Data-driven path collective variables for atomic scale transformations

Arthur France-Lanord, Hadrien Vroylandt, Fabio Pietrucci, Benjamin Rotenberg, Marco Saitta, Mathieu Salanne

ISCD, Sorbonne Université

MAterials for Energy through STochastic sampling and high peRformance cOmputing

MAESTRO

An ISCD-funded initiative bringing together mathematicians, physicists, chemists

The focus is on computational materials science: atomic-scale modeling techniques for the description of transformations in energy storage materials

11 PIs, including:

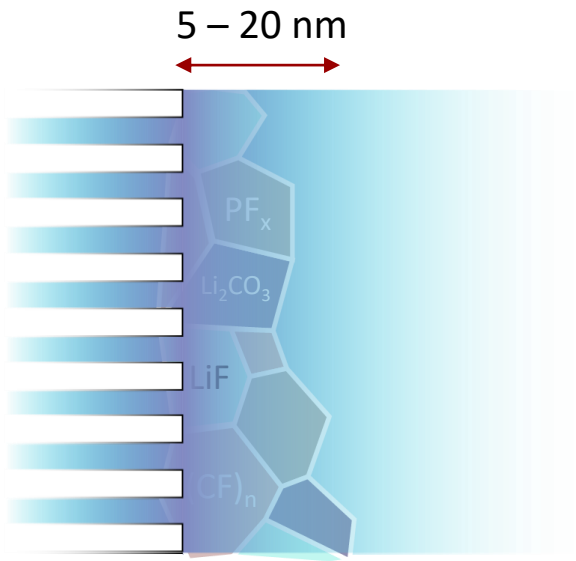
- Marco Saitta (IMPMC)
- Fabio Pietrucci (IMPMC)
- Benjamin Rotenberg (PHENIX)
- Mathieu Salanne (PHENIX)

2 Postdocs hosted at ISCD:

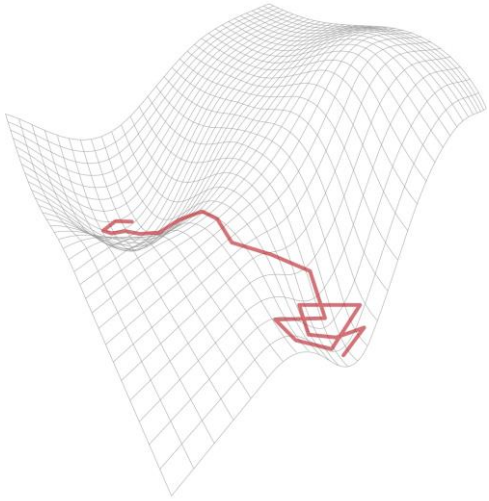
- Hadrien Vroylandt (Langevin equations)
- Arthur France-Lanord (projection quality)

Transformations in energy storage materials, and computational approach

SEI (Solid Electrolyte Interphase)

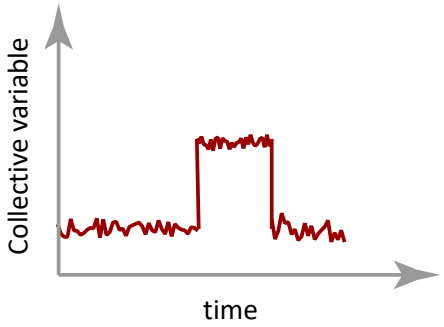


Molecular dynamics



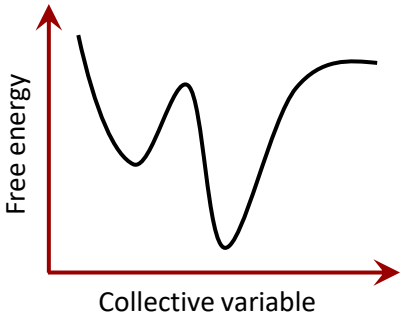
Configurational space

$\mathbb{R}^{3N}$



Collective variable

$\mathbb{R}^{M}$, M≪3N



Free energy, rates

What we want:

- Thermodynamics
- Kinetics

$$F \propto -k_B T \ln Z_\Sigma$$

$$Z_\Sigma = \int_\Sigma e^{-\beta U(\mathbf{X})} d\mathbf{X}$$

$$\mathbf{X} \in \mathbb{R}^{3N}$$

$$\mathbf{X} \rightarrow \boldsymbol{\xi}$$

$$F(\boldsymbol{\xi}) = -k_B T \ln \rho(\boldsymbol{\xi})$$

France-Lanord, Pietrucci, Saitta, Tarascon, Grimaud, and Salanne
PRX Energy (2022)

Atomic-scale modeling for transformations: collective variables $\boldsymbol{\xi}$

- A collective variable (CV) is a (usually low-dimensional) projection of the configurational space **X**

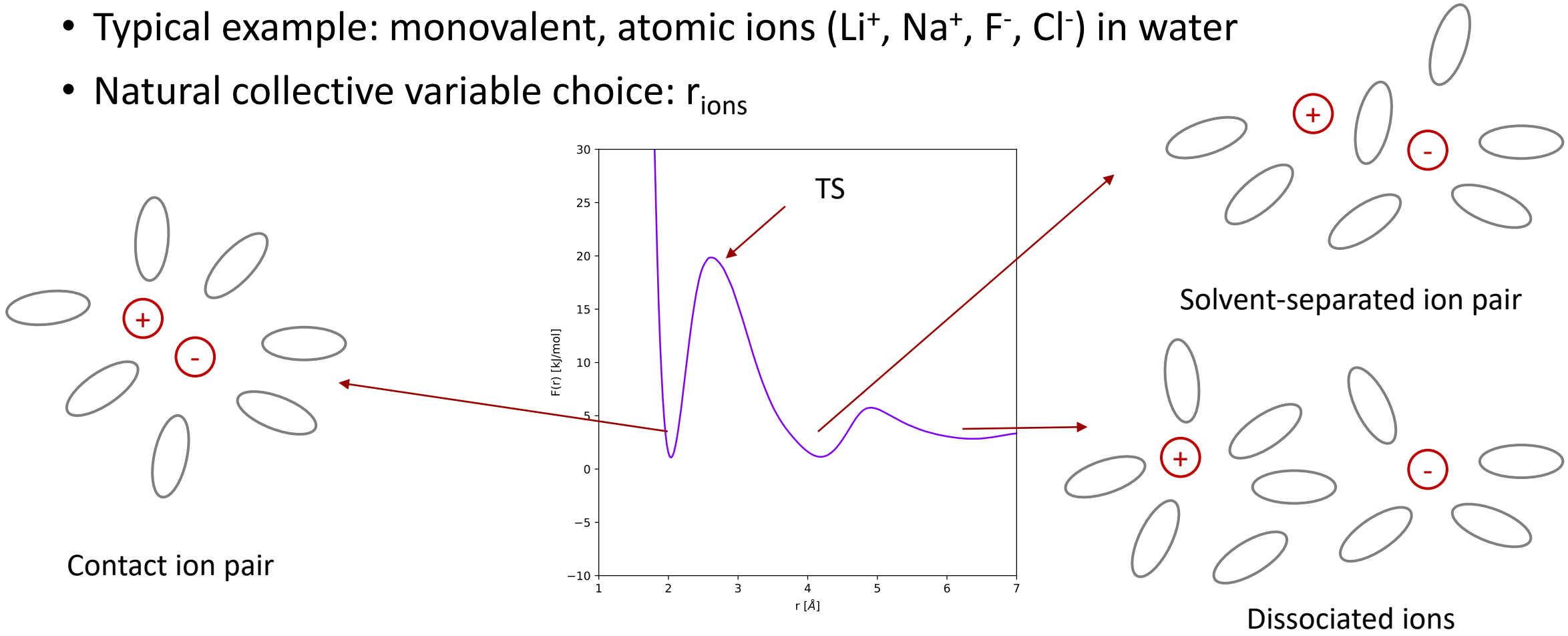
$$\mathbf{X} \rightarrow \boldsymbol{\xi} \qquad\qquad F(\boldsymbol{\xi}) = -k_B T \ln \rho(\boldsymbol{\xi})$$

- $F(\boldsymbol{\xi})$ is a free energy surface, the Boltzmann inversion of a marginal probability density $\rho(\boldsymbol{\xi})$

- CVs allow to rationalize reaction mechanisms, and to obtain free energy differences and reaction rates which can be compared to experimental data

