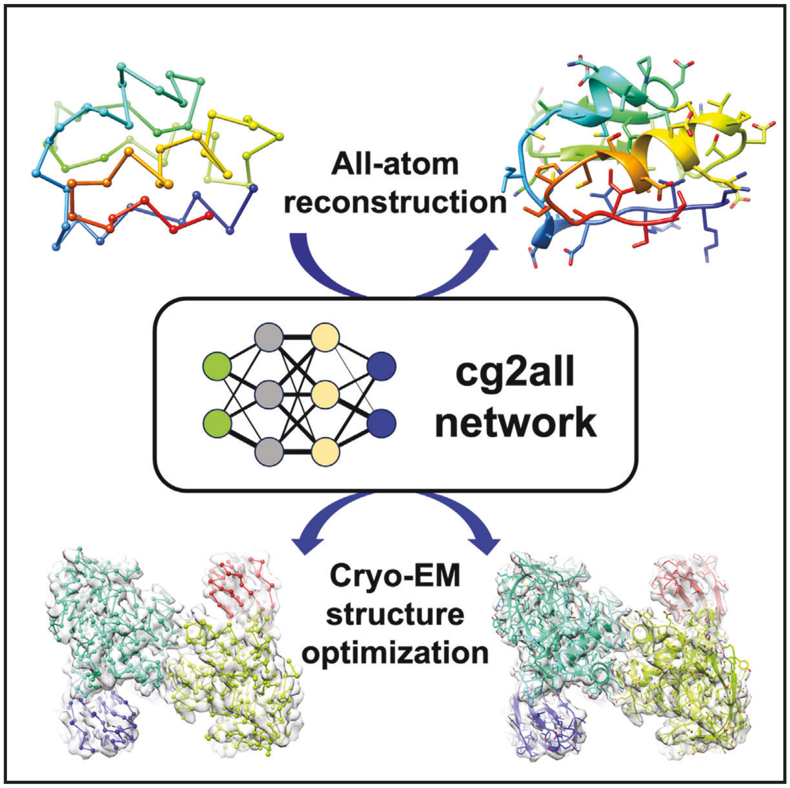
Structure

Resource

One bead per residue can describe all-atom protein structures

Graphical abstract



Highlights

- Accurate all-atom reconstruction from coarse-grained representations of proteins
- Addition of atomistic detail to lower resolution protein models
- Rapid refinement against cryo-EM maps via multi-scale protocol

Authors

Lim Heo, Michael Feig

Correspondence

mfeiglab@gmail.com

In brief

Heo and Feig describe a machine-learning-based protocol for accurate recovery of atomistic detail from coarse-grained models of proteins. The method can be applied to add atomistic details to low-resolution experimental structures. It also facilitates rapid structure determination based on cryo-EM maps.

Heo & Feig, 2024, Structure *32*, 1–15
January 4, 2024 © 2023 Elsevier Ltd.
https://doi.org/10.1016/j.str.2023.10.013

CellPress

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

Structure

CellPress

Resource

One bead per residue can describe all-atom protein structures

Lim Heo[1] and Michael Feig[1,2,*]
[1]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA
[2]Lead contact
*Correspondence: mfeiglab@gmail.com
https://doi.org/10.1016/j.str.2023.10.013

SUMMARY

Atomistic resolution is the standard for high-resolution biomolecular structures, but experimental structural data are often at lower resolution. Coarse-grained models are also used extensively in computational studies to reach biologically relevant spatial and temporal scales. This study explores the use of advanced machine learning networks for reconstructing atomistic models from reduced representations. The main finding is that a single bead per amino acid residue allows construction of accurate and stereochemically realistic all-atom structures with minimal loss of information. This suggests that lower resolution representations of proteins may be sufficient for many applications when combined with a machine learning framework that encodes knowledge from known structures. Practical applications include the rapid addition of atomistic detail to low-resolution structures from experiment or computational coarse-grained models. The application of rapid, deterministic all-atom reconstruction within multi-scale frameworks is further demonstrated with a rapid protocol for the generation of accurate models from cryo-EM densities close to experimental structures.

INTRODUCTION

Proteins play central roles in biological processes, and their behavior is often studied at the molecular level to understand biological function. Structural resolution at an atomistic level is the gold standard for experiments and computation alike. Experimental methods such as X-ray crystallography,[1,2] nuclear magnetic resonance,[3] and cryogenic electron microscopy (cryo-EM)[4] allow the construction of structural models in atomistic detail, but achieving such high resolution requires significant effort.[5,6] Computational modeling and simulations also typically require all-atom representations of the protein to achieve maximum accuracy and to gain detailed mechanistic insights.[7,8] Atomistic modeling remains computationally expensive, though, limiting practical applications,[9] even with the latest high-performance computing platforms and simulation accelerators (e.g., Anton or CUDA).[10] Similarly, it is also very demanding to train machine-learning-based methods for directly predicting atomistic models[11–13] and conformational ensembles.[14,15]

Coarse graining (CG) of protein structures is a common strategy to overcome various challenges.[7] When interpreting experimental data, reduced representations may be a natural fit to match lower experimental resolutions. In computational applications, CG models greatly increase efficiency by reducing the number of particles. CG representations may range from single beads per protein[16,17] to residue-based models[18,19] and multiple sites per amino acid residue.[20–24] Lower resolution models of experimental data often default to Cα traces. In computational applications, the choice of resolution may depend on the questions that are being investigated as model accuracy and transferability depend on the degree of CG.[7,25]

While the coordinate mapping from atomistic levels to CG representations is straightforward, the reverse mapping is in principle ill-defined because of dynamics in degrees of freedom that are not reflected in a reduced representation. For example, side chains may fluctuate for the same set of Cα coordinates. However, the reverse mapping is better defined when the goal is to map CG models to an atomistic coordinate representation of the ensemble-averaged dominant state, essentially akin to how most experimental structures of proteins are meant to be interpreted.

To recover atomistic information from CG models, all-atom reconstruction algorithms have been developed with different strategies depending on the CG representation. For a united-atom model, which omits only hydrogen atoms, missing hydrogens can be placed using their pre-defined local geometries.[26] An all-atom structure can still be generated relatively accurately and quickly from higher resolution CG models such as PRIMO or MARTINI, based on geometry-based reconstruction rules.[21,27] The reconstruction of all-atom structures from coarser representations such as Cα-traces is more complex. Methods such as PULCHRA[28] and REMO[29] convert Cα-traces to all-atom structures by first rebuilding the backbone atoms before predicting side-chain orientations. There is also an

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

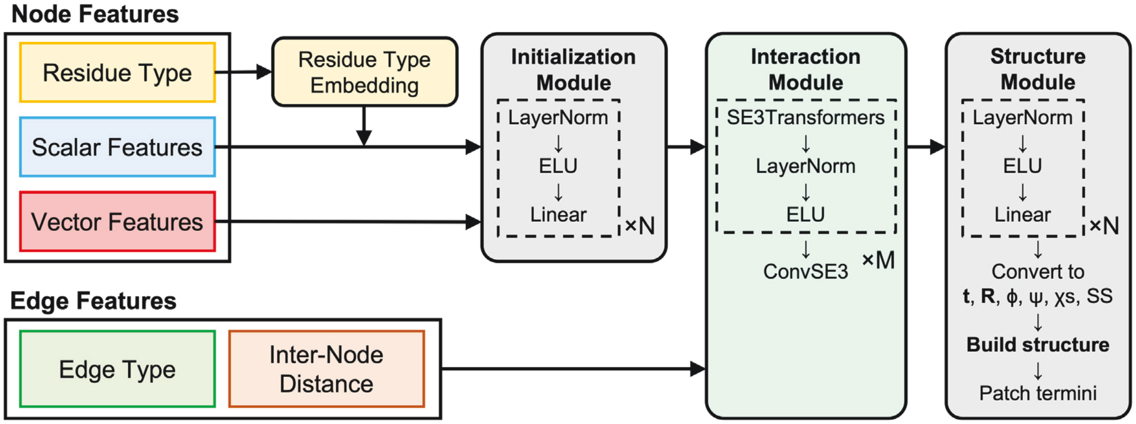


Figure 1. Architecture of coarse-grained to all-atom structure conversion model
See also Figures S1 and S2.

additional set of methods that focus only on rebuilding side chains given a protein backbone.[30] These methods typically rely on pre-defined backbone fragment libraries, side-chain rotamer libraries,[30,31] or other empirical information derived from known structures. In most cases, extensive optimization is then required to avoid clashes and find energetically optimal structures.[32] Most recent methods also adopted machine learning approaches for rebuilding side chains without rotamer libraries.[33,34] Nevertheless, the resulting reconstructions may retain significant deviations from correct all-atom structures when only $C\alpha$ atoms are available as input. The reconstructions may also vary from one run to another if they depend on stochastic optimization techniques. The relatively poor accuracy when reconstructing atomistic coordinates from lower resolutions has limited the full interpretation of experimental data that do not directly provide atomistic details and hindered effective implementations of multi-scale sampling methods that are both efficient and thermodynamically consistent with sampling at all-atom levels.[8,35,36]

In the meantime, the recent success of accurate structure prediction via machine learning methods[11–13] has demonstrated that deep neural network models can effectively learn from the large amount of known structures how to generate atomistic models just from amino acid sequences. This suggests that it should also be possible using similar approaches to reconstruct atomistic coordinates at high resolution if lower resolution structural information is available as additional input.

Inspired by AlphaFold2,[11] we trained an SE(3)-equivariant graph neural network model for reconstructing all-atom detail from lower representations. Like AF2, the model utilizes rigid-body blocks for generating 3D structures from predicted features, but the model was extended to better describe hydrogen atoms and secondary structure dependencies. The network learned structural features of backbone and side-chain atoms from known protein conformations, but also incorporates physical constraints necessary to produce realistic all-atom structures. The model is applicable to a range of CG models such as $C\alpha$-traces, traces of residue center-of-mass model, and MARTINI models.[20] It provides all-atom reconstructions at much higher accuracy than with previous methods, better than 1 Å for heavy atoms from only $C\alpha$ atoms and better than 0.5 Å with a single site at residue centers of mass. This suggests that structural details of proteins at a resolution close to experimental accuracy can be captured essentially with a single site per residue if the current knowledge of protein structures is taken advantage of via machine learning.

The all-atom reconstruction via the machine learning framework is fast and deterministic, and since gradients are available via back propagation, it is straightforward to map energies and constraints at the all-atom level directly to the CG representation. It is therefore possible to sample a residue-based model guided directly by all-atom forces via backmapping through the all-atom reconstruction network. As a proof of principle, we demonstrate practical value in the rapid refinement of all-atom coordinates against intermediate- and low-resolution cryo-EM densities. The protocol achieves comparable accuracy to traditional all-atom simulation-based approaches but with much reduced computational effort.

RESULTS

Accurate reconstruction of all-atom structures from coarse-grained representations

We trained SE(3)-equivariant machine learning models, called cg2all, to reconstruct all-atom structures of proteins from CG representations (*cf.* STAR methods section). The network architecture is shown in Figure 1. The initialization module processes input features (Figure S1; and Table S1) to encode a protein structure in a CG representation and the corresponding residue-type information (Methods S1, and Algorithm S1). An interaction module based on SE(3)-Transformers[37] exchanges scalar and vector encodings between residues to infer inter-residue relationships (Methods S1, and Algorithm S2). The model adopts a rigid-body block-based all-atom structure building method, analogous to the method used for AlphaFold2 structure building.[11] A structure module predicts values for the rigid-body block-based structure building method: translation and rotation of backbone rigid-body blocks consisting of N, $C\alpha$, and C atoms and torsion angles to place the remaining atoms of residues

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

(Methods S1, and Algorithm S3). The model was trained using 5,690 structures (PDB 6k set) or 28,914 structures (PDB 29k set) for 300 and 120 epochs, respectively, using an Adam optimizer with learning rates of 0.01 for parameters for torsion angle predictions and 0.001 for the others (*cf.* STAR Methods). The target loss function combines a target-dependent loss to minimize differences between the target structure and a reconstruction and a data-independent loss to reduce physically unrealistic structural features (e.g., atomic clashes and rotamer outliers) (*cf.* STAR Methods). Model variations with different hyperparameters were explored to determine the optimal model architecture (*cf.* STAR Methods). An ablation study was further carried out to determine optimal input features and loss function components. Only results with the optimized architecture are described subsequently.

The progress in learning protein structural features by the Cα--trace model proceeded in the order of distance from the Cα atom (Figure S2). When progress in the recovery of structural features was tracked for every 10 epochs, we observed that backbone-related features such as the Ramachandran angle, the result of translation and three-dimensional rotation of backbone rigid-body blocks, were saturated during the earlier epochs of the training. At epoch 10, the Ramachandran map already resembled that from experimental structures, and it changed little afterward. On the other hand, learning side-chain torsion angles required many more epochs. At epoch 20, predictions of χ_1 angles became reasonably accurate, and the model started to learn χ_2 angles and a little bit of χ_3 angles. At epoch 60, more states of χ_2 angles were captured, and there was progress in χ_3 angle predictions. At the end of the training at epoch 120, learning of most structural features converged. Many structural features were learnt by the model; however, a few torsion angles such as those for Arg/Lys χ_4 angles could not be learnt in the end. Consequently, the loss in structural information upon the conversion from all-atom structure to Cα-trace was most significant for structural features that were far from the Cα atom. Consequently, backbone features could be learnt quickly, while torsion angles farther from the Cα atoms were slow and sometimes incomplete. In contrast, the learning progress with the residue center-of-mass model, which contains richer input information, was much faster overall and more complete than that with the Cα-trace model (Figure S2). Most structural features started to converge at epoch 20 and were almost completed at epoch 60.

Models were generated for the reconstruction from Cα atoms, from all backbone atoms, from a single particle at the center of mass of an amino acid residue or at the center of the side chain, from the MARTINI model with several beads per residue,[20] and from the higher resolution CG model PRIMO.[21] The generated all-atom models were evaluated on a test set in terms of root-mean-square deviations (RMSDs) and side-chain torsion accuracy with respect to the original reference structures as well as MolProbity[38] scores to check stereochemical quality (Table 1). Because the machine learning model estimates internal parameters that are then used to reconstruct all-atom detail via rigid body reconstruction,[11] using the parameters from the experimental all-atom structures as input for the rigid body reconstruction provides an upper limit on the accuracy that can be achieved theoretically. Because of the reduced degree of freedom using rigid-body blocks, it was not possible to reproduce the exact distribution of the bonded geometries. (Figure S3) Thus, this ideal rigid body reconstruction resulted in slight deviations of atomic coordinates (heavy-atom RMSDs of 0.16 Å) and increase in MolProbity scores to 1.81 (Table 1), but both are still within experimental accuracy of about 0.2 Å RMSD for coordinates in X-ray structures of proteins whereas MolProbity scores below 2 are expected for structures derived from data at better than 2 Å resolution.[38–41] The reconstruction from the highest resolution CG model, PRIMO, reaches the theoretical maximum accuracy and even lower MolProbity scores are obtained with slightly fewer clashes (Table 1). That may be expected since PRIMO was designed to retain maximum information from all-atom representations. However, even with lower resolution models, it is still possible to recover accurate all-atom structures. Reconstruction from MARTINI models resulted only in a slight loss of accuracy (0.31 Å RMSD) and only slightly increased MolProbity scores. Remarkably, even a single bead per residue, located at the center of a residue or at the side-chain center, still allows accurate reconstruction of all-atom details (<0.5 Å RMSD) without significant compromise of stereochemical quality. If the CG site is located at the Cα position, as is common in many CG representations, the loss of accuracy is greater with the average heavy-atom RMSD approaching 1 Å RMSD. The reason is that it becomes more difficult to accurately position side chains if only backbone atoms are given, especially side chains on the surface that are inherently free to sample different rotamer states (Figure 2A). On the other hand, residue center-of-mass models contain information about the location of the side-chain position and therefore side chains can be placed more accurately, even on the surface (Figure 2B). In our approach, we allowed the coordinates of the input low-resolution model to vary during the all-atom reconstruction so that small errors at the CG level could be corrected automatically. For reduced models based on experimental structures, this made little difference when compared to a protocol where initial CG coordinates remained fixed (Table 1).

