EXTRACTING MOLECULAR INSIGHTS FROM CONFORMATIONAL ENSEMBLES USING MACHINE LEARNING

Oliver Fleetwood[1], Marina A. Kasimova[1], Annie M. Westerlund[1] and Lucie Delemotte[1,*]

[1] Department of Applied Physics, Science for Life Laboratory, KTH Royal Institute of Technology, Stockholm, Sweden.

*correspondence to: lucie.delemotte@scilifelab.se

ABSTRACT

Biomolecular simulations are intrinsically high dimensional and generate noisy datasets of ever-increasing size. Extracting important features in the data is crucial for understanding the biophysical properties of molecular processes, but remains a big challenge. Machine learning (ML) provides powerful dimensionality reduction tools. However, such methods are often criticized to resemble black boxes with limited human-interpretable insight.

We use methods from supervised and unsupervised ML to efficiently create interpretable maps of important features from molecular simulations. We benchmark the performance of several methods including neural networks, random forests and principal component analysis, using a toy model with properties reminiscent of macromolecular behavior. We then analyze three diverse biological processes: conformational changes within the soluble protein calmodulin, ligand binding to a G protein-coupled receptor and activation of an ion channel voltage-sensor domain, unravelling features critical for signal transduction, ligand binding and voltage sensing. This work demonstrates the usefulness of ML in understanding biomolecular states and demystifying complex simulations.

KEYWORDS

Molecular Dynamics, Machine learning, CV selection, Feature extraction, GPCR, β2 Adrenergic Receptor, Calmodulin, Voltage Sensor Domain, Neural Network, Random Forest, Autoencoder, Restricted Boltzmann Machine, PCA, MLP.

INTRODUCTION

Molecular dynamics simulations of biological systems provide a unique atomistic insight into many important biological processes such as a protein's conformational change between functional states, the folding of a soluble protein or the effect of ligand binding to a receptor. These systems can be extremely high dimensional with pairwise interactions between tens to hundreds of thousands of atoms at every snapshot in time. As system sizes, as well as the reachable timescales of simulations, have increased significantly over the last decade, a conventional simulation can now generate terabytes of raw data that needs to be condensed for human interpretation. Data-driven methods reduce the risk for researchers to overlook important properties in a simulation, or even misinterpret the computational experiment and introduce human bias in the analysis.

A system of N particles has 3N spatial degrees of freedom. Fortunately, restraints in the system due to, for example, the force field and steric hindrance, restrict biomolecules to only adopt a subset of all possible atomic rearrangements. Thus, the system typically moves on a manifold of much lower dimensionality than its actual number of degrees of freedom. Finding this manifold is not easy, and even when it is found it can still be difficult to *interpret* it. Researchers often seek *reaction coordinates* or *Collective Variables* (CVs) to describe important features of this manifold. An optimal set of CVs should help answering a specific question. For example, if the aim is to enhance sampling of transitions from one functional state to another, we typically seek the CVs which best describe the slowest motion of the system. In other situations, we might seek answers to biophysically relevant questions by chasing subtle differences on the molecular level between slightly perturbed states. Examples include establishing the molecular signatures of a protein when bound to different ligands, considering different protonation states or the effect of a point mutation. Regardless of the biological problem at hand, analysis typically involves processing a large set of high dimensional data in search of important features. This type of dimensionality reduction problem may be addressed with Machine Learning (ML) methods.

The increasing amounts of data, as well as limited time and resources for processing and analysis, is not only a challenge in the field of biomolecular simulations. ML methods have gained enormous interest in recent years and are now applied in a wide range of research areas within biology, medicine and health care (Akay and Hess, 2019; Ching et al., 2018; Zitnik et al., 2019) such as genomics (Zou et al., 2019), network biology (Camacho et al., 2018), drug discovery (Chen et al.,

2018) and medical imaging (Jang and Cho, 2019; Lundervold and Lundervold, 2019). In molecular simulations, such methods have eminently been used to enhance sampling by identifying CVs in a data-driven manner (Ahmad et al., 2019; Fleetwood et al., 2019; Hashemian et al., 2013; Jung et al., 2019; Mittal and Shukla, 2018; Ribeiro and Tiwary, 2018; Ribeiro et al., 2018; Sidky and Whitmer, 2018; Sittel and Stock, 2018; Sittel et al., 2014; Sultan et al., 2014; Trapl et al., 2019; Wehmeyer and Noé, 2018; Zhang and Chen, 2018) and as an exploratory tool to generate new protein conformations (Degiacomi, 2019). However, many tools borrowed from ML, most notably nonlinear models such as Neural Networks (NN), are criticized for their black-box resemblance which obstructs human interpretable insights (Akay and Hess, 2019; Camacho et al., 2018; Ching et al., 2018; Zitnik et al., 2019). Therefore, turning the "black box" into a "white box" is an active area of research (Montavon et al., 2018; Olah et al., 2018).

In this study we have demonstrated how to learn ensemble properties from molecular simulations and provide easily interpretable metrics of important features with prominent ML methods of varying complexity, including principal component analysis (PCA), random forests (RFs) and three types of neural networks: autoencoders, restricted Boltzmann machines (RBM) and multilayer perceptrons (MLP). For different types of datasets, we considered supervised methods, where each simulation frame belongs to a known class, as well as unsupervised methods, where information is extracted from unlabelled data. In order to show how the methods perform under various circumstances, we first evaluated them on a toy model designed to mimic real macromolecular behavior, where functional states were defined by state-dependent random displacement of atoms. This approach enabled us to understand and illustrate the different methods' properties and shortcomings in a quantitative manner. In particular, we find that all methods perform equally well on the toy model's simple setup, in which optimal inverse interatomic distances are used as an input. However, while supervised and nonlinear methods are still able to identify important features for a more complex input such as Cartesian coordinates, unsupervised and linear methods fail to do so.

We further applied our protocol for identifying key features of three biologically interesting processes that involve systems sampled using extensive conventional MD or enhanced sampling techniques: the conformational rearrangements occurring within the C-terminal domain of the Ca^{2+}-bound soluble protein calmodulin, the effect of ligand binding to a G protein-coupled receptor

and the response of an ion channel voltage-sensor domain to a change in the transmembrane potential. In all three cases, the ML methods pinpointed molecular features that are known to be important for the function of these proteins. In general, our results demonstrate the ability of ML methods to reveal valuable insights into biomolecular systems. We anticipate that this straightforward, yet powerful approach, can become useful for many researchers when demystifying complex simulations.

RESULTS AND DISCUSSION

BENCHMARKING WITH A TOY MODEL PROVIDES QUANTITATIVE MEASURES OF PERFORMANCE OF THE ML METHODS AND GUIDANCE FOR ANALYSIS OF REAL SYSTEMS.

Biomolecular systems consist of a large number of atoms or molecules. At each functional state, these fluctuate around their equilibrium positions due to thermal motion. As time progresses, a typical biomolecule transitions between a number of metastable *states*, either spontaneously, or because of changes in its environment. For real systems, identifying important features of such metastable states is desirable, and ML offers data-driven methods for this purpose.

To evaluate the performance of a certain method, one needs to truly know how the states are defined. We have designed and used a toy model that mimics a biomolecular system and for which we are able to control the parameters describing states (Figure 1.a). In short, the atoms were first randomly placed within a box at their equilibrium positions. To define new states, certain atoms were displaced from their original positions (Figure 1.a) in either a linear (Figure 1.c,d) or nonlinear manner (Figure S2)(STAR Methods and SI Methods). Artificial simulation frames were then generated by adding noise to the atomic coordinates. The accuracy of a method at *finding all* displaced (relevant) atoms was evaluated using a metric based on the mean-squared error between the estimated and the true feature importance profiles. In addition to this, we measured the ability of a method to *ignore irrelevant* atoms (i.e. minimizing false positives) (STAR methods). The way these scoring methods describe known distributions is shown in Fig. S1. In this way, the toy model allowed us to test and quantify how successfully a method identified true positives. We completed

this analysis by evaluating the performance of the different methods when considering cartesian or internal coordinates as input features, in a full or reduced set (Figure S3) and different noise levels (Figure S4).

Given a trajectory consisting of many atoms and frames, with none or very little prior knowledge of the system, a widespread first approach in the biomolecular simulation community would be to apply principal component analysis (PCA) (Fig. 2a) to all atomic coordinates and evaluate the ability of the first principal components (PCs) to identify the features that contributed most to the variance in the data. Fig. 1c and S2a illustrate some major drawbacks of this unsupervised approach. Since PCA intrinsically performs a linear mapping of input features to a low dimensional representation (Tipping and Bishop, 1999), the performance of the method is highly dependent on the choice of input coordinates (Sittel et al., 2014). For example, PCA failed to detect the important features of the toy system and even performed worse than random guessing when using raw, unaligned Cartesian coordinates as input (Fig. 1c, S2a). To remedy this, we performed PCA on internal coordinates (Fig. 1d, Fig. S2b), namely inverse interatomic distances, either using the full set of all pairwise distances or a reduced set (Fig. S3) only connecting an atom to the four atoms before and after it in the sequence (STAR Methods). The interatomic displacement that was performed to construct the states is easily described by such input coordinates, and PCA was able to successfully identify all the correct atoms for this simpler setup of the toy model (Figure 1d).

In addition to PCA, two alternative unsupervised learning methods were evaluated: a Bernoulli restricted Boltzmann Machine (RBM) (Fig. 2b) (Smolensky, 1986) and an autoencoder (AE) (Fig. 2c) (Hinton and Salakhutdinov, 2006). For every set of input features and labels, important features for a single classification decision were derived using layerwise relevance propagation (LRP) (Fig. 2f and STAR methods) (Montavon et al., 2017). RBM and AE were both poor at distinguishing important features from irrelevant ones using Cartesian coordinates (Fig. 1c). They were able, however, to perform with higher accuracy for linear as well as nonlinear toy model displacements using internal coordinates (Fig. 1d and S2b), although PCA showed better performance in this regard. While AE is suggested to be a promising unsupervised method for datasets that require nonlinear transformation, our results show that it performs either only marginally better (Cartesian coordinates) or even worse than PCA (full set of internal coordinates). This outcome indicates that AE is difficult to train on datasets of large dimensionality such as biomolecular simulations,

and that to use this method successfully one should reduce the number of input features to a minimum.

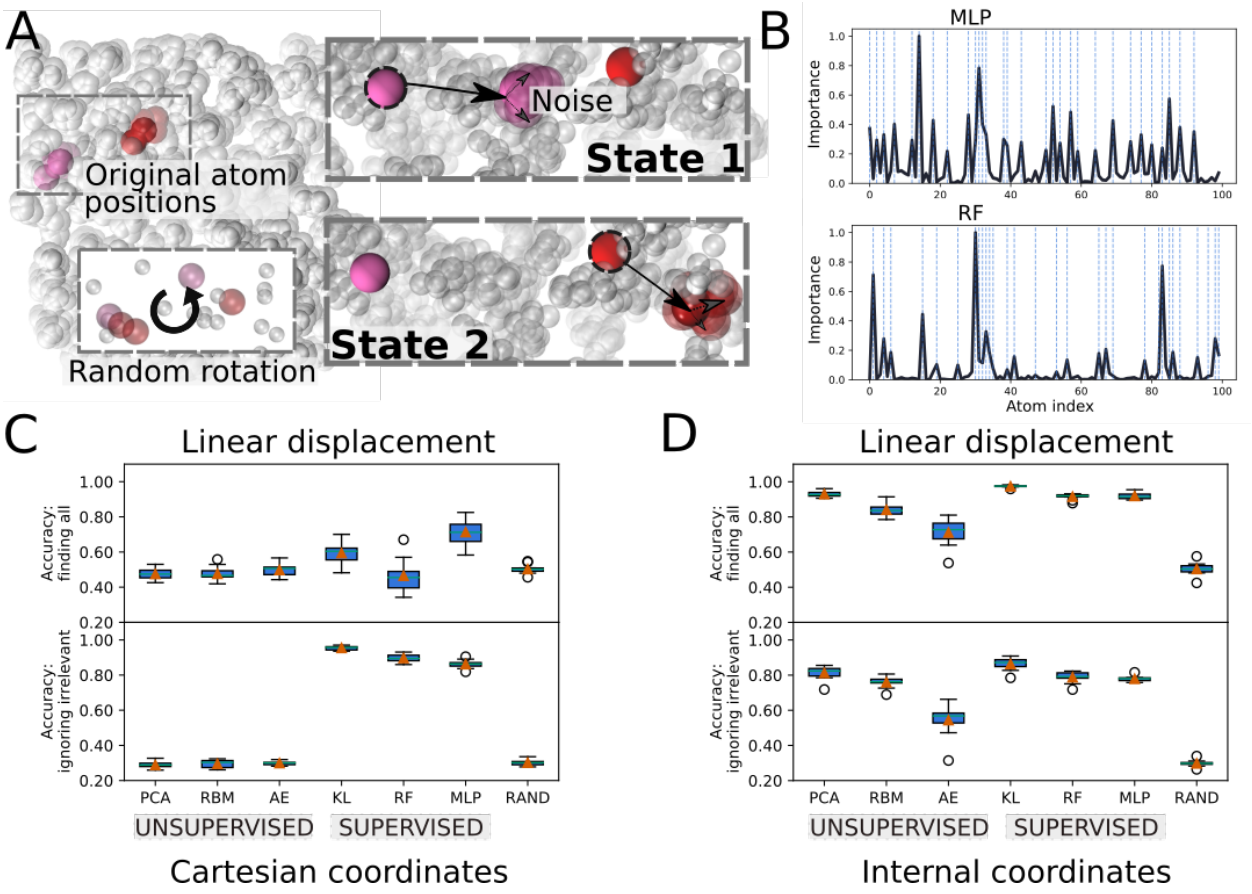


Figure 1. (a) Toy model used to benchmark different ML approaches to extract important features from simulation data. Atomic coordinates are randomly generated and a subset of the atoms, unique to every state, is displaced linearly from their initial positions. Artificial simulation frames are generated by adding noise to all atoms' positions. Since only the relative and not the absolute positions of atoms are of significance in real biological systems, the system is rotated randomly around the origin. (b) Importance per atom for single instances of the toy model. The index of the displaced atoms are highlighted as dashed vertical lines, coinciding with all peaks in importance in the case of a Multilayer Perceptron (MLP), and some of them for a Random Forest (RF). (c), (d) Boxplots of the performance of the different methods using either Cartesian coordinates (c) or inverse interatomic distances as input features (d) sampled over different instances of the toy model with linear displacement and 10% noise level. A high accuracy at finding all atoms signifies that every displaced atom has been identified as important and that other atoms have low importance. A high accuracy at ignoring irrelevant atoms signifies that only displaced atoms, although not necessarily all of them,

have been marked important (STAR methods and Fig. S1). The best performing set of hyperparameters found after benchmarking every method (Fig. S5-S10) have been used. RAND indicates random guessing.

