Supplementary Information 2 for: LightPFP: A Lightweight Route to Ab Initio Accuracy at Scale

Wenwen Li,[1, a)] Nontawat Charoenphakdee,[1, b)] Yong-Bin Zhuang,[1] Ryuhei Okuno,[1] Yuta Tsuboi,[1] So Takamoto,[1] Junichi Ishida,[2] and Ju Li[3, 4]

[1)] *Preferred Networks Inc., Tokyo, Japan.*

[2)] *Matlantis Corporation, Tokyo, Japan.*

[3)] *Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

[4)] *Department of Nuclear Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

(Dated: 26 November 2025)

[a)]Electronic mail: wenwenli@preferred.jp

[b)]Electronic mail: nontawat@preferred.jp

CONTENTS

S1. APPLICATION 1: SIMULATION OF INTERFACIAL STRUCTURES OF PT (111)/BENZENE

The main purpose of this example is to demonstrate how to build and validate the LightPFP model to simulate the solid-liquid interface using the Pt/benzene interface as a model system. By leveraging LightPFP's enhanced speed and accuracy, researchers can gain insights into solid-liquid interface behavior and expand its applications in materials science, chemistry, and nanotechnology.

S1.1. Student fine-tuning

To train the LightPFP model, we compiled a diverse set of structures with their energies, forces, and stresses labeled using the PFP. The training dataset encompassed various components: bulk Pt, Pt (111) slab, bulk benzene, and Pt (111)/benzene interface structures.

System	Methods	Num of structures	Num of Atoms
Pt	Cell compression/Stretching/Deforming Atom Displacement Vacancy Surface and MD (NPT 500–1500 K)	2249	111,290
Pt (111)	MD (NPT 500–1000 K)	640	43,680
Benzene	Cell compression/Stretching MD (NPT 300–400 K; NVT 500–1500 K)	984	177,120
Interface	MD (NPT 300–800 K; NVT 500–1750 K)	5400	2,632,800

The dataset was randomly split into training and testing datasets, with 90% comprising the training dataset and the remaining 10% constituting the testing dataset.

S1.2. Evaluation using PFP

To validate the performance of the LightPFP models, we performed MD simulations on a small Pt (111)/benzene interface system and compared the trajectories with PFP.

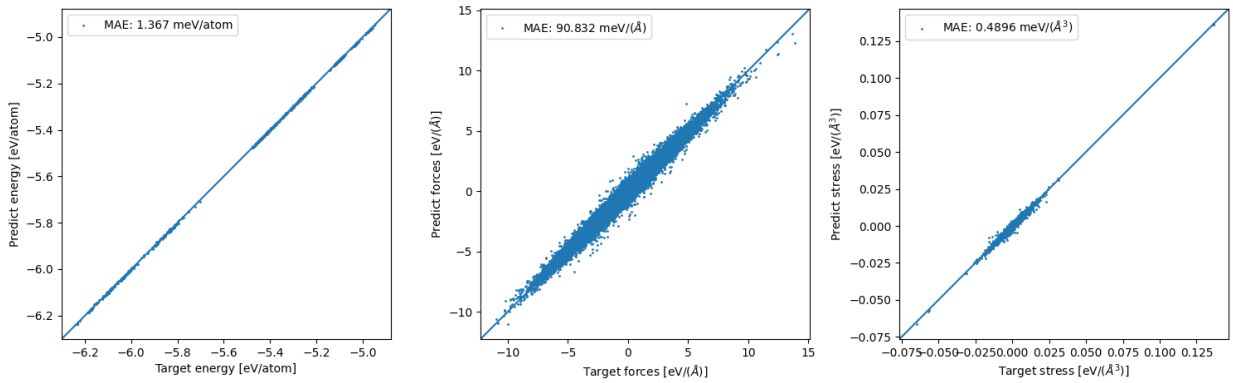

FIG. S1: Parity plots for energy, forces, and stress on the test set: PFP ground-truth values (x-axis) vs LightPFP model predictions (y-axis).

The initial simulation box has dimensions of $38.58\,\text{Å} \times 33.10\,\text{Å} \times 101.32\,\text{Å}$, with cell angles of 90°, 90°, and 120°. It consists of a total of 9,072 atoms, including 1,512 Pt atoms and 630 benzene molecules. We have chosen a relatively smaller structure in order to compare the results with PFP.

The initial structure undergoes a 20 ps equilibrium at a temperature of 300.0 K at first, using the NVT ensemble. Subsequently, NPT MD simulations are conducted for 100 ps at 300.0 K and 1 bar, utilizing the NPT ensemble. MD snapshots are saved for future analysis.

The density estimated from LightPFP NPT MD (5.08 g/cm^3) agrees well with that from PFP (5.06 g/cm^3). To calculate the radial distribution function, we used the snapshots of the MD trajectory taken after 50 ps. The resulting radial distribution function is shown in Fig S2. We discovered that the results obtained from both the LightPFP and PFP trajectories exhibit a high degree of agreement.

To characterize the atomic distribution along the z axis (normal to the interface), we sampled MD trajectory snapshots after 50 ps and generated a z-position density via Gaussian broadening (width 0.25 Å). The resulting distribution is plotted in Fig. S3.

The density profile in Fig S3 shows several peaks in the H and C elements near the Pt surface (around 20 Å), indicating that the benzene structure is significantly different from the uniform liquid phase due to adsorption with the Pt surface. As we move further away from the Pt surface, the interfacial liquid structure gradually transitions towards a uniform liquid phase. According to the figure, the thickness of the interfacial layer is approximately 15 Å.

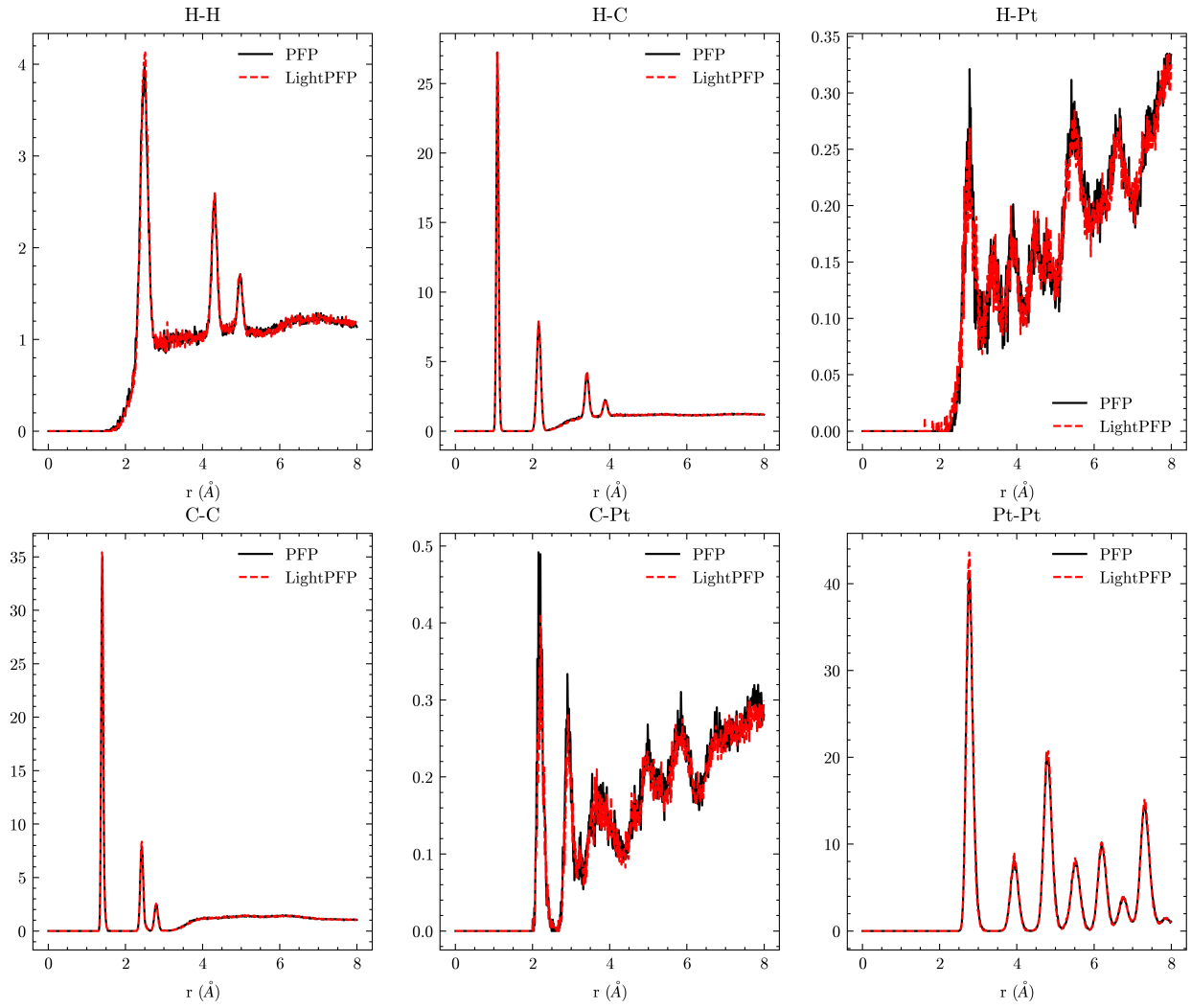

FIG. S2: Comparison of radial distribution functions from LightPFP and PFP trajectories

The position and intensity of the peaks in the LightPFP model were compared with the PFP result, and they matched each other across most regions. This alignment demonstrates that the LightPFP model effectively captures and represents the structural characteristics of the solid/liquid interface.

Symmetric peaks can be observed around $z = 100\,\text{Å}$, indicating the presence of the same liquid-solid interface due to the periodic boundary condition. However, these peaks exhibit slight fluctuations in shape due to noise originating from cell shape changes in the NPT MD simulation.

In Fig S4, we can observe the benzene molecules that have adsorbed onto the Pt surface. Specifically, in the LightPFP MD simulation, 18 benzene molecules were found to be

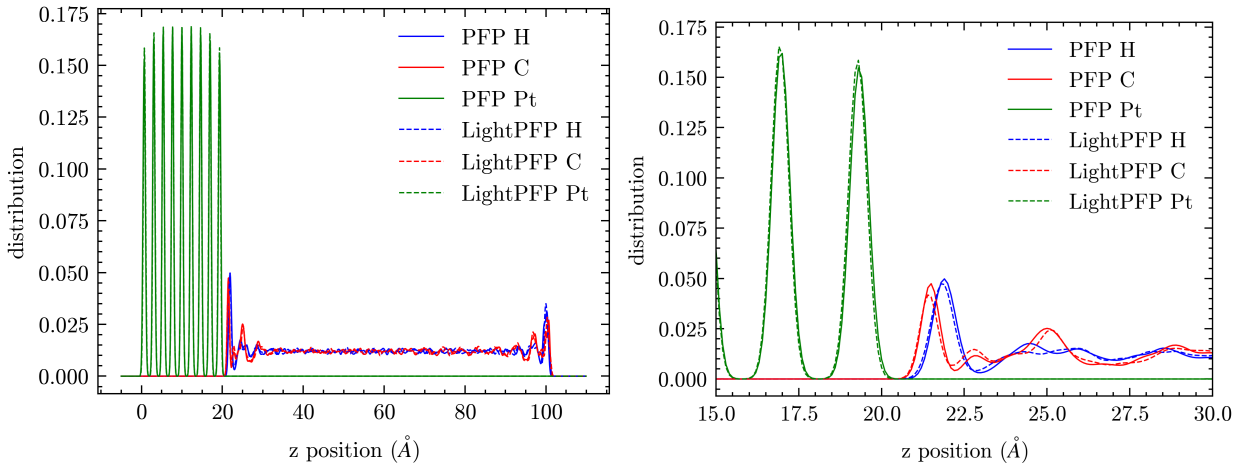

FIG. S3: Spatial distribution of Pt, C and H atoms along the z direction. (left) Whole simulation box. (right) Zoomed-in view at the interface.

