Received: 5 August 2020 | Revised: 21 January 2021 | Accepted: 23 February 2021

DOI: 10.1002/prot.26068

RESEARCH ARTICLE

Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning

Yiming Jin[1,2] | Linus O. Johannissen[1] | Sam Hay[1]

[1]Manchester Institute of Biotechnology and Department of Chemistry, The University of Manchester, Manchester, UK

[2]School of Computer Science and Engineering, Central South University, Changsha, China

Correspondence
Linus O. Johannissen and Sam Hay, Manchester Institute of Biotechnology and Department of Chemistry, The University of Manchester, 131 Princess St, Manchester M1 7DN, UK.
Email: linus.johannissen@manchester.ac.uk, (L. O. J.) and sam.hay@manchester.ac.uk (S. H.)

Funding information
Biotechnology and Biological Sciences Research Council, Grant/Award Number: BB/M017702/1

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 a user-defined conformational landscape defined by principal components analysis or specific structural features, 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.

KEYWORDS
autoencoder, Calmodulin, conformational landscape, molecular dynamics, protein

1 | 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 (eg, 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 (eg, by identifying collective variables and imposing biasing potentials) have also been developed[8-14]; a conceptually simpler and more flexible approach is to utilize 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 two simple conformational landscapes and trained to decode points within this landscape into new structures. Two test cases are used, a short homoalanine peptide and the calcium-binding protein calmodulin (CaM). Two conformational landscapes descriptors were also used, the first two principal components of a 2D-RMSD matrix and two dihedral angles that describe the relative orientation of the two CaM globular domains. We show that it is possible to predict physically plausible conformations which were not sampled during the MD simulation(s).

2 | METHODS

2.1 | 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 cut-offs. Energy minimisation was followed by 100 ps of constant volume (NVT ensemble) and 100 ps of constant pressure (NPT ensemble; 1 bar) 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, and all simulations were run at 300 K.

2.2 | Conformational landscape

Our machine learning algorithm takes a conformational landscape in the form of a series of vectors, as input. For initial development and testing, a simple conformational landscape was defined based on the 2D-RMSD matrix, a square matrix of RMSD values for every structure relative to every other structure (ie, 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. For further validation of the method we used a conformational landscape defined by a pair of dihedrals, which describe the relative conformations of each CaM globular domain.

2.3 | Machine learning

Our code and data for model 1 are available at https://github.com/Imay-King/MDMachineLearning. The protein structures (Cartesian coordinates) were first extracted from MD simulations using the MDanalysis package.[22,23] The structures were then aligned to the starting structure by minimizing the RMSD for the same atom selection (model 1: all atoms; model 2: heavy backbone atoms) subsequently used for the ML, and the Cartesian coordinates were normalized using MinMax scaling. For the complete set of protein structures with coordinates $\left(\left(x_1^1, y_1^1, z_1^1\right) \ldots \left(x_n^m, y_n^m, z_n^m\right)\right)$, where n is the number of atoms per structure and m is the total number of structures, the normalized coordinates for atom i in 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 normalization, 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. The approach is illustrated in Figure 1. The

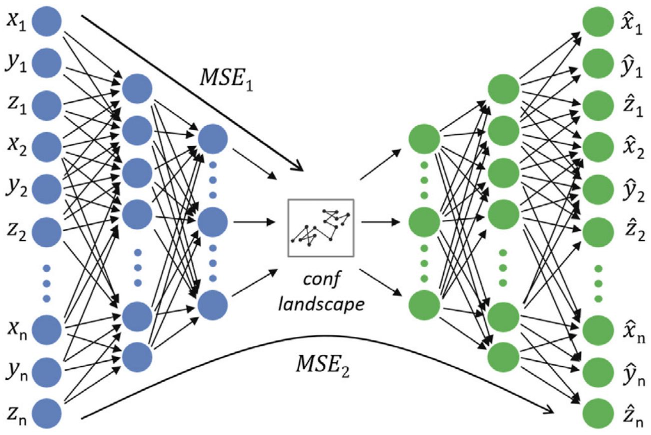

FIGURE 1 Modified autoencoder for prediction of protein structure from a user-defined protein conformational landscape, in this case defined by the first two principal components (*PC1* and *PC2*) of the 2D-RMSD matrix. The autoencoder is trained on two loss functions, *MSE$_1$* for the loss between the latent vector and *PC1*/*PC2*) and *MSE$_2$* for the loss between the target and predicted structures