The main limitation of PCA and other unsupervised learning techniques is that when the conformational states are known and each frame is labeled according to the state it belongs to, they fall short in utilizing this information. Supervised learning techniques, on the other hand, are designed for this purpose. In principle, supervised ML should be able to identify even the most subtle differences between states. We have applied three methods of different character and complexity (Fig. 1c,d and S2) that take known class labels into account. As a simple first step, we computed the Kullback Leibler (KL) (Fig. 2d) divergence (Bishop, 2006) between one state and all other states for every individual input feature. A high KL divergence indicates that the feature is good at differentiating between states and hence important. Next, we trained a Random Forest (RF) (Fig. 2e) classifier (Ho, 1995) to discriminate between states and computed feature importance using mean decrease impurity (Breiman et al., 2017). The third method we evaluated was a Multi-layer Perceptron (MLP) classifier (Fig. 2f), a prototypical neural network (Bishop, 2006), in combination with LRP (Montavon et al., 2017).

Computing the KL divergence between atomic coordinates does not transform the features in any way, and therefore failed to identify important atoms from the raw Cartesian coordinates (Fig. 1c), especially when the noise in the toy model was strong (Fig. S4a). On the other hand, it did not inaccurately score irrelevant atoms. RF also performed well for only one of the two accuracy metrics, specifically to ignore irrelevant atoms and map out some important ones, albeit not all of them (Fig. 1b,c and S2a). In many situations, this is an appealing characteristic; RF and KL provide quick overviews with a close-to-minimal set of features to describe the system and give highly interpretable importance profiles. This, however, comes at the risk of ignoring other significant features in the data. MLP, which has the ability to approximate nonlinear classification functions due to its multi-layer architecture and use of activation functions, successfully identified the majority of the important features from unaligned Cartesian coordinates (Fig. 1b,c and S2a). Its performance, compared to the other methods, was found to be superior when states were defined by nonlinear instead of linear displacement between atoms (Fig. S2a) or after increasing the strength of the noise in the toy model (Fig. S4). On the other hand, it was not as good at ignoring

irrelevant features, as it inaccurately scores some non-displaced atoms as relatively important by mapping out overfitted nonlinearities.

All three supervised methods were excellent at identifying important features using the optimal distance-based features (Fig. 1d, Fig. S2b, Fig. S3). RF and KL often outperformed MLP, although all three supervised methods were close to perfect at identifying the important features for this setup. Note that to find good discriminators between states, one should use a classifier such as MLP or RF instead of KL divergence. However, seeing that a high KL divergence correlates with a proper choice of input coordinates suggests that it can be used as a *separation score* to describe how well a few selected features are at separating states.

For every method, we chose a set of hyperparameters to tune using cross-validation (Fig S5-S10) and included the results for the parameter set with best average performance in Fig. 1 and S2-S4. For PCA, we varied the number of components to consider in the aggregated importance per feature (Fig. S5) and found that it is better to include all the PCs with corresponding eigenvalues (i.e. variance covered) within the same order of magnitude than to only include the first PC**.** For all three neural network based methods (RBM, AE and MLP) we evaluated different shapes of the hidden layers (Fig. S6-S8) and found that increasing the number of hidden nodes or layers does not necessarily increase performance in general. For MLP and AE we also considered the regularization parameter and for the RBM we evaluated the performance for different values of the learning rate. Neither had a significant impact on the overall performance in this case, with the exception of having a too high learning rate equal to one for the RBM. We also computed the KL divergence with different discretization parameters (Fig. S9) and trained RF classifiers as either a set of binary or multiclass classifiers with varying the number of estimators (Fig. S10). A proper discretization size for KL seemed to be about 1% of the range of values covered in the data. Increasing the number of estimators in RF classifiers tended to slightly increase the accuracy. Otherwise, both of these methods were moderately sensitive to the choice of hyperparameters, although a set of binary RF classifiers were, unsurprisingly, considerately better at identifying important atoms of a specific state compared to a multiclass classifier, which was only able to select the important atoms for the entire ensemble. In general, we found that a small change in a method's hyperparameter value for this application typically resulted in a small change in the

importance profile, and that even a significantly different set of hyperparameter values performed with comparable accuracy.

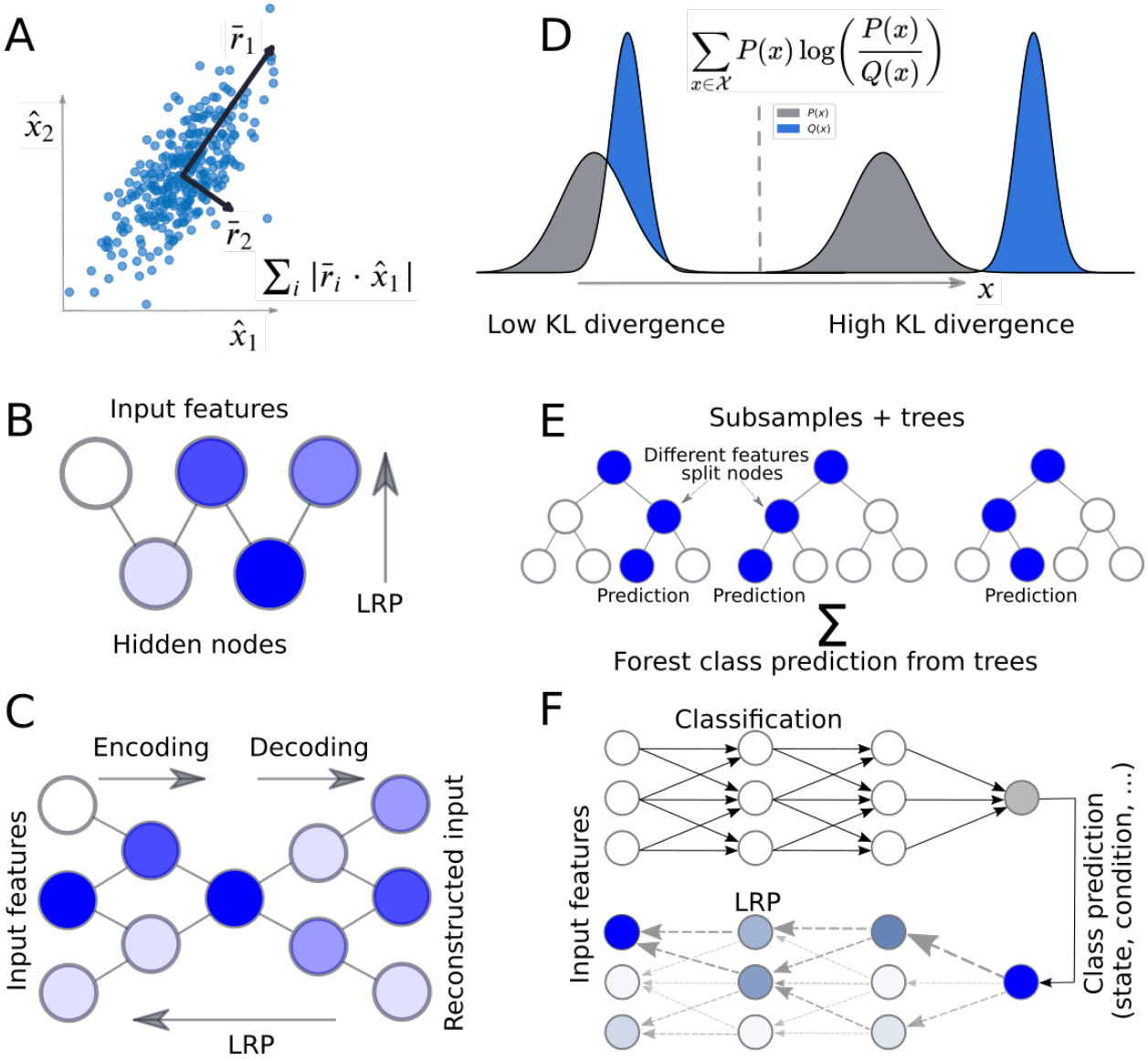


Figure 2. Machine learning models used in this study. (a) Principal Component Analysis (PCA) converts the input features to an orthogonal set of linearly uncorrelated variables called principal components (PCs) through an orthogonal transformation. The first component covers the largest variance possible. The feature importance is taken to be the components' coefficients of a feature times the variance covered by (i.e. the eigenvalues) the components. (b) Restricted Boltzmann Machine (RBM): a generative stochastic neural network trained to maximize the likelihood of the data using a graphical model with a layer of hidden nodes connected to the input nodes. Important features can be derived using Layer Relevance Propagation (LRP), an algorithm originally developed

for image classification problems (Montavon et al., 2017). (c) Autoencoder (AE): a generative neural network trained to reconstruct the features through a set of hidden layers of lower dimensionality. Important features can be derived using LRP. (d) Kullback–Leibler (KL) divergence (also called relative entropy): this metric computes the difference between two distributions $P(x)$ and $Q(x)$ along every individual feature. We compute the KL divergence between one class and all other classes as an indication of the importance of a feature. (e) Random Forest (RF) classifier. A prediction is taken as an ensemble average over many decision trees. Average relevance per feature is computed as the mean decrease impurity (Breiman et al., 2017) . (f) A Multilayer Perceptron (MLP) is a feedforward artificial neural network with fully connected layers. Important features can be derived using LRP.

We note that it is unlikely that the best performing set of parameters in this study is optimal for all biological problems. Moreover, although important atoms in the toy model were successfully identified with distance based features, it is not sufficient to assume inter-residue distances to form an adequate set of input features for all biological problems. For example, it is possible that backbone dihedral angles are better indicators of certain types of conformational changes or that local root-mean-square deviation (RMSD) to experimentally resolved structures better describes how computationally sampled states agree with the experiments. To have a toolbox of methods which are not as sensitive to the choice of input coordinates in combination with metrics such as a separation score makes it straightforward to compare different sets of features and methods hyperparameters, and avoid making poor choices. Based on the learnings from the toy model benchmarks, we outlined our protocol in the form of a checklist (see at the end of the paper) that we followed as we transitioned into biological applications.

SUPERVISED AND UNSUPERVISED LEARNING PINPOINT CONSERVED RESIDUES AND FREQUENT TARGET-PROTEIN BINDERS OF CALMODULIN.

Calmodulin is a ubiquitous messenger protein activated by calcium ions (Kuboniwa et al., 1995). The Ca^{2+} -bound state typically resembles a dumbbell with N-terminal and C-terminal domains separated by a fully helical linker (Fig. 3a). When activated, clefts of hydrophobic residues are exposed in the two domains. The exposed hydrophobic surfaces enable binding to a large number of target proteins. Through regulation of these target proteins, CaM plays an important part in many physiologically vital pathways (Kanehisa et al., 2017).

We made a qualitative comparison between different methods (extractors) applied to a dataset extracted from enhanced sampling trajectories of Ca^{2+}-bound (holo) calmodulin (CaM). The dataset contains six states found with spectral clustering of the C-terminal domain of CaM (Westerlund and Delemotte, 2018).

Fig. 3b-c show the average estimated feature importance obtained from the CaM dataset, using internal coordinates as input. The peaks obtained from the different extractors are shown as sticks on the molecular structure (Fig 3a) and highlighted with vertical dashed lines in the importance profiles (Fig 3b). Notably, the supervised methods agree well while the unsupervised methods provide noisy profiles. Moreover, the residues that were found important are either highly conserved across phylogenetic groups (Halling et al., 2016), involved in binding target proteins (Villarroel et al., 2014), or known to cause severe cardiac diseases upon mutation (Limpitikul et al., 2014). M109, M124 and M144-145, for example, are highly conserved (Halling et al., 2016) and recognized to play important roles for target-protein binding through various packings (Shukla et al., 2016). Another example is F141, a well-studied mutation linked to long QT syndrome and decrease in Ca^{2+} affinity (Crotti et al., 2013; Limpitikul et al., 2014).

The supervised methods specifically identify V108 and A128, two residues that are not typical target protein binders. However, both residues are highly conserved (Halling et al., 2016) and thus likely play stabilizing roles for exposing binding surfaces. In fact, the packing and repacking of V108 between F89 and F92 is linked to the transition between apo and holo-like C-terminal states (Shukla et al., 2016). A third surprising peak is that of V136, which is not conserved (Halling et al., 2016) and only participates in binding to some extent (Villarroel et al., 2014). This residue is, however, part of the C-terminal domain beta-sheet that was shown to be disrupted in an extra-open state of the analyzed dataset (Westerlund and Delemotte, 2018). Thus, by contrasting the output of the different methods, it is evident that the supervised methods show more distinguishable peaks and the ensemble of methods successfully uncovers important residues of the CaM C-terminal domain.

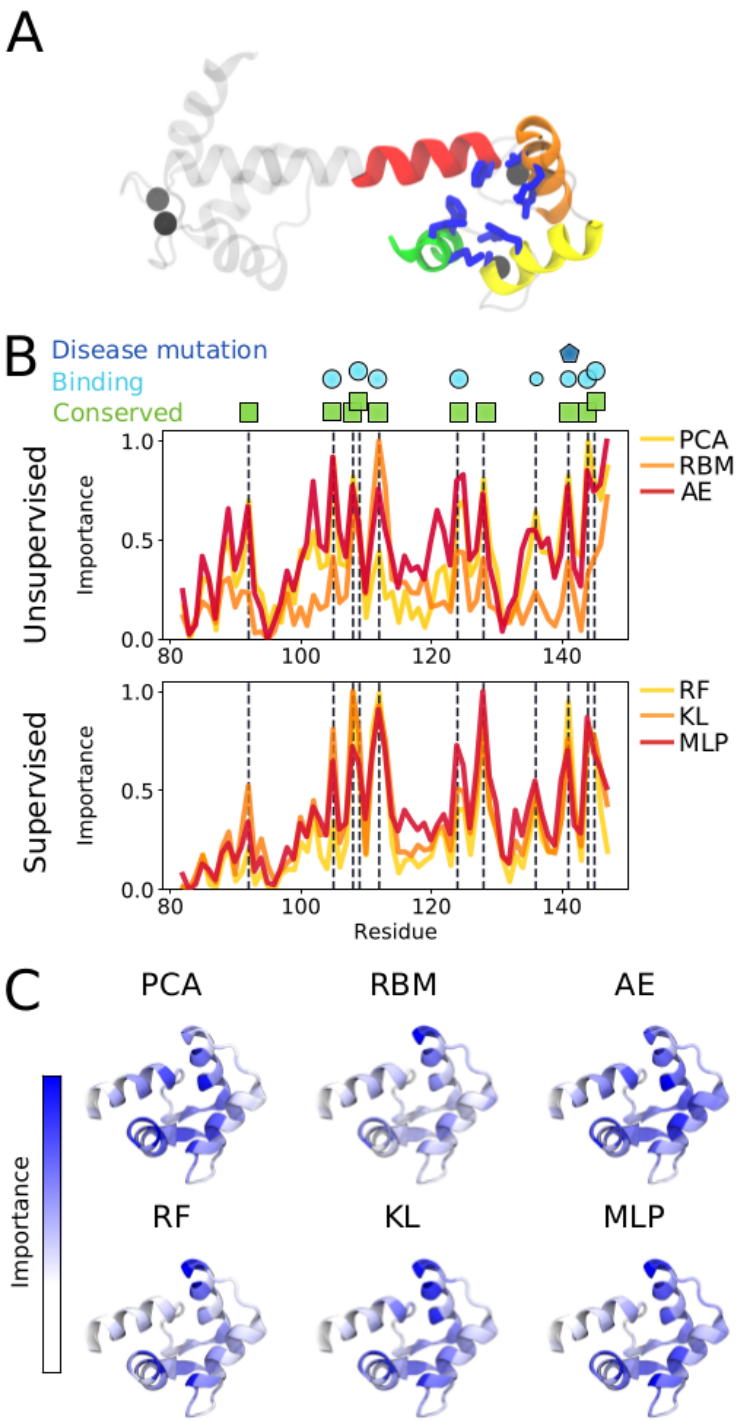


Figure 3. Important residues in the calmodulin C-terminal domain according to the feature extractors. (a) A molecular structure of Ca^{2+}-bound CaM. The helices of the C-terminal domain are represented by colored ribbons, while Ca^{2+} ions are shown as black spheres. The residues identified as important in the analysis are shown with blue sticks. (b) The importance plotted along the residue sequence of unsupervised (upper) and supervised (lower) methods. The peaks identified through

visual inspection of the supervised method results are marked by vertical dashed lines showing that most extractors agree. The peak residues are marked as being conserved (green box), involved in binding (cyan circle) or known to be mutated in a disease (blue pentagon). (c) Molecular structures of the CaM C-terminal domain with important residues identified by the six extractors highlighted in blue.

