EPFL

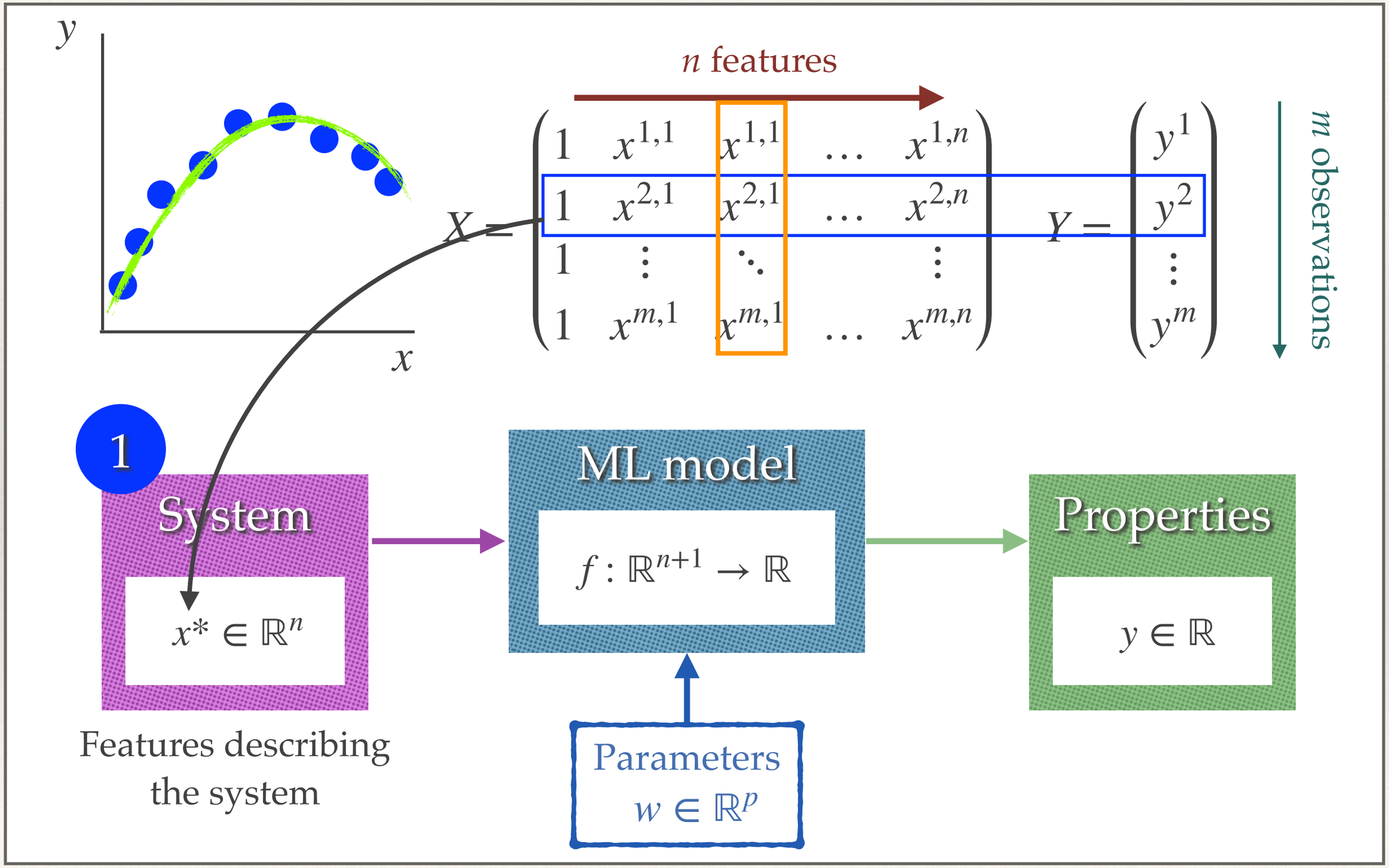
ChE-609

Seyed Mohamad Moosavi, Kevin Maik Jablonka & Berend Smit

Basics of machine learning for chemistry and materials science

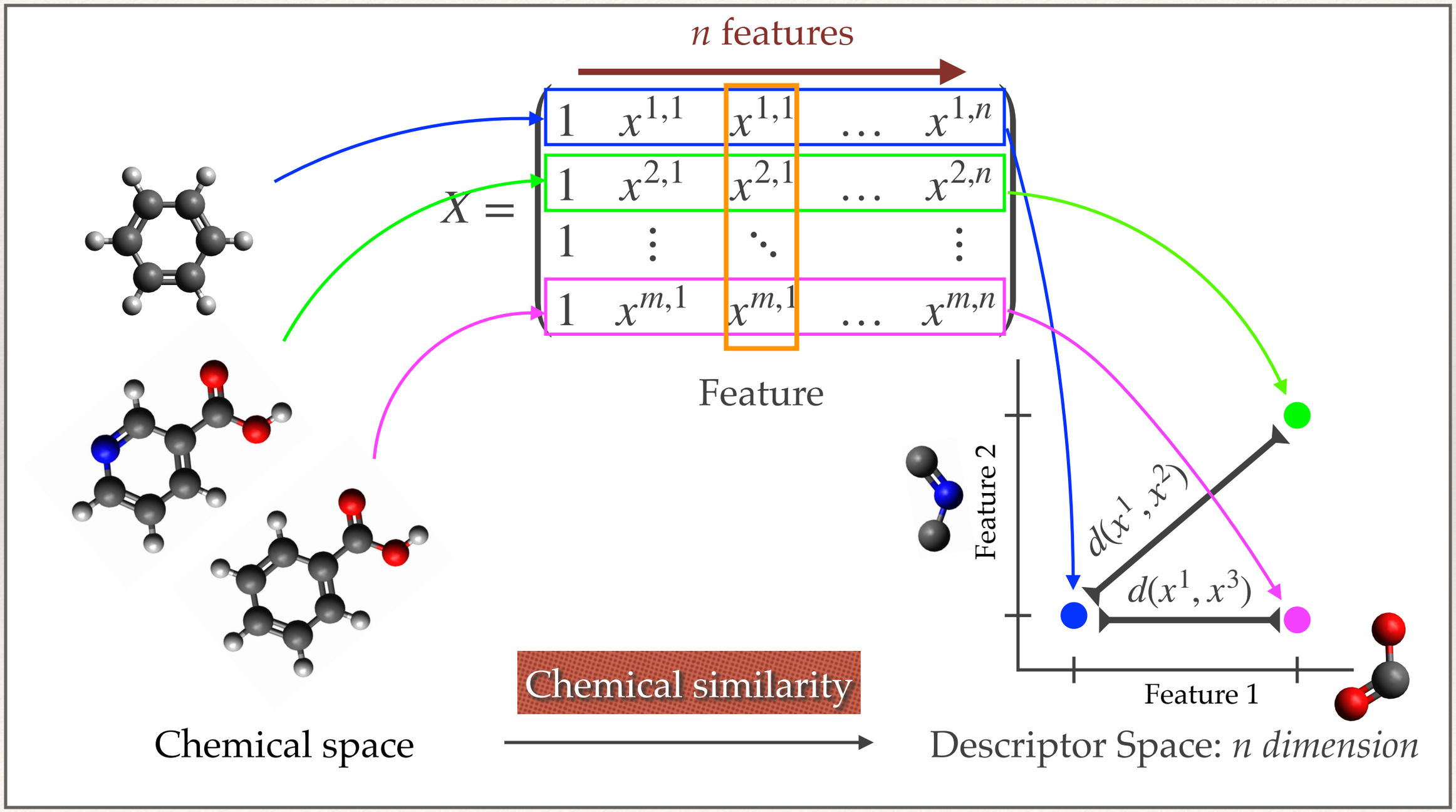
seyedmohamad.moosavi@epfl.ch @SMohMoosavi

