Statistical methods in Atomistic Computer Simulations

5. Dimensionality reduction

Michele Ceriotti

MSE 639 - EPFL - EDMX

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Outline

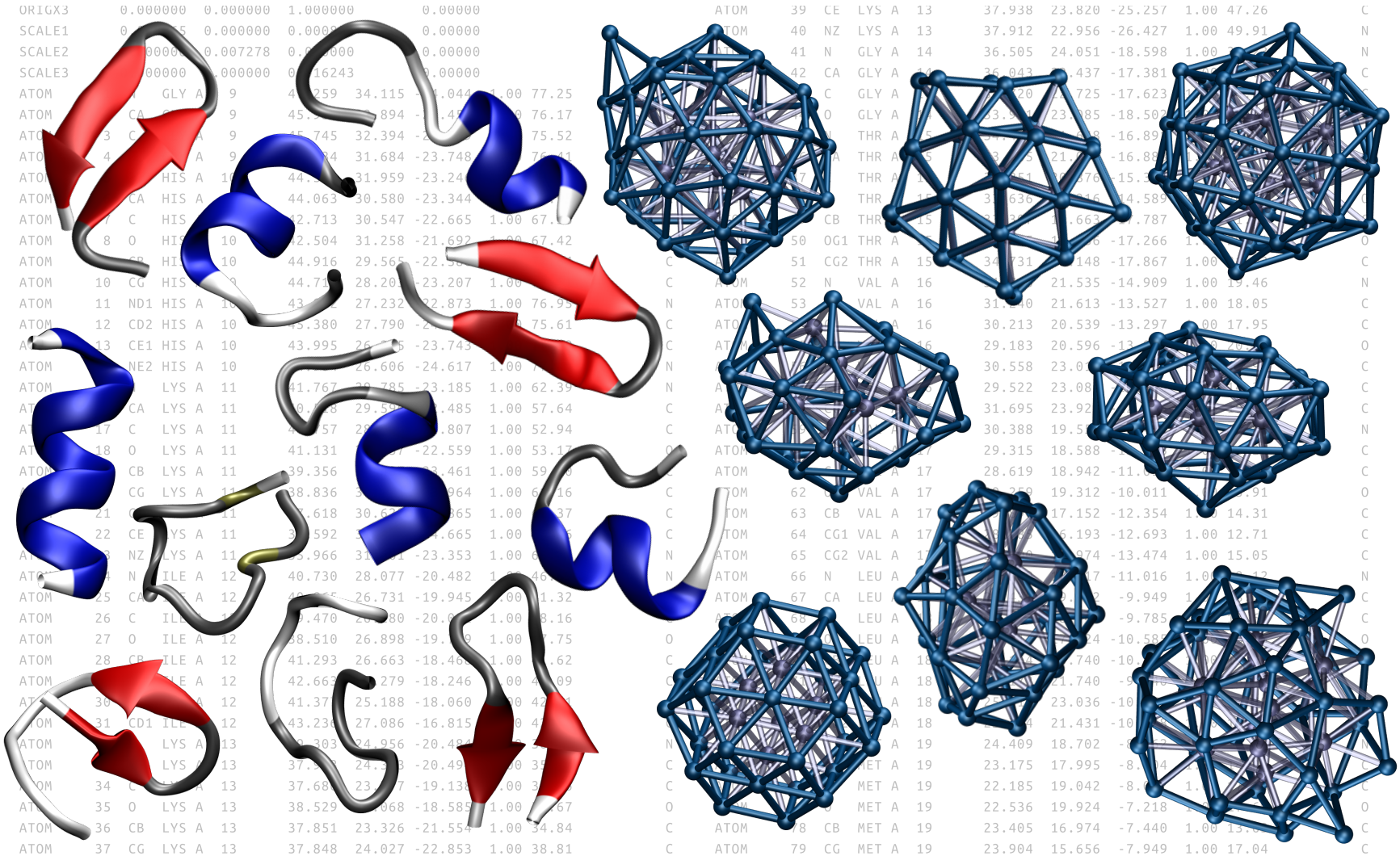
1. Sampling the canonical ensemble in atomistic simulations: Monte Carlo, correlation functions, ergodicity
2. Thermostatting MD efficiently: white and colored-noise Langevin dynamics
3. Rare events: transition state theory and the transition state ensemble
4. Statistical re-weighting, biased sampling, metadynamics
5. Linear and non-linear dimensionality reduction: determining automatically order parameters

- 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
```

High dimensional data in atomistic simulatio

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

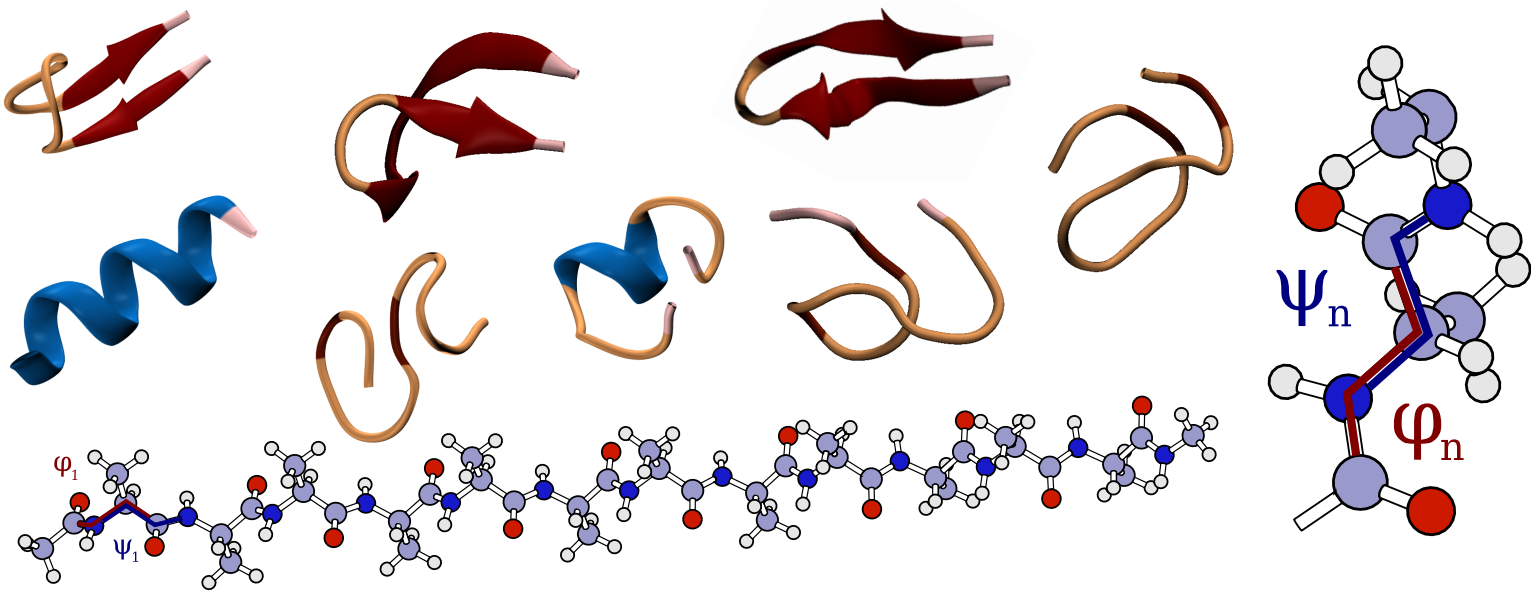


Describing structural complexity

- We are looking for **collective variables** that can describe this structural complexity
  - Discriminate between different structures
  - Follow the system across transitions
- This is not only important for post-processing
  - Good CVs make for better transition-state approximation to the rate
  - Biased MD requires coarse-grained but thorough description of the problem
- Finding these variables is time-consuming and error-prone: can we **automate** the process?

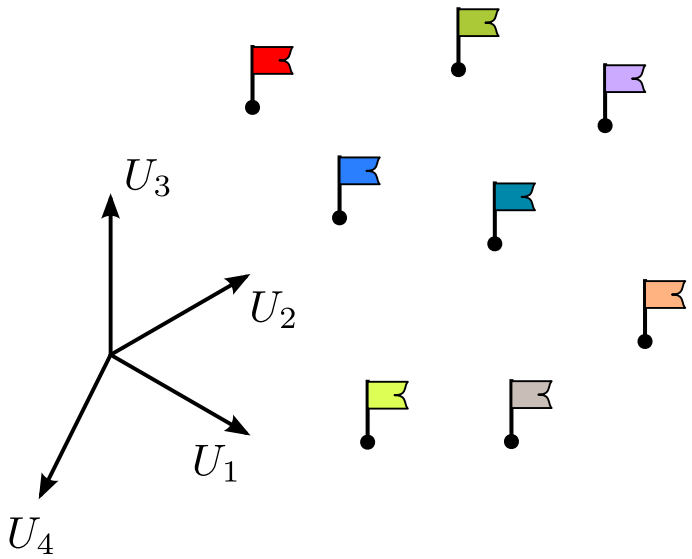
Dimensionality reduction

- 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**
  - 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**



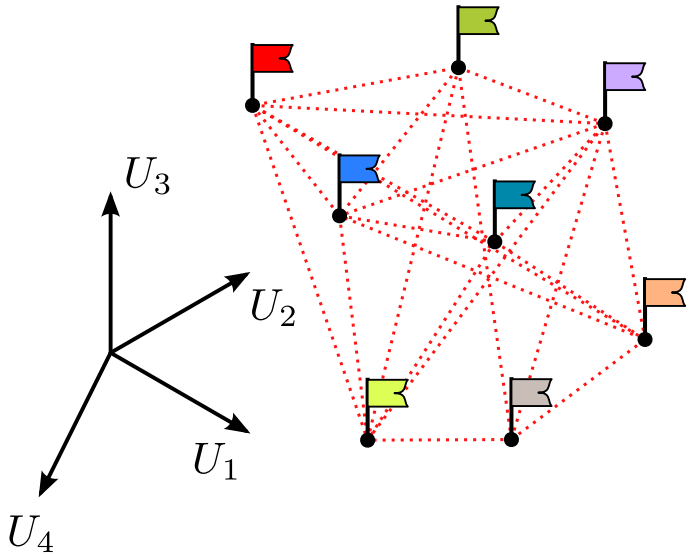
Dimensionality reduction

- 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**
  - 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**



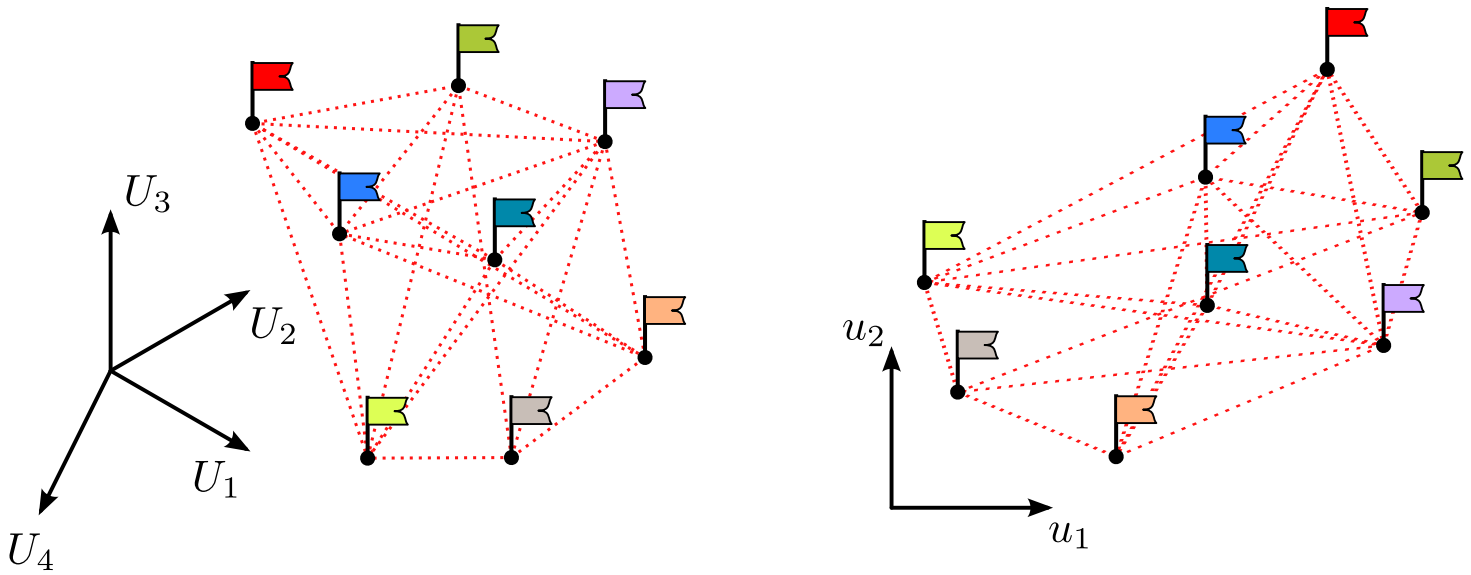
Dimensionality reduction

- 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**
  - 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**



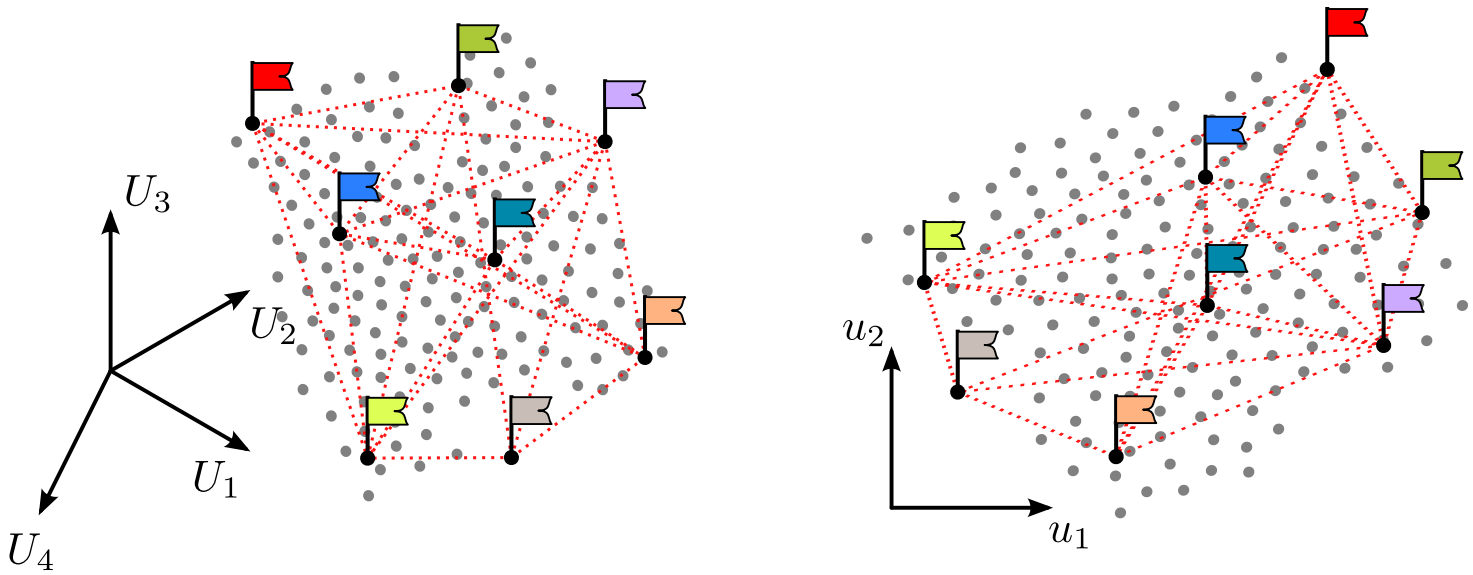
Dimensionality reduction

- 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**
  - 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**



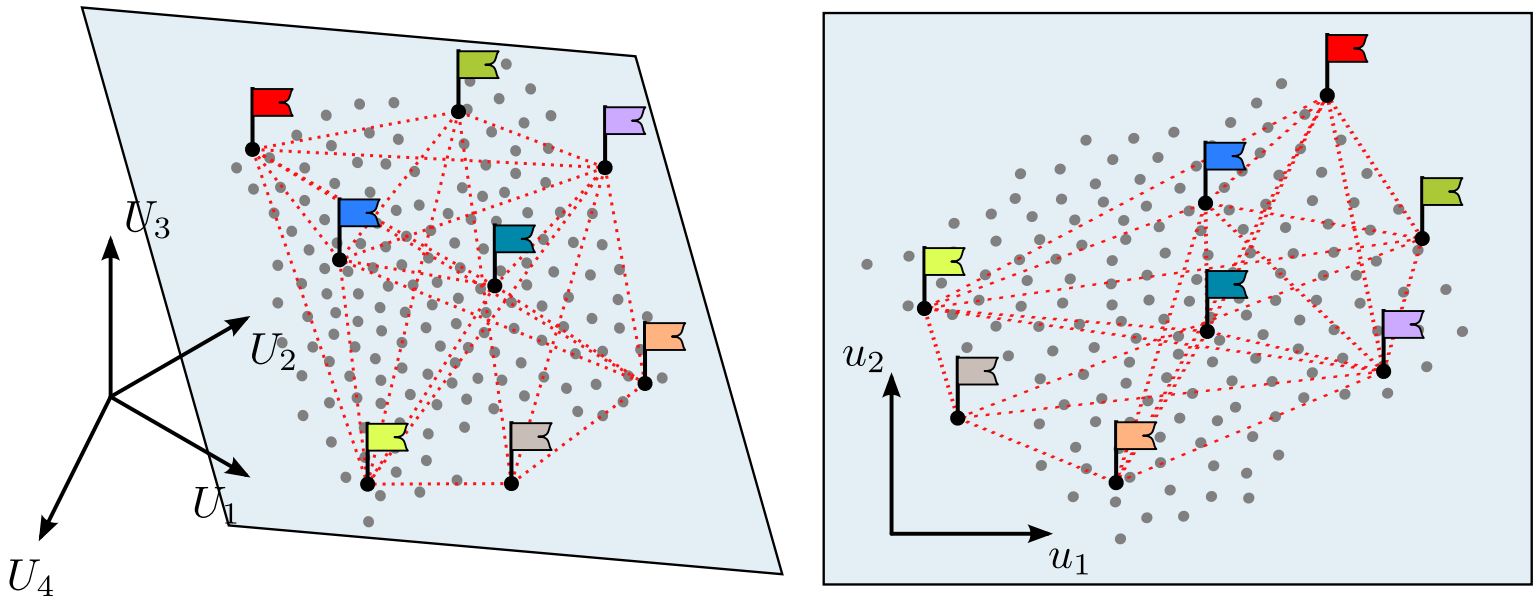
Dimensionality reduction

- 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**
  - 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

- 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**
  - 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**



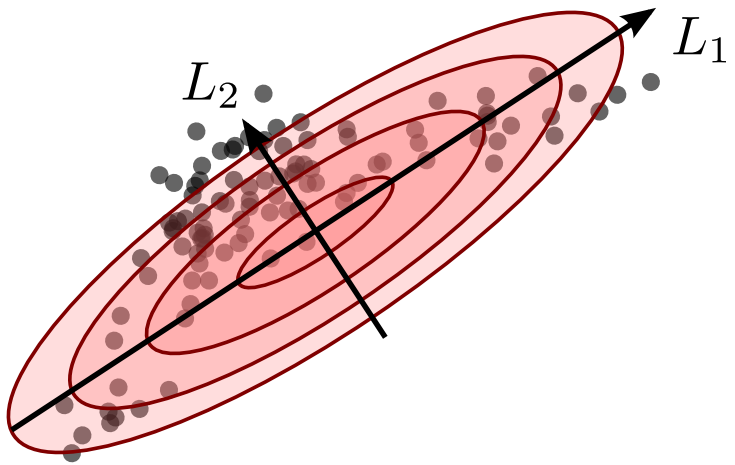
Principal component analysis

- Principal component analysis: assumes that the "important" coordinates are the linear combinations with the largest variance
  - $\{X_i\}$ are N vectors in D dimensions. Let $\mathbf{X}$ be the $N \times D$ matrix with the X_i as rows.
  - Define the $N \times N$ centering matrix $H_{ij} = \delta_{ij} - \frac{1}{n}$
  - Define the covariance matrix $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{H}\mathbf{X}$, $C_{ij} = \frac{1}{n}\sum_k \left(X_k - \bar{X}\right)_i \left(X_k - \bar{X}\right)_j$
  - Pick the d eigenvectors P_i associated with the largest eigenvalues λ_i and use them as the rows of a **linear projector** $\mathbf{P}$.
  - The low-dimensional projections are $x_i = \mathbf{P}^T X_i$



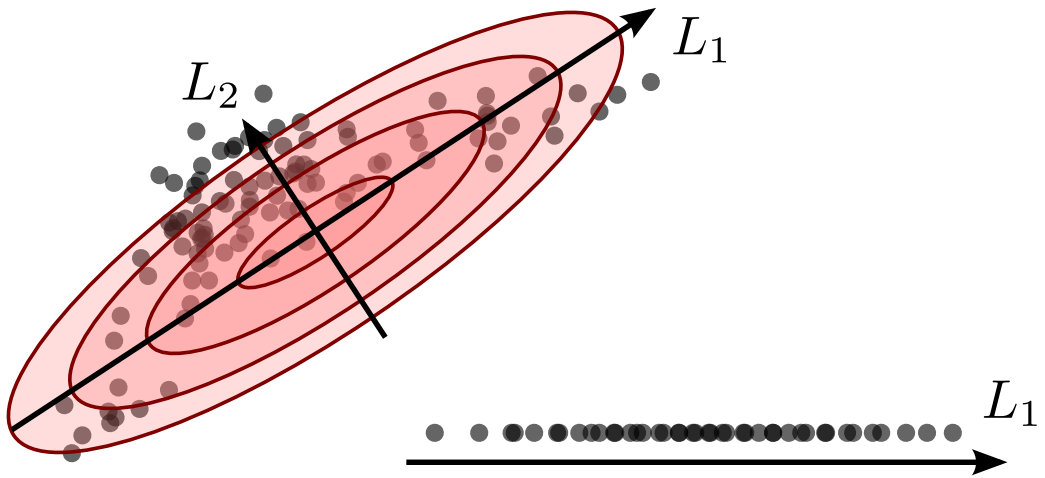
Principal component analysis

- Principal component analysis: assumes that the "important" coordinates are the linear combinations with the largest variance
  - $\{X_i\}$ are N vectors in D dimensions. Let $\mathbf{X}$ be the $N \times D$ matrix with the X_i as rows.
  - Define the $N \times N$ centering matrix $H_{ij} = \delta_{ij} - \frac{1}{n}$
  - Define the covariance matrix $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{H}\mathbf{X}$, $C_{ij} = \frac{1}{n}\sum_k \left(X_k - \bar{X}\right)_i \left(X_k - \bar{X}\right)_j$
  - Pick the d eigenvectors P_i associated with the largest eigenvalues λ_i and use them as the rows of a **linear projector** $\mathbf{P}$.
  - The low-dimensional projections are $x_i = \mathbf{P}^T X_i$



Principal component analysis

- Principal component analysis: assumes that the "important" coordinates are the linear combinations with the largest variance
  - $\{X_i\}$ are N vectors in D dimensions. Let $\mathbf{X}$ be the $N \times D$ matrix with the X_i as rows.
  - Define the $N \times N$ centering matrix $H_{ij} = \delta_{ij} - \frac{1}{n}$
  - Define the covariance matrix $\mathbf{C} = \frac{1}{n}\mathbf{X}^T\mathbf{H}\mathbf{X}$, $C_{ij} = \frac{1}{n}\sum_k \left(X_k - \bar{X}\right)_i \left(X_k - \bar{X}\right)_j$
  - Pick the d eigenvectors P_i associated with the largest eigenvalues λ_i and use them as the rows of a **linear projector P**.
  - The low-dimensional projections are $x_i = \mathbf{P}^T X_i$



Multidimensional scaling

- A literal implementation of the general idea of dimensionality reduction
  - define $\Delta_{ij} = \Delta(X_i, X_j)$ where $\Delta(X, Y)$ is a measure of similarity between points in D dimensions
  - find d-dimensional projections $\{x_i\}$ minimizing

$$\chi^2 = \sum_{ij} (\Delta_{ij} - |x_i - x_j|)^2$$

- *Classical* MDS turns this iterative optimization in an eigenvalue problem
  - Define $S_{ij} = \Delta_{ij}^2$ and $\mathbf{B} = -\frac{1}{2}\mathbf{HSH}$. Note that $\mathbf{B} = (\mathbf{HX})(\mathbf{HX})^T$
  - Compute the largest d eigenvalues of $\mathbf{B}$, λ_i and the eigenvectors $\mathbf{V}_i$
  - Make the $n \times d$ matrix whose columns are $\sqrt{\lambda_i}\mathbf{V}_i$. The rows are the x_i low-dimensional projections
- If $\Delta(X_i, X_j)$ is the Euclidean norm, classical MDS is the best *linear* projection preserving the squared distances. It corresponds to PCA, but it is more easily generalized to different dissimilarities

Cox & Cox, *Multidimensional Scaling* (CRC Press, 2010)

The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D.
