Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential

Richard Messerly

richard.messerly@lanl.gov

Los Alamos National Laboratory

Shuhao Zhang

Carnegie Mellon University

Małgorzata Makoś

Southern Methodist University

Ryan Jadrich

Los Alamos National Laboratory

Elfi Kraka

Southern Methodist University

Kipton Barros

Los Alamos National Laboratory

Benjamin Nebgen

Los Alamos National Laboratory https://orcid.org/0000-0001-5310-3263

Sergei Tretiak

Los Alamos National Laboratory https://orcid.org/0000-0001-5547-3647

Olexandr Isayev

Carnegie Mellon University https://orcid.org/0000-0001-7581-8497

Nicholas Lubbers

Los Alamos National Laboratory

Justin Smith

NVIDIA Corp. https://orcid.org/0000-0001-7314-7896

Article

Keywords:

Posted Date: June 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2383420/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

Version of Record: A version of this preprint was published at Nature Chemistry on March 7th, 2024. See the published version at https://doi.org/10.1038/s41557-023-01427-3.

Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential

Shuhao Zhang[1,2], Małgorzata Z. Makoś[3,4], Ryan B. Jadrich[2,5], Elfi Kraka[3], Kipton Barros[2,5], Benjamin T. Nebgen[2], Sergei Tretiak[2,5], Olexandr Isayev[1], Nicholas Lubbers[4*], Richard A. Messerly[2*], and Justin S. Smith[2,6*]

[1]Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[3]Computational and Theoretical Chemistry Group (CATCO), Department of Chemistry, Southern Methodist University, 3215 Daniel Avenue, Dallas, Texas 75275, USA
[4]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[5]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[6]NVIDIA Corp., San Tomas Expy, Santa Clara, CA 95051, USA
[*]nlubbers@lanl.gov, richard.messerly@lanl.gov, jusmith@nvidia.com

ABSTRACT

Reactive chemistry atomistic simulation has a broad range of applications from drug design to energy to materials discovery. Machine learning interatomic potentials (MLIPs) have become an efficient alternative to computationally expensive quantum chemistry simulations. In practice, developing reactive MLIPs requires prior knowledge of reaction networks to generate fitting data and refitting to extensive datasets for each new application. For this reason, many fields of chemistry would greatly benefit from a general reactive MLIP, i.e., an MLIP that is applicable to a broad range of reactive chemistry such that it can be applied to new systems without the need for retraining. In this work, we develop a general reactive MLIP through unbiased active learning with an atomic configuration sampler inspired by nanoreactor molecular dynamics. The resulting potential (ANI-1xnr) is then applied to study five distinct condensed-phase reactive chemistry systems: carbon solid-phase nucleation, graphene ring formation from acetylene, biofuel additives, combustion of methane and the spontaneous formation of glycine from early-earth small molecules. In all studies, ANI-1xnr closely matches experiment and/or previous studies using traditional model chemistry methods. As such, ANI-1xnr proves to be a highly general reactive MLIP for C, H, N, and O elements that does not need to be refit for each application, enabling high-throughput *in silico* reactive chemistry experimentation.

Introduction

Over the past several decades, atomic-scale simulation has become an invaluable computational tool for providing qualitative explanations of experimentally-observed phenomena. In principle, nearly all chemical and materials properties can be evaluated through molecular dynamics (MD) simulation, wherein atomic motion is dictated by integrating the second law of Newtonian physics. The quantitative predictability of MD simulation depends almost entirely on the accuracy of the underlying model used to compute the forces acting on each atom. However, standard physics-based computational models, such as classical force fields (FF) or quantum mechanics (QM), suffer from the historical trade-off between computational cost, accuracy, and generality. For this reason, a fast, accurate, and general reactive potential is of paramount importance, as it would enable MD simulation to fully realize its true potential to predict reliable reaction rates, propose efficiency reducing side reactions, and warn of dangerous conditions, all without ever entering the laboratory.

Approaches such as empirical valence bond,[1] modified embedded atom method,[2] reactive force field (ReaxFF),[3] and reactive empirical bond order[4] are often applied to describe the making and breaking of chemical bonds during reactive atomistic simulations. However, these classical reactive FF methods use pre-determined physically-inspired functional forms with a small number of model parameters to approximate the potential energy surface. While these methods have proven to be very powerful, they require application-specific parameterization, since their physically-inspired functional forms lack the flexibility to simultaneously describe a broad range of chemical systems.[5] Fitting reactive FFs also requires prior knowledge of the reaction networks to be simulated, which significantly limits the predictive capability and potentially results in human bias concerning

which reactions proceed. By contrast, QM methods (e.g., density-functional theory, DFT) are based on the underlying physics of electronic structure theory, rather than on pre-defined bonding patterns, which renders QM calculations quite transferable (i.e., they are applicable for estimating energies and forces for a wide range of systems without reparameterization). However, the computational cost of QM methods is prohibitive for many reactive MD studies, which often require MD simulations with long time-scales ($\gtrsim$ 1 ns) and/or large systems ($\gtrsim$ 1000 atoms).

Recently, machine learning (ML) interatomic potentials (MLIPs) have been proposed as an alternative to QM and FF methods for the prediction of potential energies and forces.[6–12] MLIPs aim to bridge the speed vs. accuracy vs. generality gap that has existed in chemistry for many decades. In 2007, Behler and Parrinello proposed a neural network (NN)-based ML method to represent high-dimensional potential energy surface of atomic systems.[13] In their method, atomic symmetry functions represent the local chemical environment of each atom and serve as input into elemental NNs to predict an atomic contribution to the potential energy. The potential energy is then calculated as the sum of atomic energy predictions and the force is computed as the gradient of the potential energy with respect to positions. These concepts have been applied in building a range of MLIPs for different applications.[14–17] For example, the ANAKIN-ME (a.k.a., ANI) method combined modifications to the Behler and Parrinello symmetry functions with massive datasets to construct transferable MLIPs for organic molecules containing the elements C, H, N, O, S, F, and Cl.[18,19] While the ANI MLIPs proved to be extremely general (i.e., applicable to many different functional groups and types of materials without requiring refitting) and accurate for near-equilibrium conformations of *in vacuo* organic molecules, these potentials do not address the challenges of condensed-phase (i.e., periodic systems of liquids, supercritical fluids, or solids) reactive chemistry. Similarly, most classical reactive FFs are parameterized using a small number of *in vacuo* QM calculations or experimental gas-phase data, despite the fact that their main use case is in condensed-phase high-temperature reactive MD simulations. By contrast, training an MLIP directly to condensed-phase QM data would ensure that the potential is reliable for the density ranges typically simulated with condensed-phase reactive MD.

MLIPs have been successfully applied to model chemical reactions in specific contexts,[20,21] for example, unimolecular/bimolecular gas-phase (or *in vacuo*) reaction pathways[22–24] and specific condensed-phase reactive chemistry simulations.[25] However, each application necessitates a new dataset and retraining of the MLIP.[26] This bespoke development of reactive MLIPs requires expert MLIP developers and significant compute resources to build MLIPs for each new target system, limiting the accessibility and impact of MLIPs on reactive simulations. For this reason, a highly general MLIP targeting *large classes* of condensed-phase reactive chemistry for the elements C, H, N, and O would be transformational. However, the current philosophy of fitting reactive MLIPs (and classical reactive FFs) by building a dataset of pathways for pre-determined reactions based on chemical intuition is inadequate for developing a general reactive potential. The major drawback of this approach is that the resulting MLIP is biased to only perform well on the assumed reaction network, which is unlikely to be completely exhaustive when spanning a wide range of chemical space. In this sense, a reactive MLIP trained with such a dataset mirrors many of the limitations of reactive FFs. Recent endeavors to build large datasets have yielded groundbreaking results towards a universal MLIP for a significant portion of the periodic table.[27] However, the sampling procedure of Takomoto et al. does not directly target reactive chemistry involving molecules consisting of the elements of interest in the present study, namely, C, H, N, and O. Furthermore, random sampling from MD trajectories can lead to datasets that are sparse and clustered in their coverage of phase space, especially for rare events such as chemical reactions, resulting in MLIPs with so-called "holes"[28] in the potential energy surface and overall poorer performance than targeted, model-aware sampling strategies.[18]

Active learning (AL)[29] is a class of algorithms designed to automatically sample, select, and label new data with the goal of efficiently generating a diverse and relevant dataset for producing a more robust ML model. AL aims to ameliorate human bias through automating the decision-making process for adding new data to a training dataset. Starting from a small initial bootstrap dataset, the algorithm is applied iteratively, yielding generations of datasets designed to improve upon their ancestors. AL has been applied to develop numerous MLIPs in recent years.[18,20,30–35] A sampling scheme (often MD-based, and often using an MLIP trained to the current AL dataset) is used to generate a very large pool of molecular configurations. The selection of new data from this pool is often based on an uncertainty quantification approach (such as query-by-committee[36]), aiming to include only high-uncertainty structures in the ever-growing dataset. The labeling of selected configurations with energies and forces is accomplished with QM calculations. Iterations of training, sampling, selecting, and labeling are performed until the resulting MLIP performs as desired or is no longer improving.

For the development of a reactive MLIP, existing methodologies for training, selecting, and labeling are relatively straightforward to apply. However, in the sampling step, adequately exploring reactive chemical space in an automated fashion is extremely challenging[37] because it requires the exploration of chemical variance of molecular species in tandem with structural variance associated with non-equilibrium thermodynamic processes. While recent work (performed simultaneous and independent to this study) presented an automated approach to sample transition states and minimum-energy-path structures for gas-phase (or *in vacuo*) reactions,[38] a sampling procedure designed to enable reliable condensed-phase reactive MD simulations is essential to ensure that our MLIP is robust for condensed-phase high-temperature reactive MD simulations.

Wang and co-workers developed an elegant but expensive approach for the MD-based exploration of reaction pathways

in the condensed phase, referred to as the *ab initio* nanoreactor (NR).[39,40] The NR was designed to model high-velocity molecular collisions of small molecules by using a fictitious biasing force to promote chemical reactions and the formation of new molecules, thus automatically exploring reaction pathways between arbitrary reactants and products. The NR took an intermediate stance between physically-realistic simulation and rule-based enumeration approaches. The pathways that result from energy refinement are applicable to any thermodynamic setting by providing reaction parameters (for example, concentration and temperature) as input variables to a kinetic mechanistic model. The *ab initio* NR was successfully able to predict graphene ring formation from pure acetylene as well as reaction pathways to form glycine, one of the building blocks of life today, from small early-earth molecules. Although Wang et al. clearly demonstrated the promise of the NR to discover reactive chemistry, their *ab initio* NR approach is extremely computationally expensive. Specifically, a 1-ns *ab initio* NR simulation required 132,400 graphics processing unit hours, despite using the relatively low-level Hartree-Fock method with a small basis set (3-21G).

Inspired by the work of Wang et al., here we design an AL sampling procedure based on the NR that targets arbitrary reactive chemical processes and compositions of C, H, N, and O elements, including near pure elemental systems and mixtures. Combined with the ANI model architecture and applying AL at scale, we aim to produce a robust and transferable MLIP. Figure 1 presents a summary of the nanoreactor active learning workflow and the specific applications investigated in this work with the final model, referred to as ANI-1xnr. To evaluate the reliability of ANI-1xnr in practical research scenarios, we conduct several condensed-phase reactive chemistry simulations with the ANI-1xnr potential, namely, carbon solid-phase nucleation, graphene ring formation from acetylene with varying O_2 concentrations, biodiesel ignition with different fuel additives, methane combustion, and the spontaneous formation of glycine from early-earth molecules. Across this wide range of applications, we show ANI-1xnr provides results that are consistent with chemical intuition, experimental data, QM calculations (DFT, Hartree-Fock and density-functional tight-binding, DFTB), and classical reactive MD simulations (ReaxFF and a bespoke MLIP). This study demonstrates the capability of automated chemical exploration workflows to build a general-purpose reactive potential, resulting in ANI-1xnr – the first fast, accurate and transferable potential capable of simulating a wide range of real-world reactive systems containing C, H, N, and O elements.

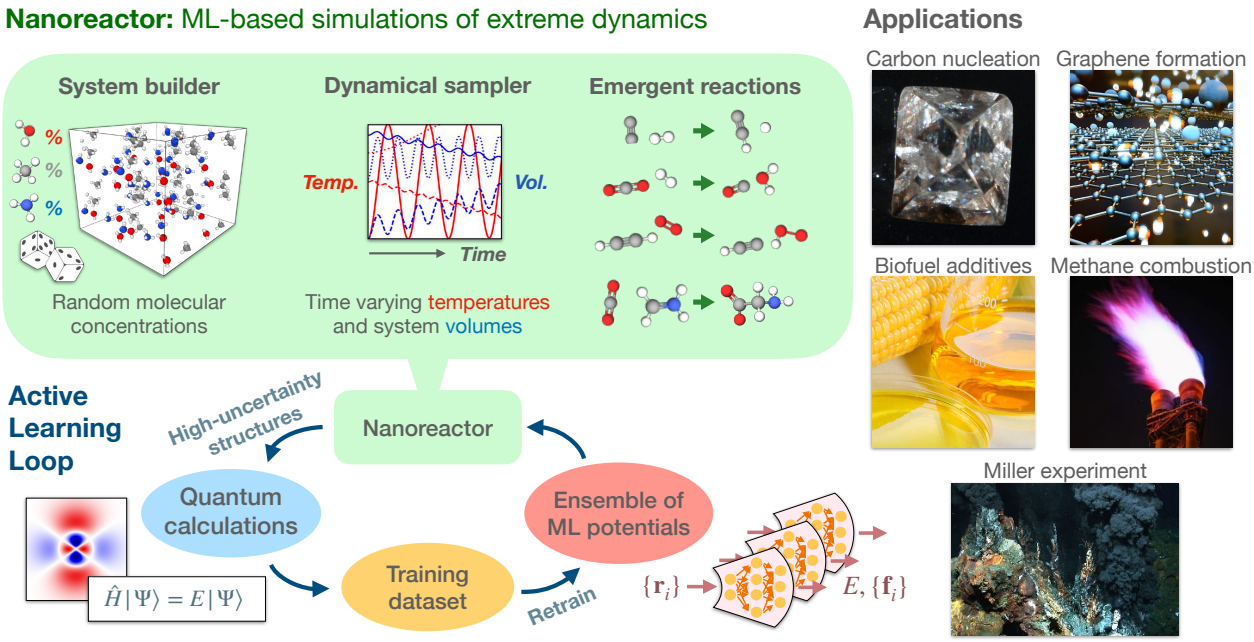


Figure 1. Summary of the nanoreactor active learning workflow and specific applications considered.

Results

We begin by evaluating the resulting training dataset produced by our nanoreactor active learning scheme. To demonstrate the state-of-the-art transferability of the ANI-1xnr potential, we then apply ANI-1xnr in five distinct case studies. First, we study the solid-phase nucleation of pure carbon at varying densities and compare our results with expected experimental structures. Second, we investigate the effect of O_2 concentration on the formation of graphene rings from acetylene, in comparison with experiment and DFTB simulations. Third, we determine the degree of ignition enhancement caused by different fuel additives

in biodiesel, in comparison with ReaxFF simulations. Fourth, we compare the ANI-1xnr potential with an application-specific MLIP for simulating methane combustion. Fifth, we develop a kinetic mechanism for the spontaneous formation of glycine from early-earth compounds and compare our mechanism with those obtained by *ab initio* MD nanoreactor and DFT-MD simulations. These five case studies were chosen to allow for a diverse comparison with common reactive chemistry methods. These case studies also serve as a real-world assessment of the transferability of ANI-1xnr, which was not developed specifically for any of the case study systems. Although we caution future users to avoid applying ANI-1xnr completely outside of the intended domain use (i.e., high-temperature condensed-phase reactive chemistry), the transferability of ANI-1xnr is truly remarkable, as we are not aware of any other general MLIP capable of modeling all five of the case study reactive systems.

Nanoreactor active learning

One of the most challenging tasks in developing a dataset for training MLIPs is knowing what atomic structures need to be included, which is only exacerbated when developing reactive MLIPs. Given the flexibility of neural network potentials (e.g., ANI-based models), it is insufficient to include in the training dataset only the reactants, products, or transition state optimized structures, or even the minimum energy path between states. Rather, for the MLIP to accurately model reaction dynamics, the training dataset must include the potential energy surface (up to some relative energy) that spans the distribution of structures sampled at the temperature of interest. For a general reactive MLIP, the temperature of interest is not known *a priori*. Therefore, we employ active learning (AL) combined with a nanoreactor (NR) sampler to generate the ANI-1xnr training dataset. In brief, the NR sampler randomly initializes condensed-phase systems of small molecules, containing $O(100)$ atoms. Then, the temperature and density oscillate during molecular dynamics to promote reactions and the formation of new products. After more than 50 iterations of AL, more than 26,000 condensed-phase systems with random C, H, N, and O compositions were selected and labeled with DFT energies and forces. Figure 1 shows a diagram of the nanoreactor active learning workflow developed in this study. See the Methods section for a more detailed description of our AL process and NR sampler.

Figure 2 a) through d) shows T-distributed Stochastic Neighbor Embeddings of atomic environment descriptors for each element calculated for a random subset of atoms in the training dataset. The blue points represent atomic environments in the ANI-1xnr dataset, while the red points represent atomic environments from the ANI-1x dataset. As expected, since ANI-1x is a non-reactive near-equilibrium molecule dataset, the ANI-1xnr dataset covers effectively all the space covered by ANI-1x. More importantly, the ANI-1xnr dataset provides pathways between many of the clusters in the ANI-1x dataset. These pathways likely correspond to reactions in a low-dimensional representation. By cross-referencing the ANI-1xnr training dataset with the existing PubChem database (only molecules with less than 10 CNO atoms), we conclude that the ANI-1xnr dataset contains 1212 unique known PubChem molecules. Figure 2 e) shows examples of known PubChem molecules that were found in the ANI-1xnr training dataset. Since all initial NR systems started from random placement of small (1 to 2 non-H atoms) molecules, the reaction pathways to produce these 1212 unique molecules must have been visited during the nanoreactor active learning process. In this way, the nanoreactor active learning process has automatically discovered hundreds, if not thousands, of reaction pathways leading to distinct molecular structures.

Carbon solid-phase nucleation

Accurate simulation of amorphous carbon systems has long been one of the top interests among chemists and materials scientists, as some distinct materials (e.g., graphene, diamond and carbon nanotubes) form from amorphous carbon under different conditions. Understanding this behavior would assist in the development of functional materials by controlling the solid-phase nucleation process. Many reactive FFs have been employed to simulate amorphous carbon in MD.[41,42] With the widespread use of ML methods, researchers are now also trying to investigate amorphous carbon systems with application-specific MLIPs.[43,44]

At present, most of the existing literature relies on selective sampling when building the training dataset for the MLIP, resulting in a bespoke potential for a specific application. Wang et al. trained an NN-based MLIP on a dataset obtained by randomly modifying low-dimensional carbon structures and taking snapshots from *ab initio* MD simulations of liquid carbon.[43] The MLIP trained on this dataset proved to be able to predict pure carbon fragments with the desired accuracy. Deringer et al. trained a Gaussian approximation potential on structures sampled from DFT-based MD on liquid carbon systems.[44] In addition to validating their Gaussian approximation potential with atomic energies and forces, they also used their potential to predict mechanical properties of the bulk system, e.g., Young's modulus.

Despite these achievements, MLIPs trained on application-specific datasets would have very poor transferability to new chemistry, as the model has only been fit to a limited number of structures and reactions. On the other hand, the AL approach presented in this work does not sample any specific form of carbon explicitly. We rely on the NR sampler and AL algorithm to automatically select physically-relevant and unbiased configurations of carbon atoms. To validate ANI-1xnr in carbon solid-phase nucleation simulations under different conditions, we perform simulations at high (3.52 g/cc), medium (2.25 g/cc), and low (0.50 g/cc) densities.

