

Navigating protein landscapes with a machine-learned transferable coarse-grained model

Received: 8 November 2023

Accepted: 11 June 2025

Published online: 18 July 2025

 Check for updates

Nicholas E. Charron^{1,2,3,4,15}, Klara Bonneau^{2,15}, Aldo S. Pasos-Trejo^{2,15}, Andrea Guljas^{2,15}, Yaoyi Chen⁵, Félix Musil², Jacopo Venturin², Daria Gusew², Iryna Zaporozhets^{2,4,6}, Andreas Krämer⁵, Clark Templeton², Atharva Kelkar⁵, Aleksander E. P. Durumeric⁵, Simon Olsson⁷, Adrià Pérez^{8,9}, Maciej Majewski^{8,9}, Brooke E. Husic¹⁰, Ankit Patel^{11,12}, Gianni De Fabritiis^{8,9,13}✉, Frank Noé^{2,5,6,14}✉ & Cecilia Clementi^{2,3,4,6}✉

The most popular and universally predictive protein simulation models employ all-atom molecular dynamics, but they come at extreme computational cost. The development of a universal, computationally efficient coarse-grained (CG) model with similar prediction performance has been a long-standing challenge. By combining recent deep-learning methods with a large and diverse training set of all-atom protein simulations, we here develop a bottom-up CG force field with chemical transferability, which can be used for extrapolative molecular dynamics on new sequences not used during model parameterization. We demonstrate that the model successfully predicts metastable states of folded, unfolded and intermediate structures, the fluctuations of intrinsically disordered proteins and relative folding free energies of protein mutants, while being several orders of magnitude faster than an all-atom model. This showcases the feasibility of a universal and computationally efficient machine-learned CG model for proteins.

Over the past 50 years, substantial developments in hardware, software and theory have advanced the simulation of macromolecules from proof of principle to in silico study of protein folding and dynamics^{1,2}. Despite this success, an ongoing challenge of the field is the accurate and efficient representation of large, biologically relevant systems. These systems are inherently multiscale: while fine-grained models must be used to describe local and fast processes, the long-timescale dynamics may be better captured at a coarse-grained (CG) resolution, which is both more computationally efficient and facilitates a more direct understanding of how macroscopic observables arise from interactions between microscopic degrees of freedom. Up to now, the most successful and widely used simulation approach is molecular dynamics

(MD) with all-atom resolution^{1,3}. Atomistic MD effectively models macromolecular changes, such as protein folding or protein–ligand binding, and predicts their thermodynamic properties. However, all-atom MD comes at extreme computational costs and requires great efforts to post-process and analyse the data^{4,5}. It is still unclear whether there is a computationally efficient CG scale that lends itself to a general and accurate simulation model. Although deep-learning methods have been wildly successful in predicting protein structure and function by reasoning over large-scale genomic and structure datasets^{6,7}, they often do not tie into a physical level of understanding. In this Article we show that deep learning can be used to develop a universal CG protein force field capable of predicting protein structures, structure transitions,

A full list of affiliations appears at the end of the paper. ✉e-mail: gianni.defabritiis@upf.edu; franknoe@microsoft.com; cecilia.clementi@fu-berlin.de

folding mechanisms, folding upon binding of an intrinsically disordered peptide, and changes of free energy upon mutation, similar to all-atom MD methods, but orders of magnitude faster.

Most MD simulation studies employ atomistic force fields fitted on a combination of quantum-chemical calculations and experimental data. Modern force fields have been shown to be qualitatively accurate for processes on nanosecond to millisecond timescales and are often quantitatively consistent with experiments^{2,8}. Recently introduced machine-learned force fields^{9–11} may capture the quantum-mechanical interactions between nuclei in the Born–Oppenheimer approximation even more accurately than conventional MD, but they also come at higher computational cost¹².

Ever since the first protein simulations, the community has striven to develop universal (CG) macromolecular models that are computationally more efficient and more simple to analyse. The feasibility of such models is justified by statistical mechanical descriptions of protein dynamics, such as energy landscape theory¹³, and results from decades of analysis of atomistic simulations⁴. These studies suggest that a protein's free energy landscape can be sufficiently described by a reduced number of collective variables with minimal loss of accuracy compared with atomistic MD. Some CG models have shown success in specific systems. These include structure-based models¹⁴, which rely on the known native structure of a protein to explore its free energy landscape, the Martini¹⁵ CG force field, which can effectively model intermolecular interactions including membrane structure formation and protein interactions, and CG force fields developed to model protein folding and conformational dynamics such as UNRES¹⁶ or AWSEM¹⁷. These models are limited to system-specific applications. For instance, Martini inaccurately models intramolecular protein dynamics, and UNRES and AWSEM often do not capture alternative metastable states.

The main hindrance to the development of an accurate biomolecular CG model is the difficulty in efficiently modelling multi-body interaction terms, which are essential to realistically represent correct protein thermodynamics and implicit solvation effects^{18,19}. In contrast, classical all-atom force fields model most non-bonded interactions as a sum of two-body terms.

Bottom-up CG force fields²⁰ are typically fit to match the equilibrium distribution of an all-atom model, so they could in principle reach atomistic-level accuracy and predictiveness. By leveraging recent developments in deep learning, it has become possible to machine-learn such many-body CG force fields using neural networks^{18,21–30}. In particular, using the variational force-matching approach^{31,32}, such force fields have been shown to accurately reproduce the all-atom distributions of CG observables for single^{18,21,22,25,27,33} and multiple proteins²⁴. Despite these advancements, a transferable CG force field that could be considered as universal, quantitative, predictive and as reliable as a modern atomistic force field is still missing³⁴.

In this Article we propose a neural network-based CG model that is truly transferable in sequence space. We learn the model parameters using a bottom-up approach from atomistic simulations of a set of proteins and then use it to successfully simulate the conformational dynamics of proteins never seen at any learning stage, with low (16–40%) sequence similarities to the training or validation protein set. This CG model is orders of magnitude faster than all-atom MD simulations, predicts metastable folded, unfolded and intermediate states comparable with all-atom MD simulations, and is consistent with experimental data for larger proteins, such as relative folding free energies of protein mutants, where converged all-atom simulations are not available. These results indicate that the CG model 'learns' to represent effective physical interactions between the CG degrees of freedom and provides strong support for the hypothesis that, using deep-learning methods, a universal CG model for realistic and predictive protein simulations at low computational cost is within reach.

Results

We generated a dataset of all-atom explicit solvent simulations of small proteins with diverse folded structures, as well as many combinations of dimers of mono- and dipeptides. Using the training data, we trained a CG force field, CGSchNet²², and conducted extensive simulations of the learned CG model on new, unseen proteins of various sizes and structures (details are presented in Fig. 1, Methods and Supplementary Section 1).

Conformational landscape of peptides and small proteins

To assess the ability of our approach for learning a transferable CG force field, we first tested how it reproduces the folding/unfolding free energy landscape of all-atom MD simulations for a set of unseen 8-peptides (Fig. 2a,b) and unseen small fast-folding proteins: the 025 mutant of chignolin (PDB 2RVD; Fig. 2c), TRPcage (2JOF; Fig. 2d), the beta–beta–alpha fold (BBA) (1FME; Fig. 2e) and the villin headpiece (1YRF; Fig. 2f). None of these proteins had sequence similarity >40% to any stretch of sequence from the training or validation datasets (Table 1 and Supplementary Section 5.3). The free energy surfaces of the CG model were obtained through parallel-tempering (PT) simulations to ensure converged sampling of the equilibrium distribution (Supplementary Section 4.2). Long constant-temperature (300 K) Langevin simulations were also performed for comparison, producing consistent results and multiple folding/unfolding events for all proteins (Supplementary Sections 6.4 and 6.9). We also obtained converged folding/unfolding reference landscapes from atomistic simulations for comparison.