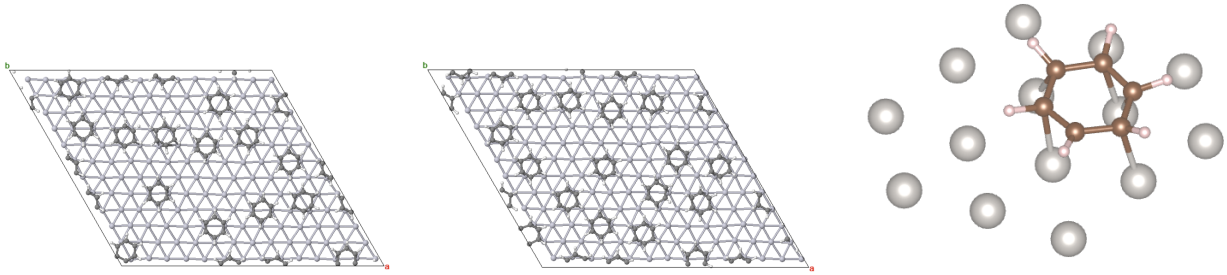

FIG. S4: The benzene adsorption on Pt surface (left) LightPFP (middle) PFP (right) The bri30 adsorption site of the benzene molecule.

adsorbed onto the Pt surface within the specified area. In the PFP MD simulation, on the other hand, there were a total of 21 benzene molecules observed to be adsorbed onto the Pt surface within the same area. In conclusion, the LightPFP MD simulation reproduced the surface coverage rate of benzene on the Pt surface well.

Fig. S4(right) illustrates the adsorption structure of a single benzene molecule. The bri30 conformation, in which the center of the benzene molecule is located on the bridge site of the Pt surface, was found to be the most energetically stable in first-principles calculations[1]. Interestingly, we observed that almost all the molecules were adsorbed in the bri30 site in both the LightPFP and PFP simulations. This result is consistent with the findings of the first-principle calculations.

S2. APPLICATION 2: MISCIBILITY OF WATER, BENZENE AND HEPTANE

In this example, we use LightPFP to investigate the miscibility of binary liquid mixtures among water, benzene and heptane via large-scale molecular dynamics simulations using LightPFP.

S2.1. Student fine-tuning

We trained LightPFP on a collection of datasets designed to cover both homogeneous and demixed liquid environments. The initial dataset comprised nine classes of configurations: (1) bulk water, (2) bulk benzene, (3) bulk heptane, (4) homogeneous water/benzene mixtures, (5) homogeneous water/heptane mixtures, (6) homogeneous heptane/benzene mixtures, and (7) explicit liquid–liquid interfaces for water/benzene, (8) water/heptane, and (9) heptane/benzene. To efficiently obtain interfacial training data, we constructed liquid–liquid interface geometries and sampled them by short MD, rather than relying on spontaneous demixing from homogeneous starting states. The latter would require prohibitively long trajectories to capture phase-separated configurations because phase separation proceeds via slow nucleation, coarsening, and domain growth. By seeding and sampling interfacial structures, the training set explicitly exposed the model to the distinct local environments present at liquid–liquid boundaries.

Starting from this initial model, we performed active learning to further improve accuracy and robustness. For each of the nine system types, we ran 20 ps MD simulations over 280-350 K, monitored model performance, and selectively augmented the training set with configurations with high disagreement w.r.t. PFP. The resulting LightPFP model was then used for the large-scale MD simulation.

S2.2. Large-scale MD simulation

Three systems are considered: (1) 17,280 water molecules with 4,320 benzene molecules; (2) 17,280 water molecules with 2,160 heptane molecules; and (3) 4,320 benzene molecules with 2,160 heptane molecules. A 1 ns molecular dynamics simulation was performed for each system. According to experimental data, water and heptane, water and benzene are immis-

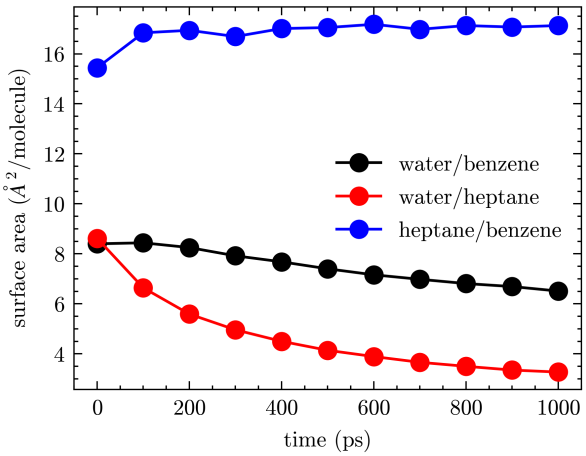

FIG. S5: Surface area of liquid-liquid interface.

cible at room temperature, while heptane and benzene are miscible[2] For the water/benzene and water/heptane mixtures, spontaneous liquid–liquid phase separation is clearly observed in the MD trajectories, with the two immiscible components forming distinct phases. In contrast, no phase separation is observed for the benzene/heptane mixture, as these liquids are mutually miscible. The MD simulation results are all consistent with experimental observations.

To quantify demixing, we analyzed the MD snapshots and computed the liquid–liquid interfacial area using OVITO's construct surface mesh modifier with the alpha-shape method[3]. For each saved frame, the two species were identified, a triangulated interface was generated, and the total interfacial area was recorded. The time evolution of this area is plotted in Fig. S5. For the water/benzene and water/heptane mixtures, the interfacial area decreases rapidly and then approaches a low, nearly steady value, indicating fast coarsening and macroscopic phase separation. In contrast, for the benzene/heptane mixture, the interfacial area remains essentially unchanged from the outset over the entire 1 ns window, consistent with the absence of demixing. In addition, the experimental data show that the solubility of heptane in water is lower than that of benzene. In MD simulations, we also found that the interface area decreases more rapidly in the heptane-water system, reflecting faster nucleation and phase transition dynamics[2,4]. Figures S6, S7, and S8 show representative interface morphologies at different times, providing a visual corroboration of these trends: the liquid–liquid interface in water/benzene and water/heptane smooths and recedes over time, whereas no well-defined interface emerges in benzene/heptane.

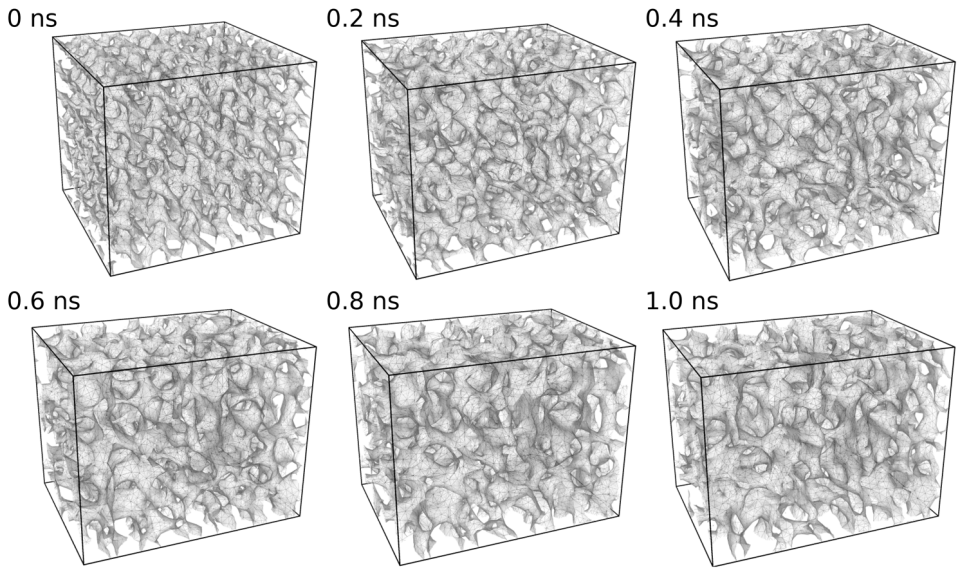

FIG. S6: Liquid-liquid interface between water and benzene

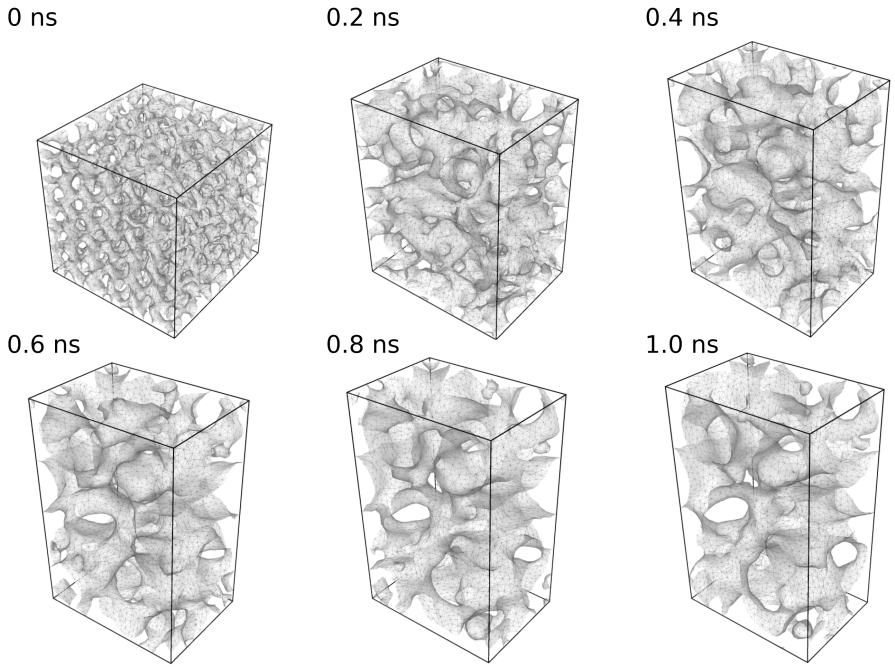

FIG. S7: Liquid-liquid interface between water and heptane

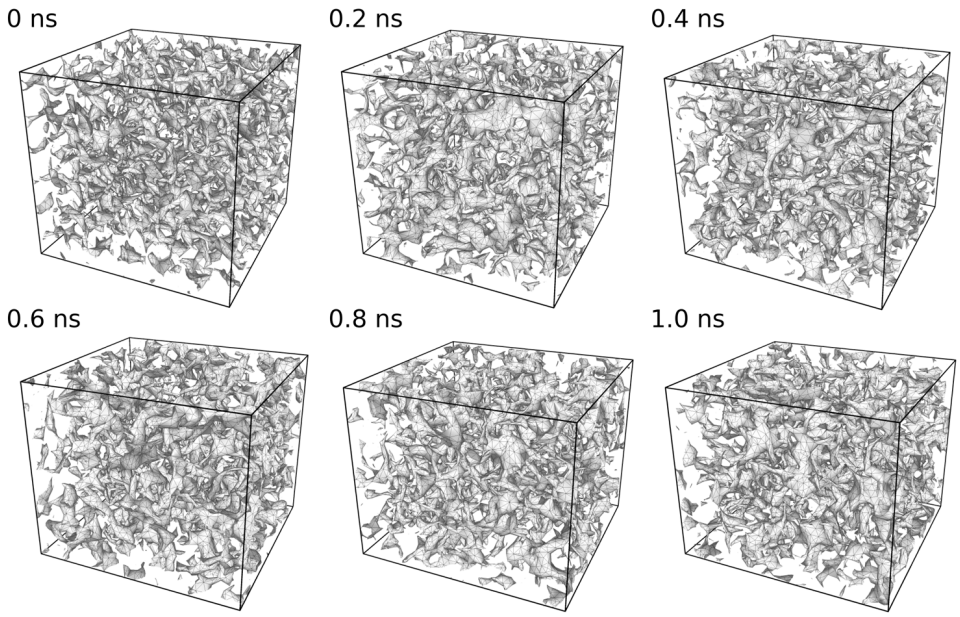

FIG. S8: Liquid-liquid interface between heptane and benzene

S3. APPLICATION 3: INTERFACE THERMAL RESISTANCE BETWEEN NI AND DPO/BP

In this example, we use LightPFP to quantify interfacial thermal transport between a Ni (111) surface and the diphenyl oxide (DPO)/biphenyl (BP) eutectic heat-transfer fluid, a widely used medium in parabolic trough concentrated solar power systems with a maximum operating temperature near 400°C. Our target property is the thermal conductivity across the metal–fluid interface, which governs heat exchange efficiency in receiver tubes and heat exchangers[5].

S3.1. Student fine-tuning

We trained LightPFP on a dataset tailored to capture both bulk and interfacial environments relevant to the Ni–DPO/BP system. The initial configurations encompassed four classes: bulk Ni, a Ni (111) slab exposing the surface, bulk liquid DPO/BP, and explicit Ni (111) and DPO/BP interfaces. Each class was sampled via short molecular dynamics and light "rattle" perturbations to diversify local environments. MD sampling covered NVT simulations at 500 K, 1000 K, and 1500 K and NPT simulations at 300 K, 400 K, and 500 K. The high-temperature NVT trajectories were chosen to generate randomized, higher-energy configurations that improve the robustness and stability of the trained model; in contrast, the NPT sampling was restricted to lower temperatures to avoid unphysical density fluctuations and cell distortions that can arise at very high temperatures under barostat control. For bulk liquid DPO/BP and for Ni (111)-DPO/BP interfaces, we also manually created multiple distinct atomic configurations prior to short MD "structure sampling." This manual seeding was used to enrich the distribution of molecular orientations in the training set, because orientational rearrangements in the liquid and at the metal–liquid interface relax slowly in MD, especially in the 300–500 K range.

