Machine Learned Coarse-Grained Protein Force Fields

Efficient calculation of slow and stationary scales in molecular dynamics

***A. S. Pasos-Trejo, N. E. Charron, F. Musil, C. Clementi**, Department of Physics, Freie Universität Berlin*

In Short

- Molecular Dynamics (MD) is a valuable tool to study biophysical processes at molecular resolution.
- Despite recent advances, the computational demand of atomistic MD does not allow the simulation of biologically relevant timescales.
- We are developing accurate and transferable coarse-grained (CG) models, which can be used to simulate large molecular systems on long (beyond millisecond) timescales.
- We use state-of-art machine-learning tools such as graph neural networks and a large set of reference data to train the models to reproduce the correct thermodynamics of the biomolecules of interest.

The dynamic interplay of protein molecules constitutes the base of biological function. Protein structures as obtained by X-ray crystallography or, more recently, as generated by generative AI models such as AlphaFold 1, provide us with a static picture. However, many biologically relevant proteins are at least partially disordered or change conformation in response to the environment (presence of other molecules, temperature, ion concentration, etc.) and protein structures alone provide a very limited view of the associated biological processes. The modeling of the dynamics at the molecular level of large protein systems over biologically relevant timescales could shed light on functional mechanisms, advance our understanding of diseases, and aid in designing novel therapeutics. The characterization of protein dynamics is at present limited on the one hand by the low time and spacial resolution of the experiments, which provide only a partial view of the molecular processes, and on the other hand by the heavy computational demand of atomistic MD simulations in explicit water, which in practice can not realistically be used to characterize processes longer than ms. In order to bridge the gap between simulation and experiment for a complete characterization of biomolecular dynamics and function, the development of computational models at a resolution coarser than atomistic is needed.

For a coarse-grained (CG) model to be predictive, it needs to be able to reproduce experimental measurements and/or the results of more accurate (first-principle) models. In particular, the matching of the thermodynamic properties of a CG and finer resolution model requires the matching of the stationary probability density of the system as a function of suitable coordinates 2. This problem has been the subject of theoretical studies in the past decade, and it has been reformulated as a variational principle. Previous work has exploited this variational formulation for the optimization of parameters of CG effective energy functions with fixed functional form, with limited success.

In the past few years, we have proposed CGNet 3, which uses a deep neural network instead of a fixed functional form to represent the CG energy. We have shown that the increased expressivity provided by the network allows us to reproduce well the thermodynamics and the folding/unfolding behavior of model proteins as obtained by extensive atomistic simulations 3. We have successfully applied this approach with different network architectures (such as CG-Schnet 4) and CG resolution 5, as illustrated in Figure 1 for the small protein Chignolin.

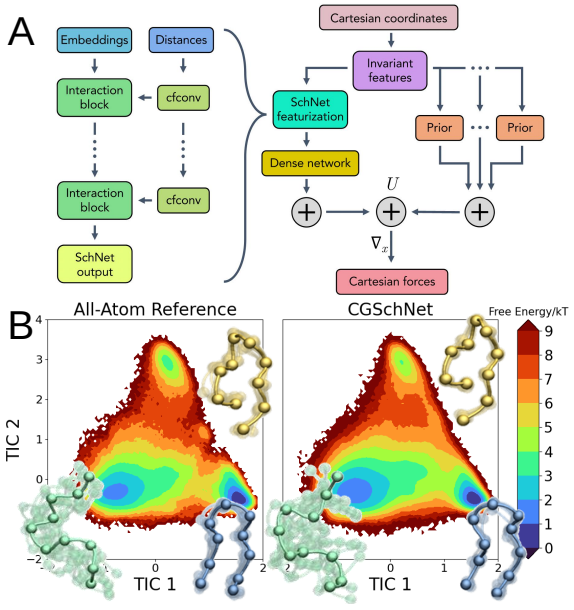

***Figure 1:** Machine-learning coarse-grained molecular dynamics. (A) Schematic of the CG-SchNet neural network architecture. (B) Application results on Chignolin peptide folding: The free energy surface in the slow TICA coordinates is reproduced by the machine-learned coarse-grained model.*

The main limitation of our previous work has been that the models were system-dependent and non-transferable in chemical space. That means that the reference data used for the training of the neural net-