Figure 3 summarizes the product of each simulation. For each of the high-, medium- and low-density carbon simulations, ANI-1xnr produces the expected structure of carbon for the respective density. Specifically, for the system with the highest

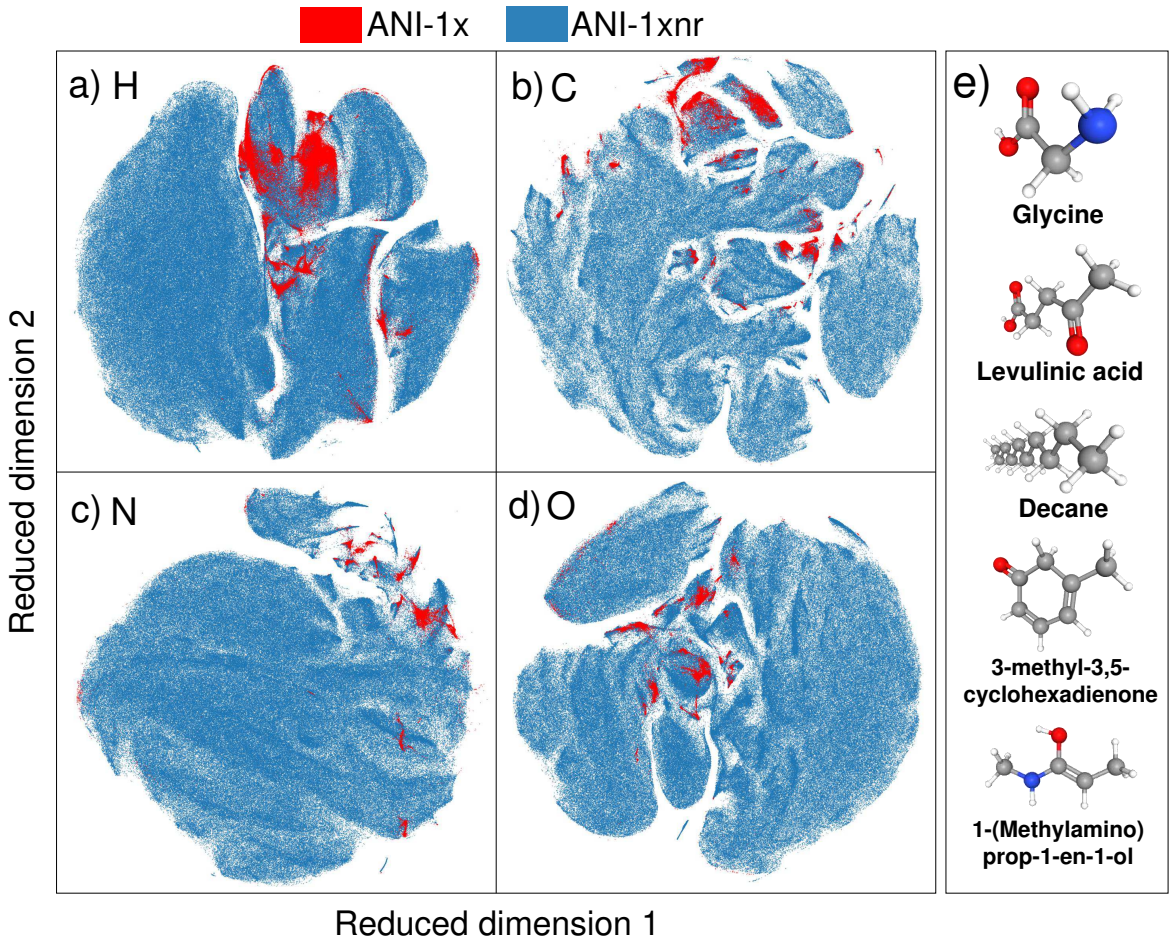


Figure 2. Inspection of the nanoreactor dataset. Panels a), b) c), and d) show 2D visualizations of the T-distributed Stochastic Neighbor Embeddings of the atomic descriptors from a random subset of the ANI-1x dataset (red) and ANI-1xnr dataset (blue). Panel e) shows five examples of the 1212 unique known PubChem molecules that formed during active learning and are present in the ANI-1xnr dataset.

density (3.52 g/cc), diamond, graphene and hexagonal diamond phase coexist after 246 ps, where 70% of carbon atoms in the simulation box forms diamond cubic crystal structure. After another 2.3 ns, the high-density system contains 86% of carbon atoms in the diamond cubic crystal structure, with very few graphene and hexagonal diamond sites. In the medium-density (2.25 g/cc) system, 31% of atoms rapidly form graphene after 8.2 ps, and the system contains 83% graphene after another 2.3 ns. Graphene sheets tend to form a stacked and more ordered graphite-like structure. The low-density (0.5 g/cc) system forms carbon atom chains after 250 ps, with 11% of atoms forming graphene sheets. After another 3 ns, the system contains 88% of atoms formed in graphene sheets. However, the graphene sheets in this low-density case are more disordered and appear to form fullerene-like closed or partially closed meshes.

Extended Data Table 1 contains the crystal lattice constants for diamond and graphite predicted from three ANI-based MLIPs and measured by experiment. ANI-1xnr reproduces the diamond cubic lattice constants with an error of only 0.01 Å. For graphite, ANI-1xnr predicts a and b lattice constants also with an error of 0.01 Å. However, the c lattice constant in graphite (along the direction of π-π stacking) is predicted with a relatively large error of 0.47 Å. Finally, ANI-1xnr accurately predicts the non-orthogonal experimental cell angles for diamond and graphite (see Supplementary Table S.I).

The relatively large error in the c lattice parameter for graphite is likely due to ANI-1xnr being a short-range potential, while long-range dispersion interactions are important for π-π stacking. We also trained an MLIP with a longer-range local cutoff (5.5/4.5 Å) than the original ANI-1xnr potential (5.2/3.5 Å), called ANI-1xnr(lr). The longer-range model performed significantly better than the original ANI-1xnr on the c lattice parameter with an error of 0.15 Å while also reducing the

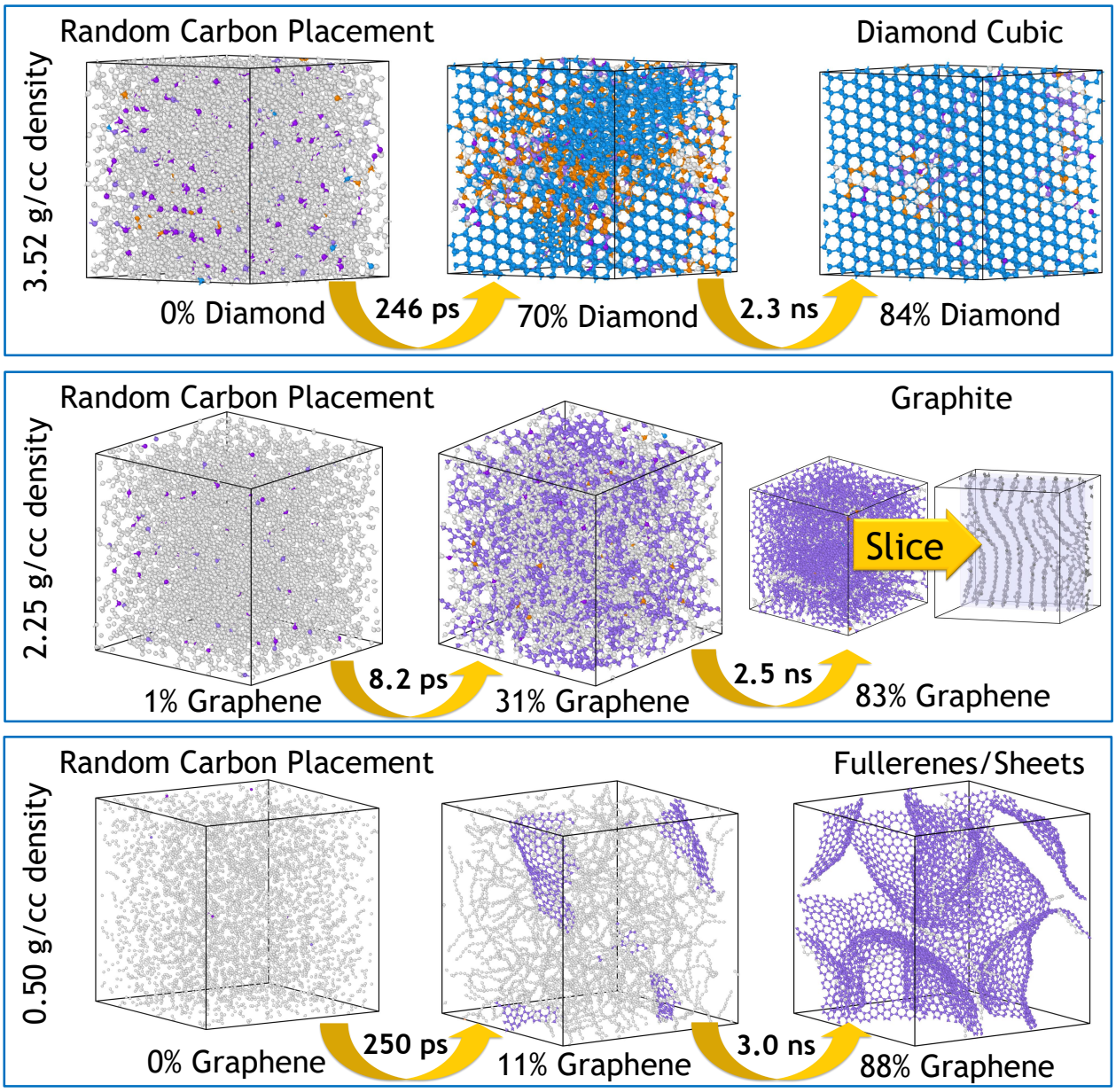


Figure 3. Results of ANI-1xnr carbon simulations starting from random carbon positions at three densities: 0.5 g/cc, 2.25 g/cc, and 3.52 g/cc.

0.01 Å error for the *a* and *b* lattice parameters (see Extended Data Table 1). However, the longer-range model performs worse on diamond cubic with an error for *a*, *b*, and *c* lattice constants of 0.09 Å. A possible explanation for this reduction in accuracy is that larger cutoffs reduce the resolution of the local atomic descriptors, which can affect accuracy in dense chemical environments. This shortcoming could be resolved by increasing the number of symmetry functions on the longer-range MLIP, but this would greatly impact the computational speed of the potential. A more optimal solution would be to add an explicit dispersion correction to ANI-1xnr that captures long-range interactions while maintaining an accurate description of the local environment.[45] We also compare the ANI-1xnr lattice constants with those from ANI-2x,[46] a model explicitly trained to small organic molecules as a baseline. ANI-2x performs poorly at predicting the lattice constants for both diamond cubic and graphite. This result is expected since the dataset used to train ANI-2x does not contain any structures similar to either of these systems. Furthermore, in contrast to the ANI-2x dataset reference calculations, the reference calculations used for building the ANI-1xnr dataset include dispersion corrections (see the Methods section for details), which are essential to reproduce the *c*

lattice parameter in graphite.

Effect of oxygen on graphene ring formation

Wang et al.[39] applied the original *ab initio* NR method to observe ring formation (i.e., the early stages of graphene formation) from a pure acetylene (C_2H_2) system. Subsequently, Lei et al.[47] presented DFTB NR simulations of acetylene in the presence of different amounts of oxygen, where O_2/C_2H_2 = 0, 0.1, ..., 1 is the ratio of added O_2 while the number of C_2H_2 molecules is fixed to 40. Graphene formation is the dominant process for pure C_2H_2, as the generation of free radicals enables the rapid growth of hydrocarbon rings. By contrast, the addition of O_2 to the system deters or, at high enough O_2/C_2H_2 ratios, completely eliminates ring formation.[47]

Similar to the work of Lei et al.,[47] we perform reactive simulations with varying ratios of C_2H_2 and O_2. In comparison with the DFTB simulations of Lei et al., ANI-1xnr enables significantly longer simulation times and larger systems. Specifically, while Lei et al. performed simulations of 0.5 ns with between 160 and 270 atoms (depending on the O_2/C_2H_2 ratio), we simulate 1000 atoms for 10 ns.

Figure 4 presents the amount of 3-, 4-, 5-, 6- , and 7-membered rings formed with respect to simulation time for 8 different O_2/C_2H_2 ratios. Increasing the oxygen ratio decreases the number of rings formed, which is in good agreement with the simulations from Lei et al. and experimental literature.[48] Furthermore, although the branching ratios (i.e., the relative production of different ring sizes) are not completely converged for all systems, the branching ratios are clearly in qualitative agreement with Lei et al. Specifically, 6-membered rings are the predominant product, followed by 5-membered and 7-membered rings at significantly lower, but nearly equal, branching ratios. However, in contrast with Lei et al., a significant number of 6-membered rings form even for an O_2/C_2H_2 ratio of 0.5. In comparison, the simulations of Lei et al. predict significant ring formation for O_2/C_2H_2 ratio up to 0.2, but negligible ring formation for an O_2/C_2H_2 ratio of 0.4. The ANI-1xnr results are in much closer agreement with experimental data, which report graphene formation for O_2/C_2H_2 ratios between 0.5 and 0.86. A clear explanation for this discrepancy is the difference in simulation timescales and system sizes achievable for ANI-1xnr compared with DFTB. Specifically, for an O_2/C_2H_2 ratio of 0.5, 6-membered rings only begin to form after 1 ns with ANI-1xnr. Considering that the DFTB simulations of Lei et al. ran for only 0.5 ns, our results suggest that 6-membered rings could form under higher oxygen ratio conditions using DFTB at longer time-scales. Although it is possible that even longer MD simulations could result in ring formation at even higher O_2/C_2H_2 ratios, this case study demonstrate the value in the significantly lower computational costs of ANI-1xnr compared to traditional methods, such as DFTB, to discover interesting phenomena that can only be observed during long time-scale simulations.

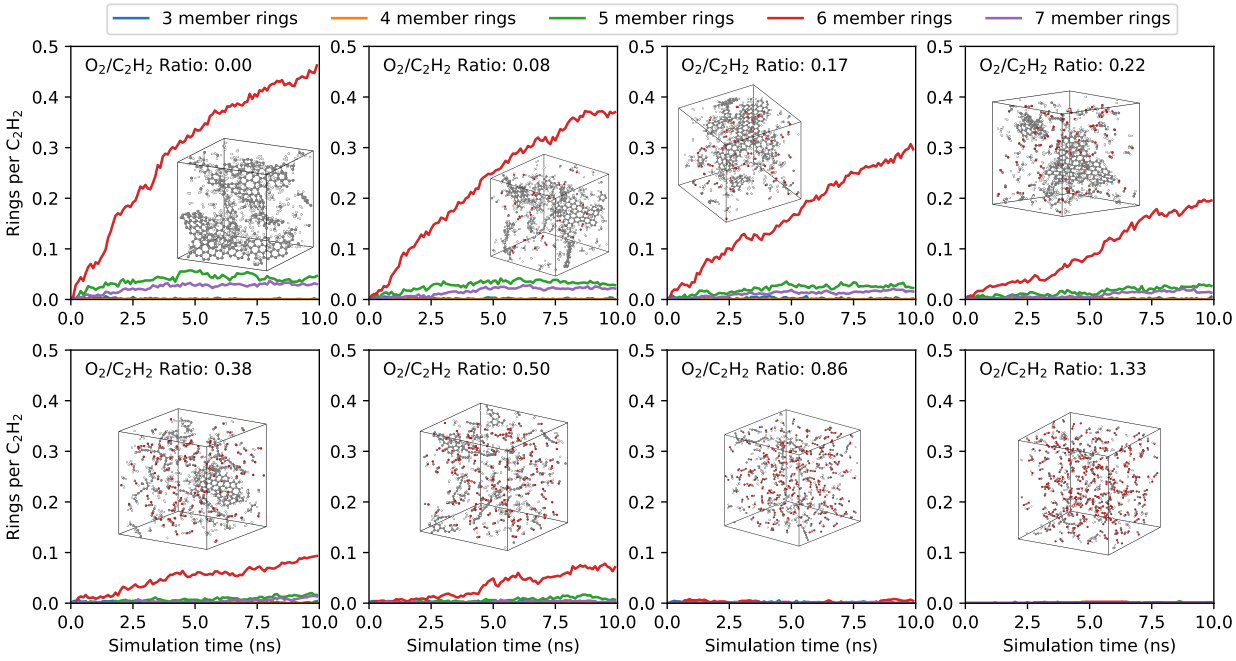


Figure 4. Comparison of 3-, 4-, 5-, 6-, and 7-membered ring formation for different ratios of C_2H_2 and O_2.

To verify that the ANI-1xnr predictions of ring formation are reliable, we compute the MLIP uncertainty (i.e., the ensemble standard deviation in energy normalized by the square-root of number of atoms) throughout the course of each simulation

(see Supplementary Figure S.1). The MLIP uncertainty is fairly constant and approximately equal to the AL energy threshold ($1.85\ \text{kcal}\cdot\text{mol}^{-1}\cdot N^{-\frac{1}{2}}$) through nearly the entire simulation (with only a few snapshots as exceptions). The relatively low and constant uncertainty confirms that each system is well-modeled by our MLIP. It is also interesting that the MLIP uncertainty decreases with increasing O_2/C_2H_2 ratio, suggesting that ANI-1xnr is most confident under a higher O_2/C_2H_2 ratio. The reason for such a tendency is that more oxidation reactions happen in the system with a larger O_2/C_2H_2 ratio, which is a common reaction in our training dataset, although typically involving species other than acetylene. By contrast, the system with a smaller O_2/C_2H_2 ratio has more ring formation, large carbon sheet formation and even phase change process, which are less common in the training set. Although C_2H_2 and O_2 are common species in the ANI-1xnr dataset (see Methods for details), the fact that the uncertainties remain slightly larger than those for any structure in the entire ANI-1xnr training set demonstrates that these specific systems were not studied directly in the NR AL sampling. Therefore, this case study serves as an assessment of the transferability of the ANI-1xnr potential.

Comparison of biofuel additives

To promote combustion processes of liquid fuel, fuel additives are utilized as detergents, oxygenates, emission depressors, corrosion inhibitors, dyes, and to increase the octane number. Chen et al. [49] performed high-temperature high-pressure MD simulations with ReaxFF to predict the mechanisms and kinetics of several fuel additives, including ethanol, 2-butanol, and methyl tert-butyl ether (MTBE). According to their results, 2-butanol was the best fuel additive at enhancing ignition while MTBE demonstrated similar ignition enhancement to 2-butanol. By contrast, ethanol was the worst fuel additive, having a negligible effect on the O_2 consumption rate and ignition delay time (IDT) compared to the clean biofuel.

In order to validate the reliability of ANI-1xnr for simulating biodiesel and to investigate the reported ignition enhancement of fuel additives, we reproduced four systems simulated by Chen et al.,[49] namely, clean biodiesel, biodiesel with ethanol as additive, biodiesel with 2-butanol as additive, and biodiesel with MTBE as additive. Figure 5 shows that the main products (CO, CO_2, and H_2O) are produced in very similar quantities to the ReaxFF simulations of Chen et al. However, the overall rate of fuel and O_2 consumption is considerably faster for ANI-1xnr compared with ReaxFF. Specifically, for all four cases, nearly all of the O_2 was consumed in the first 0.3 ns with ANI-1xnr, while there was still 20%-50% unconsumed O_2 after 2 ns with ReaxFF (for tracking plots including the entire 2 ns simulation, see Supplementary Figure S.2). The discrepancy in overall reaction rates between ANI-1xnr and ReaxFF is likely due to a difference in the underlying QM approach used to build each model (see Methods for details).

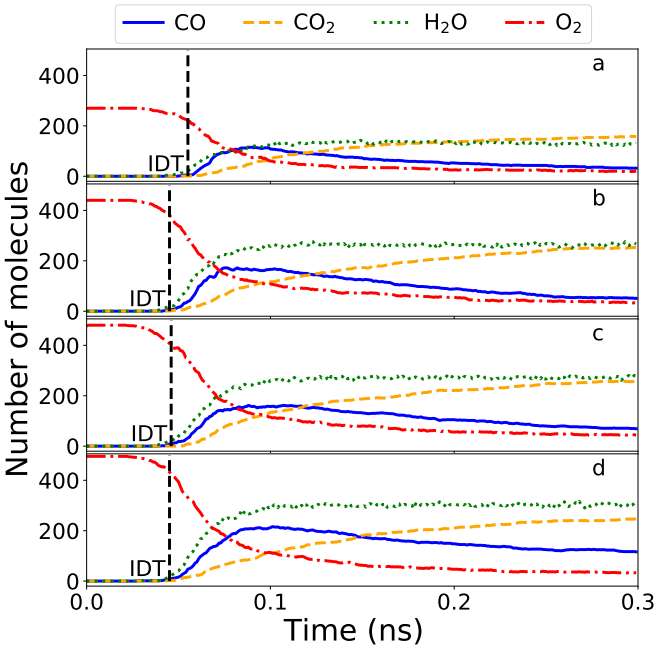


Figure 5. Tracking plot of major products (CO, CO_2, and H_2O) and O_2 for the biofuel simulations. (a) biofuel+O_2 (b) biofuel+O_2 with ethanol additive (c) biofuel+O_2 with 2-butanol additive (d) biofuel+O_2 with MTBE additive. Ignition delay time (IDT) is defined as the average time that at least five molecules of CO, CO_2, and H_2O are produced (see Supplementary Figure S.3).