An illustration: ion pairing in solution

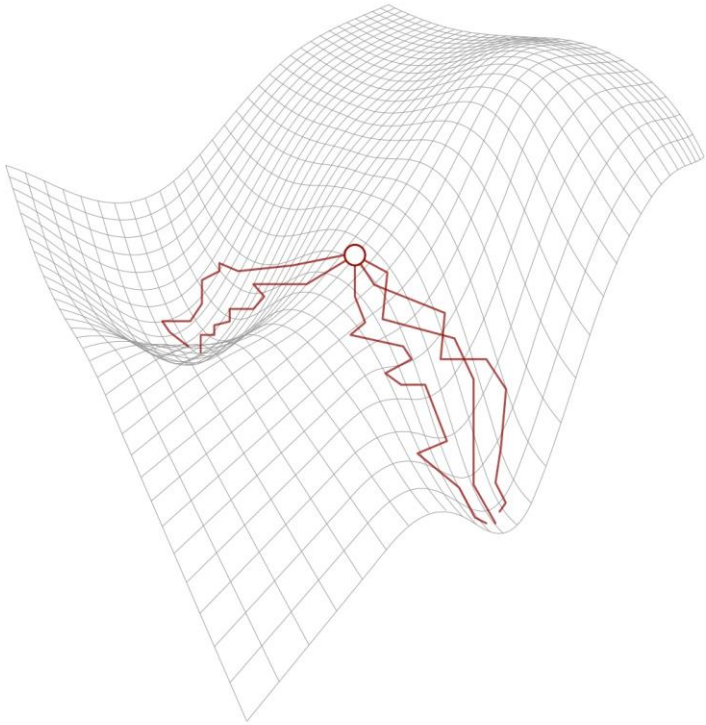
- We're interested in understanding how ions interact in solution
- Typical example: monovalent, atomic ions (Li^+, Na^+, F^-, Cl^-) in water
- Natural collective variable choice: r_{ions}



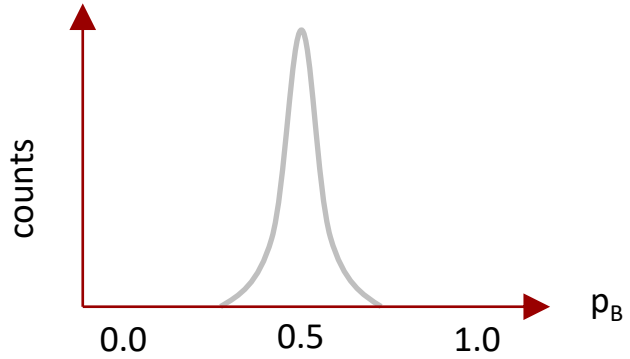
How can we know if this is a good CV?

Assessing CV quality through a committor analysis

The committor p_B is the probability that a configuration will reach basin B before reaching basin A. It is the ideal CV

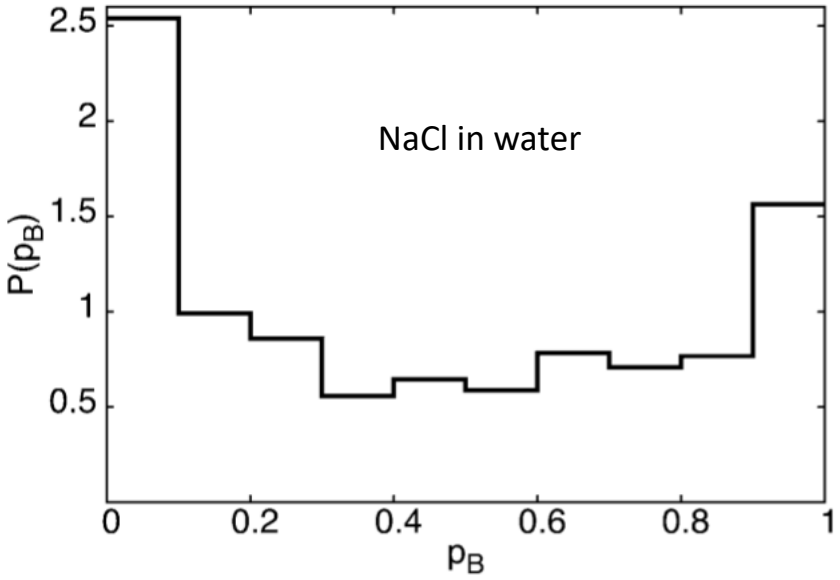


1. Select N configurations at the putative TS (here, $r_{ions}=r^*$)
2. ∀ configuration, initialize atomic velocities, and let the system evolve until it reaches a basin (here CIP or SSIP). Repeat M times
3. If a configuration is a TS, it will have equal odds to "commit" to the two basins (p_B=0.5)

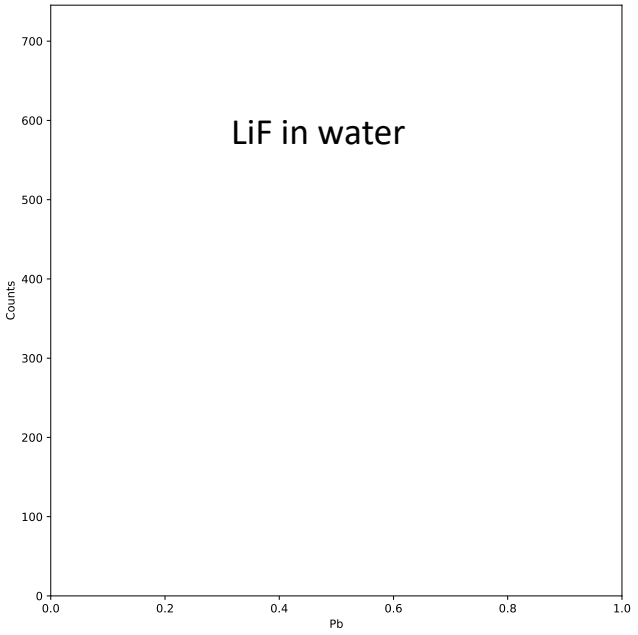


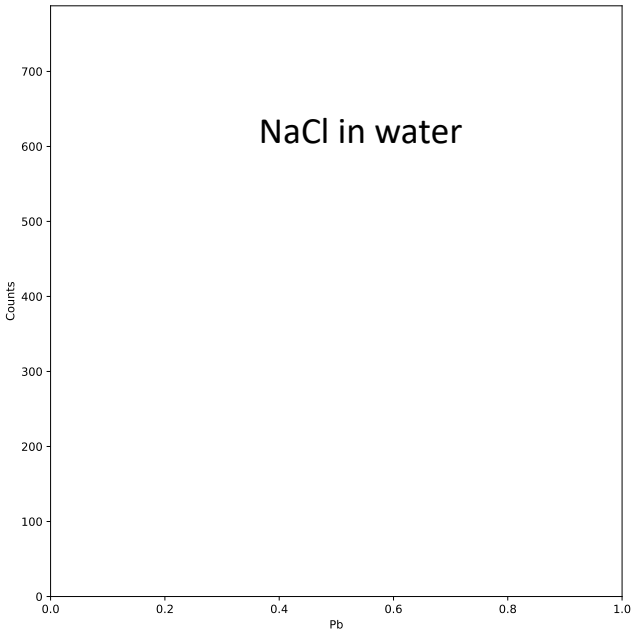
For an ideal collective variable: the p_B distribution at the putative TS should be sharply peaked at 0.5

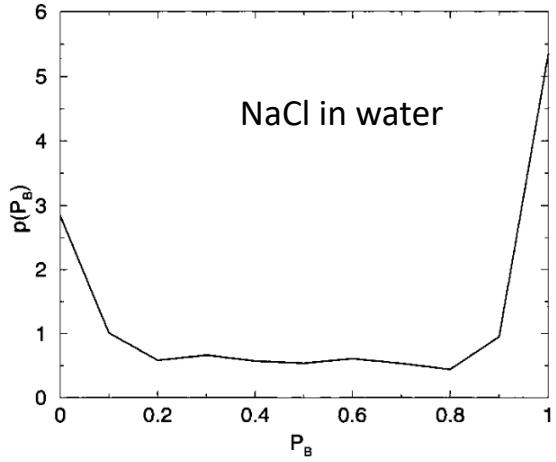
Committor analysis of monovalent ions in water



