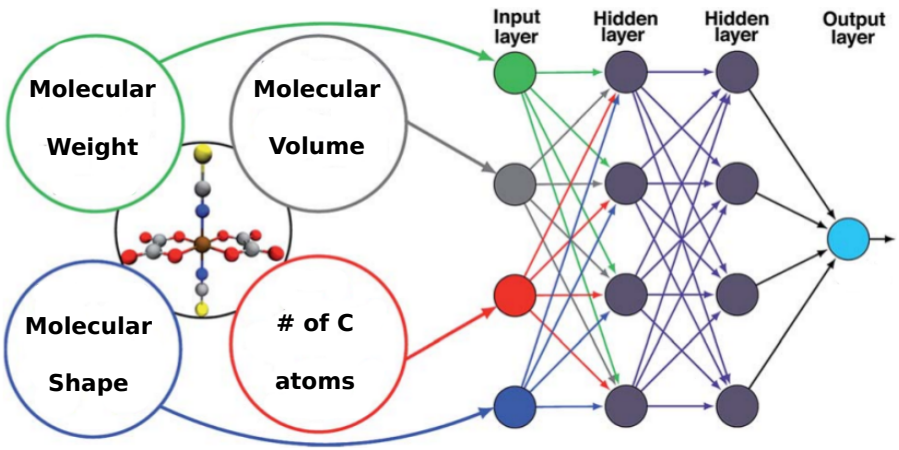


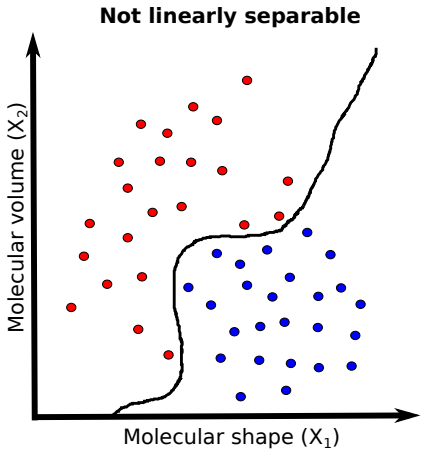
However, if the two classes are not linearly separable, either we accept the error, either we need more than a single neuron



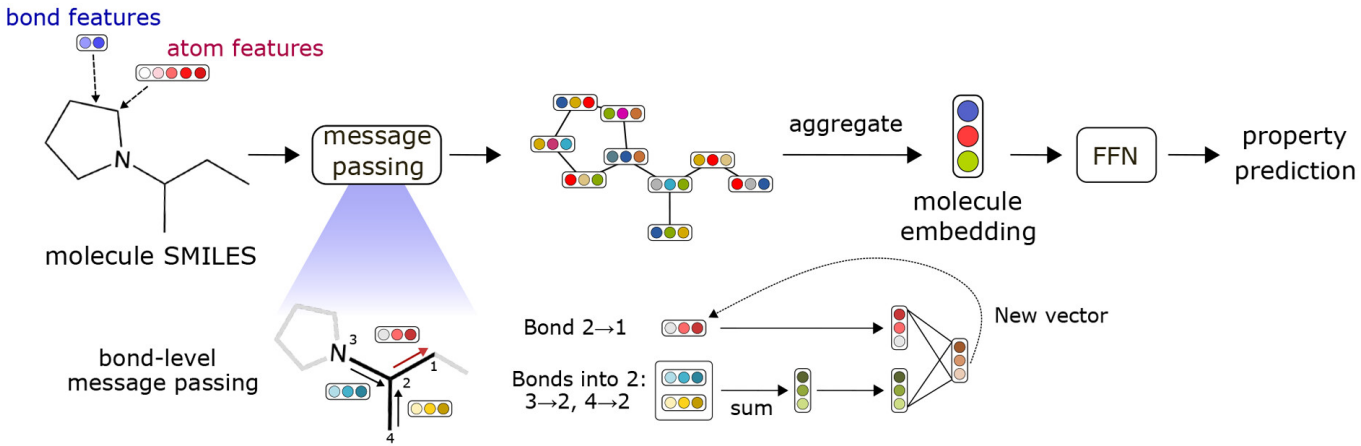
(Deep) artificial neural network

- Learns non-linear models for classification and prediction





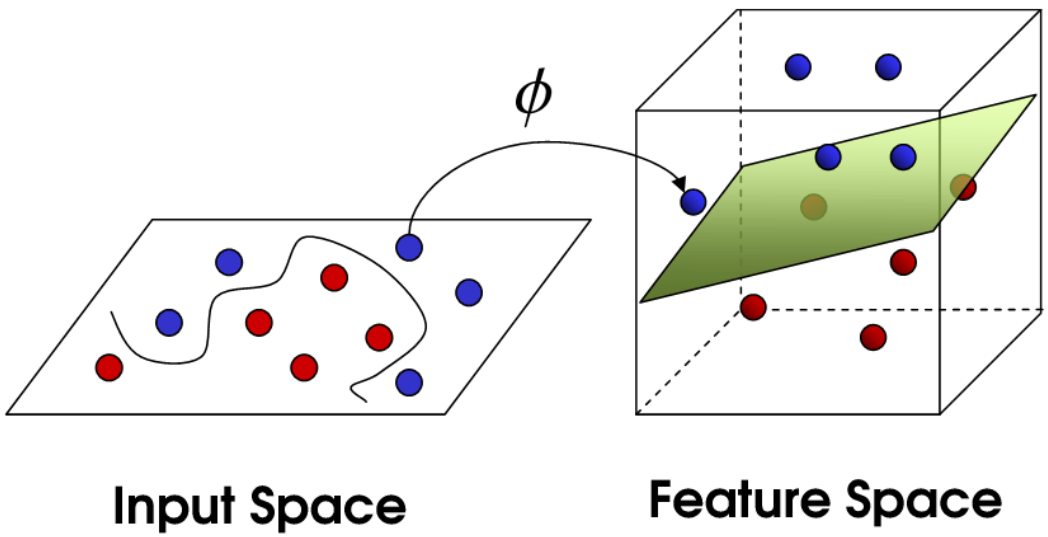
- Requires tuning a large number of hyperparameters
- Can learn new data descriptors itself



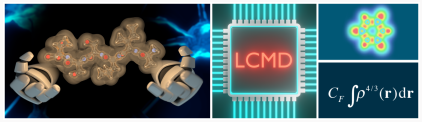
E. Heid et al., "Chemprop: A machine learning package for chemical property prediction", J. Chem. Inf. Model. **64**, 9–17 (2023)

Kernel trick:
Another way to solve non-linear problem

- Any non-linear problem can be mapped (ϕ) into a higher-dimensional feature space where it becomes linearly separable
- Computationally very expensive
- The step $\phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ is replaced by a kernel function $K(\mathbf{x}, \mathbf{x}')$

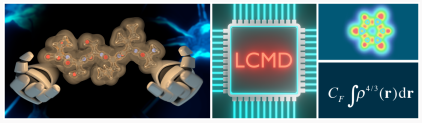


Kernel ridge regression

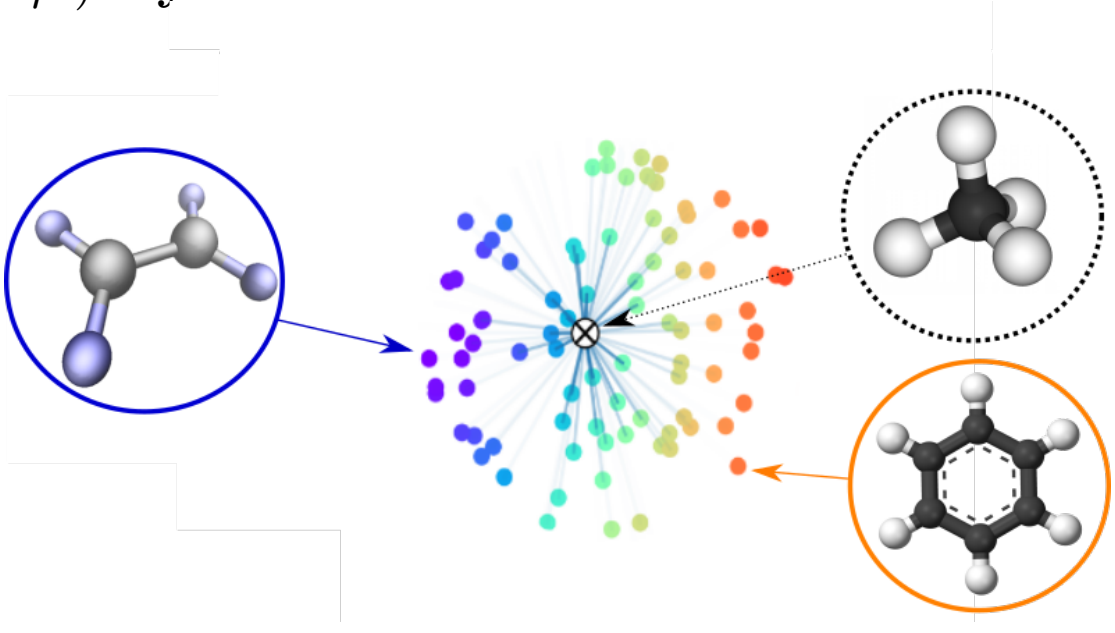


- Roughly speaking, kernels can be interpreted as similarity measures between pair of molecules
- The most widely used kernels are Gaussian $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma^2(\mathbf{x} - \mathbf{x}')^2\right)$ and Laplacian $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \sum_d^D \left|x_d - x'_d\right|\right)$

