Finding Patterns and Drawing Maps in the Configuration Space of Materials and Molecules

Michele Ceriotti, EPFL-IMX-COSMO

ML 2015 - IPAM Los Angeles

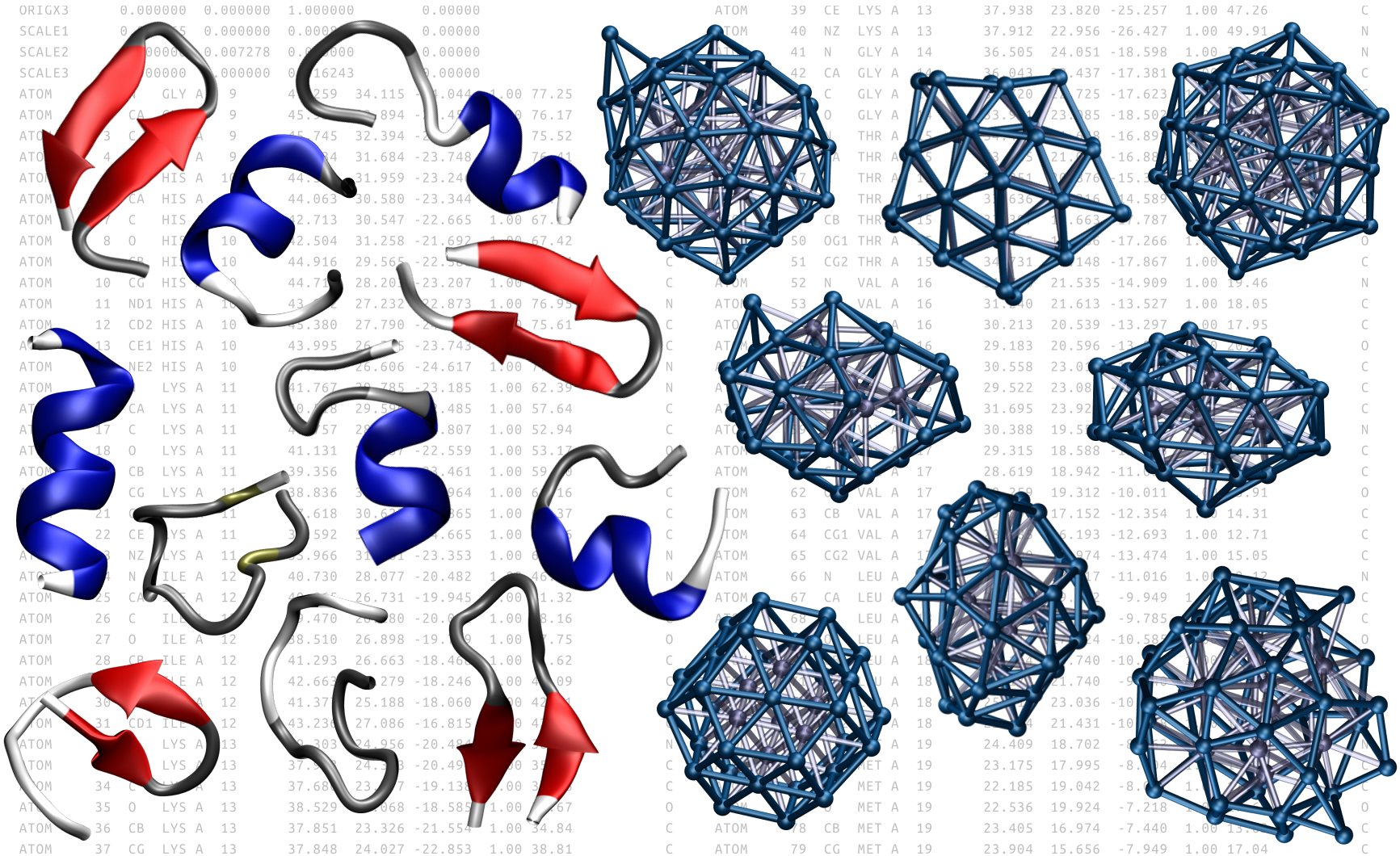
Managing data from atomistic models

- Atomistic simulations provide **too much** information
- It is hard to decipher the essential features in structurally-complex compounds, materials, proteins, etc.

```
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.008865  0.000000  0.000854        0.00000
SCALE2      0.000000  0.007278  0.000000        0.00000
SCALE3      0.000000  0.000000  0.016243        0.00000
ATOM      1  N   GLY A   9      47.259  34.115 -24.044  1.00 77.25           N
ATOM      2  CA  GLY A   9      45.958  33.894 -23.454  1.00 76.17           C
ATOM      3  C   GLY A   9      45.745  32.394 -23.509  1.00 75.52           C
ATOM      4  O   GLY A   9      46.734  31.684 -23.748  1.00 76.41           O
ATOM      5  N   HIS A  10      44.513  31.959 -23.246  1.00 73.46           N
ATOM      6  CA  HIS A  10      44.063  30.580 -23.344  1.00 70.95           C
ATOM      7  C   HIS A  10      42.713  30.547 -22.665  1.00 67.35           C
ATOM      8  O   HIS A  10      42.504  31.258 -21.692  1.00 67.42           O
ATOM      9  CB  HIS A  10      44.916  29.565 -22.586  1.00 72.64           C
ATOM     10  CG  HIS A  10      44.712  28.203 -23.207  1.00 74.51           C
ATOM     11  ND1 HIS A  10      43.872  27.232 -22.873  1.00 76.95           N
ATOM     12  CD2 HIS A  10      45.380  27.790 -24.333  1.00 75.61           C
ATOM     13  CE1 HIS A  10      43.995  26.255 -23.743  1.00 77.59           C
ATOM     14  NE2 HIS A  10      44.904  26.606 -24.617  1.00 77.04           N
ATOM     15  N   LYS A  11      41.767  29.785 -23.185  1.00 62.39           N
ATOM     16  CA  LYS A  11      40.518  29.596 -22.485  1.00 57.64           C
ATOM     17  C   LYS A  11      40.757  28.268 -21.807  1.00 52.94           C
ATOM     18  O   LYS A  11      41.131  27.357 -22.559  1.00 53.17           O
ATOM     19  CB  LYS A  11      39.356  29.487 -23.461  1.00 59.00           C
ATOM     20  CG  LYS A  11      38.836  30.831 -23.964  1.00 61.16           C
ATOM     21  CD  LYS A  11      37.618  30.622 -24.865  1.00 62.37           C
ATOM     22  CE  LYS A  11      36.592  31.732 -24.665  1.00 64.46           C
ATOM     23  NZ  LYS A  11      35.966  31.661 -23.353  1.00 65.05           N
ATOM     24  N   ILE A  12      40.730  28.077 -20.482  1.00 46.69           N
ATOM     25  CA  ILE A  12      40.865  26.731 -19.945  1.00 41.32           C
ATOM     26  C   ILE A  12      39.470  26.180 -20.031  1.00 38.16           C
ATOM     27  O   ILE A  12      38.510  26.898 -19.759  1.00 37.75           O
ATOM     28  CB  ILE A  12      41.293  26.663 -18.468  1.00 41.62           C
ATOM     29  CG1 ILE A  12      42.663  27.279 -18.246  1.00 43.09           C
ATOM     30  CG2 ILE A  12      41.377  25.188 -18.060  1.00 42.11           C
ATOM     31  CD1 ILE A  12      43.236  27.086 -16.815  1.00 42.37           C
ATOM     32  N   LYS A  13      39.303  24.956 -20.484  1.00 36.05           N
ATOM     33  CA  LYS A  13      37.973  24.393 -20.494  1.00 35.27           C
ATOM     34  C   LYS A  13      37.680  23.777 -19.138  1.00 33.46           C
ATOM     35  O   LYS A  13      38.529  23.068 -18.585  1.00 34.67           O
ATOM     36  CB  LYS A  13      37.851  23.326 -21.554  1.00 34.84           C
ATOM     37  CG  LYS A  13      37.848  24.027 -22.853  1.00 38.81           C
ATOM     38  CD  LYS A  13      37.938  23.820 -25.257  1.00 47.26           C
ATOM     39  CE  LYS A  13      37.912  22.956 -26.427  1.00 49.91           N
ATOM     40  NZ  LYS A  13      36.505  24.051 -18.598  1.00 30.23           N
ATOM     41  N   GLY A  14      36.043  23.437 -17.381  1.00 27.07           C
ATOM     42  CA  GLY A  14      34.720  22.725 -17.623  1.00 24.90           C
ATOM     43  C   GLY A  14      33.931  23.085 -18.502  1.00 24.28           O
ATOM     44  O   GLY A  14      34.467  21.648 -16.891  1.00 23.77           N
ATOM     45  N   THR A  15      33.165  21.008 -16.888  1.00 21.99           C
ATOM     46  CA  THR A  15      32.851  20.876 -15.366  1.00 20.67           C
ATOM     47  C   THR A  15      33.636  20.316 -14.589  1.00 20.91           O
ATOM     48  O   THR A  15      33.307  19.663 -17.787  1.00 21.11           C
ATOM     49  CB  THR A  15      32.445  18.636 -17.266  1.00 20.61           O
ATOM     50  OG1 THR A  15      34.731  19.148 -17.867  1.00 21.03           C
ATOM     51  CG2 THR A  15      31.766  21.535 -14.909  1.00 19.46           N
ATOM     52  N   VAL A  16      31.240  21.613 -13.527  1.00 18.05           C
ATOM     53  CA  VAL A  16      30.213  20.539 -13.297  1.00 17.95           C
ATOM     54  C   VAL A  16      29.183  20.596 -13.979  1.00 20.19           O
ATOM     55  O   VAL A  16      30.558  23.011 -13.268  1.00 17.93           C
ATOM     56  CB  VAL A  16      29.522  23.084 -12.125  1.00 14.31           C
ATOM     57  CG1 VAL A  16      31.695  23.927 -12.855  1.00 19.85           C
ATOM     58  CG2 VAL A  16      30.388  19.552 -12.431  1.00 16.80           N
ATOM     59  N   VAL A  17      29.315  18.588 -12.315  1.00 16.32           C
ATOM     60  CA  VAL A  17      28.619  18.942 -11.011  1.00 16.99           C
ATOM     61  C   VAL A  17      29.259  19.312 -10.011  1.00 15.91           O
ATOM     62  O   VAL A  17      29.896  17.152 -12.354  1.00 14.31           C
ATOM     63  CB  VAL A  17      28.785  16.193 -12.693  1.00 12.71           C
ATOM     64  CG1 VAL A  17      30.900  16.974 -13.474  1.00 15.05           C
ATOM     65  CG2 VAL A  17      27.289  18.817 -11.016  1.00 18.12           N
ATOM     66  N   LEU A  18      26.438  19.322  -9.949  1.00 18.03           C
ATOM     67  CA  LEU A  18      25.212  18.444  -9.785  1.00 17.64           C
ATOM     68  C   LEU A  18      25.012  17.524 -10.585  1.00 18.53           O
ATOM     69  O   LEU A  18      26.164  20.740 -10.401  1.00 18.68           C
ATOM     70  CB  LEU A  18      25.172  21.740  -9.940  1.00 18.30           C
ATOM     71  CG  LEU A  18      25.617  23.036 -10.575  1.00 17.40           C
ATOM     72  CD1 LEU A  18      23.754  21.431 -10.378  1.00 16.91           C
ATOM     73  CD2 LEU A  18      24.409  18.702  -8.762  1.00 15.77           N
ATOM     74  N   MET A  19      23.175  17.995  -8.504  1.00 14.82           C
ATOM     75  CA  MET A  19      22.185  19.042  -8.012  1.00 17.55           C
ATOM     76  C   MET A  19      22.536  19.924  -7.218  1.00 17.69           O
ATOM     77  O   MET A  19      23.405  16.974  -7.440  1.00 13.63           C
ATOM     78  CB  MET A  19      23.904  15.656  -7.949  1.00 17.04           C
ATOM     79  CG  MET A  19
```

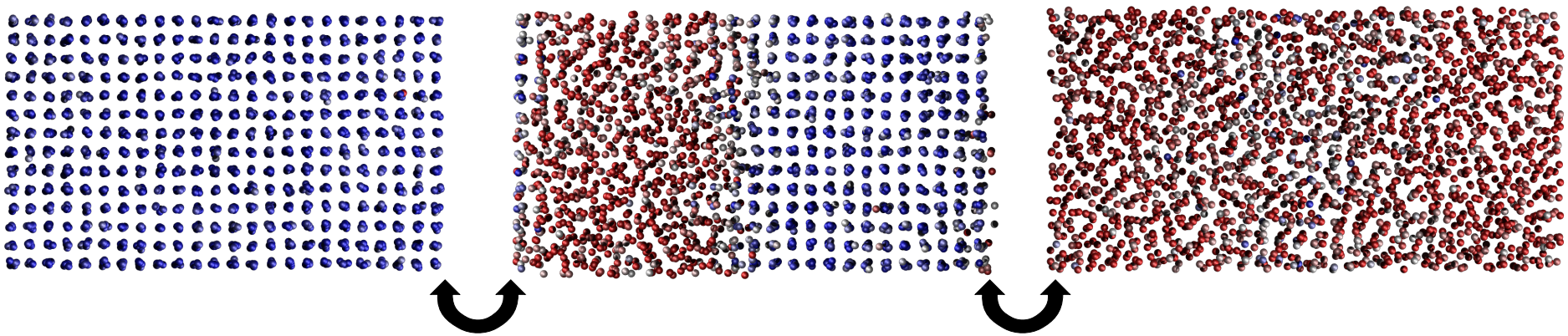
Managing data from atomistic models

- Atomistic simulations provide **too much** information
- It is hard to decipher the essential features in structurally-complex compounds, materials, proteins, etc.



Describing structural complexity

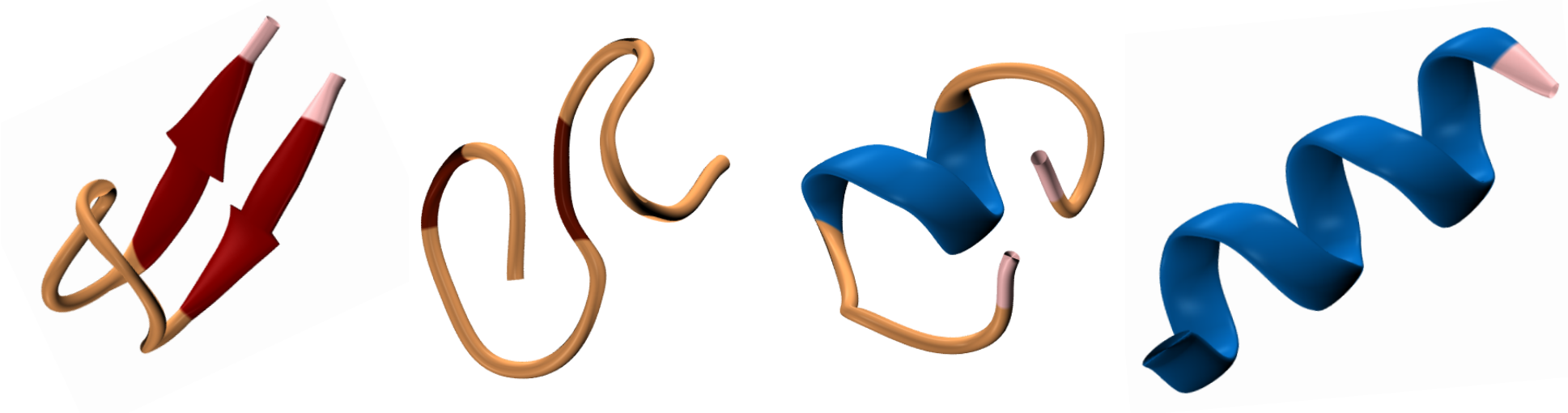
- We are looking for **collective variables** that can describe this structural complexity
  - Discriminate between different structures
  - Follow the system across transitions
- Finding these variables is time-consuming and error-prone: can we **automate** the process?



Angioletti, Ceriotti, Lee, Finnis PRB 2010

Describing structural complexity

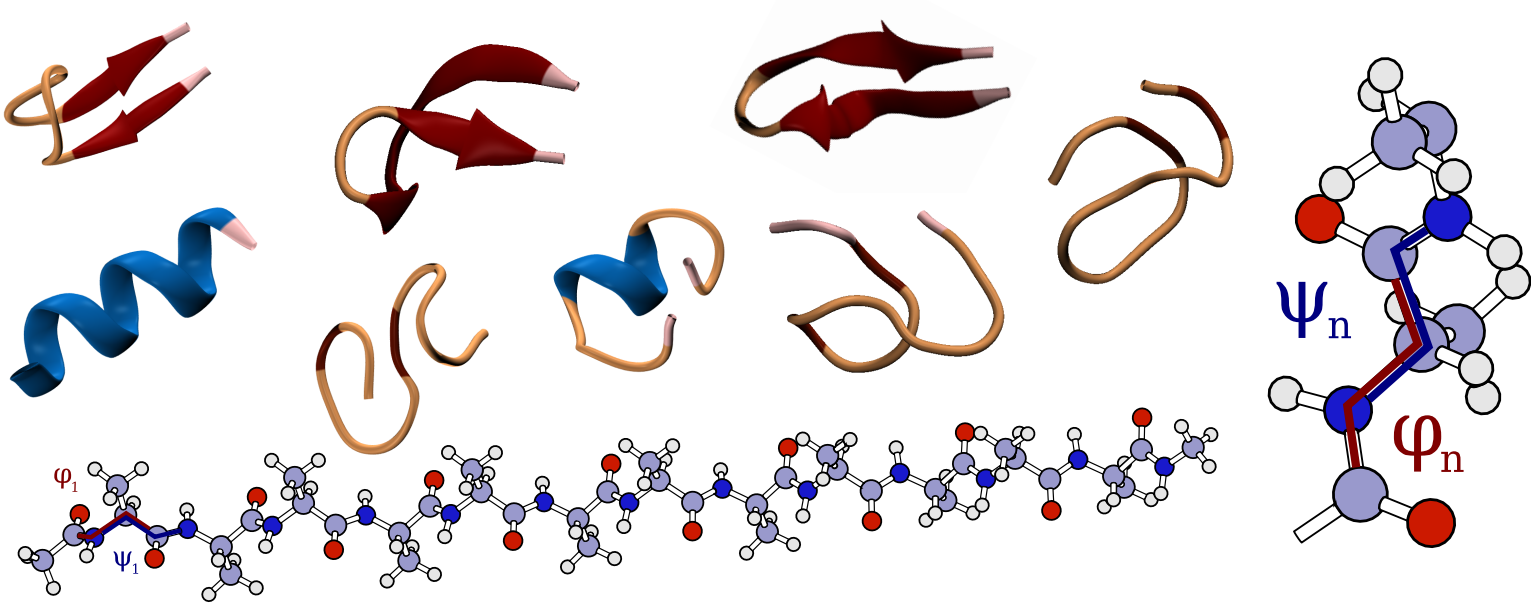
- We are looking for **collective variables** that can describe this structural complexity
  - Discriminate between different structures
  - Follow the system across transitions
- Finding these variables is time-consuming and error-prone: can we **automate** the process?



Tribello, Ceriotti, Parrinello PNAS 2010

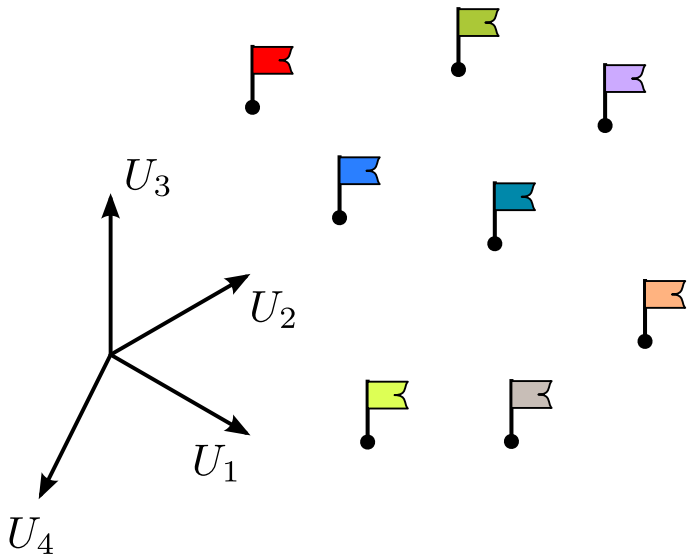
Dimensionality reduction

- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



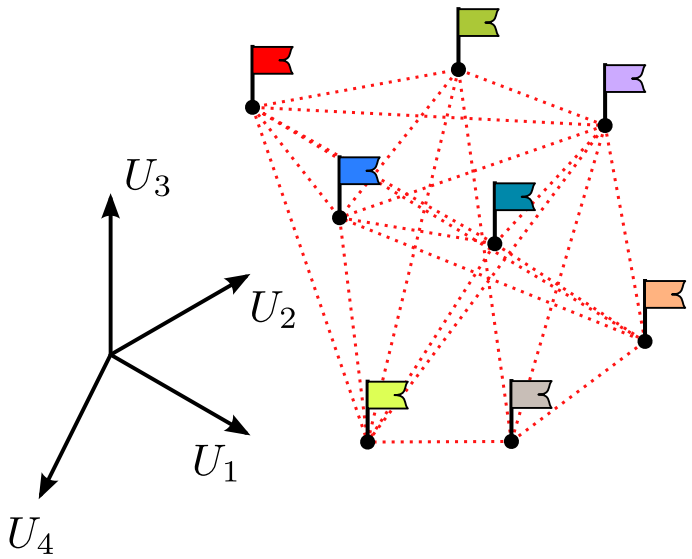
Dimensionality reduction

- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations ⇒ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



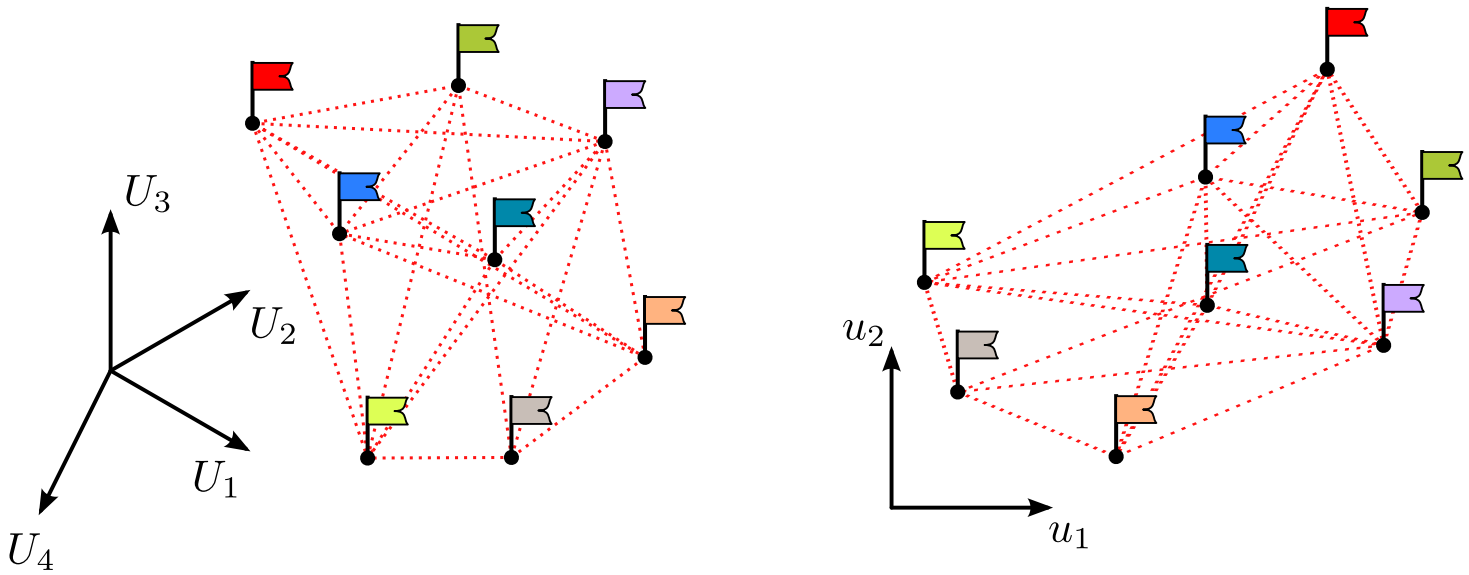
Dimensionality reduction

- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations $\Rightarrow$ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



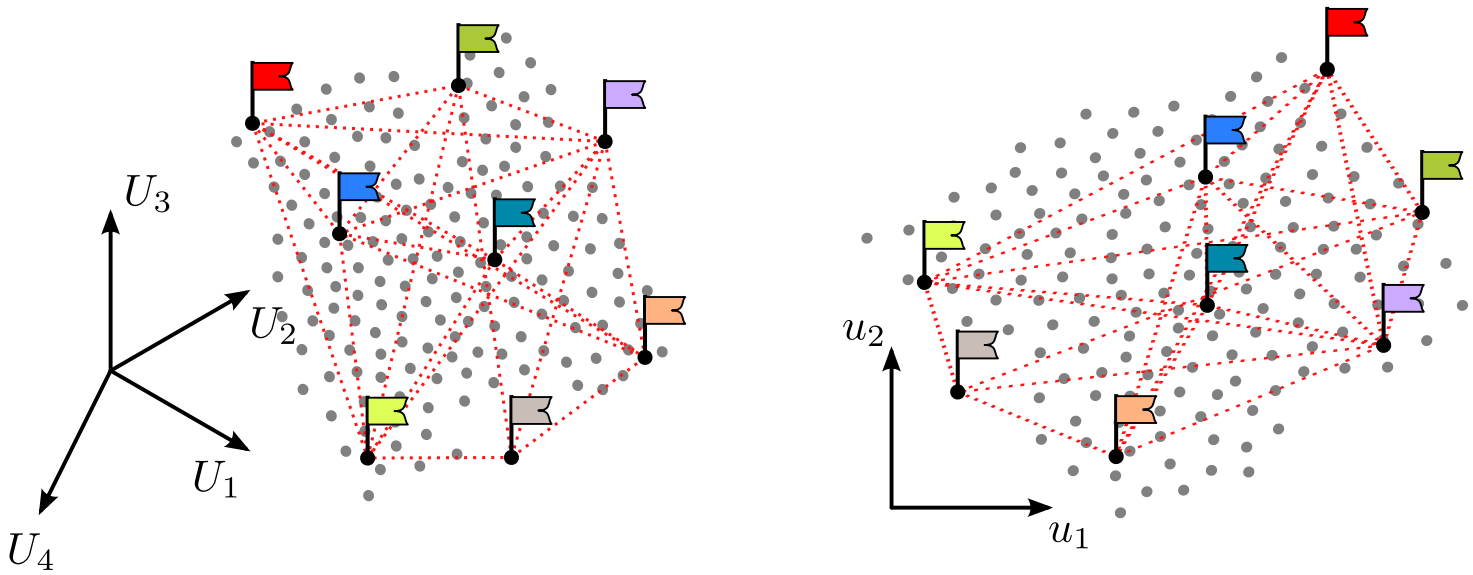
Dimensionality reduction

- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations $\Rightarrow$ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



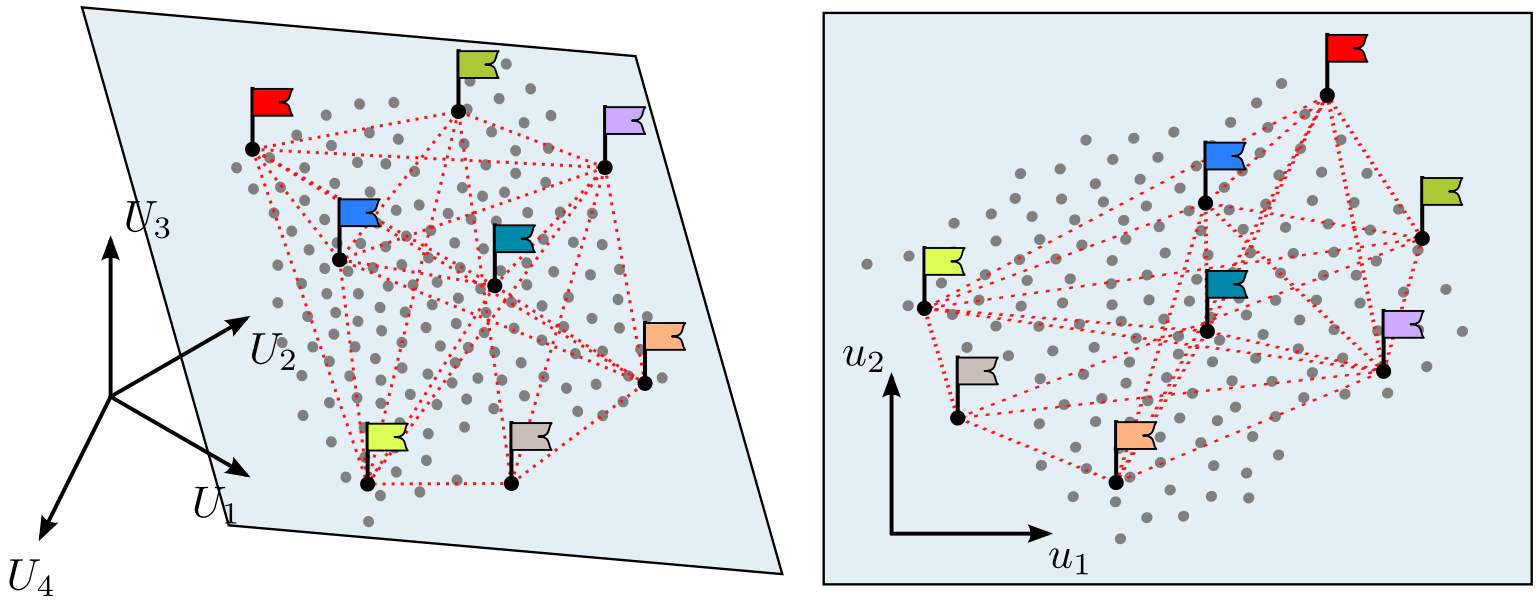
Dimensionality reduction

- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations $\Rightarrow$ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



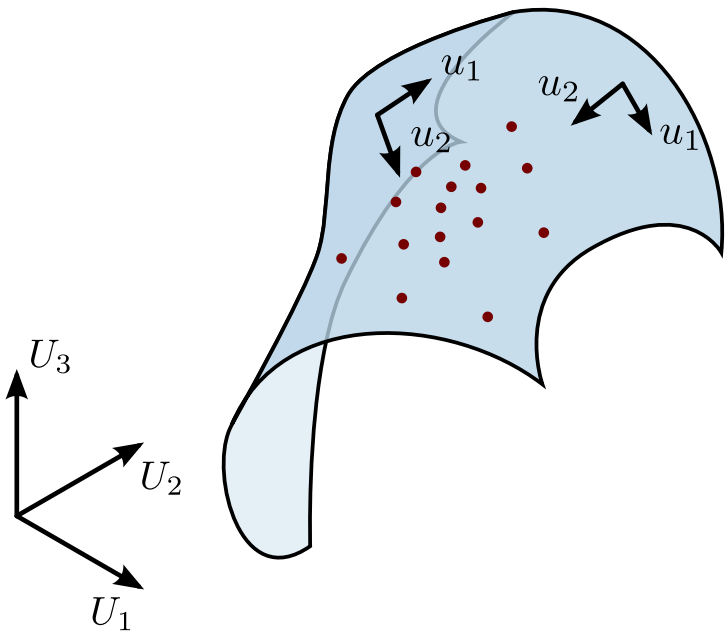
Dimensionality reduction

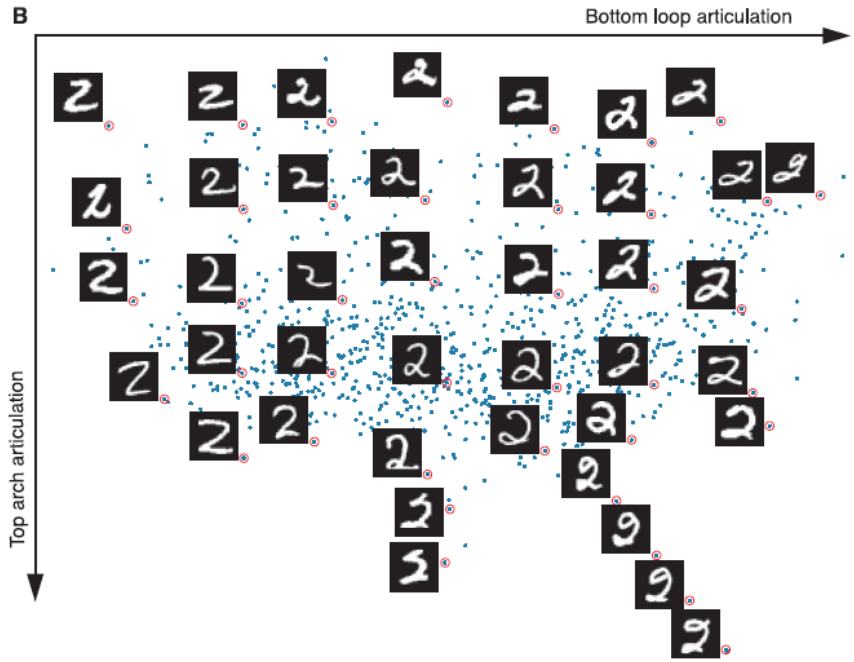
- We can describe a complex atomistic structure as a point in a high-dimensional space. Then finding CVs means finding a low-dimensional **map** to describe the accessible configurations!
  - Take a set of configurations $\Rightarrow$ high-dim. **landmark points**
  - Find the distances between the points
  - Arrange low-dim. points so that the distances are preserved
  - Locate other configurations with an **out-of-sample embedding**



