Using machine learning to map the structure and predict the properties of materials and molecules

Michele Ceriotti
http://cosmo.epfl.ch

IPAM - 27.09.2016

COSMO
COMPUTATIONAL SCIENCE AND MODELLING

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Acknowledgements

Sandip De
Piero Gasparotto
Mariana Rossi

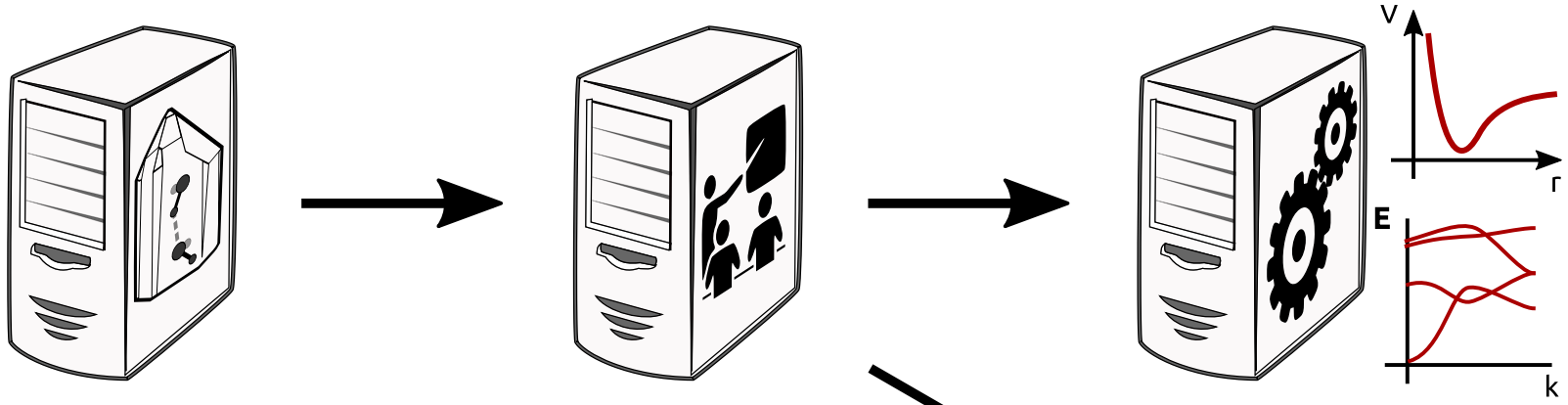
Robert Meßner
Felix Musil
Andrea Anelli

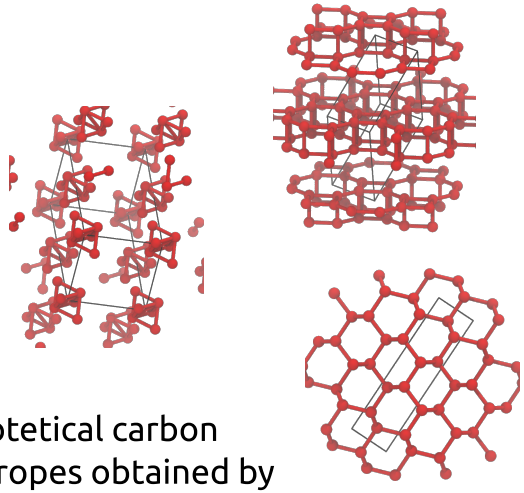
erc
European Commission
CCMX Competence Centre for Materials Science and Technology
MARVEL
NATIONAL CENTRE OF COMPETENCE IN RESEARCH
FNSNF

G. Csányi, A. Bartók, C. Baldauf, C. Pickard, G. Day, S. Goedecker

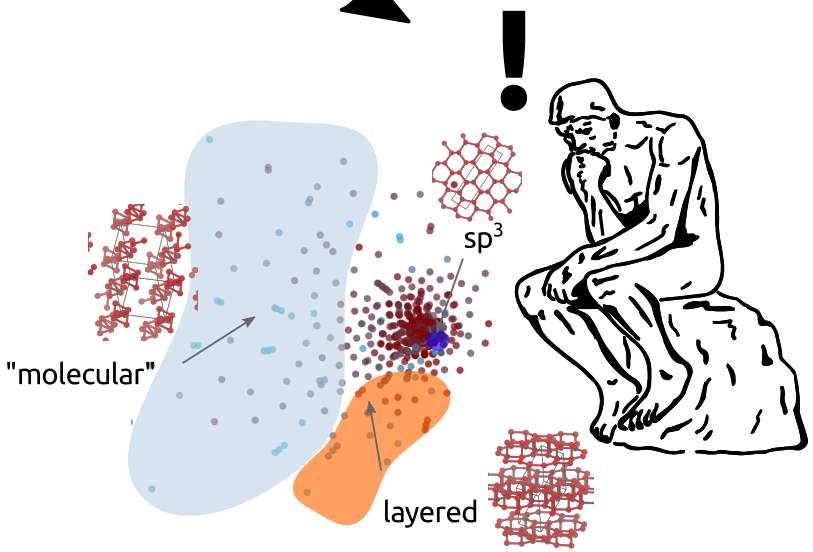
“Machine learning” and intuitive understanding

- Machine learning can be used to predict materials’ properties *or* to distill large amounts of complex data in a human-understandable form



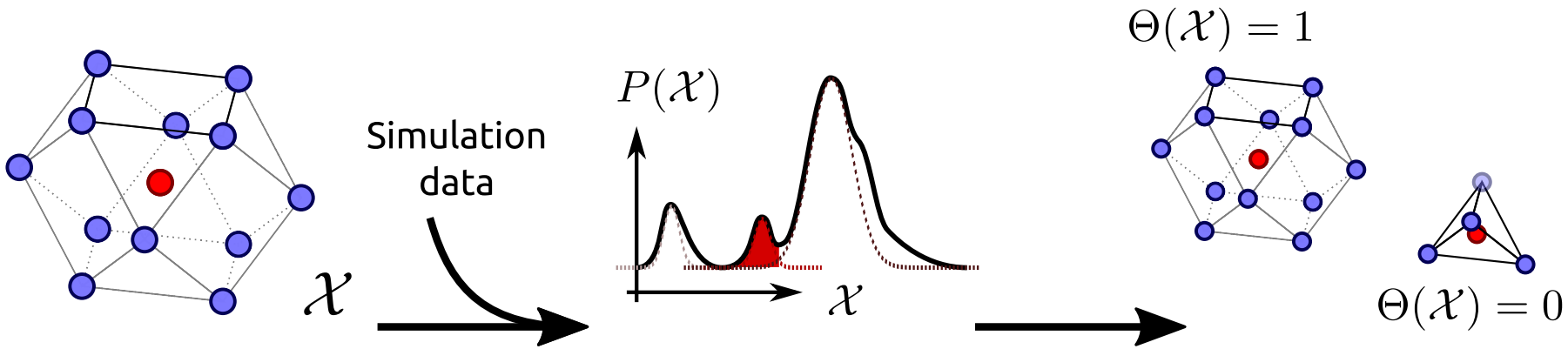


Hypotetical carbon allotropes obtained by Ab Initio Random Structure Search *(structures from C. Pickard)*



Representing patterns and mapping structures

- Understanding emergent structural complexity by analyzing simulations
  - Recognizing the building-blocks at the atomic scale
  - An effective and flexible framework for comparing structures
  - Mapping structural complexity. Key to make the best out of big data



Abstract description of chemical environments

Probabilistic analysis of structural correlations

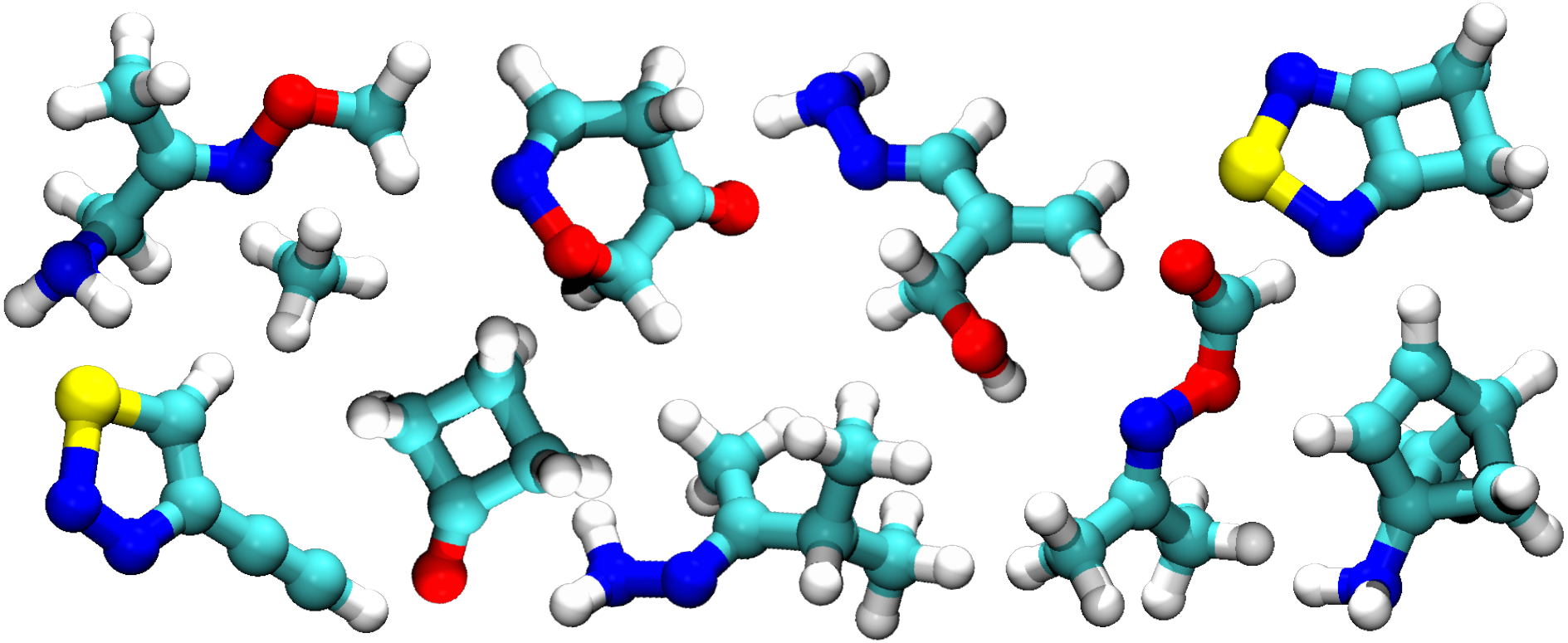
Posterior probability as a structural fingerprint

$$\Theta(\mathcal{X}) = \frac{p_0 g_0(\mathcal{X})}{\sum_k p_k g_k(\mathcal{X})}$$

Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Representing patterns and mapping structures

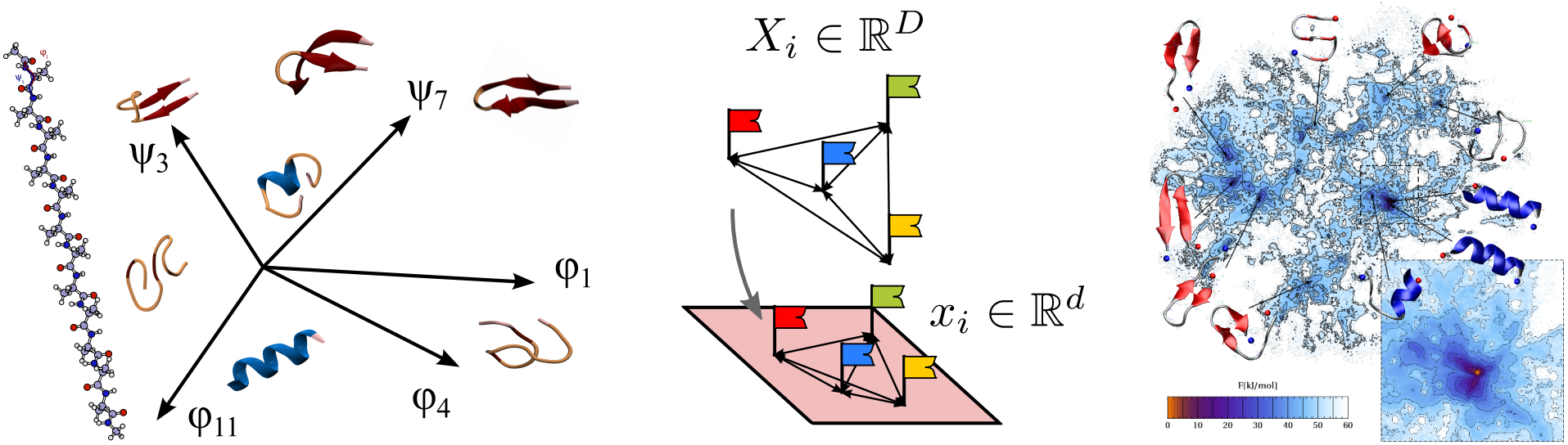
- Understanding emergent structural complexity by analyzing simulations
  - Recognizing the building-blocks at the atomic scale
  - An effective and flexible framework for comparing structures
  - Mapping structural complexity. Key to make the best out of big data



De, Bartók, Csányi, Ceriotti, PCCP (2016)

Representing patterns and mapping structures

- Understanding emergent structural complexity by analyzing simulations
  - Recognizing the building-blocks at the atomic scale
  - An effective and flexible framework for comparing structures
  - Mapping structural complexity. Key to make the best out of big data



High-dimensional descriptors of a complex system

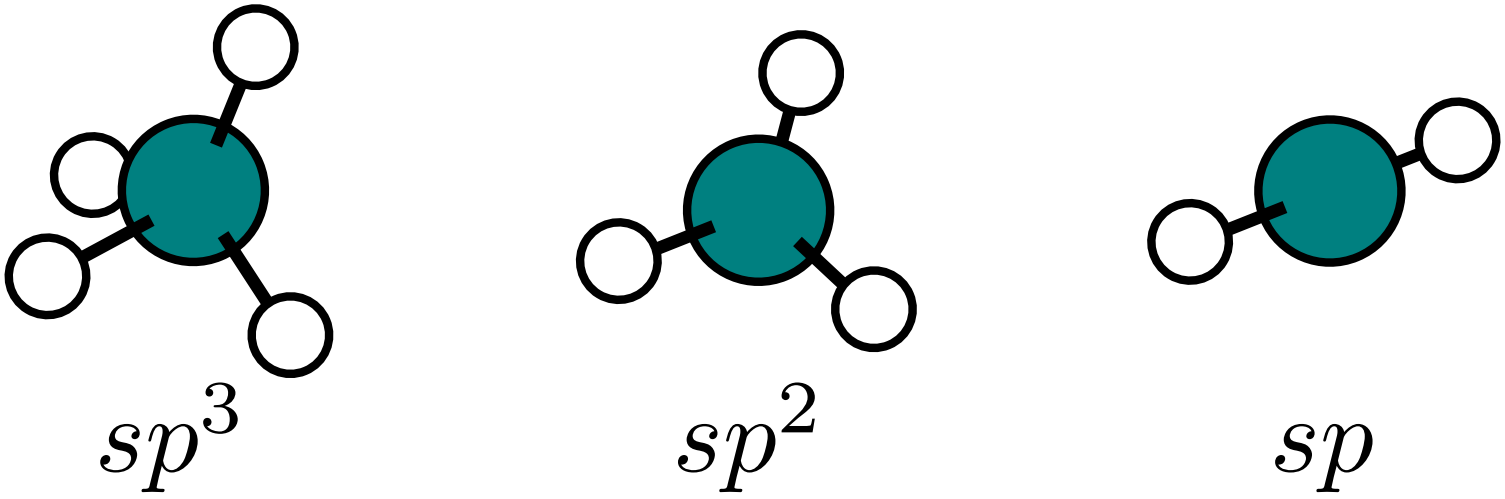
Non-linear dimensionality reduction (**sketch-map**)

Intuitive maps of structural stability

Ceriotti, Tribello, Parrinello, PNAS (2011)

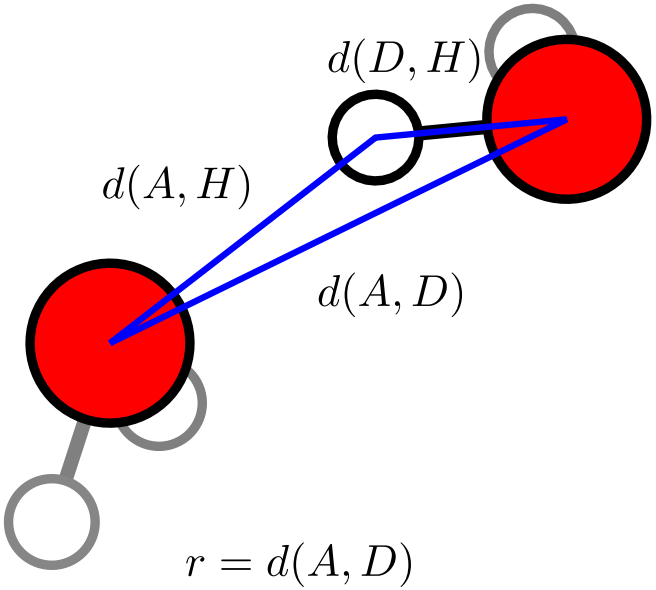
Recognizing molecular patterns

- "Chemical intuition" builds on recognizing recurring patterns in atomic configurations
- Atomistic models provide large amount of data for statistical analysis
- Automatic scheme to single out structural motifs in simulations



An Agnostic Definition of the H-Bond

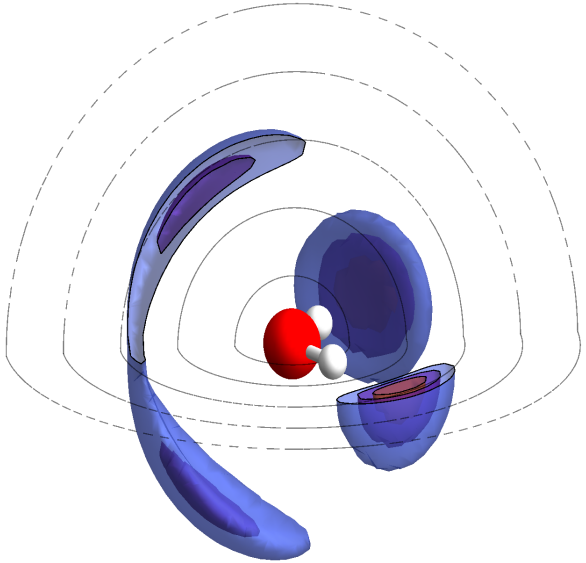
- Most general description of a H-bond geometry: 3 distances. Many heuristic definitions available
- How to recognize automatically what is a H-bond, based only on analysis of a simulation?

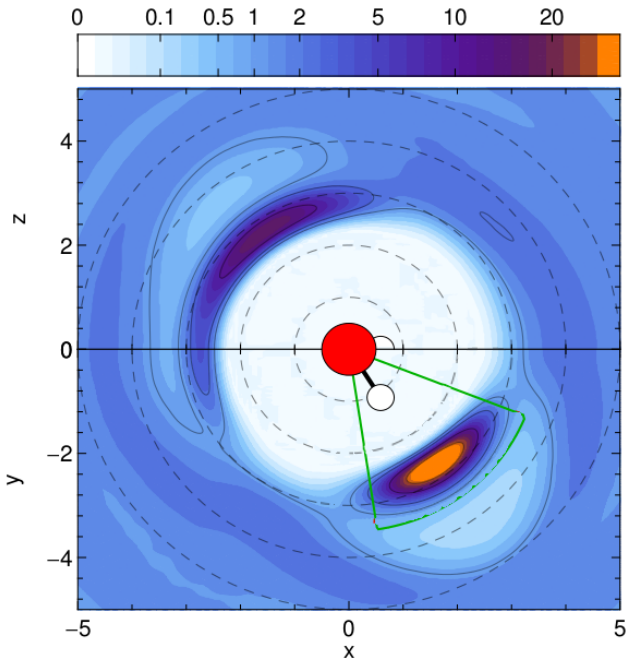


$$r = d(A,D)$$
$$\mu = d(A,H) + d(D,H)$$
$$\nu = d(D,H) - d(A,H)$$

An Agnostic Definition of the H-Bond

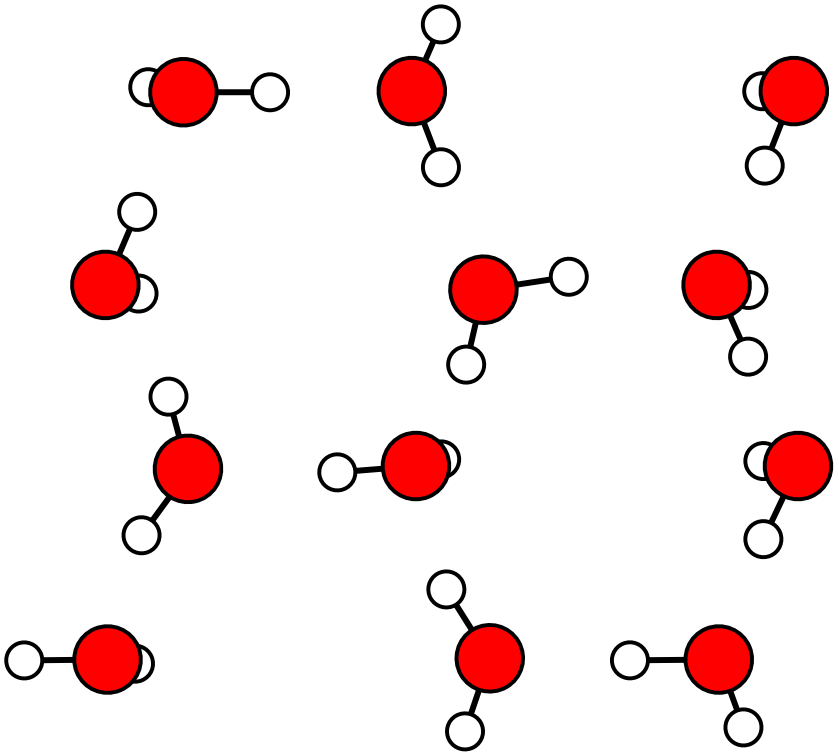
- Most general description of a H-bond geometry: 3 distances. Many heuristic definitions available
- How to recognize automatically what is a H-bond, based only on analysis of a simulation?





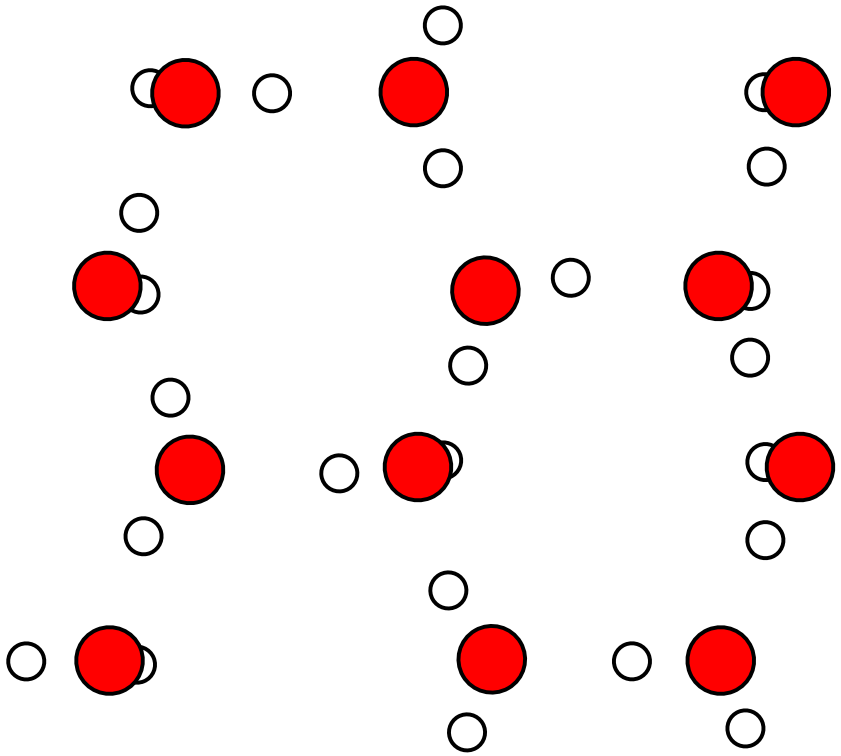
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances. Many heuristic definitions available
- How to recognize automatically what is a H-bond, based only on analysis of a simulation?



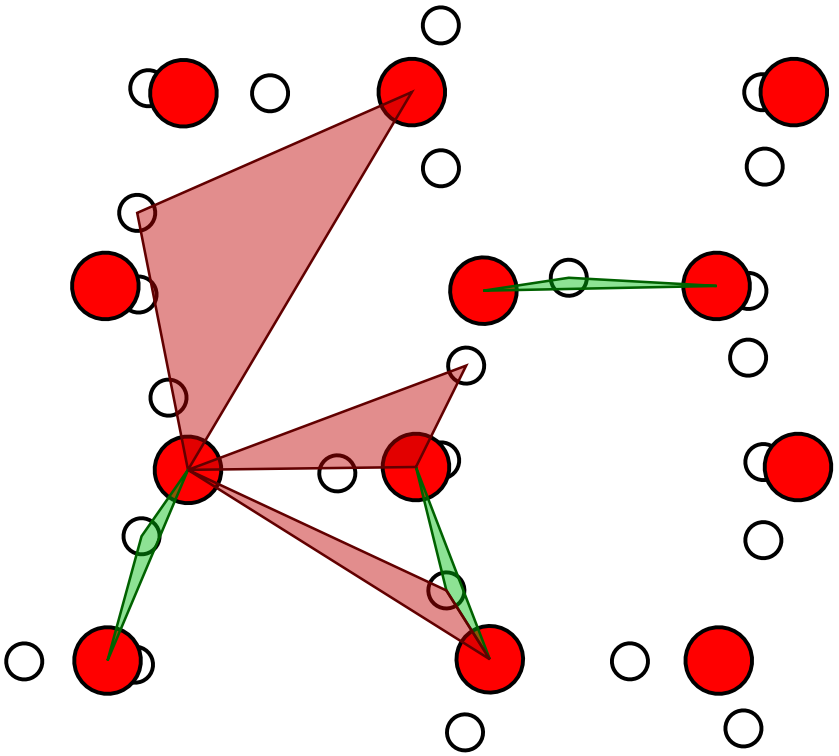
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances. Many heuristic definitions available
- How to recognize automatically what is a H-bond, based only on analysis of a simulation?



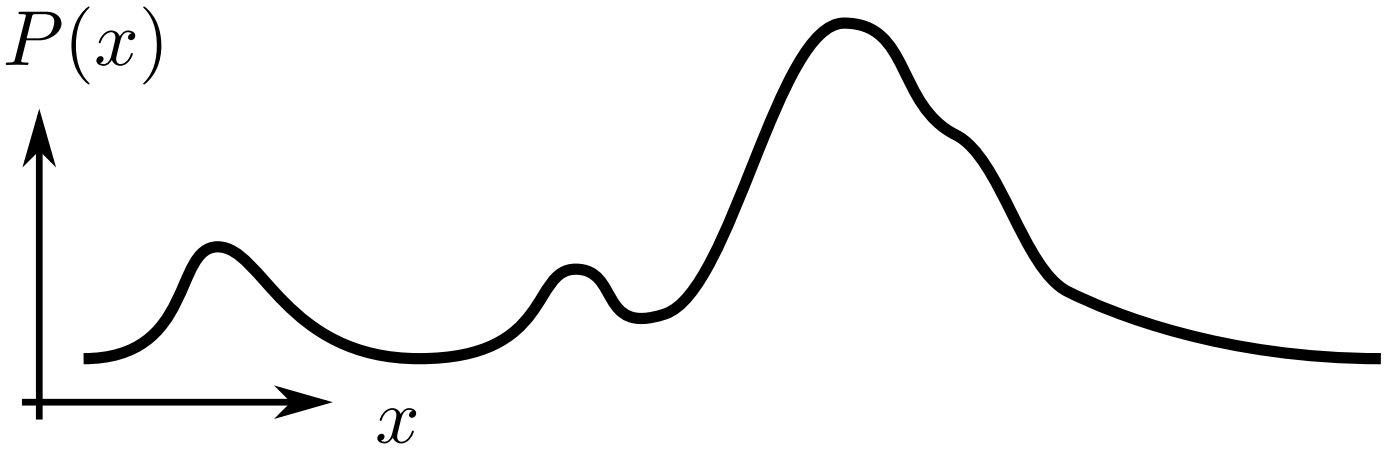
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances. Many heuristic definitions available
- How to recognize automatically what is a H-bond, based only on analysis of a simulation?



