CHEMICAL PHYSICS

Biomolecular dynamics with machine-learned quantum-mechanical force fields trained on diverse chemical fragments

Oliver T. Unke[1,2,3], Martin Stöhr[4]†‡, Stefan Ganscha[1], Thomas Unterthiner[1], Hartmut Maennel[1], Sergii Kashubin[1], Daniel Ahlin[1], Michael Gastegger[2,3,5], Leonardo Medrano Sandonas[4], Joshua T. Berryman[4], Alexandre Tkatchenko[4]*, Klaus-Robert Müller[1,2,6,7,8]*

Molecular dynamics (MD) simulations allow insights into complex processes, but accurate MD simulations require costly quantum-mechanical calculations. For larger systems, efficient but less reliable empirical force fields are used. Machine-learned force fields (MLFFs) offer similar accuracy as ab initio methods at orders-of-magnitude speedup, but struggle to model long-range interactions in large molecules. This work proposes a general approach to constructing accurate MLFFs for large-scale molecular simulations (GEMS) by training on "bottom-up" and "top-down" molecular fragments, from which the relevant interactions can be learned. GEMS allows nanosecond-scale MD simulations of >25,000 atoms at essentially ab initio quality, correctly predicts dynamical oscillations between different helical motifs in polyalanine, and yields good agreement with terahertz vibrational spectroscopy for large-scale protein-water fluctuations in solvated crambin. Our analyses indicate that simulations at ab initio accuracy might be necessary to understand dynamic biomolecular processes.

INTRODUCTION

Molecular dynamics (MD) simulations allow to determine the motion of individual atoms in chemical and biological processes, enabling mechanistic insights into molecular properties and functions, as well as providing a detailed interpretation of experimental studies. MD simulations require a reliable model of the forces acting on each atom at every time step of the dynamics (*1*). It is most desirable to obtain atomic forces from accurate solutions to the many-body Schrödinger equation, but this is only feasible for short MD simulations of few atoms for the foreseeable future (*2*). We remark that while there is always a unique exact solution to the Schrödinger equation for every atomic configuration, the proliferation of approximate empirical force fields (FFs) reflects the grand challenge of accurately capturing [and even fundamentally understanding (*3*)] interatomic interactions at all relevant length and timescales.

For larger systems, it is common practice to derive the forces from empirical models of the potential energy. Such force fields (FFs) approximate the interactions between atoms with computationally efficient, albeit rather rigid, terms and enable MD simulations of proteins at millisecond timescales (*4*).

A disadvantage of FFs is their limited accuracy due to the neglect of important quantum-mechanical effects, such as changes to hybridization states, interactions between orbitals delocalized over several atoms, or electronic correlations between distant molecular fragments. Further, many FFs require a predetermined covalent bonding structure, preventing bond breaking and formation. When additional accuracy and flexibility is required, for example, to study an enzymatic reaction, a possible alternative is quantum mechanics/molecular mechanics (QM/MM) simulations (*2*, *5*): The system is divided into a small QM region modeled with ab initio methods (e.g., substrate and active site of an enzyme) and an MM region (e.g., the remaining protein and solvent molecules) described with an FF. However, the high computational cost associated with an accurate treatment of the QM region and the fact that it is often unclear which atoms need to be included for an adequate description of the process of interest (*6*) may limit the applicability of QM/MM methods.

In recent years, machine-learned force fields (MLFFs) have emerged as an alternative means to execute MD simulations, combining the computational efficiency of traditional FFs with the high accuracy of quantum-chemistry methods (*7*). To construct an MLFF, a machine learning (ML) model is trained on ab initio reference data to predict energies and forces from atomic positions—without the need to explicitly solve the Schrödinger equation outside of the reference data. MLFFs have led to numerous insights, e.g., regarding reaction mechanisms (*8*), or the importance of quantum-mechanical effects for the dynamics of molecules (*9*) and have been successfully applied to MD simulations of small- to medium-sized systems (tens to hundreds of atoms) in gas phase (*10*) and periodic materials (e.g., metallic copper) with millions of atoms (*11*). Despite these successes, applications to large heterogeneous systems, like proteins or other biologically relevant systems, have largely remained elusive, due to the increased complexity of constructing physically informed ML architectures and obtaining reliable reference data for long-range interactions, which are known to play a key role in biomolecular dynamics (*12*, *13*). While the construction of MLFFs for oligopeptides (*14*) and

[1]Google DeepMind, Tucholskystraße 2, 10117 Berlin, Germany and Brandschenkestrasse 110, 8002 Zürich, Switzerland. [2]Machine Learning Group, Technische Universität Berlin, 10587 Berlin, Germany. [3]DFG Cluster of Excellence "Unifying Systems in Catalysis" (UniSysCat), Technische Universität Berlin, 10623 Berlin, Germany. [4]Department of Physics and Materials Science, University of Luxembourg, L-1511 Luxembourg City, Luxembourg. [5]BASLEARN — TU Berlin/BASF Joint Lab for Machine Learning, Technische Universität Berlin, 10587 Berlin, Germany. [6]Department of Artificial Intelligence, Korea University, Anam-dong, Seongbuk-gu, Seoul 02841, Korea. [7]Max Planck Institute for Informatics, Stuhlsatzenhausweg, 66123 Saarbrücken, Germany. [8]BIFOLD — Berlin Institute for the Foundations of Learning and Data, Berlin, Germany.
*Corresponding author. Email: alexandre.tkatchenko@uni.lu (A.T.); klaus-robert.mueller@tu-berlin.de (K.-R.M.)
†Present address: Department of Chemistry, Stanford University, Stanford, CA 94305, USA.
‡Present address: SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.

proteins (*15*) has been attempted previously, so far, they have not been demonstrated to yield stable and accurate dynamics over extended timescales (several nanoseconds). A more detailed overview over conventional and MLFFs can be found in section S1.

This work proposes a general approach to constructing accurate MLFFs for large-scale molecular simulations (GEMS). On the basis of the divide-and-conquer principle, MLFFs for large heterogeneous systems are trained on molecular fragments of varying size, which are still amenable to electronic-structure calculations. These fragments do not form a partition of the larger system; rather, they can be overlapping pieces, or even just be structurally related to the original system. The fragments are not used directly when evaluating the MLFF, but only during the training process to learn the relevant physicochemical interactions present in the larger system. From these fragment data (which include water or solvent molecules), the ML model infers to recompose the original system and is able to predict the full potential energy surface (PES) including interactions with solvent, which allows GEMS to successfully address the long-standing challenge of biomolecular simulations at ab initio quality (Fig. 1A). As such, GEMS refers to the general principle of running molecular simulations with MLFFs constructed in this fashion (see also fig. S21 for a schematic depiction).

While MLFFs can successfully learn local chemical interactions from small molecules (*16*), a sufficient number of larger fragments are needed to learn long-range effects necessary to generalize to larger systems and achieve high prediction accuracy (0.450 meV/atom for energies and 36.704 meV/Å for forces) with respect to the ab initio ground truth. Here, we rely on the recently proposed SpookyNet architecture (*17*), which models dispersion and electrostatics explicitly by embedding physically motivated interaction terms into the ML architecture and learning their parameters from reference data. We note that the SpookyNet model is not the first to explicitly model long-range electrostatics, and other models follow similar approaches (*18*–*21*). In addition, an empirical term for short-ranged repulsion between atomic nuclei increases the robustness of the model for strong bond distortions. SpookyNet also includes a mechanism to describe effects like nonlocal charge transfer, which other MLFFs [with some exceptions (*22*)] are typically unable to. Together, these components enable the model to generalize to larger molecules when trained on appropriate reference data. Crucially, this allows GEMS to account for cooperative, long-range effects, which is difficult or impossible for conventional FFs. While extensive reference data for small fragments are mainly used to learn a robust "baseline" representation of short-ranged interactions, additional larger fragments allow GEMS to also capture long-range interactions and the interplay between different interaction scales. In the same manner, solvent effects can also be included (by explicitly describing the interaction with solvent molecules). We demonstrate that GEMS can learn to accurately model large-scale phenomena, such as cooperative polarization effects, from such fragment data, achieving close agreement to the ab initio ground truth.

However, ultimately, the quality and reliability of an MLFF should be judged by its predictions of experimental measurements—for example, we show that GEMS is able to quantitatively reproduce experimental results regarding the helix stability of polyalanine systems at different temperatures and correctly describe the terahertz

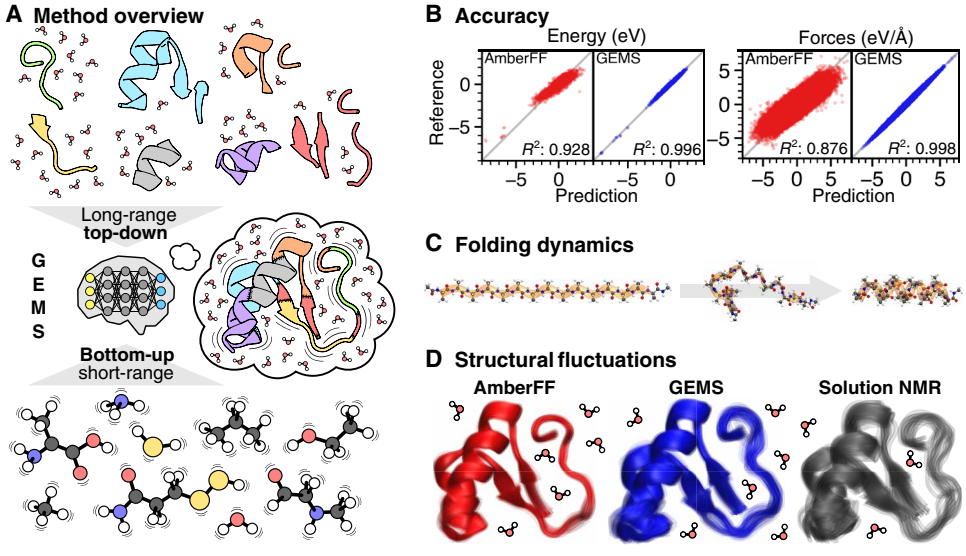

Fig. 1. Insights from GEMS simulations. (**A**) Overview of the GEMS method. Different interaction scales on the PES of a large system are learned from a combination of ab initio reference data for top-down and bottom-up fragments. The resulting model is able to accurately reconstruct the PES of the molecular system and then used to study its dynamics. (**B**) Prediction accuracy for energies and forces of $AceAla_{15}Nme$ conformations of GEMS compared to AmberFF (*24*) with respect to the PBE0/def2-TZVPP+MBD (*28*, *29*, *44*) reference. Note that AmberFF was not fitted to PBE0/def2-TZVPP+MBD reference data, so a direct comparison can only show qualitative trends. (**C**) GEMS simulations show that the folding of $AceAla_{15}Nme$ from a fully extended structure (FES) (left) to a helical conformation (right) at 300 K in gas phase occurs via intermediate conformations characterized by hydrogen bonding between backbone atoms of adjacent residues (middle). (**D**) Overlay of representative conformations (obtained from cluster analysis) sampled during an aggregated 10 ns of NPT dynamics of crambin in aqueous solution at 300 K and ambient pressure (see also fig. S13 for a 360° view of crambin highlighting the interactions relevant for its three-dimensional structure). Simulations with GEMS (blue) lead to greater structural fluctuations compared to AmberFF (red), indicating that the protein is more flexible. For comparison, 20 low energy water refined structures of crambin in dodecylphosphocholine micelles based on NMR measurements (gray) are shown as well (*47*). To allow a quantitative comparison, structures should be modeled with GEMS instead of a conventional FF when interpreting the NMR results.

infrared (IR) vibrational spectrum of a solvated 46-residue protein (crambin), which is extremely difficult to achieve using traditional empirical FFs that do not account for collective many-body interactions and therefore yield large-scale vibrational modes that are qualitative at best, typically giving a smear-out of peak structure and an exaggeration of amplitude over the 25 to 150 cm^{-1} spectral region (*23*).

GEMS is applied to MD simulations of model peptides and the 46-residue protein crambin in aqueous solution with 8205 explicit water molecules (>25,000 atoms). When comparing to conventional FFs, such as AMBER99SB-ILDN (*24*) (AmberFF), GEMS approximates energies and forces computed from density functional theory much more closely (Fig. 1B). Our findings reveal previously unknown intermediates in the folding pathway of polyalanine peptides (Fig. 1C) and a dynamical equilibrium between α- and 3_{10}-helices. In the simulations of solvated crambin, GEMS indicates that protein motions are qualitatively different, with much smoother PESs and softened vibrations when compared to computations with a conventional FF (Fig. 1D), showing contrasting short and long timescale dynamics. Low-frequency vibrational modes largely determine the free energy of proteins (*25*); hence, our results suggest that simulations at ab initio accuracy may be necessary to fully understand dynamic processes in biomolecules.