  - *Versicolor* and *Virginica* are pretty close...



Iris Setosa



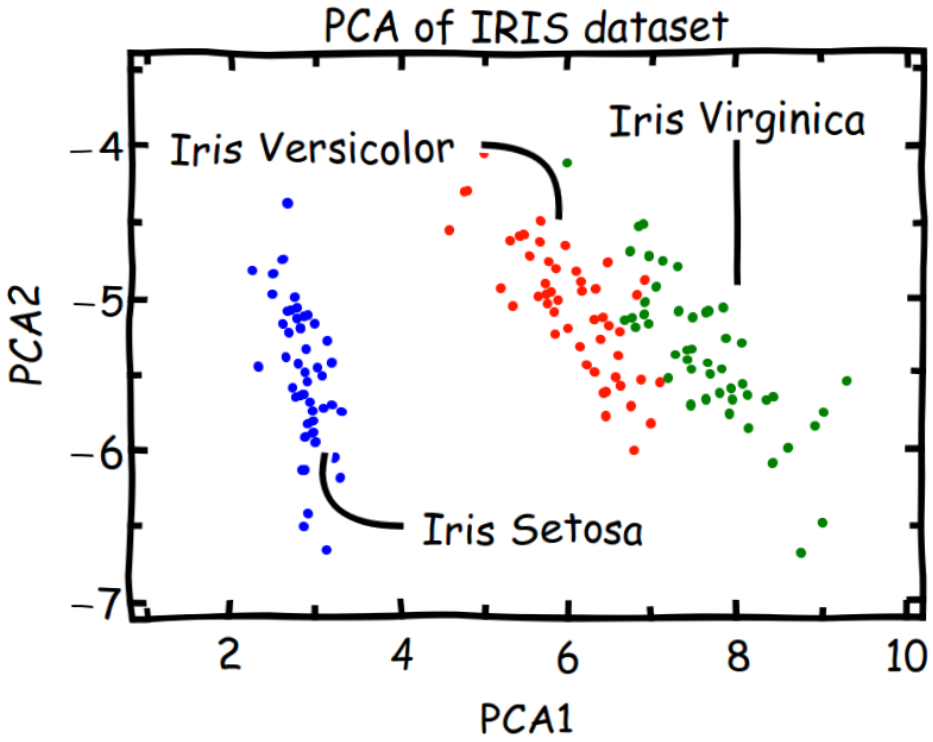
Iris Versicolor



Iris Virginica

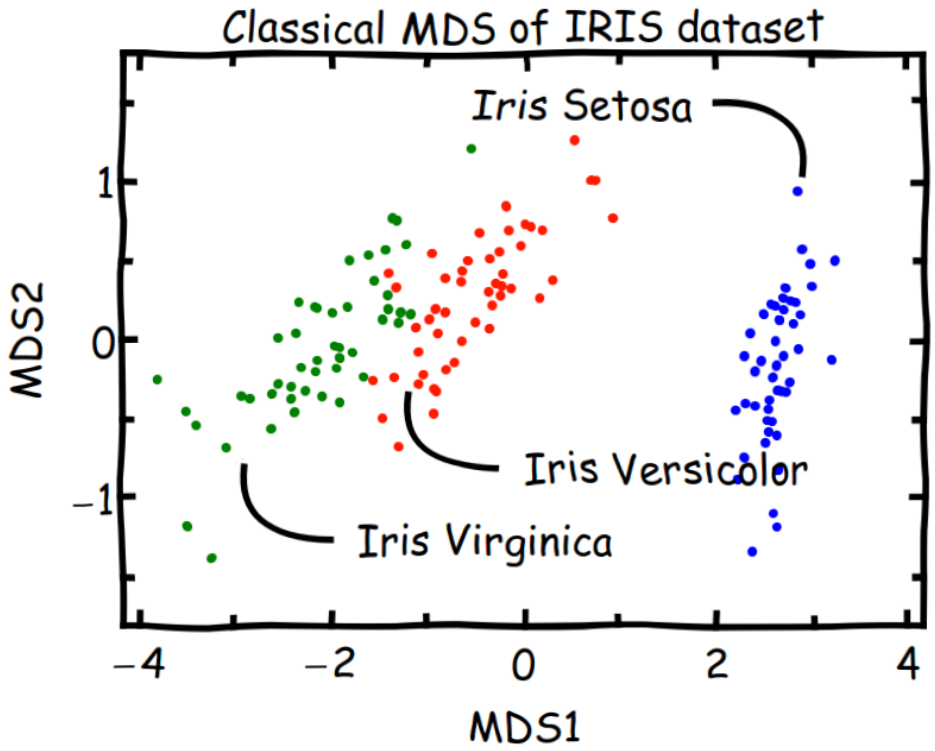
The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D.
  - *Versicolor* and *Virginica* are pretty close...



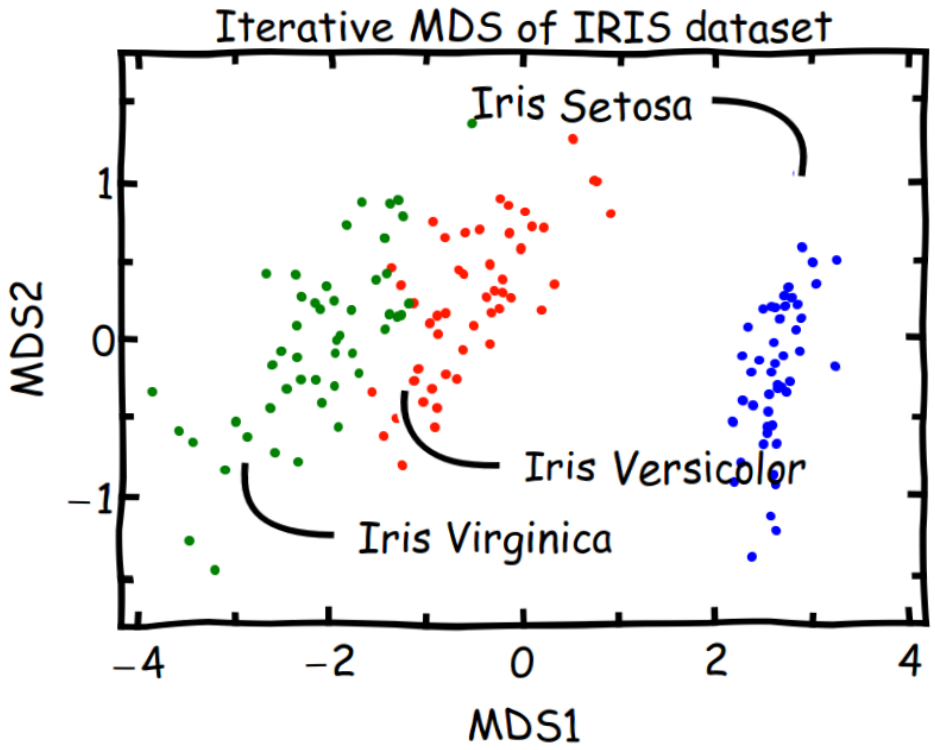
The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D.
  - *Versicolor* and *Virginica* are pretty close...



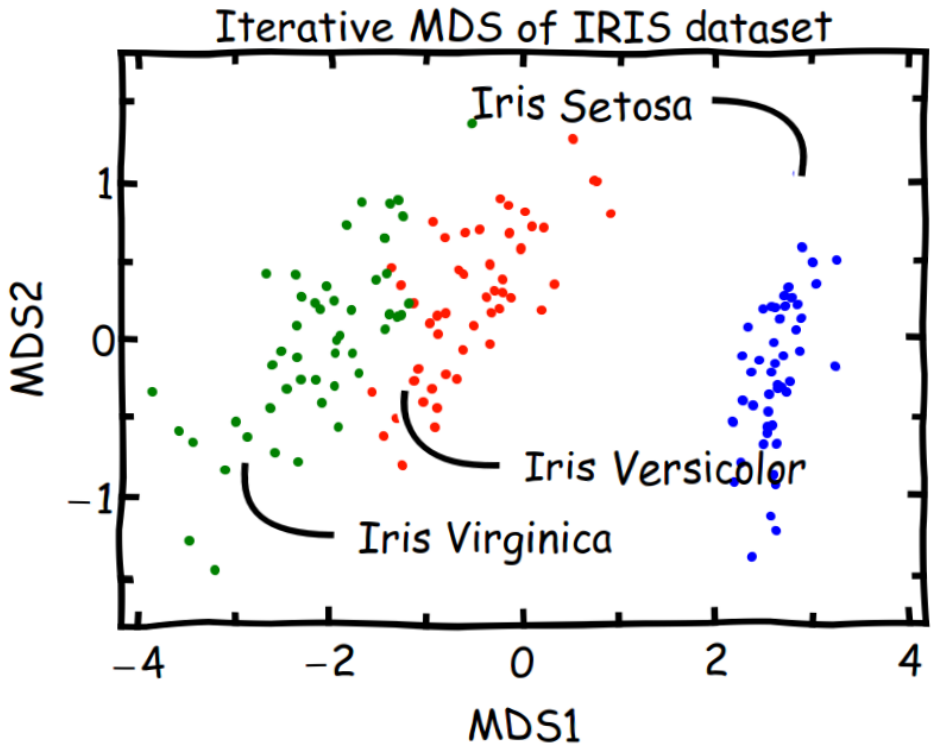
The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D. Iterative MDS does not make a big difference
  - *Versicolor* and *Virginica* are pretty close...



The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D. Iterative MDS does not make a big difference
  - *Versicolor* and *Virginica* are pretty close...



The IRIS dataset

- Measures of four morphological features (petal/sepal width/length) of a total of 150 samples of three species of iris
- 4D dataset, strong correlation between the indicators, but also spread within one species
- Apply PCA and classical MDS – equivalent modulo a rotation.
  - Clearly clustered in 2D. Iterative MDS does not make a big difference
  - *Versicolor* and *Virginica* are pretty close... and they look quite similar!



Iris Setosa



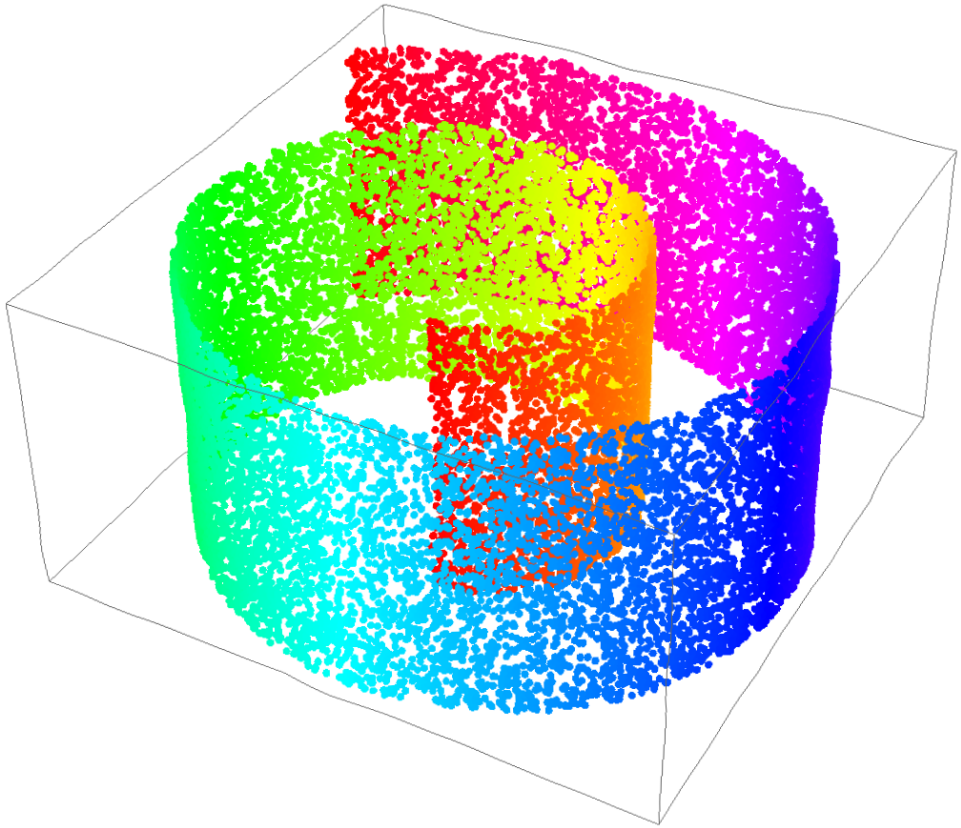
Iris Versicolor



Iris Virginica

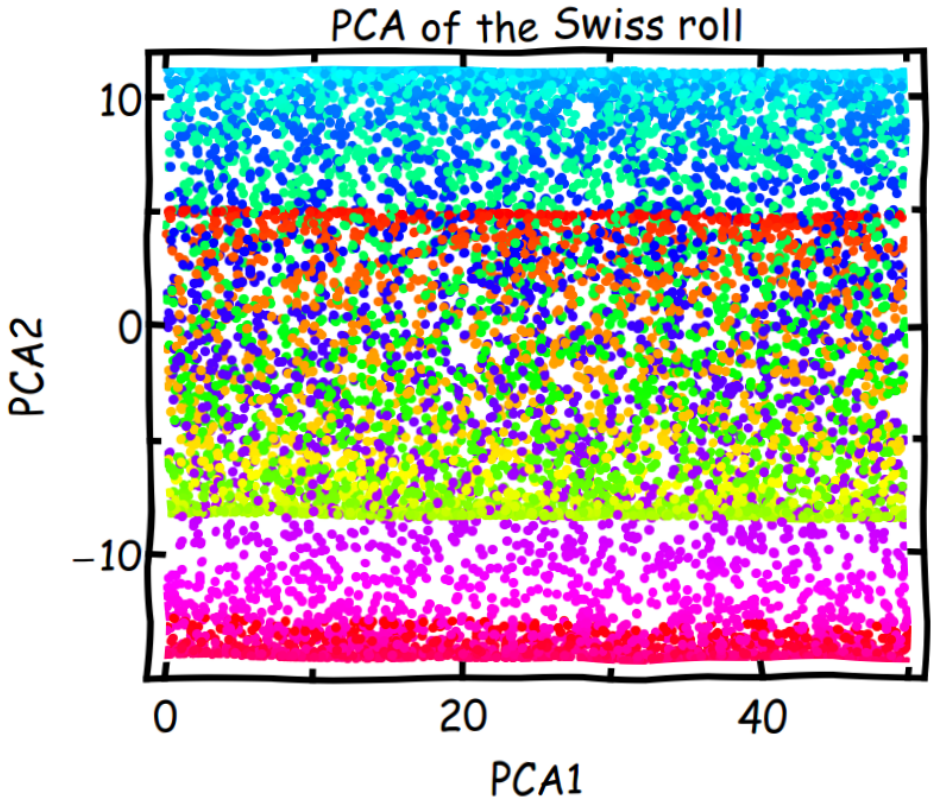
The Swiss roll dataset

- A 2D manifold embedded in three-dimensional space
- PCA cannot capture the low-dimensional structure of the manifold, because it is just a **linear** projection!
  - Linear methods work when data lie (almost) on a plane
  - One would need a method that can deal with a curved manifold which is only **locally** linear



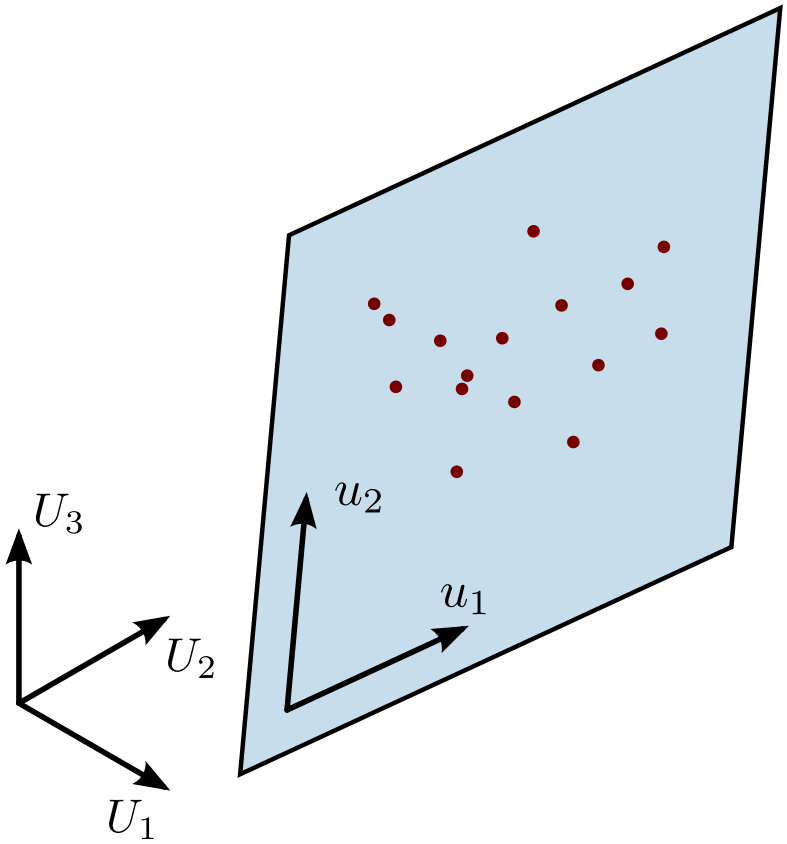
The Swiss roll dataset

- A 2D manifold embedded in three-dimensional space
- PCA cannot capture the low-dimensional structure of the manifold, because it is just a **linear** projection!
  - Linear methods work when data lie (almost) on a plane
  - One would need a method that can deal with a curved manifold which is only **locally** linear



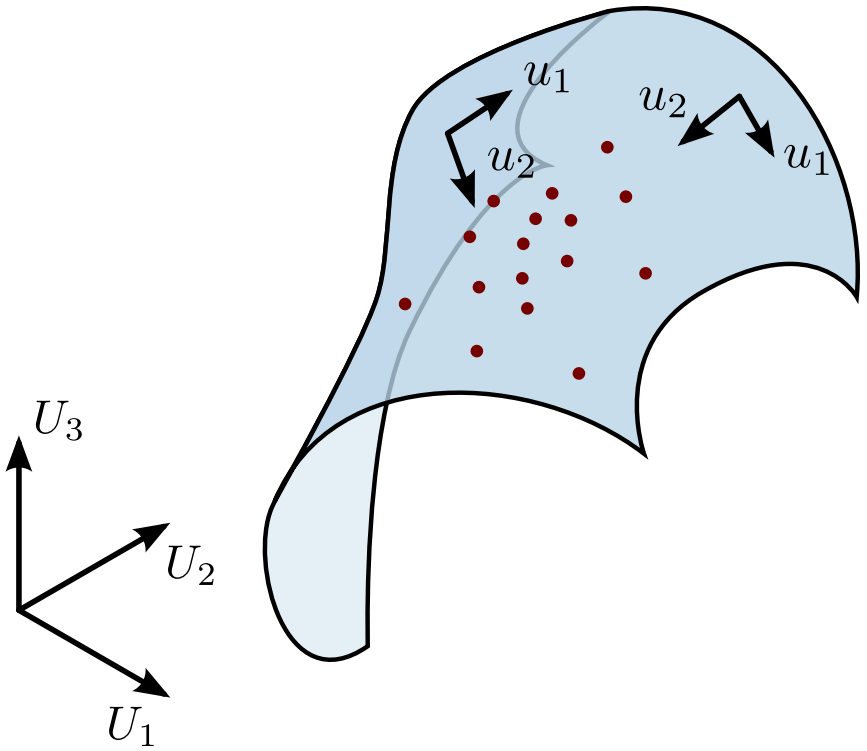
The Swiss roll dataset

- A 2D manifold embedded in three-dimensional space
- PCA cannot capture the low-dimensional structure of the manifold, because it is just a **linear** projection!
  - Linear methods work when data lie (almost) on a plane
  - One would need a method that can deal with a curved manifold which is only **locally** linear



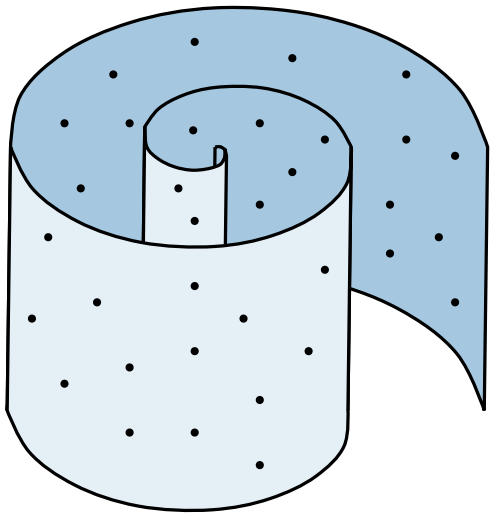
The Swiss roll dataset

- A 2D manifold embedded in three-dimensional space
- PCA cannot capture the low-dimensional structure of the manifold, because it is just a **linear** projection!
  - Linear methods work when data lie (almost) on a plane
  - One would need a method that can deal with a curved manifold which is only **locally** linear



Non-Linear DR: ISOMAP

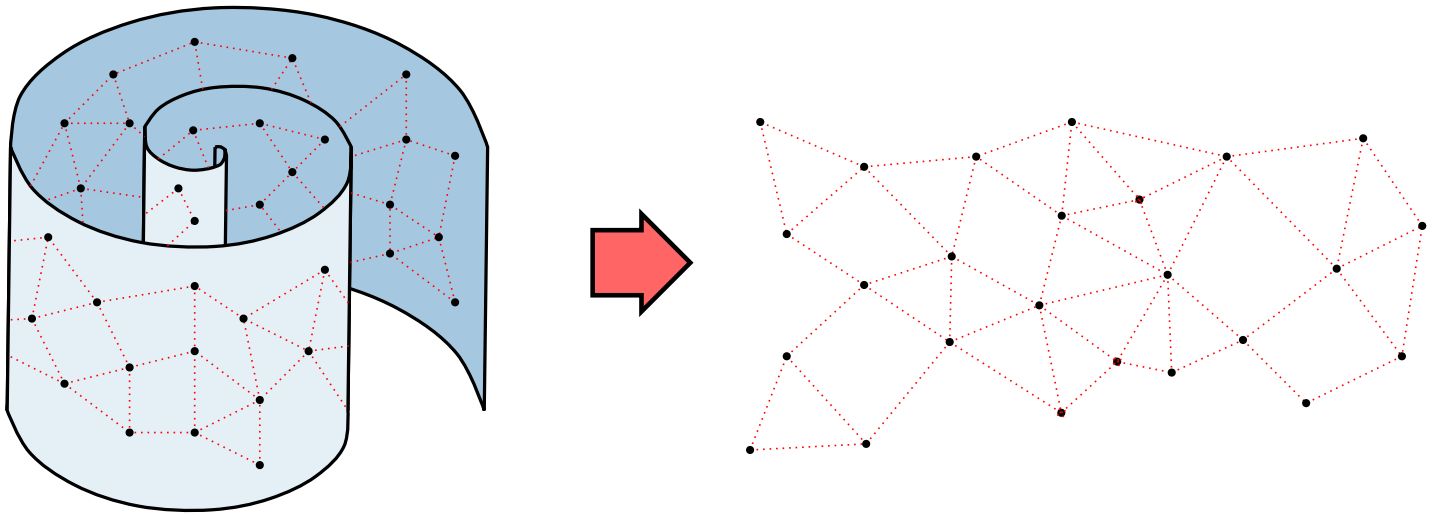
- A family of methods introduces non-linearity in the dissimilarity metric
- ISOMAP defines point-point distances based on geodesics
  - Approximate geodesics by hopping between neighbours
  - Problem: very sensitive to uneven sampling and noise. One "wrong" neighbor detection can mess up geodesics completely.



