Learning Important Features from Molecular Simulations

OLIVER FLEETWOOD, MARINA KASIMOVA, ANNIE WESTERLUND & LUCIE DELEMOTTE

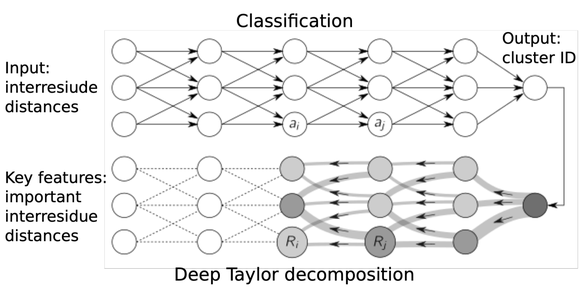
Abstract: Biomolecular simulations are intrinsically high dimensional and generate datasets of ever increasing size. Reducing the number of features in the dataset and gaining insight into the biophysical properties of molecular states is currently a big challenge that many scientists face on a regular basis. Following the recent years' rising interest in machine learning methods there are now many powerful dimensionality reduction tools easily accessible, although such methods are often criticized to resemble black boxes and provide limited human-interpretable insight. In this study we demonstrate how a number of methods from supervised as well as unsupervised machine learning can learn ensemble properties from molecular simulations and provide easily interpretable metrics of what features are actually important. In order to show which methods perform best under different circumstances we first test the performance using a toy model designed to mimic real macromolecular behavior. Finally, we apply the methods to simulations of the β_2 adrenergic receptor to gain insights into its activation mechanism and the effect of ligand binding. The results demonstrate how machine learning methods can produce valuable insights into properties of biomolecular states and we anticipate that our approach can be useful to aid many researchers in demystifying complex simulations.

Methods for classification and feature extraction

Supervised Learning

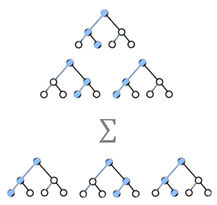
Multilayer Perceptron(MLP)

- Feedforward artificial neural network with fully connected layers.
- Implementation from sklearn [1] with a stochastic gradient descent solver.
- Important features can be derived using Deep Taylor Decomposition [2], a method originally derived for image classification.



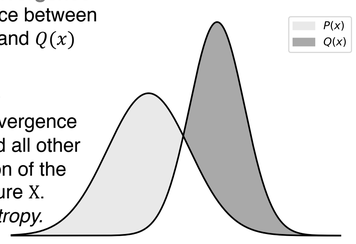
Random Forest(RF)

- Prediction is taken as the ensemble average of many decision trees.
- Average relevance per feature is computed as the mean decrease impurity.



Kullback–Leibler (KL) divergence

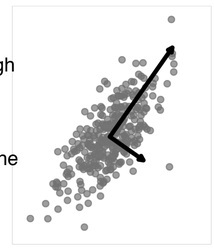
- Computes the difference between two distributions $P(x)$ and $Q(x)$ along a feature X as $-\sum_{x \in X} P(x) \log\left(\frac{Q(x)}{P(x)}\right)$.
- We compute the KL divergence between one class and all other classes as an indication of the importance of the feature X.
- Also called *relative entropy.*



Unsupervised Learning

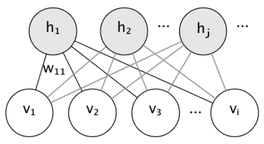
Principal Component Analysis (PCA)

- Converts the input features to an orthogonal set of linearly uncorrelated variables (principal components) through an orthogonal transformation.
- The first component has 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.



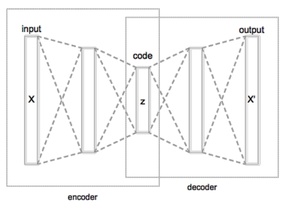
Restricted Boltzmann Machine (RBM)

- 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 are identified using Deep Taylor Decomposition.



Autoencoder (AE)

