The Confluence of Machine Learning and Multiscale Simulations

Harsh Bhatia[1], Fikret Aydin[2], Timothy S. Carpenter[2], Felice C. Lightstone[2], Peer-Timo Bremer[1], Helgi I. Ingólfsson[2], Dwight V. Nissley[3,*], Frederick H. Streitz[2,*]

[1]Computing Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550.
[2]Physical and Life Sciences (PLS) Directorate, Lawrence Livermore National Laboratory, Livermore, CA 94550.
[3]RAS Initiative, The Cancer Research Technology Program, Frederick National Laboratory, Frederick, MD 21701.

[*]Corresponding authors: Dwight V. Nissley, nissleyd@mail.nih.gov, and Frederick H. Streitz, streitz1@llnl.gov

Abstract

Multiscale modeling has a long history of use in structural biology, as computational biologists strive to overcome the time- and length-scale limits of atomistic molecular dynamics. Contemporary machine learning techniques, such as deep learning, have promoted advances in virtually every field of science and engineering and are revitalizing the traditional notions of multiscale modeling. Deep learning has found success in various approaches for distilling information from fine scale models, such as building surrogate models and guiding the development of coarse-grained potentials. However, perhaps its most powerful use in multiscale modeling is in defining latent spaces that enable efficient exploration of conformational space. This confluence of machine learning and multiscale simulation with modern high-performance computing promises a new era of discovery and innovation in structural biology.

Keywords. Deep learning, machine learning, artificial intelligence, multiscale simulations, sampling, neural networks, backmapping, coarse-grained, all-atomistic, force fields, proteins

Highlights

- Modern machine learning techniques have revitalized the traditional notions of multi scale simulations through leveraging massive amounts of data.
- Novel deep learning based methods have been developed to build coarse-grained force fields using finer-resolution simulations, as well as translating coarse-grained structures to atomistic ones.
- The use of machine learning to define relevant latent spaces that can automatically steer simulation ensembles on modern supercomputers offers new approaches for efficient exploration of conformational spaces.

1 Introduction

The time- and length-scales accessible to any given type of modeling and simulation technique are limited. Despite consistent advances of modern computing technologies, the need to utilize different simulation models persists, each with different levels of resolution and fidelity as well as varying computational requirements. Multiscale simulations are key to circumventing these limitations, as they facilitate combining information and/or models that capture different spatial or temporal scales. Multiscale frameworks address the contention between access to long- and large-scale dynamics and the computational viability of high-fidelity models. Indeed, multiscale techniques now form the backbone of scientific enquiries in structural biology(*1-12*) and almost all other areas of science and engineering(*1, 4, 5, 8-10, 12, 13*).

Multiscale approaches in the field of structural biology encompass a wide range of topics, and the study of complex membrane-protein systems is an important area of investigation and has been used for developing multiscale methods. Developing methods for distilling information from one scale, *e.g.*, all atomistic (AA) resolution to coarse-grained (CG) models or vice versa, are ubiquitous(*14, 15*). Accelerated molecular dynamics (MD) and enhanced sampling methods(*16*) are also crucial for computational modeling and simulations of complex biological systems. In this perspective, we focus on a rapidly evolving class of techniques for facilitating multiscale simulations for structural biology — those that utilize machine learning (ML).

1.1 Machine Learning for Multiscale Simulations

The past decade has seen ML technologies, in particular, deep learning (DL)(*17*), creating capability with far-reaching implications in structural biology. DL models are considered universal function approximators(*18, 19*), *i.e.*, they can approximate any complex but continuous mapping between inputs and outputs through an appropriately designed neural networks (NNs). This property obviates the need to define such mappings *a priori,* instead learning necessary function approximations through vast amounts of data. The tremendous growth in computing — consequential for simulating new data and training DL models — as well as advances in modern, higher-throughput instruments for data capturing (such as X-ray, cryo-EM and NMR) is enabling DL to play an integral role in contemporary biological applications.

DL techniques are influencing multiscale modeling and simulations in numerous ways(*20, 21*). For example, DL systems have shown great success as surrogate models(*22*) as well as in generating spatial structures from sequences of amino acids(*23, 24*) and highly accurate CG force fields for specific biological systems(*25, 26*). DL is also being used in novel ways for analyzing complex data, *e.g.*, capturing membrane lipid fingerprints at different scales(*27, 28*). ML-based techniques (including DL) are also playing key roles for steering large ensemble simulations*(11, 29, 30).*

An important and noteworthy application of DL in structural biology is the technology to accurately predict low-energy protein structures from linear sequences of amino acids. In particular, AlphaFold(*23*) has outperformed the traditional methods of predicting protein structures(*24, 31, 32*). Despite the impressive success and potential of AlphaFold(*33*), some challenges remain(*34*), such as predicting multi-protein components, metal ions, cofactors, and other ligands. To overcome these challenges, there are various efforts underway to capture protein interactions, such AlphaFold Multimer(*35*), RoseTTAFold(*36*), and ESMFold(*37*). Although such methods facilitate working across "scales" (*i.e.*, primary to tertiary structures), they are not considered multiscale techniques it the usual sense. As such, although they offer substantial promise for structural biology, they will not be discussed further in this review.

Traditional multiscale approaches have been classified as serial or parallel(*38*). Serial or sequential multiscale methods resolve or collapse degrees of freedom across scales *a priori* and utilize information from the finer scale to parameterize coarser scales and/or sampling at coarser scales to instantiate the finer scale. Parallel or concurrent multiscale methods, on the other hand, are coupled and perform cross-scale information exchange inside the running multiscale simulations, where specific region or molecules of interest are often represented at a finer scale and coupled, through specific annealing regions or using cross-scale parameters, to a coarser scale used for the bulk environment. Recent works on ML-driven ensemble-based, coupled multiscale simulations(*11, 30*) leverage the simplicity of serial multiscale methods while coupling in parallel the coarser macro model to continue improvements from concurrently running finer-scale simulations.