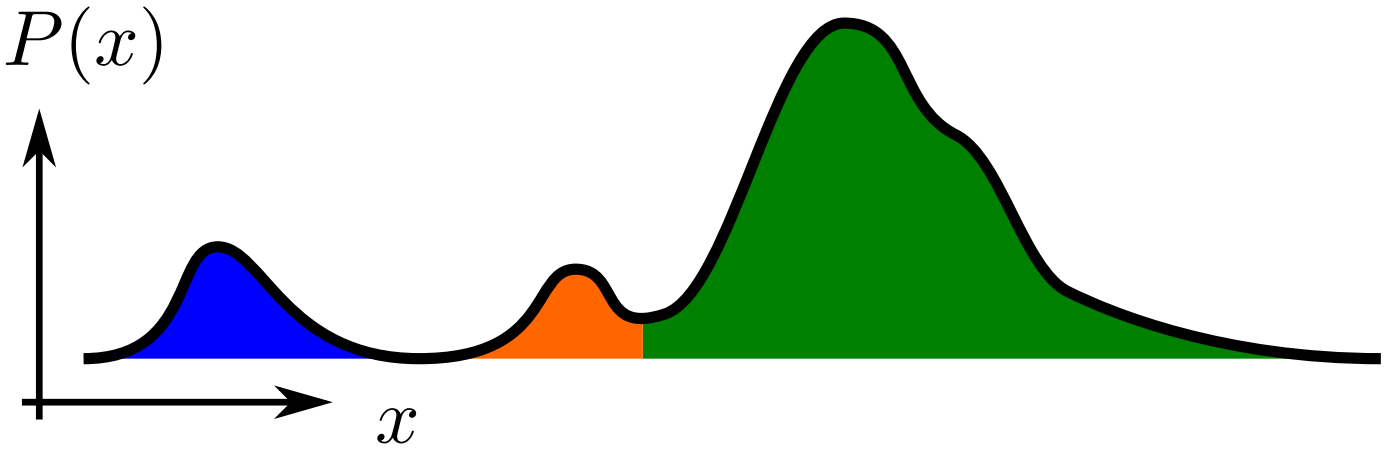
Mode Analysis of a Distribution

- A natural way of recognizing patterns in a distribution is to identify its modes, and the basin of attraction of each mode.
  - One can then fit a simple Gaussian model (with fixed centers), and use posteriors to assign fingerprints to each cluster



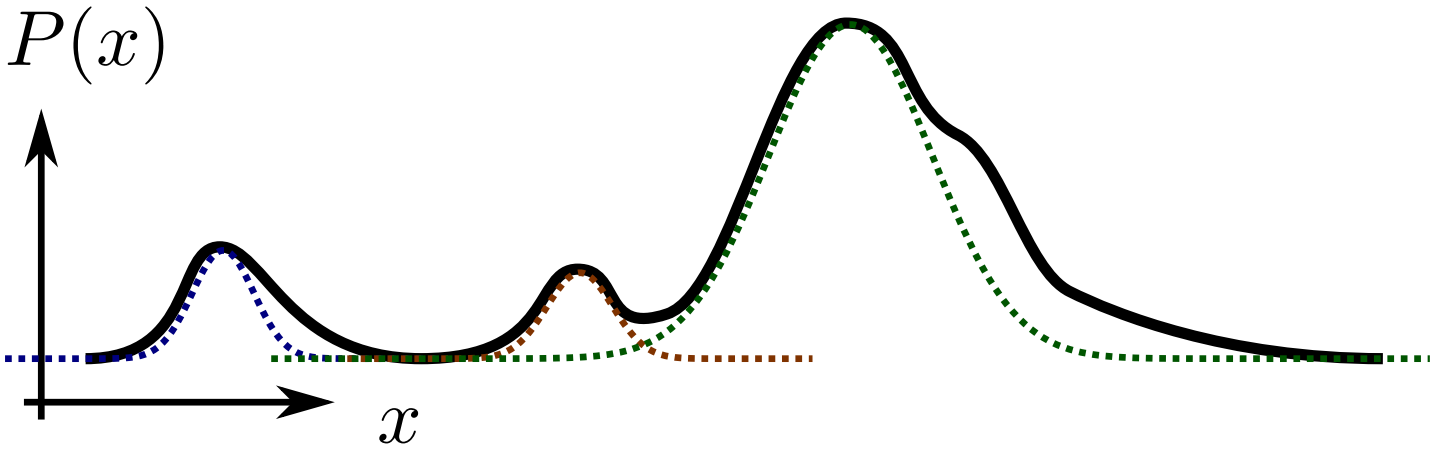
Mode Analysis of a Distribution

- A natural way of recognizing patterns in a distribution is to identify its modes, and the basin of attraction of each mode.
  - One can then fit a simple Gaussian model (with fixed centers), and use posteriors to assign fingerprints to each cluster



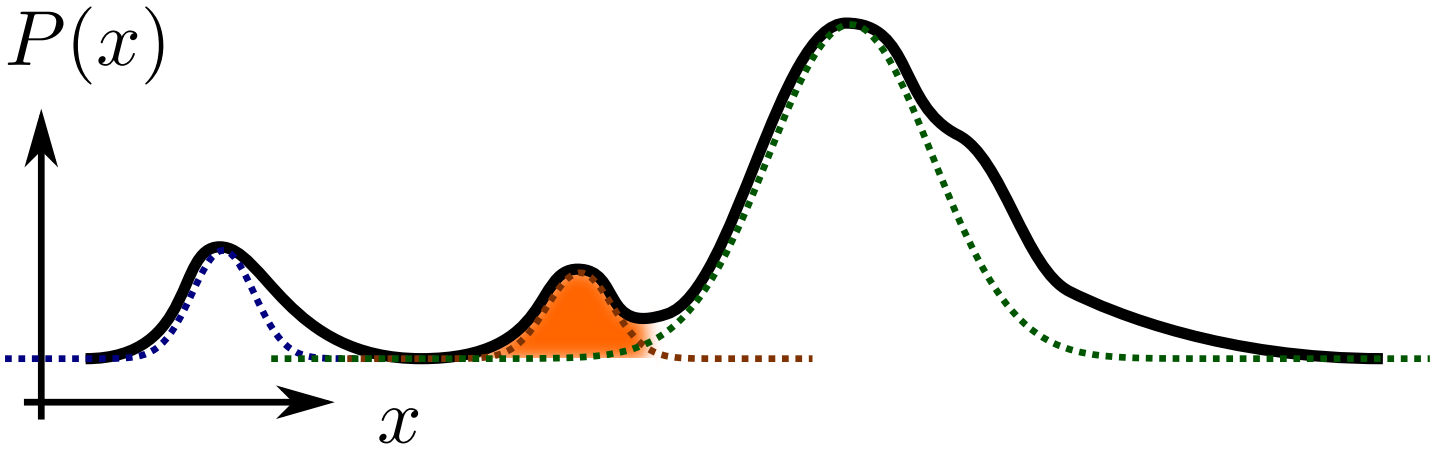
Mode Analysis of a Distribution

- A natural way of recognizing patterns in a distribution is to identify its modes, and the basin of attraction of each mode.
  - One can then fit a simple Gaussian model (with fixed centers), and use posteriors to assign fingerprints to each cluster



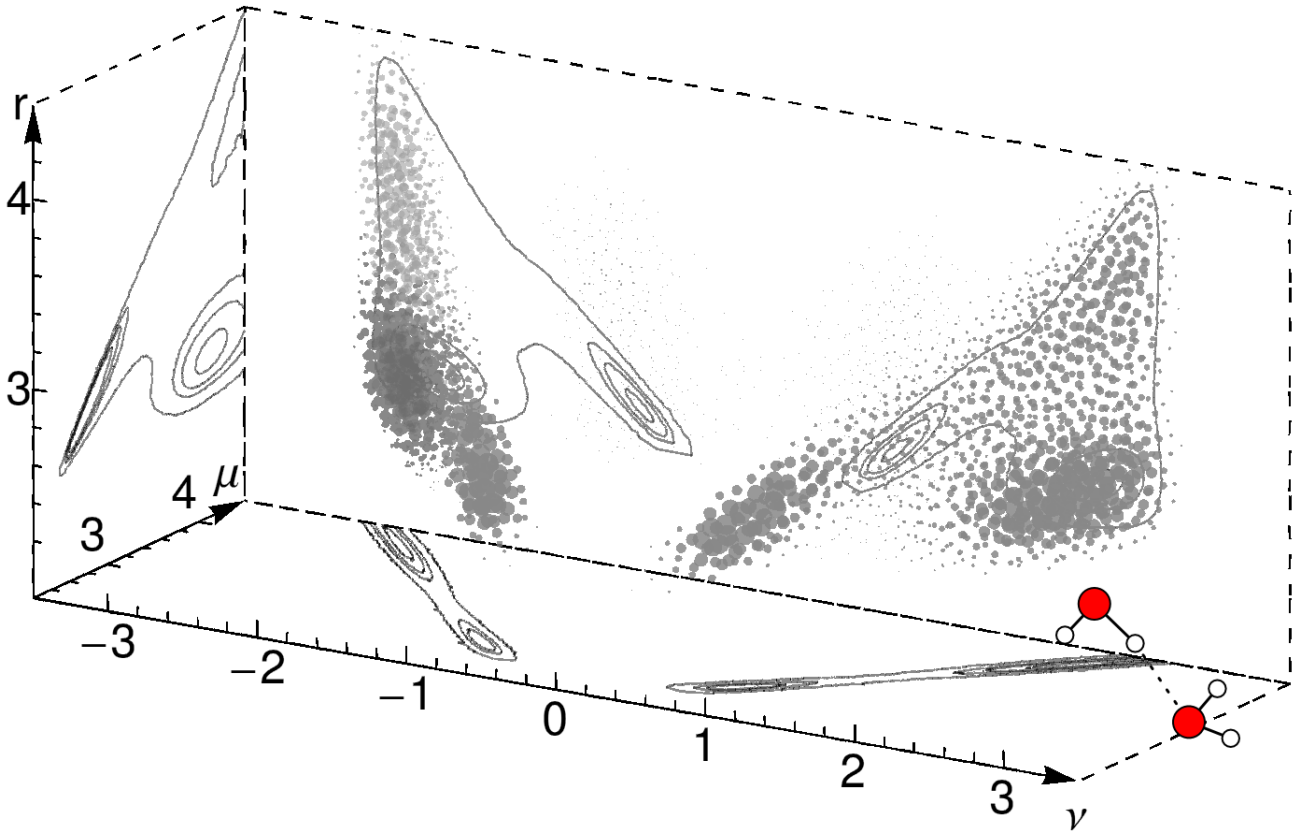
Mode Analysis of a Distribution

- A natural way of recognizing patterns in a distribution is to identify its modes, and the basin of attraction of each mode.
  - One can then fit a simple Gaussian model (with fixed centers), and use posteriors to assign fingerprints to each cluster



An Adaptive, Data-Driven Definition

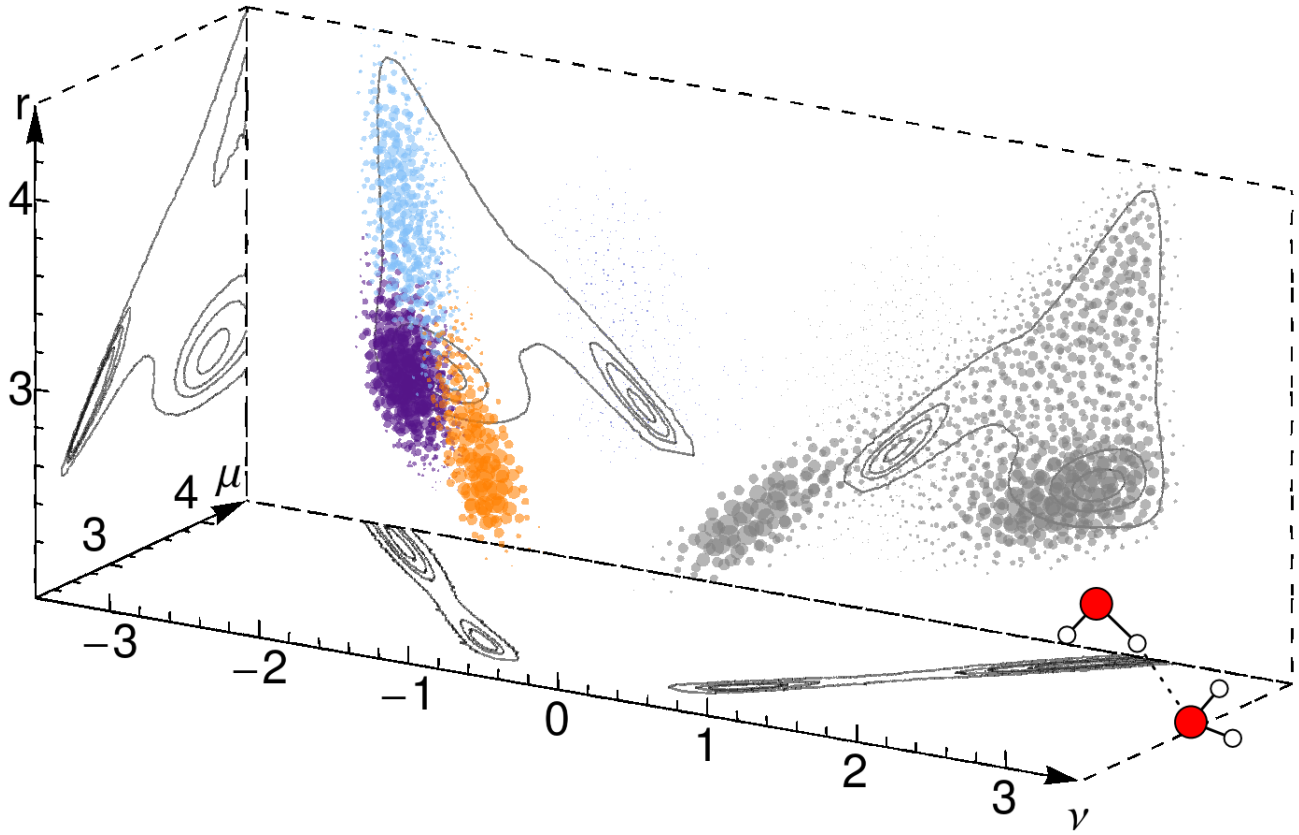
- Ab initio water PAMM recognizes multiple modes - one corresponds to the H-bond
- Adaptive, context-dependent definition - details of the model, thermodynamic conditions, type of HB



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

An Adaptive, Data-Driven Definition

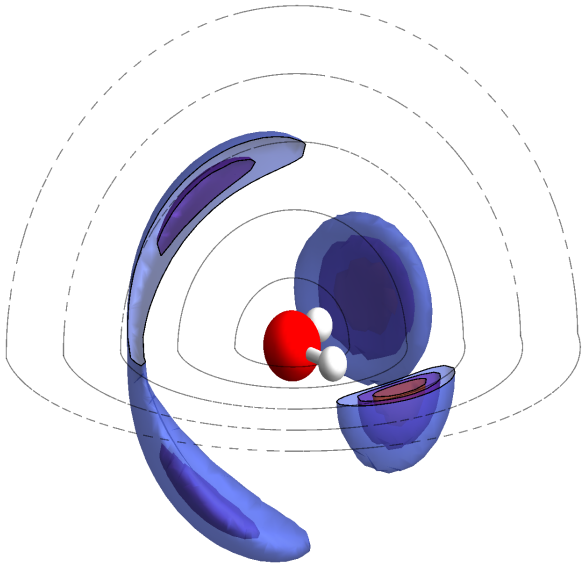
- Ab initio water PAMM recognizes multiple modes - one corresponds to the H-bond
- Adaptive, context-dependent definition - details of the model, thermodynamic conditions, type of HB

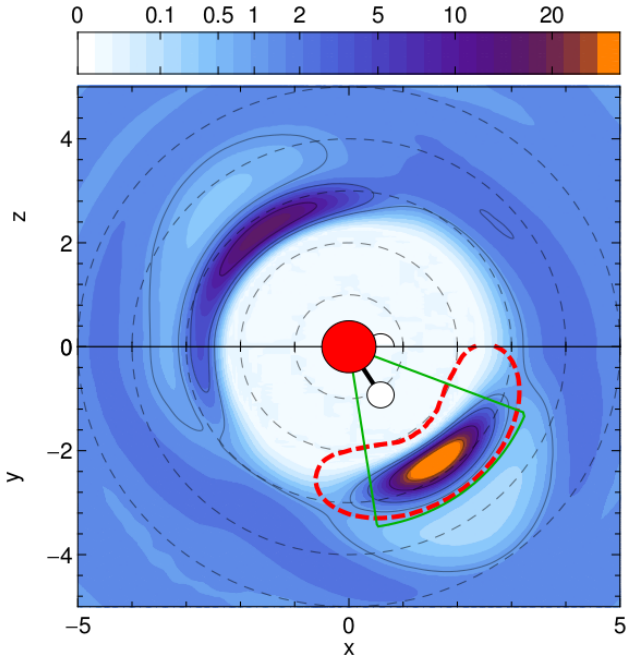


Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

An Adaptive, Data-Driven Definition

- Ab initio water PAMM recognizes multiple modes - one corresponds to the H-bond
- Adaptive, context-dependent definition - details of the model, thermodynamic conditions, type of HB

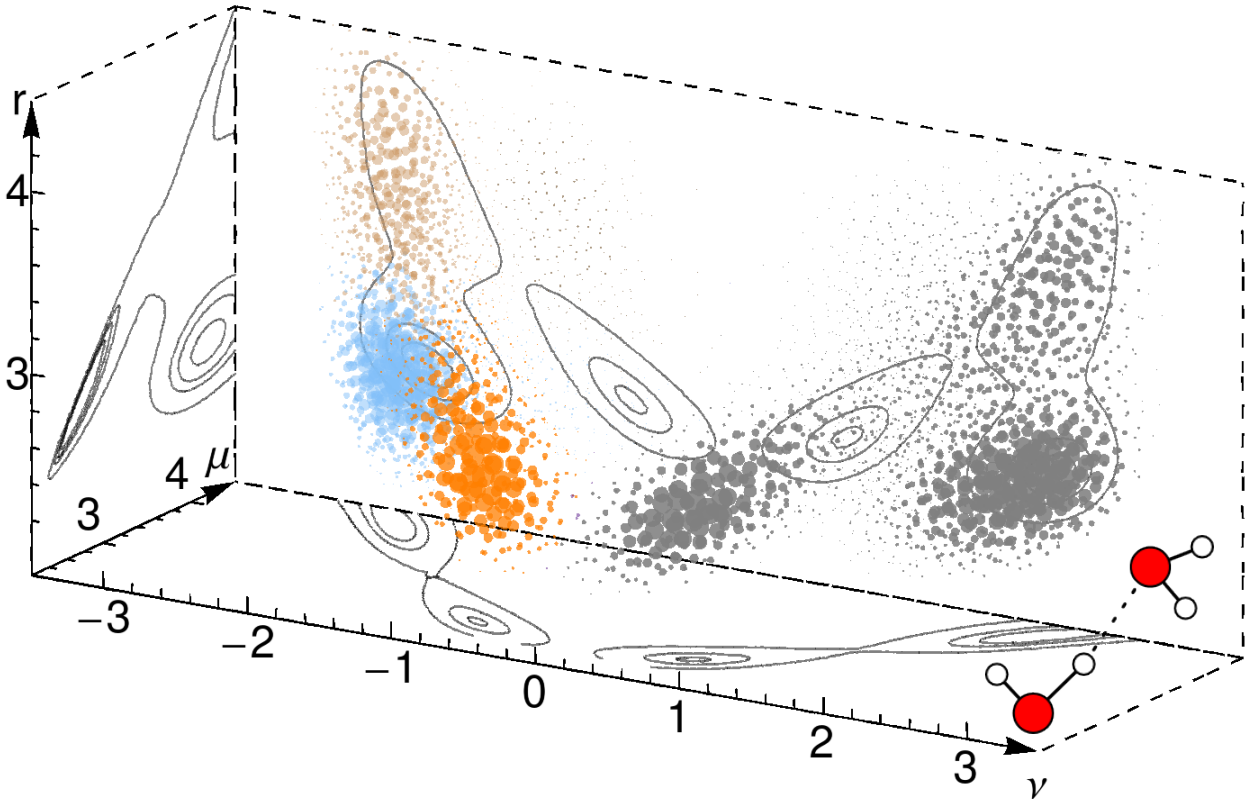




Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

An Adaptive, Data-Driven Definition

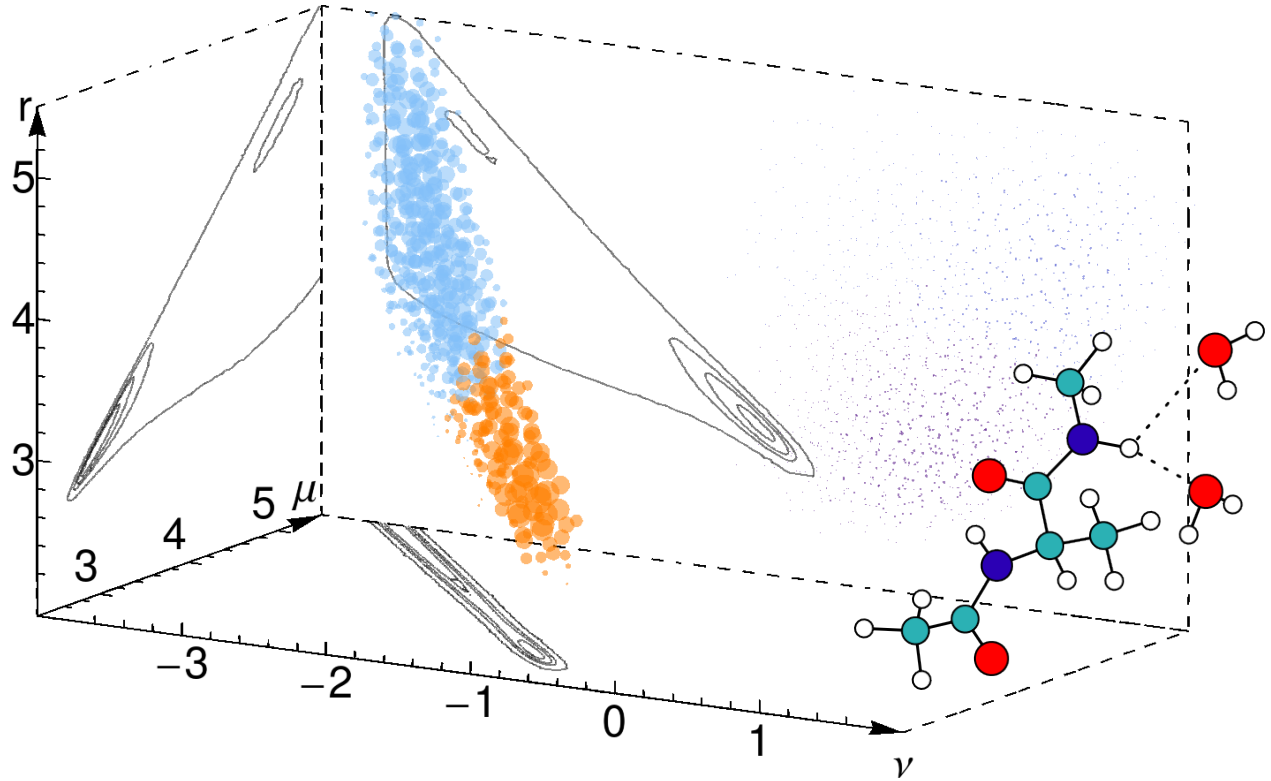
- Ab initio water PAMM recognizes multiple modes - one corresponds to the H-bond
- Adaptive, context-dependent definition - details of the model, thermodynamic conditions, type of HB



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

An Adaptive, Data-Driven Definition

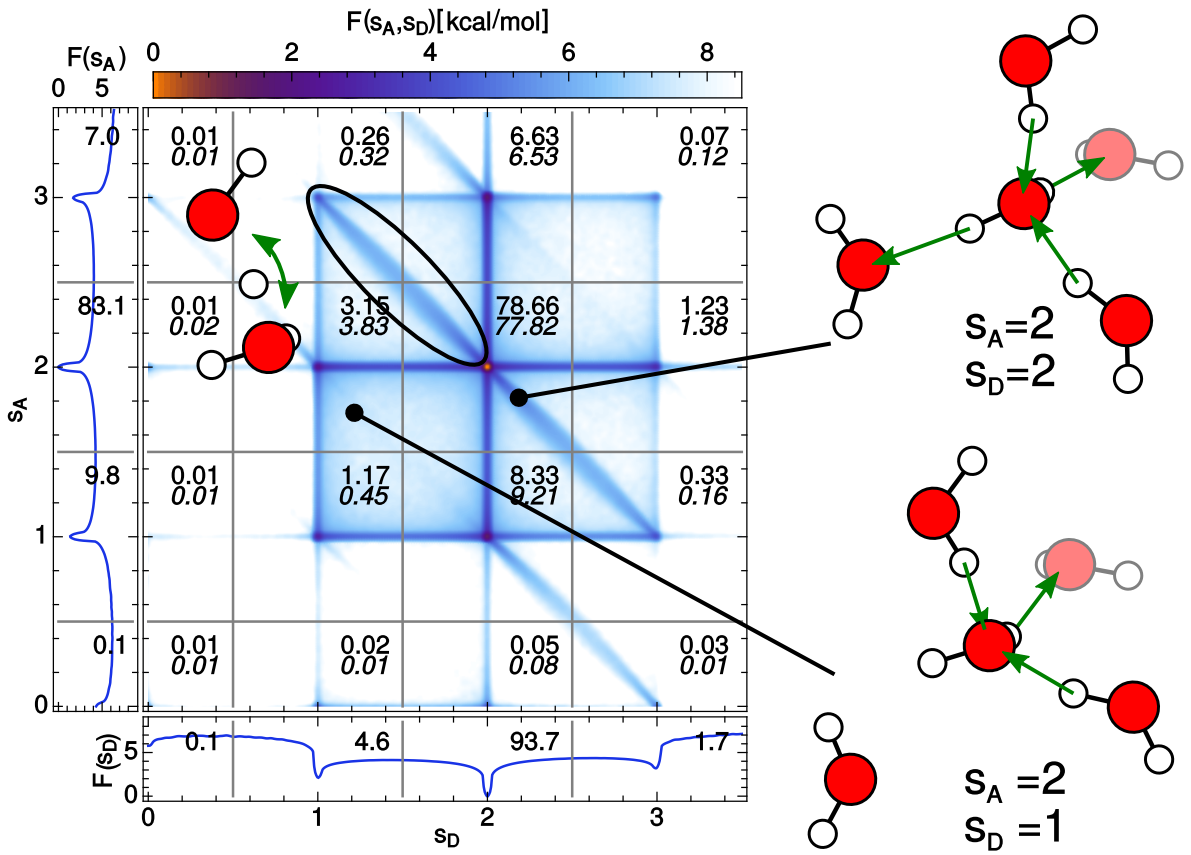
- Ab initio water PAMM recognizes multiple modes - one corresponds to the H-bond
- Adaptive, context-dependent definition - details of the model, thermodynamic conditions, type of HB



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Defects and correlations in water

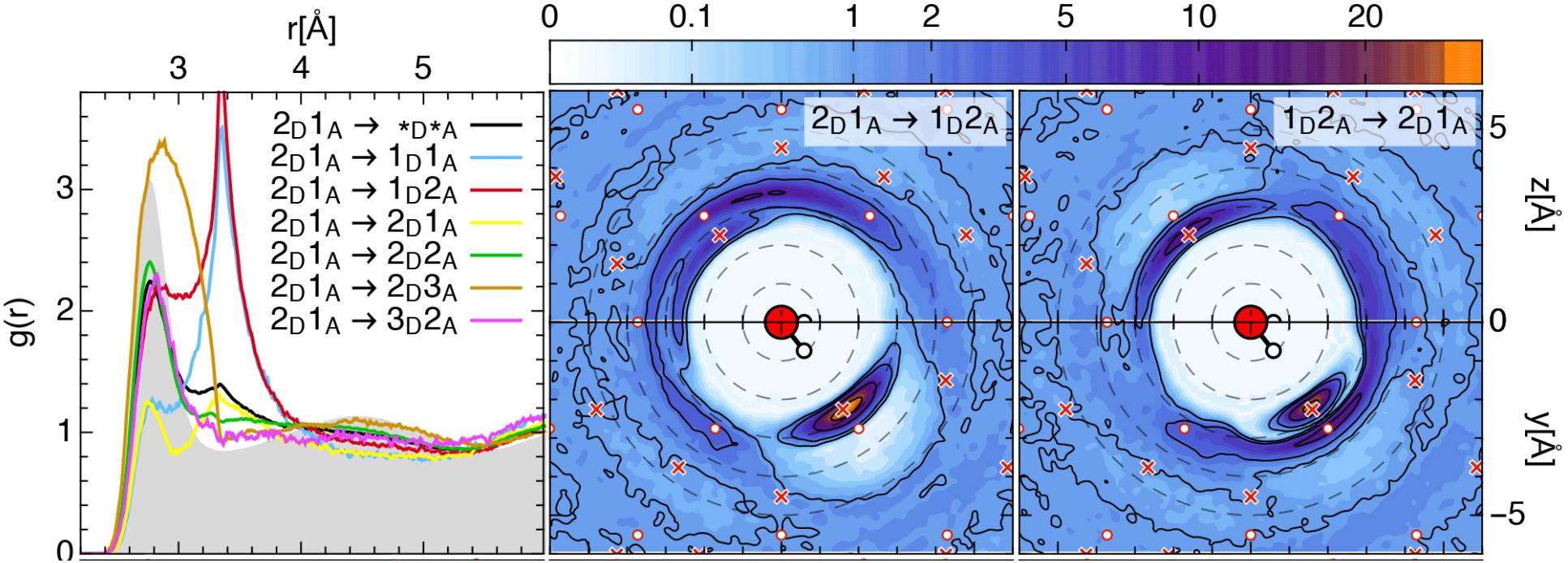
- Disentangling the correlations between topological defects in the H-bond network of water
- Angular correlations between standard 2_D2_A water correspond to ice Ih, but undercoordinated defects are similar to ice VIII



Gasparotto, Hassanali, Ceriotti, JCTC 2016

Defects and correlations in water

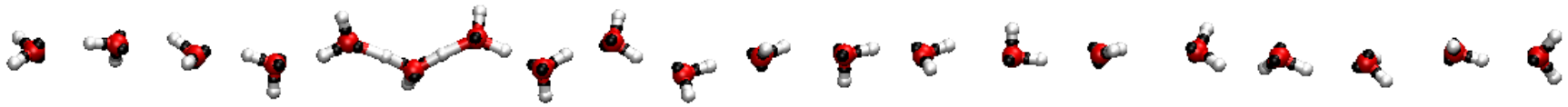
- Disentangling the correlations between topological defects in the H-bond network of water
- Angular correlations between standard 2_D2_A water correspond to ice Ih, but undercoordinated defects are similar to ice VIII



Gasparotto, Hassanali, Ceriotti, JCTC 2016

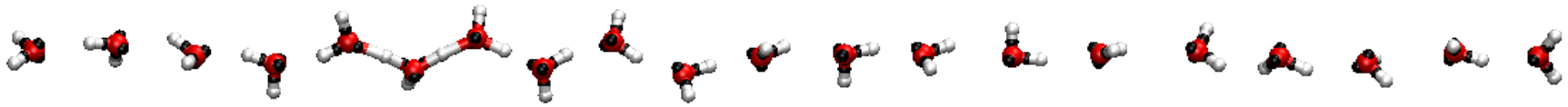
PAMM Analysis of a Water Wire

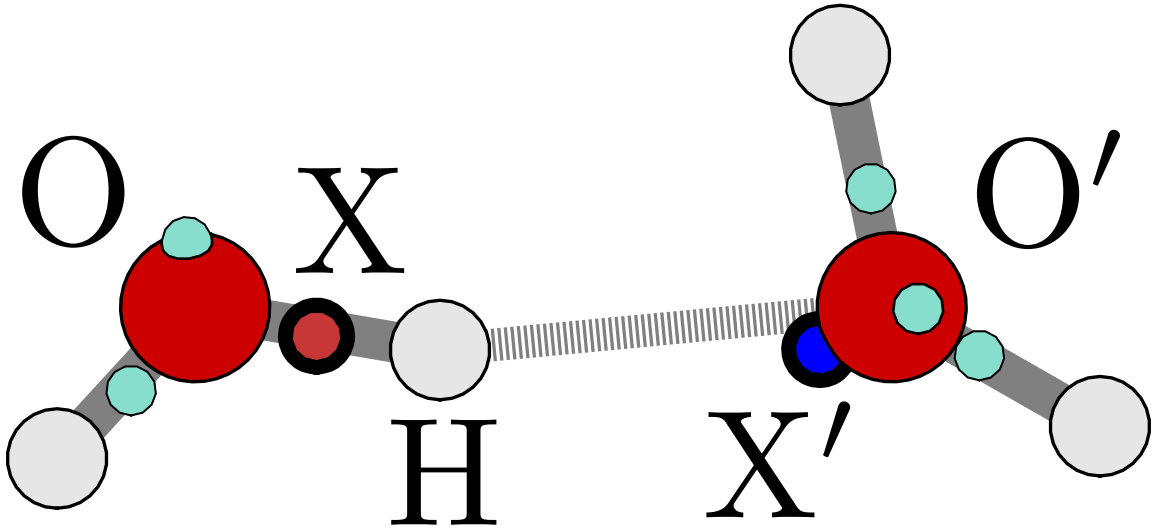
- Use PAMM to identify proton-like molecules in a wire, based on the position of Wannier centers relative to oxygens
- One can identify two separate clusters
- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire (then you can do quantum dynamics and measure diffusion rate)



PAMM Analysis of a Water Wire

- Use PAMM to identify proton-like molecules in a wire, based on the position of Wannier centers relative to oxygens
- One can identify two separate clusters
- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire (then you can do quantum dynamics and measure diffusion rate)