Starting from this base model, we performed active learning to improve accuracy and robustness over the temperature range of interest. For each configuration class, we ran short MD trajectories spanning 300–500 K under complementary conditions: 5 ps NVT, 10 ps NPT, and 30 ps reverse non-equilibrium MD (rNEMD) to explicitly sample heat-flow states. We also varied setup details (for example, MD temperature, slab and liquid layer

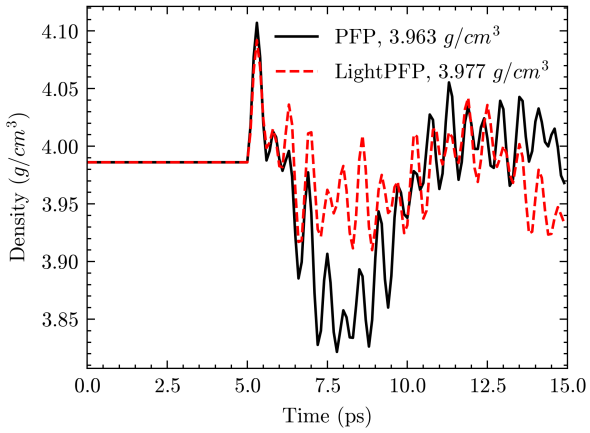

FIG. S9: Density of Ni and DPO/BP interface structure at 300 K

thickness and the rNEMD swapping frequency, etc.) to enrich coverage of interfacial motifs. Configurations exhibiting high model disagreement were iteratively labeled and added to the training set, and the query–retrain loop was repeated until validation metrics stabilized.

S3.2. Evaluation using PFP

We assessed the resulting LightPFP model against a reference PFP potential under matched conditions. The model reproduces key equilibrium and non-equilibrium observables: system densities (Fig S9) and radial distribution functions (Fig. S10) for interfacial systems, as well as steady-state temperature profiles obtained from reverse non-equilibrium MD across Ni (111)-DPO/BP interfaces (Fig. S11). The close agreement with PFP across these metrics supports the use of LightPFP for the interfacial thermal conductivity calculations reported below.

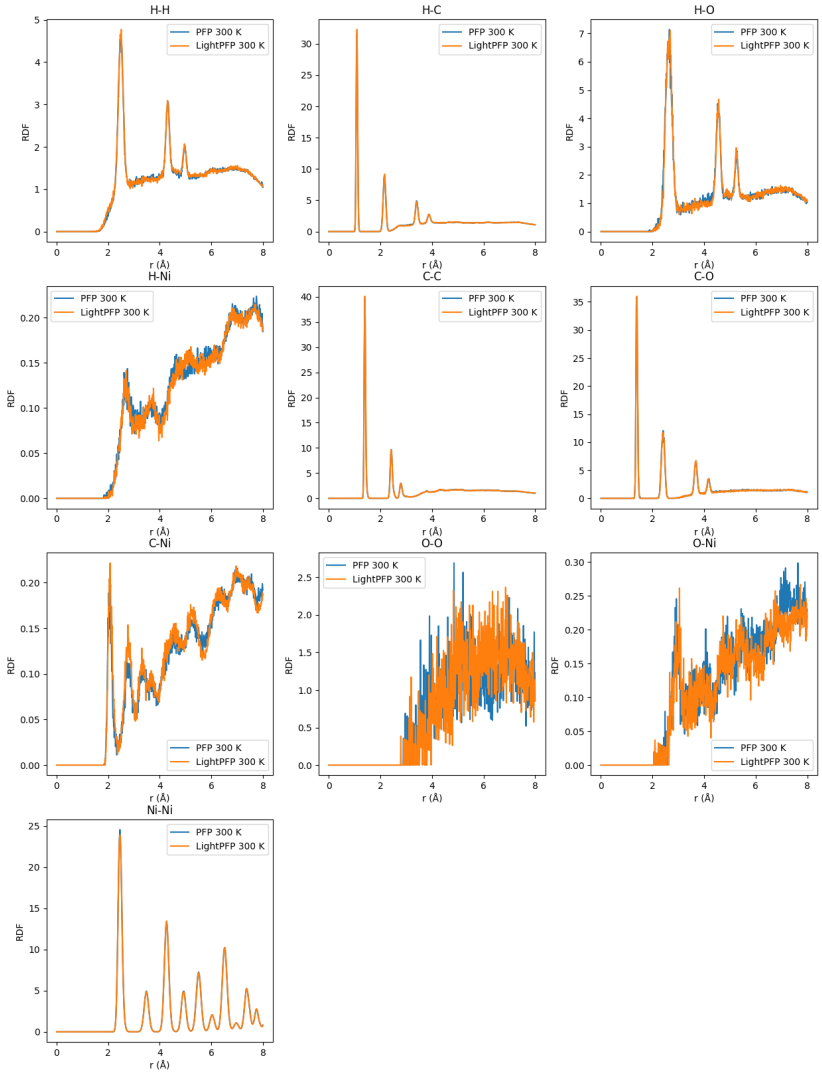

FIG. S10: Radial distribution function of Ni and DPO/BP interface structure at 300 K

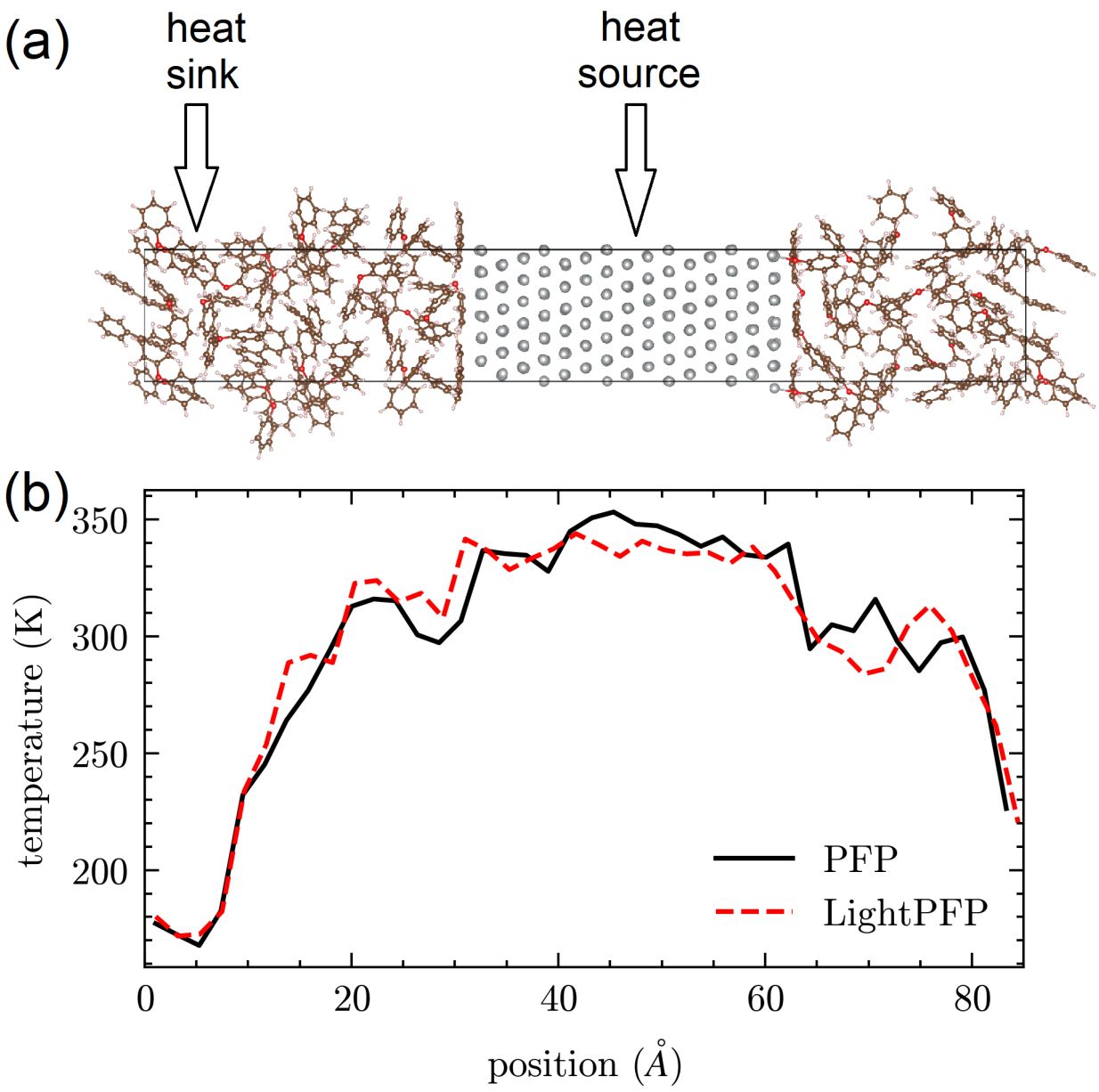

FIG. S11: Temperature profile across Ni and DPO/BP interface

S4. APPLICATION 4: VISCOSITY OF N-DECANE

In this study, we target the shear viscosity of liquid n-decane via reverse non-equilibrium molecular dynamics (rNEMD) using the Müller–Plathe momentum-exchange scheme[6]. To be more specific, a liquid is equilibrated at the target thermodynamic state and the simulation box is partitioned into slabs along the gradient direction, with two slabs designated as momentum source and sink. At fixed intervals, the particle with the largest positive flow-direction velocity in the source slab and the particle with the most negative component in the sink slab exchange their flow-direction velocities, imposing a constant momentum flux. The spatially resolved velocity profile is accumulated and time-averaged; its linear region away from the exchange slabs provides the velocity gradient. The imposed flux is computed from the cumulative exchanged momentum divided by cross-sectional area and simulation time (accounting for periodic shear planes). The shear viscosity is then $\eta = \frac{J}{\mathrm{d}v/\mathrm{d}z}$, where J is cumulative exchanged momentum, v is the velocity of atoms and z is atomic position along z axis.

Important practical considerations accompany rNEMD. Because accessible simulation times are limited, rNEMD often relies on velocity gradients far larger than those used in experiments, which can alter molecular orientation (e.g., induce partial alignment along the flow) and thereby modify the intrinsic shear response. To reduce the artificially imposed gradient, one can lower the exchange frequency or elongate the simulation box along the gradient direction. Lowering the exchange frequency weakens the driving but requires longer MD sampling to resolve the slope of the velocity profile accurately, particularly for high-viscosity liquids. Increasing the box length similarly reduces the gradient at a given momentum flux but raises computational cost. This makes uMLIP require longer time to compute viscosities for many molecules, whereas the LightPFP method becomes more useful due to its better computational efficiency. In this example, we selected n-decane as the case study and trained its LightPFP model. We computed the viscosity of n-decane at high temperatures under different computational conditions using both PFP and LightPFP as a validation of LightPFP. Because viscosity is lower at high temperature, PFP also yields good results. We then used LightPFP to compute the viscosity of n-decane at room temperature, ultimately obtaining values that closely match experiment.

S4.1. Student fine-tuning

We trained LightPFP for n-decane with a dataset focused on bulk liquid environments. Several initial liquid structures were generated with realistic densities to cover different molecule orientations. The initial dataset combined short MD segments and "rattle" perturbations to diversify local configurations: NVT trajectories at 500 K, 1000 K, and 1500 K; NPT trajectories at 300 K, 400 K, 500 K, and 600 K; and rattle displacements are sampled from normal distribution with standard deviation of 0.10 Å and 0.15 Å. As in our other applications, high-temperature NVT sampling was used to generate randomized, higher-energy configurations that improve robustness, whereas NPT sampling was focused at moderate temperatures to avoid unphysical density excursions and cell distortions that can occur under aggressive barostatting at very high temperatures.

Starting from this base model, we executed an active learning loop tailored to the viscosity task. For bulk liquid n-decane, we ran short trajectories under complementary ensembles and non-equilibrium driving: 5 ps NVT, 10–20 ps NPT, and 20–30 ps rNEMD. rNEMD simulations followed the slab-based momentum-exchange approach, with systematic variation of setup details (swap interval, swap slab thickness, and cell aspect ratio) to probe sensitivity and enrich configurational coverage. Configurations exhibiting large model disagreement were labeled and added to the training set; the query–retrain cycle was repeated until validation metrics stabilized.