A more detailed analysis on the models reconstructed from Cα-traces and residue center-of-mass models shows that backbone and side-chain torsion angles are closely matched (Figure S4). The backbone angles (Cα–C–N and C–N–Cα) are also closely matched (Figure S3), but the peptide bond (C–N) showed a somewhat larger standard deviation of 0.027 Å around the average distance of 1.322 Å in the reconstructed all-atom structures from Cα-traces (or 1.330 ± 0.048 Å from residue center-of-mass models) compared to a standard deviation of 0.008 Å around an average of 1.331 Å in the experimental structures. The greater variation in the only flexible backbone bond distance in the rigid-body reconstruction procedure likely compensated for keeping all other bonds rigid. However, one should also note that all but the very highest resolution experimental structures are solved using molecular modeling programs that bias experimental structures toward expected bond lengths.[42] This likely results in apparently reduced variations of such bonds. Finally, *cis*-peptide ω torsion angles were not produced for non-pre-proline residues (Figure S3).

The machine-learning-based all-atom reconstruction via cg2all performed significantly better than most of the previously proposed all-atom reconstruction schemes across all metrics

Table 1. Performance of conversion to all-atom structures from CG models with cg2all

		RMSD[a]		χ-angle accuracy[a,b]		MolProbity[a]			
CG representation	Maximum number of beads per residue	Backbone [Å]	Heavy atom [Å]	χ_1 [%]	χ_{1+2} [%]	Score	Clash score	Rama favor [%]	Rotamer outlier [%]
Experimental structure	–	–	–	–	–	1.25 (0.34)	3.2 (2.4)	97.9 (1.2)	1.4 (1.2)
Rigid body reconstruction[c]	–	0.03 (0.01)	0.16 (0.03)	–	–	1.81 (0.29)	12.2 (4.8)	97.4 (1.3)	1.4 (1.2)
Cα	1	0.18 (0.05)	0.96 (0.12)	86.2 (3.0)	71.4 (4.8)	2.07 (0.21)	31.2 (9.5)	97.9 (1.1)	0.8 (0.7)
Cα (fixed)[d]	1	0.17 (0.05)	0.93 (0.12)	86.5 (3.0)	71.8 (4.7)	2.13 (0.21)	34.2 (10.3)	98.0 (1.1)	1.0 (0.7)
N, Cα, C	3	0.07 (0.02)	0.83 (0.11)	89.3 (2.8)	75.7 (4.5)	2.09 (0.23)	27.6 (8.5)	97.5 (1.3)	1.1 (0.7)
N, Cα, C (fixed)[d]	3	0.06 (0.02)	0.82 (0.11)	89.4 (2.7)	75.8 (4.5)	2.07 (0.22)	27.8 (8.6)	97.9 (1.2)	1.1 (0.7)
N, Cα, C, O	4	0.04 (0.01)	0.82 (0.11)	89.6 (2.8)	75.6 (4.7)	2.08 (0.22)	26.9 (8.3)	97.4 (1.3)	1.0 (0.8)
N, Cα, C, O (fixed) [d]	4	0.00 (0.00)	0.82 (0.11)	89.7 (2.8)	75.7 (4.6)	2.05 (0.22)	27.3 (8.3)	97.9 (1.2)	1.1 (0.8)
CM[e]	1	0.22 (0.05)	0.46 (0.06)	95.4 (1.9)	85.9 (3.9)	2.00 (0.23)	20.6 (6.2)	97.3 (1.4)	1.1 (0.7)
SC[f]	1	0.29 (0.07)	0.49 (0.06)	92.8 (2.4)	85.6 (4.1)	2.13 (0.28)	22.9 (7.1)	97.0 (1.5)	1.5 (1.0)
Cα + CM[e]	2	0.11 (0.03)	0.39 (0.05)	98.0 (1.3)	88.7 (3.6)	1.97 (0.20)	22.7 (6.8)	97.7 (1.2)	0.8 (0.7)
Cα + CM[e] (fixed) [d]	2	0.10 (0.03)	0.39 (0.05)	98.0 (1.3)	88.8 (3.5)	1.99 (0.21)	23.6 (6.9)	97.7 (1.2)	0.8 (0.7)
Cα + SC[f]	2	0.13 (0.04)	0.40 (0.04)	95.2 (1.9)	88.9 (3.4)	1.93 (0.21)	20.1 (6.2)	97.7 (1.2)	0.9 (0.7)
Cα + SC[f] (fixed) [d]	2	0.12 (0.04)	0.39 (0.04)	95.1 (1.9)	89.1 (3.4)	1.96 (0.21)	21.4 (6.3)	97.6 (1.2)	0.9 (0.7)
MARTINI	5	0.08 (0.02)	0.31 (0.05)	98.8 (1.0)	93.1 (2.8)	1.88 (0.21)	17.2 (5.4)	97.5 (1.2)	0.9 (0.7)
PRIMO	8	0.04 (0.01)	0.18 (0.03)	99.9 (0.2)	99.7 (0.5)	1.72 (0.27)	10.7 (4.2)	97.5 (1.3)	1.1 (0.9)

See also Figures S3 and S4.

[a] The average reconstruction accuracy measures for the test set protein structures (n = 720) are given with their standard deviations in the parentheses.

[b] Side-chain χ-angles were considered accurate when deviations from experimental values were less than 30°.

[c] Experimental structures were reconstructed with the rigid body blocks using residue orientations and torsion angles from the experimental structures.

[d] The atomic coordinates in the input files were preserved, while the original method does not. For instance, cg2all model for "Cα (fixed)" generates output structures with the exact same Cα coordinates of the input structures. On the other hand, the original cg2all model generates slightly altered Cα coordinates.

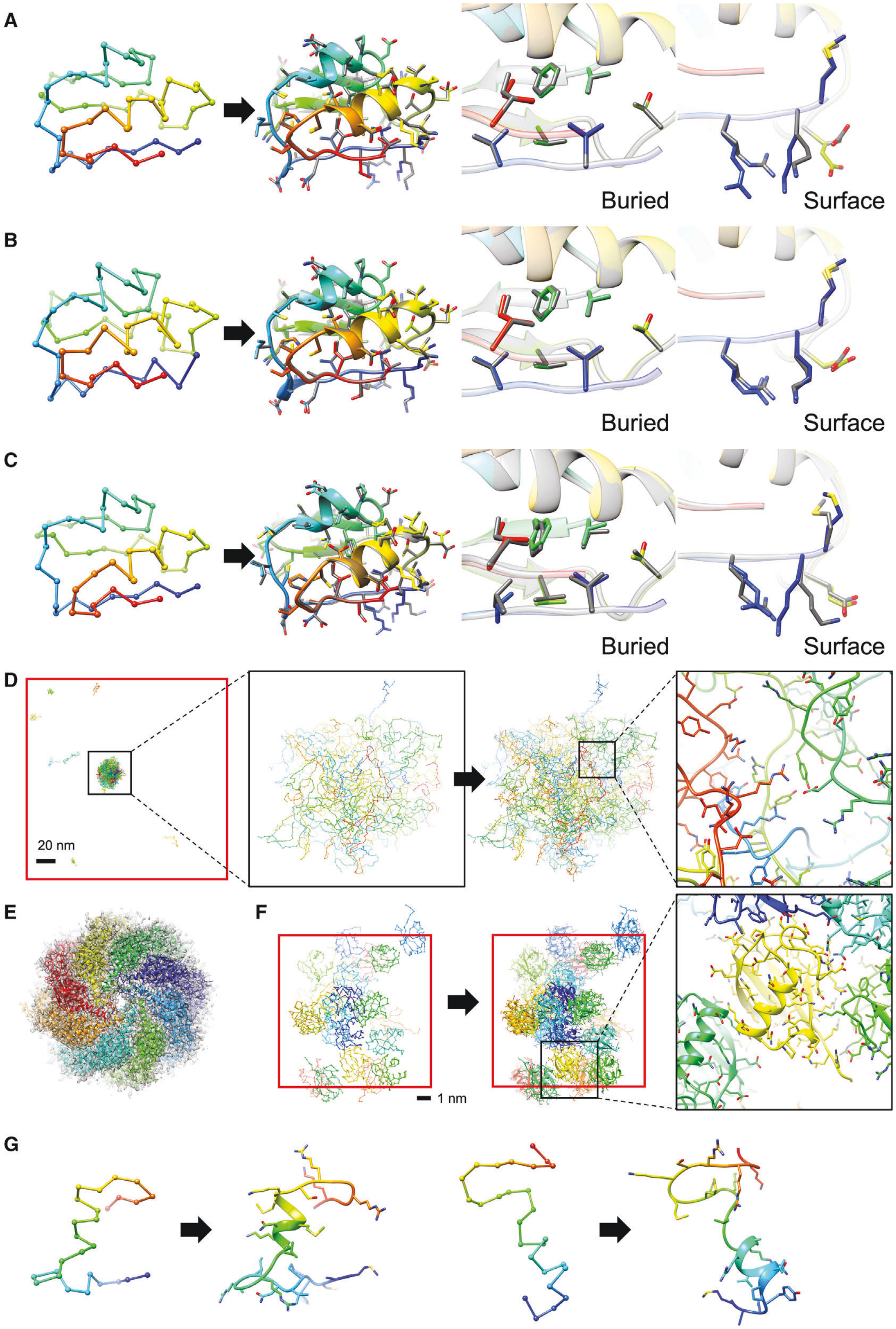
[e] A bead located at the center-of-mass of an amino acid.

[f] A bead located at the center-of-mass of side-chain atoms. For glycine, it is located at the position of Cα atom.

(Table 2). No other method achieved significantly better than 1 Å RMSD for heavy atoms, when reconstructing from a Cα-trace or a residue center-of-mass model, and even the higher resolution MARTINI model.[27] The other older methods for those CG representations also produced structures with significant clashes and higher MolProbity scores, despite energy-guided optimization to avoid clashes. We furthermore tested the widely used rotamer-based method SCRWL4[30] for placing side chains in combination with backbone atoms generated with cg2all or other methods. Using SCWRL, MolProbity scores were generally improved, even slightly over cg2all, but the accuracy decreased compared to cg2all, especially for MARTINI and center-of-mass-based reconstructions. The reason is that SCWRL's side-chain modeling only uses backbone coordinates as input and is based strictly on a rotamer library that, by design, prevents outliers that are occasionally found in experimental structures.

In comparison with other side-chain prediction methods, where the input is a complete protein backbone, cg2all performed similar or slightly better than other methods in terms of accuracy. In comparison with SCWRL, cg2all showed better reconstruction accuracy in terms of heavy-atom RMSD, χ_1 and χ_{1+2} angle accuracies. However, SCWRL generates almost no rotamer outliers as it relies on a rotamer library where outliers are prevented by design. However, this might limit the reconstruction accuracy because real structures do contain outliers. A state-of-the-art machine-learning-based method, AttnPacker,[33] reconstructed side-chain structures with high accuracy in terms of heavy-atom RMSD. However, when using their method, we found that the reconstructed models had not just numerous rotamer outliers but also highly unrealistic side-chain bond lengths and angles. (Figure S5); thus, they required additional local energy minimization to correct poor stereochemistry.

Because all-atom reconstruction via cg2all is achieved with a single forward pass without any iterative optimization, the computational cost is low, on the order of seconds (Table 3; and Figure S6). To better understand the computational performance, some additional analysis is necessary. A reconstruction run with cg2all consists of (1) loading Python libraries, (2) loading a PyTorch model, (3) reading an input PDB file followed by preprocessing, (4) a forward pass through the model, and (5) writing an output PDB file. Loading the PyTorch model takes around 2.7 s, limited by I/O speed. The computational cost for the remaining steps is linearly dependent on the number of residues of the system, with an average of less than 4 s for the test set with a single CPU thread and less time when multiple threads or a GPU was used. Using a GPU, incurred additional overhead and is not efficient for a single average-size reconstruction, but there is a significant benefit for very large systems or when processing many snapshots with alternate conformations that can be processed



(legend on next page)

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

simultaneously on a GPU. For a single reconstruction on a single thread, only PULCHRA was faster than cg2all. REMO took about the same total time, whereas other methods required much more time. For reconstruction of side-chain atoms from protein main-chain structures, cg2all and SCWRL were comparably fast. On the other hand, the neural network method AttnPacker required significantly more computational cost as its network has two orders of magnitude more parameters (208 M vs. 4.05 M), whereas the necessary local energy minimization step took additional computational time.

Accurate all-atom reconstruction from simulation models

The reconstruction of all-atom detail from reduced representations of experimental structures, as described in the previous section, may be considered an ideal scenario. We further tested how well all-atom resolution can be recovered from models generated via simulations. Two sets of models were considered. The first set consisted of snapshots that were extracted from all-atom molecular dynamics (MD) simulations; in the second set, the MD-based snapshots were further energy-minimized structures using a residue-based CG model, COCOMO.[18] Here, the all-atom MD snapshots serve as the reference for determining accuracy of the reconstruction protocol. All-atom reconstruction from Cα-traces of MD simulation snapshots was still accurate, but the accuracy became slightly worse than for experimental structures, with higher RMSD values of 1.18 Å for heavy atoms and lower side-chain torsion accuracies (Table 4). Once the snapshots were minimized with COCOMO at the CG level, the all-atom reconstruction RMSD further increased to 1.34 Å (Table 4) with respect to the initial all-atom MD snapshots. However, that may be expected because the CG minimization by itself resulted in deviations of 0.30 Å for the Cα positions from the initial all-atom MD snapshots. In either case, the reconstructed all-atom models had again low clash scores and very low rotamer outliers and were much closer to the reference atomistic structural models than those generated with other methods (Table 4; and Figure 2C). Thermal fluctuations in the simulations led to broadened bond geometry distributions (Figure S7) and more rotamer outliers compared to experimental structures. However, since cg2all was trained to generate experimental structure-like conformations, the broader distributions were not completely reproduced (Figure S7). This explains at least in part the slightly lower reconstruction accuracy for simulation-based models.

