biomolecules

MDPI

Article

Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures

Vy T. Duong [1], Elizabeth M. Diessner [1], Gianmarc Grazioli [2], Rachel W. Martin [1,3,*] and Carter T. Butts [4,*]

1 Department of Chemistry, University of California, Irvine, CA 92697, USA; vy.duong.phd@gmail.com (V.T.D.); ediessne@uci.edu (E.M.D.)
2 Department of Chemistry, San Jose State University, San Jose, CA 95192, USA; gianmarc.grazioli@sjsu.edu
3 Department of Molecular Biology & Biochemistry, University of California, Irvine, CA 92697, USA
4 Departments of Sociology, Statistics and Electrical Engineering & Computer Science, University of California, Irvine, CA 92697, USA
* Correspondence: rwmartin@uci.edu (R.W.M.); buttsc@uci.edu (C.T.B.)

Abstract: Coarse-graining is a powerful tool for extending the reach of dynamic models of proteins and other biological macromolecules. Topological coarse-graining, in which biomolecules or sets thereof are represented via graph structures, is a particularly useful way of obtaining highly compressed representations of molecular structures, and simulations operating via such representations can achieve substantial computational savings. A drawback of coarse-graining, however, is the loss of atomistic detail—an effect that is especially acute for topological representations such as protein structure networks (PSNs). Here, we introduce an approach based on a combination of machine learning and physically-guided refinement for inferring atomic coordinates from PSNs. This "neural upscaling" procedure exploits the constraints implied by PSNs on possible configurations, as well as differences in the likelihood of observing different configurations with the same PSN. Using a 1 μs atomistic molecular dynamics trajectory of Aβ_{1-40}, we show that neural upscaling is able to effectively recapitulate detailed structural information for intrinsically disordered proteins, being particularly successful in recovering features such as transient secondary structure. These results suggest that scalable network-based models for protein structure and dynamics may be used in settings where atomistic detail is desired, with upscaling employed to impute atomic coordinates from PSNs.

Keywords: coarse-grained models; molecular dynamics; protein structure networks; intrinsically disordered proteins; machine learning

Citation: Duong, V.T.; Diessner, E.M.; Grazioli, G.; Martin, R.W.; Butts, C.T. Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures. *Biomolecules* **2021**, *11*, 1788. https://doi.org/10.3390/biom11121788

Academic Editors: Thomas R. Caulfield and Alexander T. Baker

Received: 30 July 2021
Accepted: 19 November 2021
Published: 30 November 2021

1. Background

Proteins and other biological macromolecules exhibit a wide variety of complex dynamics and interactions at varying size and time scales. While atomistic molecular dynamics (MD) models currently serve as the gold standard tools for simulating dynamics at high resolution (with some inroads by quantum mechanical methods in small-scale or specialized applications), the cost of large-scale MD simulations limits their use to relatively small systems on time scales of microseconds or less. Coarse-grained (CG) models offer a means of accessing larger system sizes and longer time scales, sacrificing atomistic detail in exchange for reduced computational cost. Many "flavors" of coarse-grained simulation exist, with the most common being aggregate particle models that represent collections of atoms by single particles whose positions evolve under a suitably modified forcefield. The highly successful MARTINI model [1], for instance, represents biomolecules by "beads" corresponding roughly to one bead per four heavy atoms, with hydrogens left implicit; MARTINI and other CG MD models have proven useful in studying the structure and dynamics of large complexes, lipid phases, and other systems that are too large to be treated with atomistic MD methods [2]. An even more radical approach to coarse-graining

employs topological representations, representing molecules or molecular aggregates by network structures that encode the interactions between atoms or groups thereof, but not their positions in three-dimensional space [3,4]. Often employed for descriptive analysis of trajectories produced by MD or other methods (see, e.g., [5–7]), network representations have the advantage of retaining complex topological information involving protein structure while being highly compressive (greatly reducing the computational cost needed for, e.g., comparative analysis of long trajectories), and facilitating application of a large body of graph-theoretic measures for capturing structural properties ranging from cohesion and constraint to differences in centrality and contact rates. Recent work has also considered the generation of trajectories directly within the topological representation ("coordinate-free" simulation), allowing considerable computational savings [8,9].

While many questions can be posed directly within a CG representation, an obvious limitation of coarse-graining is that some observables of interest cannot be obtained without an additional step of "backmapping" or "upscaling" the CG trajectory to atomistic resolution. At first blush, this may seem impossible: by definition, a CG model does not resolve individual atoms. In practice, however, CG structures are often sufficiently constraining that a well-designed algorithm can infer atomic positions from them with considerable accuracy [10]. For instance, a number of upscaling methods for particle-based CG models work via a two-stage process in which initial guesses for atomic placement are made based on, e.g., random positioning [11], fragment-based [12,13], or geometry-based [14–17] initialization, followed by an energy minimization step to ensure physically realistic coordinates. This is not unrelated to protein structure prediction methods like those of [18,19], which begin with approximate structures based on local homology and subsequently refine them via minimization in a simplified force field. Such techniques have proven extremely successful in predicting the structure of globular proteins [20,21], and are widely used in enzyme discovery and engineering applications [22,23].

In the context of *topological* coarse-graining, the historical focus has been on mapping from atomistic to coarse-grained networks for purposes of analysis (e.g., [3,24–27]), with correspondingly less emphasis on the upscaling problem. Recent work, however, has suggested the potential of graph-theoretic models for molecular structure and dynamics. For instance, Grazioli et al. [9], Yu et al. [28] use Hamiltonians defined on graphs representing the structures of protein aggregates to model the equilibrium structures and kinetics of amyloid fibrils and associated aggregation states (with vertices representing individual proteins, and edges indicating bound interactions). On a smaller scale, Grazioli et al. [8] used a closely related approach to model transient structure in intrinsically disordered proteins (IDPs), using residue-level protein structure networks (PSNs) in which each vertex represents a residue and edges represent inter-residue contacts. Although we are unaware of any existing methods for upscaling such graph structures to atomic resolution, effective methods for this purpose would greatly extend the practical reach of network-based simulation models.

Our focus in this paper is this last problem: the upscaling of topological representations of macromolecular structure (and by extension, dynamics) to atomic resolution. We specifically consider the upscaling of residue-level PSNs, as this is a widely used level of network coarse-graining for proteins and poses a non-trivial challenge for atomistic refinement. To perform the mapping from network structure to atomistic structure, we exploit advances in machine learning (ML) methods, predicting atomic coordinates from topological inputs using deep neural networks. Machine learning strategies (particularly including neural networks) have become widely used in CG modeling, with past efforts focused on ML-based methods for learning or refining CG forcefields (see e.g., [24,25,29–33]). Here, we use multilayer perceptron-based (MLP) neural networks to learn pairwise interatomic distances from residue-level PSNs, allowing us to recover atomistic detail from input network structures.

In this work, we demonstrate the utility of MLP neural network models to translate coarse-grained protein structure network representations to their more finely detailed 3D

coordinate structures. We apply this to the case of IDPs, showing that the trained neural network is able to reproduce equilibrium conformations of amyloid-β protein obtained from MD simulations at atomic-level detail, also capturing its diverse transient secondary structure behavior. We additionally consider the use of further refinements (such as chirality corrections and energy minimization) to improve predictive performance. We show that this scheme can obtain a high level of accuracy, with median RMSE for predicted versus true 3D structures of approximately 2.13 Å and a high degree of correspondence for relatively folded regions of the protein. The resulting scheme provides a practical mechanism for mapping PSNs produced by generative network models to predicted atomistic structures (Figure 1), for using PSNs as an efficient tool for lossy compression of long trajectories, or other applications in which it is useful to infer atomistic information from coarse-grained topological representations.

