Exploring the frontiers of 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

ABSTRACT

Reactive chemistry atomistic simulation has a broad range of applications from drug design to energy to materials discovery. Machine learning interatomic potentials (MLIP) have become an efficient alternative to computationally expensive quantum chemistry simulations. In practice, reactive MLIPs require refitting to extensive datasets for each new application, and prior knowledge of reaction networks is required to generate fitting data. In this work, we develop a general reactive MLIP through unbiased active learning with a nanoreactor molecular dynamics inspired sampler. The resulting potential (ANI-nr) is then applied to study five distinct condensed phase reactive chemistry problems: 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-nr closely matches experiment and/or previous studies using traditional model chemistry methods, without needing to be refit for each application, which enables high-throughput *in silico* reactive chemistry experimentation.

Introduction

Atomic-scale reactive simulation is a valuable tool for chemists to plan reaction synthesis or measure material properties *in silico*. Traditionally, physics-based computational methods such as classical force fields (FF) or quantum mechanics (QM) provide the forces acting on atoms. These forces are used to simulate the motion of atoms with Newtonian physics, i.e., molecular dynamics (MD) simulation. Researchers have historically aimed to improve QM and FF algorithms to minimize computational costs and better model the underlying physics for improved accuracy. In the last few decades, low-cost reactive FF methods have contributed significantly to computational reactive chemistry research. Approaches such as empirical valence bond,[1] modified embedded atom method (MEAM),[2] reactive force field (ReaxFF),[3] and reactive empirical bond order (REBO)[4] are often applied to describe the making and breaking of chemical bonds during reactive atomistic simulations. These reactive FF methods use pre-determined physically-inspired functional forms with a small number of model parameters to approximate the potential energy surface (PES). While these methods prove 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 contributes to the expertise and labor required to perform this research, and potentially results in human bias concerning which reactions proceed. QM methods, such as density-functional theory (DFT), are transferable, that is, they are applicable for estimating energies and forces for a wide range of systems without reparameterization, since these calculations are based on the underlying physics of electronic structure theory, rather than a pre-defined bonding pattern. However, the computational cost of QM methods is prohibitive for many 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 (MLIP) have been proposed as an alternative to QM and FF methods for the prediction of potential energies and forces.[6–17] 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 PES[6] of atomic systems. In their method, atomic symmetry functions represent the local chemical environment of each atom. These symmetry functions are 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. These concepts have been applied in building a range of MLIPs for different applications, such as SSW-NN,[18] TensorMol,[19] N2P2,[20] and AMP.[21] The ANAKIN-ME (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.[22–25] While the ANI MLIPs proved to be extremely general and accurate for near-equilibrium conformations of gas phase organic molecules, these potentials do not address the challenges of condensed phase (i.e., periodic systems of liquids or solids) reactive chemistry.

MLIPs have been successfully applied to model chemical reactions in specific contexts,[26,27] for example, unimolecular/bimolecular gas-phase reaction pathways[28–34] and specific condensed-phase reactive chemistry simulations.[35] However, each application necessitates a new data set and retraining of the MLIP.[36] 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 would be transformational. However, a remaining major roadblock to developing a general reactive MLIP is that it requires humans to know *a priori* which reactions should be included to produce the ideal training data set. As such, these reactive MLIPs mirror many of the limitations of reactive FFs. Recent endeavors to build large data sets including reactions have yielded groundbreaking results for developing a general-purpose MLIP.[37] However, random sampling can produce models with poorer performance than targeted, model-aware sampling strategies.[22]

Active learning (AL)[38] is a class of algorithms designed to automatically sample, select, and label new data with the goal of efficiently generating a diverse and relevant data set 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 data set by defining an algorithm for data collection. Starting from a small initial bootstrapping data set, the algorithm is applied iteratively, yielding generations of data sets designed to improve upon their ancestors. It has been applied to develop numerous MLIPs in recent years.[22,26,39–44] A sampling scheme (often MD-based, and often using an MLIP trained to the current AL data set) 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 (UQ) approach (such as query-by-committee[45]), aiming to include only high-uncertainty structures in the ever-growing data set. The labeling of selected configurations with energies and forces is accomplished by performing QM calculations, then these new data are added to the training data set. Iterations of training, sampling, selecting, and labeling are performed until the resulting ML potential performs as desired or is no longer improving.

For reactive chemistry, 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,[46] 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) developed an automated approach to sample transition states and minimum-energy-path structures for gas-phase reactions,[47] an alternative sampling procedure designed to enable reliable condensed-phase reactive MD simulations is essential.

Wang and co-workers developed an elegant but expensive approach for the MD-based exploration of reaction pathways in the condensed phase, known as the *ab initio* nanoreactor (NR).[48,49] 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. Wang et al. applied the NR to observe graphene ring formation from pure acetylene. Wang et al. also showed that this approach was able to discover a reaction pathway from small early-earth molecules to glycine, one of the building blocks of life today. Although Wang et al. clearly demonstrated the promise of the NR to discover reactive chemistry, the current *ab initio* NR approach is extremely computationally expensive. Specifically, a 1-ns *ab initio* NR simulation required 132,400 graphics processing units (GPU) hours, despite using the relatively low-level Hartree Fock (HF) method and a minimal basis set (3-21G).

Inspired by the work of Wang et al., we design an AL sampling procedure based on the NR that targets arbitrary reactive chemical processes and compositions of H, C, 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. To evaluate the final model, which we call ANI-nr, in practical research scenarios, we conduct several condensed-phase reactive chemistry simulations with the ANI-nr 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

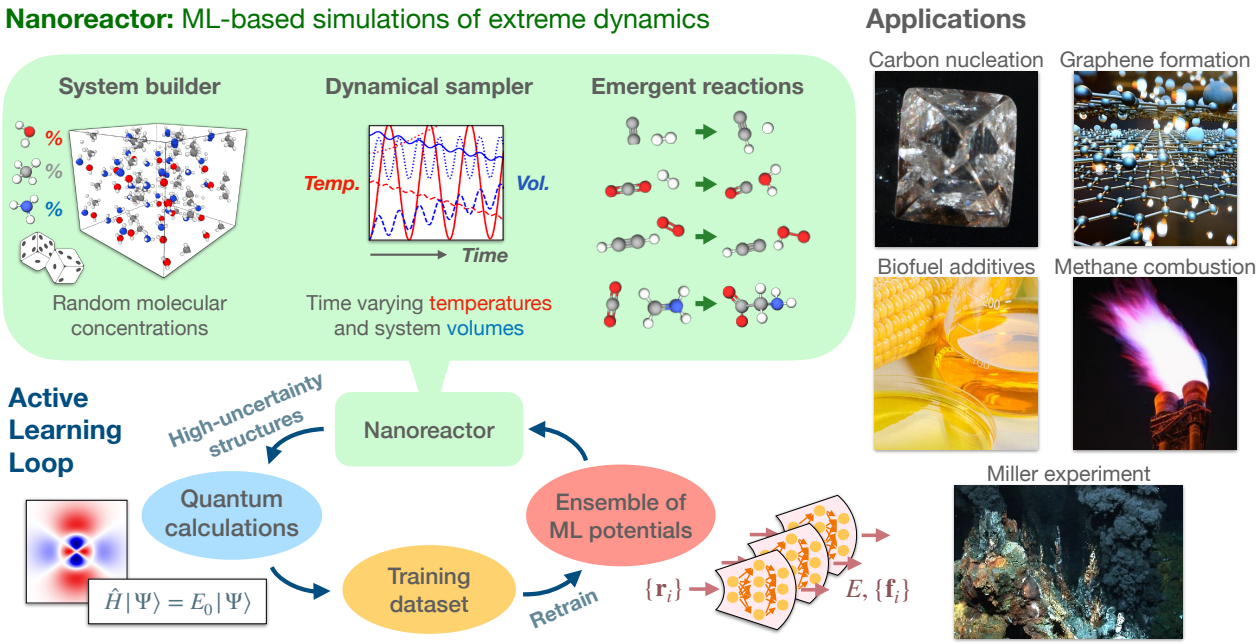

Figure 1. Summary of the nanoreactor active learning workflow and specific applications.

spontaneous formation of glycine from early-earth molecules. Across this wide range of applications, we show ANI-nr provides results that are consistent with chemical intuition, experimental data, QM calculations (HF, DFT, and density functional based tight binding, DFTB), and classical reactive MD simulations (ReaxFF and a non-transferable MLIP). This study demonstrates the capability of automated chemical exploration workflows to build a general-purpose reactive potential, resulting in ANI-nr, an accurate and transferable potential capable of simulating a wide range of real-world reactive systems.

Results

We begin by evaluating the resulting training data set produced by our active learning nanoreactor scheme. Then, to show the state-of-the-art transferability of the ANI-nr potential, we apply ANI-nr in five distinct case studies. The five case studies were chosen to allow for a diverse comparison with common reactive chemistry methods. The five case studies are: pure carbon simulated under varying pressures produces the expected structures, simulations of carbon ring formation in varying concentrations of O_2 are compared with DFTB simulations, biodiesel fuel additive simulations are compared with ReaxFF, methane combustion simulations are compared with an application-specific MLIP, and the spontaneous formation of glycine from early earth compounds is compared with *ab initio* MD nanoreactor and DFT-MD simulations.

Nanoreactor active learning

One of the most challenging tasks in developing a data set for training neural network potentials is knowing what data to include in the training data set. This problem is compounded when developing reactive models. Given the flexibility of neural networks, it is not good enough to include only the reactants, products, or transition state optimized structures, or even the minimum energy path between states, in training data sets. Rather, if the potential is to accurately model reaction dynamics, the potential energy surface up to some relative energy that corresponds with structures sampled at the temperature of interest must be included. For a more general model, the temperature of interest is not known in advance. Therefore, we employ active learning combined with a nanoreactor molecular dynamics sampler to generate the ANI-nr training data set. The nanoreactor sampler randomly initializes condensed phase systems of small molecules, then applies oscillatory temperatures and density during molecular dynamics to promote reactions and the formation of new products. Over more than 50 iterations of active learning, more than 26,000 condensed systems with random C, H, N, and O compositions were selected and labeled with density functional theory energies and forces. Figure 1 shows a diagram of the active learning process used in this work. For a detailed description of our active learning process and sampler see the methods section.

Figure 2 a) through d) shows T-distributed Stochastic Neighbor Embeddings (TSNE) of atomic environment descriptors for each element calculated for a random subset of atoms in the training data set. The blue points represent atomic environments in the ANI-nr data set, while the red points represent atomic environments from the ANI-1x data set. As expected, since ANI-1x is

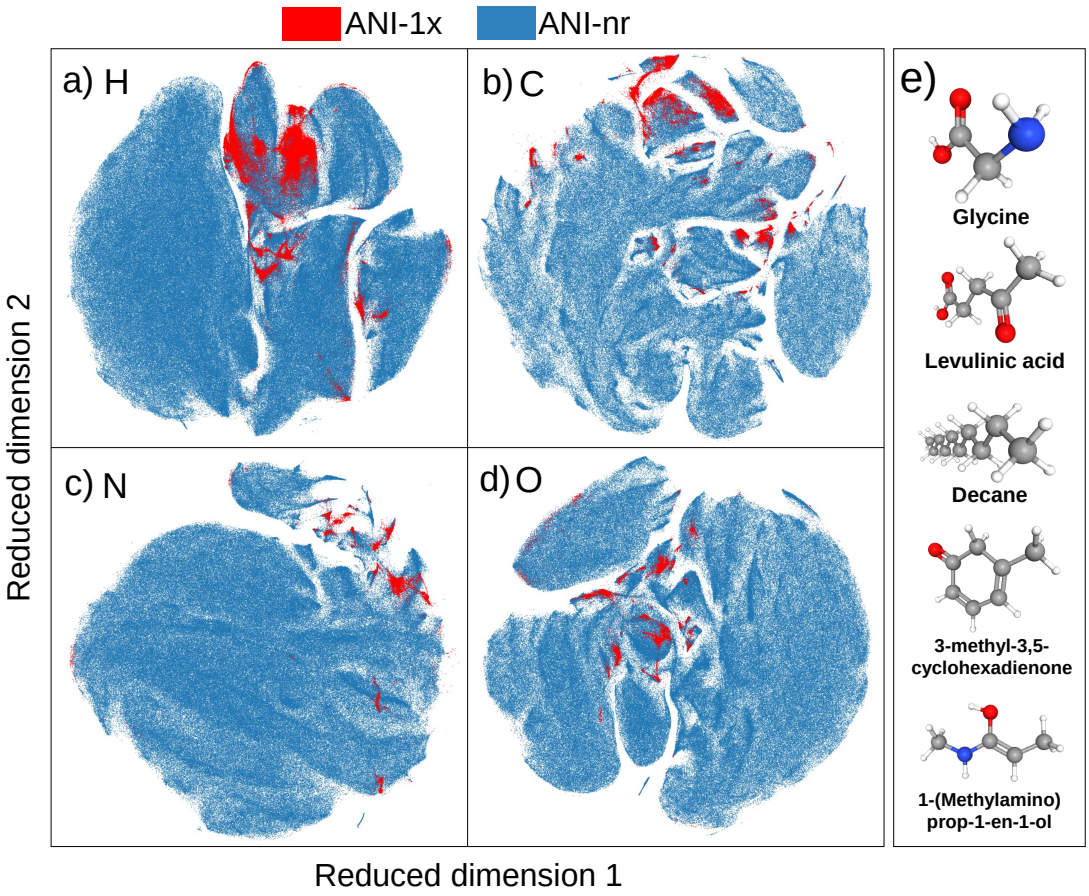

Figure 2. Inspection of the nanoreactor data set. Panels a), b) c), and d) show 2D visualizations of the T-distributed Stochastic Neighbor Embeddings (TSNE) of the atomic descriptors from a random subset of the ANI-1x data set (red) and ANI-nr data set (blue). Panel e) shows five examples of the 1212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