Ballard and Dellago, J. Phys. Chem. B 2012



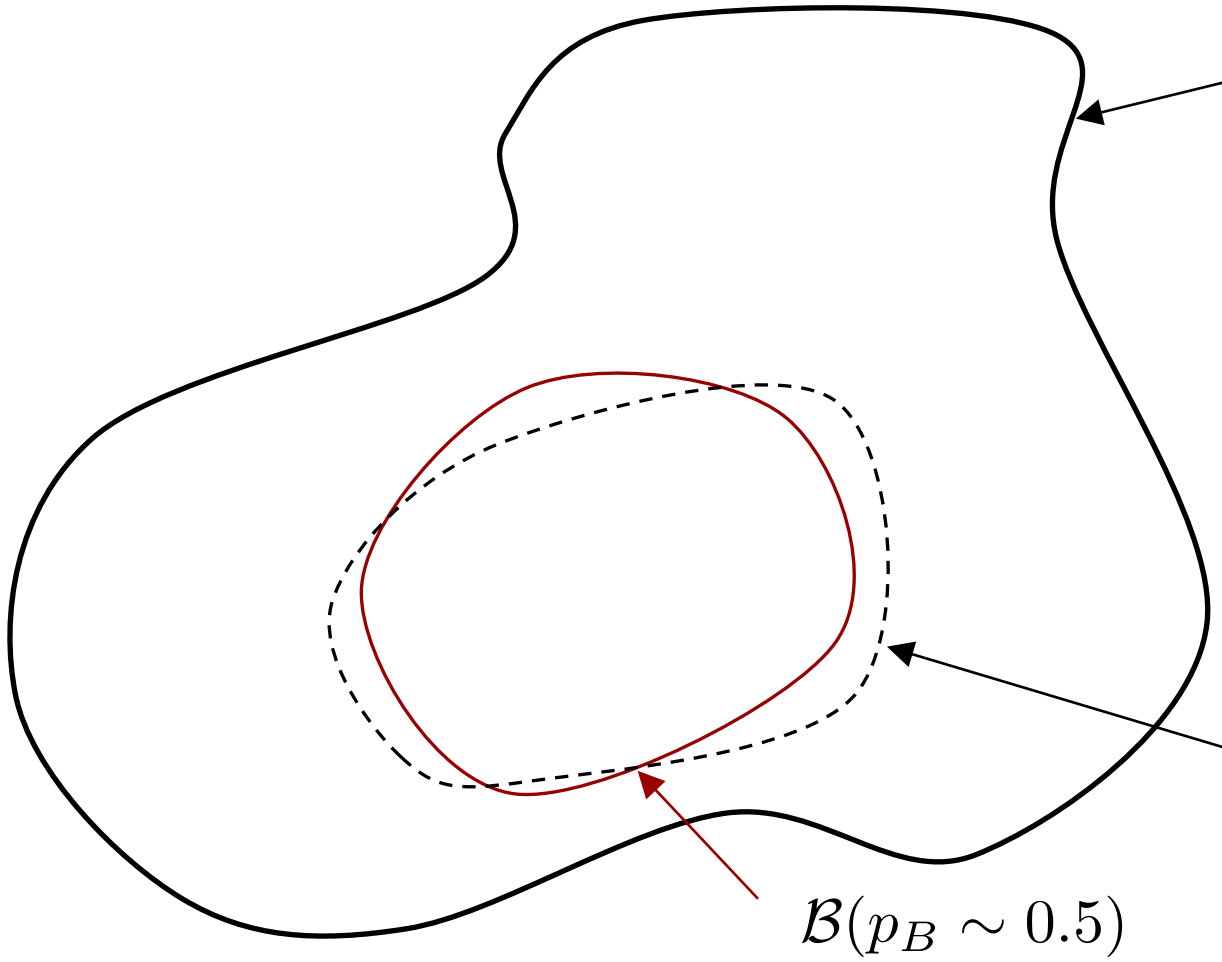




Geissler, Dellago, Chandler, J. Phys. Chem. B 1999

r_{ions} is not enough: we have to include solvent degrees of freedom too!

Committor analysis of monovalent ions in water



r_{ions} is not enough: we have to include solvent degrees of freedom too!

How can we rate CV “quality”?

One CV to rule them all

We start from a path collective variable:

$$s(\xi) = \frac{\sum_{i=1,2} i \exp -\lambda D\left[\xi, \xi_i\right]}{\sum_{i=1,2} \exp -\lambda D\left[\xi, \xi_i\right]}$$

Branduardi, Gervasio, Parrinello, JCP 2007

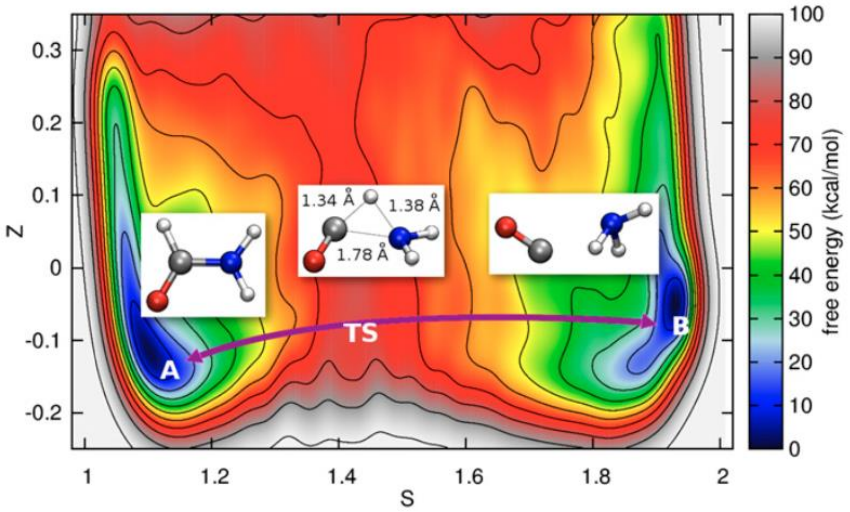
Which is, in fact, a kernel regression estimator:

$$s(\xi) = \frac{\sum_{i=1}^{N} iK(\xi, \xi_i)}{\sum_{i=1}^{N} K(\xi, \xi_i)}$$

Nadaraya-Watson estimator

Our approach is a generalization of PCVs based on data:

$$s(\xi) = \frac{\sum_{i=1}^{N} p(B|\xi_i) K\left(\xi, \xi_i\right)}{\sum_{i=1}^{N} K\left(\xi, \xi_i\right)}$$



Pietrucci and Saitta, PNAS 2015

Data-driven path collective variable

Our approach is a generalization of PCVs based on data:

$$s(\xi) = \frac{\sum_{i=1}^{N} p(B|\xi_i) K(\xi, \xi_i)}{\sum_{i=1}^{N} K(\xi, \xi_i)}$$

We go one step further, from kernel smoothing to kernel ridge regression:

$$s(\xi) = \sum_{i=1}^{N} \alpha_i K(\xi, \xi_i)$$

$$\hat{\alpha} = \arg\min_{\alpha \in \mathbb{R}^n} \|y - \mathbf{K}\alpha\|_2^2 + \lambda \alpha^T \mathbf{K} \alpha$$

- Regularization limits model complexity
- At least as good as kernel smoothing
- Many libraries implement efficient KRR, including falkon[1,2]
  - Up to billions of reference points
  - CPU/multi GPU using pytorch

1. Meanti, Carratino, Rosasco, Rudi. “Kernel methods through the roof: Handling billions of points efficiently”. NeurIPS 2020

2. Meanti, Carratino, De Vito, Rosasco. “Efficient hyperparameter tuning for large scale kernel ridge regression”, arXiv:2201.06314, 2022

Data-driven path collective variable

Practical approach:

1. Select a kernel type
2. Construct reference, training, and test sets by computing the committor of configurations
3. Optimize hyperparameters (bandwidths and regularization)

$$K(\boldsymbol{\xi}_i, \boldsymbol{\xi}) = \exp\left(-\gamma \left\|\boldsymbol{\xi}_i - \boldsymbol{\xi}\right\|_2^2\right)$$

RBF kernel, squared ℓ^2-norm

$$K(\boldsymbol{\xi}_i, \boldsymbol{\xi}) = \exp\left(-\gamma \left\|\boldsymbol{\xi}_i - \boldsymbol{\xi}\right\|_1\right)$$

Laplacian kernel, ℓ^1-norm

Three main sources of errors:

Projection

$$\mathbf{X} \rightarrow \boldsymbol{\xi}$$

Optimization

$$\exp\left(-\gamma \left\|\boldsymbol{\xi}_i - \boldsymbol{\xi}\right\|_1\right)$$

Regression

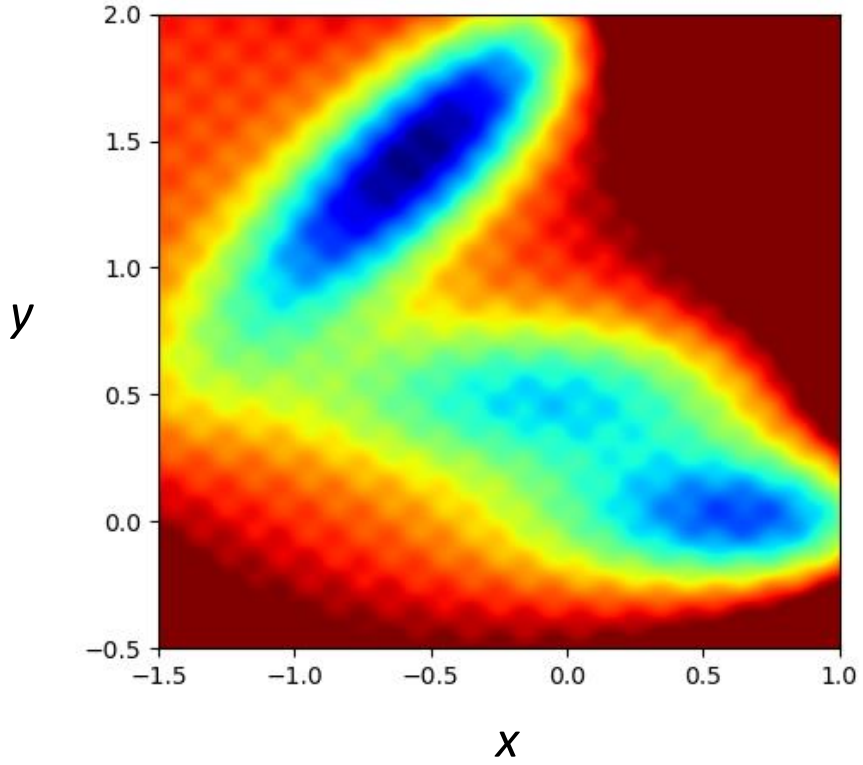
$$s(\boldsymbol{\xi}) = \sum_{i=1}^{N} \alpha_i K\left(\boldsymbol{\xi}, \boldsymbol{\xi}_i\right)$$