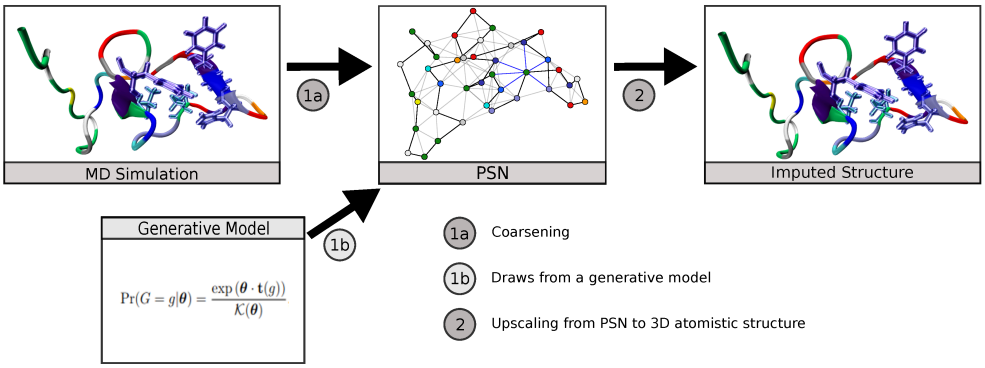

Figure 1. The ability to impute atomistic structures from network representations enables both compressive representation of structures from MD trajectories and the use of generative network models to predict distributions of atomistic structures.

The remainder of the paper is organized as follows. We introduce our approach in Section 2, including both our ML pipeline and subsequent refinement methods. Section 3 reports the results of our simulation study, Section 4 discusses further directions, and Section 5 concludes the paper.

2. Methods

Data Generation Our data come from atomistic MD trajectories of Aβ_{1-40}, a widely studied IDP implicated in the etiology of Alzheimer's disease; the atomistic trajectories and associated PSN coarsenings, respectively, serve as ground truth and inputs for the upscaling model (Figure 2). Beginning with the lowest energy monomer of the PDB structure, 2LFM [34], one Aβ_{1-40} monomer was simulated in explicit solvent for 1 µs using NAMD [35] via the following protocol: the initial monomer structure was solvated in a cubic TIP3P [36] water box of minimum margin 25 Å, and neutralized with NaCl counter-ions. This assembly was minimized for 10,000 iterations, followed by velocity initialization and 250 simulation iterations before final adjustment of the water box. A trajectory of approximately 1.1 µs was then simulated. Simulation was performed under periodic boundary conditions in NAMD with the CHARMM36m forcefield [37], using an NPT ensemble at 300K and 1 atm pressure. Temperature control was maintained by Langevin dynamics with a period of 1/ps, with Nosé-Hoover Langevin piston pressure control [38,39]. Creation of initial conditions and related data processing were performed using VMD [40].

The simulation contains 11,926 total frames/conformations, of which 72% was allocated for training, 20% for testing, and 8% for validation. When working with highly flexible, nonparametric learning methods (including approaches such as the deep learning techniques used here), it is common to employ multiple data splits for cross-validation purposes. Here, the training data is used for parameter estimation, the validation group is used for hyperparameter tuning and other optimizations to the training process, and

the test data is used to provide a held-out evaluation of the final model. Five-fold cross validation was also performed to ensure that bias was not introduced during the initial train-test division. *k*-fold cross validation (a standard technique in which the data is split at random into *k* segments, each of which is then used to produce one test-train split) guards against the risk of obtaining anomalous performance estimates due to selection of an unusual data division, and can provide additional insights into performance sensitivity by comparing results across divisions of the data. For each frame in the A$\beta_{1–40}$ simulation, a protein structure network (PSN) was calculated using a combination of VMD [40] and the `statnet` [41,42] and `bio3d` [43] libraries for R [44]).

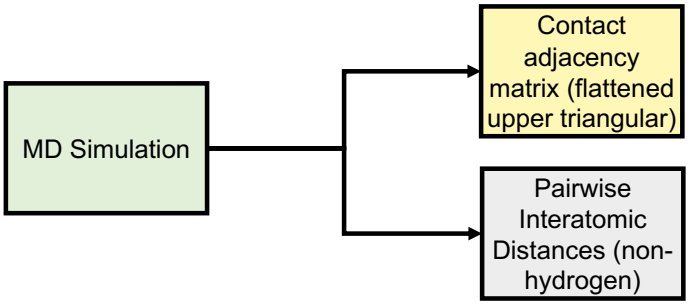

Figure 2. Data generation of input (upper triangle of PSN adjacency matrices) and output (upper triangular of PIDs) data.

Monomer states were sampled from the trajectory every 100 ps, from which residue-level protein structure networks were constructed. Vertices correspond to individual residues, with two vertices being considered adjacent if they contain respective atoms whose distance is less than or equal to 1.1 times the sum of their van der Waals radii (based on radius data from [45]). The input data used to train the neural network model consists of the flattened upper triangular matrix data extracted from the residue-level contact adjacency matrix for each conformation in the A$\beta_{1–40}$ trajectory. A contact adjacency matrix, x, is a binary square representation of the edges existing between two residues, with $x_{ij} = 0$ where there is no edge between the vertices associated with respective residues i and j, and $x_{ij} = 1$ where an edge is present. As contact is an undirected relation, x is symmetric, and only one triangle of the matrix is required for learning. Here, the upper triangle is flattened into a one-dimensional array for processing. The output data used to train the model is the corresponding set of flattened upper triangles of the pairwise interatomic distance matrices (PIDs) calculated on all non-hydrogen atoms (across all frames in the MD simulation) (Figure 2). For purposes of evaluation (as discussed below) it should be noted that when comparing predicted to observed PIDs, we define errors in terms of the pairwise distances themselves, not, e.g., the distances between equivalent atoms in the observed versus predicted structures post-alignment. For instance, let y_{ij} be the observed PID for heavy atoms i and j ($i, j \in 1, \ldots, N$), with predicted value $\hat{y}_{ij}$. Then, the squared error in PIDs is given by $\sum_{i=1}^{N} \sum_{j=i+1}^{N} (y_{ij} - \hat{y}_{ij})^2$, with RMSDs and other quantities defined accordingly.

Neural network architecture and hyperparameters After generation of input and output data, a multi-layer perceptron (MLP) neural network was utilized for training as indicated in the pipeline (Figure 3). This neural network contains four hidden layers (structured as follows), and was implemented using the machine-learning libraries Keras [46] and tensorflow [47]. The first three hidden layers consist of 2000 neurons, the fourth layer contains 8000 neurons, and the last output layer predicts the flattened upper triangle of the pairwise interatomic distance matrix for a given frame from the MD simulation (46,665 neurons) (Figure 6). Hyperparameters were optimized using the Talos Keras tuning module [48]. A Nvidia P6000 Quadro GPU card was used to train the model with the following hyperparameters: nonlinearity = relu, dropout rate = 0.2, optimization = AMSGrad, loss = mean squared error, batch size = 50, epochs = 100. Predicted output data were

initially assessed using three metrics: root-mean squared deviation/error (RMSD/RMSE), mean squared error (MSE), and mean absolute percentage error (MAPE).

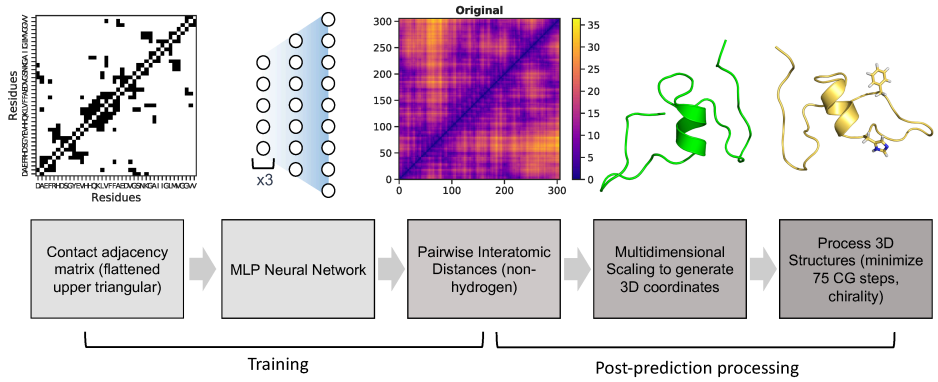

Figure 3. Pipeline of MLP neural network training and post-prediction processing.

