Title: Predicting New Protein Conformations from Molecular Dynamics Simulation Conformational Landscapes and Machine Learning

Running Title: Machine Learning on MD 2D-RMSD Landscapes

Yiming Jin,[a,b] Linus O. Johannissen[a,]*, Sam Hay[a,]*

[a] Manchester Institute of Biotechnology and Department of Chemistry, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK
[b] School of Computer Science and Engineering, Central South University, Changsha, 410000, China

Linus.Johannissen@manchester.ac.uk
Sam.Hay@manchester.ac.uk

Acknowledgements

This work was supported in part by BBSRC grant: BB/M017702/1. The authors acknowledge the assistance given by IT Services and the use of the Computational Shared Facility at The University of Manchester.

Abstract

Molecular dynamics (MD) simulations are a popular method of studying protein structure and function, but are unable to reliably sample all relevant conformational space in reasonable computational timescales. A range of enhanced sampling methods are available that can improve conformational sampling, but these do not offer a complete solution. We present here a proof-of-principle method of combining MD simulation with machine learning to explore protein conformational space. An autoencoder is used to map snapshots from MD simulations onto the conformational landscape defined by a 2D-RMSD matrix, and we show that we can predict, with useful accuracy, conformations that are not present in the training data. This method offers a new approach to the prediction of new low energy/physically realistic structures of conformationally dynamic proteins and allows an alternative approach to enhanced sampling of MD simulations.

Key words: Molecular Dynamics; Conformational Landscape; Autoencoder; Protein; Calmodulin

Introduction

Molecular dynamics (MD) simulations of proteins are a popular method of studying aspects of protein function and dynamics.[1] They require input structure(s), which are preferably experimentally determined, usually by X-ray crystallography. However, as proteins are often highly flexible, they adopt multiple conformations which interconvert over a wide range of timescales,[2, 3] which can be predominantly longer than the feasible MD simulation length of ns-μs. Enhanced sampling methods have been developed to improve the sampling of MD simulations,[4, 5] but these do not offer a complete solution to the MD sampling problem, partly because some knowledge of the system is necessary to define the coordinates (e.g. collective variables) along which sampling should be performed. Machine learning offers an alternative approach.

Machine learning (ML) has been successfully applied to the analysis of the high-dimensional data produced by MD simulations[6] and in structure prediction where an experimentally derived structure or homology model is not available.[7] Enhanced sampling techniques that use ML to guide the MD simulations (*e.g.* by identifying collective variables and imposing biasing potentials) have also been developed;[8-14] a conceptually simpler and more flexible approach is to utilise ML for the prediction of new protein conformations based on existing MD simulations, as has been recently demonstrated. This approach has recently been demonstrated using an autoencoder to encode the structural data into a low-dimensional representation, either onto the autoencoder's default latent vector[15] or using the sketch-map algorithm[16] to improve the interpretability of the low-dimensional representation.[17, 18] New structures were then predicted by decoding points on the resulting low-dimensional surface.

Here we employ a related but different approach, to use a simple, pre-defined low-dimensional conformational landscape to guide the search rather than use the machine learning algorithm define the low-dimensional representation. The aim is not to create a more robust machine learning algorithm than those discussed above, but to explore whether a very simple representation of a MD-derived conformational landscape can successfully be used to predict new, physically plausible conformations. In principle, this approach could then be used with an arbitrary representation of the conformational landscape, which can consist of structural parameters of choice such as contact matrix, backbone dihedrals (as used in ref [17, 18]) or a combination of specific parameters. For this proof-of-concept study, an autoencoder was trained to map the structures onto a simple conformational landscape, namely the first two principal components of a 2D-RMSD matrix (the matrix of RMSD values for every structure relative to every other structure), and trained to decode points along this landscape into new structures. Two test cases are used, a short homoalanine peptide and the calcium-binding protein calmodulin (CaM). We show that it is possible to predict low energy and thus physically plausible conformations which were not sampled during the MD simulation(s).

Computational Methods.