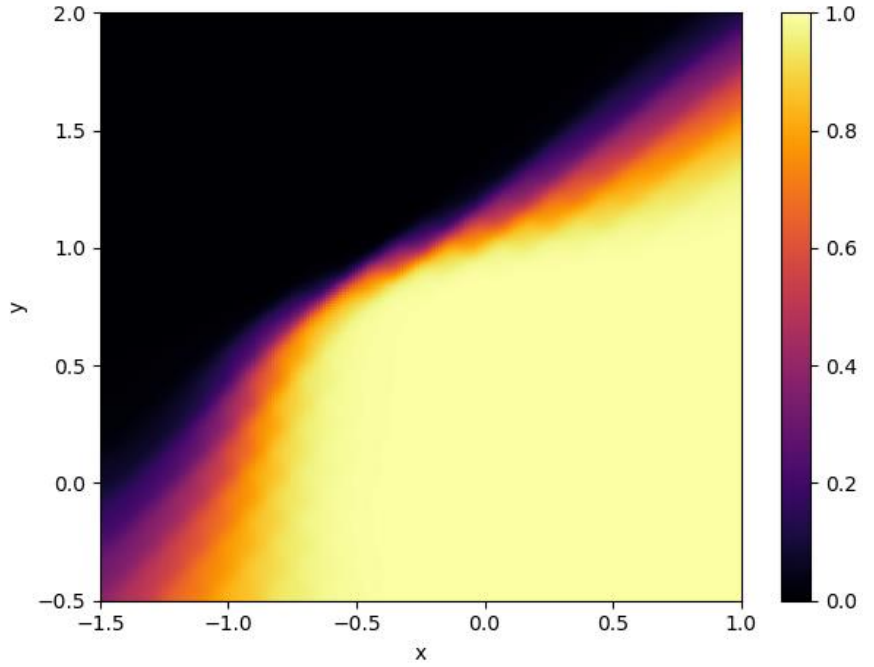
A toy model: the rugged Müller-Brown potential

$$U_{mb}(x,y) = \sum_{i=1}^{4} D_i e^{a_i(x-X_i)^2 + b_i(x-X_i)(y-Yi) + c_i(y-Y_i)^2}$$

$$U_{rmb}(x,y) = U_{mb}(x,y) + \gamma \sin(2k\pi x)\sin(2k\pi y)$$

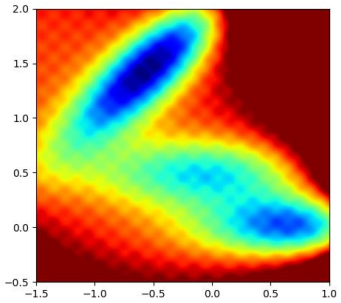
Exact committor from the overdamped Langevin generator using FEM





$p(B|(x,y))$

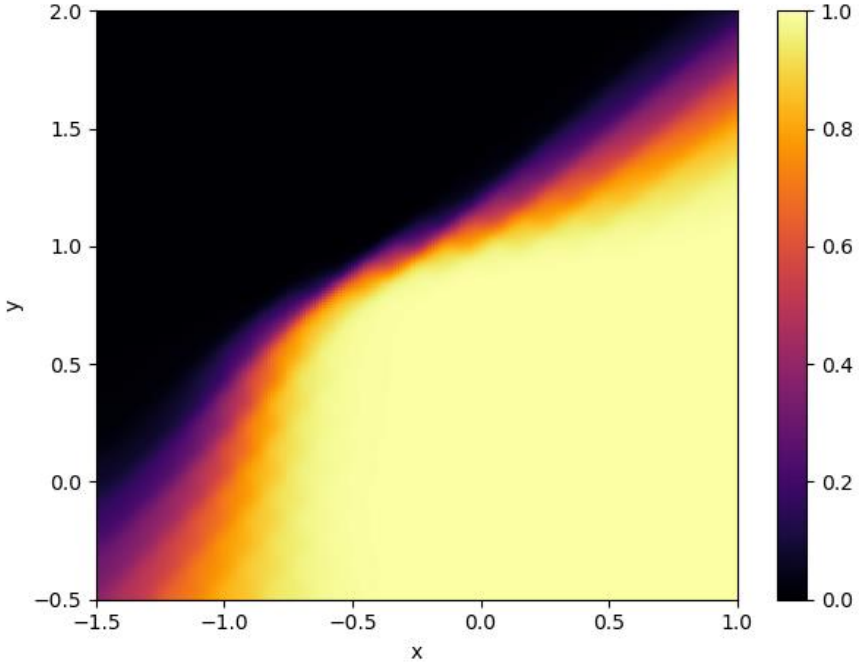
A toy model: the rugged Müller-Brown potential



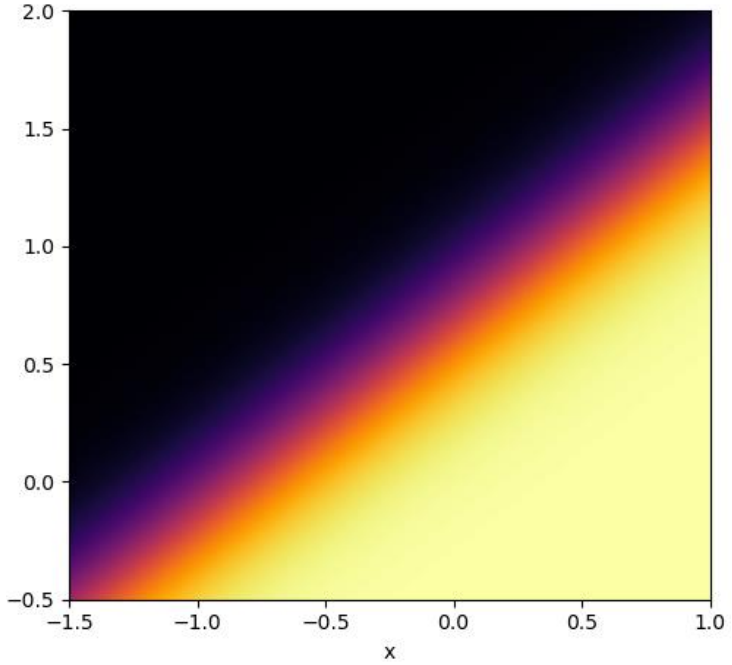
$$s(\xi) = \frac{\sum_{i=1}^{N} iK(\xi, \xi_i)}{\sum_{i=1}^{N} K(\xi, \xi_i)}$$

$$s(\xi) = \sum_{i=1}^{N} \alpha_i K(\xi, \xi_i)$$

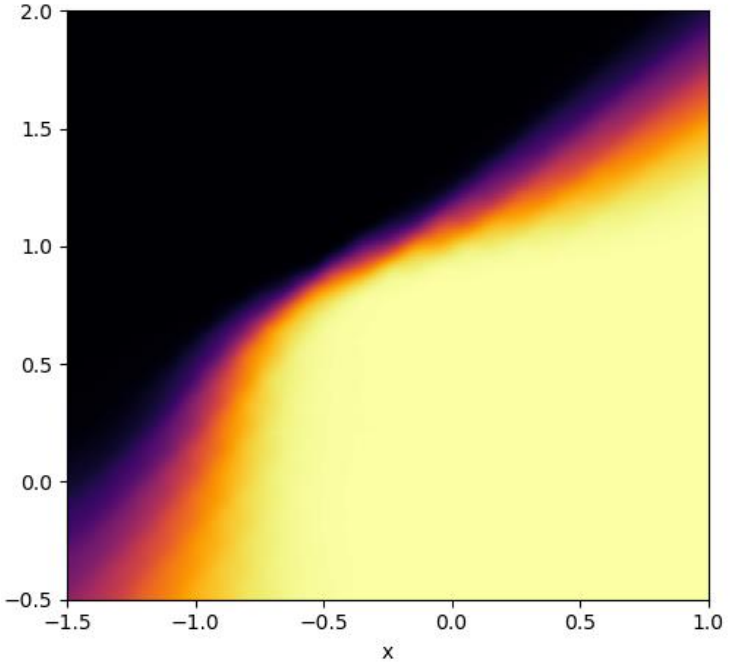
Exact committor from the overdamped Langevin generator using FEM



