Micron-scale heterogeneous catalysis with Bayesian force fields from first principles and active learning

Anders Johansson  
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA USA  
andersjohansson@g.harvard.edu

Yu Xie  
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA USA  
xiey@g.harvard.edu

Cameron J. Owen  
Harvard Department of Chemistry and Chemical Biology, Cambridge, MA USA  
cowen@g.harvard.edu

Jin Soo (David) Lim  
Harvard Department of Chemistry and Chemical Biology, Cambridge, MA USA  
limsj@g.harvard.edu

Lixin Sun  
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA USA  
lixinsun@fas.harvard.edu

Jonathan Vandermause  
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA USA  
jonathan_vandermause@g.harvard.edu

Boris Kozinsky  
Harvard John A. Paulson School of Engineering and Applied Sciences, Cambridge, MA USA  
bkoz@seas.harvard.edu

Abstract—Quantum-mechanically accurate reactive molecular dynamics (MD) at the scale of billions of atoms has been achieved for the heterogeneous catalytic system of \( \text{H}_2/\text{Pt}(111) \) using the FLARE Bayesian force field. This achievement provides accelerated time-to-solution from first principles, with Bayesian active learning enabling efficient and autonomous training of the machine learning model. The resulting model is then deployed in LAMMPS on GPUs using the Kokkos performance portability library. The Bayesian force field provides quantitative uncertainty of predictions on every atomic environment, critical for detecting configurations in large reactive simulations that are outside of the training set. Scaling benchmarks were performed using real-application MD of the \( \text{H}_2/\text{Pt}(111) \) heterogeneous catalysis on the Summit supercomputer, with simulations reaching 0.5 trillion atoms on 4556 GPU nodes.

I. JUSTIFICATION

First micrometer-scale MLFF MD simulations with peak performance of 10.5 M atom-steps/s/node, 70% improvement over state-of-the-art. MLFF benchmarks of 0.5T atoms, 25× greater than state-of-the-art. Perfect weak scaling to 40B atoms on 4096 GPU nodes. First uncertainty-aware large-scale reactive MD. Time-to-solution accelerated by orders of magnitude through Bayesian active learning.

II. PERFORMANCE ATTRIBUTES

| Performance Attribute | Our Submission |
|-----------------------|----------------|
| Category of achievement | Peak performance, scalability, time to solution from first principles |
| Performance | 10.5 M atoms-steps/s/node |
| Maximum problem size | 0.5 trillion atoms |
| Type of method used | FLARE/Kokkos via LAMMPS molecular dynamics with active learning |
| Results reported on basis of | Whole application including I/O |
| Precision reported | Double precision |
| System scale | Results measured on full-scale system 4556 GPU nodes (27336 GPUs) on OLCF Summit |
| Measurement mechanism | Timers in LAMMPS |

III. PROBLEM OVERVIEW: ELUCIDATING AND PREDICTING THE ACTIVITY OF HETEROGENEOUS REACTIONS

Interface reactions play a crucial role in determining the performance and durability of functional materials, including corrosion of metals, tribology of lubricated surfaces, mechanical properties of alloys and composites, transport across electrolyte-electrode interfaces in batteries, and heterogeneous catalysis. Heterogeneous catalysis lies at the core of sustainable chemical production, and rational catalyst design is required to move beyond the conventional trial-and-error approaches [1]. In this regard, elucidating surface reactivity is central to improve catalytic performance. But surfaces can undergo restructuring under reactive environments. Such dynamical processes necessitate elucidation and atomistic modeling of the underlying mechanisms and timescales.
Interfacial phenomena involve large length-scales and long timescales. Computational modeling at such scales presents a massive challenge, but also an incredible opportunity for combining state-of-the-art (SOTA) methods development with leadership-class computing resources. *Ab initio*, or first principles, methods based on the explicit quantum mechanical treatment of the electronic structure, such as density functional theory (DFT), are essential for establishing the structure-activity relationship of catalytic surfaces [2]. However, their poor scaling and high computational cost ($O(N^2)$ with the number of electrons) preclude their use for large-scale and long-timescale simulations. First-principles modeling of catalytic systems are often limited to static configurations or fast high-temperature reactions on small periodic structures [3] or nanoclusters [4]. Such simulations typically span picoseconds and nanometers, far behind the microsecond and micrometer time- and length-scales of real catalysts.

For complex surface structures with a wide variety of dynamic reactive sites, it is highly nontrivial to replace *ab initio* methods with cheaper and more scalable methods. Previously, cluster expansions [5] and empirical [6] / machine-learned (ML) [7] force fields have been employed. These methods rely only on the atomic positions in local environments and do not explicitly consider the electronic structure, thereby scaling linearly with the number of atoms $O(N_{\text{atom}})$ and providing several orders-of-magnitude acceleration compared to DFT.

However, the accuracy, transferability, and speed of these models have remained limited for applications in computational catalysis. One such example can be found in the modeling of H$_2$ reactivity on transition metal surfaces, a prototypical system in heterogeneous catalysis. H$_2$ reactivity has been a subject of extensive experimental and computational investigation due to its industrial importance in numerous catalytic processes (e.g., selective hydrogenation [8] and hydrogen storage [9]). On noble metals such as Pt, an H$_2$ molecule approaches the Pt surface and dissociates to form two chemisorbed H atoms, which diffuse and eventually recombine and desorb from the surface as molecular H$_2$ (Figure 1) [10].