S4.2. Evaluation using PFP

After obtaining the LightPFP model, we ran the rNEMD calculation using PFP and LightPFP. The calculation method is as follows: First, the size of initial simulation box is 17.3x17.3x51.9 angstrom contains 48 n-decane molecules. Then, the MD is performed in thee NVT ensemble for 5 ps at 400, 450 and 500 K, and then, NPT ensemble for 20 ps at the same temperature, 1 bar to achieve equilibrium status. After that, the rNEMD is performed, the simulation box is divided into 20 slabs. For each temperature, the momenta exchange is performed for each 100 or 500 fs. The rNEMD is performed for 50 ps to achieve a stable momenta profile across the simulation box.

The accuracy of LightPFP is evaluated by comparing several properties with PFP results.

TABLE S1: Density of n-decane at different temperatures

Temperature (K)	PFP (g/cm^3)	LightPFP (g/cm^3)	Error (g/cm^3)
400	0.689	0.675	0.014
450	0.647	0.650	0.003
500	0.611	0.617	0.006

As shown in table S1 at 400, 450 and 500 K is obtained from the NPT-MD part of trajectory. The predicted liquid densities agree well with PFP with an average absolute deviation of 0.0076 g/cm^3. The radial distribution functions for *n*-decane also obtained from the MD frames took from the NPT part. The results are plotted in Fig S12 showing a close agreement with PFP over the same temperature range.

At last, we compared the viscosity results calculated by LightPFP and PFP (Fig. S13(a)). First, at 450 K, both PFP and LightPFP yield the same viscosity of 0.2 mPa·s for exchange intervals of 100 fs and 500 fs. All results are consistent with experimental data[7] at 446.75 K. At 400 K, viscosities computed with a 500 fs momentum-exchange interval were higher than those with a 100 fs interval for both PFP and LightPFP. Under both intervals, PFP and LightPFP were in close agreement. The MD-derived viscosities spanned 0.23–0.51 mPa·s across the tested conditions, and the experimental value of 0.29 mPa·s falls within this range.

S4.3. Large-scale MD simulation

After confirming of the LightPFP model by comparing the high-temperatures viscosity with PFP and experimental results, we tried to run large-scale MD simulation to get more accuract viscosity at low temperature. The simulation box size is 34.6 x 34.6 x 103.8 Angstrom, which contains 384 *n*-decane molecules. We run the MD simulation at 300, 350 and 400 K. The MD protocol is same as the above one, except we prolong the rNEMD simulation to 1 ns, to get accurate statistic results. The results is shown in Fig. S13(b). The viscosity is estimated to be 0.85 mPa s at 300 K which is very close to the experiment value, 0.83 mPa s.

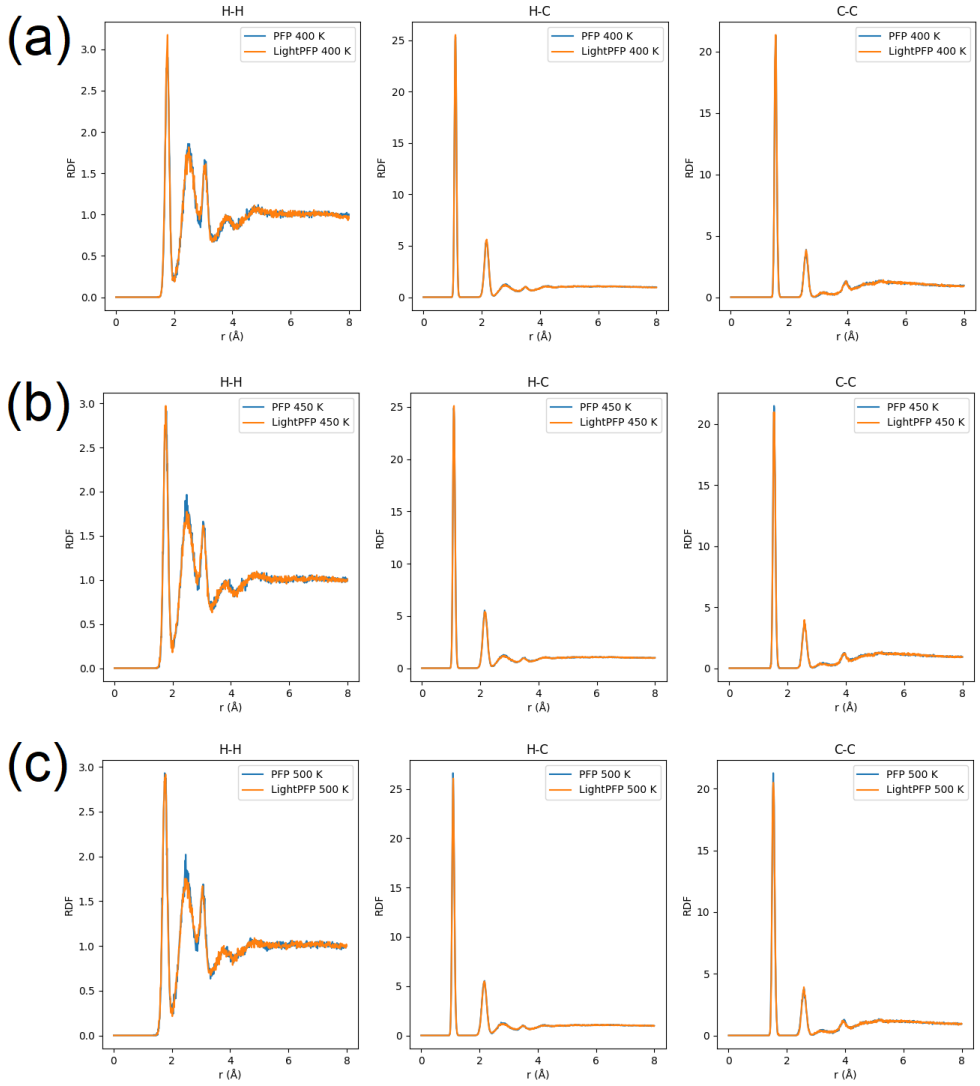

FIG. S12: Radial distribution function of n-decane. (a) 400 K (b) 450 K and (c) 500 K

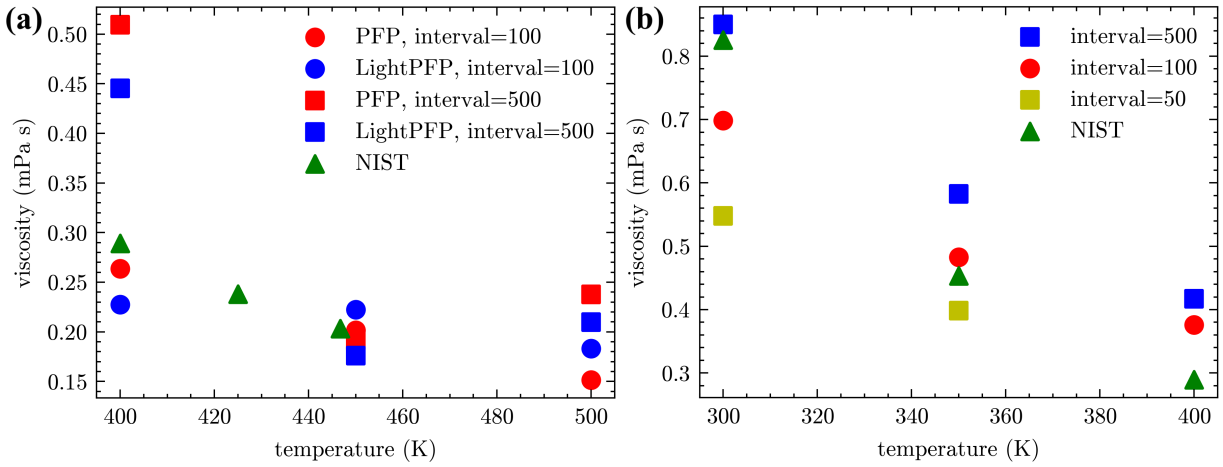

FIG. S13: Viscosity of n-decane

S5. APPLICATION 5: CRACK PROPAGATION IN GRAPHENE NANORIBBON

In this example, we assess the capability of LightPFP to describe fracture propagation in graphene nanoribbon via molecular dynamics[8].

S5.1. Student fine-tuning

The training data were constructed to span two-dimensional carbon environments under mechanical loading, from pristine structures to defect-containing systems. The initial dataset included defect-free graphene sheets and graphene nanoribbons (GNRs) with armchair (AC) and zigzag (ZZ) edges, as well as structures with a triangular hole that served as crack initiators. To sample elastic responses broadly, we (a) performed NVT MD at 500 K, 1000 K, and 1500 K; (b) applied small homogeneous deformations: ±5% strain to both diagonal (uniaxial/biaxial) and off-diagonal (simple shear) components of the simulation cell. To further diversify local environments, atomic positions were rattled with Gaussian noise of 0.1 Å standard deviation. No cracks were included in this initial dataset; thus, the base model learned from pristine bonding environments across temperatures, strain states, and edge types without explicit exposure to fracture.

Starting from this model, we carried out active learning to improve robustness in high-strain regimes. Using the initial structures (graphene sheets and AC/ZZ GNRs), we ran strain-controlled MD with a "deform extension" protocol that incrementally altered the cell shape to increase strain every fixed number of MD steps. This procedure drives the systems into strongly non-linear regimes where bond stretching, bond angle distortions, and incipient bond breaking occur. During these runs, we monitored model performance and selectively augmented the training set with configurations exhibiting large errors or anomalous forces/energies under increasing deformation. In addition, we included strained configurations of ribbons with pre-introduced triangular holes to expose the model to local stress concentration fields characteristic of crack tips and to the chemistry of bond scission in sp2 carbon. The resulting LightPFP model was then used for fracture simulations.

S5.2. Evaluation using PFP

For evaluation, we simulate crack initiation and growth in AC and ZZ GNRs at 300 K by introducing a triangular hole on one ribbon edge to serve as a notch. Uniaxial tension was applied along the ribbon axis (x direction), with the Green-Lagrange strain increased at a rate of 10^{-5} per femtosecond. Each trajectory was propagated for 50 ps, reaching a total strain of 0.5. Identical protocols were run with both the original PFP potential and LightPFP for direct comparison. The MD snapshot during the crack propagation process is shown in Fig. S14.

Both PFP and Light predict crack initiation at about 6 ps (strain is 0.06) in AC GNR. For ZZ GNR, the PFP shows crack initiation at about 6 ps while LightPFP is a little bit slower. LightPFP reproduces the spatial pattern of bond breaking and the subsequent crack path observed with PFP: bonds fail first near the notch where stress concentrates, and the crack advances into the ribbon width under continued loading. The straight crack is reproduced by LightPFP in both AC and ZZ GNR. The predicted morphology and sequence of fracture events are consistent across edge types, with no spurious branching or unphysical healing observed in LightPFP.

The stress–strain responses computed from the virial stress is plotted in Fig. S15. The curve shows close agreement between PFP and LightPFP over the entire loading history, including the elastic state where stress and strain change linearly, the crack propagation state where stress drops rapidly, and the state where stress is 0 after fracture.

Taken together, these results demonstrate that LightPFP, trained without explicit cracks and refined via strain-driven active learning, accurately captures the initiation and propagation of fractures in graphene nanoribbons. The agreement in crack onset time, crack morphology, and stress–strain behavior indicates that LightPFP attains PFP-level fidelity for fracture simulations while retaining its computational efficiency.