a non-reactive near equilibrium molecule data set, the ANI-nr data set covers effectively all the space covered by ANI-1x, plus the pathway between many of the clusters in the ANI-1x data set. The pathways that merge ANI-1x clusters represent reactions in a low dimensional representation. Further analyzing the ANI-nr data by searching all data points for molecules that also exist in the PubChem database yields 1212 unique known PubChem molecules (only molecules with < 10 CNO atoms) in the training data set. Figure 2 e) shows examples of the known PubChem molecules that exist in the active learning generated data set. Since all initial systems started from random placement of small, 1 to 2 non-H atoms, the reaction pathways to produce these 1212 unique molecules must have been visited during the active learning process. In this way, the active learning process has automatically discovered these reaction pathways.

Carbon solid-phase nucleation

Simulations of amorphous carbon have long been one of the top interests among chemists and materials scientists, as some distinctive materials like graphene, diamond and carbon nanotubes form from amorphous carbon systems under different conditions. Understanding the behavior of amorphous carbon under different conditions would help us to develop functional materials by controlling the growth process. Many reactive FFs like REBO[50] and ReaxFF[51] have been employed to simulate amorphous carbon in MD. With the widespread use of ML methods, researchers are now also trying to investigate the amorphous carbon system with trainable MLIPs.

At present, most of the existing literature is relying on selective sampling when building the training set for the MLIP, resulting in a bespoke potential for a specific application. Wang. et. al[52] trained an NN-based MLIP on a dataset obtained by randomly modifying low-dimensional carbon structures and taking snapshots from AIMD simulations of liquid carbon. The

MLIP trained on this dataset proved to be able to predict pure carbon fragments with the desired accuracy. Deringer et al.[53] trained a Gaussian approximation potential (GAP) on structures sampled from DFT-based MD on liquid carbon systems. The GAP model was validated on atomic energy and forces. Furthermore, the GAP model was used to predict some 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. To validate ANI-nr 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 the system with the highest 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, and after another 2.3 ns the 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 after another 2.3 ns the system contains 83% graphene. 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 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/partial closed meshes. We note that, for each of the high-, medium- and low-density carbon simulations, ANI-nr produces the expected structure of carbon for the respective density.

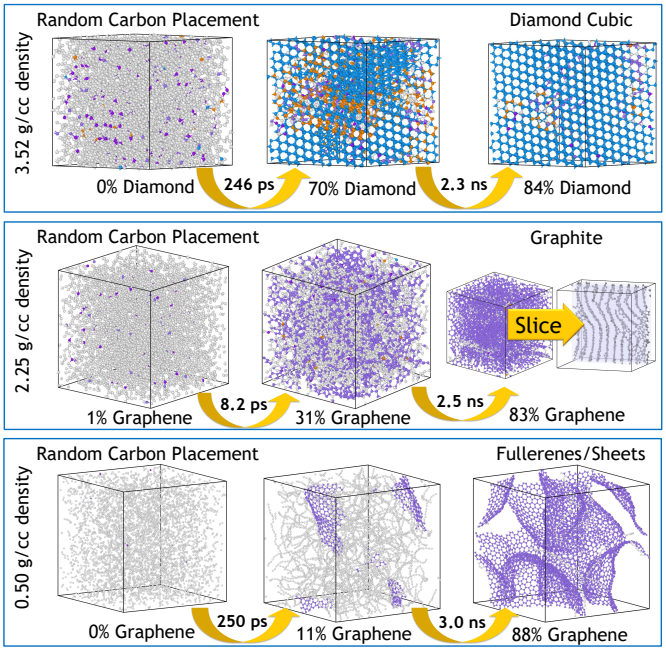

Figure 3. Results of ANI-nr carbon simulations starting from random carbon positions at three densities: 0.5g/cc, 2.25g/cc, and 3.52g/cc.

In Table 1, we present lattice constants for 3 ANI-based MLIPs and experimental lattice constants. ANI-nr reproduces the diamond cubic lattice constants with an error of only 0.01 Å. For graphite, ANI-nr predicts a and b lattice constants also with an error of 0.01 Å, however, the c lattice constant (along the direction of π-π stacking in graphite) is predicted with an error of 0.47 Å. This error is likely due to ANI-nr being a short-range model 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-nr potential (5.2/3.5 Å), called ANI-nr(lr). The longer-range model performed significantly better than the original ANI-nr on the c lattice parameter with an error of 0.15 Å while also reducing the 0.01 Å error for the a and b lattice parameters. 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

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

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

functions on the longer-range MLIP, but this would greatly impact the computational speed of the model. A more optimal solution would be to add an explicit dispersion correction to ANI-nr that captures long-range interactions while maintaining an accurate description of the local environment.[54] We also compare with ANI-2x,[55] 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 data set used to train ANI-2x does not contain any structures similar to either of these systems. Furthermore, in contrast to the ANI-2x data set reference calculations, the reference calculations used for building the ANI-nr data set includes dispersion corrections (see methods for details), which are essential to reproduce the c lattice parameter in graphite. Finally, all models accurately predict the non-orthogonal experimental cell angles for diamond and graphite (see Table S.I in Supporting Information).

Effect of oxygen on graphene ring formation

Wang et al.[48] 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.[56] 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.[56]

Similar to the work of Lei et al.,[56] 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-nr 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 simulation from Lei et al. and experimental literature.[57] However, in contrast with Lei et al., a significant number of six-member rings form even for a 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-nr 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-nr compared with DFTB. For example, 6-membered rings begin to form in the O_2/C_2H_2=0.5 system after 1 ns with ANI-nr. 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. This case study demonstrate the value in the significantly lower computational costs of ANI-nr compared to traditional methods, such as DFTB, to discover interesting phenomena that can only be observed during long time-scale simulations.

To verify that the ANI-nr predictions of ring formation are reliable, we compute the NN ensemble disagreement throughout the course of each simulation (see Figure S.1 in Supporting Information). The ensemble standard deviation in energy normalized by the square-root of number atoms (ε) is lower than the AL energy threshold ($0.03\ \text{kcal}\cdot\text{mol}^{-1}\cdot N^{-\frac{1}{2}}$) through nearly the entire simulation (with only a few snapshots as exceptions), which confirms that each system is well-modeled by our MLIP. It is also interesting that ε decreases with increasing O_2/C_2H_2 ratio, suggesting that ANI-nr 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 set, although typically included for 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,

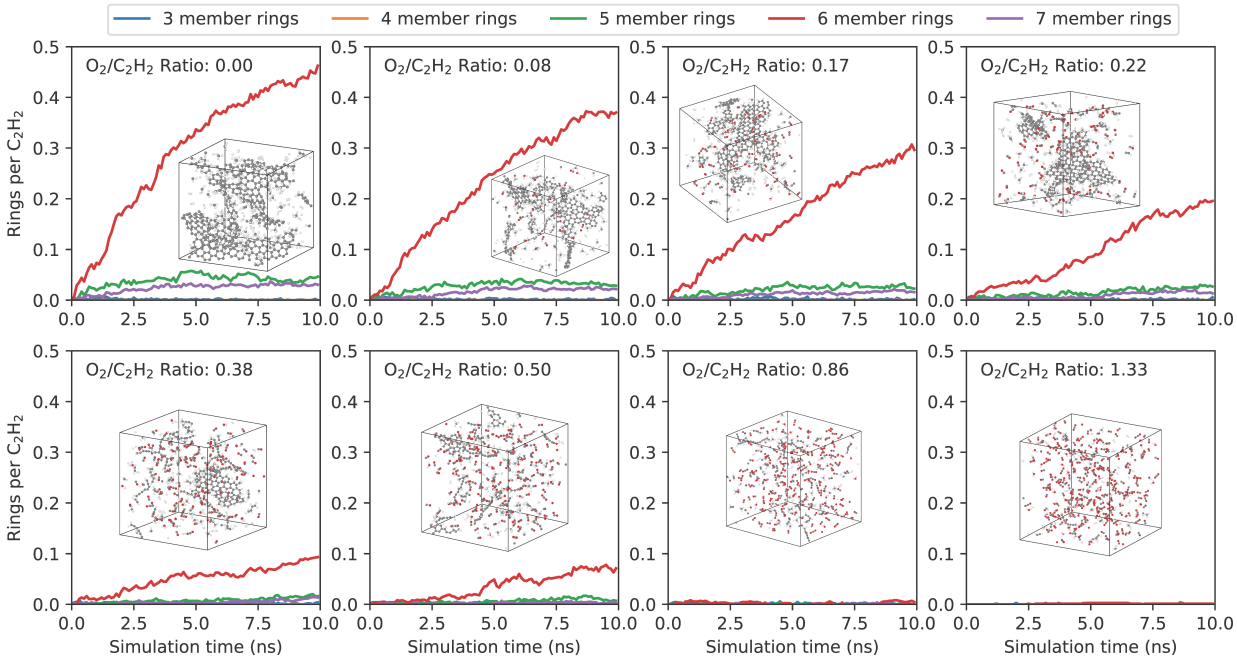

Figure 4. Comparison of 3-, 4-, 5-, 6-, and 7-member ring formation for different ratios of C_2H_2 and O_2.

which are less common in the training set.

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.[58] 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, MTBE demonstrated similar ignition enhancement to 2-butanol, but 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-nr for simulating biodiesel and to investigate the reported ignition enhancement of fuel additives, we reproduced four systems simulated by Chen et al.,[58] 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-nr compared with ReaxFF. Specifically, for all four cases, nearly all of the O_2 was consumed in the first 0.3 ns with ANI-nr, while there was still 20%-50% unconsumed O_2 after 2 ns with ReaxFF (for tracking plots including the entire 2 ns simulation, see Figure S.2 in Supporting Information).

Table 2 quantifies the similarities and differences between ANI-nr and ReaxFF results. For example, despite the quantitative difference in ignition delay times, the additive effect on ignition delay for ANI-nr 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-nr compared to ReaxFF, IDTs are highly sensitive as to how the system is initialized and to how ignition is defined (see Figure S.3 in Supporting Information). Furthermore, ANI-nr 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-nr predicts that ethanol also enhances O_2 consumption. Specifically, after the first 0.07 ns of ANI-nr 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 Figure S.4 in Supporting Information). 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-nr 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.[59] 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

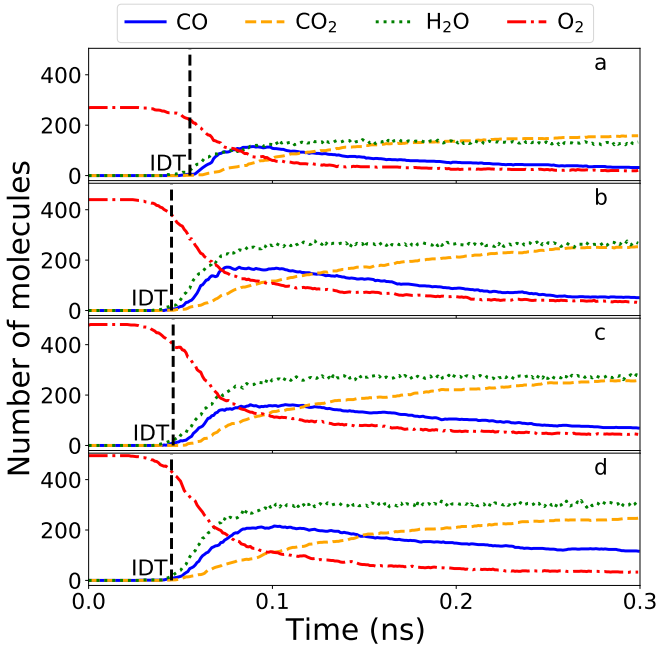

Figure 5. Tracking plot of major products (CO, CO_2, and H_2O) and O_2 for the biofuel simulations. (a) biofuel+O_2 (b) biofuel+O_2 with ethanol additive (c) biofuel+O_2 with 2-butanol additive (d) biofuel+O_2 with MTBE additive. Ignition delay time (IDT) is defined as the average time that at least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

System	Ignition delay time (ps)		O_2 consumption (%)	
	ANI-nr	ReaxFF	ANI-nr ($t = 0.07$ 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%

Table 2. Comparison between ANI-nr 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 when main products (CO, CO_2, and H_2O) are first observed. O_2 consumption is compared at 0.07 ns for ANI-nr 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.

by the initial amount of O_2 (see Figure S.5 in Supporting Information). The enhancement in OH production for ethanol is intuitive since ethanol contains a hydroxyl group with a similar bond dissociation energy (BDE) 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 ignition delay time for all three additive systems.