![Fig. 1: Mechanisms of H$_2$ reactivity on Pt(111) surface. (a) Dissociative adsorption (chemisorption) of H$_2$ on Pt(111); (b) diffusion of atomic H on Pt(111); (c) recombinative desorption of H$_2$ from the surface.](image)

A variety of force field-based methods have been employed to model H/Pt, most notably ReaxFF [11], embedded-atom method (EAM) [12], tight-binding [13], low-dimensional models [14], and ML force fields [7]. Empirical models such as ReaxFF and EAM have been shown to be inaccurate and unreliable for high-temperature long-timescale simulations [15]. Several challenges impede the applicability of available models. The first challenge for reactive force fields is the computational cost that increases with model complexity, which limits the size and time scales of simulations. Second is the limited expressiveness of their parametric functional forms, which lower their accuracy and transferability: they only consider either a bare surface interacting with a single H$_2$ molecule or H-covered surface coupled to an implicit reservoir representing the gas phase, failing to model ambient pressure conditions relevant to real-life catalysis, such as the Haber-Bosch process which operates at pressures well above 100 bar. Such conditions involve multiple gas-phase H$_2$ molecules and chemisorbed H atoms on the surface. Third, none of the existing reactive models provide any principled metric of uncertainty or reliability. Uncertainty quantification is critical for evaluating model reliability for configurations that may be outside of the training set. This is especially critical for simulations involving rare transition or migration events that occur in chemical reactions. Finally, force field models typically require laborious ad hoc manual enumeration of structural configurations for generating the training set.

The challenges outlined above hold generally for the modeling of reactive systems, not just the particular H/Pt system. Hence, progress in computations of complex reactive systems necessitates the development of flexible models for force fields that are (1) scalable and fast, reaching large sizes and long timescales of simulations needed to capture reaction kinetics, (2) chemically accurate and transferable to a range of reaction conditions and complex chemical environments, (3) endowed with predictive uncertainty quantification and reliability guarantees, and (4) quickly and autonomously trained.

IV. CURRENT STATE OF THE ART

Over the last decade, ML force fields have emerged as a promising path to highly accurate and scalable atomistic simulations. Notable developments include Behler-Parrinello neural-network potentials (HDNNP) [16], Gaussian approximation potentials (GAP) [17], spectral neighbor analysis potentials (SNAP) [18], moment tensor potentials (MTP) [19], ANI neural-network potentials [20], SchNet [21], DeePMD [22], atomic cluster expansion (ACE) [23], and NequIP [24]. These MLFFs can in principle make predictions at near *ab initio* accuracy, but none of the previous models simultaneously address all four criteria listed in Section III. Here we discuss the current state of the art in the four aspects separately.

A. Scalability and Speed

Two existing descriptor-based MLFF models demonstrated the capability to simulate billions of atoms, but only describing dynamics of periodic bulk systems. Importantly, none of the existing models describe reactive dynamics, in either hetero-
genuine or even homogeneous systems. The SNAP model was employed on bulk carbon system with peak performance 6.21 M atoms-steps/s/node [25]. Likewise, DeePMD achieved 0.3 M atoms-steps/s/node for homogeneous bulk systems - water and copper - on the OLCF Summit machine [26]. Other models, like graph neural networks, have not been shown to approach such scales due to difficulty in parallelization. On Summit, DeePMD and SNAP are able to simulate 3.9 and 20 billion atoms, respectively, using 4560 and 4650 GPU nodes, as listed in Table I.

B. Accuracy and Transferability

As mentioned in section III, existing reactive force fields based on either empirical or ML models suffer from transferability and accuracy issues. Specifically for the H/Pt heterogeneous systems, previous force fields are limited to describing low-pressure and low-temperature conditions. Latest models include a ReaxFF model from Gai et al [11] and a HDNNP from Gerrits [7]. Both models were optimized to match DFT calculations of H-covered Pt surfaces. The ReaxFF model was designed to capture H adsorption behavior, while the HDNNP can model a single H2 molecule from a molecular beam interacting with the surface. However, from our benchmarks [15], models constructed with only low-energy structures without an explicit gas-surface interface do not have sufficient accuracy in describing the ambient pressure environment, and thus cannot thermally activated states of the full catalytic reaction pathway.

C. Uncertainty Quantification at Scale

Bayesian models such as Gaussian processes (GPs) provide both mean and variance predictions, where the variance can be used to quantify uncertainty. But the computational cost of predicting the variances scales with the training set size as $O(n_s^2)$, more expensive than the $O(n_s)$ scaling of predicting the mean. This inferior scaling applies to the Bayesian regression approach used by Jinnouchi et al [28], limiting the application to short time scales and small length scales. Podryabinkin et al used D-optimality as an indicator of uncertainty of a linear descriptor-based model [29], which provides only a qualitative decision boundary but not quantitative uncertainty for each structure. In the class of neural network models, such as DeePMD, the “ensemble” technique is used to approximate uncertainty by the variance of predictions coming from multiple light-weight models [30]. This approach requires additional computational resources for multiple neural networks, and a dataset of several thousand structures, while yielding a final model still without uncertainty. In an earlier version of the FLARE model with 2+3-body kernels [31, 32], we verified uncertainty accuracy and circumvented this scaling problem by an exact mapping of the GP mean and variance to spline functions, reducing the prediction cost of energies and forces, together with their uncertainty, to a cost comparable with empirical 3-body force fields. This represents the SOTA, which we improve on using many-body kernels and sparse GPs [15], as described in Section V.