A related practical question is whether the reconstructed models from cg2all are more suitable for starting atomistic simulations. The reconstructed all-atom structure was suitable for further usages such as all-atom MD simulations. Larger systems require extensive computational cost to get their equilibrated system or systems in desired states via all-atom MD simulations. Alternatively, one may attempt to reach an equilibrium state such as liquid-liquid phase separation formation described in Figure 2D using CG simulations[18] and continue atomistic simulations from the state. We briefly examined the approach by performing atomistic MD simulations starting from reconstructed all-atom structures (Figure 3). We carried out an atomistic MD simulation for 50 ns and minimized the conformation from the last snapshot using the CG model COCOMO as a hypothetical CG simulation result for which the atomistic MD snapshot serves as a reference. The COCOMO-minimized structure was converted to an all-atom structure using our method or PULCHRA,[28] respectively. Then, the converted structures were equilibrated again and continued atomistic simulations. The simulation results were compared with another set of simulations that simply continued simulations from the last snapshot. We hypothesized that the protein structure would quickly show instabilities at the beginning of the simulation if the conversion was not producing models of sufficient quality. After conversion with cg2all, the protein structure remained stable and well-folded during the first 10 ns, as the continued simulation did. The average Cα-RMSD to the initial conformation was 1.98 Å after 10 ns (*cf.* 1.35 Å for the simulations from the last snapshot). On the other hand, with the reconstructed structure by PULCHRA, due to steric clashes, conformations deviated from the initial conformation significantly, starting at early stages of the simulations (2.94 Å Cα-RMSD with respect to the initial conformation on average after 10 ns). Consequently, residue-wise fluctuations throughout the atomistic simulation were very similar between the sets of simulations from the last snapshot and the model by our method, while an initial model from PULCHRA resulted in significantly higher fluctuations due to initial destabilization caused by steric clashes.

Figure 2. Examples of conversion from CG models to all-atom models

(A–C) Recovery of all-atom structure from Cα-trace (A), residue center-of-mass model (B), and a Cα-trace of snapshot of all-atom MD simulation after minimization with COCOMO (C). The CG model and recovered structure are shown as rainbow-colored cartoon and stick representation (from blue to red for N- to C termini). Their reference structures, an X-ray crystal structure (PDB ID: 1vjw[62]) for (A and B) and an all-atom MD snapshot (C) are shown in gray.

(D) Conversion of a COCOMO[18] simulation snapshot from a simulation of liquid-liquid phase separation (LLPS) of LAF-1 RGG peptides at 0.042 mM.[63] Each peptide consisted of 168 residues, and there were 84 monomers (14,112 residues in total). They are shown in different colors. The phase-separated particles at the CG level and a local region after the conversion are magnified to present detailed structure information (black boxes). The conversion took 33.5 s using 16 CPU threads.

(E) Building an all-atom model from a medium-resolution cryo-EM Cα-trace (PDB ID: 3iyg,[43] EMDB ID: 5148, resolution: 4.0 Å). Each chain is depicted in a different color, and the electron density is overlaid as transparent gray voxels. The density map correlation with the all-atom model was 0.723 (vs. 0.647 with the Cα-trace). The conversion of the 4,134-residue protein took 11.9 s using 16 CPU threads.

(F) Conversion of a CG MD simulation trajectory of folded proteins using COCOMO. There are 24 ubiquitins (1,824 residues in total) in the simulation box (shown in red) with a width of 10.76 nm, which resulted in a concentration of 32 mM (274.1 g/L). Each monomer is shown in a different color. A local region after the conversion is zoomed in to show atomistic details of interactions between proteins (black boxes). The conversion of 10,000 frames took 1,774 s in total using the "cuda" environment with a batch size of 4 (0.15 s/frame for the forward pass only).

(G) Models generated by idpGAN[14] for an intrinsically disordered protein (UniProt ID: Q9EP54, 27 residues). The conversion of 25,000 models took 64.6 s in total using the "cuda" environment with a batch size of 250 (1.3 ms/model for the forward pass only). Hydrogens were reconstructed but are omitted for clarity.

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

Table 2. Comparison of all-atom reconstruction accuracies with different methods

		RMSD[a]		χ-angle accuracy[a,b]		MolProbity[a]			
Input	Method	Backbone [Å]	Heavy atom [Å]	χ_1 [%]	χ_{1+2} [%]	Score	Clash score	Rama favor [%]	Rotamer outlier [%]
Cα	cg2all	0.18 (0.05)	0.96 (0.12)	86.2 (3.0)	71.4 (4.8)	2.07 (0.21)	31.2 (9.5)	97.9 (1.1)	0.8 (0.7)
	w/SCWRL[c]		1.06 (0.14)	83.2 (3.6)	70.1 (5.2)	**2.00** (0.18)	**28.6** (8.1)	97.9 (1.1)	**0.0** (0.1)
	cg2all (fixed)[d]	**0.17** (0.05)	**0.93** (0.12)	**86.5** (3.0)	**71.8** (4.7)	2.13 (0.21)	34.2 (10.3)	**98.0** (1.1)	1.0 (0.7)
	PULCHRA[e]	0.47 (0.11)	1.57 (0.14)	59.2 (4.0)	39.6 (4.5)	3.79 (0.23)	164.4 (28.5)	86.4 (3.9)	4.9 (1.7)
	w/SCWRL[c]		1.36 (0.14)	73.0 (4.3)	58.4 (5.6)	2.90 (0.20)	67.4 (17.8)	86.4 (3.9)	0.1 (0.2)
	REMO[e,f]	0.81 (0.42)	2.09 (0.41)	43.6 (4.3)	28.5 (5.4)	4.37 (0.21)	200.2 (36.8)	78.4 (7.2)	14.8 (3.8)
	w/SCWRL[c]		1.74 (0.52)	68.5 (6.5)	53.8 (7.5)	3.15 (0.25)	95.7 (43.6)	78.4 (7.2)	0.2 (0.3)
	ModRefiner[e,f]	0.66 (0.20)	1.51 (0.21)	71.8 (4.1)	55.1 (5.6)	2.38 (0.24)	56.5 (19.6)	97.0 (1.7)	0.6 (0.5)
	MODELLER	0.97 (0.88)	2.12 (0.76)	42.8 (4.0)	25.4 (4.0)	3.63 (0.19)	93.3 (15.7)	86.4 (3.4)	6.0 (1.8)
CM	cg2all	**0.22** (0.05)	**0.46** (0.06)	**95.4** (1.9)	**85.9** (3.9)	**2.00** (0.23)	**20.6** (6.2)	**97.3** (1.4)	1.1 (0.7)
	w/SCWRL[c]		1.00 (0.14)	83.9 (3.6)	70.9 (5.2)	2.01 (0.20)	25.5 (7.3)	**97.3** (1.4)	**0.0** (0.1)
	PULCHRA[e]	1.08 (0.08)	1.91 (0.12)	46.6 (4.0)	29.0 (4.1)	4.49 (0.13)	291.6 (36.2)	72.9 (4.4)	10.5 (2.3)
	w/SCWRL[c]	1.08 (0.08)	1.84 (0.12)	55.3 (4.0)	36.3 (4.5)	3.51 (0.11)	176.2 (27.6)	72.9 (4.4)	0.2 (0.3)
	MODELLER [e]	1.63 (1.03)	2.38 (0.94)	43.7 (3.9)	25.5 (3.8)	3.80 (0.18)	123.6 (19.9)	81.8 (3.4)	5.5 (1.8)
N, Cα, C, O	cg2all	0.04 (0.01)	0.82 (0.11)	89.6 (2.8)	75.6 (4.7)	2.08 (0.22)	26.9 (8.3)	97.4 (1.3)	1.0 (0.8)
	cg2all (fixed)[d]	–	0.82 (0.11)	89.7 (2.8)	75.7 (4.6)	2.05 (0.22)	27.3 (8.3)	**97.9** (1.2)	1.1 (0.8)
	SCWRL	–	0.97 (0.13)	85.6 (3.4)	73.0 (5.0)	1.97 (0.19)	26.7 (7.8)	**97.9** (1.2)	**0.0** (0.1)
	AttnPacker	–	**0.61** (0.11)	90.9 (2.9)	73.3 (5.4)	2.32 (0.25)	22.8 (7.7)	**97.9** (1.2)	3.8 (1.5)
	+local min.	–	0.67 (0.11)	**92.4** (2.8)	**80.6** (4.9)	**1.86** (0.28)	**11.9** (3.9)	**97.9** (1.2)	2.0 (1.1)
MARTINI	cg2all	**0.08** (0.02)	**0.31** (0.05)	**98.8** (1.0)	**93.1** (2.8)	**1.88** (0.21)	17.2 (5.4)	**97.5** (1.2)	0.9 (0.7)
	w/SCWRL[c]		0.98 (0.13)	85.2 (3.6)	72.6 (5.2)	2.00 (0.19)	26.1 (7.5)	**97.5** (1.2)	**0.0** (0.1)
	Backward[g]	0.84 (0.07)	1.06 (0.08)	60.9 (5.8)	46.1 (7.2)	2.71 (0.29)	**4.5** (1.7)	86.7 (4.5)	15.9 (3.8)
	w/SCWRL[c]		1.64 (0.16)	64.5 (5.5)	50.4 (6.6)	2.93 (0.22)	74.4 (21.4)	86.7 (4.5)	0.1 (0.2)

See also Figures S3–S5.

[a] The average reconstruction accuracy measures for the test set protein structures (n = 720) are given with their standard deviations in the parentheses.

[b] Side-chain χ-angles were considered accurate when their deviations from experimental values were less than 30°.

[c] Side chains were reconstructed using SCWRL4 after building a backbone structure using other methods (*e.g.*, cg2all, PULCHRA, and REMO).

[d] The atomic coordinates in the input files were preserved, while the original method does not. For instance, cg2all model for "Cα (fixed)" generates output structures with the exact same Cα coordinates of the input structures. On the other hand, the original cg2all model generates slightly altered Cα coordinates.

[e] Chains in multi-chain targets were separately converted to all-atom structures and superposed onto the original Cα-trace.

[f] Conversions of several structures failed because they cannot handle short peptides or were not completed within a reasonable time frame (12 h). Successful conversions for REMO: 684/720; for ModRefiner: 714/720.

[g] Conversion of one structure failed because short peptides (<3 residues) cannot be handled.

Addition of all-atom structural details to low-resolution models

The analysis so far shows that all-atom details can be captured essentially within experimental uncertainties at much reduced representations, up to a single bead at the center of mass of an amino acid. This is possible by drawing on the vast knowledge about protein structures via state-of-the-art machine learning. In turn, this means that all-atom structural details can be provided with high confidence for models that are initially only available at a CG level. Examples where cg2all may be used in practice are shown in Figure 2 and discussed more in the following.

Low-resolution cryo-EM structures are often reported only at the Cα level. All-atom detail can be reconstructed quickly via cg2all (Figure 2E). For a cryo-EM experimental structure (PDB ID: 3iyg[43]) with a resolution of 4.0 Å, the all-atom structure with 4,134 residues or 64,192 atoms was generated within 11.9 s, fast enough to be done on-demand when working with such structures. The generated structure had a higher density map correlation of 0.723 than that of the original Cα-trace, 0.647, but there were some clashes in the atomistic model, presumably because Cα atoms for some residues were packed too tightly in the original structure. Therefore, it may be possible to use the all-atom reconstruction as an indicator of issues with the low-resolution model itself.

We note that low-resolution cryo-EM density maps may make it challenging to trace all residues correctly,[44] especially at flexible regions.[45] If residues are missing, cg2all treats residues before and after chain breaks as C- and N termini to produce reasonable reconstructions for the residues that are resolved since cg2all can only produce atomistic coordinates for residues for which coarse-grained beads are available as input. However, for short segments, initial coarse-grained bead positions could be guessed via interpolation and subsequently refined against

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

Table 3. Average timing for all-atom reconstruction with different methods

Input	Method	Device[a]	Time [s][b]
Cα	cg2all	CPU (1)	6.4 (3.6)
		CPU (4)	4.4 (1.6)
		CPU (16)	3.9 (1.1)
		CUDA	8.7 (1.7)
		Apple Silicon	3.2 (1.8)
	cg2all+SCWRL[c]	CPU (1)	13.0
	PULCHRA	CPU (1)	0.4
	REMO	CPU (1)	6.1
	ModRefiner	CPU (1)	5211
	MODELLER	CPU (1)	98.5
N, Cα, C, O	cg2all	CPU (1)	8.2 (4.4)
	SCWRL	CPU (1)	7.0
	AttnPacker	CPU (1)	197.5 (191.6)
	+local min.		231.4
MARTINI	cg2all	CPU (1)	7.4 (4.4)
	Backward	CPU (1)	33.8

See also Figure S6.

[a] Each method was run on an Intel Xeon Silver 4214 CPUs (2.2 GHz) under Linux with 128 GB of RAM unless noted. The number of threads is given in parentheses. "CUDA" was run on the same machine but using an NVIDIA GeForce RTX 2080 Ti GPU card (11 GB of VRAM). Apple Silicon refers to an Apple Silicon M1 Pro chip with 8-core CPU, 14-core GPU, and 16 GB RAM, but only one CPU thread was used for the inference.

[b] The number of residues of the test set proteins (n = 720) ranged from 50 to 1,176 with an average and a standard deviation of 376 and 242, respectively. For cg2all, the average inference time after loading a PyTorch model is shown in parentheses.

[c] Side chains were reconstructed using SCWRL4 after building a backbone structure using cg2all.

the cryo-EM density using for example the rapid cryo-EM refinement protocol using cg2all (refer to multi-scale sampling for rapid cryo-EM refinement). On the other hand, if longer residue stretches are missing or if the connectivity is incorrect, additional modeling would be needed outside the scope of what cg2all is designed to do in order to generate at least correct coarse-grained models consistent with given experimental data.

Residue-level CG models, such as COCOMO, are increasingly being used to simulate very large systems over long time scales, for example to study protein-protein interactions or liquid-liquid phase separation. Again, cg2all can provide atomistic coordinates from the CG models (Figures 2C**, 2**D, and **2F**). Using this approach, we could obtain an all-atom structure from a snapshot of a CG model of a condensate formed by IDPs. In the example, a 14,112-residue CG system was converted to an all-atom structure with 183,624 atoms in just 33.5 s. This rendered detailed atomic interactions between peptides (e.g., salt bridges between charged side chains) inside the condensate. We note that a fully atomistic simulation of the condensation process is so far impossible to carry out.

Machine-learning-based conformational ensemble generators such as FoldingDiff[46] or idpGAN[14] may be limited to output consisting of Cα traces due to resource constraints, but all-atom models can be obtained rapidly via post-processing by cg2all (Figure 2G). Using a GPU with a large batch size allowed us to convert 25,000 IDP conformations generated via idpGAN to all-atom detail in just 64.6 s. Consequently, a practical strategy for the rapid generation of conformational ensemble via machine learning may be to focus on generating ensembles only at the CG level and leave it up to an all-atom reconstruction scheme as presented here to obtain atomistic ensembles.