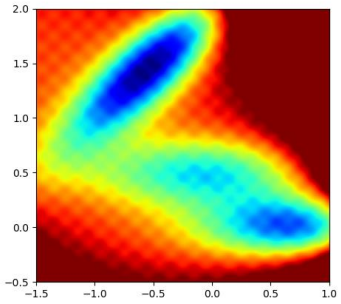
Standard path collective variable



Data-driven path collective variable

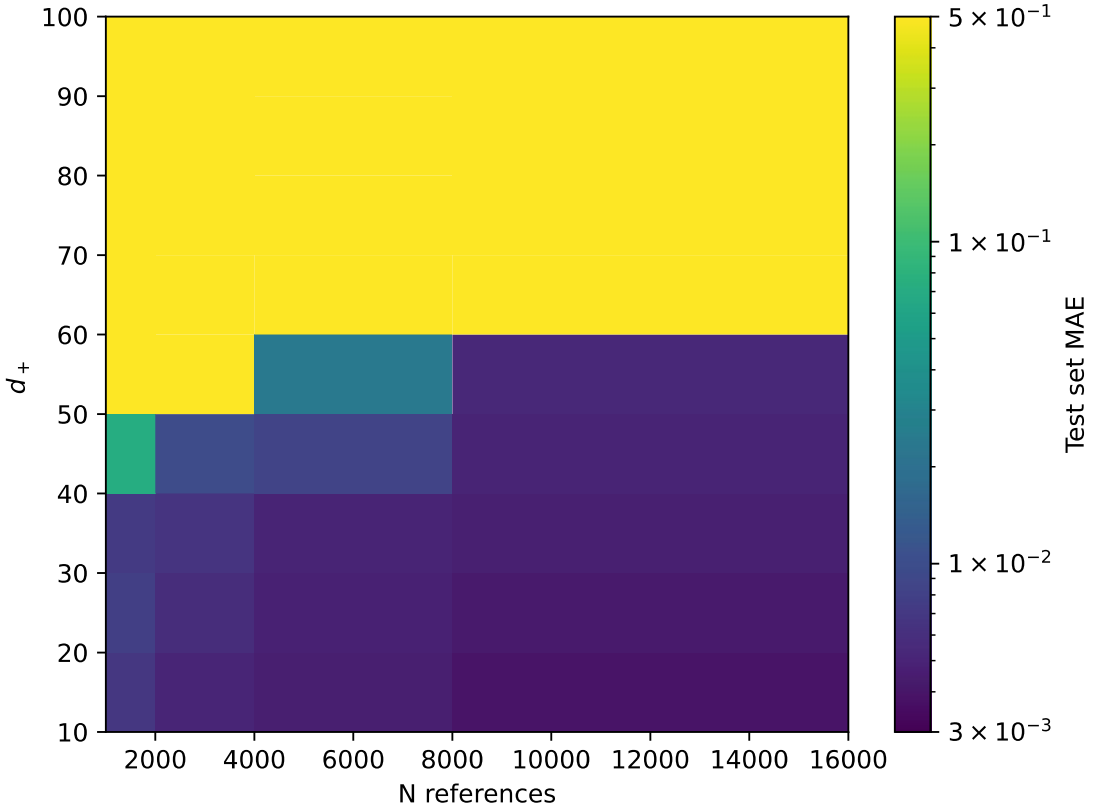


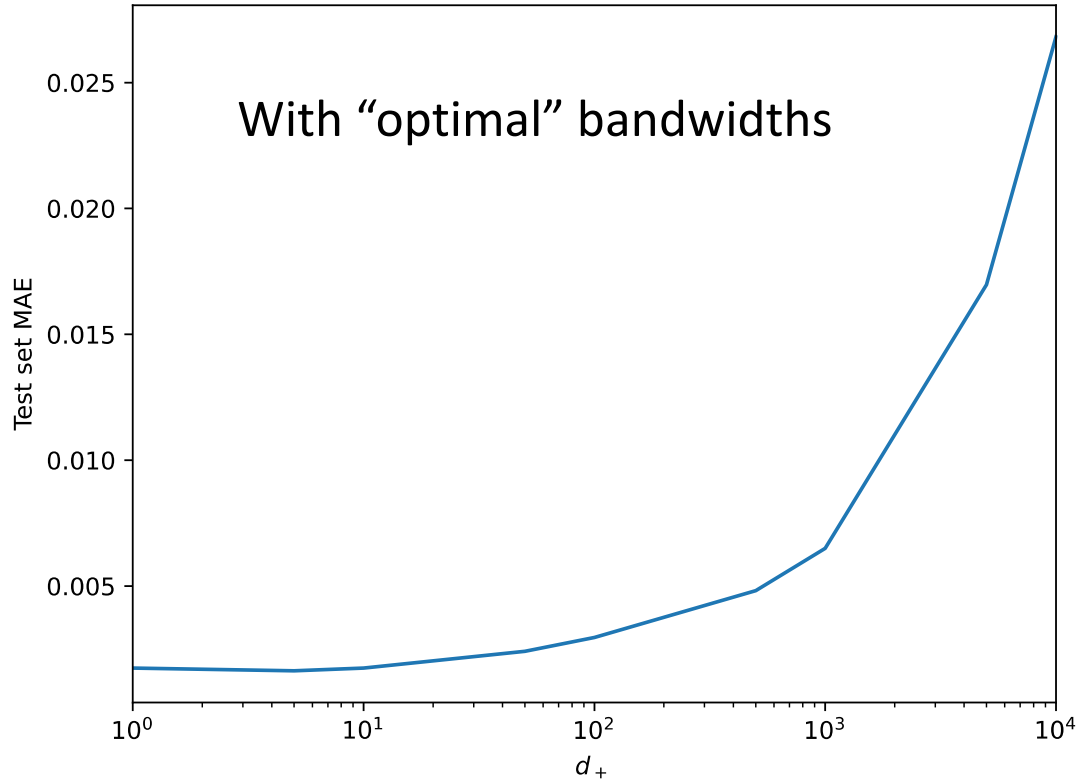
A toy model: the rugged Müller-Brown potential with parasitic dimensions



$$U(x, y, \mathbf{z}) = U_{rmb}(x, y) + \sum_{i=1}^{d_+} z_i^2$$

We add d_+ dimensions uncorrelated to the committor

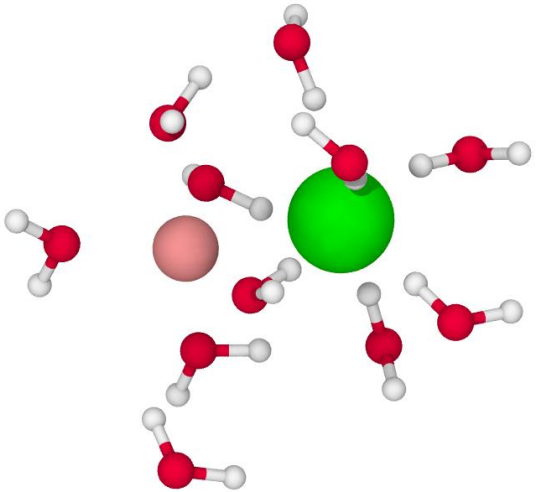




It’s not a regression problem, but an optimization problem!

Technical details

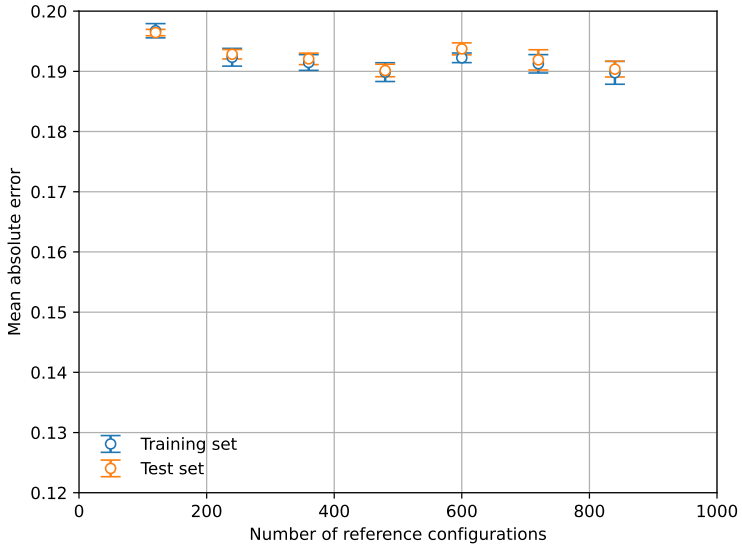
- 1 Li^+, 1 F^-, 160 water molecules
- Joung and Cheatham + SPC/E empirical potentials
- LAMMPS molecular dynamics program
- PLUMED program for committor runs
- T = 300 K, P = 1 atm
- Datasets are uniform in p_B, roughly 1000 configurations per set



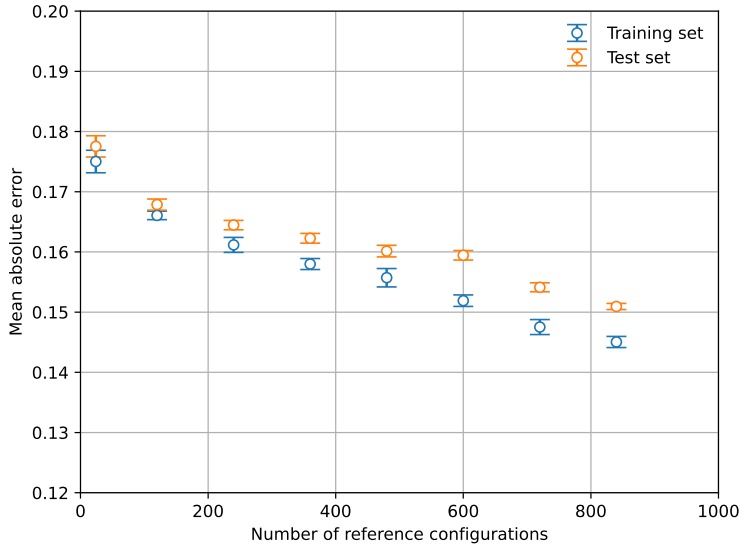
The associated LiF pair in water and its first solvation shell

