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

Shuhao Zhang\textsuperscript{1,2}, Malgorzata Z. Mako\text{\c{s}}\textsuperscript{3,4}, Ryan B. Jadrich\textsuperscript{2,5}, Elfri Kraka\textsuperscript{3}, Kipton M. Barros\textsuperscript{2}, Benjamin T. Nebgen\textsuperscript{2}, Sergei Tretiak\textsuperscript{2}, Olexandr Isayev\textsuperscript{1}, Nicholas Lubbers\textsuperscript{4}\textsuperscript{*}, Richard A. Messerly\textsuperscript{2}\textsuperscript{*}, and Justin S. Smith\textsuperscript{2,6}\textsuperscript{*}

\textsuperscript{1}Department of Chemistry, Mellon College of Science, Carnegie Mellon University, Pittsburgh, PA, 15213, USA
\textsuperscript{2}Theoretical Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
\textsuperscript{3}Computational and Theoretical Chemistry Group (CATCO), Department of Chemistry, Southern Methodist University, 3215 Daniel Avenue, Dallas, Texas 75275, USA
\textsuperscript{4}Computer, Computational, and Statistical Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
\textsuperscript{5}Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA
\textsuperscript{6}NVIDIA Corp., San Tomas Expy, Santa Clara, CA 95051, USA
\textsuperscript{*}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, developing reactive MLIPs requires prior knowledge of reaction networks to generate fitting data and refitting to extensive datasets for each new application. 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-nr) 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-nr closely matches experiment and/or previous studies using traditional model chemistry methods. ANI-nr proves to be a highly general reactive MLIP that does not need to be refit for each application, enabling high-throughput \textit{in silico} reactive chemistry experimentation.

Introduction

Atomic-scale reactive simulation is a valuable tool for chemists to plan reaction synthesis or measure material properties \textit{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,\textsuperscript{1} modified embedded atom method,\textsuperscript{2} reactive force field (ReaxFF),\textsuperscript{3} and reactive empirical bond order\textsuperscript{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. 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.\textsuperscript{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. By contrast, QM methods (e.g., density-functional theory, DFT) are based on the underlying physics of electronic structure theory, rather than on a pre-defined bonding pattern, which renders QM calculations highly 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 (MLIP) have been proposed as an alternative to QM and FF methods for the prediction of potential energies and forces.\textsuperscript{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. 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 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. 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. 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, supercritical fluids, or solids) reactive chemistry.

MLIPs have been successfully applied to model chemical reactions in specific contexts, for example, unimolecular/bimolecular gas-phase reaction pathways and specific condensed-phase reactive chemistry simulations. However, each application necessitates a new dataset and retraining of the MLIP. 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 dataset. As such, these reactive MLIPs mirror many of the limitations of reactive FFs. Recent endeavors to build large datasets including reactions have yielded groundbreaking results for developing a general-purpose MLIP. However, random sampling can produce models with poorer performance than targeted, model-aware sampling strategies.

Active learning (AL) 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 by defining an algorithm for data collection. 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. 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), aiming to include only high-uncertainty structures in the ever-growing dataset. The labeling of selected configurations with energies and forces is accomplished by performing QM calculations. Then, these new labeled data are added to the training dataset. 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 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, 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). 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, 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 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 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-nr. To evaluate the reliability of 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 O₂ 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-nr 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-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 dataset produced by our nanoreactor active learning scheme. To demonstrate the state-of-the-art transferability of the ANI-nr potential, we then apply ANI-nr 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 O₂ 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-nr 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 \textit{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.

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. This problem is compounded 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 corresponds with structures sampled at the temperature of interest. For a general reactive MLIP, the temperature of interest is not known \textit{a priori}. Therefore, we employ active learning (AL) combined with a nanoreactor (NR) sampler to generate the ANI-nr training dataset. In brief, the NR sampler randomly initializes condensed-phase systems of small molecules. 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. 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-nr 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-nr dataset.

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-nr 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-nr dataset covers effectively all the space covered by ANI-1x. More importantly, the ANI-nr 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-nr training dataset with the existing PubChem database (only molecules with less than 10 CNO atoms), we conclude that the ANI-nr dataset contains 1212 unique known PubChem molecules. Figure 2 e) shows examples of known PubChem molecules that were found in the ANI-nr 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.

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. With the widespread use of ML methods, researchers are now also trying to investigate amorphous carbon systems with application-specific MLIPs.