Post-prediction processing The predicted output data (the flattened upper triangles of the pairwise interatomic distance matrices) were first transformed into symmetric pairwise interatomic distance matrices. These were then transformed into 3D coordinate data using the multi-dimensional scaling (MDS) function from the scikit-learn python module and MDtraj [49] to generate PDB structures (Figure 3). Chimera [50] was then used to add hydrogens to predicted PDB structures, which were then further processed to remove inaccurate chiral predictions. If more than half of Cα centers were inaccurately predicted as R chiral centers (D-amino acids instead of L-amino acids), this indicated that the MDS procedure (which is reflection-invariant) predicted a reflection of the true coordinates. This was mitigated by reflecting all coordinates over the y-axis for predictions exhibiting an $\frac{R}{S}$ ratio greater than 1. If fewer than half of α-carbons exhibited R chiral centers, reflecting coordinates was unnecessary. Instead, Chimera was used to switch side chain coordinates and the α-hydrogen for all inaccurately predicted Cα chiral centers. After checking for correct chirality for each residue, all conformations were further minimized for 75 conjugate gradient steps.

The number of conjugate gradient steps was chosen by evaluating structures every subsequent 20 conjugate gradient steps for a cumulative 520 steps total. The maximum 520 conjugate gradient steps was chosen based on qualitative determination of average potential energy trends of all predicted conformations with increasing conjugate gradient minimization (Figure 4). Three superposition-based metrics (RMSD, global distance test, total score (GDT_TS), template modeling (TM) score) and one superposition-free metric (local distance difference test (LDDT)) were used to analyze any potential improvements in additional conjugate gradient steps between predicted 3D structure and the original, MD-generated 3D conformation. The RMSD metric analyzes all heavy atoms, TM score focuses primarily on Cα atoms, and GDT_TS also focuses primarily on backbone atoms. The LDDT score calculates a comparison using all-atom pairwise interatomic distances. Average values of 500 randomly chosen structures (RMSD, TM Scores, GDT_TS, and LDDT) suggest a minimization range between 50–100 conjugate gradient steps. Thus, 75 steps was chosen as the total number of conjugate gradient steps to minimize all 11,926 predicted conformations. Overall, minimization yields little improvement relative to no minimization with respect to most metrics; however, it is a necessary step to remove steric clashes and slight stereochemical errors (Figure 3, last panel).

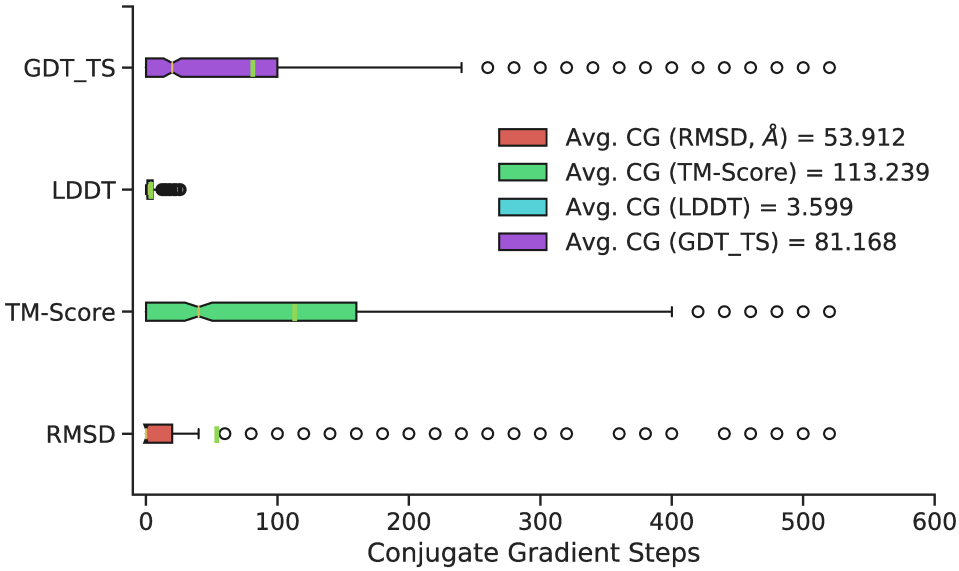

Figure 4. Distribution of the optimal number of CG steps for structure refinement, by metric, with green vertical lines representing means, and notches representing medians. Although exact optima vary by structure and metric, 50–100 steps are sufficient to provide good performance on most structures; extensive refinement beyond this point is rarely beneficial.

3. Results

3.1. Multilayer Perceptron (MLP) Neural Network Reconstructs Aβ Conformations with Atomistic Detail

Pairwise interatomic distance (PID) predictions were made for all sets of data (train, validation, test). Predictions were evaluated against the ground-truth PIDs from the MD simulation using root-mean square error/deviation (RMSE/RMSD), mean absolute error (MAE), mean absolute percentage error (MAPE). (As described above, we evaluate PID error in terms of the pairwise distances themselves, and not, e.g., imputed coordinates.) The average metrics for the test set exhibit a favorable RMSE (1.7 Å), MAE (1.17 Å), and MAPE (7.35%) (Figures 5 and 6). Five-fold cross-validation suggests bias was not arbitrarily introduced during the initial train-test split (Figure 7). Overall, average PID metrics for the validation and test set suggest the neural network was able to devise quality predictions.

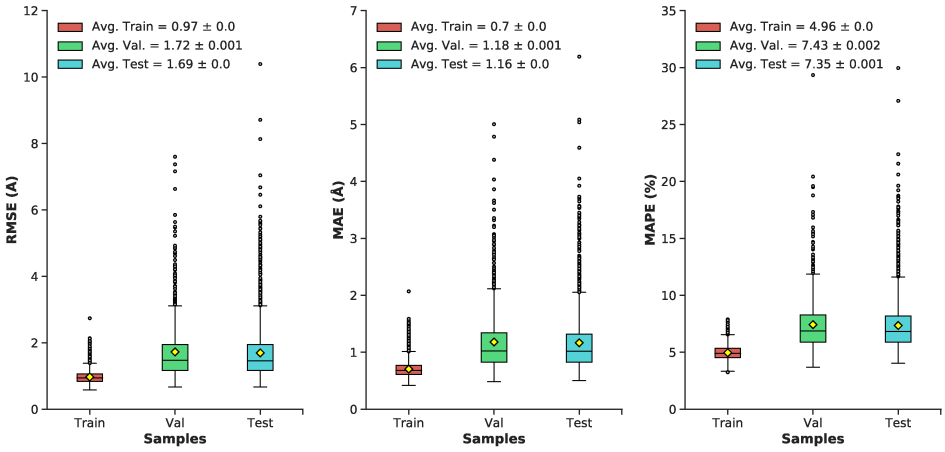

Figure 5. Boxplots of distributions for the following metrics (RMSE, MAE, MAPE) for the train, validation, and test datasets: minimum, maximum, median, outliers (grey dots), average (yellow diamond) ± standard error, lower and upper quartiles.