The discrepancy in overall reaction rates between ANI-nr and ReaxFF may be due to a difference in the underlying QM approach used to build each model. ReaxFF was primarily developed based on B3LYP calculations (supplemented with high-accuracy BDE data), while ANI-nr was trained to BLYP reference calculations. 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

Another advantage of MLIPs is that one can easily improve the performance of the potential on a given type of reaction by training to an application-specific training set. Emerging research has shown the success of application-specific MLIPs on systems like radical reactions in hydrocarbon combustions and gas-phase S_N2, etc.[60,61]

Though our ANI-nr potential was trained for a more general purpose, we compare the performance of our MLIP to

an application-specific MLIP for methane combustion under high temperatures and pressures. Simulating methane in the presence of O_2 will help determine if the ANI-nr dataset adequately samples methane combustion reactions, or if direct application-specific sampling is necessary.

Zeng et al.[35] trained an NN potential to a data set of QM recalculated fragment clusters sampled from a ReaxFF simulation of the combustion process of a mixture of CH_4 and O_2. They showed that the NN potential could then simulate the combustion process of pure methane with a reasonable mechanism. Here, we reproduced their MD simulation of methane combustion under the same conditions with our ANI-nr potential. Figure 6a shows that the ANI-nr potential produces very similar major products and species profiles to those of Zeng et al. However, by comparison with their CH_4 and O_2 consumption rates, ANI-nr predicts an overall reaction rate that is approximately a factor of four faster. Specifically, while their system required 0.5 ns of simulation time to consume half of the initial CH_4, our system required only 0.12 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.

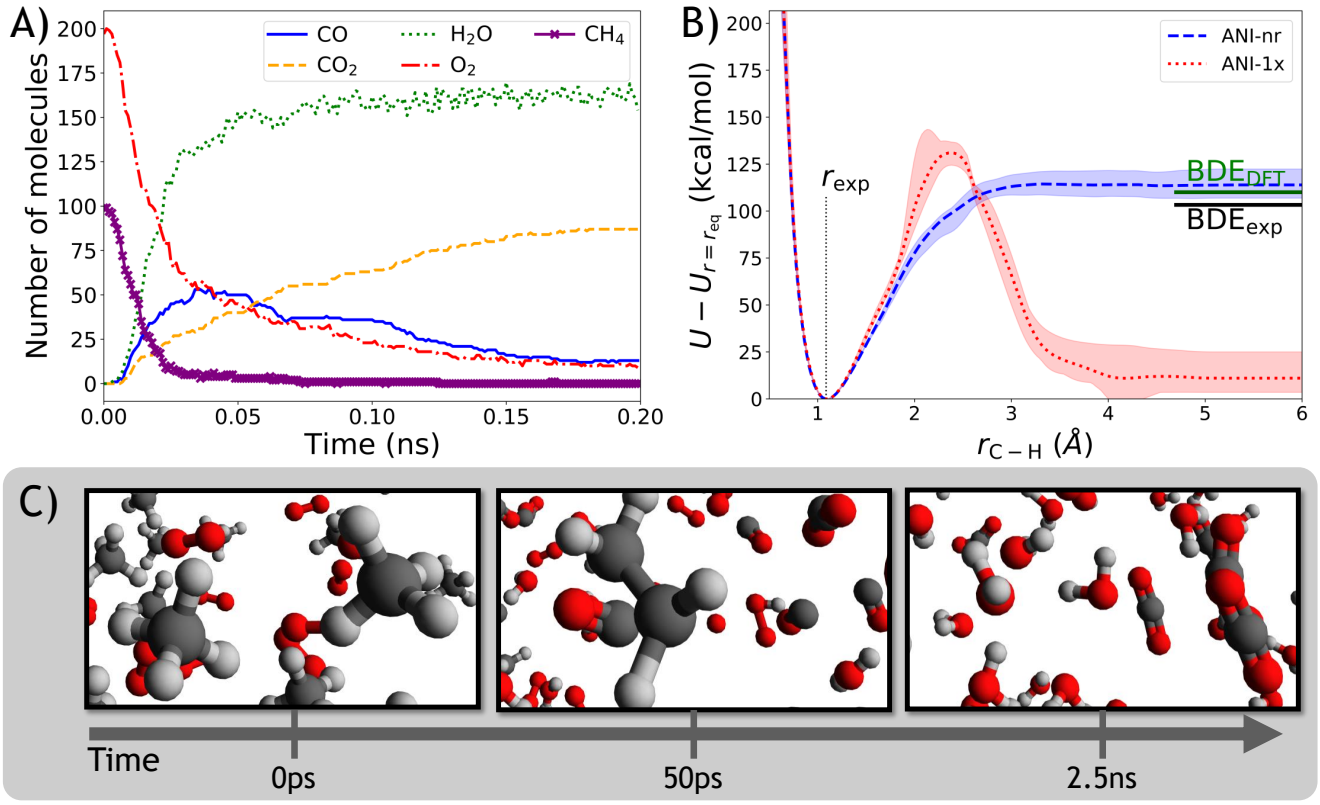

Figure 6. (a) Product molecule tracking plot of methane combustion simulation with ANI-nr. The tracking plot for the full simulation can be found in Figure S.6 of Supporting Information. (b) Bond dissociation diagram for C-H bond in methane. Comparison between ANI-nr and ANI-1x with the DFT and experimental bond dissociation energies (BDE). (c) Snapshots of initial reactants, intermediate species, and final products.

To investigate the cause for the increased rate of CH_4 consumption, Figure 6b compares the bond dissociation curve for ANI-nr with ANI-1x and the BDE values from DFT and experiment. ANI-nr is clearly a significant improvement from ANI-1x and agrees quite closely with the DFT BDE. As the ANI-nr and DFT BDEs are slightly larger than experiment, there is no evidence that a poor DFT BDE is the cause for the accelerated reactivity of ANI-nr.

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

To further demonstrate the power of our general purpose AL procedure to generate a general-use reactive dataset, we analyzed the application-specific dataset of Zeng et al. with our general MLIP, ANI-nr. Specifically, we computed the normalized ensemble standard deviation (ε) of ANI-nr for every structure in the Zeng et al. training dataset (see Figure S.7 in

Supporting Information). The ε value is smaller than the AL energy cutoff criterion of 0.03 kcal/mol/N$^{-\frac{1}{2}}$ for 99.9% of the structures in their application-specific dataset. This means that our AL approach would only consider ≈0.1% (or 72) of their 567,312 structures to be of high enough uncertainty to merit inclusion in our ANI-nr training data set. Based on these results, we conclude that the application-specific dataset does not contain a significant amount of unique information for the combustion of methane that is not already contained in our ANI-nr training dataset. Thus, poor sampling of methane combustion is not the cause for the discrepancy in overall reaction rate.

Miller experiment

In 1959, Stanley Miller designed a famous experiment to elucidate the origins of life on earth.[62] 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[63] performed relatively short (≈40 ps) near-ambient-temperature (400 K) condensed-phase (≈1 g/cc) DFT-MD simulations wherein an electric field is applied directly to "spark" chemical reactions. 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 NR study of Wang et al.[64] However, due to the low computational cost of our MLIP, we are able to run our simulations considerably longer (≈4 ns) than the AIMD simulations of Wang et al. (≈1 ns) with the same system size of 228 atoms and 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 case for Wang et al.

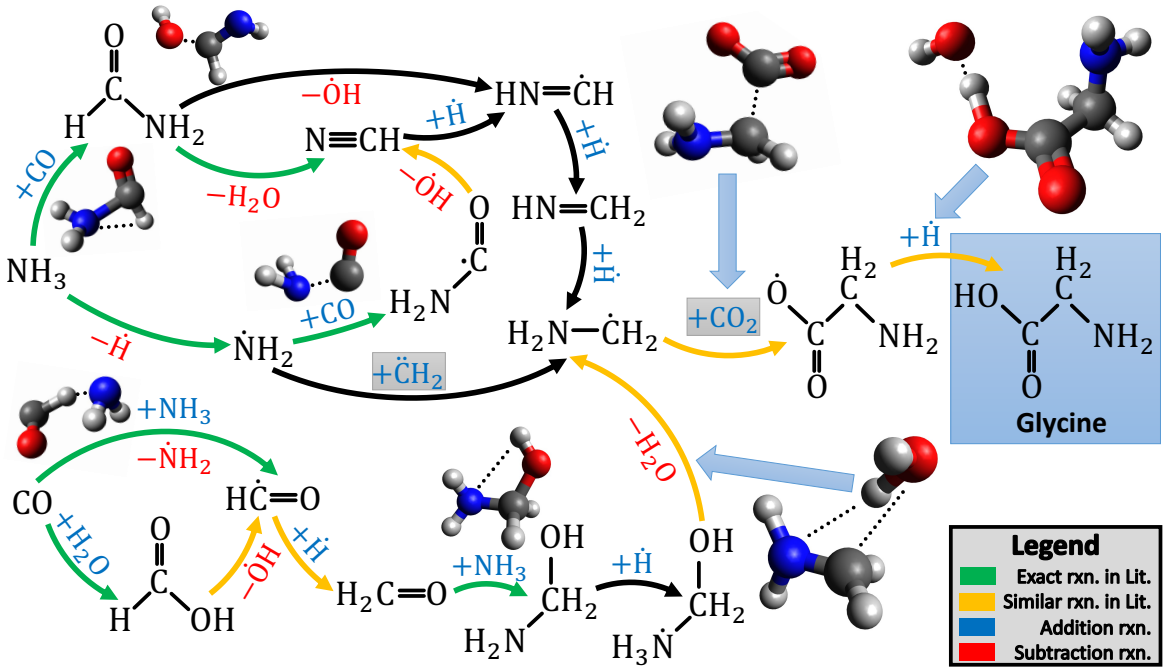

Figure 7. Formation of glycine. Note that +H does not necessarily signify a free hydrogen atom, +H is short-hand for a proton donor, e.g., NH_4, NH_3, CHO, CHNO, H_3O, H_2O. Likewise, -H does not necessarily signify dissociation of a hydrogen atom. -H is short-hand for a proton acceptor, e.g., NH_2, 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 Figure S.8 of Supporting Information. The depiction of bond orders and radical species is based simply on chemical intuition, since ANI-nr does not provide explicit orbital or electronic information (see Figure S.9 in Supporting Information for an alternative interpretation of this mechanism involving ionic species).

Figure 7 presents the ANI-nr 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. 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. and Saitta and Saija. The pathways to form the key intermediates carbon dioxide (CO_2) and methylene (CH_2) are reported in Figure S.8 in Supporting Information.

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., HF vs DFT vs MLIP), but also the difference in the simulation methodologies (i.e., our simulation did not utilize a "piston" nor induce an electric field). To verify the fidelity of ANI-nr for this system, we compared the DFT energies and forces with ANI-nr over the first 800 ps of the Miller experiment simulation (see Figure S.10 in Supporting Information).

Conclusions

In this article, we introduce a general machine learning interatomic potential (ANI-nr) for organic condensed-phase reactive molecular dynamics. ANI-nr is trained to a large data set obtained using an active learning (AL) workflow employing a new MD-based sampling algorithm to discover diverse and relevant condensed phase reactive atomistic configurations. Our novel sampler, inspired by the *ab initio* nanoreactor work of Wang et al., is the main innovation that drives these successes by building a reactive data set 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 nanoreactor sampler was the result of one or more reaction pathways which did not need to be known, specified, or analyzed ahead of time. Our new active learning approach represents a break through in the automated development of next generation reactive potentials for molecular dynamics simulation.

