Machine Learning of Biomolecular Reaction Coordinates

Simon Brandt, Florian Sittel, Matthias Ernst, and Gerhard Stock

Biomolecular Dynamics, Institute of Physics, Albert Ludwigs University, 79104 Freiburg, Germany

Supporting Information

ABSTRACT: We present a systematic approach to reduce the dimensionality of a complex molecular system. Starting with a data set of molecular coordinates (obtained from experiment or simulation) and an associated set of metastable conformational states (obtained from clustering the data), a supervised machine learning model is trained to assign unknown molecular structures to the set of metastable states. In this way, the model learns to determine the features of the molecular coordinates that are most important to discriminate the states. Using a new algorithm that exploits this feature importance via an iterative exclusion principle, we identify the essential internal coordinates (such as specific interatomic distances or dihedral angles) of the system, which are shown to represent versatile reaction coordinates that account for the dynamics of the slow degrees of freedom and explain the mechanism of the underlying processes. Moreover, these coordinates give rise to a free energy landscape that may reveal previously hidden intermediate states of the system.

As it is neither possible nor desirable to follow the motion of a large molecule along its 3N atomic coordinates, a low-dimensional representation is needed, which in some sense describes the system’s essential dynamics. For example, when we consider biomolecular processes such as protein folding, binding or aggregation, we want to explain the mechanism and the underlying structural rearrangements by using a low-dimensional reaction coordinate. To this end, a number of efficient and systematic strategies of dimensionality reduction have been developed, which aim to identify a few collective variables. Popular methods include principal component analysis, which represents a linear transformation to coordinates that maximizes the variance of the first components, time-lagged independent component analysis, which aims to maximize the time scales of the first components, and full correlation analysis, which tries to minimize the mutual information between components. Assuming a time scale separation between the slow motion of the first few components (representing the "system") and the fast motion of the remaining components (representing the "bath"), the first components of the transformation may serve as a multidimensional reaction coordinate, which can be used to calculate transition rates, to construct a Langevin model, or in various enhanced sampling techniques.

While these systematic dimensionality reduction methods are indispensable to achieve a general preprocessing of high-dimensional data, they often fall short in directly yielding appropriate reaction coordinates when complex conformational rearrangements in macromolecules are considered. This is because these processes may involve small structural changes (thus defying variance-optimizing methods) and exhibit hierarchically coupled processes on various time scales (that hamper methods maximizing time scales). Given as linear combinations of high-dimensional input coordinates, moreover, collective variables often do not necessarily point to the essential internal coordinates, e.g., important specific interatomic distances or dihedral angles. Nonetheless, given some reasonable set of collective variables, a clustering of the data in this reduced representation may be performed, which indicates the metastable conformational states of the system. By calculating the transition probabilities between these states, a Markov model of the dynamics can be constructed that describes protein dynamics in terms of memory-less jumps. Assuming that the metastable conformational states provide a well-defined representation of the low-energy regions of the free energy landscape, they should also contain valuable information on the essential internal coordinates of the system. The question is just how to retrieve this information.

In this work, we propose a systematic approach to identify a low-dimensional set of essential internal coordinates (ECs). The method is based on a combination of state-of-the-art clustering and supervised machine learning techniques (see Figure 1). We start with a data set of 3N-dimensional molecular coordinates, which can be obtained from experiment or molecular dynamics (MD) simulation. Due to inevitable mixing of overall and internal motion, Cartesian coordinates are in general not suited for dimensionality reduction and are therefore converted to internal coordinates such as interatomic distances and angles. In the following, this data set will be referred to as MD coordinates. By using a suitable dimensionality reduction technique like principal component analysis, we next construct a set of collective variables, which facilitate a low-dimensional representation of the dynamics. Employing density-based and dynamical clustering techniques, the metastable conformational states of the system are obtained. In the subsequent pivotal step, a supervised machine
learning model is trained to assign "new" MD structures (i.e., structures not used for training) to the set of metastable states. Hence the model learns to identify the features of the MD coordinates that are most important to discriminate the states. Extending previous works, we propose a new algorithm that exploits this "feature importance" via an iterative exclusion principle (see below) in order to finally obtain the desired ECs. To demonstrate the potential and the performance of the approach, we adopt two well-established model problems, folding of villin headpiece and hinge-bending functional motion of T4 lysozyme. For both systems, long MD trajectories and a detailed characterization of their metastable states are available.