Molecular Dynamics Simulations. All simulations were performed in Gromacs 2016.4[19] using the Amber FF14SB[20] force field. Each system was solvated with a water box at least 13

Å larger than the peptide/protein on each axis with counter-ions (if required) generated in AmberTools 16[21]. All calculations used a periodic boundary condition and LINCS constraints on all bonds involving hydrogen atoms, the Verlet cut-off scheme with 10 Å cutoffs. Energy minimisation was followed by 100 ps of constant volume and 100 ps of constant pressure solvent equilibration, using the Parrinello-Rahman pressure coupling with a time constant of 2 ps, and positional restraints with a force constant of 10 kJ mol^{-1} nm^{-2} applied to the protein/peptide. Constant pressure was also used for the subsequent unrestrained production run.

Conformational Landscape. Our machine learning algorithm takes a conformational landscape in the form of a series of vectors, as input. A simple conformational landscape was defined based on the 2D-RMSD matrix: the square matrix where each cell is the pairwise RMSD between structures (and the diagonal elements are therefore 0). For m total structures, the 2D-RMSD matrix is an $m \times m$ matrix, and principal components analysis the results in m eigenvectors, or principal components (PCs). We then used the top two PCs (those with the largest eigenvalues) to define a 2-dimensional conformational landscape, although the input is not limited to 2-dimensional vectors, so a more complex, multidimensional landscape can be used by using additional PCs.

Machine Learning. The protein structures (Cartesian coordinates) were extracted from MD simulations using the MDanalysis package.[22, 23] The structures were first aligned to the starting structure by minimising the RMSD for the same atom selection subsequently used in the ML, and the Cartesian coordinates were normalised using MinMax scaling. For the complete set of protein structures with coordinates $\left((x_1^1, y_1^1, z_1^1) \ldots (x_n^m, y_n^m, z_n^m)\right)$, where n is the number of atoms per structure and m is the total number of structures, the normalised coordinates are for atom i n structure k are given by:

$$x_i^k \mapsto \frac{x_i^k - \min(x)}{\max(x) - \min(x)} \tag{1}$$

Note that we also tried using z-score normalisation, which is suitable for Gaussian distributions, but this performed poorly (in terms of the final predictions) as the 2D-RMSD matrix projected onto principal components eigenvectors are not normally distributed.

Our modified autoencoder was built in Python 3.6 using Keras (https://keras.io/), an open-source deep learning library with a Tensorflow[24, 25] backend and the code is available at https://github.com/Imay-King/MDMachineLearning. The approach is illustrated in Figure 1. First, the algorithm encodes the normalised protein structure $((x_1,y_1,z_1), (x_2,y_2,z_2) \ldots (x_n,y_n,z_n))$ onto the first two PCs from the 2D-RMSD matrix using a constraint in the training process in the form of a mean square error (MSE) loss between the latent vector and pairs of PCs. A second MSE loss function is then used to train the decoder to predict a new protein structure $((\hat{x}_1,\hat{y}_1,\hat{z}_1), (\hat{x}_2,\hat{y}_2,\hat{z}_2) \ldots (\hat{x}_n,\hat{y}_n,\hat{z}_n))$ from the PC values. After the decoder has created a new protein structure, the coordinates are de-normalised using the inverse of the MinMax method (eq. 1).

Note that this approach is not limited to the first two PCs from the 2D-RMSD PCA: other dimensionality reduction techniques could be employed, and the conformational landscape could be defined using any set of structural parameters of interest. We simply chose the PCA approach here as it is a familiar method of analysing protein MD data.

Results and Discussion

Model 1. As a proof of concept, we first attempted to predict structures from a 10 ns MD simulation of the simple peptide *L*-Ala$_{13}$, with snapshots taken every 2 ps for a total of 5000 structures. This is a highly flexible peptide, with folding and unfolding of an α-helical structure observed during the simulation. A maximum heavy-atom RMSD of 7.79 Å was observed between any two structures (*i.e.* from the 2D-RMSD matrix), and a maximum RMSD of 4.85 Å relative to the average structure (SI Figure S1). Since this is a relatively small system, we included all non-hydrogen atoms in the ML (66 atoms, 198 features per conformation). From an 80/20 training/testing split of the data, using a 3-layer model we found that the best results were obtained with a combination of the Adam optimizer[26] and the ReLU activation function[27] for all layers except the last layer of the encoder and the first layer of the decoder, for which the sigmoid activation function was used instead. This gave a model that converged reasonably quickly with very similar performance for the training and testing sets: the loss (SI Figure S2) converged to 5.5×10^{-3} for the training set and 5.7×10^{-3} for the testing set. The average RMSD (± 1 standard deviation) between the predicted and target structure for the 1000 structures in the testing set is 0.73 ± 0.41 Å.