Non-linear dimensionality reduction

- Non-linear dimensionality reduction algorithms:
  - Describe curved, "locally-flat" manifolds
  - Developed by the CS community (image recognition)
  - Attempts to apply to chemical problems (PCA, ISOMAP, Diff. maps,...)
- Atomistic simulations are harder:
  - **Thermal fluctuations** are high-dimensional
  - A network of transition pathways with a complex topology

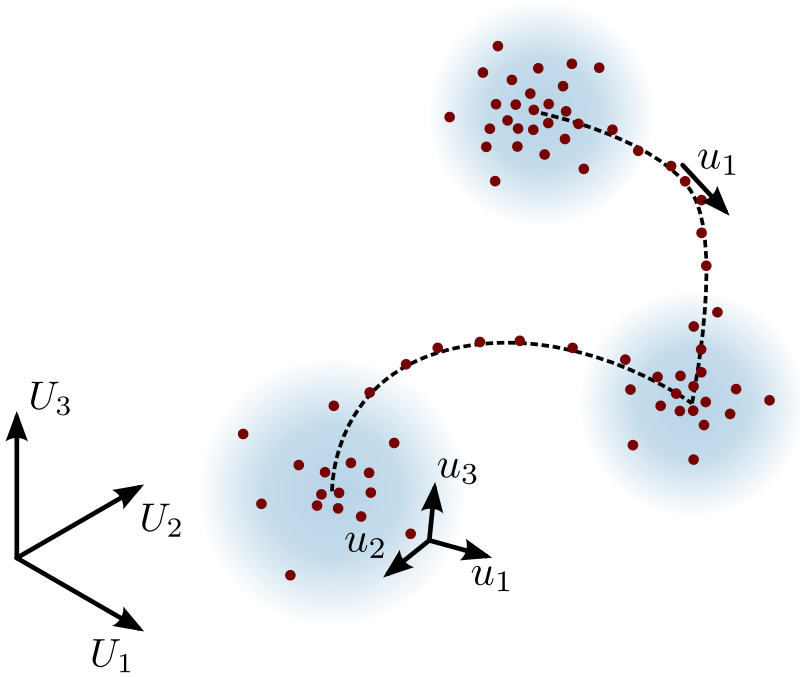


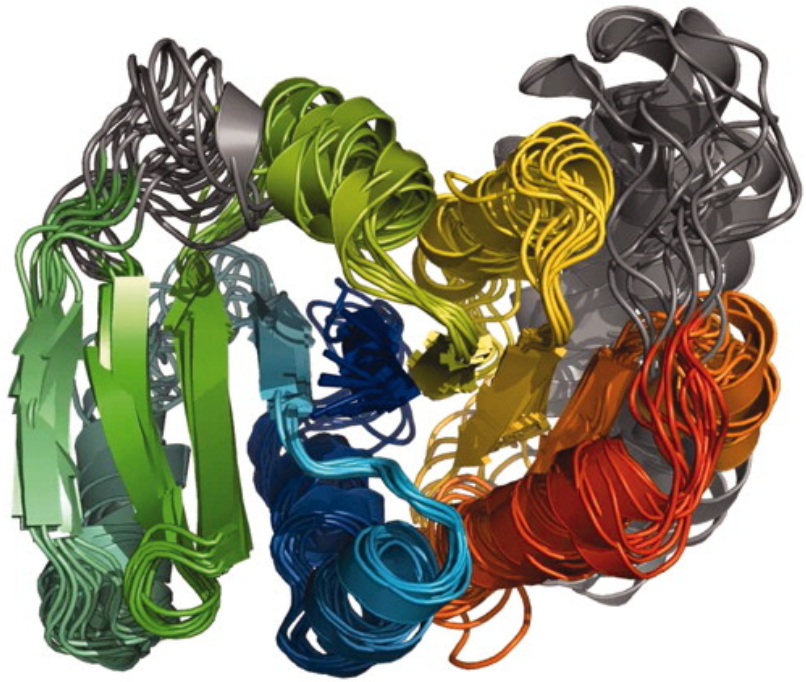


Tenenbaum *et al.*, Science (2000)

Non-linear dimensionality reduction

- Non-linear dimensionality reduction algorithms:
  - Describe curved, "locally-flat" manifolds
  - Developed by the CS community (image recognition)
  - Attempts to apply to chemical problems (PCA, ISOMAP, Diff. maps,...)
- Atomistic simulations are harder:
  - **Thermal fluctuations** are high-dimensional
  - A network of transition pathways with a complex topology



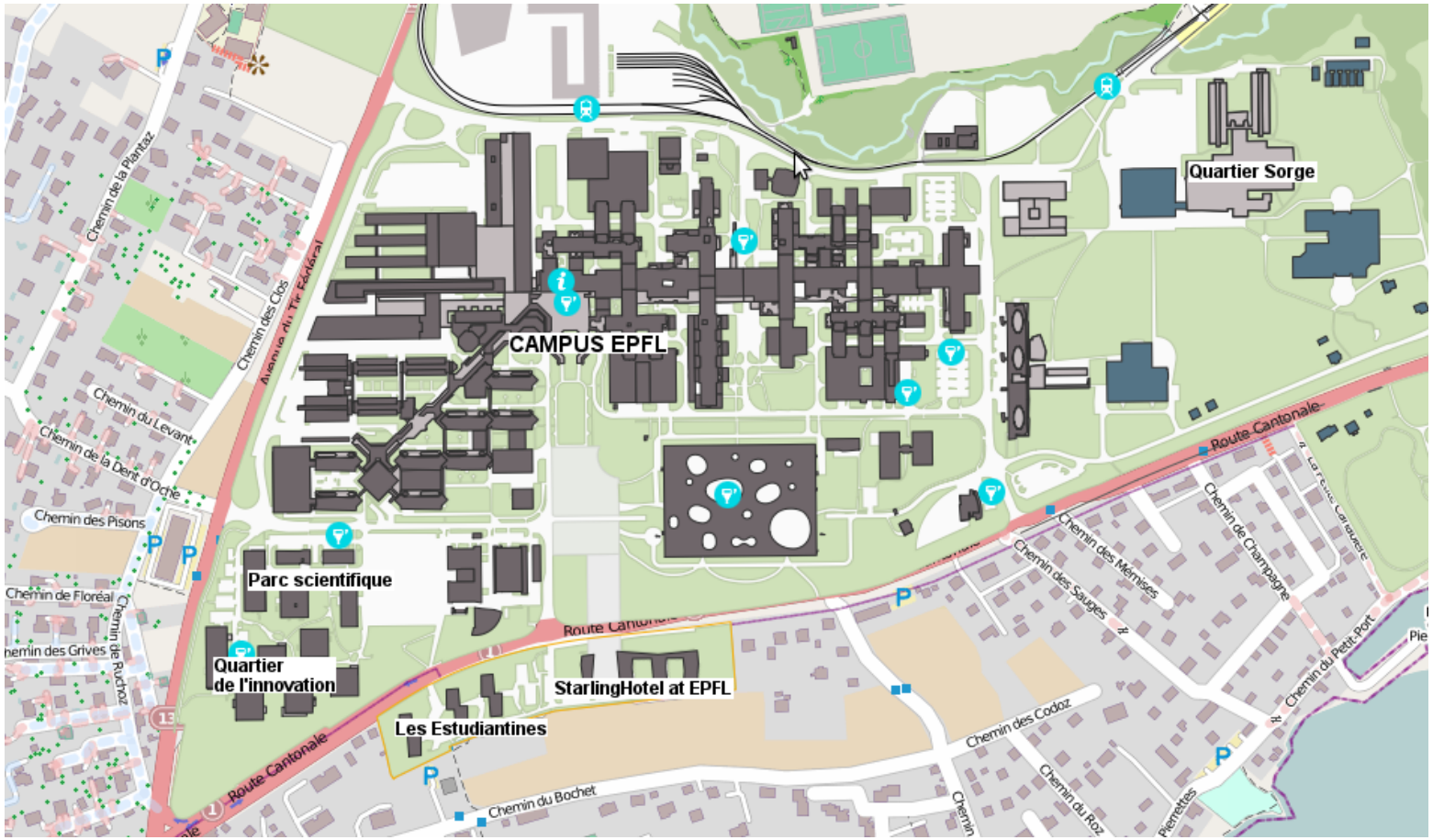


Introducing sketch-map

- Developing a more robust NLDR method
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map

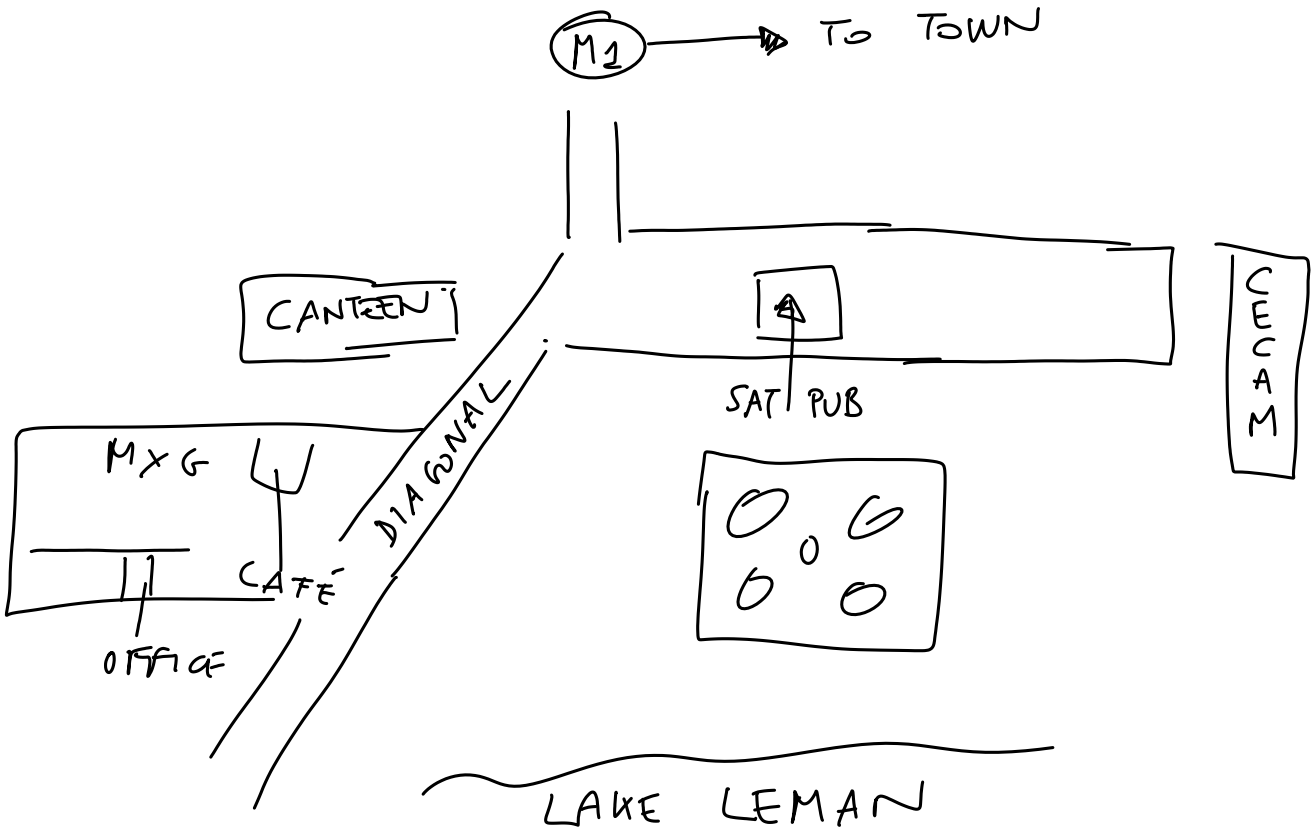
Introducing sketch-map

- Developing a more robust NLDR method
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map



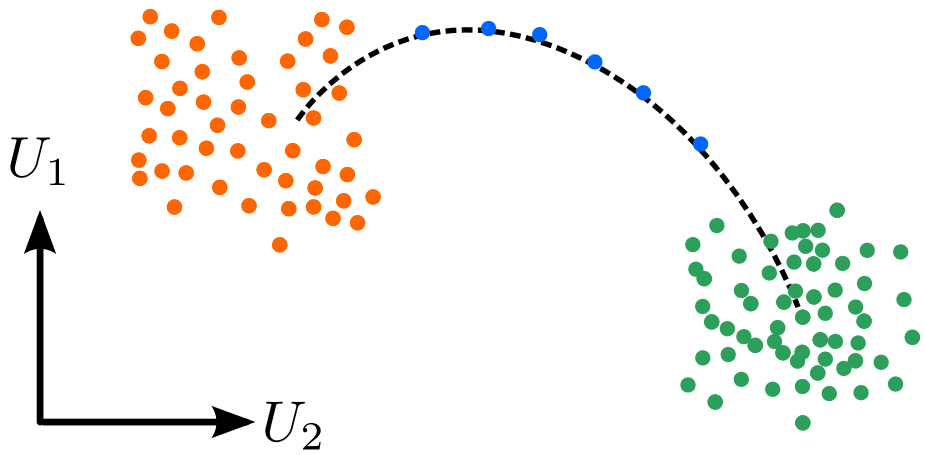
Introducing sketch-map

- Developing a more robust NLDR method
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map



Proximity matching

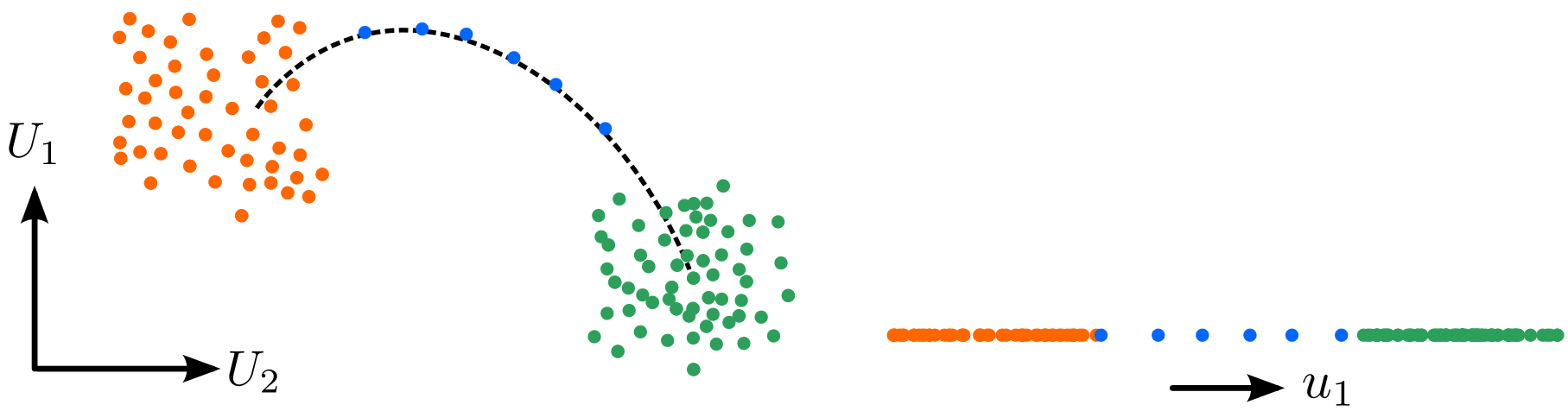
- We would like to capture the low-dimensional structure of complex transitions
- Some portions of the landscape cannot be projected by matching high and low-dimensional distances. Thermal fluctuations exhibit inherent full dimensionality!
  - Idea: simplify the task, aim for proximity matching: close↔close, far↔far



Ceriotti, Tribello, Parrinello, PNAS 2011

Proximity matching

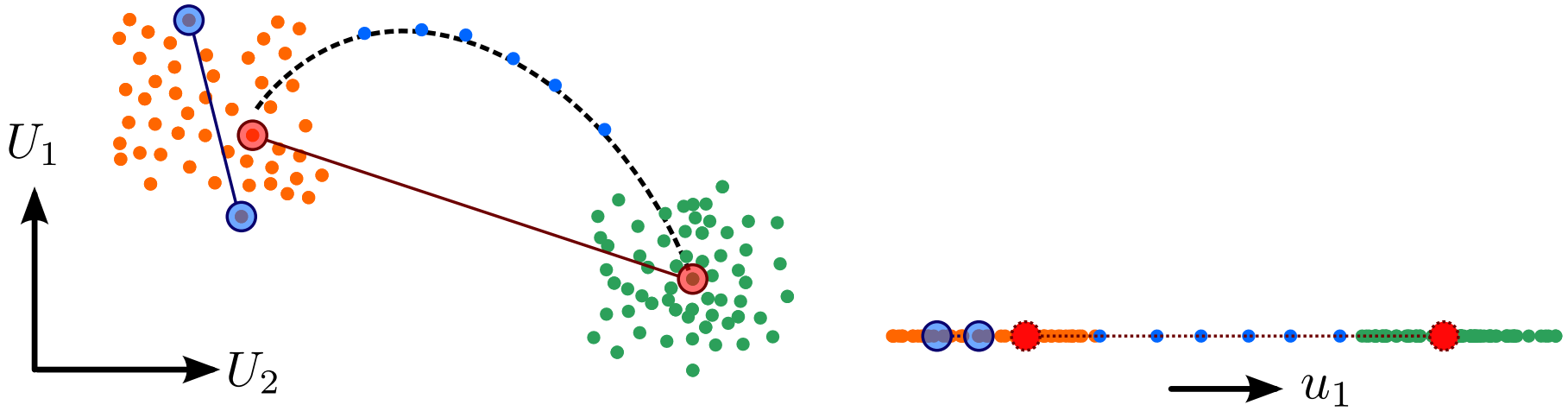
- We would like to capture the low-dimensional structure of complex transitions
- Some portions of the landscape cannot be projected by matching high and low-dimensional distances. Thermal fluctuations exhibit inherent full dimensionality!
  - Idea: simplify the task, aim for proximity matching: close↔close, far↔far



Ceriotti, Tribello, Parrinello, PNAS 2011

Proximity matching

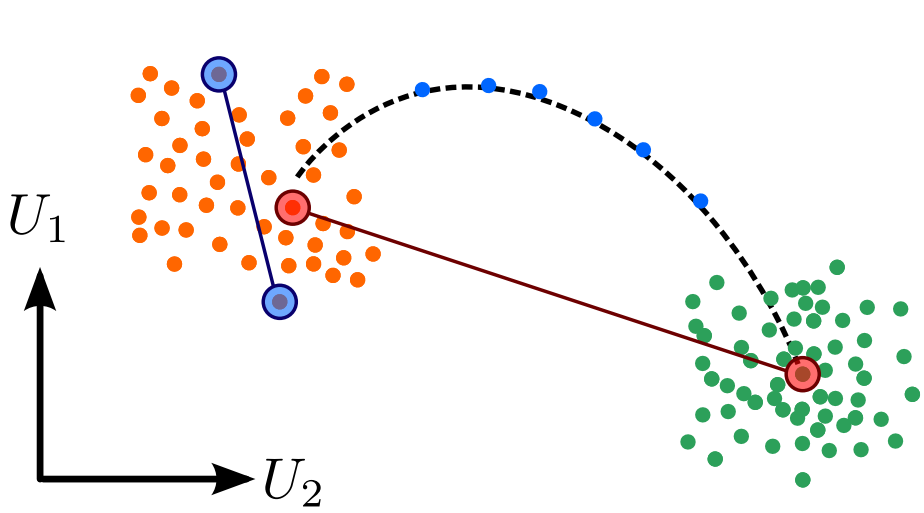
- We would like to capture the low-dimensional structure of complex transitions
- Some portions of the landscape cannot be projected by matching high and low-dimensional distances. Thermal fluctuations exhibit inherent full dimensionality!
  - Idea: simplify the task, aim for proximity matching: close↔close, far↔far

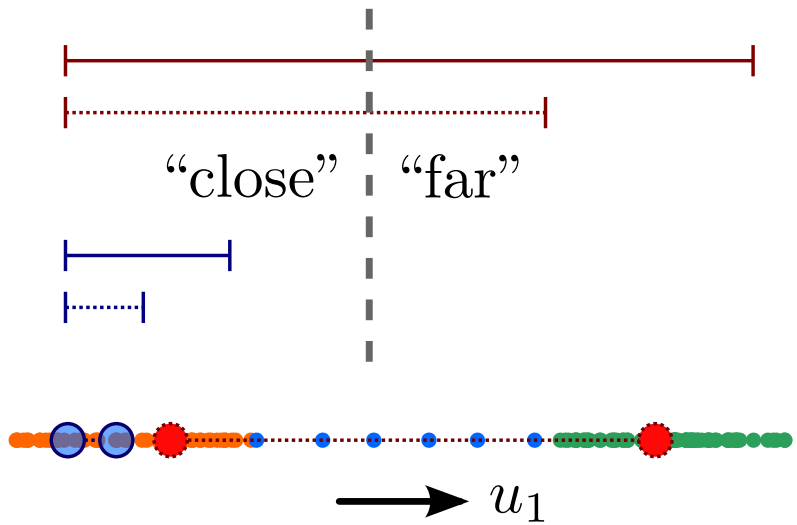


Ceriotti, Tribello, Parrinello, PNAS 2011

Proximity matching

- We would like to capture the low-dimensional structure of complex transitions
- Some portions of the landscape cannot be projected by matching high and low-dimensional distances. Thermal fluctuations exhibit inherent full dimensionality!
  - Idea: simplify the task, aim for proximity matching: close↔close, far↔far





Ceriotti, Tribello, Parrinello, PNAS 2011

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

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

Ceriotti,Tribello,Parrinello PNAS 2011

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

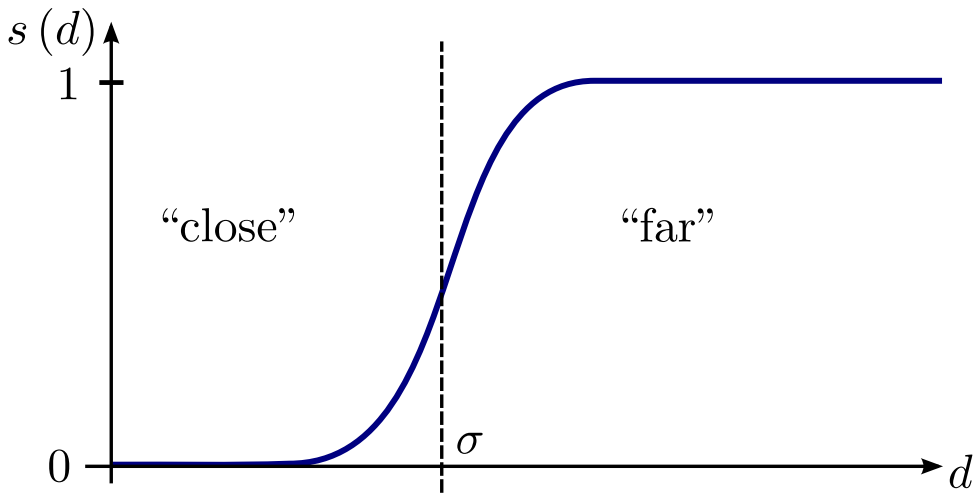
$$\chi^2 = \sum_{i,j=1}^{N} [|X_i - X_j| - |x_i - x_j|]^2$$

Ceriotti,Tribello,Parrinello PNAS 2011

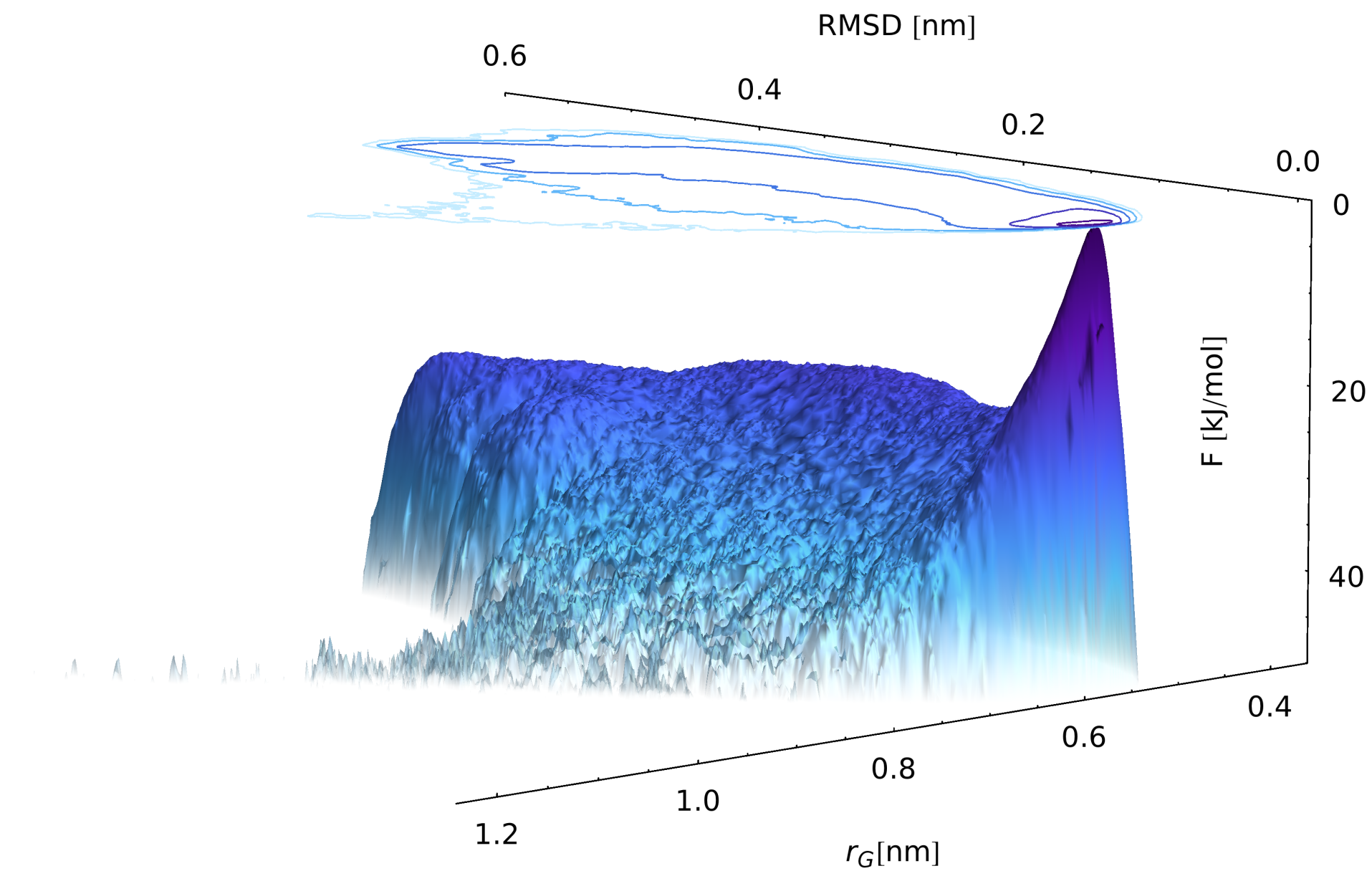
Sketch-map algorithm

- 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

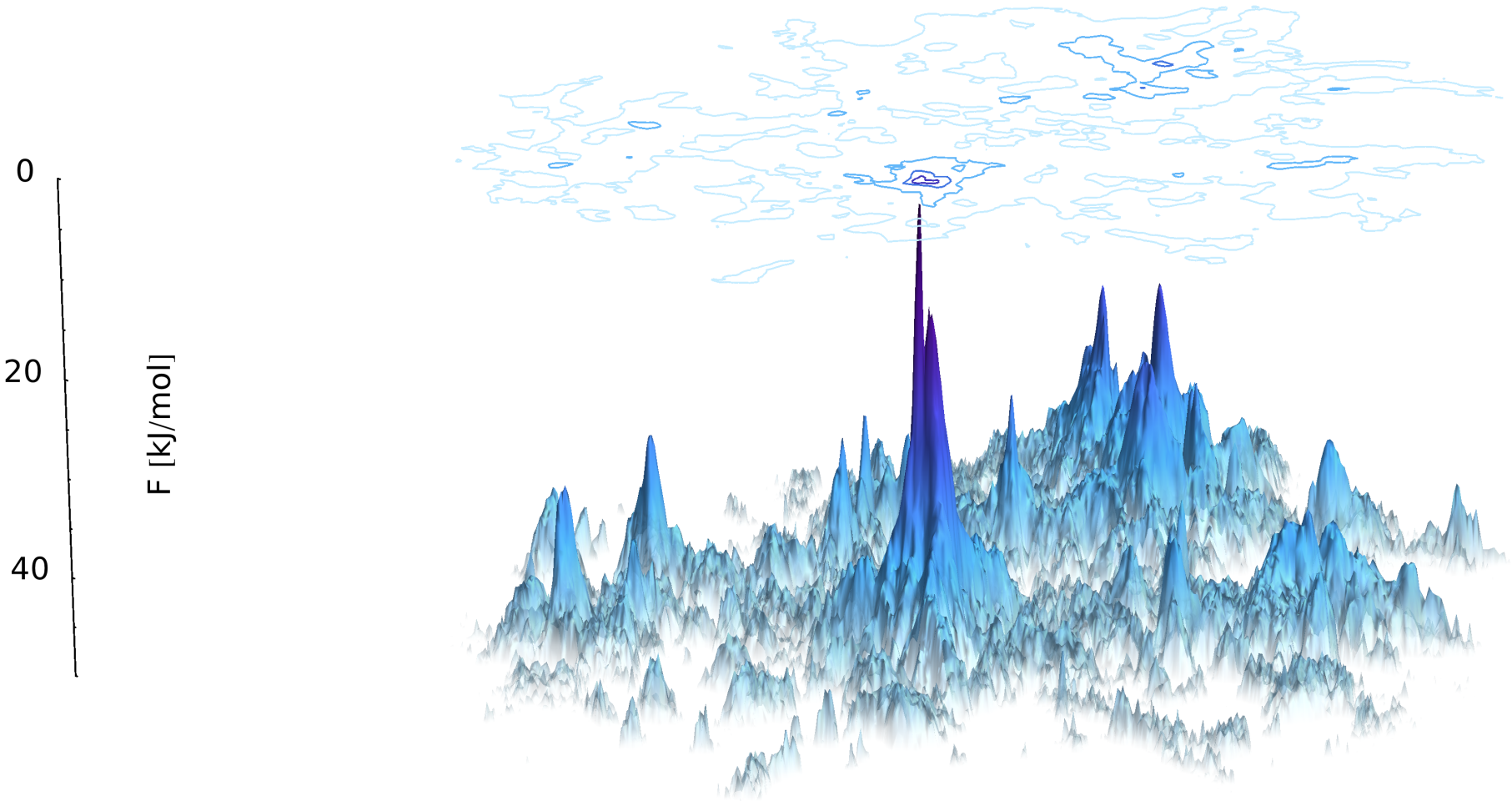
$$\chi^2 = \sum_{i,j=1}^{N} \left[s\left(|X_i - X_j|\right) - s\left(|x_i - x_j|\right) \right]^2$$