Machine Learning of Essential Internal Coordinates. Given a trajectory of MD coordinates and a set of metastable states defined by these coordinates, it is our goal to train a machine learning model that assigns new MD data to the state they most likely belong to. To this end, the model learns classification rules based on certain features of the MD coordinates, e.g., a certain distance should be bigger than a trained cutoff value, or some angles should lie in a certain region. Provided that these features can be retrieved from the model, we may get direct information about the importance of specific coordinates for discriminating the metastable states.

A machine learning strategy particularly suited for this task is the decision tree family, which represents a group of algorithms with hierarchical decision rules to classify coordinates. Here we use the XGBoost algorithm, which constructs for every given metastable state an ensemble of decision trees that allow one to assign new input data to the metastable state they most likely belong to. A decision tree can be represented by a tree-like graph that consists of nodes associated with classification rules of the MD coordinates, edges (or branches) corresponding to different outcomes of the classifications, and end nodes (or leaves), which assign some score to the corresponding path. Starting at the tree's root, a decision path is found by following the branches of the tree, until the leaf nodes finally yield the probability of a set of coordinates to belong to the state. In this way, coordinates can be classified by simply assigning them to the respective state of highest probability. For training the trees, a loss function is optimized, which grows with both the number of wrongly classified structures and the tree size (i.e., the number of decisions necessary to classify a state). While we obviously want to minimize the former, the latter is important as regularization to prevent overfitting, since the tree size directly reflects the number of fitting parameters that should be small to reduce bias in the model. See the SI for a more detailed description of the XGBoost model and the model parameters employed.

The overall accuracy of the model (i.e., the success rate to correctly assign MD structures to states) can be estimated by dividing the available data into a training set and a test set. While the former is used to train the model, the test set (containing new structures that were not used for training) is used to check whether the model is able to correctly classify the data. The model’s state-dependent accuracy is then the relative number of correctly identified structures for that state.

A major advantage of decision tree algorithms is the simplicity by which the importance of input features (i.e., the MD coordinates) can be determined. (This is in clear contrast to, e.g., deep learning networks, whose typically high numbers of intermediate nodes, weighted sums, and nonlinear transformations render this task extremely challenging.) For every class, the importance of a coordinate is given by the gain of the loss function value that is obtained from decision rules involving the coordinate. Thus, an MD coordinate that is used in decisions that lead to high gains in the loss function—and therefore to a better classification of structures—is clearly more important for characterizing the state than others. This importance criterion, however, has the problem that it is weighted relatively to the importance of all other coordinates. As a consequence, a coordinate might be considered far less important to distinguish one state than another coordinate, simply because the second one was used for the decision prior to the first one, albeit both might equally well describe the state difference.

Thus, to attribute a more meaningful importance index to a given coordinate, it must be evaluated independently from the other coordinates considered as highly important. In extension of previous works employing machine learning techniques for the construction of reaction coordinates, we introduce a novel algorithm that provides an improved estimate of the overall importance of a coordinate. The approach is straightforward: Given a trained model, sort all coordinates by their importance (as given by the XGBoost algorithm; see eq 9 in the SI), remove the coordinate with highest (or lowest) importance from the training set and reiterate the procedure (Figure 1) by retraining the model based on all remaining coordinates. At every iteration, recompute the importance of the remaining coordinates and estimate the accuracy of the trained model.