D. Time-to-solution from first principles

Time-to-solution for MLFF has previously been defined as the production simulation speed of the finally trained model, i.e. the M atom-steps/s/node [26, 25]. However, the cost of generating relevant and sufficiently diverse first-principles reference data and training the force fields can take months and requires domain knowledge and arduous human trial and error. For example, without uncertainty quantification and active learning, the SNAP potential of carbon is trained from static DFT and ab-initio MD data [53], which takes days or weeks to generate. DeePMD used a relatively large dataset of 7646 frames of DFT data for a homogeneous pure bulk copper model [30]. Training of empirical force fields, like ReaxFF, relies on manual selection of relevant atomic configurations and typically takes months of human iteration. The generation cost increases rapidly with structural and chemical complexity, as well as rare events like chemical reactions. Therefore, “time-to-solution from first principles” in this work emphasizes the entire data workflow starting from system description, selecting training configurations, performing reference DFT calculations, and training the final model.

Active learning workflows can replace this arduous human-driven process. In this approach, molecular dynamics or another sampling method is used to explore the configuration space with the fast surrogate model, and expensive ab initio methods are only called when the model uncertainty exceeds a threshold. To actively train their corresponding models mentioned above Zhang et al used the ensemble method to train DeePMD force fields [50], Podryabinkin et al used D-optimality [29], and Jinnouchi et al used Bayesian variance [28]. As these models do not produce large-scale MD with uncertainties, no further iterative model refinement is possible. Our earlier FLARE 2+3-body model was used to perform hierarchical active learning where mapped quantitative uncertainty of sequentially larger-scale MD was used to detect configurations outside the training and to refine the model [52]. In this work, we improve upon this SOTA approach by implementing many-body descriptor based kernels [34] as described in Section V.

V. INNOVATIONS REALIZED

In this work, we employ the FLARE framework to develop a fast and accurate Bayesian many-body force field and simulate the heterogeneous H/Pt catalytic reaction, demonstrating significant improvements over the SOTA performance in the following aspects:

1) The speed of our model exceeds both the non-reactive DeePMD benchmark on simple thermal dynamics in homogeneous systems by 30× for bulk water [26] and 4× for bulk copper [27], the SNAP benchmark by 1.7× for bulk carbon [25], and the reactive force field ReaxFF by at least 5×. ReaxFF benchmark was on 1M atoms on one V100 GPU node, and does not scale linearly with the system size. (Section V.B and V.C).

2) The accuracy of FLARE exceeds the only available model for H/Pt heterogeneous catalysis reactions, ReaxFF, by an
order of magnitude in the mean absolute error of force predictions relative to first-principle calculations, also evidenced by nonphysical structural evolution observed with ReaxFF at realistic chemical conditions [15].

3) Predictive uncertainty quantification is implemented in FLARE with minimal impact on speed, even in 1B-atom scale MD calculations. This capability is unique among any available machine learning potentials. (Section V-B)

4) The time-to-solution from first principles is 214.8 wall time hours, and 568.8 node hours realized by fully autonomous Bayesian active learning, which outperforms other machine learning potentials trained with AIMD data [33] by orders of magnitude. The number of first-principle calculations used for our training is 13× smaller than the DeePMD copper potential from DP-GEN [30], demonstrating a higher data efficiency. (Section V-A)

With the above implementation, we are able to model the heterogeneous H/Pt catalytic process at micrometer length scales under realistic reaction conditions, making this the largest reactive MD simulation to date.

A. Bayesian Active Learning

Utilizing the inherent uncertainty quantification of the FLARE Bayesian force field (BFF), we design the Bayesian active learning approach, enabling orders of magnitude higher efficiency in data generation than using ab-initio MD [31], [15], as detailed in Table II. As illustrated in Figure 2, the MD simulation is driven by the fast Bayesian force field, with the uncertainty calculated at each step. When the uncertainties are greater than a decision threshold \( \Delta_{\text{DFT}} \), indicating that the current structure is significantly different from those stored in the training set, the MD is interrupted and DFT is called on the uncertain structure. The results of that DFT calculation are then added to the training set, and the Bayesian force field model is updated.

At the beginning of the training procedure, when the training set is small, DFT needs to be called frequently. As the training set grows, DFT is called less frequently, until the simulation has sufficiently explored the configuration space, meaning that a diverse training set is automatically created. Due to the high data efficiency of the GP, we can obtain a trustworthy model for MD simulations, while the uncertainty quantification helps to detect unexplored region and improve the model accuracy.

B. Acceleration of the Bayesian Force Field

In this work, we use the FLARE Bayesian force field framework, a descriptor-based model with not only the energy, forces and stress, but also the uncertainty quantification associated with the predictions [15]. The model is constructed with three parts: the descriptors, the kernel, and the Gaussian process regression model.

Local structure descriptors. A local atomic environment \( \rho_i \) of an atom \( i \) consists of all the neighbor atoms within a cutoff radius \( r_{\text{cut}} \) and includes information of atomic positions and chemical species. Atomic cluster expansion (ACE) B2 descriptors [23] are used to describe the local atomic envi-

![Figure 2](image-url)  
**Fig. 2:** a. Bayesian active learning workflow driven by predictive uncertainty. b. H/Pt(111) configuration for training. c. Atoms colored by uncertainty during the training. [15]

![Figure 3](image-url)  
**Fig. 3:** Left: local atomic environment. Upper right: examples of spherical harmonics used for the descriptors. Lower right: GP regression (schematic).