SUPERVISED LEARNING OF THE B2 ADRENERGIC RECEPTOR IDENTIFIES LIGAND INTERACTION SITES AND REVEALS ALLOSTERIC COUPLING, WHILE UNSUPERVISED METHODS LEARN THE MAIN CHARACTERISTICS OF GPCR ACTIVATION.

G protein-coupled receptors (GPCRs) form a large family of receptors in the human genome and act as an important class of modern drug targets (Hauser et al., 2017). The β2 adrenergic receptor (β2AR; Fig. 4a) is a class A GPCR often used as a prototype in studies of GPCR activation. In its native state without any binding partner, the intracellular part of the β2AR undergoes interconversion between an inactive and an active state allowing a G protein to bind to the intracellular region, a process called basal activity. Agonist ligands interact with residues in the orthosteric site and shift the equilibrium between states to make the active state more energetically favorable.

We analyzed the conformational ensemble of short trajectories from a string-of-swarms simulation of β2AR activation in the presence and absence of an agonist ligand (Fleetwood et al., 2019). Unsupervised learning with PCA on the entire internal coordinate dataset revealed known hallmarks of activation (Fig. 4b,c), including significant rearrangements of helices around the G-protein binding site, especially the twist of the conserved NPxxY motif at the bottom of TM7 (Latorraca et al., 2017; Weis and Kobilka, 2018). Another protruding hallmark of GPCR activation identified by unsupervised learning was the displacement of the intracellular part of TM6, including E268, which is the residue to undergo the largest conformational change upon activation (Rasmussen et al., 2011), as well as residues L272, L275 and M279 on TM6, whose side chains are rotated upon activation (Dror et al., 2011). The third most important region consisted of residues just below the extracellular N-terminus, a disordered region consistently changing its orientation throughout the simulation. Its large variance in interactions with other residues makes it prone to be identified as important by these methods, even though its movement is not considered a typical characteristic of GPCR activation. Finally, L144 at intracellular loop 2 was shown as

important, and it is known to be implicated in β2AR/Gs coupling (Moreira, 2014; Moro et al., 1993).

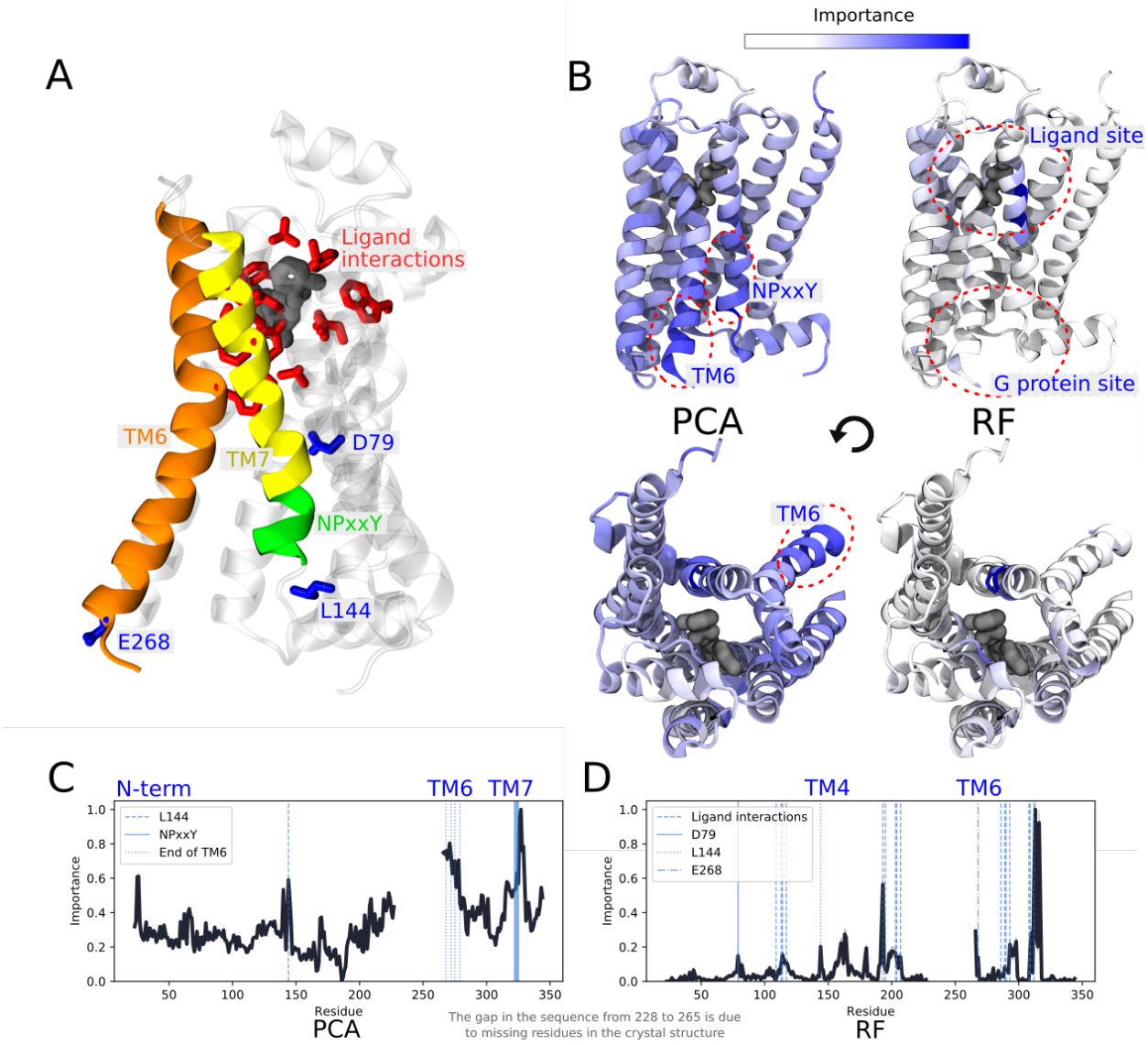


Figure 4. Important residues of the β2 adrenergic receptor using frames from a string-of-swarms simulation of the receptor's activation (Fleetwood et al., 2019). (a) A molecular structure of the β2 adrenergic receptor bound to an agonist ligand. Transmembrane helices TM6 and TM7 as well the NPxxY motif are represented as colored ribbons. Important residues known to interact with ligands are shown with red sticks while other key residues identified in the analysis are colored in blue. (b) active state structures with important residues highlighted in blue. (c) unsupervised learning with PCA. The G-protein binding site undergoes the most significant conformational change upon activation, especially TM6 and the NPxxY motif at the bottom of TM7. (d) supervised learning with

an RF classifier trained to identify protein conformations as either being ligand bound or ligand unbound. Residues known to interact with the ligand binding site (Chan et al., 2016) coincide with the residues identified as important. Important residues to discriminate between the two conditions are also located in the G-protein binding site, an indication of allosteric coupling between the intracellular part of the receptor and the orthosteric site.

Following the guidelines in our checklist (STAR Methods) we trained an RF classifier to discriminate between the ligand-bound and ligand-unbound simulation frames within the same dataset. After training, the most important residues were identified close to the ligand binding site (Fig. 4b,d), most of which are well known to interact with agonist ligands (Chan et al., 2016). Other parts of the receptor interacting directly with these residues were also identified as important, especially residues along transmembrane helix 4 (TM4). T164, for example, interacts with ligand binding site residues S203, S207 and V114. In general, residues further away from the orthosteric site were considered less important (Fig. 4d). One exception is the local peak at the D79 residue on TM2, the most conserved residue in all class A GPCRs. This residue is also thought to change protonation state during activation (Fig. 4d) (Fleetwood et al., 2019). Remarkably, important residues were further identified around the G-protein binding site: L144 and E268 at the bottom of TM6, which sample both larger as well as smaller values of displacement in the apo trajectories (Fleetwood et al., 2019). This result is indicative of allosteric coupling between the two sites and shows that the classifier is able to detect subtle local changes in the G-protein binding site that are induced by agonist binding. It suggests that ligand binding not only shifts the energy of the active state, but forces the G-protein binding site to adopt alternative conformations. The intracellular end of TM6 and the NPxxY motif do not appear as important on the other hand, presumably because they are observed in similar conformations both in the ligand bound and the ligand-unbound ensemble.

In Fig. 4b-d it can clearly be seen that RF gives a more easily-interpretable profile compared to PCA, in which the peaks are distinct and the average importance is low. This is not merely a result of the chosen methods; it is also due to the kind of insights they provide. GPCR activation involves conformational change throughout the protein, while ligand binding and allosteric signalling are more local effects. To conclude, unsupervised ML was useful on this dataset to study the general large-scale activation mechanism, whereas supervised ML successfully showed the effect of agonist binding throughout the conformational ensemble.

SUPERVISED LEARNING REVEALS STATE-DEPENDENT RESIDUES ALLOSTERICALLY COUPLED TO THE GATING CHARGE.

The voltage sensor domain (VSD) is a common module of voltage-gated ion channels responsible for reacting to changes in the membrane potential (Hille, 2001). It is usually composed of four transmembrane segments named S1 to S4, and a short surface C-terminal helix, the S4-S5 linker (Fig. 5a) (Long et al., 2005; Payandeh et al., 2011; Swartz, 2008). A few positively charged residues on the S4 segment confer voltage sensitivity to the VSD function (Fig. 5a). As they are located in a low dielectric medium, these residues are specifically sensitive to an externally applied electric field, and thus adapt their positions according to the field direction and strength (Delemotte et al., 2011; Treptow and Tarek, 2006; Treptow et al., 2009). For instance, when the membrane is depolarized they move toward the extracellular solution. This motion engages the entire S4 segment which promotes the voltage sensor activation: the S4 transition from resting (inward-facing) to the activated (outward-facing) state (Jensen et al., 2012; Khalili-Araghi et al., 2010; Schwaiger et al., 2011; Vargas et al., 2012; Yarov-Yarovoy et al., 2012). To ensure the stability of the positive charges inside the low dielectric medium, they are paired with negative charges of the three remaining segments. During activation, the positive charges switch negatively charged counterparts so that a network of salt bridges between S4 and S1-S3 is maintained in every conformational state of the VSD (Vargas et al., 2012).

In a previous study, we reported five conformational states of the Kv1.2 VSD called E (resting), A (activated) and $\Delta, \Gamma,$ B (intermediate states) (Delemotte et al., 2011, 2017). Here, we followed the provided checklist and trained a RF classifier to compare these states and elucidate their defining structural elements using more than 12,000 frames from an enhanced sampling trajectory (see STAR Methods). Fig. 5b shows the per-residue importance obtained using RF classifiers, averaged over all five states. The first top-ranked residues (R1, R2, R3 and R4) are four out of six positive charges of the S4 segment (R1, R2, R3, R4, K5 and R6). Thus, in agreement with the classical view, the positively charged residues are suggested to be crucial in delineating the five conformational states of the Kv1.2 VSD. Analysis of the per-residue importance of each state (Fig. 5c) reveals that while top-ranked charges are indeed state-specific (R4 in A; R3 and R4 in B; R2 and R3 in Γ; R1 and R2 in Δ; R1 in E), their localization is state-independent: all of them are placed near the so-called hydrophobic plug, the central part of the VSD where the response to the

external electric field is the largest (Delemotte et al., 2011; Kasimova et al., 2018; Tao et al., 2010). Other highly ranked residues pinpoint positions residues located along transmembrane segments and facing the protein lumen. Taken together, our data suggests that the interaction with the hydrophobic plug is a key aspect to distinguish between the metastable activation states in the Kv1.2 VSD.