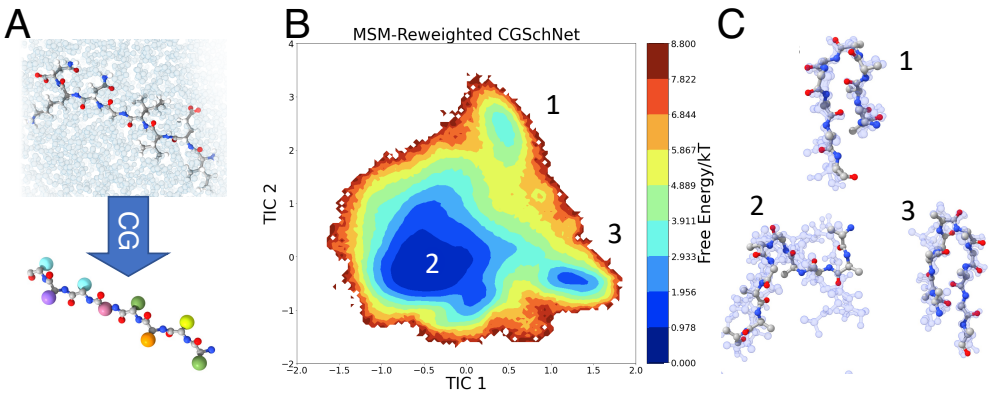

Figure 2: *Machine-learning a transferable protein force field. (A) Schematic of the CG mapping used: an atomistic structure of an 8-residue peptide in explicit solvent is mapped into a CG structure with no solvent and the side-chain represented by only the C_β atom. (B) The free energy landscape of the small protein Chignolin as obtained with a transferable machine-learned CG model trained on a set of octapeptides. (C) Representative structures from the misfolded (1), unfolded (2), and folded (3) metastable states are shown.*

work contains extensive atomistic simulation for the protein of interest. While it has been very important as a proof of principle, such models are not applicable to predict the dynamical behavior of unknown protein systems. In this project, we work on the design of a *transferable*, "universal" CG force field. Such a model could be used to simulate protein systems inaccessible to more conventional simulation (such as atomistic MD) and be combined with experimental methods for the complete characterization of the dynamics and thermodynamics of biologically relevant protein systems. We have very exciting preliminary results encouraging us to pursue this line of work: We have recently been able to reproduce the folding/misfolding/unfolding behavior of the model protein Chignolin with a model trained on a diverse and large (1000) set of atomistic simulations of short (8 residues) peptide sequences. As the sequence of Chignolin is longer (10 residue) and has less than 5% identity with the reference peptides, this result supports the feasibility of a transferable CG force field.

We are pursuing this project on multiple fronts, by building a much larger dataset of atomistic simulations for the network training, leveraging recent advances in neural network architectures (such as transformer models with attention mechanisms), and including physics-based correction terms to CG energy associated with the neural network. Each of these tasks requires significant computational resources and is only possible on high-performance computing clusters. In particular, the training of large neural network models requires the high-capacity GPUs available in the HLRN.

We believe that this approach will produce transferable CG force fields able to simulate the dynamics of multiple large proteins, with practical biomedical applications.

WWW

https://www.mi.fu-berlin.de/en/sfb1114/reasearch/projects/a04

More Information

[1] Jumper, J., Evans, R., Pritzel, A. *et al. Nature* **596**, 583–589 (2021). doi:10.1038/s41586-021-03819-2

[2] S. Izvekov and G. A. Voth. *J. Phys. Chem. B* **109**(7), 2469–2473 (2005). doi:10.1063/1.3382344

[3] J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé, and C. Clementi. *ACS Cent. Sci.* **5**(5), 755–767 (2019). doi:10.1021/acscentsci.8b00913

[4] B. E. Husic, N. E. Charron, D. Lemm, J. Wang, A. Pérez, A. Krämer, Y. Chen, S. Olsson, G. De Fabritiis, F. Noé, and C. Clementi. *J. Chem. Phys.*, **153**, 194101, (2020). doi:10.1063/5.0026133

[5] Y. Chen, A. Krämer, N. E. Charron, B. E. Husic, C. Clementi, and F. Noé. *J. Chem. Phys.*,**155**, 084101, (2021). doi:10.1063/5.0059915

Project Partners

Christof Schütte, ZIB; Frank Noé, Microsoft Research Berlin

Funding

DFG SFB 1114 Project A04

DFG Subject Area

201-02 Biophysics