Kernel. The kernel function \( k(\rho_1, \rho_2) \) quantifies the similarity between two atomic environments. We use the normalized inner product kernel: \( k_\xi(\rho_1, \rho_2) = \sigma^2 (d_1 \cdot d_2) \xi \), where \( \sigma^2 \) is the signal variance optimized by maximizing the log likelihood of SGP, and \( \xi \) is the power of the inner product.

\[
e(\rho_i) = \sum_{s \in S} k_2(\rho_1, \rho_s) \alpha_s = \sum_{s \in S} d_i^* \beta d_i, \\
V(\rho_i) = \sigma^2 k_1(\rho_1, \rho_i) - \sigma^4 \sum_{s, t \in S} k_1(\rho_i, \rho_s) (K_{S S}^{-1})_{s t} k_1(\rho_t, \rho_i) = d_i^* \gamma d_i,
\]

where the \( \alpha_s \) are constants computed from training labels and kernels between training data, \( S \) is the set of sparse
representatives of the training data set, and $K_{SS}$ is the kernel matrix between the environments in $S$. The uncertainties are

given by the square root of the variance. As a trade-off between accuracy and performance, we use the $\xi = 2$ kernel for mean and the $\xi = 1$ kernel for variance evaluations (see detailed discussions in [15], [34]).

As shown in Eq. (1), the cost of energy/forces/stress predic-
tions scales as $O(n_S n_d)$, and the uncertainty predictions as $O(n_S^2 n_d)$, where $n_d$ is the descriptor dimension, and $n_S$ the sparse training set size. For complex systems, a large number of training data are usually needed for sufficient accuracy, indicating $n_S \gg n_d$, which makes the naive evaluation of the SGP model slow. In FLARE, we introduced a key unique way of reducing the scaling of the mean and variance to $O(n_S^2)$ [15], [34]. Specifically, with inner product kernels, a highly efficient but lossless exact mapping is available via reorganization of the summation, as shown in Eq. (1). The $\beta$ and $\gamma$ tensors can be computed from the training data descriptors once and stored, and during the prediction we directly evaluate the quadratic model without the explicit summation of all sparse data $S$, making the computational cost independent of the training data set size $n_S$.

**Acceleration of force predictions.** In SGP, the evaluation of forces requires computing the gradient of the descriptors $\nabla d_i$ and is the bottleneck of the performance. We observe that in the mapped quadratic model, this can be circumvented by reordering the summation, such that the cost of the forces is reduced by a factor of $n_{\text{max}}$, the number of radial basis functions in the descriptors. This provides an overall speedup of approximately $2 \times$ compared to calculating forces directly from $\nabla d_i$.

**Acceleration of heterogeneous multi-element system.** The dimension of the ACE-B2 descriptors grows as the square of the number of chemical elements. If an atom only has neighbors of its own chemical species, such as in the Pt bulk or in the H$_2$ gas, the majority of the descriptor components will be zero. By checking the neighborhood of each atom before the descriptor calculation, we avoid computing unnecessary components and drastically reduce the size of the quadratic model in Eq. (1). For all atoms not near the surface, the dimension of the descriptor is reduced from 420 to 112, thus accelerating the quadratic model by more than an order of magnitude.

**C. Hardware acceleration with Kokkos**

Over the past few years, hardware accelerators such as graphical processing units (GPUs) have seen widespread adaptation in high-performance computing (HPC) centers around the world, with 7 out of 10 most powerful supercomputers in the world being GPU-based [35]. GPUs offer massive parallelism with a large number of cores, making them ideal for MD simulations where the forces and energies of large numbers of atoms need to be computed.

The Kokkos performance portability library [36] is developed for hardware-agnostic programming, which automatically fits different architectures and programming models such as OpenMP, CUDA and HIP during compilation. Kokkos has been tightly integrated with the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS) [37], such that MD runs entirely on GPUs with zero CPU-GPU data transfer except for I/O processes. All CPU-GPU and GPU-GPU communication is handled by LAMMPS’s well-proven routines, giving us easy access to excellent performance and scalability.

We integrate the FLARE force field with LAMMPS and optimize the performance by making extensive use of Kokkos’s hierarchical parallelism functionality, which is directly mapped to CUDA blocks, warps and threads. For most computational steps, one CUDA block of threads will work together on one atom, while the intra-block parallelism is exploited for an atom’s neighbors and/or the components of its descriptor vector. For example, each CUDA-block performs one $\beta d_i$ matrix-vector product in Eq. (1), and each warp of threads computes one component of the result with a parallel dot-product between one matrix row and the descriptor vector. The repeatedly accessed descriptor vector is loaded into the highest-performance shared memory on the GPU. The matrix-vector product is computed in chunks whenever the full-sized descriptor vector is too large to fit into shared memory.