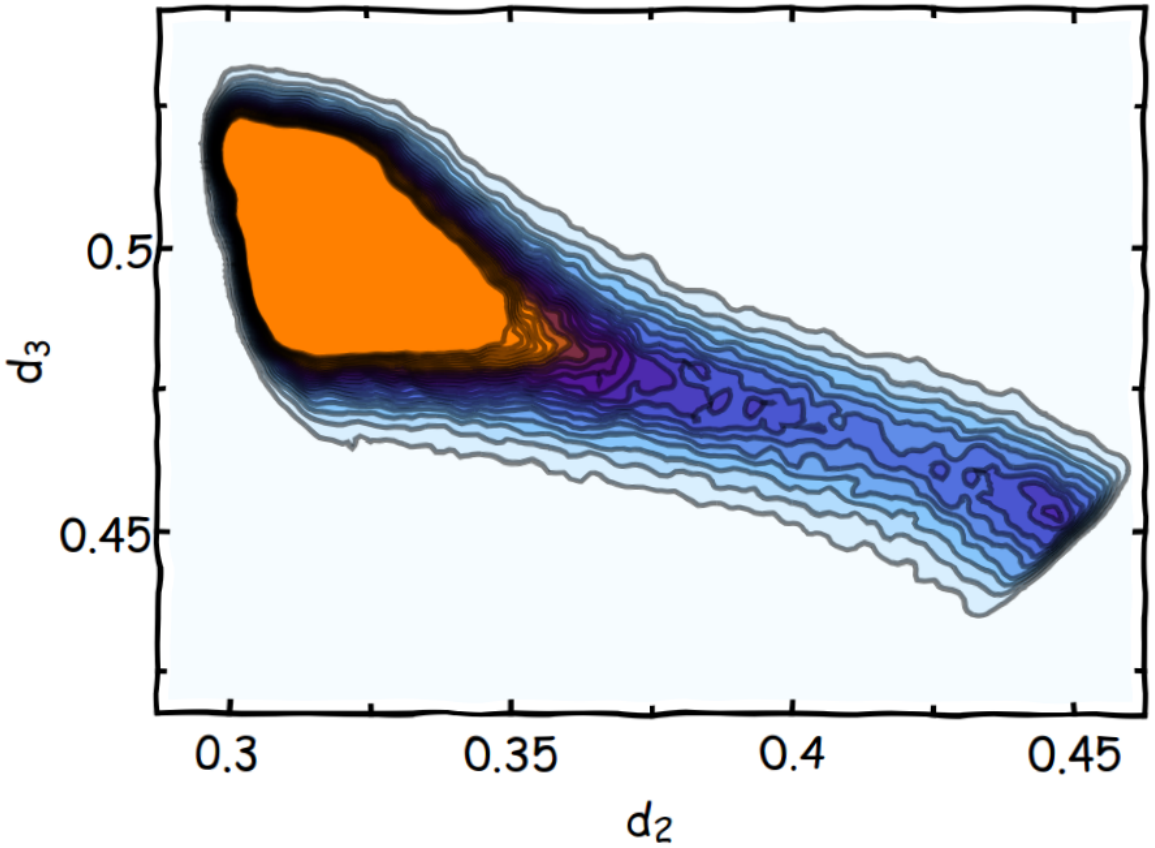




Marzari, Vanderbilt PRB 1997

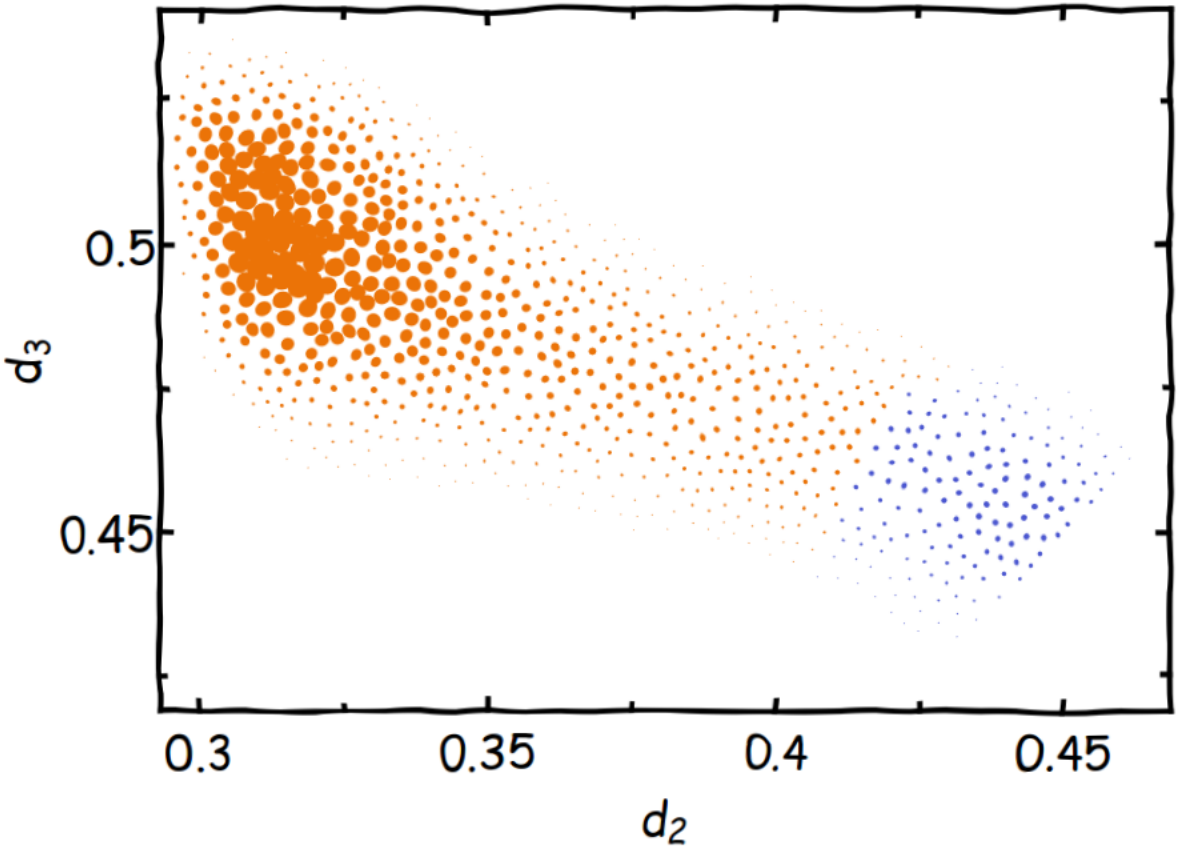
PAMM Analysis of a Water Wire

- Use PAMM to identify proton-like molecules in a wire, based on the position of Wannier centers relative to oxygens
- One can identify two separate clusters
- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire (then you can do quantum dynamics and measure diffusion rate)



PAMM Analysis of a Water Wire

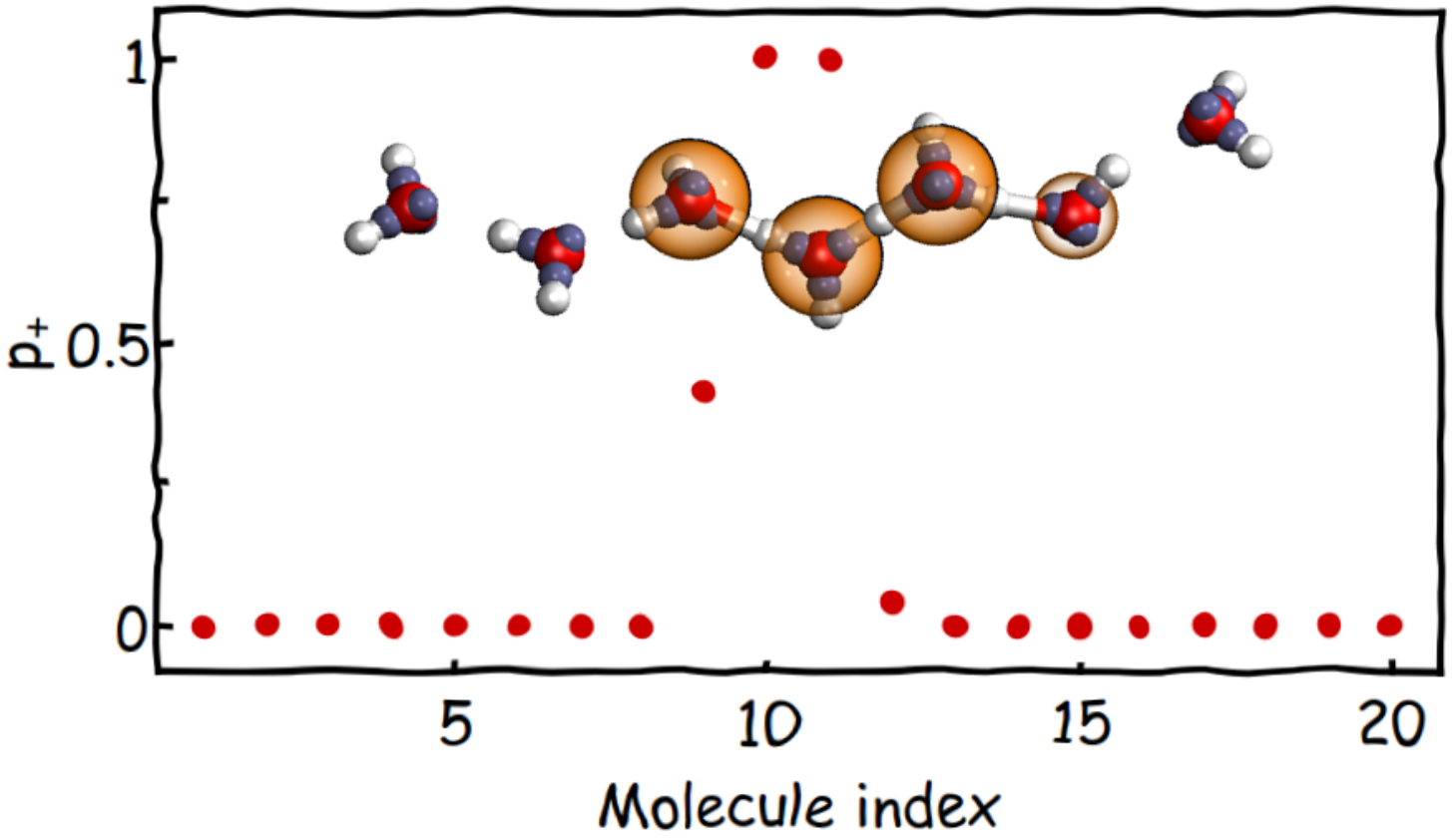
- Use PAMM to identify proton-like molecules in a wire, based on the position of Wannier centers relative to oxygens
- One can identify two separate clusters
- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire (then you can do quantum dynamics and measure diffusion rate)



Rossi, Ceriotti, Manolopoulos, JPCL 2016

PAMM Analysis of a Water Wire

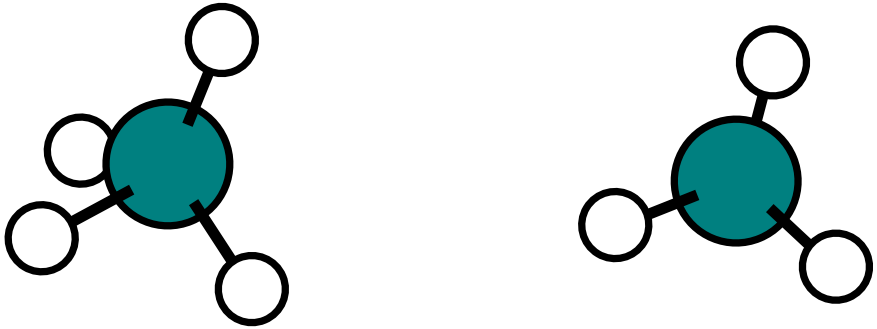
- Use PAMM to identify proton-like molecules in a wire, based on the position of Wannier centers relative to oxygens
- One can identify two separate clusters
- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire (then you can do quantum dynamics and measure diffusion rate)



Rossi, Ceriotti, Manolopoulos, JPCL 2016

Representing chemical environments

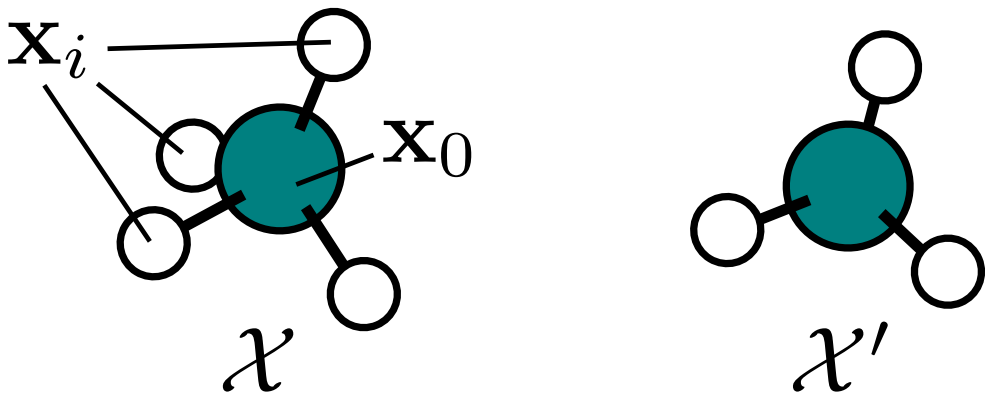
- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations



Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations

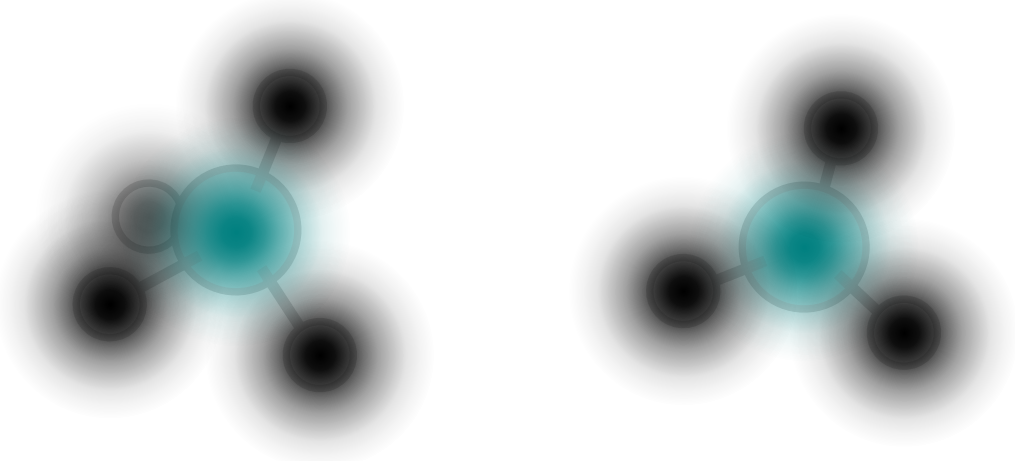


$$\{\mathbf{x}_i - \mathbf{x}_0\} \leftrightarrow \mathcal{X}$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations



$$\rho_\alpha(\mathbf{x}) = \sum_{i\in\alpha} g(\mathbf{x} - \mathbf{x}_i)$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations

$$k(\mathcal{X}, \mathcal{X}') = \int \rho(\mathbf{x})\rho'(\mathbf{x})$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations

$$k(\mathcal{X},\mathcal{X}') = \int \mathrm{d}\hat{R} \left| \int \rho(\mathbf{x})\rho'(\hat{R}\mathbf{x}) \right|^2$$

Bartók, Kondor, Csányi, PRB (2013)

Representing chemical environments

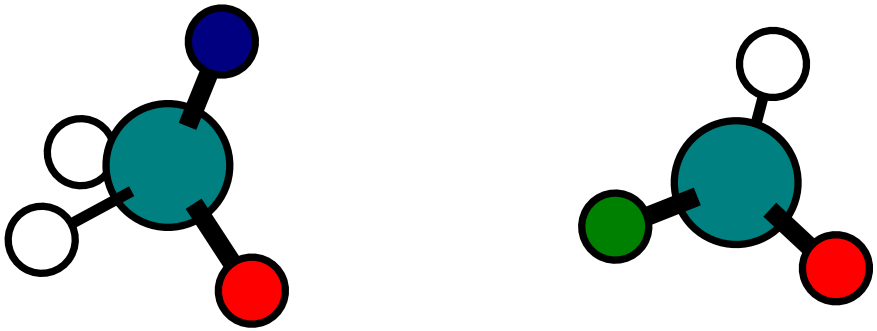
- A kernel to compare chemical environments based on the overlap of atomic densities (SOAP)
- Invariant to permutations, translations and rotations

$$k(\mathcal{X},\mathcal{X}') = \int \mathrm{d}\hat{R} \left| \int \rho(\mathbf{x})\rho'(\hat{R}\mathbf{x}) \right|^2$$

Bartók, Kondor, Csányi, PRB (2013)

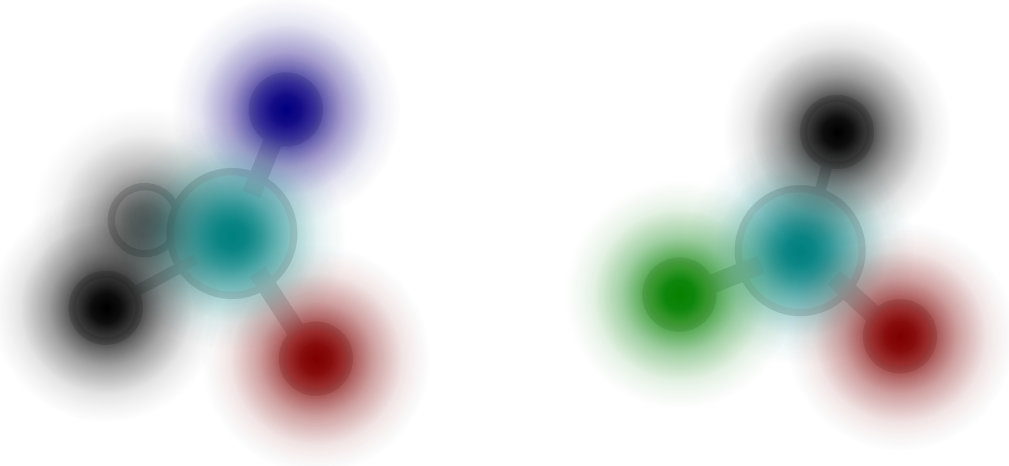
Comparing with multiple species

- What to do when comparing with multiple species?
- Density representation with multiple "channels"
- Treating alchemical as well as structural complexity on the same footings by introducing an **alchemical similarity kernel**



Comparing with multiple species

- What to do when comparing with multiple species?
- Density representation with multiple "channels"
- Treating alchemical as well as structural complexity on the same footings by introducing an **alchemical similarity kernel**

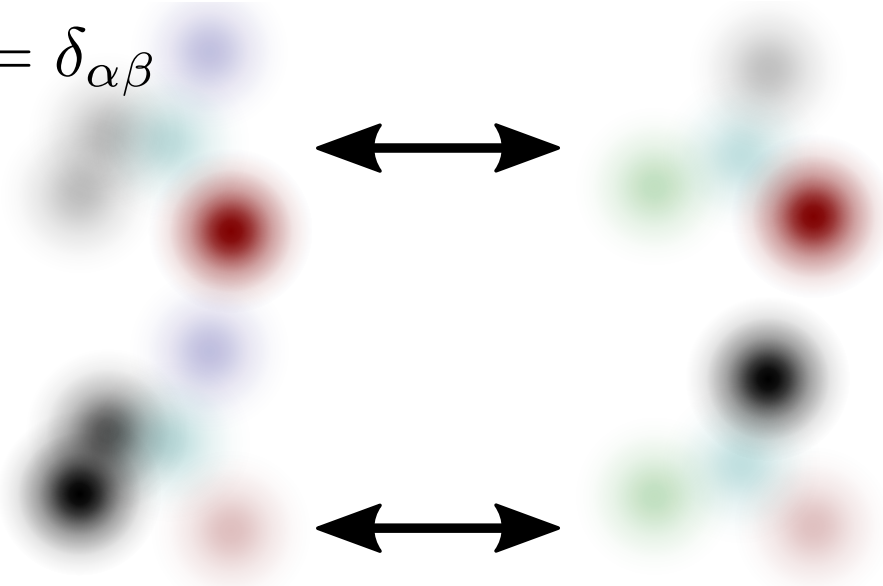


$$\rho = \sum_\alpha \rho_\alpha(\mathbf{x}) \, |\alpha\rangle$$

Comparing with multiple species

- What to do when comparing with multiple species?
- Density representation with multiple "channels"
- Treating alchemical as well as structural complexity on the same footings by introducing an **alchemical similarity kernel**

$$\langle \alpha | \beta \rangle = \delta_{\alpha\beta}$$

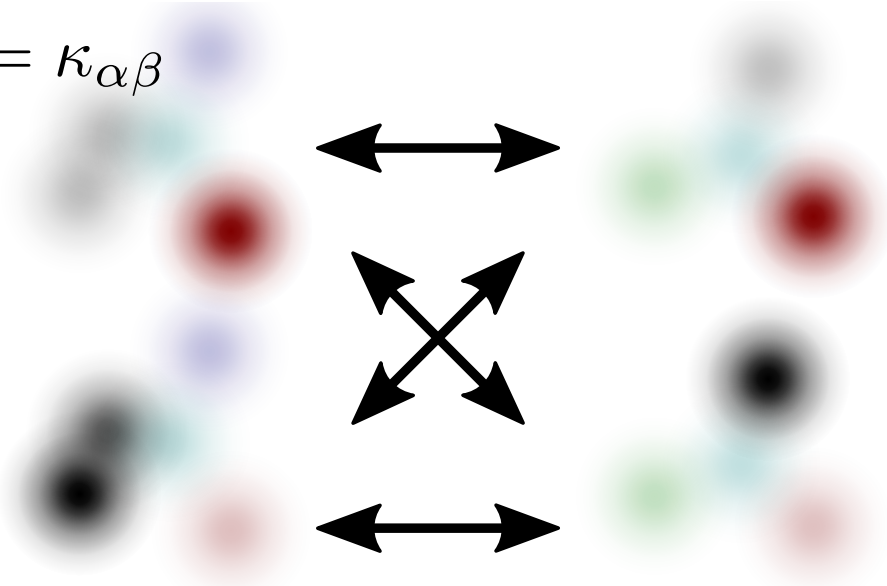


$$\int \mathrm{d}\hat{R} \left| \sum_\alpha \int \rho_\alpha(\mathbf{x}) \rho'_\alpha(\hat{R}\mathbf{x}) \right|^2$$

Comparing with multiple species

- What to do when comparing with multiple species?
- Density representation with multiple "channels"
- Treating alchemical as well as structural complexity on the same footings by introducing an **alchemical similarity kernel**

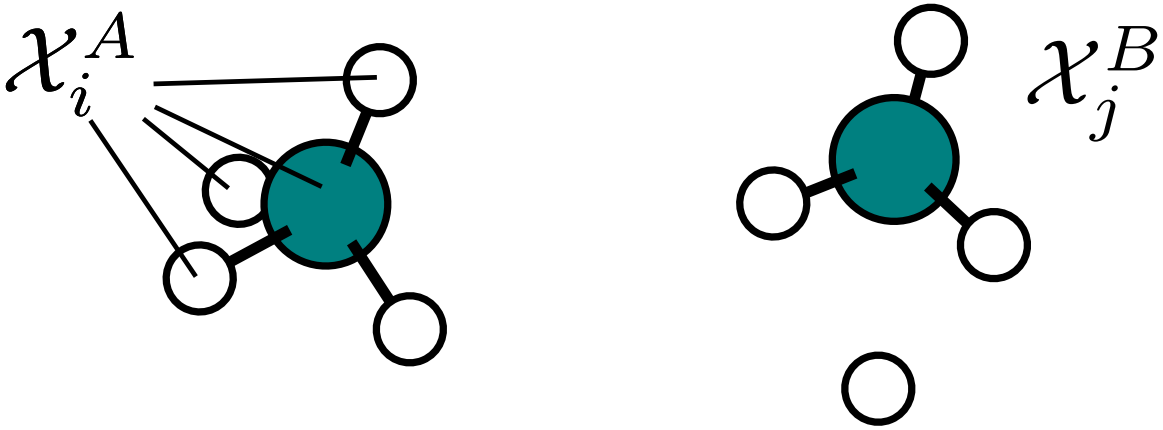
$$\langle \alpha | \beta \rangle = \kappa_{\alpha\beta}$$



$$\int \mathrm{d}\hat{R} \left| \sum_{\alpha\beta} \kappa_{\alpha\beta} \int \rho_\alpha(\mathbf{x}) \rho'_\beta(\hat{R}\mathbf{x}) \right|^2$$

From environment kernels to structural kernels

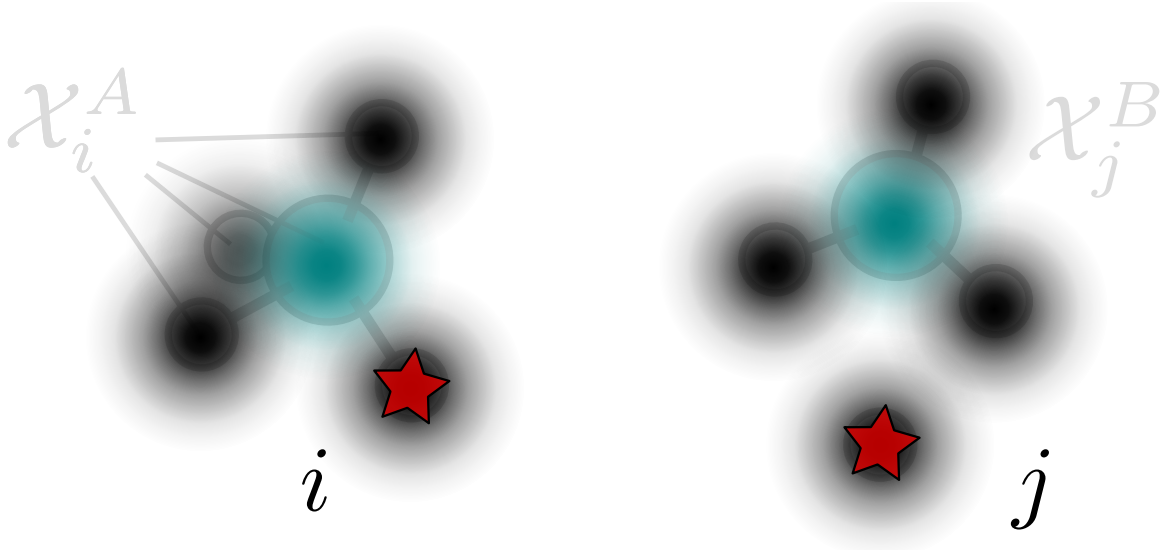
- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$



De, Bartók, Csányi, Ceriotti, PCCP (2016)

From environment kernels to structural kernels

- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$

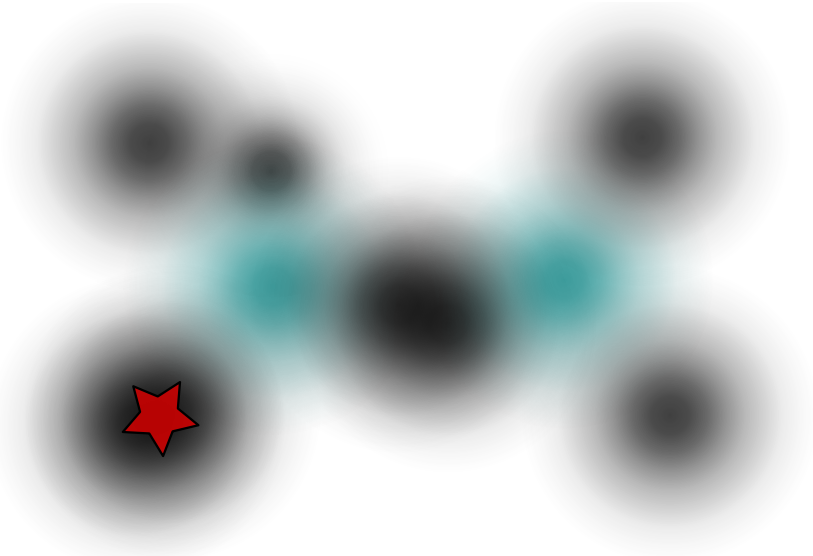


$$C_{ij}^{A,B} = k(\mathcal{X}_i^A, \mathcal{X}_j^B)$$

De, Bartók, Csányi, Ceriotti, PCCP (2016)

From environment kernels to structural kernels

- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$

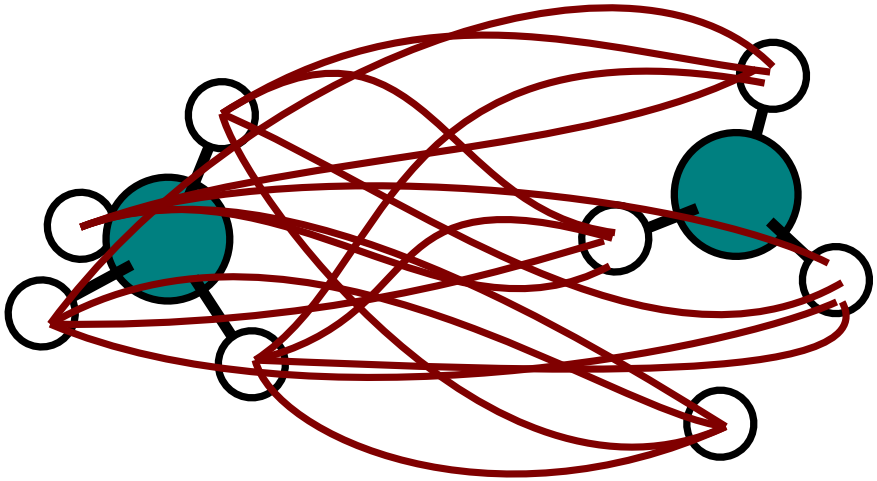


$$C_{ij}^{A,B} = k(\mathcal{X}_i^A, \mathcal{X}_j^B)$$

De, Bartók, Csányi, Ceriotti, PCCP (2016)

From environment kernels to structural kernels

- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$

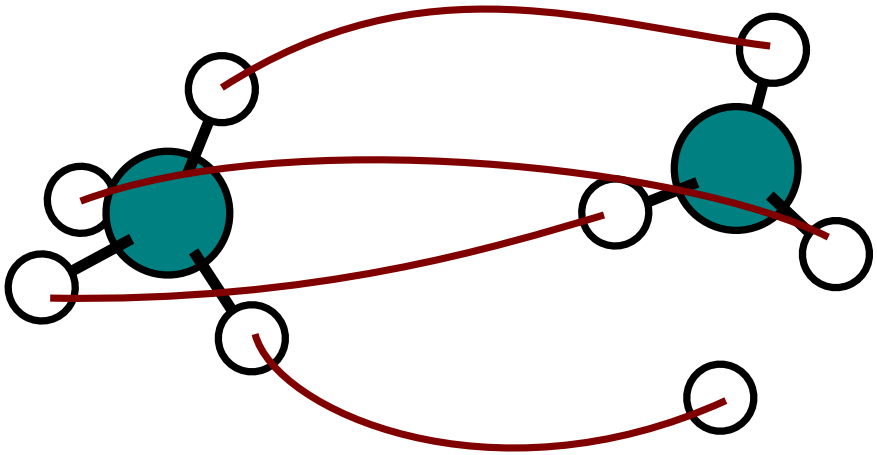


$$\bar{K}(A,B) \propto \sum_{ij} C_{ij}^{A,B}$$

De, Bartók, Csányi, Ceriotti, PCCP (2016)

From environment kernels to structural kernels

- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$

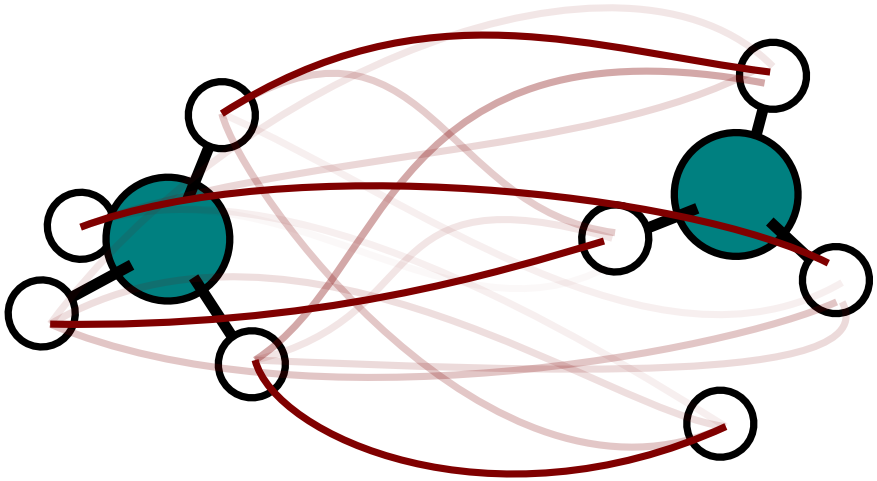


$$\hat{K}(A,B) \propto \max_\pi \sum_i C_{i\pi_i}^{A,B}$$

De, Bartók, Csányi, Ceriotti, PCCP (2016)

From environment kernels to structural kernels

- Given the kernel matrix between the environments of the two structures, we can build a global structural kernel
  1. By averaging: $\bar{K}(A,B) = \frac{1}{N^2}\sum_{ij} k\left(\mathcal{X}_i^A, \mathcal{X}_i^B\right)$
  2. By picking the best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_\pi \sum_i k\left(\mathcal{X}_i^A, \mathcal{X}_{\pi_i}^B\right)$
  3. By **regularized** best-match permutation $\hat{K}(A,B) = \frac{1}{N}\max_{\mathbf{P}} \sum_{ij} P_{ji} k\left(\mathcal{X}_i^A, \mathcal{X}_j^B\right) - \gamma \sum_{ij} P_{ij} \ln P_{ij}$