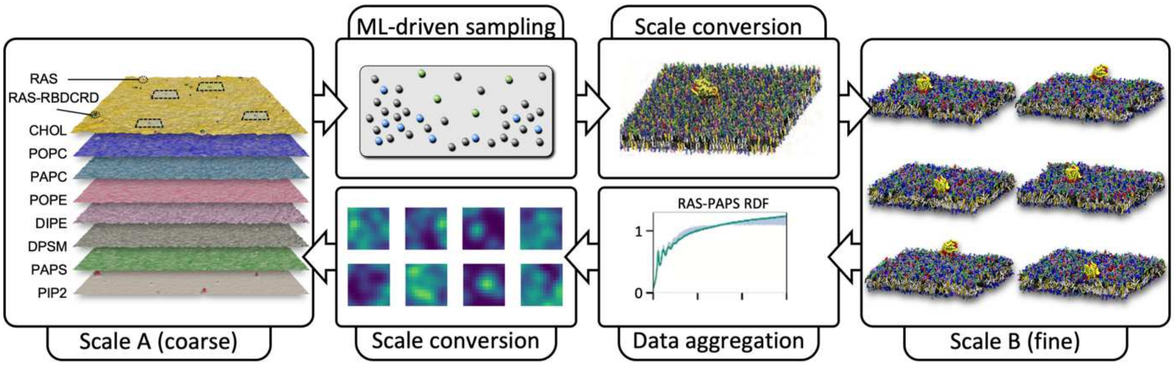

Figure 1. A contemporary pipeline for automated, ML-driven multiscale simulation in biology (adapted from the work of Ingólfsson et al.(30)). Coupling two scales encompass several operations, each of which may utilize traditional approaches or leverage ML. In this review, we discuss the existing and potential use of ML in these operations.

In this paper, we focus on two broad applications of ML for facilitating multiscale simulations in structural biology. The first is in the context of scale bridging. Several ML techniques have been proposed to transform data from one scale to another, *e.g.*, coarse-graining of AA configurations(*39-41*), as well as backmapping approaches, *e.g.*, from CG to AA(*42-44*). The second class of techniques focus on sampling and control of simulations using ML, *e.g.*, to identify when and where to promote configurations to finer scales(*23, 24*) or to stop simulations that explore uninteresting regions of phase space(*45*). Both class of techniques are key to enabling large multiscale simulations, especially when leveraging modern computing resources.

1.2 ML-driven Automation of Multiscale Simulations — Steps towards Exascale

The unprecedented scale of modern computing resources offers exciting opportunities for scientific applications; accompanied are many challenges in making efficient use of these software and hardware resources. The high-performance computing (HPC) community is moving away from large and monolithic codes to sophisticated workflows(*46-49*) that create massive simulation ensembles. Traditional metrics for scaling, such as strong and weak scaling, are getting replaced by the need for simultaneous utilization of heterogenous resources, tailored to the needs of multiscale. ML-based techniques have demonstrated immense value in facilitating such a vision through automated/semi-automated frameworks that rely on ML to generate targeted or exploratory ensembles of multiscale simulations(*11, 30, 50-54*). Multiscale frameworks powered by ML are paving the way for a new revolutionary approach for studying scientific phenomena. Such methods are likely going to be the centerpiece of computational sciences in the Exascale era.

2 Key Research Directions

Among the numerous ways ML is revolutionizing the field of multiscale simulations, we discuss two broad classes, namely, for supporting the coupling of resolution across scales and for facilitating ensemble multiscale simulations.

2.1 ML for Resolution Coupling

Transformation of data representations (from coarser to finer resolutions, or its inverse) has traditionally been performed using data-driven statistical models, guided through expert knowledge. Innovations in ML and, especially, DL, have revamped such efforts by leveraging vast amounts of data. Compared to the traditional techniques, which focus on unidirectional coupling, bidirectional coupling allows information flow from both ends, *e.g.*, from fine to coarse resolution and back, ensuring a greater degree of consistency in the multiscale simulation. Recent works have furthered the reach of DL technology through automated bidirectional coupling of resolutions. Section 2.2.1 describes a novel capability that exploits ML to enable such a coupling on the fly, assimilating information at fine and coarse resolutions, each to improve the fidelity of the other.

2.1.1 ML for Learning Force Fields

Time- and length-scales beyond the limits of AA simulations can be achieved by using CG models(*55-58*). Traditionally, CG models are derived by either analyzing finer-resolution simulations (bottom-up)(*59*) and/or by tuning interaction potentials to capture a suite of known properties (top-down). Being data driven in nature, such tasks are getting relegated to ML due to its capability to learn CG force fields.

Recent years have seen successful application of various DL methods to build CG force fields from AA training data, and many different types of DL technologies have been utilized(*40*). For example, NNs to build many-body CG potentials(*25*), to construct effective CG Hamiltonian based on high-dimensional free energy surface(*60*), NNs(*26*) or graph convolutional NNs(*41*) to generate coarse-grained free energy functions via force-matching scheme, NN-assisted particle swarm optimization method for CG force fields(*61*), and generative adversarial networks (GANs) to optimize CG force fields(*62*). The advantages of ML-driven force fields over classical methods include incorporation of many-body interactions and nonlinearities as well as the ability to optimize the model using training data at different thermodynamic states. However, ML force fields are typically constructed for specific systems under defined circumstances, and they are generally not transferrable. ML-based force fields have been applied successfully to small systems, where the training set can encompass a comfortable fraction of the entire phase space. Nevertheless, unexpected challenges may arise when extending these models to larger and more complex systems, as the training set becomes a smaller and smaller fraction of what is possible. Further developments are necessary to make ML force fields more generalizable and use them for larger and more complex molecular systems.

2.1.2 ML for Parameterization of MD Potentials

As discussed above, NNs trained on high-resolution simulation data in a bottom-up fashion have been highly successful in recent works. On the other hand, top-down approaches, which learn MD potentials directly from experimental data have been less studied, largely due to computational difficulties, such as numerical issues.

Such issues are starting to be addressed through automatic differentiation (AD) techniques(*63*). Specifically, a recently developed method, Differentiable Trajectory Reweighting (DiffTRe)(*64*), can train NNs on experimental data with lower computational effort. Leveraging the power of AD in combination with existing MD reweighting techniques, DiffTRe eliminates the need to differentiate through the simulation and provides end-to-end gradient computation. DiffTRe can also be used as a bottom-up model parameterization scheme without any additional changes by using target observables from a high-fidelity simulation instead of an experiment.