The multi-level parallelism enables saturated performance on GPUs for small systems of hundreds of atoms and extreme-scale systems with billions of atoms, thus utilizing the full power of GPUs for both long time-scale and large length-scale simulations. Descriptor-based methods like FLARE are memory-hungry because the derivative of every descriptor component of every atom with respect to each neighbor is required. The issue is avoided by dividing atoms into batches, so that only a subset of descriptors and gradients are stored at a time. The batch size is automatically calculated based on memory. Due to the multi-level parallelism, peak performance is achieved even for small batch sizes, thus LAMMPS’s neighbor lists ultimately become the limiting factor in terms of system size.

**VI. How Performance Was Measured**

**A. Scientific applications used to measure performance**

As discussed in Section III, it is critical to model catalytic reactions at conditions relevant to real-life catalysis. To model the prototypical H$_2$ reaction at ambient pressures, we consider a combined system of H$_2$ gas interacting with a Pt(111) slab (Figure 5). The Pt surface has 885 million atoms with the area of 2.2 $\mu$m $\times$ 1.9 $\mu$m, while 120 million H atoms are scattered in the space, yielding a total of 1 billion atoms.

Such large system sizes allow for meaningful direct two-phase reaction simulations at experimentally relevant conditions. Our previous simulations show that a smaller supercell with 448 H atoms only observes twenty total reactions (splitting and recombination) within 500 ps at 450 K [15]. The limited number of events is not sufficient for statistical analysis of reactivity. This problem can be mitigated by drastically increasing the length-scale with a larger Pt surface and 120
million H atoms, allowing more reactions to happen in a short period of time and providing sufficient statistics for studying the reaction rates and mechanisms.

Fig. 4: Perspective view of the 1B-atom system and its dimensions. The height of the simulation box is not shown to scale.

The large-scale H/Pt simulations were prepared in three steps. A smaller supercell with 216000 Pt and 29200 H atoms is first generated by randomly assigning H2 molecules to the vacuum and a 0.8 monolayer of H atoms adsorbed on the Pt surface. This supercell was thermalized at 450 K for 2.1 ns till the surface coverage of H approached equilibrium. The supercell was then replicated to yield the final structure.

Catalytic reactions were modeled at 450 K using the Nosé-Hoover thermostat. A time step of 0.1 fs was employed to account for the much faster vibrational mode of H compared to Pt. Snapshots of the trajectory were dumped every 0.1 ps, allowing for post-processing of the data to determine the number of reactions observed. Hydrogen coordination numbers were determined to be 1 or 0 whether another hydrogen atom was within 1 Å, where a change in this coordination number would correspond to a reaction event taking place (e.g., $1 \rightarrow 0$ being dissociation & $0 \rightarrow 1$ being recombination). Furthermore, atomic uncertainties were calculated at a frequency of 1 ps to provide insight into the reliability of the physics during the trajectory. The uncertainties observed over the course of our two simulations are provided in Figure 5.

Prior to deployment, the resulting MLFF was validated against DFT and compared with the ReaxFF model [11], as reported previously [15]. The parity plots for force predictions are shown in Figure 6. The mean absolute errors (MAE) for energies, forces, and stresses (trace) are 1.7 meV/atom, 91 & 74 meV/Å (Pt & H), and 0.6 meV/Å3 with our FLARE model, all of which are an order of magnitude lower than 33 meV/atom, 631 & 676 meV/Å (Pt & H), and 15.7 meV/Å3 obtained with the ReaxFF model, respectively.

The full transition state pathways were examined for H2 dissociative adsorption and atomic H diffusion on Pt(111) at the low-coverage limit [15]. The configuration energies are in excellent agreement between our FLARE model and DFT. ReaxFF is precluded from transition state modeling here, as it wrongly predicts the top site to be the most stable by 0.16 eV compared to the HCP hollow site.

The ultimate accuracy performance test for reactive MD simulations is the comparison with experimental measurements of reaction rates. In that regard, our first-iteration H/Pt FLARE model produced an estimate of the effective exchange activation energy of 0.25 eV in excellent agreement with the experimentally measured value of 0.23 eV [15]. This indicates that the FLARE MD simulation accurately captures all the steps of the catalytic reaction chain and retains first-principles accuracy.

B. Systems and Environment for Measurement of Performance

All simulations were performed using the Kokkos portability library in conjunction with LAMMPS MD.

All of our scaling and production MD simulations were performed on Summit, the GPU-based, leadership-class HPC system at Oak Ridge National Laboratory. Summit has 4608 nodes, each with 6 NVIDIA V100 GPUs and an IBM Power9.
We used the LAMMPS version stable_29Sep2021_update2 with our FLARE extension. The code was compiled with GCC 9.3, CUDA 11.0 and Spectrum MPI 10.4, using CUDA-aware MPI for efficient GPU-GPU communication.

C. Measurement metrics

We consider the overall time-to-solution from first principles defined in section [V] as the key measurement metric: it is the sum of training data generation, force fields training, and production time. For the first two steps, we simply measure the wall time required to run the different Bayesian active learning simulations, and then combine the resulting training data into one complete model. We compare the wall time of the active learning run with the time estimate of using AIMD for every step.

For the final molecular dynamics, we perform extensive scaling experiments for a wide range of system sizes and computational resources, to demonstrate close to the ideal linear scaling of MD. We measure strong scaling by increasing the number of nodes for a fixed system size and weak scaling by maintaining a constant per-node system size. Each simulation is run for 2 minutes (excluding the initial setup) to ensure sufficient sampling. We report the results in the unit of mega-atom-steps per second per node, as this is a convenient measure that should be constant for all simulations with linear scaling. Finally, we track the fraction of the simulation spent computing forces.