RESULTS

MLFFs for large systems trained on diverse chemical fragments

We start by generating reference data for smaller molecular fragments to train an MLFF, where the learned model accurately reflects the full large system. There are several strategies to achieve this goal. On the one hand, the model needs to be able to learn all relevant interactions that are necessary to reconstruct a complete and accurate picture of the system of interest from the fragment data. This is important to capture weak, but long-range interactions, which collectively dominate, e.g., relative energy differences of different conformations of large molecules. On the other hand, it is necessary to prevent "holes" in the PES (*18*)—regions with low potential energy corresponding to unphysical structures, e.g., featuring unnaturally large or short bond lengths. The existence of holes in the PES prevents stable MD simulations, because long trajectories eventually may become trapped by such artefacts and behave unphysically (*26*). To achieve both requirements, we propose the use of two complementary methods to construct fragments, which allow models to learn different aspects of the PES of large systems. The first method follows a top-down approach, where fragments are constructed by "cutting out" spherical regions of the system of interest, which also includes solvent molecules in the condensed phase (Fig. 2A) (*27*). By including solvent molecules in the generated fragment data, the MLFF can learn to treat solvent effects explicitly from reference data. Any dangling bonds resulting from cutting through covalent bonds are saturated with hydrogen atoms or by including a limited number of atoms beyond the cutoff radius (see Materials and Methods for details). The fragments are chosen as large as possible to sample important long-range effects, but still small enough such that reference energies and forces computed with quantum chemistry methods are accessible in a reasonable time. As our tests on polyalanine systems demonstrate (see below), the top-down fragments we choose are sufficiently large for the systems studied in this work. Although any generated top-down fragments are system-specific (except for possible

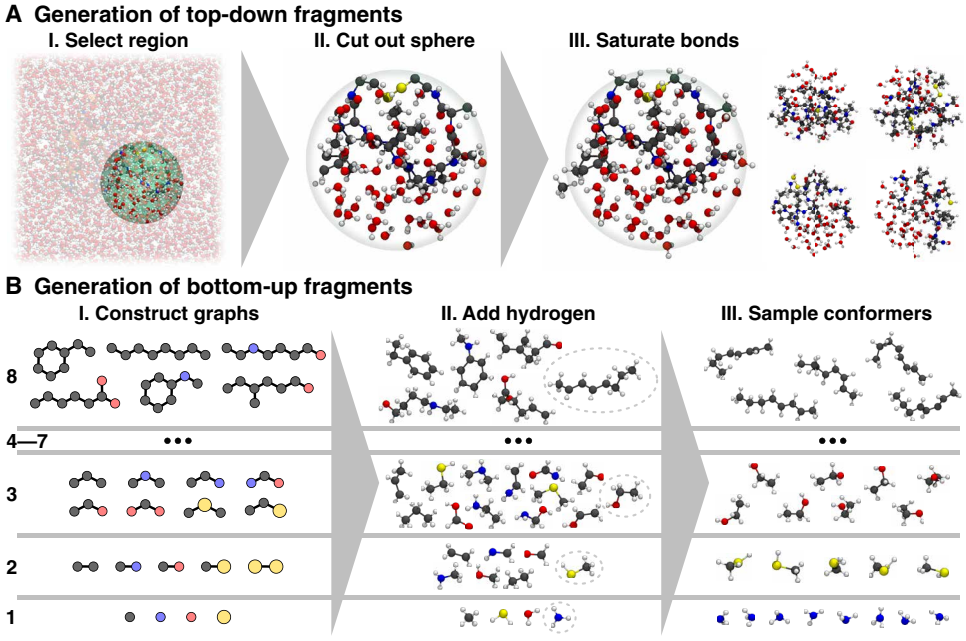

Fig. 2. Generation of top-down and bottom-up fragments. (**A**) Top-down fragments are generated by cutting out a spherical region around an atom (including solvent molecules) and saturating all dangling bonds (the right side shows four top-down fragments generated from different regions). They are crucial for learning weak but long-range interactions, which are important for the dynamics of large systems. (**B**) Bottom-up fragments are generated by constructing chemical graphs consisting of one to eight nonhydrogen atoms (not all possible graphs are shown). The graphs are then converted to three-dimensional structures by adding hydrogen atoms. Because of their small size, multiple ab initio calculations for many different conformers of each generated structure can be performed, allowing extensive sampling of the PES, which is necessary for training robust models.

structural similarities to other systems), the method to obtain them is general and can be applied to any large condensed phase system. Further, generated top-down fragments may still be used as training data for learning in new systems.

To train robust models, the top-down fragments are enriched by smaller bottom-up fragments, for which atomic forces for many different conformations can be calculated. Starting from single atoms, molecules similar to local bonding patterns of the system of interest (*16*) are systematically constructed by growing chemical graphs in a bottom-up fashion (Fig. 2B) (missing valencies are filled with hydrogen atoms, see Materials and Methods for details). By limiting the size of these fragments, it is possible to sample many different conformations, allowing models to learn the effects of strong distortions in local structural patterns, which is key to preventing holes in the PES. As a result, the combination of bottom-up and top-down fragments enables learning accurate and robust MLFFs for large systems.

"Accurate" in this context refers to the ability of the MLFF to reproduce the chosen reference method. The "true accuracy" of the MLFF (and thus also the GEMS method itself), i.e., its ability to capture the physics of a particular system of interest, is tied to the accuracy of the underlying reference method chosen to compute the training data. Here, we choose PBE0+MBD (*28*, *29*) as reference method. It explicitly includes long-range dispersion interactions, yet is sufficiently efficient to perform reference calculations for the larger top-down fragments. This level of theory has been shown to provide an accurate and reliable description in excellent agreement with high-level quantum chemistry methods and experiment for, e.g., polypeptides (*30*, *31*), supramolecular complexes (*32*), and molecular crystals with and without water (*33*, *34*), which show very similar bonding patterns as the biomolecular systems studied here. Additionally, PBE0+MBD has been found to be well suited for modeling interactions of proteins in water (*13*, *35*).

To summarize, the training data for GEMS consist of a large number of small "general" fragments (2,713,986 structures), which can be shared for a wide class of chemical systems (in this work: peptides/proteins in gas phase and aqueous solution covering interatomic distances from below 1 Å to about 12 Å), and a small number of large system-specific fragments (covering interatomic distances up to 18 Å). For example, for training GEMS for crambin, 5624 additional top-down fragments are used (see also fig. S25 for a histogram showing the size distribution of fragments for the crambin training data and fig. S26 for an overview over the distribution of pairwise distances). In total, the dataset used in this work to build MLFF models amounts to about 60 million atomic forces, ranging from 650,000 forces for sulfur to 37.6 million forces for hydrogen. The constructed fragments also contain substantial information about water interacting with protein fragments, with about 5 million water molecules in total.

Polyalanine systems

We apply GEMS to predict the properties and dynamics of several peptides consisting primarily of alanine. These are popular model systems for proteins and well studied both theoretically and experimentally. Further, by limiting the number of residues, it is still possible to perform electronic-structure calculations for the full system. Thus, the predictions of an ML model trained only on fragment data can be directly compared to reference calculations, which allows to verify the ability of GEMS to reconstruct the properties of larger systems from the chemical knowledge extracted from smaller molecules.

As a first test case, we consider the cooperativity between hydrogen bonds in polyalanine peptides capped with an N-terminal acetyl group and a protonated lysine residue at the C-terminus ($AceAla_n$-Lys + H^+). In α-helices, the local dipole moments of hydrogen bonds formed between backbone peptide groups are aligned, leading to a cooperative polarization effect (*36*). Thus, the relative stabilization energy of an α-helix compared to a fully extended structure (FES) fluctuates nontrivially with helix length and is a challenging prediction task. We find that GEMS closely agrees with the reference ab initio method, demonstrating that large-scale effects can be learned effectively from fragment data (Fig. 3A).

Alanine-based peptides have a strong tendency to form helical structures. While short isolated helices are only marginally stable in solution, $AceAla_{15}Lys + H^+$ is known to form stable helices in gas phase. Experimental results suggest that $AceAla_{15}Lys + H^+$ remains helical up to temperatures of ~725 K (*37*), allowing a direct comparison with theoretical predictions. By running GEMS simulations at different temperatures, we confirm that the peptide remains primarily helical up to 700 K, but forms a random coil at 800 K (see movie S1). An analysis of the formed hydrogen bonds reveals that the average number of α-helical hydrogen bonds decreases with increasing temperature (see fig. S12A), while the number of 3_{10}-helical hydrogen bonds remains almost constant until a sudden drop at 800 K (see Fig. 3B). This agrees with results from ab initio MD simulations at the PBE+vdW level (*38*), where a similar relationship between temperature and the stability of different kinds of hydrogen bonds was found. The long-range interactions learned from top-down fragments seem to be crucial to reproduce the experimental results, as a model that was only trained on bottom-up fragments predicts reduced thermal stability (see fig. S12B).

To investigate whether there are fundamental differences between GEMS and dynamics simulations performed with conventional FFs, we study the room temperature (300 K) folding process of a pure polyalanine peptide capped with an N-terminal acetyl group and a C-terminal N-methyl amide group ($AceAla_{15}Nme$) in gas phase. Starting from the FES, MD simulations with GEMS suggest that $AceAla_{15}Nme$ has a strong tendency to form H-bonds between peptide groups of directly adjacent residues within the first ~100 ps of dynamics. The formed arrangements exhibit a "wavy" structure and ϕ and ψ backbone dihedral angles of ~0° and ~0°, which lie in a sparsely populated region of the Ramachandran plot. These intermediates are typically short-lived with lifetimes of ~25 to 50 ps and fold readily into helical configurations via a characteristic twisting motion. There is still some controversy between theoretical and experimental results regarding the predominance of different helical conformations (*39*). We find that there may be cases where no single motif is preferred: Once a helix is formed, its structure fluctuates between pure α- and 3_{10}-helices, as well as hybrids of both helix types (see movie S2 for a complete folding trajectory). A 10-ns trajectory of the helical state suggests a dynamical equilibrium with a ~38/62% mixture of α- and 3_{10}-helices. Such a dynamical coexistence of α- and 3_{10}-helices has already been observed experimentally in alanine-rich peptides (*40*, *41*) and can be assessed for the specific example of $AceAla_{15}Nme$ in future experiments using nuclear magnetic resonance (NMR) or ultraviolet (UV)/IR spectroscopies on polyalanine folding under "clean room" gas-phase conditions. In contrast, MD simulations with the AmberFF yield qualitatively different results, suggesting that a more rigid and primarily α-helical configuration is formed from the FES without distinct structural intermediates (see

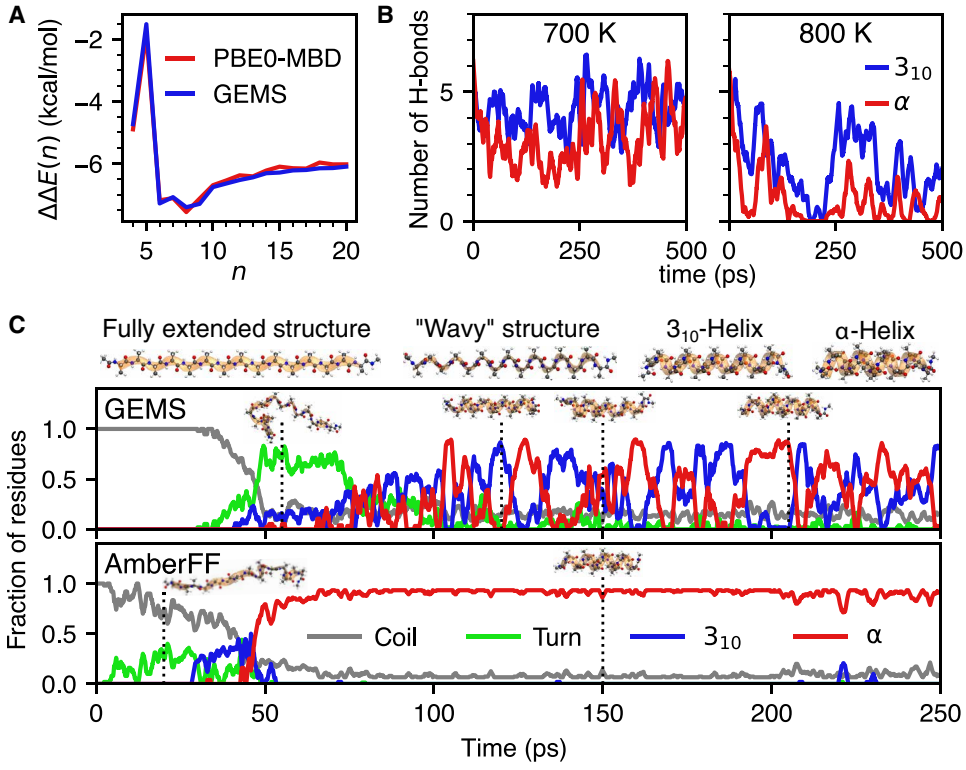

Fig. 3. Accurate simulations of polyalanine systems with GEMS. (**A**) Relative stabilization of the α-helical conformation of $AceAla_nLys + H^+$ per added alanine residue. Shown here is the double difference $\Delta\Delta E(n) = \Delta E(n) - \Delta E(n-1)$, where $\Delta E(n) = E_\alpha(n) - E_{FES}(n)$ is the relative energy of the α-helical conformation and the FES of $AceAla_n$-$Lys + H^+$ in gas phase. The prediction of GEMS (blue) is compared to ab initio reference data computed at the PBE0+MBD (*28*, *29*, *44*) level of theory. (**B**) Number of α- and 3_{10}-helical H-bonds during MD simulations of helical $AceAla_{15}Lys + H^+$ in gas phase at 700 K and 800 K with GEMS. The sharp drop in the number of H-bonds in the dynamics at 800 K indicates the formation of a random coil (see fig. S12A for an extended version of this figure with a greater range of temperatures). (**C**) Secondary structural motifs determined by STRIDE (*79*) along typical folding trajectories of $AceAla_{15}Nme$ at 300 K in gas phase. Dotted vertical lines indicate the temporal position of the shown snapshots. The trajectory computed with GEMS (top) folds via a distinct "wavy" intermediate (classified primarily as "turn") and settles into a dynamic equilibrium between 3_{10}- and α-helices. In contrast, the trajectory computed with the AmberFF (bottom) folds more directly and then stays primarily α-helical (see fig. S9 for an analysis of additional trajectories).