Results: comparing collective variables

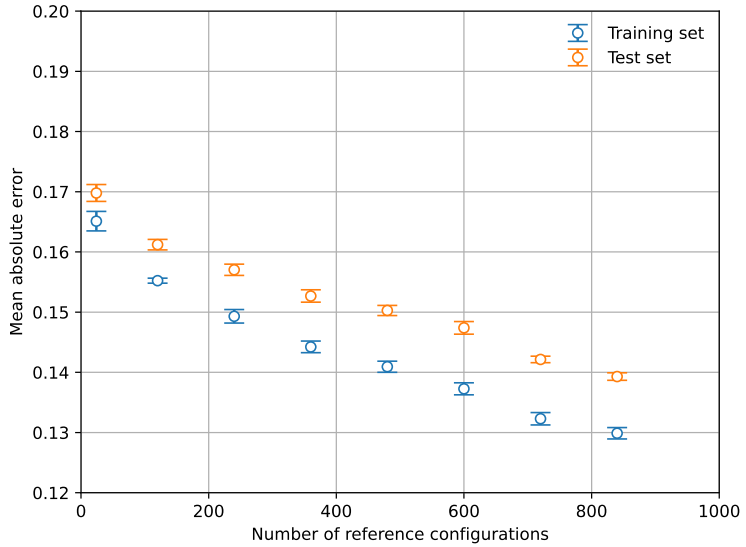
Interionic distance



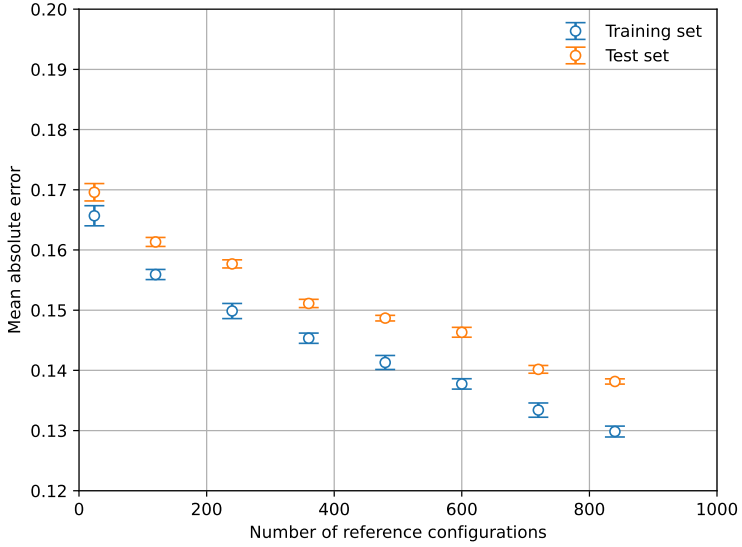
List of 8 scalars (human-made CV)