We validate the accuracy and applicability of the ANI-nr 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-nr reproduces experiment well. In other cases, where experiment is not available for comparison, ANI-nr produces results that are by and large consistent with DFT, DFTB, ReaxFF, and an application-specific MLIP, all without the need to retrain.

We are providing the resulting ANI-nr potential and data set to the community for further application and analysis. The DFT method used to calculate energies and forces for the data set was selected as an accurate approach that has been successfully employed to study a diverse range of organic reactions[65,66] while remaining affordable enough for high-throughput computations of large condensed-phase systems. For reactive chemistry, double hybrid DFT provides better accuracy, but at at 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.[23] In machine learning for language modeling, the concept of foundational models–large, general models that can be specialized to specific tasks quickly with very small amounts of data–has recently gained traction.[67] Because ANI-nr is trained to a large, general data set, it would also be interesting to consider whether it can act as a foundational model for more system-specific MLIPs when greater accuracy is required, for example, when predicting reaction rates.

Methods

ANI-nr model descriptions, training details

The ANI-nr model was trained similarly to ANI models within other contexts,[25] including materials science[40] and chemistry.[24] We use the ANI descriptors,[10] which is a modified form of the Behler and Parinello neural network descriptors.[6] ANI-nr 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 potential. The neural network architecture for each element and symmetry functions are reported in Tables S.II and S.III of Supporting Information, respectively. The model was trained using both energy and force terms in the loss function as described in previous work.[68] 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 1.0E-5.

Computational nanoreactor active learning for training set generation

The ANI-nr training data set 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 active learning process, periodic cells containing randomly placed and oriented small molecules with less than three non-H atoms and with a randomly selected composition of H, C, N, and O are generated. All training energies and forces are computed with the open-source CP2K software[69] using Kohn–Sham DFT,[70] BLYP functional,[71,72] TZV2P basis set,[73] GTH pseudopotentials,[74] D3 dispersion correction,[75] and energy cutoffs of 600 and 60 Ry, respectively, for the plain-wave and Gaussian contributions to the basis set, as recommended in previous work.[65] The overall spin multiplicity is constrained to a singlet state, consistent with previous studies that perform CP2K simulations of bulk systems containing radical species.[66] 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 Atomic Simulation Environment (ASE)[76] and the NeuroChem package. We use an uncertainty quantification (UQ) metric, i.e., the normalized ensemble standard deviation in energy and the forces,[23,40] 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 UQ metric, are included in the data set with their corresponding QM energy and forces. Below is a detailed step-by-step description of the active learning workflow:

1. Generate a bootstrap data set (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 data set 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 Table S.IV in Supporting Information):

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

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

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

5. Monitor energy and force UQ metrics every 5-50 MD steps, if the UQ values go over a pre-selected value end the simulation and add frame to a batch of new structures.

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

7. Add new data to the training data set.

8. Go back to step 2 and repeat until the potential 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 data set

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 Figures S.11, S.12, and S.13 of Supporting Information, respectively. 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.

Figure S.14 in Supporting Information shows a histogram of the sizes of all molecules that are found in the ANI-nr dataset, which includes one system up to 145 atoms. The majority are small molecules, of similar size, or slightly large, to 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 UQ estimate. Therefore, 1212 should be considered a lower-bound of molecules discovered during active learning. There are likely many more molecules formed over all AL MD sampling, which is estimated to be 100s of nanoseconds of simulation time in aggregate.

Carbon solid-phase nucleation

To investigate the formation process of diamond and graphene, we performed molecular dynamics simulations of amorphous carbon under different densities. The initial structure was built with in-house code. We fixed the total number of carbon atoms to 5000, and made three initial structures with three different densities (0.5 g/cc, 2.25 g/cc and 3.52 g/cc), by varying the simulation box length. First, we randomly select an initial position in the simulation box as the coordinate of the first carbon atom. Then, for each additional carbon atom, we generate random positions and keep the position if the distance to all previous positions is larger than twice the van der waals radius for carbon atoma (1.7 Å). We iterate this process until all 5000 carbon atoms are inserted. We run Langevin dynamics at a temperature of 2500 K for 5 ns with step length of 0.5 fs. Coordinates and properties are recorded every 50 fs. We run 8 independent trajectories for each density to verify that the correct phase is identified from different starting structures. We use the Open Visualization Tool (OVITO)[77] to distinguish phases (diamond cubic, hexagonal diamond, or graphene) in the snapshots.

Effect of oxygen on graphene ring formation

To investigate ring formation from acetylene, we performed MD simulations of 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). The initial structures were generated with PackMol.[78] Next, we use the LBFGS optimizer to obtain the energy minima structure. Then, we run Langevin dynamics simulation at 2000 K for 10 ns with a 0.5 fs time-step, and friction constant of 0.01. Snapshots and properties are recorded every 0.5 ps. We use our in-house code MolFind to identify and count ring structures of varying sizes. Considering that the distance between bonded atoms can fluctuate, we define a 0.02 Å buffer 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. The buffers are also added when analyzing other simulations.

Comparison of biofuel additives

To investigate the effect of different fuel additives on ignition performance, we performed simulations for biofuel and 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 are the same as shown in Table 2 of the ReaxFF reference paper.[58] We use Packmol to generate the initial structure such that the initial separation of all molecules is at least 2 Å. The initial density is 0.25 g/cc in all four cases. We run Langevin dynamics at a temperature of 100 K for 1 ps with steplength 0.1 fs for relaxation, then, we gradually heat up the system with a 50 K/ps rate to 3000 K. After reaching the desired temperature of 3000 K, the simulation is ran for an additional 10 ns. During the whole process (including relaxation and temperature ramping) snapshots and properties are recorded every 1 ps. We run 5 independent trajectories for each system to reduce uncertainty in species profiles.

Methane combustion

The methane combustion system is initialized with 100 methane molecules and 200 O_2 molecules. All molecules are inserted using Packmol and ensuring that all molecules are separated by at least 2.0 Å. The cubic simulation box length is 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 1ns with a time-step of 0.1 fs and with a friction constant of 0.01. Snapshots and properties are recorded every 0.1 ps.

Miller experiment

To investigate the ability to simulate complex organic system that involve biologically relevant molecules, we performed a simulation under a similar settings of the Miller experiment. We used Packmol 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 length 12.1 Å. The density is then 1.067 g/cc. The simulation was performed by Langevin dynamics with steplength 0.25 fs. We gradually increased the temperature from 0 K to 300 K in the first 100 ps, then gradually increased the temperature from 300 K to 2500 K in the next 100 ps. We held the temperature at 2500 K for 4000 ps and then gradually cooled the system from 2500 K to 300 K over the final 200 ps. Snapshots and properties are recorded every 12.5 fs (50 time steps).

Acknowledgements

The work at Los Alamos National Laboratory (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. 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 LANL Applied Machine Learning summer student 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., Shenderova, O. A., Harrison, J. A., Stuart, S. J., Ni, B. & Sinnott, S. B. 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., Hong, S., Islam, M. M., Kylasa, S. B., Zheng, Y., Shin, Y. K., Junkermeier, C., Engel-Herbert, R., Janik, M. J., Aktulga, H. M., Verstraelen, T., Grama, A. & van Duin, A. C. T. The ReaxFF reactive force-field: development, applications and future directions. *npj Comput. Mater.* **2**, 15011, DOI: 10.1038/npjcompumats.2015.11 (2016).
6. 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).
7. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330, DOI: 10.1016/j.jcp.2014.12.018 (2015).
8. 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).
9. Shapeev, A. V. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. & Simul.* **14**, 1153–1173, DOI: 10.1137/15M1054183 (2016).
10. 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).

11. Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A. & Müller, K.-R. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **30** (2017).

12. Schütt, K. T., Arbabzadah, F., Chmiela, S., Müller, K. R. & Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890, DOI: 10.1038/ncomms13890 (2017).

13. 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).

14. Unke, O. T. & Meuwly, M. PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693, DOI: 10.1021/acs.jctc.9b00181 (2019).

15. Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E. & Kozinsky, B. 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).

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

17. Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M. & Kozinsky, B. Learning local equivariant representations for large-scale atomistic dynamics. *arXiv* DOI: 10.48550/arXiv.2204.05249 (2022).

18. 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).

19. 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).

20. 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).

21. 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).

22. 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).

23. Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O. & Roitberg, A. E. 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).

24. Smith, J. S., Zubatyuk, R., Nebgen, B., Lubbers, N., Barros, K., Roitberg, A. E., Isayev, O. & Tretiak, S. 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).

25. Devereux, C., Smith, J. S., Huddleston, K. K., Barros, K., Zubatyuk, R., Isayev, O. & Roitberg, A. E. 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).

26. 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).

27. 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).

28. 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).

29. 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).

30. Lu, X., Shao, K., Fu, B., Wang, X. & Zhang, D. H. An accurate full-dimensional potential energy surface and quasiclassical trajectory dynamics of the H + H_2O_2 two-channel reaction. *Phys. Chem. Chem. Phys.* **20**, 23095–23105, DOI: 10.1039/C8CP04045A (2018).

31. Zhu, Y., Tian, L., Song, H. & Yang, M. Kinetic and dynamic studies of the $H_3^+ + CO \rightarrow H_2 + HCO^+/HOC^+$ reaction on a high-level ab initio potential energy surface. *J. Chem. Phys.* **151**, 054311, DOI: 10.1063/1.5110934 (2019).

32. 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).

33. Lu, D., Behler, J. & Li, J. Accurate global potential energy surfaces for the H + CH_3OH reaction by neural network fitting with permutation invariance. *J. Phys. Chem. A* **124**, 5737–5745, DOI: 10.1021/acs.jpca.0c04182 (2020).

34. Zuo, J., Chen, Q., Hu, X., Guo, H. & Xie, D. Theoretical investigations of rate coefficients for H + O_3 and HO_2 + O reactions on a full-dimensional potential energy surface. *J. Phys. Chem. A* **124**, 6427–6437, DOI: 10.1021/acs.jpca.0c04321 (2020).

35. 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).

36. 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).

37. Takamoto, S., Shinagawa, C., Motoki, D., Nakago, K., Li, W., Kurata, I., Watanabe, T., Yayama, Y., Iriguchi, H., Asano, Y., Onodera, T., Ishii, T., Kudo, T., Ono, H., Sawada, R., Ishitani, R., Ong, M., Yamaguchi, T., Kataoka, T., Hayashi, A., Charoenphakdee, N. & Ibuka, T. 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).

38. Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X. & Wang, X. A survey of deep active learning. *ACM Comput. Surv.* **54**, 1–40, DOI: 10.1145/3472291 (2021).

39. Sivaraman, G., Krishnamoorthy, A. N., Baur, M., Holm, C., Stan, M., Csányi, G., Benmore, C. & Vázquez-Mayagoitia, Á. 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).

40. Smith, J. S., Nebgen, B., Mathew, N., Chen, J., Lubbers, N., Burakovsky, L., Tretiak, S., Nam, H. A., Germann, T., Fensin, S. *et al.* Automated discovery of a robust interatomic potential for aluminum. *Nat. Commun.* **12**, 1257, DOI: 10.1038/s41467-021-21376-0 (2021).

41. Yoo, P., Sakano, M., Desai, S., Islam, M. M., Liao, P. & Strachan, A. 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).

42. 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).

43. 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).

44. 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).

45. 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).

46. Guan, X., Das, A., Stein, C. J., Heidar-Zadeh, F., Bertels, L., Liu, M., Haghighatlari, M., Li, J., Zhang, O., Hao, H., Leven, I., Head-Gordon, M. & Head-Gordon, T. A benchmark dataset for hydrogen combustion. *Sci.* **9**, 215, DOI: 10.1038/s41597-022-01330-5 (2022).

47. 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).

48. Wang, L.-P., Titov, A., McGibbon, R., Liu, F., Pande, V. S. & Martínez, T. J. Discovering chemistry with an ab initio nanoreactor. *Nat. Chem.* **6**, 1044–1048, DOI: 10.1038/nchem.2099 (2014).

49. 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).

50. 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).

51. 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).

52. Wang, J., Shen, H., Yang, R., Xie, K., Zhang, C., Chen, L., Ho, K.-M., Wang, C.-Z. & Wang, S. 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).

53. 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).

54. 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).