This approach leads to important insights: When discarding the coordinates of highest importance, we can estimate their overall importance independent from the other coordinates. If the accuracy remains the same, discarded and remaining coordinates more or less describe the same dynamics. Furthermore, we can infer from the states whose accuracy changes upon discarding a coordinate the dynamics described by the coordinate (and thus identify dynamical features that are important to separate states). On the other hand, when discarding the coordinate of least importance first, one can easily filter out all nonessential coordinates from the data set. These are the coordinates that, when discarded, do not lead to significant changes in the accuracy of the model. In this way, we
identify the intrinsic dimensionality of the essential internal coordinates that is necessary to unambiguously discriminate the states. In effect, we have achieved a dimensionality reduction that does not rely on linear or nonlinear combinations of coordinates, but works by simply identifying the most important coordinates of the data.

The Metastable States of HP-35 Are Determined by Six Backbone Dihedral Angles. Villin headpiece (HP-35) is a 35-residues protein fragment that represents a standard model of ultrafast protein folding. It consists of a hydrophobic core with three helices that are connected via two unstructured loops (Figure 2a). As detailed in the SI, we used a 300 μs trajectory of Piana et al.32 performed a principal component analysis on all complete backbone dihedral angle pairs (dPCA+),23 and used density-based clustering23 and the most probable path algorithm23 to obtain 12 metastable states. To visualize the secondary structure of these states, the “Ramacolor” plot in Figure 2b assigns a unique color code to all (ϕ, ψ) configurations of a residue and averages over the color values of these configurations.23 We find that most metastable states exhibit well-defined α-helices, while they clearly differ in the structure of loops and terminal residues.

Given these states, we trained a XGBoost model on the full MD data set of dihedral angles, in order to identify those angles that best describe the state’s structural differences. Figure 2c shows the resulting feature importance plot that quantifies how much a single dihedral angle affects the classification of a given metastable state. The resemblance to the Ramacolor plot is remarkable. The most important dihedral angles identified by XGBoost are clearly the ones also highlighted as state-defining by Ramacolor.

As explained above, the importance of the coordinates obtained from the feature importance plot are weighted relative to the importance of all other coordinates, which may introduce some bias to the selection scheme. To avoid this problem, the new algorithm sorts all coordinates by their importance, removes the coordinate with lowest importance from the training set and reiterates the procedure. Plotted as a function of the number of discarded coordinates, Figure 2d shows the accuracy loss of the XGBoost classifier to identify the 12 metastable states. While the accuracy is found to be remarkably stable for all states when discarding up to 60 coordinates (with the exception of state 11, whose accuracy drops faster than for the other states), it decreases sharply for most states when the remaining six coordinates are removed. These six variables, given by the backbone dihedral angles ϕ3, ϕ22, ψ13, ψ22, ψ10 and ϕ33 (ordered by decreasing importance), represent the desired ECs. Because they are necessary and sufficient to discriminate all metastable states of the system, these angles indeed describe the essential dynamics of the system. Interestingly, the accuracy loss plot in Figure 2d shows that already the single coordinate ϕ3 is sufficient to identify the most populated states 1 and 2 with high accuracy and state 5 at least roughly. To verify that the six ECs are indeed sufficient to discriminate all metastable states of HP-35, we performed a reclustering of the MD data using only these coordinates. Comparing the structures of the resulting 12 metastable states to the structures of the initial metastable states, the Ramacolor plots in Figure S1 reveal that the principal components and the ECs yield virtually the same states.

We are now in a position to discuss to what extent the above found ECs represent suitable reaction coordinates to describe...
the folding of HP-35. Discriminating the metastable states of the system, for one, these coordinates are expected to elucidate the mechanism of the considered process. In fact, we find that $\phi_3$, $\phi_2$, and $\psi_2$ account for the structure of the N-terminus that is known to be crucial to discriminate the folded and the intermediate basins, $\psi_1$ and $\psi_4$ describe the structure of the loop between helices h1 and h2 that is important to discriminate the intermediate and the unfolded basin, and $\phi_4$ reports on C-terminal motion. While the functional relevance of the N- and C-termini as well as the h1–h2 loop was recognized in previous work, these studies did not reveal the ECs. This is because collective variables such as the first principal components do not point to specific internal coordinates, but only yield linear combinations of numerous coordinates (Figure S3). We note that our machine learning-based construction of ECs can also be applied to the case of nonlinear dimensionality reduction approaches such as diffusion maps, whose collective variables are not straightforwardly described in terms of the input coordinates.