Fig. 3C). Additionally, we also investigated the dynamics with the CHARMM27 (*42*) and GROMOS96 53A5 (*43*) FFs (see fig. S11 for representative trajectories). While the dynamics with CHARMM27 are comparable to those of AmberFF (apart from typically folding slightly later during the dynamics), we could not observe helix formation when using GROMOS96 53A5 at all. However, the "wavy intermediate" observed in the GEMS simulations (where this structure seems to be metastable) is readily formed almost instantly in many GROMOS trajectories, but does not fold to a helix subsequently and instead is stable over hundreds of picoseconds. The partial agreement of different classical FFs with the GEMS trajectory can thereby be understood as the parametrization process of the individual FFs imposing a select, limited set of correct constraints. As a result, each FF correctly captures certain aspects (such as formation of the wavy intermediate or folding into an α-helix) but fails to correctly reproduce all features due to the limited flexibility of the fixed FF energy functional. Altogether, the ab initio accurate GEMS simulations provide a concrete prediction of a dynamical coexistence of α- and 3_{10}-helices in $AceAla_{15}Nme$, which is in line with experimental observations for alanine-rich polypeptides, but is not predicted by conventional FFs. It is important to point out that conventional FFs are usually parametrized for simulations in solvent, not in the gas phase. As such, their performance in gas phase is not necessarily an indicator for their performance in solution. The results shown here for different conventional FFs are meant to emphasize that differences in parametrization can lead to substantially different dynamics, which all differ qualitatively from the dynamics predicted by GEMS. We note in passing that many other FFs are available in the literature. However, most of them are based on the same restricted functional form for the bonded interactions. A comprehensive assessment of different generations of empirical FFs goes beyond the scope of our work.

For completeness, we also investigate trajectories of $AceAla_{15}Nme$ simulated with a GEMS model that was only trained on bottom-up, but not top-down, fragments. In this setting, we observe no helix formation on the investigated timescale; instead, $AceAla_{15}Nme$ typically stays structurally close to the FES during the dynamics, only rarely forming partial loop motifs (see fig. S10). This suggests that the inclusion of top-down fragments is crucial to correctly describe the folding process and is consistent with the results obtained for the thermal stability of $AceAla_{15}Lys + H^+$, where a model trained without top-down fragments predicts diminished stability of the folded state (see also fig. S12B).

As a final test for the accuracy of GEMS, we compare the predictions of the ML model to ab initio data computed at the PBE0/def2-TZVPP+MBD (*28*, *29*, *44*) level of theory. To this end, we use 1554 and 1000 $AceAla_{15}Nme$ structures sampled from densely and sparsely

populated regions (rare events) of the configurational space visited in 100 aggregated 250-ps MD trajectories (25 ns total) in the NVT ensemble at 300 K simulated with GEMS (see section S4 for details). We find that predicted energies and forces are in good agreement with the reference values in both cases. For energies and forces, correlation coefficients are $R^2 = 0.996$ and $R^2 = 0.998$, respectively, and mean absolute errors (MAEs) are 0.450 meV/atom and 36.704 meV/Å. Again, we find that the inclusion of top-down fragments during training is crucial for high accuracy, as prediction errors for a model trained only on bottom-up fragments are much larger (see fig. S8). For completeness, we also compare predictions with the conventional AmberFF (*24*) to the ab initio reference. Although AmberFF is not fitted to reproduce energies and forces from density functional theory (DFT) calculations, its predictions display correlation coefficients of $R^2 = 0.928$ (for energy) and $R^2 = 0.876$ (for forces). Nonetheless, the MAEs are much larger at 2.274 meV/atom and 329.328 meV/Å (distributions of predicted and reference energy values were shifted to have a mean of zero before computing MAEs in both cases such that constant energy offsets between different methods do not influence the results). Although a quantitative comparison between GEMS and AmberFF in this context is not meaningful, as a qualitative trend, we observe that predictions with GEMS reproduce the reference across the whole range of values without the presence of a single outlier, whereas the AmberFF systematically under- and overpredicts small and large energy values, respectively (see Fig. 1B). These findings show that GEMS gives accurate predictions even for rare configurations and the simulated MD trajectories are essentially ab initio quality (see figs. S6 and S7 for a more detailed analysis of correlations within the different subsets of configurations). This comparison between GEMS and AmberFF also suggests that reproducing relative energies of different protein conformations is an easier task than accurately capturing atomic forces that drive the biomolecular dynamics.

Crambin

GEMS enables accurate molecular simulations in the condensed phase. The 46-residue protein crambin in aqueous solution (25,257 atoms) is chosen as a model system. Crambin contains 15 of the 20 natural amino acids and forms common structural motifs such as β-sheets, α-helices, turns/loops, and disulfide bridges. To assess qualitative differences between simulations with a conventional FF (here, the AmberFF is chosen) and GEMS, we consider the power spectrum (*45*) computed from 125 ps of dynamics at a temporal resolution of 2.5 fs (Fig. 4A) (no constraints were used for bonds to hydrogen atoms in these simulations). The power spectrum is related to the internal motions of the system and reveals the dominant frequencies of molecular vibrations, which are influenced by the atomic structure and characteristic for the presence of certain functional groups. In comparison to the results obtained from the dynamics with a conventional FF, peaks in the power spectrum computed with GEMS are shifted toward lower wave numbers and lie close to the frequency ranges expected from measured IR spectra. For example, the dominant peaks above 1000 cm^{-1} correspond to bending and stretching vibrations of water molecules, which are experimentally expected at around ~1600 cm^{-1} and ~3500 cm^{-1}, respectively (*46*), which is consistent with the GEMS spectrum. In contrast, the corresponding peaks for the conventional FF are blue-shifted several hundreds of wave numbers. Additionally, peaks in the GEMS spectrum are broader, indicating that the frequencies of characteristic vibrations are influenced stronger by intermolecular interactions, hence broadening their frequency range. Long-range interactions may particularly influence slow protein motions, i.e., the low-frequency parts of the power spectrum, where notable differences between GEMS and AmberFF can be observed.

Similar to the results for $AceAla_{15}Nme$, we find that in comparison to simulations with the AmberFF, crambin is more flexible in GEMS simulations (Fig. 1D). Qualitatively, the increased flexibility seems to agree more closely to structures modeled from NMR spectroscopy measurements (see Fig. 1D), but a direct comparison of root mean square deviation (RMSD) values along GEMS and AmberFF trajectories reveals that both simulation types agree similarly well with the structural ensemble (see figs. S22 and S23). The direct comparison to the structures published in (*47*) should be regarded with caution, as the structural ensemble is not obtained by a direct experimental measurement, but rather fitted to experimentally obtained distance constraints with a conventional FF, which may introduce notable bias. The increased flexibility is also indicated by a Ramachandran map of the backbone dihedral angles of crambin (Fig. 4B), which shows that a wider range of values is sampled in simulations with GEMS (backbone bond length distributions are, however, comparable between both simulations; see fig. S15). Similar results can be observed in Fig. 5, where the GEMS simulation shows an additional mode for the torsion angle in two of the six cysteine residues that form disulfide bridges, and in fig. S18, which shows that the GEMS trajectories contain more modes of the distributions for the four torsion angles in ARG17. An illustration of different configurations is shown in Fig. 6.

This becomes even more apparent by projecting the trajectories into a low-dimensional space that allows a direct visualization of the path taken through conformational space (Fig. 4C). However, a time-resolved analysis of the trajectories reveals that structural fluctuations with GEMS are only larger on timescales in excess of ~200 ps (Fig. 4D). On shorter timescales on the other hand, the trend is reversed. Despite the fact that fluctuations in GEMS trajectories seem to be growing on larger timescales, we find that the overall structure stays close to the folded state at all times (see fig. S24), i.e., we do not observe any signs of early unfolding. We can also see the timescale dependence directly on the torsion angles: When forming a moving average over 100 time steps, short time fluctuations cancel out more for the AmberFF trajectories, which leads to sharper peaks in the distributions. As an example, fig. S20 shows the distribution of the torsion angles for the other four cysteine residues. While the original distributions (on the left) do not differ much between GEMS and AmberFF, the distributions of the averaged angles (on the right) show a much sharper peak for AmberFF. The same can be observed in fig. S19 for ARG10. This suggests that there are qualitative differences between simulations with conventional FFs and GEMS on all timescales.

To investigate whether the difference in structural fluctuations has a direct effect on experimental observables, we compute terahertz timescale IR spectra of solvated crambin simulated with GEMS and AmberFF to experimental measurements for a partially solvated crambin sample (*48*). The terahertz spectrum corresponds to slow dynamics of the folded protein in solvent, and hence, it is sensitive to the correct description of long-range interactions. We find that GEMS is able to reproduce most experimentally observed features rather well (see Fig. 7), whereas the spectrum computed from AmberFF simulations lacks the distinctive features of the experimental spectrum and is largely featureless over the range for which experimental data