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 \textit{ab initio} MD 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. trained a Gaussian approximation potential on structures sampled from DFT-based MD on liquid carbon systems. 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., \textit{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
Figure 3 summarizes the product of each simulation. For each of the high-, medium- and low-density carbon simulations, ANI-nr produces the expected structure of carbon for the respective density. Specifically, 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. 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.

![Diagram of carbon simulations at different densities](image)

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

<table>
<thead>
<tr>
<th>Crystal</th>
<th>Model</th>
<th>$a$ (Å)</th>
<th>$b$ (Å)</th>
<th>$c$ (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond</td>
<td>ANI-nr</td>
<td>3.58</td>
<td>3.58</td>
<td>3.58</td>
</tr>
<tr>
<td></td>
<td>ANI-nr(lr)</td>
<td>3.66</td>
<td>3.66</td>
<td>3.66</td>
</tr>
<tr>
<td></td>
<td>ANI-2x</td>
<td>3.75</td>
<td>3.75</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>Exp.</td>
<td>3.57</td>
<td>3.57</td>
<td>3.57</td>
</tr>
<tr>
<td>Graphite</td>
<td>ANI-nr</td>
<td>2.47</td>
<td>2.47</td>
<td>6.24</td>
</tr>
<tr>
<td></td>
<td>ANI-nr(lr)</td>
<td>2.46</td>
<td>2.46</td>
<td>6.56</td>
</tr>
<tr>
<td></td>
<td>ANI-2x</td>
<td>2.44</td>
<td>2.44</td>
<td>10.0</td>
</tr>
<tr>
<td></td>
<td>Exp.</td>
<td>2.46</td>
<td>2.46</td>
<td>6.71</td>
</tr>
</tbody>
</table>

**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.).

Table 1 contains the crystal lattice constants for diamond and graphite predicted from three ANI-based MLIPs and measured by experiment. 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 potential, while
long-range dispersion interactions are important for \( \pi-\pi \) 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 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-nr that captures long-range interactions while maintaining an accurate description of the local environment.\(^{44}\) We also compare the ANI-nr lattice constants with those from ANI-2x,\(^{45}\) 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-nr dataset include dispersion corrections (see the Methods section for details), which are essential to reproduce the \( c \) lattice parameter in graphite. Finally, all MLIPs 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.\(^{38}\) applied the original \textit{ab initio} NR method to observe ring formation (i.e., the early stages of graphene formation) from a pure acetylene (C\(_2\)H\(_2\)) system. Subsequently, Lei et al.\(^{46}\) 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.\(^{46}\)

Similar to the work of Lei et al.,\(^{46}\) 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 simulations from Lei et al. and experimental literature.\(^{47}\) 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-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. Specifically, for an O\(_2/C_2H_2\) ratio of 0.5, 6-membered rings only begin to form 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 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 Figure S.1 in Supporting Information). The MLIP uncertainty is lower than the AL energy threshold (0.03 kcal·mol\(^{-1}·N^{-1/2}\)) through nearly the entire simulation (with only a few snapshots as exceptions). The relatively low 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-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 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.

**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.\(^{48}\) 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-nr for simulating biodiesel and to investigate the reported ignition enhancement of fuel additives, we reproduced four systems simulated by Chen et al., 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. 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 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 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.

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 bond dissociation energy 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
**Figure 5.** Tracking plot of major products (CO, CO$_2$, and H$_2$O) 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$_2$O are produced (see Figure S.3 in Supporting Information).

<table>
<thead>
<tr>
<th>System</th>
<th>Ignition delay time (ps)</th>
<th>O$_2$ consumption (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ANI-nr</td>
<td>ReaxFF</td>
</tr>
<tr>
<td></td>
<td>($t=0.07$ ns)</td>
<td>($t=2$ ns)</td>
</tr>
<tr>
<td>Clean biofuel</td>
<td>55</td>
<td>239</td>
</tr>
<tr>
<td>Ethanol additive</td>
<td>45</td>
<td>126</td>
</tr>
<tr>
<td>2-butanol additive</td>
<td>46</td>
<td>110</td>
</tr>
<tr>
<td>MTBE additive</td>
<td>45</td>
<td>92</td>
</tr>
</tbody>
</table>

**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 the average time that at least five molecules of CO, CO$_2$, and H$_2$O are produced (see Figure S.3 in Supporting Information). 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.