algorithm is trained using two mean square error (MSE) loss functions simultaneously (with equal weighting). The first (*MSE_1*) minimizes the loss between the latent vector and the chosen conformational parameters (for the conformational landscape defined by PCA of the 2D RMSD matrix this is the two first PCs), and the second (*MSE_2*) minimizes the loss between the target and predicted structures. Predicted coordinates are then de-normalized using the inverse of the MinMax method (Equation 1).

3 | RESULTS AND DISCUSSION

3.1 | Model 1

As a proof of concept, we first attempted to predict structures from a 100 ns MD simulation of the simple peptide *L*-Ala_{13}, with snapshots taken every 20 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 (ie, 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 SD) 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 seven 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 predict protein secondary and tertiary structural elements from an MD simulation.

3.2 | 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; these are both NMR structures, and the first structure in the PDB file was used in each case. 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 100 ns MD simulation from each starting structure, with snapshots taken every 50 ps for a total of 4000 structures. As can be seen from the PCA and RMSD plots (Figures 3 and SI Figure S3), the two simulations converged to

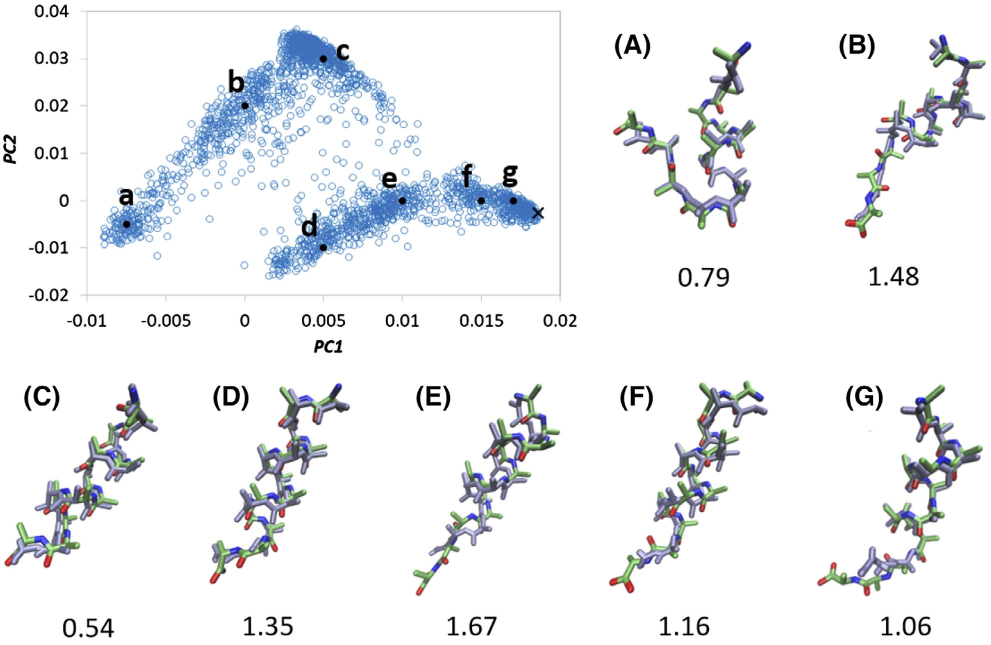

FIGURE 2 Structure predictions of the *L*-Ala_{13} peptide. Top left: *PC2* vs *PC1* plot from the 2D RMSD matrix (blue circles), with points a-g (black dots) used for testing. The black cross at PC1 ≈ 0.018 belongs to the initial, fully helical structure. For each prediction, points within (±0.002, ±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 [Color figure can be viewed at wileyonlinelibrary.com]