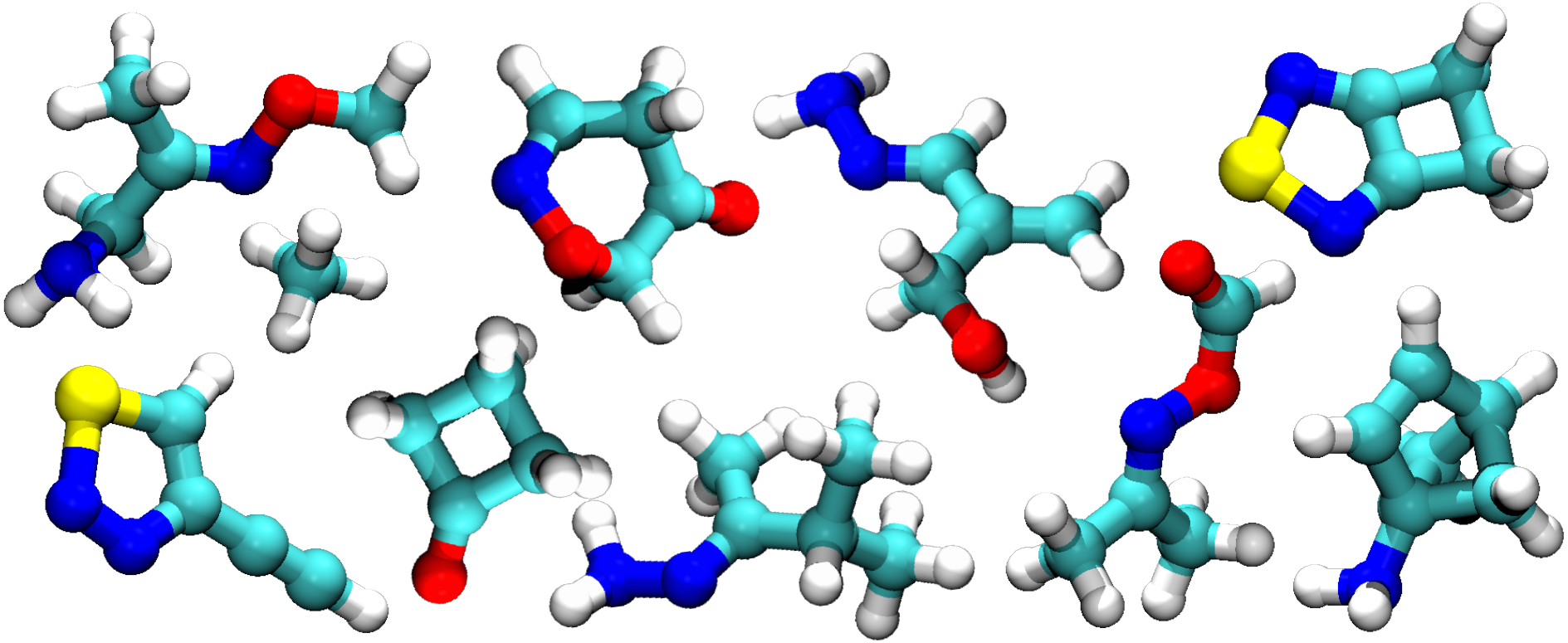


$$\hat{K}^\gamma(A,B) \propto$$
$$\max_{\mathbf{P}\in\mathcal{U}} \sum_{ij} P_{ji}(C_{ij}^{A,B} - \gamma \ln P_{ji})$$

M. Cuturi, NIPS 2013

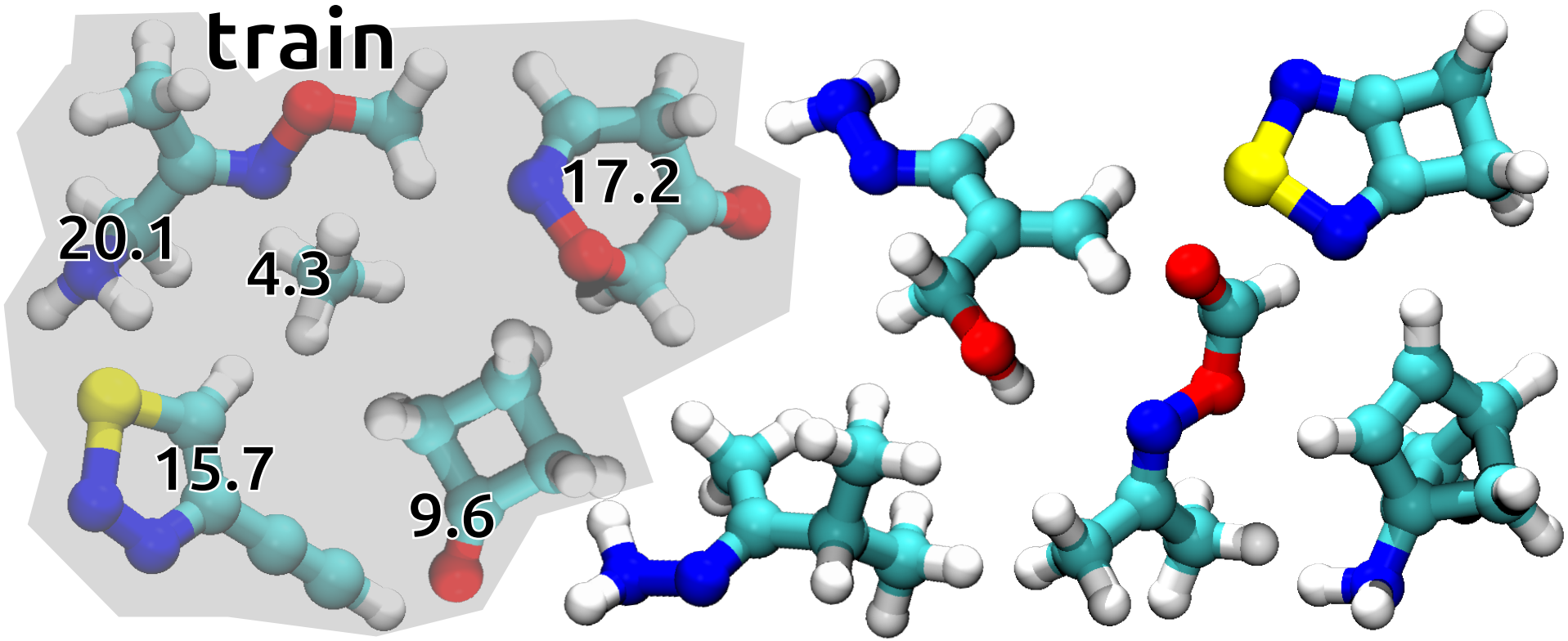
Teaching chemistry to a computer

- Using kernel-ridge regression to learn molecular properties (pioneered by Von Lilienfeld, Müller, Tkatchenko, Rupp, . . .)
- SOAP-REMatch kernel gives chemical accuracy, effortlessly (QM7b dataset)



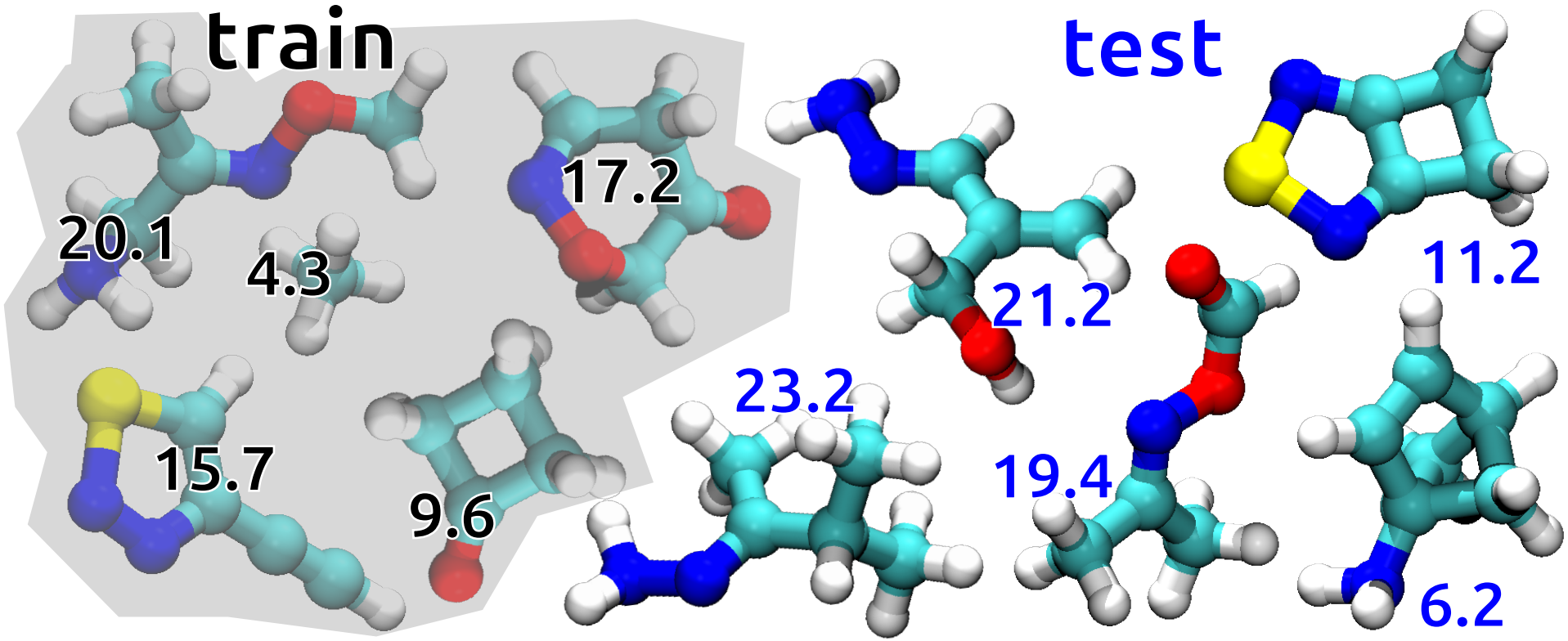
Teaching chemistry to a computer

- Using kernel-ridge regression to learn molecular properties (pioneered by Von Lilienfeld, Müller, Tkatchenko, Rupp, . . .)
- SOAP-REMatch kernel gives chemical accuracy, effortlessly (QM7b dataset)



Teaching chemistry to a computer

- Using kernel-ridge regression to learn molecular properties (pioneered by Von Lilienfeld, Müller, Tkatchenko, Rupp, . . .)
- SOAP-REMatch kernel gives chemical accuracy, effortlessly (QM7b dataset)



$$E(\mathcal{A}) = \sum c_i K(\mathcal{A}, \mathcal{A}_i)$$

Teaching chemistry to a computer

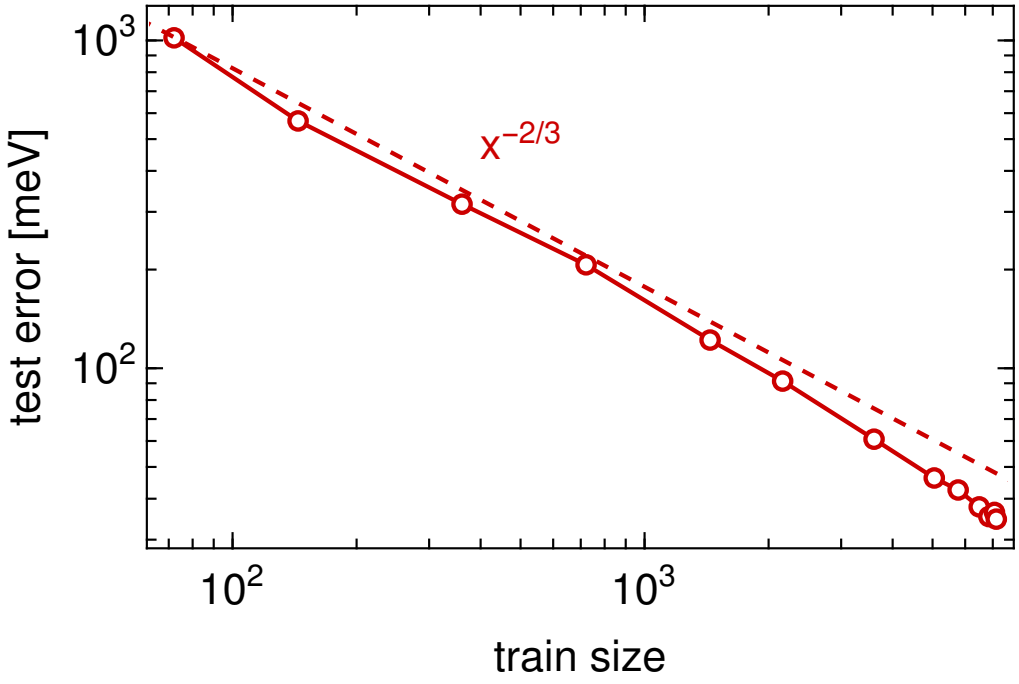
- Using kernel-ridge regression to learn molecular properties (pioneered by Von Lilienfeld, Müller, Tkatchenko, Rupp, . . .)
- SOAP-REMatch kernel gives chemical accuracy, effortlessly (QM7b dataset)

Property (eV, Å^3)	**SD**	**MAE**	**RMSE**	**MAE[1]**	**RMSE[1]**
E (PBE0)	9.70	0.04	0.07	0.16	0.36
α (PBE0)	1.34	0.05	0.07	0.11	0.18
α (SCS)	1.47	0.02	0.04	0.08	0.12
HOMO (GW)	0.70	0.12	0.17	0.16	0.22
HOMO (PBE0)	0.63	0.11	0.15	0.15	0.21
LUMO (GW)	0.48	0.12	0.17	0.13	0.21
LUMO (PBE0)	0.68	0.08	0.12	0.12	0.20

[1] data from Montavon et al. NJP 2013
http://quantum-machine.org/
De, Bartók, Csányi, Ceriotti, PCCP 2016

Learning rate & kernel hyperparameters

- Excellent learning rate up to the full dataset
- The kernel can be modified with a non-linear transform $K \leftarrow K^{\xi}$, and the KRR procedure can be regularized with a diagonal term $\sigma\mathbf{1}$. The REMatch kernel contains itself the entropy regularization parameter γ, and the SOAP kernels depend on the environment cutoff r_{max}
- Lots of room for development - e.g. on the alchemical kernel front....



Learning rate & kernel hyperparameters

- Excellent learning rate up to the full dataset
- The kernel can be modified with a non-linear transform $K \leftarrow K^{\xi}$, and the KRR procedure can be regularized with a diagonal term $\sigma\mathbf{1}$. The REMatch kernel contains itself the entropy regularization parameter γ, and the SOAP kernels depend on the environment cutoff r_{max}
- Lots of room for development - e.g. on the alchemical kernel front....

local$\longleftrightarrow$long range

r_{max} [Å] ($\gamma = 0.5$)	2.0	3.0	4.0	5.0
MAE (E, 30% test) ($\frac{kcal}{mol}$)	5.6	1.0	1.6	1.8

best match$\longleftrightarrow$average

γ	0	0.2	0.5	1.0	2.0	5.0	10	50	∞
MAE (E, 30% test) ($\frac{kcal}{mol}$)	5e3	92	1.0	0.8	0.8	0.8	0.9	1.1	3.0

Learning rate & kernel hyperparameters

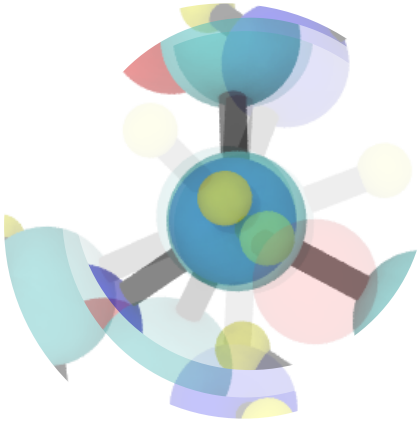
- Excellent learning rate up to the full dataset
- The kernel can be modified with a non-linear transform $K \leftarrow K^{\xi}$, and the KRR procedure can be regularized with a diagonal term $\sigma\mathbf{1}$. The REMatch kernel contains itself the entropy regularization parameter γ, and the SOAP kernels depend on the environment cutoff r_{max}
- Lots of room for development - e.g. on the alchemical kernel front....

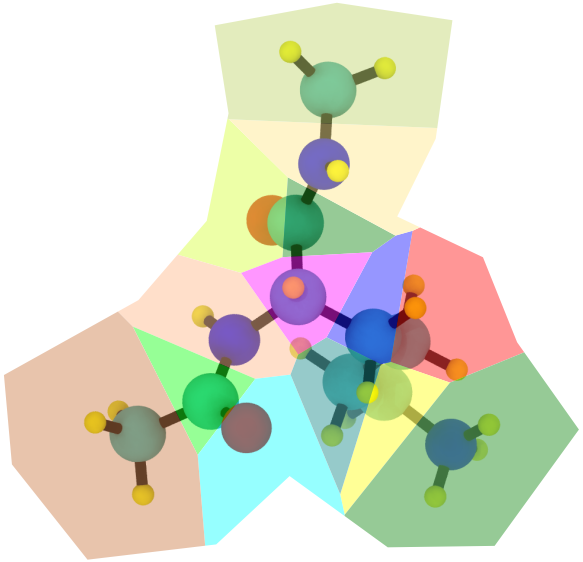
$$\kappa_{\alpha\beta} = e^{-(E_\alpha - E_\beta)^2/2} \rightarrow \textbf{MAE=0.55kcal/mol!}$$

. . . whad do we learn about electronegativity?

Growing larger

- With proper normalization, the average kernel is equivalent to an atom-centered energy decomposition $E_{\text{avg}}(A) \equiv \sum_i e(\mathcal{X}_i^A)$.
  A connection between chemical & potential learning.
- Training with the average kernel on GDB9 shows its limitations
- . . . but REMatch-ing brings you the extra mile down to $<$ 1kcal/mol MAE with 15% training set!





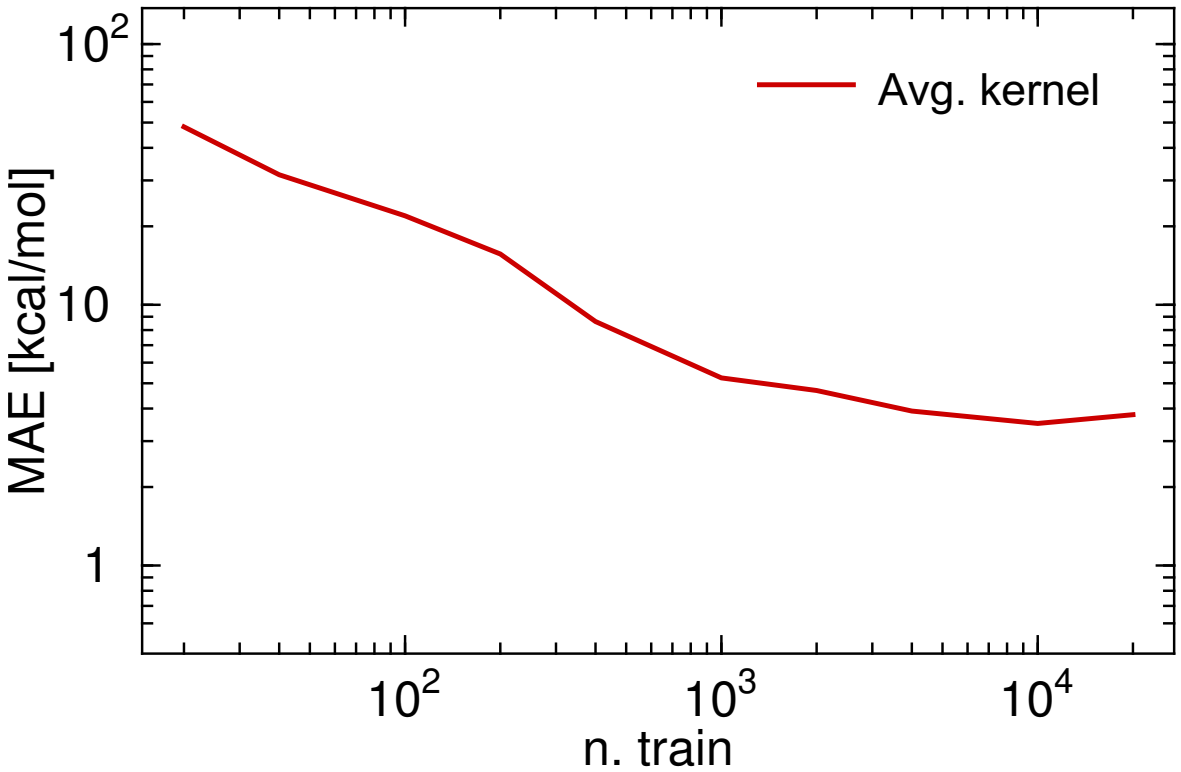
$$E(A) = E_{\text{avg}}\left(\frac{1}{N}\sum_i \mathcal{X}_i^A\right)$$

$$E(A) = \sum_i e(\mathcal{X}_i^A)$$

Work in progress w/Gabor and Albert

Growing larger

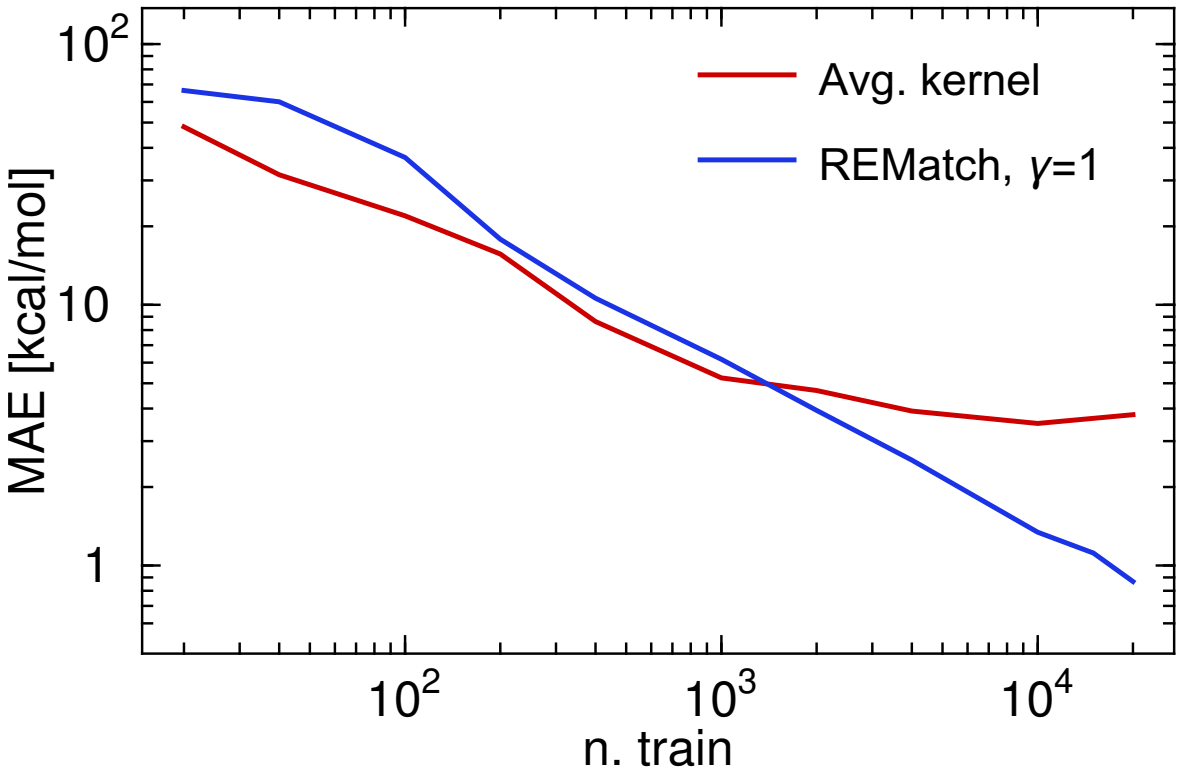
- With proper normalization, the average kernel is equivalent to an atom-centered energy decomposition $E_{\text{avg}}(A) \equiv \sum_i e\left(\mathcal{X}_i^A\right)$.
  A connection between chemical & potential learning.
- Training with the average kernel on GDB9 shows its limitations
- . . . but REMatch-ing brings you the extra mile down to $<$ 1kcal/mol MAE with 15% training set!



Work in progress w/Gabor and Albert

Growing larger

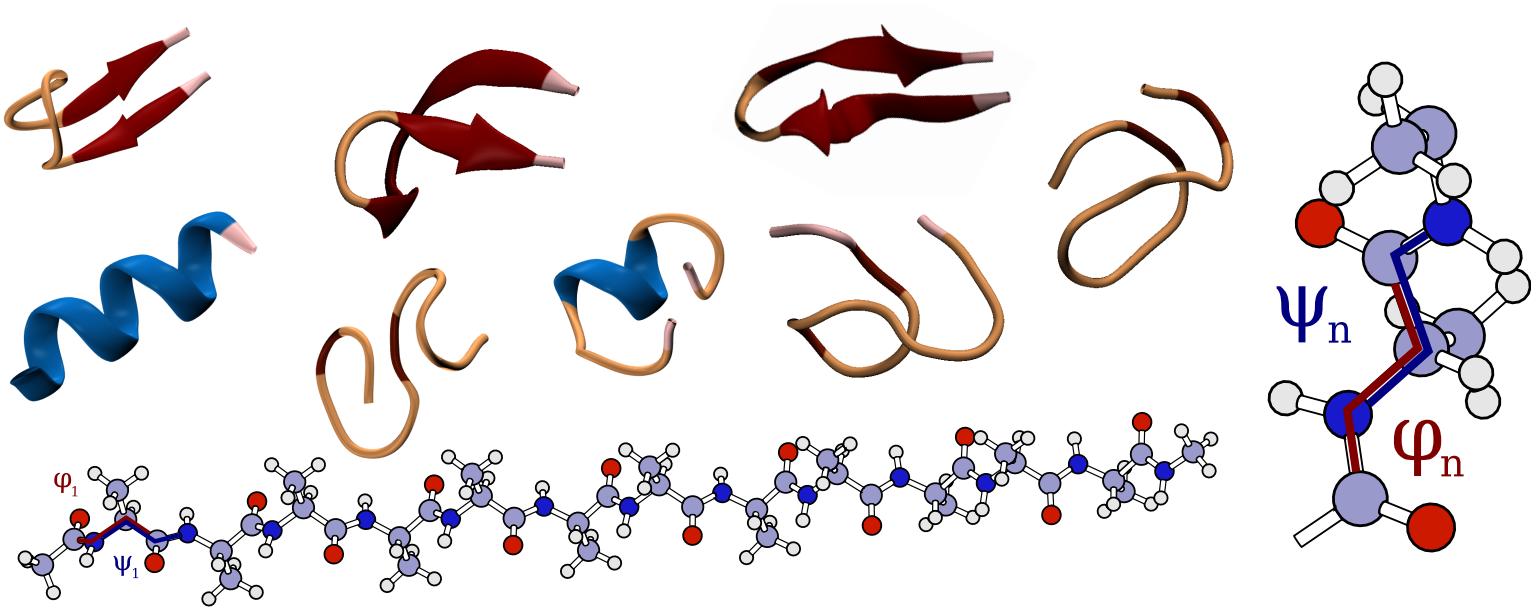
- With proper normalization, the average kernel is equivalent to an atom-centered energy decomposition $E_{\text{avg}}(A) \equiv \sum_i e\left(\mathcal{X}_i^A\right)$.
  A connection between chemical & potential learning.
- Training with the average kernel on GDB9 shows its limitations
- . . . but REMatch-ing brings you the extra mile down to $<$ 1kcal/mol MAE with 15% training set!