Tenenbaum et al., Science (2000)

Non-Linear DR: ISOMAP

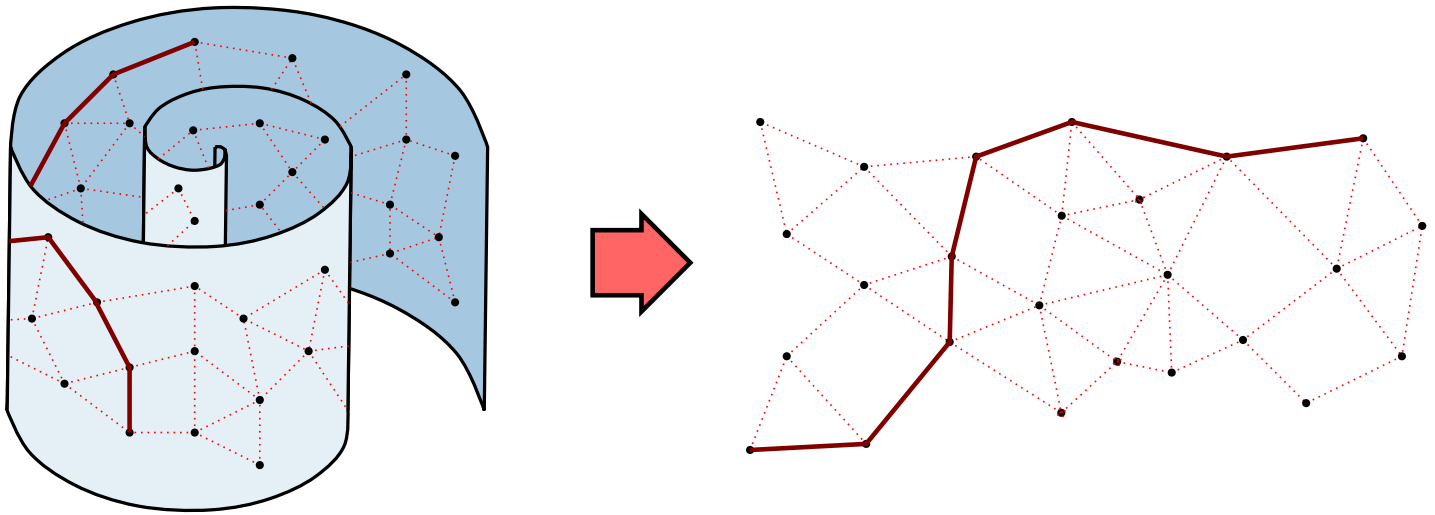
- A family of methods introduces non-linearity in the dissimilarity metric
- ISOMAP defines point-point distances based on geodesics
  - Approximate geodesics by hopping between neighbours
  - Problem: very sensitive to uneven sampling and noise. One "wrong" neighbor detection can mess up geodesics completely.



Tenenbaum et al., Science (2000)

Non-Linear DR: ISOMAP

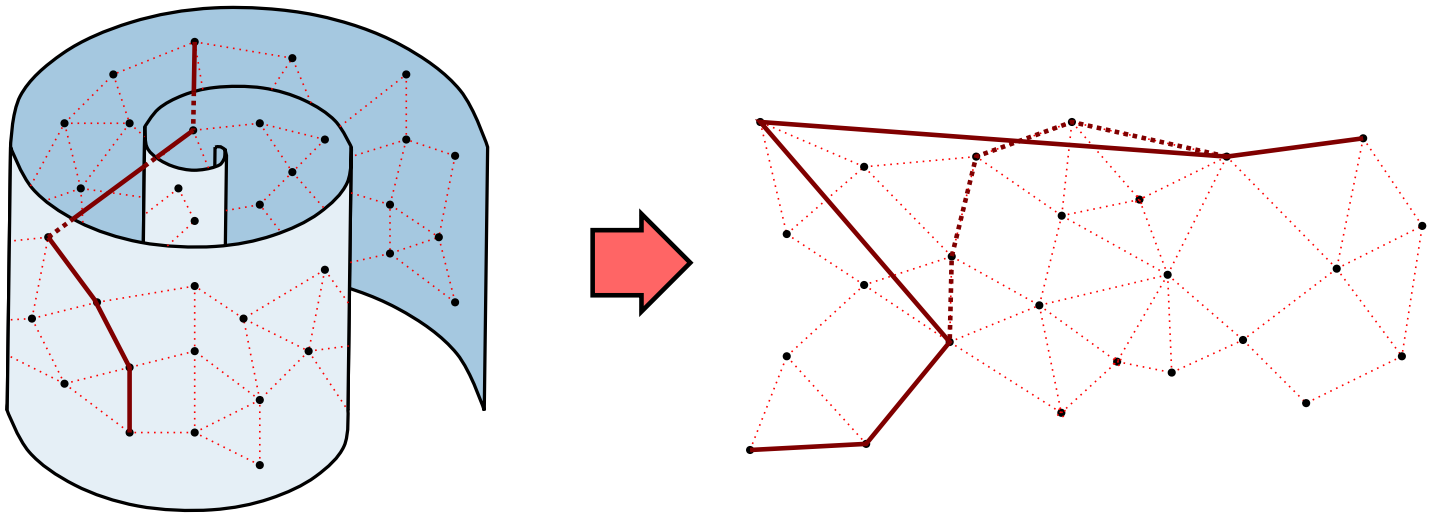
- A family of methods introduces non-linearity in the dissimilarity metric
- ISOMAP defines point-point distances based on geodesics
  - Approximate geodesics by hopping between neighbours
  - Problem: very sensitive to uneven sampling and noise. One "wrong" neighbor detection can mess up geodesics completely.



Tenenbaum et al., Science (2000)

Non-Linear DR: ISOMAP

- A family of methods introduces non-linearity in the dissimilarity metric
- ISOMAP defines point-point distances based on geodesics
  - Approximate geodesics by hopping between neighbours
  - Problem: very sensitive to uneven sampling and noise. One "wrong" neighbor detection can mess up geodesics completely.



Tenenbaum et al., Science (2000)

Non-Linear DR: ISOMAP

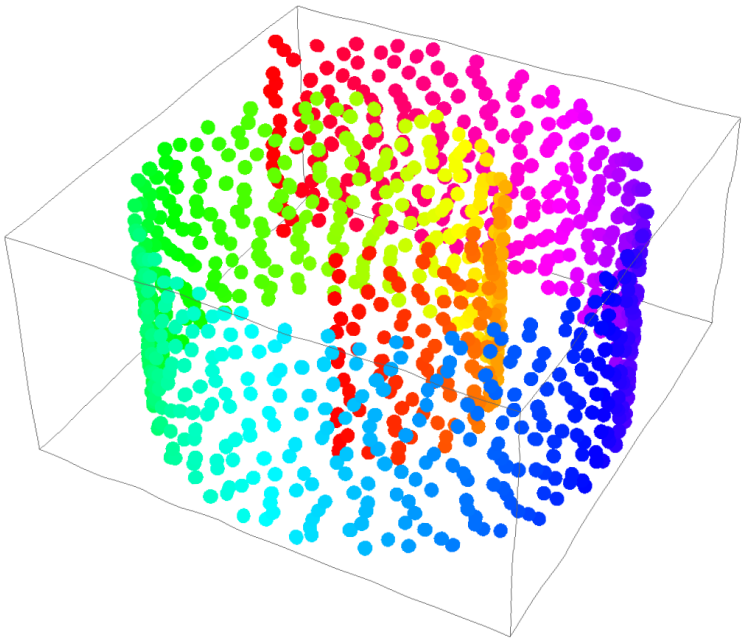
- A family of methods introduces non-linearity in the dissimilarity metric
- ISOMAP defines point-point distances based on geodesics
  - Approximate geodesics by hopping between neighbours
  - Problem: very sensitive to uneven sampling and noise. One "wrong" neighbor detection can mess up geodesics completely.

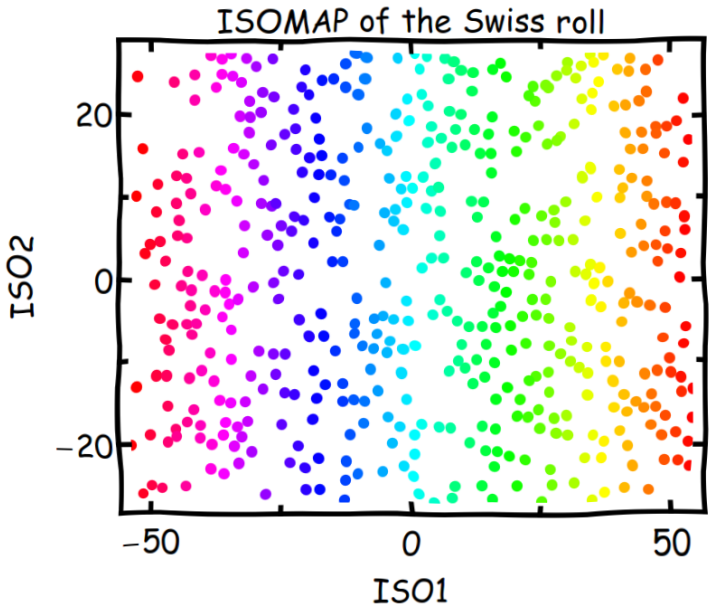
1. Define neighbour relations between points (k nearest neighbours or points within ϵ)
2. Compute the graph distance matrix as an approximant to geodesics
3. Run classical MDS.

Tenenbaum et al., Science (2000)

ISOMAP and the Swiss roll

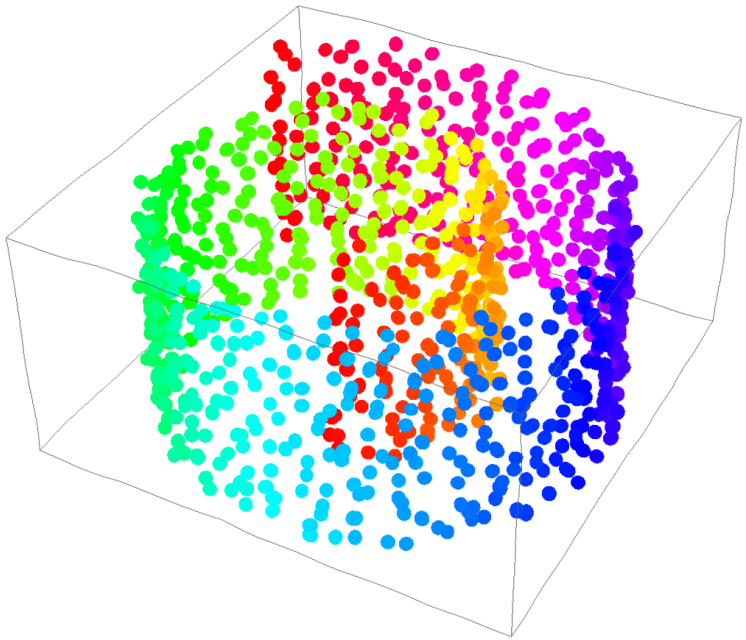
- ISOMAP works very well for the Swiss roll, as it identifies beautifully the manifold directions
- It is however very sensitive to noise, and to uneven sampling. When it fails, it fails dramatically!

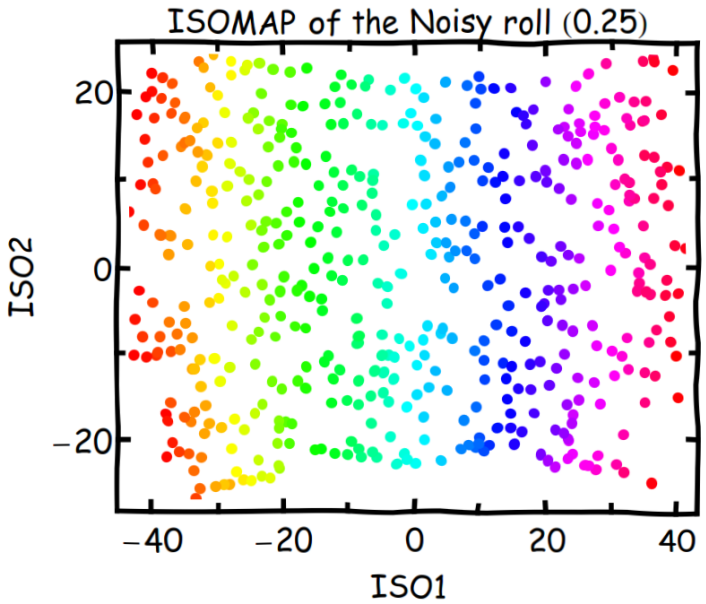




ISOMAP and the Swiss roll

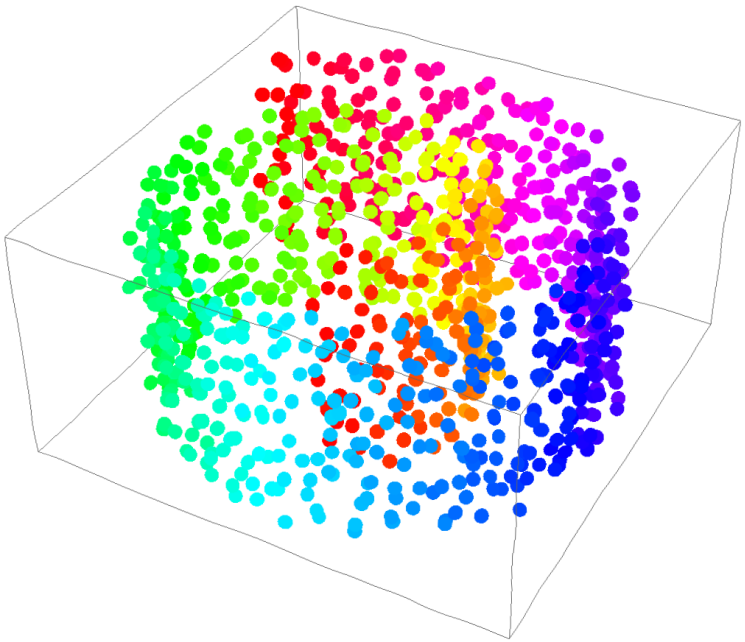
- ISOMAP works very well for the Swiss roll, as it identifies beautifully the manifold directions
- It is however very sensitive to noise, and to uneven sampling. When it fails, it fails dramatically!

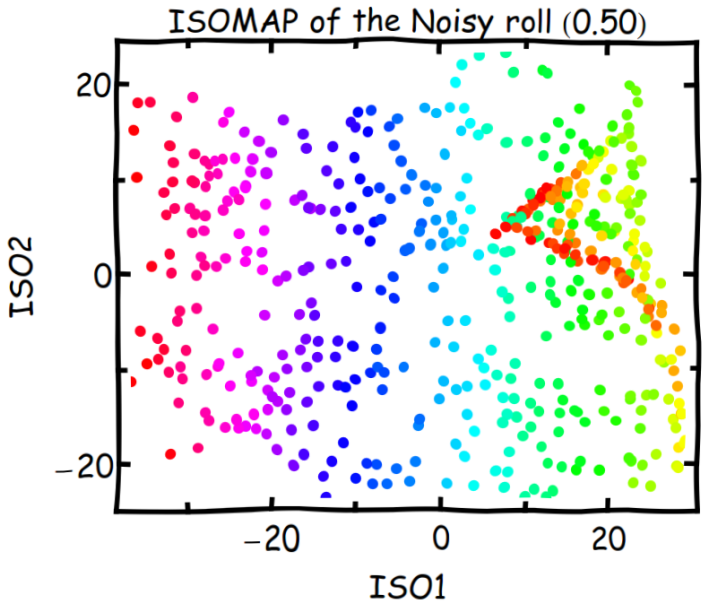




ISOMAP and the Swiss roll

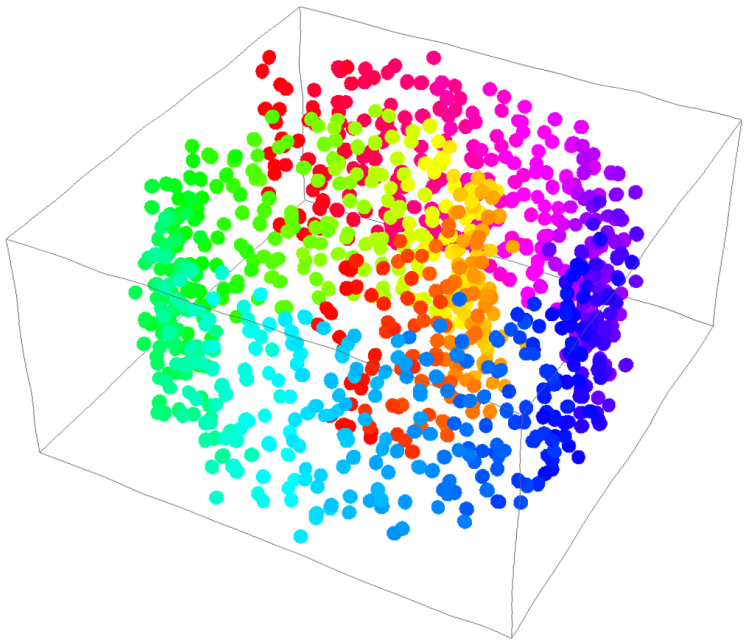
- ISOMAP works very well for the Swiss roll, as it identifies beautifully the manifold directions
- It is however very sensitive to noise, and to uneven sampling. When it fails, it fails dramatically!

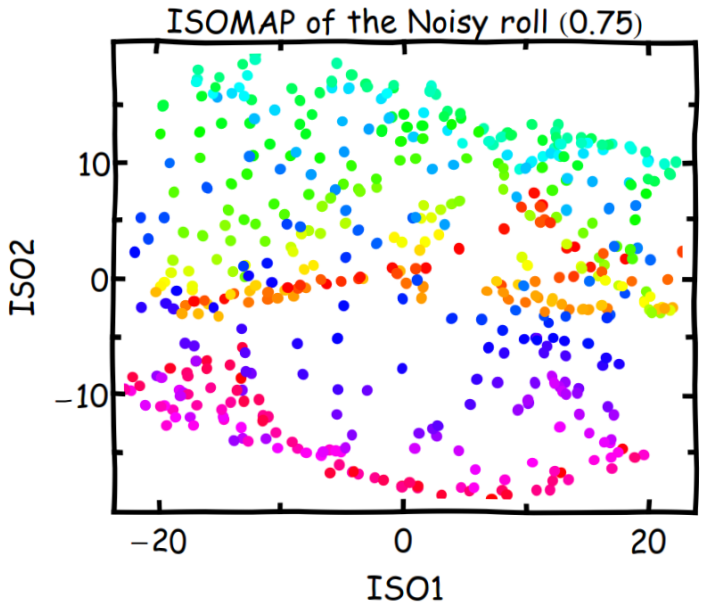




ISOMAP and the Swiss roll

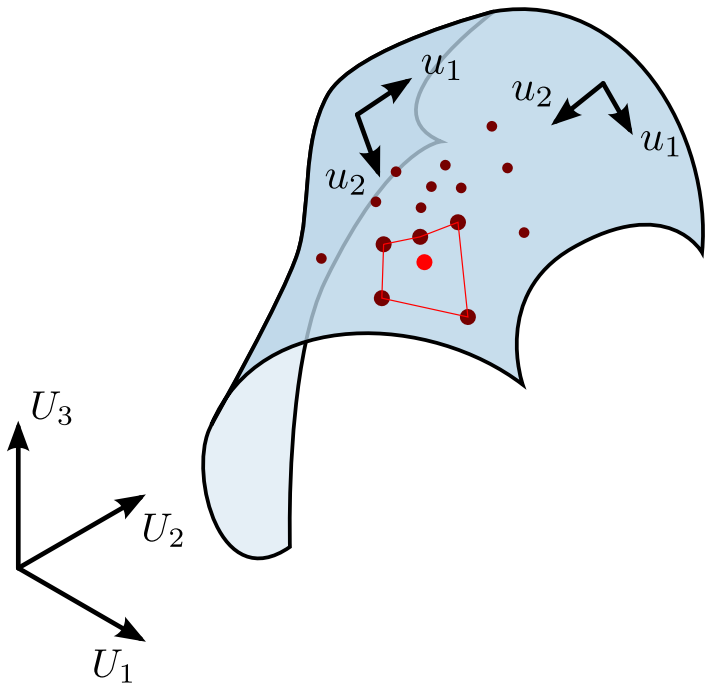
- ISOMAP works very well for the Swiss roll, as it identifies beautifully the manifold directions
- It is however very sensitive to noise, and to uneven sampling. When it fails, it fails dramatically!





Locally Linear Embedding

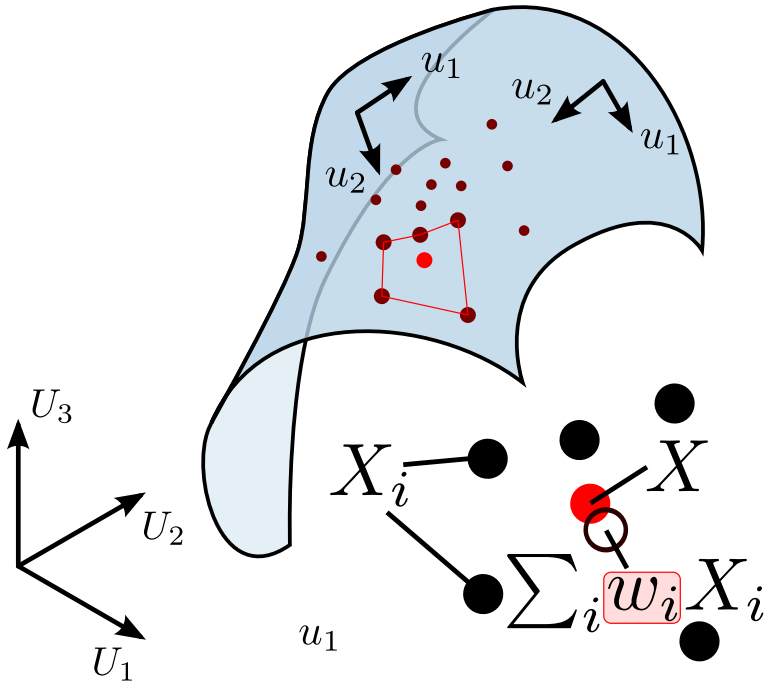
- Use the fact that if the manifold is locally flat each point can be expressed as a combination of its neighbors.
  1. Determine a neighborhood of each point X, and the weights w_i that best match X and its embedding
  2. Determine the low-dimensional points such that for each point, x an its embedding are as close as possible *keeping the weights fixed*
  3. This is again formulated as a linear algebra problem



Roweis, Saul, Science (2000)

Locally Linear Embedding

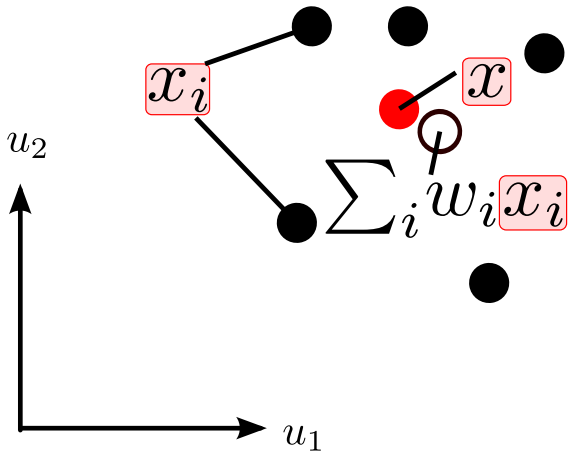
- Use the fact that if the manifold is locally flat each point can be expressed as a combination of its neighbors.
  1. Determine a neighborhood of each point X, and the weights w_i that best match X and its embedding
  2. Determine the low-dimensional points such that for each point, x an its embedding are as close as possible *keeping the weights fixed*
  3. This is again formulated as a linear algebra problem



Roweis, Saul, Science (2000)

Locally Linear Embedding

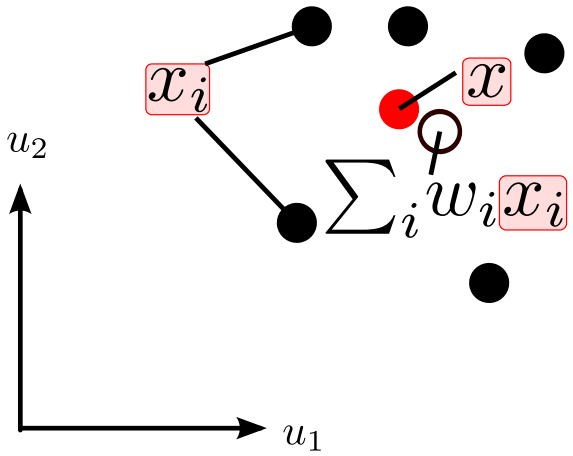
- Use the fact that if the manifold is locally flat each point can be expressed as a combination of its neighbors.