Extended Data Table 2 quantifies the similarities and differences between ANI-1xnr and ReaxFF results. Despite the quantitative difference in ignition delay times, the additive effect on ignition delay for ANI-1xnr agrees qualitatively with ReaxFF, namely, all three additives cause product formation to occur at earlier times compared to clean biodiesel (recall Figure 5). While the reduction in IDT is not as significant for ANI-1xnr compared to ReaxFF, IDTs are highly sensitive as to how the system is initialized and to how ignition is defined (see Supplementary Figure S.3). Furthermore, ANI-1xnr predicts that 2-butanol and MTBE both result in significant enhancement of O_2 consumption, similar to ReaxFF. The primary qualitative discrepancy with ReaxFF is that ANI-1xnr predicts that ethanol also enhances O_2 consumption. Specifically, after the first 0.07 ns of ANI-1xnr simulation, 50% of O_2 was consumed in the pure biofuel system, while systems with additives consumed around 60% of O_2 (for O_2 consumption plots, see Supplementary Figure S.4). By contrast, in the ReaxFF simulations both the clean biofuel and ethanol additive systems consumed around 50% of O_2 after 2 ns, while the 2-butanol additive and MTBE additive systems consumed about 70% of O_2.

While the ANI-1xnr results for ethanol are in conflict with ReaxFF, experimental work demonstrates that ethanol can actually accelerate fuel ignition at relatively high pressures, in agreement with our high-pressure simulation results.[50] Closer inspection of our results provides understanding as to how ethanol accelerates the ignition process, similar to 2-butanol and MTBE. In comparison to the pure biofuel, all three systems with additives have a higher and earlier peak in OH radical, when normalized by the initial amount of O_2 (see Supplementary Figure S.5). The enhancement in OH production for ethanol is intuitive since ethanol contains a hydroxyl group with a similar bond dissociation energy to 2-butanol. Considering the important role that the OH radical plays in ignition and combustion chemistry, the accelerated rate of OH production is consistent with a lower IDT for all three additive systems.

Methane combustion

Emerging research has shown the success of application-specific (bespoke) MLIPs on systems like radical reactions in hydrocarbon combustion and well-known gas-phase mechanisms.[51,52] Zeng et al.[25] trained an NN-based potential to a dataset of QM calculated fragment clusters sampled from a ReaxFF simulation of the combustion process of a mixture of CH_4 and O_2. They showed that their application-specific MLIP could then simulate the combustion process of methane with a reasonable mechanism. Though our ANI-1xnr potential was trained for a more general purpose, we compare the performance of our MLIP to the application-specific MLIP of Zeng et al. for methane combustion under high temperatures and pressures. Specifically, we reproduce their MD simulation of methane combustion under the same conditions with the ANI-1xnr potential. Figure 6 a) shows that the ANI-1xnr potential produces very similar major products and species profiles to those of Zeng et al. However, by comparison with the CH_4 and O_2 consumption rates of Zeng et al., ANI-1xnr predicts an overall reaction rate that is approximately a factor of 40 times faster. Specifically, while their system required 0.5 ns of simulation time to consume half of the initial CH_4, our system required only 0.012 ns. Similar to the biofuel case, the difference in the overall reaction rate is likely due to the difference in the reference DFT reaction energy barriers (see Methods for details).

Due to the extreme simulation conditions, no experimental reference data are available for comparison. However, the similar trend for species concentration with respect to time in comparison with the work of Zeng et al. indicates that our general-purpose MLIP was able to learn the relevant physics and mechanisms as well as the application-specific MLIP of Zeng et al. Also, the CH_4 and O_2 consumption curves for the ANI-1xnr model are much closer to exponential decay, which is more physically reasonable than the near-linear decay plots of Zeng et al.

As a preliminary attempt to elucidate the cause for the discrepancy between the ANI-1xnr rates and those of Zeng et al., we compare the C-H bond dissociation energies from ANI-1xnr (116.6 $\text{kcal} \cdot \text{mol}^{-1}$), our reference DFT (110.0 $\text{kcal} \cdot \text{mol}^{-1}$), the reference DFT of Zeng et al. (115.0 $\text{kcal} \cdot \text{mol}^{-1}$), and experiment (103.3 $\text{kcal} \cdot \text{mol}^{-1}$). Although an *in vacuo* energy barrier may not correspond to a condensed-phase reaction rate, this analysis demonstrates that the accelerated consumption of CH_4 is not due to a significantly under predicted C-H bond dissociation energy by ANI-1xnr or our reference DFT. However, the relatively large uncertainties in ANI-1xnr along the bond dissociation path suggest that ANI-1xnr should not be utilized for studying gas-phase (or *in vacuo*) reactions (see Supplementary Figure S.7). This is not surprising since the training dataset was generated with condensed-phase NR simulations. To further demonstrate that ANI-1xnr should only be utilized in condensed-phase systems, we report the MLIP uncertainty of ANI-1xnr for the Zeng et al. methane training dataset (consisting of small molecular clusters) and the Transition-1x reactive *in vacuo* dataset (see Supplementary Figure S.8).

Miller experiment

In 1959, Stanley Miller designed a famous experiment to elucidate the origins of life on earth.[53] Miller applied an electric field to a gaseous system consisting of simple small-molecule species (e.g., NH_3, CO, H_2O, H_2, and CH_4) and reported the formation of amino acids, such as glycine ($C_2H_5NO_2$). This revolutionary experiment led to the formation of the field of prebiotic chemistry, which aims to discover the reaction networks that produce molecules which are essential for the formation of life. In this spirit, computational studies have attempted to imitate the reaction conditions of the Miller experiment to predict the key reaction pathways that lead to the formation of glycine. Recently, Saitta and Saija performed relatively short (≈40 ps)

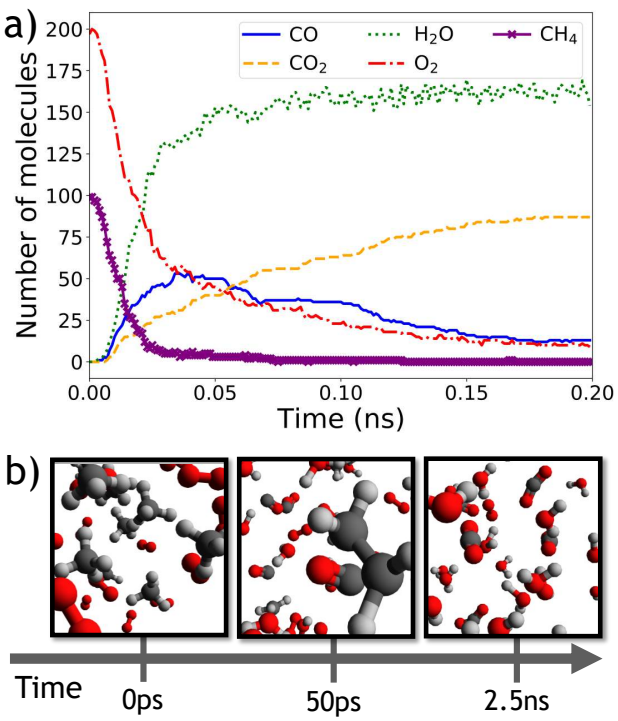


Figure 6. (a) Product molecule tracking plot of methane combustion simulation with ANI-1xnr. The tracking plot for the full simulation is provided as Supplementary Figure S.6. (b) Snapshots of initial reactants, intermediate species, and final products.

near-ambient-temperature (400 K) condensed-phase (≈1 g/cc) DFT-MD simulations wherein an electric field is applied directly to "spark" chemical reactions.[54] As our MLIP does not contain the necessary electronic information to apply an electric field, we instead encourage reactions to occur on picosecond time scales by performing high-temperature high-density MD simulations, similar to the Miller NR simulation of Wang et al.[39] Due to the low computational cost of our MLIP, we are able to run our Miller simulation considerably longer (≈4 ns) than the *ab initio* NR simulations of Wang et al. (≈1 ns) with the same system size of 228 atoms but with periodic boundary conditions. For this reason, we use a constant condensed-phase density (with corresponding pressures around 1 GPa) rather than applying an artificial "piston" to periodically compress the non-periodic gas-phase system to around 10 GPa, as was the approach employed by Wang et al.

Figure 7 presents the ANI-1xnr reaction mechanism to form glycine starting from the initial reactants. During our Miller simulation, glycine is formed three times and persists for approximately 225 fs, 375 fs, and 913 fs. Dissociation of glycine in less than 1 ps is expected, considering the relatively high temperature of this simulation. The final step to form glycine is hydrogen addition to $C_2H_4NO_2$, similar to the mechanism of Saitta and Saija. However, hydrogen addition occurs at an oxygen atom in our mechanism, rather than at the α-carbon as in the Saitta and Saija mechanism. In one instance, our Miller simulation produced the same $C_2H_4NO_2$ isomer as reported by Saitta and Saija. By contrast to the Saitta and Saija mechanism, this $C_2H_4NO_2$ isomer dissociated in our simulation rather than forming glycine. The key precursor to $C_2H_4NO_2$ is CH_4N, which is formed through several pathways. The pathway to form CH_4N that proceeds through the CH_2O intermediate is very similar to the mechanism reported by Wang et al.[39] The mechanisms to form the intermediates formaldehyde (CH_2O) and hydrogen cyanide (CHN) from the initial reactants CO, NH_3, and H_2O were nearly identical to those reported by Wang et al.[39] and Saitta and Saija.[54] The novel pathways to form the key intermediates carbon dioxide (CO_2) and methylene (CH_2) are reported in Supplementary Figure S.9.

Overall, there are several similarities between our mechanism and those of Wang et al. and Saitta and Saija. Although some differences exist between our mechanism and those reported in these previous simulation studies, this is not surprising considering not only the difference in levels of theory (i.e., Hartree-Fock vs DFT vs MLIP), but also the difference in the simulation methodologies (i.e., our Miller simulation did not utilize a “piston” nor induce an electric field). By comparing the DFT energies and forces with ANI-1xnr over the first 800 ps of the Miller simulation (see Supplementary Figure S.11), we verified that ANI-1xnr is consistent with the underlying reference DFT. Further validation of ANI-1xnr would require an extensive investigation into the accuracy of our DFT approach for the complete reaction network of this system.

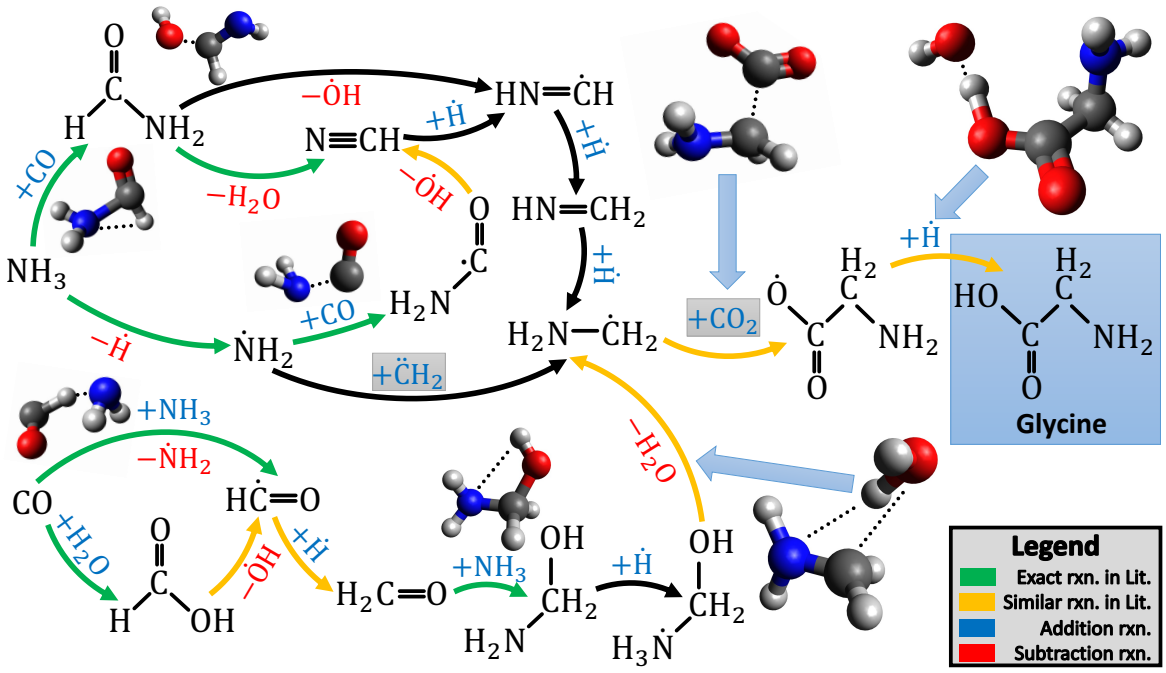


Figure 7. Formation of glycine. Note that +H does not necessarily signify a free hydrogen atom, +H is short-hand for a proton donor, e.g., NH_4, NH_3, CHO, CHNO, H_3O, H_2O. Likewise, -H does not necessarily signify dissociation of a hydrogen atom. -H is short-hand for a proton acceptor, e.g., NH_2, CO, CNO, H_2O, OH. Green arrows denote reactions previously identified by Wang et al. or Saitta and Saija. Orange arrows denote reactions that have a similar reaction in Wang et al. or Saitta and Saija. Boxes encapsulate key intermediates, whose formation mechanisms are reported in Supplementary Figure S.9. The depiction of bond orders and radical species is based simply on chemical intuition, since ANI-1xnr does not provide explicit orbital or electronic information (see Supplementary Figure S.10 for an alternative interpretation of this mechanism involving ionic species).

Conclusions

In this article, we introduce a nanoreactor active learning data generation approach and resulting general machine learning interatomic potential (ANI-1xnr) for organic condensed-phase reactive molecular dynamics. ANI-1xnr is trained to a large dataset obtained using an AL workflow employing a new MD-based sampling algorithm to discover diverse and relevant condensed-phase reactive atomistic configurations. Our nanoreactor sampler builds a reactive dataset spanning elemental compositions of C, H, N, and O under a wide range of conditions. Our AL process provided data of unprecedented diversity and relevance while also uncovering more than one thousand unique molecules from nine small-molecule initial species. Each unique molecular species formed by molecular dynamics simulation in our NR sampler was the result of one or more reaction pathways which did not need to be known, specified, or analyzed ahead of time. Our AL approach represents a break through in the automated development of next generation potentials for reactive molecular dynamics simulation.

We further validate the accuracy and applicability of the ANI-1xnr potential on five distinct condensed-phase reactive studies, namely, carbon solid-phase nucleation at different densities, high temperature graphene ring formation from acetylene with varying O_2 concentrations, ignition of biodiesel with three different fuel additives, combustion of methane, and the spontaneous formation of glycine in early-earth conditions. In carbon solid-phase nucleation and graphene ring formation studies, we show that ANI-1xnr reproduces experiment well. In other cases, where experiment is not available for comparison, ANI-1xnr produces results that are by and large consistent with modern modeling approaches, such as DFT, DFTB, ReaxFF, and an application-specific MLIP, all without the need for retraining.

We are providing the resulting ANI-1xnr potential and dataset to the community for further application and analysis. The DFT method used to calculate energies and forces for the dataset was selected as an accurate reference approach that has been successfully employed to study a diverse range of organic reactions while remaining affordable enough for high-throughput computations of large condensed-phase systems. For reactive chemistry, advanced electronic structure methods, such as double hybrid DFT, may provide better accuracy, albeit at a much greater computational cost. A highly valuable avenue for future research is to improve upon the potential through more accurate quantum chemistry calculations, perhaps using transfer

learning.[55] Another beneficial direction for future work would be to retrain an MLIP to an augmented dataset that includes the original ANI-1xnr condensed-phase dataset and *in vacuo* reactive data computed with the same DFT method (e.g., by recomputing the Transition-1x dataset or by following the original nanoreactor refinement process[39] to extract key reaction pathways from the NR simulations).[37,38] A recent advancement in ML for natural language processing is the concept of foundational models, i.e., large, general models usually trained with unlabeled data that can be specialized to specific tasks quickly with very small amounts of data.[56] Because ANI-1xnr is trained to a large, general dataset, it would also be interesting to consider whether it can act as a foundational model for more application-specific MLIPs when greater accuracy is required, for example, when predicting reaction rates. The ANI-1xnr potential is fully local, meaning long-range effects (i.e., London dispersion and Coulombic interactions) are not described explicitly. While this approximation does not appear to have a significant impact on the model reliability for the high-temperature applications studied in this work, this approximation may be inadequate for future applications. Fortunately, there are now many available options with explicit long-range terms, including charge equilibration schemes,[57] as well as cutting-edge graph neural network models[58] that implicitly account for long-range interactions. Future work should investigate the possible benefits of training these long-range ML model structures to the ANI-1xnr dataset. Similar to the recent advancement of ANI-2x,[19] we also recommend future work extend the training dataset to include additional elements, such as S, F, and Cl.

Methods

ANI-1xnr model descriptions, training details

The ANI-1xnr model was trained similarly to ANI models within other contexts,[19] including materials science[31] and chemistry.[59] We use the ANI descriptors,[60] which is a modified form of the Behler and Parinello neural network descriptors.[13] ANI-1xnr uses a local cutoff of 5.2 Å for the radial descriptors and 3.5 Å for the angular descriptors. The model is trained for the elements C, H, N, and O, each of which has its own specialized NN-based potential. The neural network architecture for each element and symmetry functions are reported as Supplementary Tables S.II and S.III, respectively. The model was trained using both energy and force terms in the loss function as described in previous work.[61] During training, we employ early stopping to prevent overfitting with learning rate annealing to ensure a high-fidelity fit. The model training is considered converged when the learning rate drops below 1.0×10^{-5}. Model performance against the held-out test dataset is presented in Supplementary Table S.IV and Supplementary Figures S.12 and S.13. Although the ANI-1xnr root-mean-squared-errors (RMSEs) are approximately an order of magnitude larger than an MLIP trained to near-equilibrium data (e.g., ANI-1x), this is not surprising since the ANI-1xnr dataset is considerably more challenging to train to, due to the wider range of atomic environments and system energies present in a reactive dataset. Furthermore, the ANI-1xnr dataset is much more general than most reactive datasets that are limited to a single system of interest, e.g., the Zeng et al. reactive dataset for $CH_4 + O_2$. By comparison, the force RMSEs of ANI-1xnr are only about 30% higher than those of the MLIP trained to the Zeng et al. single-reactive-system dataset, despite the ANI-1xnr dataset covering a significantly wider range of reactive chemistry.

Computational nanoreactor active learning for training set generation

The ANI-1xnr training dataset was generated through an iterative active learning process, where sampling of new atomic configurations is obtained with a nanoreactor-inspired MD simulation. To bootstrap the AL process, periodic cells containing randomly placed and oriented small molecules with less than three non-H atoms and with a randomly selected composition of C, H, N, and O are generated. All training energies and forces are computed with the open-source CP2K software[62] using unrestricted Kohn–Sham DFT,[63] BLYP functional,[64,65] TZV2P basis set,[66] GTH pseudopotentials,[67] D3 dispersion correction,[68] and energy cutoffs of 600 and 60 Ry, respectively, for the plane-wave and Gaussian contributions to the basis set, as recommended in previous work.[69] The overall spin multiplicity is constrained to a singlet state. Note that this is the spin for the entire box, not just a single molecule or radical. The assumption that a condensed-phase system does not accumulate a significant amount of spin is effectively an infinite-system-size approximation. This choice of spin multiplicity is consistent with previous studies that perform CP2K simulations of bulk systems containing radical species.[70]

The current AL generation MLIP is then used to drive MD sampling with random oscillations of temperature and density to promote reactions during the allotted simulation time. All MD simulations in this study are performed with the NeuroChem package and the Atomic Simulation Environment (ASE).[71] The average computational speed of our ASE-NeuroChem MD simulations was approximately 50k atomic gradients per second on a single NVIDIA Titan V GPU. We acknowledge that a more optimized code, such as LAMMPS,[72] would be significantly more computationally efficient. For example, recent work has demonstrated that highly-optimized MLIPs within LAMMPS can obtain 1-10 million atomic gradients per second on a single NVIDIA A100 GPU.[73,74] However, because our AL and case study simulations were performed on relatively small systems (no greater than 5000 atoms), it was not necessary to fully optimize our computational performance. To demonstrate that there is an opportunity to significantly improve our performance, we achieved 90k atomic gradients per second simply by increasing our simulation size to 25k atoms and, thereby, more efficiently utilizing the GPU.