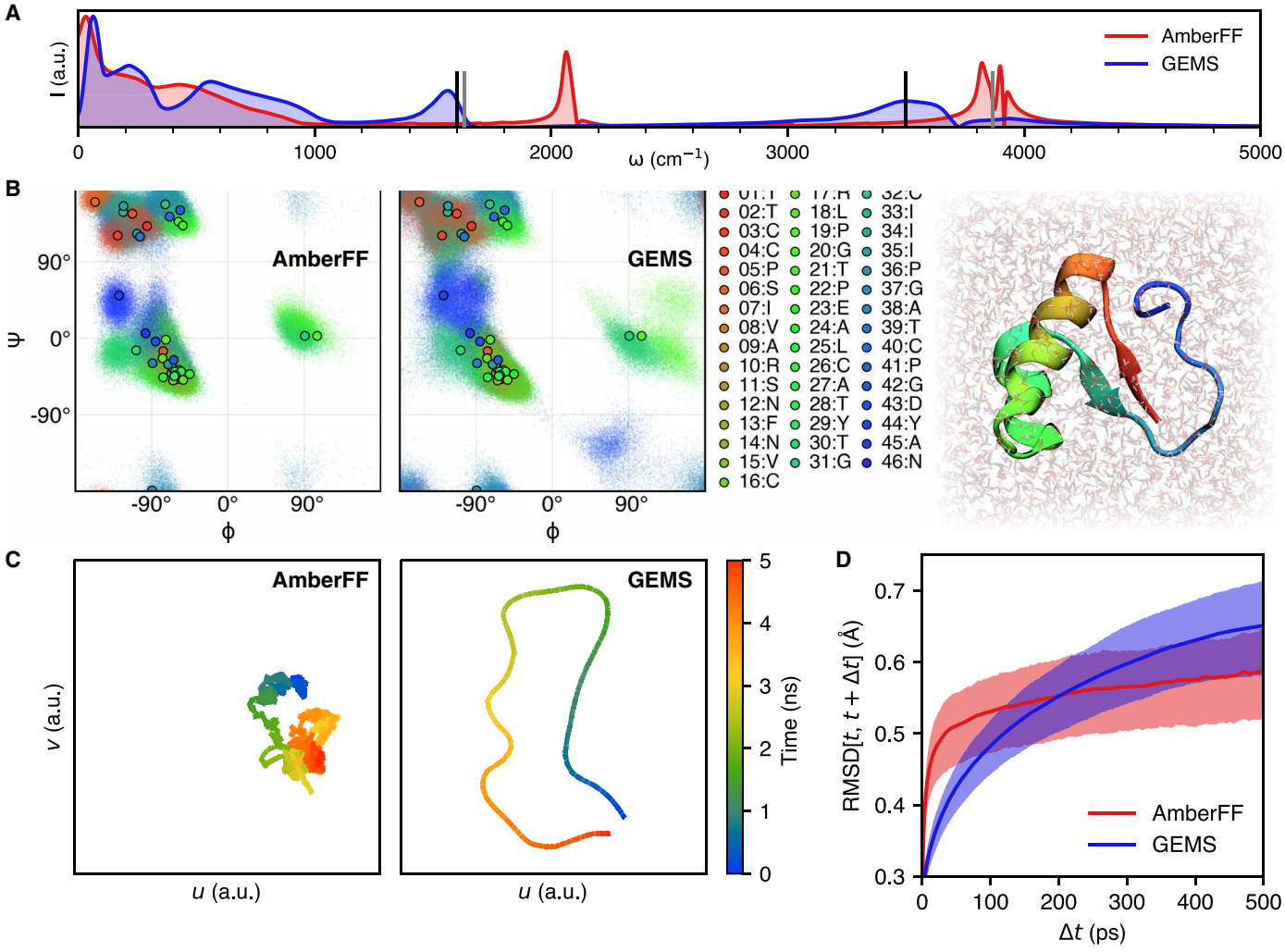

Fig. 4. Analysis of dynamics simulations of crambin in aqueous solution. (**A**) Power spectrum of crambin in water obtained from 125 ps of dynamics computed with the AmberFF (without any constraints on bonds to hydrogen) and GEMS. In the GEMS spectrum, peaks associated with bending and stretching vibrations of water molecules are closer to experimentally expected values at around ~1600 cm^{-1} and ~3500 cm^{-1} (*46*) (vertical black lines). For reference, we also show the gas-phase harmonic peaks calculated at the PBE0+MBD level of theory (vertical gray lines). (**B**) Ramachandran map for crambin (color-coded by residue number). The scatter shows the (ϕ, ψ)-dihedral angles sampled during an aggregated 10 ns of dynamics; points with black outline show values of the crystal structure (*70*) for reference. Dynamics with GEMS (right) generally sample a broader distribution compared to AmberFF (left), indicating that the protein is more flexible. (**C**) Two-dimensional Uniform Manifold Approximation and Projection (UMAP) (*80*) projection of the path through conformational space sampled during a 5-ns trajectory of crambin in aqueous solution. Compared to the trajectory computed with the AmberFF, dynamics with GEMS sample a wider distribution and are less likely to revisit previously visited regions of conformational space. (**D**) Distribution of RMSDs (excluding hydrogen atoms) between conformations sampled at times t and $t + \Delta t$. Solid lines depict the mean, whereas the shaded region indicates the area between the 25th and 75th percentiles. Dynamics with the AmberFF (red) show larger structural fluctuations on short timescales, whereas fluctuations on longer timescales are larger for dynamics computed with GEMS (blue).

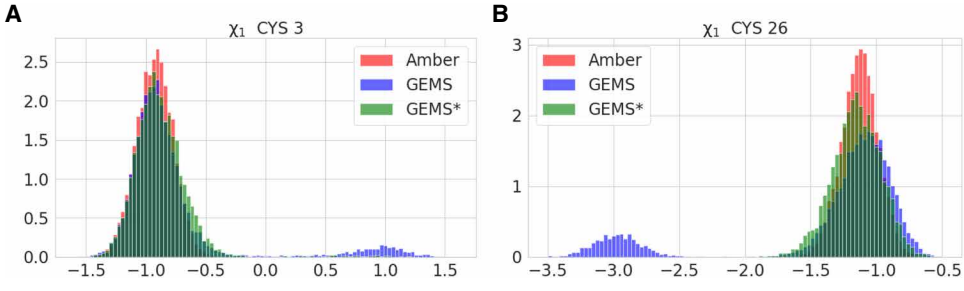

Fig. 5. Comparison of dihedral angles of disulfide bridges under different FFs. (**A**) Residue 3. (**B**) Residue 26.

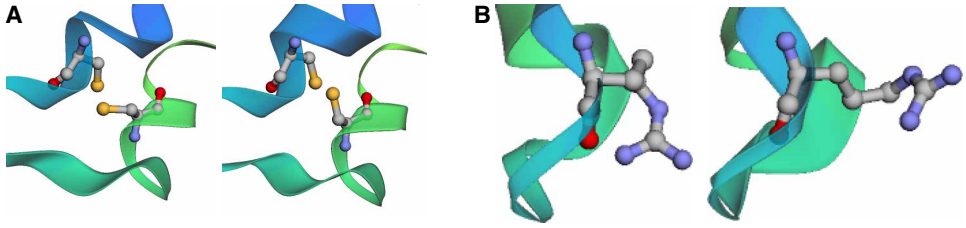

Fig. 6. Cysteine/arginine residues in crambin. (**A**) Different torsion angles χ_1 in CYS3, as observed in the GEMS trajectory. (**B**) Different ARG17 configurations, as observed in the GEMS trajectory.

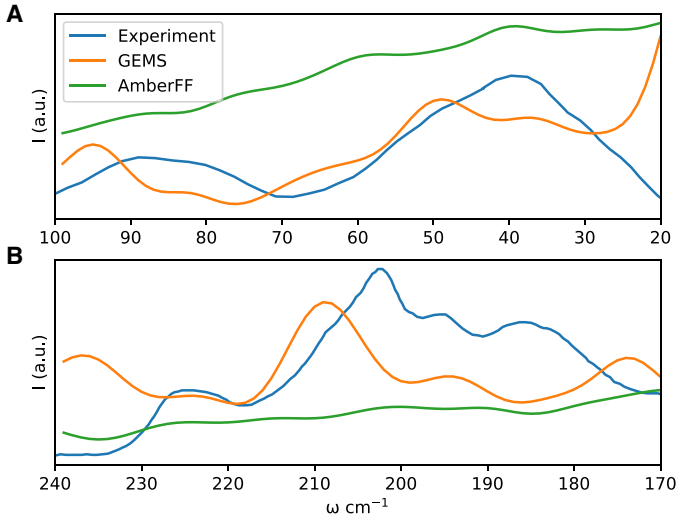

Fig. 7. IR spectrum of crambin on the terahertz timescale. (**A**) Frequency range from 20 to 100 cm^{-1}. (**B**) Frequency range from 170 to 240 cm^{-1}. Good agreement between GEMS and the experimental spectrum (*48*) is found, whereas the spectrum computed from simulations with AmberFF is smooth and largely featureless.

are available, consistent with the vibrational power spectrum shown in Fig. 4A. We remark that one cannot expect quantitative agreement with experiment, given that the solvent in the experimental system was a mixture of water and organic salts, while the simulations were run in pure water, with dielectric screening from partial solvent then recalculated post hoc (see Materials and Methods for details).

In addition, we also compare the crambin dynamics observed with GEMS to those of a model that was only trained on (general) bottom-up fragments, but not on (system-specific) top-down fragments (referred to as GEMS* in the following). While a visual inspection of the GEMS* trajectories suggests no marked differences to the regular GEMS model, a detailed analysis reveals that, while being qualitatively similar overall, some regions of the Ramachandran map are less frequently visited and appear closer to the observations for AmberFF (see fig. S17A). Further, the additional modes (compared to the AmberFF trajectory) in the distribution of the cysteine torsion angles (that the GEMS trajectory showed) vanish for GEMS* (see Fig. 5). Also, the distribution of torsion angles for ARG17 in GEMS* lies somewhere in between the observations for GEMS and AmberFF (see fig. S18). Similarly, while structural fluctuations on short timescales agree between GEMS and GEMS*, the long timescale fluctuations of GEMS* are smaller than those for GEMS and instead closer to those observed for AmberFF (see fig. S17B). As such, it appears that long-range effects learned from top-down fragments are crucial for describing the crambin dynamics on long timescales.

A model for understanding the different dynamics of crambin

Reasons for the observed qualitative differences between AmberFF and GEMS trajectories must be related to differences in the PESs. AmberFF is a conventional FF, and as such, models bonded interactions with harmonic terms. Consequently, structural fluctuations on small timescales are mostly related to these terms. Intermediate-scale conformational changes as involved in, for example, the "flipping" of the dihedral angle in the disulfide bridges of crambin, on the other hand, can only be mediated by (nonbonded) electrostatic and dispersion terms, because the vast majority of (local) bonded terms stay unchanged for all conformations. On the other hand, GEMS makes no distinction between bonded and nonbonded terms, and individual contributions are not restricted to harmonic potentials or any other fixed functional form. Consequently, it can be expected that large structural fluctuations for AmberFF always correspond to "rare events" associated with large energy barriers, whereas GEMS dynamics arise from a richer interplay between chemical bonds and nonlocal interactions.

To test this hypothesis, we introduce a simplified model of a high-dimensional PES based on superposed one-dimensional oscillators confined to a double-well potential. The potential energy is given by

$$E(\mathbf{x}) = \sum_{i=1}^{N} \left(\frac{h_i}{a_i}\right)^4 (x_i^2 - a_i^2)^2 \qquad (1)$$

where h_i is the barrier height between two minima, $2a_i$ is their separation, and x_i is the coordinate of oscillator i (see Fig. 8A). The vector $\mathbf{x} = [x_1 \ldots x_N]^\intercal$ is the "configuration" of the N-dimensional model system. We then simulate the trajectory $\mathbf{x}(t)$ with Langevin dynamics, which couples the oscillator modes to a shared heat bath that allows energy transfer between them. This setup constitutes a simplified model of PESs with different qualitative properties. We find that for a system where large conformational changes (a is large) are always associated with high energy barriers (h is large, roughly ~4 times larger than for small conformational changes), low-dimensional projections of $\mathbf{x}(t)$ resemble those of crambin simulated with the AmberFF. On the other hand, if large conformational changes are also possible with low energy barriers, the projections resemble those of the GEMS trajectory (see Figs. 4C and 8B). All computational details for these experiments are given in Materials and Methods. These results suggest that GEMS simulations substantially enhance large-scale structural transitions between distant conformations,

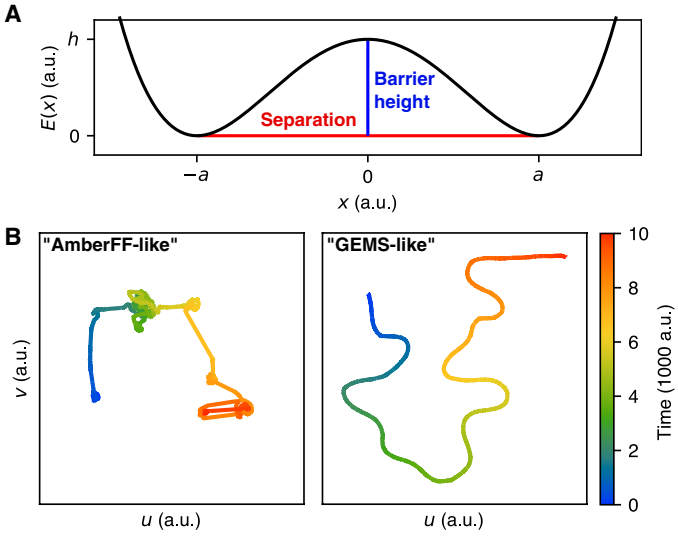

Fig. 8. Simplified model for PESs with different dynamical properties. (**A**) One-dimensional double-well potential showing the effects of the separation parameter *a* (large values of *a* correspond to large possible conformational changes) and the barrier height *h*. (**B**) Low-dimensional projections (similar to Fig. 4C) of trajectories on a PES where large conformational changes are always associated with high energy barriers ("AmberFF-like") and a PES where large conformational changes may be associated with small energy barriers (GEMS-like).

which is in agreement with the flexibility of crambin observed in NMR experiments (see Fig. 1D) (*47*). These results are also consistent with (i) the much increased flexibility of polyalanine helices observed with GEMS in comparison to the essentially rigid dynamics obtained in AmberFF simulations (see Fig. 4C) and (ii) the inability of traditional FFs to reproduce enhanced structural fluctuations of mobile protein regions observed in NMR experiments (*49*).

We would like to reiterate that this analysis addresses the PES topology of crambin in the folded state. As such, the results shown in Fig. 8 and discussed above concern larger-scale corrugation rather than global PES barriers (i.e., flexibility within structural basins rather than folding events). On the basis of the increased smoothness and flexibility found in GEMS, we expect that the correct, ab initio accurate treatment would predict a broader ensemble of folding pathways characterized by more collective, low-frequency rearrangements. This, however, does not necessarily affect the overall timescales of folding. Increased flexibility and a wider transition path ensemble accompanied by lower-frequency (slower) dynamics can preserve the (relative) phase space volumes of PES minima and the transition state (ensemble). With relative phase space volumes being the key quantity in standard reaction rate theory, the topological changes may thus not affect the overall folding rate. On the other side, conventional FFs are typically designed to reproduce properties such as folding rates to the largest extent possible within the fixed form of the classical energy functional. Simply reproducing the correct timescales, however, does not necessarily entail the correct dynamics or folding pathway, where it is known that different choices of conventional FFs produce substantially different results (*50*).