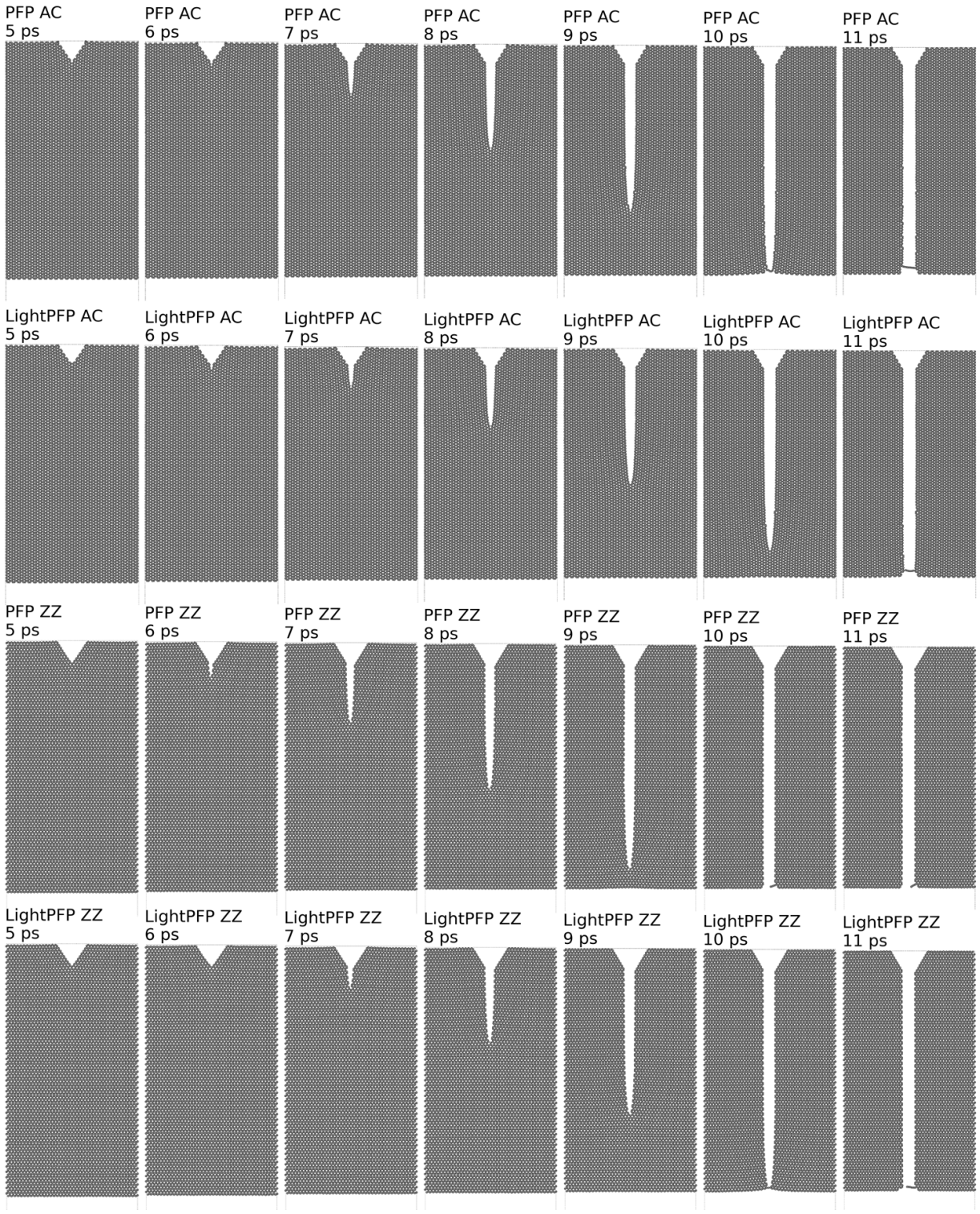

FIG. S14: Illustration of crack propagation in graphene nanoribbons. Each column displays molecular dynamics snapshots at 5, 6, 7, 8, 9, 10, and 11 ps, corresponding to strains of 0.05, 0.06, 0.07, 0.08, 0.09, 0.10, and 0.11, respectively. From top to bottom, the four rows show PFP MD of AC GNR, LightPFP MD of AC GNR, PFP MD of ZZ GNR, and LightPFP MD of ZZ GNR.

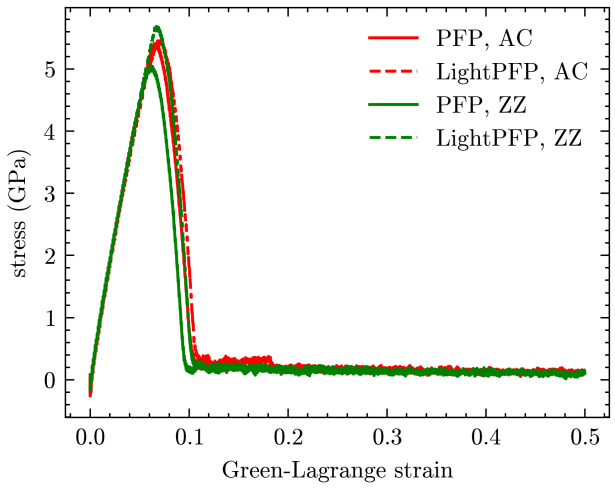

FIG. S15: Strain stress curve of AC and ZZ GNR in crack propagation process

S6. APPLICATION 6: FRICTION OF FE_2O_3 SURFACE WITH LUBRICANT AND FATTY ACID SURFACTANT

In this example, we use LightPFP to investigate lubrication and shear responses at iron oxide–organic interfaces by simulating friction between two Fe_2O_3 slabs separated by a multilayer film of fatty-acid surfactants and a squalane lubricant[9]. Specifically, the simulation box comprises a five-layer stack: an Fe_2O_3 slab, a monolayer of stearic acid or oleic acid, a squalane layer, a second monolayer of stearic or oleic acid, and a second Fe_2O_3 slab. Figure S16 shows the atomistic structure of this system. This configuration results in a complex multilayer structure, making the collection of training data more challenging.

S6.1. Student fine-tuning

We trained LightPFP on a curated dataset comprising (1) bulk Fe_2O_3 crystals, (2) Fe_2O_3 slabs, (3) bulk liquids of stearic acid, oleic acid, and squalane, (4) solid-liquid interfaces between Fe_2O_3 and each of stearic acid, oleic acid, and squalane, and (5) liquid-liquid interfaces between squalane and stearic acid, and between squalane and oleic acid. We generated diverse initial structures within each category and used them to sample training configurations. Multiple distinct initial configurations were employed to enhance the diversity of the training set.

To construct the initial dataset, we combined molecular-dynamics sampling with rattle perturbations. We performed MD in both NVT and NPT ensembles. NVT sampling at elevated temperatures (500, 1000, and 1500 K) broadened the range of conformations and interfacial arrangements, whereas NPT sampling at 300, 400, and 500 K targeted thermodynamic states relevant to friction simulations and yielded realistic organic-layer densities. To improve robustness to rare distortions, we applied Gaussian-distributed displacements (standard deviations 0.10 and 0.15 Å) to the initial structures, generating physically plausible yet diverse local environments.

Because friction can induce large molecular deformations and uncommon contact geometries that are rare under equilibrium sampling, we augmented the model via active learning focused on shear. In each iteration, we (i) fixed the bottom part of the system, (ii) imposed a controlled lateral displacement on the top part to generate steady sliding, and (iii) evolved

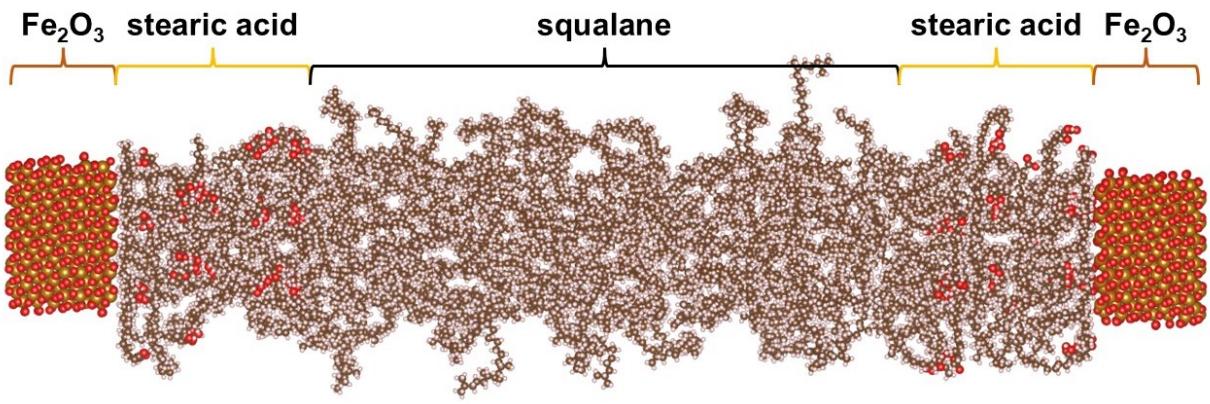

FIG. S16: An atomistic structure of Fe_2O_3 - lubricant - surfactant system

the remaining atoms with NVT dynamics. Configurations exhibiting large energy/force discrepancies were added to the training set, and the model was retrained. We conducted 10 active-learning rounds: iterations 1–5 used smaller interface cells containing both iron oxide–molecule solid–liquid interfaces and surfactant–lubricant liquid–liquid interfaces to rapidly accumulate diverse contact motifs, whereas iterations 6–10 employed larger cells to capture cooperative rearrangements under shear.

S6.2. Evaluation using PFP

The final LightPFP model was used to simulate the full five-layer stack: Fe_2O_3 slab – stearic acid – squalane film – stearic acid – Fe_2O_3 slab. We validated the model by performing friction MD with both PFP and LightPFP. The initial system contained 4,604 atoms. Sliding simulations were run for 300 ps with a top-slab velocity of 30 m/s, while the bottom slab was fixed. The system temperature was maintained at 300 K using a thermostat. The same sliding protocol was applied with PFP, and the resulting trajectories and final structures were compared.

Figure S17 shows the displacement of molecules over 300 ps of sliding. As the upper Fe_2O_3 slab moves, nearby molecules are entrained and translate in the sliding direction. Both PFP and LightPFP predict that molecules adjacent to the moving Fe_2O_3 layer travel by approximately 80 Å due to interfacial friction, with displacements that decay with depth into the surfactant and lubricant layers. Figure S18 depicts the morphology of the moving Fe_2O_3 slab and adjacent molecules in the final frame. Interfacial shear substantially stretches

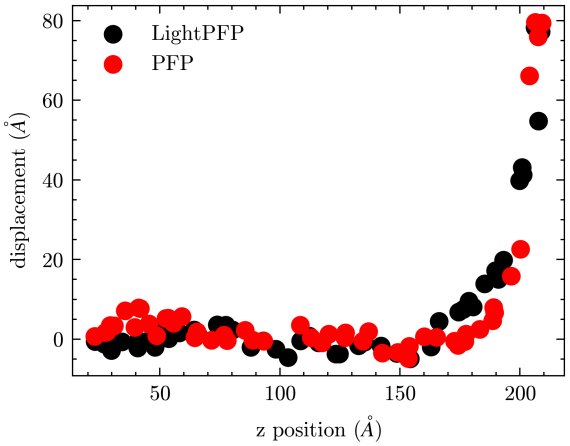

FIG. S17: Molecular displacements during 300 ps of sliding for the Fe2O3 / fatty acid / squalane / fatty acid / Fe2O3 stack.

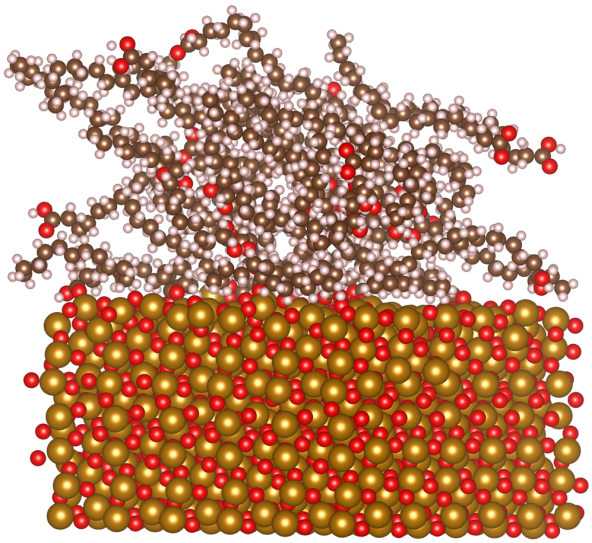

FIG. S18: Final-frame morphology after 300 ps of sliding.

the molecular chains and aligns them along the sliding direction. These results support the ability of LightPFP to capture coupled solid–organic interfacial mechanics and shear-induced ordering relevant to boundary lubrication.

S7. APPLICATION 7: DIFFUSION BEHAVIOR IN POLYMER IONIC LIQUID

In this example, we use LightPFP to investigate anion diffusion in a neat polymer ionic liquid (PIL), taking poly (ethyl vinyl imidazolium) paired with PF_6^- as a representative system[10]. Poly (ethyl vinyl imidazolium) forms a positively charged polymer network, while discrete PF_6^- anions occupy the interstitial regions of the polymer matrix, collectively giving rise to the characteristic ion-conducting behavior of PILs.

S7.1. Student fine-tuning

We curated a training set to expose LightPFP to the relevant local environments of the target PIL across a broad range of densities, temperatures, and chain conformations. To efficiently cover conformational diversity, we generated multiple initial structures rather than relying on a single long trajectory: starting from distinct packings accelerates exploration of chain orientations, local packing motifs, and ion coordination states. Structural assembly was performed with a dedicated routine that constructs mixed-oligomer boxes comprising monomers, dimers, trimers, 5-mers, and 7-mers of poly (ethyl vinyl imidazolium), together with the stoichiometric number of PF_6^- anions to satisfy charge neutrality. Molecules were placed with randomized positions and orientations to maximize initial configurational diversity.