  1. Determine a neighborhood of each point X, and the weights w_i that best match X and its embedding
  2. Determine the low-dimensional points such that for each point, x an its embedding are as close as possible *keeping the weights fixed*
  3. This is again formulated as a linear algebra problem



Roweis, Saul, Science (2000)

Locally Linear Embedding

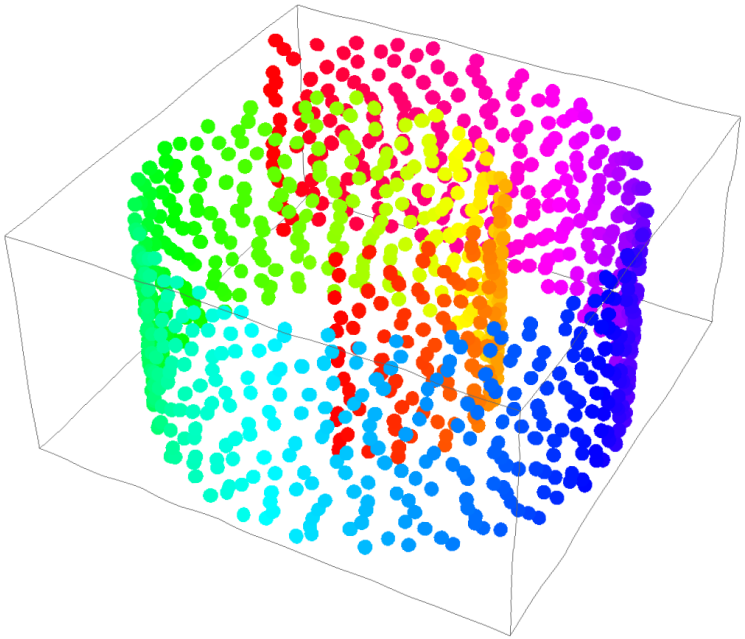
- Use the fact that if the manifold is locally flat each point can be expressed as a combination of its neighbors.
  1. Determine a neighborhood of each point X, and the weights w_i that best match X and its embedding
  2. Determine the low-dimensional points such that for each point, x an its embedding are as close as possible *keeping the weights fixed*
  3. This is again formulated as a linear algebra problem

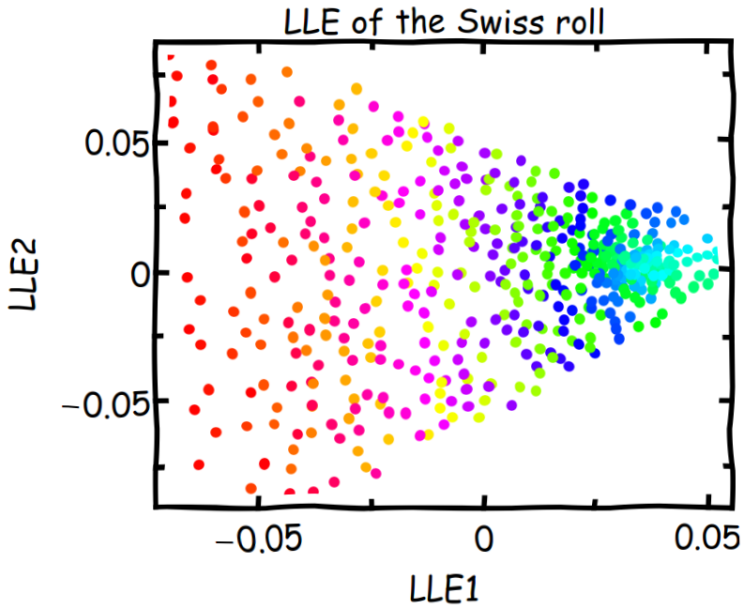


Roweis, Saul, Science (2000)

LLE and the Swiss roll

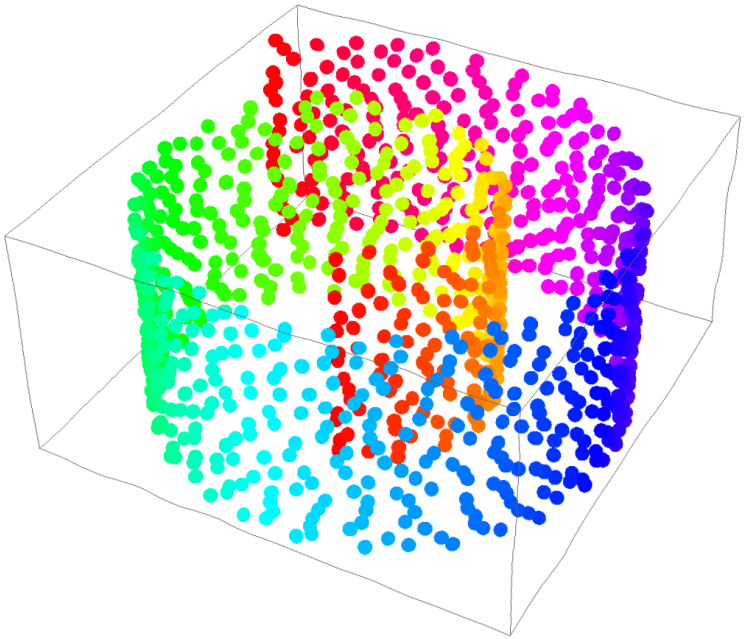
- LLE seems not to work . . . 2-but some higher variables do capture the manifold structure
- Growing noise destabilizes the embedding, and shifts to even higher-order LLE vectors the reasonable map

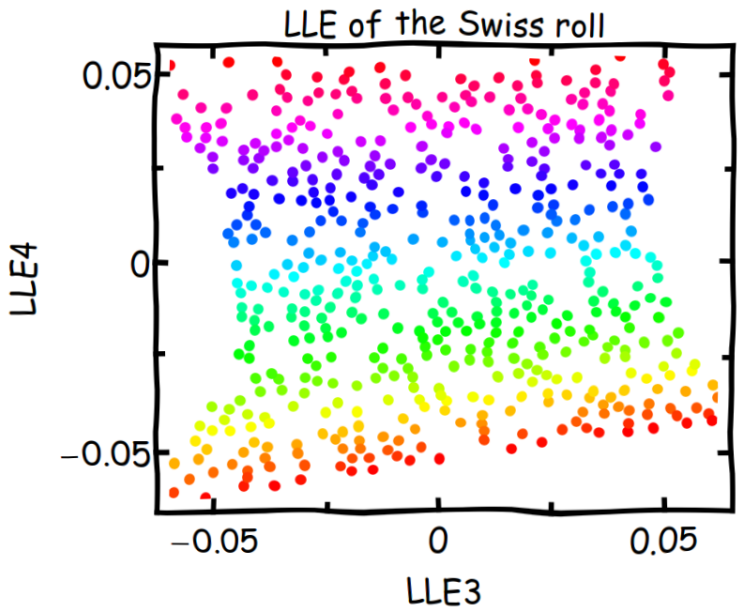




LLE and the Swiss roll

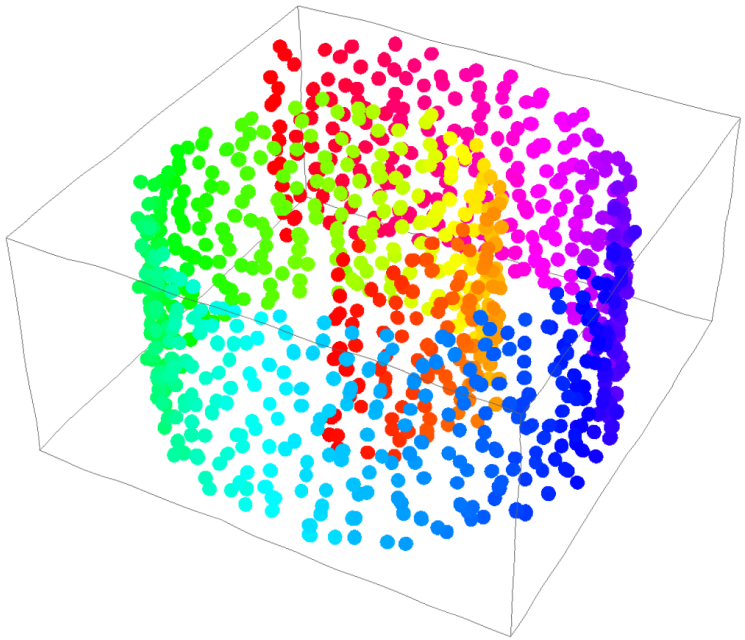
- LLE seems not to work . . . 2-but some higher variables do capture the manifold structure
- Growing noise destabilizes the embedding, and shifts to even higher-order LLE vectors the reasonable map

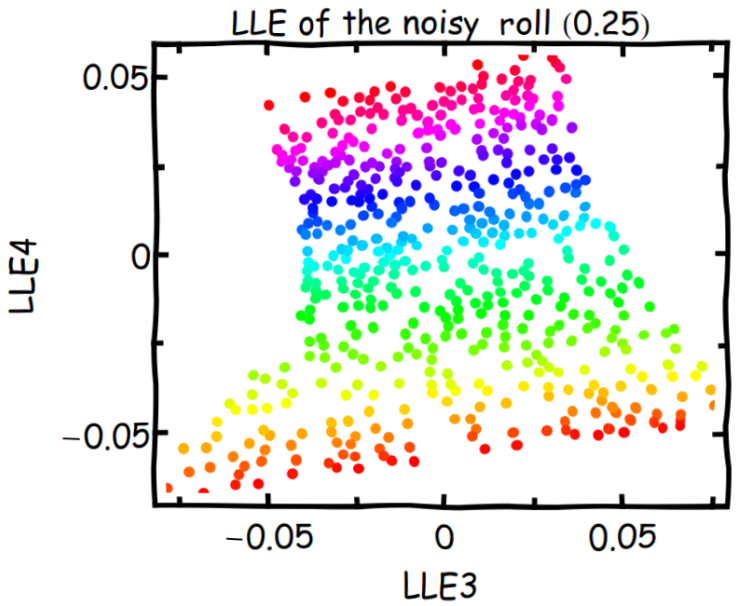




LLE and the Swiss roll

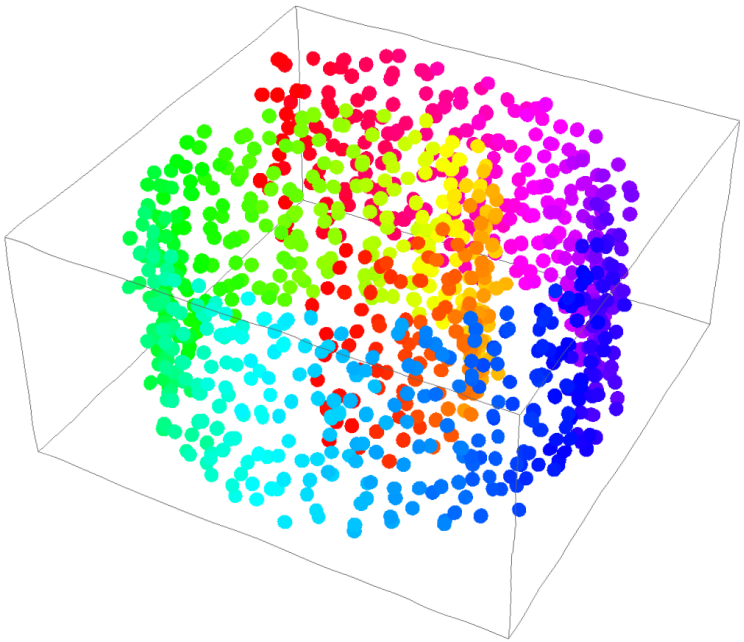
- LLE seems not to work . . . 2-but some higher variables do capture the manifold structure
- Growing noise destabilizes the embedding, and shifts to even higher-order LLE vectors the reasonable map

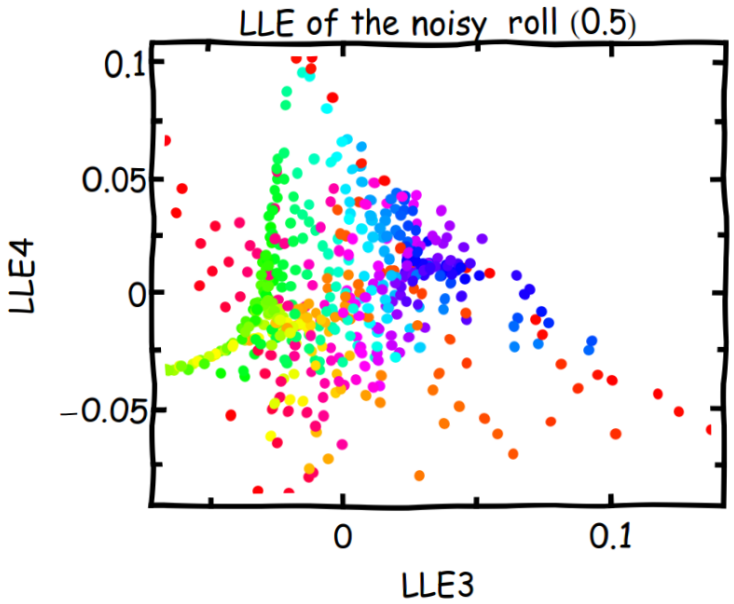




LLE and the Swiss roll

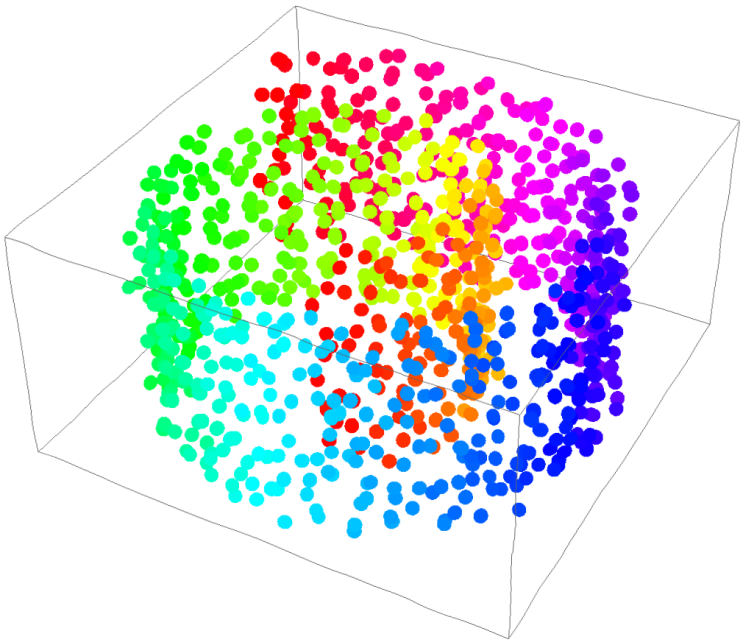
- LLE seems not to work . . . 2-but some higher variables do capture the manifold structure
- Growing noise destabilizes the embedding, and shifts to even higher-order LLE vectors the reasonable map

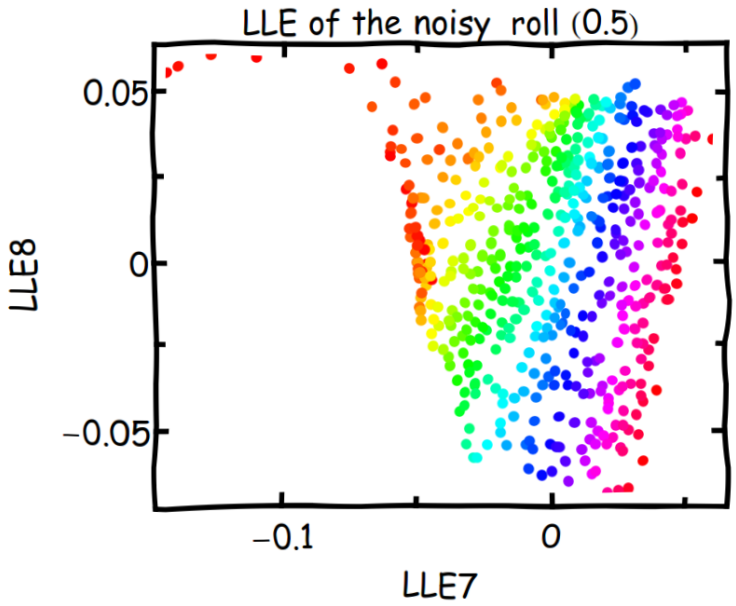




LLE and the Swiss roll

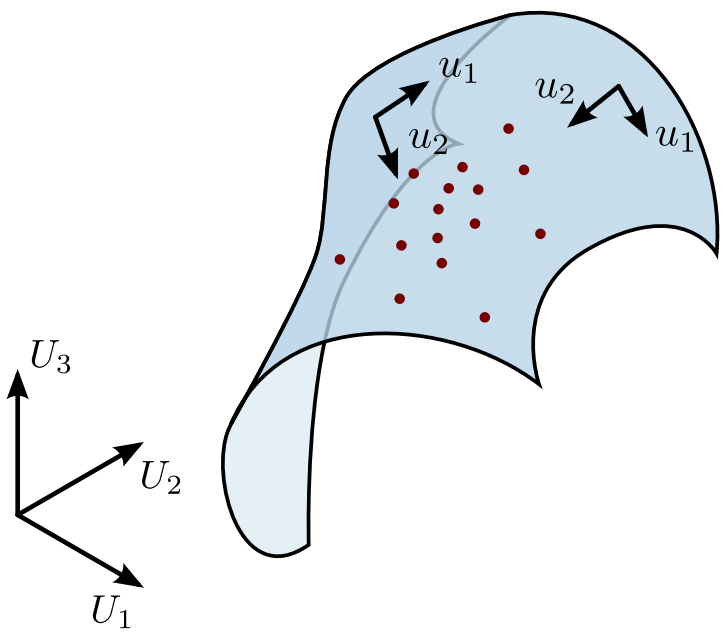
- LLE seems not to work . . . 2-but some higher variables do capture the manifold structure
- Growing noise destabilizes the embedding, and shifts to even higher-order LLE vectors the reasonable map

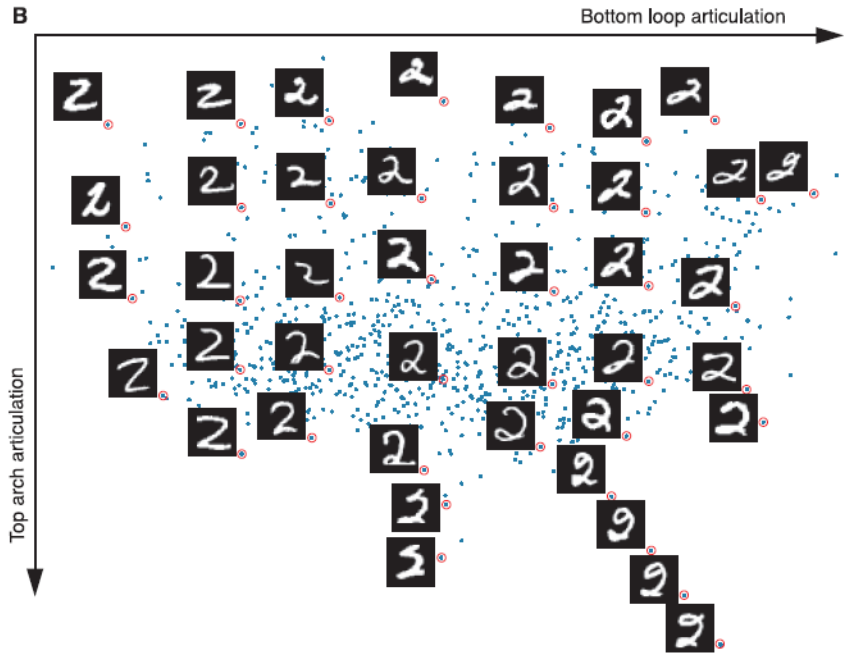




NLDR and atomistic simulations

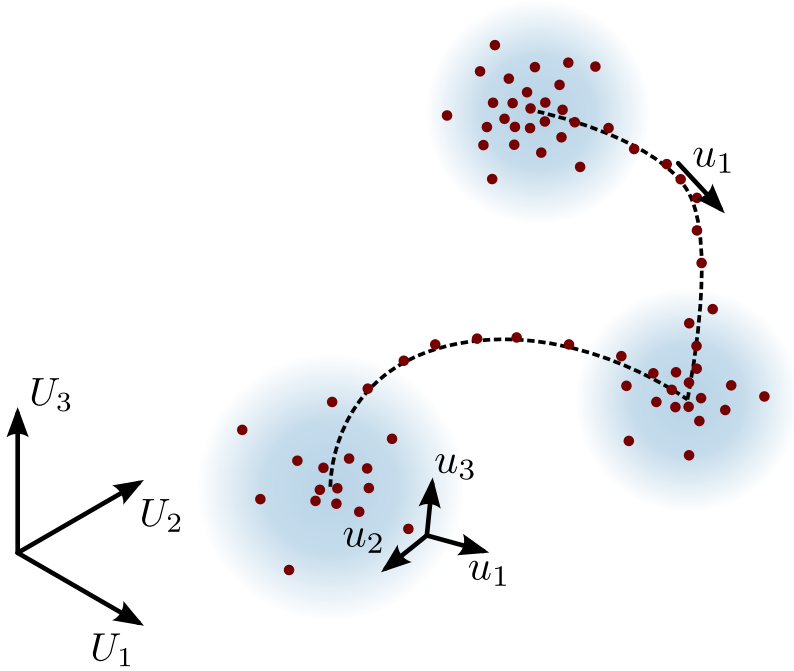
- Non-linear dimensionality reduction algorithms:
  - Describe curved, "locally-flat" manifolds
  - Developed by the CS community (image recognition)
  - Attempts to apply to chemical problems (PCA, ISOMAP, LLE,...)
- Atomistic simulations are harder:
  - **Thermal fluctuations** are high-dimensional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

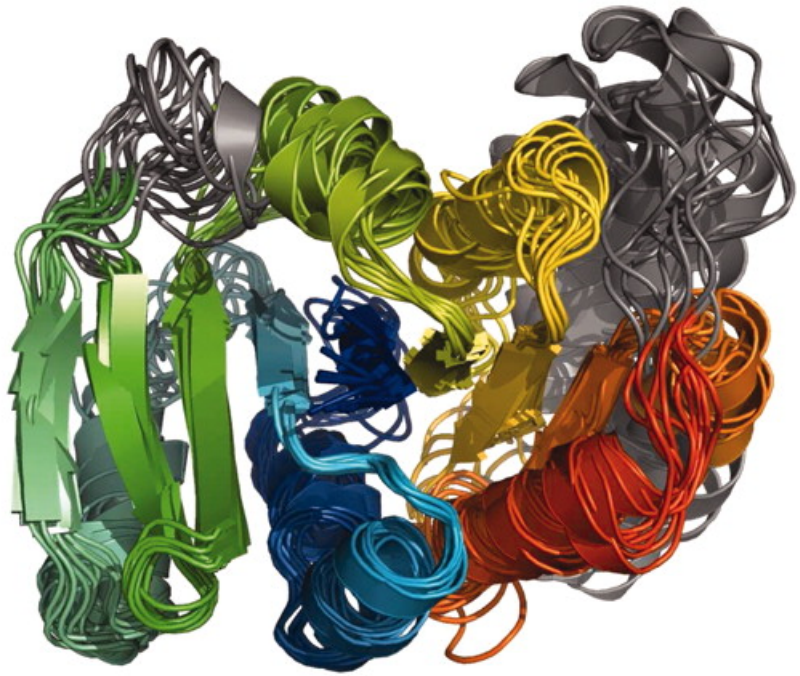




NLDR and atomistic simulations

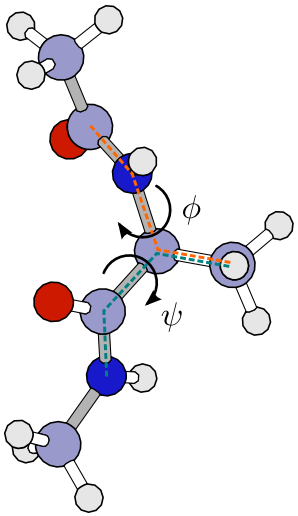
- Non-linear dimensionality reduction algorithms:
  - Describe curved, "locally-flat" manifolds
  - Developed by the CS community (image recognition)
  - Attempts to apply to chemical problems (PCA, ISOMAP, LLE,...)