VII. PERFORMANCE RESULTS

In this section, we provide details of the SOTA performance of FLARE on scalability/production and training from first principles. In Table I we compared FLARE with the two other MLFF models used for large-scale MD simulations on Summit.

A. Scalability

We performed strong scaling experiments with system sizes from $10^8$ to $10^{11}$ atoms, with up to 4096 nodes (24576 GPUs) on Summit in powers of 2. We also performed one benchmark with 0.5 trillion atoms, which is by far the largest MLFF benchmark ever performed and required 4556 nodes. The $10^8$, $10^{10}$ and $10^{11}$ atom simulations needed 8, 128 and 1024 nodes to fit in GPU memory, respectively, when FLARE was limited to use up to 2 GB for its batching of atoms. When going beyond $10^{10}$ atoms, we found that the initial, CPU-based replication of the 245200-atom structure became too expensive. Instead, the $10^{11}$ benchmarks were initiated using a faster protocol of creating a lattice of 272-atom unit cells, which resulted in a structure with a slightly different Pt surface H coverage value.

The strong scaling results are shown in Figure 7. In the top panel, the speed of FLARE is compared to the peak performance of SNAP [25] and DeePMD [27]. FLARE outperforms SNAP and DeePMD by factors of 1.7 and 5.3, respectively, when in the regime of linear scaling, achieving a peak performance of $10.5 \text{ M atom-steps/s/node}$. For our chosen PtH system, we observe that FLARE leaves the linear scaling regime earlier than for a simple bulk system like silicon. In the bottom panel, we measure the fraction of time spent inside FLARE during the strong scaling simulations. The deviation from 100 % aligns perfectly with the deviation from linear scaling and demonstrates the impact of communication and load imbalance. Communication and load imbalance are much more severe issues when running complicated heterogeneous, near-two-dimensional systems like PtH than when running simple bulk crystals.

![Graph](image_url)  
**Fig. 7:** Strong scaling from $10^8$ to $5 \cdot 10^{11}$ atoms. Top: Speed comparison with the peak performance for SNAP ($2 \cdot 10^{10}$ atoms) and DeePMD ($3.4\cdot 10^9$ atoms). Bottom: The percentage of the simulation spent computing forces.

Figure 8 shows the weak scaling of FLARE by maintaining approximately $10^7$ atoms per node and scaling from 1 to 4096 nodes. We compare the weak scaling of a simple crystal system, silicon, to that of our heterogeneous, catalytic system. We find near-perfect weak scaling in both cases, but the slightly improved scaling of the silicon simulations demonstrates the effect of load imbalance for the catalytic system and its heterogeneous geometry. Based on these observations, we decided to continuously perform time-weighted re-balancing of the spatial decomposition for long-timescale production runs.
DFT calls for different configurations at different conditions, shown in Table II, Bayesian active learning only makes 575 converge the model to a satisfactory accuracy. For example, as [31], only a small number of DFT calculations are needed to the high data efficiency of the kernel-based regression while monitoring rare events by uncertainty. Meanwhile, due to exploration of configuration space with a fast surrogate model to realize the Bayesian active learning workflow, which facilitates the principled uncertainty of the Gaussian process model to re-

C. Time to solution from first principles

As introduced in Section V, FLARE utilizes the quantitative principled uncertainty of the Gaussian process model to realize the Bayesian active learning workflow, which facilitates exploration of configuration space with a fast surrogate model while monitoring rare events by uncertainty. Meanwhile, due to the high data efficiency of the kernel-based regression [31], only a small number of DFT calculations are needed to converge the model to a satisfactory accuracy. For example, as shown in Table II, Bayesian active learning only makes 575 DFT calls for different configurations at different conditions, while the same exploration trajectory would take 138.7 years to generate with AIMD. In comparison, training a model like SNAP/HDNNP with non-accelerated AIMD data usually requires more computational resources and explores orders of magnitude shorter time-scales. FLARE also demonstrates a 13× higher data efficiency for learning the H/Pt heterogeneous system model, compared to the DP-GEN active learning scheme of DeePMD for pure copper [30].

TABLE II: Training set collected by Bayesian active learning; MD temperature $T$; total simulation time $\tau_{\text{sim}}$; total simulation and training wall time $\tau_{\text{wall}}$ (hours); number of CPU nodes $n_{\text{nodes}}$ used; number of atoms in the simulation $N_{\text{atoms}}$; total number of DFT calls made during training $N_{\text{DFT}}$; node time needed to generate an equivalent ab-initio molecular dynamics trajectory without active learning $\tau_{\text{AIMD}}$. "Master" refers to the offline training of a master potential for production that combines all the training data from Bayesian active learning.