Initial data generation combined short molecular dynamics (MD) sampling and stochastic perturbations. MD trajectories were run in both NVT and NPT ensembles. For NVT, we used elevated temperatures (500 K, 1000 K, 1500 K) to accelerate configurational decorrelation and broaden the coverage of local structures. For NPT, we sampled temperatures (400 K, 500 K, 600 K) representative of the intended application window to capture realistic densities and coordination statistics. In addition, a rattle procedure applied Gaussian displacements (standard deviations of 0.10 and 0.15 Å) to further diversify local atomic environments and improve robustness. The initial LightPFP model is trained from this dataset.

Starting from this model, we carried out active learning to enhance accuracy and stability under production conditions. Each active-learning iteration consisted of multiple short MD

TABLE S2: Density of polymer ionic liquid

Temperature (K)	PFP (g/cm^3)	LightPFP (g/cm^3)
300 K	1.394	1.389
400 K	1.374	1.372
500 K	1.351	1.356

jobs, each seeded from a newly generated structure via the above mentioned procedure. For each job, we randomly selected a temperature between 300 and 700 K, ran a 5 ps NVT trajectory followed by a 50 ps NPT trajectory, and attempted data collection every 100 MD steps. Configurations identified as poorly described by the current model were labeled and added to the training set, after which the model was retrained.

S7.2. Evaluation using PFP

To validate LightPFP for the target PIL, we performed MD comparisons against PFP. First, we generated reference PFP trajectories at 400, 500, and 600 K with NPT ensemble. Using identical initial configurations, ensemble settings, and integrator parameters, we then repeated the simulations with LightPFP and saved all trajectories for post hoc analysis.

We computed equilibrium densities from the last 10 ps NPT trajectories. The densities at 300 K, 400 K, and 500 K listed in Table S2. LightPFP densities closely track PFP across this range. The radial distribution functions (RDFs), which were accumulated over the last 10 ps of each trajectory at 300 K, 400 K, and 500 K, are shown in Fig. S19. LightPFP reproduces the positions and heights of the PFP peak, indicating consistent local coordination and packing.

To probe anion mobility, we monitored phosphorus atoms (proxies for PF_6^-) and computed mean squared displacement (MSD), as shown in Fig. S20. MSD curves from LightPFP agree well with those from PFP, indicating consistent diffusive behavior in the polymer matrix. In addition, the diffusion coefficients D(T) were extracted from the MSD and fitted to an Arrhenius form to estimate the activation energy for PF_6^- transport. The activation energy from LightPFP is 10.46 kJ/mol, in close agreement with the PFP value of 10.82 kJ/mol. We note that the limited number of temperatures constrains the precision of these

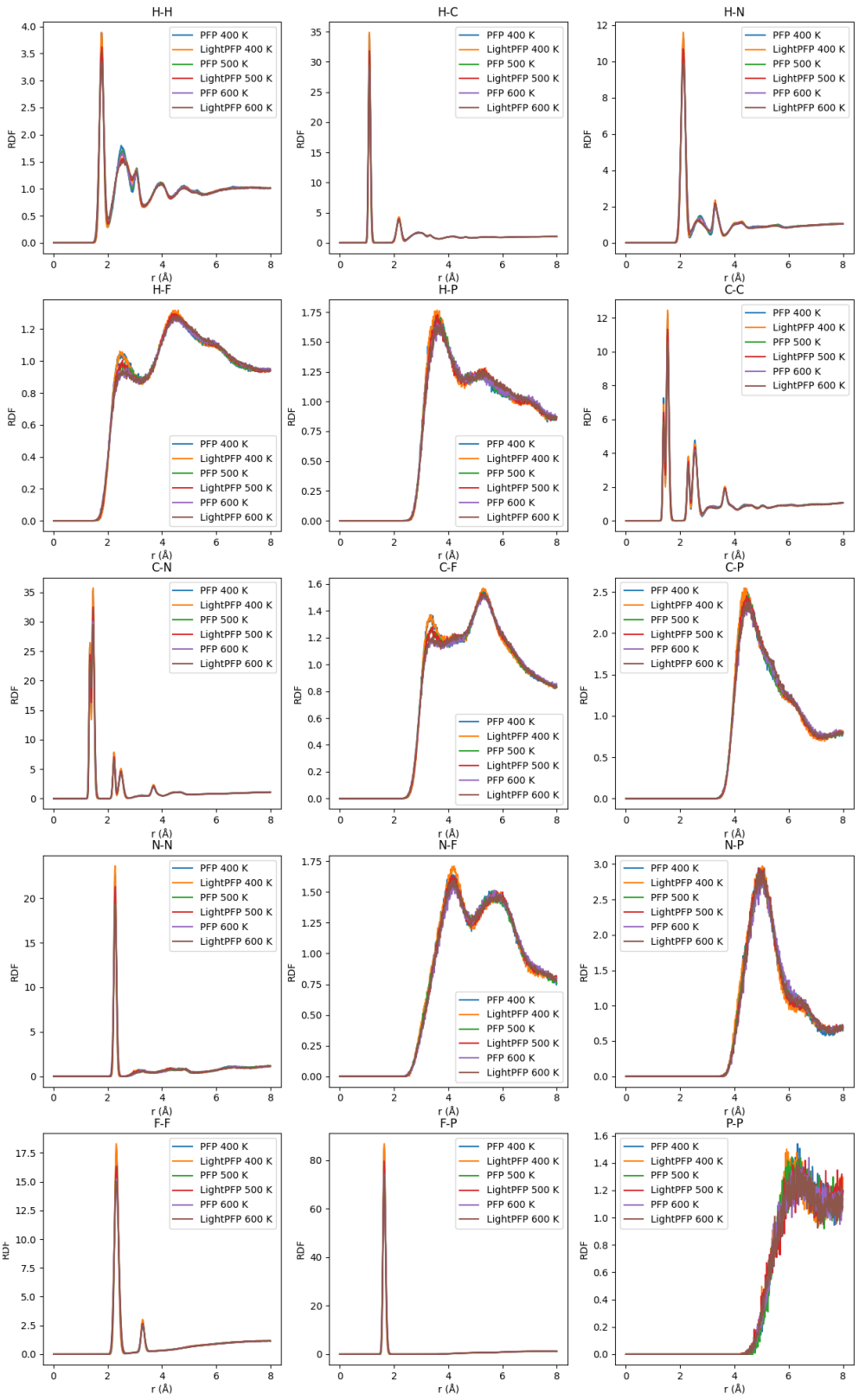

FIG. S19: Radial distribution function of polymer ionic liquid

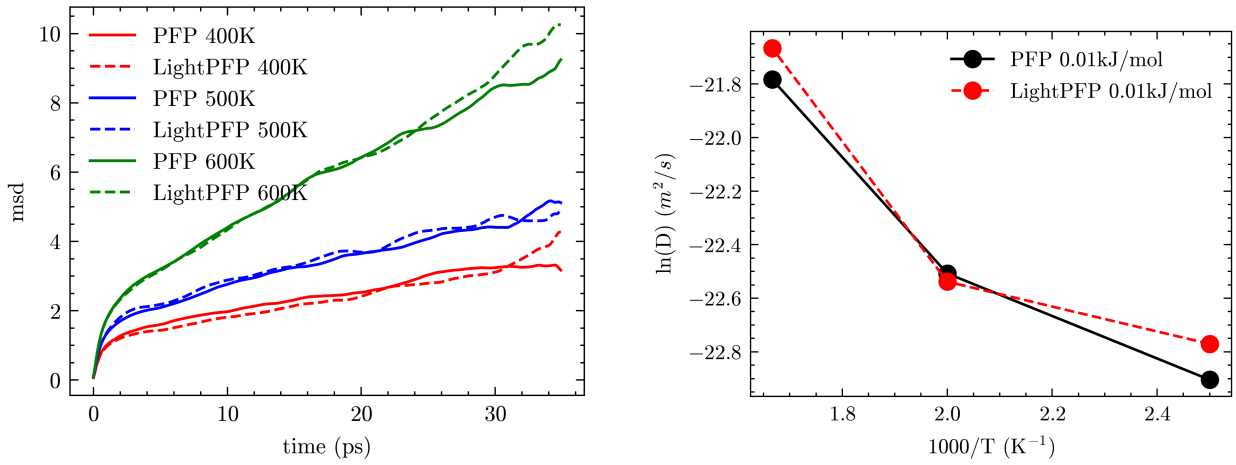

FIG. S20: Diffusion behavior of PF_6^- anion (left) mean squared displacement of PF_6^- anion at 300 K, 400 K and 500 K (right) Arrhenius plot of PF_6^- anion diffusion coefficient

estimates; the results are presented primarily to demonstrate consistency between the two models.

S8. APPLICATION 8: MECHANICAL PROPERTY OF SIO_2–P_2O_5–AL_2O_3–NA_2O GLASS

In this example, we use LightPFP to investigate composition–property relationships and the mechanical response of multicomponent oxide glasses in the SiO_2–P_2O_5–Al_2O_3–Na_2O system, as a stringent test on amorphous materials[11]. We consider glasses with compositions SiO_2:(79.69-x) mol%, P_2O_5:x mol%, Al_2O_3:13.79 mol%, and Na_2O:15.52 mol%, with x ranging from 0 to 50 mol%.

S8.1. Student fine-tuning

To build the initial training set, we combined crystalline and disordered configurations. First, crystal structures of SiO_2, P_2O_5, Al_2O_3, and Na_2O were obtained from the Materials Project to capture characteristic, energetically stable local environments (e.g., tetrahedral Si). Second, random packings were generated by placing atoms uniformly at random in periodic simulation boxes subject to a minimum interatomic separation, avoiding atomic overlap while spanning a broad space of disordered motifs. From these crystalline and random initial structures, we performed NPT MD at 500 K, 1000 K, and 1500 K, and sampled uncorrelated snapshots across densities and coordination states to train the initial LightPFP model.

Active learning was then used to refine accuracy for glassy states. For each composition, we initiated melt–quench protocols from random packings: (i) NVT MD at 2000 K for 10 ps to fully melt and randomize the network; (ii) linear cooling from 2000 K to 500 K at 0.1 K/fs to form the glass; and (iii) additional MD at 500 K to relax the structure. During these runs, we monitored model reliability and selectively augmented the training set with configurations exhibiting large force/energy discrepancies relative to the reference PFP, iterating retraining until convergence. This pipeline exposes the model to topological rearrangements (bond breaking/formation, modifier-induced non-bridging oxygens) and the broad spectrum of local environments that emerge during melt–quench.

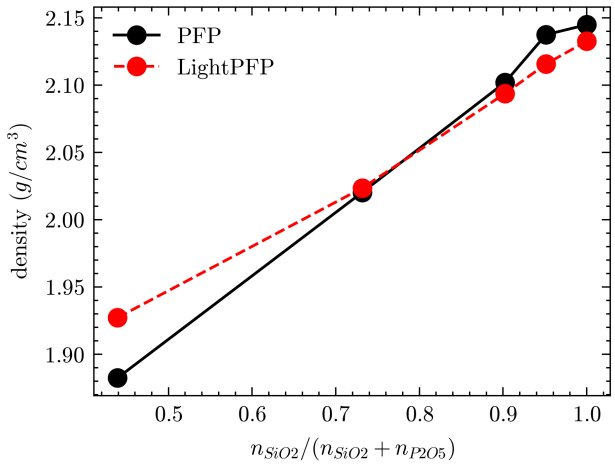

FIG. S21: Density of (82-x)SiO_2-xP_2O_5-16Al_2O_3-18Na_2O glass

S8.2. Evaluation using PFP

For evaluation, we generated glass structures at five representative compositions, i.e. (82-x)SiO_2–xP_2O_5–16Al_2O_3–18Na_2O, where x = 0, 4, 8, 22, 46. Identical melt–quench protocols were performed with both the PFP and LightPFP. Specifically, each system was equilibrated in the melt and then quenched at 0.02 K/fs to the 500K, followed by NPT relaxation to determine the density. The resulting densities are plotted in Fig. S21. In both PFP and LightPFP simulations, the density increases with SiO_2 content. The mean absolute error (MAE) of the LightPFP-predicted density across the five compositions is 0.014 g/cm^3.

In addition, We evaluated the glass structure using the RDF. Figure S22 shows the RDF for the composition 74SiO_2–16Al_2O_3–18Na_2O–8P_2O_5. LightPFP is in good agreement with PFP across the principal pair correlations.