Supervised machine learning



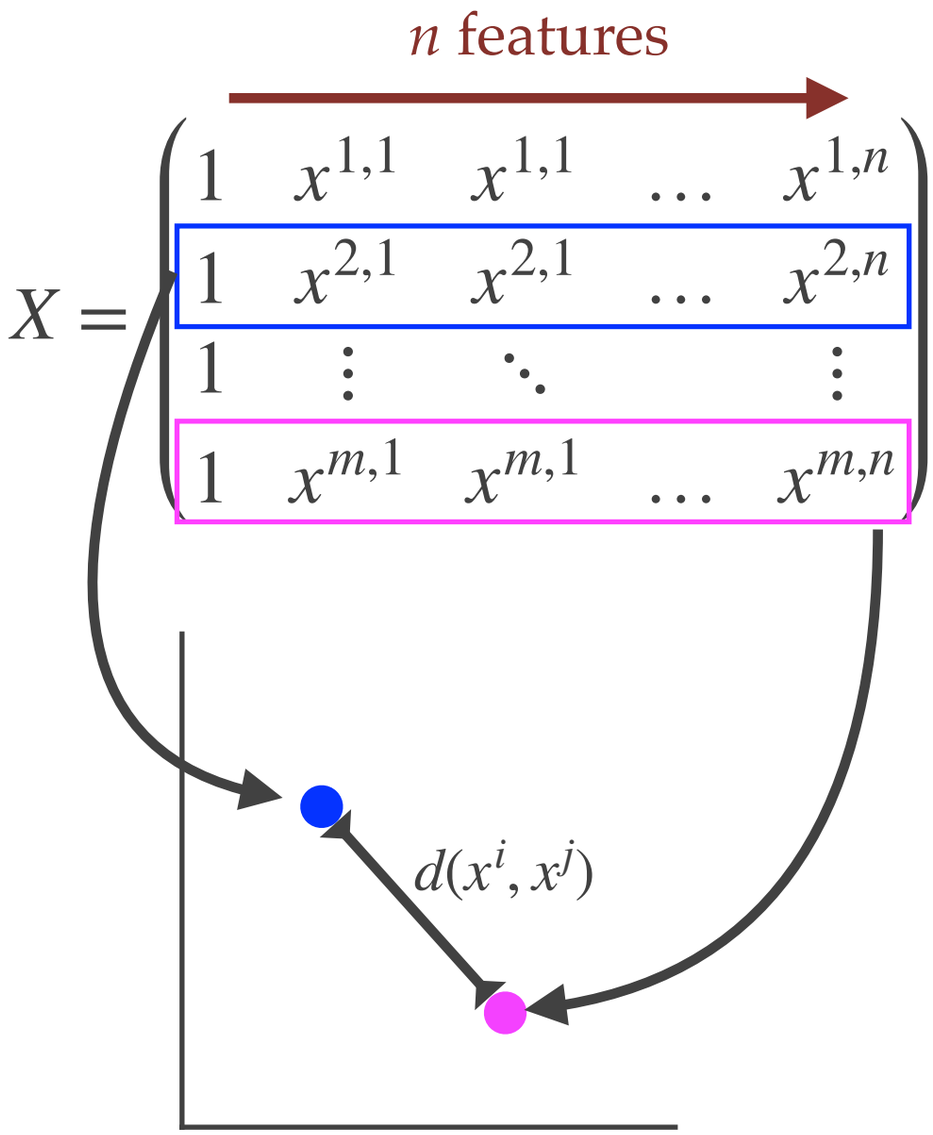
What is featurisation/a descriptor?

Encoding chemistry into numbers: “chemical space” to descriptor space



What makes a descriptor good?

Encoding chemistry into numbers



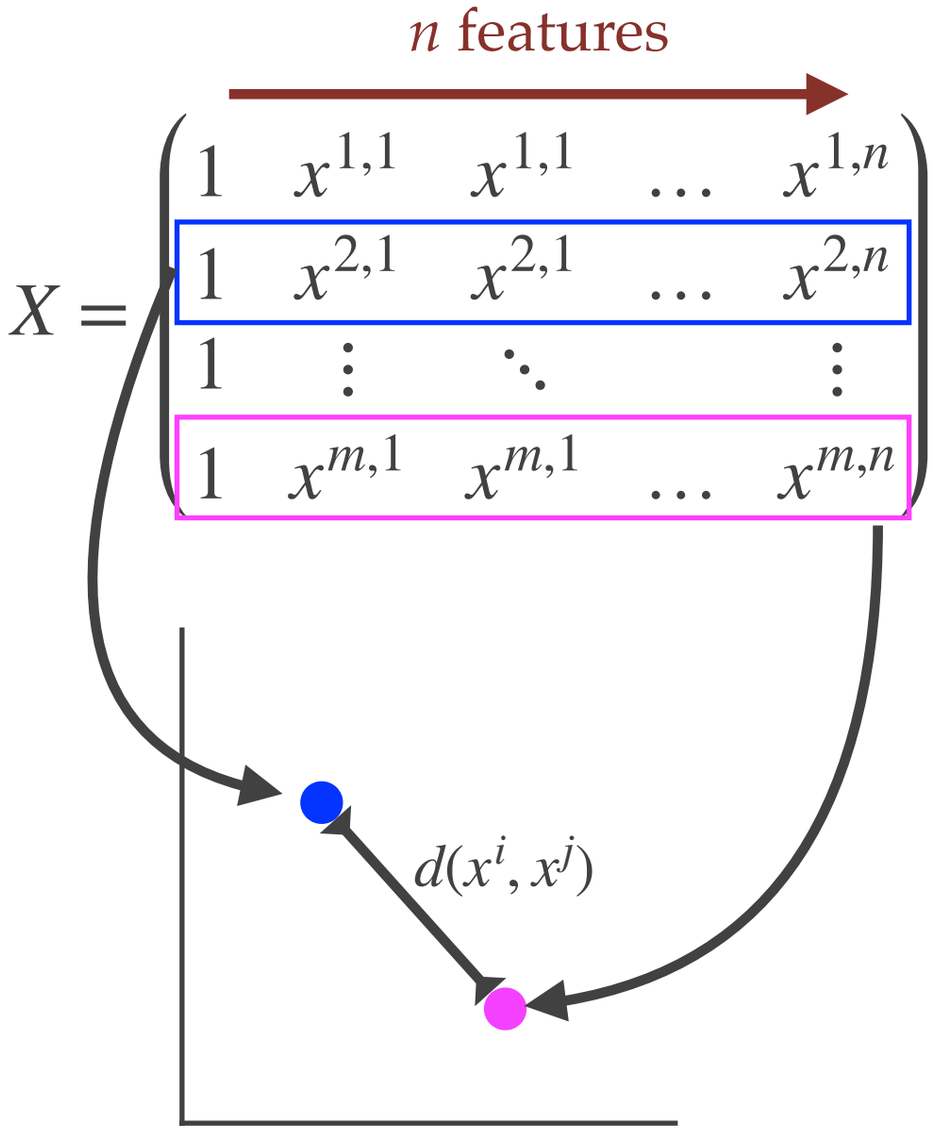
Descriptor Space: *n dimension*

Good descriptors —> obey physics

- Invariant w.r.t. symmetries
- As low dimension as possible
- Cheap to compute
- Non-degenerate
- Transferability across elements