The free energy landscapes of the two 8-peptides match the atomistic references closely (Fig. 2a,b). These peptides are mostly disordered, and their landscapes are mostly determined by the torsional dynamics contained in the prior energy term of the model, whereas the machine-learned multi-body terms have a small effect on the result. In contrast, for the four fast-folding proteins (Fig. 2c–f), the neural network must learn to predict the configurational landscape; control simulations with only the prior energy term only visit the unfolded state for these proteins (Supplementary Fig. 7). For these systems, the CG model predicts metastable folding and unfolding transitions, and the CG folded states are predicted with a fraction of native contacts Q close to 1 and low C_{α} root-mean-square deviation (r.m.s.d.) values, and they are populated with structures closely resembling the correct native state (Fig. 2c–f). For chignolin, the model is also able to stabilize the same misfolded state with misaligned TYR1 and TYR2 residues, as found in the reference atomistic simulations (Fig. 4).

For three of the four fast-folding proteins in Fig. 2, the free energy basin containing the native state is the global minimum, whereas for BBA it is a local minimum, indicating that all proteins are able to fold/unfold correctly (also Supplementary Fig. 25). However, the relative free energy differences between the metastable states do not exactly match the reference. The model performs better on chignolin, TRPcage and villin than on BBA, which contains both helical and anti-parallel β -sheet motifs. Previous CG models have also noted difficulty on this target system^{24,35}.

Extrapolation on larger proteins

To assess the ability of our CG model to fold and maintain the folded states of larger and more complicated systems, we considered the following proteins: 54-residue engrailed homeodomain (1ENH) and 73-residue de novo designed protein alpha3D (2A3D) (Fig. 3). The sizes of these proteins prevent atomistic simulations from sampling the folding/unfolding transitions in reasonable time, whereas the full free energy landscape can be easily explored by the CG model. Therefore, we simulate the folded state with the atomistic force field and compared these dynamics with those of our CG model, defining the lowest free energy minimum as the CG folded state. From extended configurations, the model simulates the folding of both proteins to their correct native structure (Fig. 3a,b).

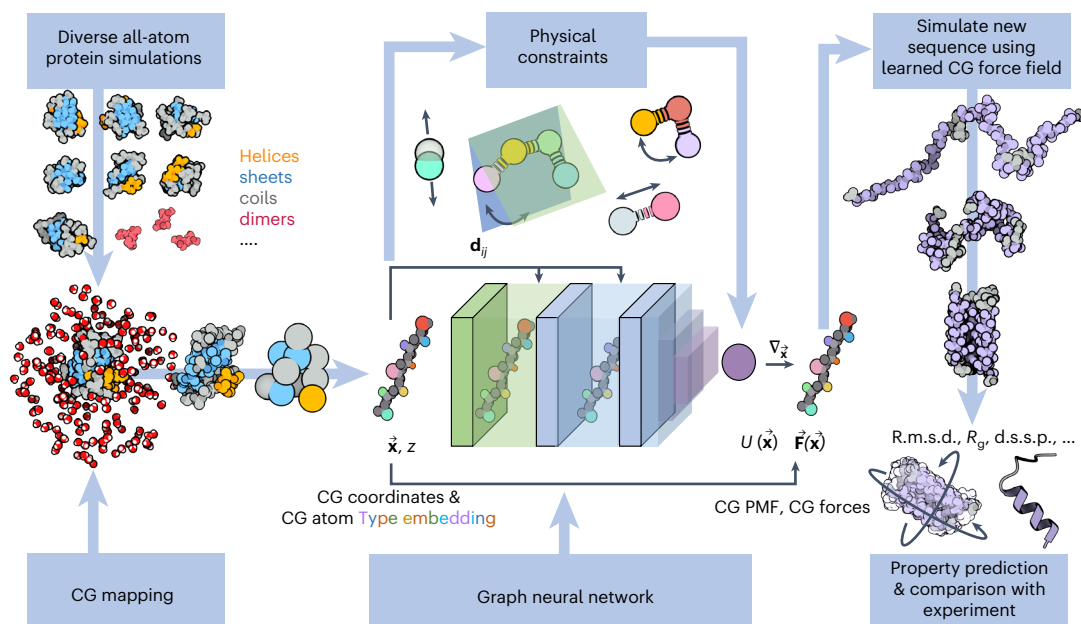


Fig. 1 | Conceptual illustration of the pipeline for building and testing a transferable, bottom-up, machine-learned, CG protein force field. Pipeline for building and testing a transferable, bottom-up, machine-learned, CG protein force field from a diverse dataset of all-atom simulations, a chosen CG resolution, and a set of basic physical prior energy terms (bonds, angles, dihedrals and purely repulsive interactions). The CG atom types z and CG coordinates $\mathbf{x}$

are transformed into pairwise distances d_{ij} are fed into the neural network architecture to predict the CG effective potential energy U and corresponding CG forces $\mathbf{F}$. The trained neural network can subsequently be used to simulate new sequences and predict observables such as root mean square deviations (RMSD), radii of gyration (R_g) or dictionaries of secondary structure in proteins (d.s.s.p.).

We also compared the C_α root-mean-square fluctuations (r.m.s.f.) within the CG folded-state free energy minimum with the reference all-atom simulations. The CG model stabilizes homeodomain in a state very close to the reference native structure, with similar terminal flexibility to the all-atom simulations (Fig. 3a, bottom left) but with slightly higher fluctuations along the length of the sequence. The difference between the folded state predicted by our model and the crystal structure (mean r.m.s.d., ~ 0.5 nm; fraction of native contacts, ~ 0.75) is similar to that of the all-atom simulations from ref. 1 (Supplementary Fig. 12), suggesting the difficulty in accurately predicting homeodomain's crystalline structure.

Our reference simulations of alpha3D show flexibility at the termini as well as between each helical bundle (Fig. 3b), similar to the CG model. The CG model also stabilizes a metastable state of alpha3D very close to the native structure corresponding to an alternative three-helix bundle topology with a different packing of the helices (a detailed analysis is provided in Supplementary Section 6.8 and Supplementary Fig. 24). Alpha3D is a protein designed de novo by iteratively stabilizing the selected native state topology, and precursors of the protein populate both the native state and the alternative topology similar to the one detected by our model³⁶.

These results show that the transferable machine-learned CG model can extrapolate to larger unseen proteins, stabilizing the correct native states and reproducing their associated backbone fluctuations. As an additional analysis, we also demonstrate the extrapolative performance of our CG model on stabilizing the folded states of two large proteins for which we only have experimentally determined structures as reference data, and on reproducing the conformational heterogeneity of a partially disordered protein. The results are discussed in detail in Supplementary Section 6.1.

Detailed analysis and comparison with other CG force fields

We compared the characteristics of the learned CG energy landscapes with the reference simulations and with three other CG force fields with similar resolutions: AWSEM¹⁷, UNRES¹⁶ and Martini¹⁵ (Supplementary

Sections 4.3, 4.4 and 4.5). We note that AWSEM is parameterized to stabilize native states¹⁷, and all presented targets should be stable at this temperature. Similarly, UNRES is parameterized with conformational data for multiple systems at several temperatures at ~ 300 K (ref. 16). The Martini force field is unable to stabilize the folded state of a protein without elastic restraints^{37,38}; here, we show Martini simulation results without native restraints to compare the force field's ability to explore the conformational landscape of a protein system without prior knowledge of the protein's structure. In Supplementary Fig. 27, we show that Martini simulations with an elastic network only allow for small fluctuations around the native structure.