Kernel ridge regression



- Roughly speaking, kernels can be interpreted as similarity measures between pair of molecules
- The most widely used kernels are Gaussian $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma^2(\mathbf{x} - \mathbf{x}')^2\right)$ and Laplacian $K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \sum_d^D \left|x_d - x'_d\right|\right)$
- The prediction for y_t is a weighted sum of kernel functions centered on each training point: $y_t = \sum_i^N K(\mathbf{x}_t, \mathbf{x}_i) w_i^\star$
- The regression coefficients represent the contribution of each training point to the target value and minimize the quadratic loss function on the training set: $\mathbf{w}^\star = (\mathbf{K} + \eta \mathbf{1})^{-1} \mathbf{y}$



Molecular representations

Molecular descriptors

"...the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number"

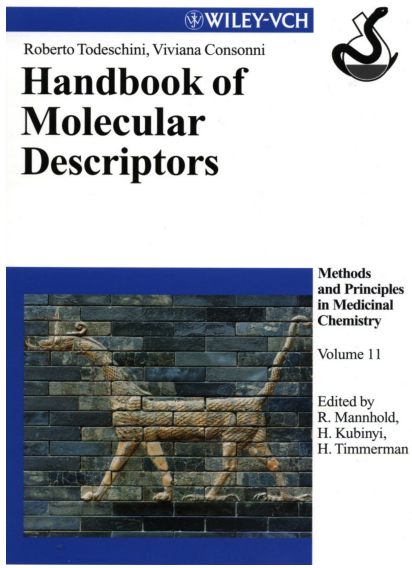
R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Vol. 11, Methods and Principles in Medicinal Chemistry (Wiley, 2000); R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, Vol. 41, Methods and Principles in Medicinal Chemistry (Wiley, 2009)

Molecular descriptors

"...the final result of a logic and mathematical procedure which transforms chemical information encoded within a symbolic representation of a molecule into a useful number"

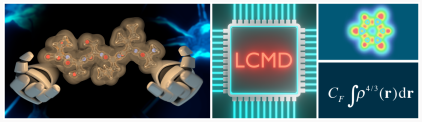
Vast majority:

- constitutional (# of atoms, MW,...)
- graph-based (bond-, distance-, adjacency matrices,...)
- mainly used in bio or chemo-informatics for QSAR applications



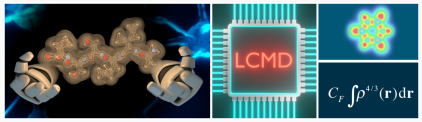
R. Todeschini and V. Consonni, *Handbook of molecular descriptors*, Vol. 11, Methods and Principles in Medicinal Chemistry (Wiley, 2000); R. Todeschini and V. Consonni, *Molecular descriptors for chemoinformatics*, Vol. 41, Methods and Principles in Medicinal Chemistry (Wiley, 2009)

Molecular representations



- Focus on QML models: physics-based, systematic, universal

Molecular representations

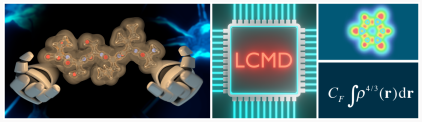


- Focus on QML models: physics-based, systematic, universal
- To better distinguish QML from QSAR, we prefer "representation" to "descriptor"

Molecular representations

- Focus on QML models: physics-based, systematic, universal
- To better distinguish QML from QSAR, we prefer "representation" to "descriptor"
- The ground state of a chemical system is defined by its Hamiltonian

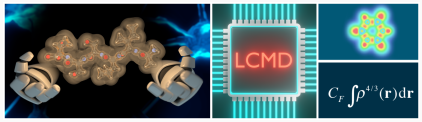
Molecular representations



- Focus on QML models: physics-based, systematic, universal
- To better distinguish QML from QSAR, we prefer "representation" to "descriptor"
- The ground state of a chemical system is defined by its Hamiltonian
- Hamiltonian depends on elemental composition and geometry (and number of electrons)

$$\hat{H}(\mathbf{Z}_i, \mathbf{R}_i) \xrightarrow{\text{QM}} E$$

Molecular representations

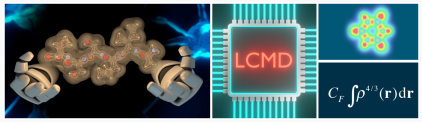


- Focus on QML models: physics-based, systematic, universal
- To better distinguish QML from QSAR, we prefer "representation" to "descriptor"
- The ground state of a chemical system is defined by its Hamiltonian
- Hamiltonian depends on elemental composition and geometry (and number of electrons)

$$\hat{H}(\mathbf{Z}_i, \mathbf{R}_i) \xrightarrow{\text{QM}} E$$

- Representation in ML plays the role of Hamiltonian/wavefunction in QM

Molecular representations



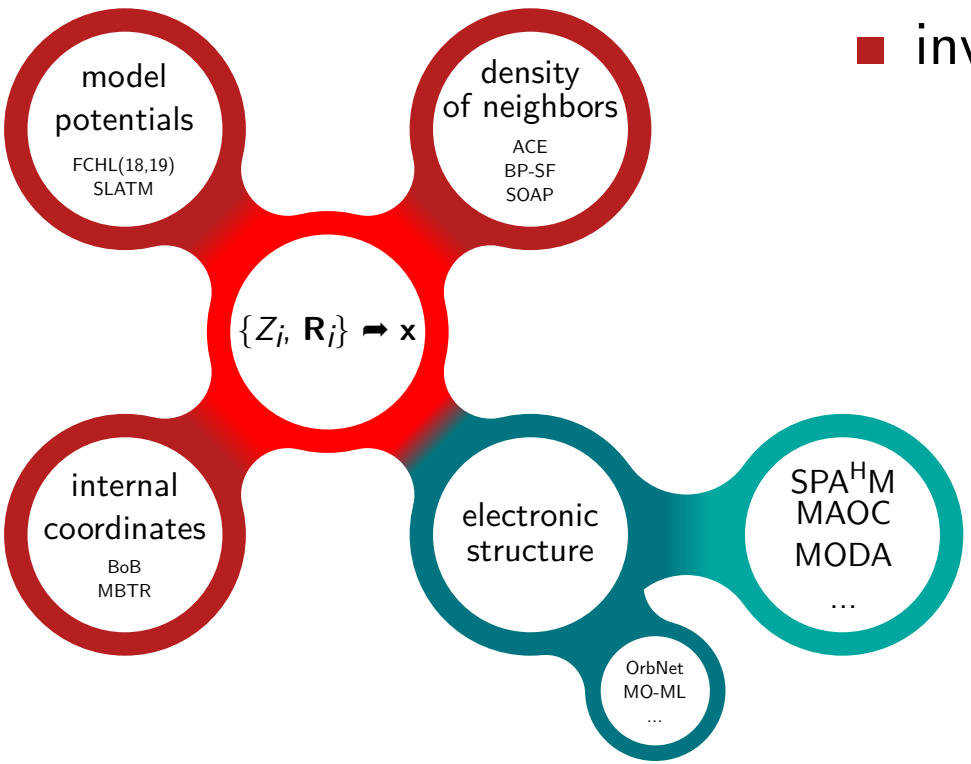
- Focus on QML models: physics-based, systematic, universal
- To better distinguish QML from QSAR, we prefer "representation" to "descriptor"
- The ground state of a chemical system is defined by its Hamiltonian
- Hamiltonian depends on elemental composition and geometry (and number of electrons)

$$\hat{H}(\mathbf{Z}_i, \mathbf{R}_i) \xrightarrow{\text{QM}} E$$

- Representation in ML plays the role of Hamiltonian/wavefunction in QM
- It should be a vector which encodes composition and geometry (and charge/spin) of a molecule

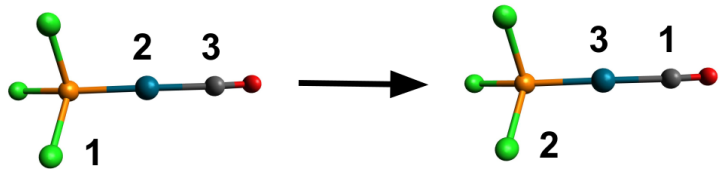
$$\text{representation}(\mathbf{Z}_i, \mathbf{R}_i) \xrightarrow{\text{ML}} E$$