55. Devereux, C., Smith, J. S., Huddleston, K. K., Barros, K., Zubatyuk, R., Isayev, O. & Roitberg, A. E. 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.

56. Lei, T., Guo, W., Liu, Q., Jiao, H., Cao, D.-B., Teng, B., Li, Y.-W., Liu, X. & Wen, X.-D. 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).

57. 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.

58. 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).

59. 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).

60. 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).

61. 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).

62. 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).

63. 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).

64. Wang, L.-P., Titov, A., McGibbon, R., Liu, F., Pande, V. S. & Martínez, T. J. Discovering chemistry with an *ab initio* nanoreactor. *Nat. Chem.* **6**, 1044–1048, DOI: 10.1038/nchem.2099 (2014).

65. 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).

66. Fetisov, E. O., Kuo, I.-F. W., Knight, C., VandeVondele, J., Van Voorhis, T. & Siepmann, J. I. First-principles Monte Carlo simulations of reaction equilibria in compressed vapors. *ACS Cent. Sci.* **2**, 409–415, DOI: 10.1021/acscentsci.6b00095 (2016).

67. Yuan, L., Chen, D., Chen, Y.-L., Codella, N., Dai, X., Gao, J., Hu, H., Huang, X., Li, B., Li, C., Liu, C., Liu, M., Liu, Z., Lu, Y., Shi, Y., Wang, L., Wang, J., Xiao, B., Xiao, Z., Yang, J., Zeng, M., Zhou, L. & Zhang, P. Florence: A new foundation model for computer vision. *arXiv* DOI: 10.48550/arXiv.2111.11432 (2021).

68. 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).

69. Kühne, T. D., Iannuzzi, M., Del Ben, M., Rybkin, V. V., Seewald, P., Stein, F., Laino, T., Khaliullin, R. Z., Schütt, O., Schiffmann, F., Golze, D., Wilhelm, J., Chulkov, S., Bani-Hashemian, M. H., Weber, V., Borštnik, U., Taillefumier, M., Jakobovits, A. S., Lazzaro, A., Pabst, H., Müller, T., Schade, R., Guidon, M., Andermatt, S., Holmberg, N., Schenter, G. K., Hehn, A., Bussy, A., Belleflamme, F., Tabacchi, G., Glöß, A., Lass, M., Bethune, I., Mundy, C. J., Plessl, C., Watkins, M., VandeVondele, J., Krack, M. & Hutter, J. 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).

70. 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).

71. 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).

72. 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).

73. 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).

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

75. 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).

76. Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C. *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).

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

78. 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).

Supporting Information: Exploring the frontiers of 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-nr	90.0	90.0	90.0
	ANI-nr(lr)	90.0	90.0	90.0
	ANI-2x	90.0	90.0	90.0
	Exp.	90.0	90.0	90.0
Graphite	ANI-nr	90.0	90.0	120.
	ANI-nr(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-nr, ANI-nr(lr), ANI-2x and experiment (Exp.).

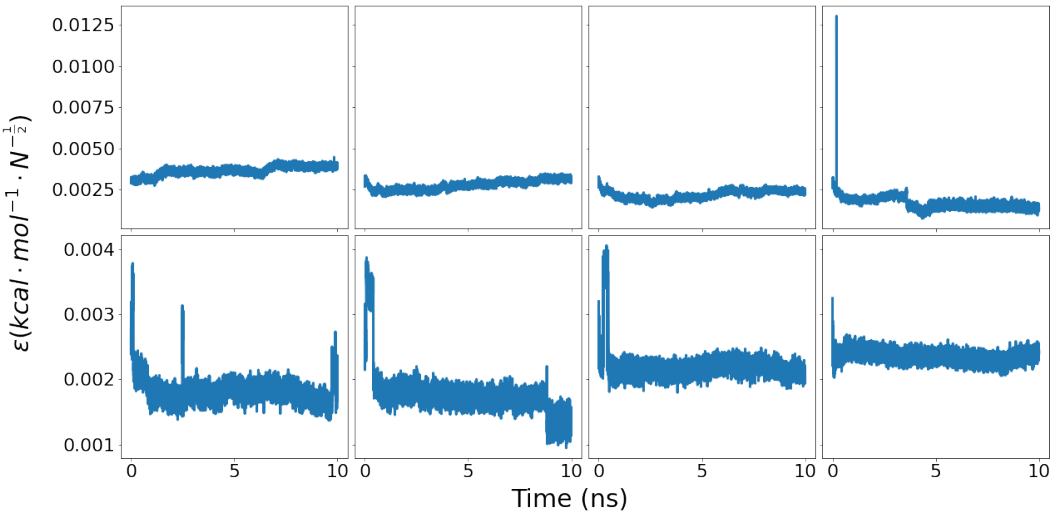

Figure S.1. Normalized ensemble standard deviation (ε) for all eight O_2/C_2H_2 ratios.

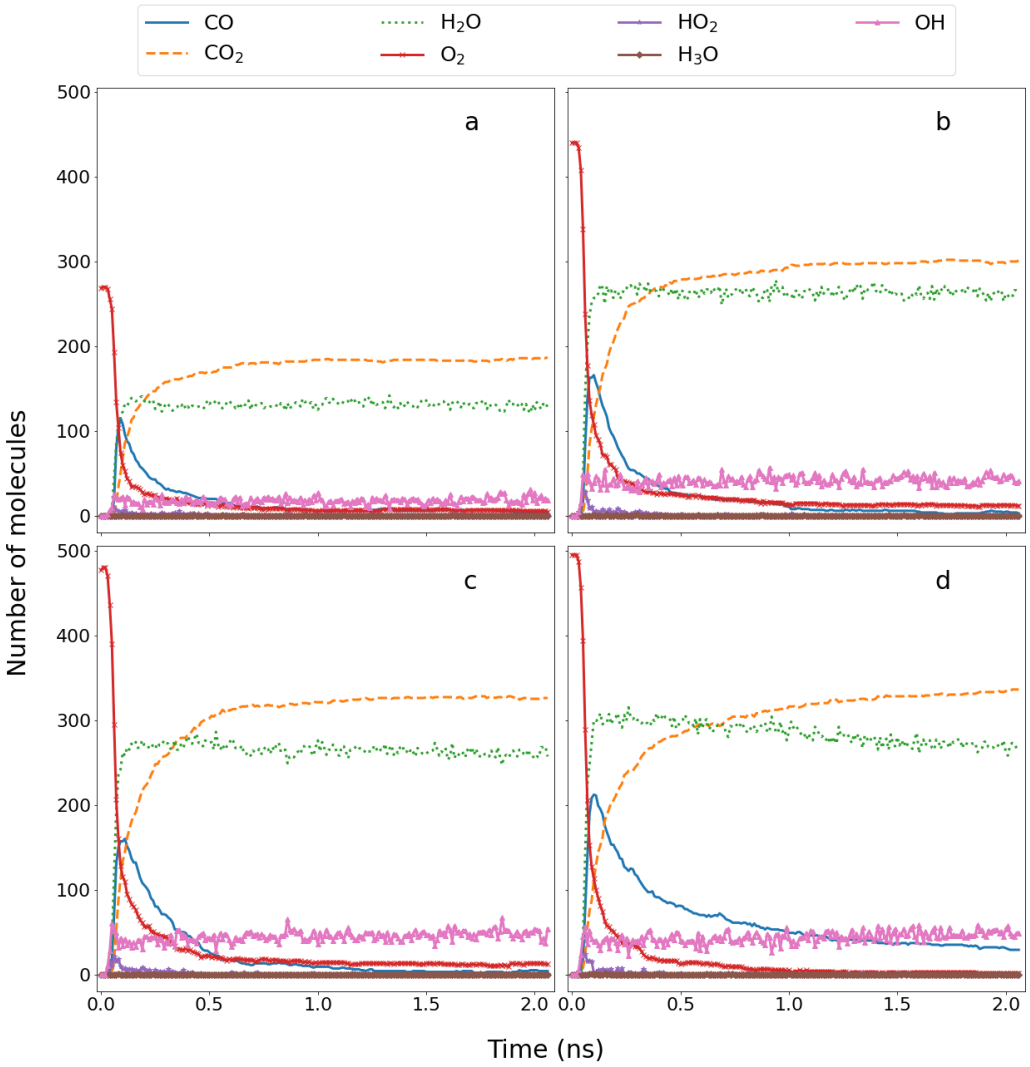

Figure S.2. Tracking plot of major products of the biofuel simulations. (a) biofuel+O_2 system (b) biofuel with ethanol+O_2 (c) biofuel with 2-butanol+O_2 (d) biofuel with MTBE+O_2.

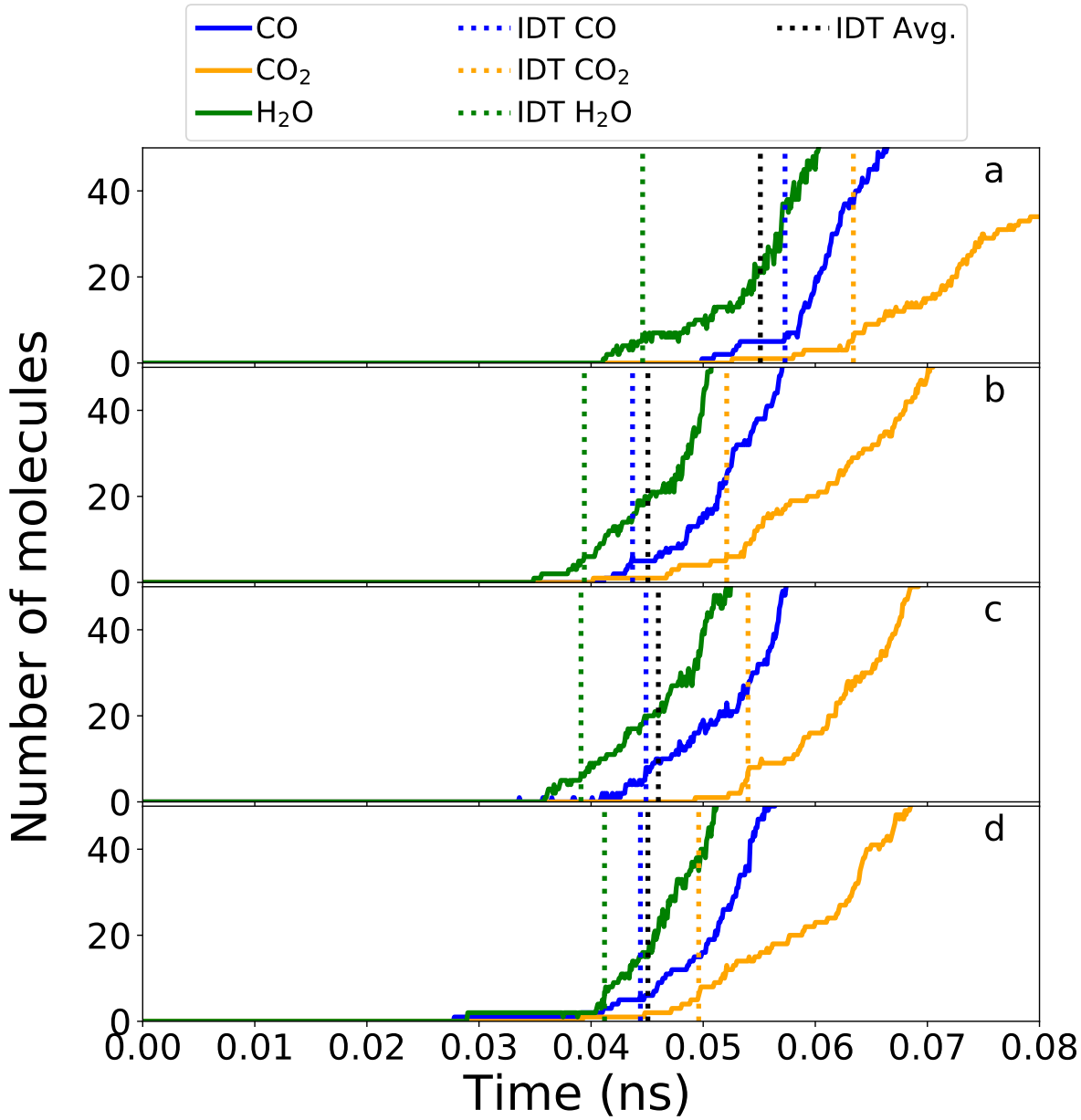

Figure S.3. Ignition delay time (IDT) for biofuel simulations based on each of the major products CO, CO_2, and H_2O. To remove anomalies when only a single molecule is produced significantly prior to "true" ignition, we define IDT as the earliest time that at least five molecules of a given product 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.