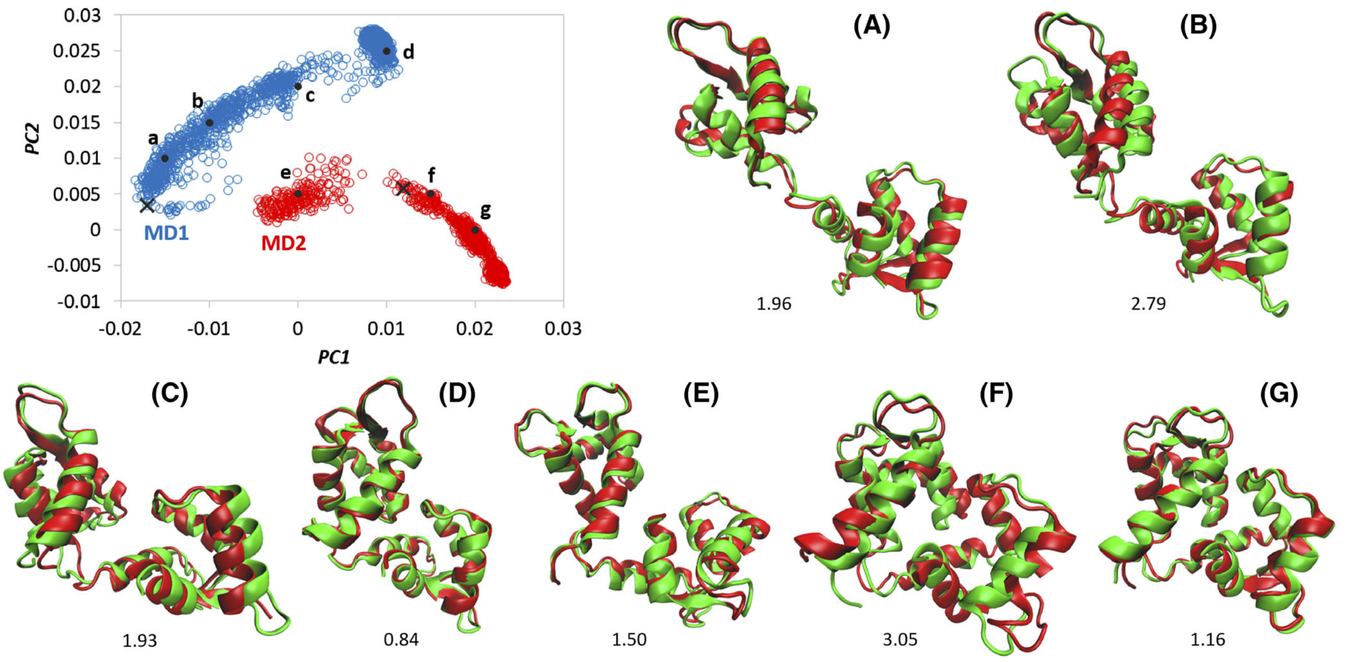

FIGURE 3 Structure predictions of CaM. Top left: *PC2* vs *PC1* plot from the 2D RMSD matrix constructed from two MD simulations (blue and red circles), with points **a-g** (black dots) used for testing. The black crosses belong to the starting structures for each simulation. For each structure, testing 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 [Color figure can be viewed at wileyonlinelibrary.com]

different conformations and the conformational space sampled in each simulation does not overlap. Here, the 100 ns simulations do not allow sufficient sampling of the CaM conformational landscape. 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 five 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 seven 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; that is, 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 seven 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.