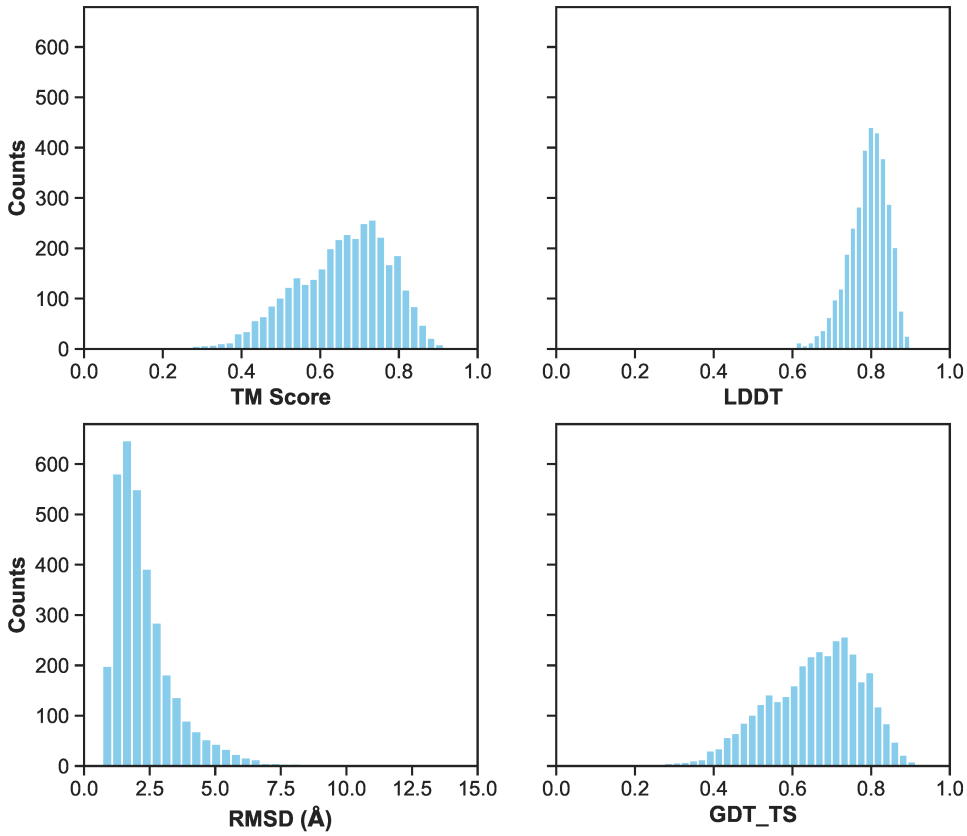

Figure 6. A histogram of metrics for the combined validation and test set of the first cross validation fold. Metrics include RMSD/RMSE in angstroms, LDDT, TM-Score, and GDT_TS. Excluding RMSD, the other three metrics range from 0 (inaccurate) to 1 (accurate) prediction. Lower RMSD values indicate more accurate predictions.

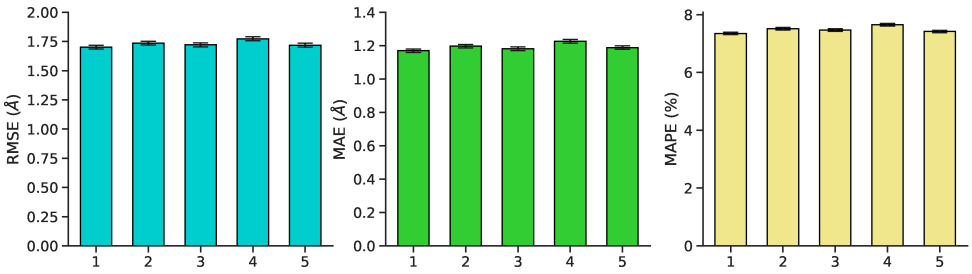

Figure 7. Mean validation performance ± standard error on RMSE, MAE, and MAPE for each of five cross-validation splits. Performance is robust to choice of fold.

To illustrate model performance, we show examples of both good and bad predictions from the test set, beginning with the positive example of frame 1133. Original and predicted pairwise interatomic distances for frame 1133 upon initial visualization, have highly comparable values (Figure 8a,b). A grayscale depiction of absolute value differences between original and predicted PIDs reveals white and light grey data points, denoting mostly low values (Figure 8d). The distribution of these data shows that approximately 98% of difference values are less than 2 Å and 88% are less than 1 Å (Figure 8c). This result from the test set highlights one of the most accurate predictions of atomistic structures made by the neural network model.

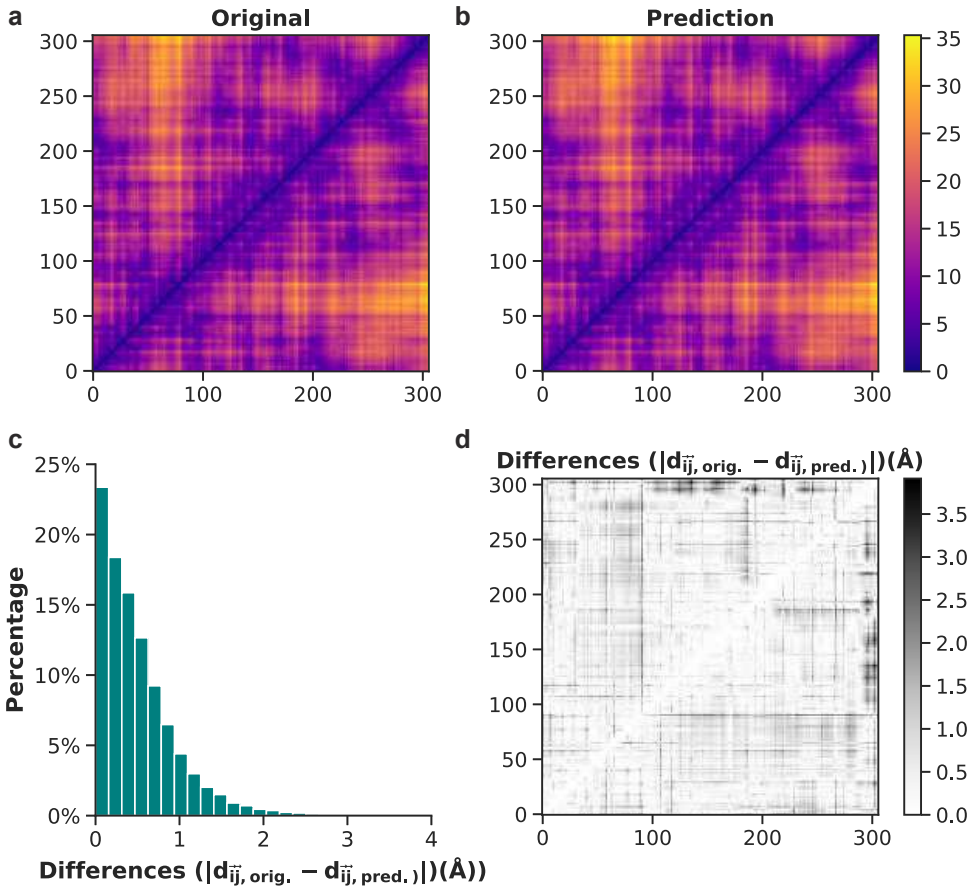

Figure 8. Comparison between original and predicted pairwise interatomic distances for frame 1133 (from the test set). (**a**) Actual distances are shown for all heavy atoms. (**b**) Heavy-atom predictions of all pairwise interatomic distance. (**c**) Histogram of differences between original and predicted Euclidean distances. (**d**) Binary plot displaying the absolute difference values between each actual and predicted distance for frame 1133.