DISCUSSION

Modeling quantum-mechanical properties of large molecules is an outstanding challenge, and it holds promise for broad application in chemistry, biology, pharmacology, and medicine. We have developed a general framework for constructing MLFFs—GEMS—for large molecular systems such as proteins by learning from ab initio reference data of small(er) fragments without the need to perform electronic-structure calculations for a whole protein—as the latter would constitute a computationally impractical task. The proposed divide-and-conquer strategy using a library of ~3 million DFT+MBD computations on fragments and using an ML model that incorporates physical constraints and long-range interactions allows to efficiently construct MLFFs that accurately reflect the quantum-mechanical energies and atomic forces in large molecules. An interesting insight of our ab initio accurate simulations is that proteins seem to be substantially more flexible than previously thought. These molecular fluctuations and associated low-frequency vibrations are expected to strongly contribute to dynamical processes such as in biomolecules (*51*).

While our work focuses exclusively on the study of peptides and proteins, the proposed framework can be applied to any atomic system too large to study with ab initio methods. We find that even small polyalanine peptides display qualitatively different dynamics when simulated with GEMS in comparison to dynamics with conventional FFs. For example, GEMS simulations suggest that the folding of $AceAla_{15}Nme$ from the FES to a helical conformation occurs via short-lived intermediates characterized by hydrogen bonding between peptide groups of adjacent residues. Once a helix is formed, its structure fluctuates between 3_{10}- and α-helices in a dynamical equilibrium. This is in stark contrast to simulations with a conventional FF, where the peptide forms a rigid α-helix without visiting a common intermediate. Future NMR or UV/IR spectroscopy experiments could confirm or disprove our predictions for polyalanine helices. These results are reminiscent of the first MD study of a protein (*52*), which showed that proteins are less rigid than previously thought (*53*). The current findings, already alluded to above, indicate that proteins might be even more flexible, and our simulations of crambin suggest that the general trend observed for peptides in gas phase also holds for proteins in solution. In particular, crambin samples a larger conformational space in GEMS simulations and its backbone dihedral angles have broader distributions. Experiments with a simplified model for the PES suggest that the increased flexibility observed in GEMS simulations is associated with low energy barriers for large conformational changes, whereas with conventional FFs, large fluctuations are always associated with large barriers. This could explain the long-standing disagreement between classical MD and the structural fluctuations of mobile protein groups observed in NMR experiments (*49*). However, structural fluctuations on short timescales are reduced in comparison to simulations with a conventional FF. These observations show that there are qualitative differences in the dynamics of proteins when they are simulated with ab initio quantum-mechanical accuracy.

A promising avenue for future work is to extend GEMS to larger systems and longer timescales, for example, by distributing GEMS simulations over multiple accelerators, which requires nontrivial modifications to the way the MLFF is evaluated. Other possible extensions to GEMS include incorporating nuclear quantum effects, which were demonstrated to substantially change the dynamics of small molecules (*54*). It is likely that similar effects can be observed for larger systems.

Let us discuss some limits of MLFFs when compared to classical MD simulations. Although MLFFs are orders of magnitude more computationally efficient than ab initio calculations, their computational

efficiency is lower than that of conventional FFs (as to be expected). For example, simulating a single timestep of NPT dynamics of crambin in aqueous solution on an NVIDIA A100 GPU with GEMS takes roughly ~500 ms, whereas GROMACS (*55*) only requires ~2 ms for a single time step with a conventional FF on similar hardware. Consequently, at this moment, GEMS simulations are limited to shorter timescales. In addition, evaluating MLFFs usually requires increased memory compared to conventional FFs, limiting the maximum system size that can be simulated with GEMS. Nonetheless, GEMS allows to simulate several nanoseconds of dynamics for systems consisting of thousands of atoms with ab initio accuracy. Furthermore, GEMS like every other MLFF may lead to unphysical dynamics, if not properly trained [see, e.g., (*26*) for a discussion of such phenomena]. As a rule, MLFF simulations should therefore always be subjected to more scrutiny than results from mechanistic FFs. In particular, the resulting trajectories need to be carefully checked for unphysical bond breaking or formation, or otherwise unphysically distorted conformations, which are prevented in traditional FFs by construction. Nevertheless it should be emphasized again that compared to simulations with a conventional FF, GEMS offers highly improved accuracy as well as enables to study chemical transformations such as the making and breaking of chemical bonds and proton transfer processes.

Another advantage of using accurate MLFFs is the availability of arbitrary derivatives—including the potential to obtain alchemical derivatives (*56*). This may enable the optimization of accessible observables, such as docking/binding energies, with respect to local (nearly isosteric) mutations. In a more conventional approach, MLFFs can be used to describe the effects of mutations via thermodynamic integration as regularly performed with classical FFs (*57*, *58*). Given the incorporation of nonlocality in the present methodology, such analyses could naturally account for longer-range phenomena like (static) allosteric effects and the inherent nonadditivity of interactions known to be relevant for the free energy of binding or stability (*59*). In a similar vein, this may allow to perform sensitivity analyses to identify allosteric hotspots and networks, which have also been speculated to play an important role for the evolutionary aspects of proteins and the biomolecular machinery (*60*). Again, we would like to stress that such studies and the above approaches are not limited to biomolecular systems. They may equally well be applied in materials design, for example, studying and optimizing point defects in solid-state systems as relevant to the design of quantum materials.

Finally, we would like to highlight a promising application of GEMS to modeling protein-protein interactions. Figure 9 shows the binding curves of the angiotensin converting enzyme 2 (ACE2) and the receptor binding domain (RBD) of the spike protein of severe acute respiratory syndrome coronavirus 1 (SARS-CoV-1) and SARS-CoV-2 variants using either AmberFF or GEMS (in gas phase). Here, as expected from experimental evidence (*61*), we observe a stronger binding of SARS-CoV-2 for both the classical FF and GEMS. However, GEMS yields a substantially stronger binding by −1.1 eV. Note that the obtained binding energies were computed for static structures in gas phase and do not account for solvation or entropic effects, nor the presence of dynamic loops at the protein surface, so they cannot be directly compared to experimental binding affinities. However, although these results are preliminary and should not be overinterpreted, they indicate the potential importance of ab initio accuracy when studying interactions between complex biological systems. We therefore would like to stress the high promise of GEMS for enabling quantum-mechanical insight in broad application domains across enzyme and protein chemistry or heterogeneous materials. Although top-down fragments in this work are system-specific, in the future, they may be generated to cover a wider range of systems and enable GEMS simulations with a chemically transferable and size-extensive "universal" MLFF.

MATERIALS AND METHODS

Construction of fragment data

Bottom-up fragments

The construction of bottom-up fragments follows an approach similar in spirit to the one described by Huang and von Lilienfeld (*16*). Ignoring hydrogen atoms, increasingly large chemical graphs with the same local bonding patterns as the system of interest are constructed until a maximum number of heavy atoms is reached. This is achieved by starting from graphs consisting of a single heavy atom. Larger graphs are constructed by successively adding additional heavy atoms and pruning graphs, which do not appear as substructures in the original system. Once the graphs are constructed, they are converted to bottom-up fragments by saturating all valencies with hydrogen atoms and generating the corresponding three-dimensional molecular structure [e.g., using Open Babel (*62*)]. For all structures, multiple conformers are sampled using either MD simulations at high temperatures or normal mode sampling (*7*). Here, we use the bottom-up fragments for solvated proteins generated in earlier work (*63*). These bottom-up fragments also contain micro-solvated structures with explicit water molecules and structures for bulk water [see (*19*) for details]. Since all similar local structures are covered by the same graphs, the bottom-up fragments optimally exploit any structural redundancies, often resulting in a surprisingly small number of fragments. For example, just 2307 chemical graphs (with a maximum of eight heavy atoms) are sufficient to cover all local bonding patterns appearing in proteins consisting of the 20 natural amino acids, even when considering different protonation states and the possibility of disulfide bridges (*19*).

Top-down fragments

Starting from an MD snapshot of the system of interest (sampled from conventional MD simulations, see below and section S3 for more details), all atoms outside a spherical region around a central atom are deleted. The cutoff radius for selecting the spherical region should be chosen as large as possible, but still resulting in fragment sizes for which reference calculations are feasible (here, we choose 8 Å). Then, any resulting dangling single bonds on heavy atoms are saturated with hydrogen atoms. Valencies situated on hydrogen atoms or corresponding to double bonds are eliminated by including the bonded atom in the original system (outside the cutoff). This process is repeated until all valencies are saturated (*27*). By choosing different central atoms, several (partially overlapping) top-down fragments can be constructed from a single configuration of the original system. Since the snapshots from which top-down fragments are constructed are sampled by MD simulations with a conventional FF, structures that are either not well described by the chosen FF or not visited during the dynamics may potentially introduce bias toward certain "interaction motifs" into the training data. Although this is partially alleviated by a thorough sampling of bottom-up fragments, care should be taken that the conventional FF used for sampling is well parametrized. Especially for nonbiological systems, where widely used FF parametrizations are much less common, it might become necessary to sample MD snapshots with, e.g., semi-empirical methods or derive top-down fragments from structures sampled in other ways.

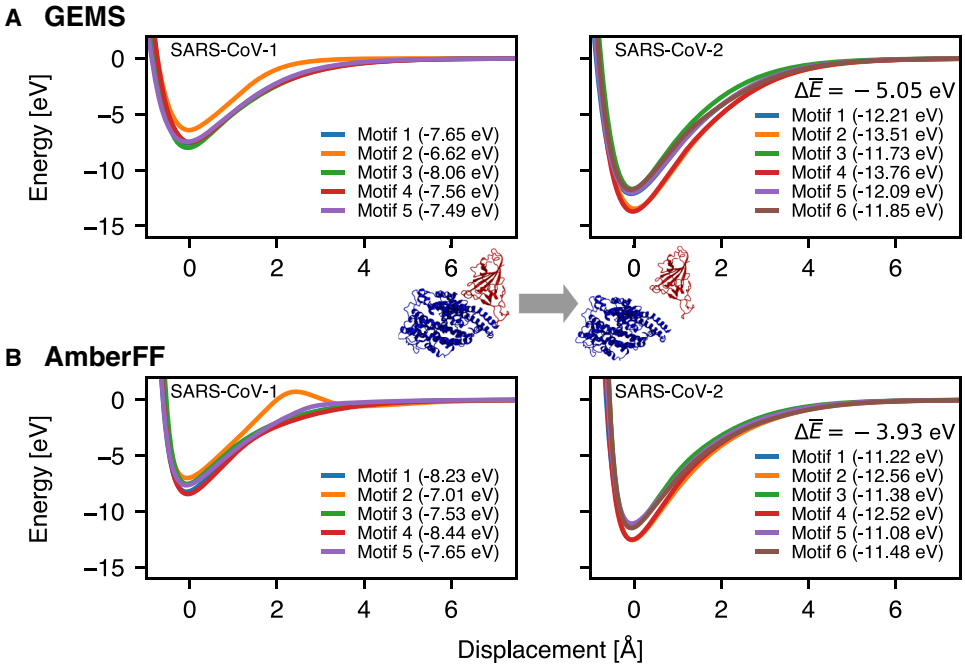

Fig. 9. Toward accurate quantum-mechanical protein-protein interactions: Gas-phase binding curves of ACE2 (blue) and RBD of the SARS-CoV spike protein (red). The ACE2 and RBD proteins are displaced along the line connecting their centers of mass relative to their equilibrium position in solution (computed with the AmberFF), keeping their internal structure fixed. Different binding motifs [taken from (*74*)] are distinguished (values in brackets are the maximum well depth for the corresponding motif). All energy values are referenced with respect to infinite separation of ACE2 and RBD. The displayed value $\Delta\bar{E}$ gives the difference in well depth (averaged over all binding motifs) between SARS-CoV-2 and SARS-CoV-1. The $\Delta\Delta\bar{E}$ between AmberFF (**B**) and GEMS (**A**) is −1.11 eV.

DFT calculations

DFT reference calculations were performed using the Psi4 software package (*64*) at the PBE0/def2-TZVPP+MBD (*28*, *29*, *44*) level of theory on Google Cloud Platform (GCP). Each fragment was run on an independent Docker container within a cloud compute engine virtual machine. We mostly used n2d-higmem-4 and n2-highmem-4 virtual machine instances with four cores, 32 GB RAM and 768 GB of disk space each, with some larger fragments being manually relaunched on higher-memory machines if they crashed with out-of-memory errors. Execution was parallelized on up to 20,000 CPU cores. Calculations where shut down if they did not complete within 21 days, which was the case for a few outliers, but median execution time per fragment was ~48 hours. For example, of the 2292 crambin fragments, 5% (120) did not finish successfully on an n2-highmem-4 machine, due to machine errors, lack of memory, or because the fragment failed to converge to a meaningful solution. The rest (2172 fragments) all finished within a week, with a median runtime of 47.4 hours (mean: 50.7 hours) (see fig. S16 for the runtime distribution). Only 33 fragments needed more than 4 days of compute to complete. In total, the successful runs required approximately 110,000 compute hours.