To assess mechanical response, we computed the elastic stiffness tensor for the representative glass 74SiO_2–16Al_2O_3–18Na_2O–8P_2O_5 with both PFP and LightPFP. Results are listed in Table S3, including major components of elastic tensor and derived bulk, shear, and Young's moduli, as well as Poisson's ratio.

These results indicate that LightPFP can robustly learn and transfer the structural and mechanical behavior of complex multicomponent oxide glasses from PFP, while accommodating substantial composition variation from silica-rich to phosphate-rich regimes.

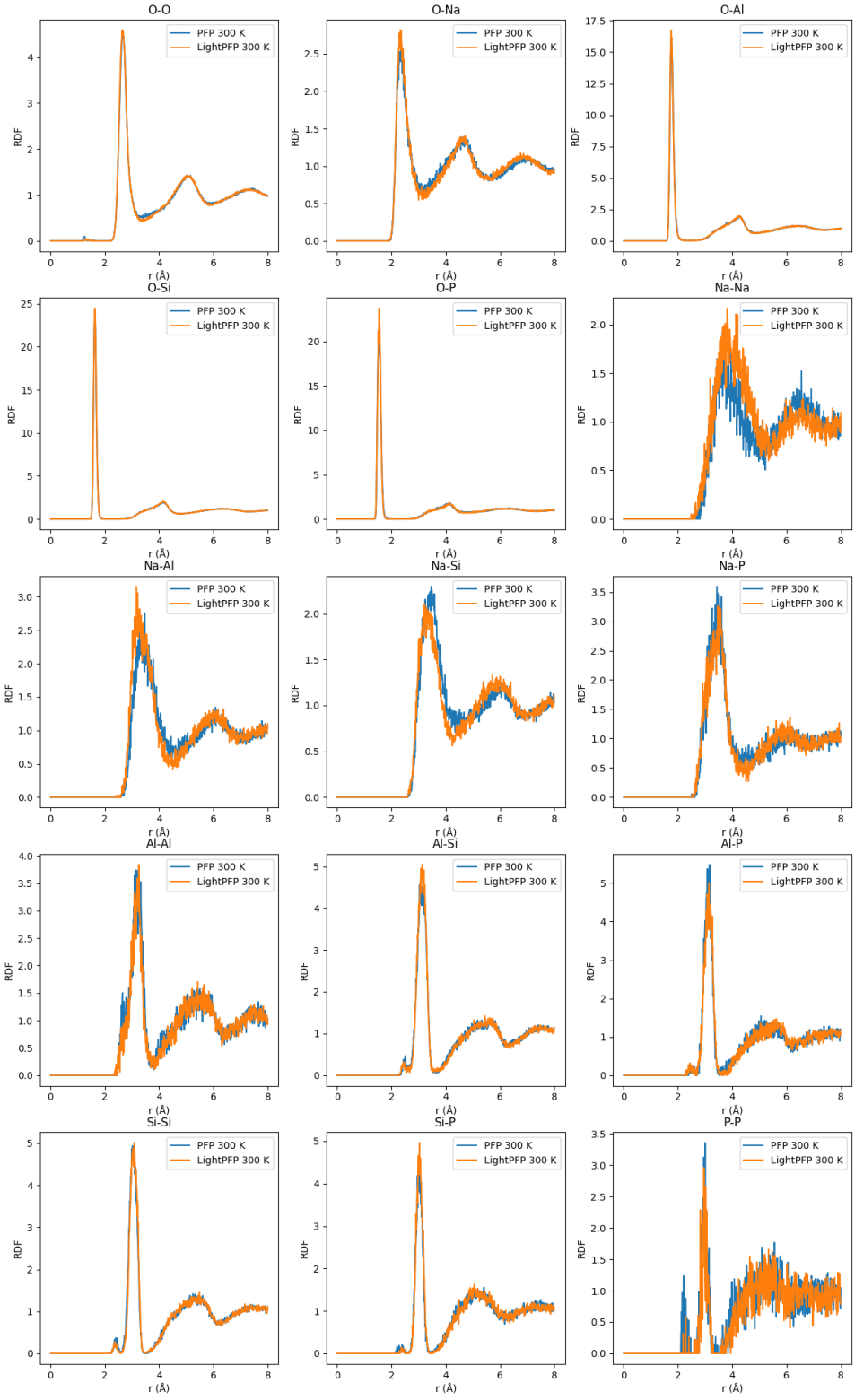

FIG. S22: Radial distribution function of $74SiO_2$-$16Al_2O_3$-$18Na_2O$-$8P_2O_5$

TABLE S3: Elastic properties of $74SiO_2$-$16Al_2O_3$-$18Na_2O$-$8P_2O_5$

	LightPFP	PFP	abs error	relative error
C11	53.312	55.138	1.825	0.033
C12	18.077	15.643	2.434	0.156
C13	18.169	15.699	2.470	0.157
C22	58.004	56.435	1.570	0.0278
C23	19.364	16.745	2.619	0.156
C33	58.283	58.618	0.335	0.00571
C44	19.709	20.521	0.812	0.0396
C55	19.283	20.065	0.782	0.0390
C66	18.493	19.606	1.113	0.0568
bulk modulus	31.163	29.580	1.583	0.0535
shear modulus	19.081	20.168	1.087	0.0539
Young's modulus	47.540	49.299	1.759	0.0357
Poisson ratio	0.246	0.222	0.0235	0.106

S9. APPLICATION 9: HETEROGENEOUS GRAIN BOUNDARY BETWEEN FCC CU AND BCC MO

In this example, we use LightPFP to investigate the energetics of grain boundaries between FCC Cu and BCC Mo via large-scale atomistic simulations.

S9.1. Student fine-tuning

We constructed initial Cu/Mo bicrystals using the cut-and-concatenate method, starting from ideal FCC Cu and BCC Mo crystals. The method is shown in Fig. S23. Simply speaking, the method takes two input crystal structures (i.e., Cu and Mo), cuts out cubic fragments at arbitrary positions and orientations, and then stitches them together to form grain-boundary–like structures. Unlike predefined low-energy grain boundaries structures, the generated structures are highly diverse. They often exhibit large lattice mismatch, and because the original periodic order is disrupted, numerous defects are introduced. This is not a drawback; rather, it is advantageous for training MLIPs, as it provides more off-equilibrium data samples and thereby improves the stability and robustness of the MLIP.

The initial database was assembled via short MD sampling on these crystal and grain boundary structures. Specifically, we performed NVT trajectories at 500 K, 1000 K, and 1500 K to introduce thermal disorder and local reconstructions, and NPT trajectories at 300 K, 400 K, 500 K, and 600 K to allow relaxation and sampling of strain-accommodated interfacial structures. The initial LightPFP model is trained on this dataset.

Starting from this initial model, we performed active learning to improve accuracy and robustness. For one initial Cu/Mo bicrystal structure generated by cut-and-concatenate method, we ran MD from 300 K to 1000 K using a two-step cycle: (1) 2 ps in the NVT ensemble to enable short-time reconstructions, followed by (2) 5 ps in the NPT ensemble to capture stress relaxation and incipient structural transformations. After that, configurations were relaxed by geometry optimization. Frames identified as low accuracy were added to the training set and the model was retrained. Iterating this procedure produced a LightPFP model that faithfully describes bulk phases and interface structure.

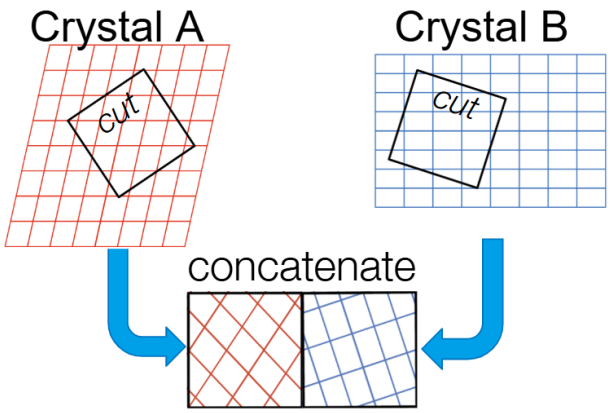

FIG. S23: Illustration of cut-and-concatenate method

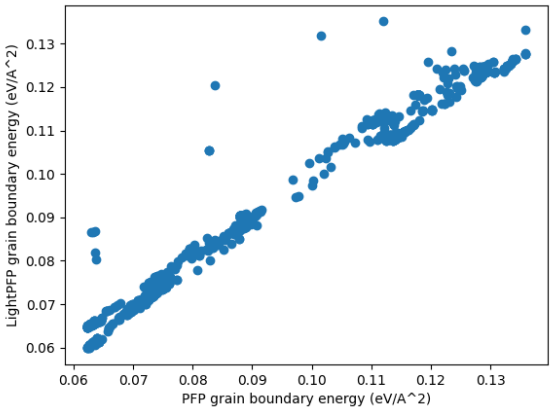

FIG. S24: Comparison of grain boundary energy of 1000 Cu/Mo interface structures calculated by PFP and LightPFP

S9.2. Evaluation using PFP

For evaluation, we generated 1000 low-strain CSL Cu/Mo grain-boundary structures using pymatgen, covering a broad range of misorientations and in-plane shifts. For each structure, we performed structure optimization first, and then computed the grain-boundary energy with both the reference PFP model and LightPFP. The grain-boundary energy was obtained by subtracting appropriate bulk reference energies for FCC Cu and BCC Mo from the total energy of the bicrystal and normalizing by the interfacial area (accounting for the two interfaces in periodic slabs). The resulting parity plot is shown in Fig. S24. Across the 1000-member test set, LightPFP achieves good agreement with PFP for the vast majority of boundaries.

Overall, these results demonstrate that LightPFP can accurately predict Cu/Mo grain-boundary energies over a broad structural space while matching the reference PFP model for most configurations, validating its use for high-throughput screening and large-scale interfacial simulations.

S10. APPLICATION 10: MICELLE SIMULATION

In this example, we investigated micelle formation using LightPFP. Micellization involves collective structural rearrangements of surfactant molecules, where hydrophilic groups orient toward the aqueous phase and hydrophobic tails aggregate to minimize their exposure to water.

As a representative system, we selected a gemini surfactant with dual hydrophobic tails and dual cationic headgroups, accompanied by bromide counter-ions (Br^-). The 12–s–12 family (where s denotes the spacer length) is known to form micelles efficiently due to its molecular architecture (see Fig. S25).

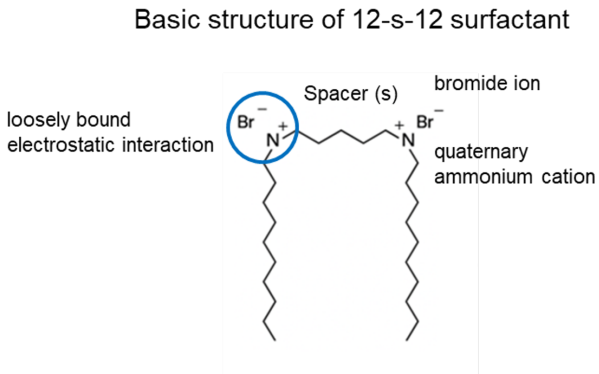

FIG. S25: Structure of 12-s-12 gemini surfactant with dual hydrophobic tails, dual cationic heads, and bromide counter-ions.

S10.1. Student fine-tuning

Initial structures for the 12–s–12 surfactant were prepared from SMILES expressions. For each spacer length s, we independently constructed initial datasets by generating multiple solvated structures with different numbers of surfactant and water molecules. Separate LightPFP models were therefore prepared for each value of s.

During active learning, small systems were used in the early iterations (0–4), medium-sized systems in iterations 5–8, and the largest systems in iterations 9–12. Within each iteration, internal loops varied the number of surfactant molecules, the water content, and the temperature (300–500 K) to enhance configurational diversity. After completing active learning, the collected dataset was used for final model training to obtain the LightPFP

model for production MD. The entire workflow required approximately half a day and was repeated for $s = 2, 6, 10$.

S10.2. Large-scale MD simulation

Production simulations consisted of 2 ns NPT runs at 350 K for $s = 2, 6, 10$. Fig. S26 shows the time evolution of the number of surfactant clusters. Initially, the cluster count equals the number of surfactant molecules; as aggregation proceeds, this count decreases. In all systems, the number of clusters gradually reduced, ultimately forming a large aggregate consistent with early-stage micelle formation. Representative initial and final snapshots are provided in Fig. S27. The final structures exhibit the expected morphology, where hydrophilic headgroups orient outward and hydrophobic tails are sequestered inside the aggregate.