Using RMSEs of PIDs as a basis for selection, we show processed 3D predictions of the lowest RMSE score representation (frame 1133, Figure 9a), the median representation (frame 7431, Figure 9b), and the highest RMSE score structure (frame 7560, Figure 9b). The prediction with the lowest RMSE (0.67 Å) exhibits more helical secondary structure compared to the median and worst predictions, which exhibit more random coil-like dynamics. RMSE of all heavy atoms for the median representation exhibits a fairly reasonable value of 1.46 Å whereas the worst PID prediction has a RMSE of 10.4 Å. Notably, the prediction for Figure 9c aligns reasonably well for the first 20 residues and the remaining residues are more poorly predicted by the neural network model. Since folded regions are inherently more data-rich for binary contact adjacency matrix representations (e.g., a 5 Å PID and a 500 Å PID both produce the same zero matrix element), it is not surprising the neural network model struggles to predict this specific overly extended conformation; however, we note that the prediction still preserves the qualitative aspects of the extended structure, and is quite accurate for the N-terminal region. The RMSEs according to 3D structure alignment between original and processed 3D structure (and not on the basis of PIDs) also show similar values: best (0.77 Å), median (2.13 Å), and worst (12.01 Å). These values are slightly higher compared to PID-based RMSEs, most likely due to introduced 3D alignment, whereas PIDs report RMSEs between all heavy atoms.

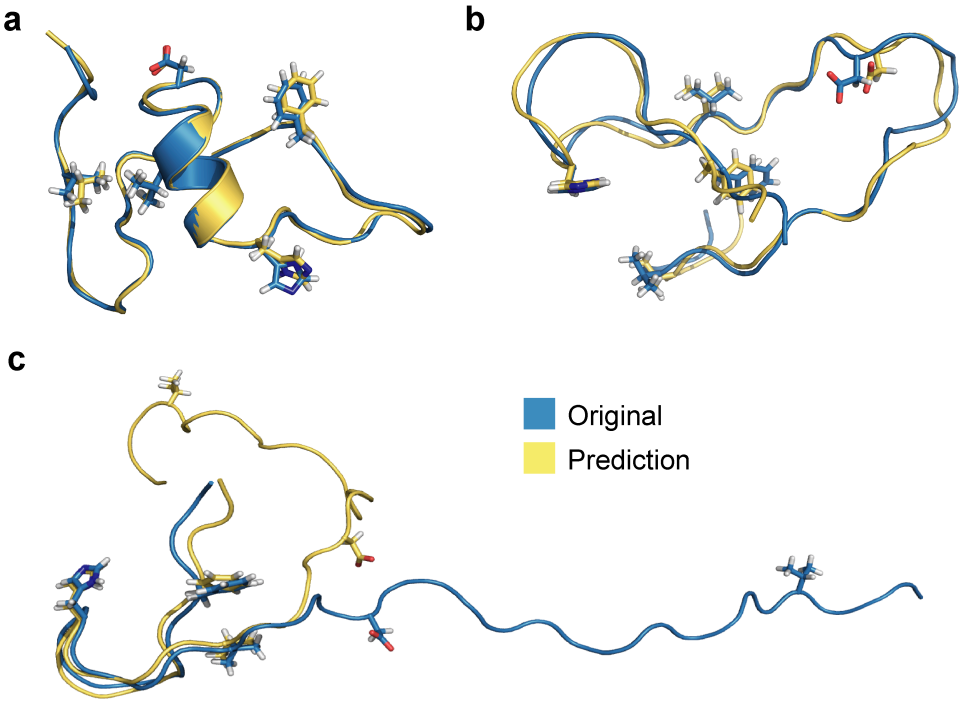

Figure 9. Alignment between original and predicted and processed 3D structures for (**a**) the best, (**b**) median, and (**c**) worst predictions based on RMSE values of PIDs.

3.2. Generation of 3D Structures and Subsequent Minimization

When multidimensional scaling maps PIDs into 3D coordinates, it does so without regard to chirality. There are thus instances in which entire conformations are D- instead of L-amino acids, a correction that can be easily identified and fixed by reflecting coordinates across the y-axis. We also corrected conformations that contained only a few instances of D-amino acids, a result of the neural network predicting slightly incorrect side chain PIDs. These chirality checks followed by minimization are necessary, computationally inexpensive processing steps required to transform PIDs into sterically reasonable 3D structures. Once corrections where fixed using Chimera, we then minimized all proteins for 75 conjugate gradient steps (a determination detailed in Methods), with a few conformations (23) requiring an additional 5 steps.

Figure 10 depicts a pre- and post-minimization of the best predicted conformation (frame 1133) in the test set. Here, we focus particularly on residues histidine 13 (His13) and phenylalanine 4 (Phe4). Both residues in the pre-minimized conformation are sterically incorrect and misplaced, whereas in the post-minimized conformation, both residues have expected canonical sterics, devoid of incorrectly positioned atoms. When these optimization techniques (stereochemical corrections and minimization) are combined with the predictive power of the MLP neural network, this method yields highly effective predictive capabilities.

After minimization, it was also imperative to compare 3D minimized predictions to their original MD simulation counterparts. Three superposition-based metrics (RMSD, TM score, GDT_TS) and one superposition-free metric (local distance difference test (LDDT)) were utilized for this evaluation. The template modeling (TM) score measures the backbone similarity between a reference protein and target protein with a range from 0 (dissimilar) to 1 (identical) [51]. RMSD is a canonical protein comparison metric, and here we parameterize it to compare all heavy atoms between native and predicted structures. LDDT utilizes pairwise interatomic distances in its methodology, focusing on local intramolecular interactions and the degree (range 0–1) of their retention in the target conformation in comparison to the native reference structure [52]. The global distance test, total score (GDT_TS) is an improvement compared to RMSD designed to assess structures with the same sequence but

different tertiary structure, with a higher score denoting better agreement (range 0–1) [53]. All four metrics are commonly used during the biennial Critical Assessment of Structure Prediction (CASP) structure prediction and assessment competition [54], and here we use these metrics to assess the predictive performance of the model.

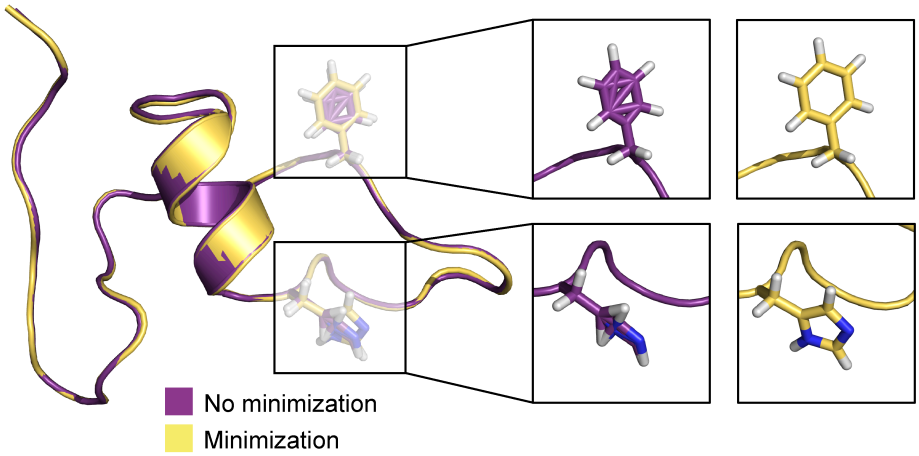

Figure 10. Comparison of pre- and post-minimized structures of the best prediction in the test set, frame 1133.

Figure 11 illustrates these metrics for the combined validation-test set. There exists a positive correlation between LDDT vs. TM scores and GDT_TS (Figure 11a,b). Between RMSDs vs. TM scores and GDT_TS, predictions exhibit a negative correlation (Figure 11c,d). Included are also the aforementioned best (yellow diamond), median (purple diamond), and worst (red diamond) PID predictions from Figure 9. Since their designation as best, median and worst were on the basis of RMSEs of PIDs and not 3D structure, it is interesting to observe the surprisingly high LDDT value of frame 7560 (the worst prediction). This suggests the neural network was able to preserve more local residue interactions despite struggling with larger more regional intramolecular interactions. TM scores exhibit values in the lower range of <0.5, whereas most GDT_TS and LDDT values occupy a range >0.5, suggesting TM scores may not be as reliable of an assessment metric for Aβ_{1-40}. The average and 95% confidence intervals suggest predicted 3D models are predicted relatively well considering the high GDT_TS average and narrow 95% confidence interval (Figure 12). The best and median test cases occupy expected 3D metrics (Figure 11). In combination with PID metrics (Figure 5), the 3D metrics demonstrate the model's ability to reasonably reconstruct the complex protein conformation of Aβ_{1-40} from coarse contact adjacency matrices.