Training the MLF

All MLFFs in this work use the recently proposed SpookyNet architecture [see (*17*) for details]. We use three different trained ML models here: one for the simulations of all polyalanine systems, one for the simulation of crambin in aqueous solution, and one for the gas-phase ACE2/SARS-CoV-1/2 RBD binding curves shown in Fig. 9 (CoV model). The polyalanine and CoV models use the recommended architectural hyperparameters of $T = 6$ interaction modules and $F = 128$ features (*17*). Because of hardware limitations when performing MD simulations for thousands of atoms, the crambin model uses $T = 3$ interaction modules and $F = 64$ features to reduce memory requirements. All models use a short-range cutoff of $r_{\mathrm{srcut}} = 10\ a_0$ (~5.29 Å). The crambin and CoV models additionally use a long-range cutoff of $r_{\mathrm{lrcut}} = 20\ a_0$ (~10.58 Å) for the computation of the analytical electrostatic and dispersion correction terms included in the SpookyNet energy prediction (to achieve sub-quadratic scaling with respect to the number of atoms). We follow the training protocol described in (*17*) for fitting the parameters to reference energies, forces, and dipole moments; however, the mean squared loss function was replaced by the adaptive robust loss described in (*65*). All models were trained on single NVIDIA V100 GPUs on GCP using the same 2,713,986 bottom-up fragments, and 45,948 (for the polyalanine model), 5624 (for the crambin model), or 129,942 (for the CoV model) top-down fragments. Typical training times are between 1 and 2 weeks, depending on the system. During training, structures were randomly drawn in equal amounts from bottom-up and top-down fragments, i.e., top-down fragments were oversampled to mitigate the imbalance in the numbers of bottom-up/top-down fragments.

MD simulations

Conventional FF

All classical MD simulations have been performed with the GROMACS 2020.3 software package using NVIDIA V100 or A100 GPUs in a Kubernetes system on GCP. Throughout this work, we have used the AMBER99SB-ILDN FF (*24*) for the conventional MD simulations. Standard amino acid definitions have been adapted to accommodate charged Lys + H^+ termini in accordance with the AMBER99SB-ILDN

parametrizations where needed. In the MD simulations of ACE2/SARS-CoV-2 RBD, the binding of the Zn^{2+} cofactor in ACE2 has been described via harmonic restraints to the experimentally determined ligands to avoid potential shortcomings in the description of the metal-ligand interaction. All solvated systems presented in this article or used for sampling representative structures for generating top-down fragments were initially resolvated, optimized to a maximum atomic force of 1000 kJ mol^{-1} nm^{-1}, and equilibrated according to the protocol detailed in section S2. Simulations for studying nonequilibrium processes (i.e., the gas-phase folding/unfolding of polyalanine systems) have been started directly from optimized structures with velocities drawn from a Maxwell-Boltzmann distribution at twice the simulation temperature [such that the average kinetic energy during the simulation corresponds to the desired temperature (*66*)]. The gas-phase simulations have thereby been realized in a pseudo–gas-phase setting as proposed and validated in (*66*). All constant temperature MD simulations have been performed using temperature coupling via stochastic velocity rescaling (*67*), and a Parrinello-Rahman barostat (*68*) has been used for NPT simulations. To speed up computations, standard MD simulations involved the commonly used constraint of bonds involving hydrogen with a time step of 2 fs, while the power spectra reported in this work have been obtained from fully unconstrained simulations with a time step of 0.5 fs. The starting structures of polyalanine systems have been generated with the Avogadro software (*69*), and the initial structure of crambin has been taken from Protein Data Bank (PDB) entry 2FD7 (*70*) (resolution, 1.75 Å) of a chemically synthesized mutant of crambin, where we used PyMOL (*71*) to remodel mutated residues to match the wild-type sequence (SER11 and VAL15). This starting structure was chosen because of its favorable validation metrics (e.g., clash score). For completeness, we compared the structure of crambin (after solvation and subsequent minimization with GROMACS) when choosing higher-resolution entries from the PDB [1EJG (*72*) and 3NIR (*73*)] as initial structure instead. The obtained structures are virtually identical with RMSDs below 1 Å (see fig. S14). Our simulations of the ACE2/SARS-CoV-1/2 RBD complex have been initiated from a set of representative conformations as identified in (*74*) or pointwise mutations thereof. Currently available experimental results on the mutations present in the β, γ, δ, and ϵ variants of SARS-CoV-2 do not indicate considerable structural changes to the spike RBD. After partial relaxation, simple pointwise mutations of the structural representatives obtained for the α-variant can thus be assumed to represent viable starting points for MD simulations of the different variants.

GEMS

All MD simulations with the GEMS method were performed using the SchNetPack (*75*) MD toolbox with a timestep of 0.5 fs and without any bond constraints. Simulations for polyalanine systems were performed on NVIDIA V100 GPUs on GCP, whereas crambin simulations were performed on NVIDIA A100 GPUs with 80 GB. To mimic experimental conditions (*37*), the simulations of $AceAla_{15}Lys + H^+$ helix stability were performed in the NVE ensemble starting from an optimized structure with initial velocities drawn from a Maxwell-Boltzmann distribution at twice the simulation temperature as explained above. The folding simulations of $AceAla_{15}Nme$ were performed in the NVT ensemble at 300 K starting from the optimized FES using the same method to assign initial velocities. Simulations of crambin in aqueous solution were performed in a simulation box with 8205 explicit water molecules in the NPT ensemble at a temperature of 300 K and a pressure of 1.01325 bar, starting from an optimized structure and initial velocities drawn from a Maxwell-Boltzmann distribution according to the simulation temperature (the first 1 ns of dynamics was discarded to allow the system to equilibrate). Constant temperature and/or pressure simulations use the Nosé-Hoover chain thermostat/barostat (*76*) implemented in SchNetPack using a chain length of 3. Note that simulations in aqueous solution with GEMS use a single MLFF to describe all (solute-solvent and solvent-solvent) interactions in a unified manner.

Comparison to experimental terahertz spectra of (*48*) (see Fig. 7)

Four MD simulations of crambin in aqueous solution (see above) of 500,000 frames each, sampled every 2.5 fs, were collected. The frames were stripped of water and aligned to calculate a mass-weighted covariance matrix. The eigenvectors of the covariance matrix were taken to indicate resonant vibrations of the molecule according to a quasi-normal mode approximation, and the eigenvalues were used to determine resonant frequencies. An IR spectrum was calculated by calculating a dipole moment for the average structure when modified by each eigenmode displacement. The scale of each displacement was chosen proportional to the inverse frequency. For better correspondence to the experimental IR spectra, which were collected in a partially solvated environment, solvent screening to a cutoff range of 3.3 Å was added for each displaced structure using the 3DRISM liquid state theory (*77*). Inclusion of solvent effects to either greater or lesser range was found to obscure the features of the resulting spectrum. The presented spectra are calculated as a sum of Gaussian peaks with the arbitrary width 4 cm^{-1} and heights assigned as the magnitude of the calculated dipole moment of the fluctuation. Classical MD simulations using the AMBER FF followed the same procedure, except that two longer simulations with 10 million frames each (sampled every 10 ps) were used.

Simplified model for crambin dynamics

Langevin dynamics for the toy model (see Eq. 1) were computed using the integrator proposed in (*78*) at a temperature of 1 a.u. (arbitrary units) with a timestep of 0.01 a.u. for a total duration of 10 000 a.u. and a friction coefficient $\gamma \approx 5.13$ a.u. (corresponding to 5% stochastic motion). To simulate a "GEMS-like" trajectory, barrier heights and separations were chosen as $h_i \sim \mathcal{U}(0.1,5.0)$ a.u. and $a_i \sim \mathcal{U}(0.1,10.0)$ a.u. For the "AmberFF-like" trajectory, the parameters for 90% of the modes were chose as $h_i \sim \mathcal{U}(0.1,1.25)$ a.u. and $a_i \sim \mathcal{U}(0.1,2.5)$ a.u., whereas for the remaining 10% of modes, they were chosen as $h_i \sim \mathcal{U}(3.75,5.00)$ a.u. and $a_i \sim \mathcal{U}(7.5,10.0)$ a.u. We find that similar results can be obtained for a range of parameter values, as long as there is a clear separation of large conformational changes with large energy barriers and small conformational changes with small energy barriers (to produce AmberFF-like trajectories), or no such separation (to produce GEMS-like trajectories).

Supplementary Materials

This PDF file includes:

Sections S1 to S6
Figs. S1 to S26
Legends for movies S1 and S2
References

REFERENCES AND NOTES

1. M. Karplus, J. A. McCammon, Molecular dynamics simulations of biomolecules. *Nat. Struct. Biol.* **9**, 646–652 (2002).
2. F. Mouvet, J. Villard, V. Bolnykh, U. Röthlisberger, Recent advances in first-principles based molecular dynamics. *Acc. Chem. Res.* **55**, 221–230 (2022).
3. R. H. French, V. A. Parsegian, R. Podgornik, R. F. Rajter, A. Jagota, J. Luo, D. Asthagiri, M. K. Chaudhury, Y.-M. Chiang, S. Granick, Long range interactions in nanoscale science. *Rev. Mod. Phys.* **82**, 1887–1944 (2010).
4. D. E. Shaw, M. M. Deneroff, R. O. Dror, J. S. Kuskin, R. H. Larson, J. K. Salmon, C. Young, B. Batson, K. J. Bowers, J. C. Chao, M. P. Eastwood, J. Gagliardo, J. P. Grossman, C. R. Ho, D. J. Ierardi, I. Kolossváry, J. L. Klepeis, T. Layman, C. McLeavey, M. A. Moraes, R. Mueller, E. C. Priest, Y. Shan, J. Spengler, M. Theobald, B. Towles, S. C. Wang, Anton, a special-purpose machine for molecular dynamics simulation. *Commun. ACM* **51**, 91–97 (2008).
5. H. M. Senn, W. Thiel, QM/MM methods for biomolecular systems. *Angew. Chem. Int.* **48**, 1198–1229 (2009).
6. H. J. Kulik, J. Zhang, J. P. Klinman, T. J. Martinez, How large should the QM region be in QM/MM calculations? The case of catechol O-methyltransferase. *J. Phys. Chem. B* **120**, 11381–11394 (2016).
7. O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko, K.-R. Müller, Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
8. U. Rivero, O. T. Unke, M. Meuwly, S. Willitsch, Reactive atomistic simulations of Diels-Alder reactions: The importance of molecular rotations. *J. Chem. Phys.* **151**, 104301 (2019).
9. H. E. Sauceda, S. Chmiela, I. Poltavsky, K.-R. Müller, A. Tkatchenko, Molecular force fields with gradient-domain machine learning: Construction and application to dynamics of small molecules with coupled cluster forces. *J. Chem. Phys.* **150**, 114102 (2019).
10. S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko, K.-R. Müller, Accurate global machine learning force fields for molecules with hundreds of atoms. *Sci. Adv.* **9**, eadf0873 (2023).
11. W. Jia, H. Wang, M. Chen, D. Lu, L. Lin, R. Car, E. Weinan, L. Zhang, *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis* (IEEE, 2020), pp. 1–14.
12. M. Rossi, W. Fang, A. Michaelides, Stability of complex biomolecular structures: Van der Waals, hydrogen bond cooperativity, and nuclear quantum effects. *J. Phys. Chem. Lett.* **6**, 4233–4238 (2015).
13. M. Stöhr, A. Tkatchenko, Quantum mechanics of proteins in explicit water: The role of plasmon-like solute-solvent interactions. *Sci. Adv.* **5**, eaax0024 (2019).
14. Z. Cheng, J. Du, L. Zhang, J. Ma, W. Li, S. Li, Building quantum mechanics quality force fields of proteins with the generalized energy-based fragmentation approach and machine learning. *Phys. Chem. Chem. Phys.* **24**, 1326–1337 (2022).
15. Y. Han, Z. Wang, A. Chen, I. Ali, J. Cai, S. Ye, J. Li, An inductive transfer learning force field (ITLFF) protocol builds protein force fields in seconds. *Brief. Bioinform.* **23**, bbab590 (2022).
16. B. Huang, O. A. von Lilienfeld, Quantum machine learning using atom-in-molecule-based fragments selected on the fly. *Nat. Chem.* **12**, 945–951 (2020).
17. O. T. Unke, S. Chmiela, M. Gastegger, K. T. Schütt, H. E. Sauceda, K.-R. Müller, SpookyNet: Learning force fields with electronic degrees of freedom and nonlocal effects. *Nat. Commun.* **12**, 7273 (2021).
18. J. Behler, Representing potential energy surfaces by high-dimensional neural network potentials. *J. Condens. Matter Phys.* **26**, 183001 (2014).
19. O. T. Unke, M. Meuwly, PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
20. A. Grisafi, M. Ceriotti, Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **151**, 204105 (2019).
21. L. Zhang, H. Wang, M. C. Muniz, A. Z. Panagiotopoulos, R. Car, W. E, A deep potential model with long-range electrostatic interactions. *J. Chem. Phys.* **156**, 124107 (2022).
22. T. W. Ko, J. A. Finkler, S. Goedecker, J. Behler, A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 1–11 (2021).
23. E. Balog, J. C. Smith, D. Perahia, Conformational heterogeneity and low-frequency vibrational modes of proteins. *Phys. Chem. Chem. Phys.* **8**, 5543–5548 (2006).
24. K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, D. E. Shaw, Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* **78**, 1950–1958 (2010).
25. M. Karplus, T. Ichiye, B. Pettitt, Configurational entropy of native proteins. *Biophys. J.* **52**, 1083–1085 (1987).
26. S. Stocker, J. Gasteiger, F. Becker, S. Günnemann, J. Margraf, How robust are modern graph neural network potentials in long and hot molecular dynamics simulations? *Mach. Learn. Sci. Technol.* **3**, 045010 (2022).
27. M. Gastegger, J. Behler, P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
28. C. Adamo, V. Barone, Toward reliable density functional methods without adjustable parameters: The PBE0 model. *J. Chem. Phys.* **110**, 6158–6170 (1999).
29. A. Tkatchenko, R. A. DiStasio Jr., R. Car, M. Scheffler, Accurate and efficient method for many-body van der Waals interactions. *Phys. Rev. Lett.* **108**, 236402 (2012).
30. F. Schubert, M. Rossi, C. Baldauf, K. Pagel, S. Warnke, G. von Helden, F. Filsinger, P. Kupser, G. Meijer, M. Salwiczek, B. Koksch, M. Scheffler, V. Blum, Exploring the conformational preferences of 20-residue peptides in isolation: Ac-Ala$_{19}$-Lys + H^+ vs. Ac-Lys-Ala$_{19}$ + H^+ and the current reach of DFT. *Phys. Chem. Chem. Phys.* **17**, 7373–7385 (2015).
31. C. Baldauf, M. Rossi, Going clean: Structure and dynamics of peptides in the gas phase and paths to solvation. *J. Phys. Condens. Matter* **27**, 493002 (2015).
32. J. Hermann, D. Alfè, A. Tkatchenko, Nanoscale π-π stacked molecules are bound by collective charge fluctuations. *Nat. Commun.* **8**, 14052 (2017).
33. J. Hoja, H.-Y. Ko, M. A. Neumann, R. Car, R. A. DiStasio, A. Tkatchenko, Reliable and practical computational description of molecular crystal polymorphs. *Sci. Adv.* **5**, eaau3338 (2019).
34. D. Firaha, Y. M. Liu, J. van de Streek, K. Sasikumar, H. Dietrich, J. Helfferich, L. Aerts, D. E. Braun, A. Broo, A. G. DiPasquale, A. Y. Lee, S. le Meur, S. O. Nilsson Lill, W. J. Lunsmann, A. Mattei, P. Muglia, O. D. Putra, M. Raoui, S. M. Reutzel-Edens, S. Rome, A. Y. Sheikh, A. Tkatchenko, G. R. Woollam, M. A. Neumann, Predicting crystal form stability under real-world conditions. *Nature* **623**, 324–328 (2023).
35. R. A. DiStasio Jr., B. Santra, Z. Li, X. Wu, R. Car, The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water. *J. Chem. Phys.* **141**, 084502 (2014).
36. C. Park, W. A. Goddard, Stabilization of α-helices by dipole−dipole interactions within α-helices. *J. Phys. Chem. B* **104**, 7784–7789 (2000).
37. M. Kohtani, T. C. Jones, J. E. Schneider, M. F. Jarrold, Extreme stability of an unsolvated α-helix. *J. Am. Chem. Soc.* **126**, 7420–7421 (2004).
38. A. Tkatchenko, M. Rossi, V. Blum, J. Ireta, M. Scheffler, Unraveling the stability of polypeptide helices: Critical role of van der Waals interactions. *Phys. Rev. Lett.* **106**, 118102 (2011).
39. I. A. Topol, S. K. Burt, E. Deretey, T.-H. Tang, A. Perczel, A. Rashin, I. G. Csizmadia, α- and 3_{10}-Helix Interconversion: A quantum-chemical study on polyalanine systems in the gas phase and in aqueous solvent. *J. Am. Chem. Soc.* **123**, 6054–6060 (2001).
40. G. L. Millhauser, C. J. Stenland, P. Hanson, K. A. Bolin, F. J. van de Ven, Estimating the relative populations of 3_{10}-helix and α-helix in Ala-rich peptides: A hydrogen exchange and high field NMR study. *J. Mol. Biol.* **267**, 963–974 (1997).
41. K. A. Bolin, G. L. Millhauser, α and 3_{10}: The split personality of polypeptide helices. *Acc. Chem. Res.* **32**, 1027–1033 (1999).
42. N. Foloppe, A. D. MacKerell Jr., All-atom empirical force field for nucleic acids: I. parameter optimization based on small molecule and condensed phase macromolecular target data. *J. Comput. Chem.* **21**, 86–104 (2000).
43. C. Oostenbrink, A. Villa, A. E. Mark, W. F. Van Gunsteren, A biomolecular force field based on the free enthalpy of hydration and solvation: The GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* **25**, 1656–1676 (2004).
44. F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.* **7**, 3297–3305 (2005).
45. M. Thomas, M. Brehm, R. Fligg, P. Vöhringer, B. Kirchner, Computing vibrational spectra from *ab initio* molecular dynamics. *Phys. Chem. Chem. Phys.* **15**, 6608–6622 (2013).
46. C. D. Craver, *The Coblentz Society Desk Book of Infrared Spectra* (National Standard Reference Data System, 1977).
47. H.-C. Ahn, N. Juranić, S. Macura, J. L. Markley, Three-dimensional structure of the water-insoluble protein crambin in dodecylphosphocholine micelles and its minimal solvent-exposed surface. *J. Am. Chem. Soc.* **128**, 4398–4404 (2006).
48. K. N. Woods, The glassy state of crambin and the THz time scale protein-solvent fluctuations possibly related to protein function. *BMC Biophys.* **7**, 1–15 (2014).
49. D. A. Case, Molecular dynamics and NMR spin relaxation in proteins. *Acc. Chem. Res.* **35**, 325–331 (2002).
50. S. Piana, K. Lindorff-Larsen, D. E. Shaw, How robust are protein folding simulations with respect to force field parameterization? *Biophys. J.* **100**, L47–L49 (2011).
51. R. M. Levy, D. Perahia, M. Karplus, Molecular dynamics of an α-helical polypeptide: Temperature dependence and deviation from harmonic behavior. *Proc. Natl. Acad. Sci. U.S.A.* **79**, 1346–1350 (1982).
52. J. A. McCammon, B. R. Gelin, M. Karplus, Dynamics of folded proteins. *Nature* **267**, 585–590 (1977).
53. D. Phillips, *Biomolecular Stereodynamics* (Adenine Press, 1981).
54. H. E. Sauceda, V. Vassilev-Galindo, S. Chmiela, K.-R. Müller, A. Tkatchenko, Dynamical strengthening of covalent and non-covalent molecular interactions by nuclear quantum effects at finite temperature. *Nat. Commun.* **12**, 442 (2021).
55. B. Hess, C. Kutzner, D. Van Der Spoel, E. Lindahl, GROMACS 4: Algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* **4**, 435–447 (2008).

56. K. Saravanan, J. R. Kitchin, O. A. Von Lilienfeld, J. A. Keith, Alchemical predictions for computational catalysis: Potential and limitations. *J. Phys. Chem. Lett.* **8**, 5002–5007 (2017).
57. F. R. Beierlein, G. G. Kneale, T. Clark, Predicting the effects of basepair mutations in DNA-protein complexes by thermodynamic integration. *Biophys. J.* **101**, 1130–1138 (2011).
58. M. Krepl, M. Otyepka, P. Banáš, J. Šponer, Effect of guanine to inosine substitution on stability of canonical DNA and RNA duplexes: Molecular dynamics thermodynamics integration study. *J. Phys. Chem. B* **117**, 1872–1879 (2013).
59. E. Di Cera, Site-specific thermodynamics: Understanding cooperativity in molecular recognition. *Chem. Rev.* **98**, 1563–1592 (1998).
60. A. S. Raman, K. I. White, R. Ranganathan, Origins of allostery and evolvability in proteins: A case study. *Cell* **166**, 468–480 (2016).
61. D. Wrapp, N. Wang, K. S. Corbett, J. A. Goldsmith, C.-L. Hsieh, O. Abiona, B. S. Graham, J. S. McLellan, Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
62. N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, G. R. Hutchison, Open Babel: An open chemical toolbox. *J. Chem.* **3**, 1–14 (2011).
63. O. Unke, M. Meuwly, Solvated protein fragments data set (2019); https://doi.org/10.5281/zenodo.2605371.
64. R. M. Parrish, L. A. Burns, D. G. A. Smith, A. C. Simmonett, A. E. DePrince, E. G. Hohenstein, U. Bozkaya, A. Y. Sokolov, R. Di Remigio, R. M. Richard, J. F. Gonthier, A. M. James, H. R. McAlexander, A. Kumar, M. Saitow, X. Wang, B. P. Pritchard, P. Verma, H. F. Schaefer, K. Patkowski, R. A. King, E. F. Valeev, F. A. Evangelista, J. M. Turney, T. D. Crawford, C. D. Sherrill, Psi4 1.1: An open-source electronic structure program emphasizing automation, advanced libraries, and interoperability. *J. Chem. Theory Comput.* **13**, 3185–3197 (2017).
65. J. T. Barron, A general and adaptive robust loss function. *Proc. IEEE Int. Conf. Comput. Vis.*, 4331–4339 (2019).
66. L. Konermann, H. Metwally, R. G. McAllister, V. Popa, How to run molecular dynamics simulations on electrospray droplets and gas phase proteins: Basic guidelines and selected applications. *Methods* **144**, 104–112 (2018).
67. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).
68. M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method. *J. Appl. Phys.* **52**, 7182–7190 (1981).
69. M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek, G. R. Hutchison, Avogadro: An advanced semantic chemical editor, visualization, and analysis platform. *J. Chem.* **4**, 1–17 (2012).
70. D. Bang, V. Tereshko, A. A. Kossiakoff, S. B. Kent, Role of a salt bridge in the model protein crambin explored by chemical protein synthesis: X-ray structure of a unique protein analogue, [V15A]crambin-α-carboxamide. *Mol. Biosyst.* **5**, 750–756 (2009).
71. W. L. DeLano, PyMOL: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr.* **40**, 82–92 (2002).
72. C. Jelsch, M. M. Teeter, V. Lamzin, V. Pichon-Pesme, R. H. Blessing, C. Lecomte, Accurate protein crystallography at ultra-high resolution: Valence electron distribution in crambin. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 3171–3176 (2000).
73. A. Schmidt, M. Teeter, E. Weckert, V. S. Lamzin, Crystal structure of small protein crambin at 0.48 Å resolution. *Acta Crystallogr. F* **67**, 424–428 (2011).
74. J. M. Delgado, N. Duro, D. M. Rogers, A. Tkatchenko, S. A. Pandit, S. Varma, Molecular basis for higher affinity of SARS-CoV-2 spike RBD for human ACE2 receptor. *Proteins* **89**, 1134–1144 (2021).
75. K. Schütt, P. Kessel, M. Gastegger, K. Nicoli, A. Tkatchenko, K.-R. Müller, SchNetPack: A deep learning toolbox for atomistic systems. *J. Chem. Theory Comput.* **15**, 448–455 (2018).
76. G. J. Martyna, M. E. Tuckerman, D. J. Tobias, M. L. Klein, Explicit reversible integrators for extended systems dynamics. *Mol. Phys.* **87**, 1117–1157 (1996).
77. L. Wilson, R. Krasny, T. Luchko, Accelerating the 3d reference interaction site model theory of molecular solvation with treecode summation and cut-offs. *J. Comput. Chem.* **43**, 1251–1270 (2022).
78. G. Bussi, M. Parrinello, Accurate sampling using Langevin dynamics. *Phys. Rev. B* **75**, 056707 (2007).
79. D. Frishman, P. Argos, Knowledge-based protein secondary structure assignment. *Proteins* **23**, 566–579 (1995).
80. L. McInnes, J. Healy, J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction. arXiv:1802.03426 [stat.ML] (2018).
81. P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, K. Schulten, Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure* **14**, 437–449 (2006).
82. G. Zhao, J. R. Perilla, E. L. Yufenyuy, X. Meng, B. Chen, J. Ning, J. Ahn, A. M. Gronenborn, K. Schulten, C. Aiken, P. Zhang, Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics. *Nature* **497**, 643–646 (2013).
83. M. I. Zimmerman, J. R. Porter, M. D. Ward, S. Singh, N. Vithani, A. Meller, U. L. Mallimadugula, C. E. Kuhn, J. H. Borowsky, R. P. Wiewiora, Citizen scientists create an exascale computer to combat COVID-19. bioRxiv 2020.06.27.175430 [Preprint] (2020). https://doi.org/10.1101/2020.06.27.175430.
84. J. E. Lennard-Jones, On the determination of molecular fields—II. From the equation of state of a gas. *Proc. R. Soc. Lond. A* **106**, 463–477 (1924).
85. M. González, *École Thématique de la Société Française de la Neutronique* (EDP Sciences, 2011), vol. 12, pp. 169–200.
86. O. T. Unke, D. Koner, S. Patra, S. Käser, M. Meuwly, High-dimensional potential energy surfaces for molecular simulations: From empiricism to machine learning. *Mach. Learn. Sci. Technol.* **1**, 13001 (2020).
87. F. Vitalini, A. S. Mey, F. Noé, B. G. Keller, Dynamic properties of force fields. *J. Chem. Phys.* **142**, 02B611_1 (2015).
88. T. A. Halgren, W. Damm, Polarizable force fields. *Curr. Opin. Struct. Biol.* **11**, 236–242 (2001).
89. A. Warshel, M. Kato, A. V. Pisliakov, Polarizable force fields: History, test cases, and prospects. *J. Chem. Theory Comput.* **3**, 2034–2045 (2007).
90. T. D. Rasmussen, P. Ren, J. W. Ponder, F. Jensen, Force field modeling of conformational energies: Importance of multipole moments and intramolecular polarization. *Int. J. Quantum Chem.* **107**, 1390–1395 (2007).
91. M. G. Darley, C. M. Handley, P. L. Popelier, Beyond point charges: Dynamic polarization from neural net predicted multipole moments. *J. Chem. Theory Comput.* **4**, 1435–1448 (2008).
92. S. M. Kandathil, T. L. Fletcher, Y. Yuan, J. Knowles, P. L. Popelier, Accuracy and tractability of a Kriging model of intramolecular polarizable multipolar electrostatics and its application to histidine. *J. Comput. Chem.* **34**, 1850–1861 (2013).
93. S. Cardamone, T. J. Hughes, P. L. Popelier, Multipolar electrostatics. *Phys. Chem. Chem. Phys.* **16**, 10367–10387 (2014).
94. O. T. Unke, M. Devereux, M. Meuwly, Minimal distributed charges: Multipolar quality at the cost of point charge electrostatics. *J. Chem. Phys.* **147**, 161712 (2017).
95. A. Warshel, R. M. Weiss, An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.* **102**, 6218–6226 (1980).
96. A. C. Van Duin, S. Dasgupta, F. Lorant, W. A. Goddard, ReaxFF: A reactive force field for hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409 (2001).
97. J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
98. P. L. Popelier, QCTFF: On the construction of a novel protein force field. *Int. J. Quantum Chem.* **115**, 1005–1011 (2015).
99. J. S. Smith, O. Isayev, A. E. Roitberg, ANI-1: An extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
100. J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed.* **56**, 12828–12840 (2017).
101. J. Behler, Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
102. S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, K.-R. Müller, Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
103. S. Chmiela, H. E. Sauceda, K.-R. Müller, A. Tkatchenko, Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).
104. K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller, Machine learning meets quantum physics, in *Lecture Notes in Physics* (Springer, 2020).
105. M. Rupp, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
106. G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, O. A. Von Lilienfeld, Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
107. K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, K.-R. Müller, Assessment and validation of machine learning methods for predicting molecular atomization energies. *J. Chem. Theory Comput.* **9**, 3404–3419 (2013).
108. K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Muller, A. Tkatchenko, Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *J. Phys. Chem. Lett.* **6**, 2326–2331 (2015).
109. L. Boninsegna, F. Nüske, C. Clementi, Sparse learning of stochastic dynamical equations. *J. Chem. Phys.* **148**, 241723 (2018).
110. F. Noé, S. Olsson, J. Köhler, H. Wu, Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science* **365**, eaaw1147 (2019).
111. J. Köhler, L. Klein, F. Noé, Equivariant flows: Exact likelihood generative learning for symmetric densities. arXiv:2006.02425 [stat.ML] (2020).