Molecular representations

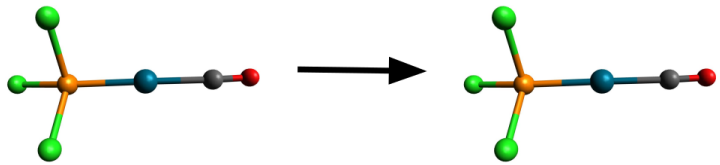


Criteria:

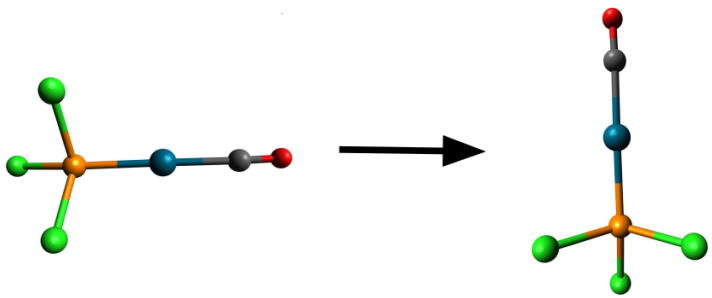
- injectivity
- continuity and differentiability
- invariance/equivariance w.r.t.:
  - permutations



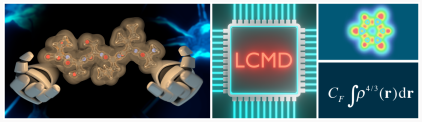
  - translations



  - rotations

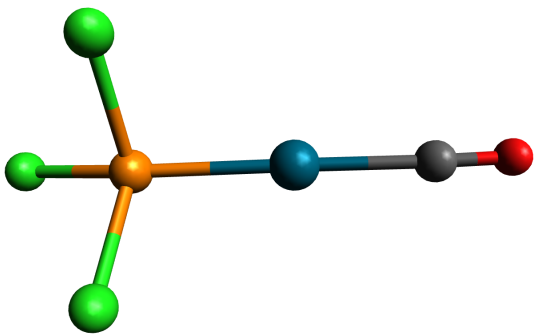


CM: the Coulomb matrix representation



- Square atom-by-atom matrix
- Off-diagonal elements correspond to the Coulomb repulsion between nuclei
- Diagonal elements remind the electronic energy of a H-like atom

$$M_{ij} = \begin{cases} \dfrac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \text{for } i \neq j \\ 0.5\, Z_i^{2.4} & \text{for } i = j \end{cases}$$

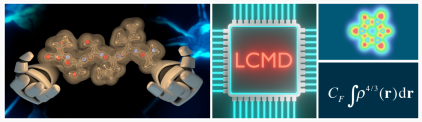


$\longrightarrow$ **M =**

	P	**Cl**	**Cl**	**Cl**	**Pd**	**C**	**O**
P	332	124	122	122	290	20	22
Cl	124	449	94	94	179	16	18
Cl	122	94	449	82	232	20	22
Cl	122	94	82	449	237	20	22
Pd	290	179	232	237	4893	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74

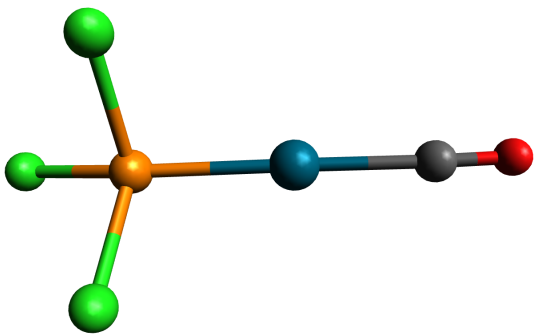
M. Rupp et al., "Fast and accurate modeling of molecular atomization energies with machine learning", Phys. Rev. Lett. **108**, 058301 (2012)

CM: the Coulomb matrix representation



- Square atom-by-atom matrix
- Off-diagonal elements correspond to the Coulomb repulsion between nuclei
- Diagonal elements remind the electronic energy of a H-like atom

$$M_{ij} = \begin{cases} \dfrac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} & \text{for } i \neq j \\ 0.5\, Z_i^{2.4} & \text{for } i = j \end{cases}$$



$\mathbf{M} =$

	P	Cl	Cl	Cl	Pd	C	O
P	332	124	122	122	290	20	22
Cl	124	449	94	94	179	16	18
Cl	122	94	449	82	232	20	22
Cl	122	94	82	449	237	20	22
Pd	290	179	232	237	4893	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74

- No well-defined ordering of the atoms in the matrix

M. Rupp et al., "Fast and accurate modeling of molecular atomization energies with machine learning", Phys. Rev. Lett. **108**, 058301 (2012)

Variants of CM

- Represent the molecule as a vector of sorted eigenvalues of the CM
- Invariant with respect to atom permutations
- Loss of information

	P	Cl	Cl	Cl	Pd	C	O
P	332	124	122	122	290	20	22
Cl	124	449	94	94	179	16	18
Cl	122	94	449	82	232	20	22
Cl	122	94	82	449	237	20	22
Pd	290	179	232	237	4893	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74

→

4954
691
367
350
220
93
8

K. Hansen et al., "Assessment and validation of machine learning methods for predicting molecular atomization energies", J. Chem. Theory Comput. **9**, 3404–3419 (2013)

Variants of CM

- Permute the matrix in order to sort the rows and the columns by their norm
- Unique CM representation
- More information than eigenspectrum

	P	Cl	Cl	Cl	Pd	C	O
P	332	124	122	122	290	20	22
Cl	124	449	94	449	179	16	18
Cl	122	94	449	82	232	20	22
Cl	122	94	82	449	237	20	22
Pd	290	179	232	237	4893	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74

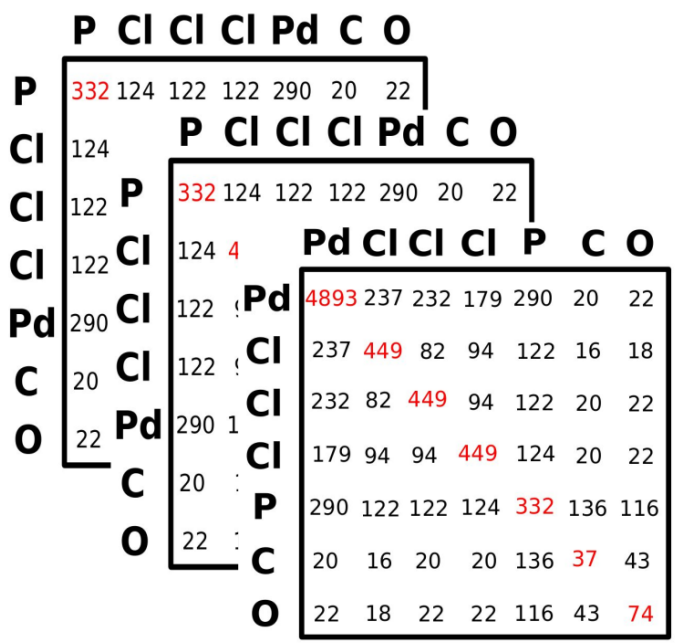
	Pd	Cl	Cl	Cl	P	C	O
Pd	4893	237	232	179	290	20	22
Cl	237	449	82	94	122	16	18
Cl	232	82	449	94	122	20	22
Cl	179	94	94	449	124	20	22
P	290	122	122	124	332	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74

K. Hansen et al., "Assessment and validation of machine learning methods for predicting molecular atomization energies", J. Chem. Theory Comput. **9**, 3404–3419 (2013)

Variants of CM

- Construct several CMs based on a random ordering of the atoms, adding a random noise ϵ to the row norms $||C||$ and determine the permutation than minimizes $||C|| + \epsilon$
- Approximate sampling of all possible valid CMs given a specific molecule
- Increased the computational costs

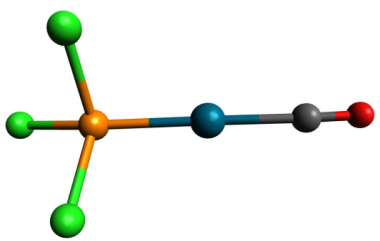
	P	Cl	Cl	Cl	Pd	C	O
P	332	124	122	122	290	20	22
Cl	124	449	94	94	179	16	18
Cl	122	94	449	82	232	20	22
Cl	122	94	82	449	237	20	22
Pd	290	179	232	237	4893	136	116
C	20	16	20	20	136	37	43
O	22	18	22	22	116	43	74



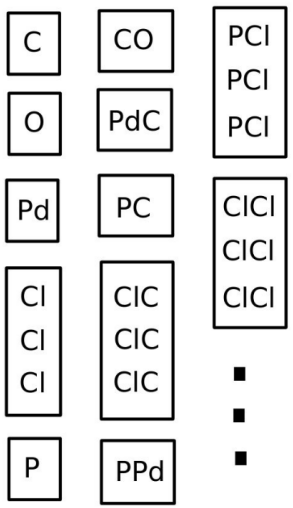
K. Hansen et al., "Assessment and validation of machine learning methods for predicting molecular atomization energies", J. Chem. Theory Comput. **9**, 3404–3419 (2013)