DMFF (Deep Modelling Force Field, or Differentiable Molecular Force Field)(*65*) is another platform that utilizes AD to comprehensively implement both conventional molecular force fields and advanced multipolar polarizable models. DMFF provides differentiable estimators for energies, forces and thermal dynamic quantities, which enable the definition of corresponding objective functions, making both bottom-up and top-down optimization workflows possible.

2.1.3 ML for Learning QM/MM Potentials

The Quantum Mechanics/Molecular Mechanics (QM/MM) approach(*66*) is a widely used method for overcoming the computational bottlenecks associated with quantum calculations for describing molecular interactions and chemical reactions. It combines a QM description of the region of interest with a realistic modelling of the surrounding environment, typically using either mechanical or electrostatic embedding schemes. Despite the reduced computational costs compared to full *ab initio* simulations, these simulations are not sufficient for resolving the phenomena occurring at longer length- and time-scales. To further reduce the computational cost, semi-empirical methods are used to describe the QM zone, though at the cost of accuracy of the simulations.

Recent works have seen significant impact of ML techniques in the QM/MM paradigm(*67*). Broadly, by replacing the simulated QM and/or MM potentials with those learned using ML can offer improved accuracy while reducing the computational cost of *ab initio* QM/MM simulations(*68-71*). For example, it has also been shown(*71*) that DL models can be trained to accurately reproduce both the QM and MM forces, as well as the differences between the *ab* *initio*(*72*) and the sub-empirical forces(*68*). Such techniques enable the calculation of accurate free-energy barriers for various solution-phase reactions in these systems(*71, 73*). DL-based potentials have also been utilized to reweight the free-energy profiles of reactions (*e.g.*, the proton transfer reaction of glycine in water) from a QM/MM approach (*69*). An adaptive QM/MM-NN framework was also proposed to perform direct MD simulations on the potential energy surface predicted by DL in order to approximate *ab initio* QM/MM MD(*72*).

Although the use of DL in QM/MM MD simulations is impactful and shows much promise, such techniques are still computationally expensive, in that the computational cost is now transferred from running the simulations to training DL models and searching for suitable ones. Furthermore, DL-based potentials, despite their documented success, still require more rigorous validations to ensure reliability, which may require developing expertise across the domains.

2.1.3 ML for Backmapping

Although some research questions can be appropriately addressed using a CG representation, many require access to AA resolution to obtain some of the observables of interest. "Backmapping" methods, which transform CG to AA representation, therefore play a vital role in studying complex biological systems. Traditional methods perform backmapping by guessing the positions of atoms (*e.g.*, based on random mapping(*74*), or geometry-(*75*) or fragment-based(*14*) positioning) followed by energy minimization to produce realistic atomic configurations.

Advances in ML now enable the development of more general methods that can efficiently and accurately map different scales for various biomolecules, such as lipids, polymers, and proteins. These ML-based methods can translate low resolution CG structures to finer, AA structures without needing system-specific information, such as molecular structures and force fields. In this context, non-DL methods, including k-nearest neighbors, Gaussian process regression, and random forests, as well as NNs have been utilized for backmapping(*42*). More recently, variational autoencoders (VAEs)(*39*), GANs(*43, 76*) and multilayer perceptrons (MLPs)(*44*) have also been successfully applied to backmap CG models to AA models (see Figure 2 for an example). These methods are capable of learning parameters associated with the system from information such as pairwise interatomic distance matrices or distance/orientation vectors, which makes them very generalizable and straightforward to implement, and they can be easily applied at different levels of coarse-graining. However, the current applications are limited to small system sizes, as these are still in early stages of development and have a high memory requirement of the underlying NNs. Another area of improvement is to marry these purely data-driven methods with existing domain knowledge, such as known interactions and restraints. As the community moves towards more sophisticated backmapping methods, these are likely to play a key role in facilitating complex multiscale simulations.

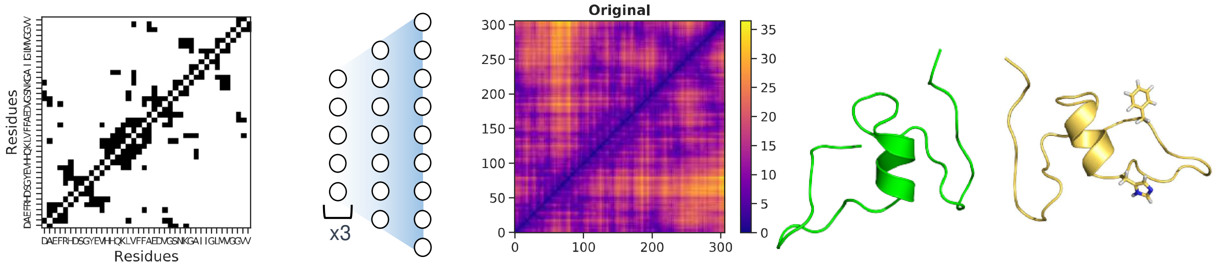
Residues
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
Residues
DAEFRHDSGYEVHHQKLVFFAEDVGSNKGAIIGLMVGGVV
x3
Original
300
250
200
150
100
50
0
0
100
200
300
35
30
25
20
15
10
5
0

Figure 2. The recent work by Duong et al.(44) uses a multilayer percepteron (MLP) — a neural network with only fully-connected layers — to learn how to predict a pairwise distance matrix of atoms, given a contact matrix (a binary matrix representing spatial proximity within some distance threshold) of coarsened proteins structure. The predicted distance matrix is then converted into 3D spatial coordinates using the multidimensional scaling (MDS) approach and resulting structure equilibrated. This approach marks an important step towards developing DL methods for backmapping; howeser, this method is limited to small molecules, in part, due to the large computational requirement of MLPs. Figure reproduced from the original article(44).

2.2 ML for Ensemble Multiscale Simulations

The fundamental idea behind ensemble multiscale is to explore by sampling some reduced representation that faithfully captures the substantially larger and more-complex phase space of the system to be simulated*(77-79)*. To this end, such ideas seek to find a few generalized degrees of freedom that can separate metastable states, drive the sampling of the phase space, and facilitate achieving longer timescales than otherwise possible(*80*).