We use an uncertainty metric, i.e., the normalized ensemble standard deviation in energy and the forces,[31,55] to gauge when the model is under performing. Snapshots of the MD that are deemed to be poorly described by the MLIP, based on the uncertainty metric, are included in the dataset with their corresponding QM energy and forces. To achieve a balance between exploration of chemical space and exploitation of the most important regions of the potential energy surface, the uncertainty thresholds vary between AL iterations, where the latter AL iterations generally have larger thresholds than earlier iterations. The final uncertainty threshold values for the normalized energy and forces were 1.85 kcal $\cdot$ mol$^{-1}\cdot N^{-\frac{1}{2}}$ and 0.3 eV/Å, respectively. Below is a detailed step-by-step description of the active learning workflow (see Figure 1 for a high-level overview):

1. Generate a bootstrap dataset (labeled with energies and forces) of 100 randomly generated periodic cells containing randomly placed and oriented small molecules including C_2, H_2, N_2, O_2, NH_3, CH_4, CO_2, H_2O, C_2H_2 with random composition.

2. Train ensemble of ANI potentials to the current training dataset using 8-fold (16 blocks) cross validation (14/1/1 - train/validation/test split) scheme.

3. Prepare for nanoreactor active learning sampling:

   Build a new random box of small molecules with random size, density, placements, orientations. Define a random schedule function for oscillating temperature (T) and density (ρ). Oscillating functional form is the same for temperature and density (see equations below), where t is time, t_{max} is a hyperparameter for the max time the simulation will run, and T_{start}, T_{end}, T_{amp}, ρ_{start}, ρ_{end}, ρ_{amp} and t_{per} are randomly selected values within a predetermined range (see Supplementary Table S.VI):

$$T(t) = T_{\text{start}} + \frac{t}{t_{\text{max}}}(T_{\text{end}} - T_{\text{start}}) + T_{\text{amp}} \sin^2\left(\frac{t}{t_{\text{per}}}\right)$$

$$\rho(t) = \rho_{\text{start}} + \frac{t}{t_{\text{max}}}(\rho_{\text{end}} - \rho_{\text{start}}) + \rho_{\text{amp}} \sin^2\left(\frac{t}{t_{\text{per}}}\right)$$

4. Run nanoreactor MD simulation using forces from current AL generation MLIP

5. Monitor energy and force uncertainty metrics every 5-50 MD steps (with MD time-step of 0.5 fs). If the uncertainty values exceed a pre-selected threshold value, end the simulation and add configuration to a new batch of structures.

6. Run QM single-point calculations on the new batch of structures to obtain energy and force labels.

7. Add new labeled data to the training dataset.

8. Go back to step 2 and repeat until the MLIP converges. We define convergence as when MLIP-driven MD sampling simulations run for O(50 ps) on average. In other words, convergence is achieved when the MLIP is confident in all new MD simulations.

Details of the resulting training dataset

The resulting training set from the AL procedure includes 26,442 simulation cells with an average system size of 139 atoms. Distributions of the system sizes, compositions, and densities can be found in Supplementary Figures S.16, S.17, and S.18, respectively. The vast majority (≈95%) of configurations in the training dataset have a density between 0.5-2.0 g/cc. While the minimum density in the dataset is ≈0.03 g/cc, less than 1% of the configurations in the dataset have a density less than 0.1 g/cc, suggesting that ANI-1xnr should not be trusted for low-density gas-phase simulations or *in vacuo* calculations.

Supplementary Figure S.19 shows a histogram of the sizes of all molecules that are found in the ANI-1xnr dataset, which includes one molecule up to 145 atoms. The majority are small molecules, of similar size or slightly larger than those from which the systems were initialized. There are many occurrences of molecules in the range of 10 to 90 atoms. The largest structures, ascertained by visual inspection, are graphene sheets. Furthermore, the 1212 unique PubChem molecules (less than 10 CNO atoms) discussed in the Results section only represent the simulation frames that were selected by the uncertainty estimate. Therefore, 1212 should be considered a lower-bound of molecules discovered during active learning. There are likely many more molecules formed over all NR AL sampling, which is estimated to be 100s of nanoseconds of MD simulation time in aggregate.

To automate the extraction of common molecular entities that formed during the AL process, we developed a NetworkX-based package called MolFind. This python software tool employs user prescribed cutoff distances for defining when two atoms are bonded or not and discovers clusters of atoms connected via bonds. The 3D molecular architecture is partially captured through a graphical representation (i.e., nodes and edges) of the bonding topology where atoms are nodes and bonds are edges.

Graphs are encoded according to the open source python package called NetworkX. The graphical representation and the NetworkX package enables (1) the counting of the number of topologically distinct molecular species in a simulation via a graph isomorphism check and (2) a comparison to known molecular entities with a specified topology. Previously, we tabulated a large database of known molecules and associated topologies by scraping the entirety of the PubChem database up to 10 non-hydrogen atoms. The existence of a species in the database is not required for MolFind to extract a bonded atomic cluster, but if found, it can affix a chemical/species name with the entity.

Carbon solid-phase nucleation

To investigate the formation process of diamond and graphene, molecular dynamics simulations were performed for amorphous carbon under different densities. Three initial system structures with three different densities (0.5 g/cc, 2.25 g/cc and 3.52 g/cc) were generated by varying the simulation box length for a constant total number of carbon atoms of 5000. The initial system structure was built with in-house code. First, the initial position for the first carbon atom in the simulation box was randomly selected. Then, random positions were proposed for each additional carbon atom. A proposed position was accepted only if the distance to all previous positions was larger than twice the van der Waals radius for carbon atoms (1.7 Å). This process was repeated until all 5000 carbon atoms were inserted. Langevin dynamics were performed at a temperature of 2500 K for 5 ns with step length of 0.5 fs. Coordinates and properties were recorded every 50 fs (100 time steps). Eight independent trajectories were run for each density to verify that the correct phase was identified from different starting structures. Different phases (diamond cubic, hexagonal diamond, or graphene) in each snapshot were distinguished with the Open Visualization Tool.[75]

Effect of oxygen on graphene ring formation

To investigate ring formation from acetylene, MD simulations were performed for eight different systems with varying O_2/C_2H_2 ratios: (0.00, 0.08, 0.17, 0.22, 0.38, 0.50, 0.86, 1.33). All systems contained 1000 atoms, resulting in a range of 150-250 C_2H_2 molecules and 0-200 O_2 molecules, depending on the O_2/C_2H_2 ratio. In order to have a nearly identical density of 0.2 g/cc for each system, the box lengths ranged between 37 Å and 44 Å. The initial structures were generated with PackMol.[76] Next, the minimum-energy structure was obtained with the LBFGS optimizer. Then, Langevin dynamics simulations were run at 2000 K for 10 ns with a 0.5 fs time step and a friction constant of 0.01. Snapshots and properties were recorded every 0.5 ps (1000 time steps). Ring structures of varying sizes were identified and counted with our in-house code MolFind. Considering that the distance between bonded atoms can fluctuate, a 0.02 Å buffer was utilized when scanning C-C bonds so that any pair of carbon atoms that has distance smaller than 1.72 Å (two times the covalent radius of carbon atom plus the buffer) were considered bonded. Similar buffers were also added when analyzing other simulations.

Comparison of biofuel additives

To investigate the effect of different fuel additives on ignition performance, MD simulations were performed for clean biofuel and biofuel with three different additives: ethanol, 2-butanol and methyl tert-butyl ether. The biofuel composition, the number of additive molecules, and the number of O_2 molecules were the same as shown in Table 2 of the ReaxFF reference paper.[49] Each system consists of approximately 2000 atoms. Initial structures were generated using Packmol such that the initial separation of all molecules was at least 2 Å. The initial density was 0.2 g/cc in all four cases, consistent with Chen et al. Langevin dynamics were run at a temperature of 100 K for 1 ps for relaxation. Then, the system temperature was gradually increased to 3000 K at a 50 K/ps heating rate. After reaching the desired temperature of 3000 K, the simulation was ran for an additional 10 ns. A fixed time step of 0.1 fs was utilized. The temperature, time step, and heating profile were the same as those utilized by Chen et al.[49] During the whole process (including relaxation and temperature ramping) snapshots and properties were recorded every 1 ps (10,000 time steps). Five independent trajectories were performed for each system to reduce uncertainty in species profiles.

ANI-1xnr was trained to BLYP reference calculations, whereas ReaxFF was primarily developed based on B3LYP calculations (supplemented with high-accuracy bond dissociation energy data). Since reaction rates are extremely sensitive to energy barriers, this difference in the DFT functional can lead to a significant difference in overall reaction rates.

Methane combustion

The methane combustion system was initialized with 100 CH_4 molecules and 200 O_2 molecules. All molecules were inserted using Packmol and ensuring that all molecules were separated by at least 2.0 Å. The cubic simulation box length was 37.60 Å, resulting in a density of 0.25 g/cc. The temperature was initialized to 3000 K by Maxwell-Boltzmann distribution. Langevin dynamics were run for 1 ns with a time-step of 0.1 fs and with a friction constant of 0.01. The initial density, number of molecules, temperature, and time step were consistent with Zeng et al.[25] Snapshots and properties were recorded every 0.1 ps (1000 time steps).

ANI-1xnr was trained to reference calculations computed with BLYP functional and TZV2P basis set, whereas Zeng et al. utilized the MN15 functional and 6-31G** basis set.[25] Since reaction rates are extremely sensitive to energy barriers, this difference in the DFT functional and basis set can lead to a significant difference in overall reaction rates.

Miller experiment

To investigate the ability to simulate complex organic system that involve biologically relevant molecules, an MD simulation was performed with a similar species composition to the Miller experiment. Packmol was utilized to randomly place 16 H_2, 14 H_2O, 14 CO, 14 NH_3 and 14 CH_4 in a cubic simulation box with edge lengths of 12.1 Å, resulting in a density of 1.067 g/cc. The simulation was run with Langevin dynamics for over 4 ns with a time step of 0.25 fs. The temperature was linearly increased from 0 K to 300 K in the first 100 ps. Then, the temperature was linearly increased from 300 K to 2500 K in the next 100 ps. The temperature was then maintained at 2500 K for 4000 ps. The system was then cooled from 2500 K to 300 K over the final 200 ps. Snapshots and properties were recorded every 12.5 fs (50 time steps).

Acknowledgements

S.Z., K.M.B., B.T.N., S.T., N.L., and R.A.M. acknowledge support from the US DOE, Office of Science, Basic Energy Sciences, Chemical Sciences, Geosciences, and Biosciences Division under Triad National Security, LLC ("Triad") contract Grant 89233218CNA000001 (FWP: LANLE3F2). S.Z and M.Z.M. gratefully acknowledge the resources of the Los Alamos National Laboratory (LANL) Applied Machine Learning summer student program. The work at LANL was supported by the LANL Directed Research and Development Funds (LDRD) and performed in part at the Center for Nonlinear Studies (CNLS) and the Center for Integrated Nanotechnologies (CINT), a US Department of Energy (DOE) Office of Science user facility at LANL. This research used resources provided by the LANL Institutional Computing (IC) Program. O.I. acknowledges support from Office of Naval Research (ONR) through Energetic Materials Program (MURI grant number N00014-21-1-2476). M.Z.M. and E.K. acknowledge funding from National Science Foundation, Grant CHE 2102461.

Data availability

Data and methods used in this study will be publicly available. Details are provided in the corresponding sections in Methods.

Code availability

The code to reproduce this study will be available upon paper acceptance.

References

1. Warshel, A. & Weiss, R. M. An empirical valence bond approach for comparing reactions in solutions and in enzymes. *J. Am. Chem. Soc.* **102**, 6218–6226, DOI: 10.1021/ja00540a008 (1980).

2. Baskes, M. Determination of modified embedded atom method parameters for nickel. *Mater. Chem. Phys.* **50**, 152–158, DOI: 10.1016/S0254-0584(97)80252-0 (1997).

3. van Duin, A. C. T., Dasgupta, S., Lorant, F. & Goddard, W. A. ReaxFF: A reactive force field for hydrocarbons. *J. Phys. Chem. A* **105**, 9396–9409, DOI: 10.1021/jp004368u (2001).

4. Brenner, D. W. *et al.* A second-generation reactive empirical bond order (REBO) potential energy expression for hydrocarbons. *J. Phys. Condens. Matter* **14**, 783–802, DOI: 10.1088/0953-8984/14/4/312 (2002).

5. Senftle, T. P. *et al.* The ReaxFF reactive force-field: development, applications and future directions. *npj Comput. Mater.* **2**, 15011, DOI: 10.1038/npjcompumats.2015.11 (2016).

6. Zuo, Y. *et al.* Performance and cost assessment of machine learning interatomic potentials. *J. Phys. Chem. A* **124**, 731–745, DOI: 10.1021/acs.jpca.9b08723 (2020).

7. Behler, J. First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem., Int. Ed.* **56**, 12828–12840, DOI: 10.1002/anie.201703114 (2017).

8. Kulichenko, M. *et al.* The rise of neural networks for materials and chemical dynamics. *J. Phys. Chem. Lett.* **12**, 6227–6243, DOI: 10.1021/acs.jpclett.1c01357 (2021).

9. Bartók, A. P. & Csányi, G. Gaussian approximation potentials: A brief tutorial introduction. *Int. J. Quantum Chem.* **115**, 1051–1057, DOI: 10.1002/qua.24927 (2015).

10. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453, DOI: 10.1038/s41467-022-29939-5 (2022).

11. Thölke, P. & Fabritiis, G. D. Equivariant transformers for neural network based molecular potentials. In *International Conference on Learning Representations* (2022).

12. Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *arXiv* DOI: 10.48550/arXiv.2204.05249 (2022).

13. Behler, J. & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401, DOI: 10.1103/PhysRevLett.98.146401 (2007).

14. Khorshidi, A. & Peterson, A. A. Amp: A modular approach to machine learning in atomistic simulations. *Comput. Phys. Commun.* **207**, 310–324, DOI: 10.1016/j.cpc.2016.05.010 (2016).

15. Yao, K., Herr, J. E., Toth, D. W., Mckintyre, R. & Parkhill, J. The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics. *Chem. Sci.* **9**, 2261–2269, DOI: 10.1039/C7SC04934J (2018).

16. Singraber, A., Morawietz, T., Behler, J. & Dellago, C. Parallel multistream training of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **15**, 3075–3092, DOI: 10.1021/acs.jctc.8b01092 (2019).

17. Kang, P.-L. & Liu, Z.-P. Reaction prediction via atomistic simulation: from quantum mechanics to machine learning. *iScience* **24**, 102013, DOI: 10.1016/j.isci.2020.102013 (2021).

18. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, DOI: 10.1063/1.5023802 (2018).

19. Devereux, C. *et al.* Extending the applicability of the ANI deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **16**, 4192–4202, DOI: 10.1021/acs.jctc.0c00121 (2020).

20. Young, T. A., Johnston-Wood, T., Zhang, H. & Duarte, F. Reaction dynamics of Diels–Alder reactions from machine learned potentials. *Phys. Chem. Chem. Phys.* **24**, 20820–20827, DOI: 10.1039/D2CP02978B (2022).

21. Jiang, B., Li, J. & Guo, H. High-fidelity potential energy surfaces for gas-phase and gas–surface scattering processes from machine learning. *J. Phys. Chem. Lett.* **11**, 5120–5131, DOI: 10.1021/acs.jpclett.0c00989 (2020).

22. Kolb, B., Zhao, B., Li, J., Jiang, B. & Guo, H. Permutation invariant potential energy surfaces for polyatomic reactions using atomistic neural networks. *J. Chem. Phys.* **144**, 224103, DOI: 10.1063/1.4953560 (2016).

23. Cooper, A. M., Hallmen, P. P. & Kästner, J. Potential energy surface interpolation with neural networks for instanton rate calculations. *J. Chem. Phys.* **148**, 094106, DOI: 10.1063/1.5015950 (2018).

24. Li, J., Song, K. & Behler, J. A critical comparison of neural network potentials for molecular reaction dynamics with exact permutation symmetry. *Phys. Chem. Chem. Phys.* **21**, 9672–9682, DOI: 10.1039/C8CP06919K (2019).

25. Zeng, J., Cao, L., Xu, M., Zhu, T. & Zhang, J. Z. Complex reaction processes in combustion unraveled by neural network-based molecular dynamics simulation. *Nat. Commun.* **11**, 5713, DOI: 10.1038/s41467-020-19497-z (2020).

26. Chen, R., Shao, K., Fu, B. & Zhang, D. H. Fitting potential energy surfaces with fundamental invariant neural network. II. Generating fundamental invariants for molecular systems with up to ten atoms. *J. Chem. Phys.* **152**, 204307, DOI: 10.1063/5.0010104 (2020).

27. Takamoto, S. *et al.* Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat. Commun.* **13**, 2991, DOI: 10.1038/s41467-022-30687-9 (2022).

28. Behler, J. Constructing high-dimensional neural network potentials: A tutorial review. *Int. J. Quantum Chem.* **115**, 1032–1050, DOI: 10.1002/qua.24890 (2015).

29. Ren, P. *et al.* A survey of deep active learning. *ACM Comput. Surv.* **54**, 1–40, DOI: 10.1145/3472291 (2021).

30. Sivaraman, G. *et al.* Machine-learned interatomic potentials by active learning: amorphous and liquid hafnium dioxide. *npj Comput. Mater.* **6**, 104, DOI: 10.1038/s41524-020-00367-7 (2020).

31. Smith, J. S. *et al.* Automated discovery of a robust interatomic potential for aluminum. *Nat. Commun.* **12**, 1257, DOI: 10.1038/s41467-021-21376-0 (2021).

32. Yoo, P. *et al.* Neural network reactive force field for C, H, N, and O systems. *npj Comput. Mater.* **7**, 9, DOI: 10.1038/s41524-020-00484-3 (2021).

33. Zaverkin, V., Holzmüller, D., Steinwart, I. & Kästner, J. Exploring chemical and conformational spaces by batch mode deep active learning. *Digit. Discov.* DOI: 10.1039/D2DD00034B (2022).

34. Young, T. A., Johnston-Wood, T., Deringer, V. L. & Duarte, F. A transferable active-learning strategy for reactive molecular force fields. *Chem. Sci.* **12**, 10944–10955, DOI: 10.1039/D1SC01825F (2021).

35. Ang, S. J., Wang, W., Schwalbe-Koda, D., Axelrod, S. & Gómez-Bombarelli, R. Active learning accelerates *ab initio* molecular dynamics on reactive energy surfaces. *Chem* **7**, 738–751, DOI: 10.1016/j.chempr.2020.12.009 (2021).

36. Seung, H. S., Opper, M. & Sompolinsky, H. *Query by Committee*, 287–294 (Association for Computing Machinery; New York, NY, Pittsburgh, Pennsylvania, July 27-29, 1992).

37. Guan, X. *et al.* A benchmark dataset for hydrogen combustion. *Sci.* **9**, 215, DOI: 10.1038/s41597-022-01330-5 (2022).

38. Schreiner, M., Bhowmik, A., Vegge, T., Busk, J. & Winther, O. Transition1x – a dataset for building generalizable reactive machine learning potentials. *arXiv* DOI: 10.48550/ARXIV.2207.12858 (2022).

39. Wang, L.-P. *et al.* Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **6**, 1044–1048, DOI: 10.1038/nchem.2099 (2014).

40. Wang, L.-P. Force field development and nanoreactor chemistry. In *Computational Approaches for Chemistry Under Extreme Conditions*, 127–159, DOI: 10.1007/978-3-030-05600-1_6 (Springer International Publishing, 2019).

41. Los, J. H., Ghiringhelli, L. M., Meijer, E. J. & Fasolino, A. Improved long-range reactive bond-order potential for carbon. I. Construction. *Phys. Rev. B* **72**, 214102, DOI: 10.1103/PhysRevB.72.214102 (2005).

42. Srinivasan, S. G., Van Duin, A. C. & Ganesh, P. Development of a ReaxFF potential for carbon condensed phases and its application to the thermal fragmentation of a large fullerene. *J. Phys. Chem. A* **119**, 571–580, DOI: 10.1021/jp510274e (2015).

43. Wang, J. *et al.* A deep learning interatomic potential developed for atomistic simulation of carbon materials. *Carbon* **186**, 1–8, DOI: 10.1016/j.carbon.2021.09.062 (2022).

44. Deringer, V. L. & Csányi, G. Machine learning based interatomic potential for amorphous carbon. *Phys. Rev. B* **95**, 094203, DOI: 10.1103/PhysRevB.95.094203 (2017).