BoB: the Bag of Bonds representation

- Inspired by NLP: bag-of-words descriptor encodes the frequency of occurrence of words in text
- Each bag corresponds to a specific type of atomic pair
- For example, all P-Cl pairs in the molecule are grouped into the bag labeled as PCl
- Crucial higher-order information (angles and dihedrals) missing



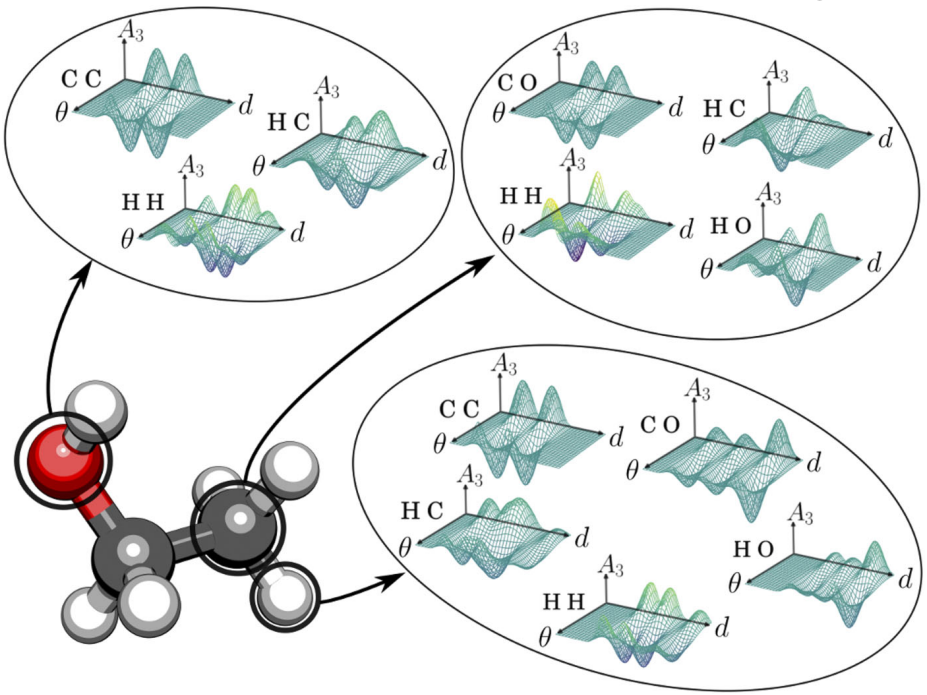
	P	Cl	Cl	Cl	Pd	C	O
P	P						
Cl	PCl	Cl					
Cl	PCl	ClCl	Cl				
Cl	PCl	ClCl	ClCl	Cl			
Pd	PPd	PdCl	PdCl	PdCl	Pd		
C	PC	ClC	ClC	ClC	PdC	C	
O	PO	ClO	ClO	ClO	PdO	CO	O



K. Hansen et al., "Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space", J. Phys. Chem. Lett. **6**, 2326–2331 (2015)

SLATM: the Spectrum of London and Axilrod–Teller–Muto potential

- Represents an atom i by accounting for all possible interactions between it and its neighboring atoms through many-body potential terms multiplied by a Gaussian distribution G and put in bags



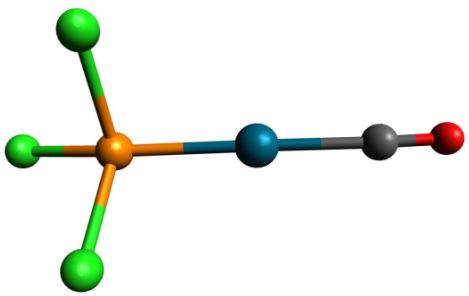
SLATM: the Spectrum of London and Axilrod–Teller–Muto potential

- Represents an atom i by accounting for all possible interactions between it and its neighboring atoms through many-body potential terms multiplied by a Gaussian distribution G and put in bags
- The one-body term: simply the nuclear charge ($x_1 = Z_i$)
- The two-body term:

$$x_2(\mathbf{r}) = \frac{1}{2}\sum_{j\neq i}^{N_{\text{at}}} G_{\sigma_2}(r - R_{ij}) \cdot \frac{Z_i Z_j}{|\mathbf{r}|^6}$$

- The three-body term:

$$x_3(\theta) = \frac{1}{3}\sum_{k\neq j\neq i}^{N_{\text{at}}} G_{\sigma_3}(\theta - \theta_{ijk}) \cdot \frac{Z_i Z_j Z_k \left[1 + \cos\theta \cos\theta_{jki} \cos\theta_{kij}\right]}{|\mathbf{R}_{ij}|^3 |\mathbf{R}_{jk}|^3 |\mathbf{R}_{ik}|^3}$$



```
[8], [17], [6], [46], [15], [8, 8], [17, 17], [6, 6], [46, 46],
[15, 15], (8, 17), (8, 6), (8, 46), (8, 15), (17, 6), (17, 46),
(17, 15), (6, 46), (6, 15), (46, 15), [8, 17, 17], [17, 8, 17],
[8, 17, 6], [8, 6, 17], [17, 8, 6], [8, 17, 46], [8, 46, 17],
[17,8, 46], [8, 17, 15], [8, 15, 17], [17, 8, 15], [8, 6, 46],
[8,46,6], [6, 8, 46], [8, 6, 15], [8, 15, 6], [6, 8, 15], [8, 46,
15], [8, 15, 46], [46, 8, 15], [17, 17, 17], [17, 17, 6], [17,6,
17], [17, 17, 46], [17, 46, 17], [17, 17, 15], [17, 15, 17], [17,
6, 46], [17, 46, 6], [6, 17, 46], [17, 6, 15], [17, 15, 6], [6,
17, 15], [17, 46, 15], [17, 15, 46], [46, 17, 15], [6,46,15], [6,
15, 46], [46, 6, 15]
```

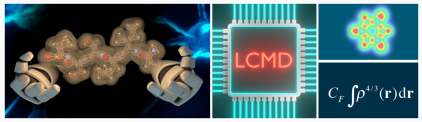
F. A. Faber et al., "Alchemical and structural distribution based representation for universal quantum machine learning", J. Chem. Phys. **148**, 241717 (2018); B. Huang and O. A. von Lilienfeld, "Quantum machine learning using atom-in-molecule-based fragments selected on the fly", Nat. Chem. **12**, 945–951 (2020)

SOAP: smooth overlap of atomic positions

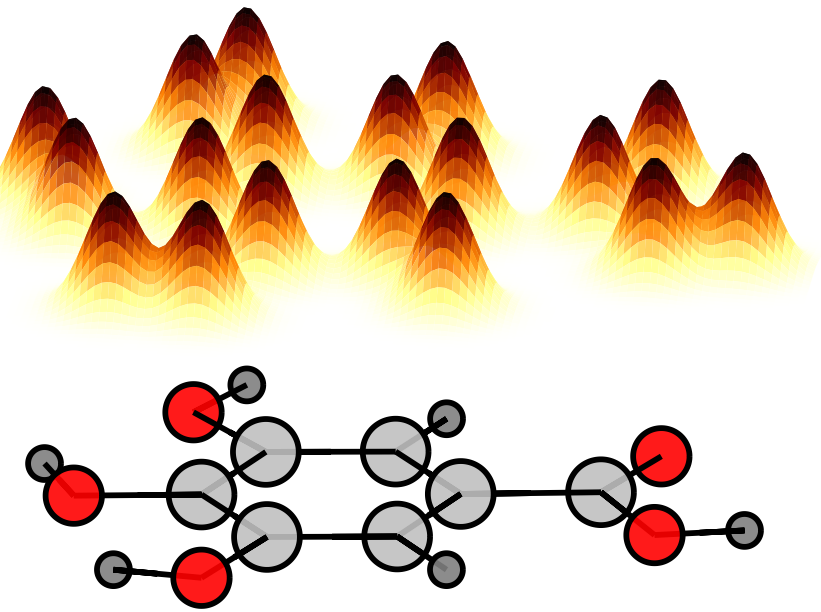
- Local similarity measure between atoms

A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments", Phys. Rev. B **87**, 184115 (2013)

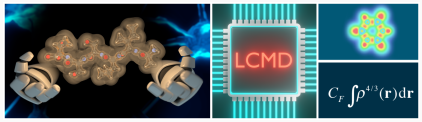
SOAP: smooth overlap of atomic positions



- Local similarity measure between atoms
- Each atom i is represented as sum of neighbor nuclei densities smoothened with a Gaussian, $\varrho_i(\mathbf{r}) = \sum_k \exp\left(-\frac{|\mathbf{R}_k - \mathbf{r}|^2}{2\sigma_\varrho^2}\right)$