Work in progress w/Gabor and Albert

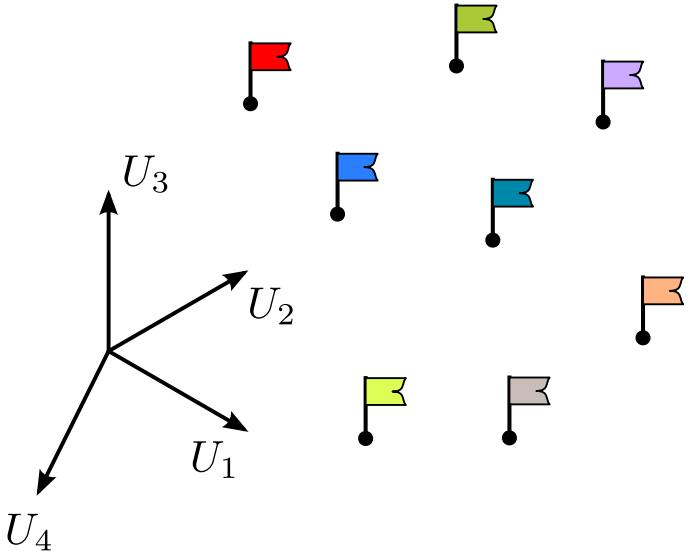
Dimensionality reduction

- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**



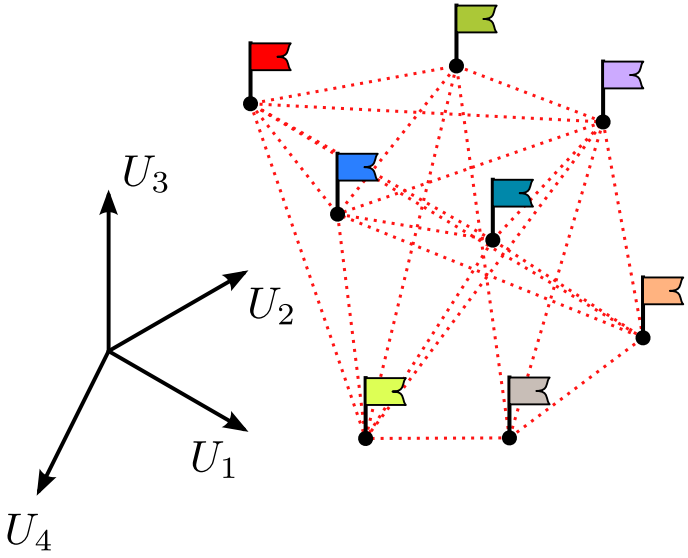
Dimensionality reduction

- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**



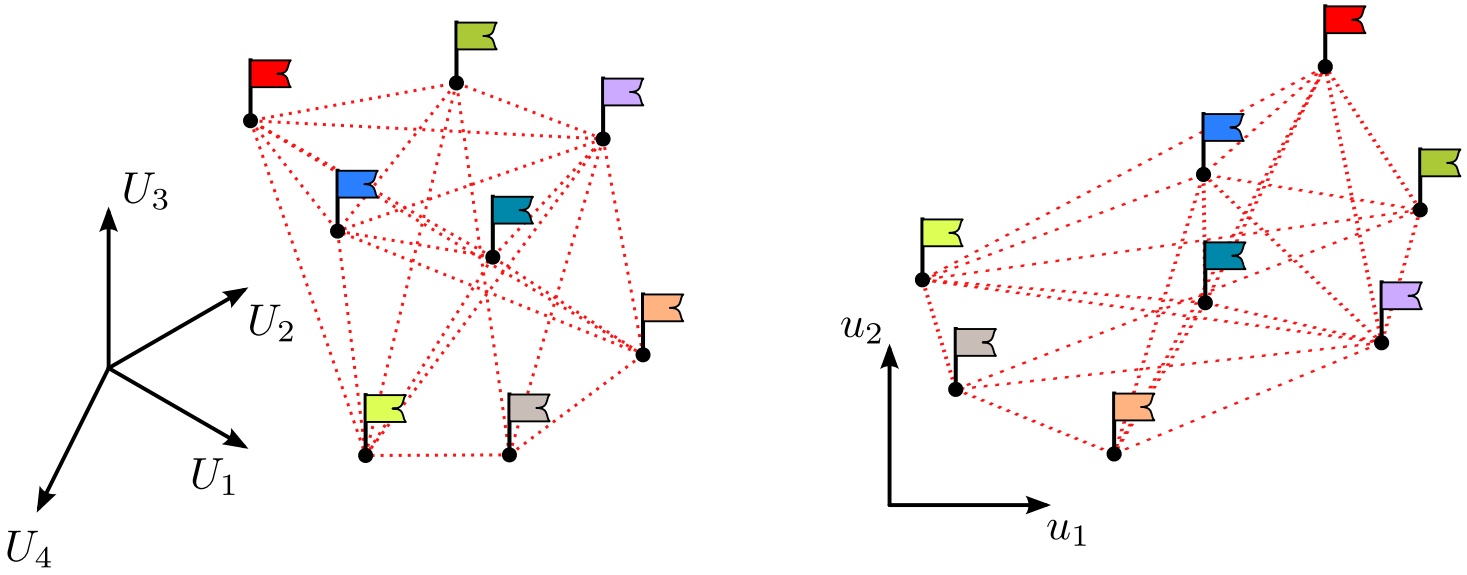
Dimensionality reduction

- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**



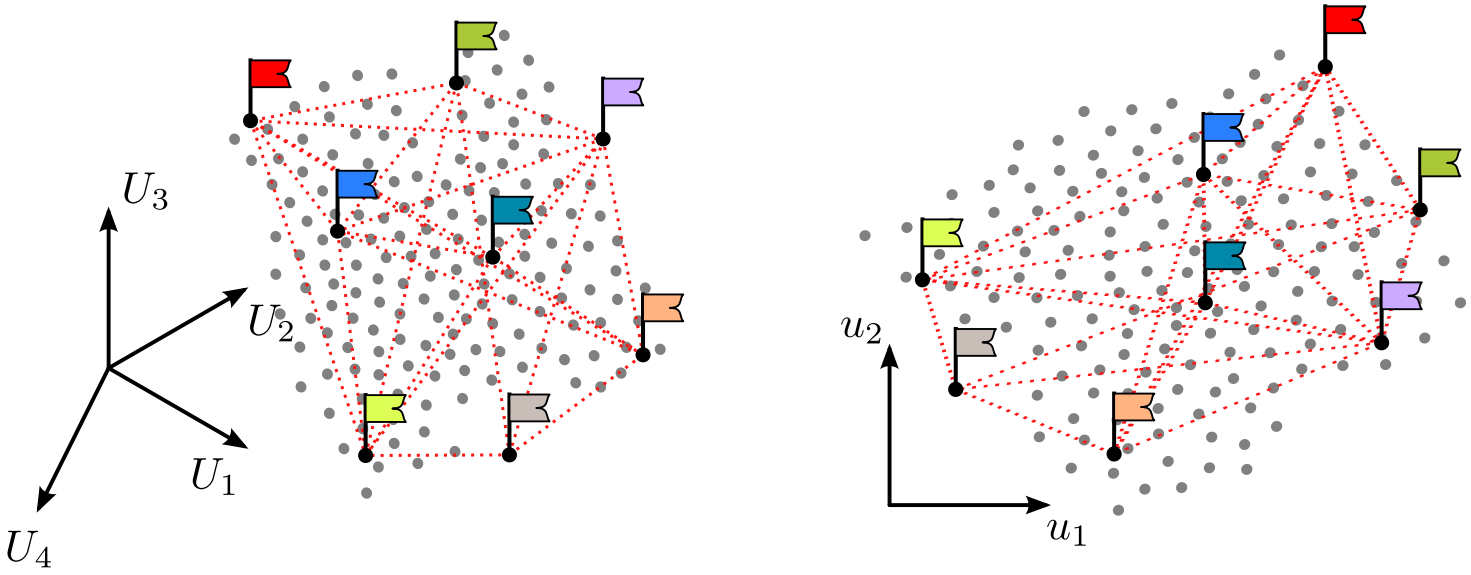
Dimensionality reduction

- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**



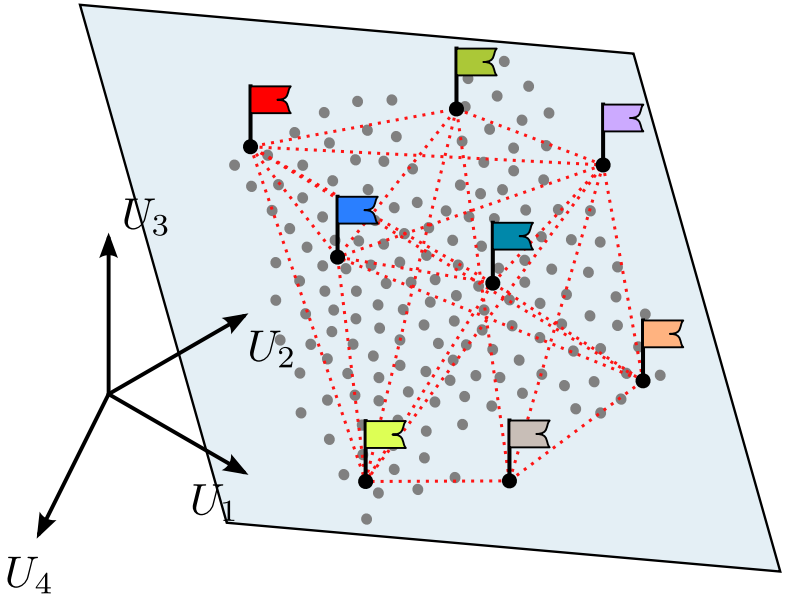
Dimensionality reduction

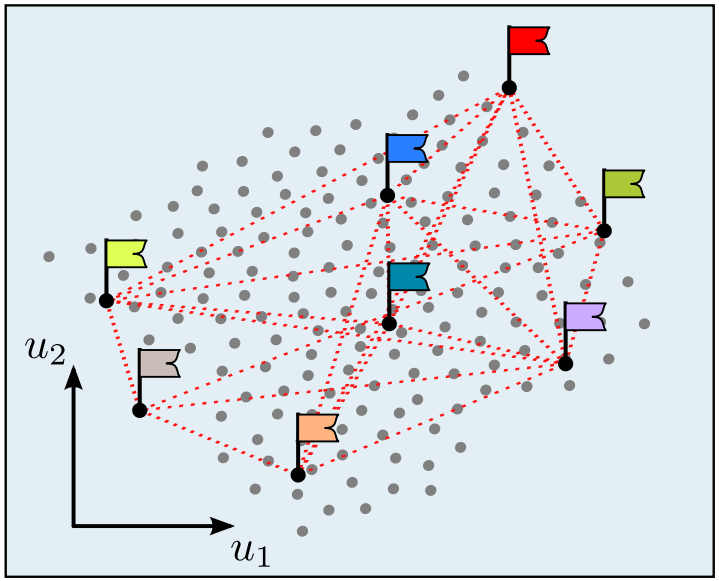
- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**



Dimensionality reduction

- Distances describe the relations between different molecules or structures. However, to visualize such relations we need to generate a low-dimensional **map** that is (approximately) consistent with the metric.
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Define a measure of dissimilarity between the points
  - Arrange low-dim. points so that the dissimilarities are preserved
  - Locate other configurations with an **out-of-sample embedding**





Sketch-map algorithm

- In "metric" MDS a stress function that measures how well distances are reproduced is minimized
- Modify the objective function to aim for proximity matching
  - Distances are transformed by **sigmoid functions** in both high and low dimension, to disregard thermal fluctuations

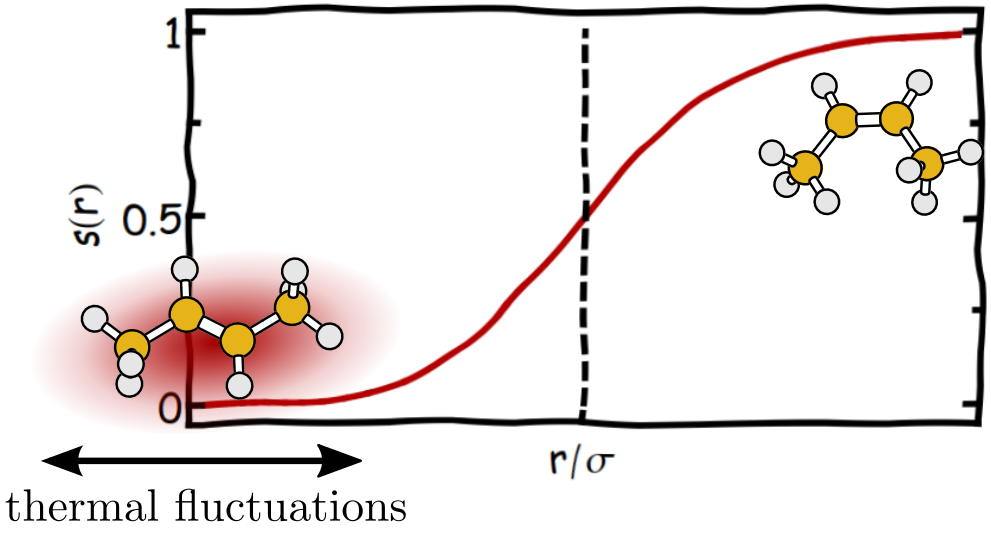
$$\chi^2 = \sum_{i,j=1}^{N} \left[\left| X_i - X_j \right| - \left| x_i - x_j \right| \right]^2$$

Ceriotti, Tribello, Parrinello, PNAS (2011); JCTC (2013)

Sketch-map algorithm

- In "metric" MDS a stress function that measures how well distances are reproduced is minimized
- Modify the objective function to aim for proximity matching
  - Distances are transformed by **sigmoid functions** in both high and low dimension, to disregard thermal fluctuations

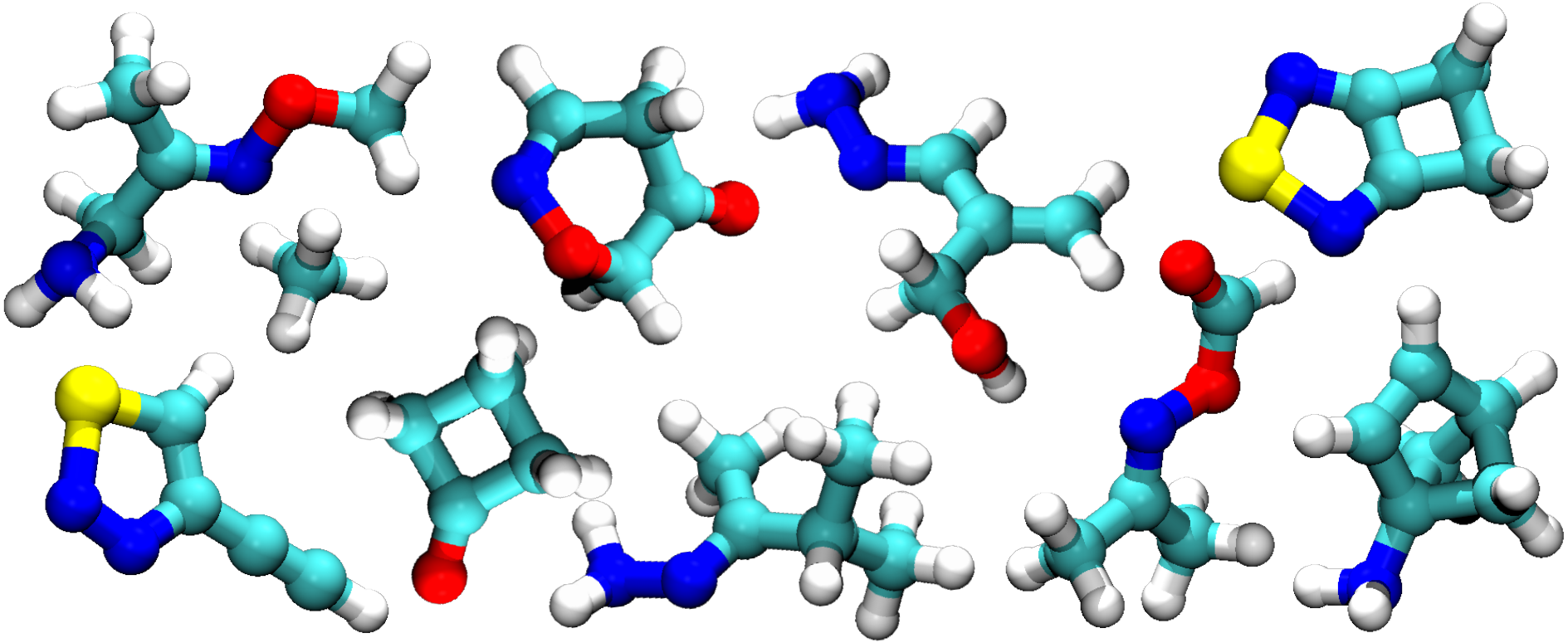
$$\chi^2 = \sum_{i,j=1}^{N} \left[s\left(\left|X_i - X_j\right|\right) - s\left(\left|x_i - x_j\right|\right) \right]^2$$



Ceriotti, Tribello, Parrinello, PNAS (2011); JCTC (2013)

A periodic table of molecules

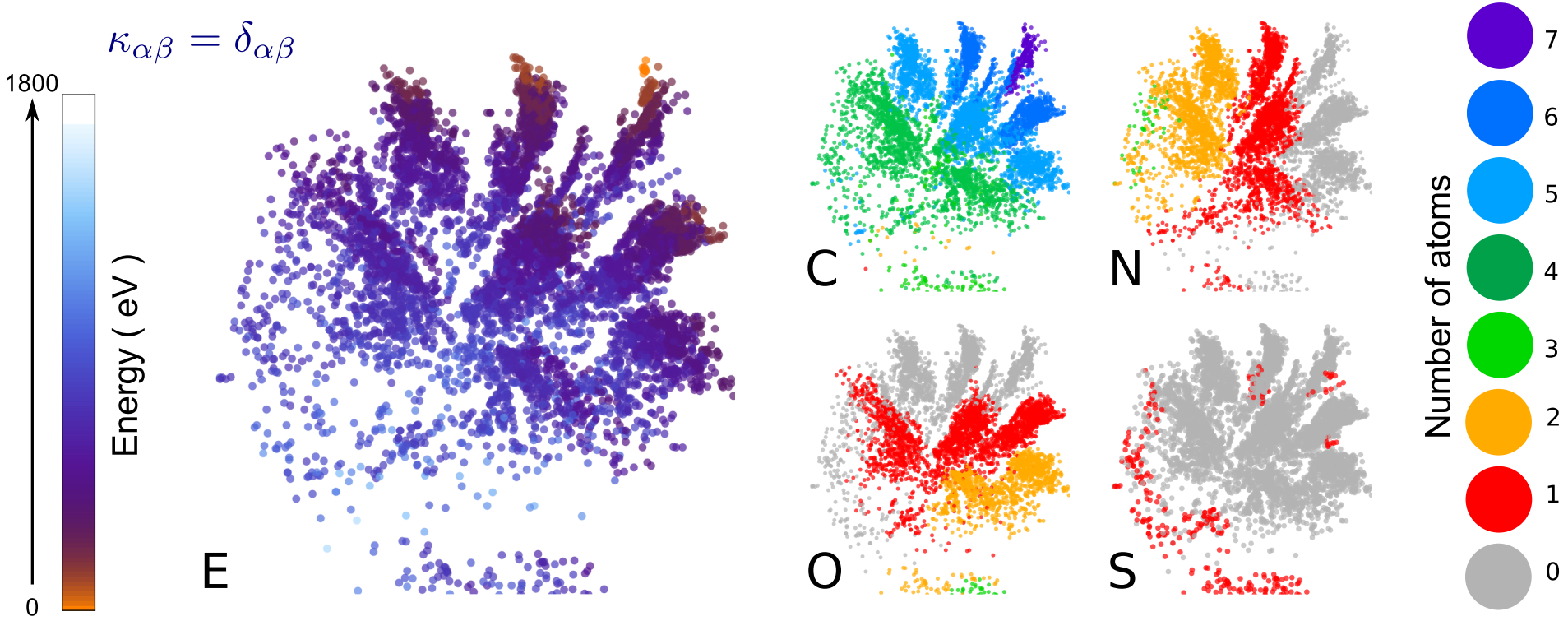
- Mapping QM7b. Variable number and nature of atoms
- Maps based on combination of local kernels represents nicely stoichiometry and energetics
- Modifying the alchemical similarity kernel modifies the metric and modulates the emphasis of the map between structure and composition



data from Montavon et al. NJP 2013

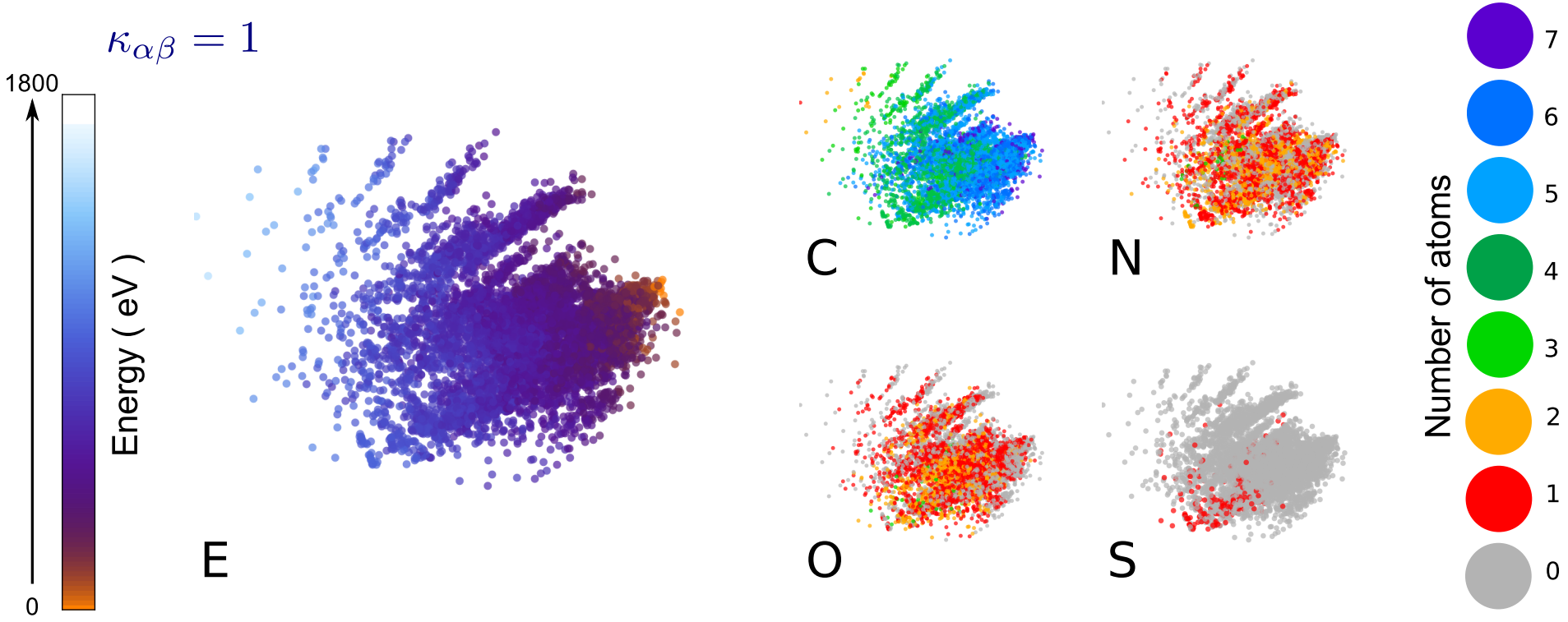
A periodic table of molecules

- Mapping QM7b. Variable number and nature of atoms
- Maps based on combination of local kernels represents nicely stoichiometry and energetics
- Modifying the alchemical similarity kernel modifies the metric and modulates the emphasis of the map between structure and composition



A periodic table of molecules

- Mapping QM7b. Variable number and nature of atoms
- Maps based on combination of local kernels represents nicely stoichiometry and energetics
- Modifying the alchemical similarity kernel modifies the metric and modulates the emphasis of the map between structure and composition

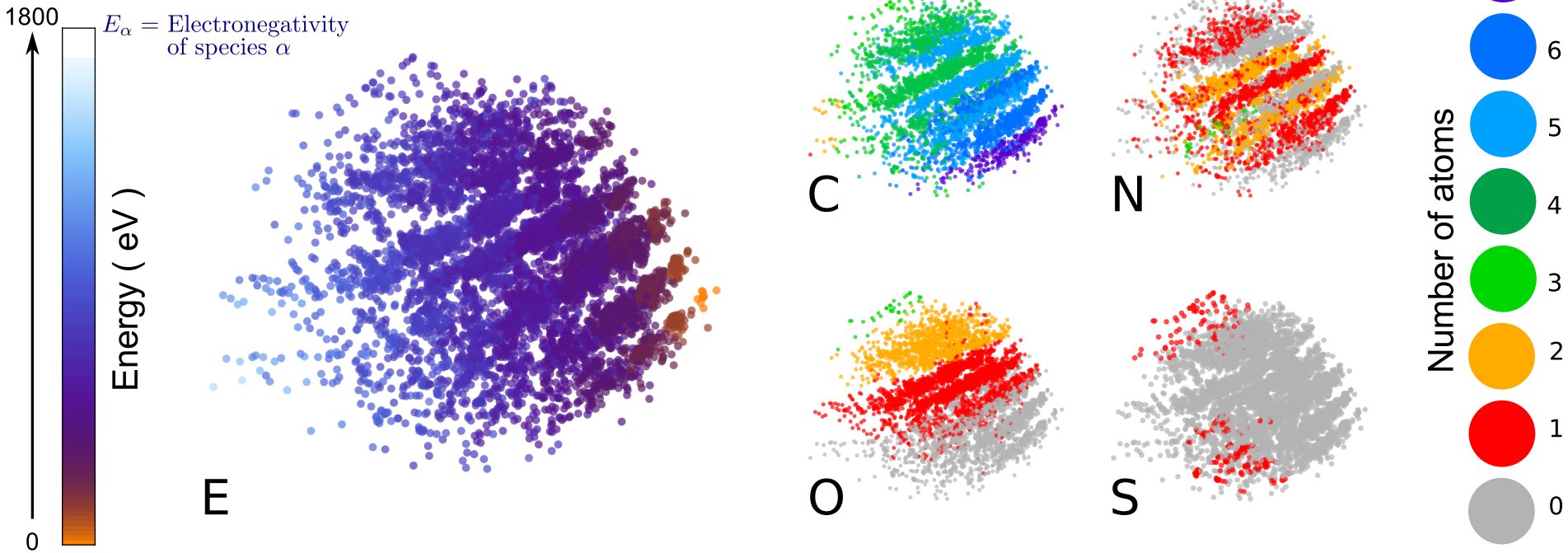


A periodic table of molecules

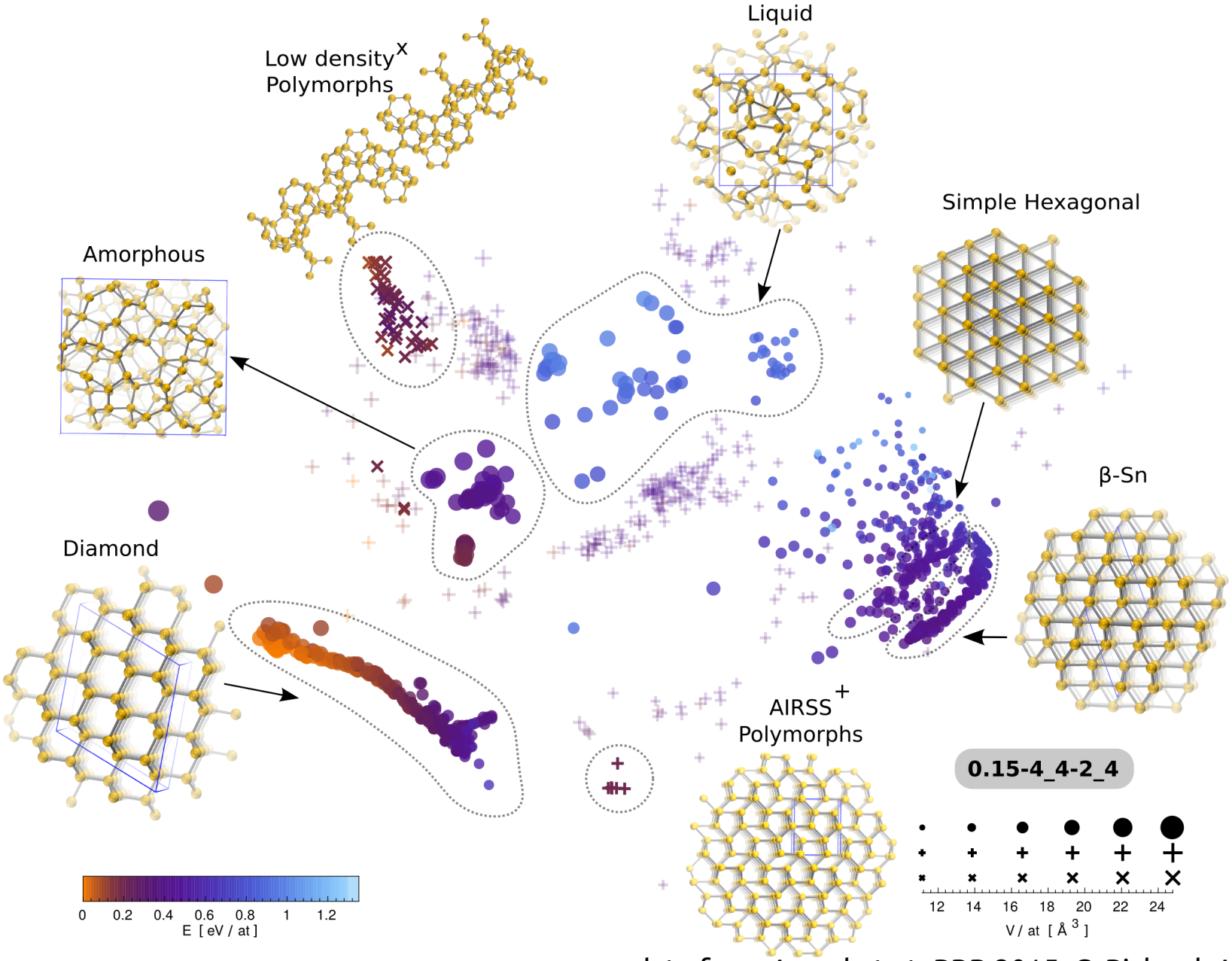
- Mapping QM7b. Variable number and nature of atoms
- Maps based on combination of local kernels represents nicely stoichiometry and energetics
- Modifying the alchemical similarity kernel modifies the metric and modulates the emphasis of the map between structure and composition

$$\kappa_{\alpha\beta} = e^{-\frac{(E_\alpha - E_\beta)^2}{2}}$$

E_α = Electronegativity of species α



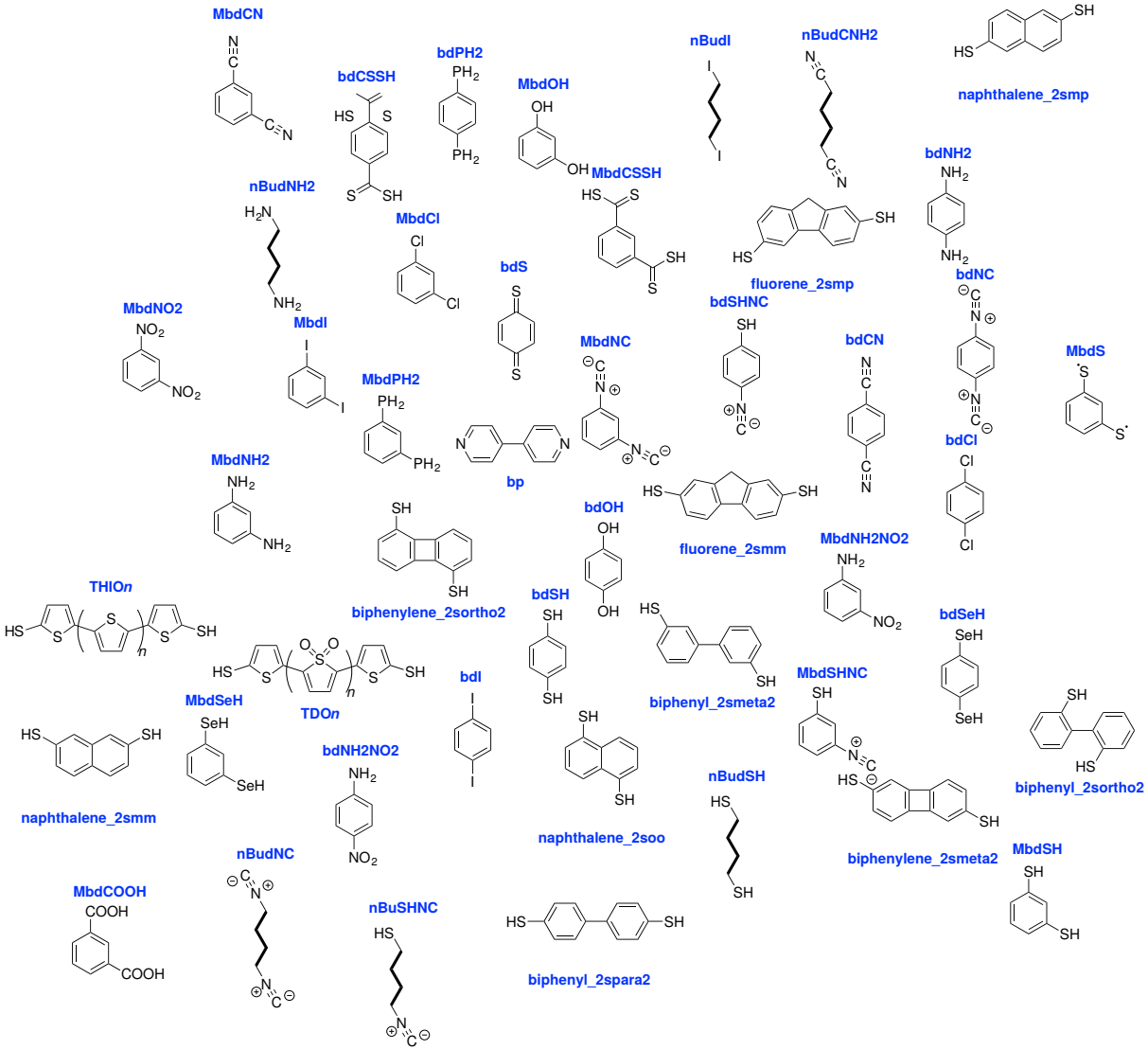
Silicon phase diagram



data from Amsel et at. PRB 2015; C. Pickard; A. Bartók

Clustering and Classification of M&M

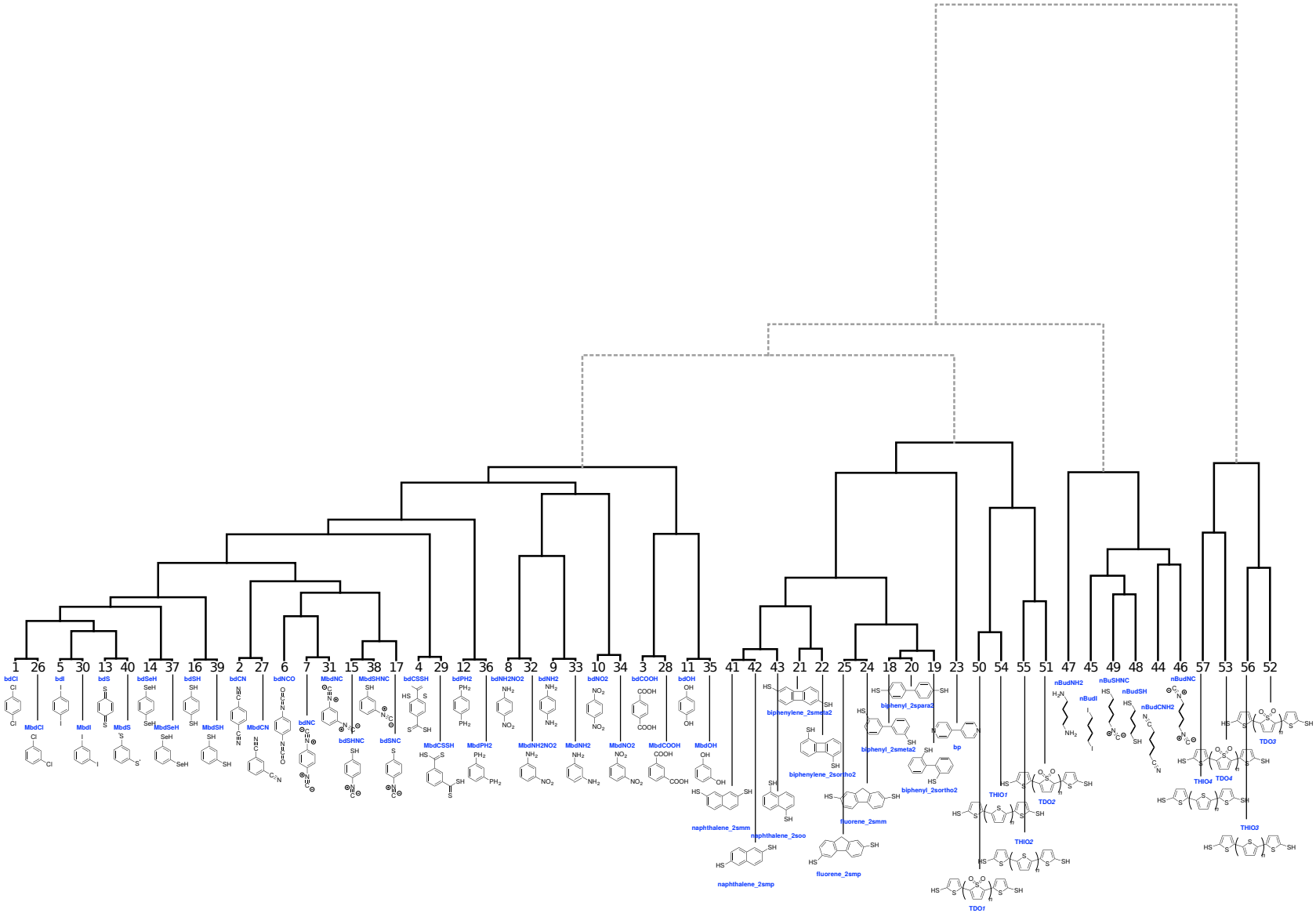
- Automatic classifications of ligands for molecular electronics - hierarchical clustering based on REMatch-SOAP