112. J. Zhang, Y. I. Yang, F. Noé, Targeted adversarial learning optimized sampling. *J. Phys. Chem. Lett.* **10**, 5791–5797 (2019).
113. D. Koner, O. T. Unke, K. Boe, R. J. Bemish, M. Meuwly, Exhaustive state-to-state cross sections for reactive molecular collisions from importance sampling simulation and a neural network representation. *J. Chem. Phys.* **150**, 211101 (2019).
114. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Zidek, A. W. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
115. J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Zidek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, D. Hassabis, Highly accurate protein structure prediction with AlphaFold. *Nature*, *596* , 583–589 (2021).
116. K. Tunyasuvunakool, J. Adler, Z. Wu, T. Green, M. Zielinski, A. Zidek, A. Bridgland, A. Cowie, C. Meyer, A. Laydon, S. Velankar, G. J. Kleywegt, A. Bateman, R. Evans, A. Pritzel, M. Figurnov, O. Ronneberger, R. Bates, S. A. A. Kohl, A. Potapenko, A. J. Ballard, B. Romera-Paredes, S. Nikolov, R. Jain, E. Clancy, D. Reimann, S. Petersen, A. W. Senior, K. Kavukcuoglu, E. Birney, P. Kohli, J. Jumper, D. Hassabis, Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
117. G. Carleo, M. Troyer, Solving the quantum many-body problem with artificial neural networks. *Science* **355**, 602–606 (2017).
118. D. Pfau, J. S. Spencer, A. G. Matthews, W. M. C. Foulkes, *Ab initio* solution of the many-electron Schrödinger equation with deep neural networks. *Phys. Rev. Res.* **2**, 033429 (2020).
119. J. Hermann, Z. Schätzle, F. Noé, Deep-neural-network solution of the electronic Schrödinger equation. *Nat. Chem.* **12**, 891–897 (2020).
120. K. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, R. J. Maurer, Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nat. Commun.* **10**, 5024 (2019).
121. M. Gastegger, A. McSloy, M. Luya, K. Schütt, R. Maurer, A deep neural network for molecular wave functions in quasi-atomic minimal basis representation. *J. Chem. Phys.* **153**, 044123 (2020).
122. O. Unke, M. Bogojeski, M. Gastegger, M. Geiger, T. Smidt, K.-R. Müller, SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *Adv. Neur. Inform. Process. Syst.* **34**, 1 (2021).
123. M. Gastegger, K. T. Schütt, K.-R. Müller, Machine learning of solvent effects on molecular spectra and reactions. *Chem. Sci.* **12**, 11473–11483 (2021).
124. M. Popova, O. Isayev, A. Tropsha, Deep reinforcement learning for *de novo* drug design. *Sci. Adv.* **4**, eaap7885 (2018).
125. M. Popova, M. Shvets, J. Oliva, O. Isayev, MolecularRNN: Generating realistic molecular graphs with optimized properties. arXiv:1905.13372 [cs.LG] (2019).
126. N. W. Gebauer, M. Gastegger, K. T. Schütt, *NeurIPS 2018 Workshop on Machine Learning for Molecules and Materials* (2018).
127. N. Gebauer, M. Gastegger, K. Schütt, Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Adv. Neural. Inf. Process. Syst.*, 7566–7578 (2019).
128. M. Hoffmann, F. Noé, Generating valid Euclidean distance matrices. arXiv:1910.03131 [cs.LG] (2019).
129. R. Winter, F. Montanari, A. Steffen, H. Briem, F. Noé, D.-A. Clevert, Efficient multi-objective molecular optimization in a continuous latent space. *Chem. Sci.* **10**, 8016–8024 (2019).
130. N. W. Gebauer, M. Gastegger, S. S. Hessmann, K.-R. Müller, K. T. Schütt, Inverse design of 3d molecular structures with conditional generative neural networks. *Nat. Commun.* **13**, 973 (2022).
131. F. Strieth-Kalthoff, F. Sandfort, M. H. Segler, F. Glorius, Machine learning the ropes: Principles, applications and directions in synthetic chemistry. *Chem. Soc. Rev.* **49**, 6154–6168 (2020).
132. A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K. A. Persson, Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
133. C. C. Fischer, K. J. Tibbetts, D. Morgan, G. Ceder, Predicting crystal structure by merging data mining with quantum mechanics. *Nat. Mater.* **5**, 641–646 (2006).
134. D. P. Tabor, L. M. Roch, S. K. Saikin, C. Kreisbeck, D. Sheberla, J. H. Montoya, S. Dwaraknath, M. Aykol, C. Ortiz, H. Tribukait, C. Amador-Bedolla, C. J. Brabec, B. Maruyama, K. A. Persson, A. Aspuru-Guzik, Accelerating the discovery of materials for clean energy in the era of smart automation. *Nat. Rev. Mater.* **3**, 5–20 (2018).
135. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
136. P. De Luna, J. Wei, Y. Bengio, A. Aspuru-Guzik, E. Sargent, Use machine learning to find energy materials. *Nature* **552**, 23–27 (2017).
137. G. L. Hart, T. Mueller, C. Toher, S. Curtarolo, Machine learning for alloys. *Nat. Rev. Mater.* **6**, 730–755 (2021).
138. F. Noé, A. Tkatchenko, K.-R. Müller, C. Clementi, Machine learning for molecular simulation. *Annu. Rev. Phys. Chem.* **71**, 361–390 (2020).
139. O. A. von Lilienfeld, K.-R. Müller, A. Tkatchenko, Exploring chemical compound space with quantum-based machine learning. *Nat. Rev. Chem.* **4**, 347–358 (2020).
140. B. Huang, O. A. von Lilienfeld, *Ab initio* machine learning in chemical compound space. *Chem. Rev.* **121**, 10001–10036 (2021).
141. O. A. von Lilienfeld, K. Burke, Retrospective on a decade of machine learning for chemical discovery. *Nat. Commun.* **11**, 4895 (2020).
142. A. Tkatchenko, Machine learning for chemical discovery. *Nat. Commun.* **11**, 4125 (2020).
143. J. A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller, A. Tkatchenko, Combining machine learning and computational chemistry for predictive insights into chemical systems. *Chem. Rev.* **121**, 9816–9872 (2021).
144. A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, A. Laio, Unsupervised learning methods for molecular simulation data. *Chem. Rev.* **121**, 9722–9758 (2021).
145. M. Meuwly, Machine learning for chemical reactions. *Chem. Rev.* **121**, 10218–10239 (2021).
146. J. Westermayr, P. Marquetand, Machine learning for electronically excited states of molecules. *Chem. Rev.* **121**, 9873–9926 (2021).
147. B. J. Frey, D. Dueck, Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
148. A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, The atomic simulation environment—A Python library for working with atoms. *J. Phys. Condens. Matter* **29**, 273002 (2017).
149. J. A. Morrone, R. Car, Nuclear quantum effects in water. *Phys. Rev. Lett.* **101**, 017801 (2008).
150. G. Hura, J. M. Sorenson, R. M. Glaeser, T. Head-Gordon, A high-quality x-ray scattering experiment on liquid water at ambient conditions. *J. Chem. Phys.* **113**, 9140–9148 (2000).

Acknowledgments: We thank M. Brenner for insightful comments. **Funding:** O.T.U. acknowledges funding from the Swiss National Science Foundation (grant no. P2BSP2_188147). M.G. works at the BASLEARN—TU Berlin/BASF Joint Lab for Machine Learning, cofinanced by TU Berlin and BASF SE. M.S. and A.T. acknowledge financial support from the Fond National de la Recherche Luxembourg (FNR) under AFR PhD grant "CNDTEC (11274975)" as well as from the European Research Council via ERC Consolidator Grant "BeStMo (725291)." This work was supported in part by the German Ministry for Education and Research under grant nos. 01IS14013A-E, 01GQ1115, 01GQ0850, 01IS18025A, 031L0207D, and 01IS18037A. K.R.M. was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (no. 2019-0-00079, Artificial Intelligence Graduate School Program, Korea University and no. 2022-0-00984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation). **Authors contributions:** O.T.U.: Writing—original draft, conceptualization, investigation, writing—review and editing, methodology, data curation, validation, supervision, formal analysis, software, and visualization. M.S.: Writing—original draft, conceptualization, investigation, validation, formal analysis, software, and visualization. S.G.: Writing—original draft, investigation, resources, formal analysis, and software. T.U.: Investigation, writing—review and editing, methodology, resources, formal analysis, software, and visualization. H.M.: Writing—original draft, investigation, methodology, formal analysis, and software. S.K.: Data curation, validation, and software. D.A.: Software. M.G.: Conceptualization, investigation, writing—review and editing, methodology, validation, and software. L.M.S.: Investigation, resources, and formal analysis. J.T.B.: Investigation, writing—review and editing, methodology, validation, formal analysis, and visualization. A.T.: Writing—original draft, conceptualization, investigation, writing—review and editing, methodology, resources, funding acquisition, data curation, validation, supervision, formal analysis, project administration, and visualization. K.-R.M.: Writing—original draft, conceptualization, writing—review and editing, methodology, funding acquisition, data curation, validation, supervision, project administration, and visualization. **Competing interests:** All authors affiliated with Google DeepMind disclose their financial relationships with Alphabet Inc. The authors declare no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. The DFT data are available at Zenodo: https://zenodo.org/records/10720941.

Submitted 10 December 2023
Accepted 29 February 2024
Published 5 April 2024
10.1126/sciadv.adn4397