Figure 4 shows the free energy landscapes of the four small fast-folding proteins from Fig. 2 as a function of the two slowest time-lagged independent component analysis (TICA) coordinates³⁹, generated from extensive MD simulations using the reference all-atom model, CGSchNet, AWSEM, UNRES and Martini. The all-atom landscapes exhibit the most structure and have the most metastable states, whereas the CG landscapes are smoother. CGSchNet explores much of the all-atom free energy landscape and it clearly resolves folded and unfolded states as well as other metastable states, whereas this behaviour is rarely observed with the other CG models. Often, AWSEM, UNRES and Martini only stabilize a single metastable state, which is either folded or unfolded. This behaviour is expected, as both the AWSEM and UNRES force fields have been primarily parameterized for stabilizing proteins with a more pronounced fold rather than whole free energy landscapes of less stable proteins. Interestingly, there is also appreciable similarity between the all-atom reference and our machine-learned CGSchNet in structural ensembles besides the folded state prediction. For chignolin, all three all-atom main states (folded, misfolded and unfolded) are visited by both CGSchNet and AWSEM, but these are clearly metastable only with CGSchNet. AWSEM explores but does not stabilize the additional states, and UNRES and Martini do not fold chignolin at all. In the landscapes of TRPcage, BBA and villin, these differences are even more striking, as CGSchNet captures several of the metastable states of the all-atom reference, in particular those

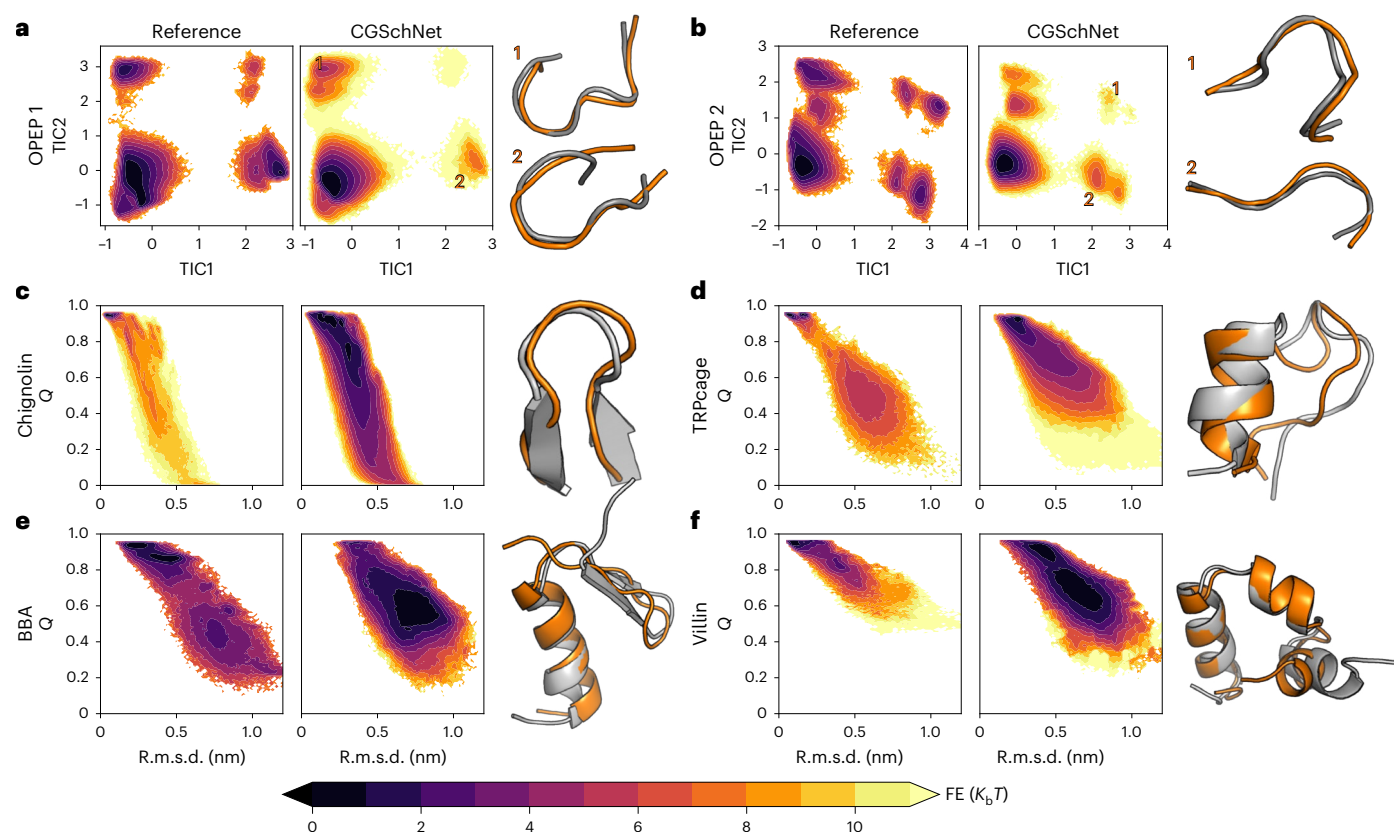


Fig. 2 | Transferable CGSchNet performance on test peptides and proteins. **a–f**, 8-residue peptide DYGC SIHP (**a**), 8-residue peptide SLEAGGRG (**b**), chignolin (**2RVD**) (**c**), TRPcage (**2JOF**) (**d**), BBA (**1FME**) (**e**) and villin (**1YRF**) (**f**). Each subfigure shows the two-dimensional (2D) free energy (FE) surface of the CG model (CGSchNet) and reference atomistic simulations at 300 K as a function of the first two TICA coordinates³⁹ for the two 8-peptides and as a function of the fraction

of native contacts, Q , and the C_{α} r.m.s.d. to the native state for the four small proteins. The structures shown are sampled from the most folded-like metastable basin (or labelled metastable basins for the 8-peptides) for CGSchNet (orange) and atomistic (grey) models. CG free energy landscapes are obtained through Multistate Bennett Acceptance Ratio (MBAR)-reweighted⁶¹ parallel-tempering simulations (details are provided in Supplementary Sections 4.2 and 6.4).

Table 1 | Maximum sequence similarities of test proteins to the proteins used in the model training

Test protein	Length (amino acids)	Sequence similarity to train (%)
DYGC SIHP	8	38
SLEAGGRG	8	50
Chignolin (2RVD)	10	40
TRPcage (2JOF)	20	35
BBA (1FME)	28	29
Villin (1YRF)	35	26
Homeodomain (1ENH)	54	20
SH3 (2NUZ)	55	24
CI2 (2CI2)	65	18
PaaA2 (3ZBE)	71	17
Alpha3D (2A3D)	73	19
S6 (1RIS)	97	16

Details on sequence similarity are in Supplementary Section 5.3.

with a partial fold, but these states are not explored by the other CG force fields. Nevertheless, substantial differences in the unstructured states between the reference and CGSchNet indicate that there is still room for improvement in our CG model.

A quantitative comparison between the folded states obtained with an all-atom model for these proteins and the CG models considered

here is presented in Supplementary Section 5.4. Not only does our model better populate and stabilize the native state structure than the other CG models, but it is comparable to a reference all-atom model (Supplementary Fig. 12). In particular, in the case of homeodomain (**1ENH**), the folded-like metastable state visited by the atomistic model is at a Q -value of around 0.6, lower than in our CG model.

It is important to note that our model is not designed primarily for structure prediction, but rather for the exploration of free energy landscapes for protein systems through CG MD. Unsurprisingly, structures predicted by AlphaFold3⁶ for these test proteins are very close to the corresponding crystal structures, as AlphaFold models are primarily trained to recover PDB structures.