data from C. Corminboeuf

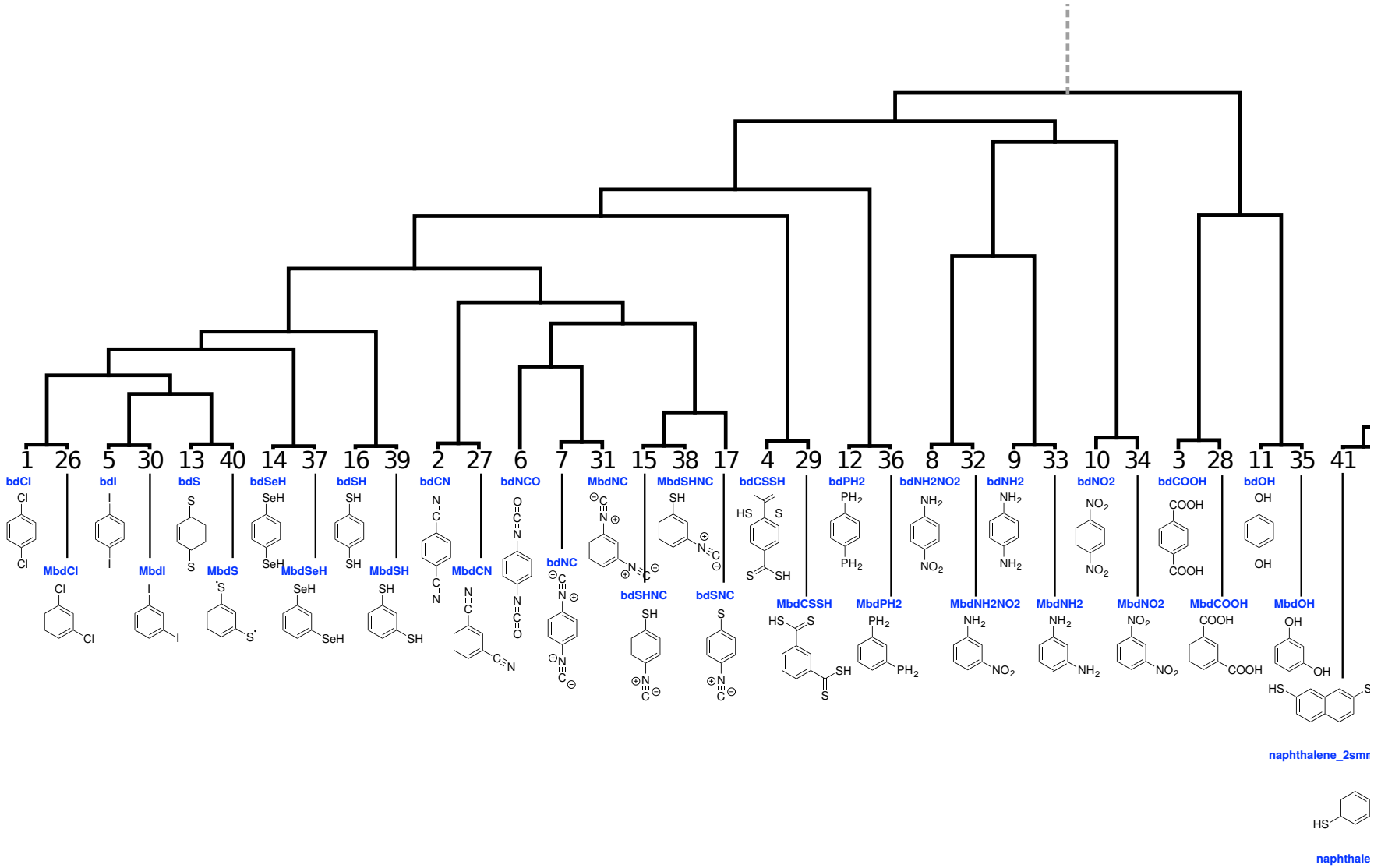
Clustering and Classification of M&M

- Automatic classifications of ligands for molecular electronics - hierarchical clustering based on REMatch-SOAP



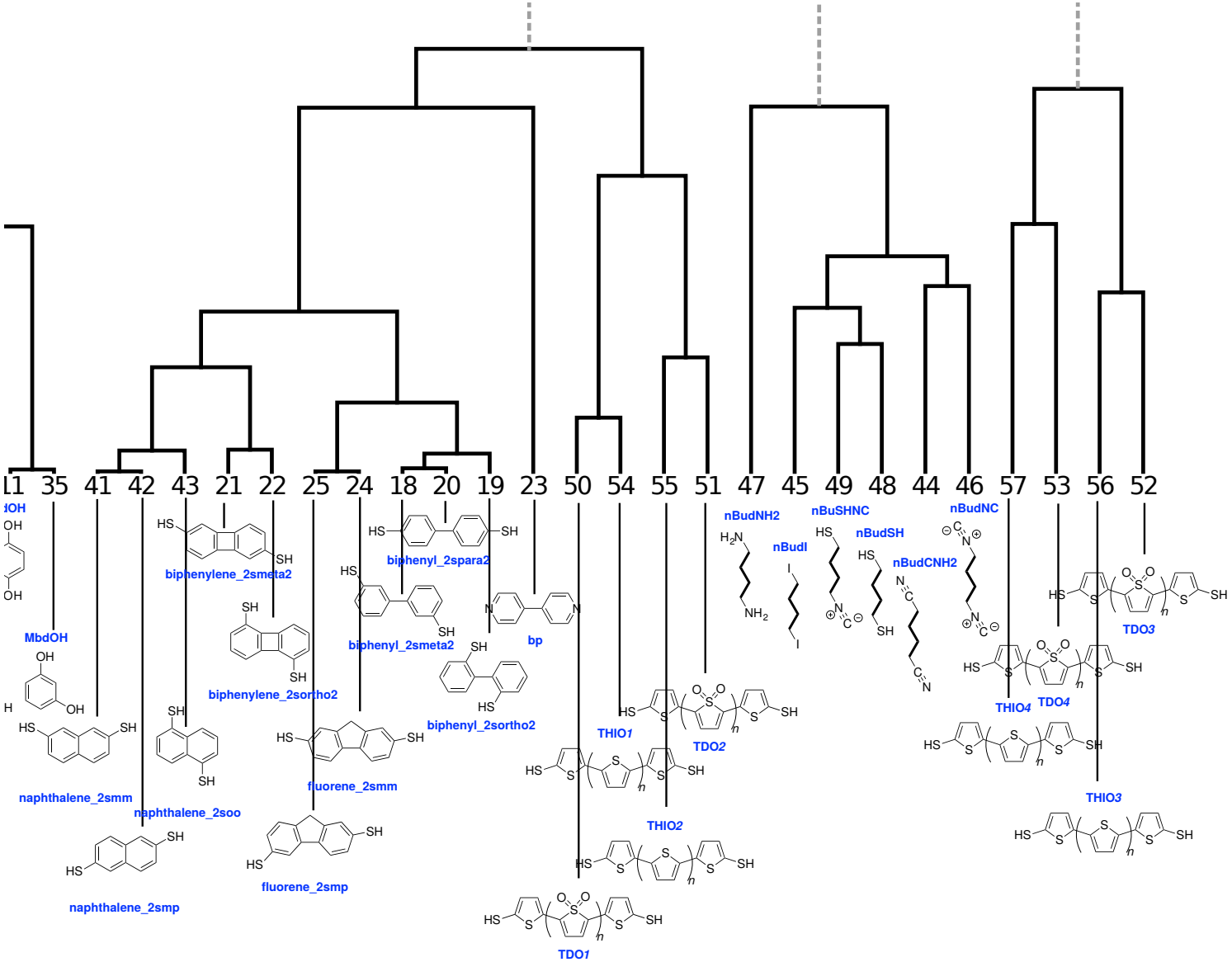
Clustering and Classification of M&M

- Automatic classifications of ligands for molecular electronics - hierarchical clustering based on REMatch-SOAP



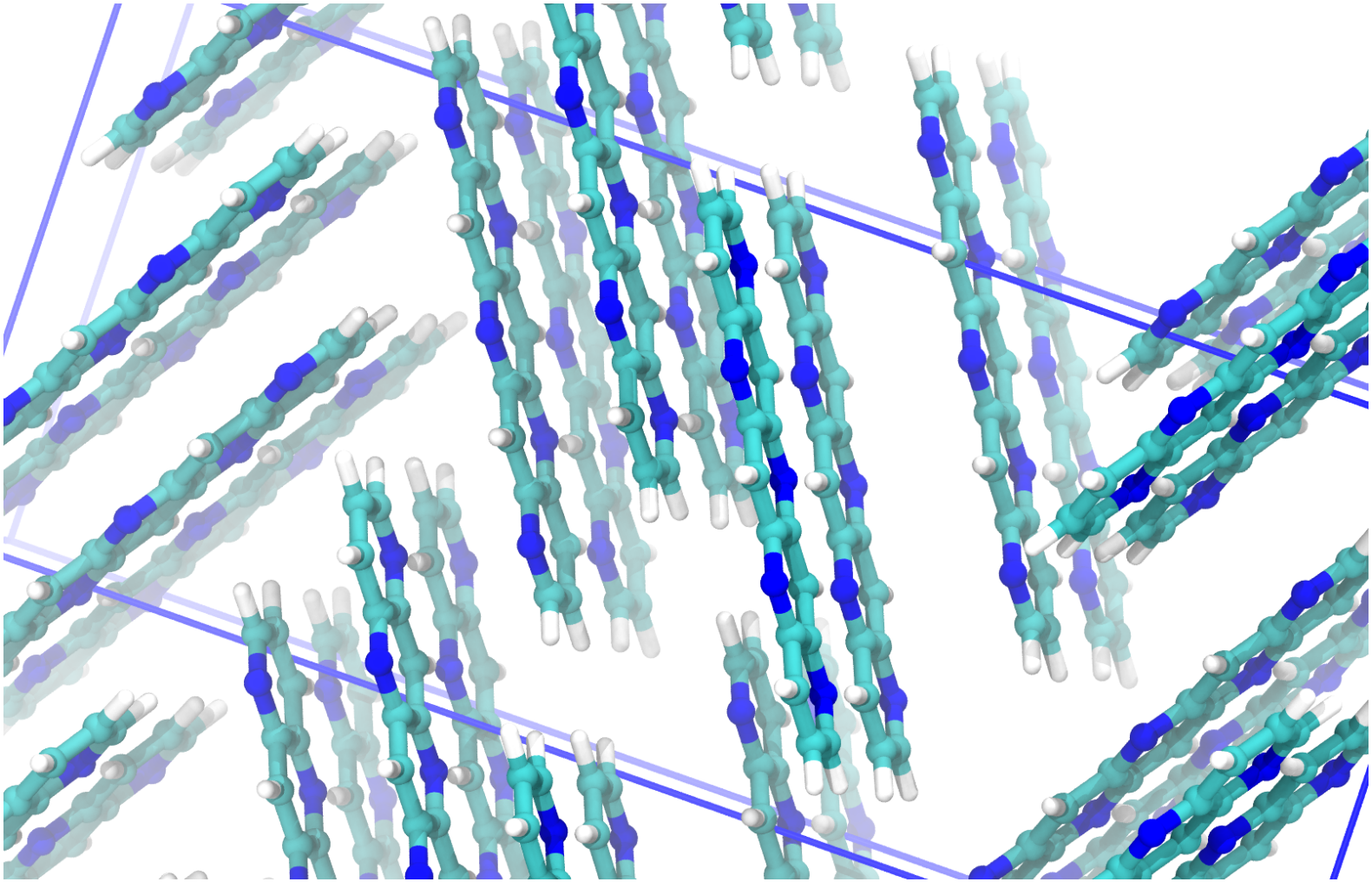
Clustering and Classification of M&M

- Automatic classifications of ligands for molecular electronics - hierarchical clustering based on REMatch-SOAP



Azapentacene - Structure & Properties

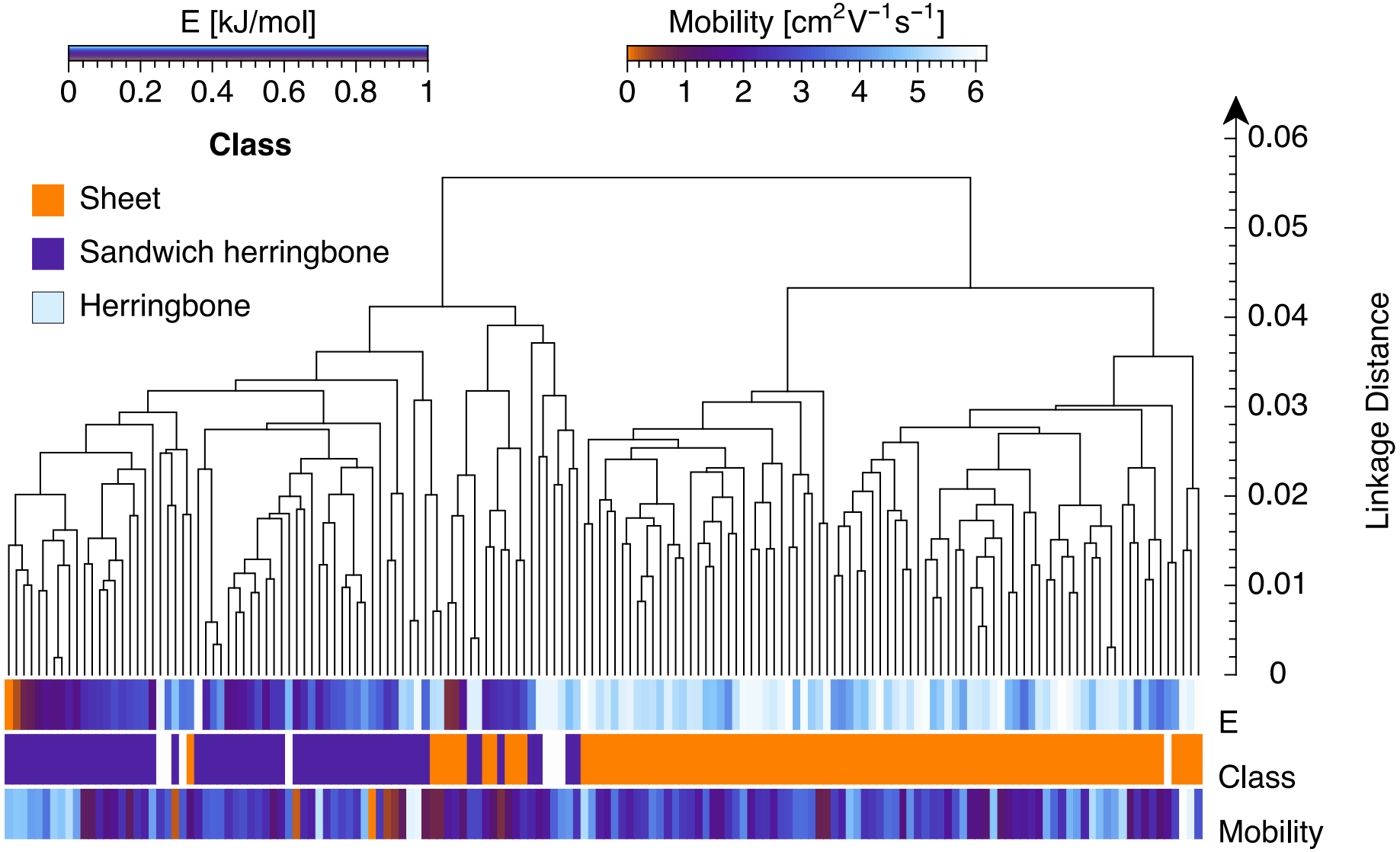
- Azapentacene: a candidate molecular electron transporter
- Dual challenge: enumeration and classification of polymorphs, and prediction of stability and electron mobility



Data from G. Day and J. Yang

Structural Classification

- Clustering/sketch-maps based on REMatch-SOAP correlate well with qualitative classification of packing motifs, and with properties

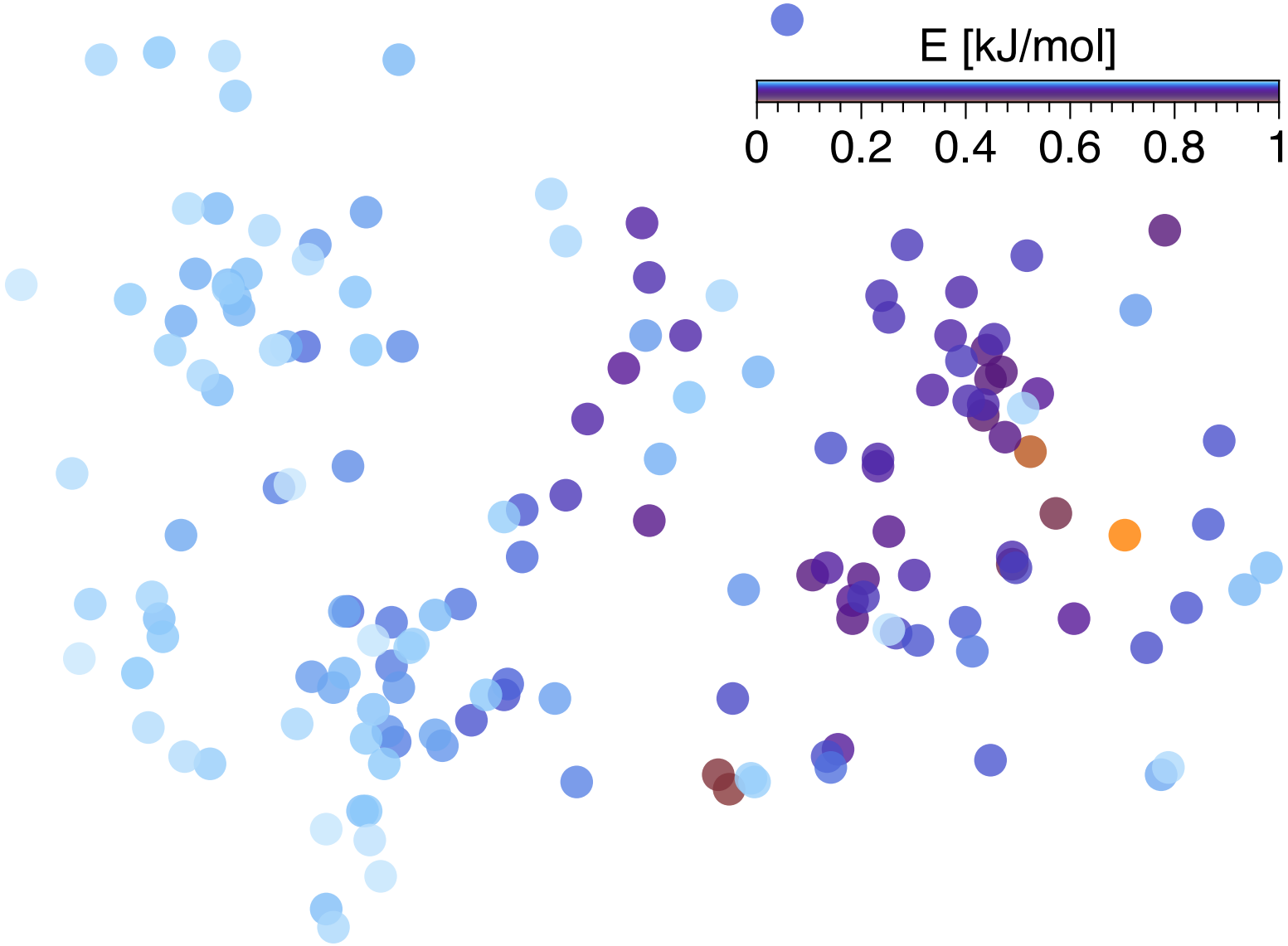


Structural Classification

- Clustering/sketch-maps based on REMatch-SOAP correlate well with qualitative classification of packing motifs, and with properties

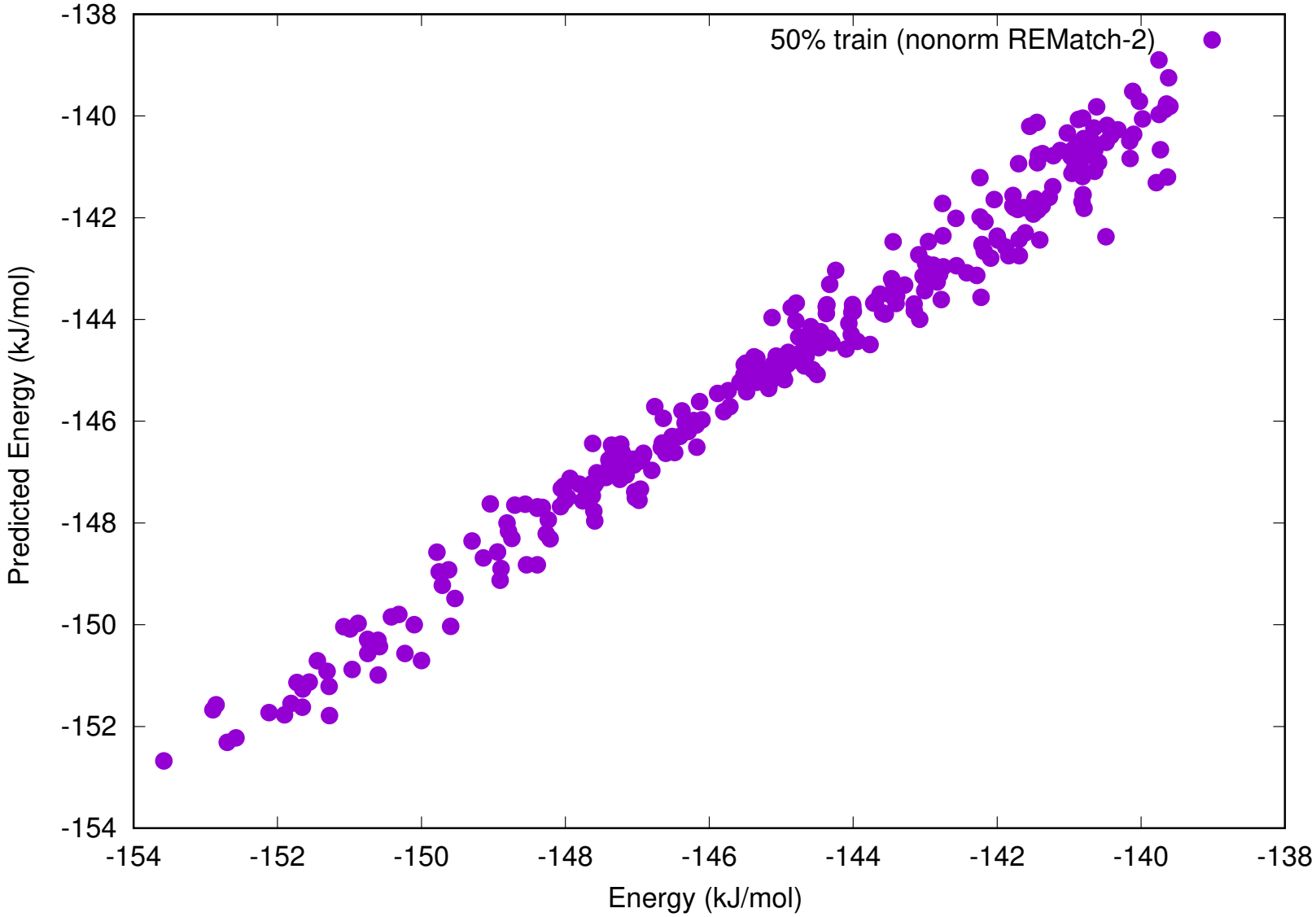
Structural Classification

- Clustering/sketch-maps based on REMatch-SOAP correlate well with qualitative classification of packing motifs, and with properties



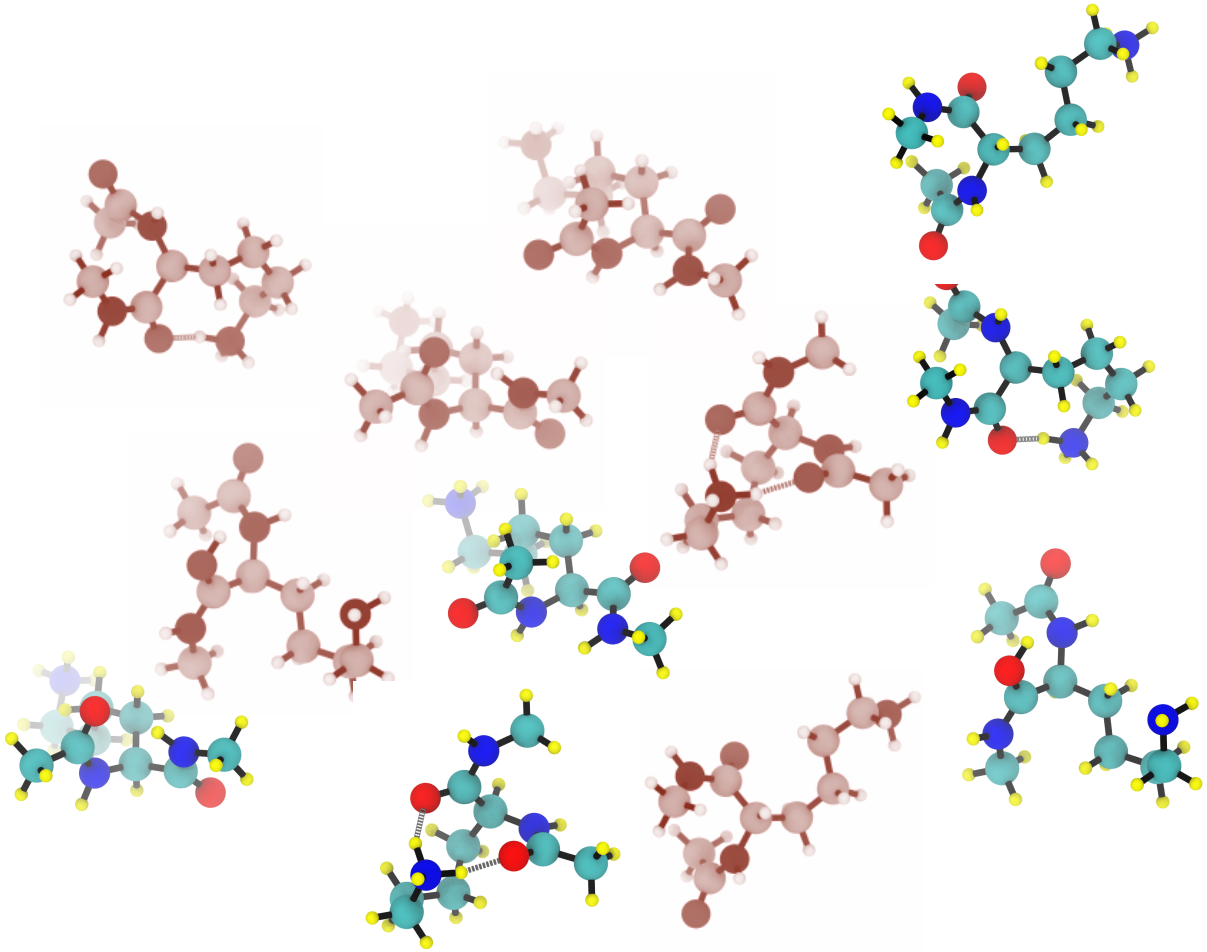
Structural Classification

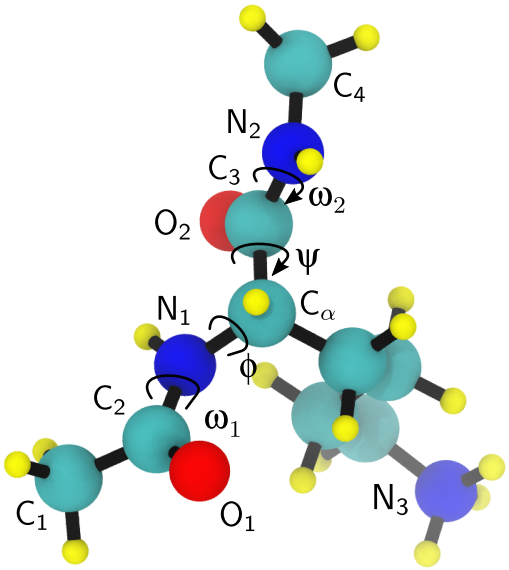
- Clustering/sketch-maps based on REMatch-SOAP correlate well with qualitative classification of packing motifs, and with properties



Structure and stability of oligopeptides

- Representing geometry and stability of (locally) stable conformers in a large database of aminoacids and dipeptides

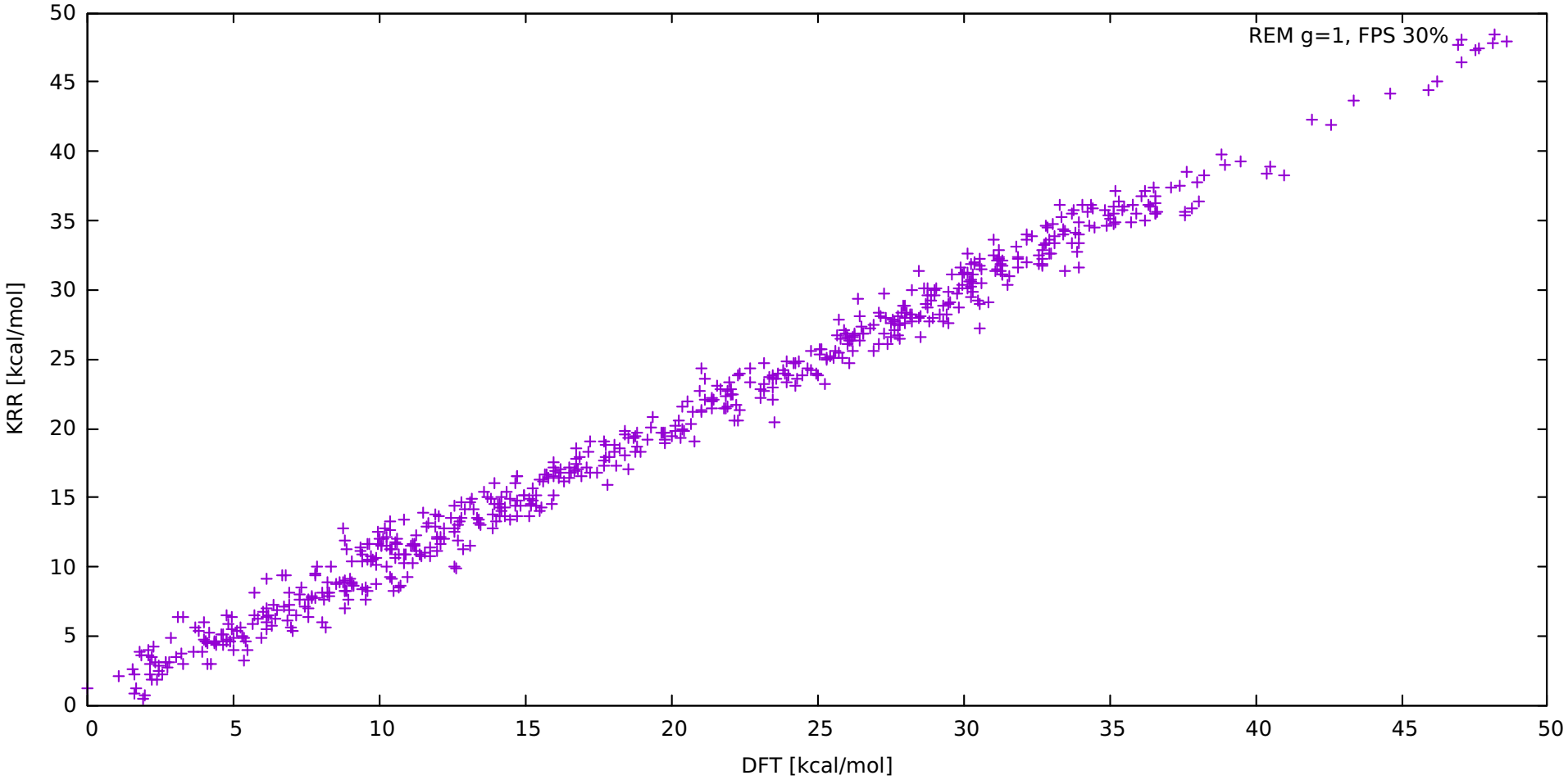




data from: Ropo, Blum, Baldauf, Scientific Data (2016)