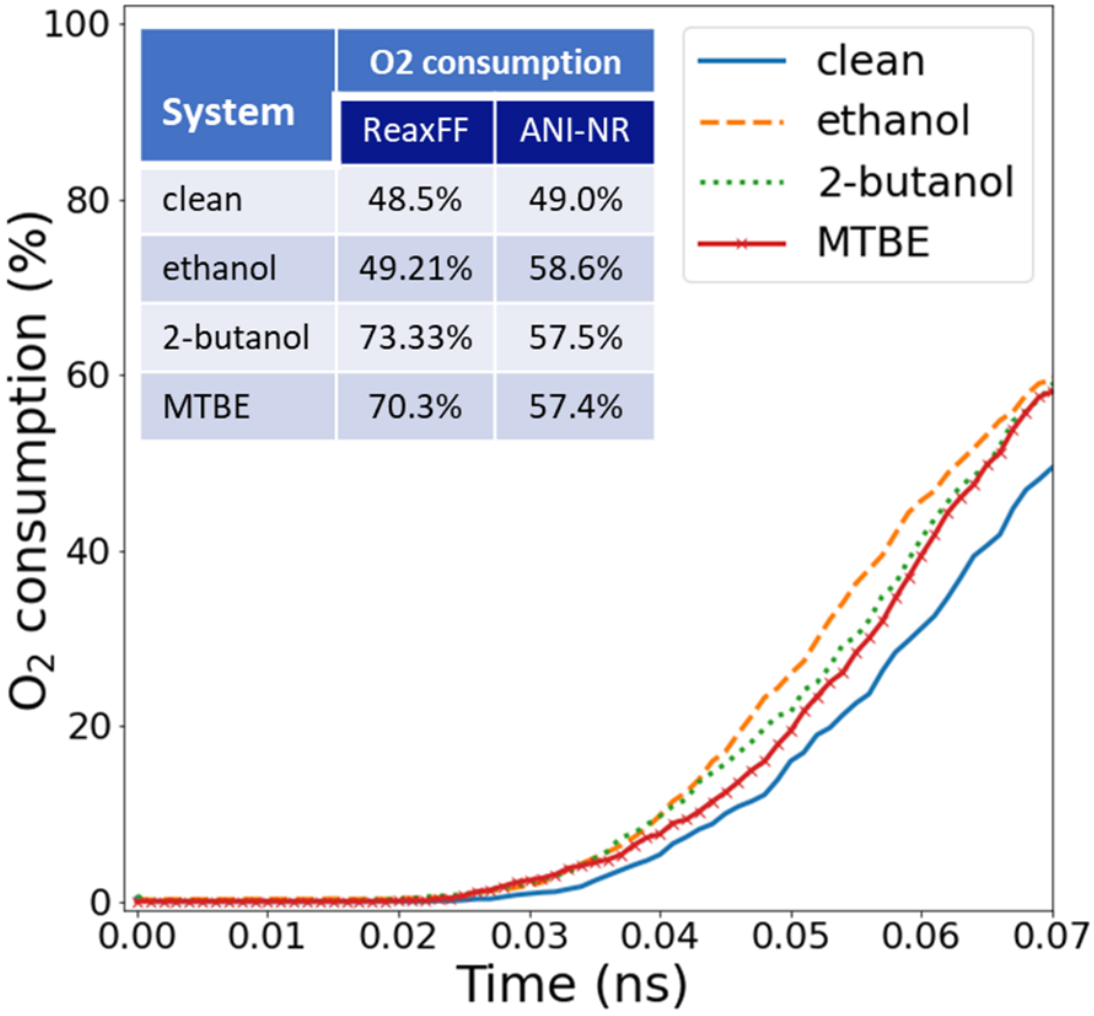

Figure S.4. O_2 consumption (%) during combustion for clean biofuel compared with three different fuel additives, namely, ethanol, 2-butanol, and MTBE. Insert compares O_2 consumption for ANI-nr (at 0.07 ns) with ReaxFF (at 2 ns). Curves are smoothed by averaging over 5 independent trajectories.

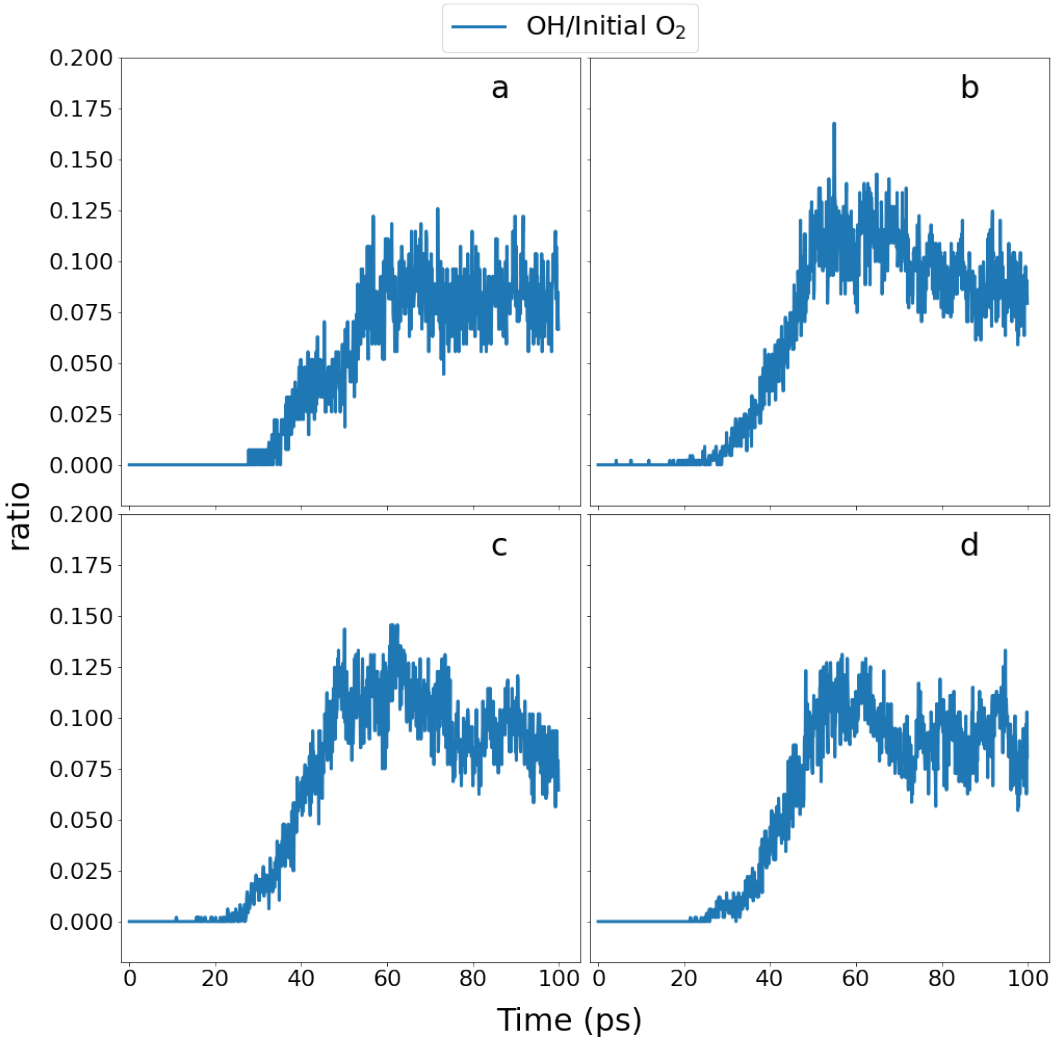

Figure S.5. Ratio of OH to initial O_2. (a) biofuel+O_2 system (b) biofuel with ethanol+O_2 (c) biofuel with 2-butanol+O_2 (d) biofuel with MTBE+O_2

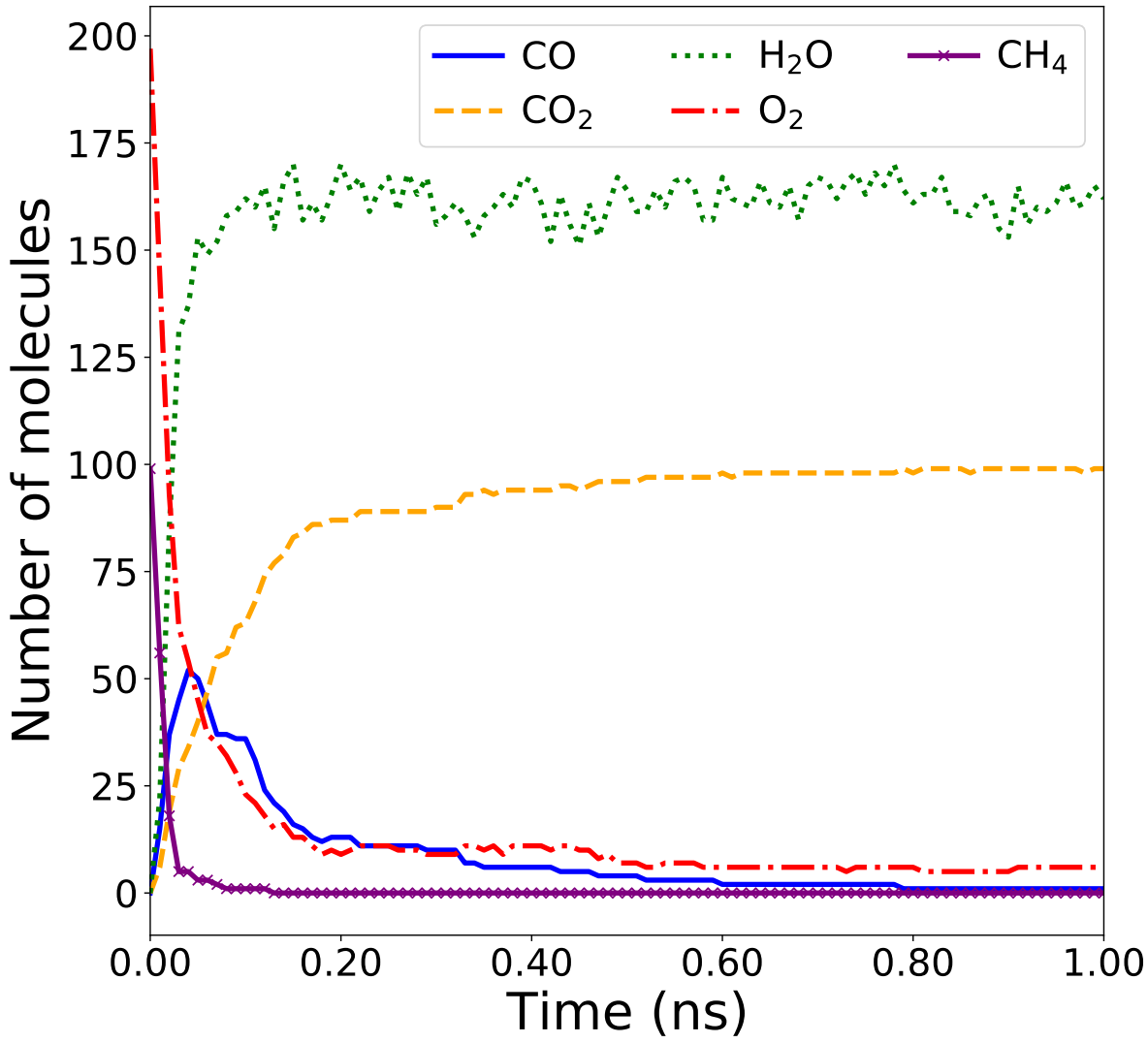

Figure S.6. Tracking plot of major products of the methane combustion simulations.