To further test whether this approach can successfully predict structures that are distinct from those in the training set, we repeated the predictions for 7 structures in different regions of the PC1/PC2 plot (Figure 1), again using an 80/20 split, but each time excluding any structures within ±0.002 along PC1 (34% of the variance) or ±0.003 along PC2 (17% of the variance) from the training set. The average RMSD observed for these predicted structures is 1.20 ± 0.31 Å, compared to 1.00 ± 0.49 Å without any exclusions. The structural features of each conformation are predicted successfully (Figure 2) and this simple example therefore demonstrates the feasibility of this approach to predicting protein secondary and tertiary structural elements from an MD simulation.

Model 2. As a more biochemically relevant example we turned to CaM, which is known to adopt several distinct conformational states.[28, 29] We chose yeast CaM in a compact target peptide- and Ca^{2+}-bound form (PDB ID: 2LHI) and a less compact Ca^{2+}-bound form (2LHH) as the starting points for two MD simulations. Both simulations were carried out without Ca^{2+} or target peptide to encourage significant conformational change during the simulation. Since this is a much larger system than the *L*-Ala$_{13}$ peptide, we only used the backbone atoms for ML and analysis (585 atoms, 1755 data points per protein structure). We ran two 10 ns MD simulation starting from each crystal structure, with snapshots taken every 5 ps for a total of 4000 structures. As can be seen from the PCA and RMSD plots (Figures 3 and S3), the two simulations converged to different conformations and the conformational space sampled in each simulation does not overlap. Here, the 10 ns simulations do not allow sufficient sampling of the CaM conformational landscape, which highlights the sampling problem with MD simulation of proteins. The maximum RMSD between any two structures across both simulations (from the 2D-RMSD matrix) is 17.8 Å and the maximum RMSD relative to the average structure is 13.6 Å. Using the same 3-layer autoencoder as for Model 1, with an 80/20 training/testing split, the loss converged to 3.44×10^{-3} for the training set and 3.39×10^{-3} for the testing set, and for the 800 structures in the testing set the average RMSD between the predicted structure and the target was 0.90 ± 0.71 Å, which is similar to that observed for the L-Ala$_{13}$ peptide. We also tested the effect of different numbers of layers in the autoencoder (SI Figure S4), with very similar results, although the loss convergence was significantly less smooth with 5 layers.

As before, we then tested whether our algorithm can predict structures that are distinct from those in the training set by repeating the prediction for 7 structures in different regions of the PCA plot (Figure 3 **a-g**). For each of the predicted structures, the training set consisted of the MD simulation from which that particular target structure did not originate; *i.e.* when predicting structures taken from the MD1 simulation the model was trained only on structures from MD2, and *vice versa*. We again experimented with the effect of different numbers of layers in the autoencoder, and found that overall the 3-layer model performed best (SI Figure S5). For the 7 predicted structures, the average RMSD relative to the target structures is 1.89 ± 0.81 Å, and even for the worst predictions (**b** and **f**) the overall gross structural features were successfully predicted.

The target CaM structures in Figure 3 are compared to the most similar structure from the training set in Table 1. The predicted structures have conformations that are not found in the training set, and in each case the RMSD to the target structure is smaller than the minimum RMSD to the structures in the training set. The two structures with the biggest improvement (**a** and **e**) are shown in Figure 4, which illustrates the ability of the autoencoder to predict structures that lie significantly outside of the range of the training set. Further, the 7 target structures span a range of physiologically-relevant ‘open’ and ‘closed’ conformational states that interconvert via a relatively complex series of domain rotations and formation/breaking of the central α-helix. It is perhaps then surprising that it is possible to describe this conformational space in only two PCs of a PCA analysis. For larger proteins this may not be sufficient, but our method is extensible to an arbitrary number of PCs, which would allow more complex conformational space to be mapped in higher dimensions.