2.2.1 ML-driven Sampling Frameworks

DL methods are now replacing the use of the traditional "collective variables" with data-driven *"latent spaces"*. For example, Gaussian process models(*81*), NNs for Bayesian learning(*82*), autoencoders(*83*), variational models(*84*), reinforcement learning(*85*), and MLPs to learn stochastic neighbor embedding(*86*) have all been successfully utilized for enhanced sampling. DL has also been utilized to sample collective variables through time-lagged autoencoders(*87*) or to generate Markov state models(*88, 89*).

Recently, AI-driven multiscale simulations have been used to study the mechanisms of SARS-CoV-2 spike dynamics(*29, 90*). Built upon a ML-driven workflow(*91*), these approaches use ML to learn which regions in the conformational phase space are sampled sufficiently and initiate new simulations in the undersampled regions. This AI-driven sampling approach facilitated a wider exploration of the conformational space of the relevant proteins and, along with several associated computational improvements, exhibited strong scaling on the *Summit* supercomputer(*92*). While effective for sampling known phase spaces, a current limitation of these techniques is that the sampling is performed offline, prior to running the simulations. Unable to adjust as the simulations are being executed, these methods fall short in offering a fully automated framework for multiscale simulations. Techniques that take additional steps towards automation will be discussed in the next section.

The recent work on Boltzmann generators(*93*) represents an important step toward learning the equivalent of "reaction coordinates" through DL. For sampling the equilibrium states of many-body systems, the key idea is to learn a mapping between the energy function of the chosen many-body system and a simple distribution, such as a Gaussian, that can be sampled easily. The samples drawn from the Gaussian are mapped to the configuration space using DL and appropriately weighted to provide statistical insights into the system. Perhaps the most important benefit of this approach is the ability to interpolate in the learnt latent space. As shown by the authors(*93*), linearly interpolated samples are likely to correspond to realistic transition pathways between metastable states. Although powerful in principle, the current scale handled by this method is modest (less than a thousand atoms); more work is needed to expand the scope to larger and more-realistic systems.

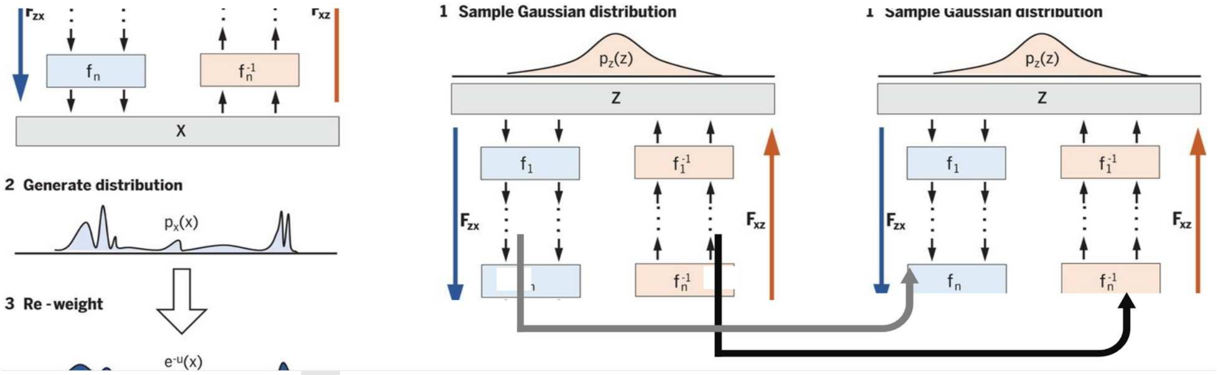

Figure 3. Boltzmann generators(93) are a new ML-driven approach to drive sampling of transition pathways between metastable states. Using a DL solution, this approach learns a mapping from a simple Gaussian distribution to the distribution of energy function and performs sampling in the former rather than the latter. The learned latent space can also be interpolated to generate realistic transitions. Image adapted from the original publication(93).

2.2.2 ML-driven Automated Simulation Frameworks

To date, the most sophisticated combination of DL and multiscale simulations is MuMMI (Multiscale Machine Learned Modeling Infrastructure)(*11, 30*), which expands upon the philosophy of using DL for enhanced sampling by utilizing it in a dynamic manner(*51, 53*) and combines it with a scalable workflow technology(*50, 52*) to deliver an automated multiscale simulation framework, which has been demonstrated at unprecedented scales, creating ensembles of over 100,000 simulations with the simultaneous use of 36,000 GPUs.

The key innovation in MuMMI, specifically when compared to other enhanced sampling methods and other multiscale frameworks, is that ML sits at the core of MuMMI and drives the multiscale simulation *dynamically* (see Figure 4). This ML-driven framework offers two key advantages. First, the ML-based sampling is dynamic, *i.e.*, the ML model projects the data onto a latent space, investigates it, and samples relevant configurations in real time as new data is being generated. The dynamic sampling lifts the paradigm of multiscale simulations out of the need to work with known or predefined phase spaces, and thus facilitates the exploration of new hypotheses and the effects of controlled steering. Second, the dynamic and integrated use of ML enables the computation of appropriate weights for these samples which are then used for *in situ* computation and aggregation of ensemble statistics. This *in situ* mechanism allows capture of the insights from the finer scale into the coarser scale immediately and without delay, steering the simulations towards improved parameterization and “more interesting” regions of the phase space. The MuMMI framework is agnostic to the specific type of DL models, and two technologies have been demonstrated: a VAE(*51*) and a deep metric learning approach(*53*).

The open-source framework MuMMI has been utilized to explore the RAS/RAF/MAPK signaling pathway, whose dysregulation is consequential for cancer(*11, 30*). Originally developed to utilize two scales, MuMMI was later extended to enable simultaneous pairwise coupling of three resolutions: a continuum model simulation(*94*), CG MD simulations(*55, 57, 95*), and AA MD simulations(*96*). The continuum model allowed for realistic simulation of protein constellations and diverse local lipid environments. ML-guided selections from the continuum model were then simulated at the CG scale, simultaneously refining the continuum scale while resolving RAS and RAS-RAF lipid fingerprints and different lipid dependent structural configurations. Further sampling at the AA scale captured the secondary structure adaptation of the proteins upon interaction with the membranes, updating the CG parameters and resolving the membrane adaptation of RAS and RAS-RAF.