Sketch-mapping the folding landscape of a

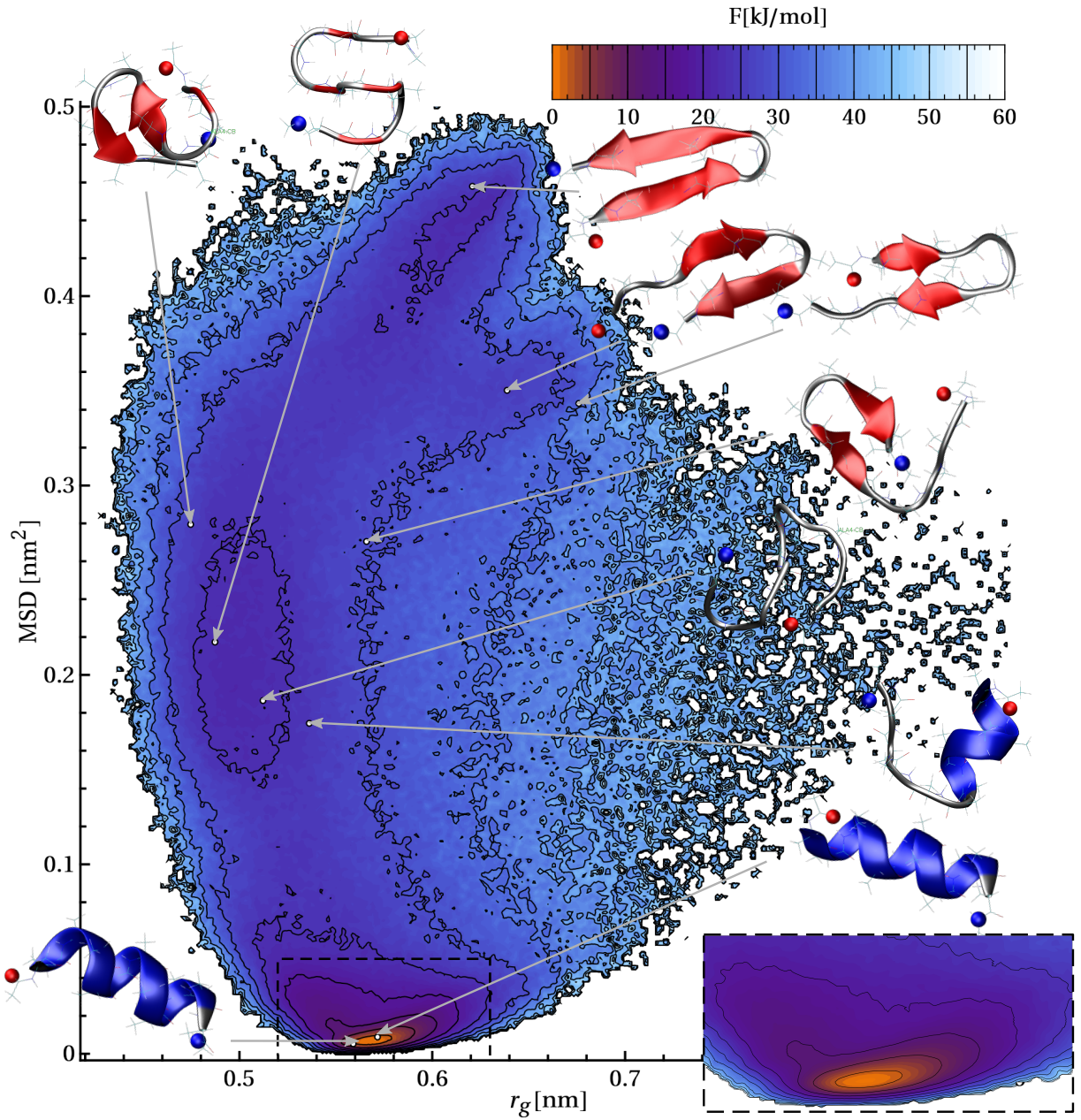


Sketch-mapping the folding landscape of a



Sketch-mapping the folding landscape of a

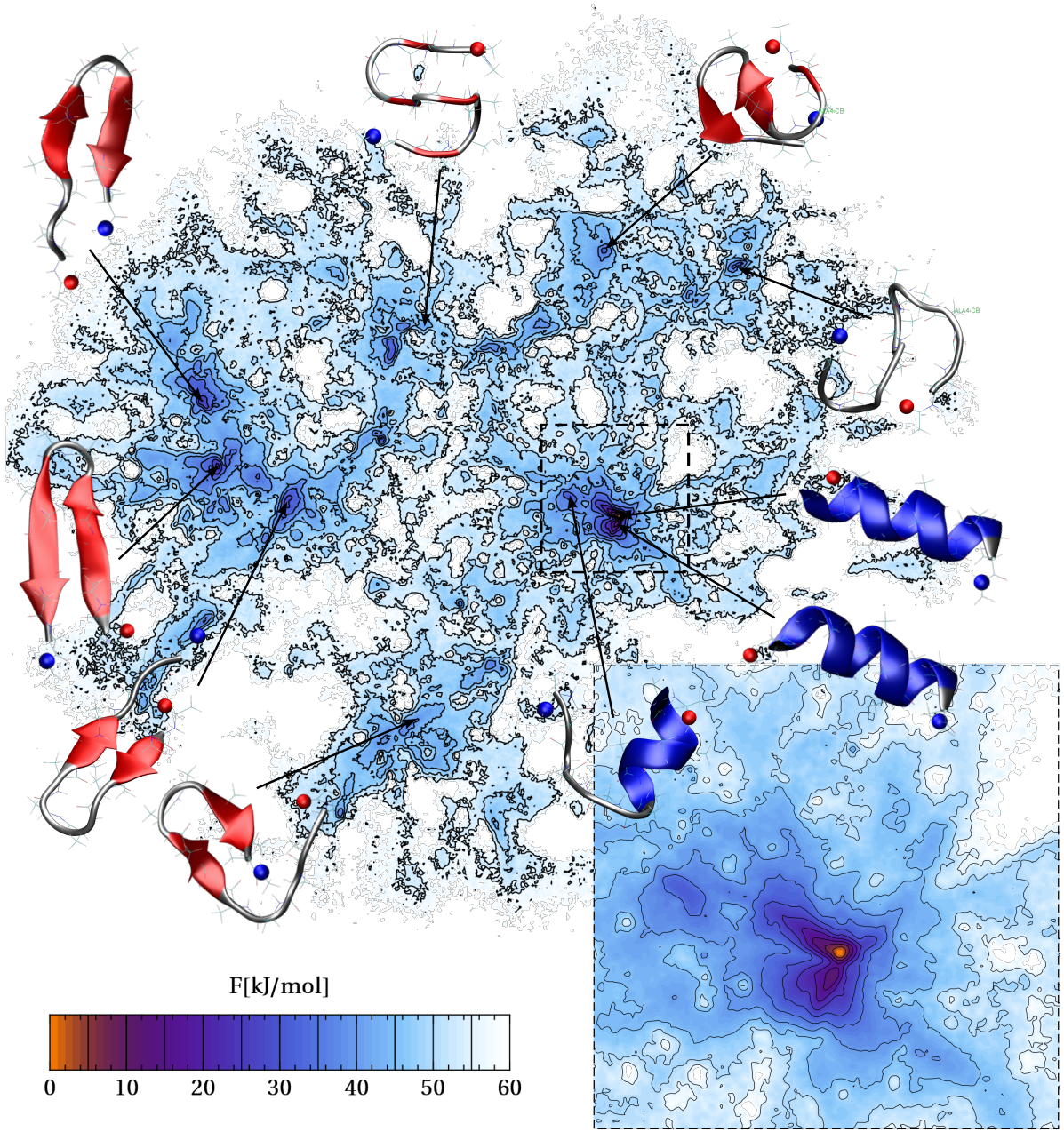
"Conventional" CVs recognize the folded state, but many meta-stable structures overlap with each other



Ceriotti, Tribello, Parrinello, PNAS 2011

Sketch-mapping the folding landscape of a

Sketch-map CVs give a very detailed picture, where each meta-stable configuration is clearly singled out

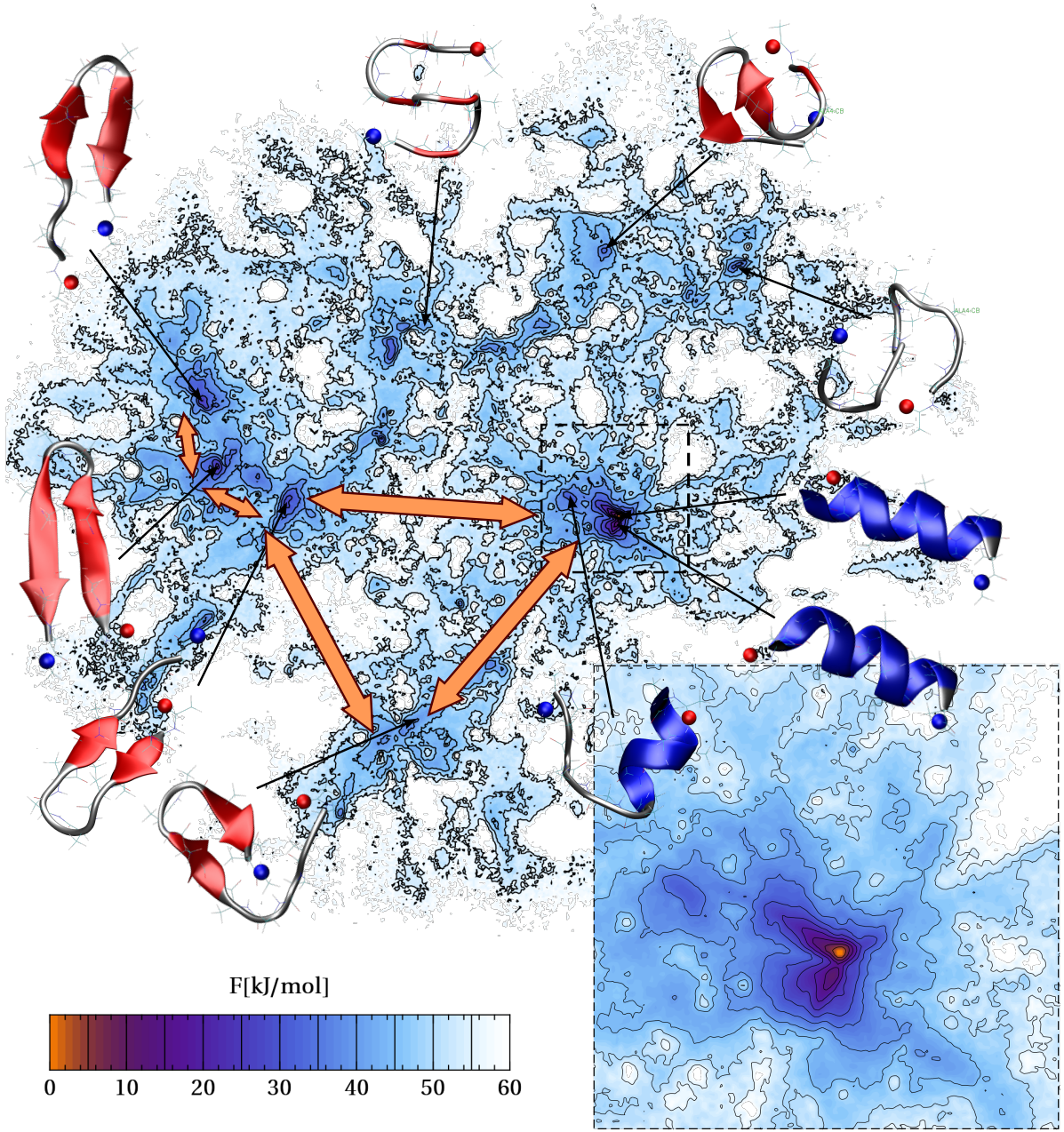


Ceriotti, Tribello, Parrinello, PNAS 2011

Sketch-mapping the folding landscape of a

Sketch-map CVs give a very detailed picture, where each meta-stable configuration is clearly singled out