SOAP: smooth overlap of atomic positions

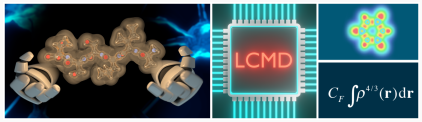


- Local similarity measure between atoms
- Each atom i is represented as sum of neighbor nuclei densities smoothened with a Gaussian, $\varrho_i(\mathbf{r}) = \sum_k \exp\left(-\frac{|\mathbf{R}_k - \mathbf{r}|^2}{2\sigma_\varrho^2}\right)$
- Similarity is the overlap averaged over rotations

$$\bar{K}_{ij} = \int \mathrm{d}\hat{R} \left[\int \varrho_i(\mathbf{r}) \varrho_j(\hat{R}\mathbf{r}) \, \mathrm{d}^3\mathbf{r}\right]^2, \quad K_{ij} = \left[\frac{\bar{K}_{ij}}{\sqrt{\bar{K}_{ii}\bar{K}_{jj}}}\right]^\zeta$$

A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments", Phys. Rev. B **87**, 184115 (2013)

SOAP: smooth overlap of atomic positions



- Local similarity measure between atoms
- Each atom i is represented as sum of neighbor nuclei densities smoothened with a Gaussian, $\varrho_i(\mathbf{r}) = \sum_k \exp\left(-\frac{|\mathbf{R}_k - \mathbf{r}|^2}{2\sigma_\varrho^2}\right)$
- Similarity is the overlap averaged over rotations

$$\bar{K}_{ij} = \int \mathrm{d}\hat{R} \left[\int \varrho_i(\mathbf{r}) \varrho_j(\hat{R}\mathbf{r}) \, \mathrm{d}^3\mathbf{r}\right]^2, \quad K_{ij} = \left[\frac{\bar{K}_{ij}}{\sqrt{\bar{K}_{ii}\bar{K}_{jj}}}\right]^\zeta$$

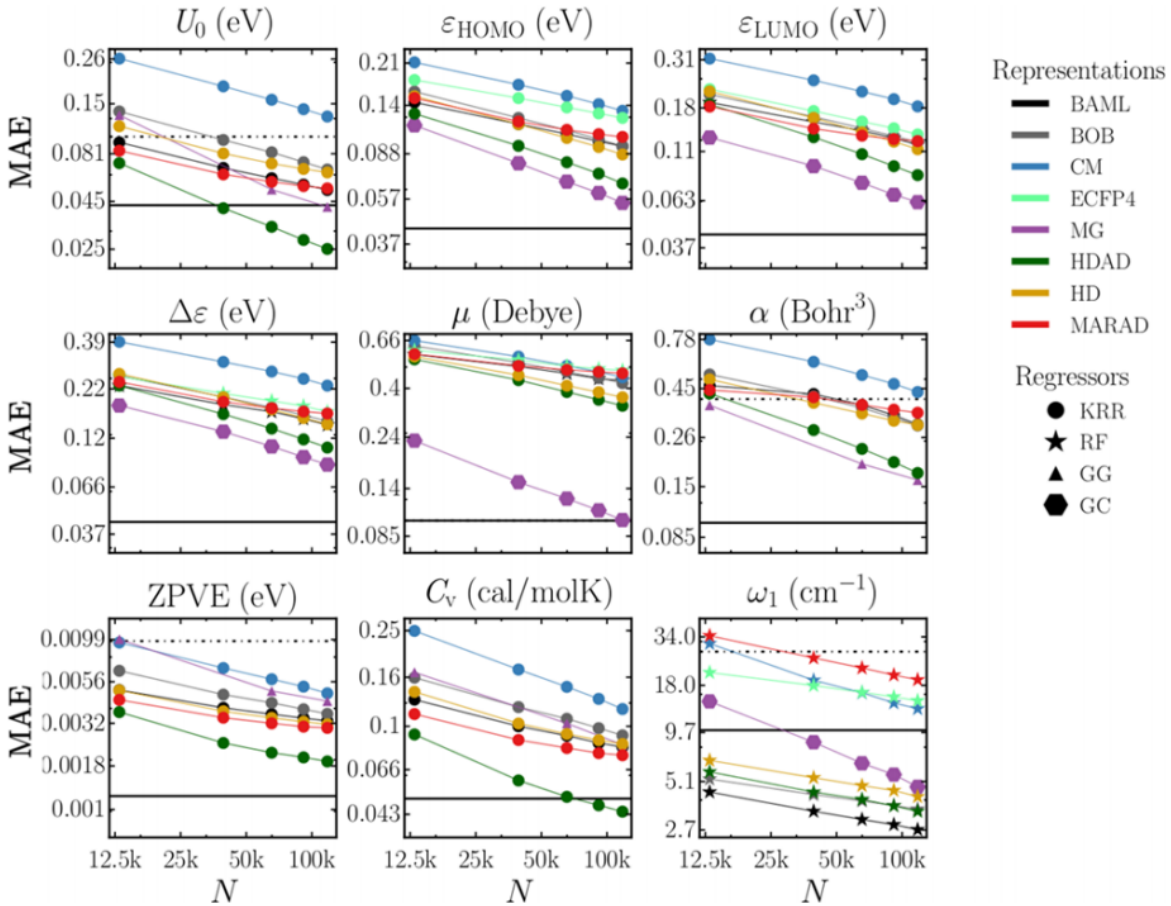
- In practice, ϱ_i is decomposed onto atom-centered basis leading to power spectrum representations $\mathbf{p}$

$$K_{ij} = \left[\frac{\mathbf{p}_i^\intercal \mathbf{p}_j}{\sqrt{\mathbf{p}_i^\intercal \mathbf{p}_i \cdot \mathbf{p}_j^\intercal \mathbf{p}_j}}\right]^\zeta = \left[\bar{\mathbf{p}}_i^\intercal \bar{\mathbf{p}}_j\right]^\zeta, \quad p_{i,nn'\ell} = \sqrt{\frac{8\pi^2}{2\ell+1}} \sum_m c_{i,n\ell m} c^*_{i,n'\ell m}$$

A. P. Bartók, R. Kondor, and G. Csányi, "On representing chemical environments", Phys. Rev. B **87**, 184115 (2013)

Few examples: ML applied to chemical properties

Simple molecular properties

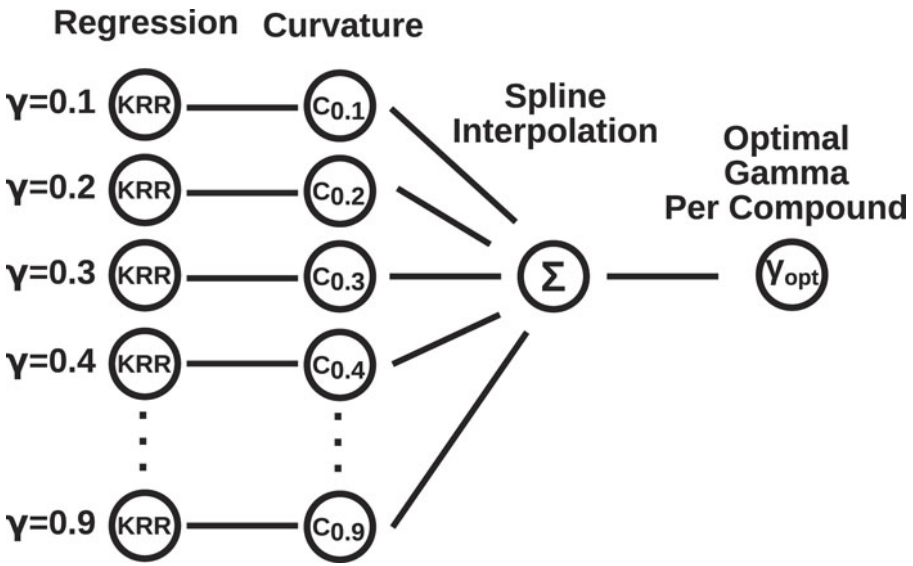


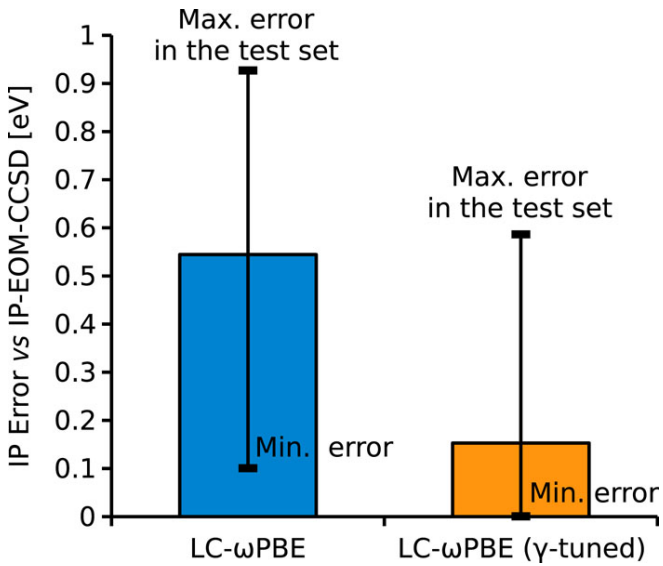
QM9 database (molecules with up to 9 heavy atoms)