Finally, since cg2all can efficiently parallelize all-atom reconstructions on GPUs, entire CG MD simulation trajectories could be rapidly converted to atomistic detail. For example, the conversion of 10,000 frames of a 1,824-residue system takes 1,774 s on a GPU with a batch size of 4 (Video S1). This allows not just the consideration of all-atom detail when sampling with CG models, but also, vice versa, suggests that CG models could be used for lossy data compression. This has been proposed before based on the higher resolution CG model PRIMO,[47] but much greater compression can be achieved if only a single particle per residue is used. For example, a 3.4 GB all-atom trajectory in the DCD format could be compressed into a 210 MB trajectory with single beads, such as Cα or center-of-mass, resulting in a 94% compression ratio. Such high degree of compression could greatly facilitate the public sharing of extensive atomistic trajectories that otherwise remains a significant resource challenge.[48,49]

Multi-scale sampling for rapid cryo-EM refinement

To further demonstrate the potential of cg2all, we turn to the refinement of models against cryo-EM densities. A typical challenge involves the flexible fitting of initial models from crystallography or structure prediction to intermediate- to low-resolution density maps, for which direct atomistic model building is difficult due to insufficient information.[50] A number of protocols are commonly used such as Coot,[51] Isolde,[52] and Direx.[53] The most effective methods to date employ sampling via atomistic simulations, such as the molecular dynamics flexible fitting (MDFF) protocol.[54] This approach is successful but may take on the order of hours to days because of the computational cost of the simulations. Here, we explore the sampling of CG models guided by a density map correlation energy function based on reconstructed all-atom representations that is possible with cg2all. Sampling at the CG level avoids the kinetic barriers that hinder sampling at the atomistic level, whereas using an energy penalty based on atomistic coordinate reconstructions ensures that the optimized CG model is maximally compatible with the experimental data.

The multi-scale approach based on cg2all outperformed local optimization protocols such as energy minimization at the all-atom representation using an atomistic energy function or energy minimization at the CG level using a CG energy function alone across the entire range of map resolutions (Figure 4). With a high-resolution (3 Å) electron density map, local energy minimization of an all-atom model or a CG model is trapped in a local energy minimum because of a rugged energy landscape. Furthermore, local energy minimization of a CG model cannot exploit the high-resolution information from the electron density map. However, our multi-scale approach effectively optimizes structures by minimizing at the CG level where kinetic barriers are low or absent while still targeting the high-resolution data via all-atom reconstruction. During the minimization process, the gradient resulting from the discrepancy between an all-atom model and the target electron density map is backpropagated

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

Table 4. Performance of conversion to all-atom structures from Cα-traces of simulation snapshots

		RMSD[a]		χ-angle accuracy[a,b]		MolProbity[a]			
Input	Method	Backbone [Å]	Heavy atom [Å]	χ_1 [%]	χ_{1+2} [%]	Score	Clash score	Rama favor [%]	Rotamer outlier [%]
Cα from all-atom MD snapshots	MD snapshots					1.45 (0.23)	0.8 (0.5)	93.5 (2.1)	2.7 (1.2)
	cg2all	**0.25** (0.04)	1.18 (0.12)	77.7 (3.4)	58.7 (4.6)	2.31 (0.23)	37.1 (9.8)	**96.4** (1.5)	1.0 (0.7)
	w/SCRWL[c]		1.27 (0.14)	76.2 (3.6)	58.1 (4.8)	**2.20** (0.19)	**32.5** (8.0)	**96.4** (1.5)	**0.0** (0.1)
	cg2all (fixed)[d]	**0.25** (0.04)	**1.15** (0.12)	**77.9** (3.4)	**58.8** (4.6)	2.42 (0.24)	41.0 (10.1)	96.3 (1.5)	1.4 (0.8)
	PULCHRA[e]	0.52 (0.16)	1.70 (0.14)	54.2 (4.0)	32.1 (3.9)	3.84 (0.20)	162.7 (25.5)	85.1 (3.8)	5.2 (1.6)
	w/SCWRL[c]		1.51 (0.15)	68.0 (4.2)	49.7 (4.8)	2.93 (0.17)	67.1 (15.5)	85.1 (3.8)	0.1 (0.2)
	REMO[e,f]	0.94 (0.47)	2.21 (0.46)	44.5 (4.5)	26.4 (4.8)	4.42 (0.21)	197.5 (37.2)	75.1 (7.9)	15.6 (3.9)
	w/SCWRL[c]		1.95 (0.53)	63.5 (5.9)	45.5 (6.5)	3.22 (0.24)	102.4 (45.0)	75.1 (7.9)	0.2 (0.3)
Cα after minimization with COCOMO[g]	cg2all	0.43 (0.13)	1.34 (0.19)	**73.7** (4.2)	**54.8** (5.1)	2.55 (0.23)	44.5 (12.5)	**94.4** (1.9)	1.1 (0.8)
	w/SCWRL[c]		1.43 (0.19)	72.3 (4.1)	53.5 (5.2)	**2.38** (0.18)	**36.1** (10.2)	**94.4** (1.9)	**0.1** (0.1)
	cg2all (fixed)[d]	**0.42** (0.13)	**1.32** (0.19)	73.3 (4.3)	54.3 (5.0)	2.70 (0.24)	50.1 (13.3)	94.1 (2.0)	1.7 (0.9)
	PULCHRA[e]	0.61 (0.18)	1.79 (0.18)	52.5 (4.1)	31.0 (4.1)	3.87 (0.19)	156.8 (25.3)	83.5 (3.9)	5.6 (1.7)
	w/SCWRL[c]		1.62 (0.19)	65.8 (4.4)	47.1 (5.1)	2.95 (0.15)	65.1 (15.3)	83.5 (3.9)	0.1 (0.2)
	REMO[e,f]	1.16 (0.49)	2.40 (0.53)	43.3 (4.7)	25.7 (4.8)	4.50 (0.22)	194.5 (40.1)	69.9 (9.5)	17.4 (4.3)
	w/SCWRL[c]		2.23 (0.59)	59.4 (6.1)	41.3 (6.6)	3.33 (0.25)	116.1 (48.2)	69.9 (9.5)	0.2 (0.3)

See also Figure S7.

[a] The average reconstruction accuracy measures for the test set protein structures (n = 720) are given with their standard deviations in the parentheses.

[b] Side-chain χ-angles were considered accurate when deviations from experimental values were less than 30°.

[c] Side chains were reconstructed using SCWRL4 after building a backbone structure using other methods (*e.g.*, cg2all, PULCHRA, and REMO).

[d] The atomic coordinates in the input files were preserved, while the original method does not. For instance, cg2all model for "Cα (fixed)" generates output structures with the exact same Cα coordinates of the input structures. On the other hand, the original cg2all model generates slightly altered Cα coordinates.

[e] Multi-chain targets were converted chain-by-chain to all-atom structures and superposed onto the original Cα-trace.

[f] Conversions of several structures failed because short peptides could not be handled.

[g] All-atom MD simulation snapshots were considered as the ground truth. The average structure change in Cα-RMSD after minimization using COCOMO model was 0.30 Å.

to the CG level through the cg2all network. This enables the utilization of high-resolution information from the electron density map at the CG level. Because all-atom models are compared with the high-resolution map, CG beads are less likely to become trapped in a local minimum where a correct all-atom structure could not be generated, as the gradient arising from the discrepancy at the all-atom resolution would guide the CG beads away from such a local minimum toward the correct positions. We believe that this multi-scale approach conceptually retains the accuracy of all-atom representation while achieving better performance by moving on a smooth landscape during minimization at the CG level. Here, accuracy refers to comparisons of the generated models with the experimental reference structures and agreement with EM maps. The optimized structures obtained via the cg2all-based multi-scale approach reached comparable Cα-RMSD values to the full MDFF protocol (0.36 vs. 0.35 Å), slightly lower cross-correlation coefficients (CCC, 0.866 vs. 0.881) and slightly larger heavy-atom RMSD values (0.88 vs. 0.74 Å). Importantly, the similar accuracy with cg2all vs. MDFF is achieved in much shorter time (minutes vs. hours). An example for the high model accuracy that can be achieved with cg2all is shown in Figure 5. In the example, several regions in the initial AlphaFold2 model located outside of the 5 Å resolution electron density (indicated by red arrows) with a heavy-atom RMSD of 2.26 Å and a CCC of 0.829. Using the MDFF protocol with the electron density map, the model was optimized to a heavy-atom RMSD of 0.80 Å and has a higher CCC of 0.941. Our multi-scale approach optimized the model to a comparable accuracy, a heavy-atom RMSD of 0.85 Å, and a CCC of 0.936, even though it performed the actual optimization at the CG level. We note that our multi-scale approach achieved the comparable accuracy in 8.9 min, while the MDFF protocol took 8.8 h.

When the electron density map has much lower resolution, such as 10 Å, optimization at the all-atom level becomes less effective, even using MDFF, whereas the CG-based-optimized refinement, via cg2all, still allows structure refinement, and still within minutes. This opens up the possibility for high-throughput model refinement of many lower resolution maps, for example to fit models to maps of dynamics conformational ensembles captured via cryo-EM.

In this proof-of-concept demonstration, we employed a naive CG energy function that only prevents severe clashes between CG beads. Distance restraints between CG beads were applied to keep the protein structure folded; however, this also limited the potential for improvement as it prohibited partial structure unfolding and refolding.[55,56] In future work, this will be addressed by introducing a more sophisticated CG energy function that is capable of not only maintaining folded structures but also allowing significant structural changes.

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

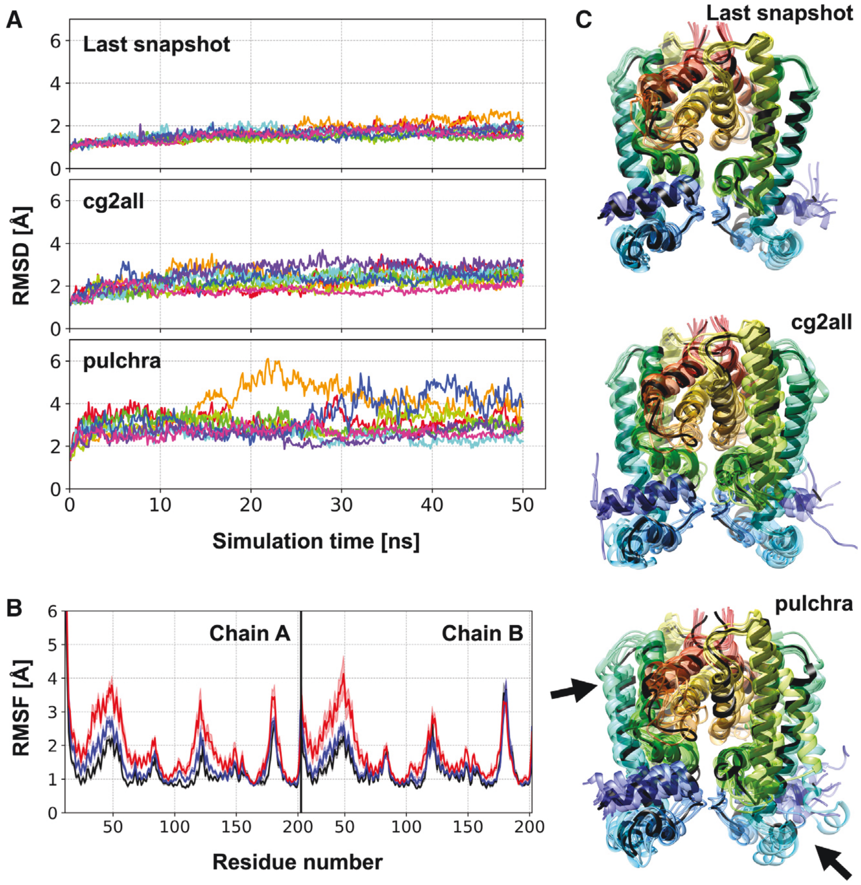


Figure 3. Stability of all-atom MD simulations continued from reconstructed all-atom models

(A) The last snapshot of a dimeric protein (PDB ID: 2ibd) all-atom simulation was locally minimized using COCOMO model, and the minimized Cα-trace was converted to an all-atom model using cg2all and PULCHRA. Then, eight replicas of all-atom simulations were performed starting from the reconstructed all-atom models after an equilibration step. (A) Cα-RMSD trajectories with respect to their starting model. Each trajectory is colored differently.

(B) Residue-wise root-mean-square-fluctuation (RMSF) for the last snapshot, cg2all, and PULCHRA model are shown in black, blue, and red lines. Standard errors of the value are shown with transparent shades. For the RMSF evaluation, the first 10 ns was discarded as the equilibration process.

(C) Ensemble of structures after 10 ns of all-atom simulations (transparent rainbow colors) are compared with their starting structure (black). Highly deviated regions in the PULCHRA simulations are indicated by black arrows.

DISCUSSION

The results presented here show that all-atom details of proteins can be captured essentially within experimental uncertainties with only a single bead per amino acid, especially when placed at the residue center of mass, but perhaps also with a more traditional Cα-trace representation. This is possible now because of advances in machine learning that allow vast information from known structures in the PDB, to be applied toward different objectives, in this case, the accurate reconstruction of all-atom features from low-resolution models. We focus here on reconstructions from widely used low-resolution representations as they are of significant practical relevance, but note that the correspondence between reduced and atomistic coordinate representations could also be improved by optimizing the CG representation itself.[57]

It should be reiterated, that the all-atom reconstructions obtained here focus on finding a chemically realistic representative structure for the most likely time- and ensemble-averaged conformation. Conformational dynamics in degrees of freedoms that are not captured in the reduced representation are effectively averaged out. The resulting structures are therefore experiment-like structures where dynamics may be described only in the form of B-factors. We do not predict B-factors here for the reconstructed structures, but note that other methods are available for estimating B-factors from given atomistic structures.[58]

The approach taken here was initially motivated by recent advances in protein structure prediction methods, but it is different in terms of input as well as the final objective. Sequence alignments or template libraries are not used here; instead, a lower resolution model serves as input. On the other hand, even although the ultimate goal of providing physically realistic, accurate atomistic structures is essentially the same, structure prediction methods aim at providing the best model for the likely native state whereas the method introduced here aims at generating atomistic coordinates for any conformation, whether energetically favorable or not. This suggests that recent advances in machine learning have broad implications for structural biology that reach far beyond just the prediction of native structures from sequence.

There are immediate applications in adding accurate atomistic coordinates to CG representations. Low-resolution protein structure models based on experiments as well as CG models from simulations can thus be interpreted in atomistic detail. This is especially relevant for the generation of structures and ensembles via machine learning where the addition of atomistic detail often presents a significant burden during model training.