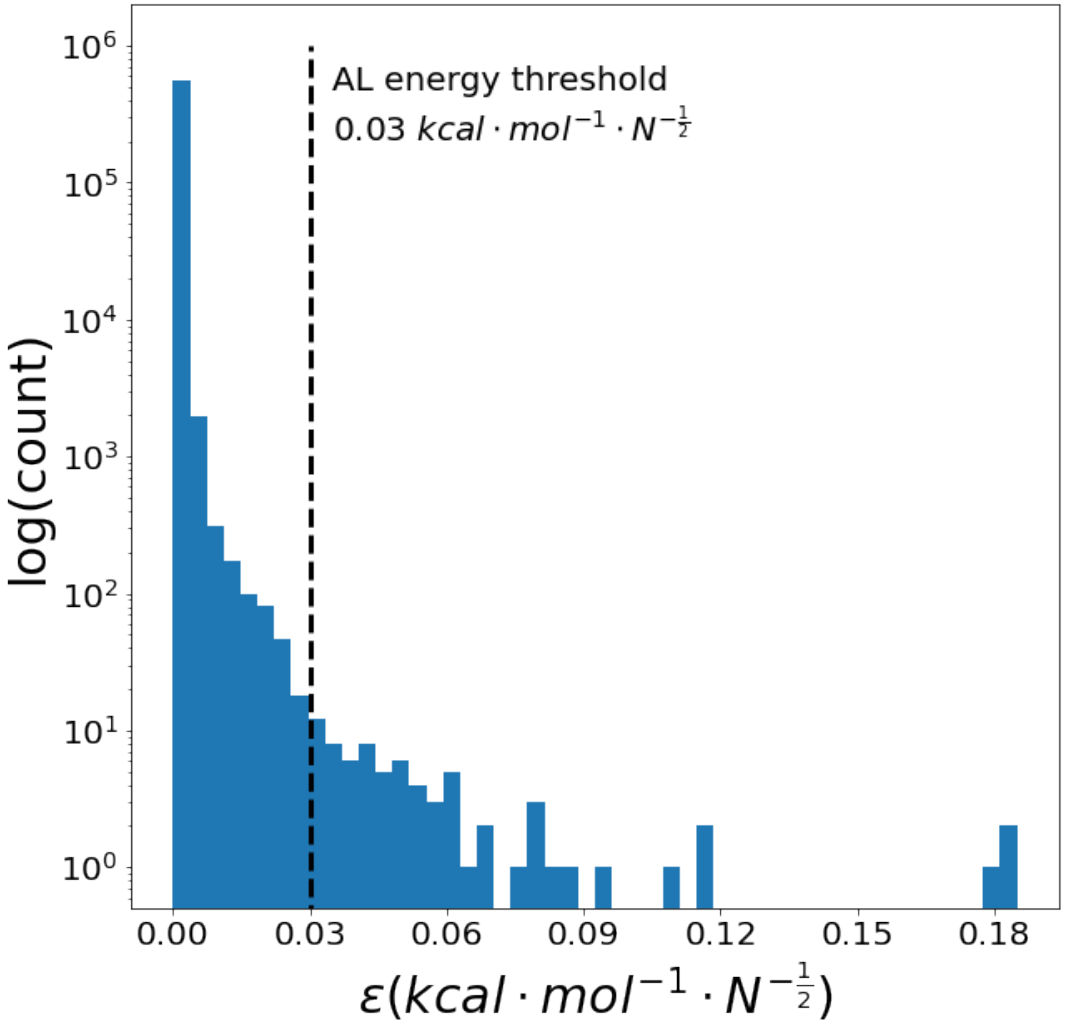

Figure S.7. Distribution of normalized ensemble standard deviation in energy for the ANI-nr model. Among 567312 available structures in their training set, only 72 structures have normalized energy uncertainty larger than 0.03 kcal $\cdot$ mol^{-1} $\cdot N^{-\frac{1}{2}}$

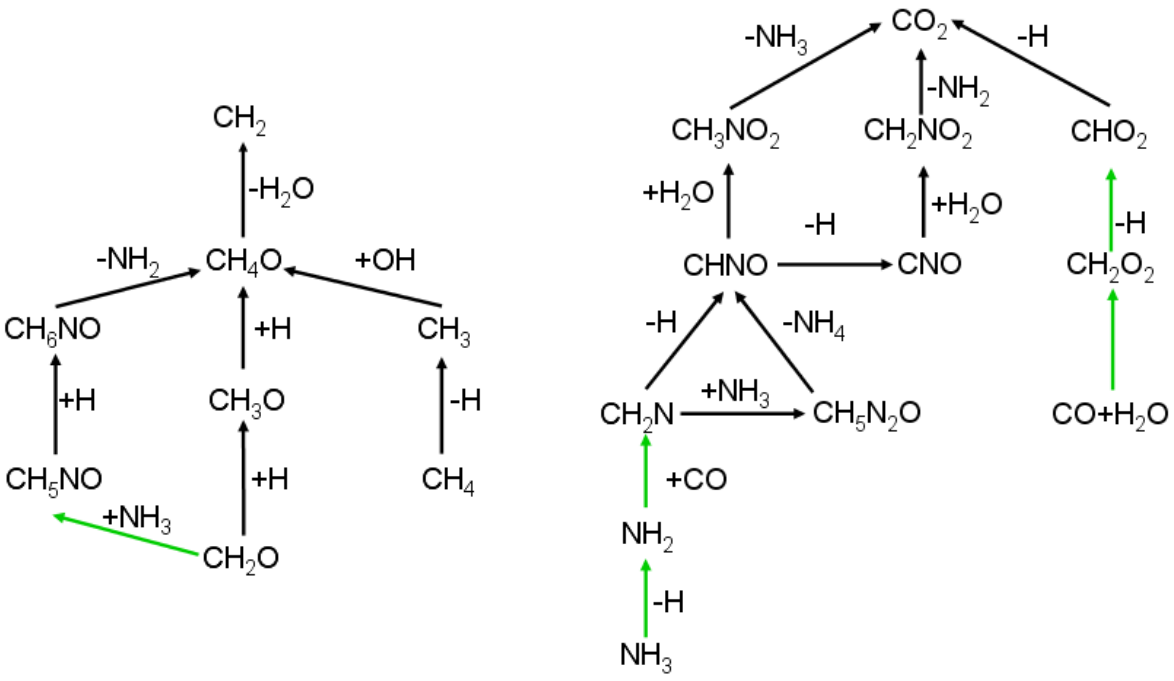

Figure S.8. Formation of intermediates CH_2 and CO_2 from initial reactants (NH_3, CO, CH_4, H_2O). Mechanism to form CH_2O is found in Figure 7 in the main text. Green arrows denote reactions previously identified by Wang et al. or Saitta and Saija. Orange arrows denote reactions that have closely-related reactions in Wang et al. or Saitta and Saija.

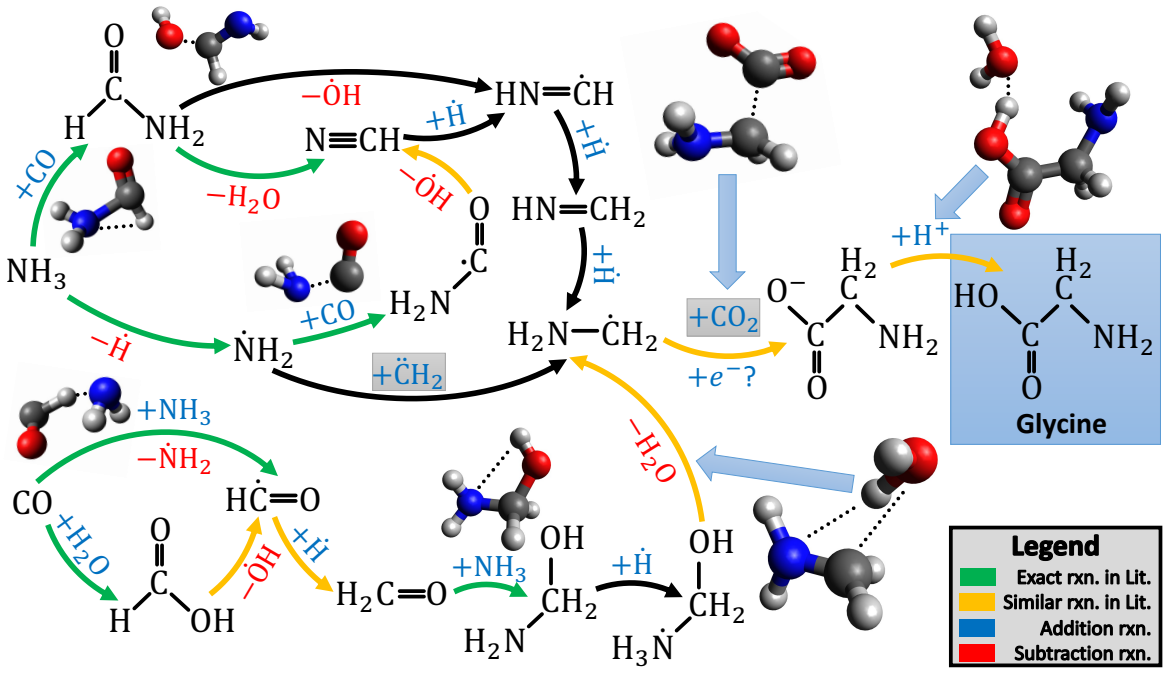

Figure S.9. An alternative mechanism for the formation of glycine in the ANI-nr Miller simulation. In this pathway, the final step to form glycine involves H-abstraction from H_3O, which is likely a cationic species (H_3O^+). The penultimate species ($C_2H_4NO_2^-$) formed prior to glycine, therefore, 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-nr does not provide explicit orbital or electronic information.

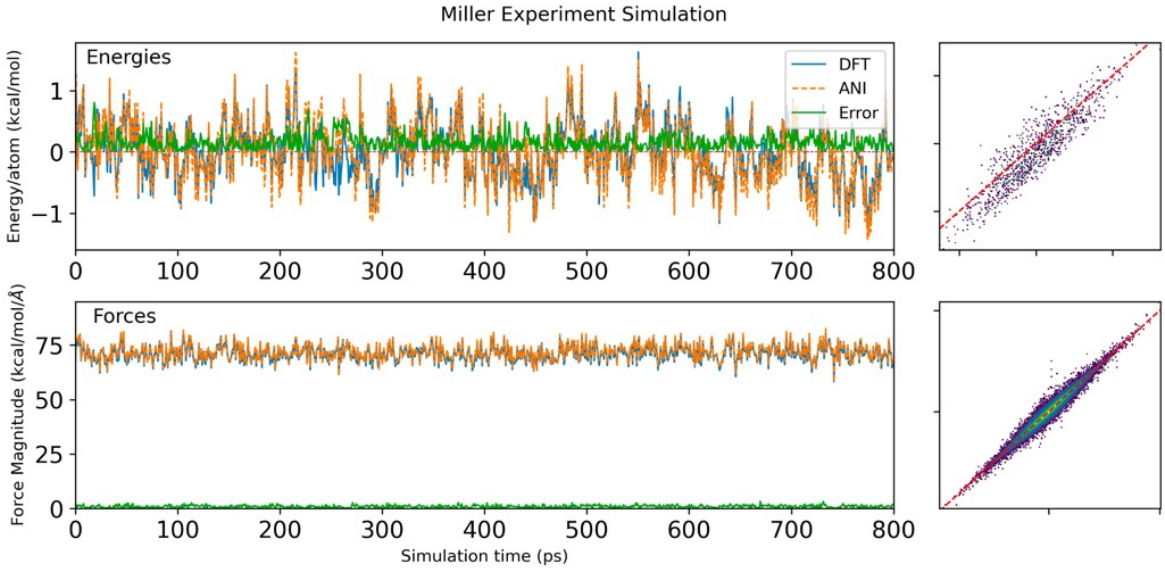

Figure S.10. Validation of Miller Experiment simulation. Comparison between DFT energies and forces with ANI-nr for the first 800 ps.

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-nr 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-nr 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-nr symmetery function parameters

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.IV. Parameters for nanoreactor oscillations in temperature and density.