As explained in the introduction, further common requirements on reaction coordinates include covering a significant part of the overall variance of the data (which is optimized in principal component analysis, here dPCA+33) and to achieve a time scale separation with respect to the fast bath fluctuations (which is the aim of time-lagged independent component analysis, TICA15). To test whether the ECs of HP-35 satisfy the former requirement, we calculated the variance of the six most important coordinates in the various approaches. We find that dPCA+ covers ~45% of the total variance, TICA ~ 13%, and the ECs ~ 25%. Since these numbers are quite different while the structural information given by the various coordinates is rather similar (Figure S3), we conclude that a selection of ECs based purely on cumulative variance may be misleading. As a measure of the time scales described by the most important coordinates, Figure 2 compares the autocorrelation functions obtained for (e) the ECs and (f) the first principal components of dPCA+. Remarkably, we find that the decay times of the ECs are in part significantly longer than the results for the principal components, and are overall also longer than the TICA decay times (Figure S2).

Finally, we wish to highlight that the time scales of the ECs can be directly related to a corresponding physical process. For example, $\psi_1$ is associated with the relative positions of helices h1 and h2 and, as such, describes the overall folding/unfolding transition, which corresponds to the slowest process of the system.23 On the other hand, $\phi_2$ and $\psi_2$ account for relatively fast fluctuations of the N-terminal, which effect a twisting motion described by $\phi_1$ that leads to the destabilization of the protein.39 The latter is associated with relatively slow rearrangements of $\psi_1$ and $\phi_3$, and finally leads to unfolding involving again $\psi_1$. Highlighting the hierarchical coupling of fast and slow motions, ECs provide a direct view of the dynamics, which is very difficult to achieve from collective coordinates that exclusively maximize variance or time scales.

Detection of a Hidden Locking Mechanism of T4L. As an example of complex functional motion, we consider T4 lysozyme (T4L), for which a 50 μs MD trajectory at 300 K is available (see SI for details). T4L is a 164-residue enzyme whose interaction with the substrate involves a prominent hinge-bending motion of its two domains, which resembles the opening and closing of the mouth of a “Pac-Man” (Figure 3a). Employing some coordinate that reflects this motion (e.g., the distance $d_{31,142}$ between residues Thr21 and Thr142), the associated free energy profile exhibits two minima corresponding to the open and closed states, which are separated by an energy barrier $\Delta G \lesssim 2k_B T$ (Figure 3b). However, when modeling the transition rate by the standard expression $k = 1/\tau = k_0 e^{-\Delta G/k_B T}$, it becomes obvious that the small energy barrier cannot account for the long (~10 μs) observed transition time $\tau$ of T4L. In fact, a recent targeted MD simulation study revealed that trying to directly enforce the open-closed transition does not recover the two-state behavior of T4L, meaning that the opening distance does not represent a suitable reaction coordinate.34 The comprehensive simulations and analyses rather showed that this transition is triggered by a locking mechanism, by which the side chain of Phe4 changes from a solvent-exposed to a hydrophobically buried state.34 This mechanism, which, e.g., can be represented by the distance $d_{4,60}$ between residues Phe4 and Lys60, leads to a barrier of ~6 $k_B T$ between the locked and unlocked states (Figure 3c) and results in a four-state description of T4L.34 It should be stressed that systematic dimensionality reduction approaches such as various types of PCA did not indicate this mechanism. In this respect, the hinge-bending motion of T4L represents a prime example of complex functional dynamics whose underlying
mechanism is not straightforwardly described by some obvious reaction coordinate.

To challenge the machine learning algorithm, here we intentionally adopt the naive view of T4L as a two-state problem and study whether the approach can provide additional insight. To this end, we characterize the system using distance coordinates of all residue–residue distances that are considered contacts (i.e., have a distance of less than 4.5 Å), define open and closed states of T4L based on the opening distance $d_{142}$ and train an XGBoost model. In particular, at every iteration we remove the most important coordinate from the data set to force XGBoost to reclassify the importance of alternative descriptors. Table 1 shows the discarded distances of the first few iterations of the algorithm. As expected, the first few most important distances connect the two sides of the binding pocket (see Figure 3a) and therefore directly monitor the open–closed transition of T4L. Surprisingly, though, the next three distances all involve residue Phe4, which is located in the hinge of the protein. The last column lists the feature importance obtained by XGBoost in the first iteration, which ranks the relevance of distances involving Phe4 much lower.