Can be used effectively for accelerated dynamics:
field-overlap metadynamics

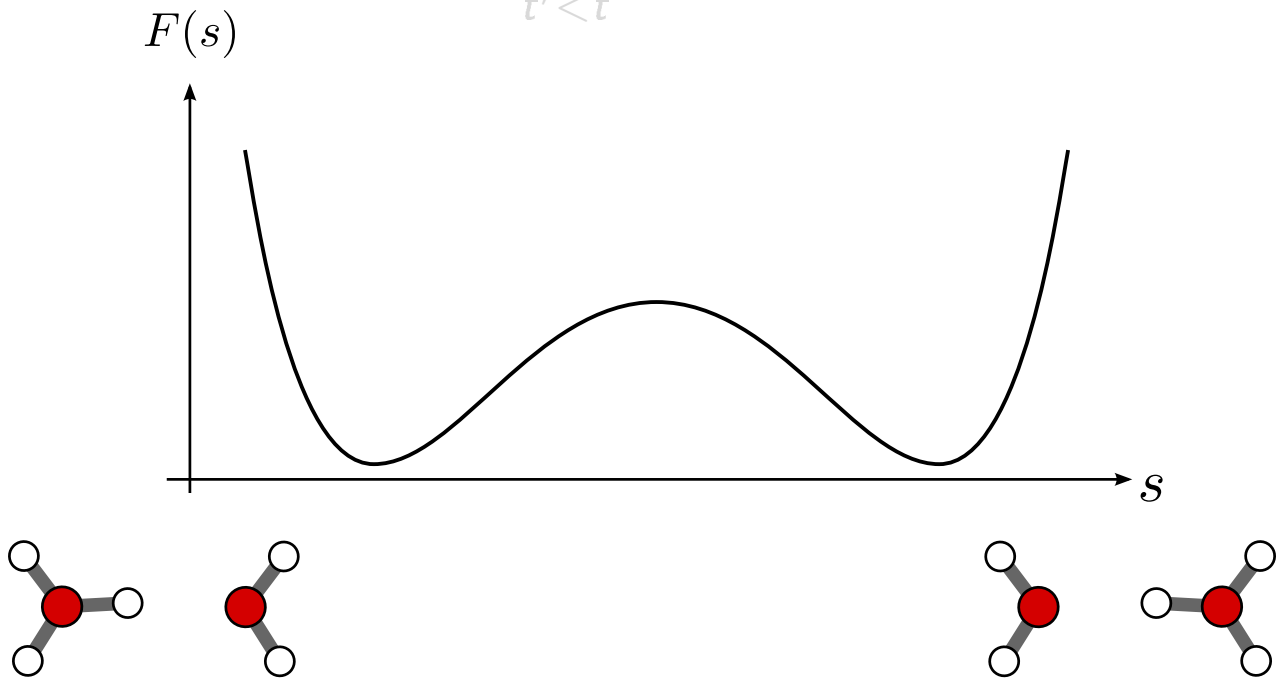


Tribello, Ceriotti, Parrinello, PNAS 2012

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

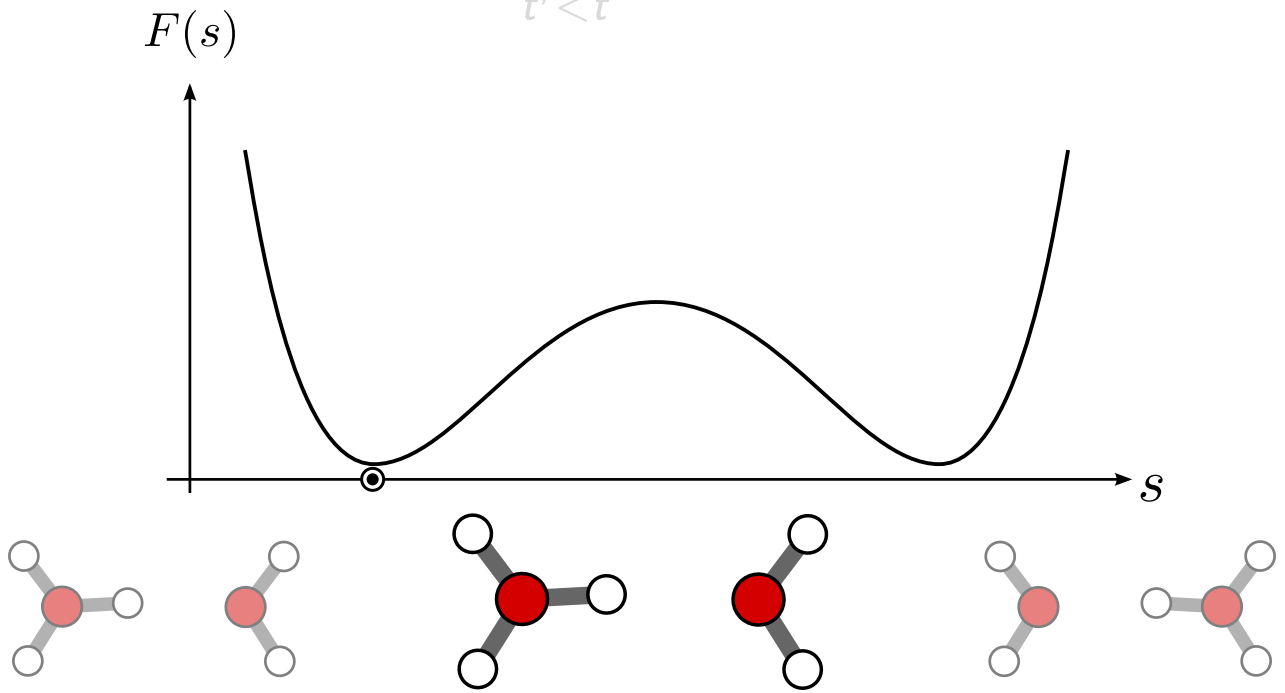


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

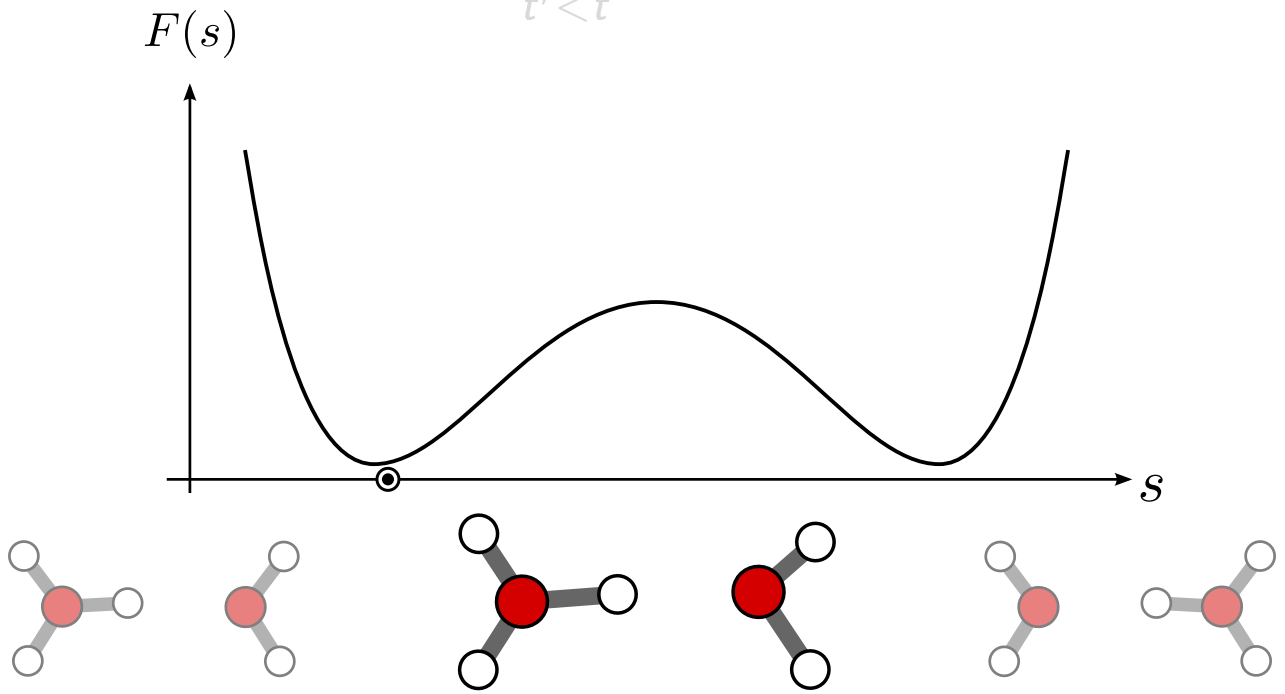


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

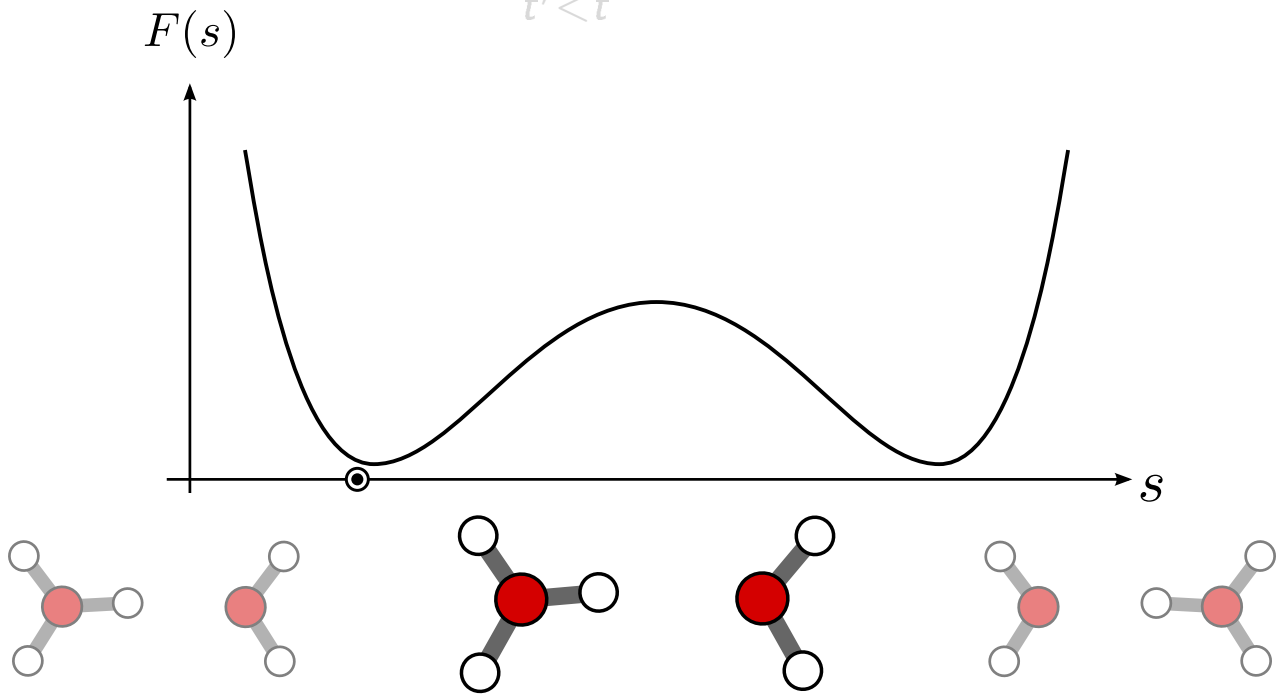


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

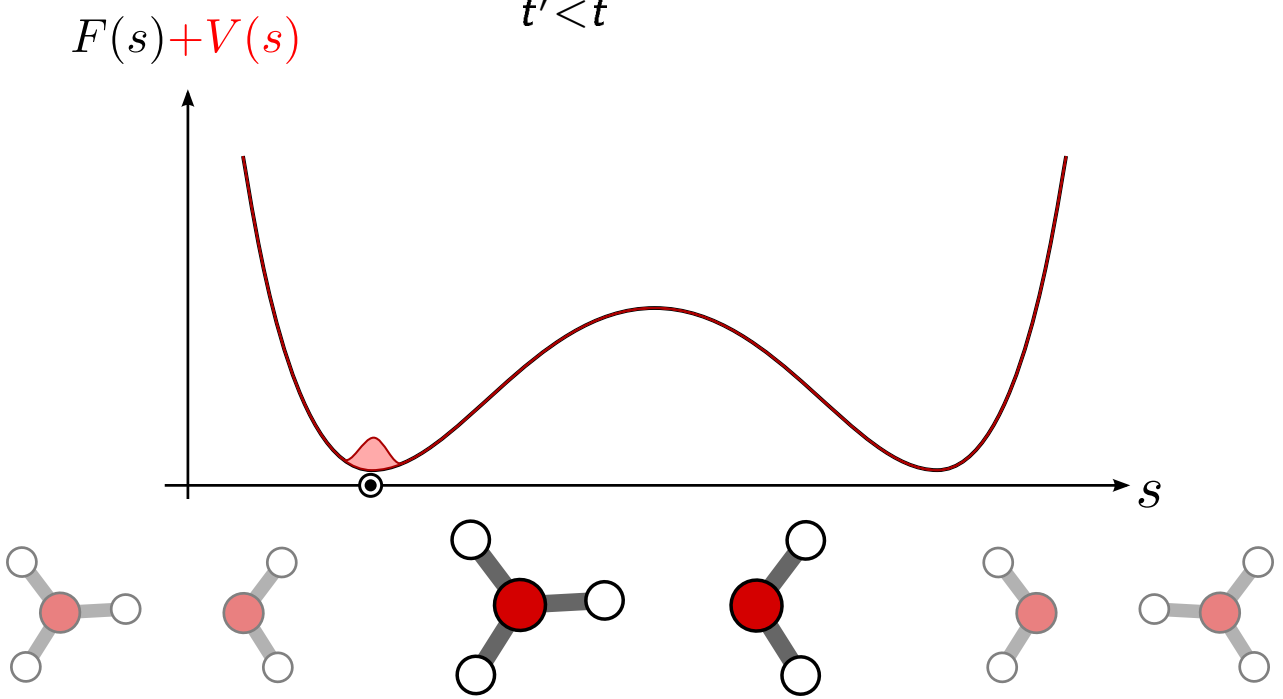


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

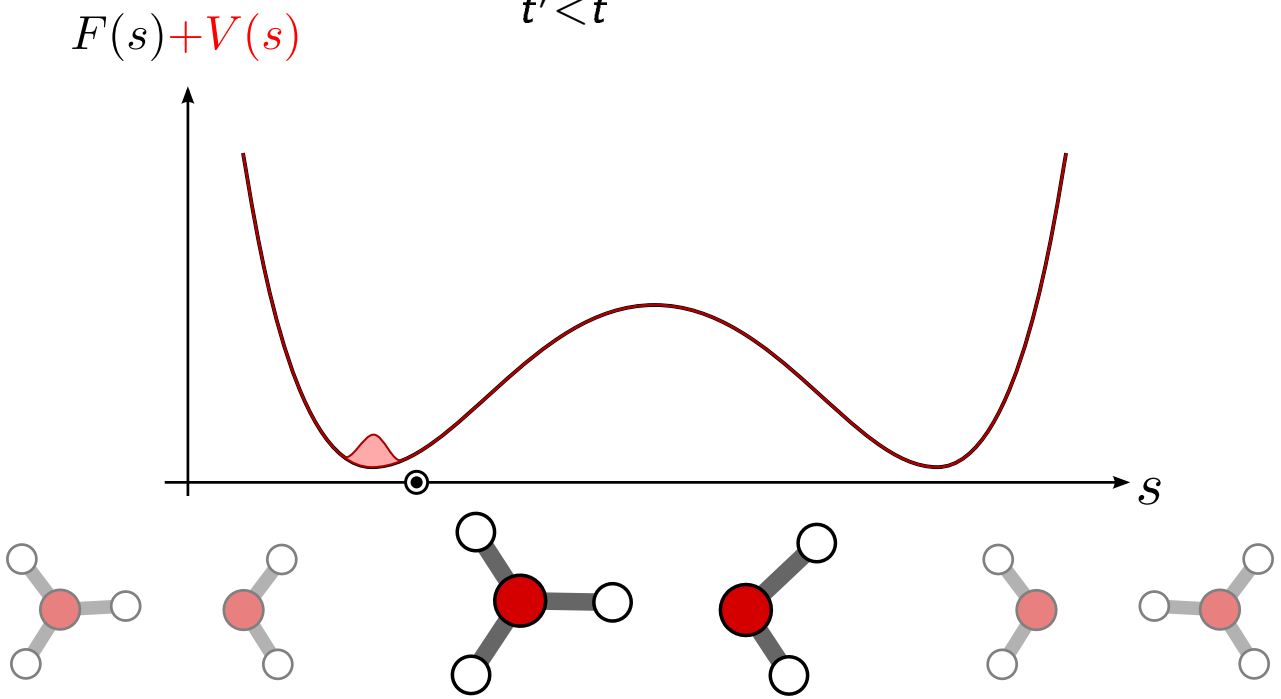


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

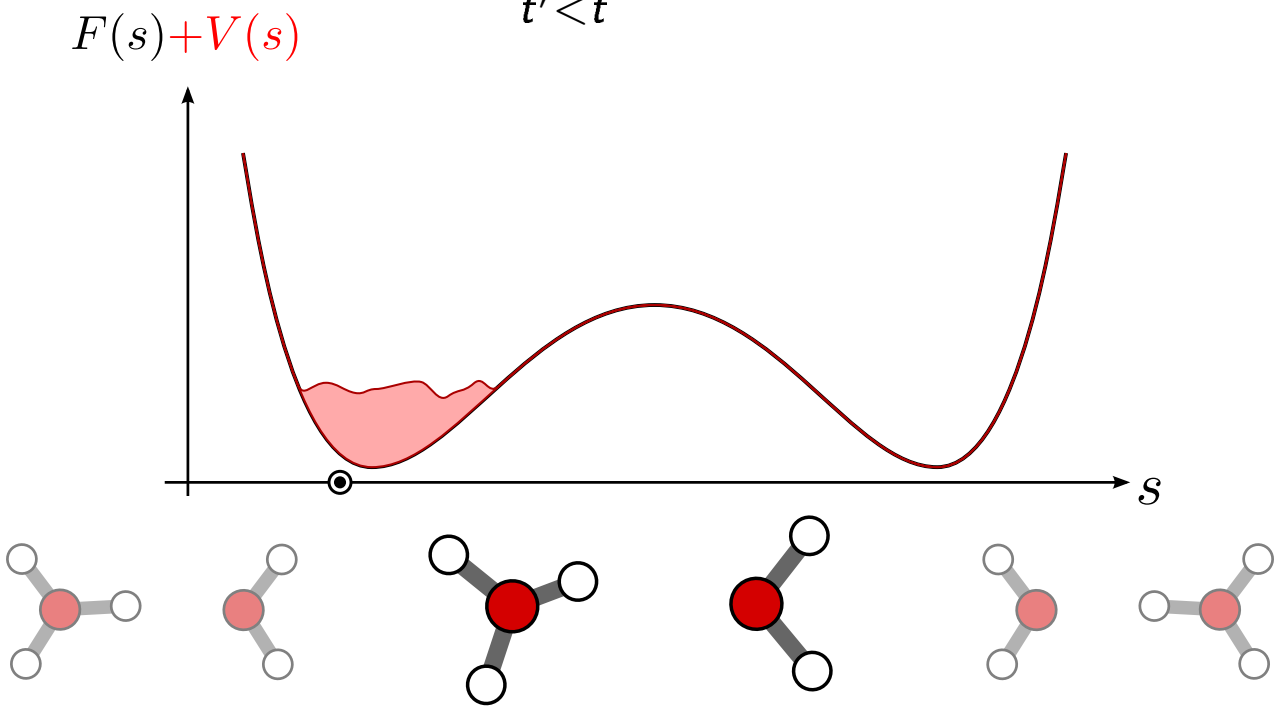


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

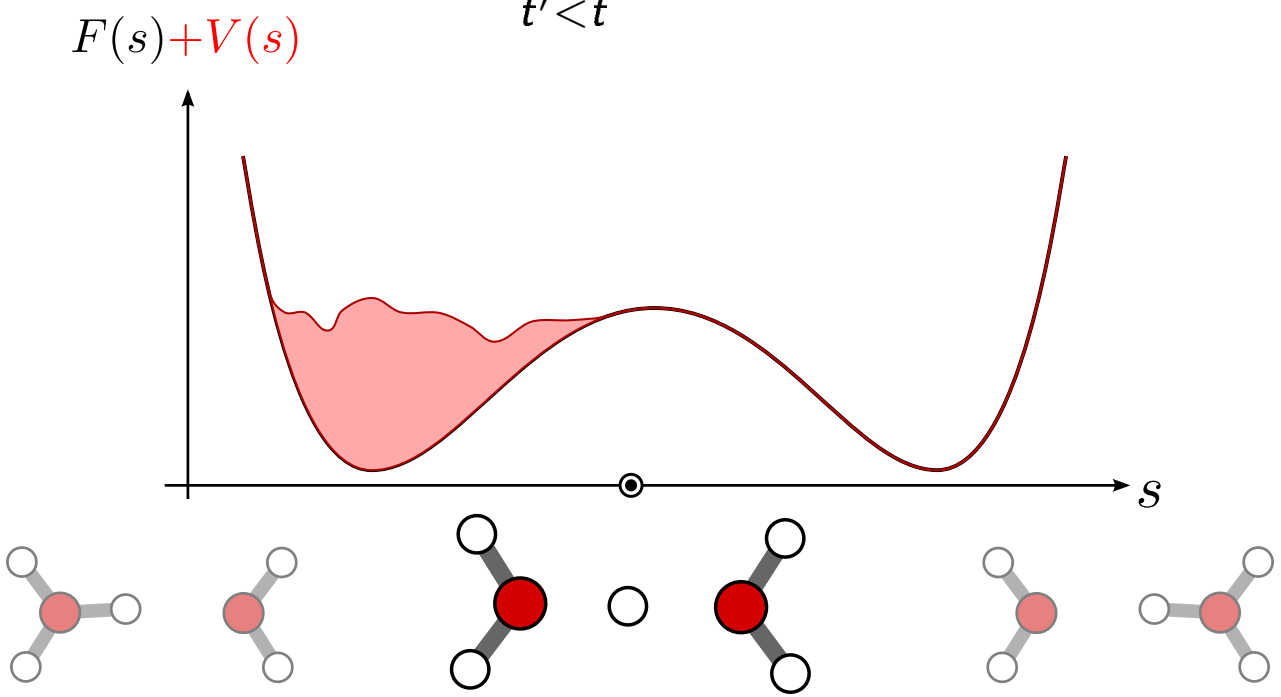


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

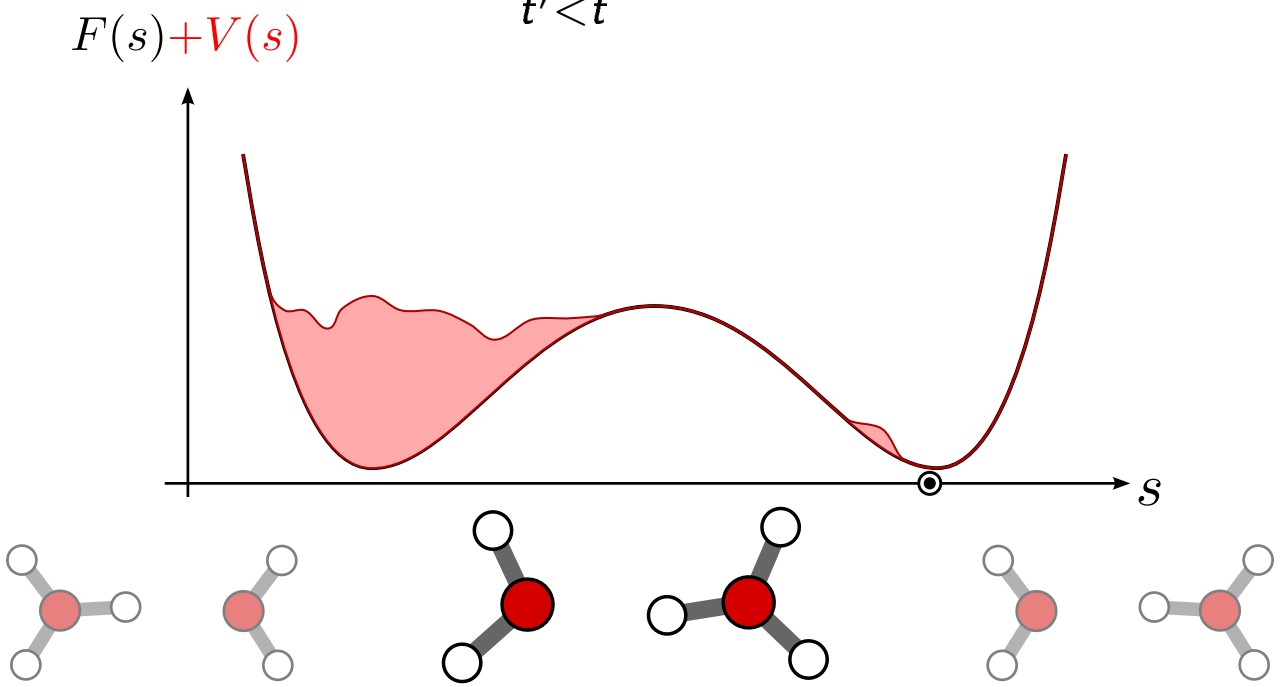


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Accelerating rare events

- When there are high **free energy barriers** it takes very long simulations to see interesting events
- Use a **bias** to discourage the system from staying in the same configuration and explore more efficiently (e.g. metadynamics)

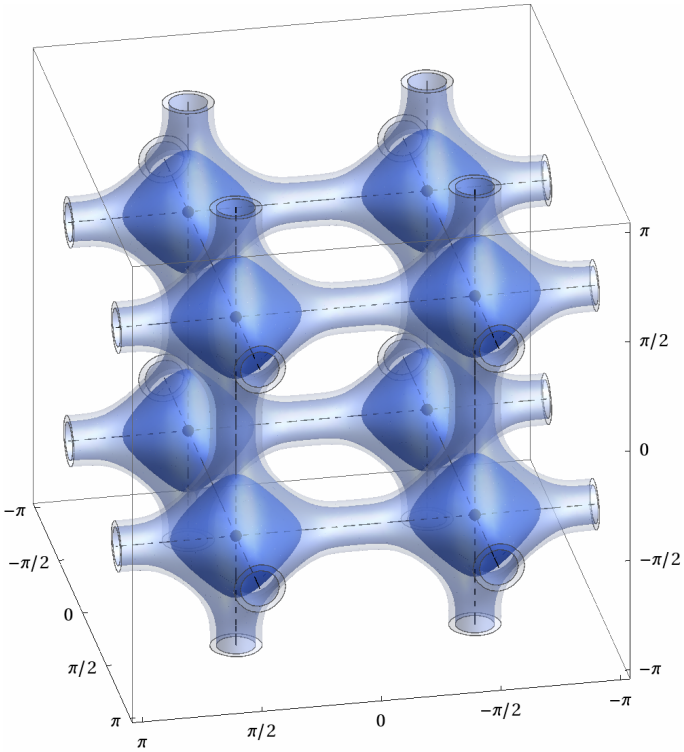
$$V(x,t) = \sum_{t'<t} w \exp -\frac{(s(x) - s_{t'})^2}{2\sigma^2}$$

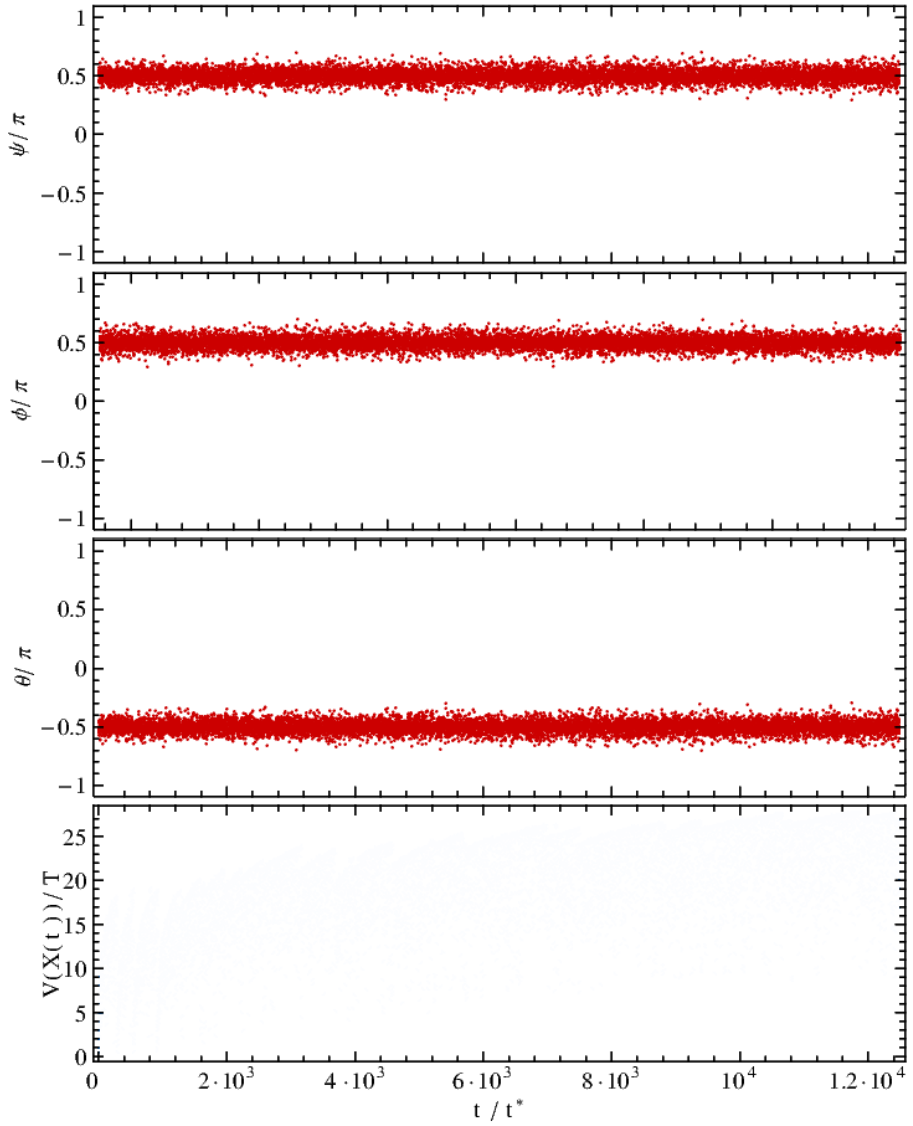


Laio, Parrinello, PNAS 2002, Huber, Torda, van Gunsteren, JCAMD 1994

Metadynamics and Hidden CVs

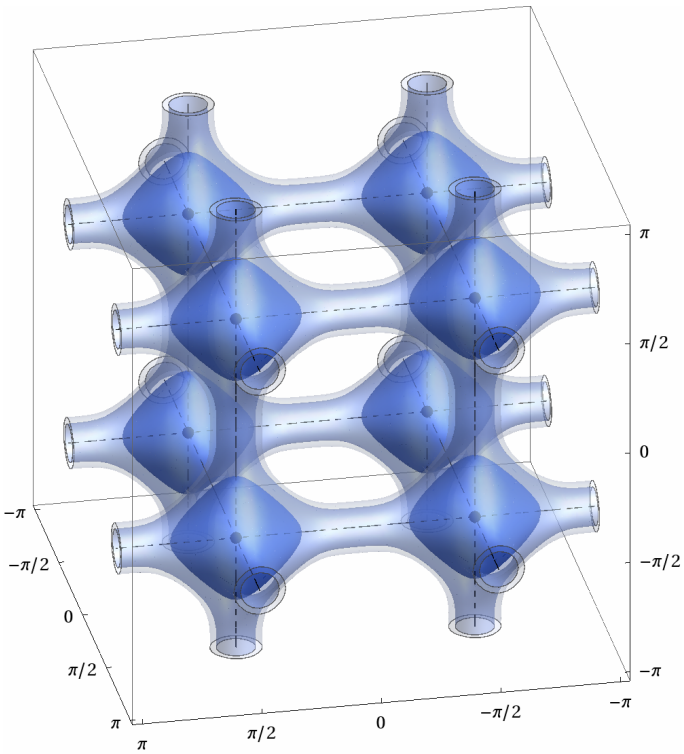
- 2d-metadynamics: the orthogonal degree of freedom is not sampled at all!
- **Field-overlap** using (2d!) **sketch-map** coordinates thoroughly samples in all directions

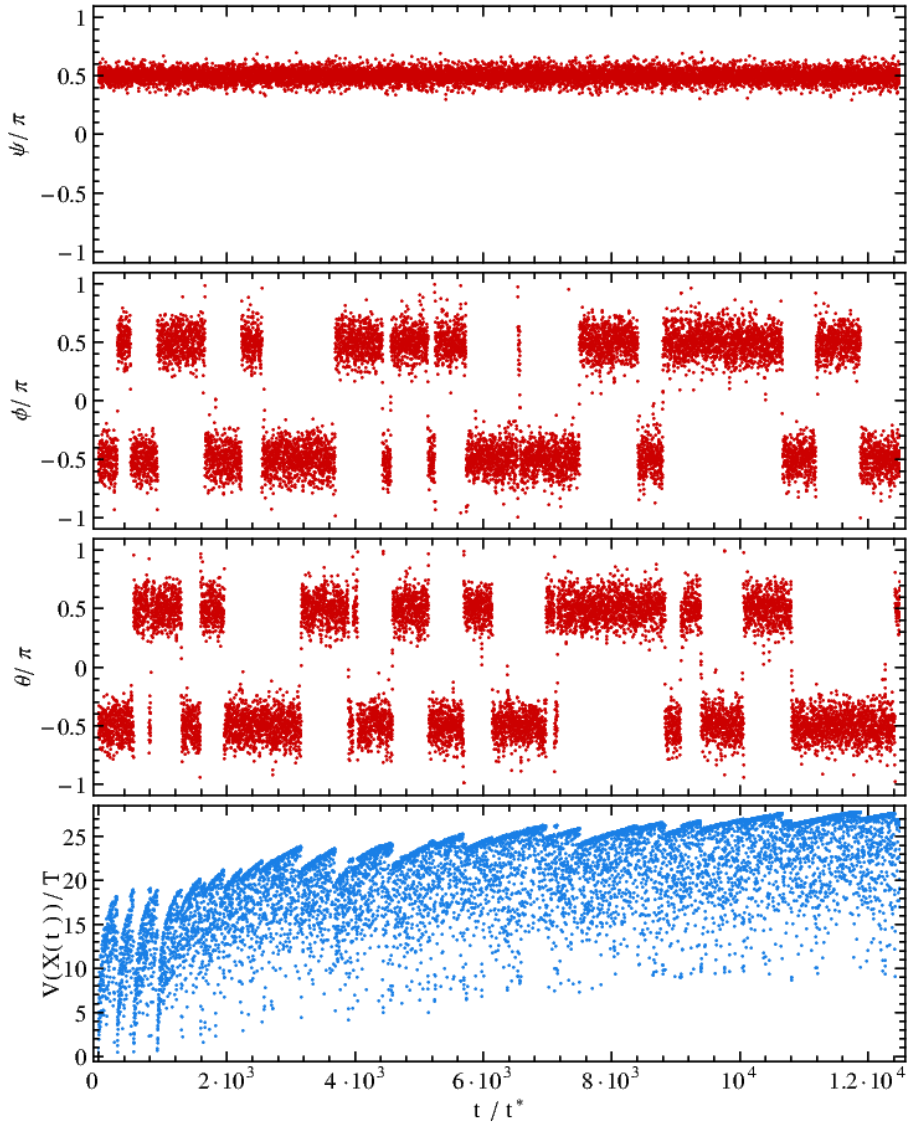




Metadynamics and Hidden CVs

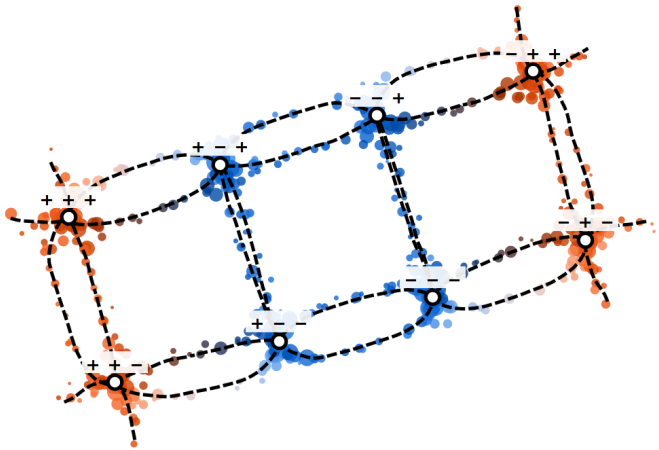
- 2d-metadynamics: the orthogonal degree of freedom is not sampled at all!
- **Field-overlap** using (2d!) **sketch-map** coordinates thoroughly samples in all directions

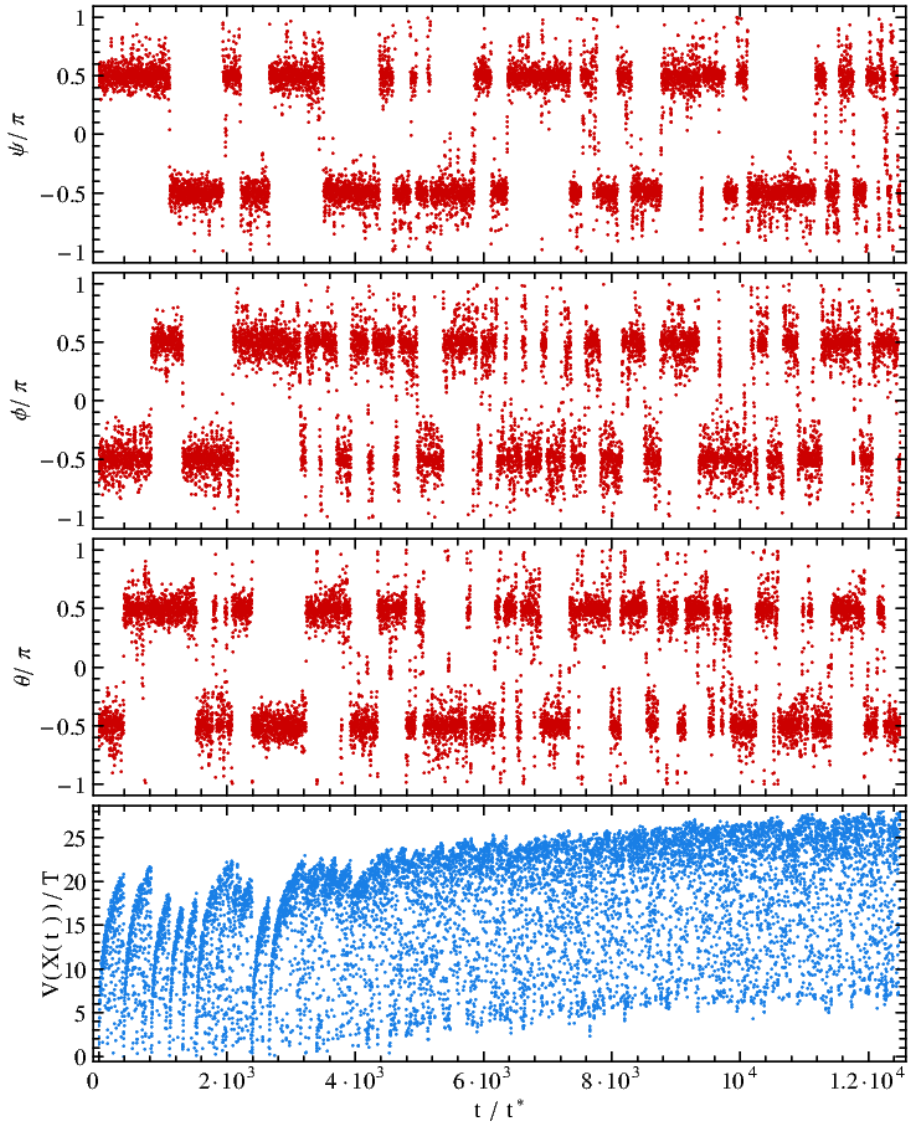




Metadynamics and Hidden CVs

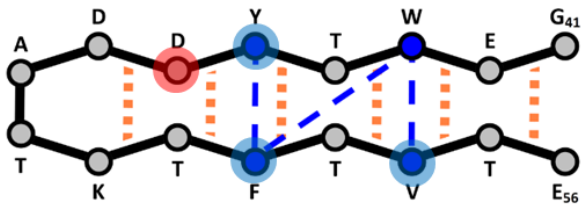
- 2d-metadynamics: the orthogonal degree of freedom is not sampled at all!
- **Field-overlap** using (2d!) **sketch-map** coordinates thoroughly samples in all directions



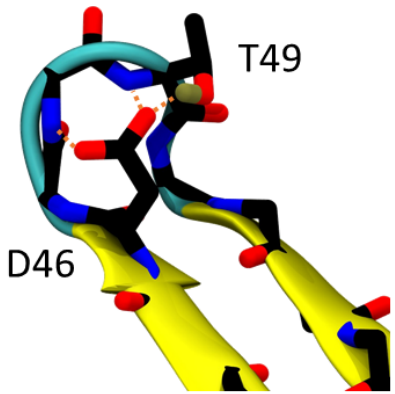


β-hairpin Mutants

- Sketch map can be used to compare the stability of different systems
- Consider a 16-residue hairpin-folding protein fragment
  - Many structures are locally stable other than the "folded" hairpin
  - We do a series of point mutations, and see how they affect stability



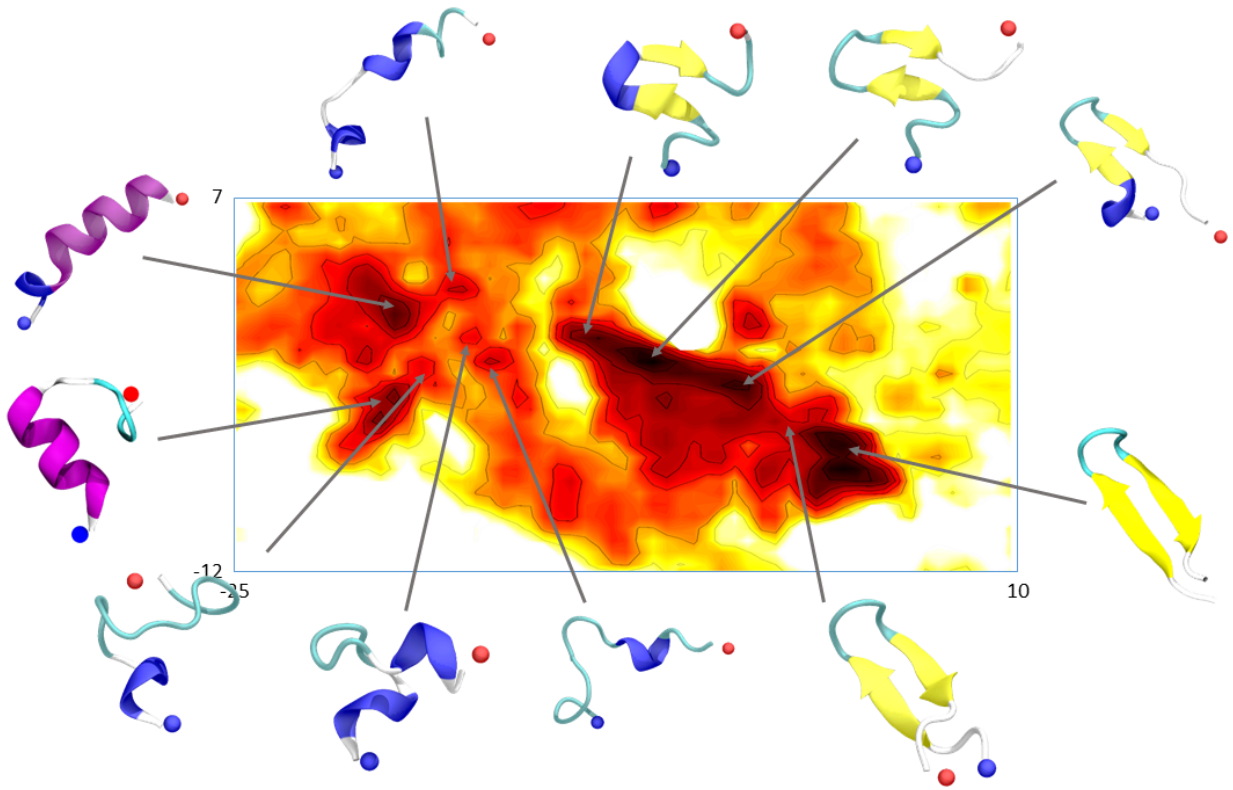
β-hairpin



Ardevol, Tribello, Ceriotti, Parrinello, JCTC (2015)

β-hairpin Mutants

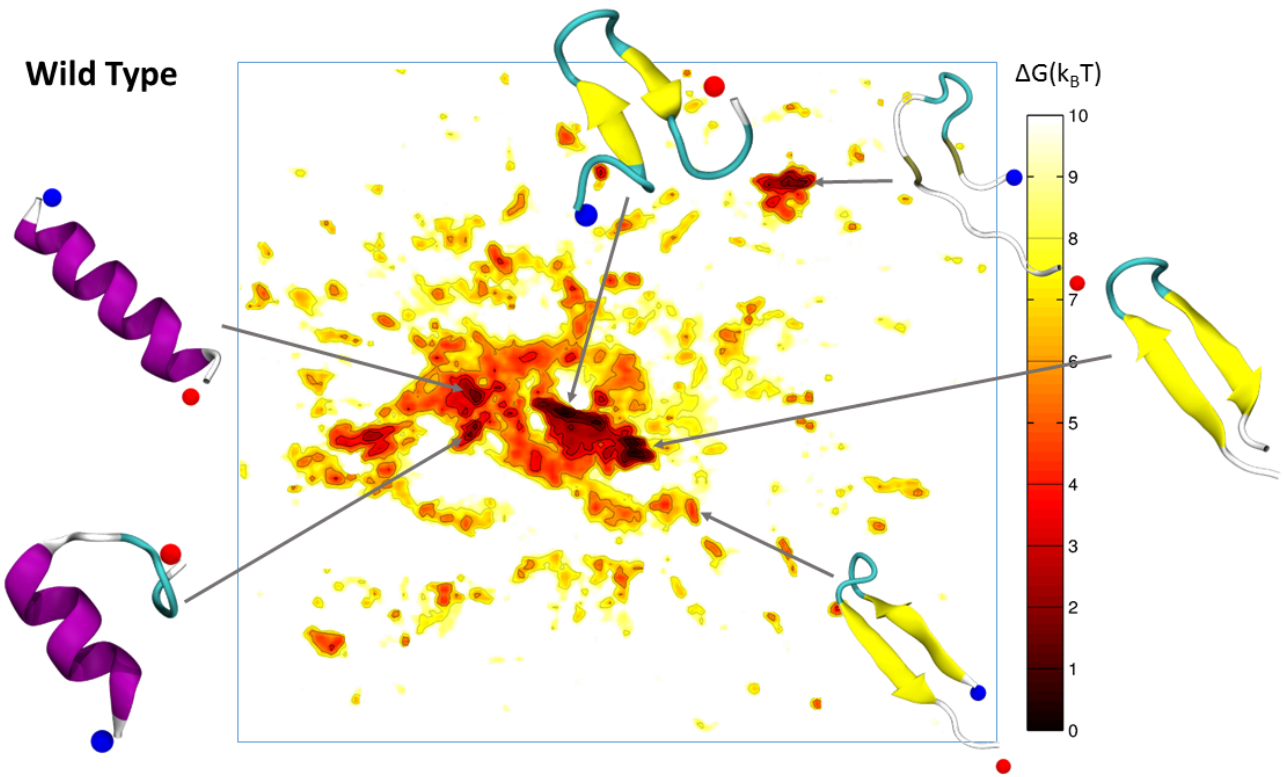
- Sketch map can be used to compare the stability of different systems
- Consider a 16-residue hairpin-folding protein fragment
  - Many structures are locally stable other than the "folded" hairpin
  - We do a series of point mutations, and see how they affect stability



Ardevol, Tribello, Ceriotti, Parrinello, JCTC (2015)

β-hairpin Mutants

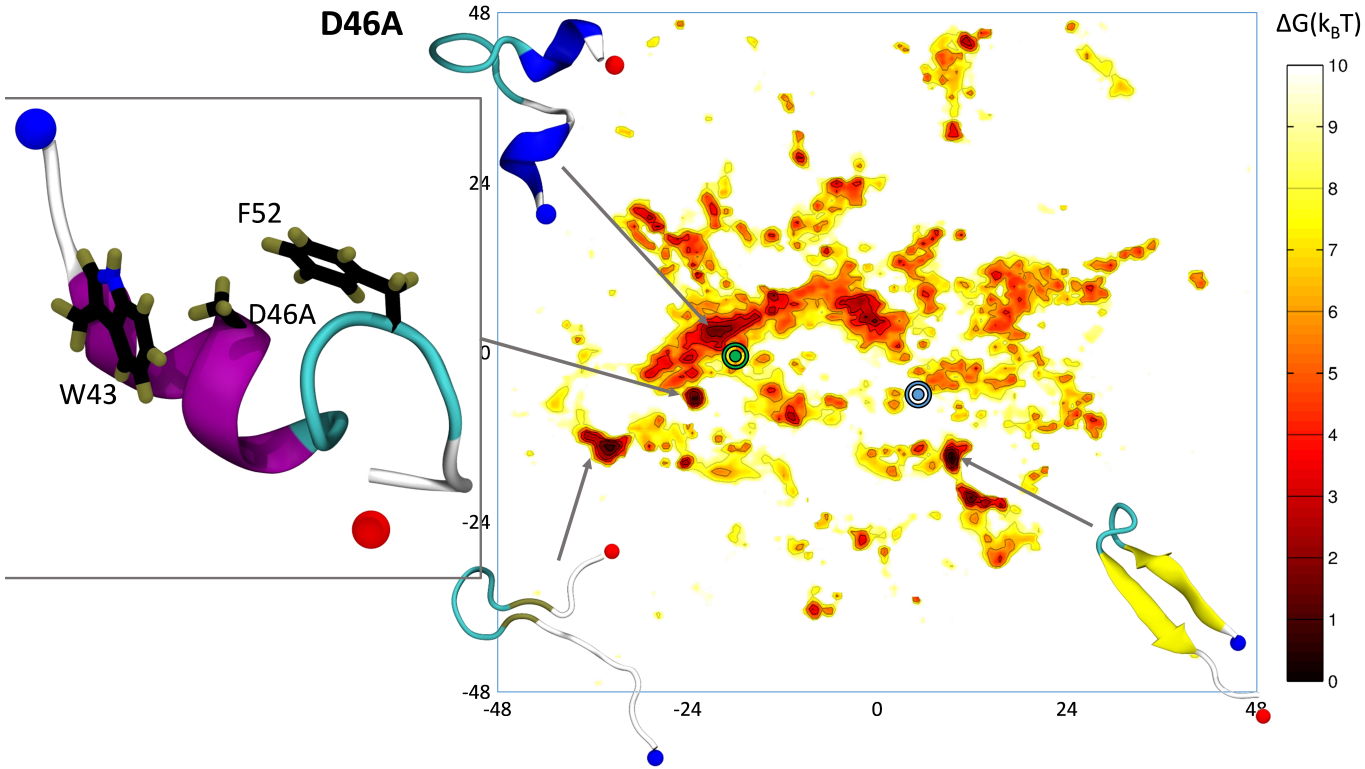
- Sketch map can be used to compare the stability of different systems
- Consider a 16-residue hairpin-folding protein fragment
  - Many structures are locally stable other than the "folded" hairpin
  - We do a series of point mutations, and see how they affect stability



Ardevol, Tribello, Ceriotti, Parrinello, JCTC (2015)