Model 2 does not include side chain residues in the autoencoder, which is more computationally efficient and also forces the PCA to describe gross tertiary structure/conformational space without the added complication of multiple side chain conformations. It is possible to include side chains (as was the case in Model 1), but again this will likely require the use of more PCs to adequately describe the conformational space of interest. Alternatively, the predicted structures could be used to generate new MD simulation input geometries by building in the sidechains using e.g. rotamer libraries,[30-33] through partial structural alignments with the original MD simulation data, or by using techniques such as steered MD[34, 35] to rapidly drive the MD simulation to new predicted conformations.

We chose to rebuild the sidechains of the predicted CaM structures using the protein sidechain prediction algorithm in SCWRL4.[36] To benchmark this approach, we initially rebuilt the sidechains for structures **a-g** in Figure 3, which resulted in an average RMSD between the rebuilt and original structures of 2.99 ± 0.02 Å (SI Table S1). However, since structures taken from an MD simulation are often high-energy structures with non-optimal sidechain-sidechain interactions (at 300 K the majority of conformations do not sit at the bottom of the potential energy well) that SCWRL4 is not designed to reproduce, we then energy minimised the sidechains of both the original and rebuild structures using the Amber14 force field in Amber using implicit solvation (5000 steps of steepest descent with a harmonic constraint of 500 kcal mol^{-1} A^{-2} on the backbone atoms). This decreased the average RMSD to 1.25 ± 0.80 Å, suggesting that this approach is able to rebuild the sidechains and generate structures that are physically realistic, with a strong correspondence between the original and rebuilt structures.

Next, we used our ML approach to predict 99 structures regularly sampled over the PC1/PC2 plot of the CaM MD simulations (Figure 5), this time combining MD1 and MD2 for training. The sidechains of the predicted structures were again built using SCWRL4 and energy minimised. 29 of the structures could not be energy minimised due to internal strain or clashes, but these lie outside of the conformational landscape covered by the MD simulations (white dots in Figure 5). For example, the gap between the two clusters from MD2 (MD2A and MDB in Figure 5) suggests that two regions of relatively stable conformations are separated by a high-energy, low-population region, and this coincides with a highly strained structure which could not be successfully energy minimised. Additionally, the starting structure for MD2 lies in the MD2A cluster (see Figure 3), but it converges to the lower-right portion of MD2B suggesting that MD2B is a more stable area of the conformational landscape, which is consistent with the cluster of low-energy structures around the lower-right portion of MD2B. Inspection of the 70 structures that were successfully energy minimised shows that the lower energy (more stable) structures tend to overlap with region of the PC1/PC2 conformation space that is sampled during one of the MD simulations (Figure 5). This demonstrates that this approach can successfully distinguish between relatively low energy and higher energy conformations. Predicted structures with relatively low energies (light and dark blue dots in Figure 5) that lie outside of regions of conformational space sampled during the MD simulations may provide interesting new conformations that could be used as input structures for additional simulations, thus enhancing the total conformational space sampled.

In summary, we have demonstrated a proof-of-principle method of combining MD simulation with machine learning to explore protein conformational space. An autoencoder is used to map snapshots from MD simulations onto the first two PCs from the 2D-RMSD matrix, and we show that we can predict, with useful accuracy, conformations that are not present in the training data. It is also straight-forward to assess the relative likelihood of each predicted conformation. This method offers a new approach to the prediction of new physically realistic structures of conformationally dynamic proteins and allows an alternative approach to enhanced sampling of MD simulations, by rapidly generating new structures from which additional, long-scale MD simulations can be initiated for a more efficient search through conformational space.