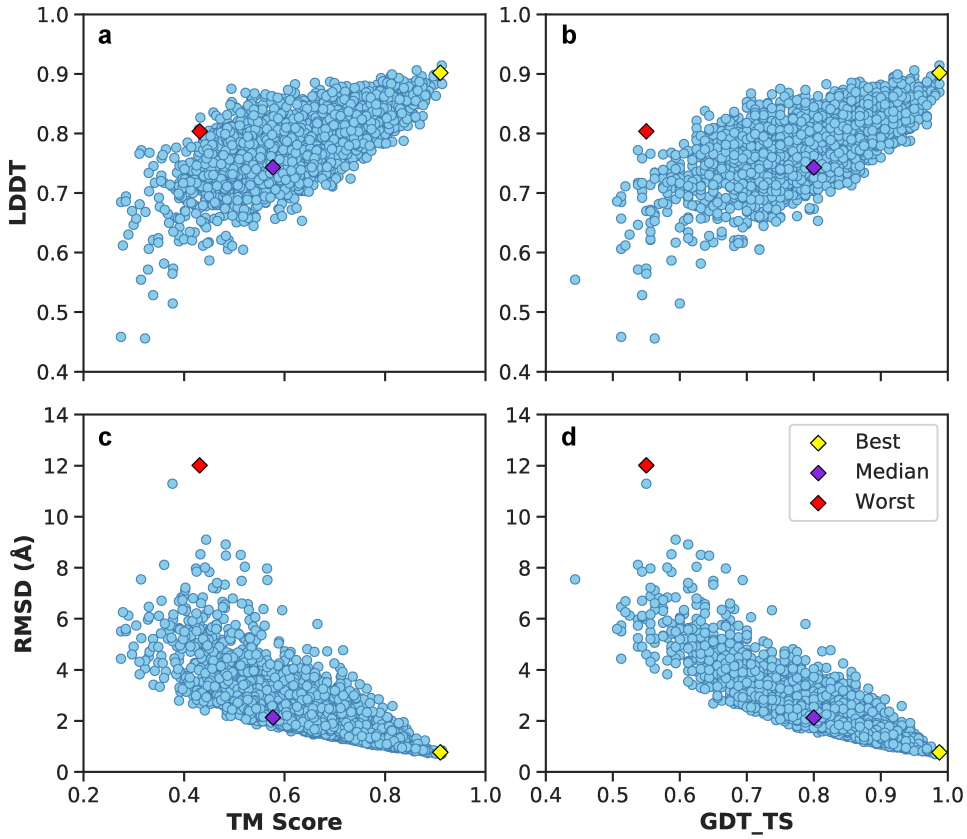

Figure 11. Juxtaposition of 3D structural metrics of the combined validation-test set: TM score, LDDT, GDT_TS, and RMSD. In addition, best, median, and worst predictions are shown based on PIDs. (**a**) LDDT vs. TM score metrics of the validation-test set. (**b**) LDDT vs. GDT_TS score metrics of the validation-test set. (**c**) RMSD vs. TM score metrics of the validation-test set. (**d**) RMSD vs. GTD_TS score metrics of the validation-test set.

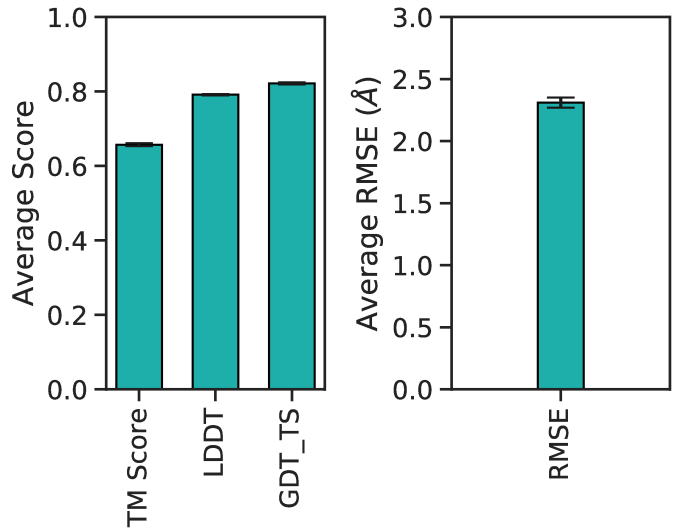

Figure 12. Barplot of average 3D accuracy metrics and corresponding 95% confidence intervals per score type.

4. Discussion

In this work, we have implemented a custom MLP neural network modeling approach to reconstruct atom-level representations of Aβ_{1-40} from residue-level PSNs. Although this particular neural upscaling model is specific to amyloid-β, the MLP neural network

model can be retrained to other biomolecular systems using inputs derived from a variety of different sources (e.g., MD simulations, NMR ensembles, etc.). Given training data in the form of PIDs, the approach used here can be used to obtain comparable predictive models for PSNs associated with any protein system (and, assuming appropriate modification of the coarsening level, alternative PSN definitions such as that of [3]). Although further work is needed to investigate the range of conditions under which these models will work well, the success of the Aβ_{1-40} model (a relatively non-trivial case, due to the presence of highly variable and often unfolded conformations) bodes well for performance in other systems. More broadly, an obvious extension of this approach is the creation of more general models for broader classes of protein systems, and for multiple levels of PSN coarse-graining. The success of deep learning in producing predictors with good generalization performance over large ranges of inputs (e.g., images with widely varying content) suggests that such general-purpose PSN upscaling tools are an achievable goal.

Although previous reverse mapping methods (e.g., random placement, geometric-based, etc.) are able to reconstruct atomistic models, they do so typically from coarse-grained force field models based on particle representations (e.g., MARTINI [1]), or from partially observed atomic coordinates (e.g., [55]). The advantage of a MLP neural network is the ability to learn and fine-tune parameters specific to the system under investigation from minimal information (here, PSN adjacency matrices) and without the requirement that the available information be geometric in character. This opens the door to the use of "coordinate-free", network based simulation methods [9] to explore the behavior of complex biomolecules while still retaining the ability to map results back to a conventional, spatial representation. Such network simulations can be produced using a network Hamiltonian based on connectivity patterns rather than atomistically detailed spatial interactions, allowing the prediction and simulation of large ensembles and/or long timescale trajectories that might otherwise be computationally expensive to model.

In the literature, another class of neural networks, specifically variational autoencoders (VAE), has been used primarily on single small molecules and bulk-phase simulations as test cases for reverse mapping [56]. This VAE methodology, although not tested on proteins, could possibly be adapted for such systems; however, we are able to demonstrate successful backmapping with a non-variational MLP neural network architecture, indicating that variational structure is not essential. To better generalize our neural upscaling technique to protein systems of different sizes, convolutional neural network architectures similar to AlphaFold [57] could be also be incorporated and trained to predict regions (e.g., N $\times$ N residue regions). With an ever-growing body of architectures whence to choose, there would seem to be considerable room for experimentation with alternative approaches.

As noted at the outset, there is considerable work on the problem of imputing atomistic structures from either coarse-grained or incomplete spatial information. Although our focus here is on the extreme setting where such information is unavailable, further enhancements may be possible in settings where both types of information can be employed. While this is not possible for, e.g., predicted PSNs arising from network models, it may be of use in cases where a combination of partial spatial information and incomplete contact maps are available, as from, e.g., incomplete NMR data or crosslinking mass spectrometry experiments. Models for such cases are an interesting direction for further work.