- Atomistic simulations are harder:
  - **Thermal fluctuations** are high-dimensional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

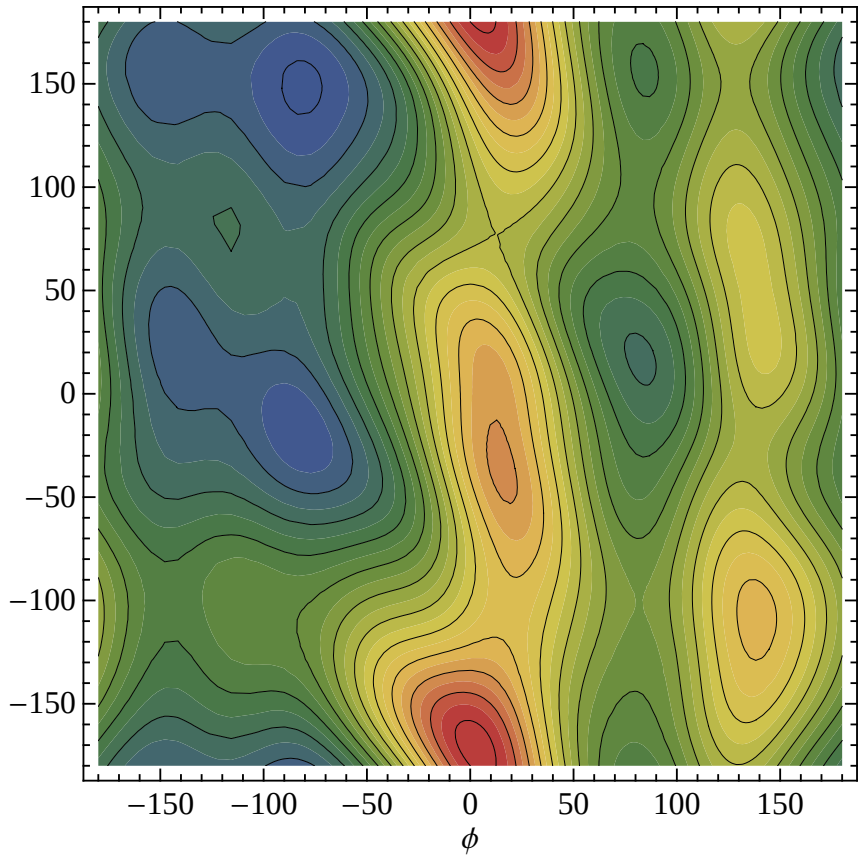




Features of a folding landscape

- The free energy landscape for ala_2 contains low-energy basins and a spider web of transition pathways
- Reconnaissance metadynamics[1] for ala_{12}: similar distribution of points for any pair of dihedrals

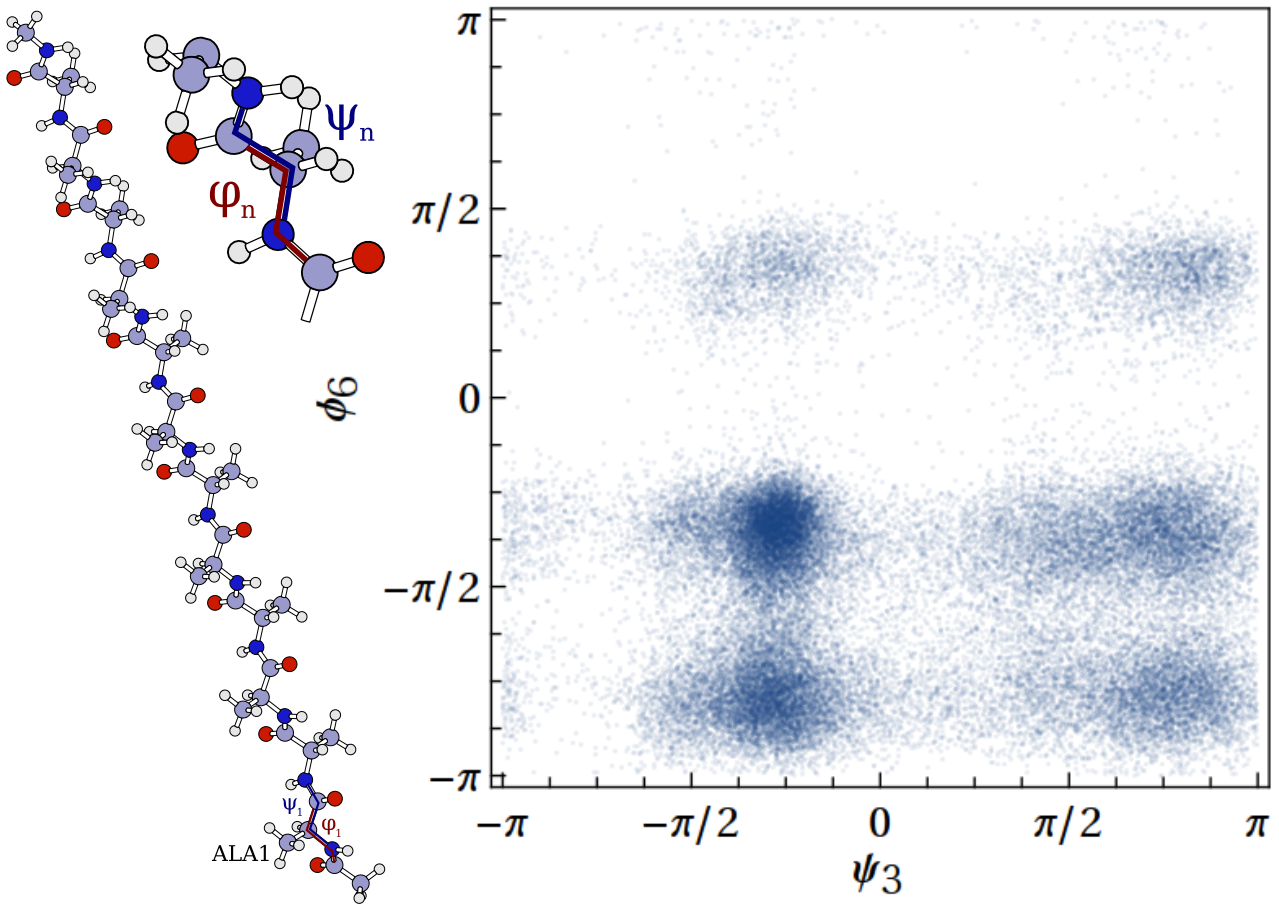




[1] Tribello, Ceriotti, Parrinello, PNAS 2010

Features of a folding landscape

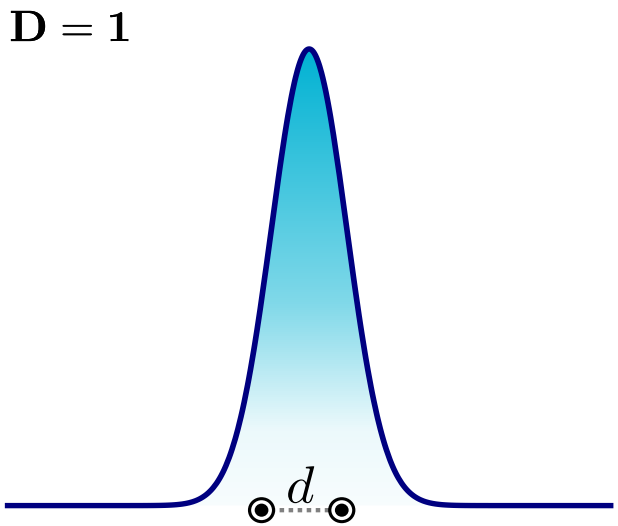
- The free energy landscape for ala_2 contains low-energy basins and a spider web of transition pathways
- Reconnaissance metadynamics[1] for ala_{12}: similar distribution of points for any pair of dihedrals



[1] Tribello, Ceriotti, Parrinello, PNAS 2010

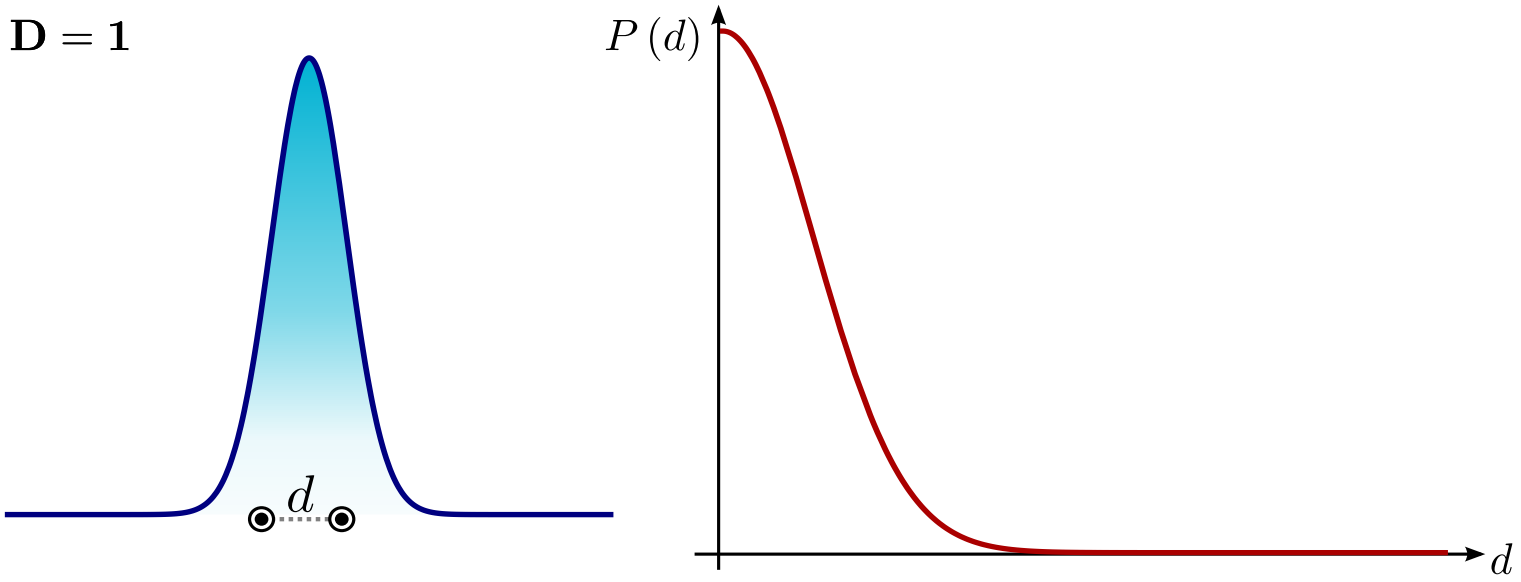
Matching fluctuations: a space odyssey

- Inherent problem when projecting **full-dimensional features**
- Take for instance the distribution of distances between points taken from a D-dimensional Gaussian
- This is a disaster for distance matching! It is *impossible* to match the distances for a 24-dimensional Gaussian using a 3d Gaussian!



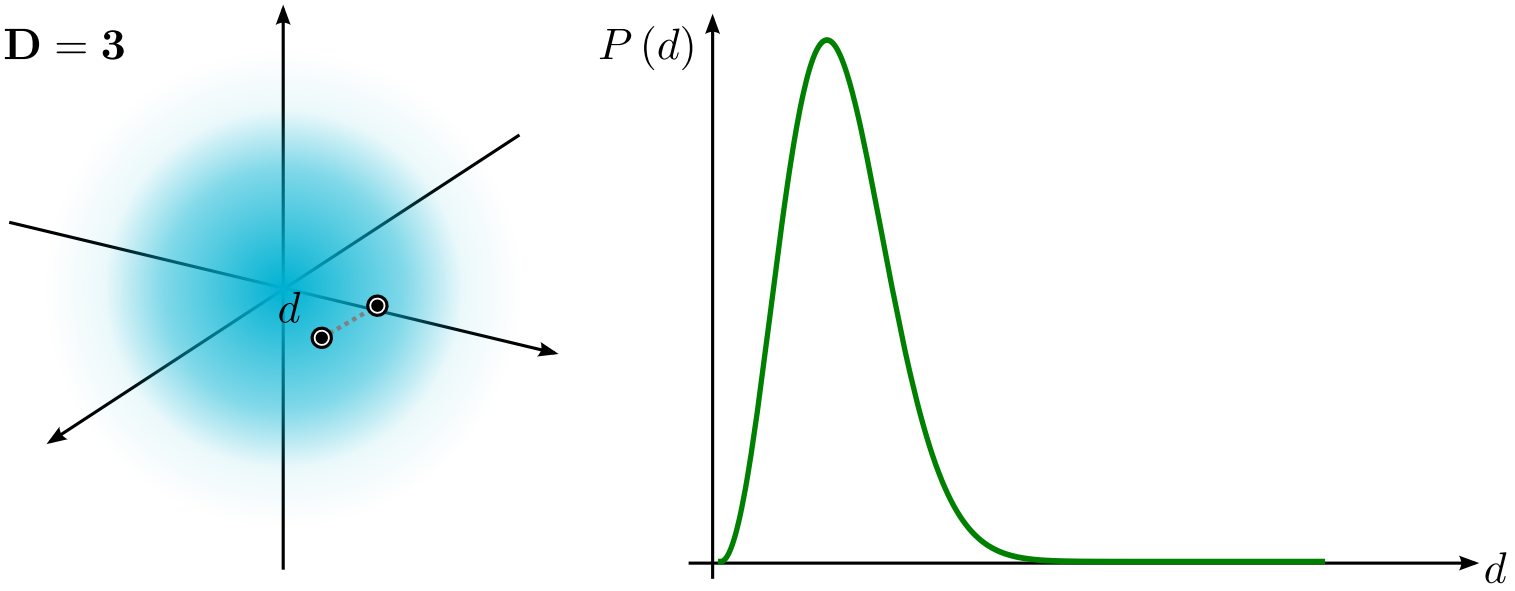
Matching fluctuations: a space odyssey

- Inherent problem when projecting **full-dimensional features**
- Take for instance the distribution of distances between points taken from a D-dimensional Gaussian
- This is a disaster for distance matching! It is *impossible* to match the distances for a 24-dimensional Gaussian using a 3d Gaussian!



Matching fluctuations: a space odyssey

- Inherent problem when projecting **full-dimensional features**
- Take for instance the distribution of distances between points taken from a D-dimensional Gaussian
- This is a disaster for distance matching! It is *impossible* to match the distances for a 24-dimensional Gaussian using a 3d Gaussian!

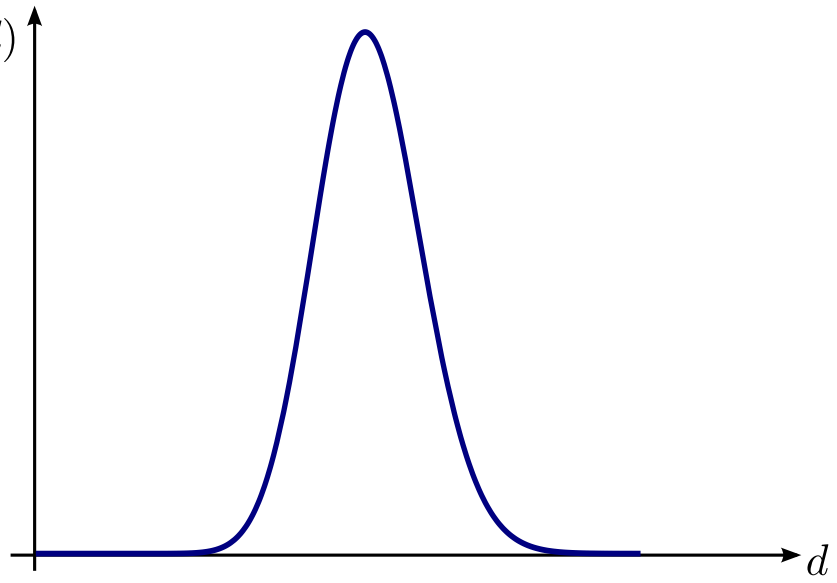


Matching fluctuations: a space odyssey

- Inherent problem when projecting **full-dimensional features**
- Take for instance the distribution of distances between points taken from a D-dimensional Gaussian
- This is a disaster for distance matching! It is *impossible* to match the distances for a 24-dimensional Gaussian using a 3d Gaussian!

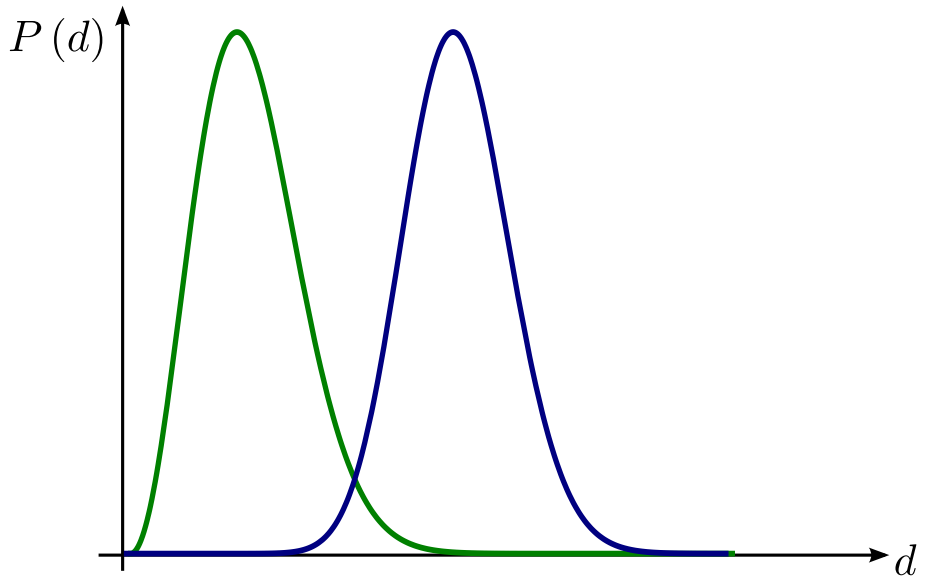
$\mathbf{D = 24}$





Matching fluctuations: a space odyssey

- Inherent problem when projecting **full-dimensional features**
- Take for instance the distribution of distances between points taken from a D-dimensional Gaussian
- This is a disaster for distance matching! It is *impossible* to match the distances for a 24-dimensional Gaussian using a 3d Gaussian!



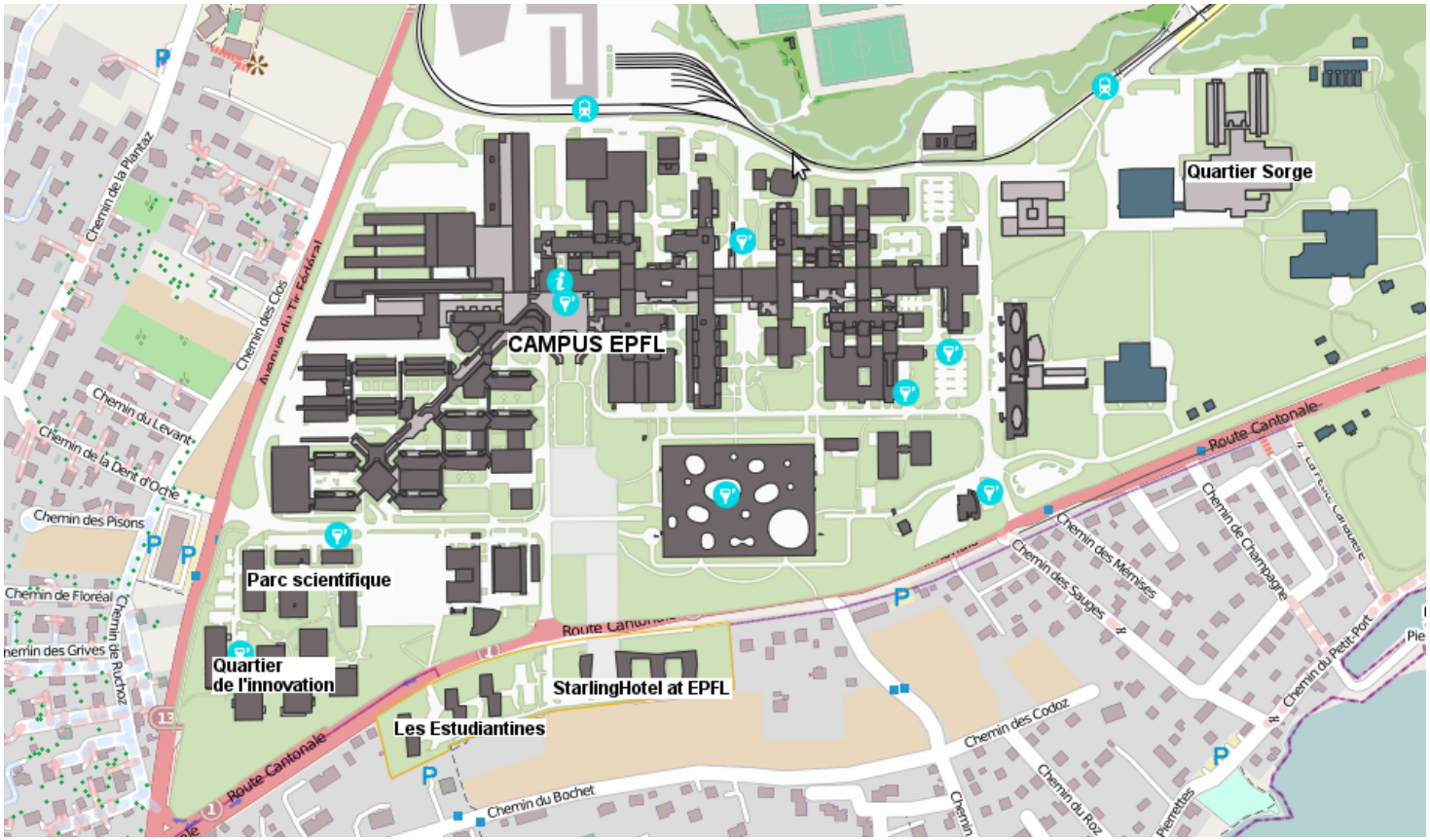
No need for a perfect map

- Developing a NLDR method which is more robust and suited for trajectory data
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map



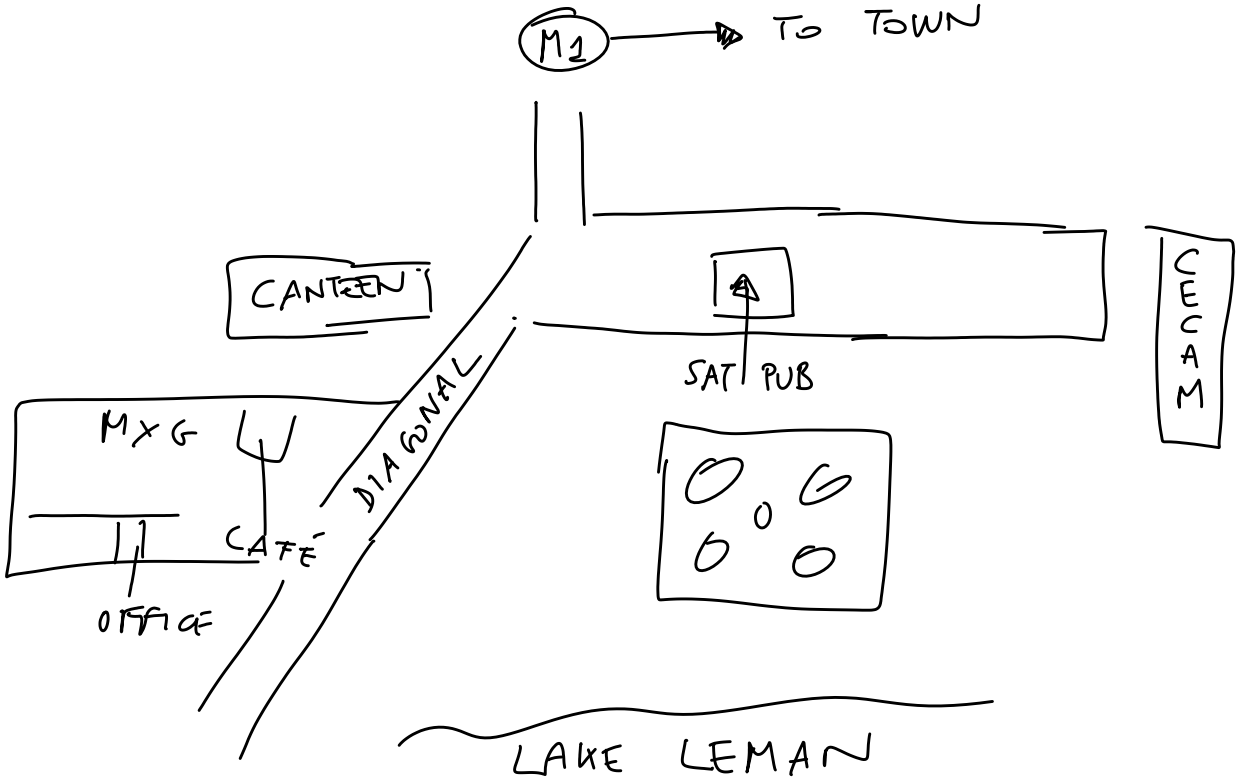
No need for a perfect map

- Developing a NLDR method which is more robust and suited for trajectory data
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map



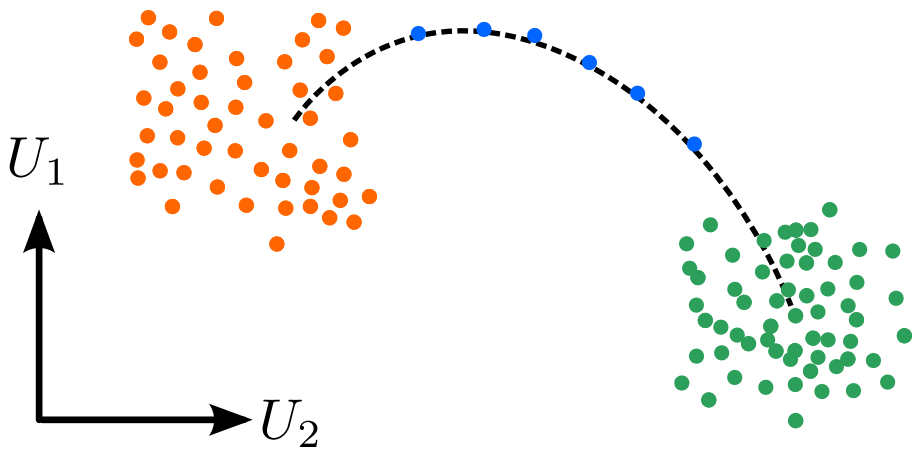
No need for a perfect map

- Developing a NLDR method which is more robust and suited for trajectory data
  - Basic idea: we don't need a precise, isometric map.