References

1. Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E., Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol* **2009,** *19* (2), 120-7.
2. Schuler, B.; Hofmann, H., Single-molecule spectroscopy of protein folding dynamics--expanding scope and timescales. *Curr Opin Struct Biol* **2013,** *23* (1), 36-47.
3. Henzler-Wildman, K.; Kern, D., Dynamic personalities of proteins. *Nature* **2007,** *450* (7172), 964-72.
4. Maximova, T.; Moffatt, R.; Ma, B.; Nussinov, R.; Shehu, A., Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput Biol* **2016,** *12* (4), e1004619.
5. Yang, Y. I.; Shao, Q.; Zhang, J.; Yang, L.; Gao, Y. Q., Enhanced sampling in molecular dynamics. *J Chem Phys* **2019,** *151* (7), 070902.
6. Fleetwood, O.; Kasimova, M. A.; Westerlund, A. M.; Delemotte, L., Molecular Insights from Conformational Ensembles via Machine Learning. *Biophys J* **2020,** *118* (3), 765-780.
7. Kandathil, S. M.; Greener, J. G.; Jones, D. T., Recent developments in deep learning applied to protein structure prediction. *Proteins* **2019,** *87* (12), 1179-1189.
8. Wang, Y.; Lamim Ribeiro, J. M.; Tiwary, P., Machine learning approaches for analyzing and enhancing molecular dynamics simulations. *Curr Opin Struct Biol* **2020,** *61*, 139-145.
9. Noe, F.; Olsson, S.; Kohler, J.; Wu, H., Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **2019,** *365* (6457).
10. Wu, H.; Mardt, A.; Pasquali, L.; Noe, F. In *Deep generative markov state models*, Advances in Neural Information Processing Systems, 2018; pp 3975-3984.
11. Ribeiro, J. M. L.; Bravo, P.; Wang, Y.; Tiwary, P., Reweighted autoencoded variational Bayes for enhanced sampling (RAVE). *J Chem Phys* **2018,** *149* (7), 072301.
12. Chen, W.; Ferguson, A. L., Molecular enhanced sampling with autoencoders: On-the-fly collective variable discovery and accelerated free energy landscape exploration. *J Comput Chem* **2018,** *39* (25), 2079-2102.
13. Bonati, L.; Zhang, Y. Y.; Parrinello, M., Neural networks-based variationally enhanced sampling. *Proc Natl Acad Sci U S A* **2019,** *116* (36), 17641-17647.
14. Shamsi, Z.; Cheng, K. J.; Shukla, D., Reinforcement Learning Based Adaptive Sampling: REAPing Rewards by Exploring Protein Conformational Landscapes. *J Phys Chem B* **2018,** *122* (35), 8386-8395.
15. Degiacomi, M. T., Coupling Molecular Dynamics and Deep Learning to Mine Protein Conformational Space. *Structure* **2019,** *27* (6), 1034-1040 e3.

16. Ceriotti, M.; Tribello, G. A.; Parrinello, M., Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc Natl Acad Sci U S A* **2011,** *108* (32), 13023-8.
17. Lemke, T.; Peter, C., EncoderMap: Dimensionality Reduction and Generation of Molecule Conformations. *J Chem Theory Comput* **2019,** *15* (2), 1209-1215.
18. Lemke, T.; Berg, A.; Jain, A.; Peter, C., EncoderMap(II): Visualizing Important Molecular Motions with Improved Generation of Protein Conformations. *J Chem Inf Model* **2019,** *59* (11), 4550-4560.
19. Van Der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., GROMACS: fast, flexible, and free. *J Comput Chem* **2005,** *26* (16), 1701-18.
20. Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C., ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J Chem Theory Comput* **2015,** *11* (8), 3696-713.
21. Case, D.; Betz, R.; Cerutti, D. S.; Cheatham, T.; Darden, T.; Duke, R.; Giese, T. J.; Gohlke, H.; Götz, A.; Homeyer, N.; Izadi, S.; Janowski, P.; Kaus, J.; Kovalenko, A.; Lee, T.-S.; LeGrand, S.; Li, P.; Lin, C.; Luchko, T.; Kollman, P. *Amber 16*, University of California: San Francisco, 2016.
22. Michaud-Agrawal, N.; Denning, E. J.; Woolf, T. B.; Beckstein, O., MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem* **2011,** *32* (10), 2319-27.
23. Gowers, R. J.; Linke, M.; Barnoud, J.; Reddy, T. J. E.; Melo, M. N.; Seyler, S. L.; Domanski, J.; Dotson, D.; Buchoux, S.; Kenney, I. M.; Beckstein, O., MDAnalysis: A Python Package for the Rapid Analysis of MolecularDynamics Simulations. In *Proceedings of the 15th Python in Science Conference*, Austin, Texas, United States, 2019.
24. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; Levenberg, J.; Monga, R.; Moore, S.; Murray, D. G.; Steiner, B.; Tucker, P.; Vasudevan, V.; Warden, P.; Wicke, M.; Yu, Y.; Zheng, X., TensorFlow: A system for large-scale machine learning. **2016**, arXiv:1605.08695.
25. Rampasek, L.; Goldenberg, A., TensorFlow: Biology's Gateway to Deep Learning? *Cell Syst* **2016,** *2* (1), 12-4.
26. Kingma, D. P.; Ba, J., Adam: A Method for Stochastic Optimization. **2014**, arXiv:1412.6980.
27. Nair, V.; Hinton, G. E., Rectified Linear Units Improve Restricted Boltzmann Machines. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010.
28. Chou, J. J.; Li, S.; Klee, C. B.; Bax, A., Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol* **2001,** *8* (11), 990-7.
29. Ogura, K.; Kumeta, H.; Takahasi, K.; Kobashigawa, Y.; Yoshida, R.; Itoh, H.; Yazawa, M.; Inagaki, F., Solution structures of yeast Saccharomyces cerevisiae