An important feature is that deterministic neural network architectures are not just very efficient but also allow gradients to be backpropagated all the way from the final output (i.e., atomistic conformations) to the input (i.e., CG conformations). In essence, this provides an avenue for tightly coupled bidirectional multiscaling. This approach again neglects the full dynamics at the atomistic level and instead emphasizes ensemble-averaged conformations as the representative atomistic states. Therefore, this framework is most suitable for interpreting time- and ensembled-averaged experimental data, especially data at lower resolutions. As one important application, we highlight the refinement of models against cryo-EM densities based on sampling at a CG level but with energy penalty functions evaluated at the atomistic level from reconstructed all-atom conformations. Finally, we recognize that it is necessary to consider

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

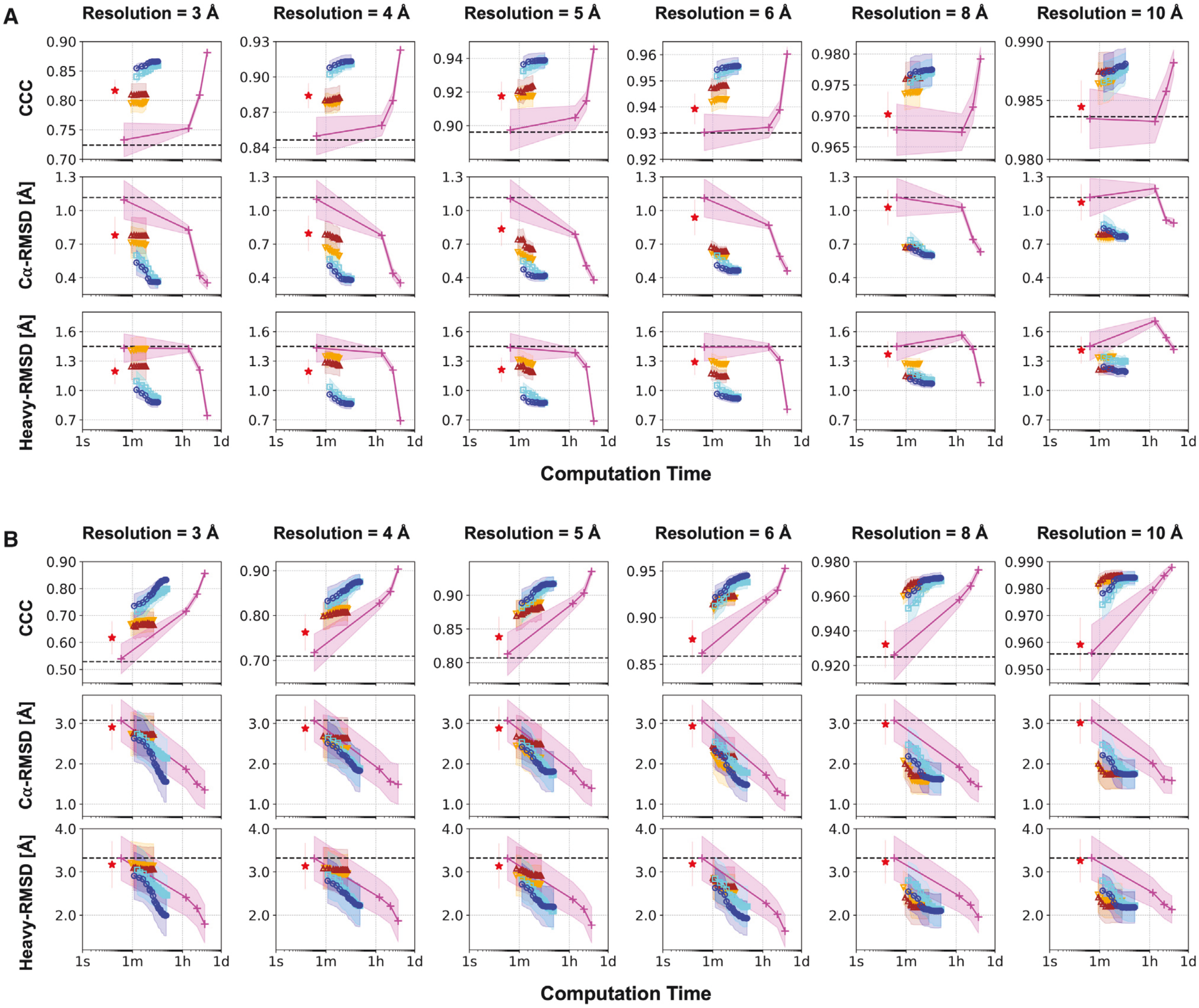


Figure 4. Refinement of initial models against cryo-EM density maps

(A and B) Initial models were obtained via AlphaFold2 (A) or ESM-Fold (B). Model quality in terms of cross-correlation coefficient (CCC) against target cryo-EM density maps, Cα and heavy atom RMSDs were analyzed as a function of computation time for several protocols: (1) optimizations using cg2all models for residue center-of-mass model (blue circles) and Cα-trace (cyan circles) followed by all-atom energy minimization, (2) optimization at the residue center-of-mass (brown triangles) or Cα-trace representation (orange triangles) followed by all-atom energy minimization, (3) all-atom energy minimization only (red star), and (4) MDFF samplings (magenta "+"). The initial model qualities are shown as black dashed lines. The average values (n = 9 for AlphaFold2 models and n = 6 for ESM-Fold models) for each metric and computation time are shown. Shaded background indicates standard errors of the mean.

conformational sampling of all degrees of freedom, e.g., via traditional simulations[59,60] or machine learning approaches[57,61] to achieve full thermodynamic consistency across different levels of resolution. Therefore, it will be ultimately necessary to learn how to generate not just a single ensemble-averaged structure, but entire conformational ensembles that are consistent with a given CG representation.[57]

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

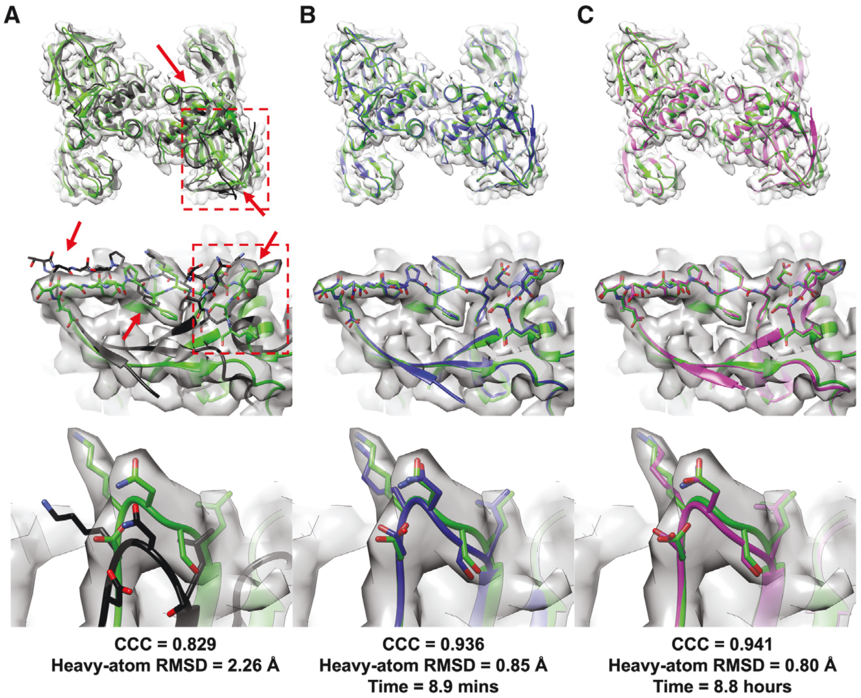


Figure 5. Refinement against a cryo-EM density map for 3isr

(A–C) The target tetrameric structure (PDB ID: 3isr[64]) is shown in green cartoon representation. Its synthetic electron density map was created at 5 Å resolution using "molmap" command in UCSC Chimera,[65] which is based on EMAN2's pdb2mrc program,[66] and depicted as transparent gray surface at a density level of 0.3. The initial model structure generated by AlphaFold-Multimer[67] (A), its optimized structure using cg2all model for Cα-trace and followed by all-atom minimization (B), and another optimized model via all-atom minimization using MDFF (C) are shown in black, blue, and red. Overall tetrameric structures are shown in the top panels, and they are gradually zoomed in (red boxes) to highlight a region where significant deviations (indicated by red arrows) were optimized in the middle and bottom panels.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.str.2023.10.013.

ACKNOWLEDGMENTS

Funding was provided by the National Institutes of Health (NIGMS) grant R35 GM126948 and the National Science Foundation grant MCB 2210228.

AUTHOR CONTRIBUTIONS

L.H. and M.F. designed the research, L.H. performed and analyzed the work, and L.H. and M.F. jointly wrote the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: July 19, 2023
Revised: September 16, 2023
Accepted: October 30, 2023
Published: November 23 ,2023

REFERENCES

1. Shi, Y. (2014). A glimpse of structural biology through X-ray crystallography. Cell *159*, 995–1014. https://doi.org/10.1016/j.cell.2014.10.051.

2. Jones, N. (2014). Crystallography: Atomic secrets. Nature *505*, 602–603. https://doi.org/10.1038/505602a.

3. Wüthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. J. Biol. Chem. *265*, 22059–22062. https://doi.org/10.1016/S0021-9258(18)45665-7.

4. Cheng, Y. (2015). Single-Particle Cryo-EM at Crystallographic Resolution. Cell *161*, 450–457. https://doi.org/10.1016/j.cell.2015.03.049.

5. Nogales, E. (2016). The development of cryo-EM into a mainstream structural biology technique. Nat. Methods *13*, 24–27. https://doi.org/10.1038/nmeth.3694.

6. Garman, E.F. (2014). Developments in x-ray crystallographic structure determination of biological macromolecules. Science *343*, 1102–1108. https://doi.org/10.1126/science.1247829.

7. Kmiecik, S., Gront, D., Kolinski, M., Wieteska, L., Dawid, A.E., and Kolinski, A. (2016). Coarse-Grained Protein Models and Their Applications. Chem. Rev. *116*, 7898–7936. https://doi.org/10.1021/acs.chemrev.6b00163.

8. Saunders, M.G., and Voth, G.A. (2013). Coarse-graining methods for computational biology. Annu. Rev. Biophys. *42*, 73–93. https://doi.org/10.1146/annurev-biophys-083012-130348.

9. Feig, M., Chocholoušová, J., and Tanizaki, S. (2006). Extending the horizon: towards the efficient modeling of large biomolecular complexes in atomic detail. Theor. Chem. Acc. *116*, 194–205. https://doi.org/10.1007/s00214-005-0062-4.

10. Lane, T.J., Shukla, D., Beauchamp, K.A., and Pande, V.S. (2013). To milliseconds and beyond: challenges in the simulation of protein folding. Curr. Opin. Struct. Biol. *23*, 58–65. https://doi.org/10.1016/j.sbi.2012.11.002.

11. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2.

12. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., et al. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. Science *373*, 871–876. https://doi.org/10.1126/science.abj8754.

13. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., et al. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science *379*, 1123–1130. https://doi.org/10.1126/science.ade2574.

14. Janson, G., Valdes-Garcia, G., Heo, L., and Feig, M. (2023). Direct generation of protein conformational ensembles via machine learning. Nat. Commun. *14*, 774. https://doi.org/10.1038/s41467-023-36443-x.

15. Noé, F., Olsson, S., Köhler, J., and Wu, H. (2019). Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. Science *365*, eaaw1147. https://doi.org/10.1126/science.aaw1147.

16. Dutagaci, B., Nawrocki, G., Goodluck, J., Ashkarran, A.A., Hoogstraten, C.G., Lapidus, L.J., and Feig, M. (2021). Charge-driven condensation of RNA and proteins suggests broad role of phase separation in cytoplasmic environments. Elife *10*, e64004. https://doi.org/10.7554/eLife.64004.

17. Yu, I., Mori, T., Ando, T., Harada, R., Jung, J., Sugita, Y., and Feig, M. (2016). Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. Elife *5*, e19274. https://doi.org/10.7554/eLife.19274.

18. Valdes-Garcia, G., Heo, L., Lapidus, L.J., and Feig, M. (2023). Modeling Concentration-dependent Phase Separation Processes Involving Peptides and RNA via Residue-Based Coarse-Graining. J. Chem. Theory Comput. *19*, 669–678. https://doi.org/10.1021/acs.jctc.2c00856.

19. Dignon, G.L., Zheng, W., Kim, Y.C., Best, R.B., and Mittal, J. (2018). Sequence determinants of protein phase behavior from a coarse-grained model. PLoS Comput. Biol. *14*, e1005941. https://doi.org/10.1371/journal.pcbi.1005941.

20. Monticelli, L., Kandasamy, S.K., Periole, X., Larson, R.G., Tieleman, D.P., and Marrink, S.J. (2008). The MARTINI Coarse-Grained Force Field: Extension to Proteins. J. Chem. Theory Comput. *4*, 819–834. https://doi.org/10.1021/ct700324x.

21. Gopal, S.M., Mukherjee, S., Cheng, Y.M., and Feig, M. (2010). PRIMO/PRIMONA: a coarse-grained model for proteins and nucleic acids that preserves near-atomistic accuracy. Proteins *78*, 1266–1281. https://doi.org/10.1002/prot.22645.

22. Gray, J.J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C.A., and Baker, D. (2003). Protein-protein docking with simultaneous optimization of rigid-body displacement and side-chain conformations. J. Mol. Biol. *331*, 281–299. https://doi.org/10.1016/s0022-2836(03)00670-3.

23. Kolinski, A. (2004). Protein modeling and structure prediction with a reduced representation. Acta Biochim. Pol. *51*, 349–371.

24. Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J., and Zhang, Y. (2015). The I-TASSER Suite: protein structure and function prediction. Nat. Methods *12*, 7–8. https://doi.org/10.1038/nmeth.3213.

25. Kar, P., and Feig, M. (2014). Recent advances in transferable coarse-grained modeling of proteins. Adv. Protein Chem. Struct. Biol. *96*, 143–180. https://doi.org/10.1016/bs.apcsb.2014.06.005.

26. Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999). Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. J. Mol. Biol. *285*, 1735–1747. https://doi.org/10.1006/jmbi.1998.2401.

27. Wassenaar, T.A., Pluhackova, K., Böckmann, R.A., Marrink, S.J., and Tieleman, D.P. (2014). Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. J. Chem. Theory Comput. *10*, 676–690. https://doi.org/10.1021/ct400617g.

28. Rotkiewicz, P., and Skolnick, J. (2008). Fast procedure for reconstruction of full-atom protein models from reduced representations. J. Comput. Chem. *29*, 1460–1465. https://doi.org/10.1002/jcc.20906.

29. Li, Y., and Zhang, Y. (2009). REMO: A new protocol to refine full atomic protein models from C-alpha traces by optimizing hydrogen-bonding networks. Proteins *76*, 665–676. https://doi.org/10.1002/prot.22380.

30. Krivov, G.G., Shapovalov, M.V., and Dunbrack, R.L., Jr. (2009). Improved prediction of protein side-chain conformations with SCWRL4. Proteins *77*, 778–795. https://doi.org/10.1002/prot.22488.

31. 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. (2017). The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. *13*, 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.

32. Xu, D., and Zhang, Y. (2011). Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization. Biophys. J. *101*, 2525–2534. https://doi.org/10.1016/j.bpj.2011.10.024.

33. McPartlon, M., and Xu, J. (2023). An end-to-end deep learning method for protein side-chain packing and inverse folding. Proc. Natl. Acad. Sci. USA *120*, e2216438120. https://doi.org/10.1073/pnas.2216438120.

34. Misiura, M., Shroff, R., Thyer, R., and Kolomeisky, A.B. (2022). DLPacker: Deep learning for prediction of amino acid side chain conformations in proteins. Proteins *90*, 1278–1290. https://doi.org/10.1002/prot.26311.