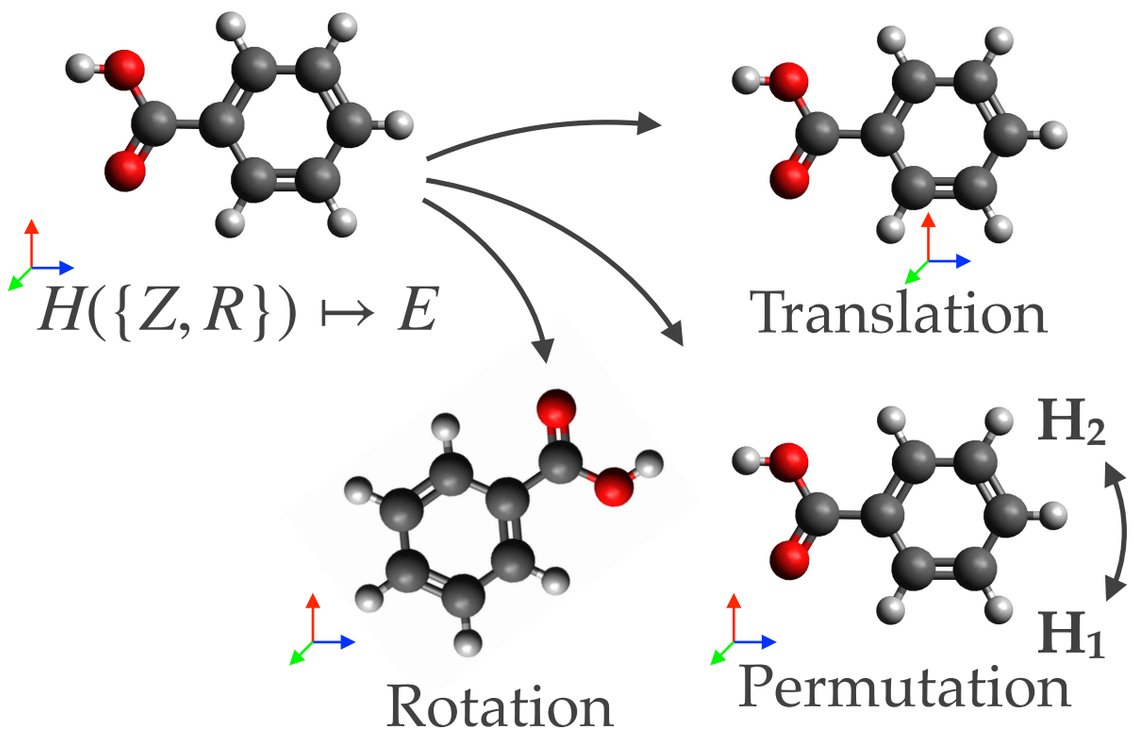
What makes a descriptor good?

Encoding chemistry into numbers



Good descriptors:

- Invariant w.r.t. symmetries
- As low dimension as possible
- Cheap to compute
- Non-degenerate
- Transferability across elements

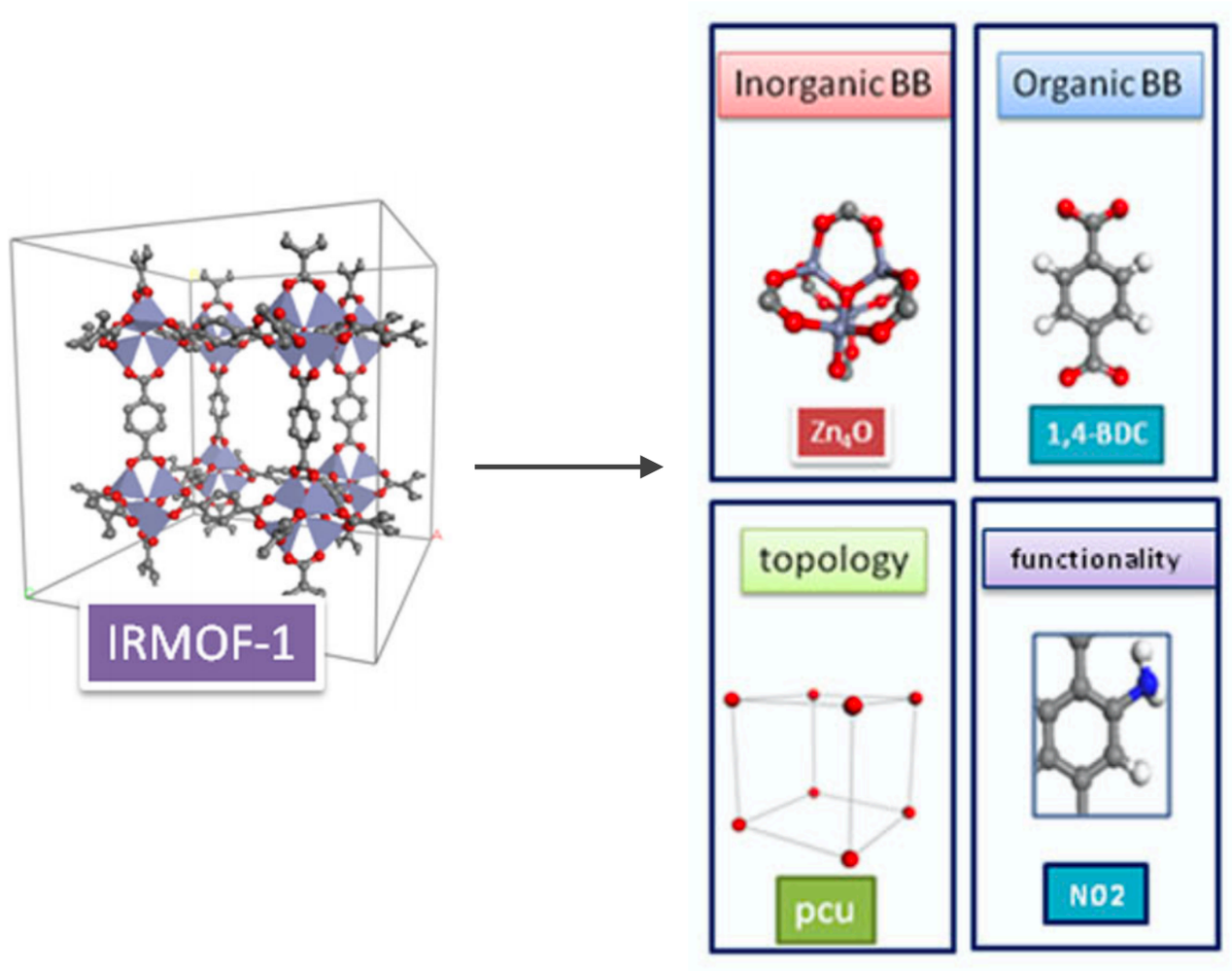


Ad hoc descriptors or properties:

Based on chemical intuition

In principle, can work but often not generalisable

One hot featurisation



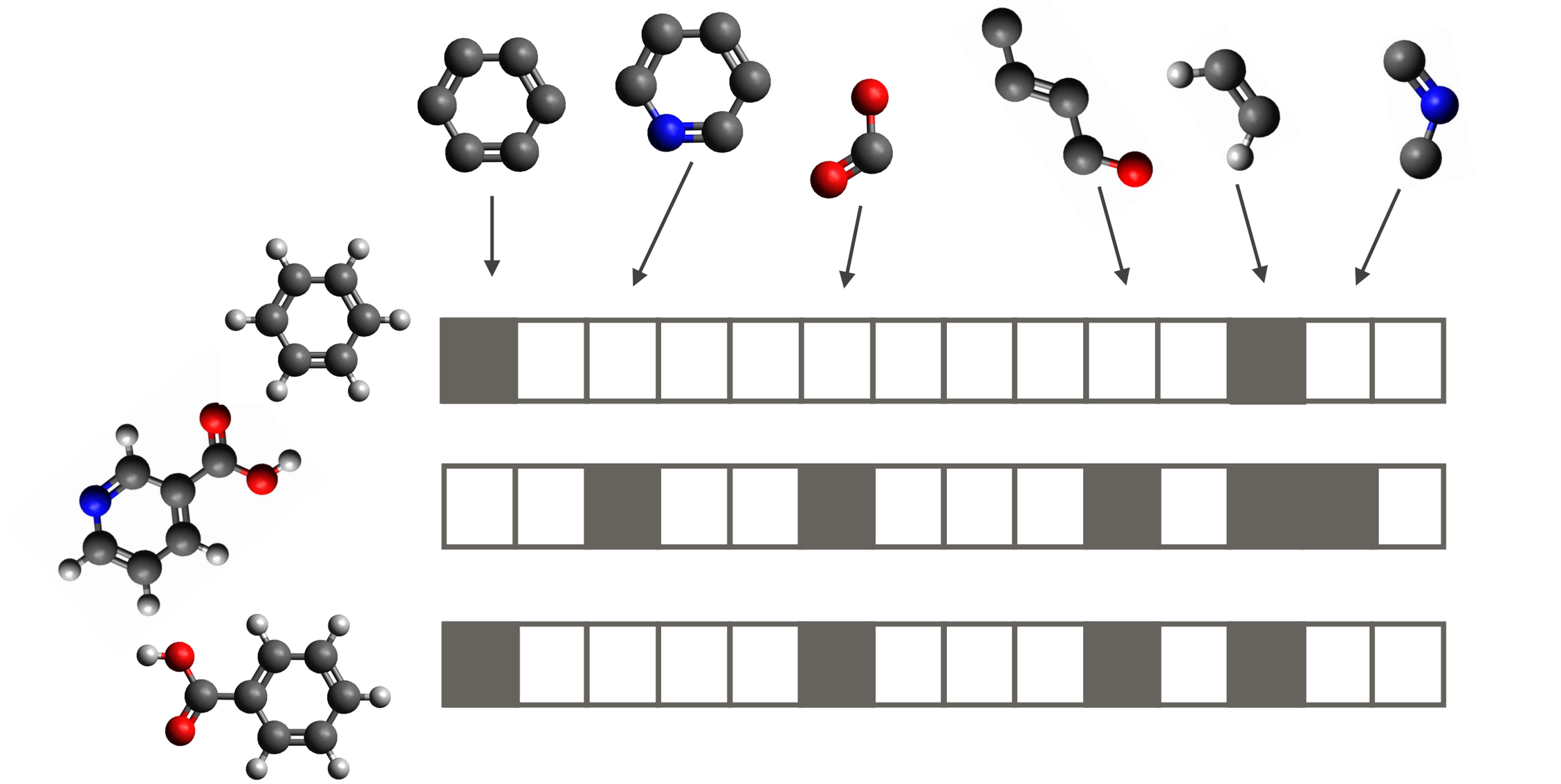
Computed properties:

- Atom identity
- Maximum positive charge
- Minimum negative charge
- etc.

Borboudakis, et al. npj computational materials (2017)
Anderson, et al. Chemistry of materials (2018)

Fragment based descriptors: fingerprints

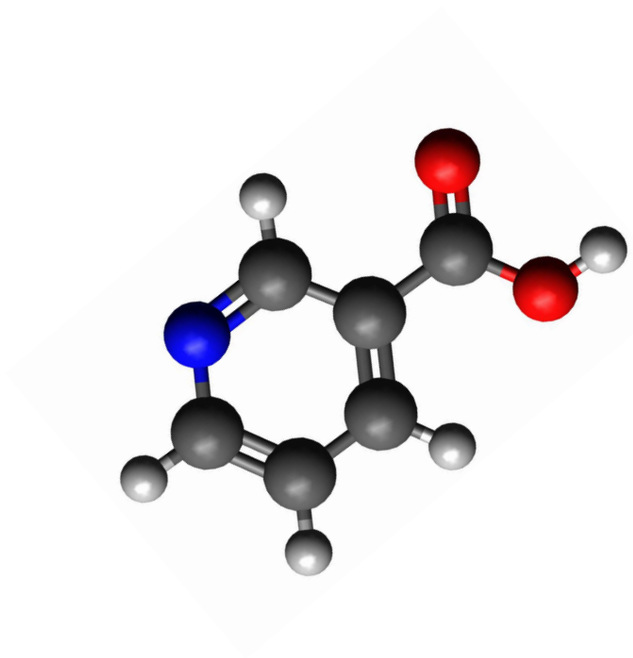
Binary vectors for molecular similarity
Varying length, e.g., FP2 fingerprint has 1024 bits



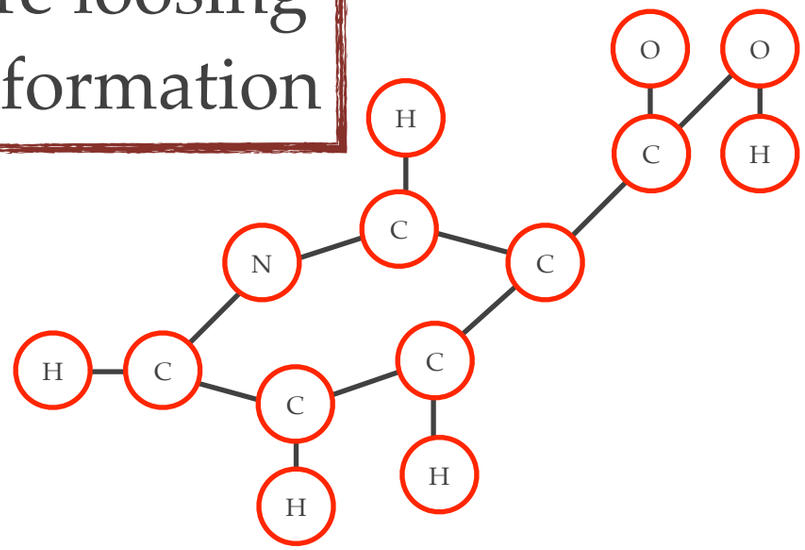
Rogers, D.; Hahn, M. “Extended-Connectivity Fingerprints.” J. Chem. Inf. and Model. **50**:742-54 (2010)

Connectivity based descriptors: RACs

Notice we are loosing geometric information



Molecular graph



Autocorrelations $\underset{scope}{^{start}}P_d^{diff} = \sum_{i}^{start} \sum_{j}^{scope} (P_i - P_j)\delta(d_{i,j}, d)$