calmodulin in calcium- and target peptide-bound states reveal similarities and differences to vertebrate calmodulin. *Genes Cells* **2012,** *17* (3), 159-72.

30. Bhuyan, M. S.; Gao, X., A protein-dependent side-chain rotamer library. *BMC Bioinformatics* **2011,** *12 Suppl 14*, S10.
31. Francis-Lyon, P.; Koehl, P., Protein side-chain modeling with a protein-dependent optimized rotamer library. *Proteins* **2014,** *82* (9), 2000-17.
32. Towse, C. L.; Rysavy, S. J.; Vulovic, I. M.; Daggett, V., New Dynamic Rotamer Libraries: Data-Driven Analysis of Side-Chain Conformational Propensities. *Structure* **2016,** *24* (1), 187-199.
33. Wang, Q.; Canutescu, A. A.; Dunbrack, R. L., Jr., SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc* **2008,** *3* (12), 1832-47.
34. Izrailev, S.; Stepaniants, S.; Balsera, M.; Oono, Y.; Schulten, K., Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophys J* **1997,** *72* (4), 1568-81.
35. Leech, J.; Prins, J. F.; Hermans, J., SMD: Visual steering of molecular dynamics for protein design. *IEEE Computational Science and Engineering* **1996,** *3* (4), 38-45.
36. Krivov, G. G.; Shapovalov, M. V.; Dunbrack, R. L., Jr., Improved prediction of protein side-chain conformations with SCWRL4. *Proteins* **2009,** *77* (4), 778-95.

Table 1. RMSD (in Å) between target structures and the most similar structure from training set and predicted structure

Target Structure	Lowest RMSD from training set	RMSD to predicted structure
a	7.01	1.96
b	5.08	2.79
c	4.45	1.93
d	3.62	0.84
e	4.89	1.50
f	3.74	3.05
g	3.69	1.16

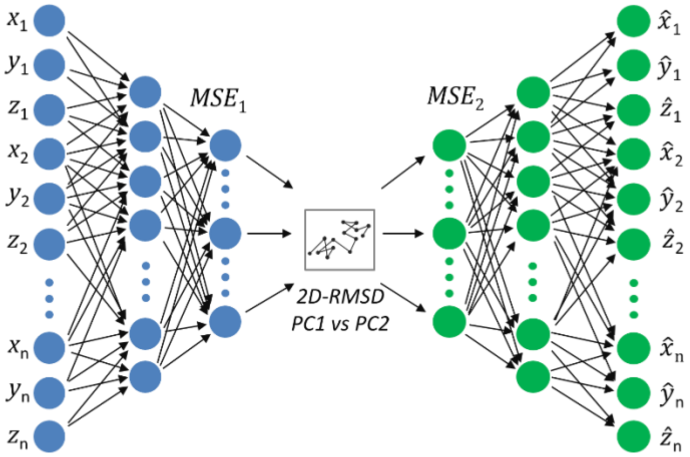

Figure 1. Modified autoencoder for prediction of protein structure from protein conformational landscape. The encoder (blue circles) is trained using a first loss function MSE_1 to reproduce the conformational landscape defined by the first two principal components (*PC1* and *PC2*) of the 2D-RMSD matrix, and the decoder (green) is trained using a second loss function MSE_2 to predict a protein structure from a pair of (*PC1*,*PC2*) values.