Beyond globular protein folding

Folding upon binding of an intrinsically disordered peptide

To test our CG model's extrapolative ability beyond protein folding, we consider the PUMA-MCL-1 system as a case study of concerted folding and binding of an intrinsically disordered peptide (IDP). The disordered BH3 motif of the PUMA ligand undergoes coupled folding and binding to the induced myeloid leukaemia cell differentiation protein MCL-1⁴⁰. Starting from an extended structure, we simulated the PUMA peptide with our transferable CG model either alone or in the presence of the folded MCL-1 protein. Figure 5 reports the evolution of the C_{α} r.m.s.d. of the ligand to its helical (folded) state during the simulations in both cases. The trajectories of the isolated PUMA (Fig. 5a,b, light blue) exhibit large r.m.s.d. fluctuations, indicating that the peptide remains disordered on its own. By contrast, the peptide simulated in the presence of the MCL-1 protein (orange) rapidly drops to an average r.m.s.d. value of ~2.5 Å, indicating

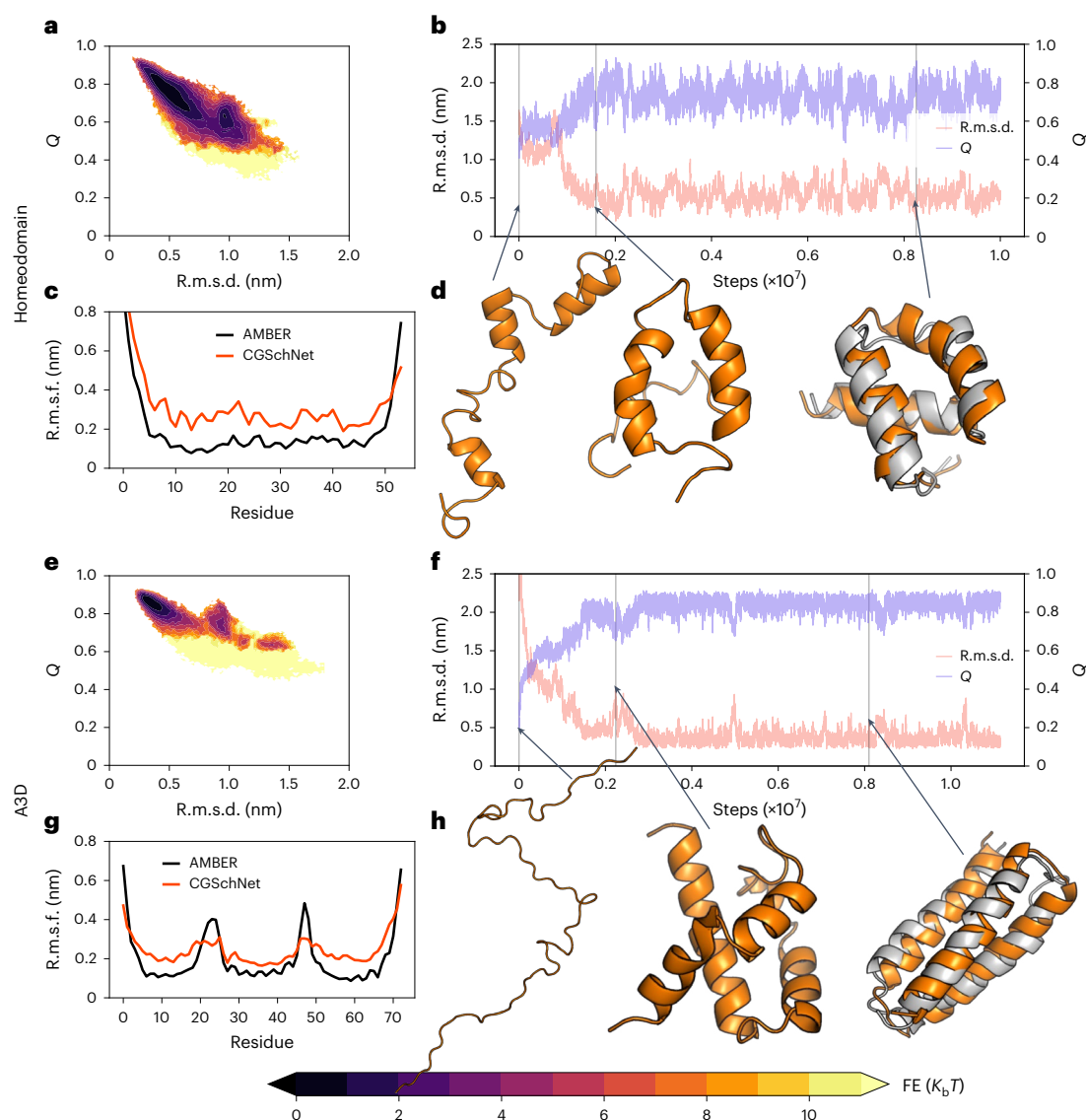


Fig. 3 | Extrapolative performance of CGSchNet on two large proteins withheld from the training and validation sets. a–h, 54-residue homeodomain (PDB 1ENH; **a–d**) and 73-residue alpha3D (PDB 2A3D; **e–h**). C_{α} r.m.s.f. values of the folded state to the crystal structure are shown (taken from a trajectory window that remained folded for more than one million MD steps; Supplementary Section 5.2) in comparison to reference (AMBER) all-atom simulations (**c,g**),

along with an exploration of the free energy surface as a function of the fraction of native contacts, Q , and the C_{α} r.m.s.d. to the native state, obtained through MBAR-reweighted⁶¹ PT simulations (**a,e**). A folding trajectory starting from a completely elongated structure is shown (**b,f**), with orange structures illustrating folding and the crystal structure shown in grey (**d,h**).

induced folding of the peptide by the binding partner. The final simulation snapshot reported on the top right of Fig. 5a reinforces this result. Details about these simulations are provided in Supplementary Section 5.5.

As a control, we simulated the unfolded PUMA ligand with a protein that is not known to induce its folding, ubiquitin (PDB 1D3Z). Here, although the peptide remains close to the protein, in none of the simulated trajectories does it fold into a stable helix, as indicated by the much larger deviations of the r.m.s.d. trace (in purple), and by the final simulation snapshot on the right of Fig. 5b. Together, these results indicate that, when simulated with our CG model, the PUMA peptide forms a stable helix only when in the presence of its correct binding partner, MCL-1. Although the CG model was trained partially on interacting mono/dipeptide pairs (Supplementary Section 1.2), the training data contain no protein–protein complexes such as MCL-1/PUMA. Despite this, the model learns nontrivial interactions that can correctly model the PUMA peptide both alone and in the presence of its correct binding partner.

Mutational analysis of ubiquitin

We illustrate the extrapolative power of our chemically transferable CG model in estimating folding free energy changes upon mutations, comparable to experimentally measured $\Delta\Delta G$ values, as described in Methods and Supplementary Section 5.6. Such mutational analysis is straightforward using our CG model, because the identity of an amino acid is solely defined by the type of the C_{β} bead in our model (or the C_{α} bead for GLY): mutations can be performed simply by changing these bead types as illustrated in Fig. 6a. We chose ubiquitin as a test system, given its extensive and available experimental data, focusing specifically on the set of conservative mutations investigated by Went and Jackson⁴¹. We note that ubiquitin has only 18% sequence similarity with any proteins in the training/validation datasets.

Figure 6b presents the comparison between the experimental $\Delta\Delta G$ values from ref. 41 and those obtained by our model as described in Supplementary Section 5.6. There is a strong correlation between our results and the experimental values; the Pearson correlation coefficient obtained