β-hairpin Mutants

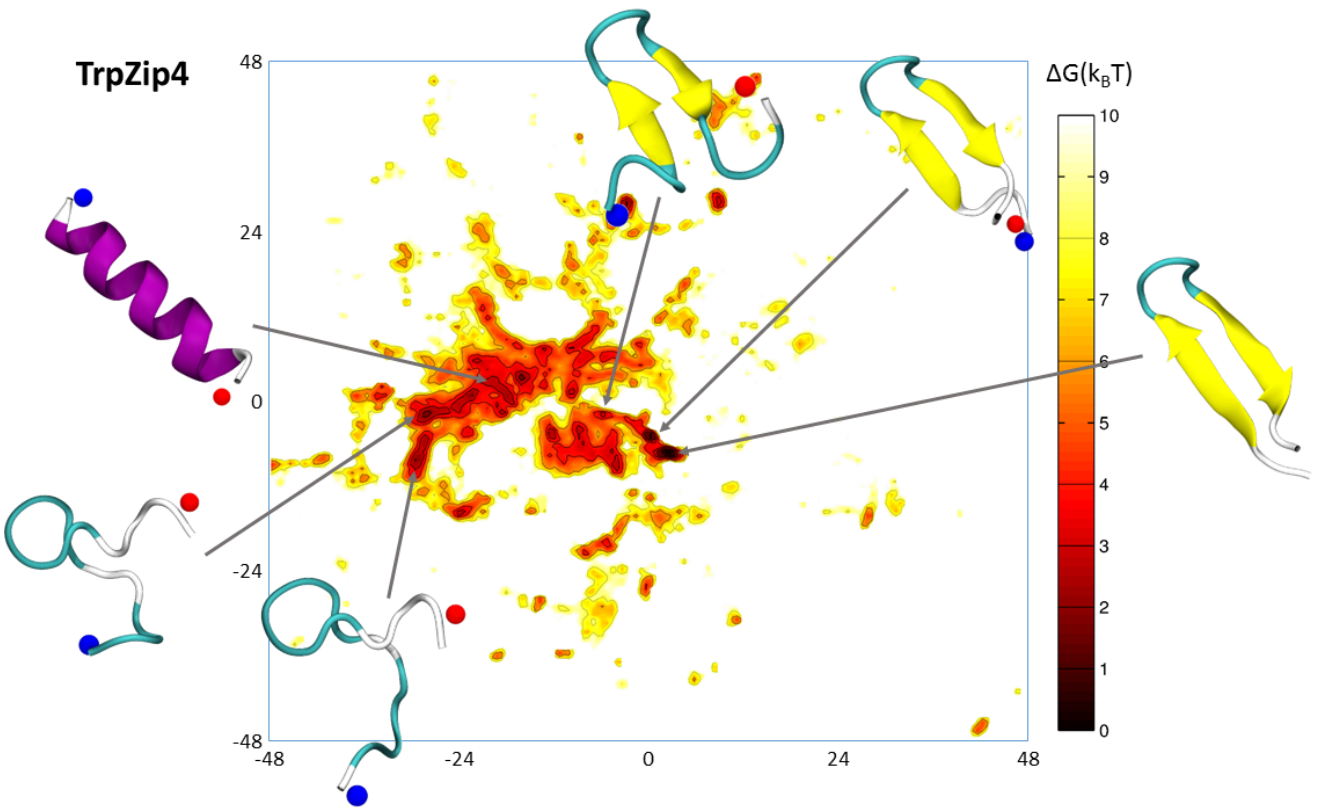
- Sketch map can be used to compare the stability of different systems
- Consider a 16-residue hairpin-folding protein fragment
  - Many structures are locally stable other than the "folded" hairpin
  - We do a series of point mutations, and see how they affect stability



Ardevol, Tribello, Ceriotti, Parrinello, JCTC (2015)

β-hairpin Mutants

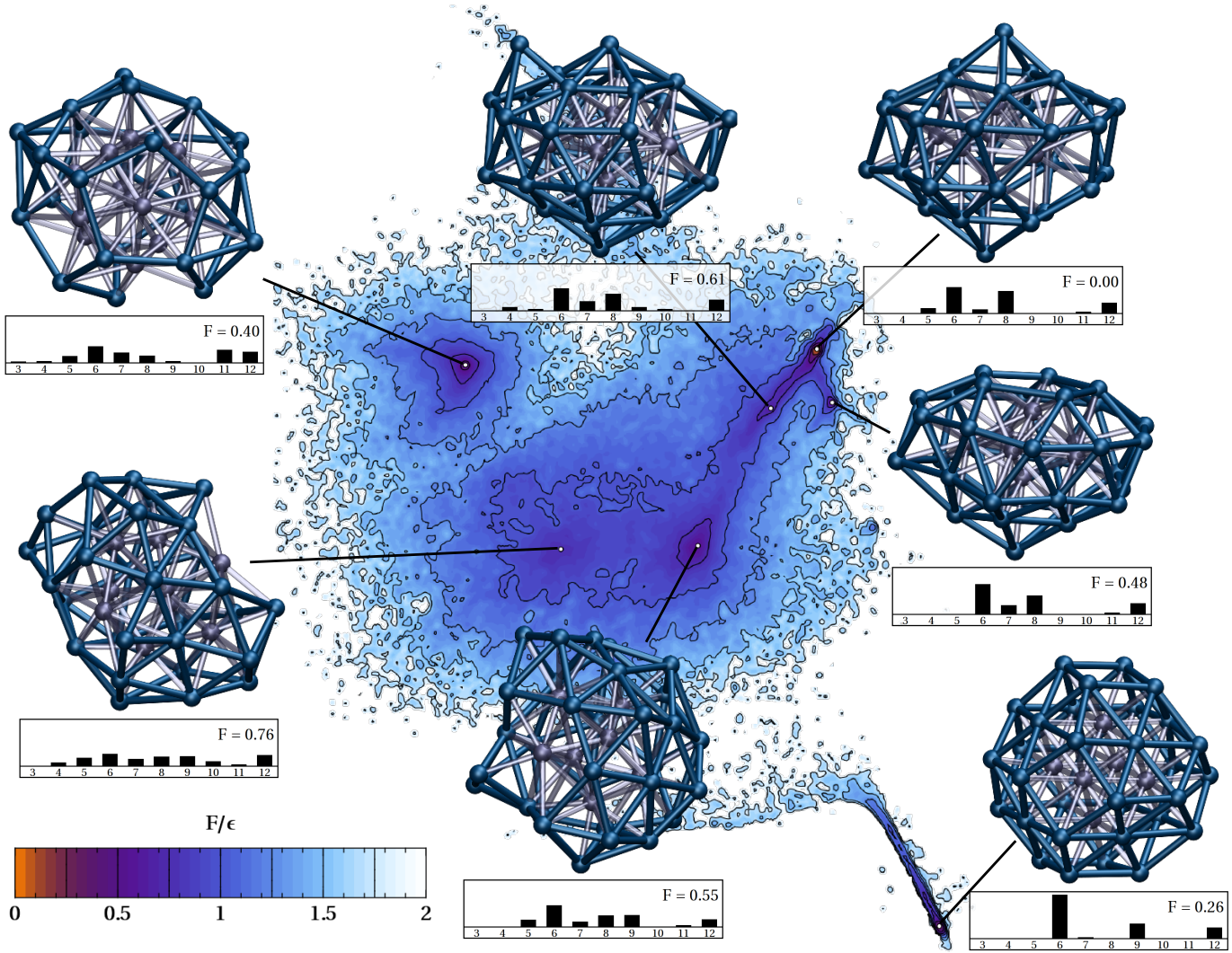
- Sketch map can be used to compare the stability of different systems
- Consider a 16-residue hairpin-folding protein fragment
  - Many structures are locally stable other than the "folded" hairpin
  - We do a series of point mutations, and see how they affect stability



Ardevol, Tribello, Ceriotti, Parrinello, JCTC (2015)

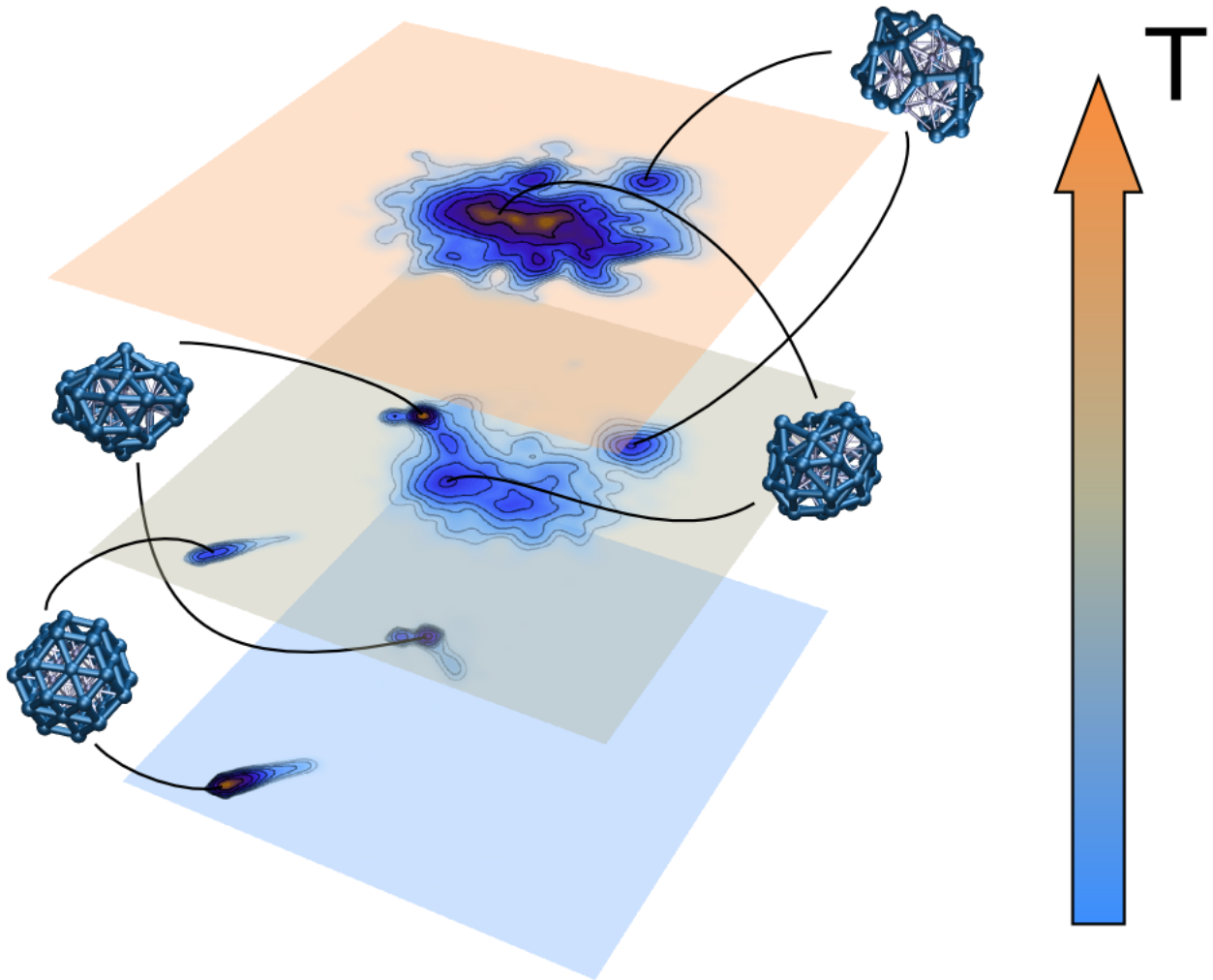
From clusters to defects in the bulk

- Start building a map for a Lennard-Jones cluster
- The same map describes the cluster across phase transitions
- ... and can even recognize the nature of defects in a bulk system!



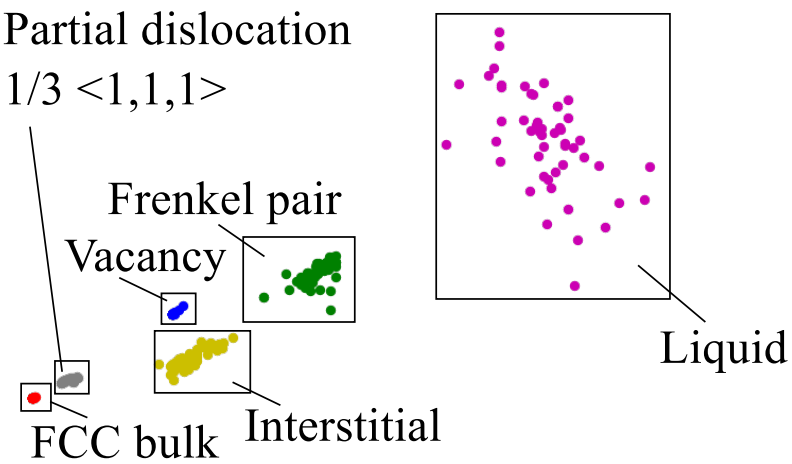
From clusters to defects in the bulk

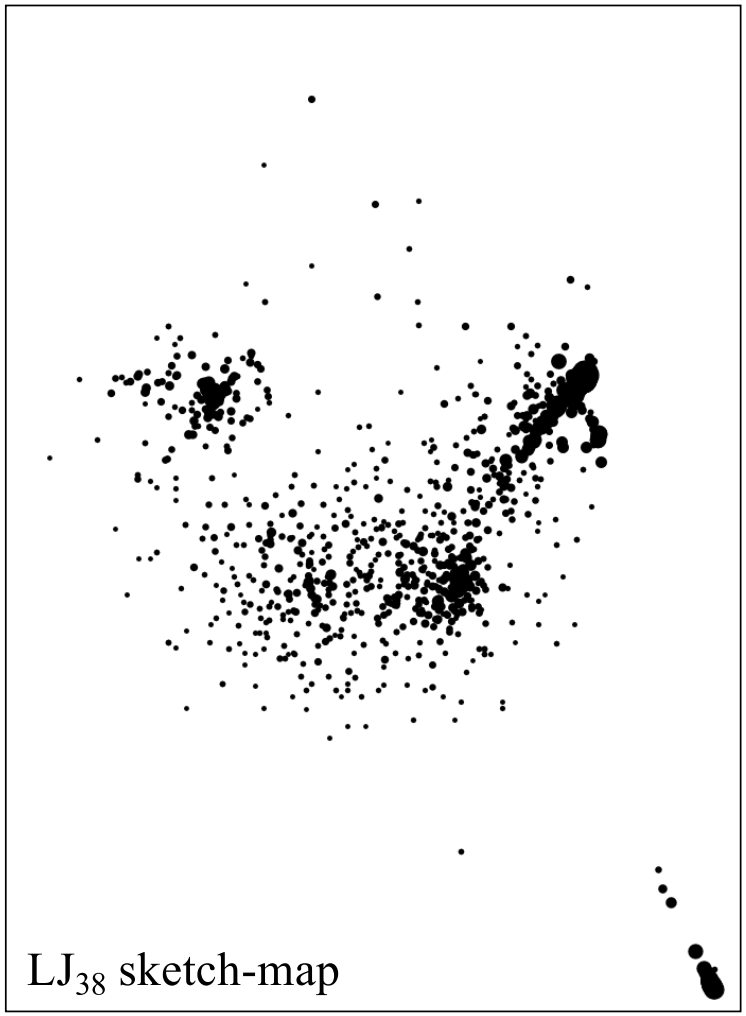
- Start building a map for a Lennard-Jones cluster
- The same map describes the cluster across phase transitions
- ... and can even recognize the nature of defects in a bulk system!



From clusters to defects in the bulk

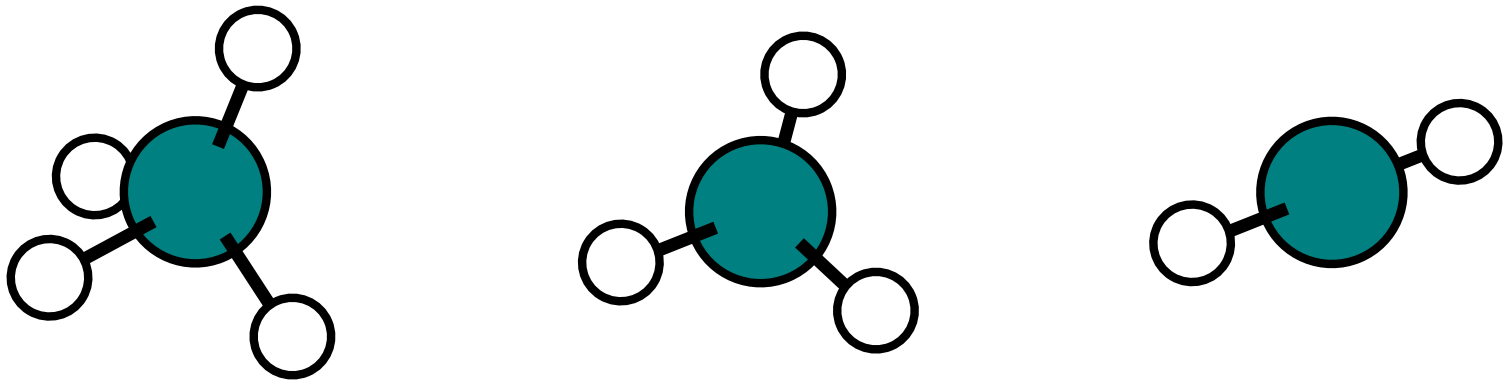
- Start building a map for a Lennard-Jones cluster
- The same map describes the cluster across phase transitions
- ... and can even recognize the nature of defects in a bulk system!





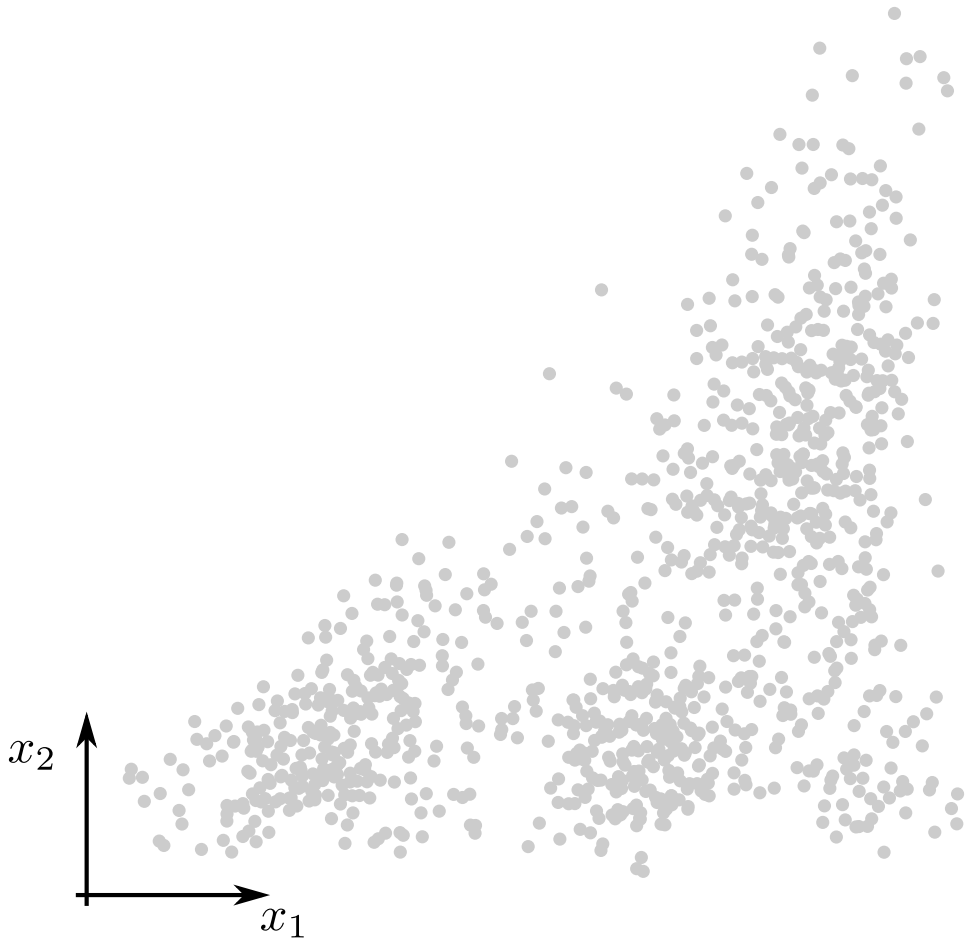
Recognizing molecular patterns

- We still need an effective high-dimensional description to start with
- "Chemical intuition" builds on recognizing recurring patterns in atomic configurations
- Automatic scheme to single out structural motifs in atomistic simulations



Probabilistic Analysis of Molecular Motifs

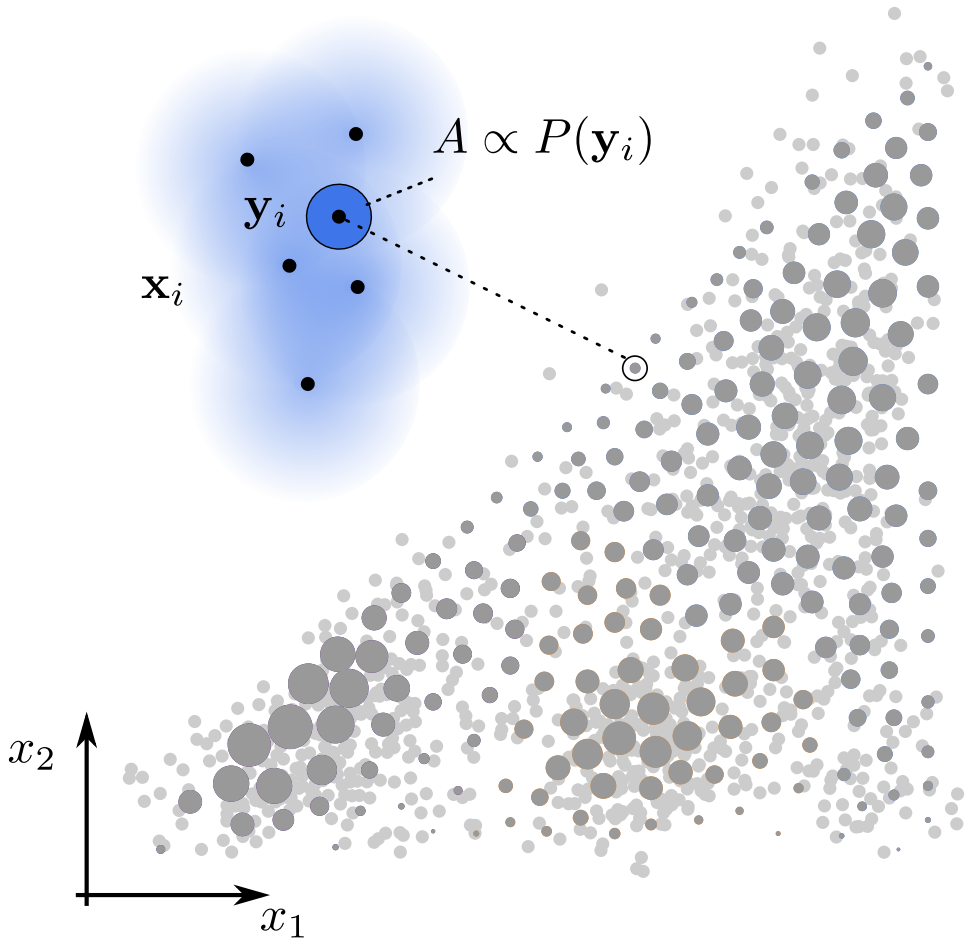
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

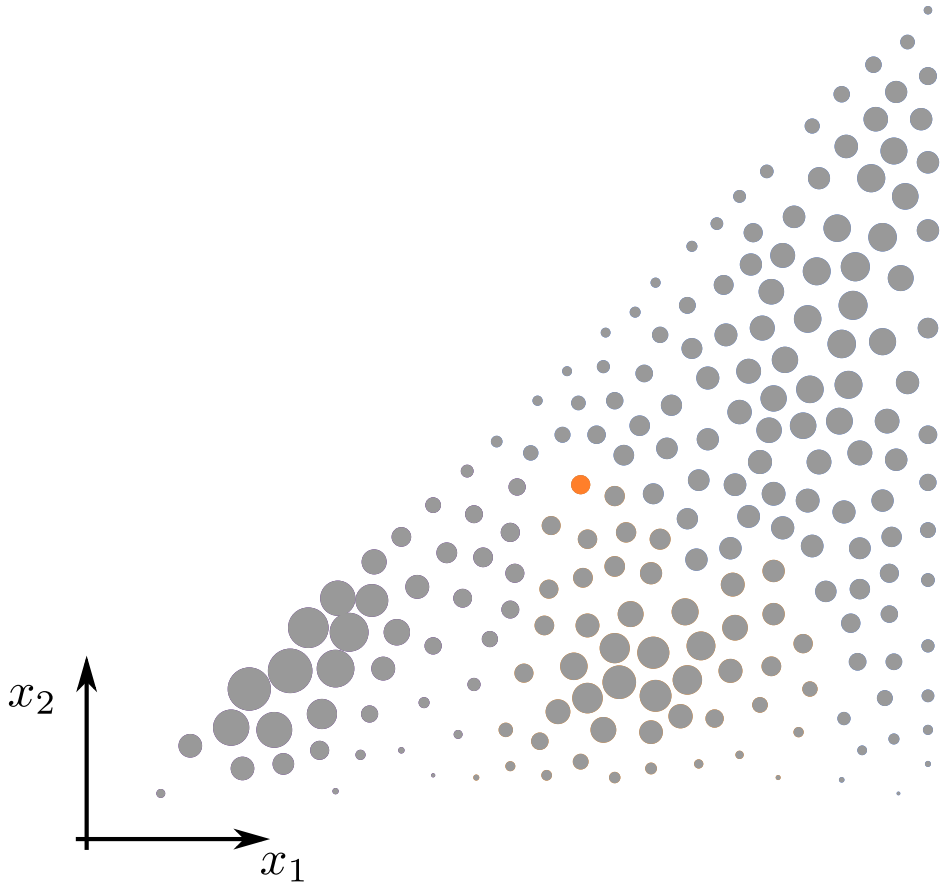
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

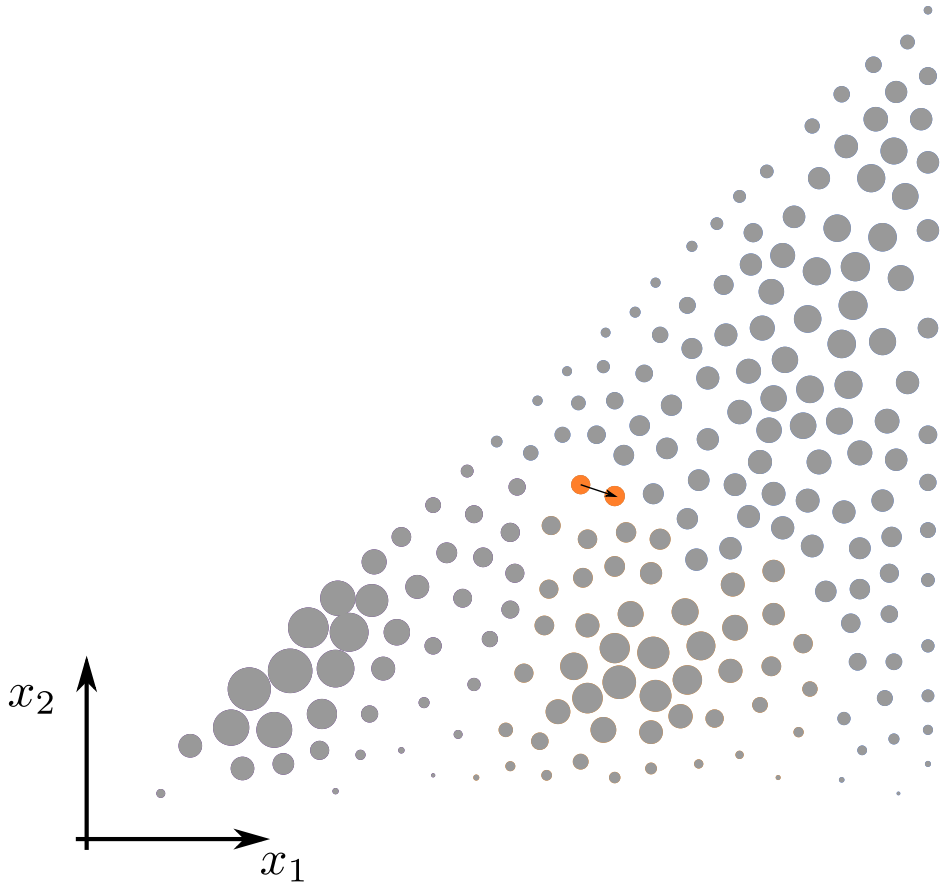
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

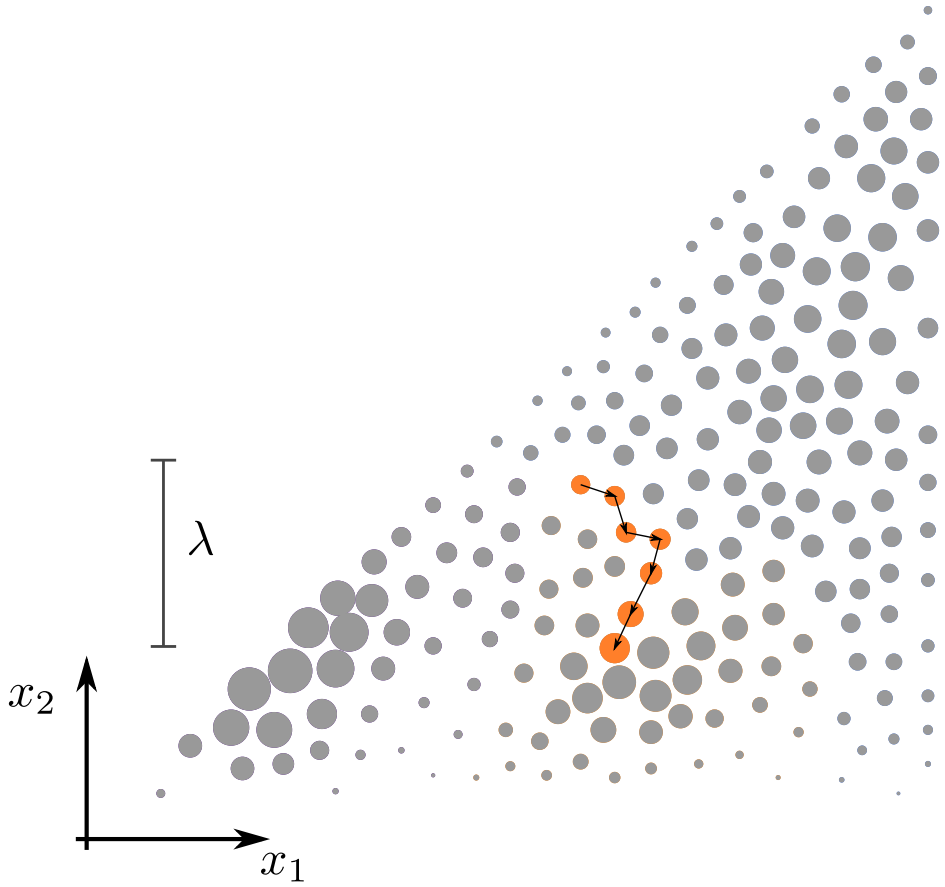
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

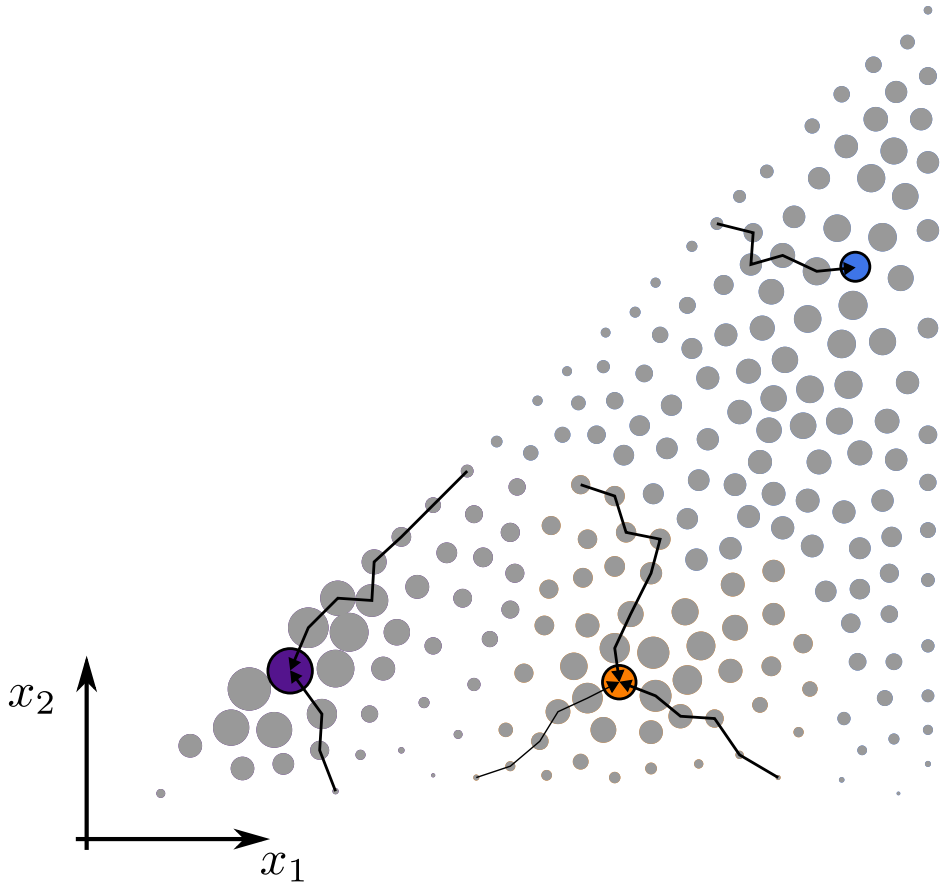
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