* Active-learning workflow started with a pre-trained model.

| System | $T$ (Kelvin) | $\tau_{\text{sim}}$ (ps) | $\tau_{\text{wall}}$ (hours) | $n_{\text{nodes}}$ | $N_{\text{atoms}}$ | $N_{\text{DFT}}$ | $\tau_{\text{AIMD}}$ (years) |
|--------|-------------|-----------------|------------------|-------------|----------------|----------------|------------------|
| H$_2$  | 1500        | 5.0             | 1.2              | 1           | 54             | 24             | 4.1              |
| H$_2$  | 2100        | 10.0            | 3.5              | 2           | 108            | 2              | 8.6              |
| Pt(111) | 300         | 4.0             | 1.2              | 2           | 54             | 4              | 3.5              |
| Pt    | 1500        | 10.0            | 1.7              | 2           | 108            | 6              | 3.4              |
| H/Pt  | 1500        | 3.7             | 61.4             | 2           | 73             | 216            | 75.0             |
| H/Pt$^*$ | 2100       | 26.5            | 143.1            | 3           | 92             | 323            | 44.1             |
| Master | -           | 2.7             | 1                | -           | -              | -              | -                |
| Total  | -           | 59.2            | 214.8            | -           | -              | 575            | 138.7            |

Quantitative uncertainty of each local environment inherent in the the Bayesian force field allows the models to be improved in accuracy and robustness in hierarchical fashion, even using uncertainty of large-scale production runs. As an example, the initial force field [13], trained on bulk Pt at $T=1500K$, Pt(111) at 300K, H$_2$ at 1500K, was deployed in 0.5M atom MD simulations for 24 hours on Longhorn during the Texascale Day at TACC. In the large-scale simulation, suspicious non-physical phenomena were captured by a spike in the uncertainty. Visual inspection revealed Pt desorption from the surface and clustering of H atoms, meaning that the attractive and repulsive regions of the Pt and H potentials were not sufficiently sampled during training with the small system size. Subsequently, two additional training runs with Bayesian active learning were employed, targeting local configurations observed in large-
scale MD runs that had high uncertainty. For improving the repulsive interaction of H-H, we augmented the training set with high-pressure H\textsubscript{2} gas with twice the density and a higher temperature of 2100 K. For the attractive interaction of Pt-Pt, we also included data on H/Pt system at 2100 K, allowing the Pt slab to melt over the course of \~27 ps.

VIII. IMPLICATIONS

Simulating dynamics of complex reactive systems on atomistic level is a grand challenge in computational materials science and chemistry. The FLARE method, when deployed on leadership-class resources, opens possibilities to simulate thermal and reactive dynamics at previously inaccessible length and time scales while maintaining first-principles accuracy. We have demonstrated excellent scaling to hundreds of billions of atoms for a heterogeneous reactive system, decreased total time-to-solution from first principles, and uncertainty quantification enabling more robust and principled MLFF generation and deployment. Importantly, FLARE offers a dramatic step forward in the direct comparison of accurate reactive MD simulations to experimental observations, enabled by the efficient scaling achieved by these MLFFs to billion-atom system sizes. By employing fast prediction and uncertainty quantification, the workflow can be applied to learn models and simulate systems with increasing chemical and structural complexity, including alloys, biological enzymes, polymers, electrode interfaces and, notably, heterogeneous catalysis. In metallic alloys simulations with billions of atoms can help answer long-standing questions about fatigue and grain boundary segregation that govern mechanical properties of structural materials. Computations of ionic and thermal currents can be used to estimate non-equilibrium transport properties in liquid and solid-state batteries and realistic thermoelectric materials, in the presence of interfaces and nontrivial microstructure. Understanding heterogeneous phenomena such as shock waves and crack propagation are also in particular need of large-scale fast MD simulations.

Reactive processes can benefit greatly from FLARE at scale, especially those that have complicated reaction networks, where multiple mechanisms may exist but are not known. Fast MD simulations may be used to reveal these mechanisms, while at the same time learning the relevant energies and forces. In the case demonstrated in this work, catalytic reaction mechanisms of H\textsubscript{2} interacting with the Pt surface at high coverage are complicated and were discovered automatically by the active learning algorithm, which led to the highly accurate prediction of overall reaction rates. Robust determination of the mechanisms could lead to more efficient material design, e.g. new catalysts for methane activation to biofuels and other industrial reactions. In these processes, the structure of the active site is critically important, but may be exhibit dynamic restructuring in response to changing reaction conditions. Catalytic processes like these present excellent opportunities for FLARE as they involve complicated reaction mechanisms over a range of timescales, which can be discovered using active learning, bypassing the need for any \textit{a priori} assumptions or expert chemical intuition. This advantage signals a departure from previous efforts relying on chemical expertise to manually enumerate elementary reaction mechanisms and perform kinetic analyses with prior assumptions of rate-limiting steps. We emphasize that progress in deciphering and simulating reactions necessarily requires both fast and scalable MD and uncertainty quantification needed to efficiently explore the configuration space and obtain reliable large-scale models.

In addition to the scientific impacts, a notable practical advantage of uncertainty-driven active learning is that it represents a step towards full automation, allowing a broader community of scientists to approach chemical or material property predictions from first principles. Users without expert chemical intuition would be able to train a MLFF automatically, and perform uncertainty-aware molecular dynamics. This process is drastically accelerated with the inclusion of active learning in model creation, decreasing the total time-to-solution from inception of the scientific goal to material property prediction from MD at scale. Where previously it took months to develop force fields that were not guaranteed to be accurate, with FLARE the models can take a few days to train without human input and provide internal reliability metrics.

Efficient parallel implementation enabling full GPU utilization for both small and large systems opens the possibility to use MD to study extremely long time scales in smaller systems, as well as to explore extremely large systems, with up trillion-atom scale on leadership-class machines. Due to the adoption of the Kokkos GPU performance portability library, we ensure long-lasting impact, as the optimizations of our models will translate across future GPU architectures adopted for supercomputers, including NVIDIA, Intel and AMD hardware, without requiring low-level code redevelopment.