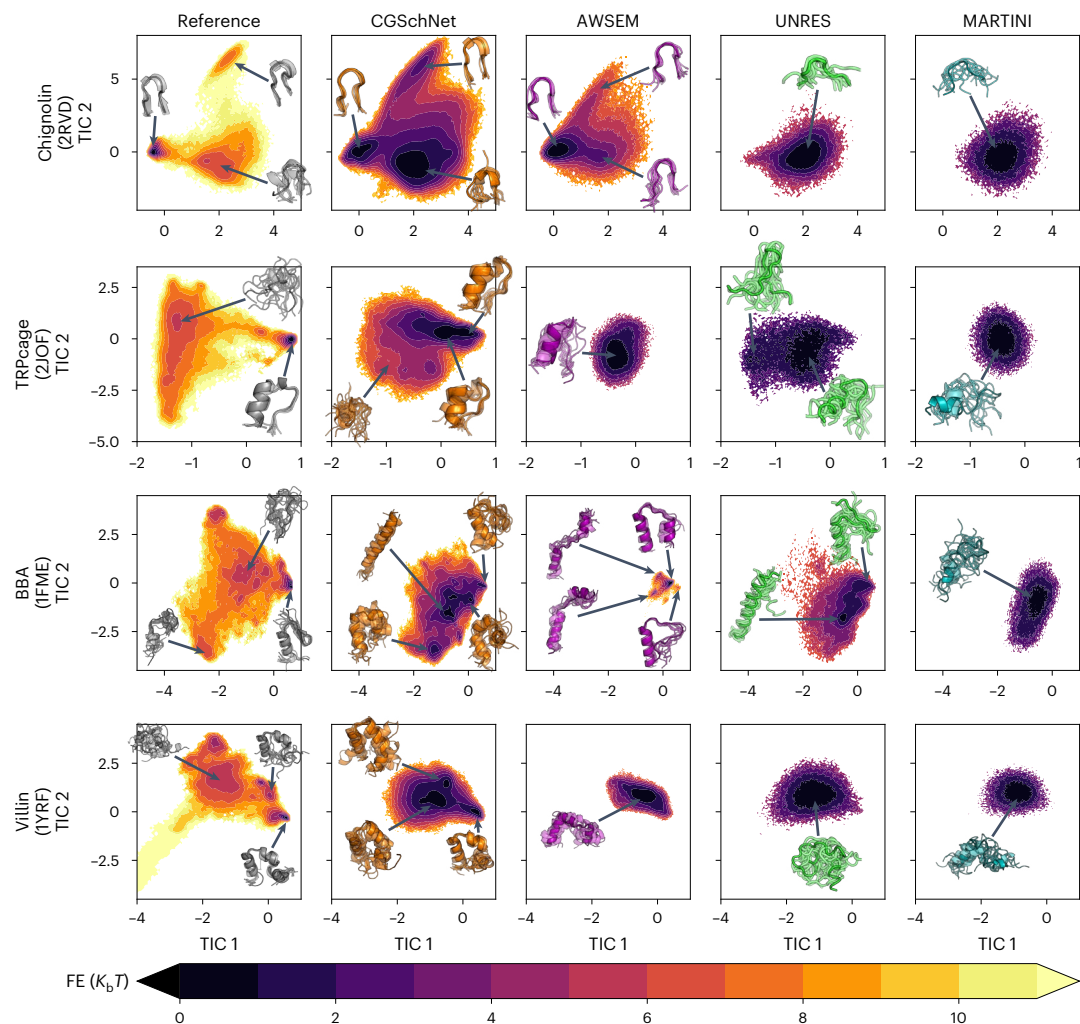


Fig. 4 | Free energy landscapes as a function of the first two TICA coordinates for four small, fast folding proteins. Results are presented for the reference atomistic simulations, for Langevin simulations of our CG model (CGSchNet),

and for three different CG force fields of a comparable resolution to our CG model, AWSEM, UNRES and Martini. Each landscape includes representative structures from different metastable minima.

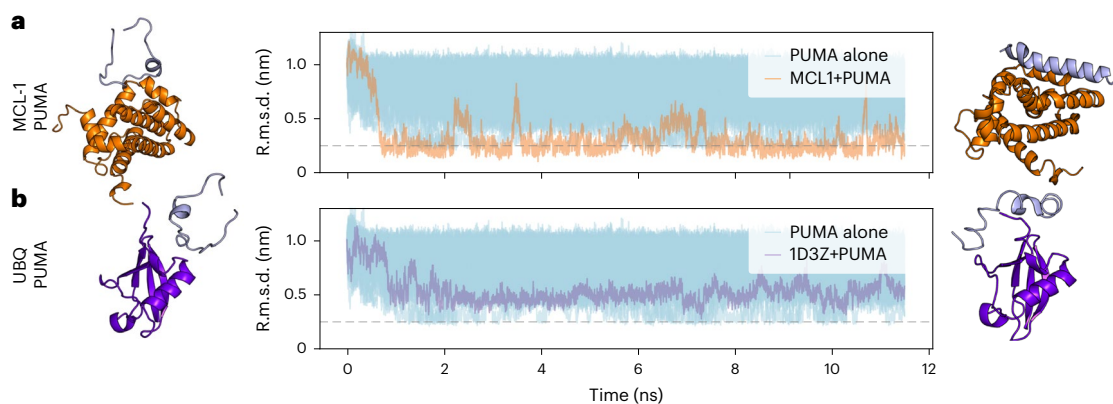


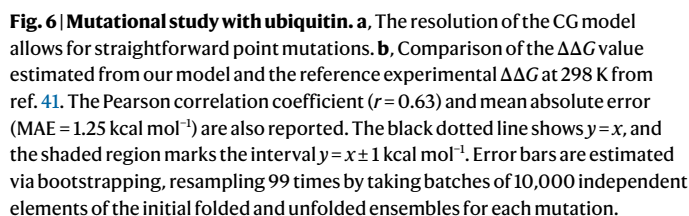
Fig. 5 | Simulation of the PUMA peptide. a, Time series of the r.m.s.d. of the PUMA ligand to its reference helical (folded) structure from the PDB (2ROC) when simulated alone (light blue traces) or close to the folded MCL-1 protein (orange traces). On the left and on the right are shown the starting structure

and a structure at the end of the simulation, respectively. **b,** As in **a** but when the PUMA peptide is simulated close to ubiquitin (PDB 1D3Z), which is not its correct binding partner. The horizontal dashed line marks the 0.25 nm threshold for comparison.

with our CG model is comparable with what can be obtained with all-atom approaches^{42,43}. This result indicates that the model has learned the general physical interactions among the residues at the CG resolution, thereby allowing useful predictions on new systems such as mutation effects.

Discussion and conclusion

We have shown that it is possible to machine-learn a transferable, bottom-up, CG effective force field that can be used for MD simulation on proteins with little sequence overlap with the systems used for



In contrast to models such as AlphaFold⁶, our model is not a protein structure prediction tool. Rather, it explores the complete free energy surface of the systems of interest, including but not limited to the folded protein structure. The ultimate goal of our model would be to reproduce the thermodynamics of our systems consistently with the underlying atomistic model, but there are some protein targets, such as BBA, where our model predicts the folded state as being less stable than the unfolded. Yet, even in these systems, our model predicts a metastability in the folded region of BBA's free energy landscape, whereas other CG models do not. Further improving the reliability of free energy predictions is an important future aim that will require both expanding the training dataset and further method development.

The particular neural network chosen here (SchNet⁴⁹) is quite simple. It consists of a series of continuous-filter convolutions and does not include an attention mechanism, nor explicit long-range interaction terms. This architectural choice was motivated by the goal of developing a ‘proof-of-concept’ model, that can be trained and simulated as fast as possible while still yielding the desired results. More sophisticated architectural choices could produce better-performing CG models. In particular, the lack of long-range interactions in our model may affect the model performance on much larger and multi-protein systems, where electrostatic interactions may play an important role^{50,51}. Multiple approaches for including long-range interactions^{52–55} and/or attention mechanisms⁵⁶ have been recently proposed for all-atom resolutions and could be incorporated into our modelling framework in the future.

It is also important to note that our CG model was trained at a specific thermodynamic condition. Transferability in temperature/pressure or other additional environmental parameters is therefore not expected at this point. In particular, the temperature dependence of the CG effective energy is highly nontrivial, as it really represents a free energy with an entropic component⁵⁷. An explicit dependence of the model on thermodynamic conditions could, in principle, be included in the framework^{58,59}. However, in practice, its training would require the curation of a substantially larger dataset encompassing multiple simulations at multiple thermodynamic conditions, which would probably require even more large-scale computational resources than those used in this work.

The results presented here were obtained with a model that, although aggressively coarse-grained with respect to an explicit water atomistic model, still retains the full backbone heavy atoms and an additional atom per side chain (excluding GLY). We have not yet investigated alternative resolutions for building a transferable model, and we expect transferability to be strongly tied to the chosen resolution. Although different methods have been proposed for the simultaneous optimization of CG effective energy and CG mapping⁶⁰, it remains unclear if and which additional resolutions allow for the optimal design of a transferable and quantitatively accurate model. We believe that the results presented here open the way to a systematic investigation of this point.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41557-025-01874-0>.