Structure and stability of oligopeptides

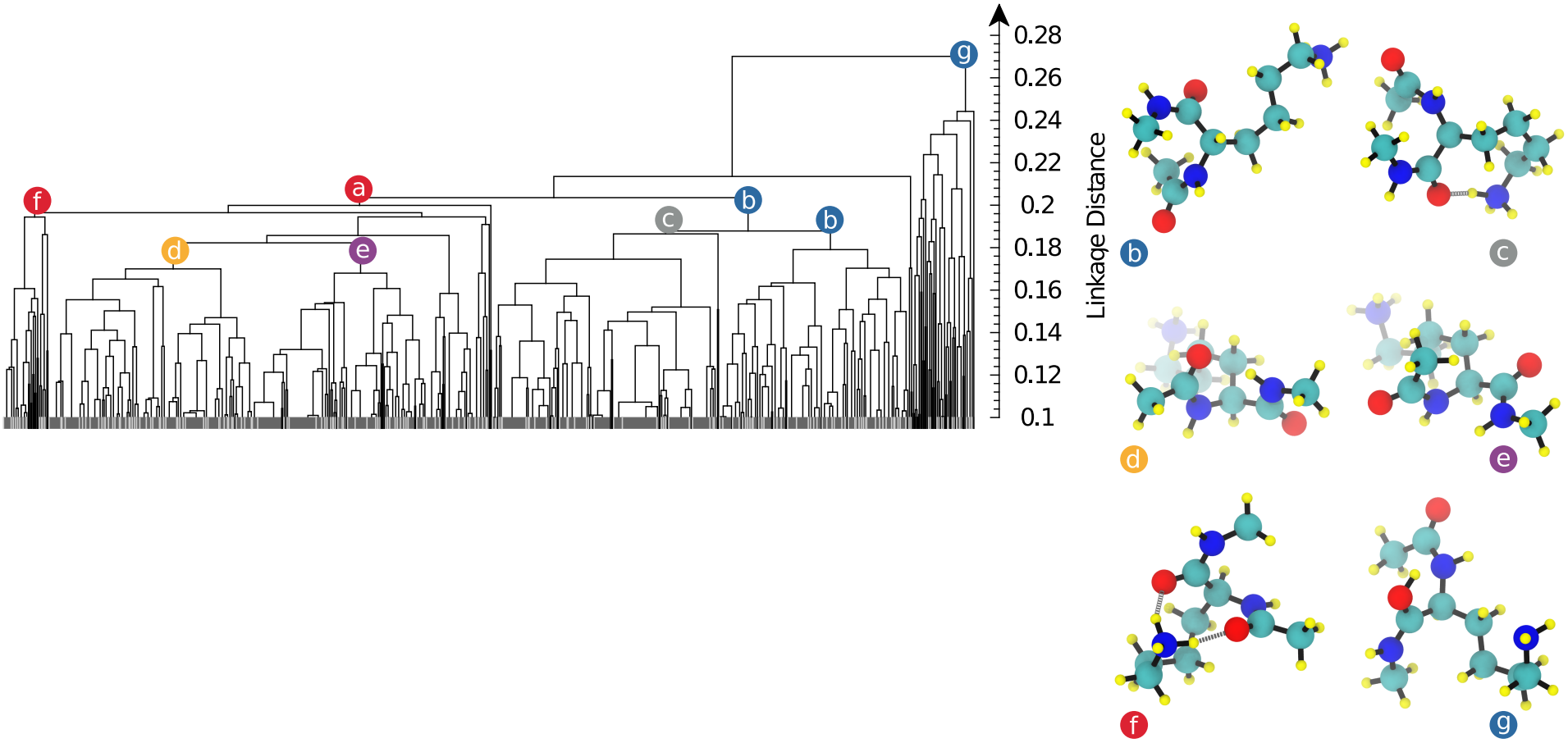
- Representing geometry and stability of (locally) stable conformers in a large database of aminoacids and dipeptides



0.9kcal/mol MAE with 270 structures out of 900

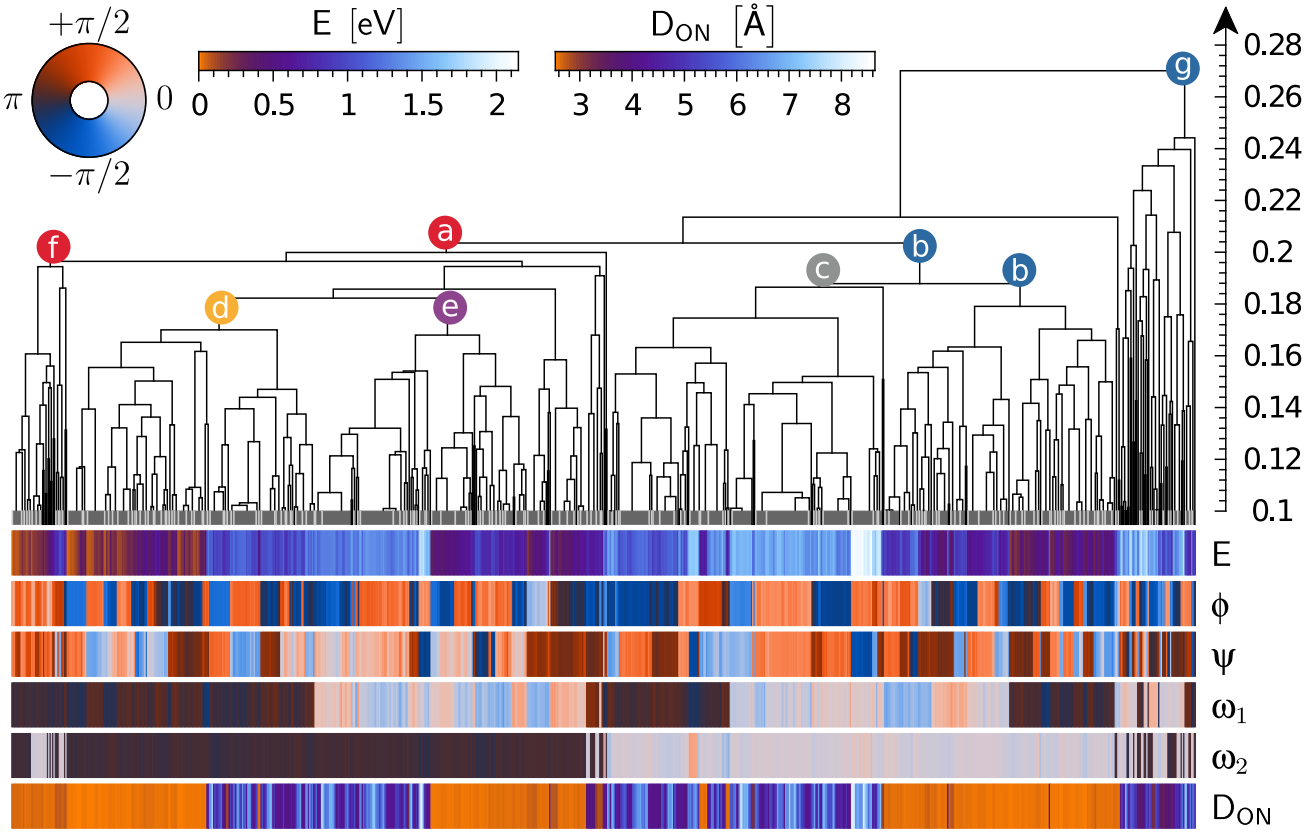
Structure and stability of oligopeptides

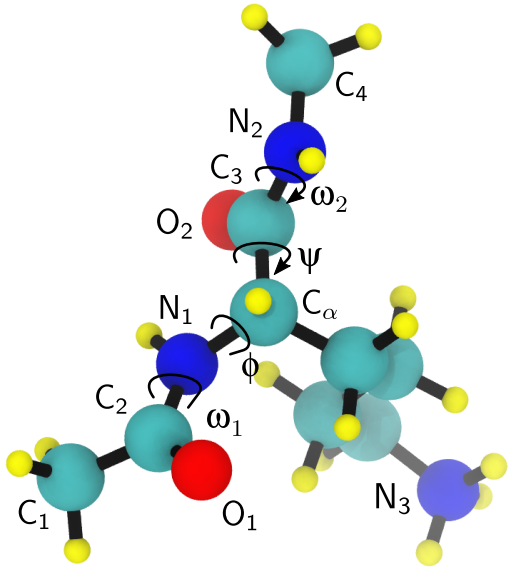
- Off-the-shelf hierarchical clustering shows structure in the dataset. . .
- Clusters clearly correlate with structural parameters (semi-trivial) and properties (non-trivial!)
- Sketch-maps give a more comprehensive picture of the relation between structures in the landscape, and complement nicely a clustering analysis



Structure and stability of oligopeptides

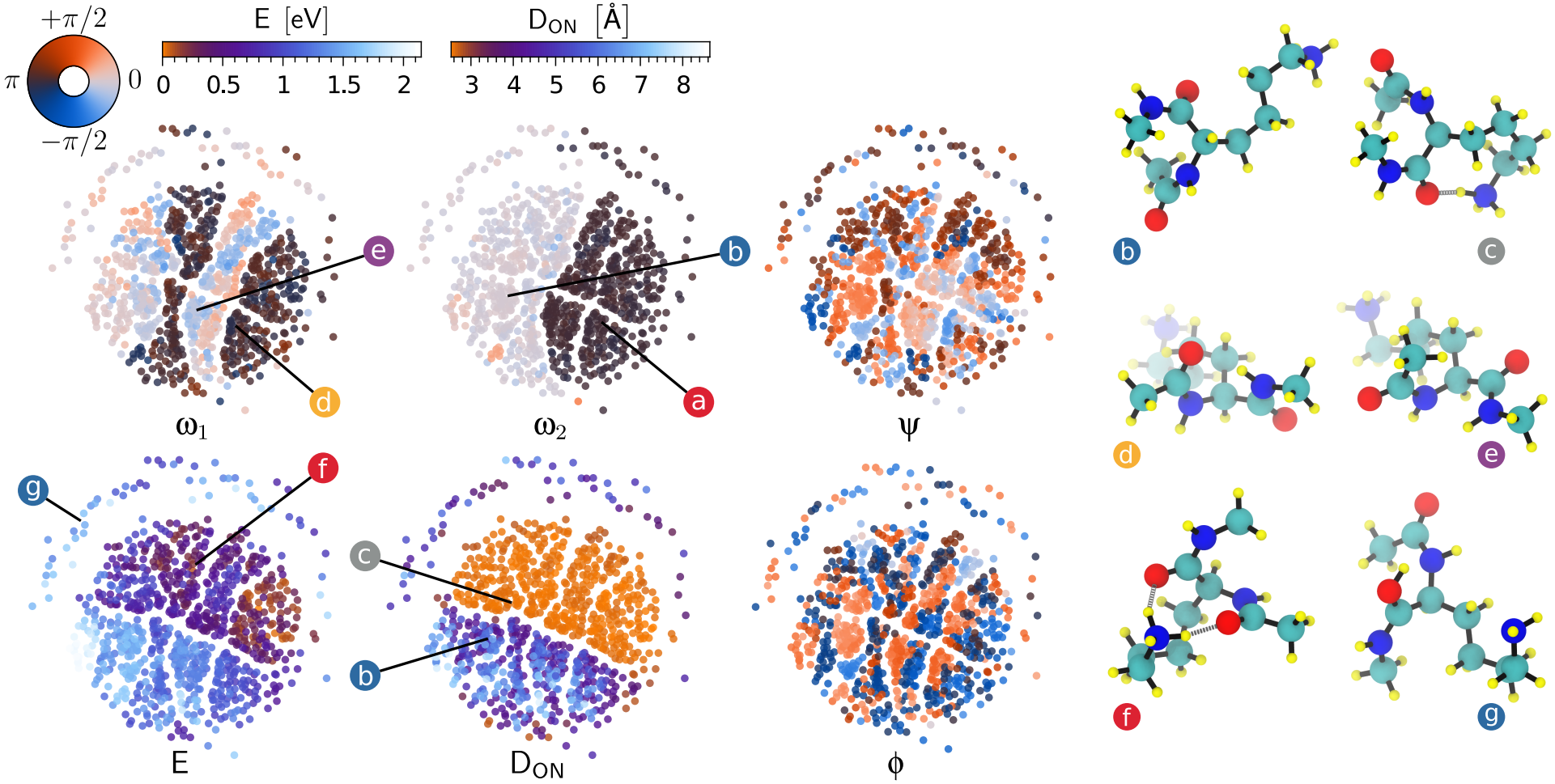
- Off-the-shelf hierarchical clustering shows structure in the dataset. . .
- Clusters clearly correlate with structural parameters (semi-trivial) and properties (non-trivial!)
- Sketch-maps give a more comprehensive picture of the relation between structures in the landscape, and complement nicely a clustering analysis





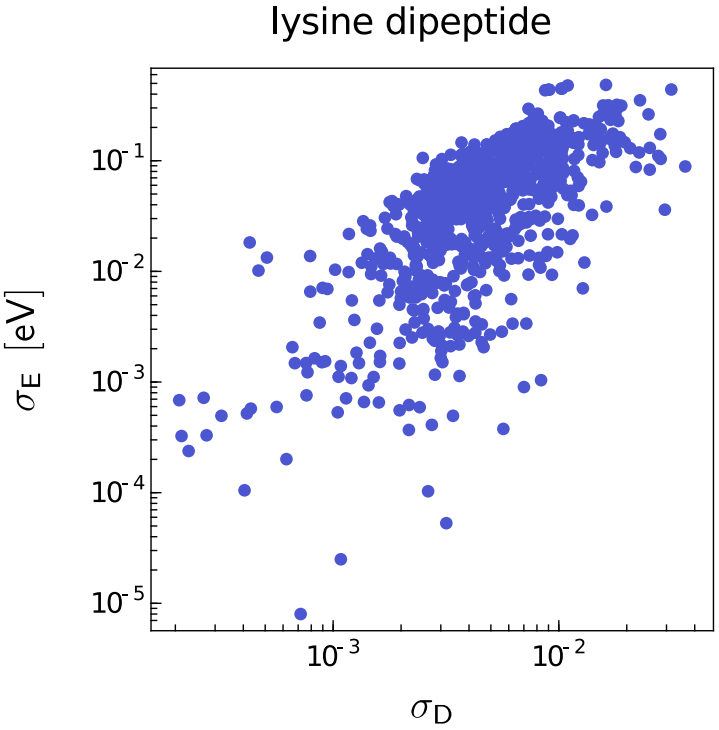
Structure and stability of oligopeptides

- Off-the-shelf hierarchical clustering shows structure in the dataset. . .
- Clusters clearly correlate with structural parameters (semi-trivial) and properties (non-trivial!)
- Sketch-maps give a more comprehensive picture of the relation between structures in the landscape, and complement nicely a clustering analysis



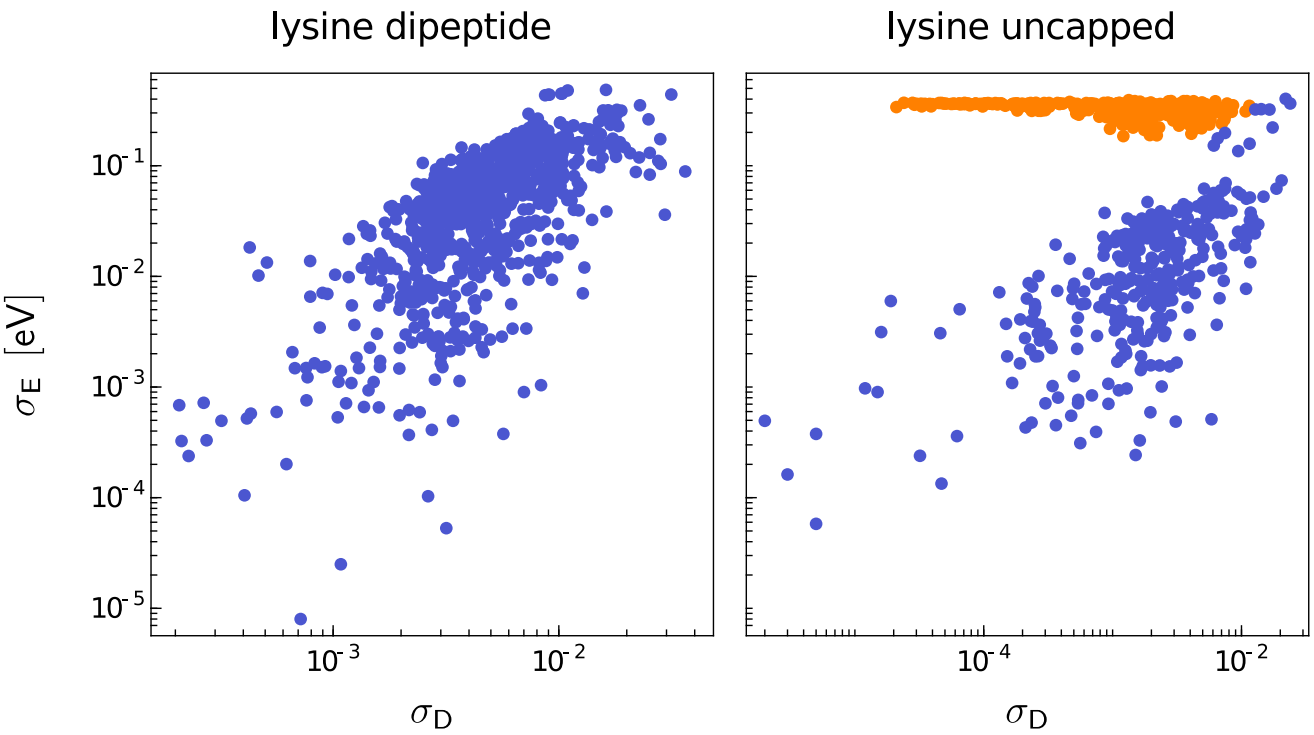
Structure and stability of oligopeptides

- Checking for data integrity in databases is going to be one of the challenges in pushing high-throughput studies
- Analysis of similarity makes it possible to detect duplicates. Distances within clusters should roughly correlate with spread of properties.
- Lack of correlations is a sign of inconsistency. We could detect and resolve half a dozen instances within the aminoacids database



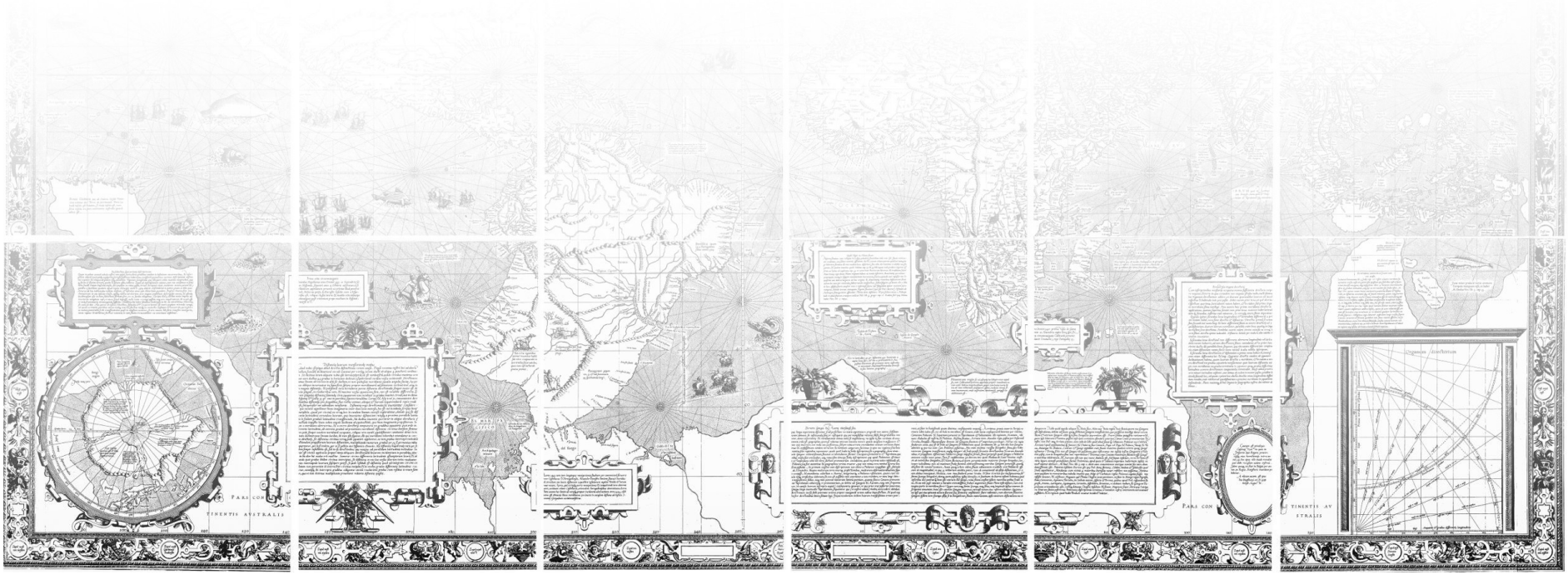
Structure and stability of oligopeptides

- Checking for data integrity in databases is going to be one of the challenges in pushing high-throughput studies
- Analysis of similarity makes it possible to detect duplicates. Distances within clusters should roughly correlate with spread of properties.
- Lack of correlations is a sign of inconsistency. We could detect and resolve half a dozen instances within the aminoacids database



Outlook

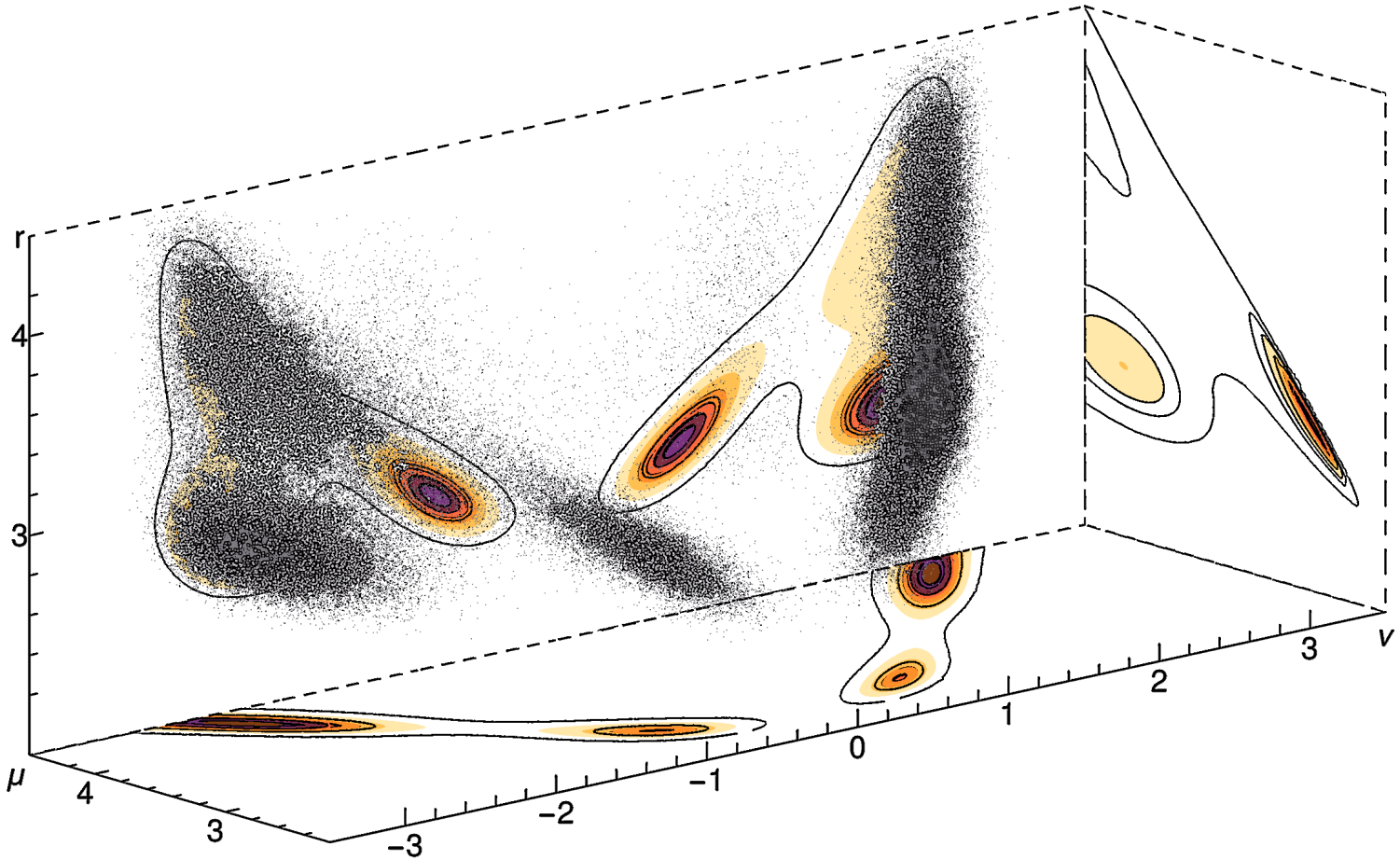
- A comprehensive framework for the analysis of atomistic simulations
  - A similarity kernel based on a combination of environment descriptors
  - Recognizing recurring patterns by a probabilistic analysis
  - Non-linear mapping of complex (free)-energy landscapes
- The same framework can be used for molecules and materials, to predict properties, represent intuitively databases, detect outliers and resolve data inconsistencies
- (Development) code available on http://epfl-cosmo.github.io



http://sketchmap.org/

Probabilistic Analysis of Molecular Motifs

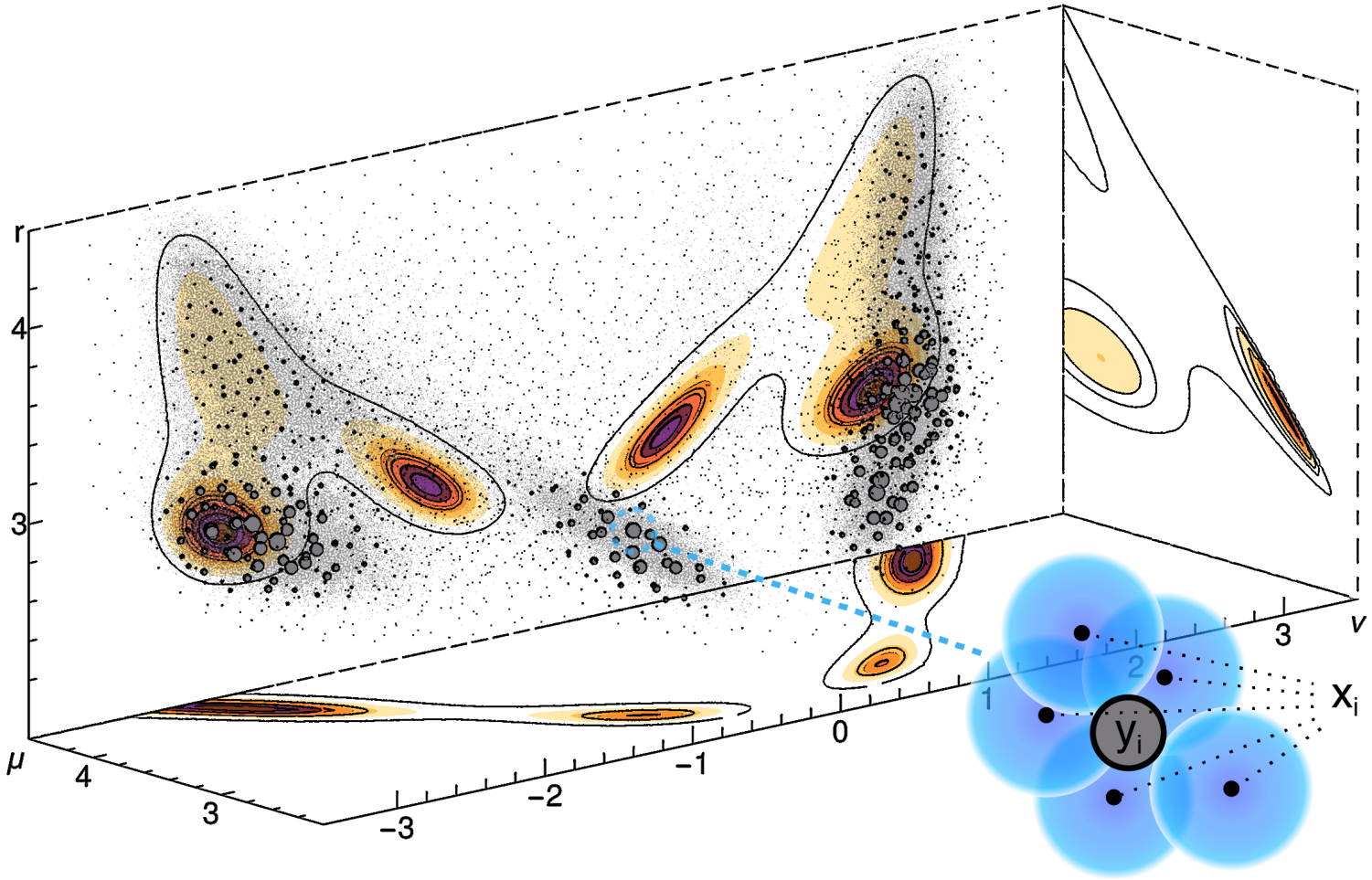
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

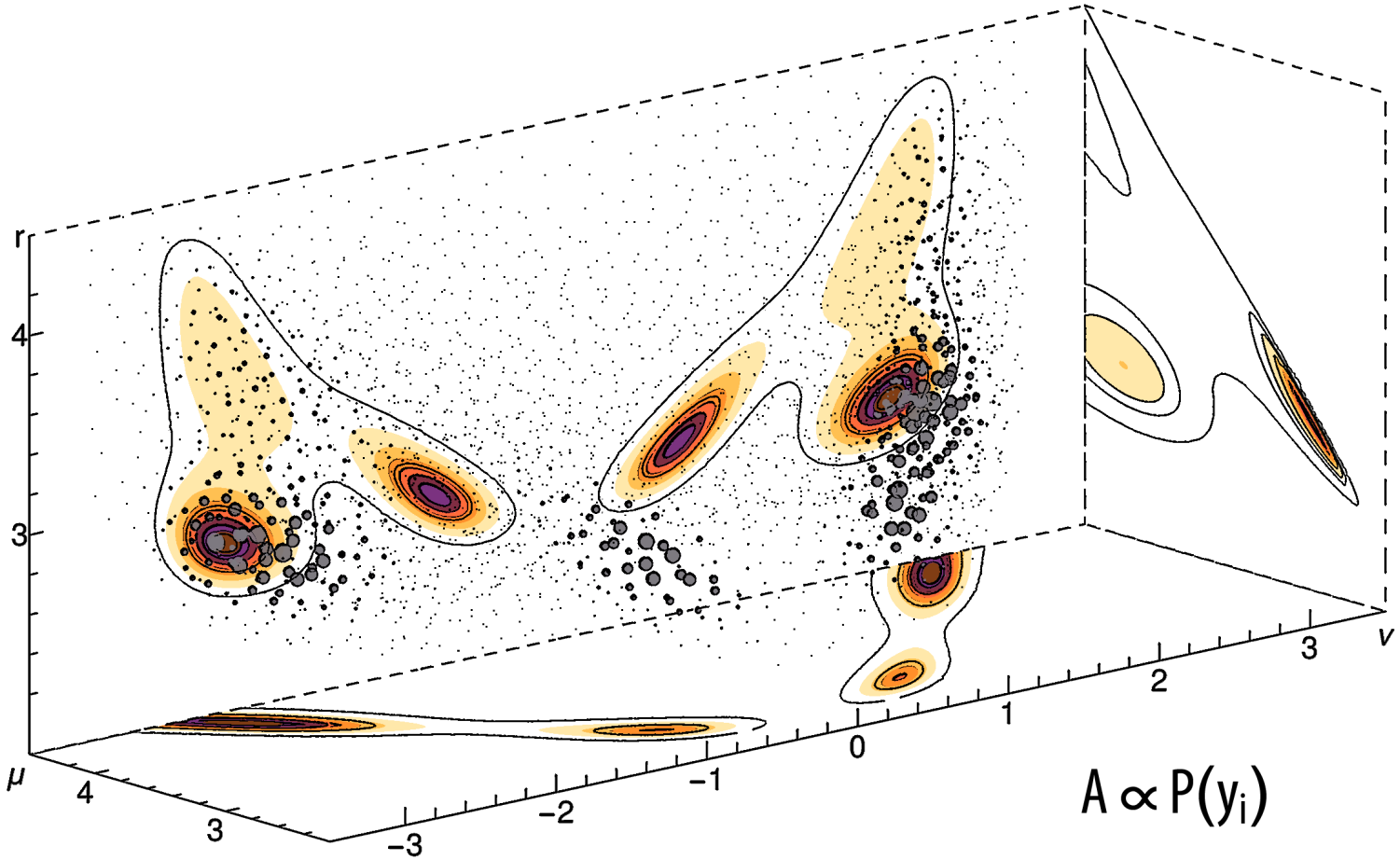
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