FLARE is made available as open-source software under the MIT license, which encourages adoption by academic and commercial research and development efforts. Multiple research groups in universities, national labs and industry (representative letters attached) are already adopting the new MLFF models to perform large-scale accurate MD for a variety of real-world applications, including understanding and design of advanced semiconductor and catalytic materials.

Finally, the applications of the fast scalable FLARE models are not limited to molecular dynamics. Combining them with grand-canonical Monte Carlo simulations will allow for rapid explorations of phase diagrams of complex alloys. Meta-dynamics, umbrella sampling and other enhanced sampling techniques based on the new MLFF can be used to estimate free energy surfaces and reaction rates. In the domain of soft and biological materials, coarse grained models can readily be developed using the FLARE active learning and force matching formalism, allowing for even longer time scales to be examined.

IX. ACKNOWLEDGEMENT

We would like to acknowledge helpful discussions with Stan Moore from Sandia Natl. Lab. Computational resources
on Summit were provided by the Oak Ridge Leadership Computing Facility. The authors acknowledge computing resources provided by the Harvard University FAS Division of Science Research Computing Group and by the Texas Advanced Computing Center (TACC) at The University of Texas at Austin.

REFERENCES

[1] M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[2] B. Hammer and J. K. Nørskov, "Theoretical surface science and catalysis—calculations and concepts," Advances in Catalysis, vol. 45, pp. 71–129, 2000.
[3] T. Cheng, H. Xiao, and W. A. Goddard, "Full atomistic reaction mechanism with kinetics for co reduct on cut(100) from ab initio molecular dynamics free-energy calculations at 298 k," Proceedings of the National Academy of Sciences, vol. 114, no. 8, pp. 1795–1800, 2017.
[4] C.-Q. Xu, M.-S. Lee, Y.-G. Wang, D. C. Cantu, J. Li, V.-A. Glezakou et al., "Structural rearrangement of au–pd nanoparticles under reaction conditions: An ab initio molecular dynamics study," ACS Nano, vol. 11, no. 2, pp. 1649–1658, 2017.
[5] S. Hoppe, Y. Li, L. V. Moskaleva, and S. Müller, "How silver segregation stabilizes Iod surface gold oxide: A cluster expansion study combined with ab initio md simulations," Physical Chemistry Chemical Physics, vol. 19, pp. 14845–14853, 2017.
[6] T. P. Sentle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin et al., "The reaxff reactive force-field: Development, applications and future directions," npj Computational Materials, vol. 2, no. 1, p. 15011, 2016.
[7] N. Gerrits, "Accurate simulations of the reaction of h 2 on a curved pt crystal through machine learning," Journal of Physical Chemistry Letters, vol. 12, no. 51, pp. 12157–12164, 2021.
[8] M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[9] B. Hammer and J. K. Nørskov, "Theoretical surface science and catalysis—calculations and concepts," Advances in Catalysis, vol. 45, pp. 71–129, 2000.
[10] M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[11] L. Gai, Y. K. Shin, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[12] T. Tokumasu, M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[13] F. Ahmed, R. Nagumo, R. Miura, S. Ai, H. Tsuboi, N. Hatakeyama et al., "Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with bayesian inference," Physical Review Letters, vol. 122, no. 22, p. 225701, 2019.
[14] E. V. Podryabinkin and A. V. Shapoval, "Active learning of linearly parametrized interatomic potentials," Computational Materials Science, vol. 140, pp. 171–180, 2017.
[15] Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang et al., "Optimizing the kinetic for adatom diffusion on metal surfaces: A quantum chemical study of activation energy barriers," Applied Surface Science, vol. 501, pp. 026302, 2020.
[16] T. Tokumasu, M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[17] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[18] T. Tokumasu, M. L. Personick, M. M. Montemore, E. Kaxiras, R. J. Madix, J. Biener, and C. M. Friend, "Catalyst design for enhanced sustainability through fundamental surface chemistry," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, vol. 374, no. 2061, p. 20150077, 2016.
[19] J. T. Williman, A. S. Williams, K. Nguyen-Cong, A. P. Thompson, M. A. Wood, A. B. Belonoshko et al., "Quantum accurate snap carbon potential for md shock simulations," in AIP Conference Proceedings, vol. 253, no. 1, AIP Publishing LLC, 2020, p. 070055.
[20] Y. Xie, J. Vandermause, L. Sun, A. Cepellotti, and B. Kozinsky, "Baysian force fields from active learning for simulation of inter-dimensional transformation of stannene," npj Computational Materials, vol. 7, no. 1, pp. 1–10, 2021.
[21] J. T. Williman, A. S. Williams, K. Nguyen-Cong, A. P. Thompson, M. A. Wood, A. B. Belonoshko et al., "Quantum accurate snap carbon potential for md shock simulations," in AIP Conference Proceedings, vol. 253, no. 1, AIP Publishing LLC, 2020, p. 070055.
[22] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[23] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[24] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[25] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[26] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[27] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[28] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[29] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.
[30] J. Vandermause, S. B. Fornes, M. Raju, A. C. T. van Duin, and S. Raman, "Atomistic mechanism with kinetics for co reduction on cu(100) from ab initio analysis—calculations and concepts," Advances in Catalysis, vol. 71–129, 2000.