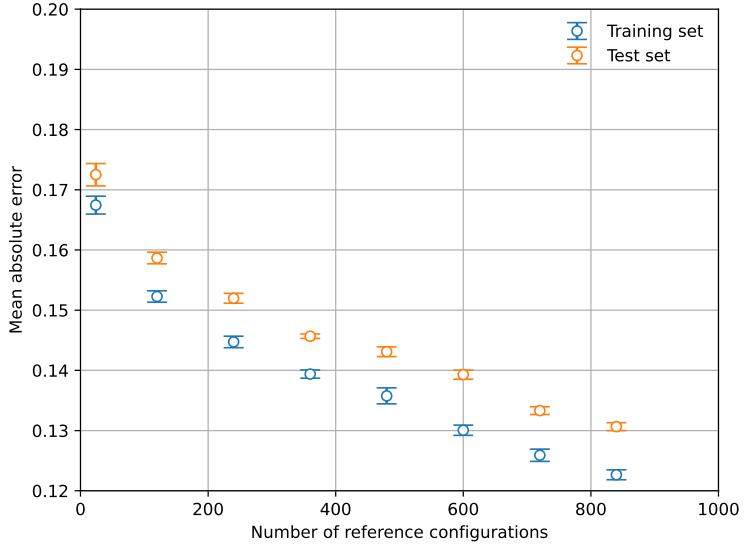
Local PIV

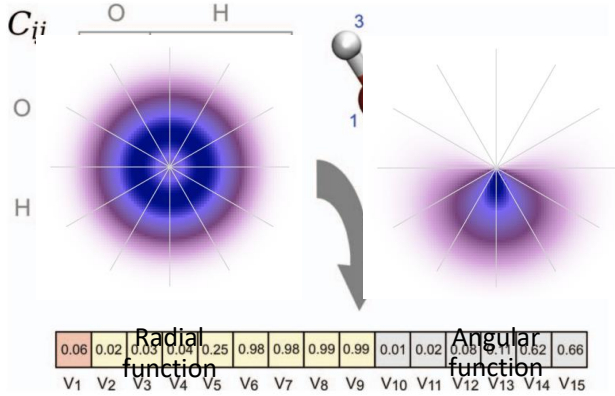


Standard ACSFs



Polynomial ACSFs for organic matter



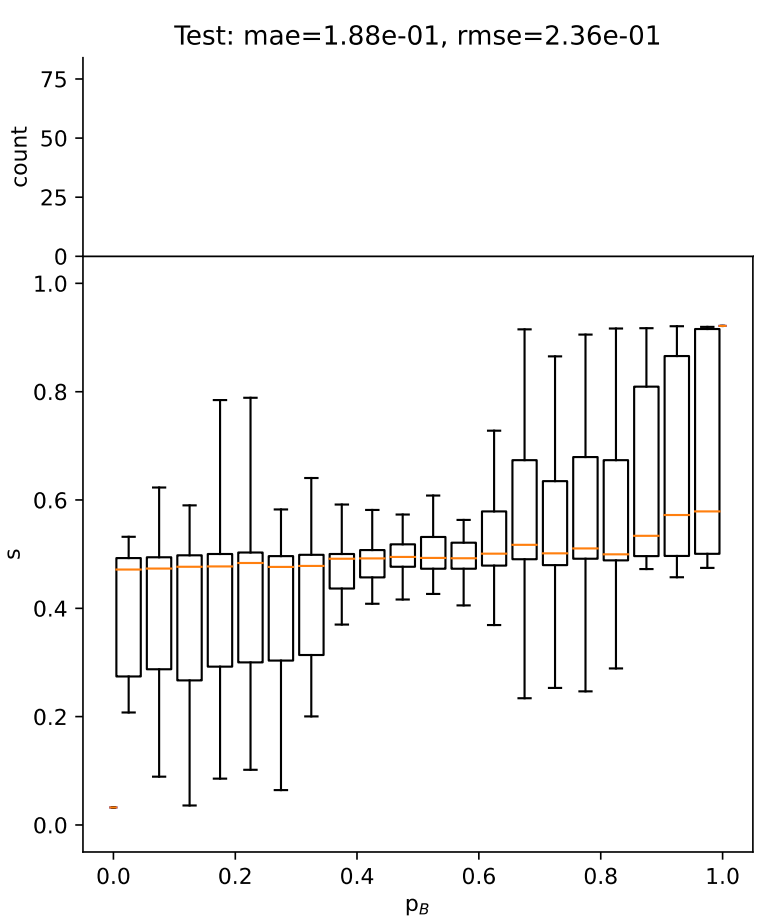


Behler and Parrinello, 2007

Gaussian Moment Tensor, 2016

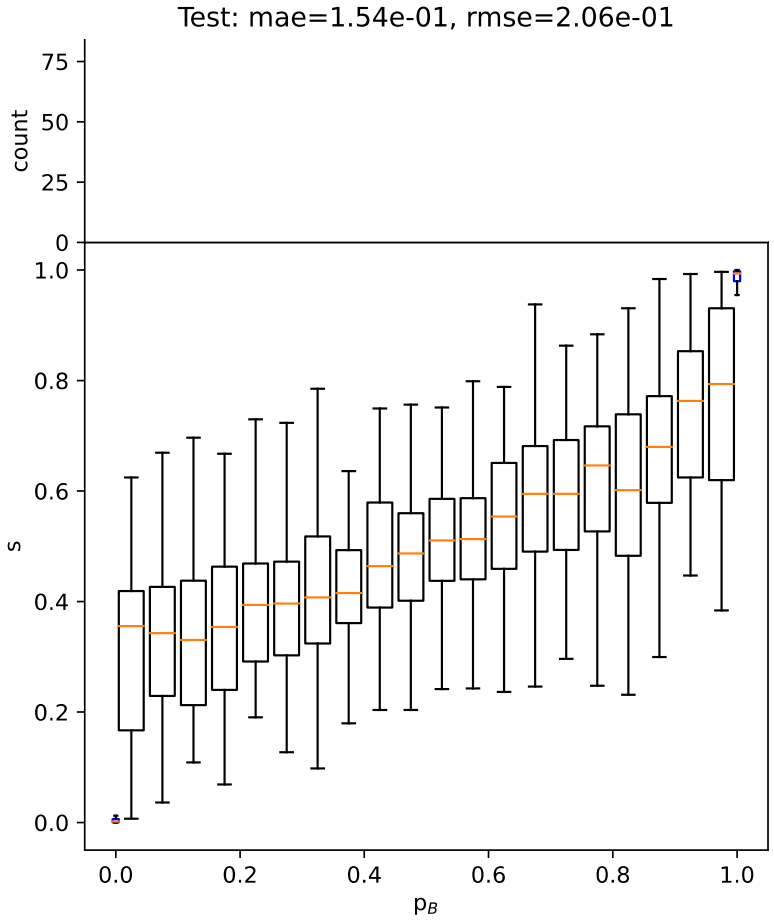
Bircher, Singraber, Dellago, ML:ST 2021

Results: regression plots



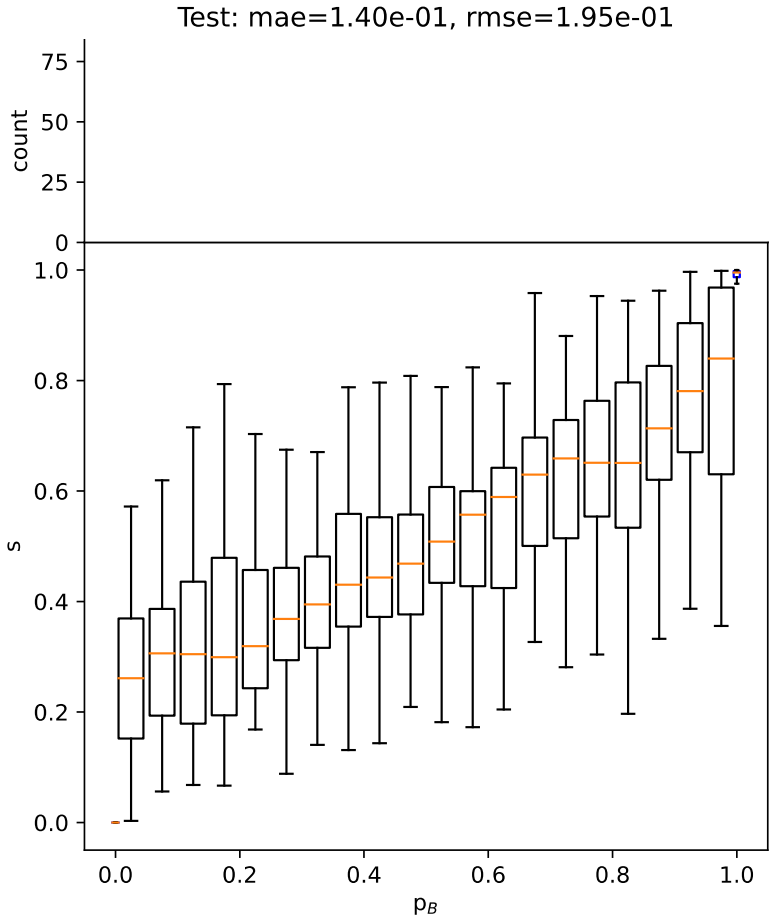
r_{ions}

d=1



r_{ions}, $N_O(Li)$, $N_H(Li)$, $N_O(F)$, $N_H(F)$, N_B, $\psi_s(Li)$, $\psi_s(F)$

d=8

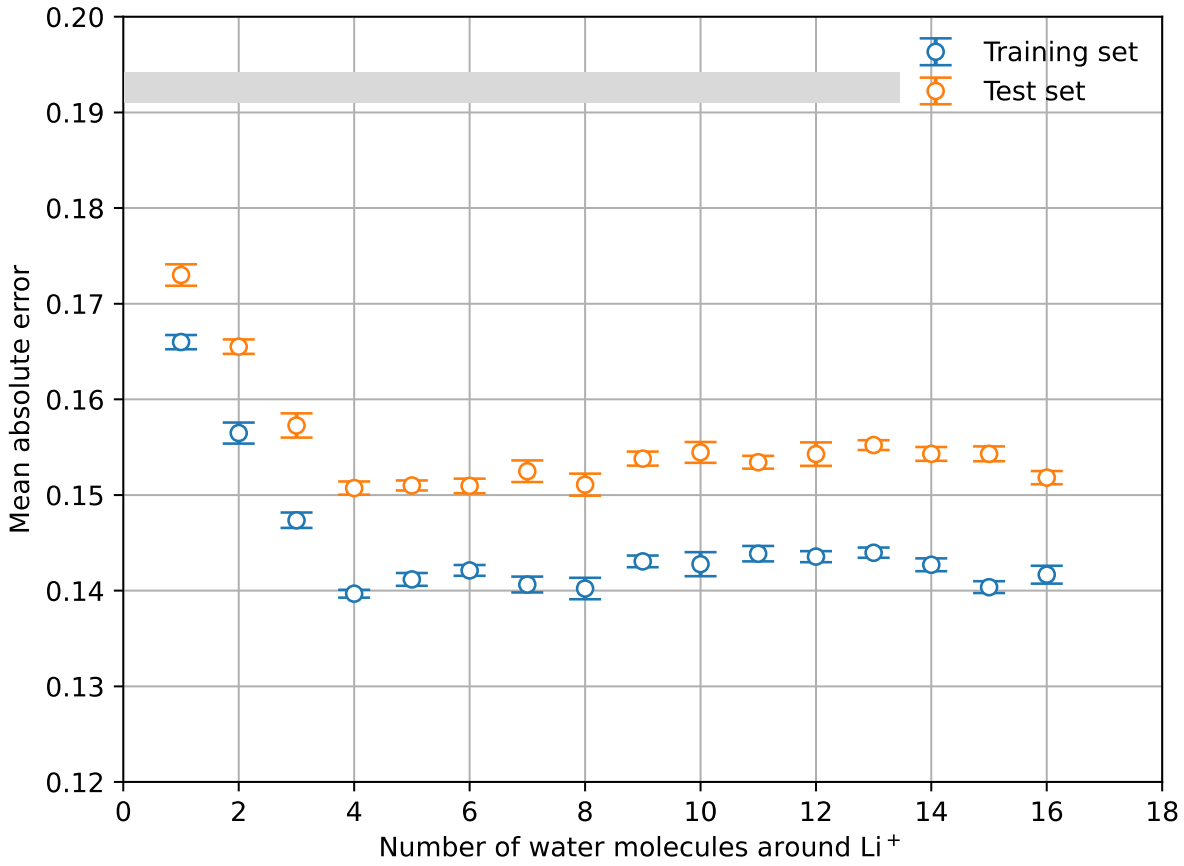


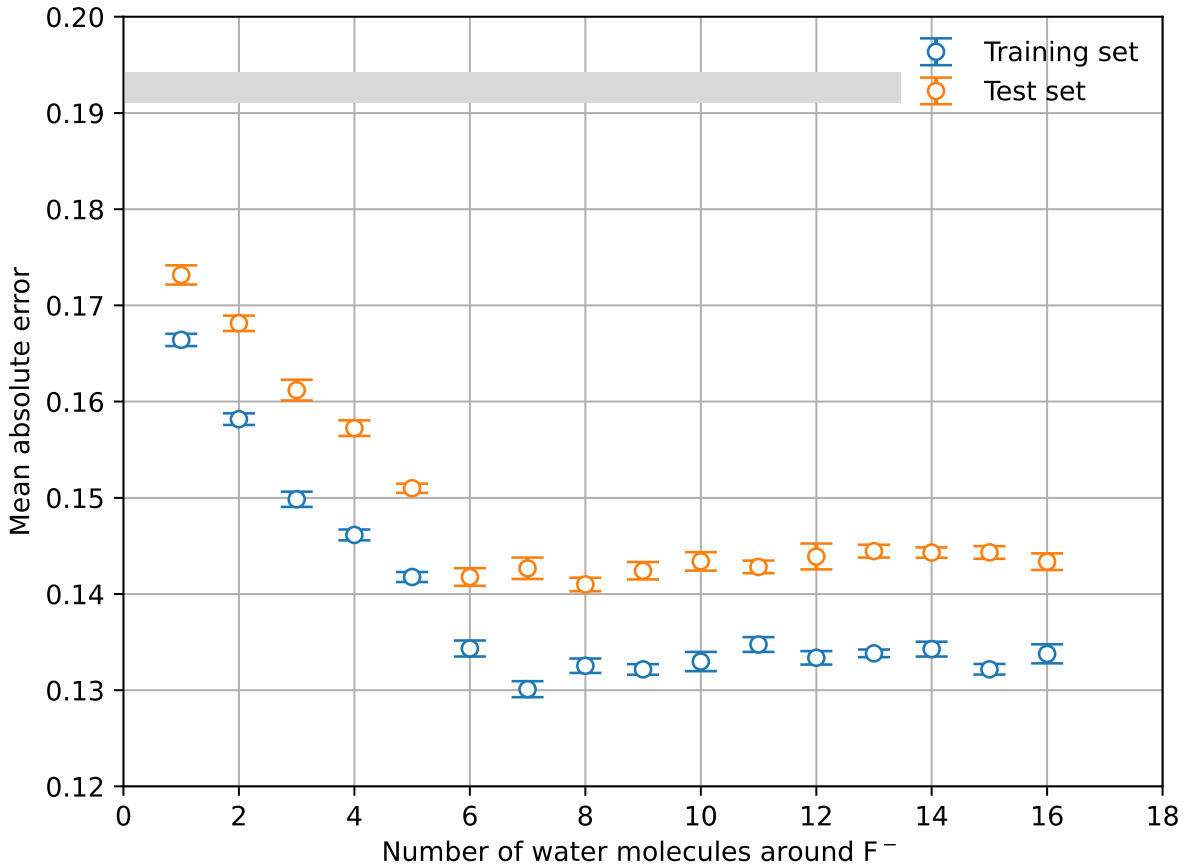
Local PIV

d=42

Results: extracting information further away from the ions

Strategy: construct a PIV including more and more water molecules according to their distance to one of the ions