  - We need the computational equivalent of a hand sketched map



Proximity matching

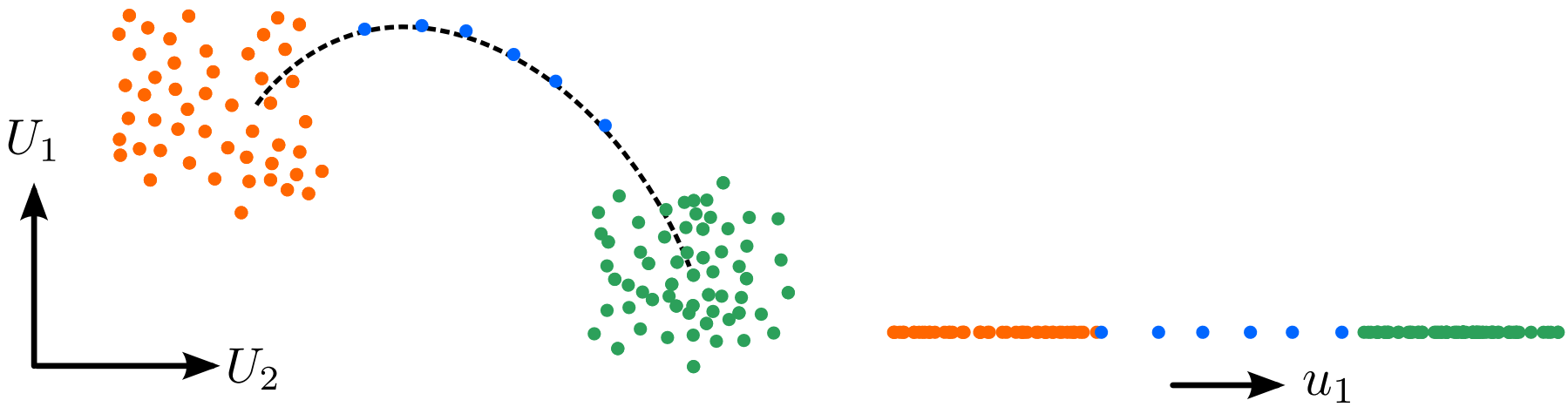
- We would like to capture the low-dimensional structure of complex transitions
- How to deal with full-dimensional thermal fluctuations? Portions of the landscape cannot be projected by matching high and low-dimensional distances.
  - Idea: simpler task, aim for **proximity matching**: close↔close, far↔far



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

Proximity matching

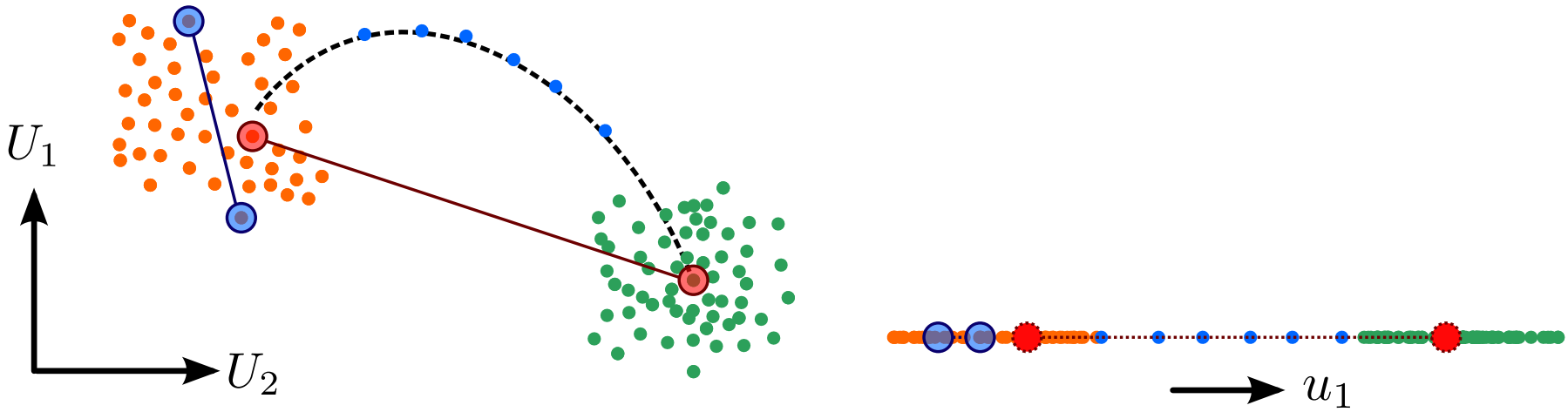
- We would like to capture the low-dimensional structure of complex transitions
- How to deal with full-dimensional thermal fluctuations? Portions of the landscape cannot be projected by matching high and low-dimensional distances.
  - Idea: simpler task, aim for **proximity matching**: close↔close, far↔far



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

Proximity matching

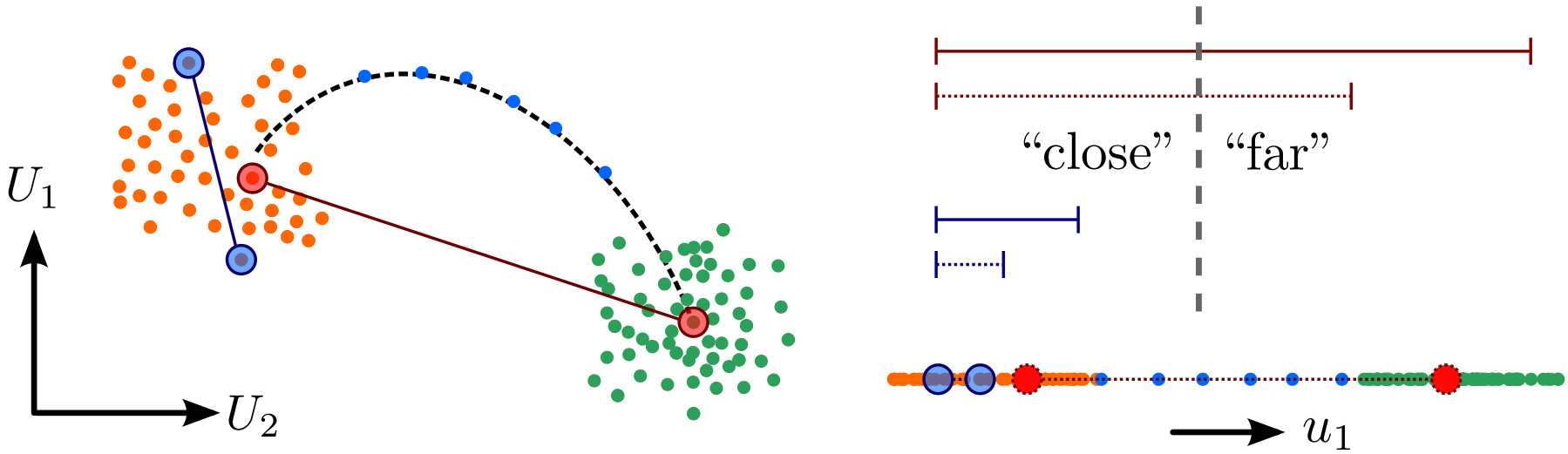
- We would like to capture the low-dimensional structure of complex transitions
- How to deal with full-dimensional thermal fluctuations? Portions of the landscape cannot be projected by matching high and low-dimensional distances.
  - Idea: simpler task, aim for **proximity matching**: close↔close, far↔far



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

Proximity matching

- We would like to capture the low-dimensional structure of complex transitions
- How to deal with full-dimensional thermal fluctuations? Portions of the landscape cannot be projected by matching high and low-dimensional distances.
  - Idea: simpler task, aim for **proximity matching**: close↔close, far↔far



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

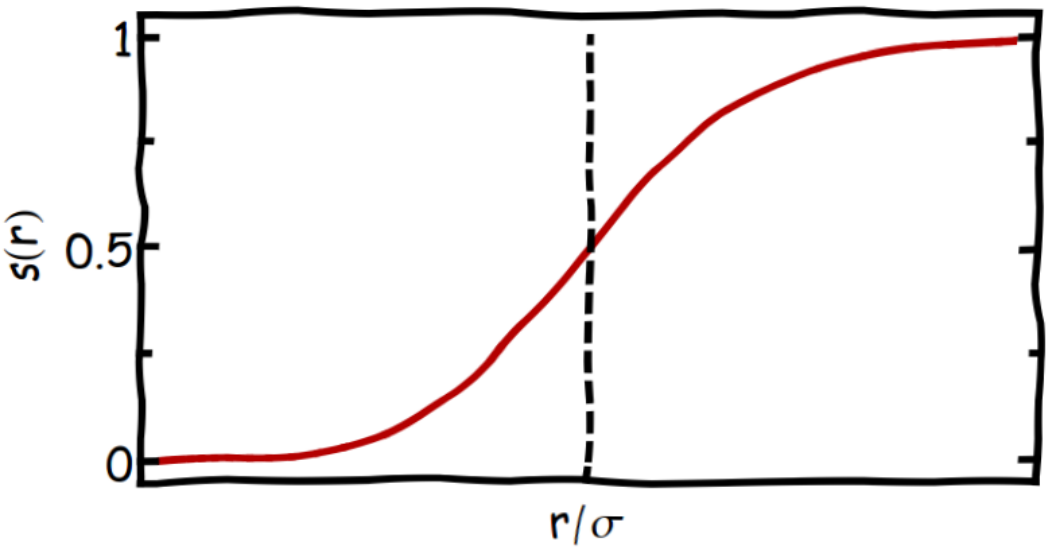
$$\chi^2 = \sum_{i,j=1}^{N} \left[|X_i - X_j| - |x_i - x_j| \right]^2$$

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

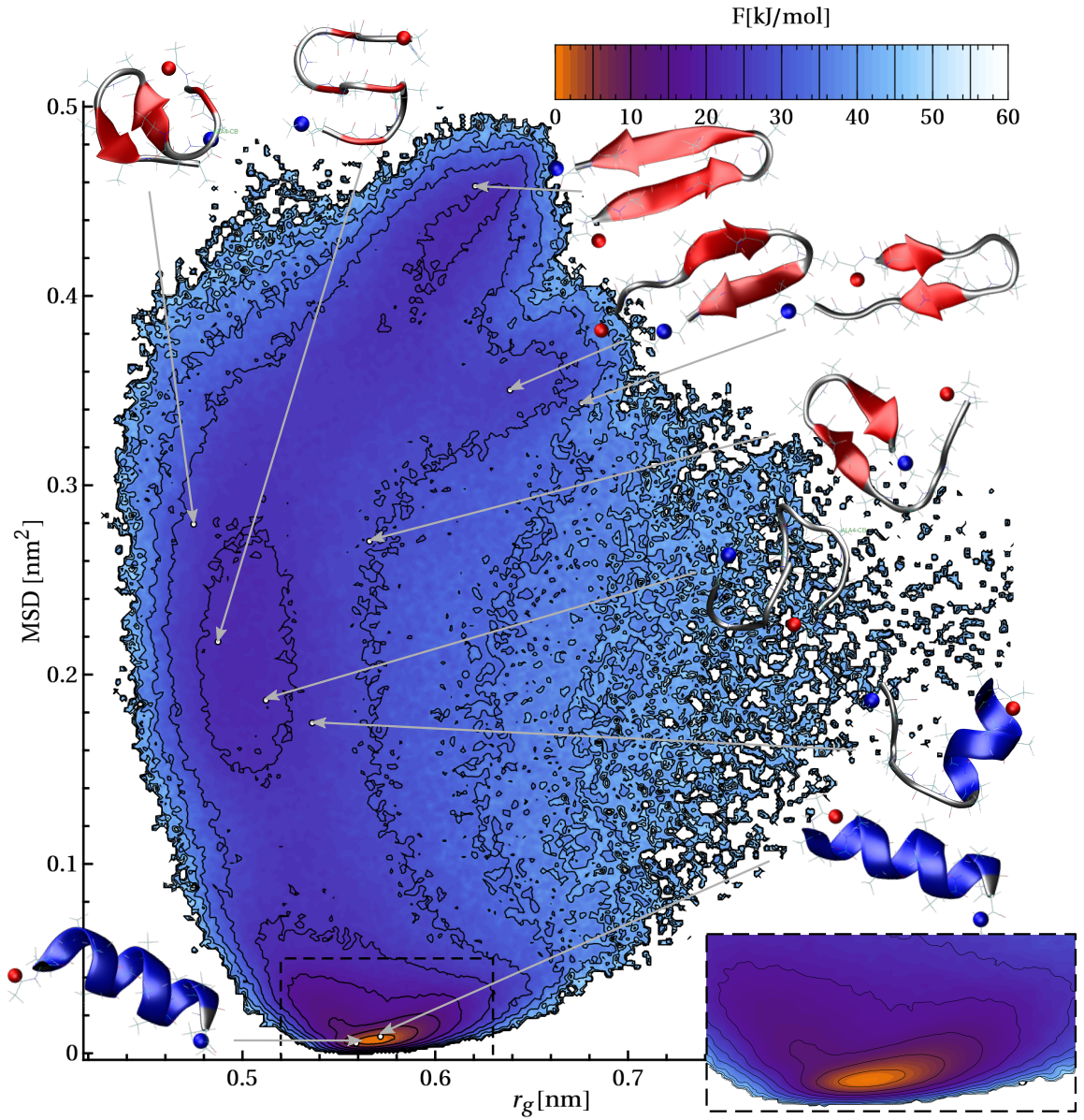
$$\chi^2 = \sum_{i,j=1}^{N} \left[s\left(|X_i - X_j|\right) - s\left(|x_i - x_j|\right) \right]^2$$



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

The folding landscape of ala_{12}

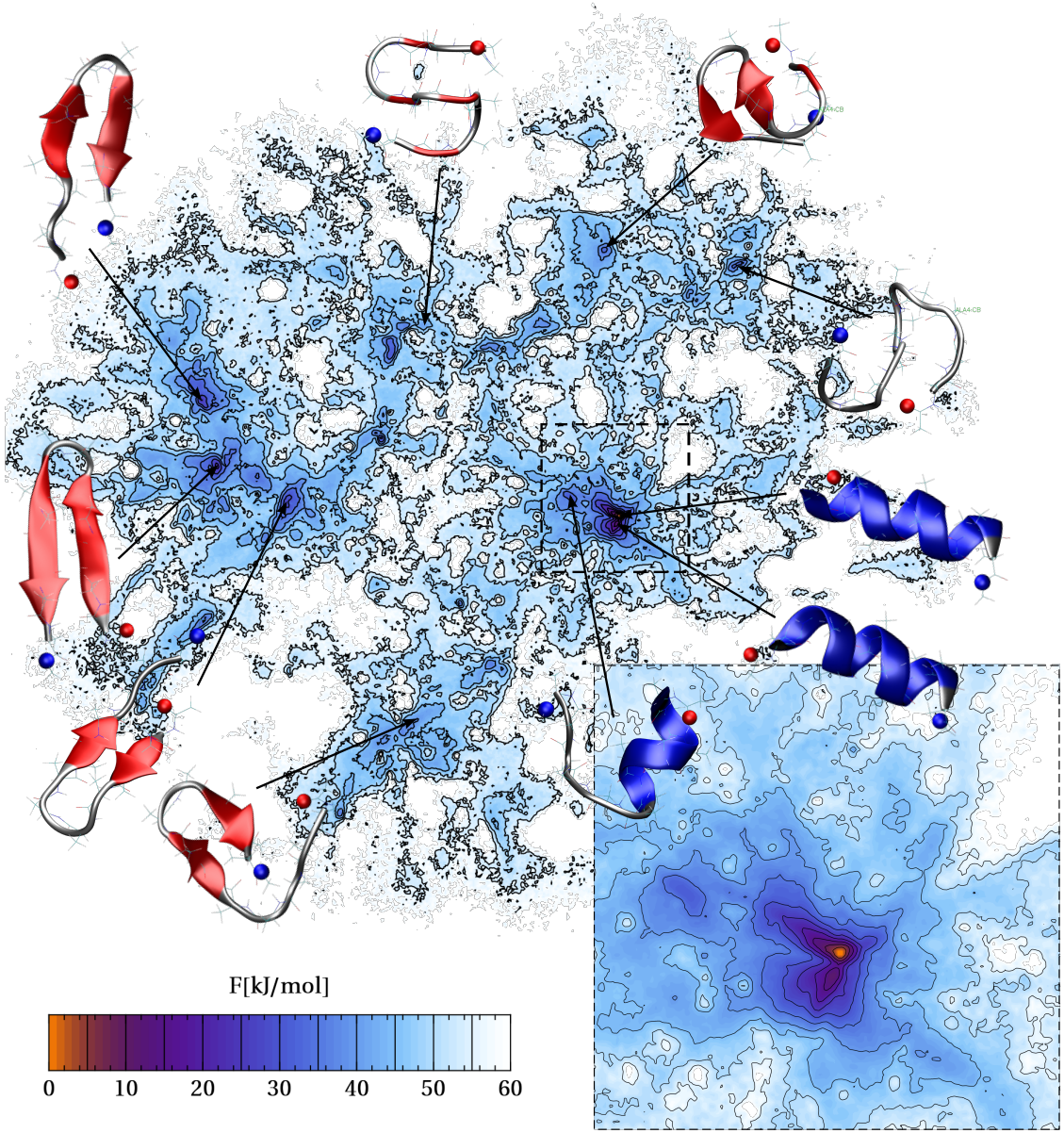
"Conventional" CVs recognize the folded state, but many meta-stable structures overlap with each other



Ceriotti, Tribello, Parrinello, PNAS (2011); Tribello, Ceriotti, Parrinello PNAS (2012)

The folding landscape of ala_{12}

Sketch-map CVs give a very detailed picture, where each meta-stable configuration is clearly singled out[1]

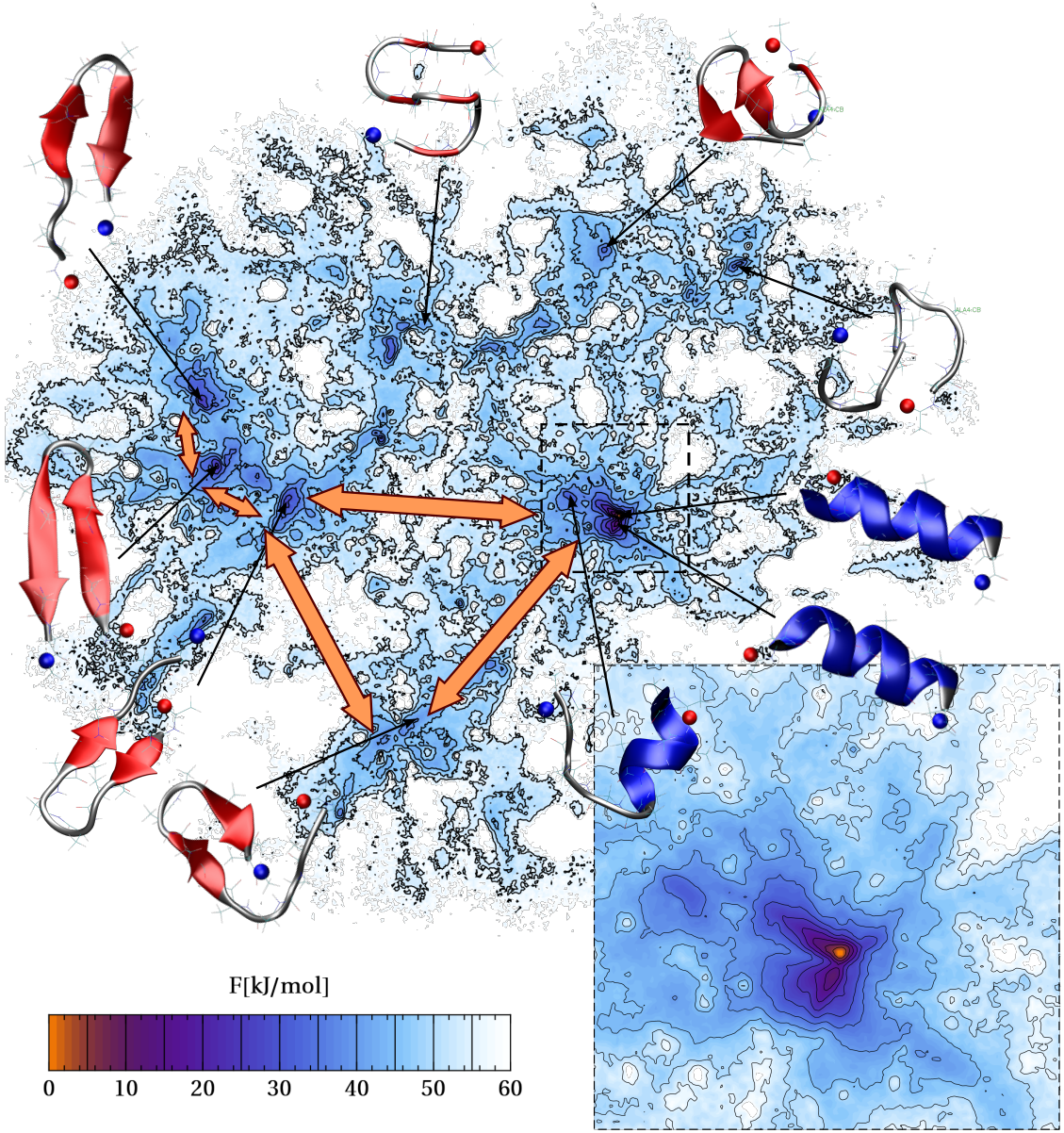


Ceriotti, Tribello, Parrinello, PNAS (2011); Tribello, Ceriotti, Parrinello PNAS (2012)

The folding landscape of ala_{12}

Sketch-map CVs give a very detailed picture, where each meta-stable configuration is clearly singled out[1]

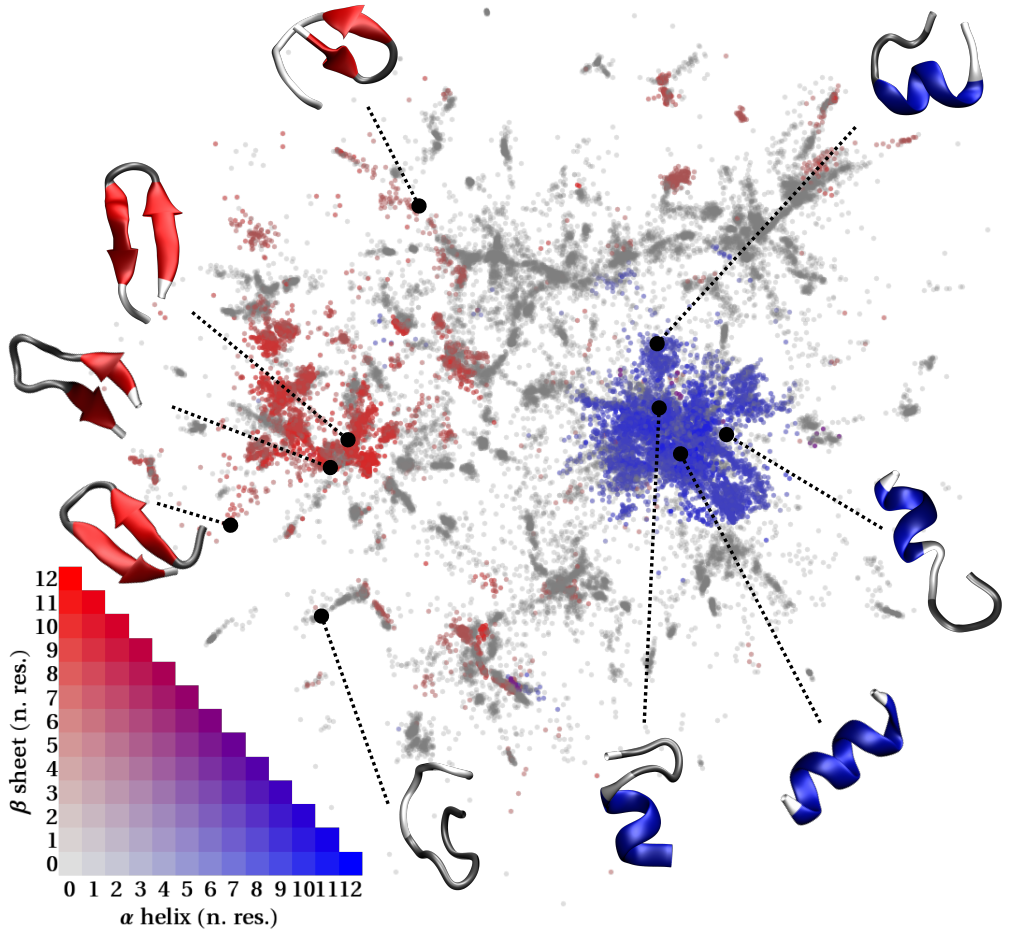
Can be used effectively for accelerated dynamics:
field-overlap metadynamics



Ceriotti, Tribello, Parrinello, PNAS (2011); Tribello, Ceriotti, Parrinello PNAS (2012)

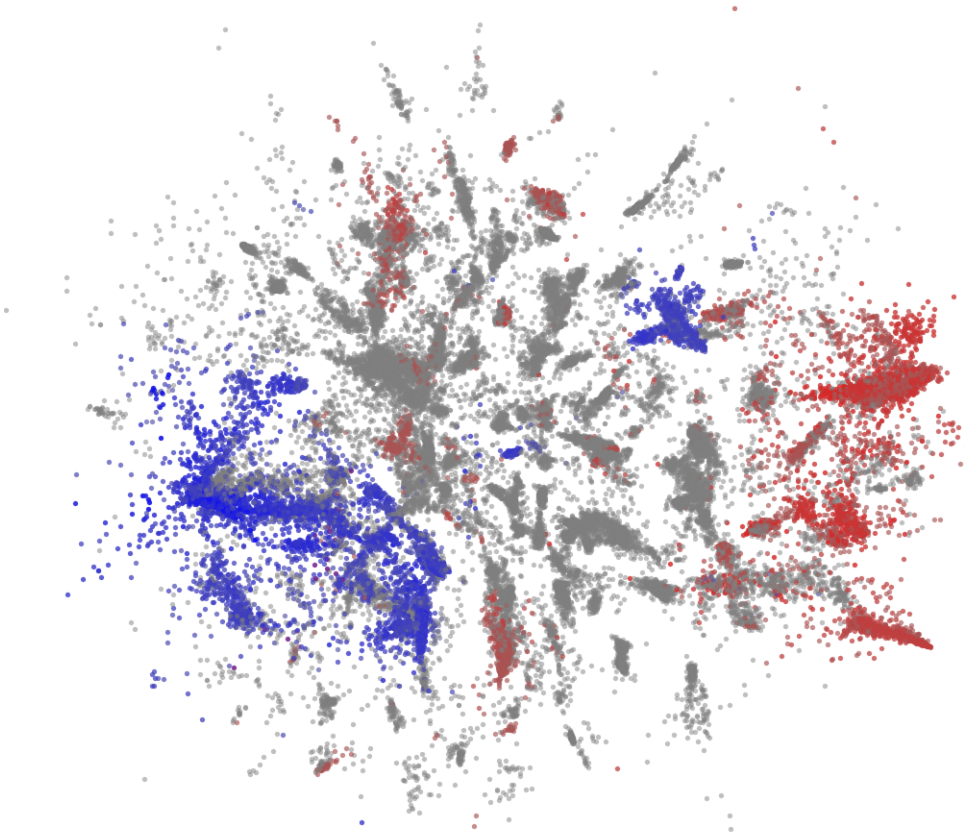
Sketch-map and secondary structure

- Same qualitative features of the hi-D description (basins + network of transitions)
  - Sketch-map CVs correlate nicely with the **secondary structure**
  - Qualitatively similar picture if using contact-maps distance



Sketch-map and secondary structure

- Same qualitative features of the hi-D description (basins + network of transitions)
  - Sketch-map CVs correlate nicely with the **secondary structure**
  - Qualitatively similar picture if using contact-maps distance



Out-of-sample embedding

- In order to use (N)LDR as CV, one needs a way to project an arbitrary point X to low dimension.