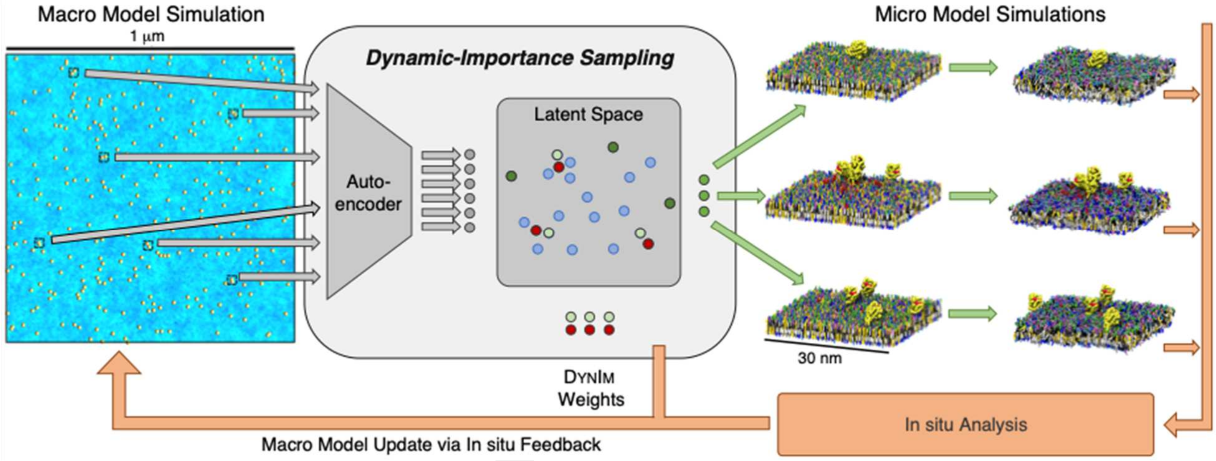

Figure 4. Dynamic-Importance Sampling framework(51) integrated as part of MuMMI(11) steers the multiscale simulation using a ML model that investigates the data from the coarse scale (macro model) simulation. Configurations that are likely to produce new insights (through a diversity sampling in a machine learned latent space) are identified and selected to be promoted to the finer scale (micro model). Through a dynamic reweighting of ths statistics from the finer scale, the parameterization of the coarse scale model are improved, creating an in situ feedback process. MuMMI framework is arbitrarily scalable(50, 52) and agnostic to the specific type of DL model(53) as well as other types of modeling techniques(50). Expanding the generalizability of MuMMI to other biological systems as well as other application domains beyond biology is feasible but remains to be a task for the future. Image used with permission from Bhatia et al.(51).

3 The Way Forward

There exist several more-specific areas where ML can play a key role in structural biology. For example, advances in experimental technologies also provide resourceful data, but it is usually disparate from simulated data. Multimodal ML techniques could prove useful in joint processing of such data as well as by incorporating additional constraints from experimental data into ML models. The use of ML in choosing reaction coordinates is also in early stages and is likely to see rapid research and development. Overall, current trends indicate that ML techniques play an even more central role in the development and execution of multiscale simulation techniques, particularly as current limitations are alleviated.

A particularly impactful use of ML that can potentially revolutionize the field of multiscale simulations is concurrent transfer of information across scales through *in situ* feedback. The ability to refine an approximate parameterization by providing feedback from running simulations lifts the requirement that simulation models be parameterized in final form *a priori*. Feedback can enable a truly explorational and hypotheses-driven approach by offering a means to learn appropriate parameterizations at run time, potentially leading more easily to insights not previously imagined. However, these methods need to be demonstrated on a broader class of systems and parameters, and active research into such ideas will prove fruitful in the future.

Broadly, there are two key limitations of current ML methods in the context of multiscale simulations — computational cost and generalizability across systems. Both will be mitigated to a great extent by natural progressions in the maturity of a new technology, making the future bright for the confluence of ML and multiscale methods. The community is making significant advances in addressing scalability concerns by designing new algorithms and developing efficient software frameworks. Additionally, specialized hardware for DL is substantially reducing the computational cost of training and will soon become easy to access at affordable costs. Generalizability, on the other hand, poses a bigger challenge owing largely to the highly complex nature of physical, chemical, and biological interactions inherent in scientific systems. Current DL models in sciences are trained for specific tasks and on a small number of systems. Going forward, the community could instead draw inspiration from Foundation Models(*97*), which are trained on large amount of broad data and can be adapted for a wide range of tasks.

The age of big data has opened tremendous opportunities for discovery and innovation in all fields, scientific or otherwise. Data-driven techniques such as ML and (especially) DL are accelerating the development and validation of new hypotheses — a process which is at the heart of the scientific method. The advent of Exascale Computing, where compute is ubiquitous and data access convenient, is enabling simulation ensembles of unprecedented scale– hundreds of thousands of simulations and hundreds of TB storage. In the future, the modeling and simulation community will find itself increasingly reliant on large-scale, sophisticated workflow technologies to undertake and manage these large computational studies. As a result, scientific workflows that can be steered automatically and efficiently using ML will become integral to large scientific studies.

Acknowledgements

The authors acknowledge many useful discussions with the ADMIRRAL team at the Frederick National Laboratory for Cancer Research. This work has been supported in part by the Joint Design of Advanced Computing Solutions for Cancer (JDACS4C) program established by the U.S. Department of Energy (DOE) and the National Cancer Institute (NCI) of the National Institutes of Health (NIH). This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Release LLNL-JRNL-839645.

Disclosures

The authors declare no conflict of interest.