Atomic properties χ, Z, T, S, I, α

Bond distance

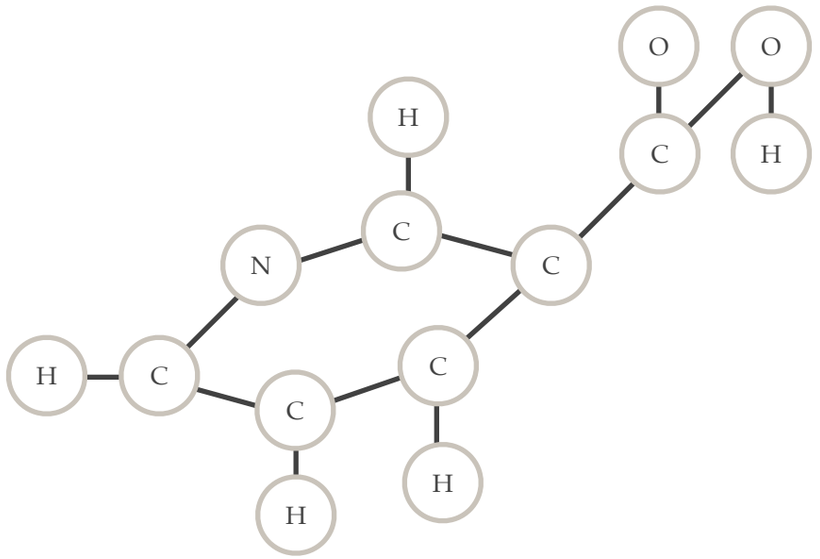
Janet, Jon Paul, and Heather J. Kulik. The Journal of Physical Chemistry (2017)

Connectivity based descriptors: RACs

Autocorrelations $${}^{start}_{scope}P^{diff}_{d} = \sum_{i}^{start} \sum_{j}^{scope} (P_i - P_j)\delta(d_{i,j}, d)$$

Examples:

$${}^{[N]}_{all}\chi^{diff}_{1} = \sum_{i}^{[N]} \sum_{j}^{all} (\chi_i - \chi_j)\delta(d_{i,j},1)$$

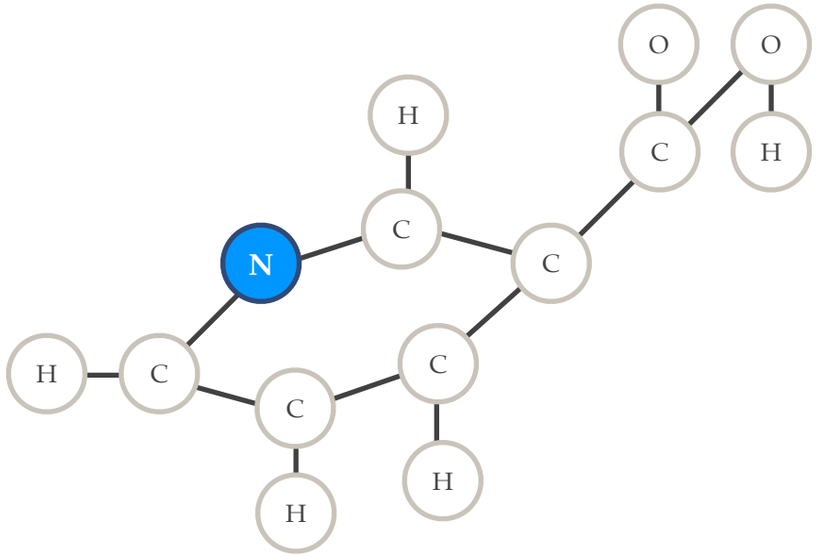


Connectivity based descriptors: RACs

Autocorrelations $${}^{start}_{scope}P^{diff}_{d} = \sum_{i}^{start} \sum_{j}^{scope} (P_i - P_j)\delta(d_{i,j}, d)$$

Examples:

$${}^{[N]}_{all}\chi^{diff}_{1} = \sum_{i}^{[N]} \sum_{j}^{all} (\chi_i - \chi_j)\delta(d_{i,j},1)$$



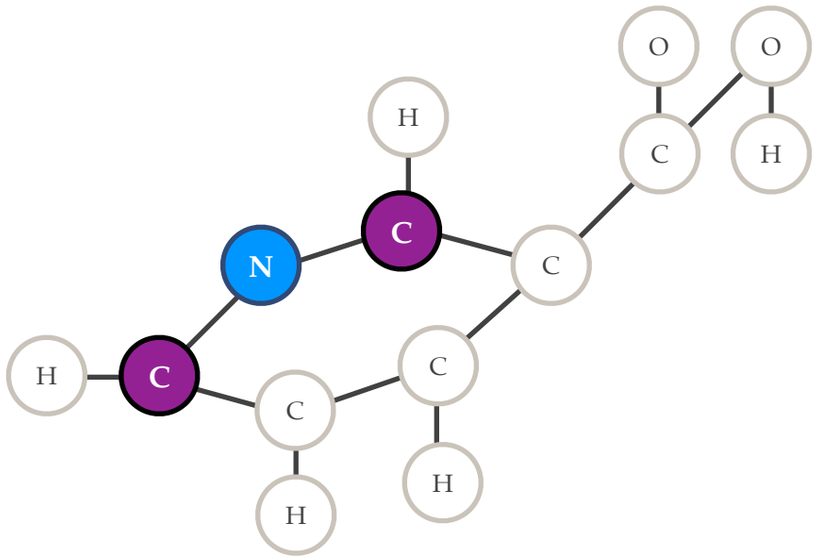
Connectivity based descriptors: RACs

Autocorrelations $${}^{start}_{scope}P_d^{diff} = \sum_i^{start} \sum_j^{scope} (P_i - P_j)\delta(d_{i,j}, d)$$

Examples:

1 bond distance

$${}^{[N]}_{all}\chi_1^{diff} = \sum_i^{[N]} \sum_j^{all} (\chi_i - \chi_j)\delta(d_{i,j},1)$$



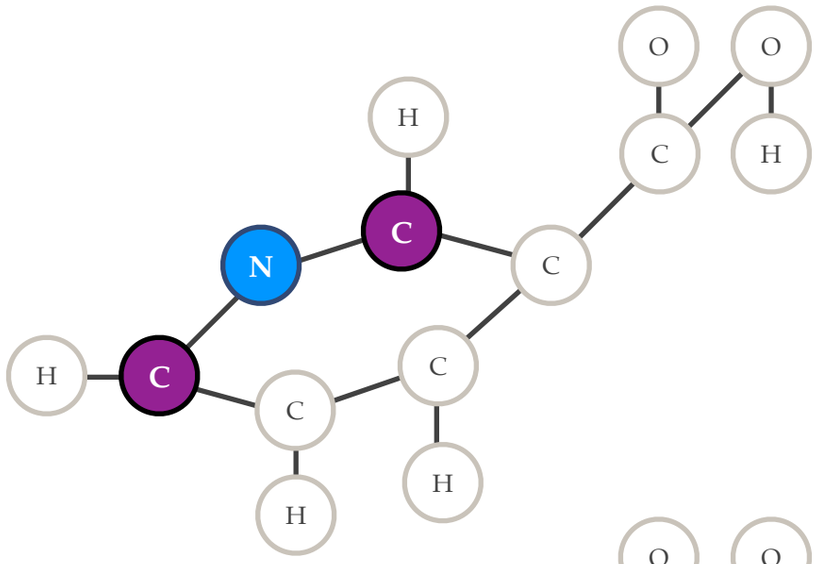
Connectivity based descriptors: RACs

Autocorrelations $${}^{start}_{scope}P_d^{diff} = \sum_i^{start} \sum_j^{scope} (P_i - P_j)\delta(d_{i,j}, d)$$