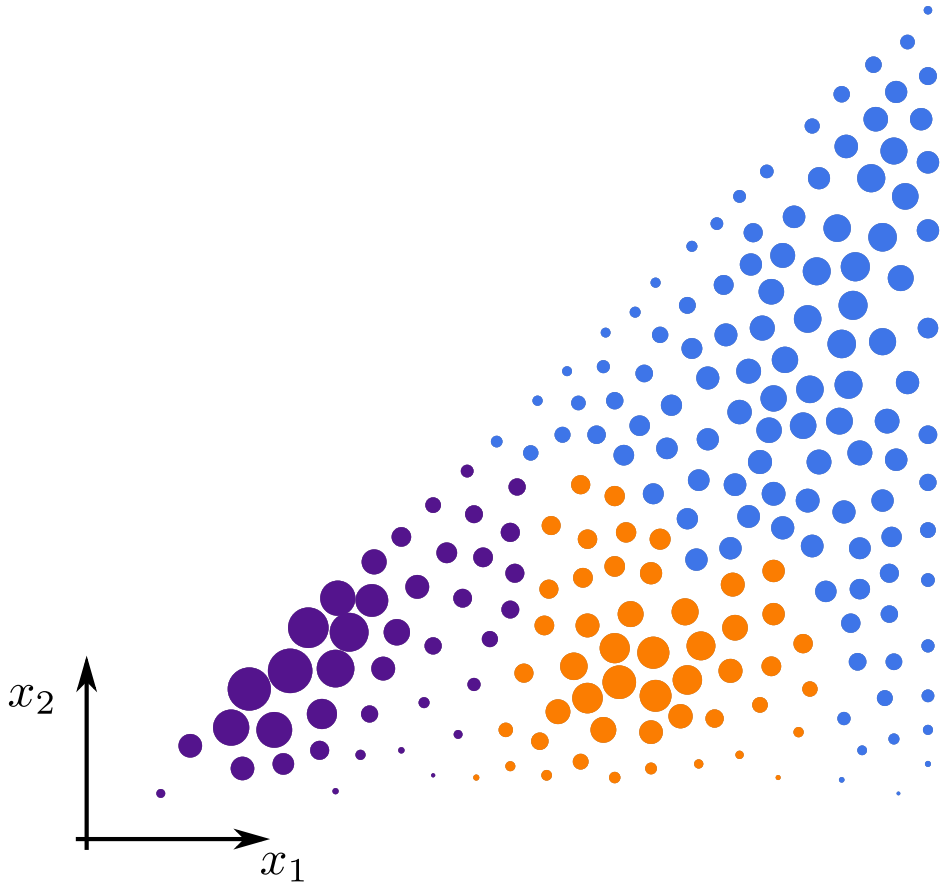
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

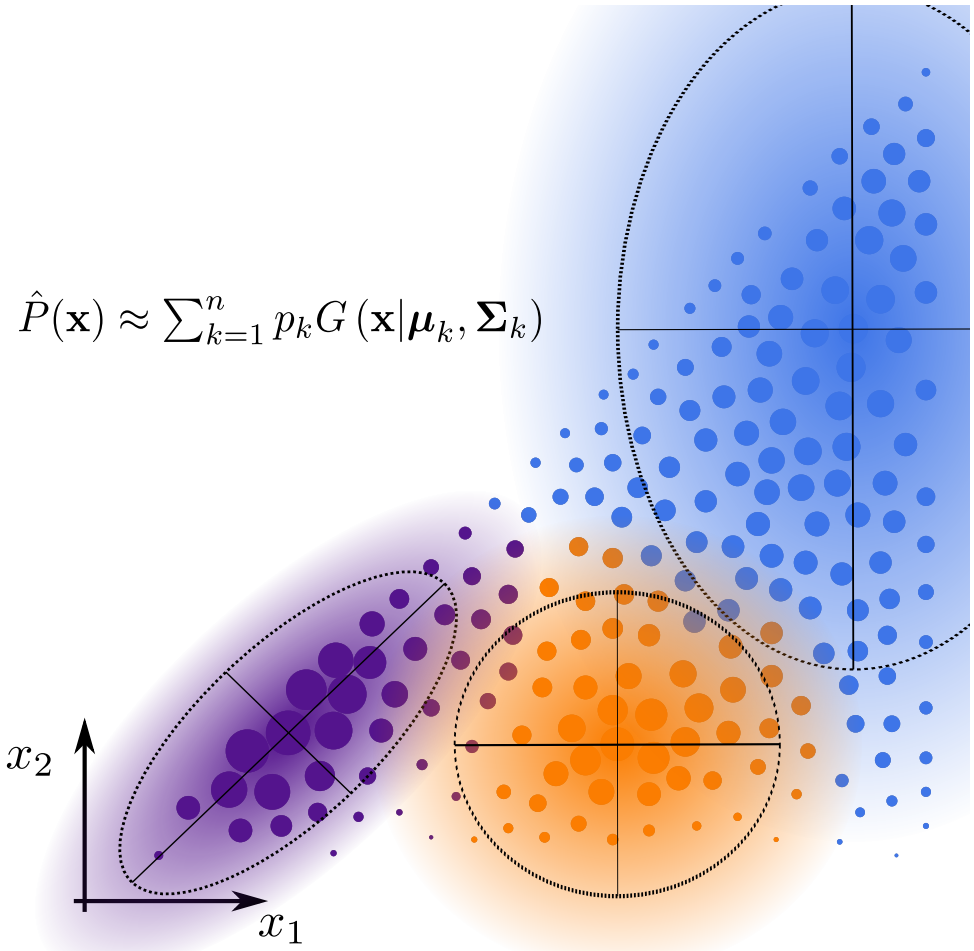
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

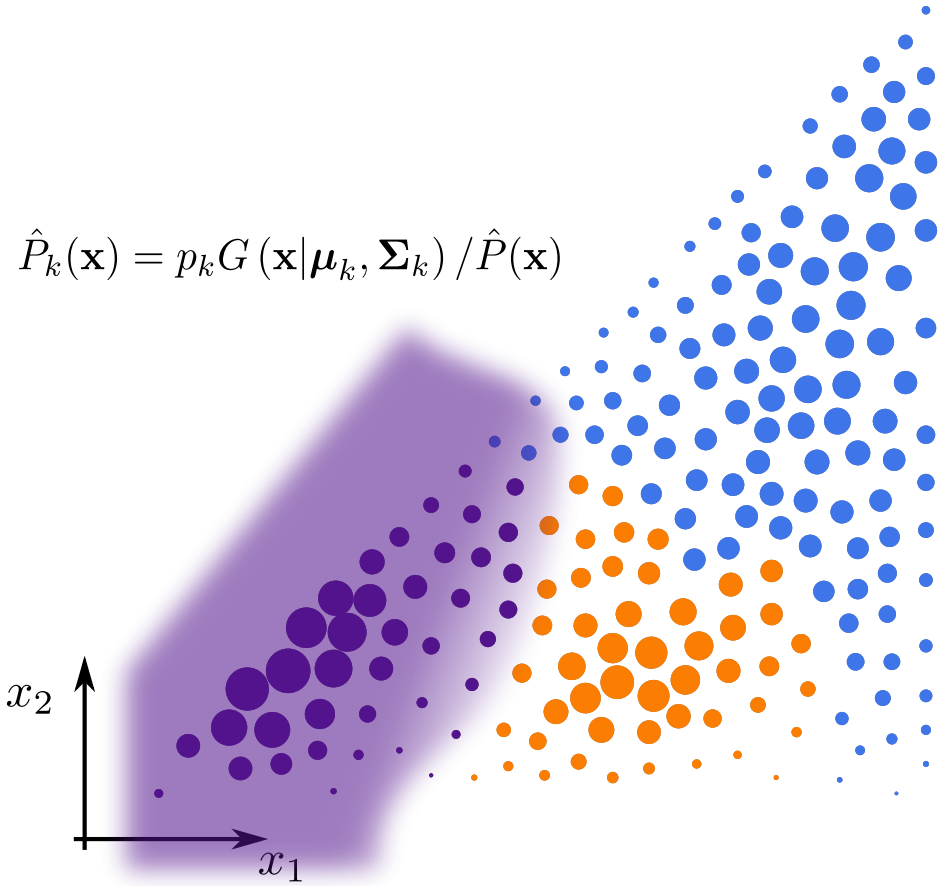
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

Probabilistic Analysis of Molecular Motifs

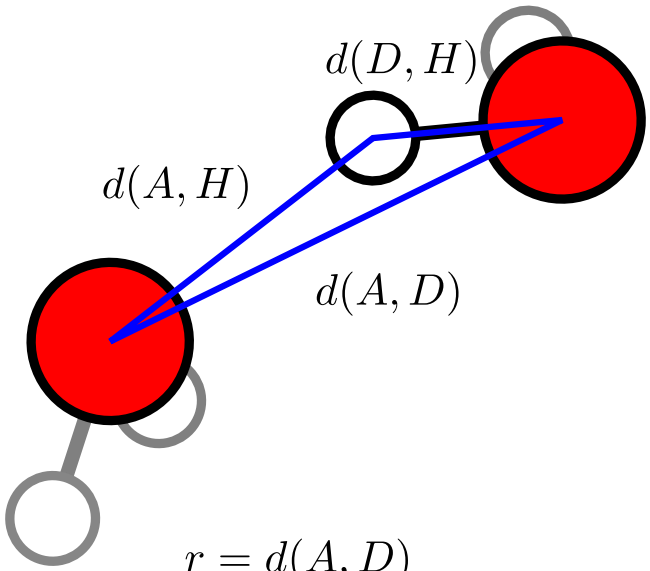
- Evaluate the probability distribution of molecular structures
- Cluster it around the modes of the distribution
- Naturally gives a fuzzy and continuous partitioning of configuration space



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

An Agnostic Definition of the H-Bond

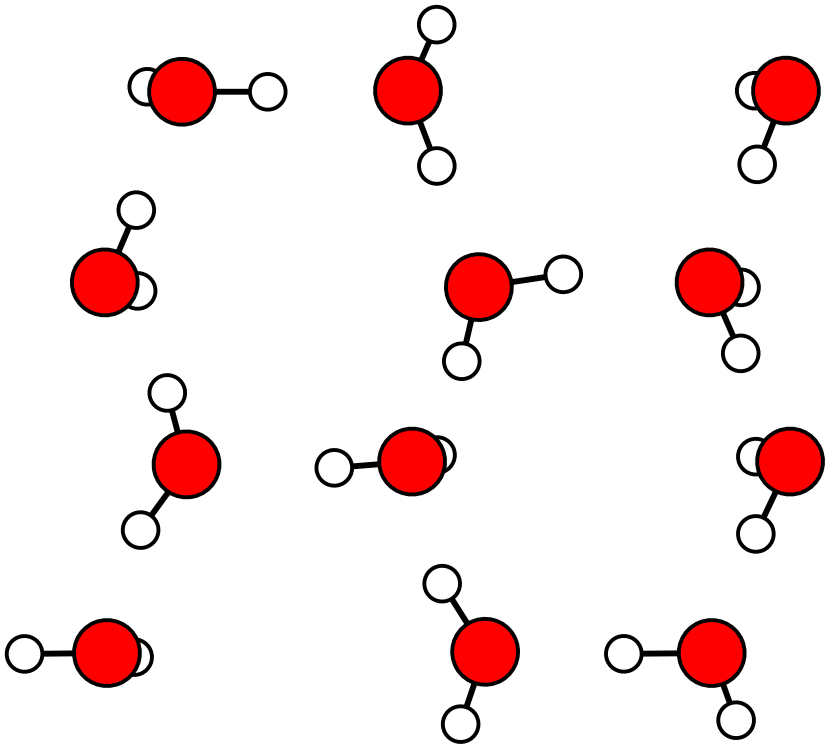
- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



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

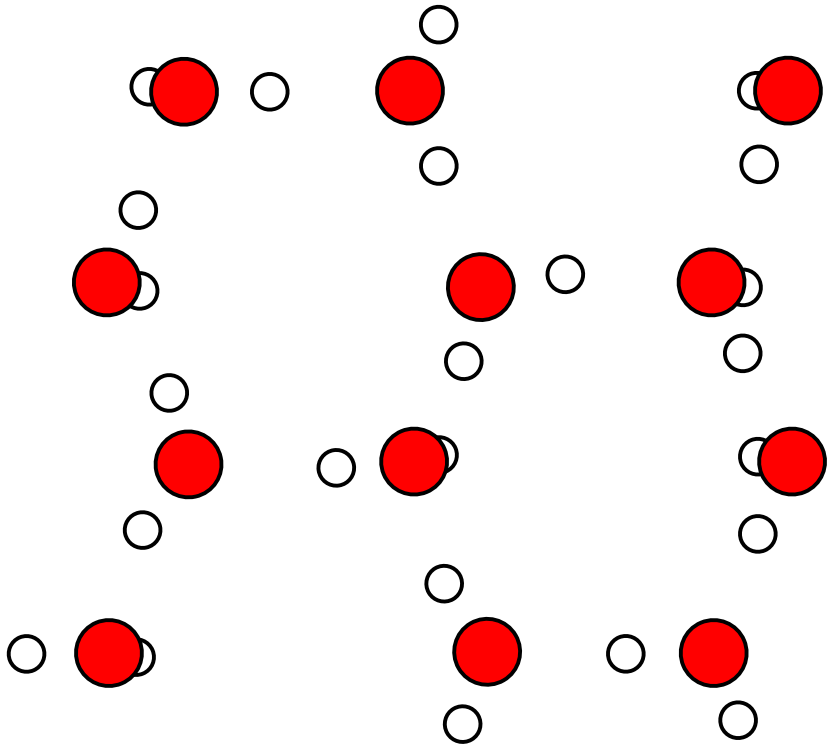
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



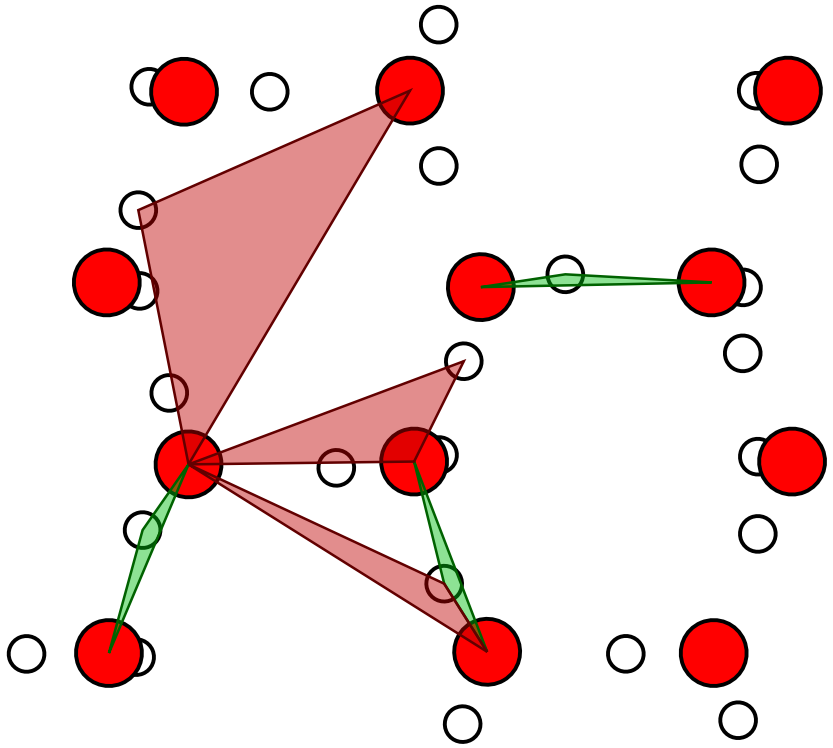
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



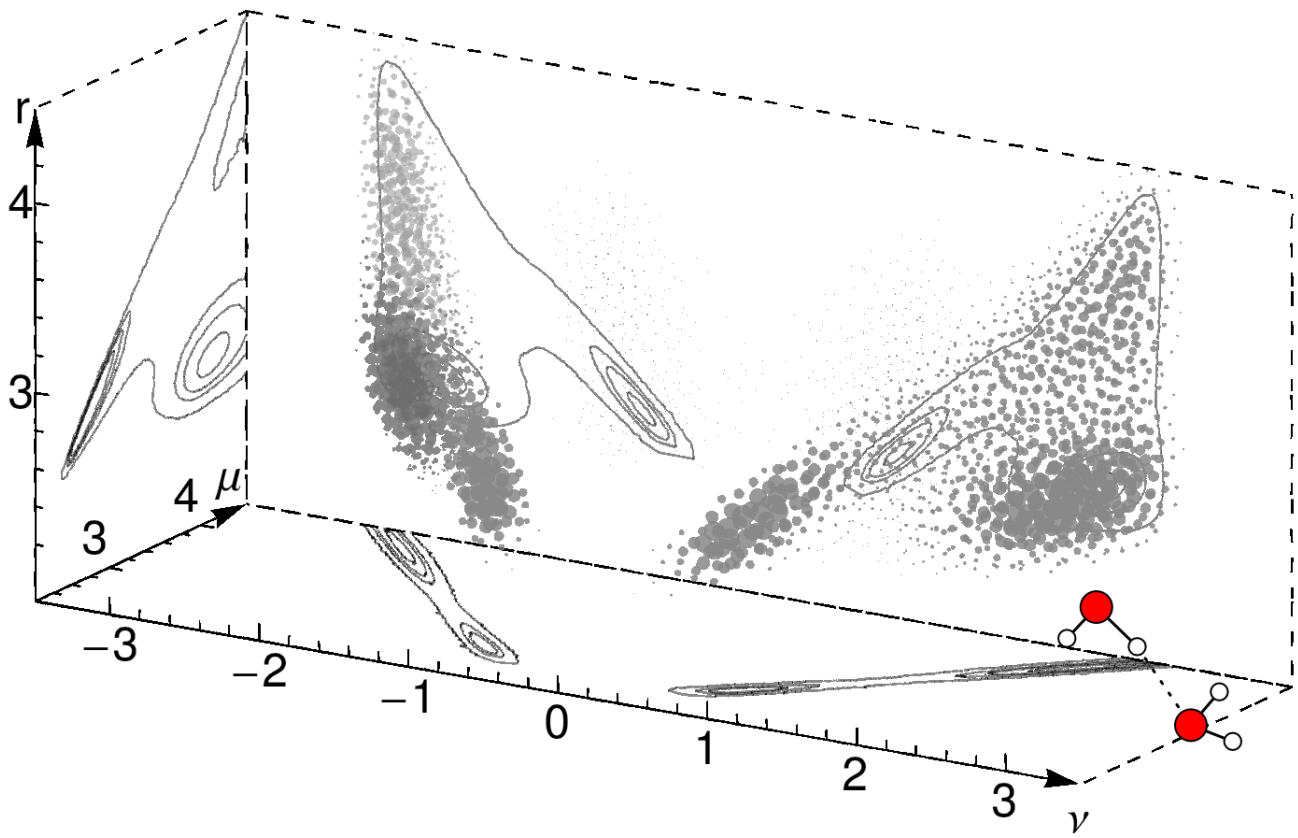
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



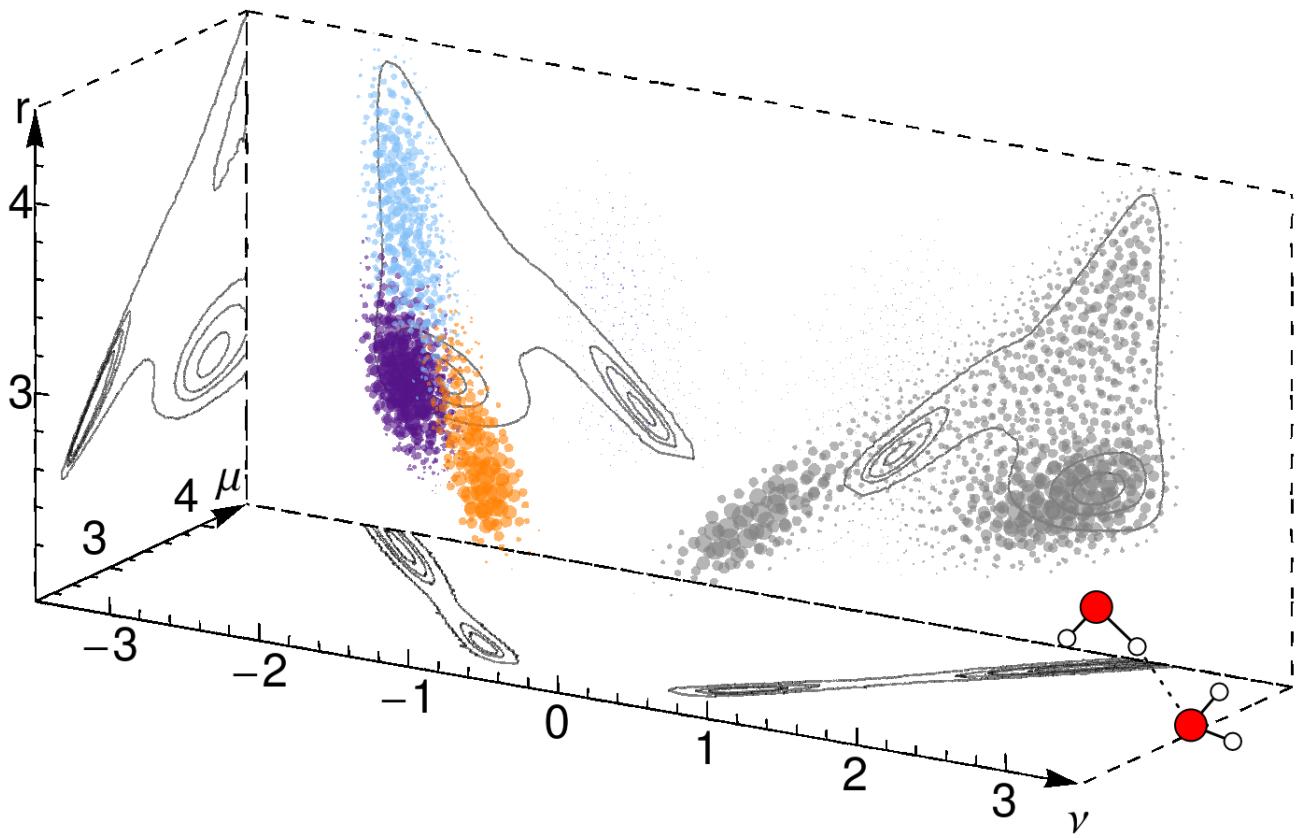
An Agnostic Definition of the H-Bond

- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



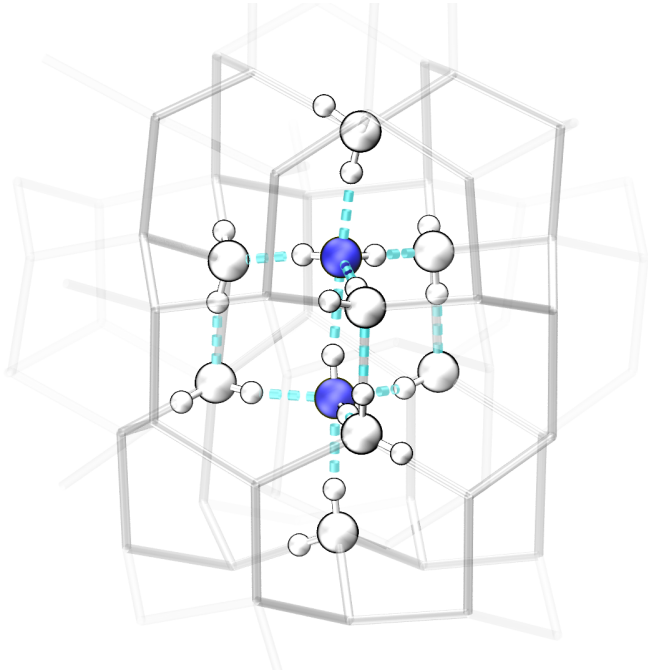
An Agnostic Definition of the H-Bond

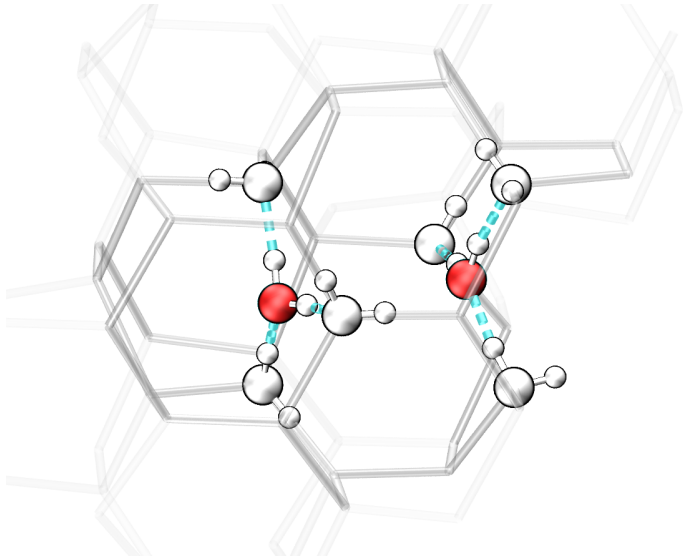
- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy



An Agnostic Definition of the H-Bond

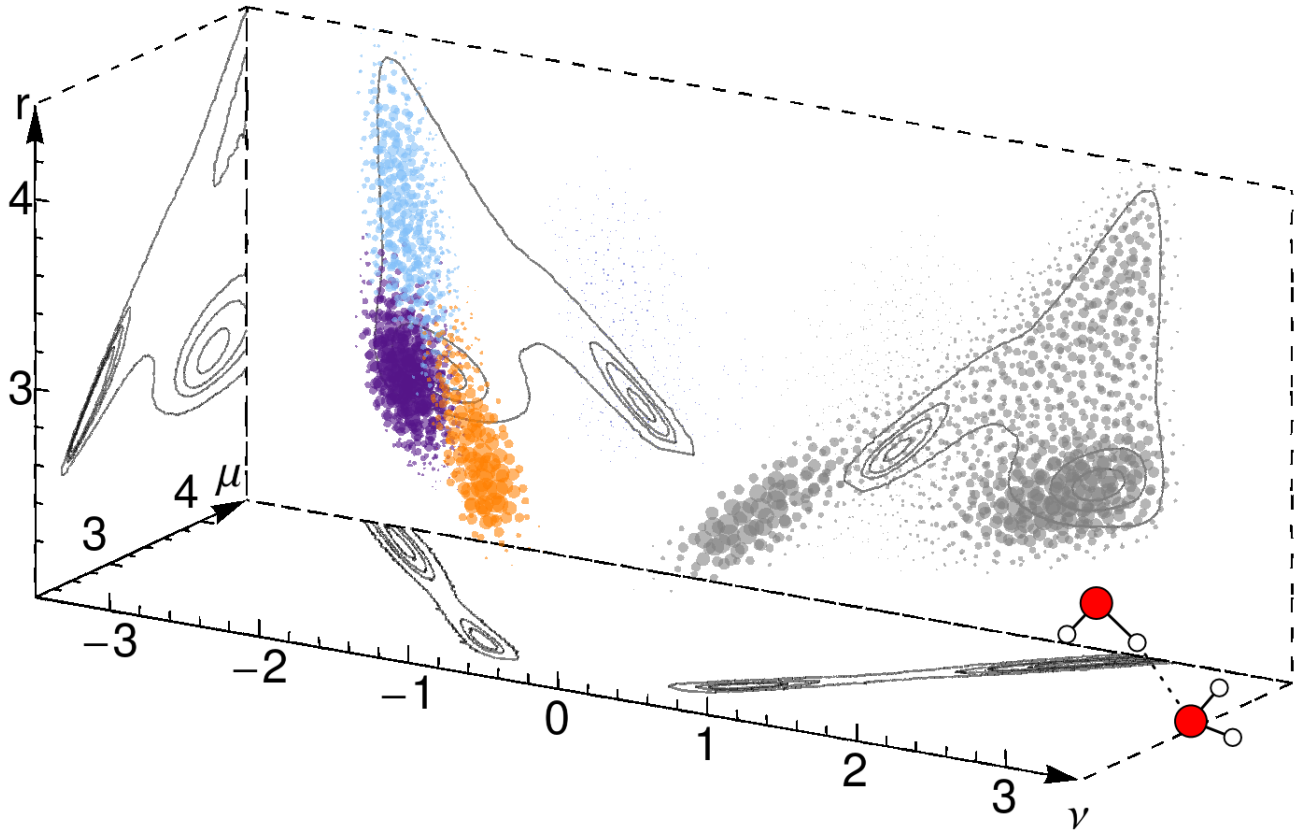
- Most general description of a H-bond geometry: 3 distances
- PAMM recognizes multiple modes - one corresponds to the H-bond
- PAMM H-bond fingerprints can be used as HB counts, but are adaptive, unbiased and fuzzy





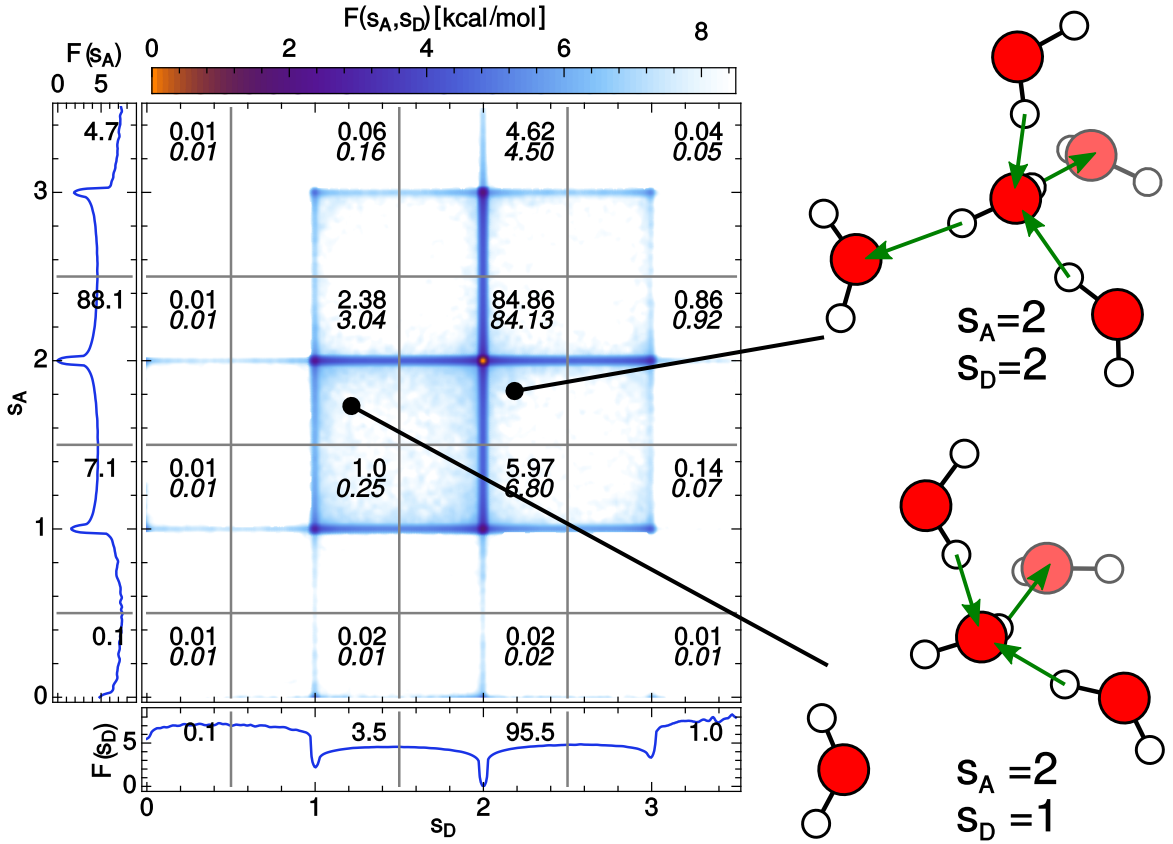
A H-bonding Environment Map

- PAMM gives for each (D, H, A) triplet a H-bond count s_{DHA}
- Define integrated H-bond counts, e.g. $s_{\mathrm{D}} = \sum_{\mathrm{H,A}} s_{\mathrm{DHA}}$
- An "identity card" of H-bonding
- Nuclear quantum effects trigger transient proton jumps