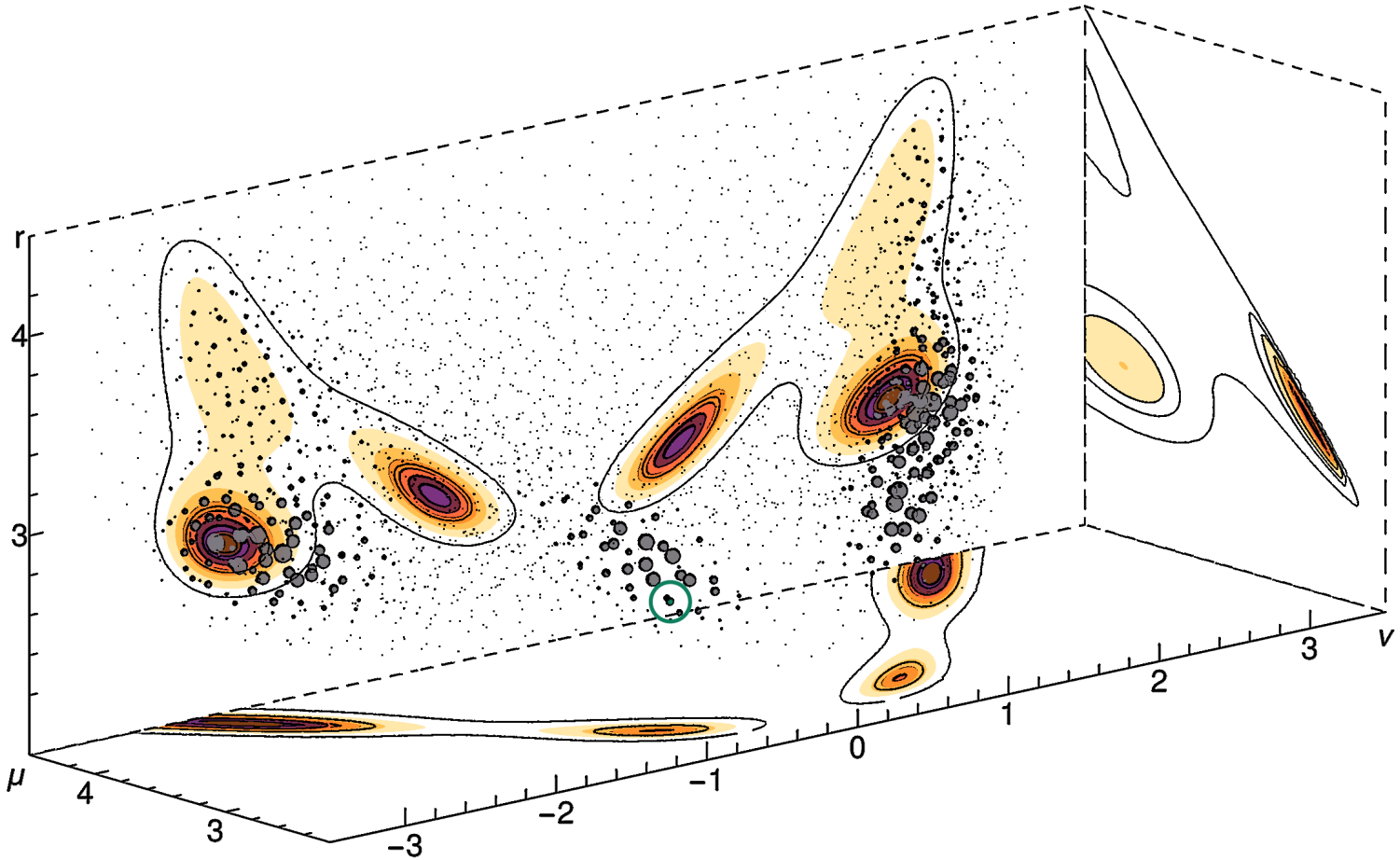
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

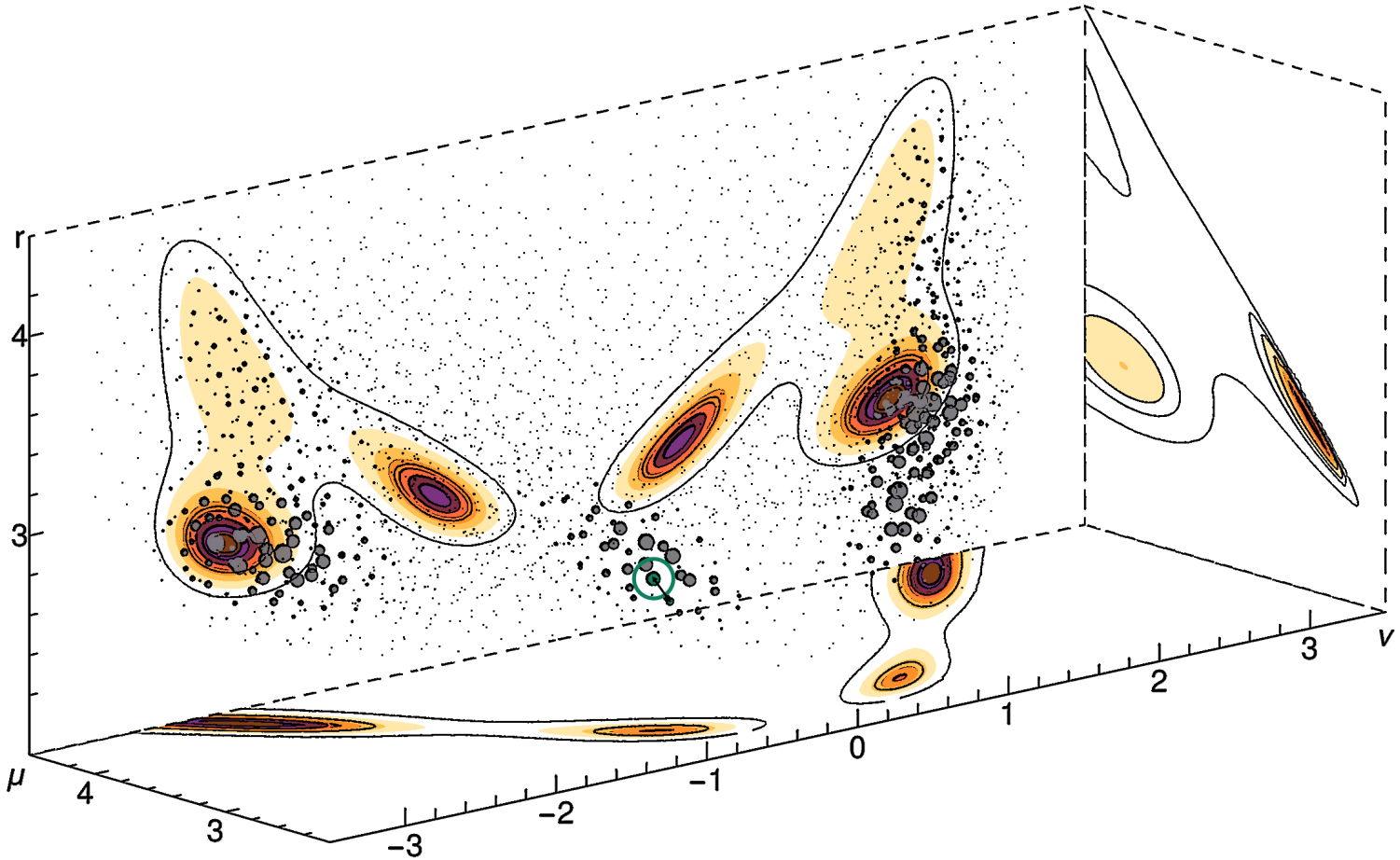
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

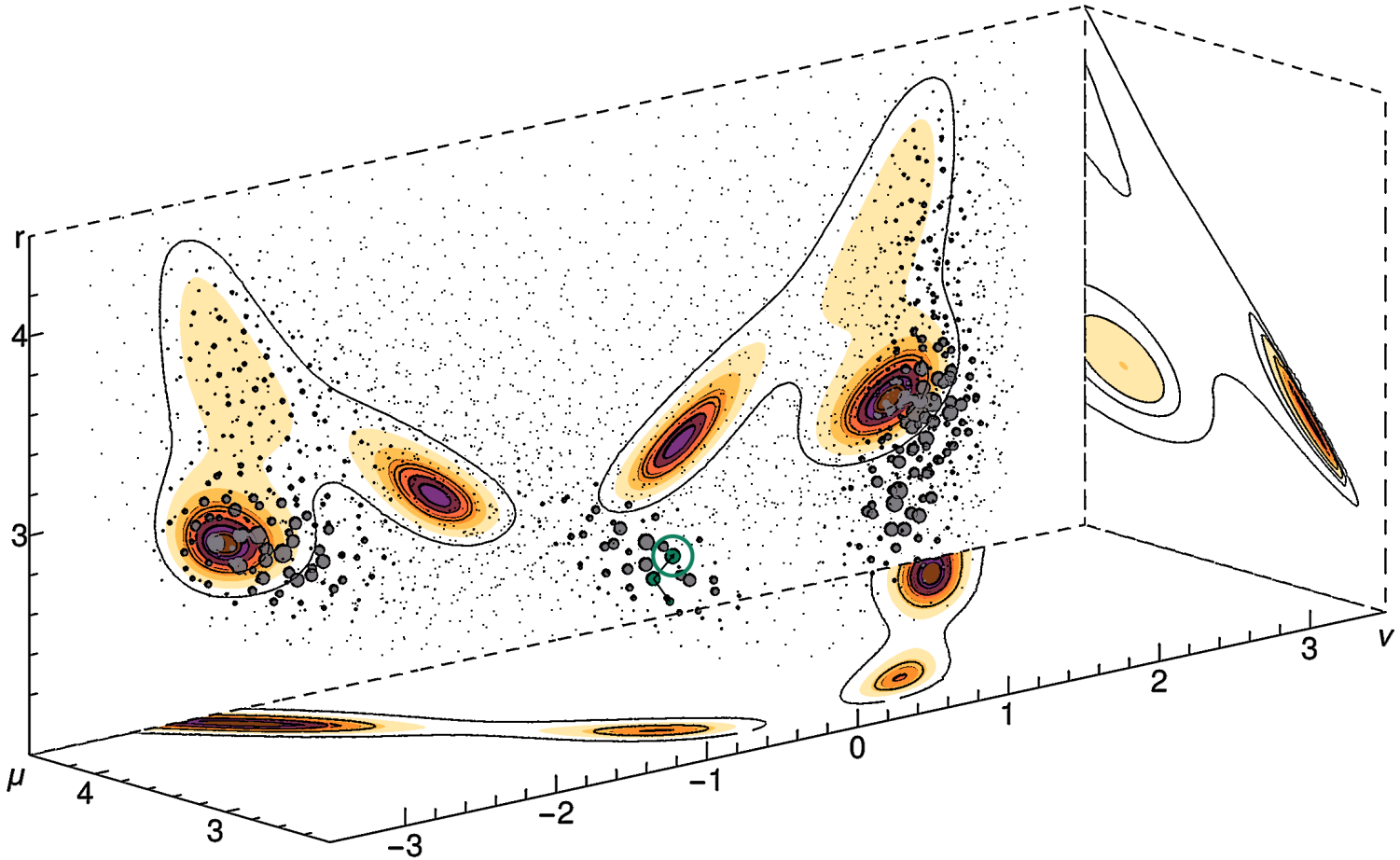
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

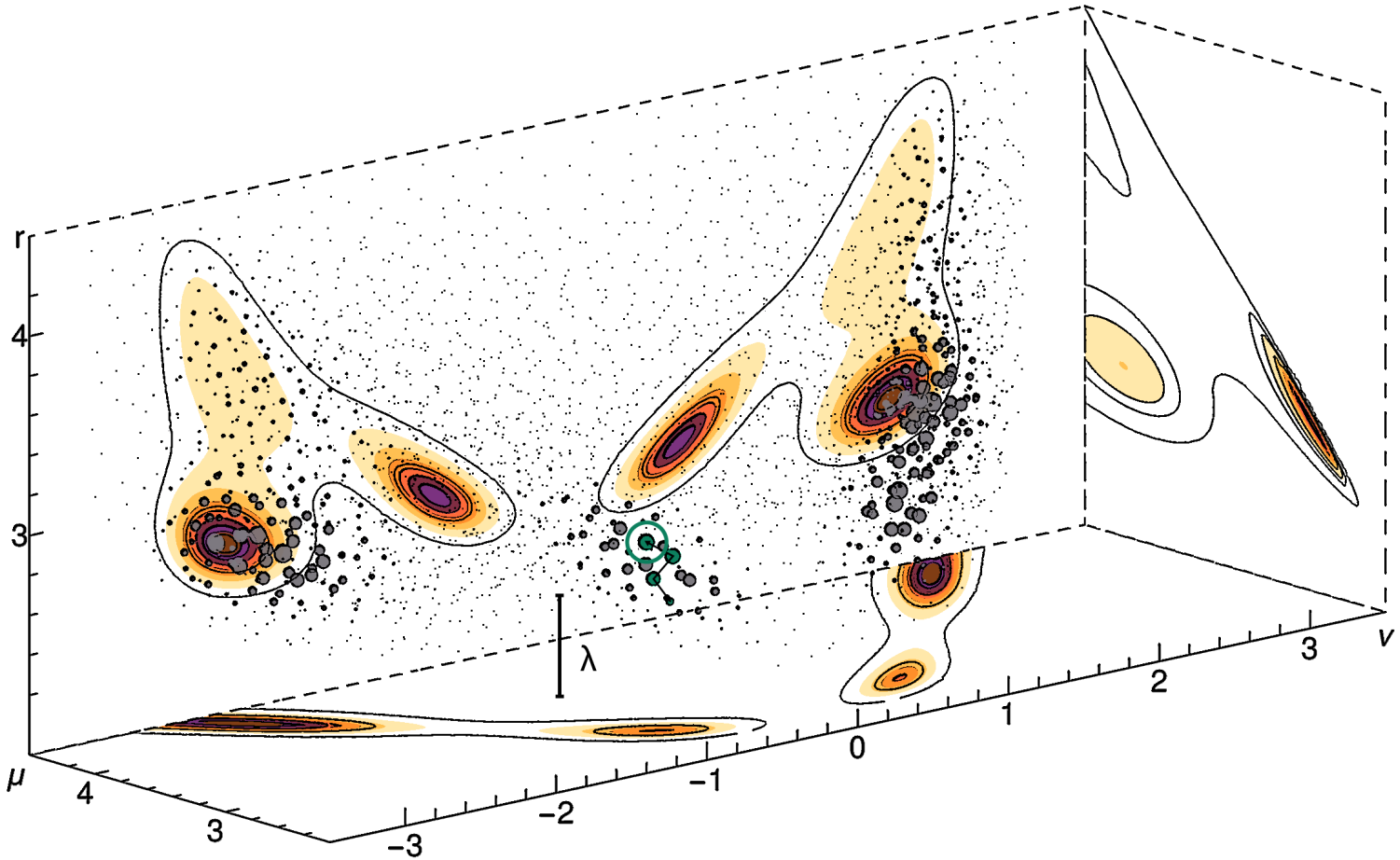
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

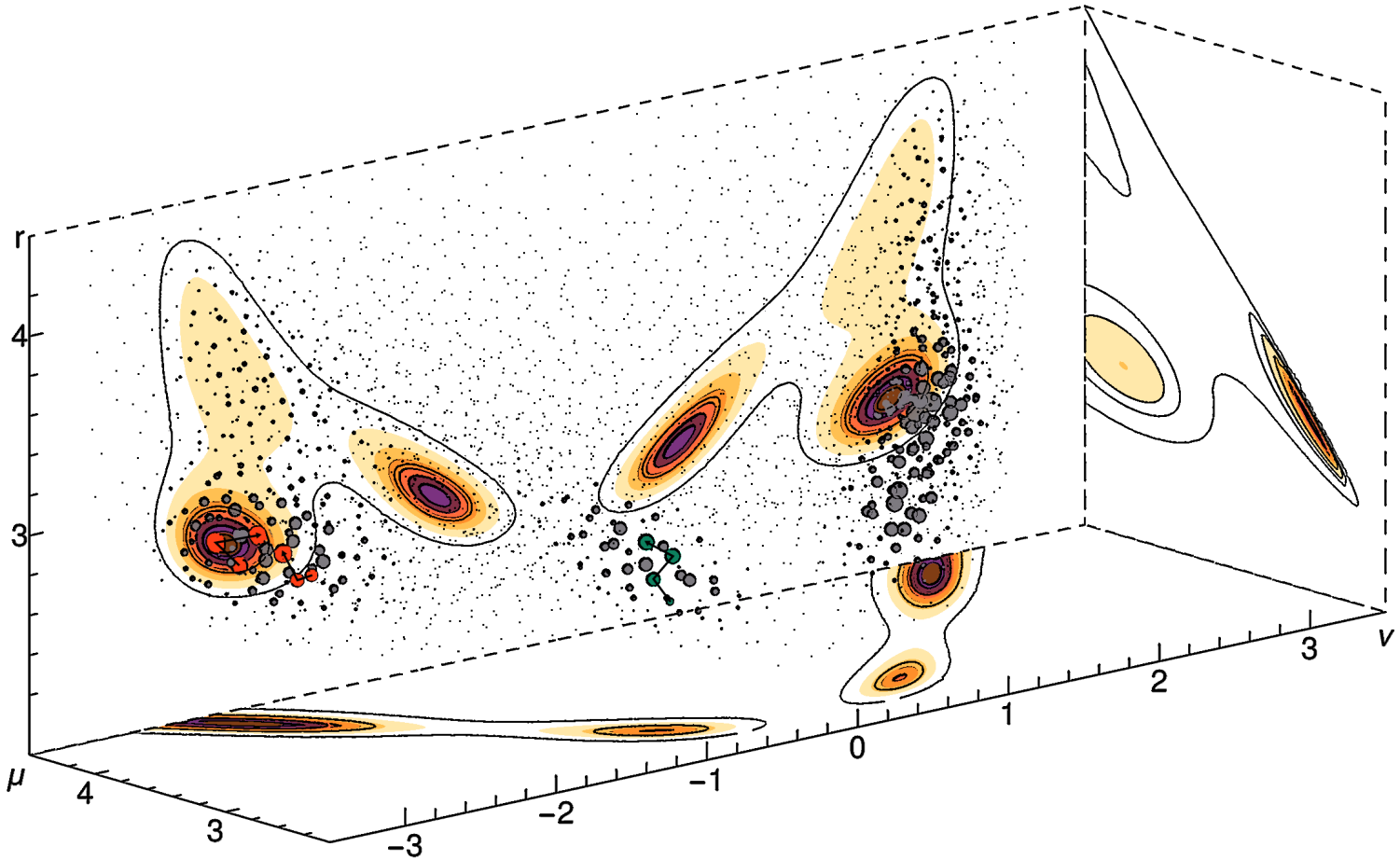
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

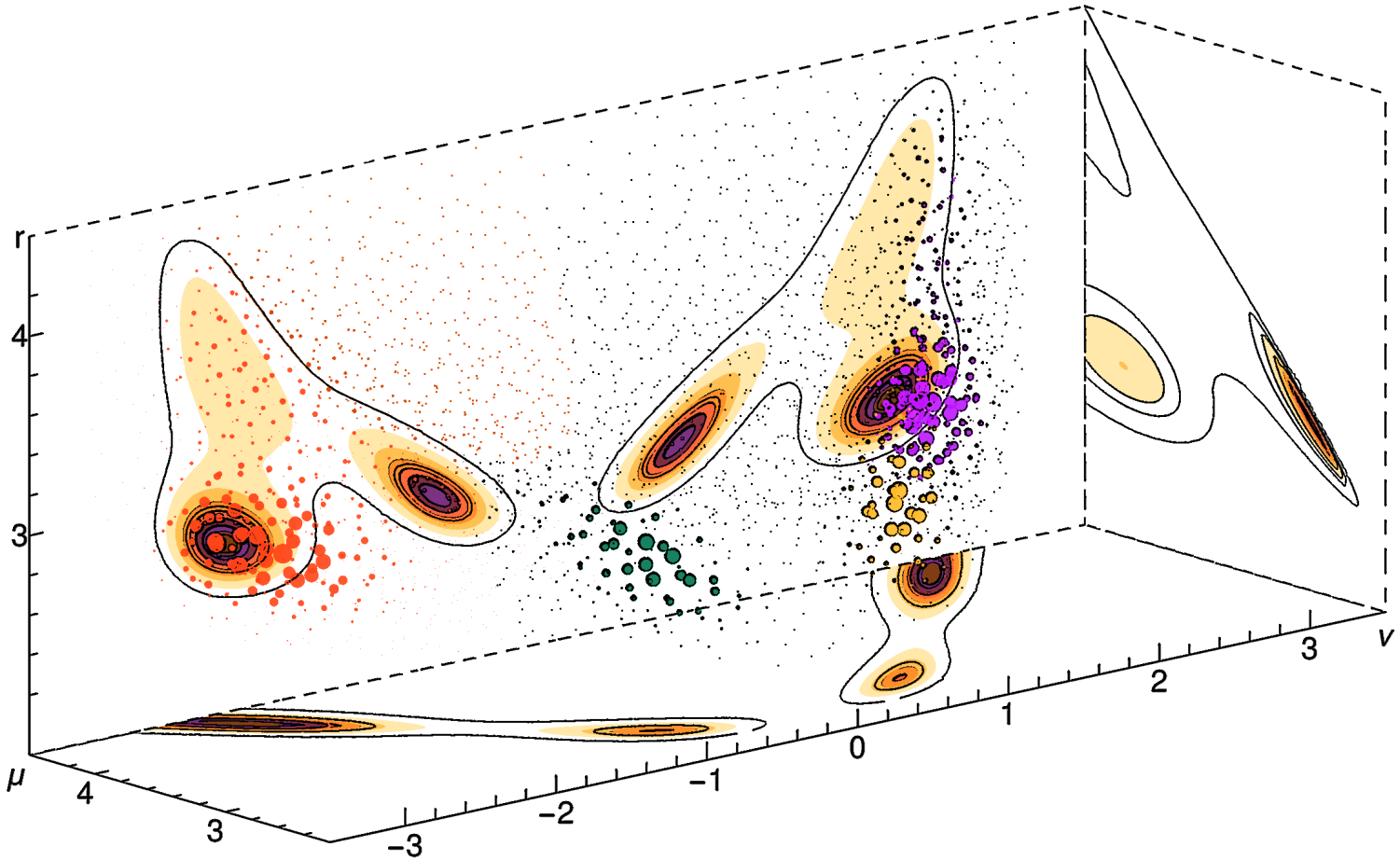
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

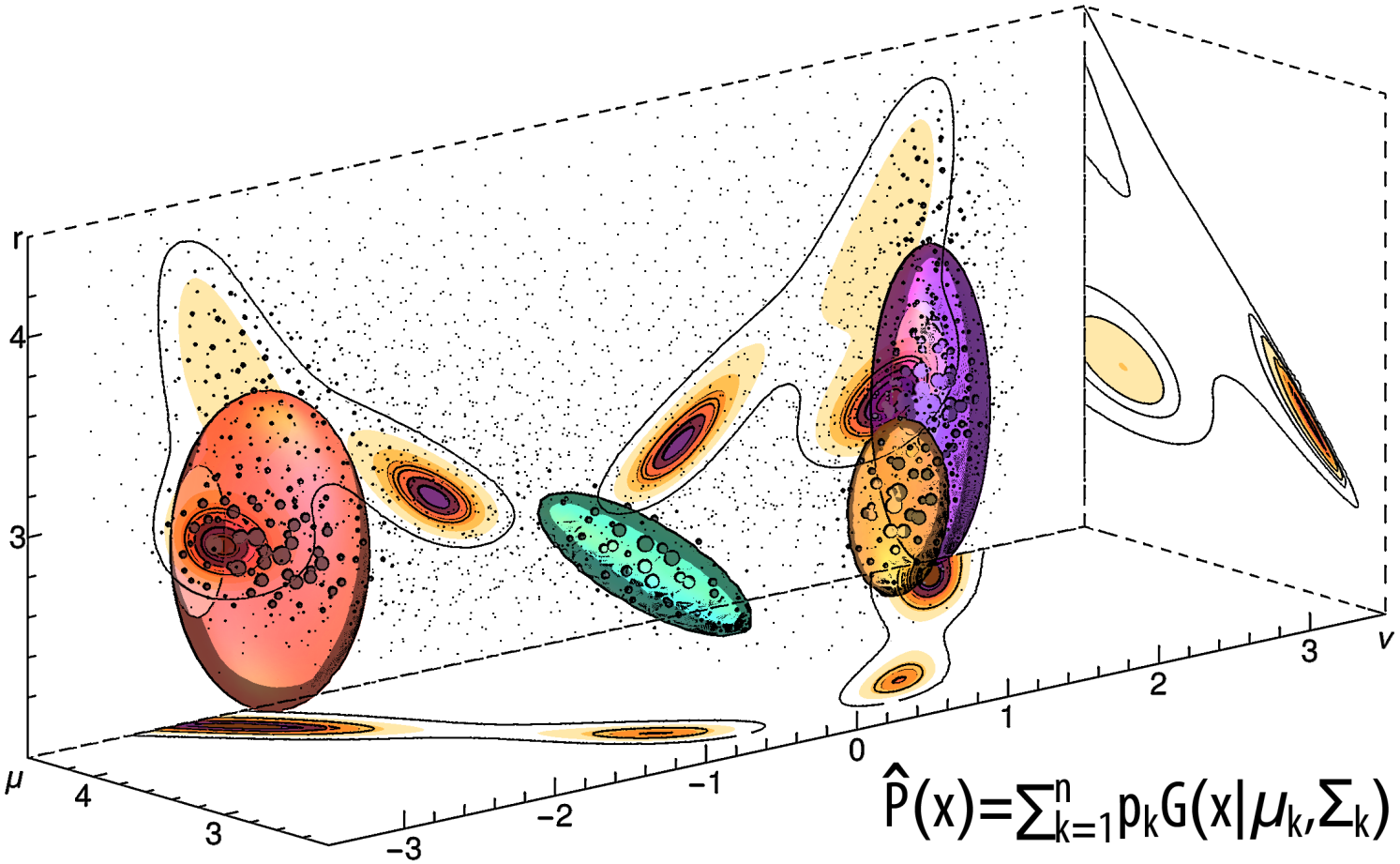
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space

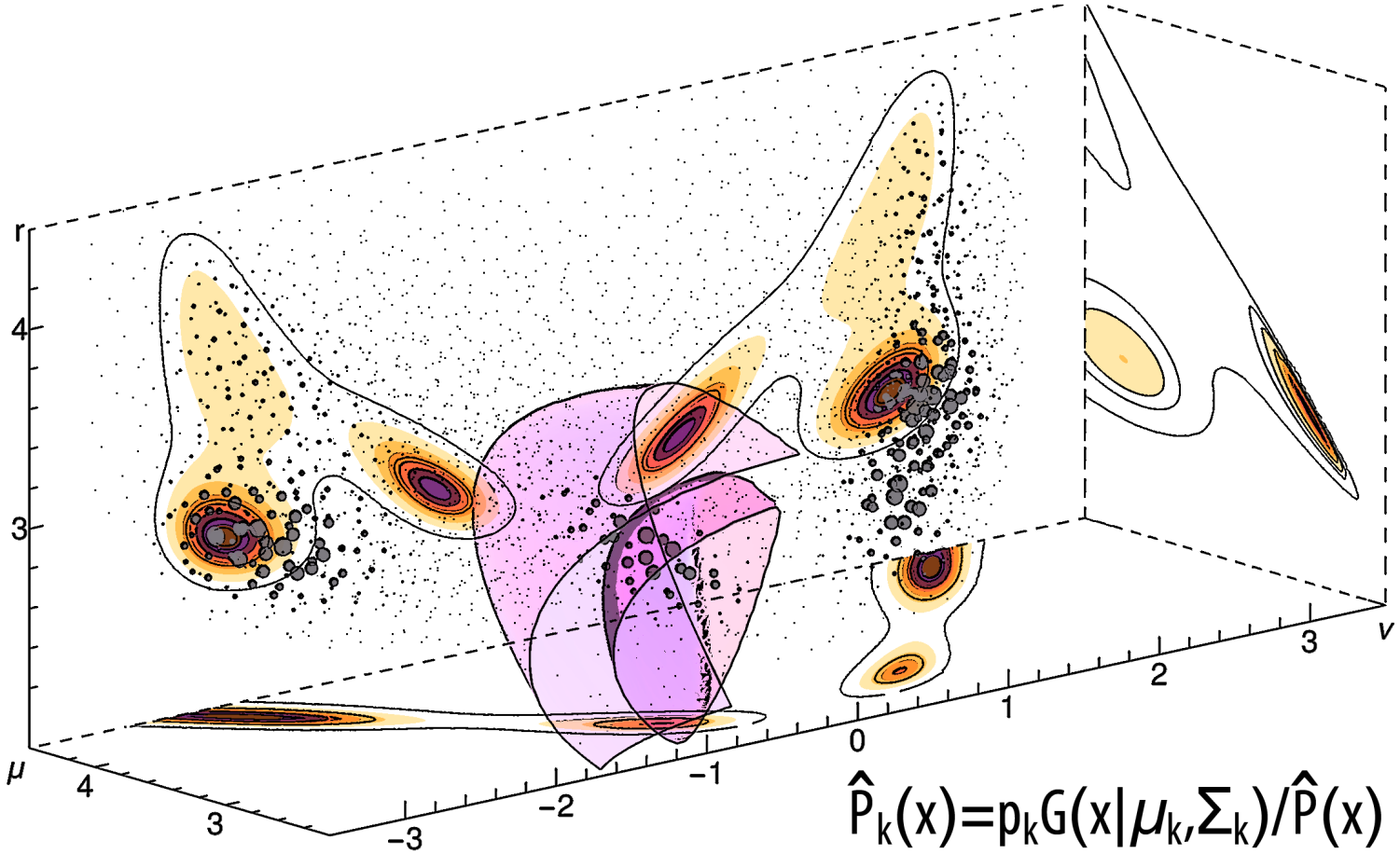


$$\hat{P}(x)=\sum_{k=1}^{n} p_k G(x|\mu_k,\Sigma_k)$$

Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)

Probabilistic Analysis of Molecular Motifs

- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy, continuous partitioning of configuration space



$$\hat{P}_k(x) = p_k G(x|\mu_k, \Sigma_k)/\hat{P}(x)$$

Piero Gasparotto & Michele Ceriotti, JCP 174110, 141 (2014)