Finally, we note that non-neural network methods can also be applied to the upscaling problem. In preliminary experiments (not shown), we found that a kernelized ordinary least squares predictor [58] was able to obtain relatively good results (mean RMSD of approximately 2.4 Å, mean median ARE approximately 8% on interatomic distances (PIDs) under 10-fold cross-validation). Though the model was outperformed by the neural network architecture described here, and we did not therefore pursue it further, there may be situations in which non-neural network classes of predictors will prove useful. This would also seem to be a promising area for further investigation.

5. Conclusions

Direct predictions of PID metrics demonstrate the predictive capabilities of the MLP neural network to reconstruct all-atom representations of proteins from binary contact adjacency matrices. Example conformations of the best, median and worst PID-based predictions in the test set illustrate the MLP performance. In the worst prediction (frame 7560), the RMSD between the N-terminal halves of the original vs. predicted is still quite favorable (0.98 Å). Chirality corrections and conjugate gradient minimization were vital post-prediction processing steps in generating stereochemically reasonable 3D structures. Three-dimensional accuracy metrics, in particular GDT_TS—the main assessment metric in the CASP competition—suggest the neural network performed well given the average values and 95% confidence intervals. In totality, we are able to illustrate the viability of the MLP neural network architecture in this transformation experiment. This work exemplifies neural network-based techniques capable of extracting useful, meaningful data from coarse-grained models.

Author Contributions: C.T.B. performed MD simulation of Aβ_{1-40}, data processing, and PSN creation; V.T.D. created the NN architecture, trained the NN, processed and analyzed model performance; G.G. contributed to study design; E.M.D. contributed to data visualization; C.T.B. and R.W.M. conceived and designed the study. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by NSF awards DMS-1361425, IIS-1939237, and SES-1826589, and NASA award 80NSSC20K0620.

Data Availability Statement: Not applicable.

Acknowledgments: We thank Ray Luo for co-advising of Vy Duong and use of computational resources. This research was also supported by computational resources supported by UCI Calit2 Think Tank.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Marrink, S.J.; Risselada, H.J.; Yefimov, S.; Tieleman, D.P.; De Vries, A.H. The Martini Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824. [CrossRef] [PubMed]
2. Capelli, R.; Gardin, A.; Empereur-mot, C.; Doni, G.; Pavan, G.M. A Data-Driven Dimensionality Reduction Approach to Compare and Classify Lipid Force Fields. *J. Phys. Chem. B* **2021**, *125*, 7785–7796. [CrossRef] [PubMed]
3. Benson, N.C.; Daggett, V. A Chemical Group Graph Representation for Efficient High-Throughput Analysis of Atomistic Protein Simulations. *J. Bioinform. Comput. Biol.* **2012**, *10*, 1250008. [CrossRef] [PubMed]
4. Mustoe, A.M.; Al-Hashimi, H.M.; Brooks, C.L. Coarse Grained Models Reveal Essential Contributions of Topological Constraints to the Conformational Free Energy of RNA Bulges. *J. Phys. Chem. B* **2014**, *118*, 2615–2627. [CrossRef] [PubMed]
5. Wong, E.K.; Prytkova, V.; Freites, J.A.; Butts, C.T.; Tobias, D.J. Molecular Mechanism of Aggregation of the Cataract-Related γD-Crystallin W24r Variant from Multiscale Atomistic Simulations. *Biochemistry* **2019**, *58*, 3691–3699. [CrossRef]
6. Cross, T.J.; Takahashi, G.R.; Diessner, E.M.; Crosby, M.G.; Farahmand, V.; Zhuang, S.; Butts, C.T.; Martin, R.W. Sequence Characterization and Molecular Modeling of Clinically Relevant Variants of the SARS-CoV-2 Main Protease. *Biochemistry* **2020**, *9*, 3741–3756. [CrossRef]
7. Demakis, C.; Childers, M.C.; Daggett, V. Conserved Patterns and Interactions in the Unfolding Transition State Across SH3 Domain Structural Homologues. *Protein Sci.* **2021**, *30*, 391–407. [CrossRef]
8. Grazioli, G.; Martin, R.W.; Butts, C.T. Comparative Exploratory Analysis of Intrinsically Disordered Protein Dynamics Using Machine Learning and Network Analytic Methods. *Front. Mol. Biosci.* **2019**, *6*, 42. [CrossRef]
9. Grazioli, G.; Yu, Y.; Unhelkar, M.H.; Martin, R.W.; Butts, C.T. Network-Based Classification and Modeling of Amyloid Fibrils. *J. Phys. Chem. B* **2019**, *123*, 5452–5462. [CrossRef]
10. Ferrie, J.J.; Petersson, E.J. A Unified De Novo Approach For Predicting The Structures Of Ordered And Disordered Proteins. *J. Phys. Chem. B* **2020**, *124*, 5538–5548. [CrossRef]
11. Rzepiela, A.J.; Schäfer, L.V.; Goga, N.; Risselada, H.J.; De Vries, A.H.; Marrink, S.J. Reconstruction of Atomistic Details from Coarse-Grained Structures. *J. Comput. Chem.* **2010**, *31*, 1333–1343. [CrossRef]
12. Hess, B.; León, S.; Van Der Vegt, N.; Kremer, K. Long Time Atomistic Polymer Trajectories from Coarse Grained Simulations: Bisphenol-A Polycarbonate. *Soft Matter* **2006**, *2*, 409–414. [CrossRef]