While 2 ns is insufficient to observe complete micelle formation, these results demonstrate that LightPFP can capture the essential aggregation behavior. Notably, the use of the R2SCAN mode was critical for stability: when PFP's PBE+D3 mode was used instead, the simulations became unstable and eventually collapsed, likely due to the poorer description of water properties (e.g., RDFs, viscosity, density) at the PBE level. The improved accuracy of R2SCAN for aqueous systems appears to contribute significantly to the robustness of the MD simulations.

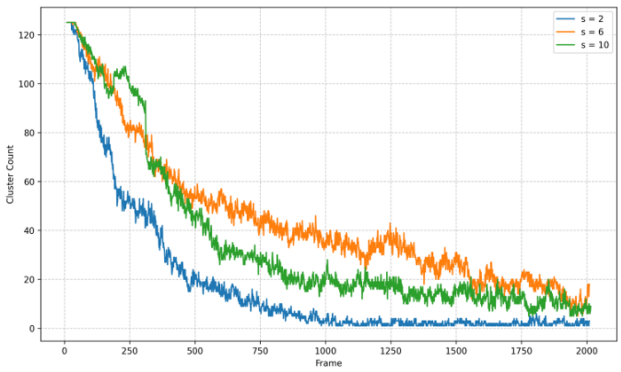

FIG. S26: Evolution of the number of clusters over time during 2 ns NPT simulations at 350 K for surfactants with spacer lengths s = 2, 6, and 10.

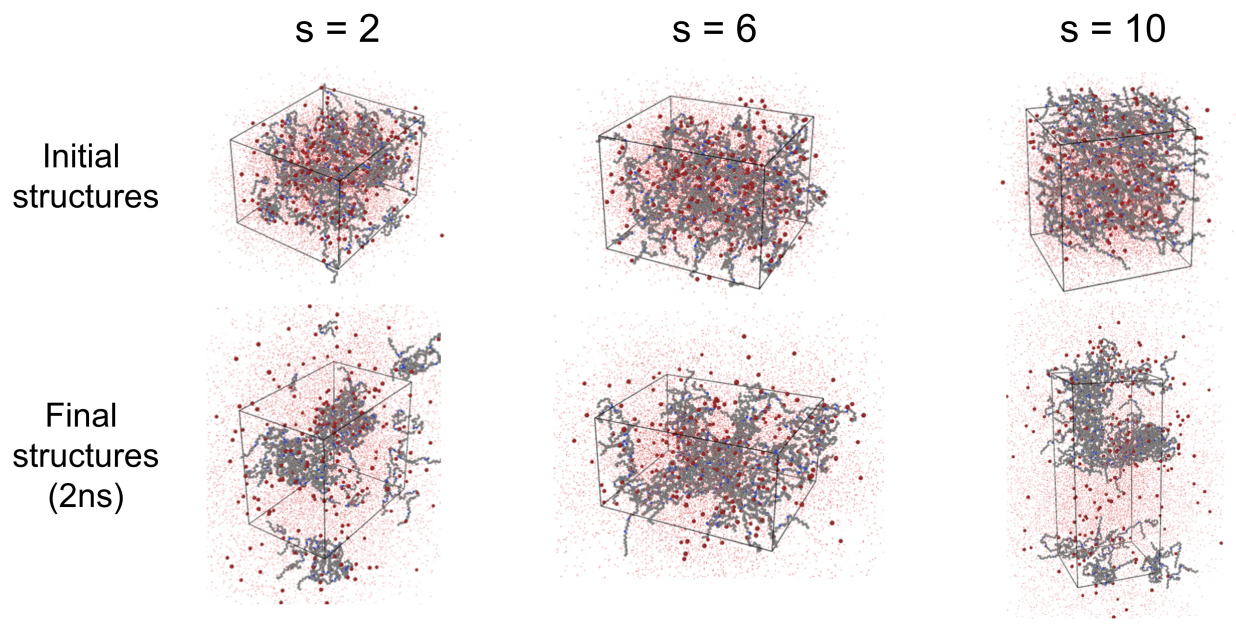

FIG. S27: Snapshots from MD simulations showing initial and final structures for systems with spacer lengths $s = 2$, 6, and 10, demonstrating micelle formation.

S11. APPLICATION 11: CHEMICAL MECHANICAL POLISHING OF SI SURFACE

To illustrate the practical application of LightPFP, we present an example reproducing abrasive rolling during the chemical mechanical polishing (CMP) of crystalline silicon by a silica particle[12]. The simulation box consists of a Si(100) slab and a spherical SiO_2 particle placed above the slab. The simulation protocol follows the two-stage loading scheme commonly employed in molecular dynamics (MD) studies of CMP: first, an external normal ("down") load is applied to the silica particle to bring it into contact with the silicon surface; then, a tangential ("driving") load is applied to the same particle to induce rolling and sliding motion across the slab.

S11.1. Student fine-tuning

The system construction and initial training dataset are intentionally minimal. Crystalline Si and SiO_2 structures are generated, from which spherical SiO_2 clusters are cut. Si and SiO_2 slabs are then prepared with vacuum layers, and a representative solid/solid interface is assembled through the simple cut-and-concatenate procedure introduced in Section S9. To populate the initial dataset, we sample crystalline structures using MD, rattle, compression, deformation, and vacancy methods, while non-crystalline and interfacial structures are sampled using MD and rattle only. This combination of bulk, surface, cluster, and interface configurations provides chemically diverse yet computationally inexpensive coverage for the first LightPFP fitting.

The active-learning protocol is used to generate a more robust LightPFP model. Active learning uses the same MD protocol as described above. The structure with large discrepancy with PFP will be detected and collected for further training. The active cycle is organized as a simple curriculum across particle sizes (small $\rightarrow$ medium $\rightarrow$ large), so that the model learns contact physics at increasing mechanical intensity. After the active acquisition phase the accumulated dataset is used for a single, consolidated retraining step to produce the final LightPFP model.

S11.2. Evaluation using PFP

For validation, we compare LightPFP with a PFP reference model using a small simulation cell containing 6133 atoms, following the same two-stage loading protocol. Four loading conditions were tested with down/driving forces of (5/10 eV/Å), (10/10 eV/Å), (5/20 eV/Å), and (10/20 eV/Å), each simulated for 200 ps. The number of removed atoms was determined by counting those displaced more than 2 Å during the MD trajectory, as shown in Figure S28. Both LightPFP and PFP exhibit consistent trends, with the number of removed atoms increasing in the order: (10/10) < (5/10) < (5/20) < (10/20). However, LightPFP consistently predicts a higher number of removed atoms than PFP under all loading conditions. This difference likely arises from subtle variations in surface interaction strength. Despite this quantitative deviation, the qualitative agreement in trend demonstrates that LightPFP accurately reproduces the underlying contact and material removal physics while offering improved sensitivity to interfacial dynamics.

S11.3. Large-scale MD simulation

Finally, we demonstrate scalability by simulating the polishing of Si by a 5 nm-diameter silica particle in a dry environment. This production system comprises 59,266 atoms and is evolved for 0.6 ns under a 50 eV/Å normal load and a 100 eV/Å driving load. The snapshots of the MD trajectory are presented in Figure S29, showing the motion of the SiO_2 particle and the corresponding polishing process.

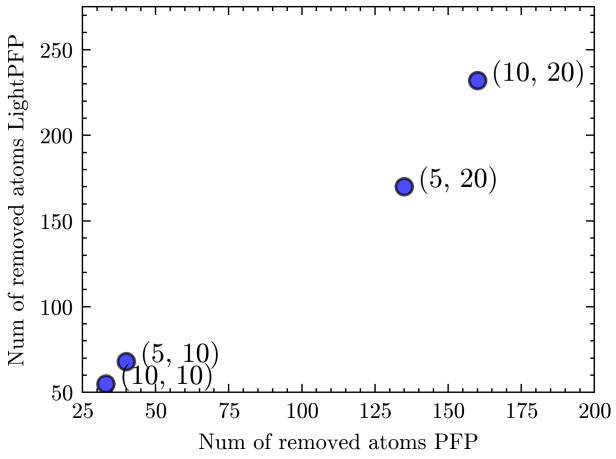

FIG. S28: Number of removed Si atoms from Si slab during MD simulation of chemical mechanical polishing. The down/driving force applied on particle is annotated.

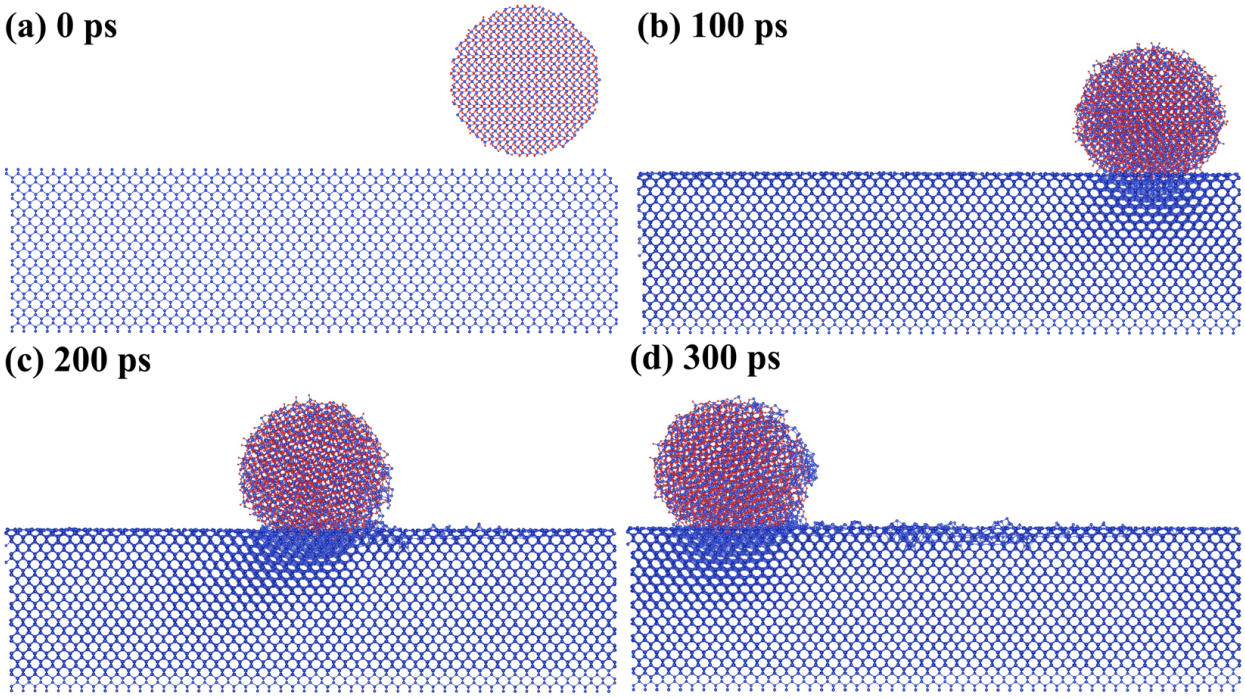

FIG. S29: Time evolution of the molecular dynamics simulation illustrating the chemical mechanical polishing (CMP) of crystalline Si by a 5 nm SiO_2 particle. Panels (a)–(d) correspond to 0, 100, 200, and 300 ps, respectively, showing progressive rolling motion, and surface atom removal on the Si slab.

REFERENCES

[1]C. Morin, D. Simon, and P. Sautet, The Journal of Physical Chemistry B **108**, 5653 (2004).

[2]National Institute of Standards and Technology (NIST) and International Union of Pure and Applied Chemistry (IUPAC), "Iupac-nist solubility database," Accessed: 2025-10-13.

[3]A. Stukowski, Jom **66**, 399 (2014).

[4]National Center for Biotechnology Information, "Pubchem compound summary for heptane," PubChem, section: Solubility. Accessed 2025-10-13.

[5]I. Carrillo-Berdugo, J. Navas, and R. Grau-Crespo, The Journal of Chemical Physics **160** (2024).

[6]B. Hess, The Journal of chemical physics **116**, 209 (2002).

[7]"Nist chemistry webbook," (2025), entry: n-Decane (CAS RN 124-18-5).

[8]G. S. Jung, H. Myung, and S. Irle, Machine Learning: Science and Technology **4**, 035001 (2023).

[9]M. Doig, C. P. Warrens, and P. J. Camp, Langmuir **30**, 186 (2014).

[10]X. Luo, H. Liu, and S. J. Paddison, ACS Applied Polymer Materials **3**, 141 (2020).

[11]Y. Qian, B. Song, J. Jin, G. I. Prayogo, K. Utimula, K. Nakano, R. Maezono, K. Hongo, and G. Zhao, Journal of the American Ceramic Society **105**, 6604 (2022).

[12]L. Si, D. Guo, J. Luo, X. Lu, and G. Xie, Journal of Applied Physics **109** (2011).