Examples:

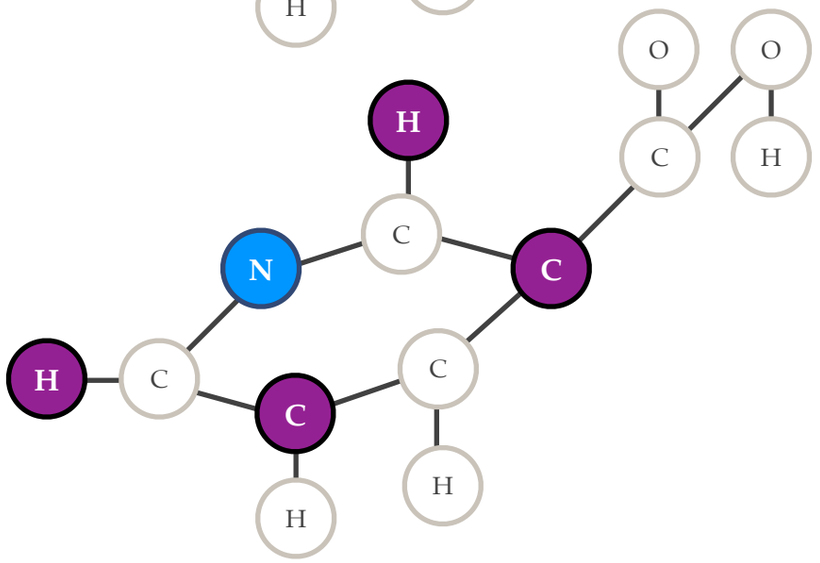
1 bond distance

$${}^{[N]}_{all}\chi_1^{diff} = \sum_i^{[N]} \sum_j^{all} (\chi_i - \chi_j)\delta(d_{i,j},1)$$



2 bond distance

$${}^{[N]}_{all}\chi_2^{diff} = \sum_i^{[N]} \sum_j^{all} (\chi_i - \chi_j)\delta(d_{i,j},2)$$



RACs for MOFs, hands on session in the afternoon

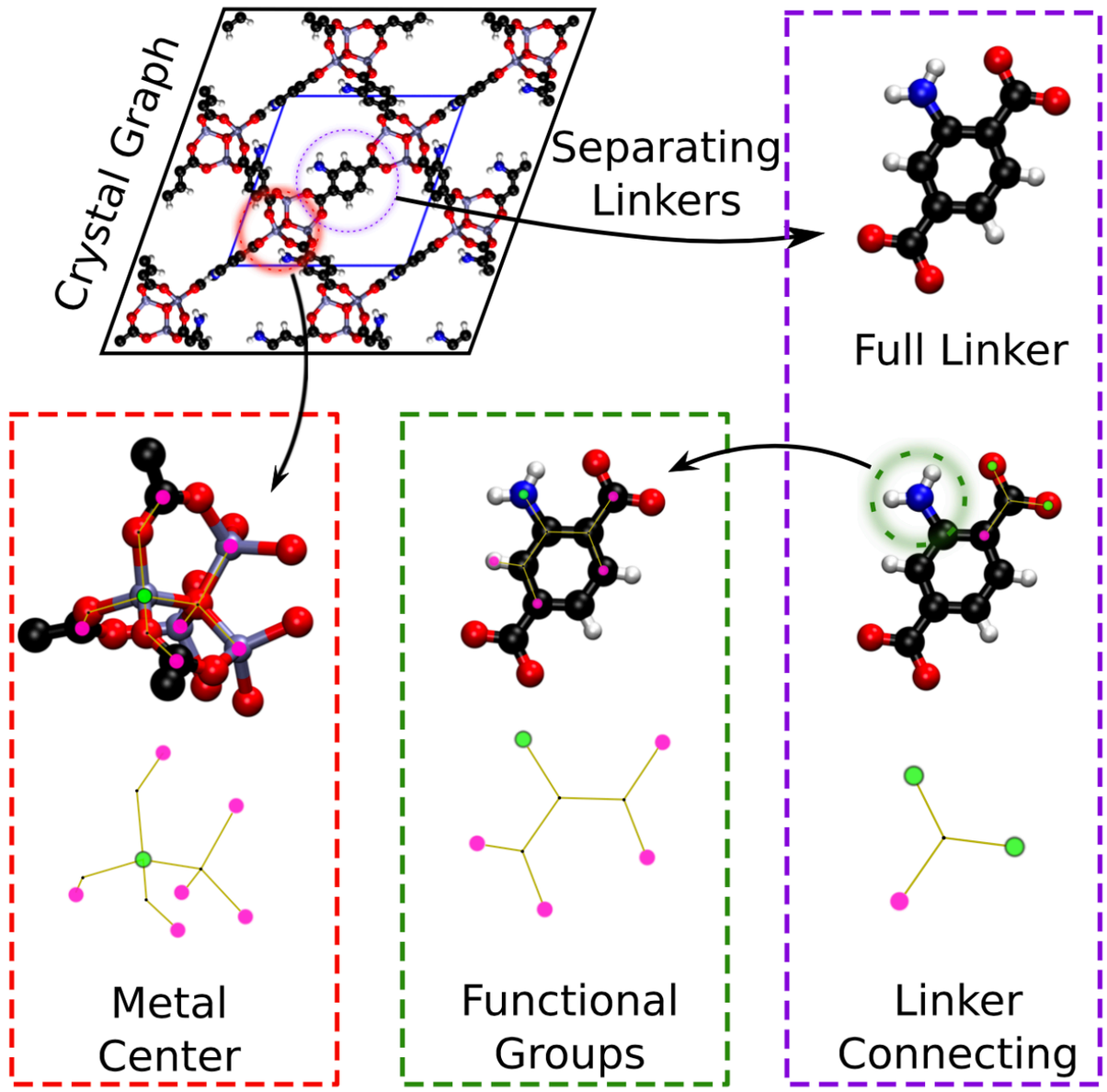
molSimplify

$$ {}^{start}_{scope}P^{diff}_{d} = \sum_{i}^{start\ scope} \sum_{j} (P_i - P_j)\delta(d_{i,j}, d) $$

$$ {}^{start}_{scope}P^{prod}_{d} = \sum_{i}^{start\ scope} \sum_{j} (P_i \times P_j)\delta(d_{i,j}, d) $$

$$ {}^{start}_{scope}R^{type}_{depth} $$

$$ \chi, Z, T, S, I, \alpha $$



https://github.com/hjkgrp/molSimplify

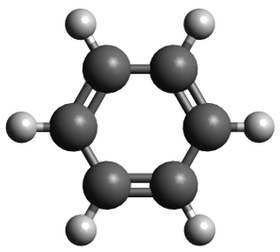
13

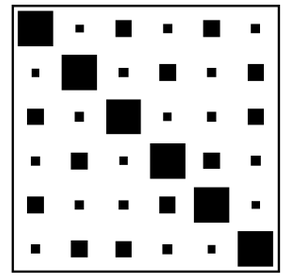
Encoding geometry: Coulomb matrix

Inspired by how quantum mechanics works:

$$H(\{Z,R\}) \xrightarrow{\Psi} E \qquad \Leftrightarrow \qquad \{Z,R\} \xrightarrow{\text{ML}} E$$