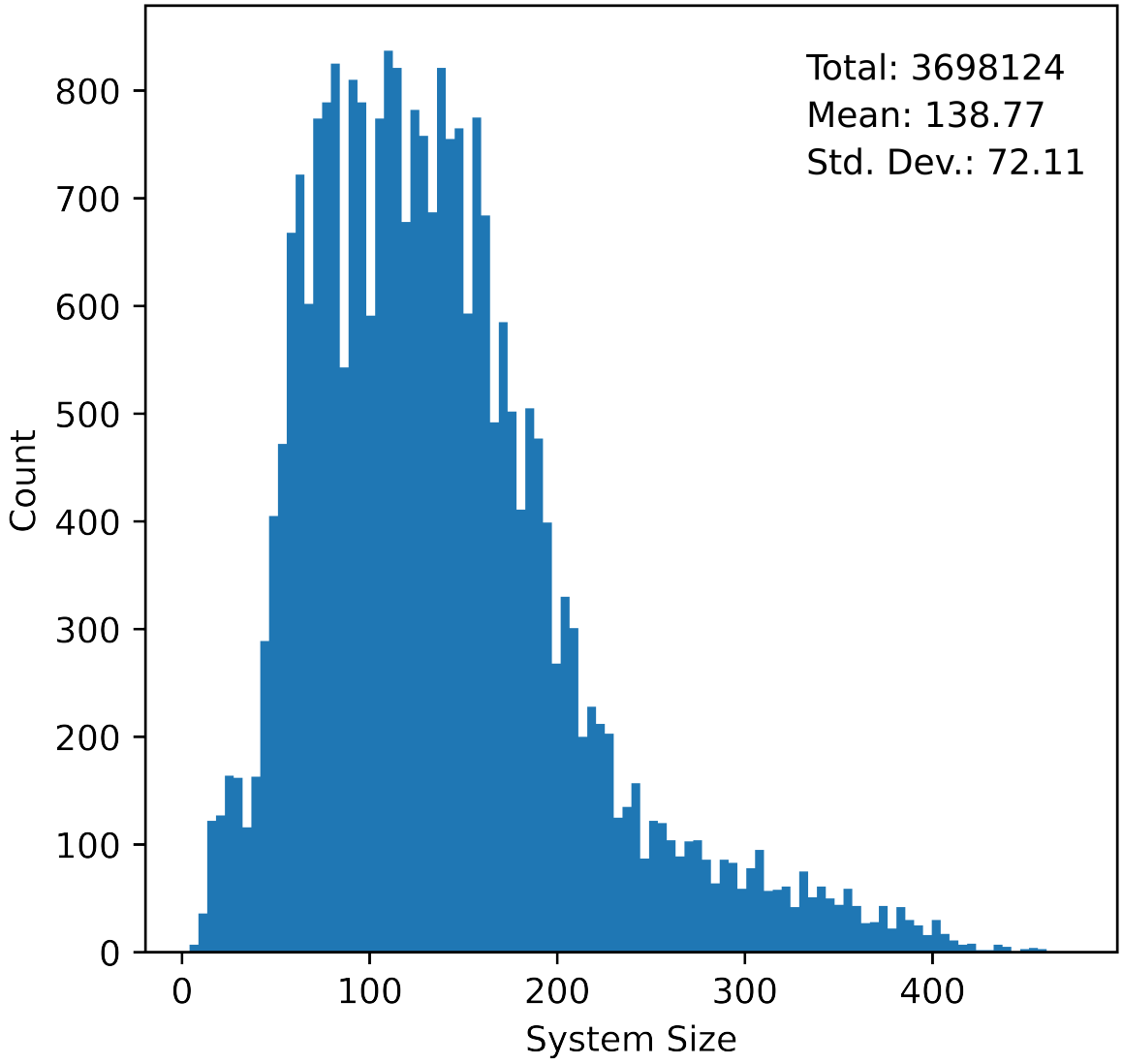

Figure S.11. Histogram of the system size (i.e., number of atoms) per system in the ANI-nr training data set.

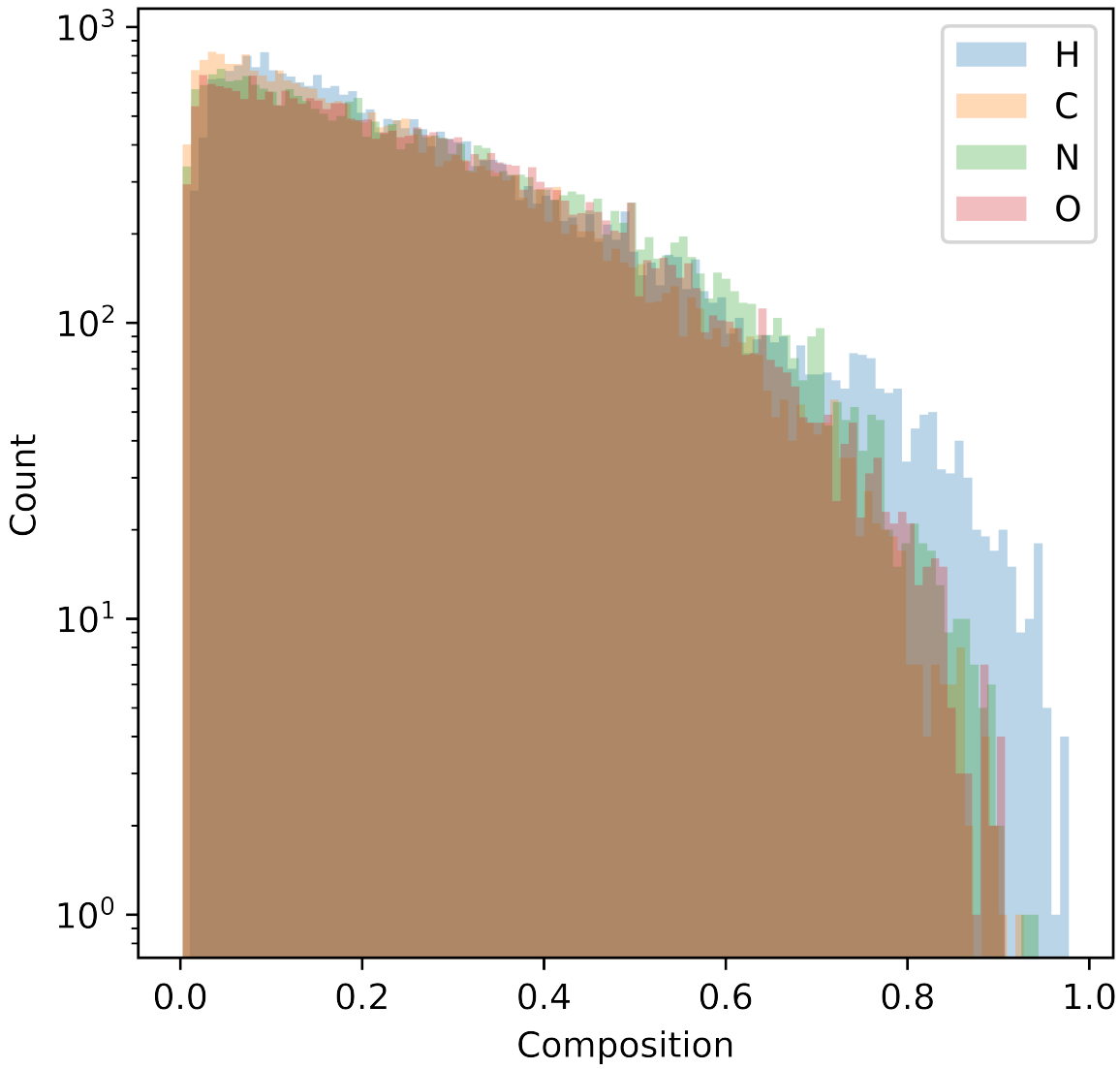

Figure S.12. Histogram of the system composition of all systems in the training data set, colored by element.

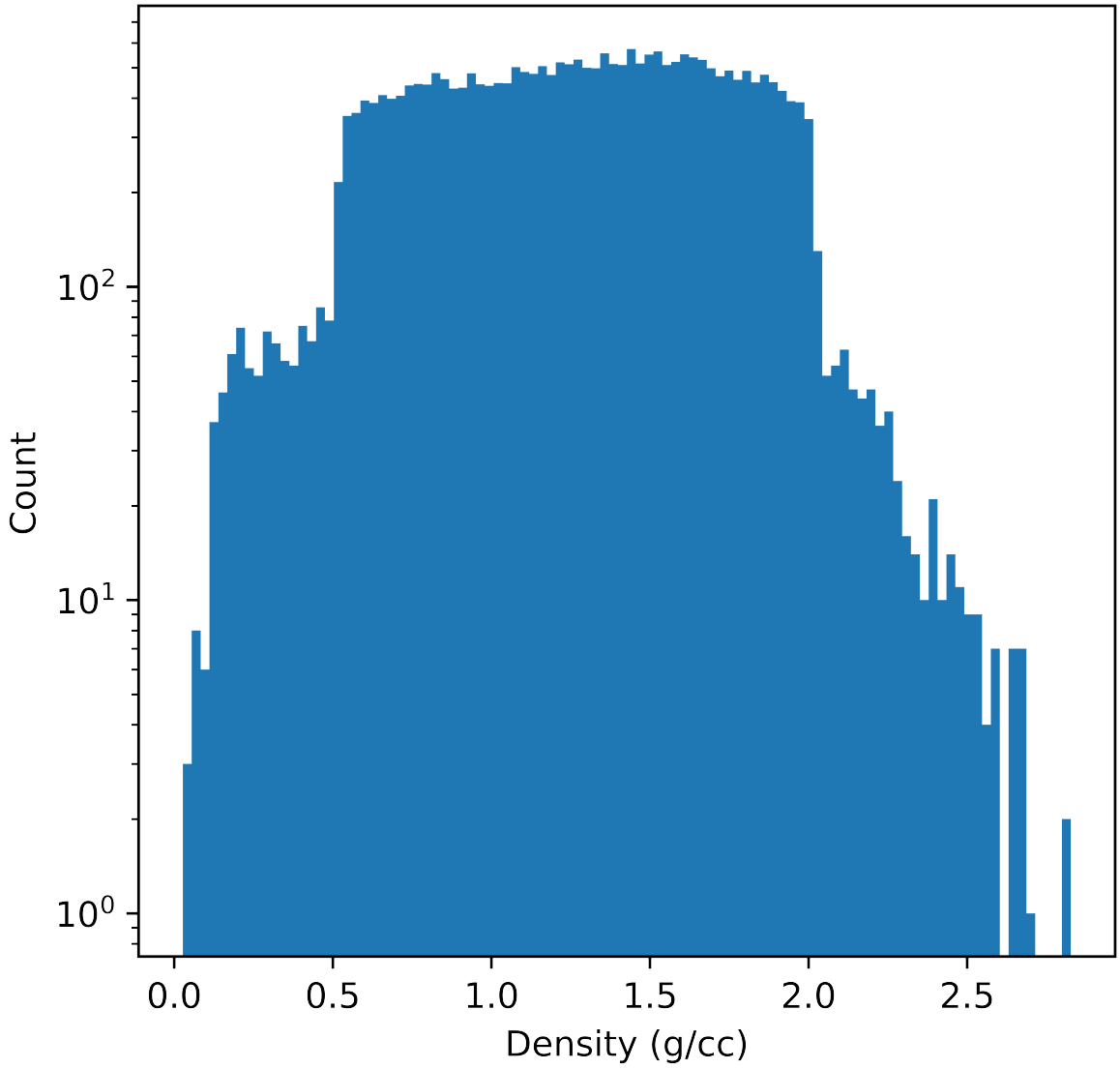

Figure S.13. Histogram of the mass density (g/cc) of all systems in the training data set.

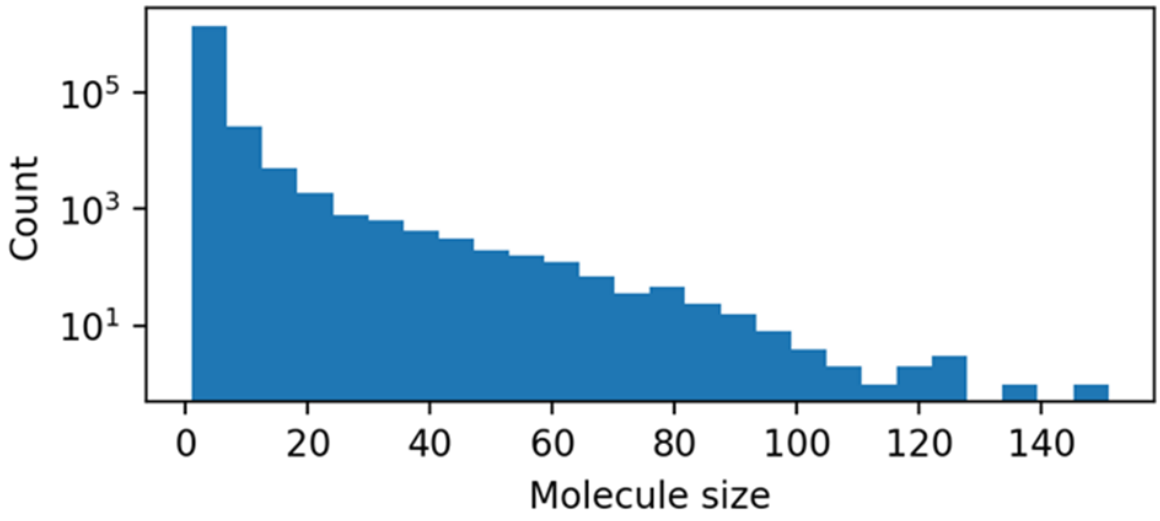

Figure S.14. Distribution of the molecule size (i.e., number of heavy atoms) in the ANI-nr training set.