F. A. Faber et al., "Prediction errors of molecular machine learning models lower than hybrid DFT error", J. Chem. Theory Comput. **13**, 5255–5264 (2017)

Density functional optimization

- Average energy curvature is $C^N_{\text{avg}} \int_{N-1}^{N} \frac{\partial^2 E}{\partial x^2}(x)\,\mathrm{d}x = \varepsilon^N_{\text{HOMO}} - \varepsilon^{N-1}_{\text{LUMO}}$
- Minimization of C is a criterion for the optimal tuning of range-separated hybrid density functionals (γ)
- Pipeline: predict curvature for different γ and find the optimal γ with spline interpolation

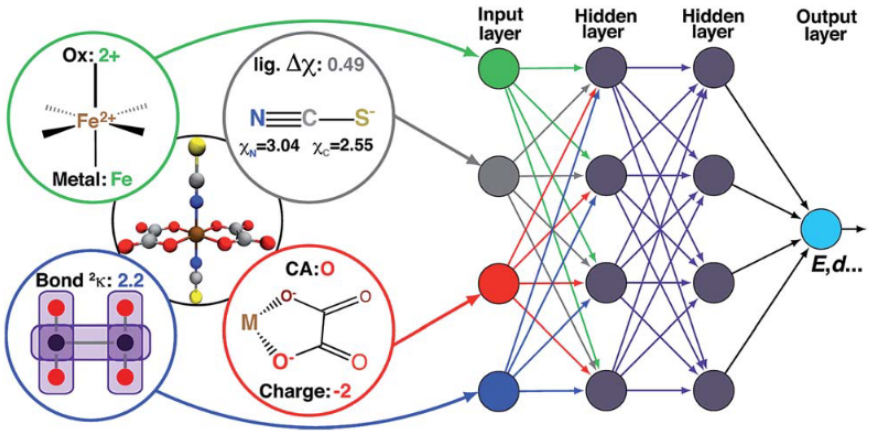


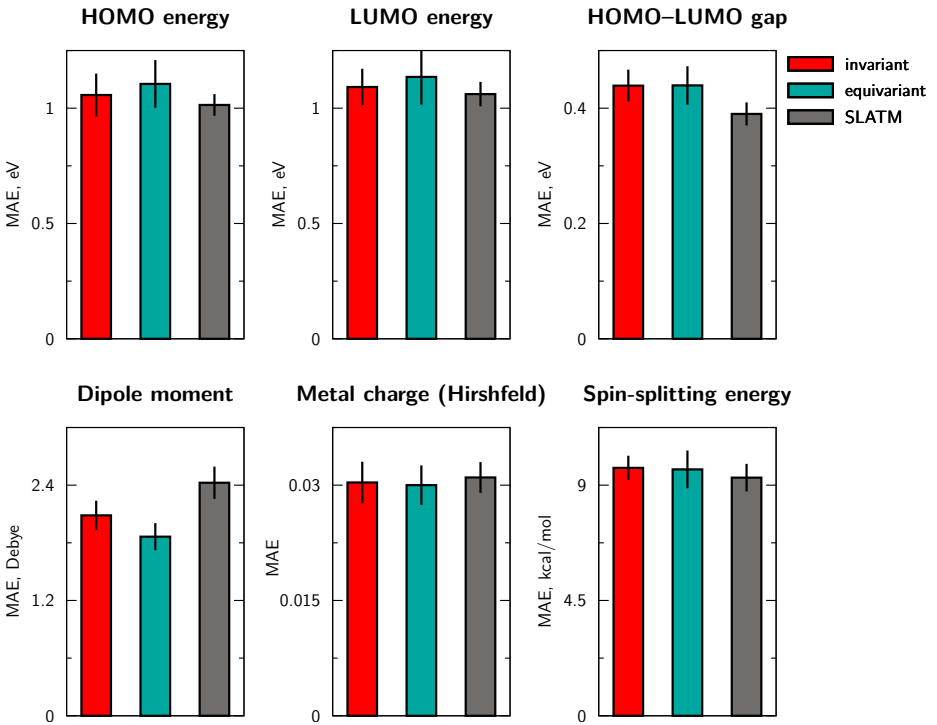


A. Fabrizio, B. Meyer, and C. Corminboeuf, "Machine learning models of the energy curvature vs particle number for optimal tuning of long-range corrected functionals", J. Chem. Phys. **152**, 154103 (2020)

Organometallic properties

Spin-state ordering, sensitivity to HF exchange, spin-state specific bond lengths





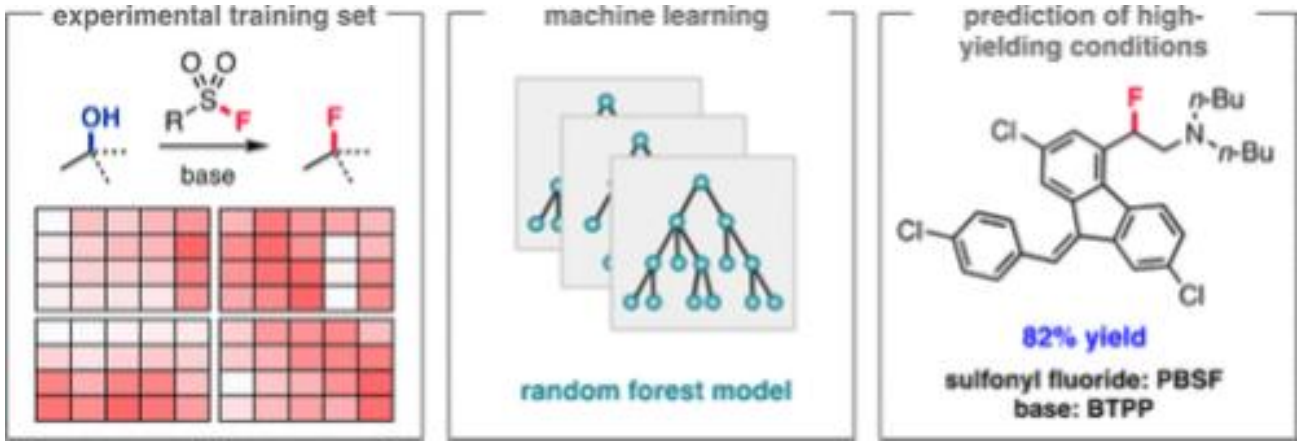
- Graph theory based descriptors: Atomic connectivity, Kier index. Preferred to 3-d structural information (e.g., Coulomb Matrix).
- Complex Based: Metal identity, oxidation state, empirical pairwise Pauling electronegativity, ...
- Atomic descriptors: Charge, mass, ...

- KRR + molecular representations (local, global, structure, electronic...)
- Invariant/equivariant tensor field neural network

J. P. Janet and H. J. Kulik, "Predicting electronic structure properties of transition metal complexes with neural networks", Chem. Sci. **8**, 5137–5152 (2017); Y. Cho et al., "Automated prediction of ground state spin for transition metal complexes", Digital Discovery **3**, 1638–1647 (2024)

Homogeneous catalysis

Prediction of the reaction performance in C–N cross-coupling and deoxylfluorination with sulfonyl fluorides from experimental training set constructed by ultra-high-throughput nanoscale experimentation

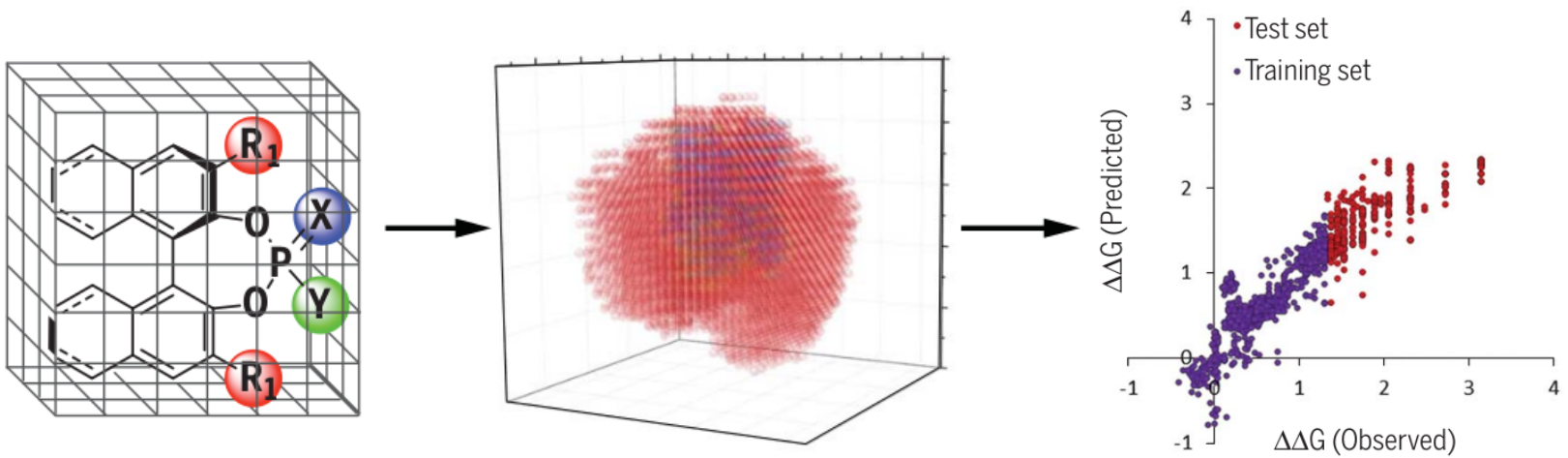


- Molecular descriptors: Molecular volume, Surface area, Molecular weight, E_{HOMO}, E_{LUMO}
- Atomic descriptors: Electrostatic charge and NMR shift
- Vibrational descriptors: Frequency and intensity