References

- Lindorff-Larsen, K., Piana, S., Dror, R. O. & Shaw, D. E. How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
- Plattner, N., Doerr, S., De Fabritiis, G. & Noé, F. Complete protein–protein association kinetics in atomic detail revealed by molecular dynamics simulations and Markov modelling. *Nat. Chem.* **9**, 1005–1011 (2017).
- Voelz, V. A., Bowman, G. R., Beauchamp, K. & Pande, V. S. Molecular simulation of ab initio protein folding for a millisecond folder NTL9(1–39). *J. Am. Chem. Soc.* **132**, 1526–1528 (2010).
- Noé, F. & Clementi, C. Collective variables for the study of long-time kinetics from molecular trajectories: theory and methods. *Curr. Opin. Struct. Biol.* **43**, 141–147 (2017).
- Husic, B. E. & Pande, V. S. Markov state models: from an art to a science. *J. Am. Chem. Soc.* **140**, 2386–2396 (2018).
- Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
- Lewis, S. et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. Preprint at *bioRxiv* <https://www.biorxiv.org/content/10.1101/2024.12.05.626885v1> (2024).
- Casasnovas, R., Limongelli, V., Tiwary, P., Carloni, P. & Parrinello, M. Unbinding kinetics of a p38 map kinase type II inhibitor from metadynamics simulations. *J. Am. Chem. Soc.* **139**, 4780–4788 (2017).
- Unke, O. T. et al. Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments. *Sci. Adv.* **10**, eadn4397 (2024).
- Kozinsky, B., Musaelian, A., Johansson, A. & Batzner, S. Scaling the leading accuracy of deep equivariant models to biomolecular simulations of realistic size. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis* article no. 2, 1–12 (ACM, 2023).
- Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
- Rufa, D. A. et al. Towards chemical accuracy for alchemical free energy calculations with hybrid physics-based machine learning/molecular mechanics potentials. Preprint at *bioRxiv* <https://doi.org/10.1101/2020.07.29.227959> (2020).
- Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
- Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: what determines the structural details of the transition state ensemble and ‘en-route’ intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953 (2000).
- Souza, P. C. T. et al. Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nat. Methods* **18**, 382–388 (2021).
- Liwo, A. et al. A general method for the derivation of the functional forms of the effective energy terms in coarse-grained energy functions of polymers. III. Determination of scale-consistent backbone-local and correlation potentials in the UNRES force field and force-field calibration and validation. *J. Chem. Phys.* **150**, 155104 (2019).
- Davtyan, A. et al. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B* **116**, 8494–8503 (2012).
- Wang, J. et al. Multi-body effects in a coarse-grained protein force field. *J. Chem. Phys.* **154**, 164113 (2021).
- Zaporozhets, I. & Clementi, C. Multibody terms in protein coarse-grained models: a top-down perspective. *J. Phys. Chem. B* **127**, 6920–6927 (2023).
- Jin, J., Pak, A. J., Durumeric, A. E. P., Loose, T. D. & Voth, G. A. Bottom-up coarse-graining: principles and perspectives. *J. Chem. Theory Comput.* **18**, 5759–5791 (2022).
- Wang, J. et al. Machine learning of coarse-grained molecular dynamics force fields. *ACS Cent. Sci.* **5**, 755–767 (2019).
- Husic, B. E. et al. Coarse graining molecular dynamics with graph neural networks. *J. Chem. Phys.* **153**, 194101 (2020).
- Ding, X. & Zhang, B. Contrastive learning of coarse-grained force fields. *J. Chem. Theory Comput.* **18**, 6334–6344 (2022).
- Majewski, M. et al. Machine learning coarse-grained potentials of protein thermodynamics. *Nat. Commun.* **14**, 5739 (2023).
- Köhler, J., Chen, Y., Krämer, A., Clementi, C. & Noé, F. Flow-matching: efficient coarse-graining of molecular dynamics without forces. *J. Chem. Theory Comput.* **19**, 942–952 (2023).
- Chennakesavalu, S., Toomer, D. J. & Rotskoff, G. M. Ensuring thermodynamic consistency with invertible coarse-graining. *J. Chem. Phys.* **158**, 124126 (2023).
- Krämer, A. et al. Statistically optimal force aggregation for coarse-graining molecular dynamics. *J. Phys. Chem. Lett.* **14**, 3970–3979 (2023).
- Wellawatte, G. P., Hocky, G. M. & White, A. D. Neural potentials of proteins extrapolate beyond training data. *J. Chem. Phys.* **159**, 085103 (2023).
- Airas, J., Ding, X. & Zhang, B. Transferable implicit solvation via contrastive learning of graph neural networks. *ACS Cent. Sci.* **9**, 2286–2297 (2023).
- Marloes Arts, V. G. S. et al. Two for one: diffusion models and force fields for coarse-grained molecular dynamics. *J. Chem. Theory Comput.* **19**, 6151–6159 (2023).
- Izvekov, S. & Voth, G. A. A multiscale coarse-graining method for biomolecular systems. *J. Phys. Chem. B* **109**, 2469–2473 (2005).
- Noid, W. G. et al. The multiscale coarse-graining method. I. A rigorous bridge between atomistic and coarse-grained models. *J. Chem. Phys.* **128**, 244114 (2008).
- Chen, Y. et al. Machine learning implicit solvation for molecular dynamics. *J. Chem. Phys.* **155**, 084101 (2021).
- Durumeric, A. E. P. et al. Machine learned coarse-grained protein force-fields: are we there yet? *Curr. Opin. Struct. Biol.* **79**, 102533 (2023).
- Liwo, A., Czaplewski, C., Pillardy, J. & Scheraga, H. A. Cumulant-based expressions for the multibody terms for the correlation between local and electrostatic interactions in the united-residue force field. *J. Chem. Phys.* **115**, 2323–2347 (2001).
- Bryson, J. W., Desjarlais, J. R., Handel, T. M. & Degradó, W. F. From coiled coils to small globular proteins: design of a native-like three-helix bundle. *Protein Sci.* **7**, 1404–1414 (1998).
- Periole, X., Cavalli, M., Marrink, S.-J. & Ceruso, M. A. Combining an elastic network with a coarse-grained molecular force field: structure, dynamics and intermolecular recognition. *J. Chem. Theory Comput.* **5**, 2531–2543 (2009).
- Poma, A. B., Cieplak, M. & Theodorakis, P. E. Combining the MARTINI and structure-based coarse-grained approaches for the molecular dynamics studies of conformational transitions in proteins. *J. Chem. Theory Comput.* **13**, 1366–1374 (2017).
- Pérez-Hernández, G., Paul, F., Giorgino, T., De Fabritiis, G. & Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **139**, 015102 (2013).