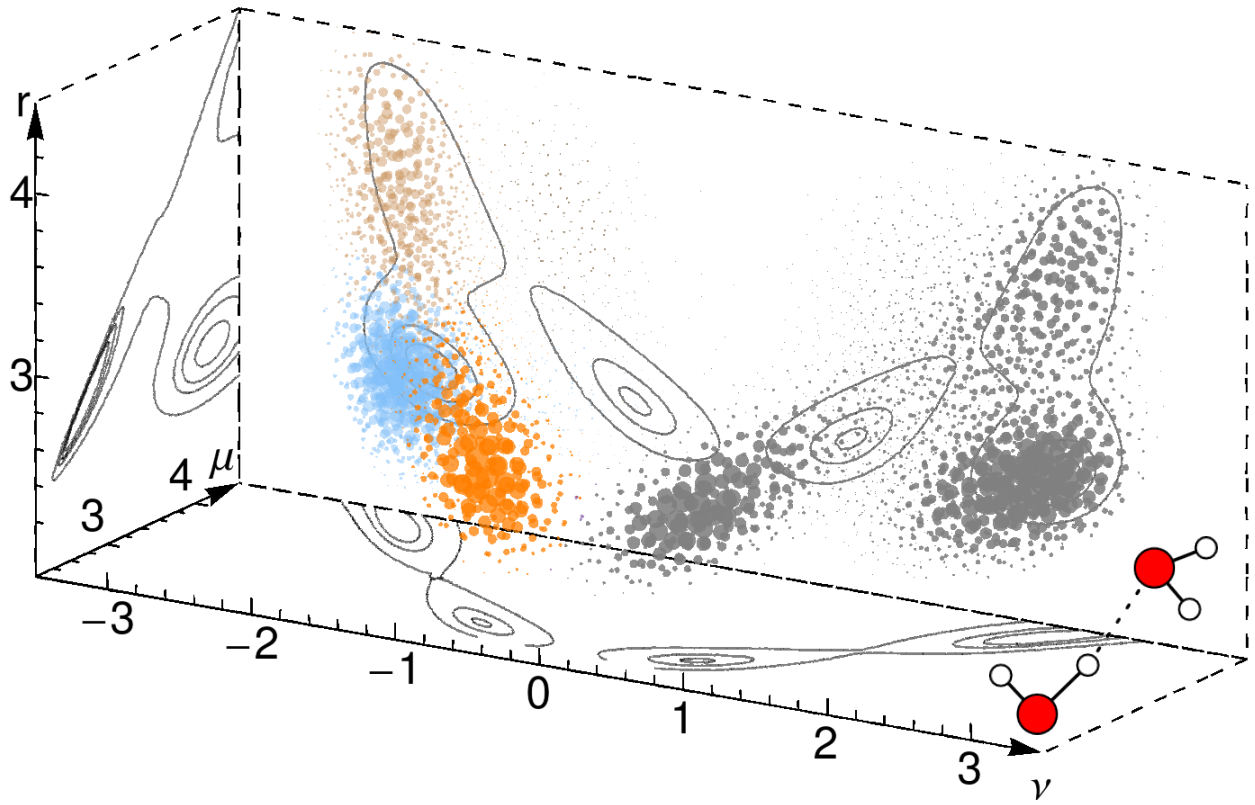
A H-bonding Environment Map

- PAMM gives for each (D, H, A) triplet a H-bond count s_{DHA}
- Define integrated H-bond counts, e.g. $s_D = \sum_{H,A} s_{DHA}$
- An "identity card" of H-bonding
- Nuclear quantum effects trigger transient proton jumps



A H-bonding Environment Map

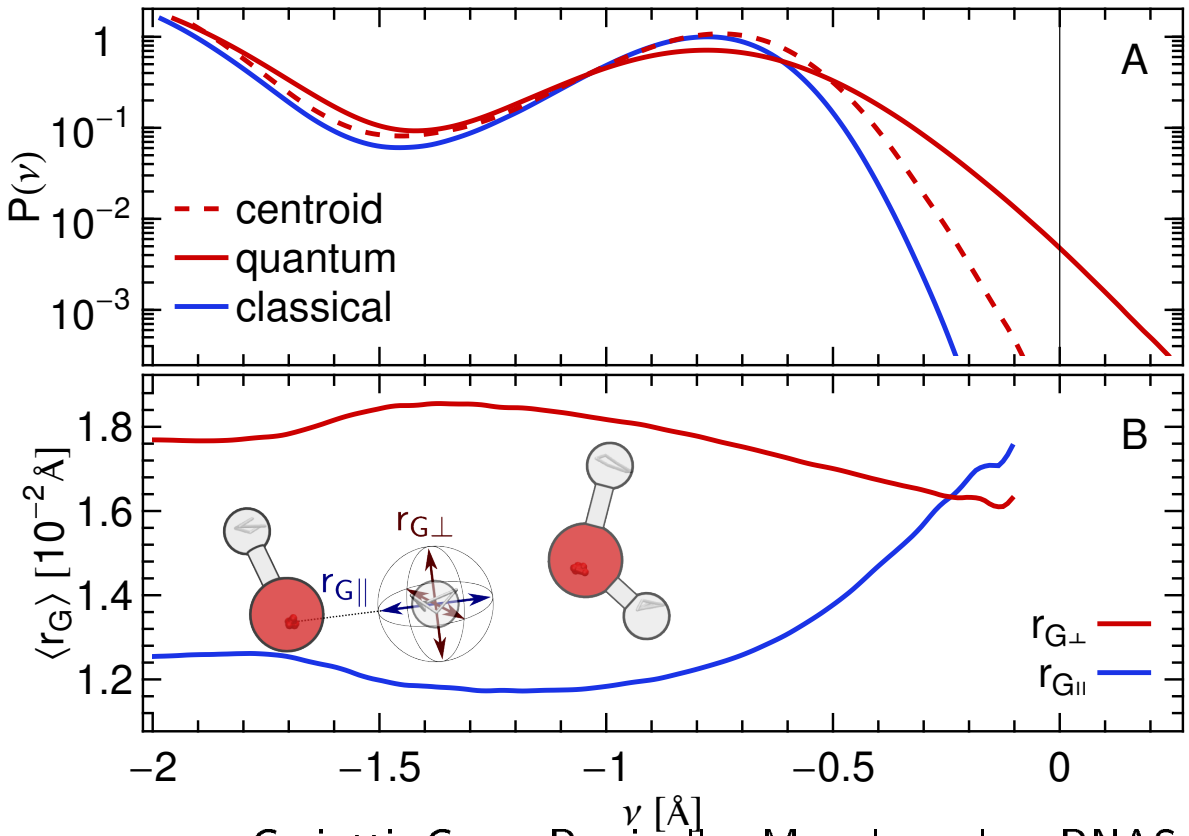
- PAMM gives for each (D, H, A) triplet a H-bond count s_{DHA}
- Define integrated H-bond counts, e.g. $s_{\mathrm{D}} = \sum_{\mathrm{H,A}} s_{\mathrm{DHA}}$
- An "identity card" of H-bonding
- Nuclear quantum effects trigger transient proton jumps



Ceriotti, Cuny, Parrinello, Manolopoulos, PNAS 110, 15591 (2013)

A H-bonding Environment Map

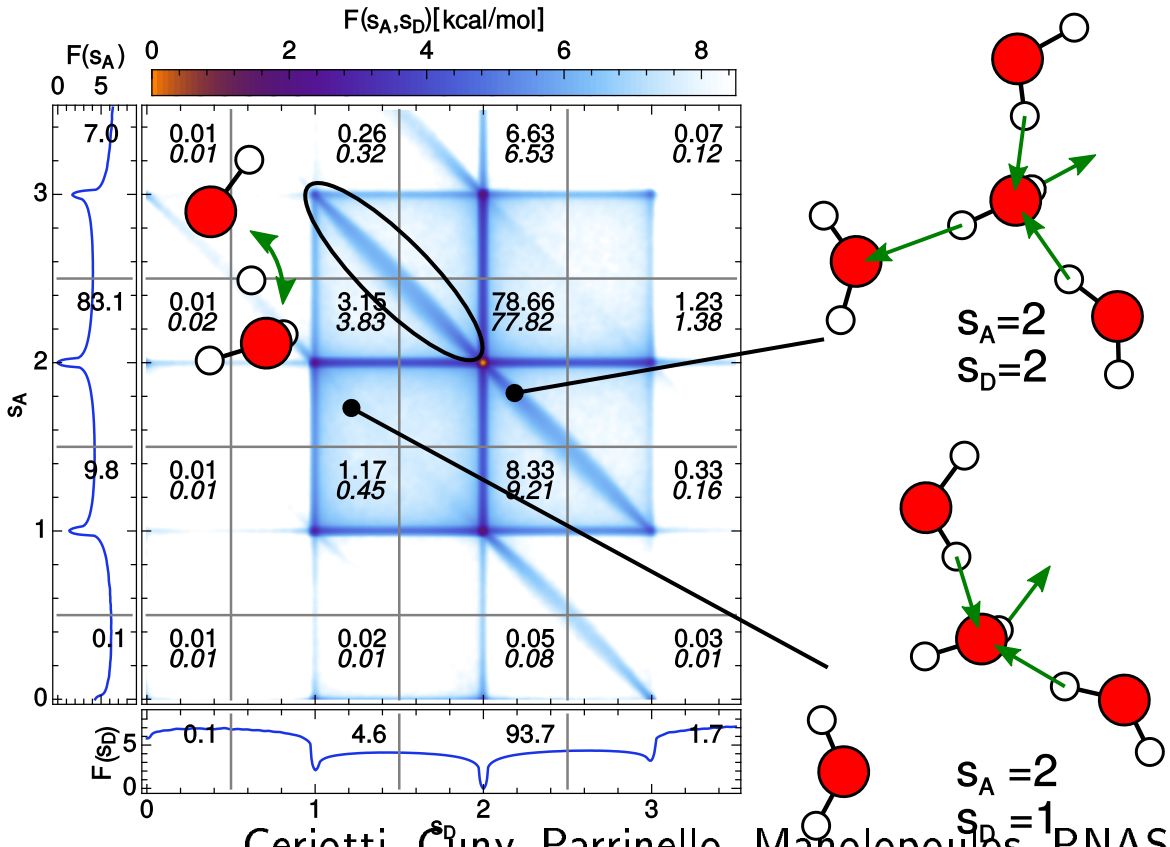
- PAMM gives for each (D, H, A) triplet a H-bond count s_{DHA}
- Define integrated H-bond counts, e.g. $s_{\mathrm{D}} = \sum_{\mathrm{H,A}} s_{\mathrm{DHA}}$
- An "identity card" of H-bonding
- Nuclear quantum effects trigger transient proton jumps



Ceriotti, Cuny, Parrinello, Manolopoulos, PNAS 110, 15591 (2013)

A H-bonding Environment Map

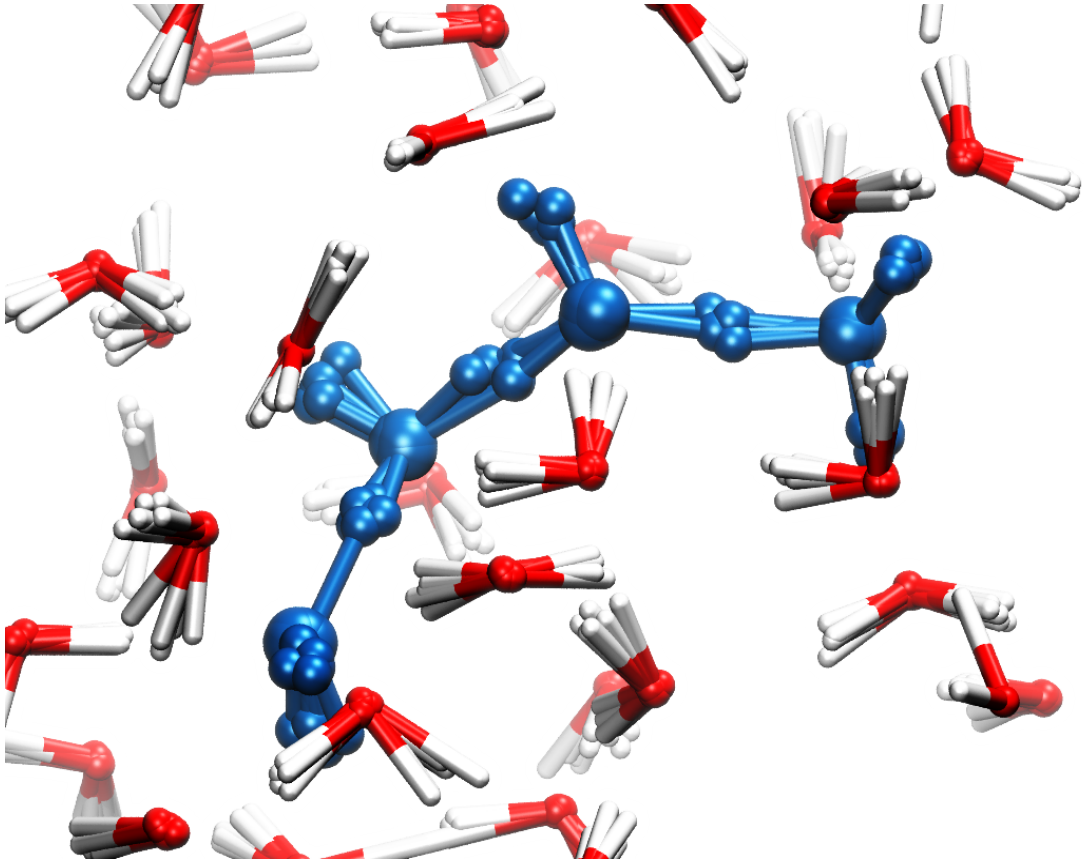
- PAMM gives for each (D, H, A) triplet a H-bond count s_{DHA}
- Define integrated H-bond counts, e.g. $s_D = \sum_{H,A} s_{DHA}$
- An "identity card" of H-bonding
- Nuclear quantum effects trigger transient proton jumps



Ceriotti, Cuny, Parrinello, Manolopoulos, PNAS 110, 15591 (2013)

What's an Excess Proton?

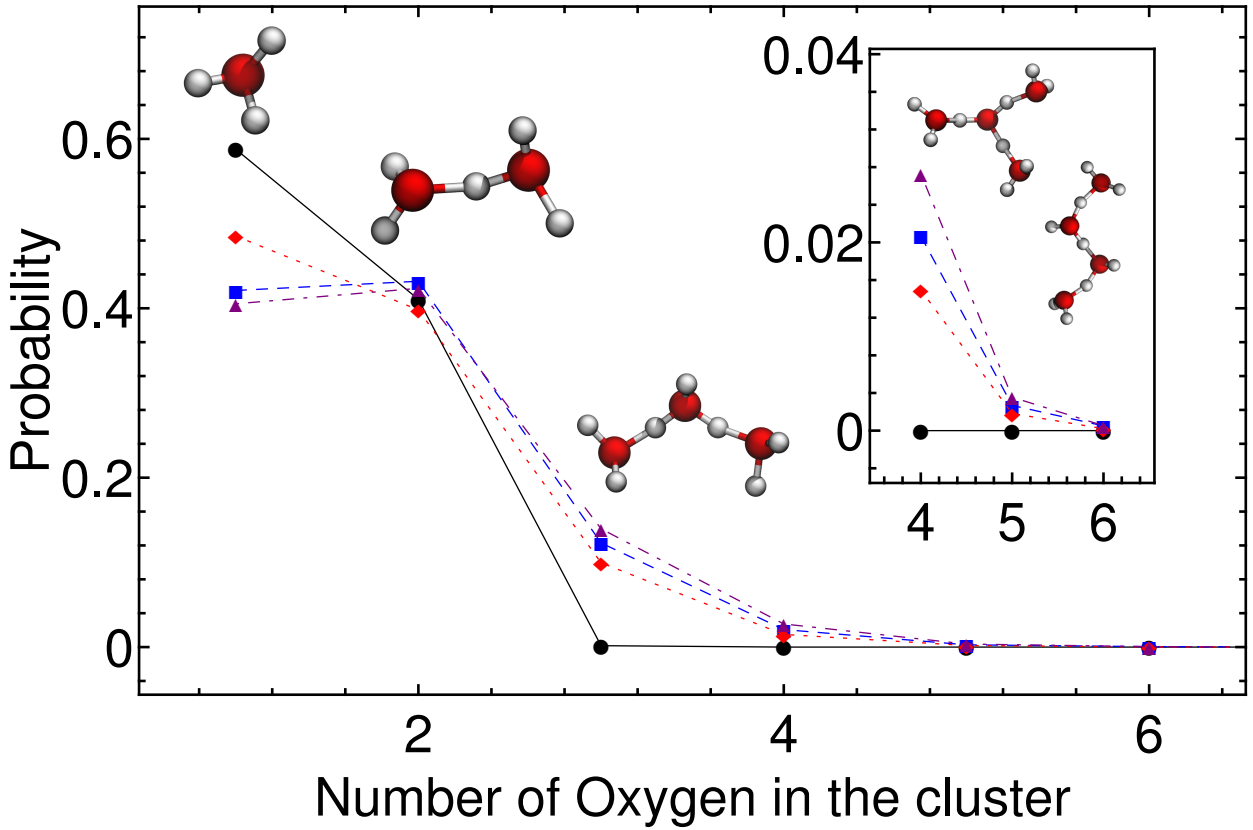
- Due to quantum fluctuations, it is not trivial to recognize an excess proton
- One can recognize protons by a cumbersome cluster analysis
- Can we exploit explicit knowledge of the electronic structure?



Giberti, Hassanali, Ceriotti, Parrinello J. Phys. Chem B , 15591 (2014)

What's an Excess Proton?

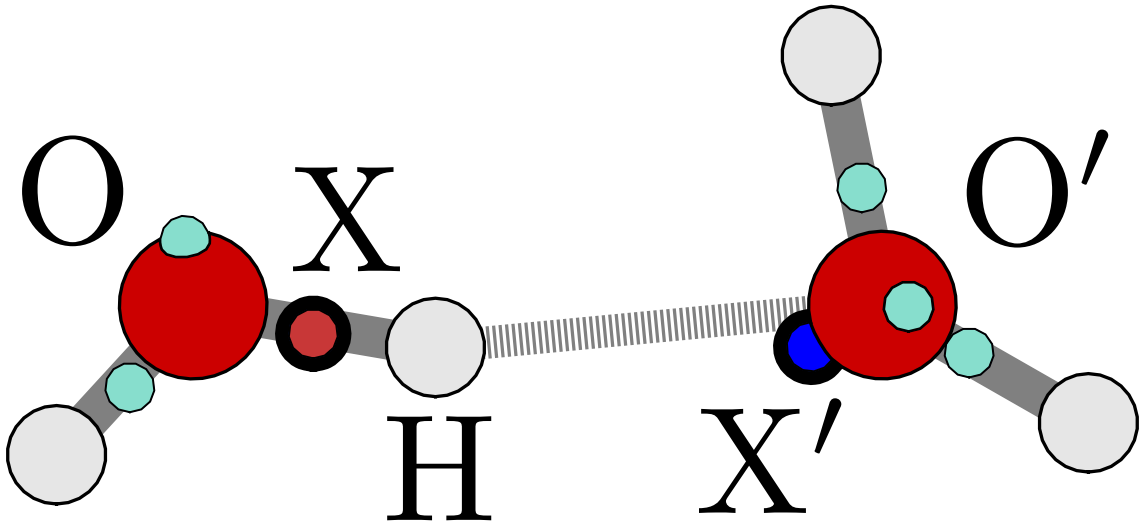
- Due to quantum fluctuations, it is not trivial to recognize an excess proton
- One can recognize protons by a cumbersome cluster analysis
- Can we exploit explicit knowledge of the electronic structure?



Giberti, Hassanali, Ceriotti, Parrinello J. Phys. Chem B , 15591 (2014)

What's an Excess Proton?

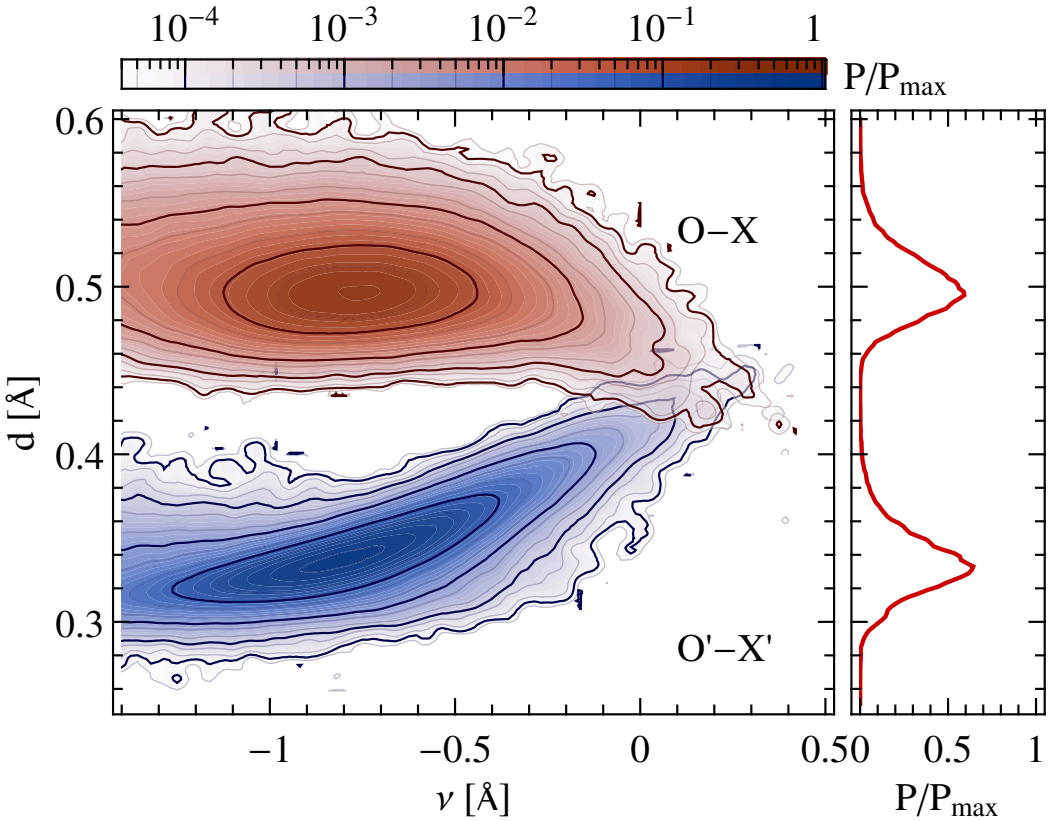
- Due to quantum fluctuations, it is not trivial to recognize an excess proton
- One can recognize protons by a cumbersome cluster analysis
- Can we exploit explicit knowledge of the electronic structure?



Giberti, Hassanali, Ceriotti, Parrinello J. Phys. Chem B , 15591 (2014)

What's an Excess Proton?

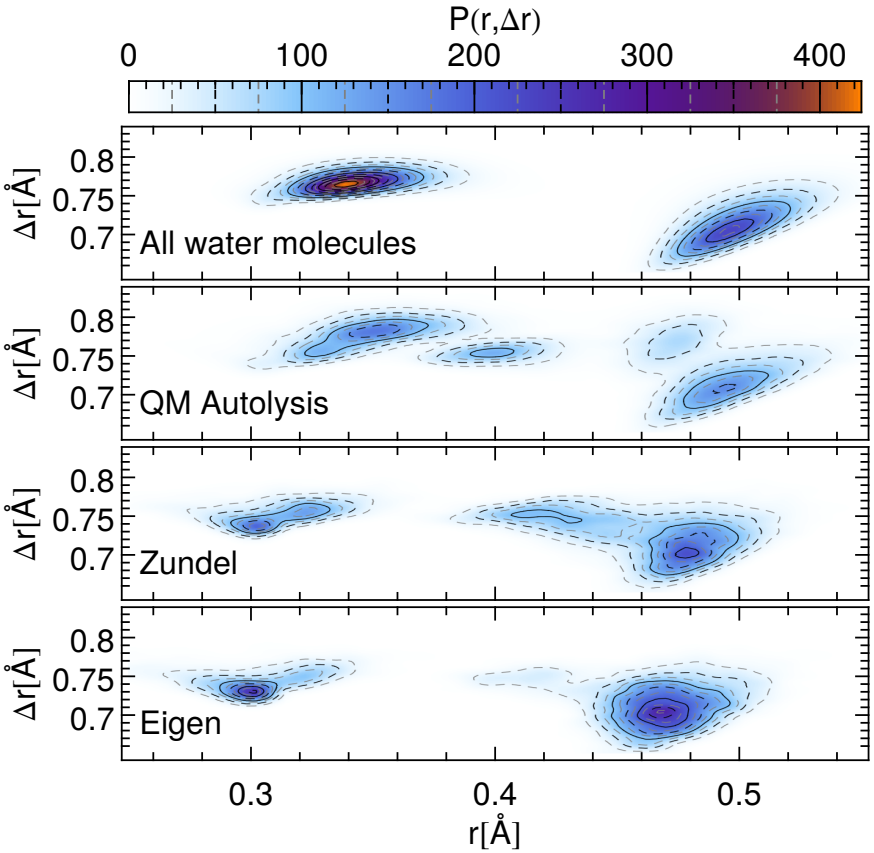
- Due to quantum fluctuations, it is not trivial to recognize an excess proton
- One can recognize protons by a cumbersome cluster analysis
- Can we exploit explicit knowledge of the electronic structure?



Giberti, Hassanali, Ceriotti, Parrinello J. Phys. Chem B , 15591 (2014)

What's an Excess Proton?

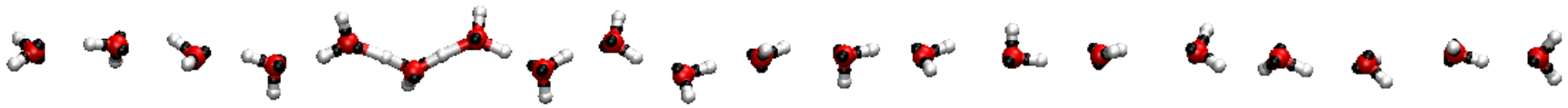
- Due to quantum fluctuations, it is not trivial to recognize an excess proton
- One can recognize protons by a cumbersome cluster analysis
- Can we exploit explicit knowledge of the electronic structure?



Giberti, Hassanali, Ceriotti, Parrinello J. Phys. Chem B , 15591 (2014)

PAMM Analysis of a Water Wire

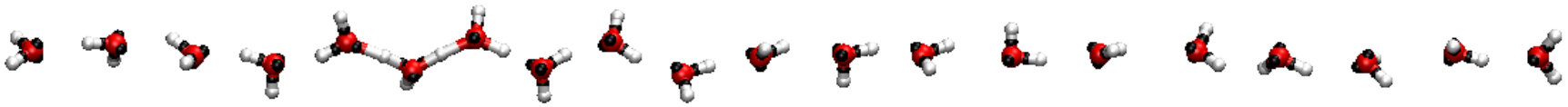
- Can we use PAMM to identify proton-like water molecules in a wire?
- The key variables are the distances of the "intermediate" Wannier centers from the O atoms
- Despite the very weak "proton" signal (1/20!) one can recognize two clear clusters

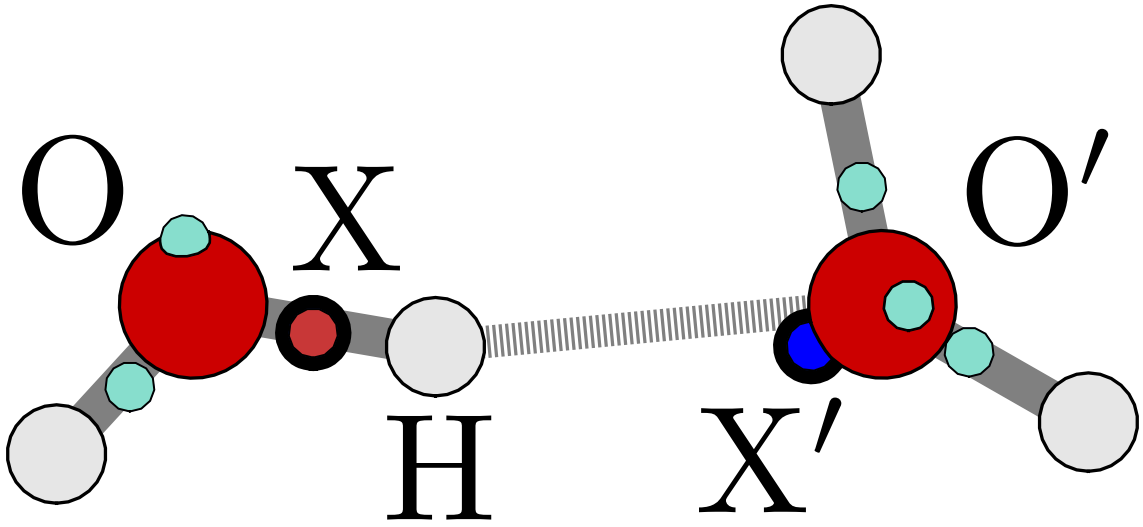


Mariana Rossi & MC, Work in Progress...

PAMM Analysis of a Water Wire

- Can we use PAMM to identify proton-like water molecules in a wire?
- The key variables are the distances of the "intermediate" Wannier centers from the O atoms
- Despite the very weak "proton" signal (1/20!) one can recognize two clear clusters

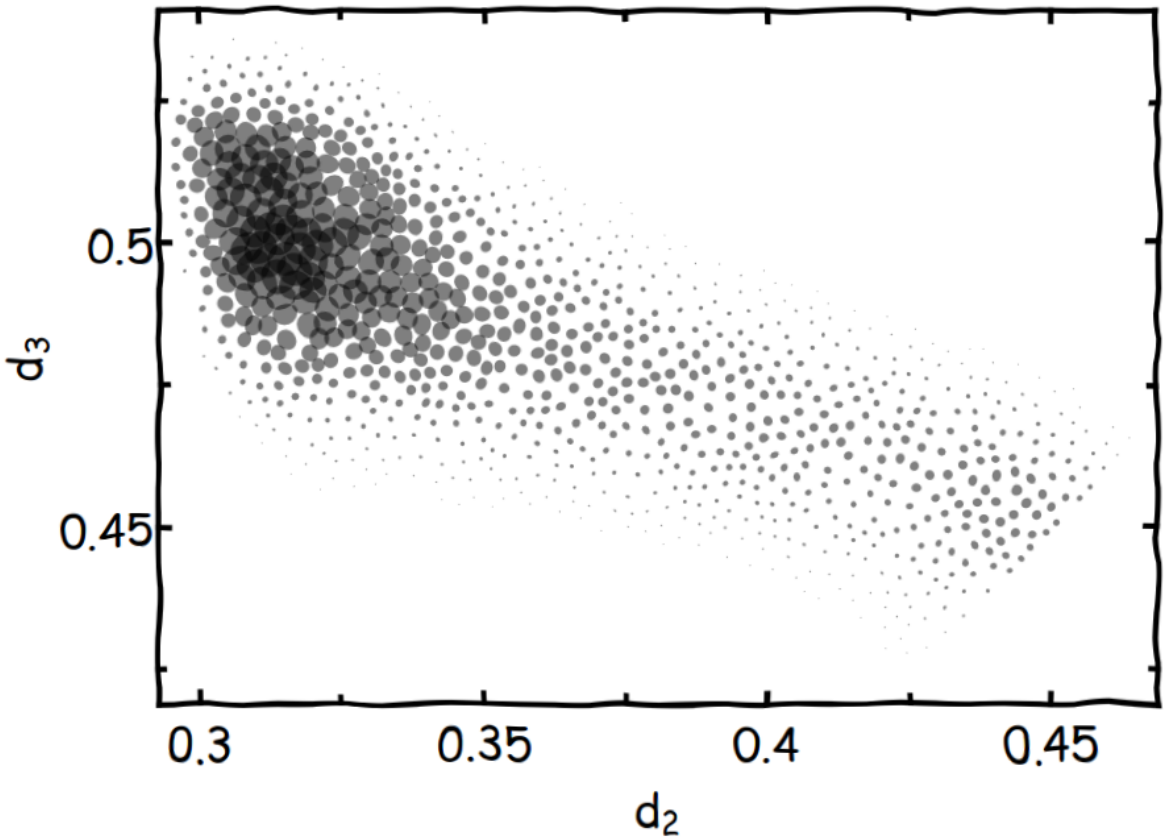




Mariana Rossi & MC, Work in Progress...