45. Rezajooei, N., Thien Phuc, T. N., Johnson, E. & Rowley, C. A neural network potential with rigorous treatment of long-range dispersion. *ChemRxiv* DOI: 10.26434/chemrxiv-2022-mdz85 (2022).

46. Devereux, C. *et al.* Extending the applicability of the ani deep learning molecular potential to sulfur and halogens. *J. Chem. Theory Comput.* **16**, 4192–4202, DOI: 10.1021/acs.jctc.0c00121 (2020). PMID: 32543858.

47. Lei, T. *et al.* Mechanism of graphene formation via detonation synthesis: A DFTB nanoreactor approach. *J. Chem. Theory Comput.* **15**, 3654–3665, DOI: 10.1021/acs.jctc.9b00158 (2019).

48. Sorensen, C., Nepal, A. & Singh, G. P. Process for high-yield production of graphene via detonation of carbon-containing material (2016). US Patent 9,440,857.

49. Chen, Z., Sun, W. & Zhao, L. Combustion mechanisms and kinetics of fuel additives: A ReaxFF molecular simulation. *Energy Fuels* **32**, 11852–11863, DOI: 10.1021/acs.energyfuels.8b02035 (2018).

50. Cooper, S. P., Mathieu, O., Schoegl, I. & Petersen, E. L. High-pressure ignition delay time measurements of a four-component gasoline surrogate and its high-level blends with ethanol and methyl acetate. *Fuel* **275**, 118016, DOI: 10.1016/j.fuel.2020.118016 (2020).

51. Brickel, S., Das, A. K., Unke, O. T., Turan, H. T. & Meuwly, M. Reactive molecular dynamics for the [Cl–CH3–Br]-reaction in the gas phase and in solution: a comparative study using empirical and neural network force fields. *Electron. Struct.* **1**, 024002, DOI: 10.1088/2516-1075/ab1edb (2019).

52. Li, J., Chen, J., Zhang, D. H. & Guo, H. Quantum and quasi-classical dynamics of the OH + CO $\rightarrow$ H + CO2 reaction on a new permutationally invariant neural network potential energy surface. *J. Chem. Phys.* **140**, 044327, DOI: 10.1063/1.4863138 (2014).

53. Miller, S. L. & Urey, H. C. Organic compound synthesis on the primitive earth. *Sci. (New York, N.Y.)* **130**, 245–251, DOI: 10.1126/science.130.3370.245 (1959).

54. Saitta, A. M. & Saija, F. Miller experiments in atomistic computer simulations. *Proc. Natl. Acad. Sci. U S A* **111**, 13768–13773, DOI: 10.1073/pnas.1402894111 (2014).

55. Smith, J. S. *et al.* Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, DOI: 10.1038/s41467-019-10827-4 (2019).

56. Yuan, L. *et al.* Florence: A new foundation model for computer vision. *arXiv* DOI: 10.48550/arXiv.2111.11432 (2021).

57. Ko, T. W., Finkler, J. A., Goedecker, S. & Behler, J. A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398, DOI: 10.1038/s41467-020-20427-2 (2021).

58. Lubbers, N., Smith, J. S. & Barros, K. Hierarchical modeling of molecular energies using a deep neural network. *J. Chem. Phys.* **148**, 241715, DOI: 10.1063/1.5011181 (2018).

59. Smith, J. S. *et al.* The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. data* **7**, 134, DOI: 10.1038/s41597-020-0473-z (2020).

60. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci. J.* **8**, 3192–3203, DOI: 10.1039/C6SC05720A (2017).

61. Smith, J. S., Lubbers, N., Thompson, A. P. & Barros, K. Simple and efficient algorithms for training machine learning potentials to force data. *arXiv* DOI: 10.48550/arXiv.2006.05475 (2020).

62. Kühne, T. D. *et al.* CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys* **152**, 194103, DOI: 10.1063/5.0007045 (2020).

63. Kohn, W. & Sham, L. J. Self-Consistent Equations Including Exchange and Correlation Effects. *Phys. Rev.* **140**, A1133–A1138, DOI: 10.1103/PhysRev.140.A1133 (1965).

64. Becke, A. D. Density-functional exchange-energy approximation with correct asymptotic behavior. *Phys. Rev. A* **38**, 3098–3100, DOI: 10.1103/PhysRevA.38.3098 (1988).

65. Lee, C., Yang, W. & Parr, R. G. Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density. *Phys. Rev. B* **37**, 785–789, DOI: 10.1103/PhysRevB.37.785 (1988).

66. VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *J. Chem. Phys.* **127**, 114105, DOI: 10.1063/1.2770708 (2007).

67. Goedecker, S., Teter, M. & Hutter, J. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B* **54**, 1703–1710, DOI: 10.1103/PhysRevB.54.1703 (1996).

68. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104, DOI: 10.1063/1.3382344 (2010).

69. Jadrich, R. B., Ticknor, C. & Leiding, J. A. First principles reactive simulation for equation of state prediction. *J. Chem. Phys.* **154**, 244307, DOI: 10.1063/5.0050676 (2021).

70. Fetisov, E. O. *et al.* First-principles Monte Carlo simulations of reaction equilibria in compressed vapors. *ACS Cent. Sci.* **2**, 409–415, DOI: 10.1021/acscentsci.6b00095 (2016).

71. Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002, DOI: 10.1088/1361-648X/aa680e (2017).

72. Thompson, A. P. *et al.* LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comp. Phys. Comm.* **271**, 108171, DOI: 10.1016/j.cpc.2021.108171 (2022).

73. Musaelian, A. *et al.* Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579, DOI: 10.1038/s41467-023-36329-y (2023). Number: 1 Publisher: Nature Publishing Group.

74. Chigaev, M. *et al.* Lightweight and effective tensor sensitivity for atomistic neural networks. *J. Chem. Phys.* **158**, 184108, DOI: 10.1063/5.0142127 (2023).

75. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO–the open visualization tool. *Model. Simul. Mat. Sci. Eng.* **18**, 015012 (2009).

76. Martínez, L., Andrade, R., Birgin, E. G. & Martínez, J. M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164, DOI: 10.1002/jcc.21224 (2009).

Extended Data: Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential

Shuhao Zhang[1,2], Małgorzata Z. Makoś[3,4], Ryan B. Jadrich[2,5], Elfi Kraka[3], Kipton M. Barros[2], Benjamin T. Nebgen[2], Sergei Tretiak[2], Olexandr Isayev[1], Nicholas Lubbers[4*], Richard A. Messerly[2*], and Justin S. Smith[2,6*]

[1]Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[3]Computational and Theoretical Chemistry Group (CATCO), Department of Chemistry, Southern Methodist University, 3215 Daniel Avenue, Dallas, Texas 75275, USA
[4]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[5]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[6]NVIDIA Corp., San Tomas Expy, Santa Clara, CA 95051, USA
[*]nlubbers@lanl.gov, richard.messerly@lanl.gov, jusmith@nvidia.com

Crystal	Model	a (Å)	b (Å)	c (Å)
Diamond	ANI-1xnr	3.58	3.58	3.58
	ANI-1xnr(lr)	3.66	3.66	3.66
	ANI-2x	3.75	3.75	3.64
	Exp.	3.57	3.57	3.57
Graphite	ANI-1xnr	2.47	2.47	6.24
	ANI-1xnr(lr)	2.46	2.46	6.56
	ANI-2x	2.44	2.44	10.0
	Exp.	2.46	2.46	6.71

Extended Data Table 1. Optimized crystal lattice constants (a, b, c) for diamond and graphite phases. Comparison between ANI-1xnr, ANI-1xnr(lr), ANI-2x and experiment (Exp.).

System	Ignition delay time (ps)		O_2 consumption (%)	
	ANI-1xnr	ReaxFF	ANI-1xnr ($t = 0.07$ ns)	ReaxFF ($t = 2$ ns)
Clean biofuel	55	239	49.0%	48.5%
Ethanol additive	45	126	58.6%	49.21%
2-butanol additive	46	110	57.5%	73.33%
MTBE additive	45	92	57.4%	70.3%

Extended Data Table 2. Comparison between ANI-1xnr and ReaxFF ignition delay time (IDT) and O_2 consumption for clean biofuel and biofuel with each of the three additives. Ignition delay time is defined as the average time that at least five molecules of CO, CO_2, and H_2O are produced. O_2 consumption is compared at 0.07 ns for ANI-1xnr and at 2 ns for ReaxFF, i.e., the time that the O_2 consumption for the clean biofuel is approximately equal for both models.

Supplementary Information: Exploring the frontiers of condensed-phase chemistry with a general reactive machine learning potential

Shuhao Zhang[1,2], Małgorzata Z. Makoś[3,4], Ryan B. Jadrich[2,5], Elfi Kraka[3], Kipton M. Barros[2], Benjamin T. Nebgen[2], Sergei Tretiak[2], Olexandr Isayev[1], Nicholas Lubbers[4*], Richard A. Messerly[2*], and Justin S. Smith[2,6*]

[1]Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
[2]Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[3]Computational and Theoretical Chemistry Group (CATCO), Department of Chemistry, Southern Methodist University, 3215 Daniel Avenue, Dallas, Texas 75275, USA
[4]Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[5]Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[6]NVIDIA Corp., San Tomas Expy, Santa Clara, CA 95051, USA
[*]nlubbers@lanl.gov, richard.messerly@lanl.gov, jusmith@nvidia.com

Crystal	Model	α (^{o})	β (^{o})	γ (^{o})
Diamond	ANI-1xnr	90.0	90.0	90.0
	ANI-1xnr(lr)	90.0	90.0	90.0
	ANI-2x	90.0	90.0	90.0
	Exp.	90.0	90.0	90.0
Graphite	ANI-1xnr	90.0	90.0	120.
	ANI-1xnr(lr)	90.0	90.0	120.
	ANI-2x	90.4	89.7	120.
	Exp.	90.0	90.0	120.

Table S.I. Optimized crystal angles (α, β, γ) for diamond and graphite phases. Comparison between ANI-1xnr, ANI-1xnr(lr), ANI-2x and experiment (Exp.).

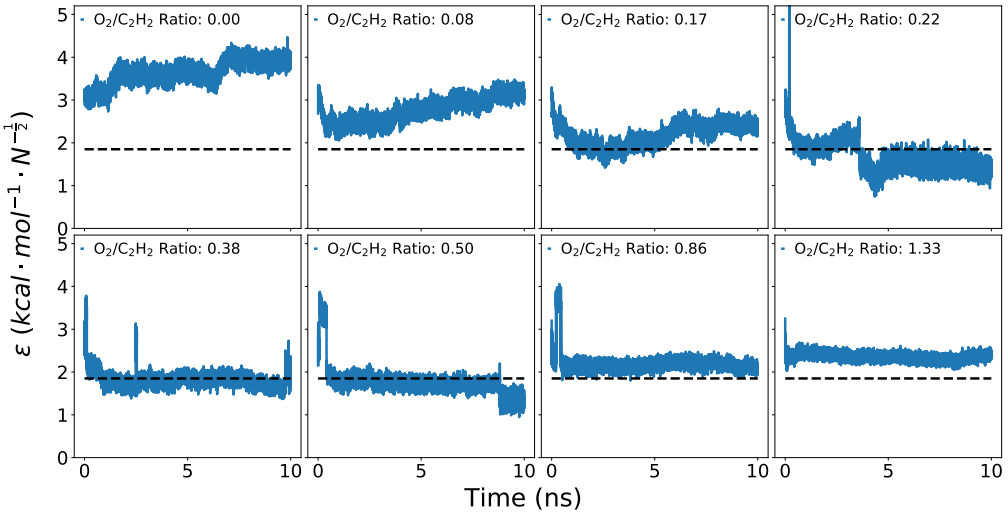


Figure S.1. ANI-1xnr ensemble standard deviation in energy normalized by the square-root of number atoms (ε) for all eight O_2/C_2H_2 ratios. Dashed black line corresponds to AL energy threshold of 1.85 kcal $\cdot$ mol$^{-1}\cdot N^{-\frac{1}{2}}$.

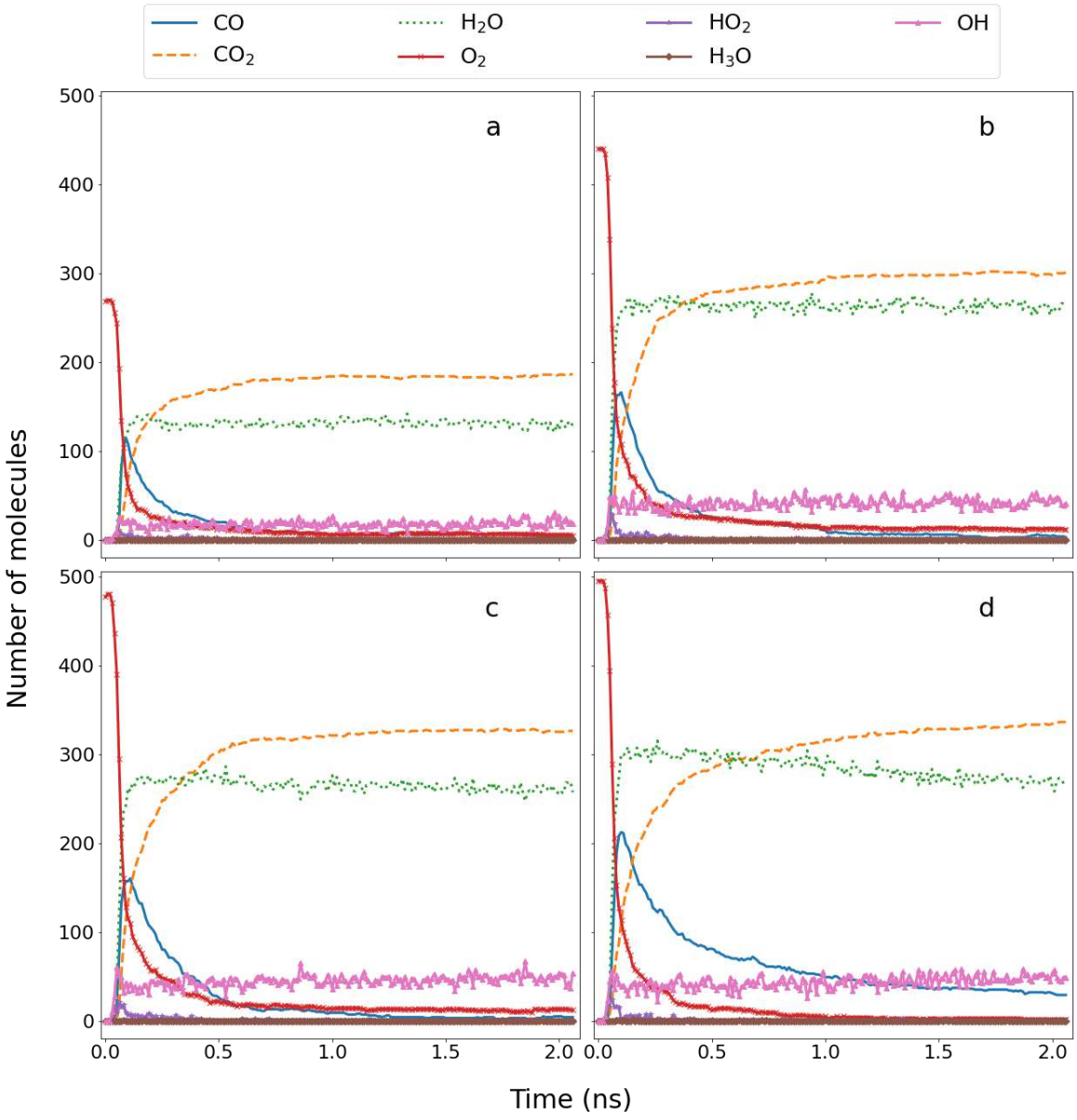


Figure S.2. Tracking plot of major products of the biofuel simulations. (a) biofuel+O_2 system (b) biofuel with ethanol+O_2 (c) biofuel with 2-butanol+O_2 (d) biofuel with MTBE+O_2.

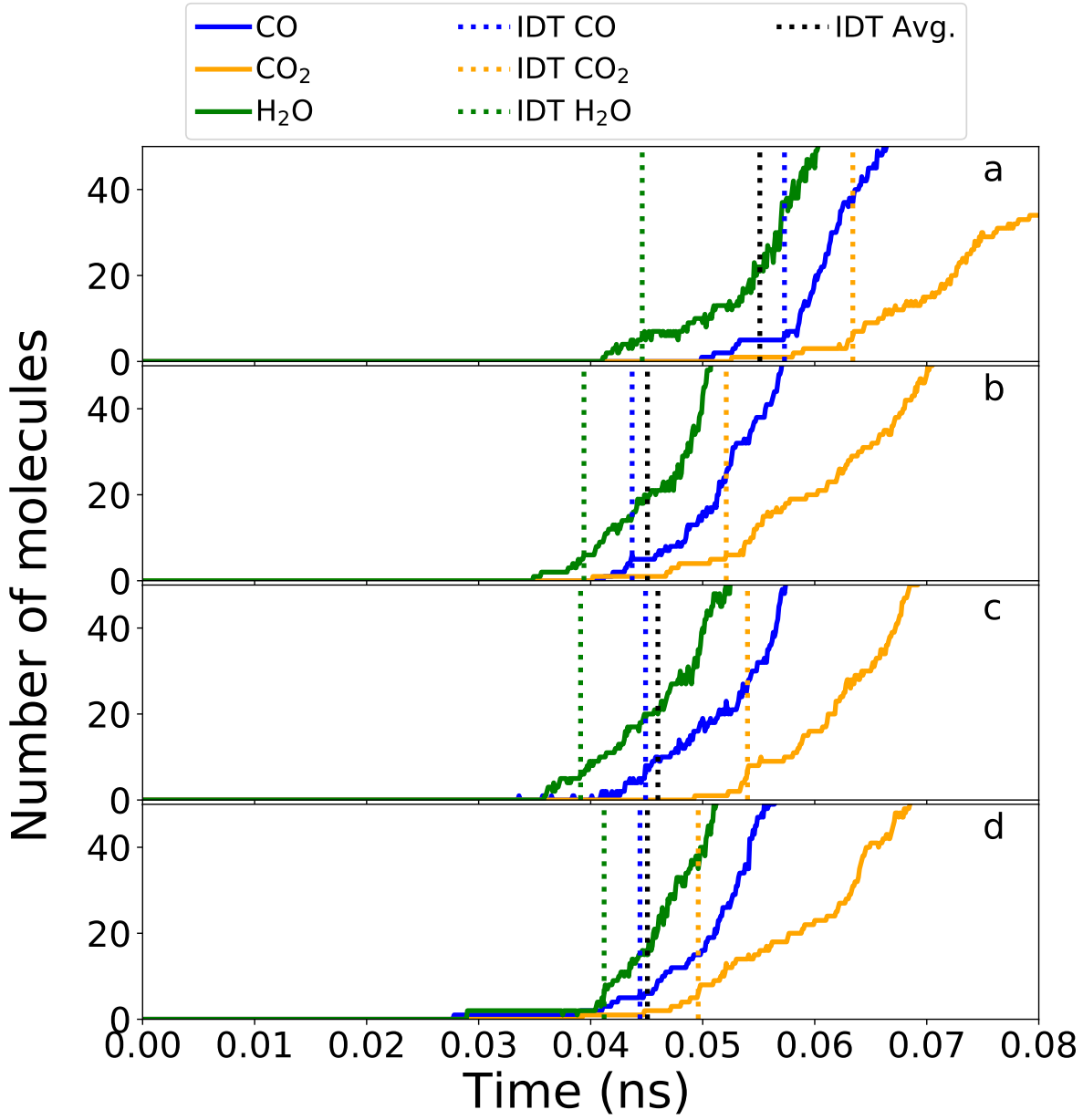


Figure S.3. Ignition delay time (IDT) for biofuel simulations based on each of the major products CO, CO_2, and H_2O. To remove anomalies when only a single molecule is produced significantly prior to "true" ignition, we define IDT as the earliest time that at least five molecules of a given product are produced. The manuscript uses the average IDT value between CO, CO_2, and H_2O. (a) biofuel+O_2 system (b) biofuel with ethanol+O_2 (c) biofuel with 2-butanol+O_2 (d) biofuel with MTBE+O_2.