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

The target CaM structures in Figure 3 are compared with 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. Further, the seven 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 (the conformational landscape is read in as an array which is not limited to

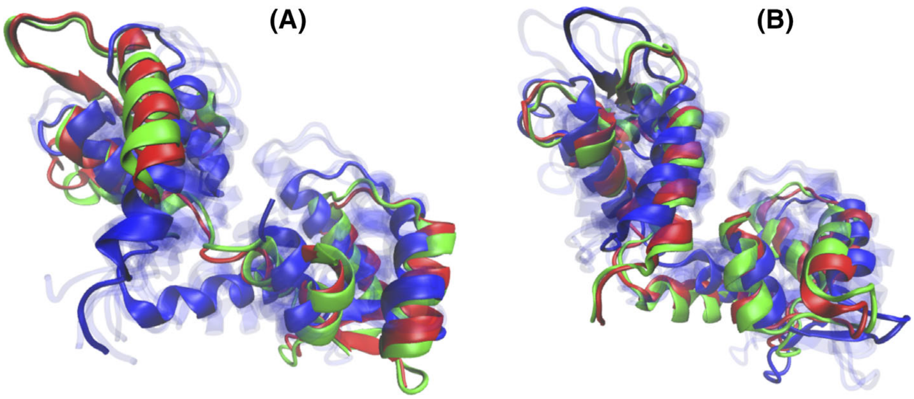

FIGURE 4 Overlay of the predicted (red) and target (green) structures a and e from Figure 3, with the most similar structure from the corresponding training set (blue), and a range of structures from the training set (the structrues in Table 1; transparent blue) [Color figure can be viewed at wileyonlinelibrary.com]

2 dimensions) which would allow more complex conformational space to be mapped in higher dimensions.

It is important to note that by necessity (so that target structures can be defined for comparison), the PC space for each prediction analyzed so far (Figure 3 and Table 1) originated from the PCA of the 2D RMSD matrix for the entire simulation (training + test data). This means that some sampling information in the test data are retained in the principal components of the training data. We will address this point below. Firstly, we address the issue that model 2 does not include the sidechains in the machine learning, because this 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. In principle there are several ways to use the predicted backbone structures for additional modeling: input geometries can be generated by building in the sidechains using 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 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 typically high-energy structures with non-optimal sidechain-sidechain interactions (at 300 K only a small minority of conformations sit at the bottom of the potential energy well) that SCWRL4 is not designed to reproduce, we then energy minimized the sidechains of both the original and rebuild structures using the FF14SB 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.