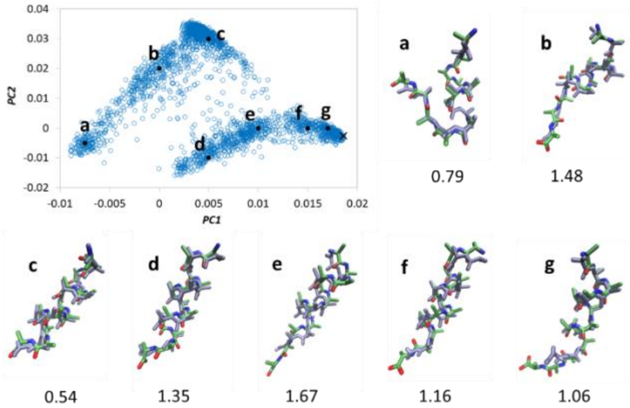

Figure 2. Structure predictions of the *L*-Ala13 peptide. Top left: *PC2* vs *PC1* plot from the 2D RMSD matrix (blue circles), with points a-g (black dots) used for testing. The black cross at PC1 ≈ 0.018 belongs to the initial, fully helical structure. For each prediction, points within (±0.002, ±0.003) of the (*PC1*,*PC2*) value were excluded from the training set. Overlays of the original structure (green, red and blue atoms) with the predicted structure (light blue) are shown for each point (**a-g**), with the RMSD in Å shown below.