List of Annotated References (should be within the last 2 years)

Papers of outstanding interest (**)

1. *(64)* (2021, ML for Top-down Parameterization): *This work leverages the power of automatic differentiation to significantly reduce the computational cost of training DL models for learning MD potentials from experimental data.*

2. (*29*) (2019, Gordon Bell Award winner): *This work utilized AI with a large-scale workflow and, through demonstration on the Summit supercomputer, won the 2019 Special Prize for High Performance Computing-Based COVID-19 Research.*

3. *(93)* (2021, Boltzmann Generator paper): *This works demonstrates the viability of DL techniques to learn energy functionals of complex biological structures, paving a way to explore these landscapes in much simpler forms.*

4. (*11*) (2021, MuMMI paper): *MuMMI demonstrates an automated, ML-driven workflow to steer massive ensemble-multiscale simulations, together with in situ feedback. This work captures new insights into lipid-protein interactions crucial for cancer biology.*

5. (*51*) (2021, MuMMI ML paper): *This work presents the dynamic-importance sampling framework that places DL at the core of workflow technologies, facilitating massive multiscale simulations to create large ensembles sampled for exploration.*

Papers of special interest (*)

6. (*68*) (2021, ML for QM/MM): *This work demonstrates that DL models can act as high-quality and inexpensive surrogates for expensive QM and semi-empirical methods, and lays out a viable approach for promising (QM) ML/MM MD approaches.*

7. (*44*) (2021, Backmapping using DL): *This work demonstrates the viability of simple MLP networks to backmap protein structures from coarsened to atomistic levels.*

8. (*52*) (2019, MuMMI Supercomputing winner): *This award-winning paper at Supercomputing Conference 2019 presents innovations in workflow technology to facilitate large ensemble simulations using ML.*

9. (*17*) (2015, Deep Learning survey): *Despite being a few years old, this review paper by some of the pioneers of deep learning remains an exceptional resource equally for DL practitioners as well as for application scientists looking to foray into DL.*

References

1. R. E. Miller, E. B. Tadmor, A unified framework and performance benchmark of fourteen multiscale atomistic/continuum coupling methods. *Modelling and Simulation in Materials Science and Engineering* **17**, 053001 (2009).
2. G. S. Ayton, G. A. Voth, Multiscale simulation of protein mediated membrane remodeling. *Seminars in Cell & Developmental Biology* **21**, 357-362 (2010).
3. V. Tozzini, Multiscale Modeling of Proteins. *Accounts of Chemical Research* **43**, 220-230 (2010).
4. B. Chopard, J.-L. Falcone, A. G. Hoekstra, J. Borgdorff, in *Unconventional Computation,* C. S. Calude, J. Kari, I. Petre, G. Rozenberg, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011), pp. 2-8.
5. A. Hoekstra, B. Chopard, P. Coveney, Multiscale modelling and simulation: a position paper. *Phil. Trans. R. Soc. A* **372**, 20130377 (2014).
6. V. V. Krzhizhanovskaya, D. Groen, B. Bozak, A. G. Hoekstra, Multiscale Modelling and Simulation Workshop:12 Years of Inspiration. *Procedia Computer Science* **51**, 1082-1087 (2015).
7. G. Enkavi, M. Javanainen, W. Kulig, T. Róg, I. Vattulainen, Multiscale Simulations of Biological Membranes: The Challenge To Understand Biological Phenomena in a Living Substance. *Chemical Reviews* **119**, 5607-5774 (2019).
8. R. G. Huber *et al.*, in *Lipid-Protein Interactions: Methods and Protocols,* J. H. Kleinschmidt, Ed. (Springer New York, New York, NY, 2019), pp. 1-30.
9. E. van der Giessen *et al.*, Roadmap on multiscale materials modeling. *Modelling and Simulation in Materials Science and Engineering* **28**, 043001 (2020).
10. D. Bishara, Y. Xie, W. K. Liu, S. Li, A State-of-the-Art Review on Machine Learning-Based Multiscale Modeling, Simulation, Homogenization and Design of Materials. *Archives of Computational Methods in Engineering*, (2022).
11. H. I. Ingólfsson *et al.*, Machine Learning-driven Multiscale Modeling Reveals Lipid-Dependent Dynamics of RAS Signaling Protein. *Proc. Natl. Acad. Sci. USA* **119**, (2022).
12. A. Brandt, in *Multiscale and Multiresolution Methods,* T. J. Barth, T. Chan, R. Haimes, Eds. (Springer Berlin Heidelberg, Berlin, Heidelberg, 2002), pp. 3-95.
13. V. V. Krzhizhanovskaya, D. Groen, B. Bozak, A. G. Hoekstra, Multiscale Modelling and Simulation Workshop: 12 Years of Inspiration. *Procedia Computer Science* **51**, 1082-1087 (2015).
14. C. Peter, K. Kremer, Multiscale simulation of soft matter systems – from the atomistic to the coarse-grained level and back. *Soft Matter* **5**, 4357-4366 (2009).
15. P. J. Stansfeld, M. S. P. Sansom, From coarse grained to atomistic: a serial multiscale approach to membrane protein simulations. *Journal of Chemical Theory and Computation* **7**, 1157-1166 (2011).
16. X. Gong, Y. Zhang, J. Chen, Advanced Sampling Methods for Multiscale Simulation of Disordered Proteins and Dynamic Interactions. *Biomolecules* **11**, (2021).
17. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**, 436-444 (2015).
18. G. Cybenko, Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems* **2**, 303-314 (1989).
19. K. Hornik, Approximation capabilities of multilayer feedforward networks. *Neural Networks* **4**, 251-257 (1991).
20. F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, Machine Learning for Molecular Simulation. *Annual Review of Physical Chemistry* **71**, 361-390 (2020).
21. G. C. Y. Peng *et al.*, Multiscale Modeling Meets Machine Learning: What Can We Learn? *Archives of Computational Methods in Engineering* **28**, 1017-1037 (2021).

22. J. C. S. Kadupitiya, F. Sun, G. Fox, V. Jadhao, Machine learning surrogates for molecular dynamics simulations of soft materials. *Journal of Computational Science* **42**, 101107 (2020).

23. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589 (2021).

24. M. Edich, D. C. Briggs, O. Kippes, Y. Gao, A. Thorn, The impact of AlphaFold2 on experimental structure solution. *Faraday Discussions*, (2022).

25. L. Zhang, J. Han, H. Wang, R. Car, W. E, DeePCG: Constructing coarse-grained models via deep neural networks. *The Journal of Chemical Physics* **149**, 034101 (2018).

26. J. Wang *et al.*, Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Central Science* **5**, 755-767 (2019).

27. F. Aydin *et al.*, in *arXiv*. (arXiv:2207.06630, 2022).

28. K. Georgouli *et al.*, in *arXiv*. (arXiv:2207.04333, 2022).

29. L. Casalino *et al.*, AI-driven multiscale simulations illuminate mechanisms of SARS-CoV-2 spike dynamics. *The International Journal of High Performance Computing Applications* **35**, 432-451 (2021).