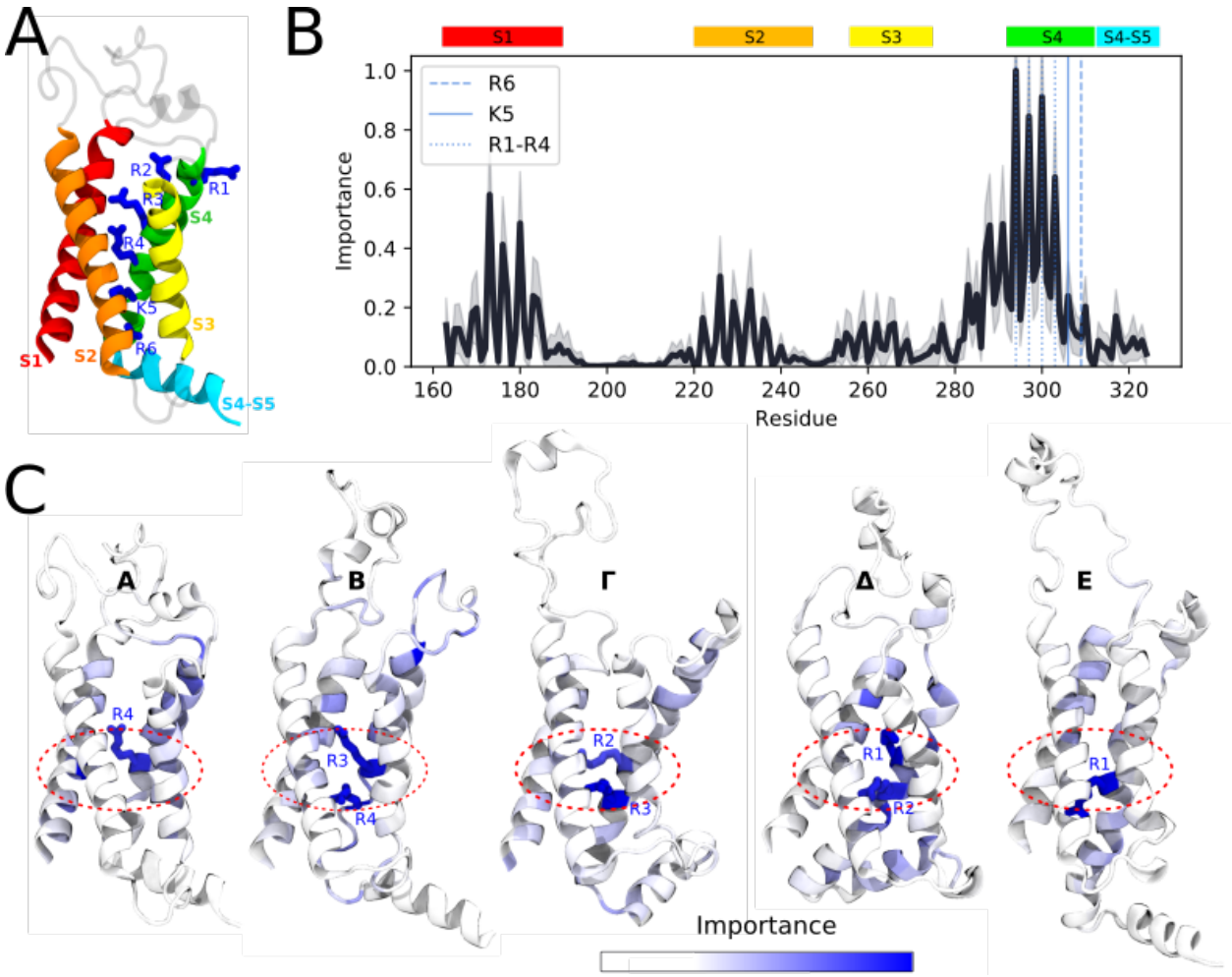


Figure 5. Identifying key residues for Kv1.2 voltage sensor function. (a) Cartoon representation of the Kv1.2 voltage sensor. The four membrane segments S1-S4 and the surface S4-S5 helix are shown. S4 positive charges are represented as blue sticks. (b) Per-residue importance obtained using RF classifiers and averaged over the five states of the Kv1.2 VSD. The colored rectangles denote the helical segments. S4 charges (R1, R2, R3, R4, K5 and R6) are marked with the blue dashed lines. The average importance and its standard deviation are shown as the black line and the shadow around it, respectively. (c) Per-state per-residue importance. The top-ranked residues are shown as sticks. The area of the hydrophobic plug is marked as a red dashed zone.

SUPERVISED AND UNSUPERVISED LEARNING PLAY DIFFERENT ROLES IN IDENTIFYING IMPORTANT FEATURES.

Supervised and unsupervised methods are designed for different purposes and are not directly comparable. Sometimes the labels are inherent to the computational experiment, such as in the example of comparing ligand bound and ligand unbound simulations of β2AR, or when comparing the conformational ensemble of a wildtype protein with a specific mutant. For such cases, supervised learning techniques may reveal subtle differences invisible to unsupervised techniques. The performance evaluated with the toy model also shows that supervised learning generally outperforms unsupervised learning on labelled datasets.

Standalone unsupervised learning becomes more attractive when simulation data cannot be labeled directly. In order to use supervised learning for such cases, one would need to preprocess the data, for example by manual labelling or with a clustering method, to assign each frame to a state. Given adequate states, supervised learning will provide useful insights about the data, as was done for calmodulin and the Kv1.2 VSD in this study. Unsupervised learning techniques can also be used as prior feature extractors to supervised classifiers or provide a projection onto a low dimensional space where it is straightforward to cluster structures into states. A strength of having different tools for computing the importance per feature within the same framework is that it becomes straightforward to combine supervised and unsupervised learning to give a coherent view of what features are important in a system. It is simple to extend this framework to include more methods as long as there is an algorithm to derive the importance per input feature.

For practical applications, the best choice of method depends on the type of data available. Regardless, the results should be averaged over many runs and cross-validated to reduce the variance induced by stochastic solvers, as well as the risk of introducing inaccuracies due to overfitting. However, unless a good set of input features is known or can be validated, nonlinear ML methods such as neural networks will be more reliable than linear methods in general.

FEATURE SELECTION AND MODEL PARAMETER CHOICE CAN ENHANCE PERFORMANCE, BUT ARE NOT OBSTACLES FOR INTERPRETABILITY.

We used inverse interatomic distances as features for the biological systems, which puts more emphasis on local changes than mere distances, and normalized the values to enable the different ML methods to perform well (STAR Methods). The choice of internal coordinates may seem straightforward following the results obtained from the toy model. For large systems, however, it may be computationally intractable to consider all internal coordinates as features and, in the case of limited amount of simulation frames, can also increase the chance of overfitting the feature extractors. In addition to performing cross-validation to overcome this problem, distances which did not reach below a certain cutoff in the trajectory were pre-filtered in the case of CaM and VSD (STAR Methods). This also has a physical significance since atoms close to each other have a stronger interaction energy. Long-range allosteric effects may still be discovered and propagated through nonlinear methods.

In principle, any intuitively interesting reaction coordinate can be included in the set of input features. In protein folding, dihedral angles are often reported as important, and should thus be included. Other collective variables such as hydration, ion binding, etc. may be considered as part of the input feature dataset. We note here too that we have only taken structural (and not dynamic) information into account, following the nature of the dataset we had at hand. To account for dynamics, for example a time-lagged autoencoder (Wehmeyer and Noé, 2018) could be used on continuous MD simulation trajectories, such that a neural network is trained to reconstruct features using features from previous times as input, and derive the resulting importance profiles from the trained model.

For supervised learning one should be aware of the difference between choosing a multiclass RF or MLP in contrast to using KL or RF with a one-vs-the-rest splitting of the data. One-vs-rest splitting in learning enables identifying important features for specific states as opposed to only reporting the important features across the entire ensemble (Fig. S8-S10). Furthermore, the difference between multiclass and one-vs-rest may have a major effect when dealing with unbalanced datasets. In this regard, algorithms such as LRP are powerful since they allow us to

compute the feature importance for every individual frame (Supplementary Movie 1) and thereby enhance interpretability at a very fine-grained level.

Fig. S6-S10 demonstrate that the performance of the methods are affected by choosing different hyperparameters, such as the regularization constraints for a neural network or the size of the discretization parameters to compute the KL divergence. To include too many parameters in a model by adding an unnecessarily large number of hidden nodes in a MLP or to neglect parts of the dataset by considering too few PCs are both examples of what may reduce performance. We found it particularly challenging to obtain high accuracy with the AE and expect that additional effort in parameter tuning may improve its performance. Seeing that they are both unsupervised neural network based methods, it seems reasonable for the AE to obtain a performance comparable to the RBM, even though the AE used in this study, unlike the RBM or for example a variational autoencoder (Kingma and Welling, 2013), does not make any strong assumptions about the distribution of input or hidden layer latent variables. Incidentally, an optimally trained AE with sigmoid activation and a single hidden layer is strongly related to PCA (Bourlard and Kamp, 1988; Chicco et al., 2014) which suggests that the two methods may yield similar performance under this parameter setup. In general, neural network models have a many tunable hyperparameters as well as a large number of weights that need to be fitted, while simpler methods such as PCA and KL have fewer parameters. This drawback of the complicated models needs to be taken into account for practical applications.

Putting some effort into crude parameter tuning should be a rewarding effort while excessive fine tuning may not reveal any new insights. Moreover, even though it is most likely possible to further improve the performance of these methods with extended efforts of parameter tuning, the optimal set of hyperparameters for a certain classifier likely depends on the choice of input features and the biological system. From a practical point of view, it seems instead to be more useful to follow the guidelines presented here using an adequate configuration of every ML method, ensure that it is robust to cross-validation and apply it to investigate the performance of different input features. Our checklist, which covers the most important aspects of interpreting molecular simulations with ML (STAR methods), makes this process easier.

The fact that many models will yield the same conclusion irrespective of hyperparameter setup increases the credibility of these methods for practical purposes. For neural network based models,

changing the network architecture leads to trained models with completely different underlying weight matrices. As shown in this study, networks of different shape and regularization constraints nevertheless identify the same important features in a simulation system. Although, similar arguments hold for other methods, this is a particularly important result due to the enormous interest and scepticism towards deep learning. This contradicts any criticism of the black-box nature of such methods: the biological properties propagated from the output to the input are clearly consistent and traceable, as well as interpretable.

MACHINE LEARNING AS A PILLAR IN INTERPRETING OF BIOMOLECULAR SYSTEMS.

We have applied several methods borrowed from machine learning to identify and rank biologically relevant residues in molecular simulations of proteins. The possible applications of this approach spans from interpreting large amounts of simulation or experimental ensemble data to selecting CVs for enhanced sampling simulations. Compared to the computationally expensive MD simulations, these insights are cheap: all results in this study were computed on commodity hardware using software published in the public repository *demystifying* on github.

Humans did not evolve to interpret large sets of high dimensional data by mere visual inspection. ML methods were, on the contrary, specifically designed to process big datasets. Nevertheless, despite the increasing interest and faith put into ML, current state-of-the-art methods are not at a stage where computers are able to set up, run and analyze simulations autonomously. Instead we have shown how statistical models and algorithms commonly used to solve ML problems provide a powerful toolbox to efficiently make data-driven interpretations of biomolecular systems. As the timescales accessible by simulations increase and the popularity of ML tools continue to thrive across many scientific areas, we anticipate that our approach can be useful to aid many researchers in demystifying complex simulations.

CHECKLIST FOR INTERPRETING MOLECULAR SIMULATIONS WITH MACHINE LEARNING

1. Identify the problem to investigate
2. Decide if you should use supervised or unsupervised machine learning (or both)
    - ❑ The best choice depends on what data is available and the problem at hand
    - ❑ If you chose unsupervised learning, consider also clustering the simulation frames to label them and perform supervised learning
3. Select a set of features and scale them
    - ❑ For many processes, protein internal coordinates are adequate. To reduce the number of features, consider filtering distances with a cutoff
    - ❑ Consider other features that can be expressed as a function of internal coordinates you suspect to be important for the process of interest (dihedral angles, cavity or pore hydration, ion or ligand binding, etc.)
4. Chose a set of ML methods to derive feature importance
    - ❑ To quickly get a clear importance profile with little noise, consider RF or KL for supervised learning. RF may perform better for noisy data.
    - ❑ For unsupervised learning, consider PCA, which is relatively robust when conducted on internal coordinates
    - ❑ To find all important features, including those requiring nonlinear transformations of input features, also use neural network based approaches such as MLP. This may come at the cost of more peaks in the importance distribution
    - ❑ Decide if you seek the average importance across the entire dataset (all methods), the importance per state (KL, a set of binary RF classifiers or MLP), or the importance per single configuration (MLP, RBM, AE)
    - ❑ Choose a set of hyperparameters which gives as reasonable trade-off between performance and model prediction accuracy
5. Ensure that the selected methods and hyperparameter choice perform well under cross-validation
6. Average the importance per feature over many iterations
7. Check that the distribution of importance has distinguishable peaks
8. To select low-dimensional, interpretable CVs for plotting and enhanced sampling, inspect the top-ranked features
9. For a holistic view, average the importance per residue or atom and visualize the projection on the 3d system
10. If necessary, iterate over steps 3-9 with different features, ML methods and hyperparameters

ACKNOWLEDGEMENTS

This work was supported by grants from the Gustafsson Foundation and Science for Life Laboratory to LD. The simulations were performed on resources provided by the Swedish National Infrastructure for Computing (SNIC) at PDC Centre for High Performance Computing (PDC-HPC).

AUTHOR CONTRIBUTIONS

O.F., M.K. and L.D. conceptualized the project. O.F., M.K. and A.W. wrote the code and performed different parts of data curation and formal analysis: A.W. for CaM, O.F. for β2AR and M.K. for the VSD. O.F. and M.K. designed and analyzed the toy model. L.D. supervised the project. O.F., M.K. and A.W. wrote the first draft of the paper. All authors reviewed and edited the paper.

REFERENCES

Ahmad, M., Helms, V., Kalinina, O.V., and Lengauer, T. (2019). Relative principal components analysis: application to analyzing biomolecular conformational changes. J. Chem. Theory Comput. *15*, 2166–2178.

Akay, A., and Hess, H. (2019). Deep learning: current and emerging applications in medicine and technology. IEEE J. Biomed. Health Inform. *23*, 906–920.

Bishop, C.M. (2006). Pattern recognition and machine learning.

Bourlard, H., and Kamp, Y. (1988). Auto-association by multilayer perceptrons and singular value decomposition. Biol. Cybern. *59*, 291–294.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (2017). Classification and regression trees (Routledge).

Camacho, D.M., Collins, K.M., Powers, R.K., Costello, J.C., and Collins, J.J. (2018). Next-Generation Machine Learning for Biological Networks. Cell *173*, 1581–1592.

Chan, H.C.S., Filipek, S., and Yuan, S. (2016). The Principles of Ligand Specificity on beta-2-adrenergic receptor. Sci. Rep. *6*, 34736.

Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., and Blaschke, T. (2018). The rise of deep learning in drug discovery. Drug Discov. Today *23*, 1241–1250.

Chicco, D., Sadowski, P., and Baldi, P. (2014). Deep autoencoder neural networks for gene ontology annotation predictions. In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics – BCB '14, (New York, New York, USA: ACM Press), pp. 533–540.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface *15*.

Crotti, L., Johnson, C.N., Graf, E., De Ferrari, G.M., Cuneo, B.F., Ovadia, M., Papagiannis, J., Feldkamp, M.D., Rathi, S.G., Kunic, J.D., et al. (2013). Calmodulin mutations associated with recurrent cardiac arrest in infants. Circulation *127*, 1009–1017.

Degiacomi, M.T. (2019). Coupling molecular dynamics and deep learning to mine protein conformational space. Structure *27*, 1034-1040.e3.

Delemotte, L., Tarek, M., Klein, M.L., Amaral, C., and Treptow, W. (2011). Intermediate states of the Kv1.2 voltage sensor from atomistic molecular dynamics simulations. Proc Natl Acad Sci USA *108*, 6109–6114.

Delemotte, L., Kasimova, M.A., Klein, M.L., Tarek, M., and Carnevale, V. (2015). Free-energy landscape of ion-channel voltage-sensor-domain activation. Proc Natl Acad Sci USA *112*, 124–129.