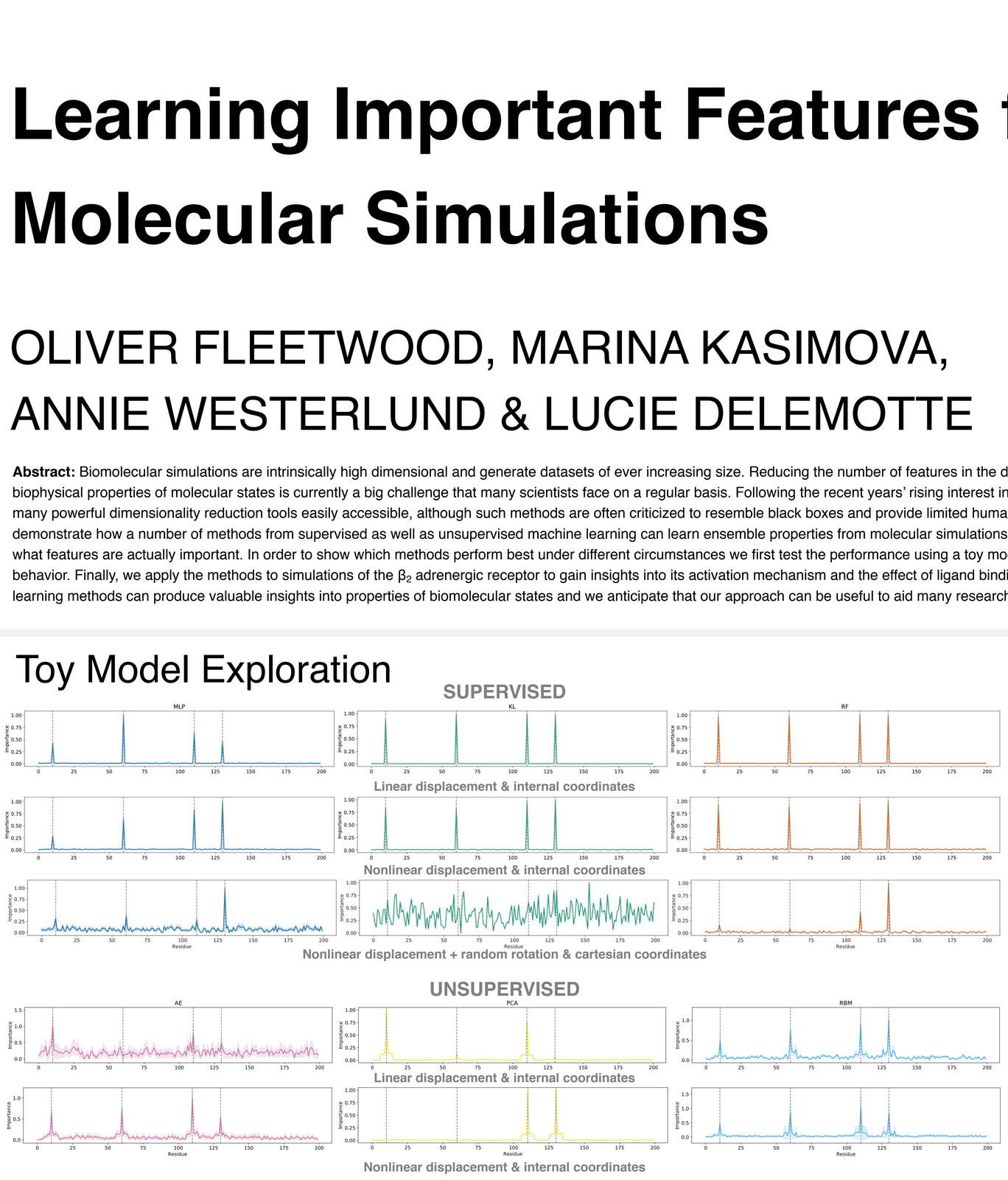
- A generative neural network trained to reconstruct the features through a set of hidden layers of lower dimensionality.
- Important features are identified using Deep Taylor Decomposition.



References

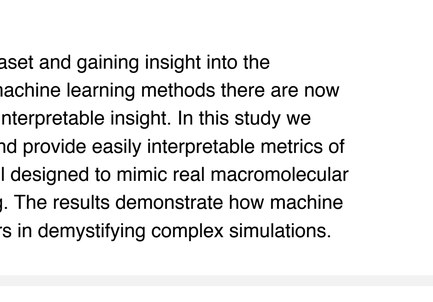
1. Scikit-learn: Machine Learning in Python, Pedregosa et al., *JMLR* 12, pp. 2825-2830, 2011.
2. G. Montavon, S. Lapuschkin, A. Binder, W. Samek, K.-R. Müller, "Explaining NonLinear Classification Decisions with Deep Taylor Decomposition", *Pattern Recognition*, 65:211–222, 2017.
3. Westerlund, Annie M., and Lucie Delemotte. "Effect of Ca2+ on the promiscuous target-protein binding of calmodulin." *PLoS computational biology* 14.4 (2018): e1006072.
4. Dror R.O., Arlow D.H., Maragakis P., Mildorf T.J., Pan A.C., Xu H., Borhani D.W., Shaw D.E., "Activation mechanism of the beta2-adrenergic receptor", *Proc. Natl. Acad. Sci. USA.* 2011;108(46).
5. "File:Autoencoder structure.png." Wikimedia Common