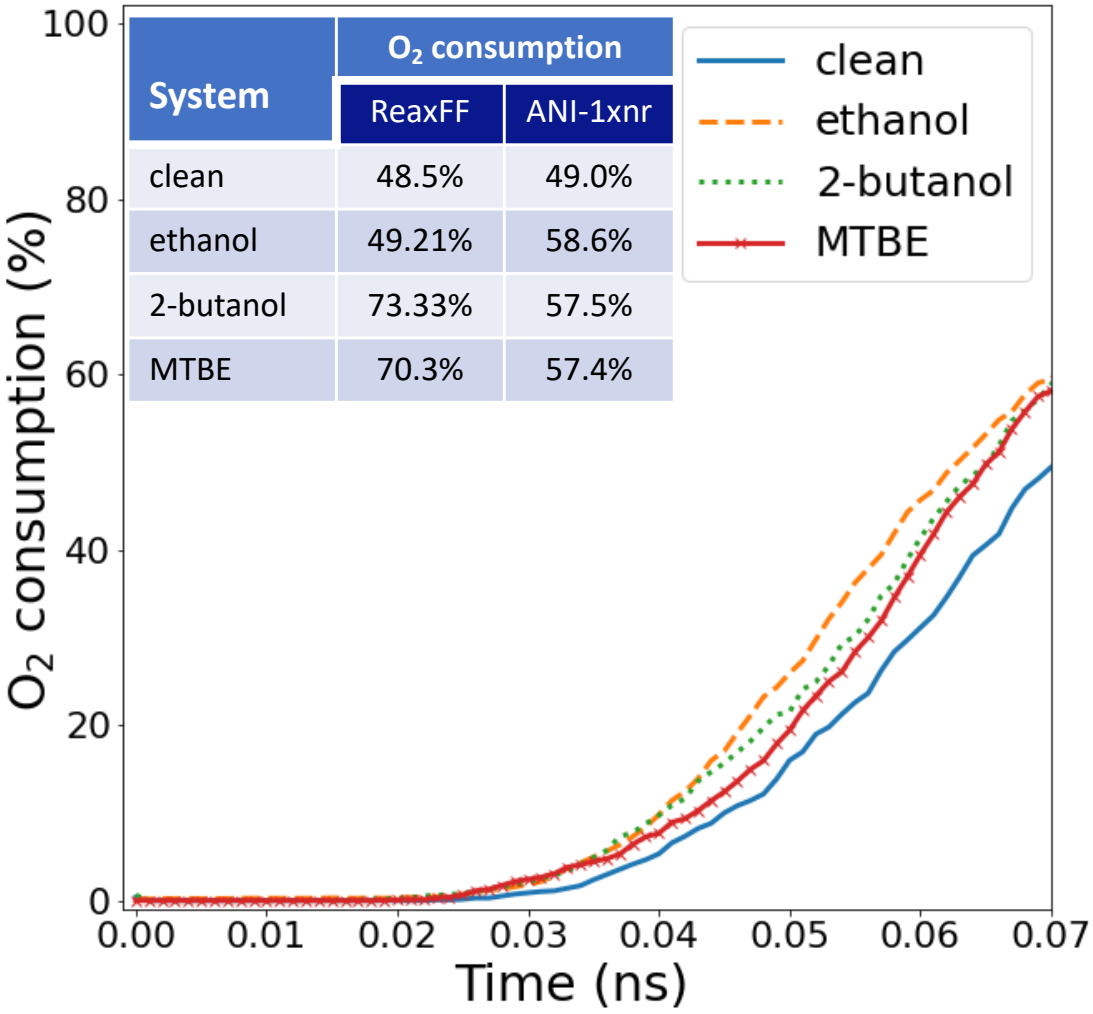


Figure S.4. O_2 consumption (%) during combustion for clean biofuel compared with three different fuel additives, namely, ethanol, 2-butanol, and MTBE. Insert compares O_2 consumption for ANI-1xnr (at 0.07 ns) with ReaxFF (at 2 ns). Curves are smoothed by averaging over 5 independent trajectories.

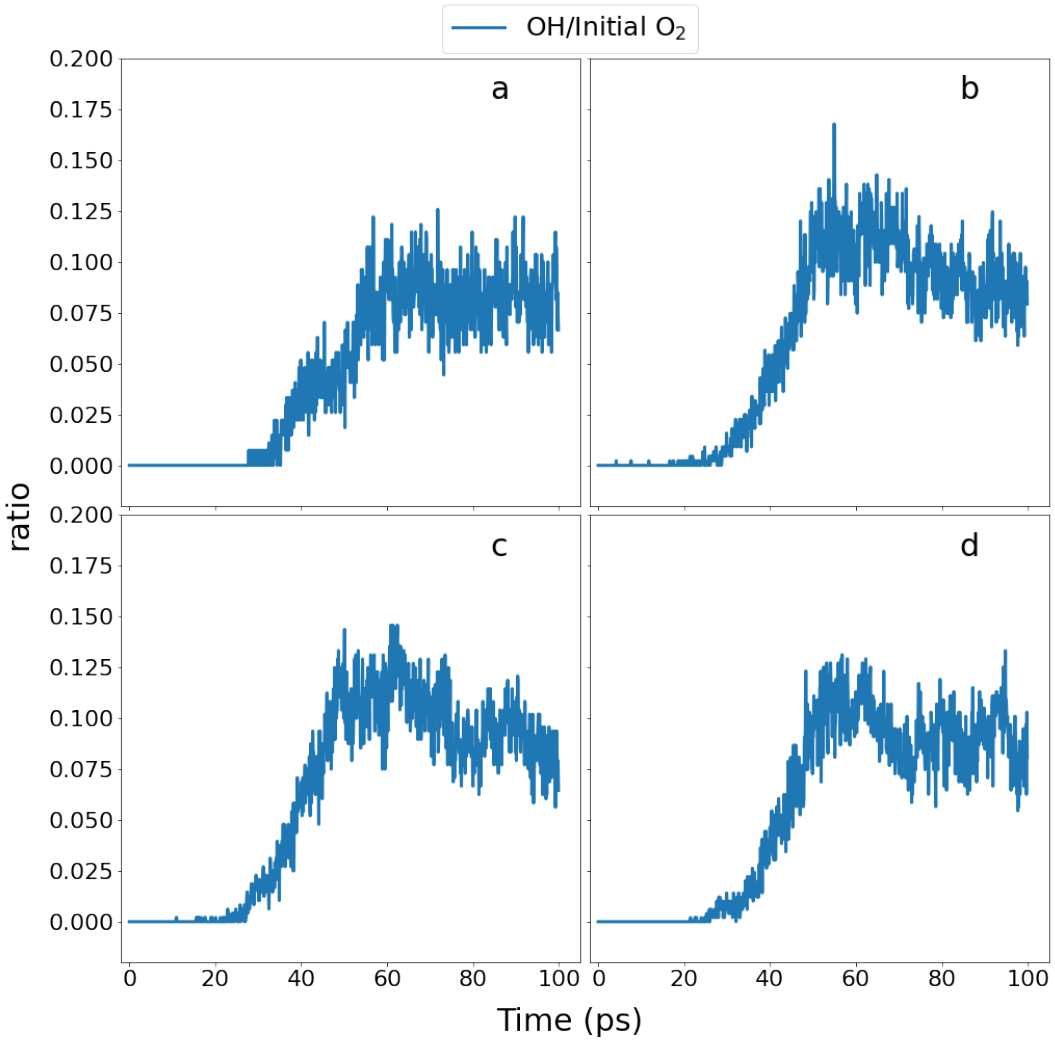


Figure S.5. Ratio of OH to initial O_2. (a) biofuel+O_2 system (b) biofuel with ethanol+O_2 (c) biofuel with 2-butanol+O_2 (d) biofuel with MTBE+O_2

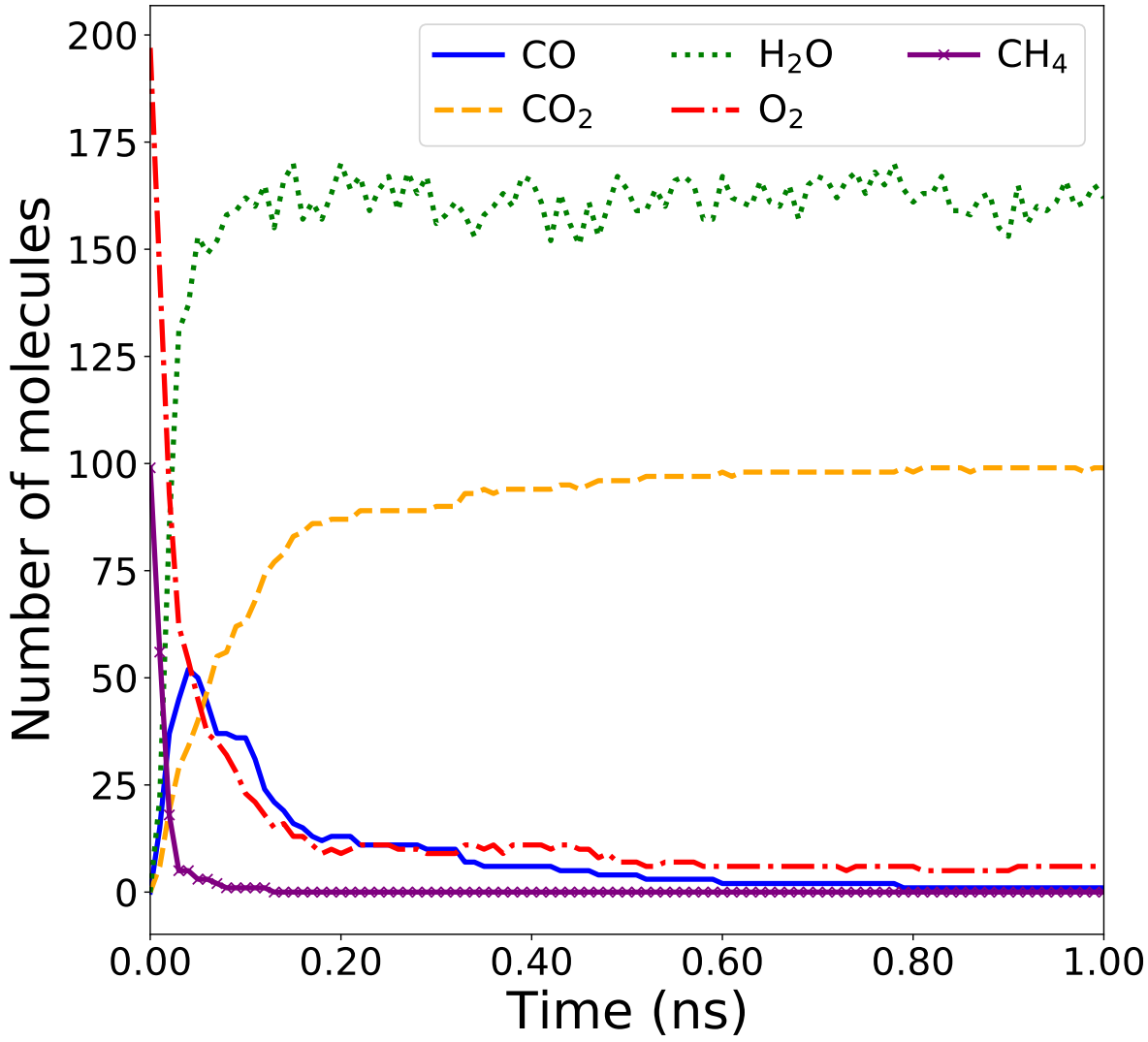


Figure S.6. Tracking plot of major products of the methane combustion simulations.

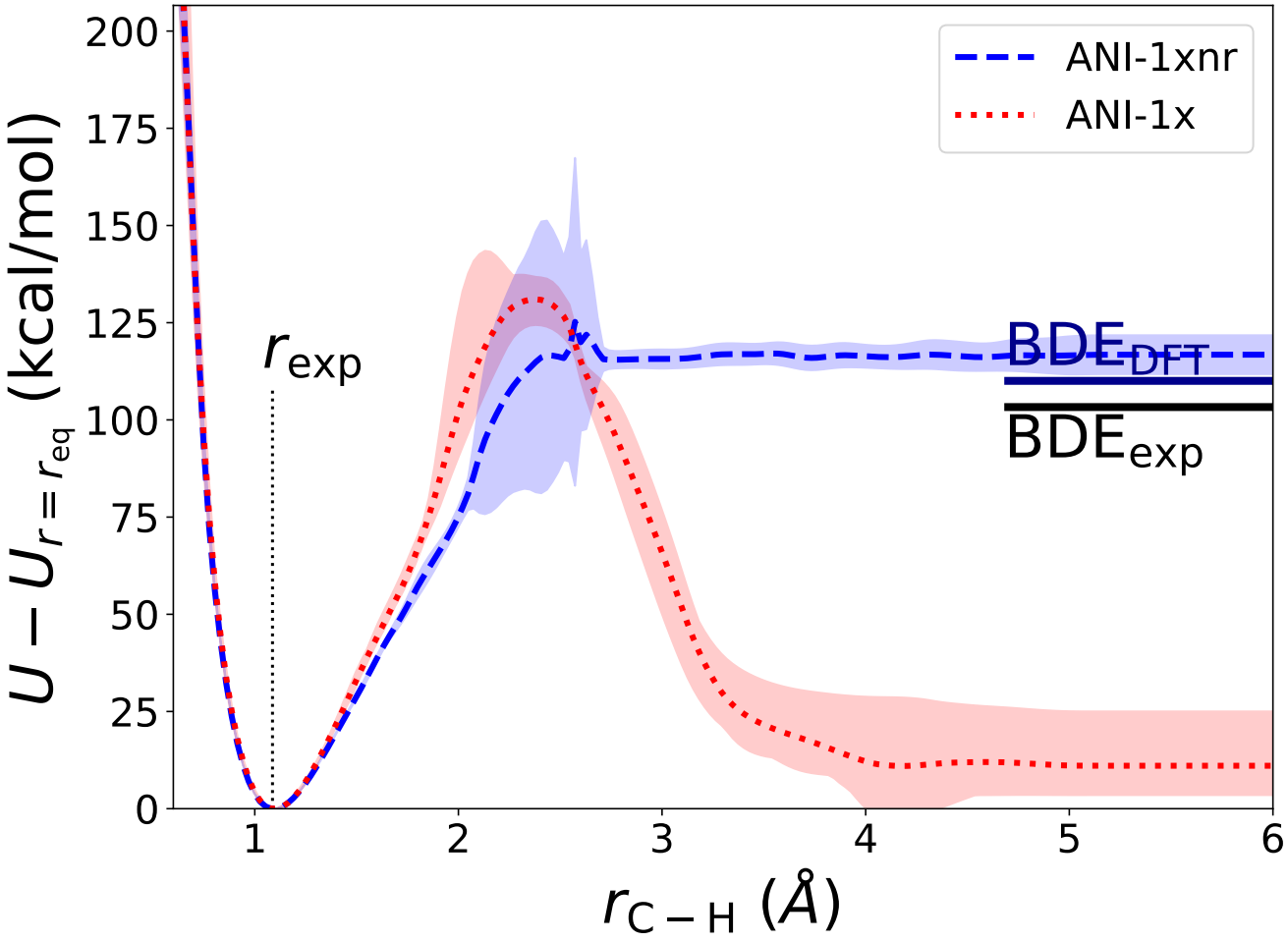


Figure S.7. Bond dissociation diagram for C-H bond in methane. Comparison between ANI-1xnr and ANI-1x with the DFT and experimental bond dissociation energies (BDE). Shaded regions depict the uncertainty, i.e., the ensemble standard deviation.

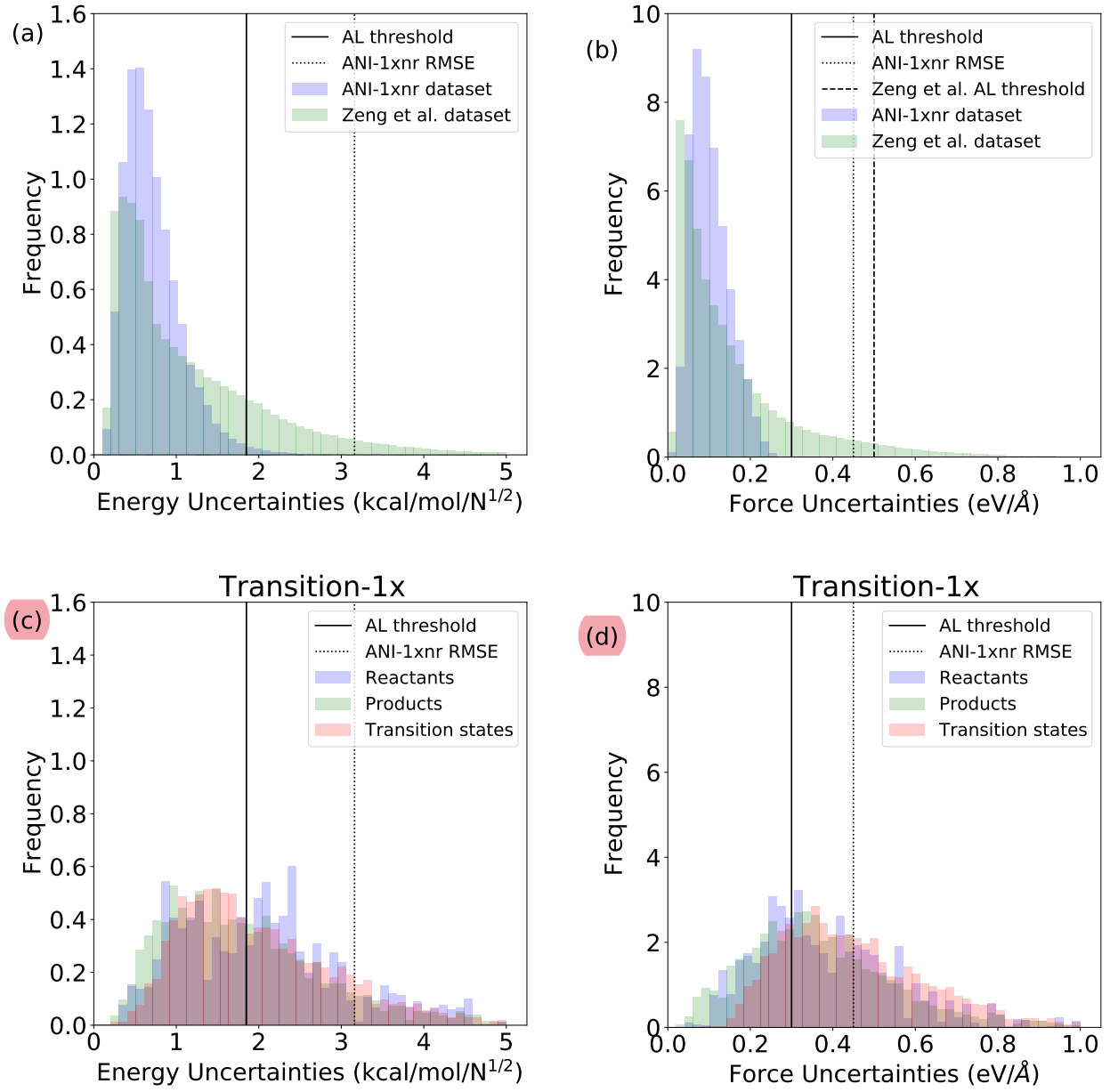


Figure S.8. Distribution of ANI-1xnr ensemble uncertainties in (a/c) energy and (b/d) force for the (a/b) ANI-1xnr dataset, (a/b) Zeng et al. dataset for methane, and (c/d) Transition-1x reactive dataset. Energy uncertainty is defined as the ANI-1xnr ensemble standard deviation for energy normalized by the square-root of number atoms. Force uncertainty is the ANI-1xnr ensemble standard deviation for force averaged over all atoms and Cartesian coordinates. Most of the ANI-1xnr ensemble uncertainties for the Zeng et al. dataset are comparable to those of the ANI-1xnr training dataset. Specifically, the ANI-1xnr energy and force uncertainties are smaller than the AL thresholds of 1.85 kcal $\cdot$ mol$^{-1} \cdot N^{-\frac{1}{2}}$ and 0.3 eV/Å, respectively, for approximately 77% of the ≈567000 structures in the Zeng et al. dataset. Note that the AL selection criterion of Zeng et al. was based solely on the force uncertainty. The ANI-1xnr force uncertainty is larger than the Zeng et al. AL threshold of 0.5 eV/Å for only 5% of the structures in the Zeng et al. dataset, suggesting that ANI-1xnr has a similar confidence on this dataset as the application-specific MLIP of Zeng et al. Note that with such a high AL threshold, it is quite likely that there are some non-physical structures in the Zeng et al. dataset. Furthermore, the Zeng et al. dataset also consists of structures sampled with ReaxFF without curation. It is also important to recall that the Zeng et al. training dataset consists of small clusters of molecules extracted from a condensed-phase MD simulation. Therefore, atoms near the center of the cluster are in condensed-phase environments while atoms on the border of the cluster are effectively in a gas-phase environment. By contrast, the unilateral high uncertainties for the Transition-1x dataset demonstrate that ANI-1xnr is not intended for *in vacuo* reactive chemistry. Note that the distribution of Transition-1x uncertainties is nearly the same for reactants, transition states, and products, demonstrating that the issue is the vacuum environment rather than unsampled transition states.