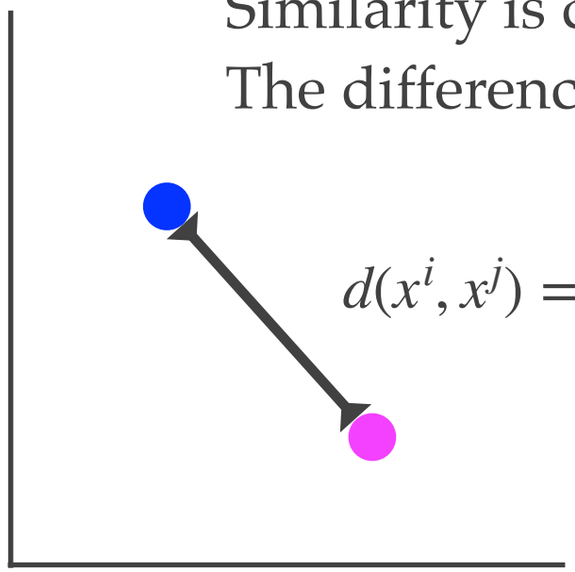
$$M_{ij} = \begin{cases} 0.5\, Z_i^{2.4} & i = j \\ \dfrac{Z_i\, Z_j}{(|\mathbf{r}_i - \mathbf{r}_j|)} & i \neq j \end{cases}$$





Similarity is defined as:

The difference in eigenvalues of **M**s between two systems

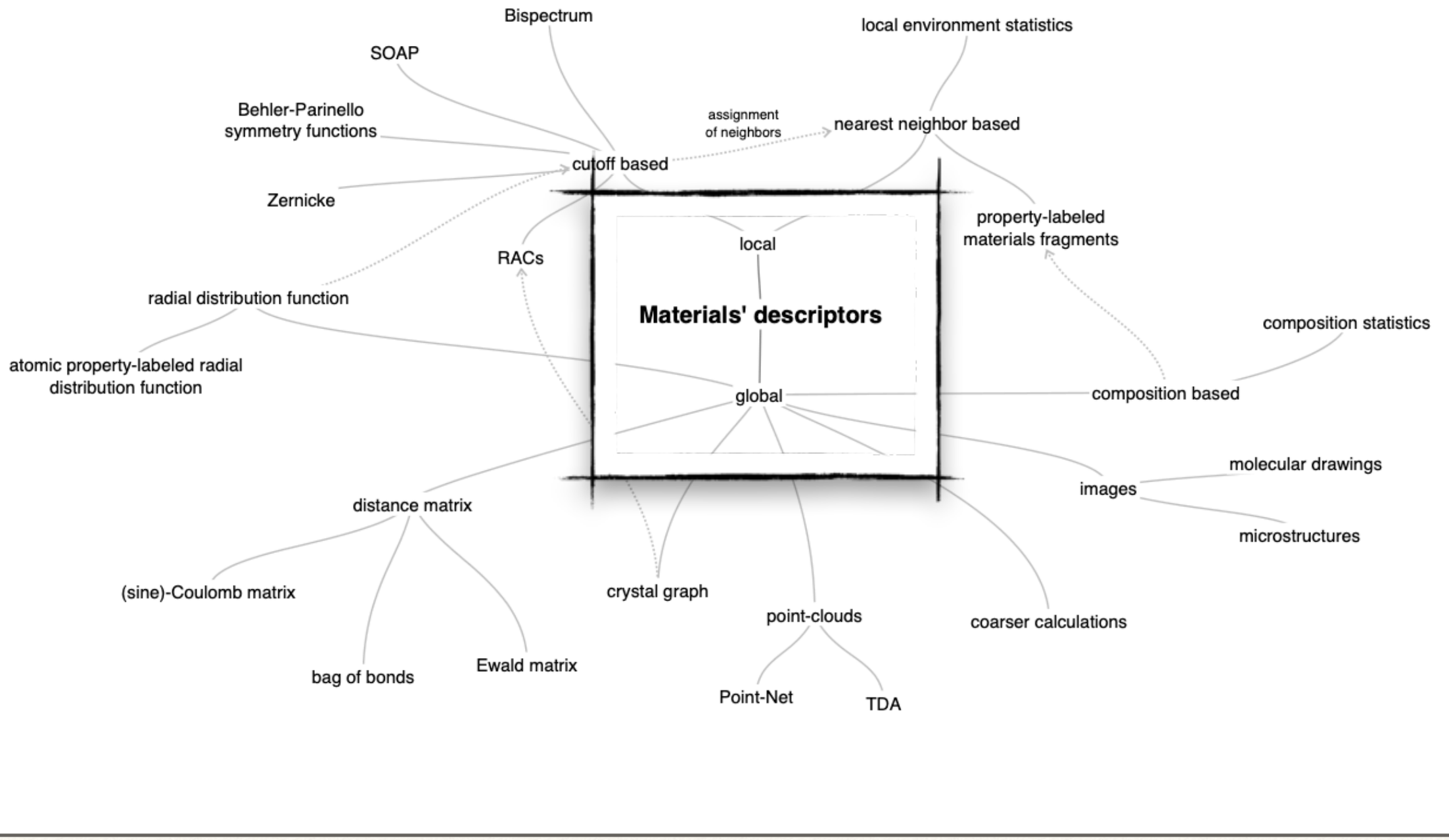


$$d(x^i, x^j) = d(\epsilon^i, \epsilon^j) = \sqrt{\sum_I |\epsilon^i - \epsilon^j|}$$

Rupp, Matthias, et al. "Fast and accurate modeling of molecular atomization energies with machine learning." Physical review letters 108.5 (2012): 058301.

And there exist many more!

The choice must be motivated with the application



Maik Jablonka, Kevin, et al. "Big-Data Science in Porous Materials: Materials Genomics and Machine Learning." arXiv (2020): arXiv-2001.

Encoding local environments: symmetry functions

Chemical Locality Assumption: decomposing property into local environments

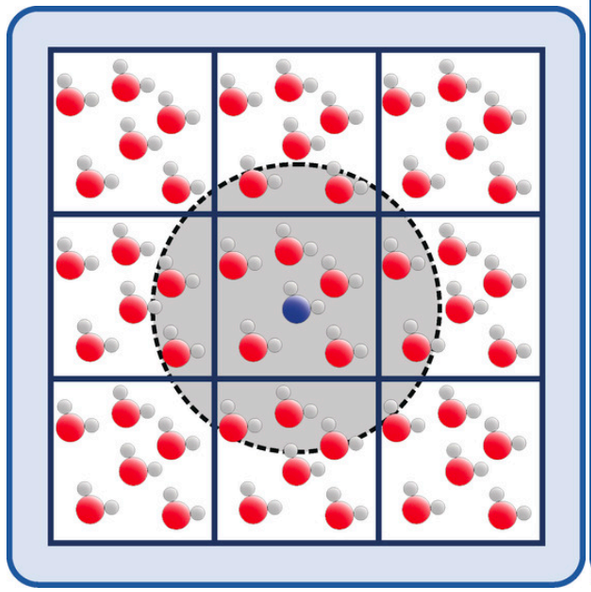
$$\text{property}(\text{descriptor}) = \sum_{i}^{\text{atoms}} \text{models}_i(\text{descriptor}_i)$$