35. Flores, S.C., Bernauer, J., Shin, S., Zhou, R., and Huang, X. (2012). Multiscale modeling of macromolecular biosystems. Briefings Bioinf. *13*, 395–405. https://doi.org/10.1093/bib/bbr077.

36. Predeus, A.V., Gul, S., Gopal, S.M., and Feig, M. (2012). Conformational sampling of peptides in the presence of protein crowders from AA/CG-multiscale simulations. J. Phys. Chem. B *116*, 8610–8620. https://doi.org/10.1021/jp300129u.

37. Fuchs, F.B., Worrall, D.E., Fischer, V., and Welling, M. (2020). SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.2006.10503.

38. Chen, V.B., Arendall, W.B., 3rd, Headd, J.J., Keedy, D.A., Immormino, R.M., Kapral, G.J., Murray, L.W., Richardson, J.S., and Richardson, D.C. (2010). MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr. D Biol. Crystallogr. *66*, 12–21. https://doi.org/10.1107/S0907444909042073.

39. Daopin, S., Davies, D.R., Schlunegger, M.P., and Grütter, M.G. (1994). Comparison of two crystal structures of TGF-beta2: the accuracy of refined protein structures. Acta Crystallogr. D Biol. Crystallogr. *50*, 85–92. https://doi.org/10.1107/S090744499300808X.

40. Chen, V.B., Wedell, J.R., Wenger, R.K., Ulrich, E.L., and Markley, J.L. (2015). MolProbity for the masses–of data. J. Biomol. NMR *63*, 77–83. https://doi.org/10.1007/s10858-015-9969-9.

41. Paxman, J.J., and Heras, B. (2017). Bioinformatics Tools and Resources for Analyzing Protein Structures. In Proteome Bioinformatics, S. Keerthikumar and S. Mathivanan, eds. (Springer New York), pp. 209–220. https://doi.org/10.1007/978-1-4939-6740-7_16.

42. Berkholz, D.S., Shapovalov, M.V., Dunbrack, R.L., Jr., and Karplus, P.A. (2009). Conformation dependence of backbone geometry in proteins. Structure *17*, 1316–1325. https://doi.org/10.1016/j.str.2009.08.012.

43. Cong, Y., Baker, M.L., Jakana, J., Woolford, D., Miller, E.J., Reissmann, S., Kumar, R.N., Redding-Johanson, A.M., Batth, T.S., Mukhopadhyay, A., et al. (2010). 4.0-Å resolution cryo-EM structure of the mammalian chaperonin TRiC/CCT reveals its unique subunit arrangement. Proc. Natl. Acad. Sci. USA *107*, 4967–4972. https://doi.org/10.1073/pnas.0913774107.

44. Terashi, G., and Kihara, D. (2018). De novo main-chain modeling for EM maps using MAINMAST. Nat. Commun. *9*, 1618. https://doi.org/10.1038/s41467-018-04053-7.

45. Benjin, X., and Ling, L. (2020). Developments, applications, and prospects of cryo-electron microscopy. Protein Sci. *29*, 872–882. https://doi.org/10.1002/pro.3805.

46. Wu;, K.E., Yang;, K.K., Berg;, R.v.d., Zou, J.Y., Lu;, A.X., and Amini, A.P. (2022). Protein Structure Generation via Folding Diffusion. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.2209.15611.

47. Cheng, Y.M., Gopal, S.M., Law, S.M., and Feig, M. (2012). Molecular dynamics trajectory compression with a coarse-grained model. IEEE/ACM

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

Trans. Comput. Biol. Bioinform. *9*, 476–486. https://doi.org/10.1109/TCBB.2011.141.

48. Tiemann, J.K.S., Szczuka, M., Bouarroudj, L., Oussaren, M., Garcia, S., Howard, R.J., Delemotte, L., Lindahl, E., Baaden, M., Lindorff-Larsen, K., et al. (2023). MDverse: Shedding Light on the Dark Matter of Molecular Dynamics Simulations. Preprin at. bioRxiv *1*, 2023.05.02.538537. https://doi.org/10.1101/2023.05.02.538537.

49. Feig, M., Abdullah, M., Johnsson, L., and Pettitt, B. (1999). Large Scale Distributed Data Repository: Design of a Molecular Dynamics Trajectory Database. Fut Gen Comput Sys *16*, 101–110. https://doi.org/10.1016/S0167-739X(99)00039-4.

50. Malhotra, S., Träger, S., Dal Peraro, M., and Topf, M. (2019). Modelling structures in cryo-EM maps. Curr. Opin. Struct. Biol. *58*, 105–114. https://doi.org/10.1016/j.sbi.2019.05.024.

51. Casañal, A., Lohkamp, B., and Emsley, P. (2020). Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. Protein Sci. *29*, 1055–1064. https://doi.org/10.1002/pro.3791.

52. Croll, T.I. (2018). ISOLDE: a physically realistic environment for model building into low-resolution electron-density maps. Acta Crystallogr. D Struct. Biol. *74*, 519–530. https://doi.org/10.1107/S2059798318002425.

53. Wang, Z., and Schröder, G.F. (2012). Real-space refinement with DireX: From global fitting to side-chain improvements. Biopolymers *97*, 687–697. https://doi.org/10.1002/bip.22046.

54. Trabuco, L.G., Villa, E., Schreiner, E., Harrison, C.B., and Schulten, K. (2009). Molecular dynamics flexible fitting: a practical guide to combine cryo-electron microscopy and X-ray crystallography. Methods *49*, 174–180. https://doi.org/10.1016/j.ymeth.2009.04.005.

55. Heo, L., and Feig, M. (2018). What makes it difficult to refine protein models further via molecular dynamics simulations? Proteins *86*, 177–188. https://doi.org/10.1002/prot.25393.

56. Heo, L., and Feig, M. (2018). Experimental accuracy in protein structure refinement via molecular dynamics simulations. Proc. Natl. Acad. Sci. USA *115*, 13276–13281. https://doi.org/10.1073/pnas.1811364115.

57. Chennakesavalu, S., Toomer, D.J., and Rotskoff, G.M. (2023). Ensuring thermodynamic consistency with invertible coarse-graining. J. Chem. Phys. *158*, 124126. https://doi.org/10.1063/5.0141888.

58. Bramer, D., and Wei, G.W. (2018). Blind prediction of protein B-factor and flexibility. J. Chem. Phys. *149*, 134107. https://doi.org/10.1063/1.5048469.

59. Tozzini, V. (2010). Multiscale modeling of proteins. Acc. Chem. Res. *43*, 220–230. https://doi.org/10.1021/ar9001476.

60. Ayton, G.S., Noid, W.G., and Voth, G.A. (2007). Multiscale modeling of biomolecular systems: in serial and in parallel. Curr. Opin. Struct. Biol. *17*, 192–198. https://doi.org/10.1016/j.sbi.2007.03.004.

61. Durumeric, A.E.P., Charron, N.E., Templeton, C., Musil, F., Bonneau, K., Pasos-Trejo, A.S., Chen, Y., Kelkar, A., Noé, F., and Clementi, C. (2023). Machine learned coarse-grained protein force-fields: Are we there yet? Curr. Opin. Struct. Biol. *79*, 102533. https://doi.org/10.1016/j.sbi.2023.102533.

62. Macedo-Ribeiro, S., Darimont, B., Sterner, R., and Huber, R. (1996). Small structural changes account for the high thermostability of 1[4Fe-4S] ferredoxin from the hyperthermophilic bacterium Thermotoga maritima. Structure *4*, 1291–1301. https://doi.org/10.1016/s0969-2126(96)00137-2.

63. Elbaum-Garfinkle, S., Kim, Y., Szczepaniak, K., Chen, C.C.H., Eckmann, C.R., Myong, S., and Brangwynne, C.P. (2015). The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. Proc. Natl. Acad. Sci. USA *112*, 7189–7194. https://doi.org/10.1073/pnas.1504822112.

64. Stein, A.J., Bigelow, L., Trevino, D., Buck, K., and Joachimiak, A. (2009). The Crystal Structure of a Putative Cysteine Protease from Cytophaga Hutchinsonii to 1.9Å (Protein Data Bank). https://doi.org/10.2210/pdb3ISR/pdb.

65. Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612. https://doi.org/10.1002/jcc.20084.

66. Tang, G., Peng, L., Baldwin, P.R., Mann, D.S., Jiang, W., Rees, I., and Ludtke, S.J. (2007). EMAN2: an extensible image processing suite for electron microscopy. J. Struct. Biol. *157*, 38–46. https://doi.org/10.1016/j.jsb.2006.05.009.

67. Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., et al. (2022). Protein Complex Prediction with AlphaFold-Multimer. Preprin at. bioRxiv *1*. https://doi.org/10.1101/2021.10.04.463034.

68. Heo, L., and Feig, M. (2023). github.com/huhlim/cg2all. https://doi.org/10.5281/zenodo.10011305.

69. Heo, L., and Feig, M. (2023). cg2all Data Set. https://doi.org/10.5281/zenodo.8273738.

70. Hintze, B.J., Lewis, S.M., Richardson, J.S., and Richardson, D.C. (2016). Molprobity's ultimate rotamer-library distributions for model validation. Proteins *84*, 1177–1189. https://doi.org/10.1002/prot.25039.

71. Wang, G., and Dunbrack, R.L., Jr. (2003). PISCES: a protein sequence culling server. Bioinformatics *19*, 1589–1591. https://doi.org/10.1093/bioinformatics/btg224.

72. Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577–2637. https://doi.org/10.1002/bip.360221211.

73. McGibbon, R.T., Beauchamp, K.A., Harrigan, M.P., Klein, C., Swails, J.M., Hernández, C.X., Schwantes, C.R., Wang, L.P., Lane, T.J., and Pande, V.S. (2015). MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. Biophys. J. *109*, 1528–1532. https://doi.org/10.1016/j.bpj.2015.08.015.

74. Huang, J., Rauscher, S., Nawrocki, G., Ran, T., Feig, M., de Groot, B.L., Grubmüller, H., and MacKerell, A.D., Jr. (2017). CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat. Methods *14*, 71–73. https://doi.org/10.1038/nmeth.4067.

75. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. *13*, e1005659. https://doi.org/10.1371/journal.pcbi.1005659.

76. Jorgensen, W.L., Chandrasekhar, J., Madura, J.D., Impey, R.W., and Klein, M.L. (1983). Comparison of simple potential functions for simulating liquid water. J. Chem. Phys. *79*, 926–935. https://doi.org/10.1063/1.445869.

77. Liu, D.C., and Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. Math. Program. *45*, 503–528. https://doi.org/10.1007/bf01589116.

78. Ba, J.L., Kiros, J.R., and Hinton, G.E. (2016). Layer Normalization. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1607.06450.

79. Clevert, D.-A., Unterthiner, T., and Hochreiter, S. (2016). Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1511.07289.

80. He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1512.03385.

81. Nair, V., and Hinton, G.E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the 27th International Conference on Machine Learning (ICML10), pp. 807–814. https://icml.cc/Conferences/2010/papers/432.pdf.

82. Zhou;, Y., Barnes;, C., Lu;, J., Yang;, J., and Li, H. (2020). On the Continuity of Rotation Representations in Neural Networks. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1812.07035.

83. Falcon, W.; The_PyTorch_Lightning_team (2022). PyTorch Lightning. Zenodo *1*. https://doi.org/10.5281/zenodo.7859091.

84. Lu, L. (2020). Dying ReLU and Initialization: Theory and Numerical Examples. Commun. Comput. Phys. *28*, 1671–1706. https://doi.org/10.4208/cicp.OA-2020-0165.

85. Schumacher, M.A., Pearson, R.F., Møller, T., Valentin-Hansen, P., and Brennan, R.G. (2002). Structures of the pleiotropic translational regulator Hfq and an Hfq-RNA complex: a bacterial Sm-like protein. EMBO J. *21*, 3546–3556. https://doi.org/10.1093/emboj/cdf322.

86. Badger, J., Sauder, J.M., Adams, J.M., Antonysamy, S., Bain, K., Bergseid, M.G., Buchanan, S.G., Buchanan, M.D., Batiyenko, Y., Christopher, J.A., et al. (2005). Structural analysis of a set of proteins resulting from a bacterial genomics project. Proteins *60*, 787–796. https://doi.org/10.1002/prot.20541.

87. Shi, W., Basso, L.A., Santos, D.S., Tyler, P.C., Furneaux, R.H., Blanchard, J.S., Almo, S.C., and Schramm, V.L. (2001). Structures of purine nucleoside phosphorylase from Mycobacterium tuberculosis in complexes with immucillin-H and its pieces. Biochemistry *40*, 8204–8215. https://doi.org/10.1021/bi010585p.

88. Boutz, D.R., Cascio, D., Whitelegge, J., Perry, L.J., and Yeates, T.O. (2007). Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. J. Mol. Biol. *368*, 1332–1344. https://doi.org/10.1016/j.jmb.2007.02.078.

89. Singleton, M., Isupov, M., and Littlechild, J. (1999). X-ray structure of pyrrolidone carboxyl peptidase from the hyperthermophilic archaeon Thermococcus litoralis. Structure *7*, 237–244. https://doi.org/10.1016/s0969-2126(99)80034-3.

90. Im, Y.J., Kim, J.I., Shen, Y., Na, Y., Han, Y.J., Kim, S.H., Song, P.S., and Eom, S.H. (2004). Structural analysis of Arabidopsis thaliana nucleoside diphosphate kinase-2 for phytochrome-mediated light signaling. J. Mol. Biol. *343*, 659–670. https://doi.org/10.1016/j.jmb.2004.08.054.

91. Hondoh, H., Kuriki, T., and Matsuura, Y. (2003). Three-dimensional structure and substrate binding of Bacillus stearothermophilus neopullulanase. J. Mol. Biol. *326*, 177–188. https://doi.org/10.1016/s0022-2836(02)01402-x.

92. Tanaka, Y., Nakagawa, N., Kuramitsu, S., Yokoyama, S., and Masui, R. (2005). Novel reaction mechanism of GTP cyclohydrolase I. High-resolution X-ray crystallography of Thermus thermophilus HB8 enzyme complexed with a transition state analogue, the 8-oxoguanine derivative. J. Biochem. *138*, 263–275. https://doi.org/10.1093/jb/mvi120.

93. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.

94. Mukherjee, S., and Zhang, Y. (2009). MM-align: a quick algorithm for aligning multiple-chain protein complex structures using iterative dynamic programming. Nucleic Acids Res. *37*, e83. https://doi.org/10.1093/nar/gkp318.

95. Kingma, D.P., and Ba, J. (2017). Adam: A Method for Stochastic Optimization. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1412.6980.

96. Loshchilov, I., and Hutter, F. (2017). SGDR: Stochastic Gradient Descent with Warm Restarts. Preprin at. arXiv *1*. https://doi.org/10.48550/arXiv.1608.03983.

97. Qi, Y., Lee, J., Singharoy, A., McGreevy, R., Schulten, K., and Im, W. (2017). CHARMM-GUI MDFF/xMDFF Utilizer for Molecular Dynamics Flexible Fitting Simulations in Various Environments. J. Phys. Chem. B *121*, 3718–3723. https://doi.org/10.1021/acs.jpcb.6b10568.