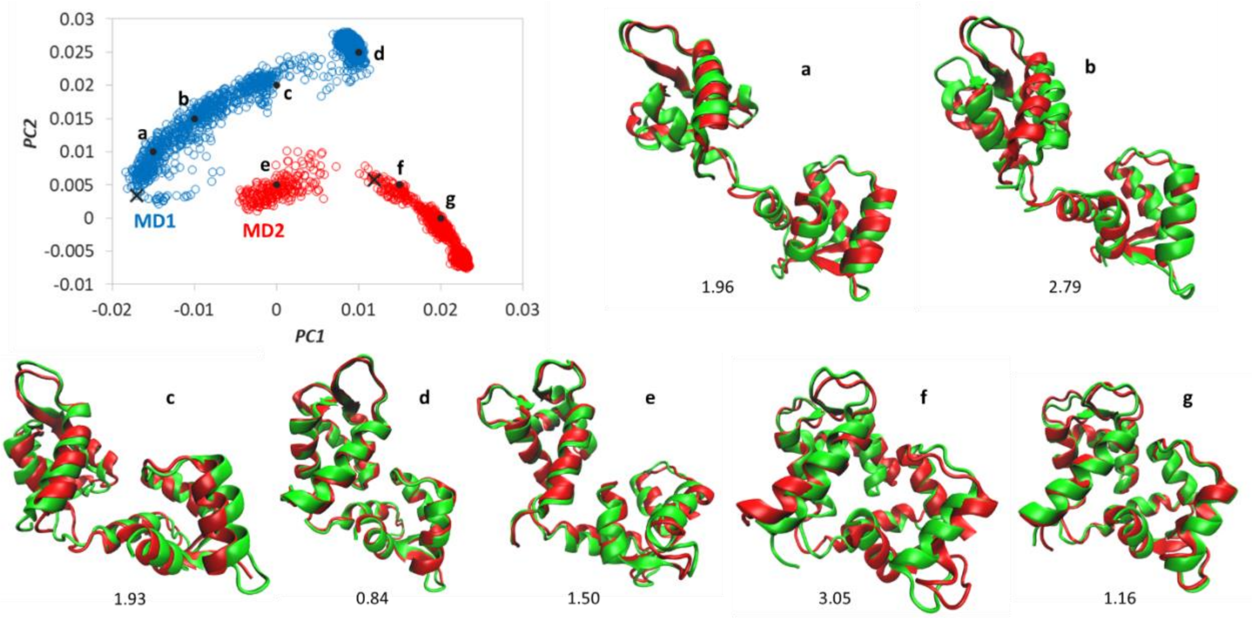

Figure 3. Structure predictions of CaM. Top left: *PC2* vs *PC1* plot from the 2D RMSD matrix constructed from two MD simulations (blue and red circles), with points **a-g** (black dots) used for testing. The black crosses belong to the starting structures for each simulation. For each structure, testing only included the MD simulation that the structure was not taken from (MD2 for **a-d**, MD1 for **e-g**). Overlays of the original structure (green) with the predicted structure (red) are shown for each point (**a-g**), with the RMSD in Å shown below.

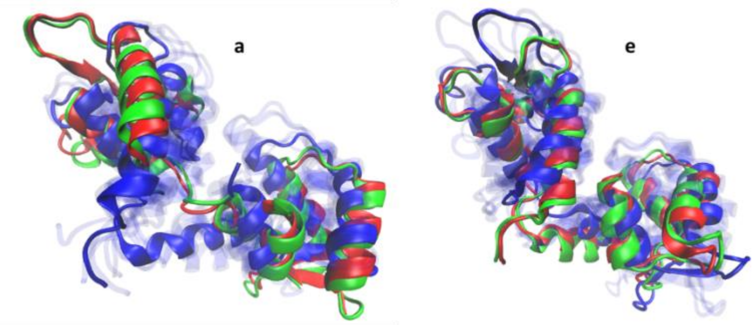

Figure 4. Overlay of the predicted (red) and target (green) structures **a** and **e** from Figure 3, with the most similar structure from the corresponding training set (blue), and a range of structures from the training set (the structrues in Table 1; transparent blue).

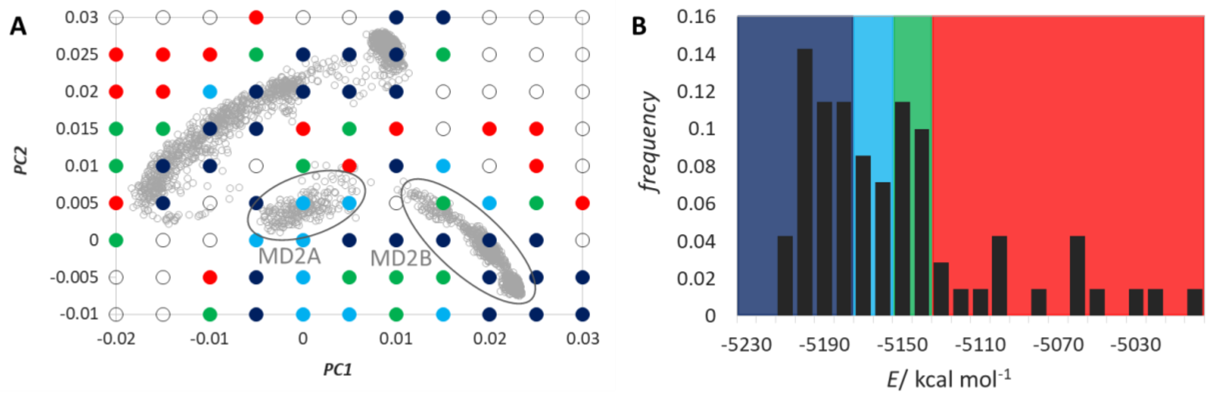

Figure 5. (**A**) Relative potential energies of predicted CaM conformations (large dots) overlaid on the conformational landscape sampled during MD simulations (small grey dots), and (**B**) energy distribution of the structures after energy minimisation of the sidechains. The colours of the large dots in **A** correspond to the coloured areas in **B**, and the white dots are those conformations that did not successfully energy minimise. The grey ovals indicate two distinct clusters, MD2A and MD2B, from MD2.