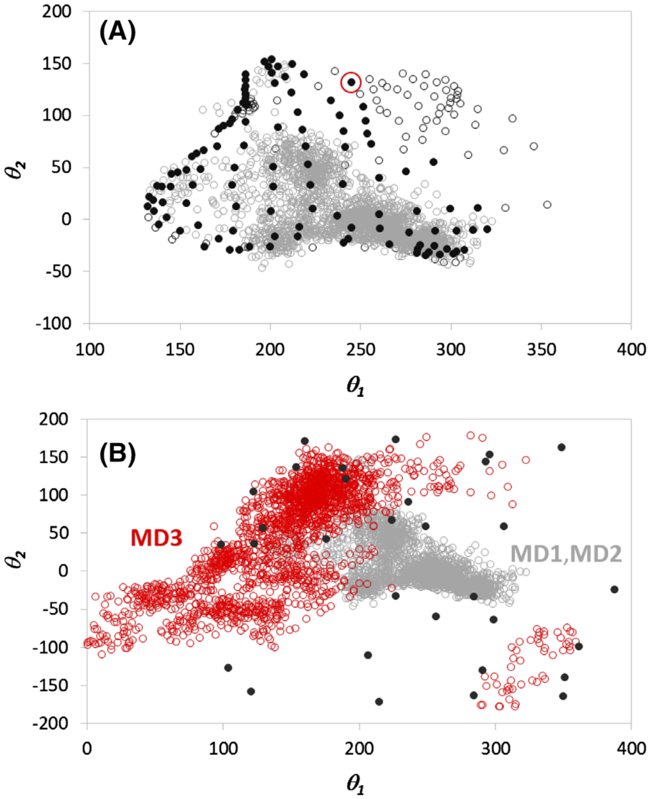

FIGURE 5 Conformational sampling on the (θ_1, θ_2) landscape: (A) structures from the combined MD1 and MD2 simulations (gray circles), predicted structures after successful sidechain reconstruction (black dots) and predicted structures with unsuccessful sidechain reconstruction (black open circles). The red circle indicates the structure chosen for additional MD simulations. (B) Conformational sampling during MD1 and MD2 (gray circles) compared to the new MD simulation (MD3, red circles) and the NMR ensemble (PDB 1LKJ, black dots) [Color figure can be viewed at wileyonlinelibrary.com]

From the NMR structure of apo CaM (PDB ID: 1LKJ) we can see that the ensemble of structures covers more conformational space than is sampled during MD1 and MD2 simulations, due to extensive domain motion (SI Figure S6). However, the first two PCs of the 2D RMSD matrix does not adequately capture this sampling (SI Figure S7A), suggesting that in this case 2D RMSD captures more intra-domain structural changes than domain motion. In order to predict new structures, we therefore chose to employ a different conformational landscape defined by two dihedral angles, θ_1 and θ_2, which describe the relative orientation of the two CaM globular domains (SI Figure S7). We defined a regular grid of (θ_1,θ_2) values and mapped this over the conformational space described by (θ_1,θ_2) for the combined MD1 and MD2 simulations (SI Figure S8). The combined MD1 and MD2 simulation data were used for training using this new (θ_1,θ_2) conformational landscape descriptor and the regular grid was used for subsequent prediction. The (θ_1,θ_2) values of the predicted structures were often observed to differ from their target values, so that the majority of predicted structures subsequently lie near or within the conformational space of MD1 and MD2 (Figure 5(A)). This suggests that the autoencoder will not arbitrarily predict a structure in a region of conformational space for which there is insufficient data for successful extrapolation. There is, however, a large region of predicted structures with distinct (θ_1,θ_2) values. Only some of these could be successfully energy minimized after sidechain reconstruction using SCWRL4, with others failing due to steric clashes. Clearly there is room for improvement here, for example, using structural cost-functions based on C_α-distances and dihedral angles as employed by the EncoderMap algorithm,[18] or possibly by performing the ML with the entire protein (without sidechains removed). Nevertheless, using this method we were able to identify a predicted structure, indicated by a red circle in Figure 5(A), which is more similar to a structure from the *apo* CaM NMR ensemble (PDB 1LKJ) than to any of the structures from MD1 or MD2 (SI Figure S9). Starting from this predicted structure (with side chains reconstructed) as the input geometry, we ran an additional 100 ns MD simulation, which results in a much greater coverage of conformational space compared to the initial MD simulations (Figure 5(C)).

4 | CONCLUSIONS

In summary, we have demonstrated a proof-of-principle method of combining MD simulation with machine learning to explore a user-defined, arbitrary conformational landscape. An autoencoder maps snapshots from MD simulations onto the conformational landscape, and we show that we can predict, with useful accuracy, conformations that are not present in the training data. This method allows the prediction of new physically realistic structures of conformationally dynamic proteins that can be used for enhanced sampling of MD simulations, by rapidly generating new structures from which additional MD simulations can be initiated for a more efficient search through conformational space.

ACKNOWLEDGMENTS

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.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26068.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available at: https://github.com/Imay-King/MDMachineLearning and from the corresponding authors upon reasonable request.