98. Phillips, J.C., Hardy, D.J., Maia, J.D.C., Stone, J.E., Ribeiro, J.V., Bernardi, R.C., Buch, R., Fiorin, G., Hénin, J., Jiang, W., et al. (2020). Scalable molecular dynamics on CPU and GPU architectures with NAMD. J. Chem. Phys. *153*, 044130. https://doi.org/10.1063/5.0014475.

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
List of PDB IDs and corresponding PDB files	This paper	https://doi.org/10.5281/zenodo.10011305
Software and algorithms		
Cg2all source code and model parameter files	This paper	https://doi.org/10.5281/zenodo.8273738 https://github.com/huhlim/cg2all

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Michael Feig (mfeiglab@gmail.com)

Materials availability

This study did not generate any reagents.

Data and code availability

- No new data has been generated in this study.
- The source code and model parameter files of the cg2all method are available at https://github.com/huhlim/cg2all.[68] It can be locally installed using a PIP command, "pip install git+ http://github.com/huhlim/cg2all". The list of PDB IDs and the corresponding PDB files (both cleaned up and side chain conformation augmented structure files) are available at https://zenodo.org/record/8273739[69]
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Experimental datasets

Two sets of experimental structures were clustered with a maximum mutual sequence identity of 70%, and one structure for each cluster was left. The original Top8000[70] and the PISCES[71] sets consisted of a single chain per Protein DataBank (PDB) entry, however, we used all protein chains in each PDB entry instead. PDB entries with more than 1,200 residues were excluded. As a result, there were 7,130 and 30,354 entries in PDB 6k and 29k sets, respectively. To have common subsets for model validation and test, 720 entries were randomly selected among the common entries for each validation and test sets. The remaining 5,690 and 28,914 structures were used as PDB 6k and 29k training sets, respectively.

The Top8000 structures were further analyzed to obtain statistics of bonded geometries for building secondary structure-dependent rigid body (SS-dep) blocks. Three-state secondary structure for every residue was assigned by the DSSP algorithm[72] using the MDTraj Python library package.[73] Bond lengths, angles, and improper dihedral angles defined by amino acid topologies of the CHARMM36m force field[74] were calculated. Their averaged values for each secondary structure type were then used to build SS-dep blocks.

Simulation datasets

We sampled an ensemble of structures by performing all-atom molecular dynamics (MD) simulations using OpenMM.[75] A protein structure was placed at the center of a periodic rectangular box with at least 10 Å distance from any protein atom to any dimension of the box edges. The remaining space in the simulation box was filled with the CHARMM version of TIP3P water molecules.[76] Some water molecules were randomly replaced with sodium or chloride ions to neutralize the system and achieve a total ion concentration of 0.015 M. The CHARMM36m force field[74] was applied to describe the system throughout the series of simulations. The system was locally minimized with the l-BFGS-b algorithm[77] in the presence of harmonic positional restraints on every Cα atom with a force constant of 0.5 kcal/mol/Å^2. Then, the systems were gradually heated to 298.15 K and equilibrated via Langevin dynamics simulations for 1 ns with a friction coefficient of 0.01/ps and a 2-fs integration time step. The NVT ensemble and the NpT ensemble at 1 bar with a Monte Carlo barostat were applied during the heating and equilibration steps, respectively. An ensemble of protein conformations was sampled from a 50 ns-long Langevin dynamics simulation with a friction coefficient of 1/ps and a 2-fs integration time step at

298.15 K and 1 bar. Five snapshots of protein conformations were picked up for further tests by selecting frames for every 10 ns. In total, 3,600 all-atom conformations were generated for 720 test set structures for further tests.

During minimization with COCOMO,[18] harmonic positional restraints were applied to every bead of the CG model with a force constant of 1.0 kcal/mol/Å^2 to keep the original all-atom conformations as the respective ground truth conformations. Upon minimization, the conformations were distorted from their all-atom structures by 0.30 Å on average.

Input features

Input features extracted from a CG model are summarized in Table S1. In total, 57 (17 from the local geometries and 40 from the residue type embedding) scalar and four (or more for multiple site CG models) vector node features and three scalar edge features were used as input features. Pseudo-bond angles and -torsion angles were encoded using cosine and sine functions to account for periodicity. The number of neighboring nodes (e.g., <10 Å) and the presence of a previous or next residue (to distinguish terminal residues and chain breaks) were added as scalar features. The residue type was converted to 40 scalars via a trainable embedding layer and concatenated with the scalar features. For vector features, unit bond vectors as illustrated in Figure S1 were used for conversion from multiple site coarse-grained models such as a MARTINI model,[20] vectors from a BB bead to SC beads were additionally used to incorporate side chain information. The edge connection type (connection via a peptide bond, an inter-residue contact through space, or a disulfide bond) were used via one-hot encoding as edge features.

Neural network model

At the core of SE(3)-equivalent neural network model, the SE(3)-Transformers architecture[37] was adopted. The input node features were processed via N (=4 for the baseline model) linear layer blocks to produce hidden features of 64 scalar and 32 vector values. (Methods S1, and Algorithm S1) A LayerNorm[78] and an exponential linear unit (ELU) activation function[79] were used to normalize the norm of features for each degree, while keeping their phase. (NormSE3) The first linear layer without bias terms (LinearNoBias) projects the input features to hidden features with the numbers of channels of 64 and 32 for scalars and vectors, respectively. The linear layer blocks consisting of a LayerNorm, an ELU activation function, and a linear layer without bias terms are repeated N-1 times, while the numbers of channels for each degree of features are maintained. Until this step, interactions between nodes are not considered yet, and features are processed for each node and each degree. In the interaction module, we facilitated the original SE(3)-Transformers architecture with M (=4 for the baseline model) SE(3)-Transformers blocks with a LayerNorm and an ELU activation function for radial profiles (AttentionBlockSE3) to communicate between nodes through edges and evaluate tensor products between different degrees of hidden features. (Methods S1 and Algorithm S2) For an AttentionBlockSE3, we used eight attention heads and 32 hidden channels of scalars ($l = 0$), vectors ($l = 1$), and rank-two tensors ($l = 2$). A normalization layer (NormSE3) was followed by to stabilize the training. At the end of the interaction module, an SE(3)-equivariant convolutional layer with a LayerNorm and an ELU activation function for radial profiles (ConvSE3) was used to produce the output of the module. Throughout the module, the number of channels for each degree of hidden features were maintained between blocks: 64 for scalars and 32 for vectors. Finally, the input representations of the module were added to the output representations for a skip connection.[80]

We used four linear layers and four SE(3)-Transformers blocks for the baseline model. Alternatively, fewer and greater numbers of blocks were examined to evaluate performance dependencies in the model size. We adopted ELU activation functions as replacements of rectified linear unit (ReLU) functions,[81] which were used for the original SE(3)-Transformers work.[37] For hidden features of the SE(3)-Transformers blocks, features up to the degree of 2 were passed within a block, and a lower value (1) was also tested. For a protein structure, we used a subgraph with a crop size of 256 residues as the baseline and tested if a larger crop (384 residues) could improve the performance. Secondary structure-dependent rigid-body blocks were not used for the baseline, but they were adopted as optional features.

Structure module

The structure module further processed the output of the interaction module to predict values for building all-atom structures via N (=4 for the baseline model) linear layer blocks. (Methods S1 and Algorithm S3) It was analogous to the initialization module but projected hidden features with greater numbers of channels to predict 16 scalar (or 20 if SS-dependent rigid body blocks were used) and 3 vector values for the followed all-atom structure building. An all-atom structure was built using the predicted values. For this process, a similar procedure that is used by AlphaFold2 was adopted.[11] AlphaFold2 used rigid-body blocks of backbone atoms (N, Cα, and C atoms), the backbone oxygen atom, and side chain heavy atoms that were segmented by rotatable torsion angles. During the step of structure building, backbone rigid-body blocks for residues were oriented first using predicted translations (***t***) and three-dimensional rotations (**R**). The remaining blocks were placed in the order of the bond connectivity from the backbone block to the tip of each side chain by rotating them by predicted torsion angles (φ, ψ, and χs). In our work, we extended the procedure to build all-atoms including hydrogen atoms. Thus, 16 predicted scalar values were used as sine and cosine values of the torsion angles (φ, ψ, 4 χs, and 2 hydrogen atom-only torsion angles) to construct rotation matrices. In addition, secondary structure-dependent rigid body blocks were used as we found that some blocks have distinguished bonded geometries (bond lengths and angles and improper dihedrals) depending on the secondary structure. Thus, secondary structure prediction (SS) was also made by the module. Four predicted scalar values determined which rigid body blocks to use among blocks for helix, sheet, coil, or SS-independent blocks. One output vector predicted a translation vector (***t***), which places the Cα atom from the representative bead for a residue. We note that we used a 6D representation[82] using the remaining two predicted vectors, which describes three-dimensional rotation using two unit

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

vectors and the Gram-Schmidt process, instead of the quaternion-based approach used by AlphaFold2, as it gave better performance in terms of accuracy and convergence during training for many SO(3) prediction tasks because of continuity in the rotation representations.[82] For the final step, both N- and C-termini were patched by replacing backbone hydrogen (HN) and oxygen (O) with the terminal amino (NH_3-) and carboxyl (-COO^-) groups, respectively, using the predicted position of the atoms.

Model training

Each model was trained for 300 and 120 epochs with the PDB 6k and 29k training sets, respectively, using a batch size of 4. This corresponds to 426,600 and 867,360 total training steps. An input data graph with more than 256 nodes for training was randomly cropped to a consecutive 256 residue graph by obtaining its subgraph. An Adam optimizer was used with learning rates of 0.01 for parameters for scalar features of the structure module and 0.001 for the others. The learning rates were linearly increased for the first ten epochs and then exponentially decreased with a multiplicative factor of 0.995. Clipping was used to prevent extreme gradients, and gradient checkpointing reduced memory consumption. Models were implemented in PyTorch, and they were trained using PyTorch Lightning.[83] The training of models was carried out on two NVIDIA GeForce RTX 2080 Ti GPU cards (11 GB of VRAM) with Distributed Data Parallel (DDP) to use multiple GPUs and eight CPU threads. We trained three models for each model variant to obtain statistics.

We fine-tuned the trained models for an additional 10 epochs to obtain variants where coordinates of the input coarse-grained structures are forced to remain unchanged. This allowed us to compare with other methods that maintain coordinates in the input structures. For example, sidechain prediction methods such as SCWRL4[30] and AttnPacker[33] only predict sidechain conformations, leaving the backbone structures unchanged. It also allowed us to test to what degree cg2all may be able to correct for slight errors in the reduced representations.

Loss function

Models were trained with a loss function (Equation 1) that consisted of training data-dependent loss functions and data-independent (physics-based) loss functions:

$$L = \begin{array}{ll} 5.0L_{FAPE,C\alpha} + 1.0L_{BB} + 1.0L_R + 1.0L_{backbone\ geometry}\ (+0.1L_{SS}) & (Data - dependent) \\ \quad + 5.0L_{torsion} + 1.0L_{v_{cntr}} & \\ + 5.0L_{atomic\ clash} + 0.1L_{torsion\ energy} + 1.0L_{side\ chain\ geometry} & (Physics - based) \end{array} \qquad \text{(Equation 1)}$$

The FAPE, Cα loss ($L_{FAPE,C\alpha}$) was first introduced by AlphaFold2,[11] and we used a variant that uses only Cα atoms for its evaluation with a distance clamp value of 10 Å. (Methods S1 and Algorithm S4) The backbone loss (L_{BB}) also contributed to correctly placing backbone rigid bodies (Equation 2).

$$L_{BB}\left(r_{BB}^{truth}\right) = \left|r_{BB} - r_{BB}^{truth}\right| + \left|1 - \overrightarrow{u_{C\alpha\rightarrow N}} \bullet \overrightarrow{u_{C\alpha\rightarrow N}^{truth}}\right| + \left|1 - \overrightarrow{u_{C\alpha\rightarrow C}} \bullet \overrightarrow{u_{C\alpha\rightarrow C}^{truth}}\right| \qquad \text{(Equation 2)}$$

where $\overrightarrow{u}$ represents a unit vector. The loss function for three-dimensional rotation representation consisted of the similarity of two vectors for the 6D representation[82] against their truths and an auxiliary term that ensures their vector sizes close to 1. (Equation 3)

$$L_R\left(\overrightarrow{R_0}, \overrightarrow{R_1} \middle| \overrightarrow{R_0^{truth}}, \overrightarrow{R_1^{truth}}\right) = \left|\left(\overrightarrow{R_0} - \overrightarrow{R_0^{truth}}\right) + \left(\overrightarrow{R_1} - \overrightarrow{R_1^{truth}}\right)\right| + 0.01\left(\left|\|\overrightarrow{R_0}\| - 1\right| + \left|\|\overrightarrow{R_1}\| - 1\right|\right) \qquad \text{(Equation 3)}$$

where $\overrightarrow{R_0}$ and $\overrightarrow{R_1}$ are vectors for the 6D representation. A loss function helped the model having correct bonded geometries for backbone atoms (Equation 4). It penalized deviations of peptide bond distances ($b_{C,+N}$) and peptide bond including bond angles ($\theta_{C\alpha,C,+N}$ and $\theta_{C,+N,+C\alpha}$).

$$L_{backbone\ geometry}\left(b_{C,+N}, \theta_{C\alpha,C,+N}, \theta_{C,+N,+C\alpha} \middle| b_{C,+N}^{truth}, \theta_{C\alpha,C,+N}^{truth}, \theta_{C,+N,+C\alpha}^{truth}\right) = \left|b_{C,+N} - b_{C,+N}^{truth}\right| + \frac{1}{2}\left(\left|\theta_{C\alpha,C,+N} - \theta_{C\alpha,C,+N}^{truth}\right| + \left|\theta_{C,+N,+C\alpha} - \theta_{C,+N,+C\alpha}^{truth}\right|\right) \qquad \text{(Equation 4)}$$

When the secondary structure-dependent rigid body blocks were used, a cross entropy loss for secondary structure prediction was additionally used.