  - PCA has a very natural linear projector solution: $x^T = \mathbf{L}X^T$
  - There are specialized solutions for different NLDR methods
- A general approach: "generalized" path coordinates

$$x(X) = \sum_i x_i e^{-|X-X_i|/\lambda} \Big/ \sum_i e^{-|X-X_i|/\lambda}$$

  - Problem: this is a convex embedding so points away from everything will map to the center of the landmark projections

Spiwok, Králová, J. Chem. Phys. 2011

Out-of-sample embedding

- In order to use (N)LDR as CV, one needs a way to project an arbitrary point X to low dimension.
  - PCA has a very natural linear projector solution: $x^T = \mathbf{L}X^T$
  - There are specialized solutions for different NLDR methods
- A general approach: "generalized" path coordinates

$$x(X) = \sum_i x_i e^{-|X-X_i|/\lambda} \Big/ \sum_i e^{-|X-X_i|/\lambda}$$

  - Problem: this is a convex embedding so points away from everything will map to the center of the landmark projections

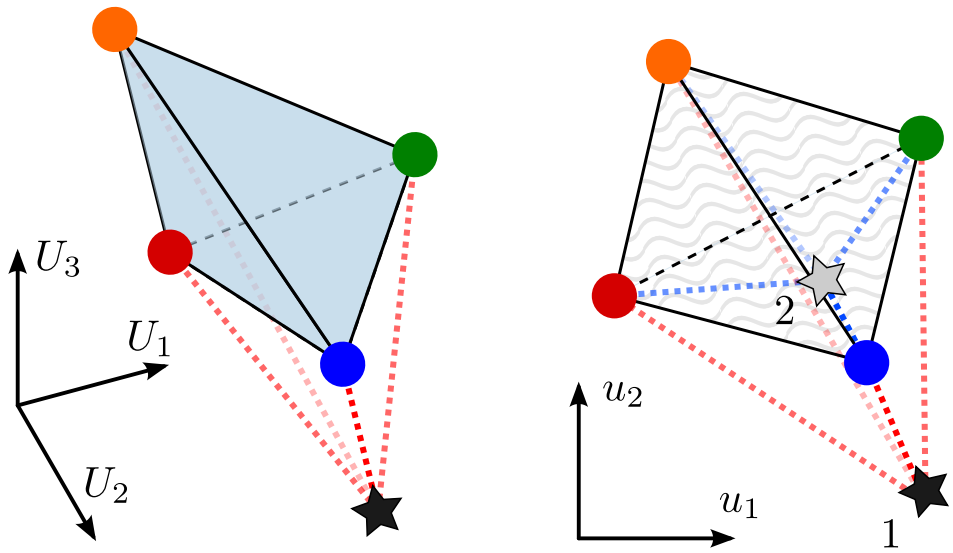
Spiwok, Králová, J. Chem. Phys. 2011

Out-of-sample embedding

- In order to use (N)LDR as CV, one needs a way to project an arbitrary point X to low dimension.
  - PCA has a very natural linear projector solution: $x^T = \mathbf{L}X^T$
  - There are specialized solutions for different NLDR methods
- A general approach: "generalized" path coordinates

$$x(X) = \sum_i x_i e^{-|X-X_i|/\lambda} \Big/ \sum_i e^{-|X-X_i|/\lambda}$$

  - Problem: this is a convex embedding so points away from everything will map to the center of the landmark projections



Spiwok, Králová, J. Chem. Phys. 2011

Sketch-map based collective variables

- Sketch-map only provides projections x_i for the **landmark points** X_i
- One can work out a very natural **out-of-sample embedding** for a new point X
  - Introduce a "stress function" based on a set of landmarks and their projections

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

  - The d-dimensional projection of the point X can be defined as the $\bar{x}$ which minimizes $\chi^2(x, X)$

$$\bar{x}(X) = x : \min \chi^2(x, X)$$

Sketch-map based collective variables

- Sketch-map only provides projections x_i for the **landmark points** X_i
- One can work out a very natural **out-of-sample embedding** for a new point X
  - Introduce a "stress function" based on a set of landmarks and their projections

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

  - The d-dimensional projection of the point X can be defined as the $\bar{x}$ which minimizes $\chi^2(x, X)$

$$\bar{x}(X) = x : \min \chi^2(x, X)$$

Sketch-map based collective variables

- Sketch-map only provides projections x_i for the **landmark points** X_i
- One can work out a very natural **out-of-sample embedding** for a new point X
  - Introduce a "stress function" based on a set of landmarks and their projections

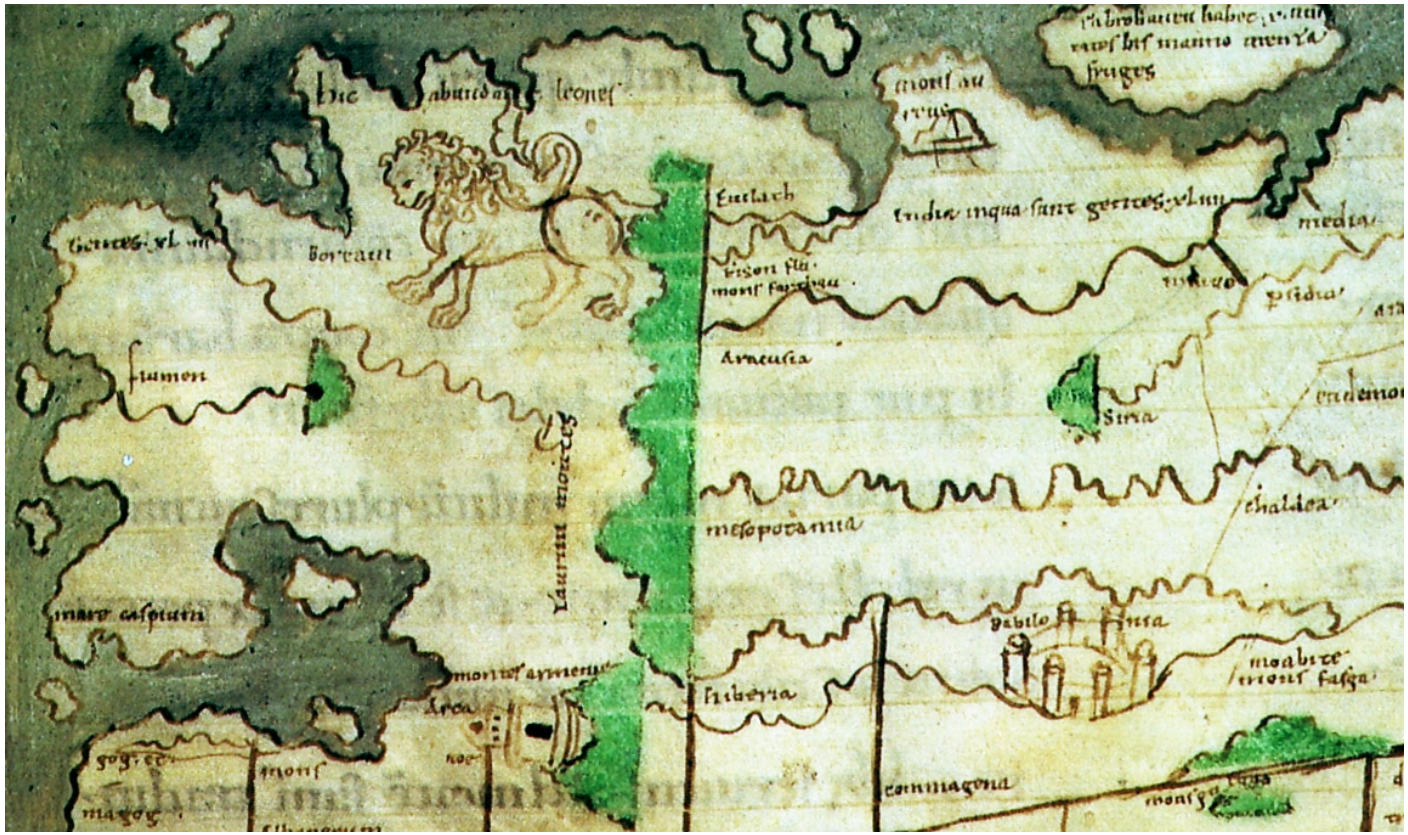
$$\chi^2(x, X) = \sum_{i=1}^{N} \left[s(|X - X_i|) - s(|x - x_i|) \right]^2$$

  - The d-dimensional projection of the point X can be defined as the $\bar{x}$ which minimizes $\chi^2(x, X)$

$$\bar{x}(X) = x : \min \chi^2(x, X)$$

Mapping in no man's land

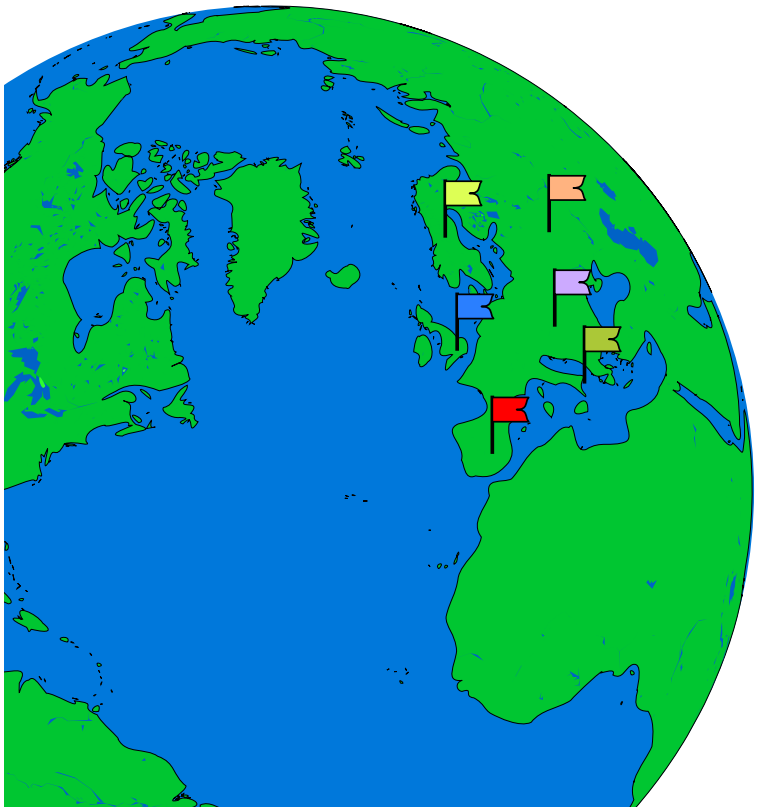
- Sketch map can describe configurations that are in "no man's land", far from any landmark point!
  - We can build a useful map from rough preliminary sampling.
  - We can compare different systems using the same map.

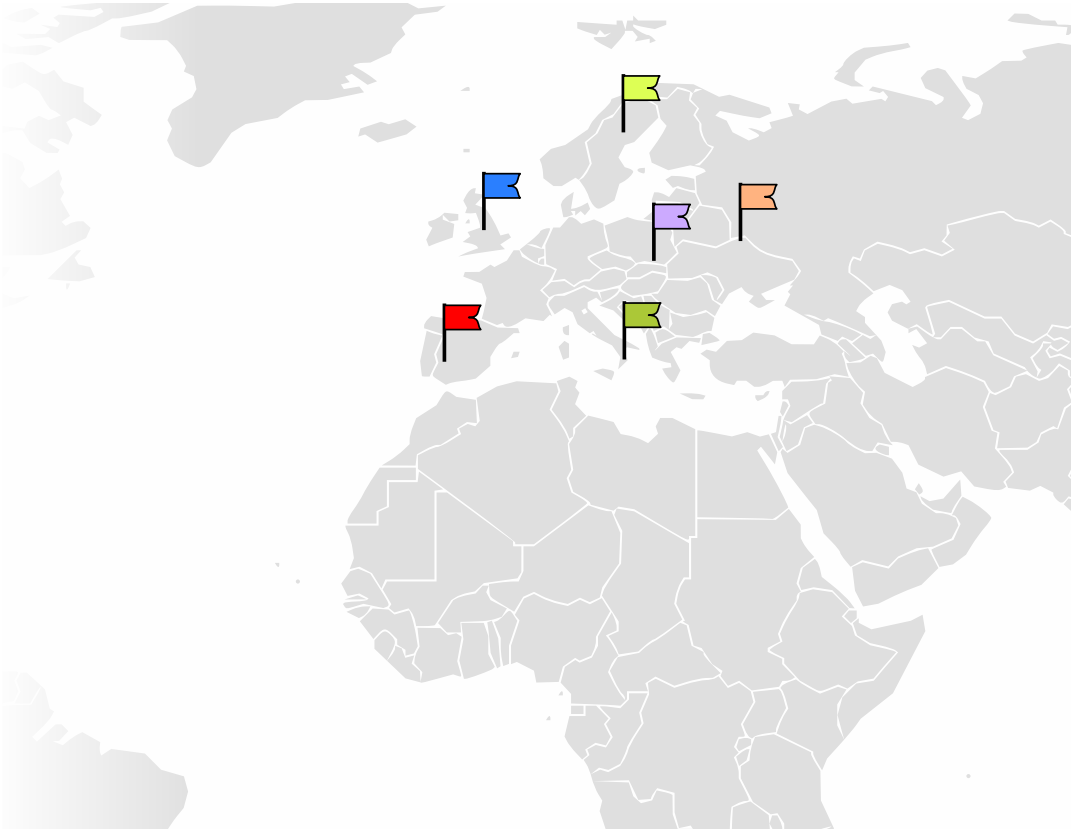


Ceriotti, Tribello, Parrinello, JCTC (2013)

Mapping in no man's land

- Sketch map can describe configurations that are in "no man's land", far from any landmark point!
  - We can build a useful map from rough preliminary sampling.
  - We can compare different systems using the same map.

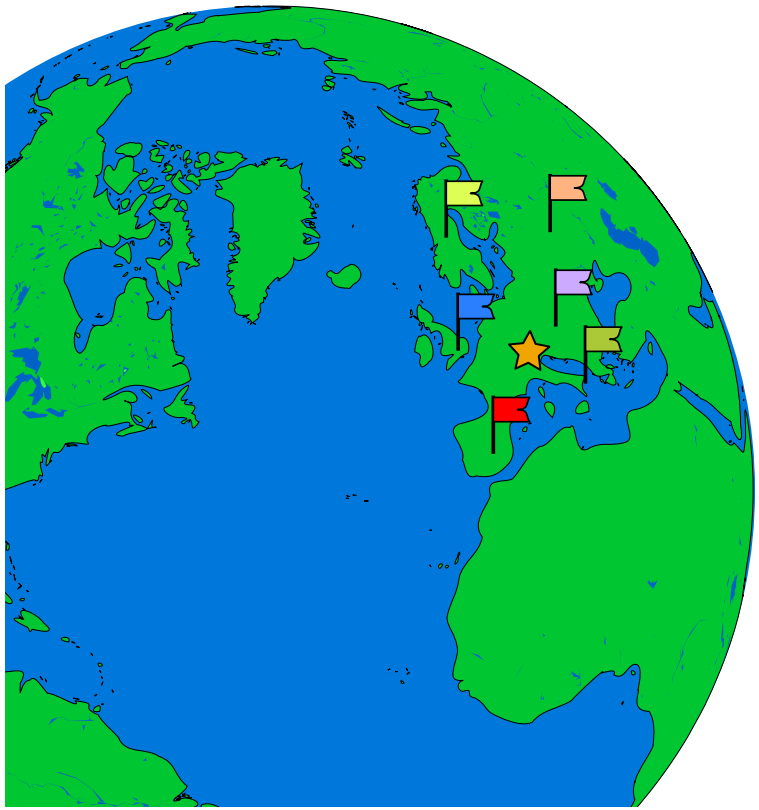


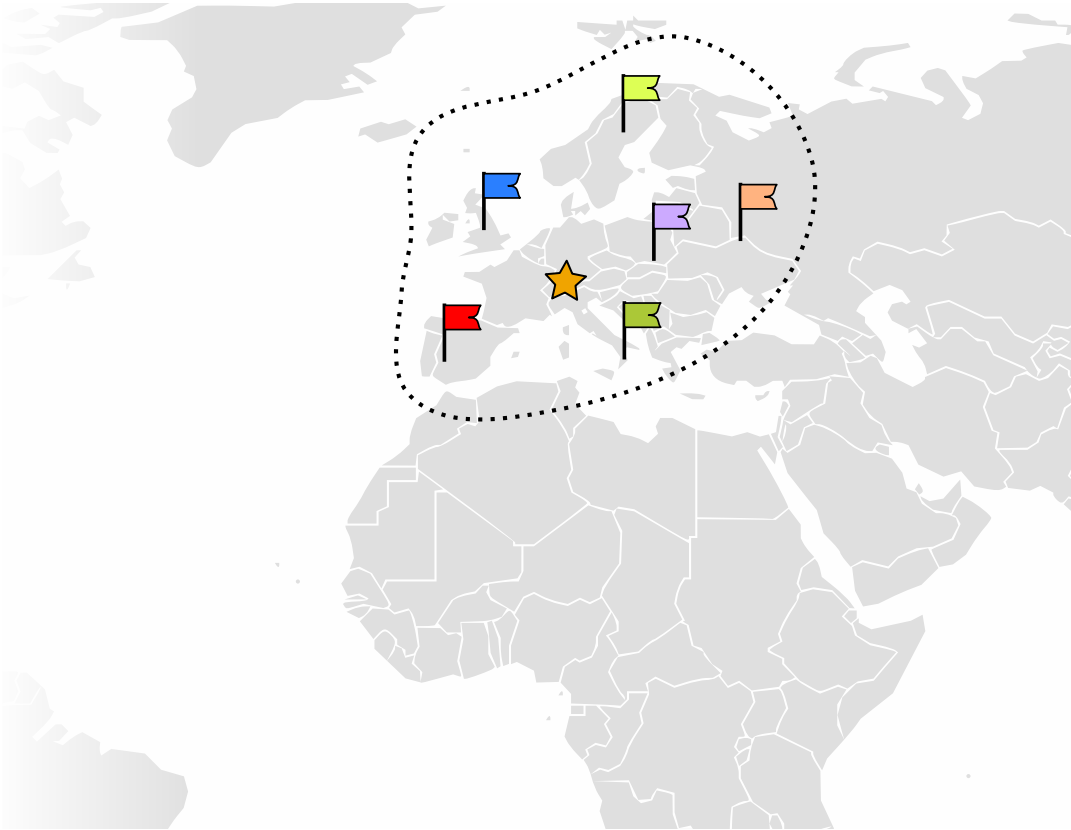


Ceriotti, Tribello, Parrinello, JCTC (2013)

Mapping in no man's land

- Sketch map can describe configurations that are in "no man's land", far from any landmark point!
  - We can build a useful map from rough preliminary sampling.
  - We can compare different systems using the same map.