Delemotte, L., Kasimova, M.A., Sigg, D., Klein, M.L., Carnevale, V., and Tarek, M. (2017). Exploring the complex dynamics of an ion channel voltage sensor domain via computation. BioRxiv.

Dror, R.O., Arlow, D.H., Maragakis, P., Mildorf, T.J., Pan, A.C., Xu, H., Borhani, D.W., and Shaw, D.E. (2011). Activation mechanism of the β2-adrenergic receptor. Proc Natl Acad Sci USA *108*, 18684–18689.

Fleetwood, O., Matricon, P., Carlsson, J., and Delemotte, L. (2019). Energy landscapes reveal agonist's control of GPCR activation via microswitches. BioRxiv.

Halling, D.B., Liebeskind, B.J., Hall, A.W., and Aldrich, R.W. (2016). Conserved properties of individual Ca2+-binding sites in calmodulin. Proc Natl Acad Sci USA *113*, E1216-25.

Hashemian, B., Millán, D., and Arroyo, M. (2013). Modeling and enhanced sampling of molecular systems with smooth and nonlinear data-driven collective variables. J. Chem. Phys. *139*, 214101.

Hauser, A.S., Attwood, M.M., Rask-Andersen, M., Schiöth, H.B., and Gloriam, D.E. (2017). Trends in GPCR drug discovery: new agents, targets and indications. Nat. Rev. Drug Discov. *16*, 829–842.

Hille, B. (2001). Ion channels of excitable membranes. Sinauer Associates. Sunderland, MA.

Hinton, G.E., and Salakhutdinov, R.R. (2006). Reducing the dimensionality of data with neural networks. Science *313*, 504–507.

Ho, T.K. (1995). Random decision forests. Proceedings of 3rd International Conference on Document Analysis and Recognition *1*, 278–282.

Humphrey, W., Dalke, A., and Schulten, K. (1996). VMD: visual molecular dynamics. J. Mol. Graph. *14*, 33–38, 27.

Jang, H.-J., and Cho, K.-O. (2019). Applications of deep learning for the analysis of medical data. Arch. Pharm. Res.

Jensen, M.Ø., Jogini, V., Borhani, D.W., Leffler, A.E., Dror, R.O., and Shaw, D.E. (2012). Mechanism of voltage gating in potassium channels. Science *336*, 229–233.

Jung, H., Covino, R., and Hummer, G. (2019). Artificial Intelligence Assists Discovery of Reaction Coordinates and Mechanisms from Molecular Dynamics Simulations. ArXiv.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res. *45*, D353–D361.

Kasimova, M.A., Lindahl, E., and Delemotte, L. (2018). Determining the molecular basis of voltage sensitivity in membrane proteins. J. Gen. Physiol. *150*, 1444–1458.

Khalili-Araghi, F., Jogini, V., Yarov-Yarovoy, V., Tajkhorshid, E., Roux, B., and Schulten, K. (2010). Calculation of the gating charge for the Kv1.2 voltage-activated potassium channel. Biophys. J. *98*, 2189–2198.

Kingma, D.P., and Ba, J. (2014). Adam: A Method for Stochastic Optimization. ArXiv.

Kingma, D.P., and Welling, M. (2013). Auto-Encoding Variational Bayes. ArXiv.

Kuboniwa, H., Tjandra, N., Grzesiek, S., Ren, H., Klee, C.B., and Bax, A. (1995). Solution structure of calcium-free calmodulin. Nat. Struct. Biol. *2*, 768–776.

Latorraca, N.R., Venkatakrishnan, A.J., and Dror, R.O. (2017). GPCR dynamics: structures in motion. Chem. Rev. *117*, 139–155.

Limpitikul, W.B., Dick, I.E., Joshi-Mukherjee, R., Overgaard, M.T., George, A.L., and Yue, D.T. (2014). Calmodulin mutations associated with long QT syndrome prevent inactivation of cardiac L-type Ca(2+) currents and promote proarrhythmic behavior in ventricular myocytes. J. Mol. Cell. Cardiol. *74*, 115–124.

Long, S.B., Campbell, E.B., and Mackinnon, R. (2005). Crystal structure of a mammalian voltage-dependent Shaker family K+ channel. Science *309*, 897–903.

Lundervold, A.S., and Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. *29*, 102–127.

McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.-P., Lane, T.J., and Pande, V.S. (2015). Mdtraj: A modern open library for the analysis of molecular dynamics trajectories. Biophys. J. *109*, 1528–1532.

Mittal, S., and Shukla, D. (2018). Recruiting machine learning methods for molecular simulations of proteins. Mol. Simul. 1–14.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., and Müller, K.-R. (2017). Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit. *65*, 211–222.

Montavon, G., Samek, W., and Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. Digit. Signal Process. *73*, 1–15.

Moreira, I.S. (2014). Structural features of the G-protein/GPCR interactions. Biochim. Biophys. Acta *1840*, 16–33.

Moro, O., Lameh, J., Högger, P., and Sadée, W. (1993). Hydrophobic amino acid in the i2 loop plays a key role in receptor-G protein coupling. J. Biol. Chem. *268*, 22273–22276.

Olah, C., Satyanarayan, A., Johnson, I., Carter, S., Schubert, L., Ye, K., and Mordvintsev, A. (2018). The building blocks of interpretability. Distill *3*.

Payandeh, J., Scheuer, T., Zheng, N., and Catterall, W.A. (2011). The crystal structure of a voltage-gated sodium channel. Nature *475*, 353–358.

Pedregosa, F., Varoquaux, G., and Gramfort, A. (2011). Scikit-learn: Machine learning in Python. … of Machine Learning ….

Raschka, S. (2017). BioPandas: Working with molecular structures in pandas DataFrames. JOSS *2*, 279.

Rasmussen, S.G.F., DeVree, B.T., Zou, Y., Kruse, A.C., Chung, K.Y., Kobilka, T.S., Thian, F.S., Chae, P.S., Pardon, E., Calinski, D., et al. (2011). Crystal structure of the β2 adrenergic receptor-Gs protein complex. Nature *477*, 549–555.

Ribeiro, J.M.L., and Tiwary, P. (2018). Achieving Reversible Ligand-Protein Unbinding with Deep Learning and Molecular Dynamics through RAVE. BioRxiv.

Ribeiro, J.M.L., Bravo, P., Wang, Y., and Tiwary, P. (2018). Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). J. Chem. Phys. *149*, 072301.

Schwaiger, C.S., Bjelkmar, P., Hess, B., and Lindahl, E. (2011). 3_{10}-helix conformation facilitates the transition of a voltage sensor S4 segment toward the down state. Biophys. J. *100*, 1446–1454.

Shukla, D., Peck, A., and Pande, V.S. (2016). Conformational heterogeneity of the calmodulin binding interface. Nat. Commun. *7*, 10910.

Sidky, H., and Whitmer, J.K. (2018). Learning free energy landscapes using artificial neural networks. J. Chem. Phys. *148*, 104111.

Sittel, F., and Stock, G. (2018). Perspective: Identification of collective variables and metastable states of protein dynamics. J. Chem. Phys. *149*, 150901.

Sittel, F., Jain, A., and Stock, G. (2014). Principal component analysis of molecular dynamics: on the use of Cartesian vs. internal coordinates. J. Chem. Phys. *141*, 014111.

Smolensky, P. (1986). Information Processing in Dynamical Systems: Foundations of Harmony Theory.

Sultan, M.M., Kiss, G., Shukla, D., and Pande, V.S. (2014). Automatic selection of order parameters in the analysis of large scale molecular dynamics simulations. J. Chem. Theory Comput. *10*, 5217–5223.

Swartz, K.J. (2008). Sensing voltage across lipid membranes. Nature *456*, 891–897.

Tao, X., Lee, A., Limapichat, W., Dougherty, D.A., and MacKinnon, R. (2010). A gating charge transfer center in voltage sensors. Science *328*, 67–73.

Tipping, M.E., and Bishop, C.M. (1999). Probabilistic Principal Component Analysis. J. Royal Statistical Soc. B *61*, 611–622.

Trapl, D., Horvacanin, I., Mareska, V., Ozcelik, F., Unal, G., and Spiwok, V. (2019). Anncolvar: approximation of complex collective variables by artificial neural networks for analysis and biasing of molecular simulations. Front. Mol. Biosci. *6*, 25.

Treptow, W., and Tarek, M. (2006). Environment of the gating charges in the Kv1.2 Shaker potassium channel. Biophys. J. *90*, L64-6.

Treptow, W., Tarek, M., and Klein, M.L. (2009). Initial response of the potassium channel voltage sensor to a transmembrane potential. J. Am. Chem. Soc. *131*, 2107–2109.

Vargas, E., Yarov-Yarovoy, V., Khalili-Araghi, F., Catterall, W.A., Klein, M.L., Tarek, M., Lindahl, E., Schulten, K., Perozo, E., Bezanilla, F., et al. (2012). An emerging consensus on voltage-dependent gating from computational modeling and molecular dynamics simulations. J. Gen. Physiol. *140*, 587–594.

Villarroel, A., Taglialatela, M., Bernardo-Seisdedos, G., Alaimo, A., Agirre, J., Alberdi, A., Gomis-Perez, C., Soldovieri, M.V., Ambrosino, P., Malo, C., et al. (2014). The ever changing moods of calmodulin: how structural plasticity entails transductional adaptability. J. Mol. Biol. *426*, 2717–2735.

Wehmeyer, C., and Noé, F. (2018). Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. J. Chem. Phys. *148*, 241703.

Weis, W.I., and Kobilka, B.K. (2018). The Molecular Basis of G Protein-Coupled Receptor Activation. Annu. Rev. Biochem. *87*, 897–919.

Westerlund, A.M., and Delemotte, L. (2018). Effect of Ca2+ on the promiscuous target-protein binding of calmodulin. PloS Comput. Biol. *14*, e1006072.

Yarov-Yarovoy, V., DeCaen, P.G., Westenbroek, R.E., Pan, C.-Y., Scheuer, T., Baker, D., and Catterall, W.A. (2012). Structural basis for gating charge movement in the voltage sensor of a sodium channel. Proc Natl Acad Sci USA *109*, E93-102.

Zhang, J., and Chen, M. (2018). Unfolding hidden barriers by active enhanced sampling. Phys. Rev. Lett. *121*, 010601.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M.M. (2019). Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Inf. Fusion *50*, 71–91.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. (2019). A primer on deep learning in genomics. Nat. Genet. *51*, 12–18.

METHODS

KEY RESOURCES TABLE

Reagent or Resource	Source	identifier
Deposited data		
Calmodulin trajectory	(Westerlund and Delemotte, 2018)	https://doi.org/10.1371/journal.pcbi.1006072
β2 trajectory	(Fleetwood et al., 2019)	https://doi.org/10.1101/627026
VSD trajectory	(Delemotte et al., 2017)	https://doi.org/10.1073/pnas.1416959112
Extracted data for this paper	This paper	https://drive.google.com/open?id=19V1mXz7Yu0V_2JZQ8wtgt7aZusAKs2Bb
Software		
demystifying	This paper	https://github.com/delemottellab/demystifying
VMD	(Humphrey et al., 1996)	http://www.ks.uiuc.edu/Research/vmd/
Python 2.7	Open source software	https://www.python.org/
scikit-learn	(Pedregosa et al., 2011)	https://scikit-learn.org/
MDTraj	(McGibbon et al., 2015)	http://mdtraj.org/

Numpy	Open source software	http://www.numpy.org/
biopandas	(Raschka, 2017)	http://rasbt.github.io/biopandas/

PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) converts the input features to a set of linearly uncorrelated variables (principal components; PCs) through an orthogonal transformation (Tipping and Bishop, 1999). The first component covers the largest variance possible in the dataset. The feature importance is taken to be the weighted average of the components of the PCs up to a certain threshold in accumulated variance with the variance covered (i.e. the eigenvalues) by the components as weights (Fig. 2a). The importance was normalized between 0 and 1.

RESTRICTED BOLTZMANN MACHINE

A restricted Boltzmann machine (RBM) (Smolensky, 1986) is a generative stochastic neural network trained to maximize the likelihood of the data using a graphical model with a layer of hidden nodes connected to the input nodes (Fig. 2b). We used an implementation from scikit-learn of a Bernoulli RBM. Training was performed with stochastic maximum likelihood. To find important features, LRP was performed from the output of the hidden layer to the input layer.

Note that in all computations, the input features were scaled between values of 0 and 1. The RBM used in this study assumes a probabilistic interpretation of the input features with values between 0 and 1, where the upper limit indicates that a specific input node is active. When feeding inverse distances to the RBM, larger input values correspond to residues in contact with each other, this models makes physical sense as long as the values are scaled properly. Cartesian coordinates, even after scaling, do not fulfill the assumptions made about the input features and have been taken into consideration only for the sake of consistency.

LAYER WISE RELEVANCE PROPAGATION (DEEP TAYLOR DECOMPOSITION)

Layer wise relevance propagation (LRP), or deep Taylor decomposition when applied to networks with ReLU activation (Montavon et al., 2017), is a method originally developed for image classification problems with the aim to produce a decomposition (an explanation) of a prediction

that can be visualized in the same way as the input data. Given a trained neural network, it is possible to perform a backward pass through the network (Fig. 2f), keeping track of which nodes were activated in a certain classification decision, and redistribute the relevance in one layer to the connected nodes in the previous layer. The redistribution of relevance works such that nodes which contributed most in a previous layer acquire more of the relevance. The importance of a feature was derived by computing the average relevance over all input sets in the training set and normalizing the importance for all features and frames to have an upper bound of 1. Unlike with the other methods in this study, it is possible to compute the importance of a feature for a specific simulation frame in this way.

AUTOENCODER

The autoencoder (AE) is a generative neural network trained to reconstruct the features through a set of hidden layers of lower dimensionality (Fig. 2c) (Hinton and Salakhutdinov, 2006). In short, the model can be broken into two parts: encoding and decoding. In the encoding part, the layers decrease in size causing the network to ignore noise during training. In the decoding part the layers increase in size up to the output layer, which consists of reconstructed input features.

The AE was implemented by training a scikit-learn MLP regressor (Pedregosa et al., 2011) with the same data as input and output. The encoding layers and the decoding layers were always set to be of the same shape. To find important features, LRP was performed from the reconstructed output to the actual input.

KULLBACK LEIBLER DIVERGENCE

The Kullback Leibler (KL) divergence (Fig. 2d) (Bishop, 2006), also called relative entropy, is a measure of the difference between two distributions $P(x)$ and $Q(x)$ along a feature X: $\sum_{x \in X} P(x) \, log \frac{P(x)}{Q(x)}$. The importance of a feature was set to be a symmetric version of the KL divergence, $\frac{1}{2} P(x) \, log \frac{P(x)}{Q(x)} + \frac{1}{2} Q(x) \, log \frac{Q(x)}{P(x)}$, between one state and all other states along that feature. All importance scores were normalized between 0 and 1.