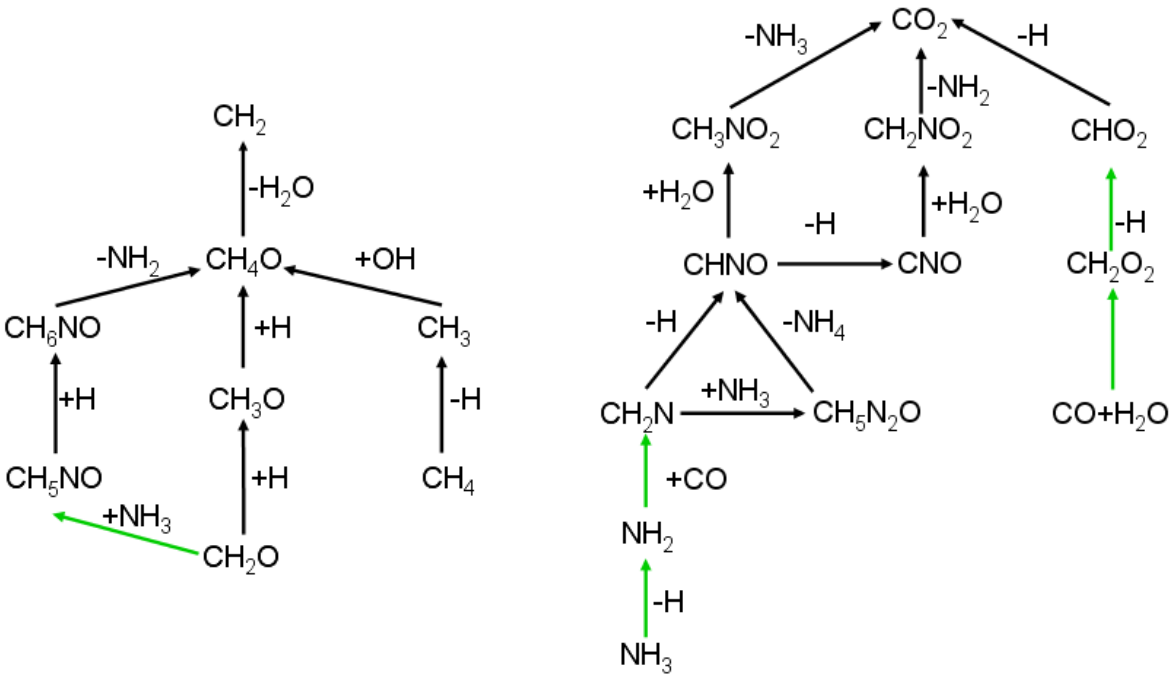


Figure S.9. Formation of intermediates CH_2 and CO_2 from initial reactants (NH_3, CO, CH_4, H_2O). Mechanism to form CH_2O is found in Figure 7 in the main text. Green arrows denote reactions previously identified by Wang et al. or Saitta and Saija. Orange arrows denote reactions that have closely-related reactions in Wang et al. or Saitta and Saija.

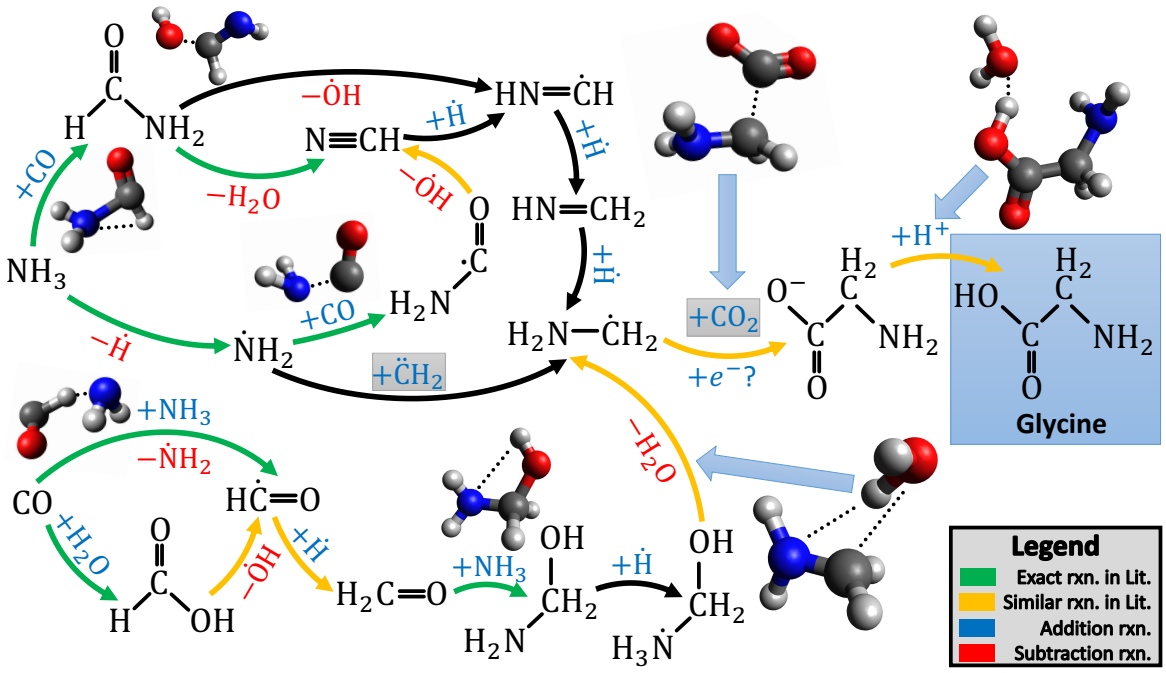


Figure S.10. An alternative mechanism for the formation of glycine in the ANI-1xnr Miller simulation. In this pathway, the final step to form glycine involves H-abstraction from H_3O, which is likely a cationic species (H_3O^+). The penultimate species ($C_2H_4NO_2^-$) formed prior to glycine, therefore, cannot be unambiguously labeled as an anion or a radical. The uncertainty regarding the ionic nature of this mechanism illustrates an issue with electron-agnostic ML potentials. The depiction of bond orders, charges on ions, and radical species is based simply on chemical intuition, since ANI-1xnr does not provide explicit orbital or electronic information.

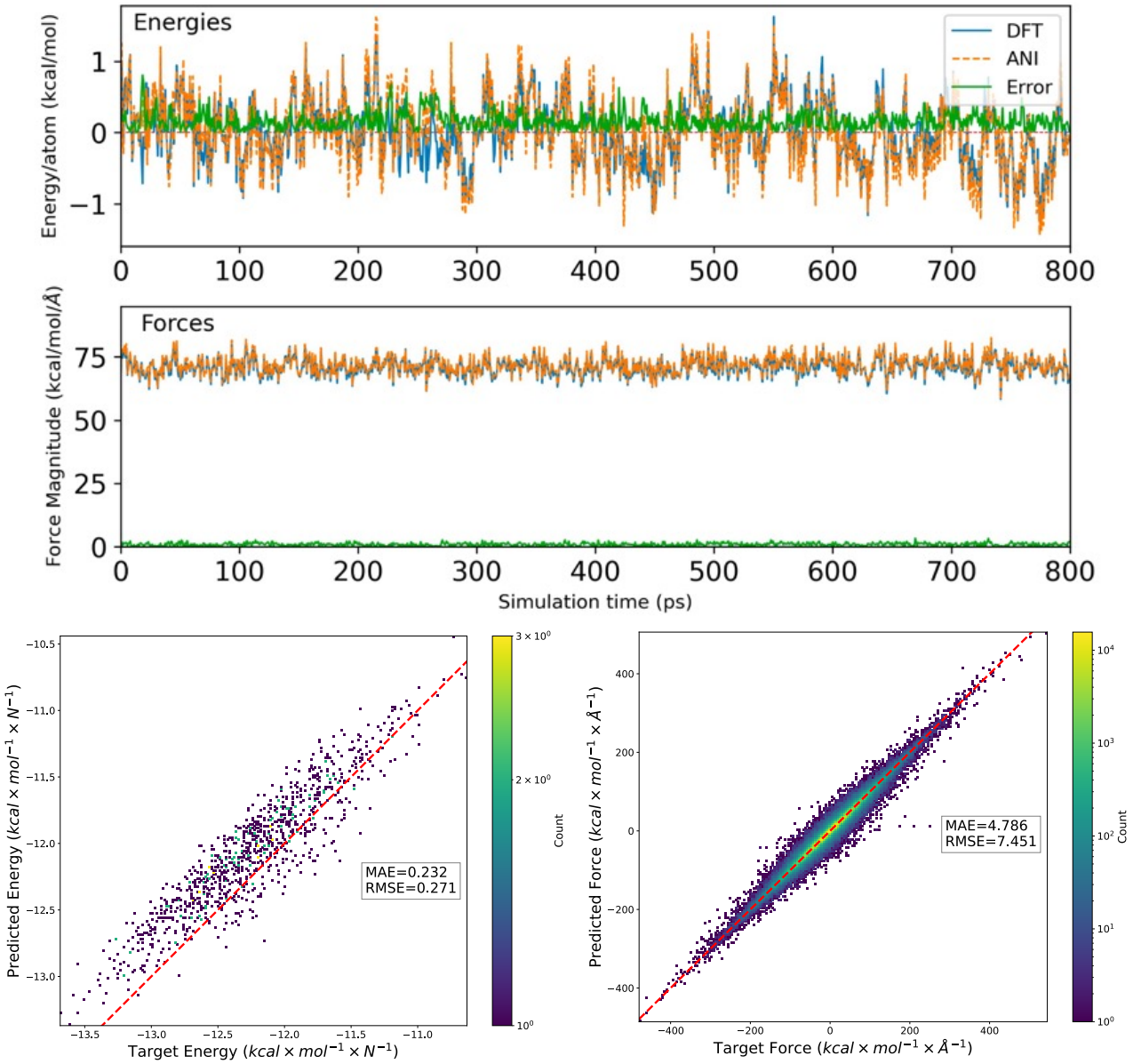


Figure S.11. Validation of Miller Experiment simulation. Comparison between DFT energies and forces with ANI-1xnr for the first 800 ps. Data were not used in training of ANI-1xnr.

The ANI neural networks used in this work were implemented in the NeuroChem C++/CUDA software package. A batch size of 32 was used while training the ANI-1xnr model. A weight of 1.0 was used on both the energy and force loss term. Learning rate annealing was used during training, starting at a learning rate of 0.001 and converging at a learning rate of 0.00001. The ADAM update algorithm is used during training. The network architecture is provided in Table S.II. The symmetry function parameters are provided in Table S.III.

	H		C		N		O	
Layer ID	Nodes	Activation	Nodes	Activation	Nodes	Activation	Nodes	Activation
1	256	CELU	224	CELU	192	CELU	192	CELU
2	192	CELU	190	CELU	160	CELU	160	CELU
3	160	CELU	160	CELU	128	CELU	128	CELU
4	1	Linear	1	Linear	1	Linear	1	Linear

Table S.II. ANI-1xnr neural network architecture

Radial Cutoff (Radial) (Å)	5.2
Radial Cutoff (Angular) (Å)	3.5
Radial Eta ($Å^{-2}$)	65.7
Radial Shift (Å)	0.500000,0.646875,0.793750,0.940625, 1.087500,1.234375,1.381250,1.528125, 1.675000,1.821875,1.968750,2.115625, 2.262500,2.409375,2.556250,2.703125, 2.850000,2.996875,3.143750,3.290625, 3.437500,3.584375,3.731250,3.878125, 4.025000,4.171875,4.318750,4.465625, 4.612500,4.759375,4.906250,5.053125
Angular Zeta (-)	14.1
Angular Angular Shift (rad.)	0.39269908,1.1780972, 1.9634954,2.7488936
Angular Eta ($Å^{-2}$)	10.1
Angular Radial Shift (rad.)	0.500,0.875,1.250,1.625, 2.000,2.375,2.750,3.125

Table S.III. ANI-1xnr symmetery function parameters

Property	RMSE	MAE
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.43 ± 0.22	0.1756 ± 0.0061
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.34 ± 0.25	6.306 ± 0.074

Table S.IV. Model performance against held-out test dataset. Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported as the average of eight ensemble models with the corresponding standard deviation.

Property	RMSE	MAE
Ensemble member 1		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.37	0.178
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.19	6.24
Ensemble member 2		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.23	0.173
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.39	6.33
Ensemble member 3		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.26	0.170
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.39	6.38
Ensemble member 4		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.22	0.169
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	9.95	6.19
Ensemble member 5		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.45	0.177
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.17	6.25
Ensemble member 6		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.23	0.175
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.21	6.33
Ensemble member 7		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.24	0.172
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.57	6.43
Ensemble member 8		
Energy $(\text{kcal}\cdot\text{mol}^{-1}\cdot N^{-1})$	0.91	0.189
Force $(\text{kcal}\cdot\text{mol}^{-1}\cdot \text{Å}^{-1})$	10.81	6.29

Table S.V. Ensemble member performance against held-out test dataset. Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported for each individual ensemble member. Recall that each ensemble member is evaluated against a different held-out dataset. Therefore, although ensemble member 8 has a higher energy RMSE (with an MAE that is similar to other ensemble members), this appears to be the result of some large outliers, potentially from some poor DFT results (see Figure S.14).

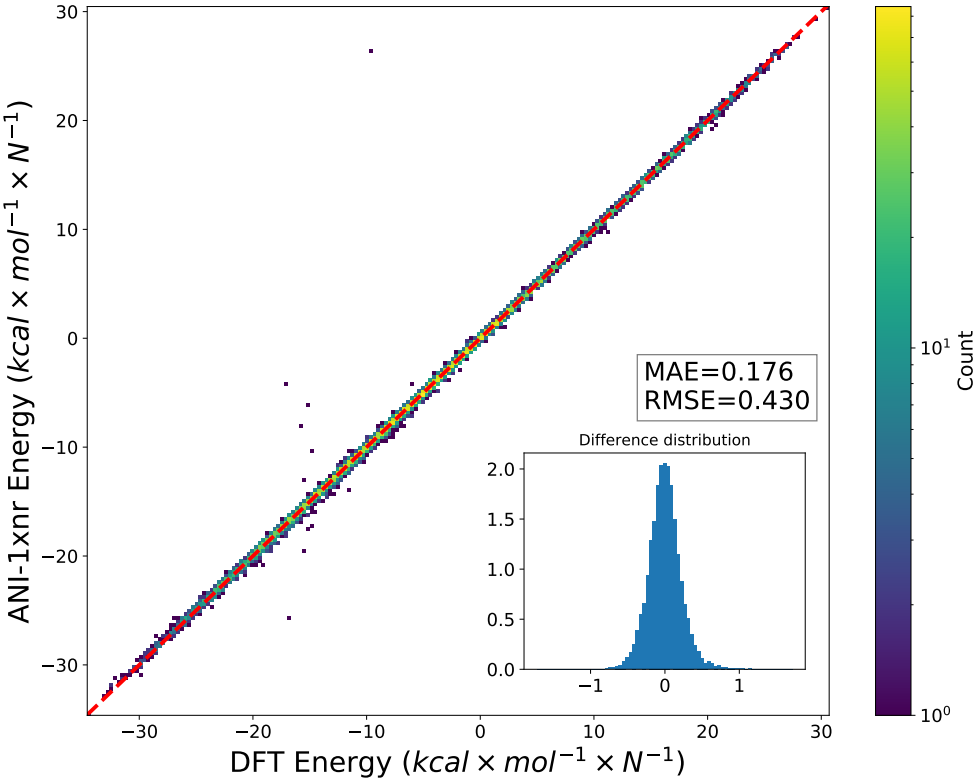


Figure S.12. Energy correlation plot. Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported as the average of eight ensemble models.

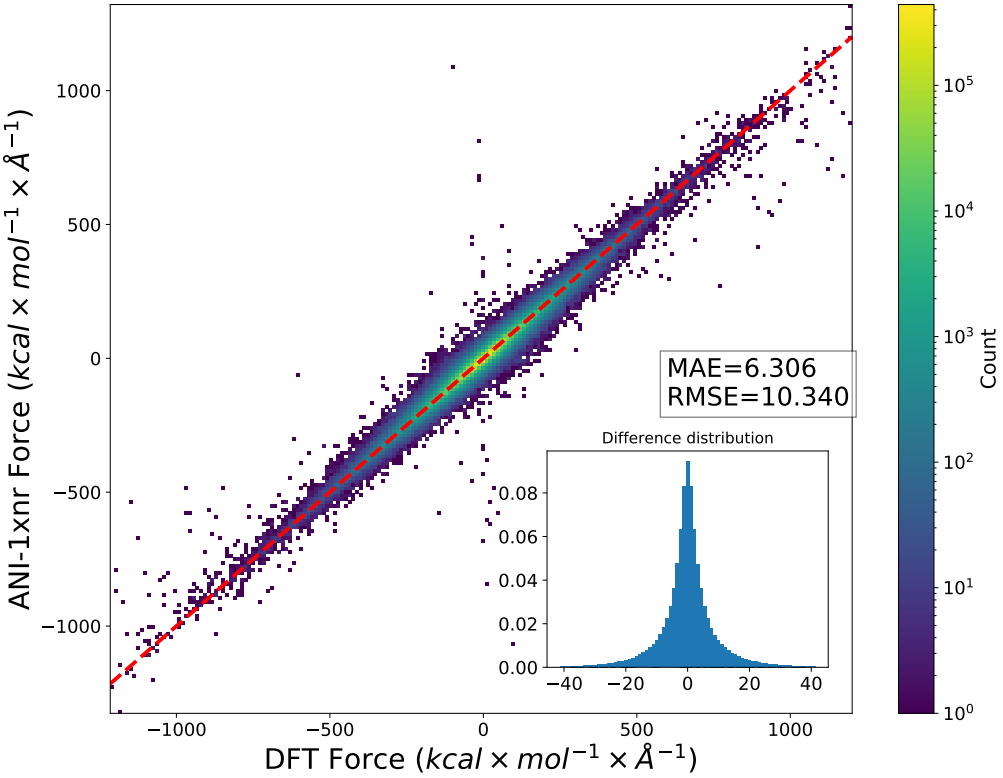


Figure S.13. Force correlation plot. Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported as the average of eight ensemble models.