30. H. I. Ingólfsson *et al.*, Machine Learning-driven Multiscale Modeling, bridging the scales with a next generation simulation infrastructure. *The Journal of Chemical Theory and Computation*, (2023). **Under review**.

31. A. Perrakis, T. K. Sixma, AI revolutions in biology. *EMBO reports* **22**, e54046 (2021).

32. N. Bouatta, P. Sorger, M. AlQuraishi, Protein structure prediction by AlphaFold2: are attention and symmetries all you need? *Acta Crystallogr D Struct Biol* **77**, 982-991 (2021).

33. E. Callaway, What's next for the AI proteinfolding revolution. *Nature* **604**, 234—238 (2022).

34. P. B. Moore, W. A. Hendrickson, R. Henderson, A. T. Brunger, The protein-folding problem: Not yet solved. *Science* **375**, 507-507 (2022).

35. R. Evans *et al.*, Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034 (2022).

36. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871-876 (2021).

37. Z. Lin *et al.*, Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.2007.2020.500902 (2022).

38. G. S. Ayton, W. G. Noid, G. A. Voth, Multiscale modeling of biomolecular systems: in serial and in parallel. *Current Opinion in Structural Biology* **17**, 192-198 (2007).

39. W. Wang, R. Gómez-Bombarelli, Coarse-graining auto-encoders for molecular dynamics. *npj Computational Materials* **5**, 125 (2019).

40. P. Gkeka *et al.*, Machine Learning Force Fields and Coarse-Grained Variables in Molecular Dynamics: Application to Materials and Biological Systems. *Journal of Chemical Theory and Computation* **16**, 4757-4775 (2020).

41. B. E. Husic *et al.*, Coarse graining molecular dynamics with graph neural networks. *The Journal of Chemical Physics* **153**, 194101 (2020).

42. Y. An, S. A. Deshmukh, Machine learning approach for accurate backmapping of coarse-grained models to all-atom models. *Chemical Communications* **56**, 9312–9315 (2020).

43. W. Li, C. Burkhart, P. Polińska, V. Harmandaris, M. Doxastakis, Backmapping coarse-grained macromolecules: An efficient and versatile machine learning approach. *The Journal of Chemical Physics* **153**, 041101 (2020).

44. V. T. Duong, E. M. Diessner, G. Grazioli, R. W. Martin, C. T. Butts, Neural Upscaling from Residue-Level Protein Structure Networks to Atomistic Structures. *Biomolecules* **11**, 1788 (2021).

45. Y. Nagai, M. Okumura, K. Kobayashi, M. Shiga, Self-learning hybrid Monte Carlo: A first-principles approach. *Physical Review B* **102**, 041124 (2020).

46. I. Altintas *et al.*, in *The 16th International Conference on Scientific and Statistical Database Management*. (2004), pp. 423-424.

47. E. Deelman *et al.*, Pegasus: a Workflow Management System for Science Automation. *Future Generation Computer Systems* **46**, 17-35 (2015).

48. D. H. Ahn *et al.*, in *IEEE/ACM Workflows in Support of Large-Scale Science (WORKS).* (IEEE, Dallas, Texas, 2018), pp. 10-19.

49. T. Ben-Nun, T. Gamblin, D. S. Hollman, H. Krishnan, C. J. Newburn, in *2020 IEEE/ACM International Workshop on Performance, Portability and Productivity in HPC (P3HPC).* (2020), pp. 57-69.

50. H. Bhatia *et al.*, in *The International Conference for High Performance Computing, Networking, Storage and Analysis.* (ACM, 2021), pp. 10.

51. H. Bhatia *et al.*, Machine Learning Based Dynamic-Importance Sampling for Adaptive Multiscale Simulations. *Nature Machine Intelligence* **3**, 401—409 (2021).

52. F. Di Natale *et al.*, in *The International Conference for High Performance Computing, Networking, Storage and Analysis.* (ACM, Denver, Colorado, 2019), pp. 57.

53. H. Bhatia *et al.*, A Biology-Informed Similarity Metric for Simulated Patches of Human Cell Membrane. *Machine Learning: Science and Technology* **3**, (2022).

54. J. Zhang, M. Chen, Unfolding Hidden Barriers by Active Enhanced Sampling. *Phys Rev Lett* **121**, 010601 (2018).

55. S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman, A. H. de Vries, The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *The Journal of Physical Chemistry B* **111**, 7812-7824 (2007).

56. L. Monticelli *et al.*, The MARTINI Coarse-Grained Force Field: Extension to Proteins. *Journal of Chemical Theory and Computation* **4**, 819-834 (2008).

57. R. Alessandri, F. Grünewald, S. J. Marrink, The Martini Model in Materials Science. *Advanced Materials* **33**, 2008635 (2021).

58. S. J. Marrink *et al.*, Two decades of Martini: Better beads, broader scope. *WIREs Computational Molecular Science* **n/a**, e1620 (2022).

59. J. Jin, A. J. Pak, A. E. P. Durumeric, T. D. Loose, G. A. Voth, Bottom-up Coarse-Graining: Principles and Perspectives. *Journal of Chemical Theory and Computation*, (2022).

60. T. Lemke, C. Peter, Neural Network Based Prediction of Conformational Free Energies - A New Route toward Coarse-Grained Simulation Models. *Journal of Chemical Theory and Computation* **13**, 6213-6221 (2017).

61. K. K. Bejagam, S. Singh, Y. An, S. A. Deshmukh, Machine-Learned Coarse-Grained Models. *The Journal of Physical Chemistry Letters* **9**, 4667-4672 (2018).

62. A. E. P. Durumeric, G. A. Voth, Adversarial-residual-coarse-graining: Applying machine learning theory to systematic molecular coarse-graining. *The Journal of Chemical Physics* **151**, 124110 (2019).

63. A. G. Baydin, B. A. Pearlmutter, A. A. Radul, J. M. Siskind, Automatic differentiation in machine learning: a survey. *The Journal of Machine Learning Research* **18**, 1-43 (2018).

64. S. Thaler, J. Zavadlav, Learning neural network potentials from experimental data via Differentiable Trajectory Reweighting. *Nat Commun* **12**, 6884 (2021).

65. X. Wang *et al.*, DMFF: An Open-Source Automatic Differentiable Platform for Molecular Force Field Development and Molecular Dynamics Simulation. *ChemRxiv*, (2022).