40. Rogers, J. M. et al. Interplay between partner and ligand facilitates the folding and binding of an intrinsically disordered protein. *Proc. Natl Acad. Sci. USA* **111**, 15420–15425 (2014).
41. Went, H. M. & Jackson, S. E. Ubiquitin folds through a highly polarized transition state. *Protein Eng Des. Sel.* **18**, 229–237 (2005).
42. Zhang, Z. et al. Predicting folding free energy changes upon single point mutations. *Bioinformatics* **28**, 664–671 (2012).
43. Gapsys, V., Michielssens, S., Seeliger, D. & Groot, B. L. Accurate and rigorous prediction of the changes in protein free energies in a large scale mutation scan. *Angew. Chem. Int. Ed.* **55**, 7364–7368 (2016).
44. Vendruscolo, M. & Domany, E. Pairwise contact potentials are unsuitable for protein folding. *J. Chem. Phys.* **109**, 11101–11108 (1998).
45. Ejtehadi, M., Avall, S. & Plotkin, S. Three-body interactions improve the prediction of rate and mechanism in protein folding models. *Proc. Natl Acad. Sci. USA* **101**, 15088–15093 (2004).
46. Scherer, C. & Andrienko, D. Understanding three-body contributions to coarse-grained force fields. *Phys. Chem. Chem. Phys.* **20**, 22387–22394 (2018).
47. Kar, P. & Feig, M. Recent advances in transferable coarse-grained modeling of proteins. *Adv. Protein Chem. Struct. Biol.* **96**, 143–180 (2014).
48. Hills Jr, R. D., Lu, L. & Voth, G. A. Multiscale coarse-graining of the protein energy landscape. *PLoS Comput. Biol.* **6**, 1000827 (2010).
49. Schütt, K. T., Sauceda, H. E., Kindermans, P.-J., Tkatchenko, A. & Müller, K.-R. SchNet: a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
50. Sheinerman, F. B. & Honig, B. On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318**, 161–177 (2002).
51. Zhang, Z., Witham, S. & Alexov, E. On the role of electrostatics in protein–protein interactions. *Phys. Biol.* **8**, 035001 (2011).
52. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).
53. Unke, O. T. et al. SpookyNet: learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 7273 (2021).
54. Kosmala, A., Gasteiger, J., Gao, N. & Günnemann, S. Ewald-based long-range message passing for molecular graphs. In *Proc. International Conference on Machine Learning* (PMLR, 2023).
55. Caruso, A. et al. Extending the RANGE of graph neural networks: relaying attention nodes for global encoding. Preprint at <https://arxiv.org/abs/2502.13797> (2025).
56. Thölke, P. & Fabritius, G. D. Equivariant transformers for neural network based molecular potentials. In *10th International Conference on Learning Representations* (Curran Associates, 2022).
57. Kidder, K. M., Szukalo, R. J. & Noid, W. Energetic and entropic considerations for coarse-graining. *Eur. Phys. J. B* **94**, 153 (2021).
58. Krishna, V., Noid, W. G. & Voth, G. A. The multiscale coarse-graining method. IV. Transferring coarse-grained potentials between temperatures. *J. Chem. Phys.* **131**, 024103 (2009).
59. Pretti, E. & Shell, M. S. A microcanonical approach to temperature-transferable coarse-grained models using the relative entropy. *J. Chem. Phys.* **155**, 094102 (2021).
60. Wang, W. & Gómez-Bombarelli, R. Coarse-graining auto-encoders for molecular dynamics. *npj Comput. Mater.* **5**, 125 (2019).
61. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* **129**, 124105 (2008).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2025

¹Department of Supercomputing, Zuse Institute Berlin, Berlin, Germany. ²Department of Physics, Freie Universität Berlin, Berlin, Germany. ³Department of Physics and Astronomy, Rice University, Houston, TX, USA. ⁴Center for Theoretical Biological Physics, Rice University, Houston, TX, USA. ⁵Department of Mathematics and Computer Science, Freie Universität Berlin, Berlin, Germany. ⁶Department of Chemistry, Rice University, Houston, TX, USA. ⁷Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden. ⁸Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona, Spain. ⁹Acellera Labs, Barcelona, Spain. ¹⁰Lewis Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA. ¹¹Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA. ¹²Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. ¹³Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain. ¹⁴AI4Science, Microsoft Research, Berlin, Germany. ¹⁵These authors contributed equally: Nicholas E. Charron, Klara Bonneau, Aldo S. Pasos-Trejo, Andrea Guljas.

✉ e-mail: gianni.defabritius@upf.edu; franknoe@microsoft.com; cecilia.clementi@fu-berlin.de

Methods

We generated a dataset of all-atom explicit solvent simulations of 50 CATH domains⁶², representing small proteins with diverse folded structures, as well as ~1,200 dimers of mono- and dipeptides. We stored all instantaneous forces on the protein atoms, performed basic force aggregation on a CG backbone representation of the proteins²⁷, and trained a CG force field, CGSchNet²², which combines a deep GNN with physically motivated prior terms. We then conducted a series of extensive Langevin and PT simulations of the learned CG model on new, unseen proteins of various sizes and structures to demonstrate its capabilities and limitations. Wherever feasible, we also performed extensive all-atom MD simulations for the test systems and analysed them with Markov state modelling^{4,5} for comparison.

Neural network model

Our model was built using the deep-learning Python packages PyTorch⁶³ and PyTorch Geometric⁶⁴. Building on previous efforts^{22,27,33}, we chose to model the optimizable term of our CG effective energy with the GNN architecture CGSchNet, which is based on a previous architecture, SchNet⁴⁹. See Supplementary Section 2 for a detailed description of network architecture, hyperparameter choices and training routines. The ability to directly learn species-dependent interactions and CG bead-wise features from data represents the primary advantage of using a convolutional GNN such as CGSchNet in this pursuit. More recent GNN architectures^{65,66} may be used as an alternative. However, we note that newer architectures can require more computational resources and more extensive hyperparameter searches, thereby creating substantial training barriers given the large number of MD simulation frames and the system sizes used in training.

Loss function

We designed our CG model within the framework of variational force-matching^{31,32}. In practice, we optimize the parameters $\{\theta\}$ of a network representing the effective energy $\tilde{U}_{\text{CG}}(\mathbf{R}; \{\theta\})$, where $\mathbf{R}$ are the CG coordinates, by minimizing a loss function in the form

$$\chi^2[\tilde{\mathbf{F}}_{\text{CG},\Delta}(\mathbf{R}; \{\theta\})] = \left\langle \frac{1}{3N} \sum_{j=1}^N \left\| [\mathcal{M}_F \mathbf{f}_{AA}(\mathbf{r})]_{\Delta,j} - \tilde{\mathbf{F}}_{\text{CG},\Delta}(\mathbf{R}; \{\theta\})_j \right\|^2 \right\rangle_{\mathbf{r}} \quad (1)$$

Here, N is the number of CG atoms in the system. In equation (1), $\tilde{\mathbf{F}}_{\text{CG},\Delta}(\mathbf{R}; \{\theta\})$ are the forces associated with the CG effective energy, $\tilde{\mathbf{F}}_{\text{CG}}(\mathbf{R}; \{\theta\}) = -\nabla_{\mathbf{R}} \tilde{U}_{\text{CG}}(\mathbf{R}; \{\theta\})$, after subtraction of the ‘prior forces’:

$$\tilde{\mathbf{F}}_{\text{CG},\Delta}(\mathbf{R}; \{\theta\}) = \tilde{\mathbf{F}}_{\text{CG}}(\mathbf{R}; \{\theta\}) - \mathbf{F}_{\text{prior}}(\mathbf{R}) \quad (2)$$

where $\mathbf{F}_{\text{prior}}(\mathbf{R}) = -\nabla_{\mathbf{R}} \tilde{U}_{\text{prior}}(\mathbf{R})$ and $\tilde{U}_{\text{prior}}(\mathbf{R})$ is a pre-fit ‘prior energy’ term. The atomistic force is similarly modified. The definition of a prior energy is discussed in the next section and it has been shown to play an important role in constructing stable and accurate neural network-based CG models by enforcing asymptotic physical behaviour in regions of phase space not covered adequately by training/validation datasets obtained through all-atom MD^{21,22,33,34}. In equation (1), the operator $\mathcal{M}_F$ projects the atomistic forces $\mathbf{f}_{AA}(\mathbf{r})$, as a function of the atomistic coordinates $\mathbf{r}$, in CG space. We have shown in previous work that a careful choice of $\mathcal{M}_F$ is crucial to the optimization of the CG model²⁷. In this Article, $\mathcal{M}_F$ is defined for each CG site as the sum of forces on the preserved atom and neighbouring hydrogen atoms connected via constrained bonds²⁷.

CG resolution and prior energy

A good choice of the prior energy model should be connected to the resolution chosen to define the CG model³⁴. Previous non-transferable CG studies^{22,25,27} have utilized a resolution that retains only the C_α atoms for each amino acid. However, when considering 20 naturally occurring amino acids, the type enumeration for common local energy terms,

such as four-body dihedral interactions, becomes very large. Efforts in the past⁴⁸ have attempted to mitigate such scaling, but this can lead to potentially limiting or overly biasing expressions for the associated prior energies.