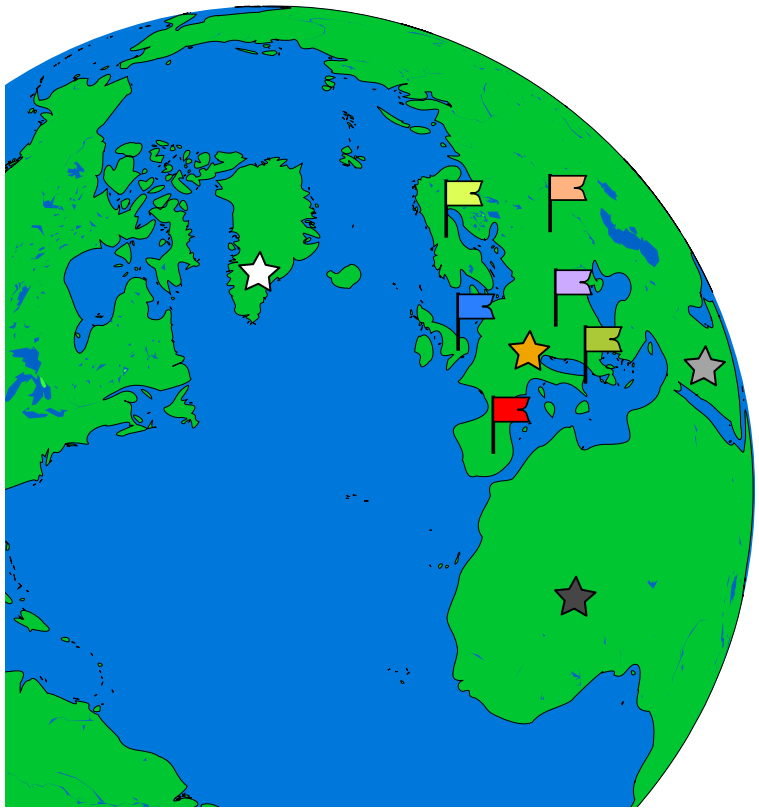


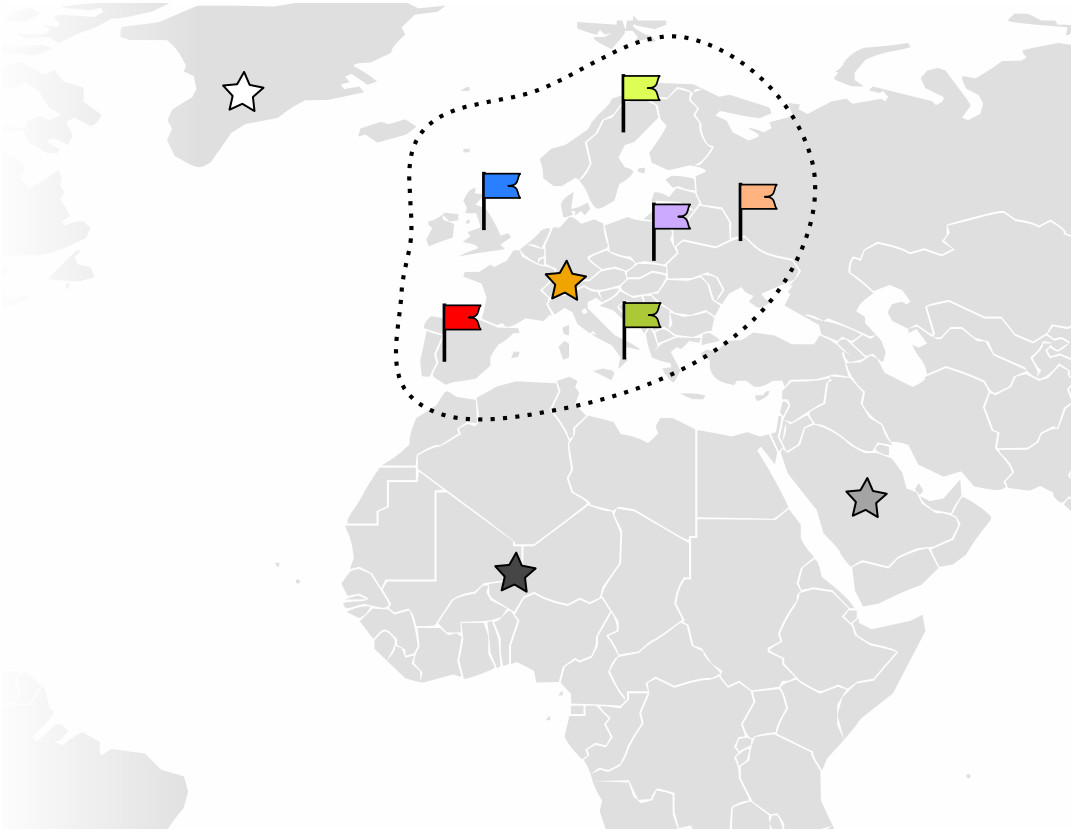


Ceriotti, Tribello, Parrinello, JCTC (2013)

Mapping in no man's land

- Sketch map can describe configurations that are in "no man's land", far from any landmark point!
  - We can build a useful map from rough preliminary sampling.
  - We can compare different systems using the same map.

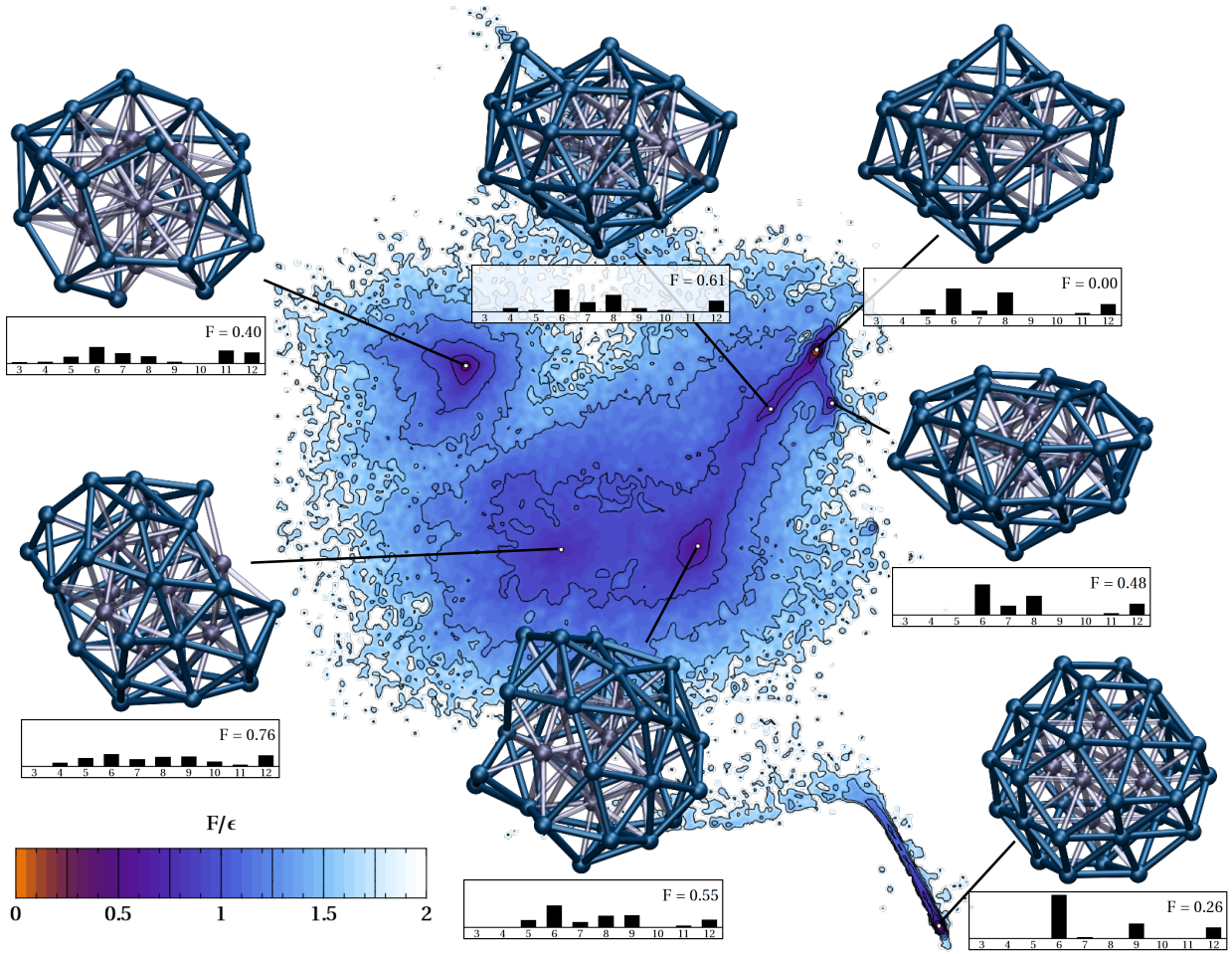




Ceriotti, Tribello, Parrinello, JCTC (2013)

From clusters to defects in the bulk

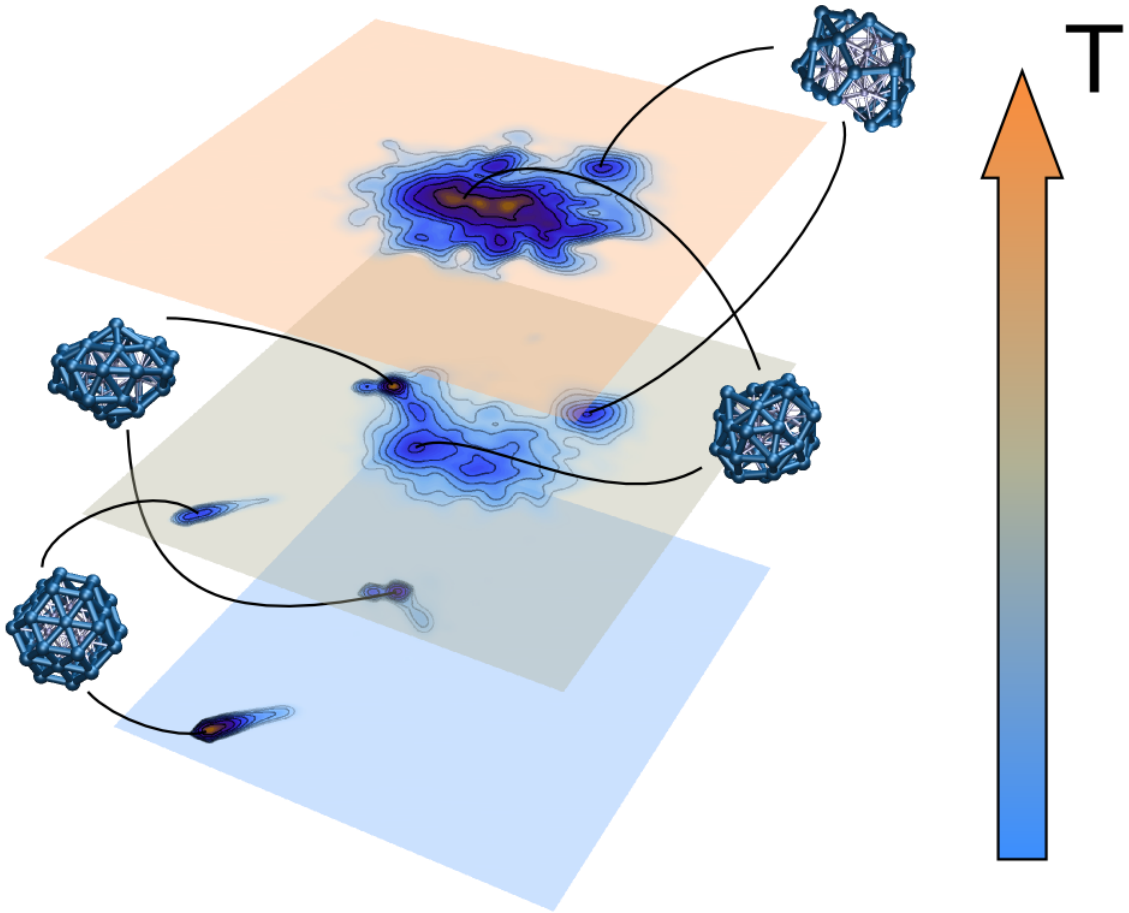
- Start building a map for a Lennard-Jones cluster
- The same map describes the cluster across phase transitions
- ... and can even be used to identify defects in a bulk system!



Ceriotti, Tribello, Parrinello, JCTC (2013)

From clusters to defects in the bulk

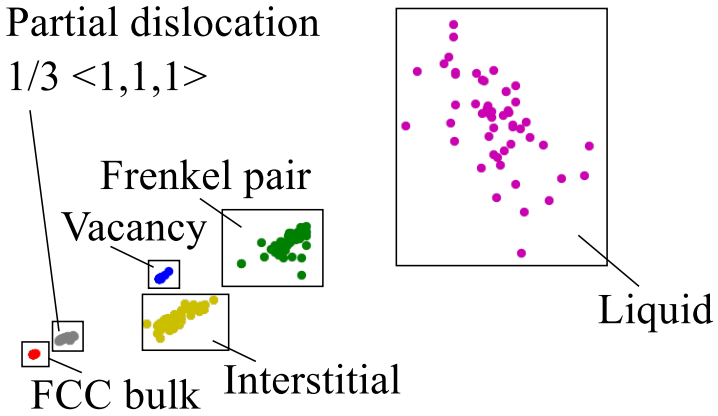
- Start building a map for a Lennard-Jones cluster
- The same map describes the cluster across phase transitions
- ... and can even be used to identify defects in a bulk system!

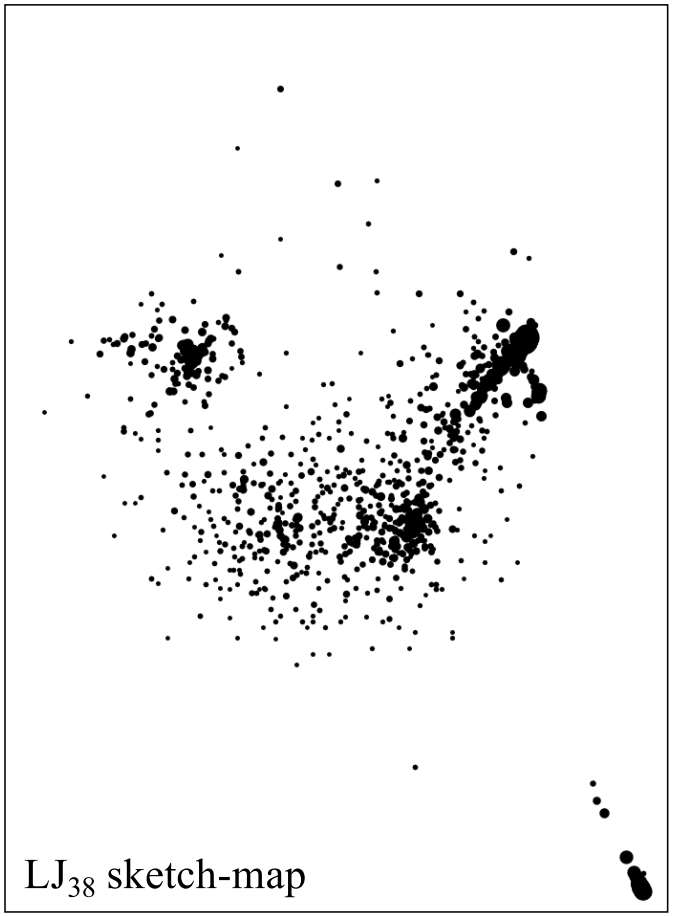


Ceriotti, Tribello, Parrinello, JCTC (2013)

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 be used to identify defects in a bulk system!

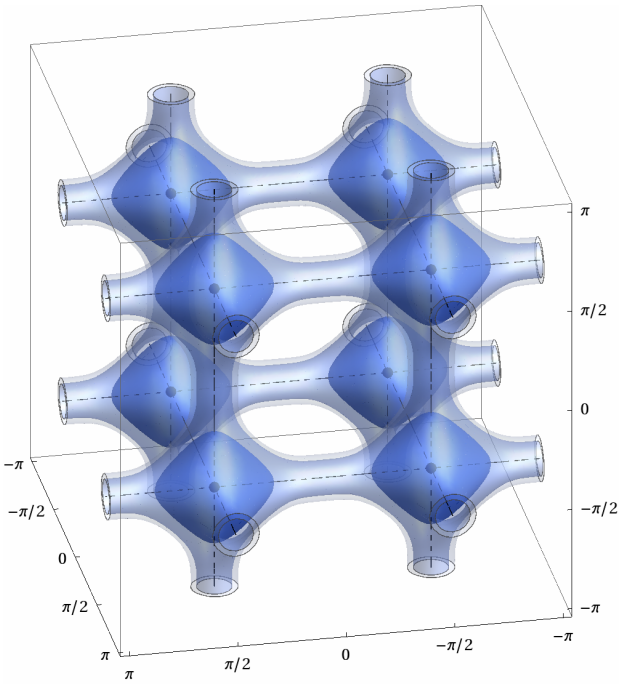


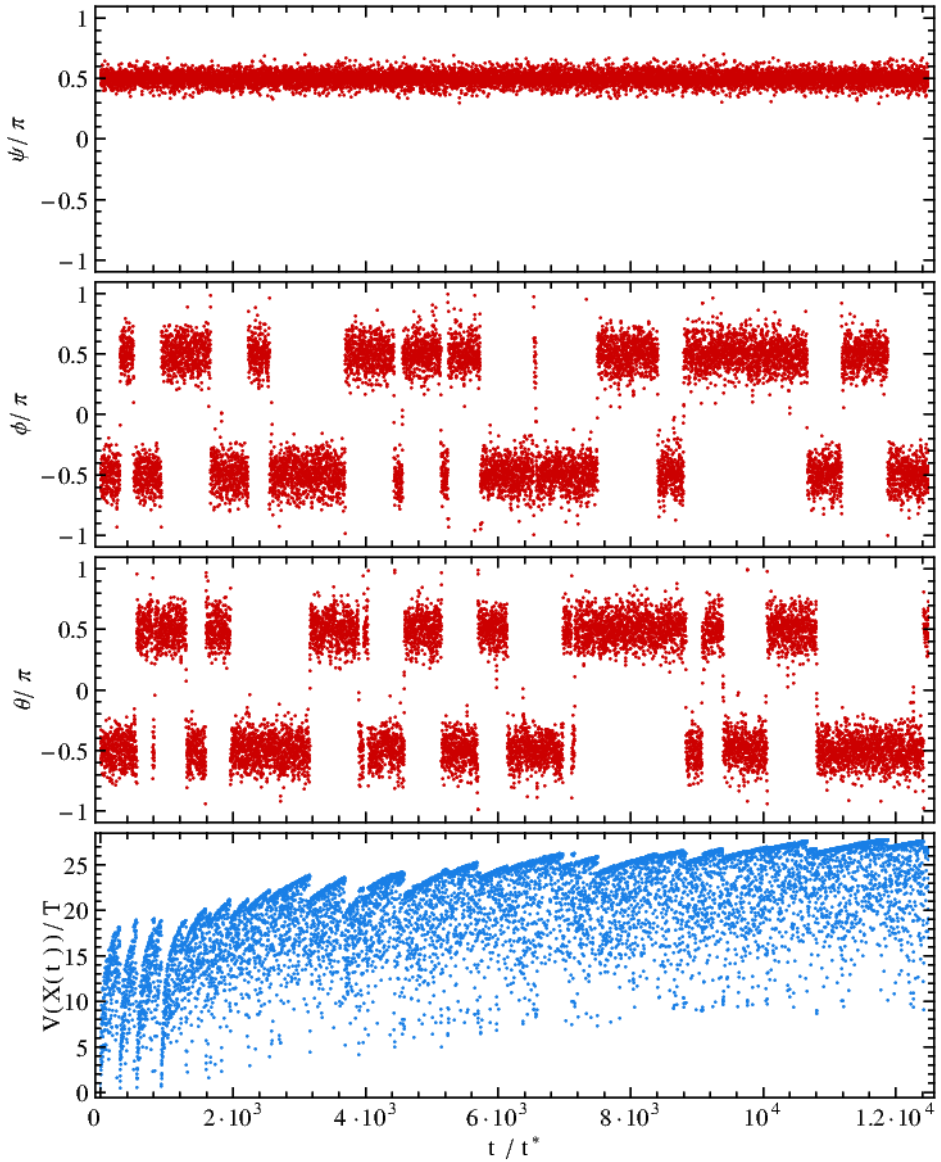


Ceriotti, Tribello, Parrinello, JCTC (2013)

FOM and the 3D example

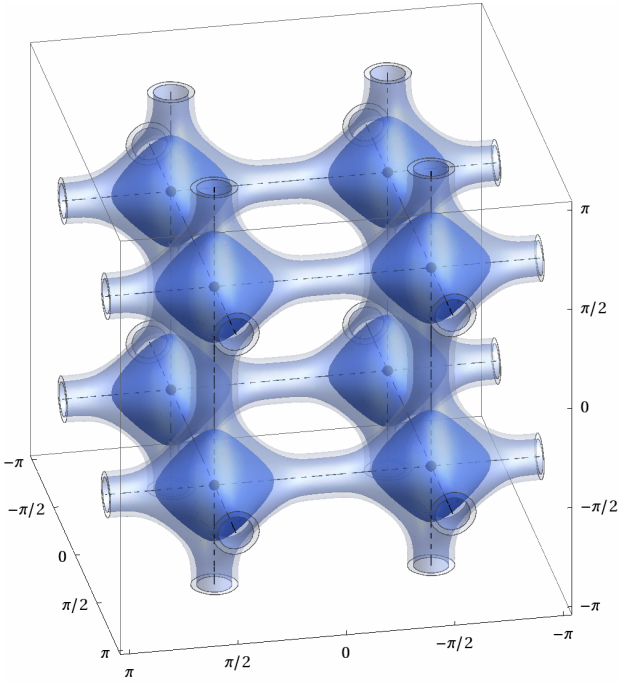
- 2d-metadynamics: the orthogonal degree of freedom is not sampled at all
- **Field-overlap** using (2D!) **sketch-map** coordinates thoroughly samples in all directions

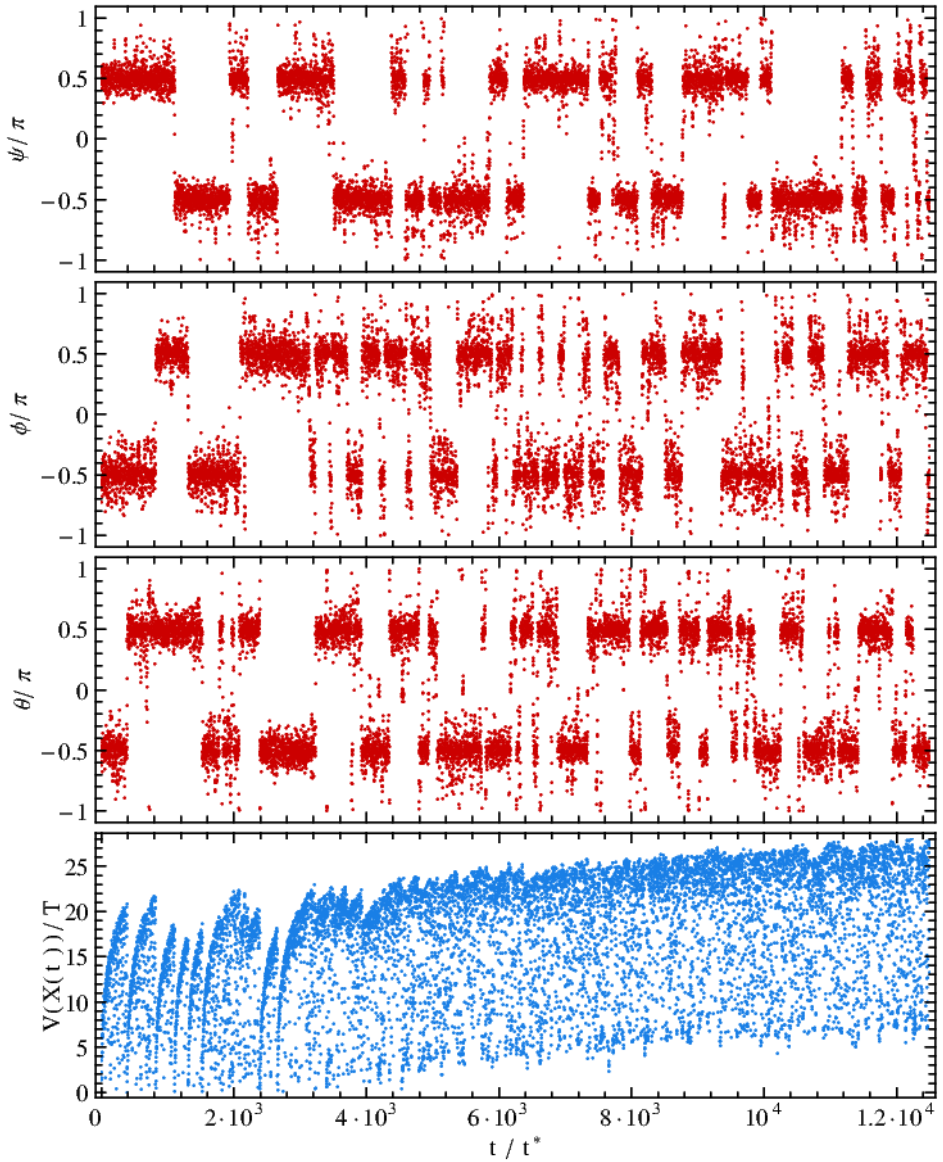




FOM and the 3D example

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





Wrap up and take home

- A common problem in atomistic simulations is how to perform **dimensionality reduction**, and describe a complex problem with few order parameters
- One would like to do this automatically – we can borrow ideas from the machine-learning community
  - PCA/Classical MDS are robust but **linear** techniques – limited effectiveness when the atomistic data do not lie on a hyperplane
  - ISOMAP, works well for "locally flat", densely sampled data. Clearly sensitive to **noise**!
  - Sketch map targets specifically the features of atomistic simulation data.
- Another problem one should keep in mind: out-of-sample embedding. Should be *continuous* and *predictive*

Bibliography

Cox & Cox, "Multidimensional Scaling"
Tenenbaum et al., Science (2000); Roweis, Saul, Science (2000)
Ceriotti, Tribello, Parrinello, PNAS (2011)
http://epfl-cosmo.github.io/sketchmap