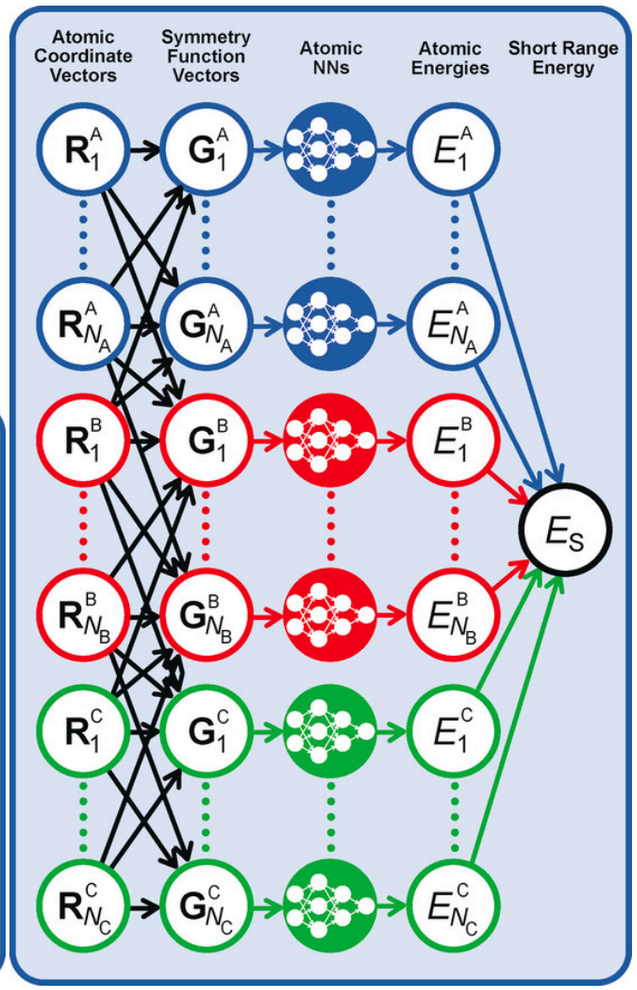
Energy can be decomposed into atomic contributions
—> this approach is used to describe PES
—> Scalable to large systems
—> differentiability of descriptors is essential

$$E_S = \sum_{\nu=1}^{N_{atoms,\nu}} \sum_{\mu=1}^{N_{elem}} E_\mu^\nu$$

$$f_{\text{cut}}(r_{ij}) = \begin{cases} \frac{1}{2}\left[\cos\left(\pi \frac{r_{ij}}{r_c}\right) + 1\right] & \text{for } r_{ij} \le r_{\text{cut}} \\ 0 & \text{for } r_{ij} > r_{\text{cut}} \end{cases}$$

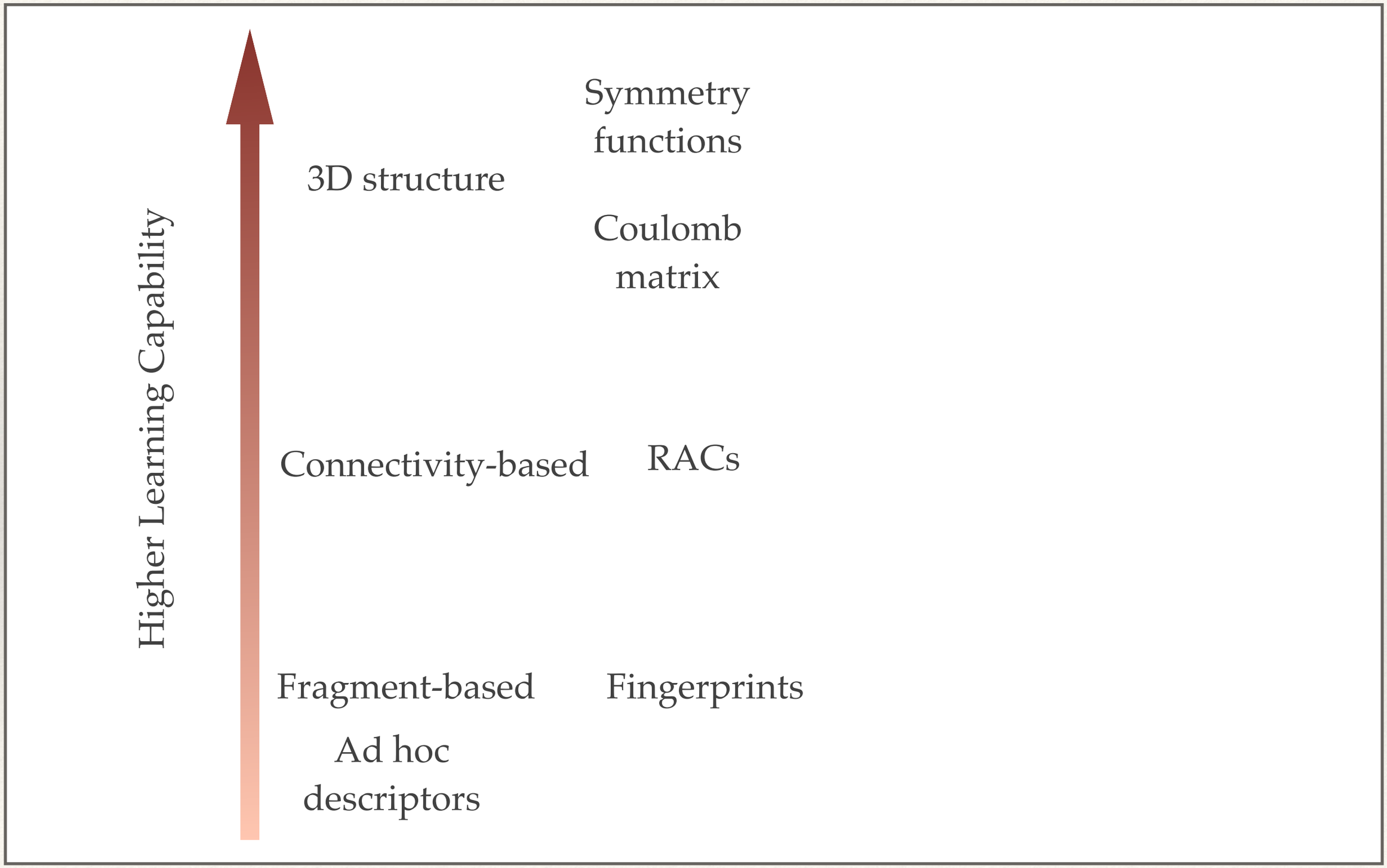
$$G_i^2 = \sum_j \exp\left[-\eta_i (r_{ij} - r_{si})^2\right] f_{cut}(r_{ij})$$





Behler, Jörg, and Michele Parrinello. "Generalized neural-network representation of high-dimensional potential-energy surfaces." *Physical review letters* 98.14 (2007): 146401.

Complexity and richness



Summary of featurisation

- The aim is to map chemical space to numbers, such that:
  - Chemical similarity is preserved
  - Physics obeyed
- Many kinds of representation exist
  - Global vs. Local
  - Richness and complexity
- Should choose representation based on the application