**Methane combustion**

Emerging research has shown the success of application-specific MLIPs on systems like radical reactions in hydrocarbon combustion and gas-phase S$_N$2.$^{50,51}$ 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. The purpose of this case study is to determine if the ANI-nr dataset adequately samples methane combustion reactions, or if direct application-specific sampling is necessary.

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. We reproduce their MD simulation of methane combustion under the same conditions with the ANI-nr potential. Figure 6 a) shows that the ANI-nr 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-nr 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.

**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) Snapshots of initial reactants, intermediate species, and final products.

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, where Zeng et al. utilized the MN15 functional and 6-31G** basis set. As a preliminary test of this hypothesis, we compare the C-H bond dissociation energies from ANI-nr (116.6 kcal·mol$^{-1}$) with the reference DFT (110.0 kcal·mol$^{-1}$) and experimental (103.3 kcal·mol$^{-1}$) values. Although a gas-phase 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 an underpredicted C-H bond dissociation energy by ANI-nr. However, the relatively large uncertainties in ANI-nr along the bond dissociation path suggest that ANI-nr should not be utilized for studying gas-phase reactions (see Figure S.7 in Supporting Information). This is not surprising since the training dataset was generated with condensed-phase NR simulations.

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 well 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 nanoreactor active learning procedure to generate a general-use reactive MLIP, we analyzed the application-specific dataset of Zeng et al. with ANI-nr. Specifically, we computed the MLIP uncertainty (i.e., the normalized ensemble standard deviation) of ANI-nr for every structure in the Zeng et al. training dataset (see Figure S.8 in Supporting Information). The MLIP uncertainty is smaller than the AL energy threshold (0.03 kcal·mol$^{-1}$·N$^{-1}$) for 99.9% of the structures in their application-specific dataset. This means that our AL approach would only consider $\approx$0.1% (or 72) of their 567,312 structures to be of high enough uncertainty to merit inclusion in our ANI-nr training dataset. 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. Miller applied an electric field to a gaseous system consisting of simple small-molecule species (e.g., NH₃, CO, H₂O, H₂, and CH₄) and reported the formation of amino acids, such as glycine (C₂H₄NO₂). 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 (∼4 ns) 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 Miller NR simulation of Wang et al. However, 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-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 C₂H₄NO₂, 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 C₂H₄NO₂ isomer as reported by Saitta and Saija. By contrast to the Saitta and Saija mechanism, this C₂H₄NO₂ isomer dissociated in our simulation rather than forming glycine. The key precursor to C₂H₄NO₂ is CH₄N, which is formed through several pathways. The pathway to form CH₄N that proceeds through the CH₂O intermediate is very similar to the mechanism reported by Wang et al. The mechanisms to form the intermediates formaldehyde (CH₂O) and hydrogen...
cyanide (CHN) from the initial reactants CO, NH₃, and H₂O were nearly identical to those reported by Wang et al. and Saitta and Saija. The novel pathways to form the key intermediates carbon dioxide (CO₂) and methylene (CH₂) are reported in Figure S.9 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., 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). 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 simulation (see Figure S.11 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 dataset 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 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 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 O₂ 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 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 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, double hybrid DFT provides better accuracy, but 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. 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. Because ANI-nr is trained to a large, general dataset, 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. The ANI-nr 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, as well as cutting-edge graph neural network models 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-nr dataset.

**Methods**

**ANI-nr model descriptions, training details**

The ANI-nr model was trained similarly to ANI models within other contexts, including materials science and chemistry. We use the ANI descriptors, which is a modified form of the Behler and Parinello neural network descriptors. 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-based 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. 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 $1.0 \times 10^{-5}$. Model performance against the held-out test dataset is presented in Table S.IV and Figures S.12 and S.13 of Supporting Information.
Computational nanoreactor active learning for training set generation

The ANI-nr 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 using Kohn–Sham DFT, BLYP functional, TZV2P basis set, GTH pseudopotentials, D3 dispersion correction, 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. The overall spin multiplicity is constrained to a singlet state, consistent with previous studies that perform CP2K simulations of bulk systems containing radical species. 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. We use an uncertainty metric, i.e., the normalized ensemble standard deviation in energy and the forces, 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. Below is a detailed step-by-step description of the active learning workflow:

1. Generate a bootstrap dataset (labeled with energies and forces) of 100 randomly generated periodic cells containing randomly placed and oriented small molecules including C2, H2, N2, O2, NH3, CH4, CO2, H2O, C2H2 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 ($\rho$). Oscillating functional form is the same for temperature and density (see equations below), where $t$ is time, $t_{\text{max}}$ is a hyperparameter for the max time the simulation will run, and $T_{\text{start}}$, $T_{\text{end}}$, $T_{\text{amp}}$, $\rho_{\text{start}}$, $\rho_{\text{end}}$, $\rho_{\text{amp}}$ and $t_{\text{per}}$ are randomly selected values within a predetermined range (see Table S.V in Supporting Information):
     \[
     T(t) = T_{\text{start}} + \frac{t}{t_{\text{max}}} (T_{\text{end}} - T_{\text{start}}) + T_{\text{amp}} \sin^2 \left( \frac{t}{t_{\text{per}}} \right) \\
     \rho(t) = \rho_{\text{start}} + \frac{t}{t_{\text{max}}} (\rho_{\text{end}} - \rho_{\text{start}}) + \rho_{\text{amp}} \sin^2 \left( \frac{t}{t_{\text{per}}} \right)
     \]
   - Run nanoreactor MD simulation using forces from current AL generation MLIP

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

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

6. Add new labeled data to the training dataset.

7. 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 Figures S.14, S.15, and S.16 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.17 in Supporting Information shows a histogram of the sizes of all molecules that are found in the ANI-nr 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.

**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.

**Effect of oxygen on graphene ring formation**

To investigate ring formation from acetylene, MD simulations were performed for eight different systems with varying $\text{O}_2/\text{C}_2\text{H}_2$ ratios: (0.00, 0.08, 0.17, 0.22, 0.38, 0.50, 0.86, 1.33). The initial structures were generated 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 $\text{O}_2$ molecules were the same as shown in Table 2 of the ReaxFF reference paper. 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 run 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. 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.

**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. Snapshots and properties were recorded every 0.1 ps (1000 time steps).
**Miller experiment**

To investigate the ability to simulate complex organic systems 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 H₂, 14 H₂O, 14 CO, 14 NH₃ and 14 CH₄ 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**

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**


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
<table>
<thead>
<tr>
<th>Crystal</th>
<th>Model</th>
<th>$\alpha$ ($^\circ$)</th>
<th>$\beta$ ($^\circ$)</th>
<th>$\gamma$ ($^\circ$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diamond</td>
<td>ANI-nr</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>ANI-nr(lr)</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>ANI-2x</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td></td>
<td>Exp.</td>
<td>90.0</td>
<td>90.0</td>
<td>90.0</td>
</tr>
<tr>
<td>Graphite</td>
<td>ANI-nr</td>
<td>90.0</td>
<td>90.0</td>
<td>120.</td>
</tr>
<tr>
<td></td>
<td>ANI-nr(lr)</td>
<td>90.0</td>
<td>90.0</td>
<td>120.</td>
</tr>
<tr>
<td></td>
<td>ANI-2x</td>
<td>90.4</td>
<td>89.7</td>
<td>120.</td>
</tr>
<tr>
<td></td>
<td>Exp.</td>
<td>90.0</td>
<td>90.0</td>
<td>120.</td>
</tr>
</tbody>
</table>

**Table S.I.** Optimized crystal angles ($\alpha$, $\beta$, $\gamma$) for diamond and graphite phases. Comparison between ANI-nr, ANI-nr(lr), ANI-2x and experiment (Exp.).
Figure S.1. ANI-nr ensemble standard deviation in energy normalized by the square-root of number atoms ($\varepsilon$) 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$_2$O. 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$_2$O. (a) biofuel+O$_2$ system (b) biofuel with ethanol+O$_2$ (c) biofuel with 2-butanol+O$_2$ (d) biofuel with MTBE+O$_2$. 