RANDOM FOREST CLASSIFIER

A random forest (RF) classifier (Ho, 1995) (Fig. 2e) is an ML model in which a prediction is taken as an ensemble average over many decision trees, in which the features considered to make a

classification decision are randomized. Important features were identified by computing the forest's mean decrease impurity (Breiman et al., 2017), which is essentially a measure of how often a feature is used to split a node in the underlying decision trees, using the scikit-learn (Pedregosa et al., 2011) implementation. To compute the feature importance for a specific state, the RF was alternatively used as a one-vs-the-rest binary classifier trained to distinguish one state from all the others.

MULTILAYER PERCEPTRON CLASSIFIER

A Multilayer perceptron (MLP) classifier (Fig. 2f) is a feedforward artificial neural network with fully connected layers (Bishop, 2006). An implementation from scikit-learn (Pedregosa et al., 2011) of an MLP classifier was used throughout this study, trained with the Adam solver (Kingma and Ba, 2014). To find important features, LRP was performed from the output to the input layer.

TOY MODEL

To construct model systems, atoms were randomly placed in a box to set up an initial configuration. To define states, a set of atoms unique to every state, typically few atoms compared to the size of the system, were labelled as important and displaced relative to the initial configuration. The toy model equivalent of simulation frames in a trajectory were generated by displacing every labeled atom's position considering uniform noise and randomly rotating the entire system around the origin. In silico systems are typically rotation invariant but aligning different configurations is often not trivial. In addition to the number of atoms and frames in the generated trajectory, the number of states, the number of important atoms per state and the strength of the noise were all configurable parameters of the toy model.

Moreover, we constructed two kinds of displacements: a linear displacement of atoms from their original position (Fig. 1a) as well as a nonlinear displacement method (Figure S2 and SI Methods). In the linear method, the magnitude and direction of the displacement was the same for all atoms. In the nonlinear method, one atom per state was displaced from its equilibrium position, the size and the direction of the displacement depending on the state index. Subsequent atoms were rotated around the position of the previous atom in the same state (see SI Methods). From a biomolecular perspective, this can be seen as a model of how interresidue contacts are broken in a linear (for

example with the displacement of one subdomain away from another) or nonlinear manner (for example with the twist of a helix).

In order to obtain statistics, we generated 10 toy models with 100 atoms in a box with a box length stretching from -1 to 1. In total there were 3 states with 10 displaced atoms per state. The magnitude of the displacement was 0.1 and the strength of the noise was 0.01 except for Fig. S3 where it was increased to 0.05. 1200 frames were generated for each state. For every ML method we computed the performance for different combinations of hyperparameters (Fig. S5-S10). For every instance of the toy model, the feature importance was computed 10 times. The parameters which gave the best performance were included in Fig. 1 and Fig. S2-S4.

When using inverse distances as features, the relevance per atom was computed by summing the average importance of all distances involving an atom and normalizing the values between 0 and 1. To take into account the smallest number of input features possible, we sought to avoid considering interatomic distances multiple times. We did so by including the distances from atom I to atoms i+1, i+2, i+3 and i+4 only; this is akin to triangulation and suffices to reconstruct the coordinates of the entire system. Cartesian coordinates generated 3N input features. To compute the importance per atom, the importance of every triplet of xyz coordinates was summed and normalized.

ACCURACY QUANTIFICATION

We used two scores to estimate a method's performance on the toy model. Given a measured distribution of importance $\bar{\phi}$, and a true distribution, $\bar{\psi}$, we computed the accuracy at *finding all* important residues as:$1 - |\bar{\psi} - \bar{\phi}|^2 \,/\, (|\bar{\psi}|^2 + |\bar{\phi}|^2)$. This mean-square-error (mse) based metric gave high scores to methods which identified all displaced atoms in the toy model and could tolerate some noise. The *ignoring irrelevant* accuracy score, designed to give high scores to methods which only identified truly important atoms without necessarily identifying all of them, was defined as $\bar{\phi} \cdot \bar{\psi} / |\bar{\phi}|$. To illustrate the scoring method, the accuracy scores of various distributions are highlighted in Fig. S1.

CALMODULIN ANALYSIS

The dataset consisted of six CaM C-terminal states extracted with spectral clustering from a mix of regular MD (3600 ns) and temperature enhanced simulations (570 ns replica exchange MD and

460 ns replica exchange solute tempering) (Westerlund and Delemotte, 2018). We used inverse Cα distances between C-terminal domain residues and a total of 7600 frames, yielding a dataset size of 7600 x 2145. The data was first shuffled in blocks of 100 frames and filtered prior to feature extraction. Only residue pairs with Cα distance less than 1.0 nm in at least one frame and more than 1.0 nm in another were kept. This cutoff-filtering decreased the number of features to 688. All features were normalized with scikit-learn's (Pedregosa et al., 2011) min-max scaler. To pick the number of components in PCA, a variance cutoff of 75 % was used. The AE and MLP both had hidden layers of 120 and 100 nodes, respectively. Both were trained with an Adam optimizer and used RelU activation functions. The RF classifier used 500 estimators and was trained with a one-vs-rest approach to obtain important features for every individual state. We used 3-fold cross-validation and averaged the residue profile over 5 independent iterations.

GPCR ANALYSIS

The dataset contained the output coordinates from every tenth short 10 ps trajectory initiated from points along the most probable transition path between β2AR's active and inactive state. Trajectories from every tenth of the 200-300 iterations of a string of swarms simulation (Fleetwood et al., 2019) were used. Inverse Cα distances between all 284 protein residues were normalized with scikit-learn's min-max scaler and were chosen as input features.

All PCs until the corresponding eigenvalue differed by at least an order of magnitude compared to the next eigenvalue were used to perform PCA. An RF classifier with 1000 estimators was trained to discriminate between ligand bound and unbound frames. In total 2674 frames were used in training with 4-fold cross validation. The final importance profiles were averaged over 30 independent iterations.

VSD ANALYSIS

The metadynamics trajectory with the five VSD states was taken from our previous study (Delemotte et al., 2015, 2017). The five states were previously identified using the network of salt bridges between the S4 positive residues and their negative counterparts as a collective variable (Delemotte et al., 2011). The overall length of the trajectory was 12175 frames with 1087, 1900, 2281, 4359, and 2548 frames corresponding to E, Δ, Γ, B and A states, respectively. Inverse heavy-atoms distances between all VSD residues were used as input features (13041 in total). Distances

larger than 0.7 nm or smaller than 0.5 nm throughout the entire trajectory were filtered out yielding a final number of 3077 input features. To identify important residues, an RF classifier was trained with 100 estimators and 3-fold cross validation. The final residue importance profile was averaged over 5 independent iterations.

SOFTWARE AVAILABILITY

Software used in this work is available for download in the Delemottelab github repository (https://github.com/delemottelab/demystifying).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures and 10 figures and can be found with this article online.

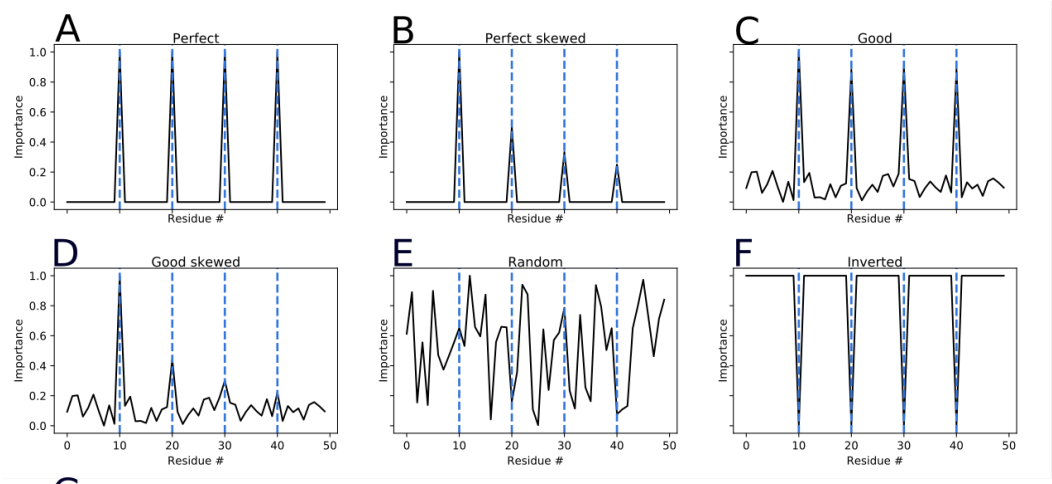
SUPPLEMENTARY METHODS

NONLINEAR DISPLACEMENT ALGORITHM FOR CONSTRUCTING THE TOY MODEL

Assuming that we have *M* states with *N* displaced atoms per state and that the magnitude of displacements is *d*, the states of the nonlinear toy model were defined as follows:

1. For every atom in state $m = 0, \ldots, M\text{-}1$ and $n = 0, .., N\text{-}1$ with position $\underline{r}_{n,m}$:
    a. if n is 0: displace the atom with the vector $\{md, 0, d(1\text{-}m)\}$
    b. else:
        i. Rotate the atom around the z axis with the angle $(-1)^n d/r_{nn'}$ where $rr_{nn'} = |\underline{r}_{n,m} - \underline{r}_{n-1,m}|_{x,y}$
        ii. Rotate the atom around the z axis with the angle $(-1)^m d/r_{mm'}$ where $r_{nn'} = |\underline{r}_{n,m} - \underline{r}_{n-1,m}|_{y,z}$

SUPPLEMENTARY FIGURES



G

Type	Name	Finding all	Ignoring irrelevant
A	Perfect	1.00	1.00
B	Perfect skewed	0.77	1.00
C	Good	0.91	0.43
D	Good skewed	0.66	0.29
E	Random	0.15	0.06
F	Inverted	0.00	0.00

Figure S1. Accuracy score values for example distributions. The vertical dashed lines highlight the residues of high importance in the true distribution. (a) A distribution of perfect accuracy. All relevant residues have an importance equal to 1 and all other residues have 0 importance. (b) A skewed version of the perfect distribution. All irrelevant residues have 0 importance and the relevant residues have varying peak heights. (c)-(d) The perfect distributions in (a) and (b) with random noise added to the distributions. (e) A distribution of random noise between 0 and 1. (f) The inverse of the true distribution has zero accuracy for both metrics. (g) Summary of the scores for all distributions. The mse-based method 'finding all' is sensitive to peak height whereas the 'ignoring irrelevant' accuracy method is not. On the other hand, the 'ignoring irrelevant' method is more sensitive to noise.

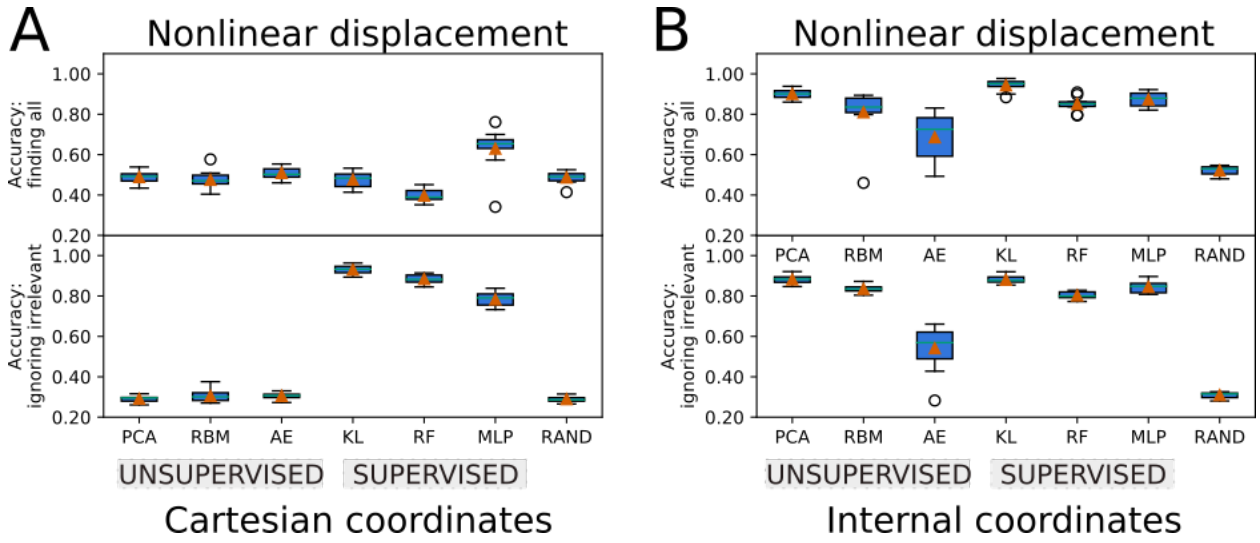


Figure S2. Non-linear displacement toy model used to benchmark the ML methods' ability to extract important features. (a), (b) Boxplots showing the accuracy of the different methods using either flattened Cartesian coordinates (a) or inverse interatomic distances as features (b). The best performing set of hyperparameters found after benchmarking (Fig. S6-S10) have been used.

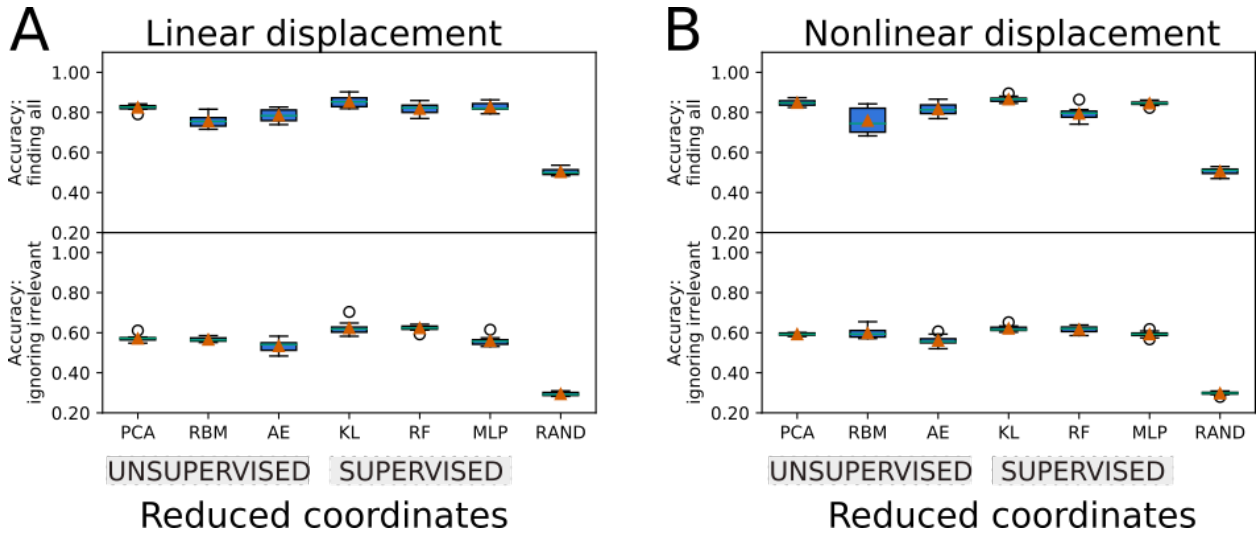


Figure S3. Toy models used to benchmark the ML methods' ability to extract important features using a reduced set of internal coordinates. (a), (b) Boxplots showing the accuracy of the different methods using either a linear (a) or nonlinear (b) toy model). The best performing set of hyperparameters found after benchmarking every method (Fig. S6-S10) have been used.