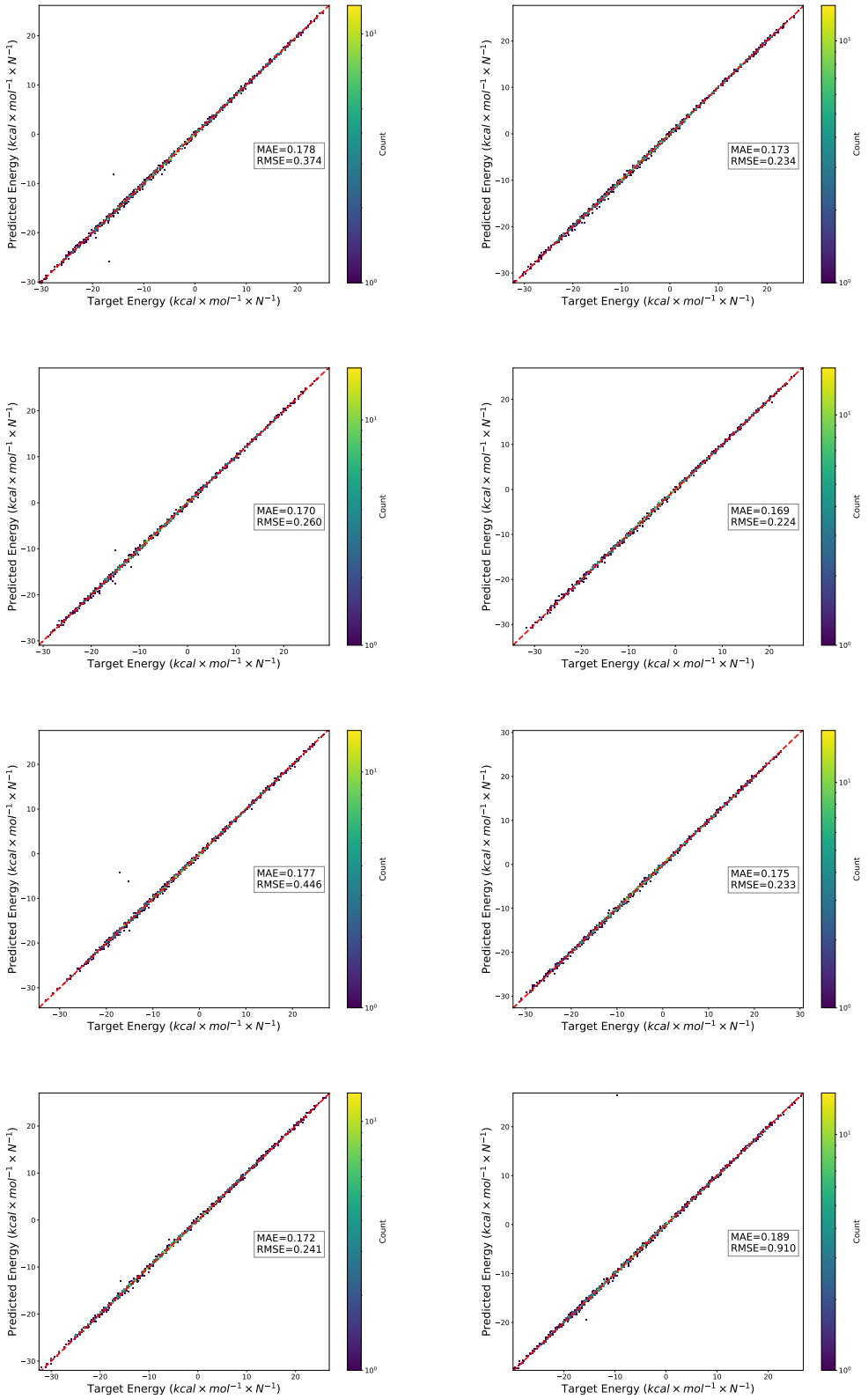


Figure S.14. Energy correlation plots for all ensemble members (members 1-8 from top-left to bottom-right). Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported for each individual ensemble member. Recall that each ensemble member is evaluated against a different held-out dataset. Therefore, although ensemble member 8 has a higher energy RMSE (with an MAE that is similar to other ensemble members), this appears to be the result of some large outliers, potentially from some poor DFT results.

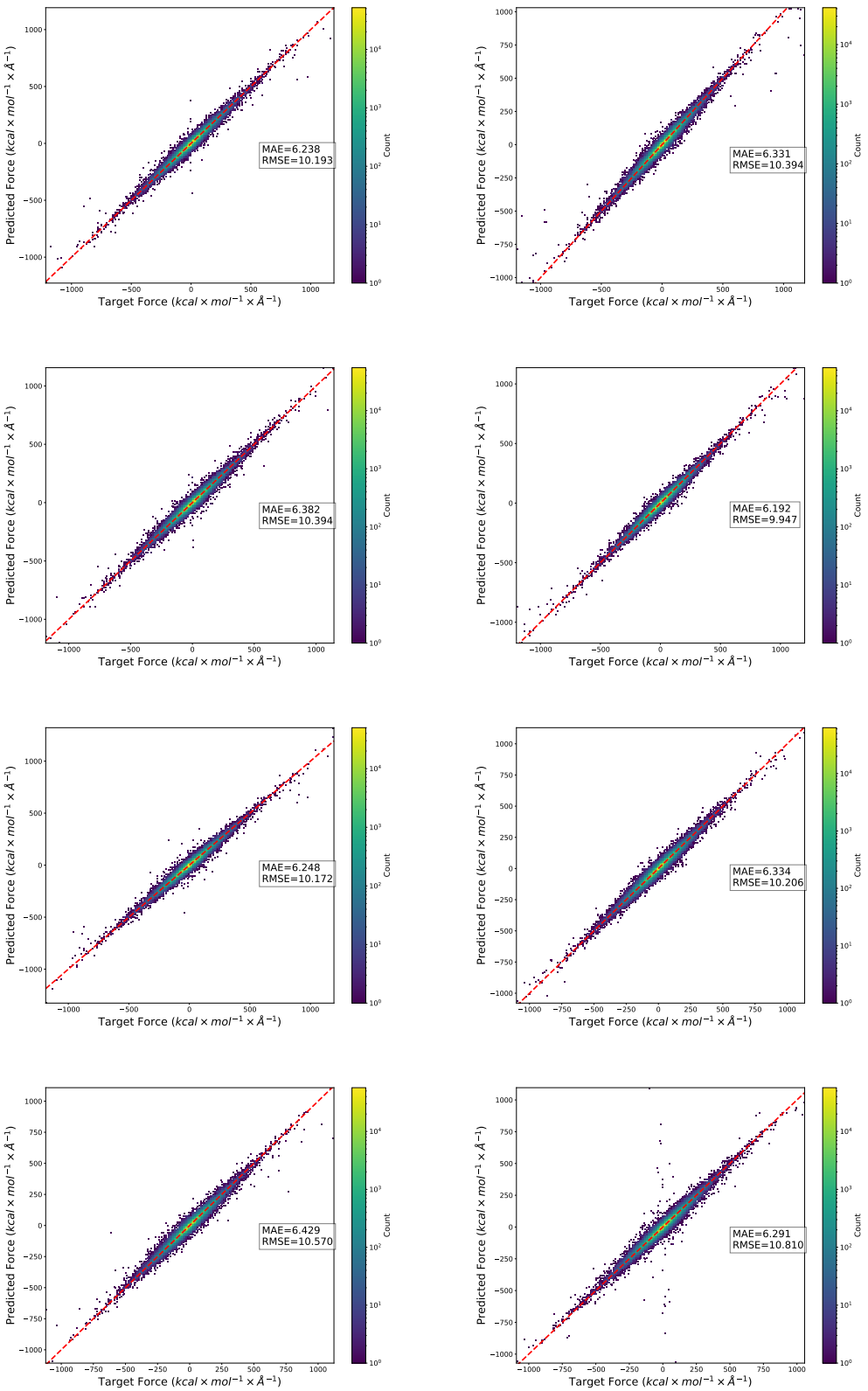


Figure S.15. Force correlation plots for all ensemble members (members 1-8 from top-left to bottom-right). Root-mean-squared-error (RMSE) and mean-absolute-error (MAE) are reported for each individual ensemble member. Recall that each ensemble member is evaluated against a different held-out dataset.

Parameter	Range
T_{start}	1000 - 3000 K
T_{end}	100 - 2000 K
T_{amp}	0 - 2000 K
ρ_{start}	0.1 - 2 g/cc
ρ_{end}	0.5 - 2 g/cc
ρ_{amp}	0 - 0.75 g/cc
t_{per}	T: 2 - 50 ps; ρ: 0.5 - 50 ps

Table S.VI. Parameters for nanoreactor oscillations in temperature and density.

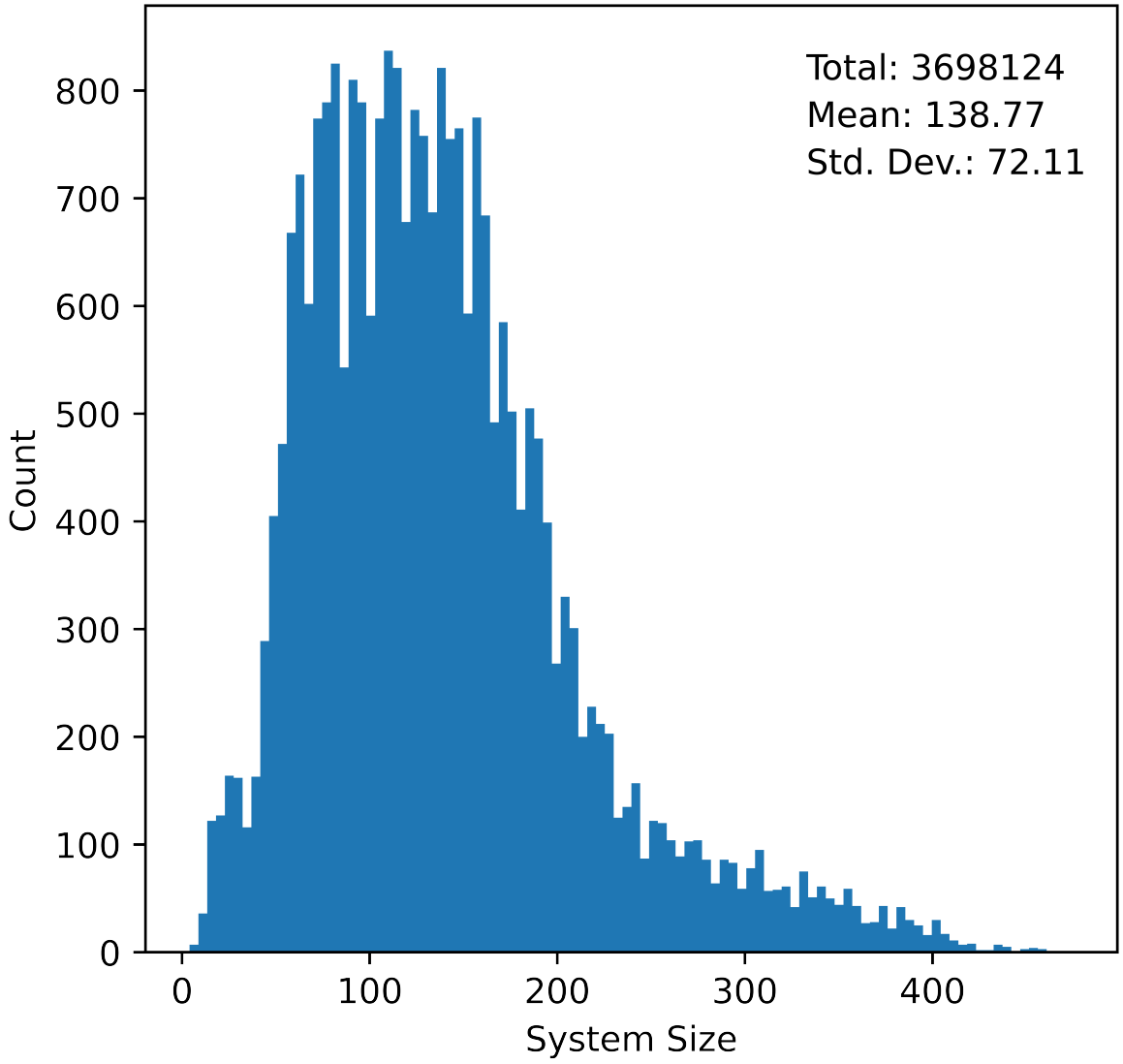


Figure S.16. Histogram of the system size (i.e., number of atoms) per system in the ANI-1xnr training data set.

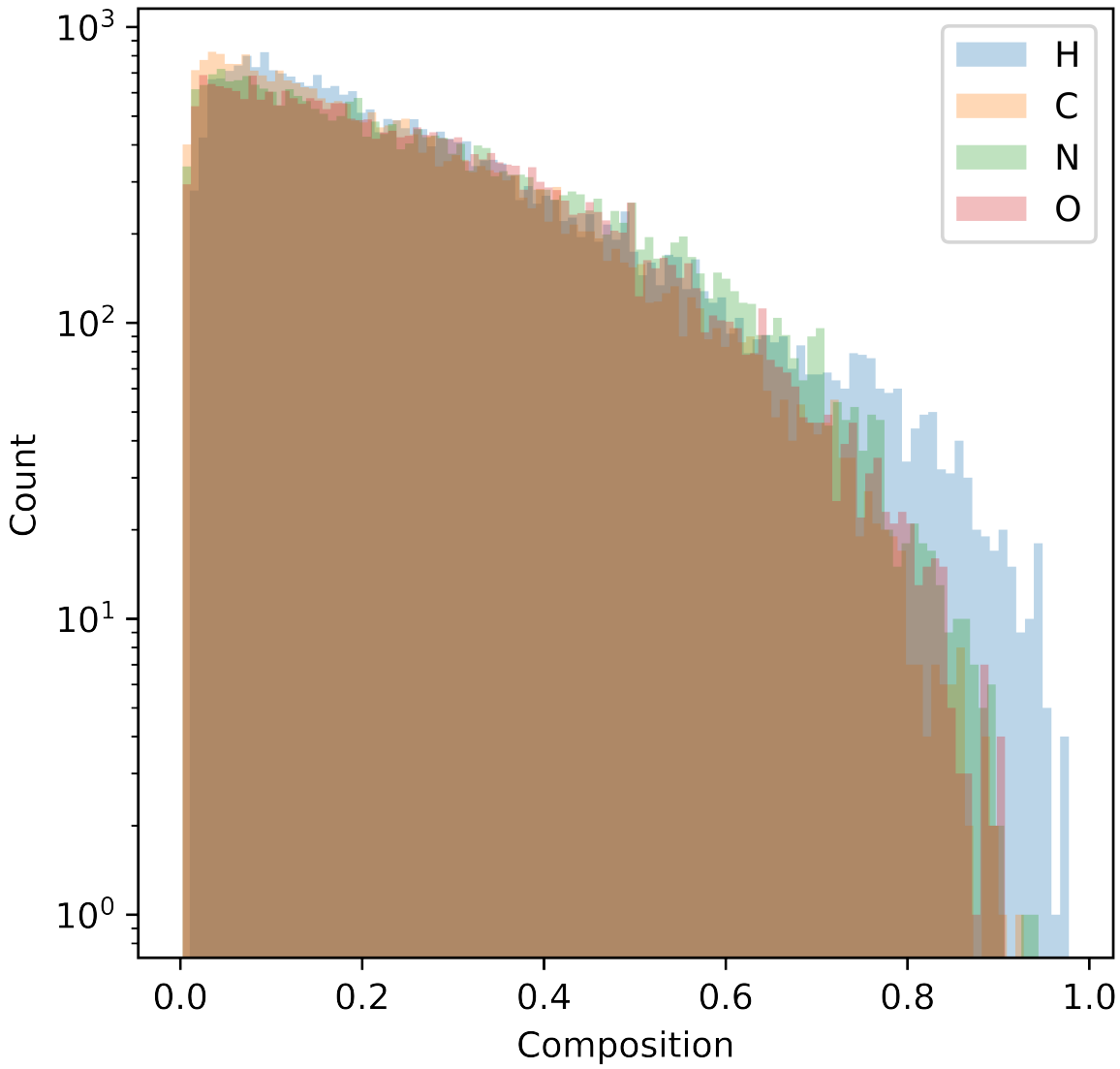


Figure S.17. Histogram of the system composition of all systems in the training data set, colored by element.

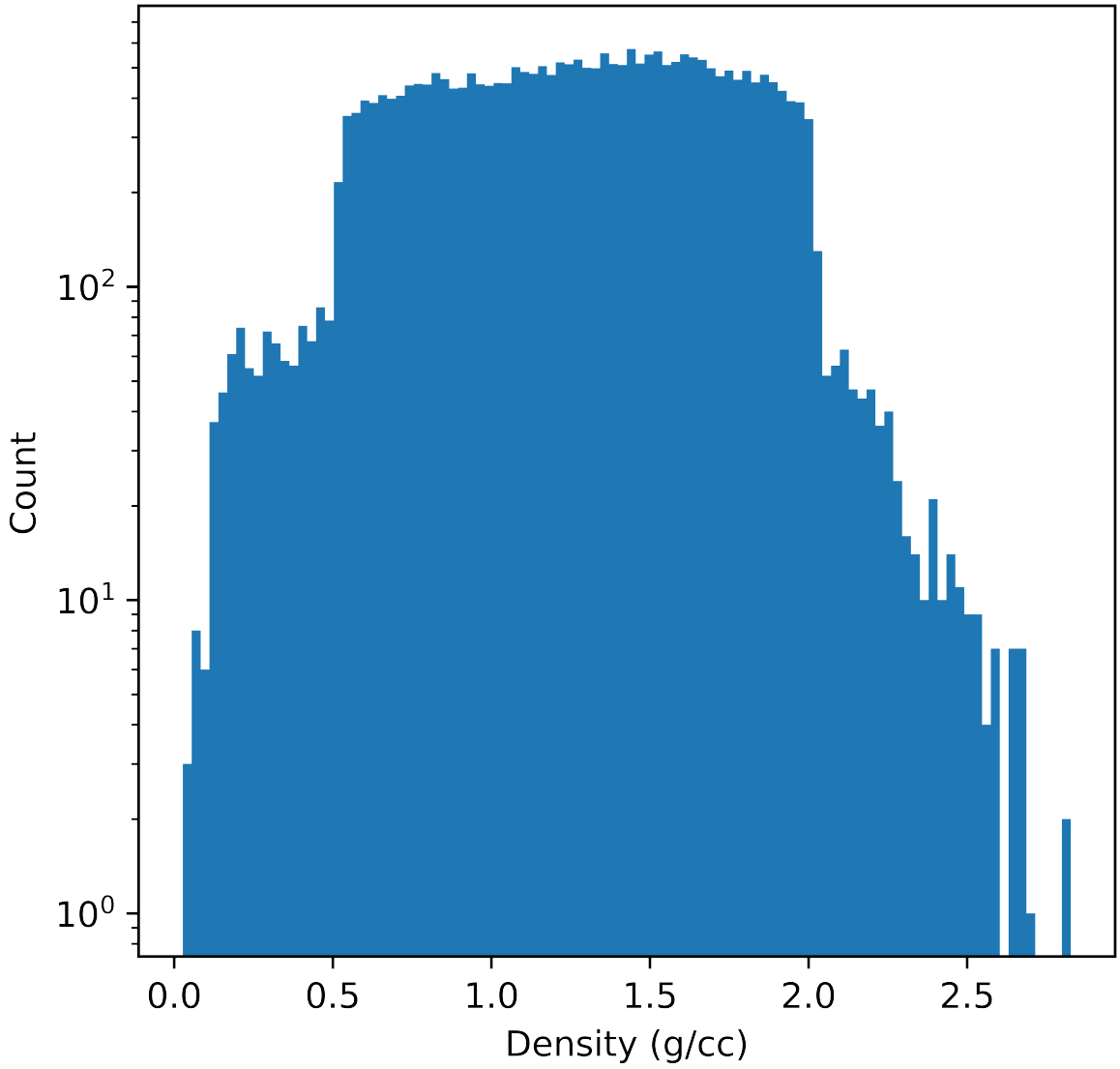


Figure S.18. Histogram of the mass density (g/cc) of all systems in the training data set.

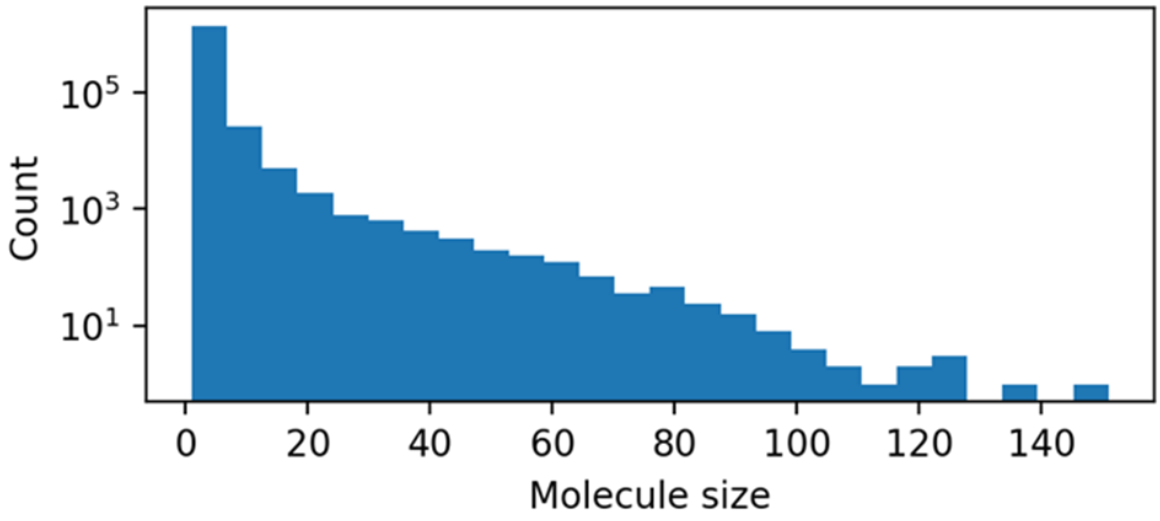


Figure S.19. Distribution of the molecule size (i.e., number of heavy atoms) in the ANI-1xnr training set.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- CodeDemoForNature.zip