PAMM Analysis of a Water Wire

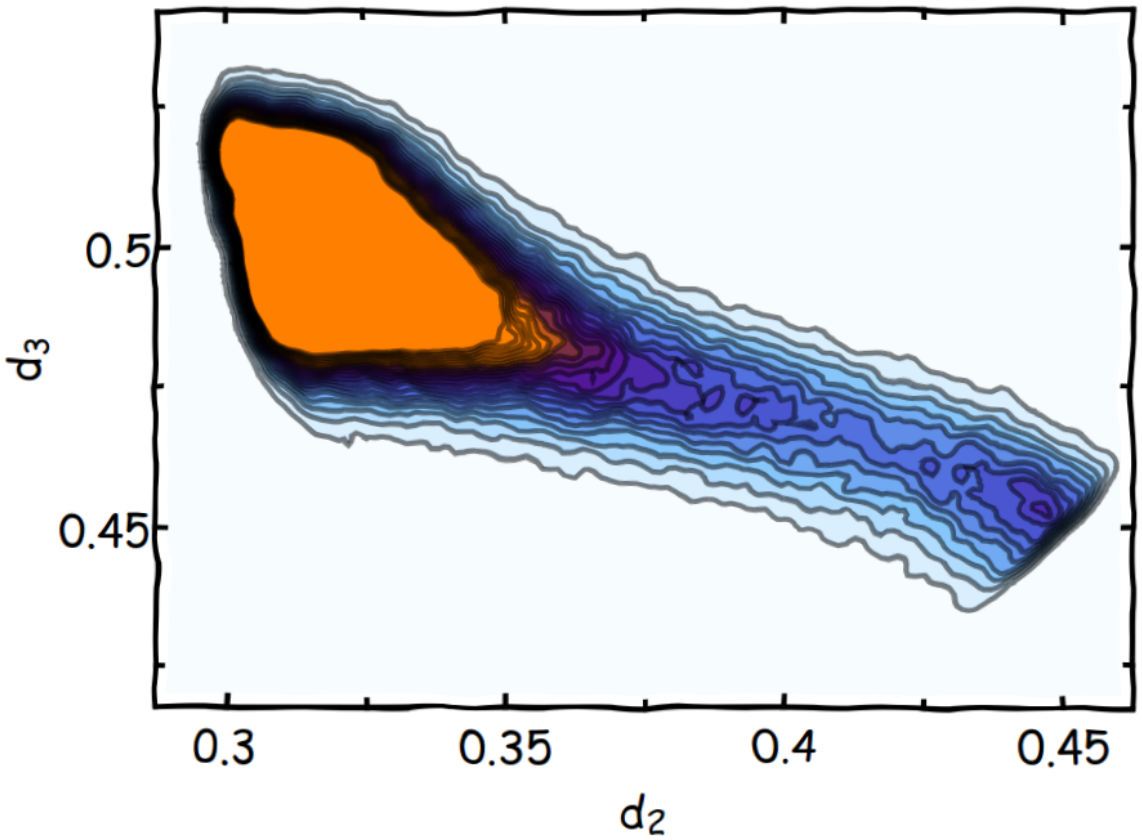
- Can we use PAMM to identify proton-like water molecules in a wire?
- The key variables are the distances of the "intermediate" Wannier centers from the O atoms
- Despite the very weak "proton" signal (1/20!) one can recognize two clear clusters



Mariana Rossi & MC, Work in Progress...

PAMM Analysis of a Water Wire

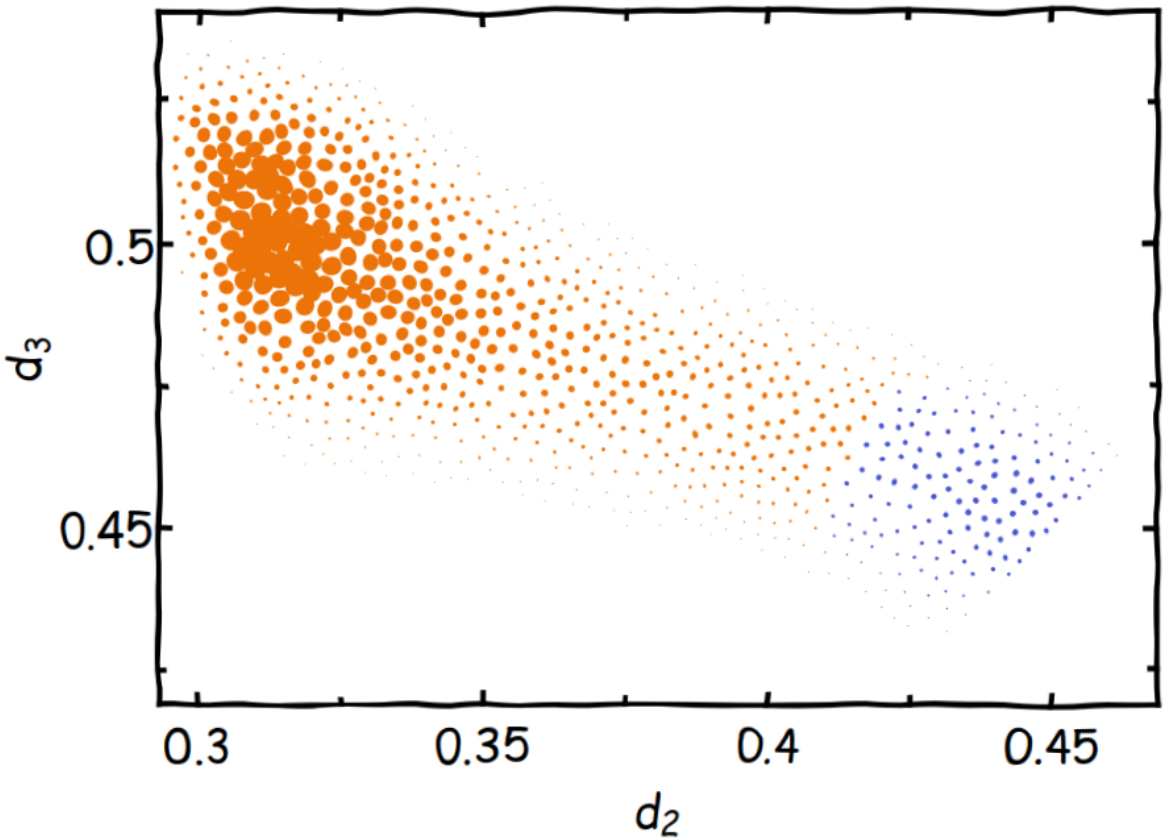
- Can we use PAMM to identify proton-like water molecules in a wire?
- The key variables are the distances of the "intermediate" Wannier centers from the O atoms
- Despite the very weak "proton" signal (1/20!) one can recognize two clear clusters



Mariana Rossi & MC, Work in Progress...

PAMM Analysis of a Water Wire

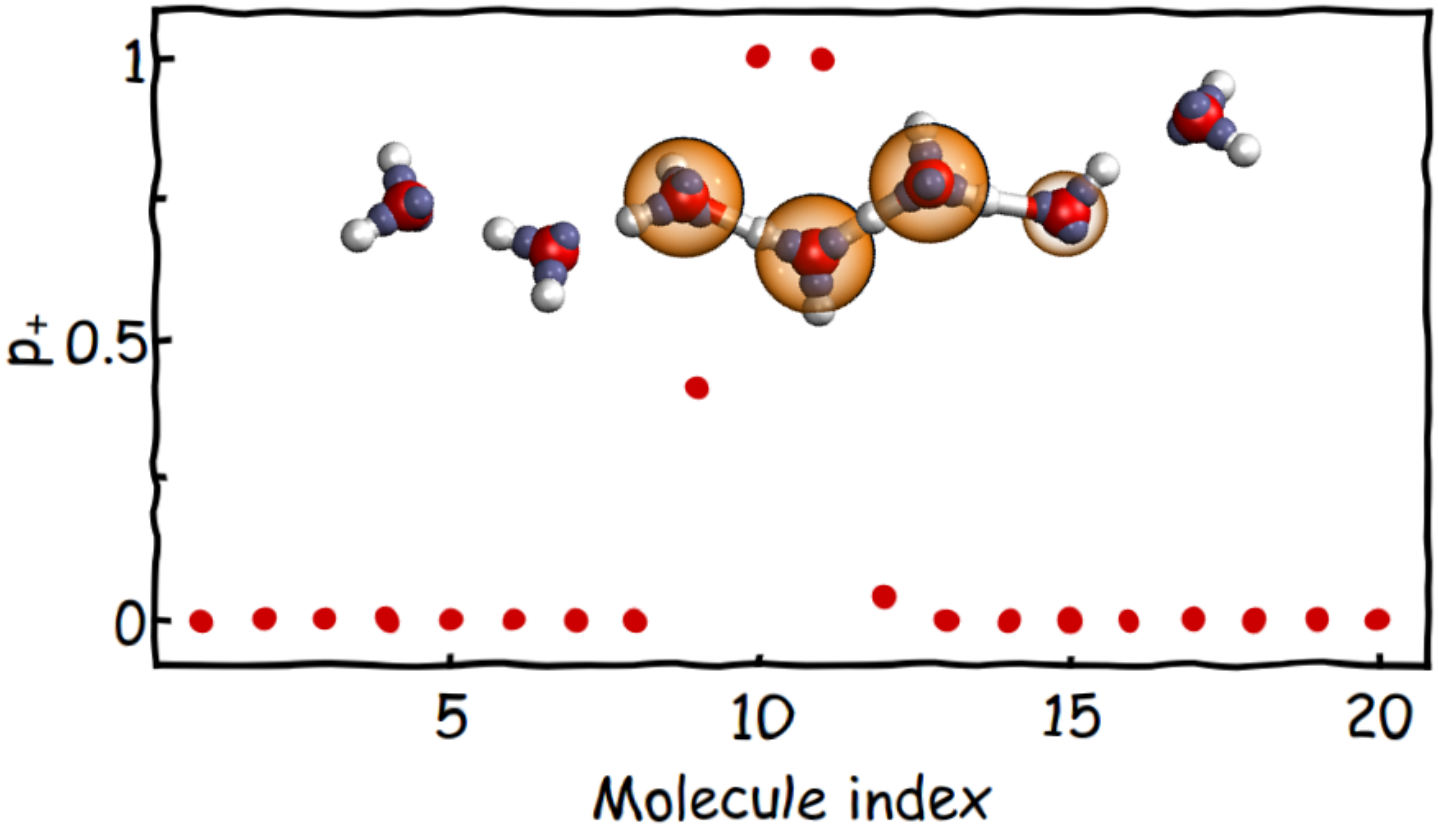
- Can we use PAMM to identify proton-like water molecules in a wire?
- The key variables are the distances of the "intermediate" Wannier centers from the O atoms
- Despite the very weak "proton" signal (1/20!) one can recognize two clear clusters



Mariana Rossi & MC, Work in Progress...

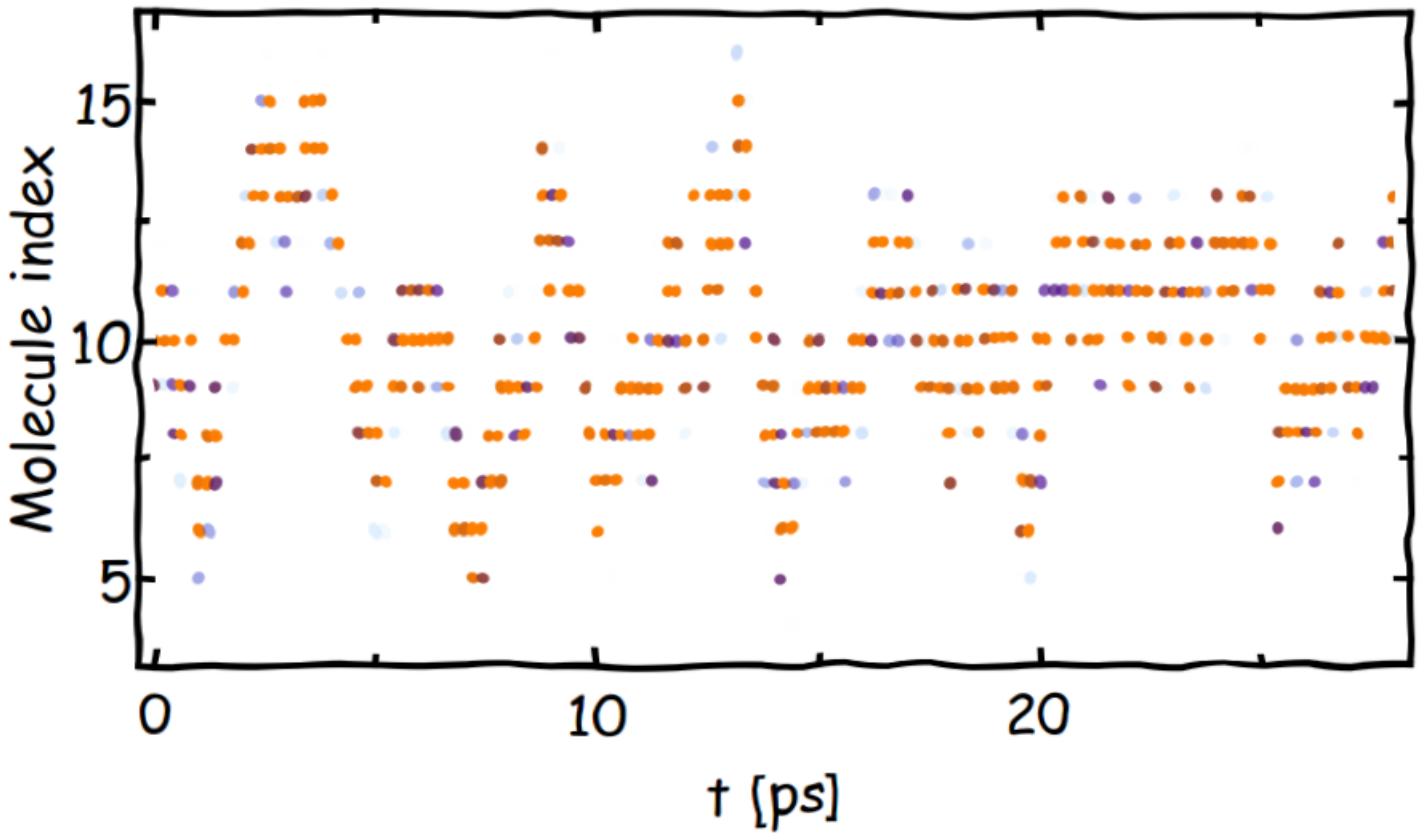
PAMM Protons are Delocalized

- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire
- Center-of-charge, or "oxygen-coordination" definitions of the proton are either discontinuous or contaminated by dipole fluctuations
- The center of the PAMM proton is a more satisfactory description



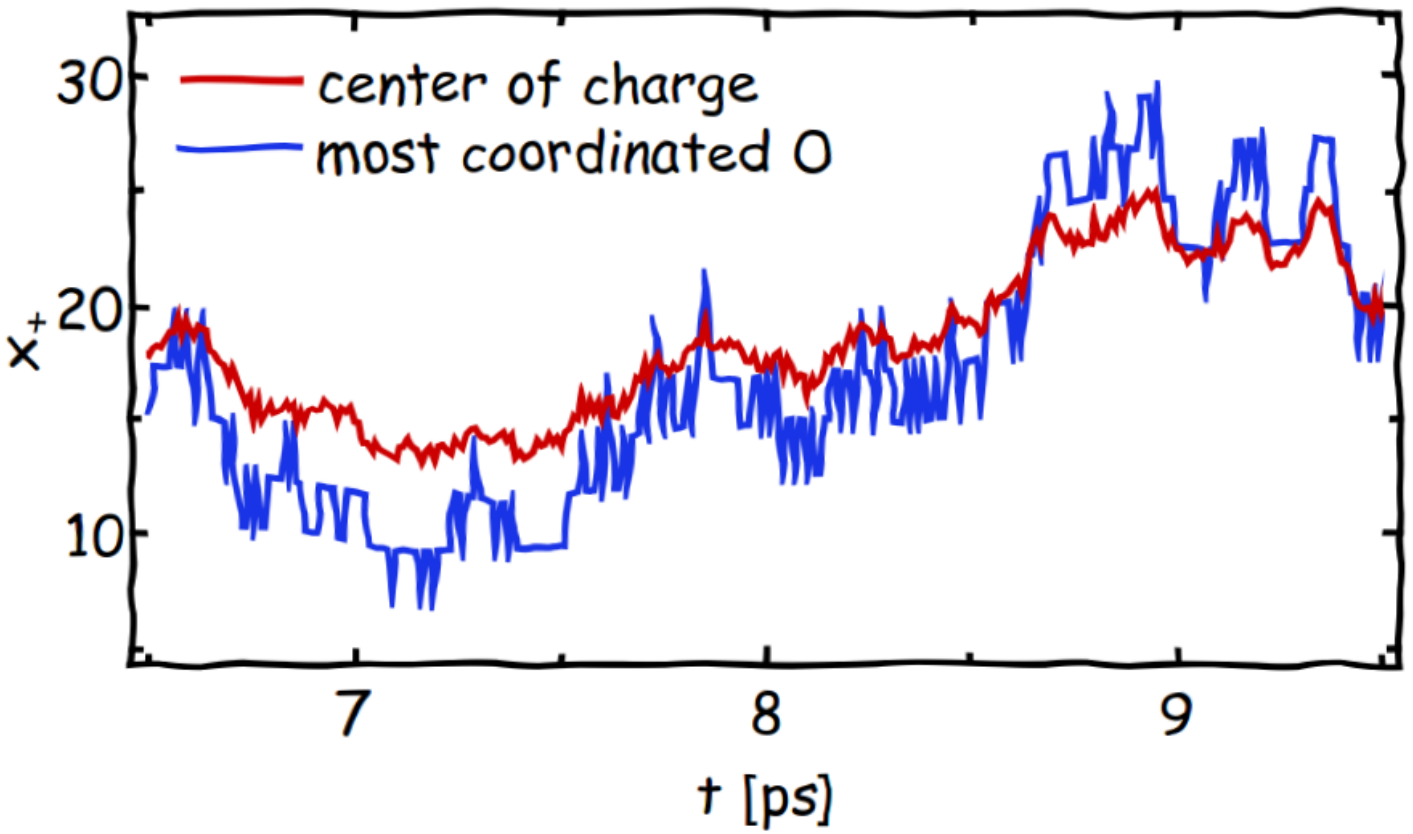
PAMM Protons are Delocalized

- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire
- Center-of-charge, or "oxygen-coordination" definitions of the proton are either discontinuous or contaminated by dipole fluctuations
- The center of the PAMM proton is a more satisfactory description



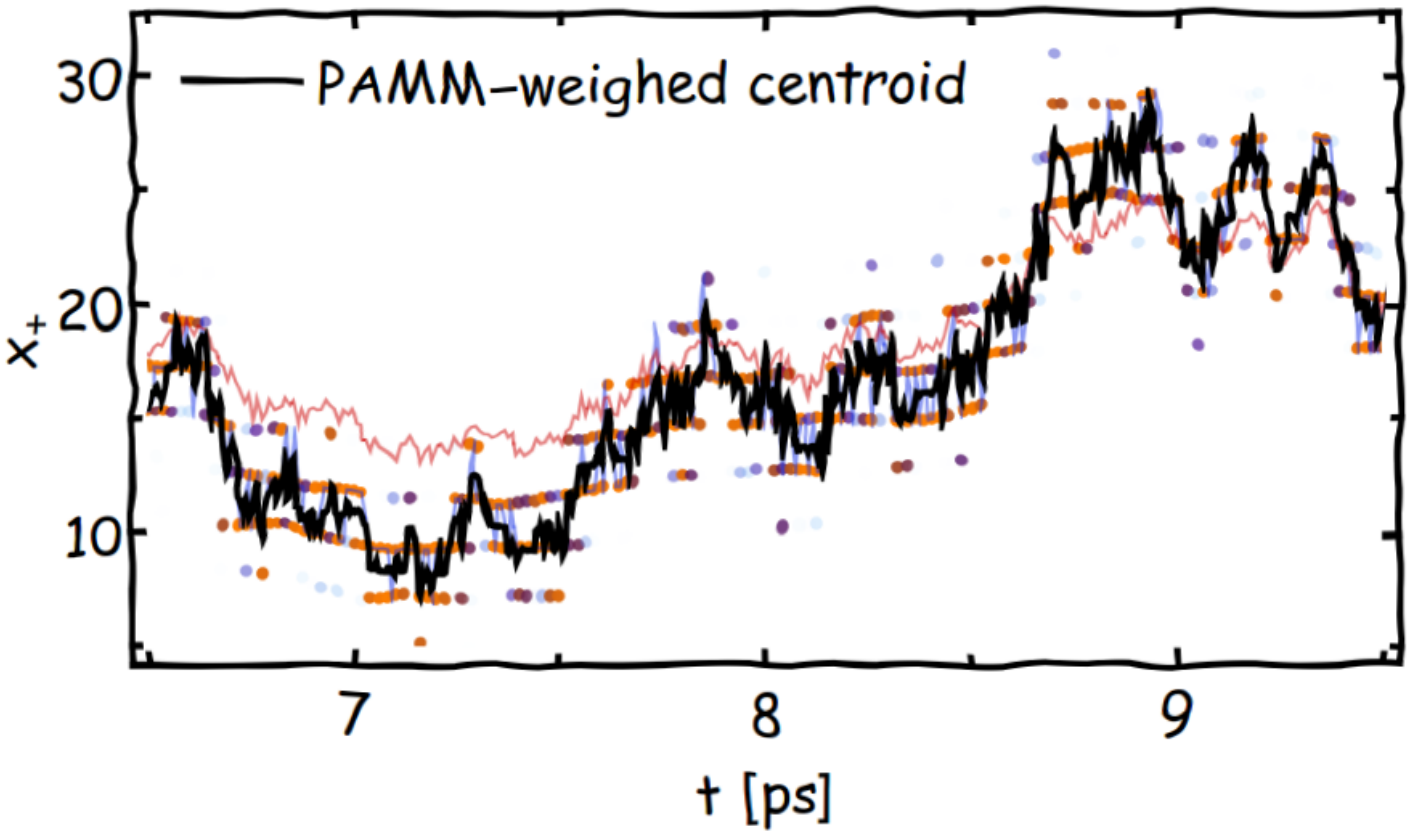
PAMM Protons are Delocalized

- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire
- Center-of-charge, or "oxygen-coordination" definitions of the proton are either discontinuous or contaminated by dipole fluctuations
- The center of the PAMM proton is a more satisfactory description



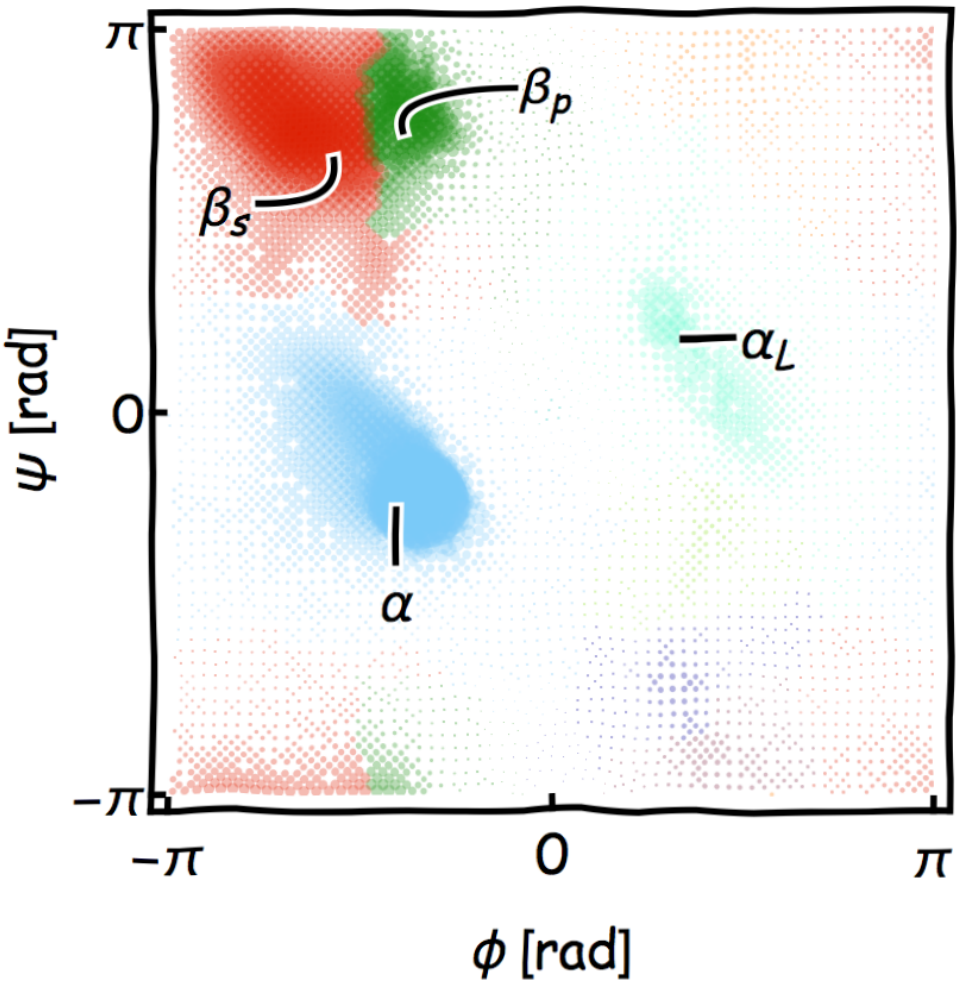
PAMM Protons are Delocalized

- PAMM fingerprints recognize a "proton wavepacket" moving in a concerted way along the wire
- Center-of-charge, or "oxygen-coordination" definitions of the proton are either discontinuous or contaminated by dipole fluctuations
- The center of the PAMM proton is a more satisfactory description



Machine-learning the Ramachandran plot

- Use data from the PDB, and "learn" with PAMM the stable patterns of proteins in dihedral space



Acknowledgements

Piero Gasparotto

CCMX
Competence Centre for
Materials Science and Technology

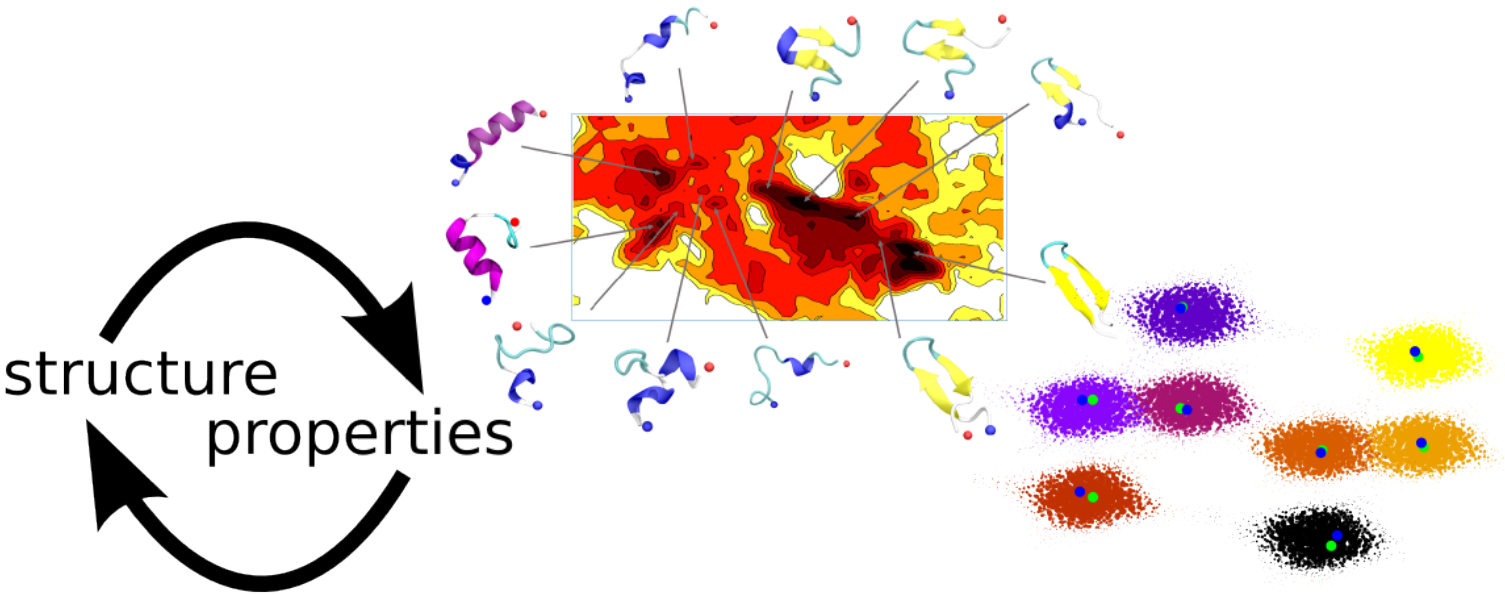
Michele Parrinello
Gareth Tribello

CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Mariana Rossi

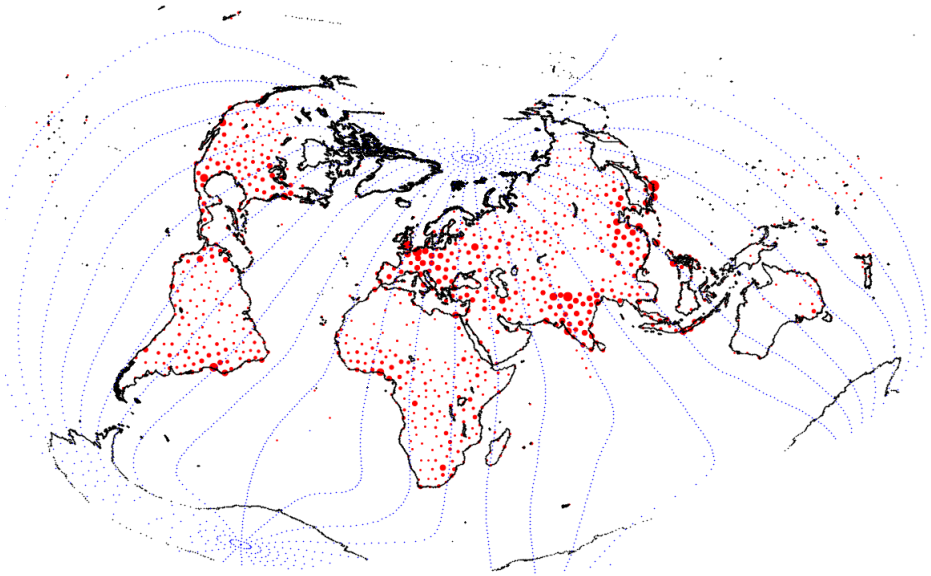
Challenges for machine learning

- Machine learning can be used to process (large and/or high-dimensional) simulation data in a simpler-to-understand form
  - Use the ML description to improve simulations
  - Can we formalize "intuitive understanding" as a goal for ML descriptors?



Challenges for machine learning

- Machine learning can be used to process (large and/or high-dimensional) simulation data in a simpler-to-understand form
  - Use the ML description to improve simulations
  - Can we formalize "intuitive understanding" as a goal for ML descriptors?



http://epfl-cosmo.github.io/sketchmap

http://github.com/epfl-cosmo/pamm