<table>
<thead>
<tr>
<th>iteration</th>
<th>discarded distance</th>
<th>feature importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Thr21-Gln141</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Glu22-Thr142</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>Glu22-Gln141</td>
<td>10</td>
</tr>
<tr>
<td>4</td>
<td>Phe4-Lys60</td>
<td>50</td>
</tr>
<tr>
<td>5</td>
<td>Phe4-Ala63</td>
<td>42</td>
</tr>
<tr>
<td>6</td>
<td>Phe4-Ile29</td>
<td>56</td>
</tr>
<tr>
<td>7</td>
<td>Arg8-Phe67</td>
<td>65</td>
</tr>
</tbody>
</table>

"While the first three iterations of the algorithm yield distances that directly measure this transition, the subsequent iterations reveal distances involving residue Phe4, which is located in the hinge of the protein. The last column lists the feature importance obtained by XGBoost in the first iteration, which ranks the relevance of distances involving Phe4 much lower.

two XGBoost models assuming a four-state system, were trained using either only backbone dihedral angles or only interresidue distances. While the latter plot clearly shows that a few distances are required to distinguish the intermediate states 2 and 3, the former readily reveals that dihedral angles are not well suited to describe the conformational dynamics of T4L, because they describe the intermediate states only with very low accuracy. The same conclusion was reached in ref 34 after a detailed discussion of PCA and clustering results using both types of coordinates.

We have described a supervised machine learning approach to reduce the dimensionality of a complex molecular system. The method relies on three components: choice of appropriate input coordinates that allow one to discriminate metastable conformational states (see Figure 4), accurate definition of these states as obtained by density based and dynamical clustering (see Figure 2), and the XGBoost-based algorithm using an iterative scheme to determine the most (or least) important coordinates (see Table 1). The final outcome of the procedure is an accuracy loss plot that reveals the essential internal coordinates (ECs) such as specific interatomic distances or dihedral angles. We note that commonly used collective variables in general cannot point out important internal coordinates.

Using two well-established models of protein dynamics, HP-35 and T4L, we have shown that the ECs provide versatile reaction coordinates, which account for the dynamics of the slow degrees of freedom and directly describe the mechanism of the process. In the case of HP-35, for example, the ECs have been found to exhibit slow time scales (overall slower than available collective variables), which could be directly related to some functional rearrangement of the system. Finally we wish to point out that our machine learning-based approach is able to...
to discover new features that are not obvious from the initially provided collective variables or metastable states. In the case of T4L, for example, the free energy landscape along the ECs reveals additional metastable states that account for a new reaction mechanism, which has only recently been detected independently with considerable effort.  

Due to the general structure of the machine learning ansatz, our approach allows for a variety of extensions, including use of dynamical information in the training of the model, modeling of nonequilibrium data, and semiautomatic detection of reaction coordinates by screening a catalogue of typical order parameters used in protein dynamics analysis.

**ASSOCIATED CONTENT**

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jpclett.8b00759. MD and XGBoost methods, details of dPCA+ and TICA of HP-35, and comparison of metastable states obtained from principal components and from XGBoost (PDF)

**AUTHOR INFORMATION**

Corresponding Author

*E-mail: stock@physik.uni-freiburg.de.

ORCID

Gerhard Stock: 0000-0002-3302-3044

**Contributor Information**

*Contributed equally.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

We thank Benjamin Lickert for instructive and helpful discussions and D. E. Shaw Research for sharing their trajectories of HP-35. This work has been supported by the Deutsche Forschungsgemeinschaft (Sto 247/11) and the bwUniCluster and bwForCluster computing initiatives of the State of Baden-Württemberg.

**REFERENCES**