For this work we chose to retain the following five atoms for each residue: backbone N, C_α , C, O and side-chain C_β . We label different atoms with an integer atom type, with the C_β atom having a residue-dependent atom type number. In the case of GLY residues, which do not contain a C_β , we retain only four atoms—N, C_α , C and O—and assign the residue-dependent atom type number to the C_α atom. Supplementary Fig. 3 provides a graphical description of the CG resolution and atom type labelling.

This five-bead-per-residue CG mapping is not unprecedented—the successful AWSEM¹⁷ CG force field, which retains the C_α , the C_β and the O atoms (as well as virtual sites for N and C atoms), utilizes a comparable resolution. This choice of CG resolution allows for a direct interpretation of secondary structures and leads to intuitive prior energy choices (for example, physical bond/angle terms, physical dihedral angles and so on). A description of the terms in the prior force field is provided in Supplementary Section 3.1.

It is important to stress that if the prior energy is used alone (without the trainable neural network energy term $\tilde{U}_{\text{CG}}(\mathbf{R}; \{\theta\})$), it is completely incapable of stabilizing any secondary or tertiary protein structures (Supplementary Fig. 7 presents the results from control, prior-only simulations). The function of the prior energy is only to prevent the model from visiting configurationally nonphysical regions (for example, configurations involving overlapping atoms), with little to no additional bias on the configurational landscape. To illustrate the relative importance of each prior on model stability/ability to access unphysical configurations, we include a prior ablation study, where the effect of removing each prior subinteraction, one by one, is investigated on a chignolin-specific model at the same five-bead-per-residue resolution (Supplementary Fig. 10).

Training data

Three strategies were used to create the training dataset. First, to capture sequence and secondary/tertiary structure diversity in proteins, we constructed a dataset of all-atom simulations of 50 protein domains in their native state from the CATH⁶² database (Supplementary Section 1.1 presents the domain selection procedure). Each simulation represents 100,000 frames of all-atom MD data in which the forces and positions of the solute are saved. In addition to this dataset of folded CATH simulations, we constructed a second dataset wherein ~1,200 mono/dipeptide dimer systems were simulated using umbrella sampling with dimer centre-of-mass distances as a reaction coordinate, and each system consists of 27,000 frames. Although this dataset does not contain direct secondary/tertiary structure information, it contains valuable asymptotic force information for atoms that are brought very close together through the nature of the enhanced sampling strategy. The necessity for both the CATH and dimer datasets was demonstrated by an ablation study in which we systematically remove both entire datasets and selected samples (Supplementary Fig. 20). Finally, additional frames were constructed from the previously described CATH and dimers datasets by taking every 50th frame of each simulation and additively combining bead positions with a Gaussian noise of mean 0 and standard deviation 0.5 Å (Supplementary Section 6.6 provides details of the hyperparameter selection). These distorted ‘decoy’ frames were combined with a zero delta-force label and used as additional training data, and are designed to prevent uncontrolled neural network extrapolation on distorted high-energy configurations that may arise transiently during CG simulation. Due to the induced distortion, the prior alone predicts a high baseline energy on the corresponding areas of phase space. Effectively, the decoy training data helps to ensure that the network does not attempt to predict strong forces in configurations that should be dominated by the prior terms. A comprehensive discussion on the training and validation datasets that are used to parameterize the model is provided in Supplementary Section 1.

Mutational analysis

From the CG model, we can estimate $\Delta\Delta G$ values by treating the effect of a single point mutation as a perturbation to the wild-type energy^{67,68}. Under the assumption that the mutation does not significantly perturb the density of states, the effect of the mutation on protein stability can be estimated from perturbation theory^{67,68} (a detailed analysis is provided in Supplementary Section 5.6).

Data availability

All training data, simulation data, and trained models are available at <https://doi.org/10.5281/zenodo.15465782> (ref. 69). Source data are provided with this paper.

Code availability

Code for the generation of the training datasets is available at <https://doi.org/10.5281/zenodo.15465782> (ref. 69). The scripts for model training and running simulations are available at <https://doi.org/10.5281/zenodo.15482457> (ref. 70). Model training and simulation was done using the *mlcg* package (<https://github.com/ClementiGroup/mlcg>). Additional tools for the prior pipeline can be found in the *mlcg-tk* package (<https://github.com/ClementiGroup/mlcg-tk>). Finally, analysis tools can be found in the *Proteka* package (<https://github.com/ClementiGroup/proteka>).

References

62. Sillitoe, I. et al. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res.* **43**, 376–381 (2015).
63. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. In *Proc. Neural Information Processing Systems* Vol. 32 (eds Wallach, H. et al.) article no. 721, 8026–8037 (Curran Associates, 2019).
64. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. In *Proc. ICLR Workshop on Representation Learning on Graphs and Manifolds* (Curran Associates, 2019).
65. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
66. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In *Proc. Advances in Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 11423–11436 (Curran Associates, 2022).
67. Matysiak, S. & Clementi, C. Optimal combination of theory and experiment for the characterization of the protein folding landscape of S6: how far can a minimalist model go? *J. Mol. Biol.* **343**, 235–248 (2004).
68. Matysiak, S. & Clementi, C. Minimalist protein model as a diagnostic tool for misfolding and aggregation. *J. Mol. Biol.* **363**, 297–308 (2006).
69. Charron, N., Bonneau, K., Pasos-Trejo, A. & Guljas, A. Navigating protein landscapes with a machine-learned transferable coarse-grained model (data and codes). *Zenodo* <https://doi.org/10.5281/zenodo.15465782> (2025).
70. Charron, N. et al. ClementiGroup/mlcg: 0.0.3. *Zenodo* <https://doi.org/10.5281/zenodo.15482457> (2025).

Acknowledgements

We thank all members of the Clementi and Noé groups for their help in different phases of this work. We gratefully acknowledge funding from the European Commission (grant no. ERC CoG 772230 ‘ScaleCell’), the International Max Planck Research School for Biology and Computation (IMPRS–BAC), the Bundesministerium für Bildung und Forschung BMBF (Berlin Institute for Learning and Data, BIFOLD, and project FAIME 01IS24076), the Berlin Mathematics Center MATH+ (AA1-6, EF1-2) and the Deutsche Forschungsgemeinschaft DFG (NO 825/2, NO 825/3, NO 825/4; SFB/TRR 186, Project A12; SFB 1114, Projects B03, B08 and A04; SFB 1078, Project C7; and RTG 2433, Projects Q05 and Q04), the National Science Foundation (CHE-1900374 and PHY-2019745) and the Einstein Foundation Berlin (Project 0420815101). We gratefully acknowledge the computing time provided on the supercomputer Lise at NHR@ZIB as part of the NHR infrastructure, and on the supercomputer JUWELS operated by the Jülich Supercomputing Centre. We thank volunteers at GPUGRID.net for contributing computational resources and Acellera for funding.

Author contributions

C.C. and F.N. conceived the project. N.E.C., K.B., A.S.P.-T. and A.G. developed the machine-learning models, trained the models, performed simulations and analysed the data. Y.C., F.M., J.V., D.G., I.Z., A. Krämer, C.T., A. Kelkar and A.E.P.D. contributed to the development of the models and data analysis. N.E.C., F.M., A. Kelkar, Y.C., C.T., K.B., A.G., A.S.P.-T., J.V. and A.E.P.D. wrote and tested the software. A. Krämer, C.T., Y.C., S.O., A. Pérez, M.M., B.E.H. and G.D.F. generated part of the reference data. C.C., F.N., G.D.F. and A. Patel supervised the project. C.C. and F.N. drafted the original version of the paper with the input and approval of all authors.

Funding

Open access funding provided by Freie Universität Berlin.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41557-025-01874-0>.

Correspondence and requests for materials should be addressed to Gianni De Fabritiis, Frank Noé or Cecilia Clementi.

Peer review information *Nature Chemistry* thanks Pratyush Tiwary and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.