Toy Model Exploration



Linear displacement

Nonlinear displacement

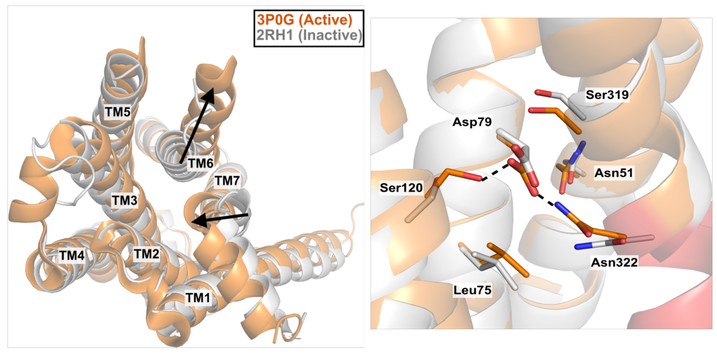


Nonlinear displacement + random rotation

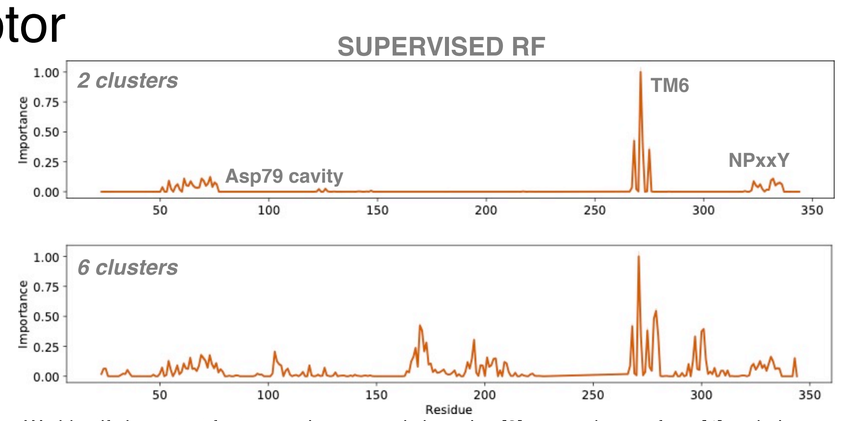
Learnings

- Simply computing the KL divergence may perform well as long as the features are well chosen.
- When using cartesian coordinates, KL performs poorly.
- More sophisticated machine learning methods (neural networks, random forests etc.) can successfully learn the important features without being as sensitive to the choice of coordinates.
- Regarding the unsupervised learning techniques, RBMs and autoencoders can pick up even nonlinear features.
- PCA on the other hand, typically performs well on identifying linear displacements.

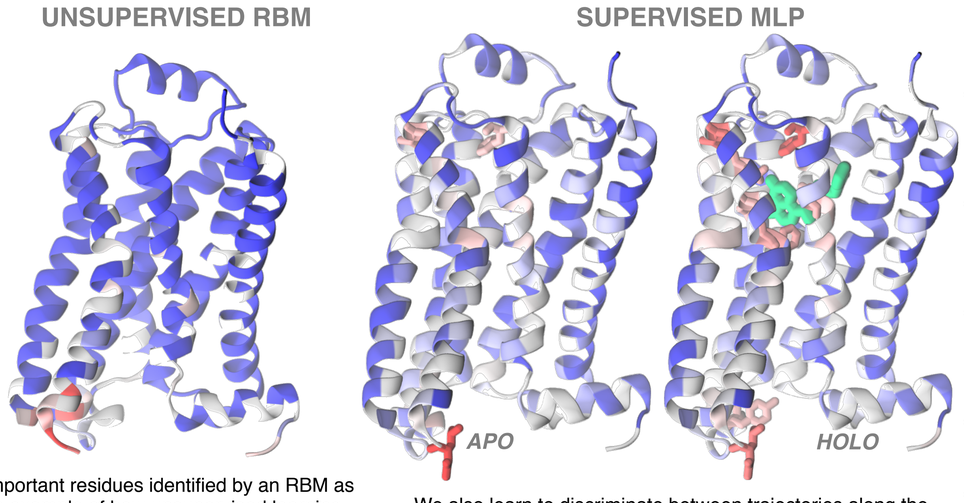
Applications to the β_2 adrenergic receptor



Hallmarks of β_2 activation include the outward movement of transmembrane helix 6 (TM6), the collapse of the cavity around the highly conserved residue Asp79$^{2.50}$, as well as the twist of the conserved NPxxY motif at the bottom of TM7.



We identify important features using spectral clustering [3] on a trajectory from [4] to derive a number of clusters along the activation path and see which features are important for cluster classification. Similar profiles are obtained for MLP and KL.



Important residues identified by an RBM as an example of how unsupervised learning can help demystify molecular simulations. Least important residues are highlighted in blue and most important residues in red.

We also learn to discriminate between trajectories along the activation path with (holo) and without (apo) an agonist ligand bound. Least important residues for classification are highlighted in blue and most important residues in red.

Learnings

- All methods identify that the G-protein binding site undergoes the most significant conformational change, especially TM6.
- The importance profiles change upon increase of the number of clusters, revealing more features important for activation.
- Regarding the effect of ligand binding we see that there are important residues in the extracellular region close to the ligand binding site.
- The corresponding input features are more strongly activated when a ligand is present.
- There is also a strong signal in the G-protein binding site, which illustrates how it is allosterically coupled to the ligand binding site.

KTH ROYAL INSTITUTE OF TECHNOLOGY

Oliver Fleetwood
PhD Student
Delemottelab

Science for Life Laboratory
Stockholm
Sweden

Tel: +46 - 708613650
Email: oliver.fleetwood@scilifelab.se