D. T. Ahneman et al., "Predicting reaction performance in C–N cross-coupling using machine learning", Science **360**, 186–190 (2018); M. K. Nielsen et al., "Deoxyfluorination with sulfonyl fluorides: Navigating reaction space with machine learning", J. Am. Chem. Soc. **140**, 5004–5008 (2018)

ML applied to homogeneous catalysis

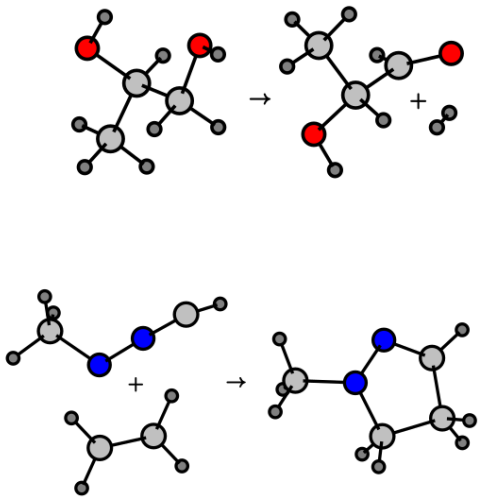
- Exploit machine learning algorithms to accurately predict selective catalysts for the chiral phosphoric acid-catalyzed thiol addition to N-acylimines reactions
- ML models used: Support Vector Machines (SVM) and Neural Network
- Descriptors: average steric occupancy (ASO), computed electrostatic parameters and NBO charges.

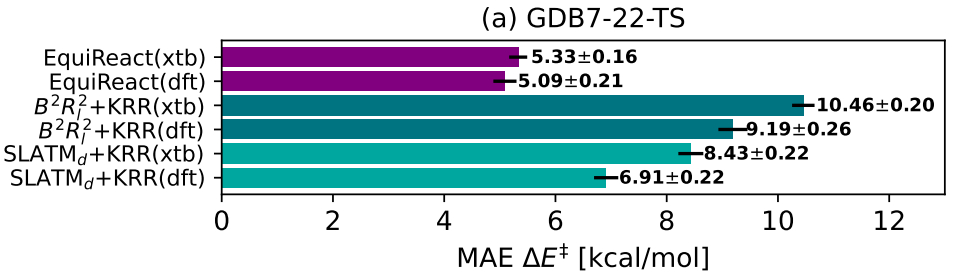


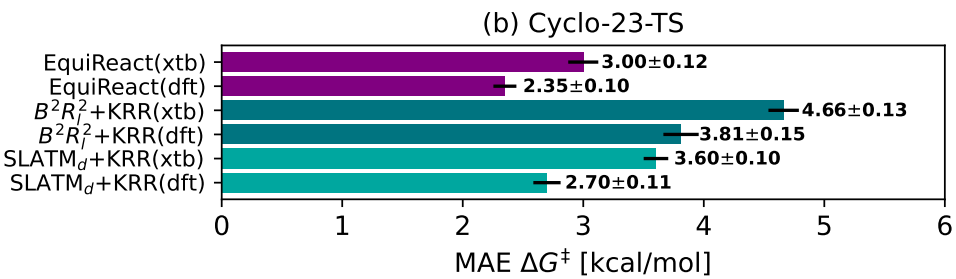
A. F. Zahrt et al., "Prediction of higher-selectivity catalysts by computer-driven workflow and machine learning", Science **363**, eaau5631 (2019)

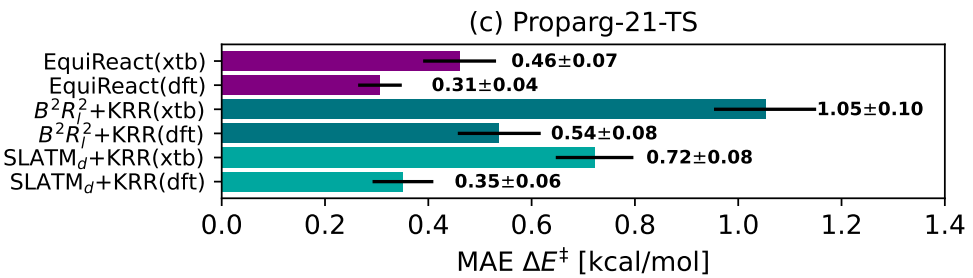
Reaction barrier prediction

- QML for reactions is more challenging
- Involve $2+$ components (reactant(s) and product(s))
- Representations design: transition state instead of Hamiltonian









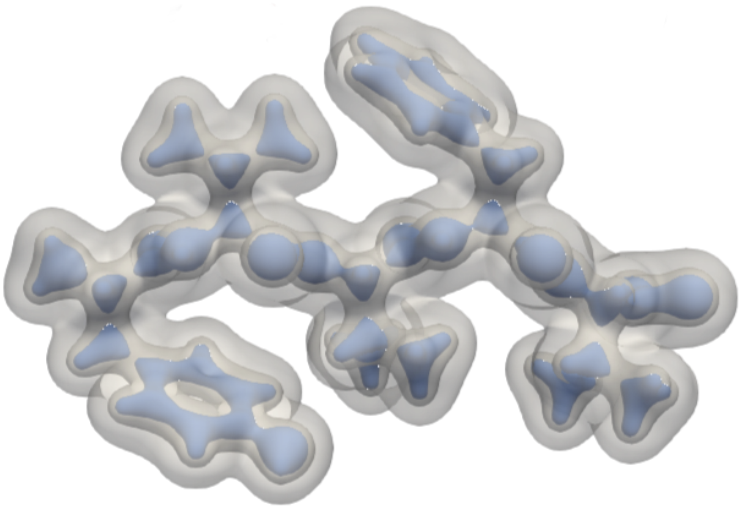
P. van Gerwen et al., "Benchmarking machine-readable vectors of chemical reactions on computed activation barriers", Digital Discovery **3**, 932–943 (2024)

Chemical properties: Beyond

There are many other chemical properties to predict, and not only scalar properties...

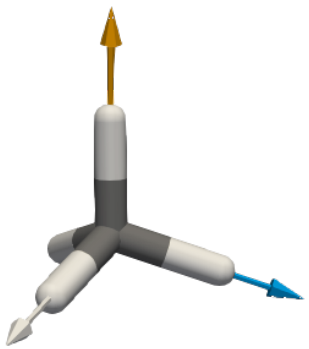
Tensorial properties

Scalar fields



Electron Density

Vectorial fields



Molecular Force Field

A. Fabrizio et al., "Electron density learning of non-covalent systems", Chem. Sci. **10**, 9424–9432 (2019); J. Behler, "Four generations of high-dimensional neural network potentials", Chem. Rev. **121**, 10037–10072 (2021)

Bibliography

Bibliography and future reading

Reviews:

- F. Musil et al., "Physics-inspired structural representations for molecules and materials", Chem. Rev. **121**, 9759–9815 (2021)
- J. A. Keith et al., "Combining machine learning and computational chemistry for predictive insights into chemical systems", Chem. Rev. **121**, 9816–9872 (2021)
- B. Huang and O. A. von Lilienfeld, "Ab initio machine learning in chemical compound space", Chem. Rev. **121**, 10001–10036 (2021)
- O. T. Unke et al., "Machine learning force fields", Chem. Rev. **121**, 10142–10186 (2021)

ML books:

- C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, (The MIT Press, 2005)
- T. Hastie, J. Friedman, and R. Tibshirani, *The Elements of Statistical Learning* (Springer, 2001)
- T. Mitchell, *Machine Learning* (McGraw-Hill, 1997)
- I. Witten and E. Frank, *Data mining – Practical Machine Learning Tools and Techniques* (2005)

Quiz

Quiz 1

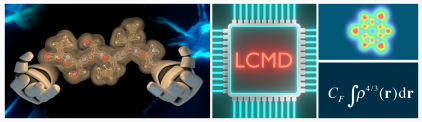
You want to machine-learn the HOMO energies of molecules.

Which family of machine learning methods would you use: Unsupervised or Supervised?

If Unsupervised, would you used a Clustering or a Dimensionality Reduction algorithm?