Moreover, there were two loss functions for correctly reconstructing side chain atoms. The torsion angle loss tried to reduce discrepancies between the predicted torsion angles ($\{\theta_k\}$) and their truth values ($\{\theta_k^{truth}\}$) (Equation 5).

$$L_{torsion}\left(\{\theta_k\} \middle| \{\theta_k^{truth}, \theta_k^{truth,alts}\}\right) = \sum_{\theta_k\ is\ defined} \left(1 - max\left(cos\left(\theta_k - \theta_k^{truth}\right), cos\left(\theta_k - \theta_k^{truth,alts}\right)\right)\right) \qquad \text{(Equation 5)}$$

Some side chain torsion angles that have periodicity: χ_2 angles of Phe and Tyr, methyl groups in Ala, Ile, Leu, Met, Thr, and Val, the side chain amino group in Lys, and the side chain carboxyl groups in Asp and Glu. For those torsion angles, alternative truth values ($\{\theta_k^{truth,alts}\}$) due to their periodicity were considered for the calculation. Furthermore, solvent-exposed side chains can have multiple valid conformations, while their experimental structures usually presented only one of them. For the training of models that convert from a Cα-based model, putative side chain conformations were generated prior to the training and used as additional $\{\theta_k^{truth,alts}\}$. The procedure for putative side chain conformation generation is described in more detail below. The other loss function for side chain atoms depended on vectors from the Cα atom to the center of mass of a residue ($\overrightarrow{v_{cntr}}$). Their deviation in their vector size to their truth values and their cosine similarity was used to learn side chain orientations at a lower resolution.

$$L_{v_{cntr}}\left(\overrightarrow{v^{truth}_{cntr}}\right) = \left| \left\|\overrightarrow{v_{cntr}}\right\| - \left\|\overrightarrow{v^{truth}_{cntr}}\right\| \right| + \left|1 - \overrightarrow{u_{cntr}} \bullet \overrightarrow{u^{truth}_{cntr}}\right| \qquad \text{(Equation 6)}$$

In addition to these data-dependent loss functions, three physics-based terms were introduced to improve the geometric properties of the reconstructed models. These terms relied on the CHARMM36m force field.[74] As two atoms rarely overlapped within the sum of their atomic radii, atomic clashes were penalized (Equation 7).

$$L_{atomic\ clash}(r) = \sum_{i<j, r_{ij}<14\text{Å}} \sum_{a \in i} \sum_{b \in j\ exclude 1-2,3,4 pairs} \sqrt{\varepsilon_a \varepsilon_b} \times \left(\min(0, d_{ab} - \sigma_a - \sigma_b)\right)^2 \qquad \text{(Equation 7)}$$

where ε and σ are the depth of the potential and the distance at which the potential becomes zero in the Lennard-Jones potential. Torsion energy terms ($L_{torsion\ energy}$) of the force field was applied to penalize disfavored torsion angles (Equation 8). Torsion angles that can be defined within a residue were considered for evaluation, thus, torsion angles that span two consecutive residues (e.g., ψ, φ, and ω angles) were not subjected to the loss functions. In order to preserve torsion angle distributions, a torsion energy clamp ($E^{clamp}_{torsion\ energy}$) was used with a value of 0.6 kcal/mol.

$$L_{torsion\ energy}(\{\theta_k\}) = \sum_{\theta_k\ is\ defined} \max\left(0, E_{torsion\ energy}(\theta_k) - E^{min}_{torsion\ energy}(\theta_k) - E^{clamp}_{torsion\ energy}\right) \qquad \text{(Equation 8)}$$

The final physical loss term was applied for two types of bonds that connect rigid body blocks in special ways, namely for proline ring closure and disulfide bonds. For these bonds, equilibrium bond lengths of 1.455 and 2.029 Å, respectively, were targeted (Equation 9).

$$L_{side\ chain\ geometry}\left(b_{Pro,N,CD}, b_{SSBOND}\right) = \left|b_{Pro,N,CD} - b^{o}_{Pro,N,CD}\right| + \left|b_{SSBOND} - b^{o}_{SSBOND}\right| \qquad \text{(Equation 9)}$$

Augmentation of putative side chain conformations

Alternative possible side chain conformations were generated prior to the training using first SCWRL4,[30] followed by REDUCE[26] and local energy minimization using the CHARMM36m force field.[74] Experimental structures, especially those determined by X-ray crystallography, have only one conformation in PDB in most of the entries even though there can be alternative coordinates. However, as proteins are not static molecules, they can have diverse conformation especially for solvent-exposed side chains. As such, reconstruction to those alternative conformations should not be penalized unless they are unfavorable. Because we aimed to reconstruct an all-atom conformation including side chains for a given Cα-trace of a protein, we generated putative side chain conformation. For a protein structure, side chain structures were predicted on the protein's backbone using SCWRL4, which uses a rotamer library and optimizes combinations of rotamer states. Then, all hydrogens were attached to the predicted structure with an optimization of torsion angles for the side chain amide groups in Asn and Gln and the imidazole ring in His. Finally, the structures were subjected to energy minimization using the l-BFGS-b algorithm for up to 1,000 steps with the CHARMM36m force field using OpenMM.[75] To prevent extensive deviation of backbone positions, harmonic positional restraints were applied on every N, Cα, C, O, and Cβ atoms with a force constant of 1.0 kcal/mol/Å^2. Torsion angles from the energy minimized structure were used as an additional set of $\{\theta_k^{truth,alts}\}$ for the torsion angle loss (Equation 5).

Hyperparameter optimization and ablation studies

Several features were introduced in our neural network models and their training, and their contributions were evaluated via ablation studies. (Figure S8) First of all, we used a hybrid loss function that consisted of data-dependent and -independent (or physics-based) loss functions (Equation 1). The physics-based loss functions were introduced to learn the characteristics of a protein molecule more efficiently and to complement insufficient data points. The atomic clash loss penalized inter-atomic clashes and effectively lowered the clash score. Side chain modeling as rotamer outliers could be suppressed by introducing the torsion energy loss function (Figure S9). Without the torsion energy loss function, side chain rotamer states could not be clearly separated, and some inferences resulted in rotamer outliers. Average MolProbity scores[38] by models with and without those physics-based loss functions were different by 0.254 (2.292 vs. 2.546). Furthermore, side chain conformations were augmented prior to the model training to account for their putative heterogenic conformations due to their conformational flexibility. Because some side chains (especially solvent-exposed ones) can have alternative conformations in addition to the experimentally resolved one, both conformations should not be penalized unless there are atomic clashes. This side chain augmentation additionally helped suppress generating rotamer outliers (Figure S9). If there were side chains with very similar input features but in different rotamer states in the training set, models could be trained to predict in the average of the rotamer states unless the side chain augmentation was used, and this could result in the inference of rotamer outliers. For example, for a homodimer with pseudo-C$_2$ symmetry, some side chain conformations may vary in different monomers, while overall Cα-traces were almost identical. As illustrated in Figure S9B, if an Arginine from a monomer has a χ_1 angle of -180°, while the other Arginine in the symmetry has a χ_1 angle of -60°, training with these data would result in predicting their averaged value, -120°, which is a rotamer outlier. It is because the averaged torsion angle loss has a minimum at the value. With

Please cite this article in press as: Heo and Feig, One bead per residue can describe all-atom protein structures, Structure (2023), https://doi.org/10.1016/j.str.2023.10.013

the consideration of both possible conformations in the torsion angle loss function via the side chain augmentation, the loss function has multiple minima (e.g., −180 and −60° in this example), and the inference of rotamer outliers is diminished.

We tested two neural network parameters that were related to the amount of information passed between layers: 1) the choice of the activation function (ELU vs. ReLU); and 2) the maximum degree for the SE(3)-Transformers (l = 2 vs. 1). When the ELU activation function was replaced with the ReLU function, which was originally used in SE(3)-Transformers, the overall quality of reconstructed models dropped slightly. It was probably because the use of ReLU function deactivated some neurons, which is known as the "dying ReLU problem",[84] and the neurons could not be efficiently utilized. Regarding the maximum degree for the SE(3)-Transformers,[37] it was beneficial to use up to the degree of 2 features, even although the input and output features utilized only features up to a degree of 1 (scalars and vectors). Fuchs et al. observed that there was big improvement when they switched the maximum degree from 1 to 2.[37] We observed a similar trend especially in features for which relationships between other residues were important such as atomic clashes or side chain angle accuracies. On the other hand, Ramachandran angles and rotamer outlier ratios were not affected as much since they could be predicted well with only localized information. Presumably, the use of higher degree hidden features provided inter-residue information, and this resulted in better predictions.

Finally, we examined if more training data, larger model, and secondary structure-dependent rigid body blocks (SS-dep blocks) could improve the performance. When we increased the training dataset size from 5,690 (6k) to 28,914 (29k) with the baseline model, there was marginal improvement. Presumably, the baseline model did not have enough capacity for learning with the bigger training data. Larger models with more layers showed comparable results to the baseline model with the smaller training dataset. (Figure S8) However, the performance with smaller models dropped significantly. We observed that training of models with eight linear layer blocks was unstable and occasionally resulted in poor performance. When larger models were trained using the bigger training dataset, the bigger data could be learned by larger models as they had enough capacity to learn them. They outperformed in terms of clash score and side chain torsion accuracies. Similarly, the use of larger crops (384 residues) slightly improved the MolProbity score. The use of SS-dep blocks contributed to accurately model backbone bonded geometries including bond lengths, angles, and Ramachandran angles. The best performance was achieved by aggregating all these components.

Structure refinement against cryo-EM density map via the cg2all network

The cg2all network enabled local optimization of a CG representation using scoring functions at both the CG and atomistic representations. (Methods S1, and Algorithm S5) In the algorithm, an atomistic structure is generated from a CG structure via a cg2all network for the CG representation. An objective function can be defined as a function of both atomistic and the CG representation. Once the objective function is evaluated, the score is backpropagated to get derivatives of the CG structure. Then, the CG structure is updated using the derivative. We applied this algorithm to optimize incorrect protein model structures against cryo-EM density maps as an example usage. Nine protein structures were arbitrarily selected from the test set, and their biological assemblies were set as target structures: 1kq1[85] (369 residues, 6-mer), 1vim[86] (760 residues, 4-mer), 1g2o[87] (786 residues, 3-mer), 2ibp[88] (814 residues, 2-mer), 1a2z[89] (880 residues, 4-mer), 1s57[90] (906 residues, 6-mer), 3isr[64] (1,149 residues, 4-mer), 1j0h[91] (1,176 residues, 2-mer), and 1wur[92] (1,848 residues, 10-mer). For those experimental structures, synthetic electron density maps were generated using "molmap" command in UCSF Chimera[65] that employs EMAN2's "pdb2mrc" program[66] at resolutions of 3, 4, 5, 6, 8, and 10 Å. Initial models for local optimization against the electron density maps were predicted by AlphaFold-Multimer[67] with multiple sequence alignments from the ColabFold API[93] and without structural templates using ESMFold.[13] We tested our local optimization protocol that utilized an objective function based on both CG and atomistic representations and compared with alternative protocols, including local optimizations at either CG or atomistic representations and molecular dynamics flexible fitting (MDFF) protocol.[54]

The objective function for local optimization against electron density map consisted of four objective functions in either atomistic or the CG representation. The first one was the electron density map potential taken from MDFF (Equation 10):

$$U(\{\overrightarrow{\boldsymbol{x}}\}) = \sum_i w_i \max\left(1, 1 - \frac{\Phi\left(\overrightarrow{\boldsymbol{x}}_i\right) - \Phi_{\mathrm{thr}}}{\Phi_{\max} - \Phi_{\mathrm{thr}}}\right) \quad \text{(Equation 10)}$$

where $\Phi(x_i)$ refers to the values of the density map.

For this test, we set Φ_{thr} to zero and w_i to corresponding atom's atomic mass. The potential was evaluated at the atomistic representation using predicted coordinates from a Cα-trace using the cg2all model network. The second one was backbone bonded potential at the atomistic representation (Equation 4) The third one was a simple CG potential that evaluated pseudo-bond length and angle potential energies (Equation 11) and soft-core van der Waals potential energy (Equation 12).

$$U_{\mathrm{bonded}} = \left(\frac{b_{C\alpha-C\alpha} - \overline{b_{C\alpha-C\alpha}}}{\sigma(b_{C\alpha-C\alpha})}\right)^2 + \left(\frac{\theta_{C\alpha-C\alpha-C\alpha} - \overline{\theta_{C\alpha-C\alpha-C\alpha}}}{\sigma(\theta_{C\alpha-C\alpha-C\alpha})}\right)^2 \quad \text{(Equation 11)}$$

$$U_{vdW} = \sum_{ij} \left(\min(0, d_{ij} - d_{ij,\min})\right)^2 \quad \text{(Equation 12)}$$

Parameters for the potential functions were obtained from a statistical analysis on the Top8000 structure. As the final one, Cα-Cα distance restraints were applied to residue pairs for which the distances in the initial model were closer than 10 Å. For this experiment, we set relative weights to 1:1:0.1:100.

Initial protein model structures were locally optimized at either residue center-of-mass or Cα-trace representations using a cg2all network-based optimization algorithm (Methods S1 and Algorithm S1). For this test, we superposed the initial structures onto their target experimental structures using MM-align[94] as initial fits to electron density maps, and the superposition was iteratively updated as well by optimizing the overall structural translation and rotation against the density maps. A structure was optimized using the Adam optimizer[95] for 1,000 or 2,500 steps for AlphaFold and ESMFold models, respectively, and intermediate snapshots were recorded for every 100 steps. The learning rate was updated every step using a cosine annealing scheduler,[96] which changed the learning rate from 0.005 to 0.0005 for 200 steps. From the snapshots, a structure with the highest cross-correlation coefficients (CCC) to the target density map was selected as an optimized structure. As alternatives, we performed local optimization with a CG representation using only CG-level objective functions. For the electron density map potential, we used the total mass of a residue as w_i instead. Optimization at a CG representation was carried out in the same way. Then, atomistic structures were generated from Cα-traces using the cg2all network, and the highest CCC structure was selected as an optimized structure using the CG representation. For the local optimization protocol at atomistic resolution, we took the initial minimization step of MDFF protocol implemented by CHARMM-GUI[97] and modified it not to use positional restraints. It performed local energy minimization for 1,000 steps in vacuum using the CHARMM36m force field using NAMD,[98] the MDFF electron density map potential, and restraints for secondary structure elements, chirality, and for fixing *cis*-peptide bonds. We also applied the local optimization protocol at atomistic representation to optimized structures from cg2all network-based optimization and optimization at the CG representations for better agreement of solvent exposed sidechains. As the final option, we carried out the full MDFF protocol implemented by CHARMM-GUI.[97]

QUANTIFICATION AND STATISTICAL ANALYSIS

The performance of the machine-learning model was assessed by analyzing its performance on independent test sets that are distinct from training and validation sets as described in the method details section. Performance metrics reported in Tables 1, 2, 3, and 4 as well as additional results shown in Figures S2–S9 were averaged over test sets consisting of 720 structures. Standard deviations are reported in Tables 1, 2, 3, and 4 to indicate the statistical variation of each metric. In addition, the results from the ablation study in Figure S8 are averaged over three independent training runs. CryoEM refinement results were averaged over nine and six different structures for AlphaFold2 and ESMFold models, respectively. To assess uncertainties, standard errors of the mean were calculated are indicated in Figure 4. We did not perform optimization of ESMFold models for 1wur (ESMFold could not model a structure due to its large size), 1j0h and 1s57 (very poor initial model quality; Cα-RMSDs of 56.1 and 46.0 Å, respectively). We performed statistical analysis using in-house Python scripts based on functions from the NumPy package. Details of the experiments can be found in the figure legends and table footnotes.

ADDITIONAL RESOURCES

cg2all is demonstrated at https://huggingface.co/spaces/huhlim/cg2all and https://colab.research.google.com/github/huhlim/cg2all/blob/main/cg2all.ipynb. A Google Colab notebook for local optimization with cryo-EM density map is available at https://colab.research.google.com/github/huhlim/cg2all/blob/main/cryo_em_minimizer.ipynb.