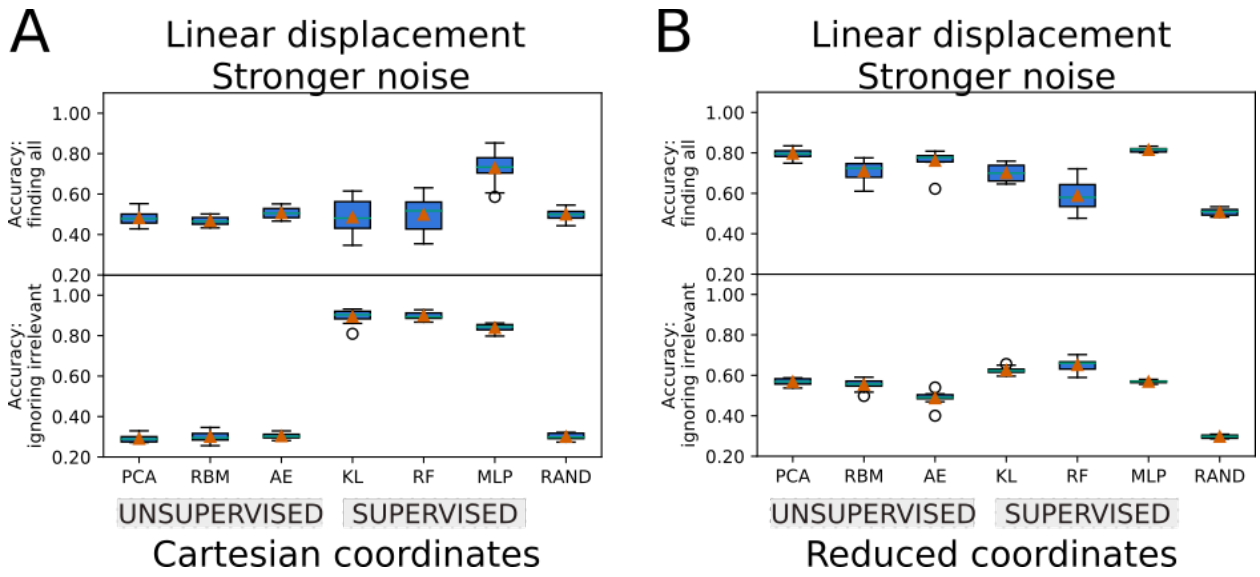


Figure S4.Toy model used to benchmark the ML methods' ability to extract important features, using a noise level equal to 50 percent of the strength of the displacement. (a), (b) Boxplots showing the accuracy of the different methods using either flattened Cartesian coordinates (a) or a reduced set of inverse interatomic distances as features (b). The best performing set of hyperparameters found after benchmarking every method (Fig. S6-S10) have been used.

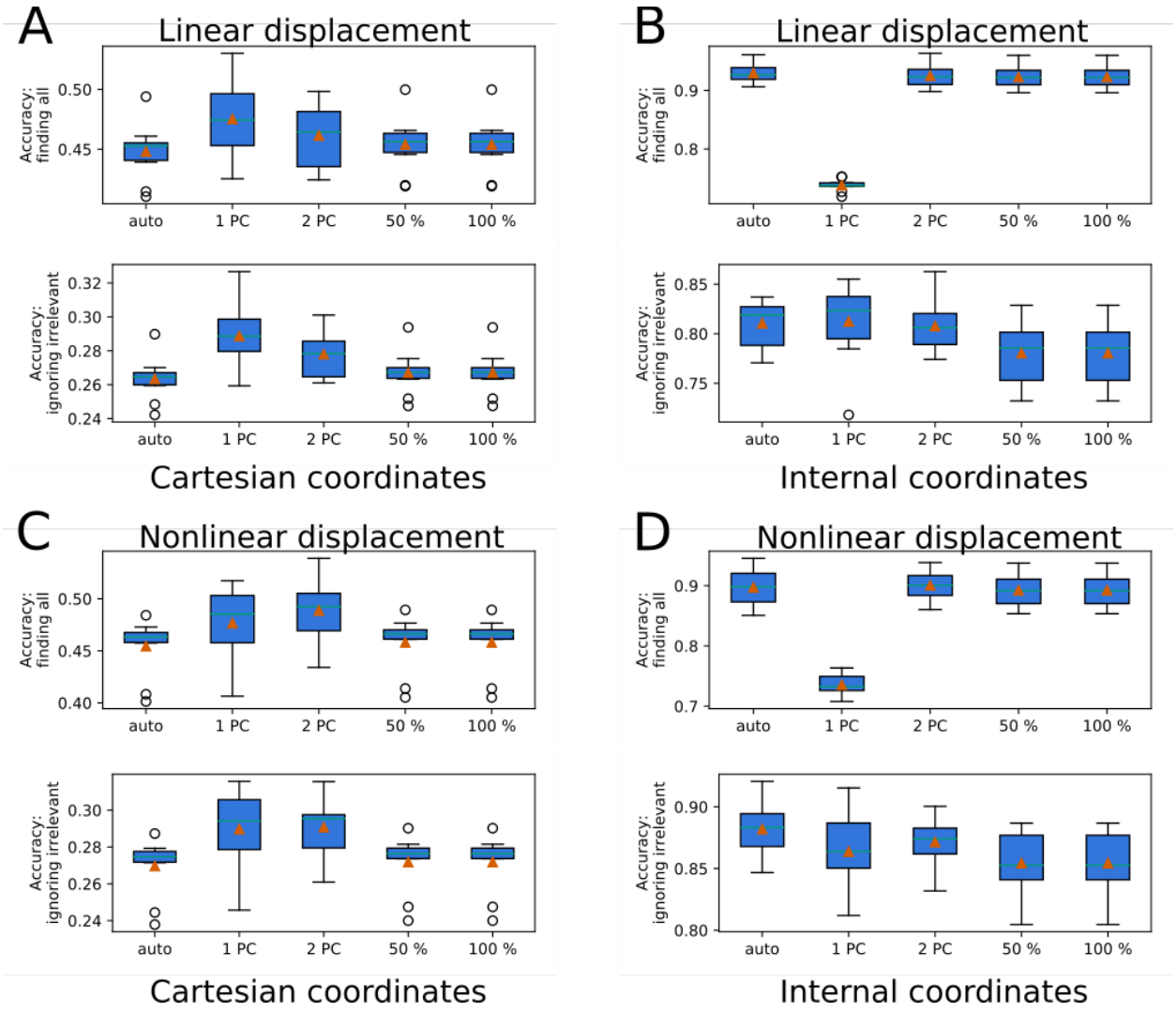


Figure S5. Hyperparameter tuning of PCA applied on the toy model. The number of PCs to include has either been determined by automatically finding ('auto') the first PC for which the next corresponding eigenvalue differs by at least an order of magnitude, by selecting a fixed number of PCs or by summing the variance covered by PCs until it exceeds a threshold given in percent. (a) Using cartesian coordinates and linear displacements, (b) using inverse distances and linear displacements, (c) using cartesian coordinates and nonlinear displacements, and (d) using inverse distances and nonlinear displacements.

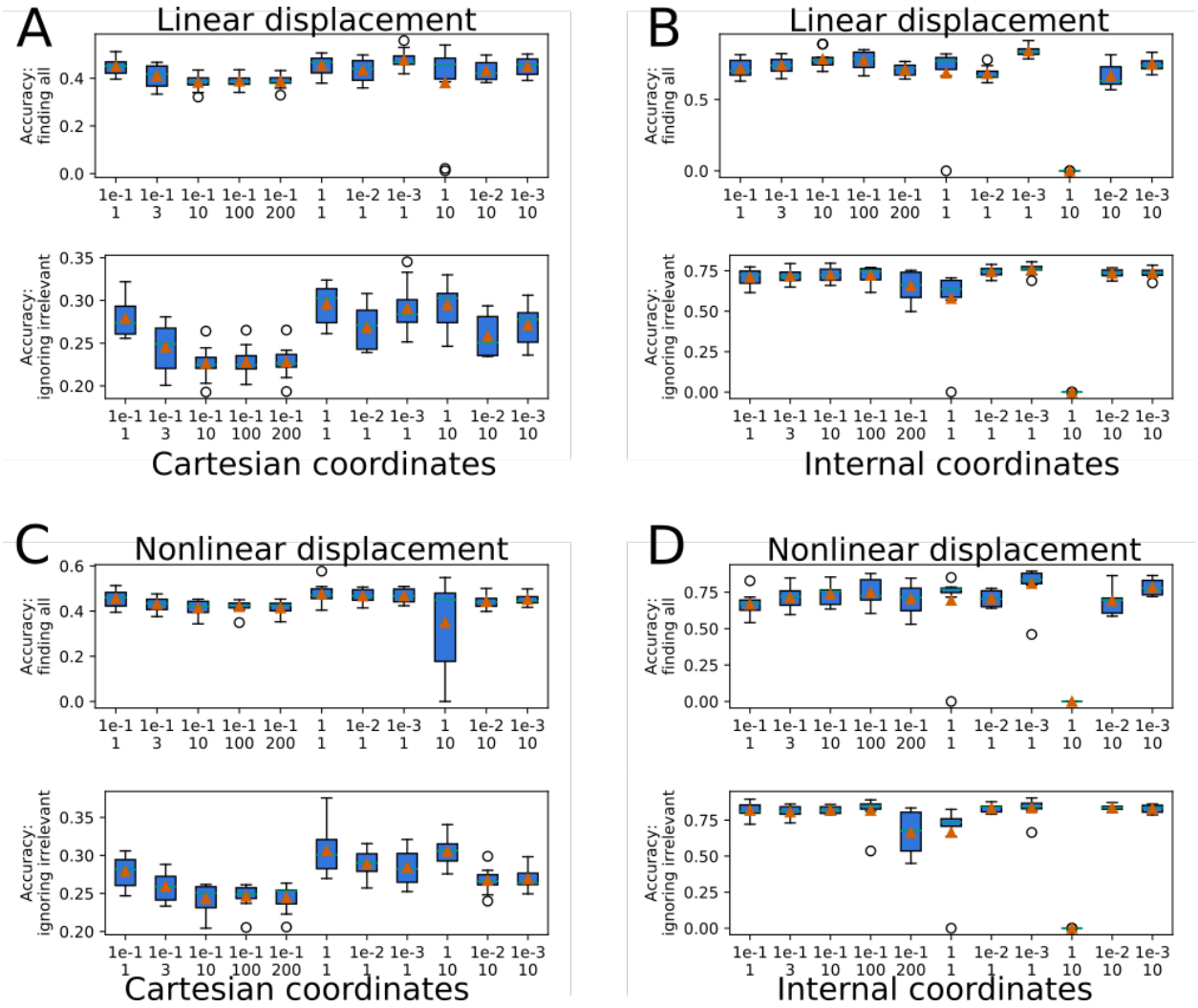


Figure S6. Hyperparameter tuning of RBM applied on the toy model for different learning rates (first row on x-axis) and number of nodes in the hidden layer (second row on x-axis). (a) Using cartesian coordinates and linear displacements, (b) using inverse distances and linear displacements, (c) using cartesian coordinates and nonlinear displacements, and (d) using inverse distances and nonlinear displacements.

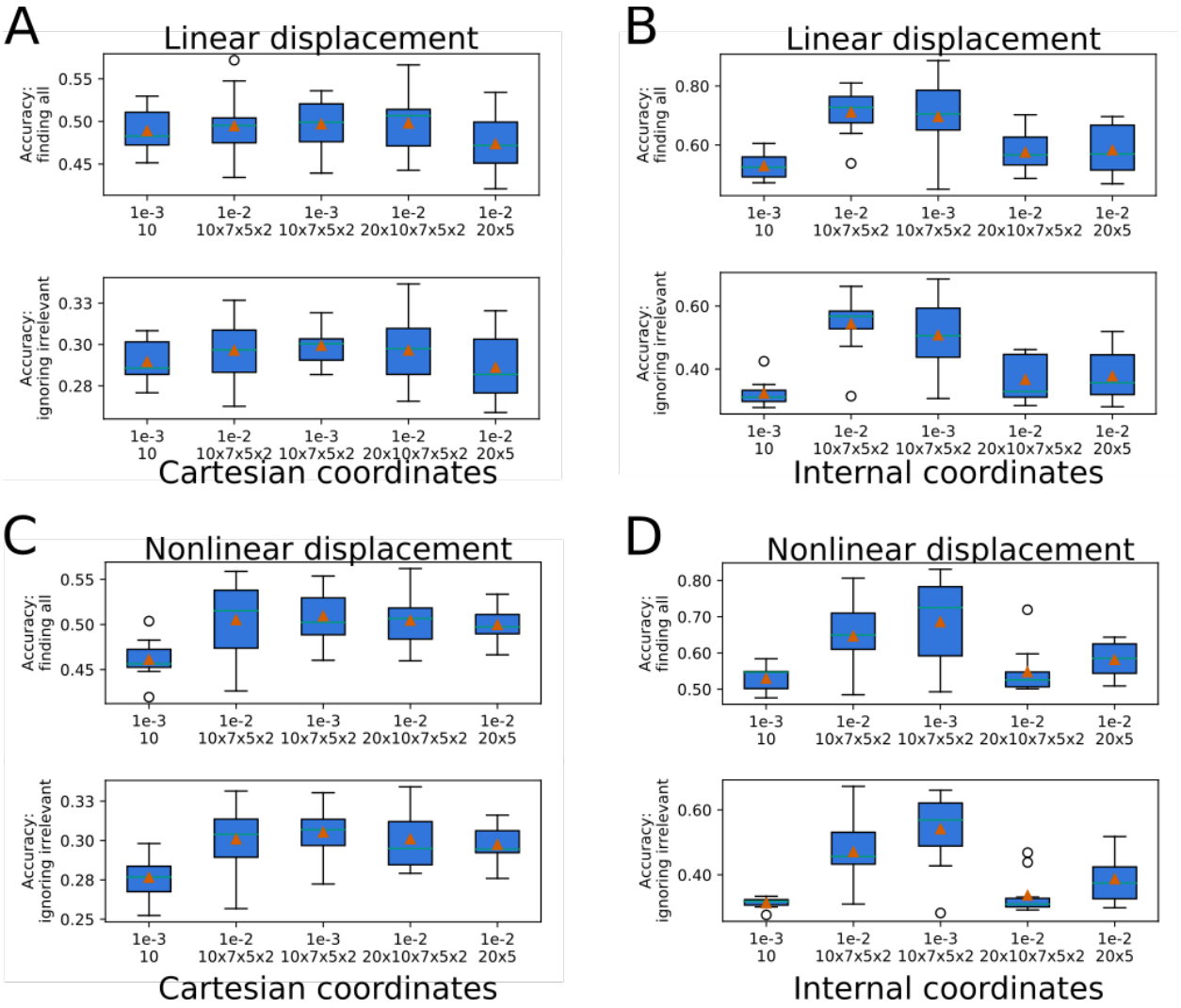


Figure S7. Hyperparameter tuning of AE applied on the toy model for different combinations of regularization parameters (first row on x-axis) and hidden layer architectures (second row on x-axis). (a) using cartesian coordinates and linear displacements. (b) using inverse distances and linear displacements. (c) using cartesian coordinates and nonlinear displacements. (d) using inverse distances and nonlinear displacements.