13. Peter, C.; Kremer, K. Multiscale Simulation of Soft Matter Systems–From the Atomistic to the Coarse-Grained Level and Back. *Soft Matter* **2009**, *5*, 4357–4366. [CrossRef]
14. Gopal, S.M.; Mukherjee, S.; Cheng, Y.M.; Feig, M. PRIMO/PRIMONA: A Coarse-Grained Model for Proteins and Nucleic Acids That Preserves Near-Atomistic Accuracy. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 1266–1281. [CrossRef] [PubMed]
15. Brocos, P.; Mendoza-Espinosa, P.; Castillo, R.; Mas-Oliva, J.; Pineiro, Á. Multiscale Molecular Dynamics Simulations of Micelles: Coarse-Grain for Self-Assembly and Atomic Resolution for Finer Details. *Soft Matter* **2012**, *8*, 9005–9014. [CrossRef]
16. Wassenaar, T.A.; Pluhackova, K.; Böckmann, R.A.; Marrink, S.J.; Tieleman, D.P. Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *J. Chem. Theory Comput.* **2014**, *10*, 676–690. [CrossRef] [PubMed]
17. Machado, M.R.; Pantano, S. Sirah Tools: Mapping, Backmapping and Visualization of Coarse-Grained Models. *Bioinformatics* **2016**, *32*, 1568–1570. [CrossRef]
18. Bonneau, R.; Tsai, J.; Ruczinski, I.; Chivian, D.; Rohl, C.; Strauss, C.E.M.; Baker, D. Rosetta in CASP4: Progress in Ab Initio Protein Structure Prediction. *Proteins Struct. Funct. Bioinform.* **2001**, *45*, 119–126. [CrossRef]
19. Zhang, Y. Template-Based Modeling and Free Modeling by I-TASSER in CASP7. *Proteins Struct. Funct. Bioinform.* **2007**, *69*, 108–117. [CrossRef]
20. Tyka, M.D.; Keedy, D.A.; André, I.; DiMaio, F.; Song, Y.; Richardson, D.C.; Richardson, J.S.; Baker, D. Alternate States of Proteins Revealed by Detailed Energy Landscape Mapping. *J. Mol. Biol.* **2011**, *405*, 607–618. [CrossRef]
21. Pearce, R.; Zhang, Y. Toward the Solution of the Protein Structure Prediction Problem. *J. Biol. Chem.* **2021**, *297*, 100870. [CrossRef] [PubMed]
22. Smith, S.T.; Meiler, J. Assessing Multiple Score Functions in Rosetta for Drug Discovery. *PLoS ONE* **2020**, *15*, e0240450. [CrossRef]
23. Alford, R.F.; Leaver-Fay, A.; Jeliazkov, J.R.; O'Meara, M.J.; DiMaio, F.P.; Park, H.; Shapovalov, M.V.; Renfrew, P.D.; Mulligan, V.K.; Kappel, K.; et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048. [CrossRef] [PubMed]
24. Webb, M.A.; Delannoy, J.Y.; de Pablo, J.J. Graph-Based Approach to Systematic Molecular Coarse-Graining. *J. Chem. Theory Comput.* **2018**, *15*, 1199–1208. [CrossRef] [PubMed]
25. Chakraborty, M.; Xu, C.; White, A.D. Encoding and Selecting Coarse-Grain Mapping Operators with Hierarchical Graphs. *J. Chem. Phys.* **2018**, *149*, 134106. [CrossRef] [PubMed]
26. Unhelkar, M.H.; Duong, V.T.; Enendu, K.N.; Kelly, J.E.; Tahir, S.; Butts, C.T.; Martin, R.W. Structure Prediction and Network Analysis of Chitinases from the CApe Sundew, DRosera Capensis. *Biochim. Biophys. Acta Gen. Subj.* **2017**, *1861*, 636–643. [CrossRef] [PubMed]
27. Duong, V.T.; Unhelkar, M.H.; Kelly, J.E.; Kim, S.H.; Butts, C.T.; Martin, R.W. Protein Structure Networks Provide Insight into Active Site Flexibility in Esterase/Lipases from the Carnivorous Plant *Drosera Capensis*. *Integr. Biol.* **2018**, *10*, 768–779. [CrossRef]
28. Yu, Y.; Grazioli, G.; Unhelkar, M.; Martin, R.W.; Butts, C.T. Network Hamiltonian Models Reveal Pathways to Amyloid Fibril Formation. *Nat. Sci. Rep.* **2020**, *10*, 15668. [CrossRef]
29. Bejagam, K.K.; Singh, S.; An, Y.; Deshmukh, S.A. Machine-Learned Coarse-Grained Models. *J. Phys. Chem. Lett.* **2018**, *9*, 4667–4672. [CrossRef]
30. Boninsegna, L.; Gobbo, G.; Noé, F.; Clementi, C. Investigating Molecular Kinetics by Variationally Optimized Diffusion Maps. *J. Chem. Theory Comput.* **2015**, *11*, 5947–5960. [CrossRef]
31. Lemke, T.; Peter, C. Neural Network Based Prediction of Conformational Free Energies—A New Route Toward Coarse-Grained Simulation Models. *J. Chem. Theory Comput.* **2017**, *13*, 6213–6221. [CrossRef] [PubMed]
32. Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N.E.; De Fabritiis, G.; Noé, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2019**, *5*, 755–767. [CrossRef]
33. Zhang, L.; Han, J.; Wang, H.; Car, R.; E, W. Deepcg: Constructing Coarse-Grained Models via Deep Neural Networks. *J. Chem. Phys.* **2018**, *149*, 034101. [CrossRef] [PubMed]
34. Vivekanandan, S.; Brender, J.R.; Lee, S.Y.; Ramamoorthy, A. A Partially Folded Structure of Amyloid-Beta (1–40) in an Aqueous Environment. *Biochem. Biophys. Res. Commun.* **2011**, *411*, 312–316. [CrossRef] [PubMed]
35. Phillips, J.C.; Braun, R.; Wang, W.; Gumbart, J.; Tajkhorshid, E.; Villa, E.; Chipot, C.; Skeel, R.D.; Kalé, L.; Schulten, K. Scalable Molecular Dynamics with NAMD. *J. Comput. Chem.* **2005**, *26*, 1781–1802. [CrossRef] [PubMed]
36. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
37. Huang, J.; Rauscher, S.; Nawrocki, G.; Ran, T.; Feig, M.; de Groot, B.L.; Grubmüller, H.; MacKerell, A.D.J. CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins. *Nat. Method* **2017**, *14*, 71–73. [CrossRef]
38. Martyna, G.J.; Tobias, D.J.; Klein, M.L. Constant Pressure Molecular Dynamics Algorithms. *J. Chem. Phys.* **1994**, *101*, 4177–4189. [CrossRef]
39. Feller, S.E.; Zhang, Y.; Pastor, R.W.; Brooks, B.R. Constant Pressure Molecular Dynamics Simulation: The Langevin Piston Method. *J. Chem. Phys.* **1995**, *103*, 4613–4621. [CrossRef]
40. Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14*, 27–28, 33–38. [CrossRef]
41. Handcock, M.S.; Hunter, D.R.; Butts, C.T.; Goodreau, S.M.; Morris, M. Statnet: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *J. Stat. Softw.* **2008**, *24*, 1548. [CrossRef] [PubMed]

42. Butts, C.T. network: A Package for Managing Relational Data in R. *J. Stat. Softw.* **2008**, *24*, 1–36. [CrossRef]
43. Grant, B.J.; Rodrigues, A.P.; ElSawy, K.M.; McCammon, J.A.; Caves, L.S. Bio3D: An R Package for the Comparative Analysis of Protein Structures. *Bioinformatics* **2006**, *22*, 2695–2696. [CrossRef]
44. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2018.
45. Alvarez, S. A Cartography of the Van Der Waals Territories. *Dalton Trans.* **2013**, *42*, 8617–8636. [CrossRef]
46. Chollet, F.; Gibson, A.; Allaire, J.J.; Rahman, F.; Branchaud-Charron, F.; Lee, T.; de Marmiesse, G. Keras. 2015. Available online: https://keras.io (accessed on 2 January 2020).
47. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A System for Large-Scale Machine Learning. In Proceedings of the 12th Symposium on Operating Systems Design and Implementation 16, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
48. Kotila, M. Autonomio v.0.3.2 User Manual—Autonomio Latest Documentation. 2017. Available online: https://autonom.io (accessed on 2 January 2020)
49. 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.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. [CrossRef]
50. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A Visualization System for Exploratory Research and Analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612. [CrossRef]
51. Zhang, Y.; Skolnick, J. Scoring Function for Automated Assessment of Protein Structure Template Quality. *Proteins Struct. Funct. Bioinform.* **2004**, *57*, 702–710. [CrossRef]
52. Mariani, V.; Biasini, M.; Barbato, A.; Schwede, T. lddt: A Local Superposition-Free Score for Comparing Protein Structures and Models Using Distance Difference Tests. *Bioinformatics* **2013**, *29*, 2722–2728. [CrossRef] [PubMed]
53. Zemla, A. LGA: A Method for Finding 3d Similarities in Protein Structures. *Nucleic Acids Res.* **2003**, *31*, 3370–3374. [CrossRef]
54. Kryshtafovych, A.; Schwede, T.; Topf, M.; Fidelis, K.; Moult, J. Critical Assessment of Methods of Protein Structure Prediction (CASP)—Round Xiii. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 1011–1020. [CrossRef]
55. Kaźmierkiewicz, R.; Liwo, A.; Scheraga, H.A. Energy-based Reconstruction of a Protein Backbone from Its Alpha-carbon Trace by a Monte-carlo Method. *J. Comput. Chem.* **2002**, *23*, 715–723. [CrossRef] [PubMed]
56. Wang, W.; Gómez-Bombarelli, R. Coarse-Graining Auto-Encoders for Molecular Dynamics. *Npj Comput. Mater.* **2019**, *5*, 125. [CrossRef]
57. Senior, A.W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A.W.; Bridgland, A.; et al. Improved Protein Structure Prediction Using Potentials from Deep Learning. *Nature* **2020**, 577, 706–710. [CrossRef] [PubMed]
58. Scholkopf, B.; Smola, A.J. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2001.