S.5/S.22
Figure S.4. \( \text{O}_2 \) consumption (%) during combustion for clean biofuel compared with three different fuel additives, namely, ethanol, 2-butanol, and MTBE. Insert compares \( \text{O}_2 \) 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\textsubscript{2}, (a) biofuel+O\textsubscript{2} system (b) biofuel with ethanol+O\textsubscript{2} (c) biofuel with 2-butanol+O\textsubscript{2} (d) biofuel with MTBE+O\textsubscript{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-nr 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-nr ensemble standard deviation in energy normalized by the square-root of number atoms ($\epsilon$). Among 567312 available structures in their training set, only 72 structures have normalized energy uncertainty larger than the AL energy threshold of 0.03 kcal $\cdot$ mol$^{-1} \cdot N^{-\frac{1}{2}}$. 

\[ \epsilon = 0.03 \text{ kcal} \cdot \text{mol}^{-1} \cdot N^{-\frac{1}{2}} \]
Figure S.9. Formation of intermediates CH$_2$ and CO$_2$ from initial reactants (NH$_3$, CO, CH$_4$, H$_2$O). Mechanism to form CH$_2$O is found in Figure 7 in the main text. Green arrows denote reactions previously identified by Wang et al. or Saiita and Saija. Orange arrows denote reactions that have closely-related reactions in Wang et al. or Saiita and Saija.
Figure S.10. 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$_3$O, which is likely a cationic species (H$_3$O$^+$). The penultimate species (C$_2$H$_4$NO$_2^-$) 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.11. 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.

<table>
<thead>
<tr>
<th>Layer ID</th>
<th>H</th>
<th>C</th>
<th>N</th>
<th>O</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>256 CELU 224 CELU 192 CELU 192 CELU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>192 CELU 190 CELU 160 CELU 160 CELU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>160 CELU 160 CELU 128 CELU 128 CELU</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1 Linear 1 Linear 1 Linear 1 Linear</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table S.II. ANI-nr neural network architecture**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Radial Cutoff (Radial) (Å)</td>
<td>5.2</td>
</tr>
<tr>
<td>Radial Cutoff (Angular) (Å)</td>
<td>3.5</td>
</tr>
<tr>
<td>Radial Eta (Å⁻²)</td>
<td>65.7</td>
</tr>
<tr>
<td>Radial Shift (Å)</td>
<td>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</td>
</tr>
<tr>
<td>Angular Zeta (-)</td>
<td>14.1</td>
</tr>
<tr>
<td>Angular Angular Shift (rad.)</td>
<td>0.39269908,1.1780972,1.9634954,2.7488936</td>
</tr>
<tr>
<td>Angular Eta (Å⁻²)</td>
<td>10.1</td>
</tr>
<tr>
<td>Angular Radial Shift (rad.)</td>
<td>0.500,0.875,1.250,1.625,2.000,2.375,2.750,3.125</td>
</tr>
</tbody>
</table>

**Table S.III. ANI-nr symmetry function parameters**
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.

<table>
<thead>
<tr>
<th>Property</th>
<th>RMSE</th>
<th>MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Energy (kcal mol⁻¹ N⁻¹)</td>
<td>0.43 ± 0.22</td>
<td>0.1756 ± 0.0061</td>
</tr>
<tr>
<td>Force (kcal mol⁻¹ Å⁻¹)</td>
<td>10.34 ± 0.25</td>
<td>6.306 ± 0.074</td>
</tr>
</tbody>
</table>
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.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_{\text{start}}$</td>
<td>1000 - 3000 K</td>
</tr>
<tr>
<td>$T_{\text{end}}$</td>
<td>100 - 2000 K</td>
</tr>
<tr>
<td>$T_{\text{amp}}$</td>
<td>0 - 2000 K</td>
</tr>
<tr>
<td>$\rho_{\text{start}}$</td>
<td>0.1 - 2 g/cc</td>
</tr>
<tr>
<td>$\rho_{\text{end}}$</td>
<td>0.5 - 2 g/cc</td>
</tr>
<tr>
<td>$\rho_{\text{amp}}$</td>
<td>0 - 0.75 g/cc</td>
</tr>
<tr>
<td>$t_{\text{per}}$</td>
<td>$T$: 2 - 50 ps; $\rho$: 0.5 - 50 ps</td>
</tr>
</tbody>
</table>

**Table S.V.** Parameters for nanoreactor oscillations in temperature and density.
Figure S.14. Histogram of the system size (i.e., number of atoms) per system in the ANI-nr training data set.
Figure S.15. Histogram of the system composition of all systems in the training data set, colored by element.
Figure S.16. Histogram of the mass density (g/cc) of all systems in the training data set.
Figure S.17. Distribution of the molecule size (i.e., number of heavy atoms) in the ANI-nr training set.