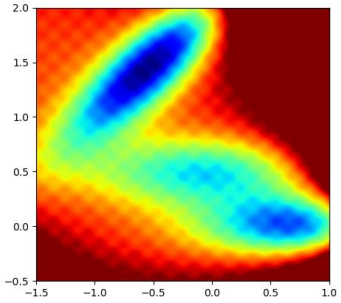
Conclusions

- We introduced a methodology to quantitatively compare collective variables
- Based on a KRR of the committor probability
- One-dimensional, differentiable CV
- For LiF in water, better CVs with solvent information, but challenging to extract something beyond the first solvation shell
- Complex representations perform better than human-made CVs, but there's room for improvement (compactness, optimization, under/over-completeness)

Ongoing work

- Nucleation/crystallization/precipitation applications
- Symmetry/locality approximation to the committor to reduce the dimensionality (akin to the local energy approximation of ML potentials)

A toy model: the rugged Müller-Brown potential embedded in a 5d space



$$x_1 = (x + 0.1y)^2$$

$$x_2 = y - 2x + 3$$

$$x_3 = \sqrt{4|xy|}$$

$$x_4 = x^3 - y^2$$

$$x_5 = xy^4$$

The MAE is roughly the same as for the non-transformed potential

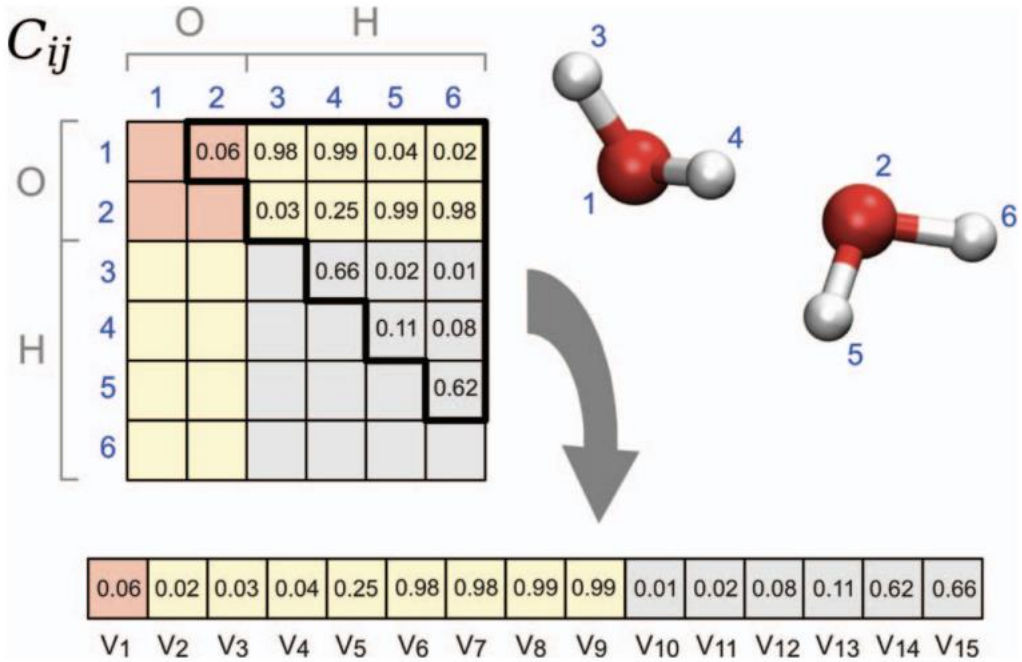
Optimization process: loss function and data splitting

Committor on *S* dataset	Reference (the actual refs.)
Committor on *J* dataset	Training (loss function)
Basins *A* and *B* datasets	Test (model performance)

- Data is binned according to p_B
- Uniform distribution over p_B
- Datasets of different sizes: from 1 to 35 config/bin/ensemble

Loss function to minimize during optimization: either MAE or RMSE on training set

The Permutation Invariant Vector



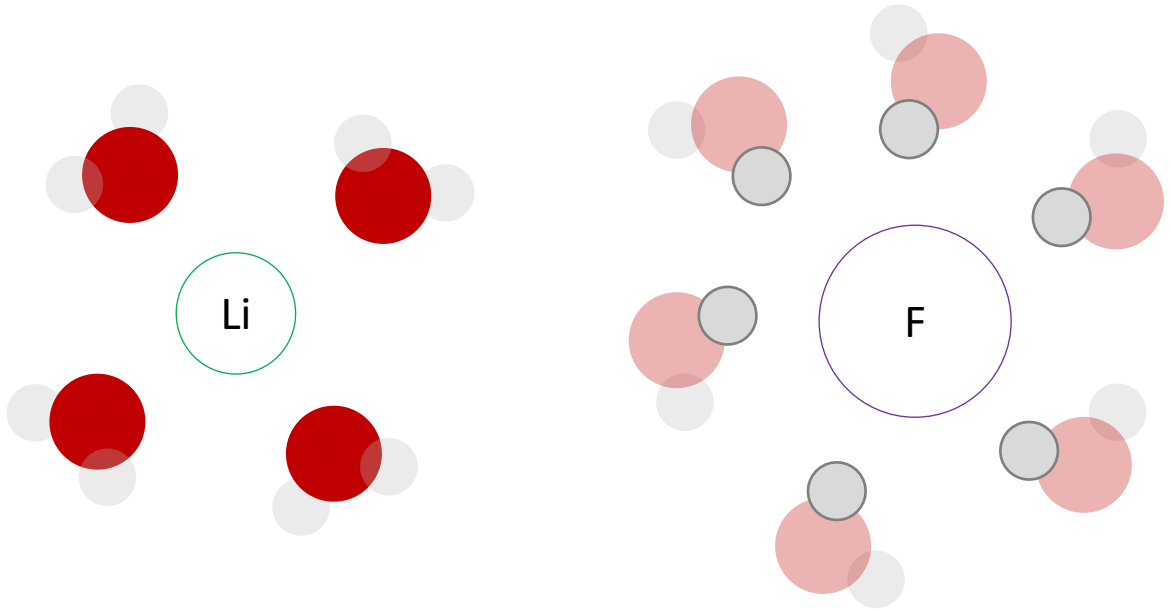
- A high-dimensional CV built on pairwise distances
- If all atoms are included, the representation is (over)complete
- The PIV is invariant by permutation of atoms of the same elements, and by rigid translation/rotations of the whole system

Gallet and Pietrucci, JCP 2013

Pipolo, Salanne, Ferlat, Klotz, Saitta, Pietrucci, PRL 2017

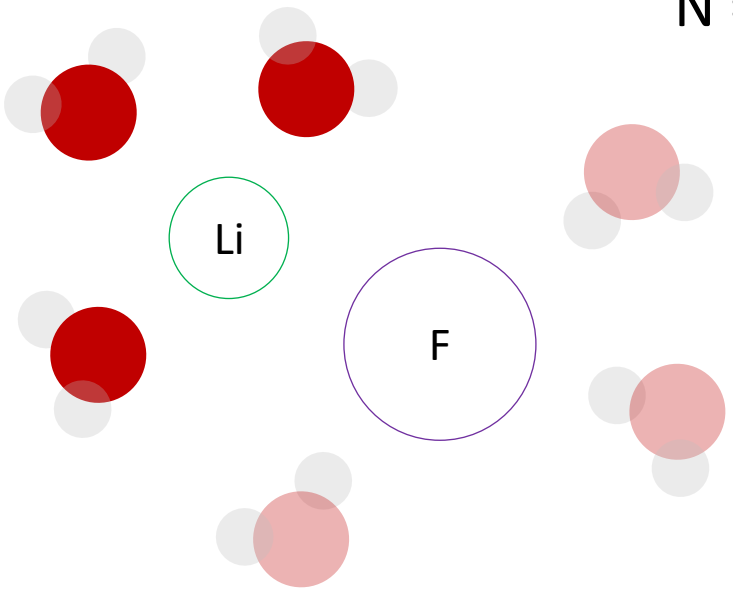
The Permutation Invariant Vector: local variants

- We investigate local variants of the full PIV, centered on the ions
- Highly reduced dimensionality



Subsystem of Li, F, 4 closest O's from Li, 6 closest H's from F

N = 3



Subsystem of Li, F, and the N closest O's from Li