66. E. Brunk, U. Rothlisberger, Mixed Quantum Mechanical/Molecular Mechanical Molecular Dynamics Simulations of Biological Systems in Ground and Electronically Excited States. *Chem Rev* **115**, 6217-6263 (2015).

67. Y. J. Zhang, A. Khorshidi, G. Kastlunger, A. A. Peterson, The potential for machine learning in hybrid QM/MM calculations. *J Chem Phys* **148**, 241740 (2018).

68. L. Boselt, M. Thurlemann, S. Riniker, Machine Learning in QM/MM Molecular Dynamics Simulations of Condensed-Phase Systems. *J Chem Theory Comput* **17**, 2641-2658 (2021).

69. L. Shen, J. Wu, W. Yang, Multiscale Quantum Mechanics/Molecular Mechanics Simulations with Neural Networks. *J Chem Theory Comput* **12**, 4934-4946 (2016).

70. J. Zeng, T. J. Giese, S. Ekesan, D. M. York, Development of Range-Corrected Deep Learning Potentials for Fast, Accurate Quantum Mechanical/Molecular Mechanical Simulations of Chemical Reactions in Solution. *J Chem Theory Comput* **17**, 6993-7009 (2021).

71. X. Pan *et al.*, Machine-Learning-Assisted Free Energy Simulation of Solution-Phase and Enzyme Reactions. *J Chem Theory Comput* **17**, 5745-5758 (2021).

72. L. Shen, W. Yang, Molecular Dynamics Simulations with Quantum Mechanics/Molecular Mechanics and Adaptive Neural Networks. *J Chem Theory Comput* **14**, 1442-1455 (2018).

73. T. J. Giese, J. Zeng, S. Ekesan, D. M. York, Combined QM/MM, Machine Learning Path Integral Approach to Compute Free Energy Profiles and Kinetic Isotope Effects in RNA Cleavage Reactions. *J Chem Theory Comput* **18**, 4304-4317 (2022).

74. A. J. Rzepiela *et al.*, Reconstruction of atomistic details from coarse-grained structures. *Journal of Computational Chemistry* **31**, 1333-1343 (2010).

75. T. A. Wassenaar, K. Pluhackova, R. A. Böckmann, S. J. Marrink, D. P. Tieleman, Going Backward: A Flexible Geometric Approach to Reverse Transformation from Coarse Grained to Atomistic Models. *Journal of Chemical Theory and Computation* **10**, 676-690 (2014).

76. M. Stieffenhofer, M. Wand, T. Bereau, Adversarial reverse mapping of equilibrated condensed-phase molecular structures. *Machine Learning: Science and Technology* **1**, 045014 (2020).

77. G. M. Torrie, J. P. Valleau, Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics* **23**, 187-199 (1977).

78. A. Laio, M. Parrinello, Escaping free-energy minima. *Proceedings of the National Academy of Sciences* **99**, 12562-12566 (2002).

79. D. M. Zuckerman, L. T. Chong, Weighted Ensemble Simulation: Review of Methodology, Applications, and Software. *Annual Review of Biophysics* **46**, 43-57 (2017).

80. F. Noé, in *Machine Learning Meets Quantum Physics,* K. T. Schütt *et al.*, Eds. (Springer International Publishing, Cham, 2020), pp. 331-372.

81. L. Mones, N. Bernstein, G. Csányi, Exploration, Sampling, And Reconstruction of Free Energy Surfaces with Gaussian Process Regression. *Journal of Chemical Theory and Computation* **12**, 5100-5110 (2016).

82. M. Schöberl, N. Zabaras, P.-S. Koutsourelakis, Predictive collective variable discovery with deep Bayesian models. *The Journal of Chemical Physics* **150**, 024109 (2019).

83. W. Chen, A. R. Tan, A. L. Ferguson, Collective variable discovery and enhanced sampling using autoencoders: Innovations in network architecture and error function design. *The Journal of Chemical Physics* **149**, 072312 (2018).

84. O. Valsson, M. Parrinello, Variational Approach to Enhanced Sampling and Free Energy Calculations. *Physical Review Letters* **113**, 090601 (2014).

85. Z. Shamsi, K. J. Cheng, D. Shukla, Reinforcement Learning Based Adaptive Sampling: REAPing Rewards by Exploring Protein Conformational Landscapes. *The Journal of Physical Chemistry B* **122**, 8386-8395 (2018).

86. J. Rydzewski, O. Valsson, Multiscale Reweighted Stochastic Embedding: Deep Learning of Collective Variables for Enhanced Sampling. *The Journal of Physical Chemistry A* **125**, 6286-6302 (2021).

87. C. Wehmeyer, F. Noé, Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *The Journal of Chemical Physics* **148**, 241703 (2018).

88. H. Wu, A. Mardt, L. Pasquali, F. Noe, paper presented at the Proceedings of the 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018.

89. H. Wu, F. Paul, C. Wehmeyer, F. Noé, Multiensemble Markov models of molecular thermodynamics and kinetics. *Proceedings of the National Academy of Sciences* **113**, E3221-E3230 (2016).

90. A. Dommer *et al.*, #COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol. *bioRxiv*, 2021.2011.2012.468428 (2021).

91. H. Lee *et al.*, DeepDriveMD: Deep-Learning Driven Adaptive Molecular Simulations for Protein Folding. *2019 IEEE/ACM Third Workshop on Deep Learning on Supercomputers (DLS)*, 12-19 (2019).
92. O. R. N. Laboratory. (2019).
93. F. Noé, S. Olsson, J. Köhler, H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
94. L. Stanton *et al.*, in *arXiv*. (arXiv:2112.08651 2021).
95. X. Zhang *et al.*, ddcMD: A fully GPU-accelerated molecular dynamics program for the Martini force field. *J Chem Phys* **153**, 045103 (2020).
96. C. A. López *et al.*, Asynchronous Reciprocal Coupling of Martini 2.2 Coarse-Grained and CHARMM36 All-Atom Simulations in an Automated Multiscale Framework. *Journal of Computational and Theoretical Chemistry* **18**, 5022—5045 (2022).
97. R. Bommasani *et al.*, in *arXiv*. (arXiv:2108.07258, 2021).