If Supervised, would you used a Regression or a Classification algorithm?

Quiz 1



You want to machine-learn the HOMO energies of molecules.

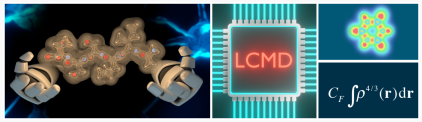
Which family of machine learning methods would you use: Unsupervised or Supervised?

If Unsupervised, would you used a Clustering or a Dimensionality Reduction algorithm?

If Supervised, would you used a Regression or a Classification algorithm?

Supervised learning, Regression algorithm

Quiz 2



You want to machine-learn the atomization energies of a set of molecules.

How do you need to describe your molecules for this purpose?

Quiz 2

You want to machine-learn the atomization energies of a set of molecules.

How do you need to describe your molecules for this purpose?

Need to describe the molecules with molecular representations (and not with common molecular descriptors used in QSAR/QSPR approches) that encode the basic information contained in the molecular Hamiltonian. For example: SLATM

Quiz 3

A company provides you a list of 10000 molecules and asks to choose good catalysts for a given reaction.

You only have the budget for testing experimentally 1000 compounds.

Which steps would you follow?

Quiz 3

A company provides you a list of 10000 molecules and asks to choose good catalysts for a given reaction.

You only have the budget for testing experimentally 1000 compounds.

Which steps would you follow?

First, we need to describe our molecules with molecular representations or descriptors.

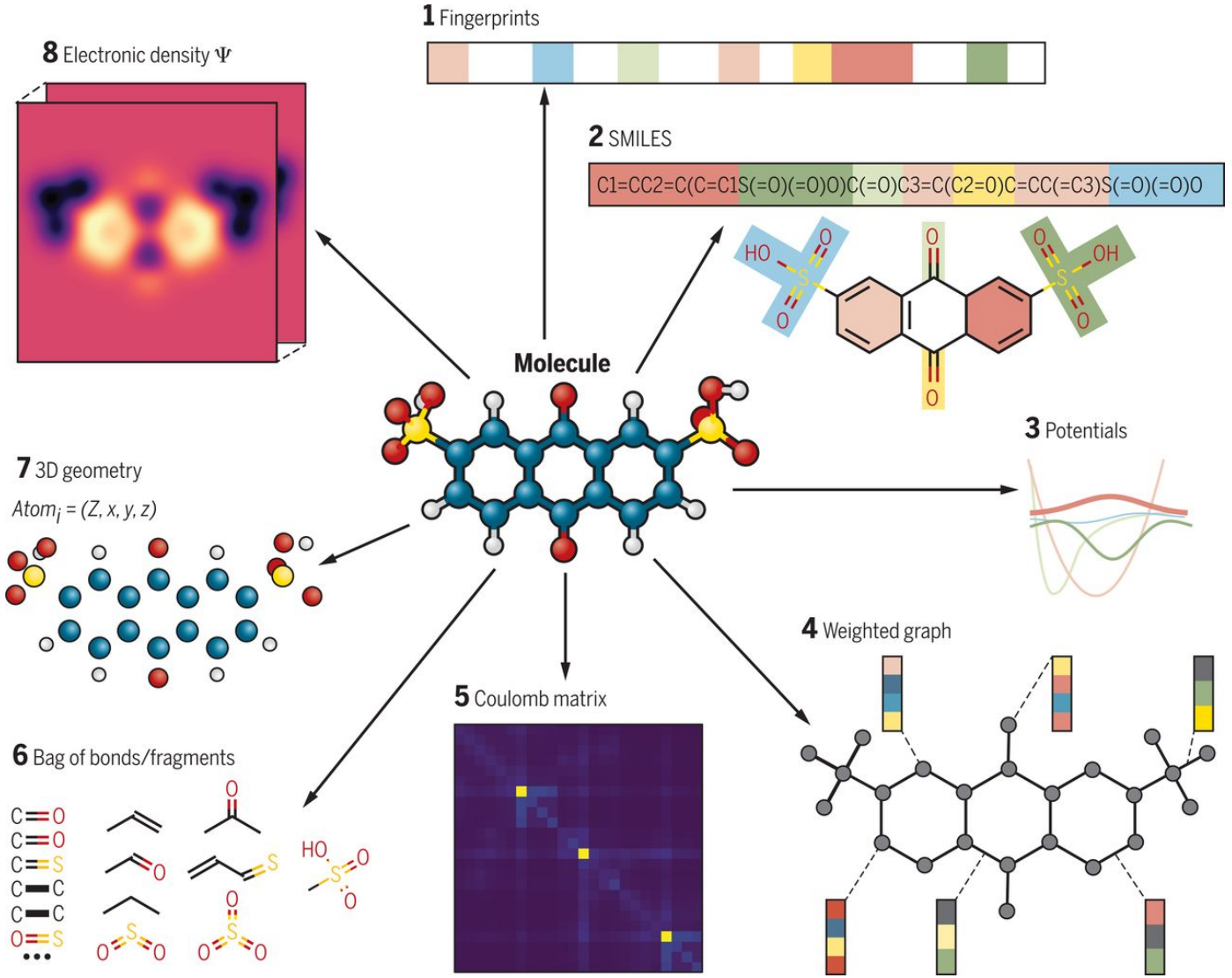
Then, we can use a clustering algorithm and pick the molecules from different clusters found to obtain a representative subset of 1000 molecules.

We run the experiments.

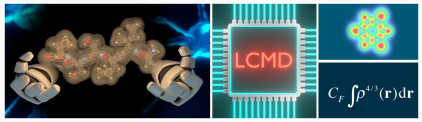
Then, we train a model on those 1000 compounds and predict the key property on the rest of the set.

Appendix

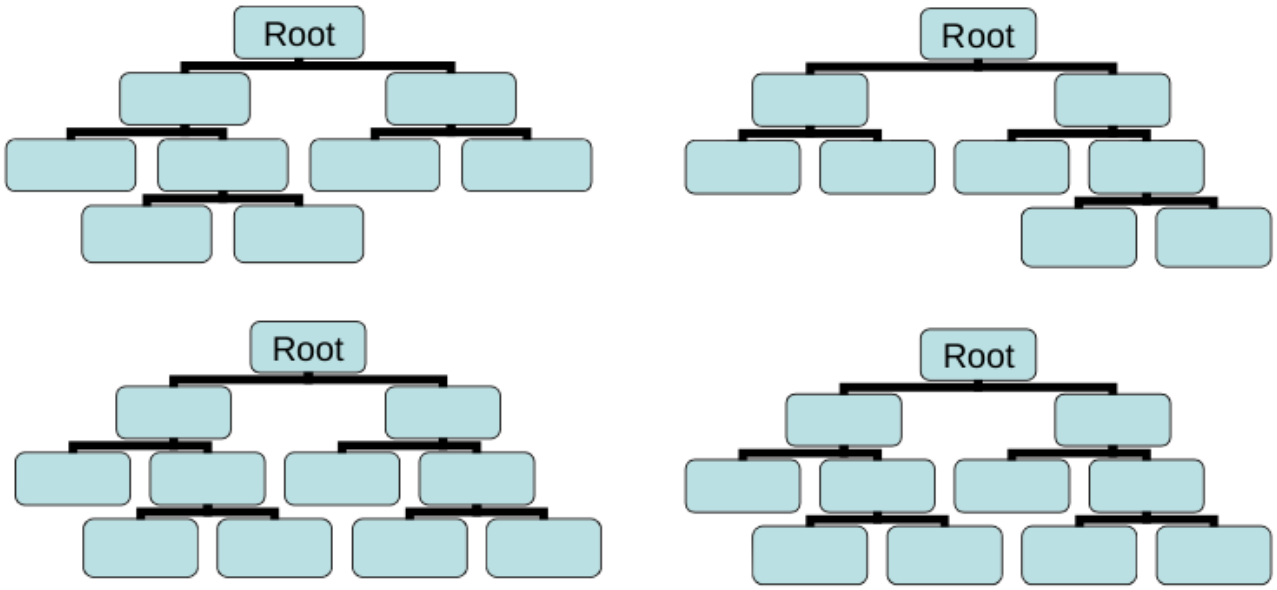
Molecular representations: beyond this introduction



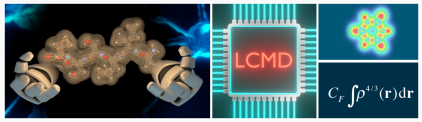
An ensemble ML method: Random Forests



- Ensemble methods combine the predictions of several base estimators built with a given ML algorithm in order to improve generalizability / robustness over a single model
- Random forest algorithm is an ensemble technique that combines multiple decision trees
- Predictions are made by majority vote of the individual trees
- Better generalization than an individual decision tree due to randomness



An ensemble ML method: Random Forests



- It runs efficiently on large databases
- It can handle thousands of input variables without variable deletion
- It gives estimates of what variables are important although less interpretative than single DT
- Performance increases with the number of trees until it saturates