ORCID

Sam Hay https://orcid.org/0000-0003-3274-0938

REFERENCES

1. Klepeis JL, Lindorff-Larsen K, Dror RO, Shaw DE. Long-timescale molecular dynamics simulations of protein structure and function. *Curr Opin Struct Biol*. 2009;19(2):120-127.
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):964-972.
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 YI, Shao Q, Zhang J, Yang L, Gao YQ. Enhanced sampling in molecular dynamics. *J Chem Phys*. 2019;151(7):070902.
6. Fleetwood O, Kasimova MA, Westerlund AM, Delemotte L. Molecular insights from conformational ensembles via machine learning. *Biophys J*. 2020;118(3):765-780.
7. Kandathil SM, Greener JG, Jones DT. Recent developments in deep learning applied to protein structure prediction. *Proteins*. 2019;87(12): 1179-1189.
8. Wang Y, Lamim Ribeiro JM, 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):1147.
10. Wu H, Mardt A, Pasquali L, Noe F. In Deep Generative Markov State Models, Advances in Neural Information Processing Systems. 2018.
11. Ribeiro JML, 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 AL. 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 YY, Parrinello M. Neural networks-based variationally enhanced sampling. *Proc Natl Acad Sci U S A*. 2019;116(36):17641-17647.
14. Shamsi Z, Cheng KJ, 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 MT. Coupling molecular dynamics and deep learning to mine protein conformational space. *Structure*. 2019;27(6):1034-1040.e3.
16. Ceriotti M, Tribello GA, Parrinello M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc Natl Acad Sci U S A*. 2011;108(32):13023.
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 AE, Berendsen HJ. GROMACS: fast, flexible, and free. *J Comput Chem*. 2005;26(16):1701-1718.

20. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput*. 2015; 11(8):3696-3713.
21. Case D, Betz R, Cerutti DS, et al. *Amber 16*. San Francisco: University of California; 2016.
22. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: a toolkit for the analysis of molecular dynamics simulations. *J Comput Chem*. 2011;32(10):2319-2327.
23. Gowers RJ, Linke M, Barnoud J, et al. MDAnalysis: a python package for the rapid analysis of MolecularDynamics simulations. In Proceedings of the 15th Python in Science Conference, Austin, Texas. 2019.
24. Abadi M, Barham P, Chen J, et al. TensorFlow: a system for large-scale machine learning, arXiv:1605.08695, 2016.
25. Rampasek L, Goldenberg A. TensorFlow: biology's gateway to deep learning? *Cell Syst*. 2016;2(1):12-14.
26. Kingma DP, Ba J. A Method for Stochastic Optimization. 2014, arXiv: 1412.
27. Nair V, Hinton GE. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on Machine Learning*. Israel: Haifa; 2010;807-814.
28. Chou JJ, Li S, Klee CB, Bax A. Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat Struct Biol*. 2001;8(11):990-997.
29. Ogura K, Kumeta H, Takahasi K, et al. 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):159-172.
30. Bhuyan MS, 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):2000-2017.
32. Towse CL, Rysavy SJ, Vulovic IM, Daggett V. New dynamic Rotamer libraries: data-driven analysis of side-chain conformational propensities. *Structure*. 2016;24(1):187-199.
33. Wang Q, Canutescu AA, Dunbrack RL Jr. SCWRL and MolIDE: computer programs for side-chain conformation prediction and homology modeling. *Nat Protoc*. 2008;3(12):1832-1847.
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):1568-1581.
35. Leech J, Prins JF, Hermans J. SMD: visual steering of molecular dynamics for protein design. *IEEE Comput. Sci. Eng*. 1996;3(4):38-45.
36. Krivov GG, Shapovalov MV, Dunbrack RL Jr. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins*. 2009;77(4): 778-795.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Jin Y, Johannissen LO, Hay S. Predicting new protein conformations from molecular dynamics simulation conformational landscapes and machine learning. *Proteins*. 2021;89:915–921. https://doi.org/10.1002/prot.26068