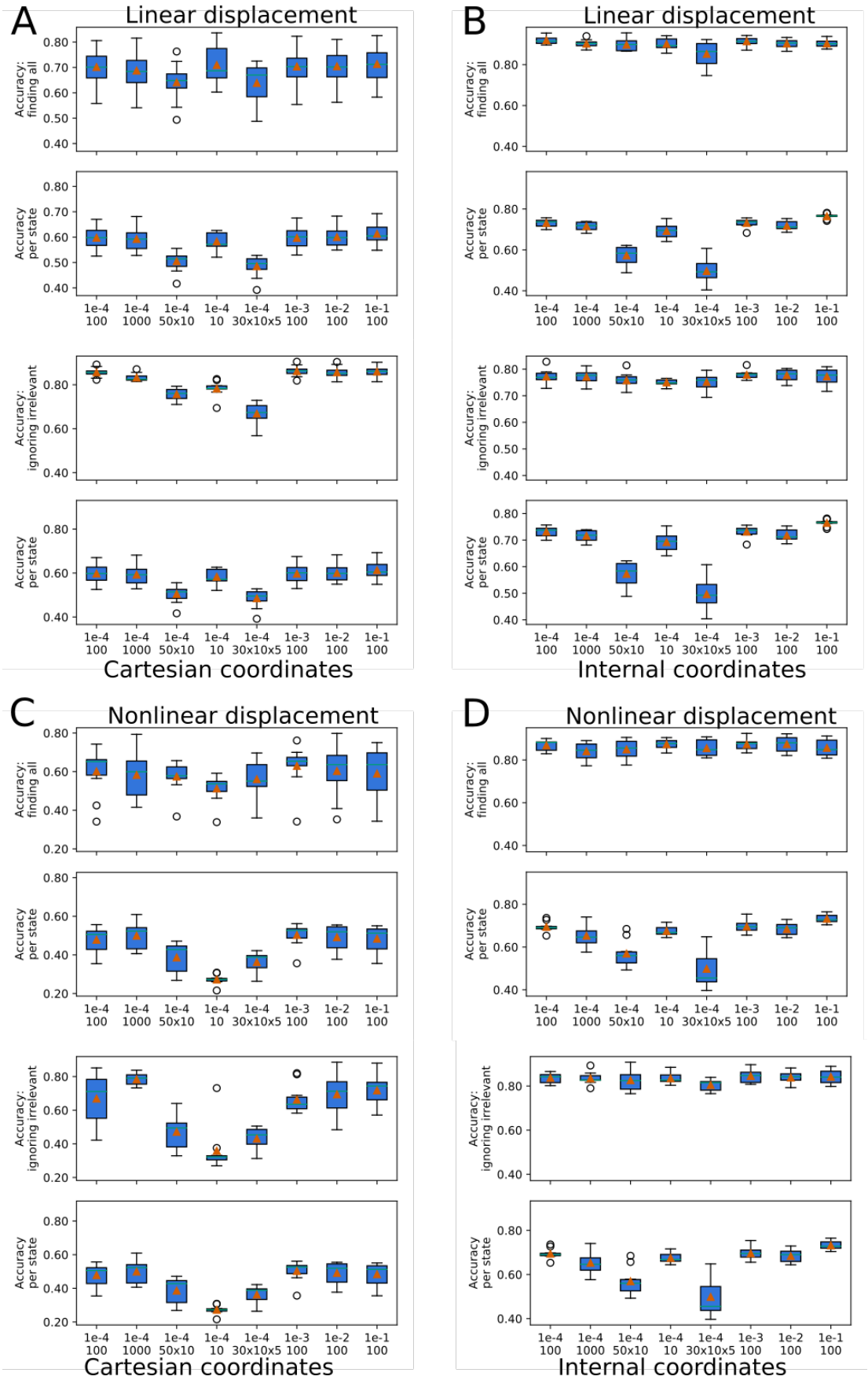

A
Linear displacement
B
Linear displacement
Accuracy: finding all
Accuracy per state
Accuracy: ignoring irrelevant
Accuracy per state
1e-4 100
1e-4 1000
1e-4 50x10
1e-4 10
1e-4 30x10x5
1e-3 100
1e-2 100
1e-1 100
Cartesian coordinates
Internal coordinates
C
Nonlinear displacement
D
Nonlinear displacement
Cartesian coordinates
Internal coordinates

Figure S8. Hyperparameter tuning of MLP applied on the toy model for different combinations of regularization parameters, α, (first row on x-axis) and hidden layer architectures (second row on x-axis). (a) Using cartesian coordinates and linear displacements, (b) using inverse distances and linear displacements, (c) using cartesian coordinates and nonlinear displacements, and (d) using inverse distances and nonlinear displacements. The accuracy at identifying all displaced atoms in all states as well as the accuracy at identifying displaced atoms in a specific state is shown.

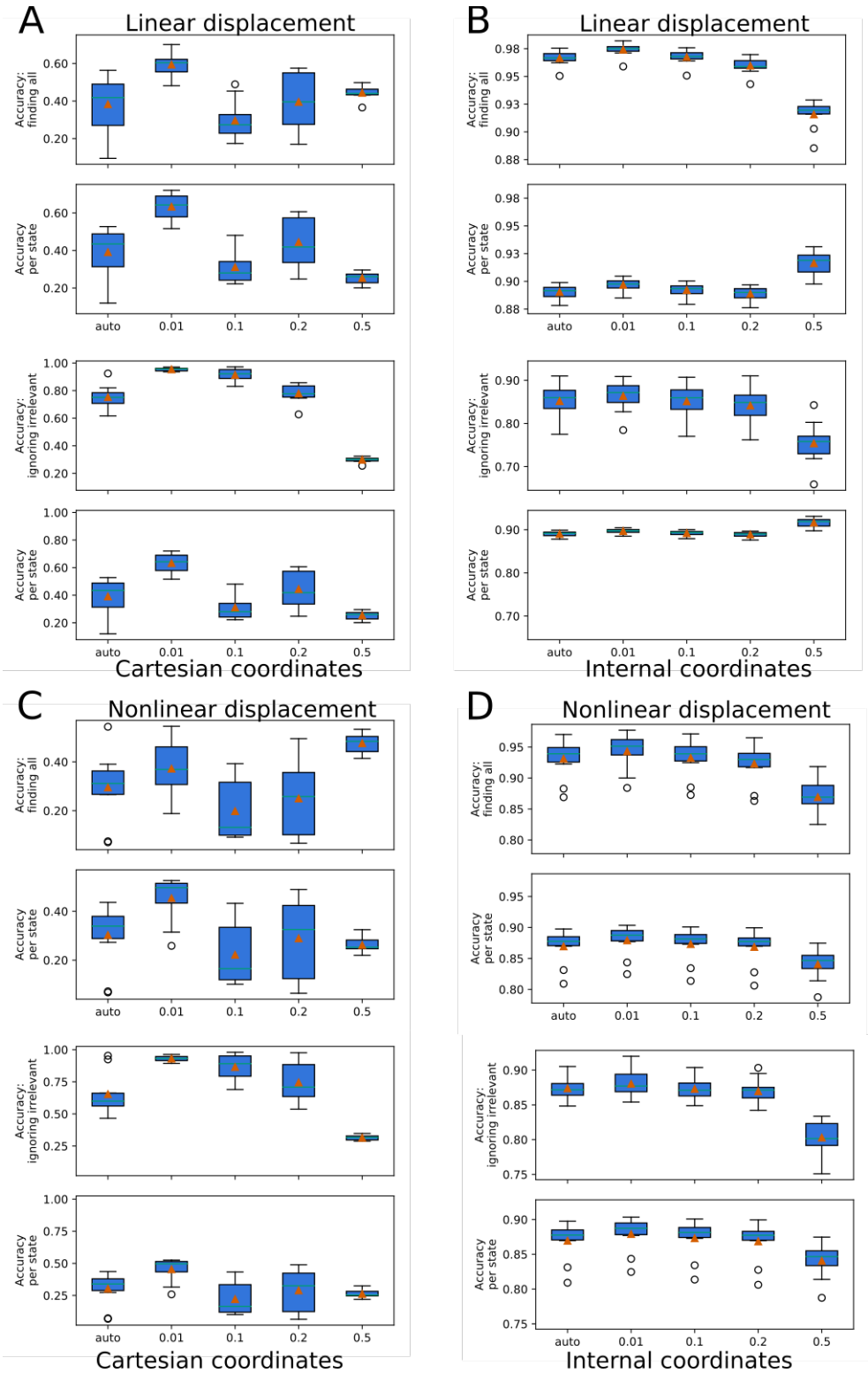


Figure S9. Hyperparameter tuning of KL divergence applied on the toy model for different bin sizes used for discretization. (a) Using cartesian coordinates and linear displacements, (b) using inverse distances and linear displacements, (c) using cartesian coordinates and nonlinear displacements and

(d) using inverse distances and nonlinear displacements. The accuracy at identifying all displaced atoms in all states as well as the accuracy at identifying displaced atoms in a specific state is shown.

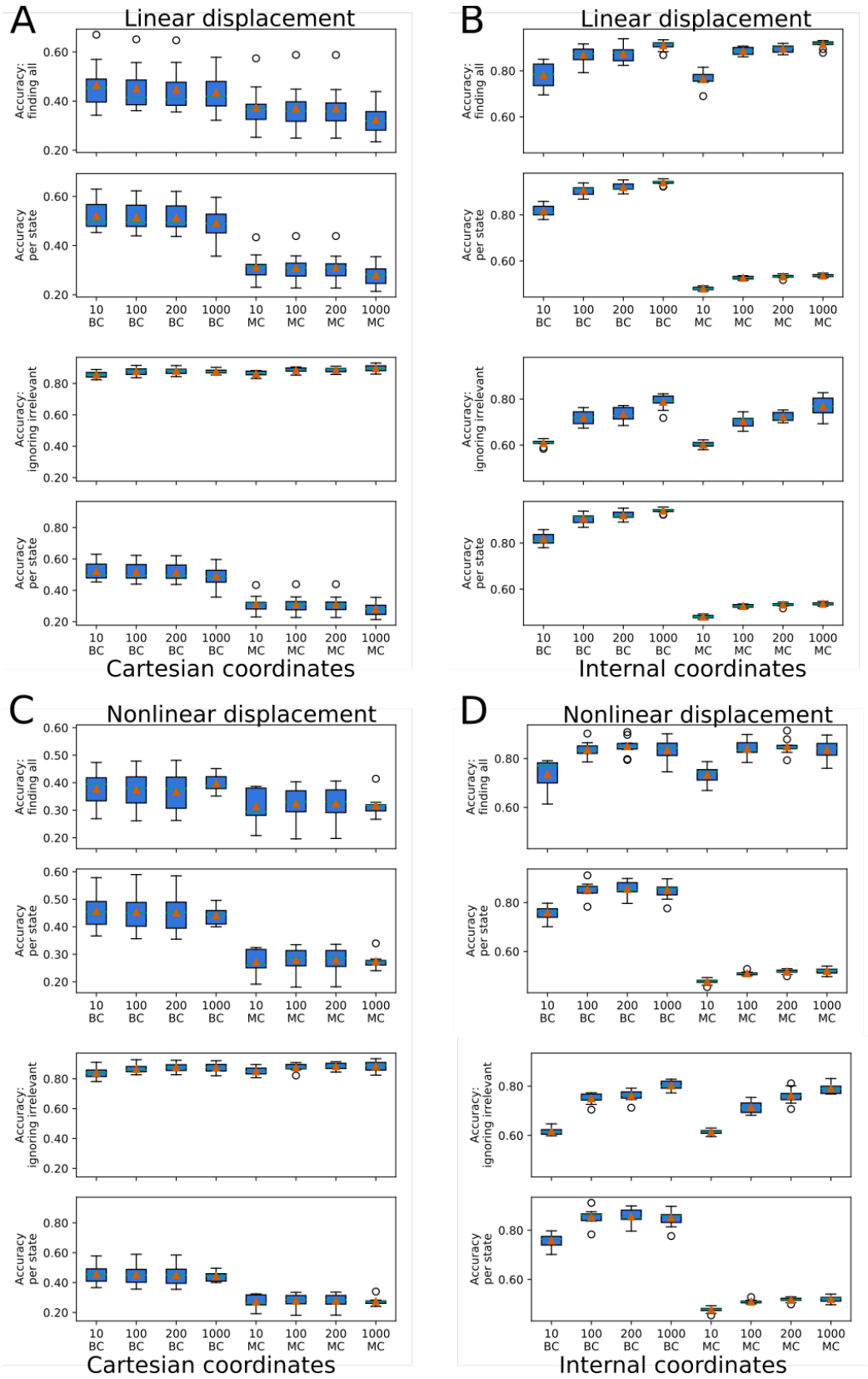

A
Linear displacement
B
Linear displacement
Accuracy: finding all
Accuracy per state
Accuracy: ignoring irrelevant
Accuracy per state
10 BC
100 BC
200 BC
1000 BC
10 MC
100 MC
200 MC
1000 MC
Cartesian coordinates
Internal coordinates
C
Nonlinear displacement
D
Nonlinear displacement
Cartesian coordinates
Internal coordinates

Figure S10. Hyperparameter tuning of RF applied on the toy model for different number of estimators in the model (first row on x-axis) and if it was trained as a single multiclass classifier (MC) or a set of binary classifiers (BC; second row on x-axis). (a) Using cartesian coordinates and linear displacements, (b) using inverse distances and linear displacements, (c) using cartesian coordinates and nonlinear displacements, and (d) using inverse distances and nonlinear displacements. The accuracy at identifying all displaced atoms in all states as well as the accuracy at identifying displaced atoms in a specific state is shown.