Eliminating Hallucinations in Machine Learning Interatomic Potentials

Executive Summary

Machine learning interatomic potentials (MLIPs) are revolutionizing materials science by providing accurate low-cost methods to predict the properties of materials. Their accuracy and cost make them the first atomic scale method that is able to widely replace fundamental physical measurements in materials research and development. Despite the significant growth of open-source datasets for training general purpose MLIPs, pretrained MLIPs often fail when applied to industrial materials problems. The chemical and structure complexity of materials yields a combinatorially large design space that causes pretrained models to often hallucinate due to the likelihood models are evaluated outside of their training datasets. Physics Inverted Materials systematically eliminates hallucinations in MLIPs through a novel active learning algorithm to enable broad deployment of MLIPs in industrial settings. The algorithm eliminates hallucinations by dynamically expanding datasets and fine-tuning models that are more accurate than pretrained MLIPs. Our integrated software solution ensures that any simulation run using PHIN's MLIPs are accurate and trustworthy.

Background: Machine Learning Interatomic Potentials

Modeling materials at the atomic scale is critical to determining the mechanisms that govern materials behavior. Industrial materials are characterized by fundamental materials properties such as ion conductivity, charge transfer, decomposition reactions, or phase changes. These properties are critical for building a researcher's intuition about how a material behaves. As the accuracy and generalizability of atomic scale models has improved, scientists have been able to study these properties digitally. The ability to model many of these properties shifted dramatically with the emergence of density functional theory (DFT) codes and generalized pseudopotentials in the 1990's that made accurate modeling possible for computational scientists. However, using DFT calculations for finite temperature simulations is expensive, costing thousands of dollars for relatively simple problems. Machine learning can overcome the high cost of DFT by modeling the relationships of the DFT inputs – atomic positions – and its outputs – energies, forces, stresses, and charge density. The result is machine learning interatomic potentials (MLIPs) that predict the energy, forces, and stresses from the atomic postitions. In comparison to classical interatomic potentials that also reduce the cost of DFT, MLIPs are able to attain much higher accuracies, making MLIPs the first viable pathway to accurate and low-cost materials modeling.

MLIPs have evolved considerably in the twenty years since the first Behler-Parrinello Neural Networks were developed. These first MLIPs used descriptors to embed atomic environments as feature vectors that neural networks used to predict the energies of atoms. While many of the principles embedded in atomic descriptors have persisted, the explicit featurization of physical structure has evolved to use extremely flexible graph neural networks. Instead of a fixed local environment, state of the art models operate on the graph of a local atomic environment to create representations that approximate any potential energy surface (PES). Many graph neural network (GNN) architectures have been proposed that learn these representations in slightly different ways by adding or removing physical constraints to balance its accuracy, data-efficiency, and cost.

Graph neural network MLIPs paved the way for the first Universal MLIPs (uMLIPs). The highly generalizable GNN architectures were capable of accurately predicting properties across a vast range of chemical systems. The first models were trained on The Materials Project, an open-source database of periodic structures calculated with density functional theory (DFT). The Materials Project database curates equilibrium structures of hundreds of thousands of materials across 89 elements in the periodic table. While not the first uMLIP, the MACE models demonstrated incredible capability of modeling materials across a vast range of chemical and materials properties [1].

The impressive generalizability of MACE was a breakthrough that ushered in a wave of pretrained models that could run simulations for most materials. However, it quickly became appartent that pretrained MLIPs would hallucinate. While some hallucinations are catastrophic, causing simulations to "explode", most go unnoticed. While irritating, catastrophic hallucinations are less problematic because you know there is a problem. The small errors from silent hallucinations, however, often go unnoticed leading to potentially large errors in downstream material property predictions. For instance, we found that pretrained models with R^2 values of 0.96 lead to errors in melting point predictions of 500 ^{o}C. It is necessary to increase R^2 values above 0.99 to achieve accuracies of 50 ^{o}C. While many uMLIPs can achieve accuracies greater than R^2 of 0.99 on training datasets, it is very challenging to ensure this on generic simulations as they are likely to be outside their training datasets. Identifying when models decrease in accuracy and fine-tuning them is the key step in eliminating hallucinations.

Fine-tuning is the process of retraining a uMLIP to increase its accuracy for a material property of interest. It has been successful at enabling MLIPs that can accurately predict materials properties that had been previously impossible to simulate [cite]. Knowing and developing the fine-tuning dataset is, however, a large problem. The quality of the fine-tuning dataset determines the final accuracy of the material properties calculated, while the size determines the economics of generating it in the first place. Experts therefore painstakingly determine the required structures that represent the important information for a simulation.

To access the benefit of MLIPs for accurately modeling the properties of materials will require addressing both the prevalence of silent hallucinations and the difficulty in building fine-tuning datasets. Both can both be addressed with efficient uncertainty quantification techniques.

Uncertainty Quantification in MLIPs

Uncertainty quantification (UQ) is a key method that can overcome the problems plaguing widespread adoption of MLIPs by copredicting the accuracy of its predictions. It addresses silent hallucinations by identifying when a model has inaccurate predictions and enables building fine-tuning datasets by labeling inaccurate data that can improve model predictions.

Due to the improtance of UQ, extensive research has gone into developing different approaches [cite], which can be categorized into three major types of UQ in MLIPs are Gaussian Process, ensemble, and heuristic. Due to the importance of UQ in accurate materials modeling, each of these has numerous variants and implementations with approaches even combining them.

Gaussian Process UQ involves adding a Gaussian Process model as the readout layer. A Gaussian Process defines a distribution over possible functions, specified by its mean and covariance (kernel) functions. This allows GP models to not only predict function values but also quantify the uncertainty of those predictions, making it suitable for complex, unknown relationships. While powerful and offering probabilistic predictions, Gaussian Process models are limited by their poor scalability to large datasets leading to work into sparsifying the representation (e.g. FLARE). Still, GP has been widely used in efficiently generating datasets for individual systems [cite] and used extensively for accelerating AIMD simulations in VASP.

Ensemble UQ methods predict model uncertainty by training multiple independent models. The average of the predictions from these models provides the estimated value, while their standard deviation serves as a measure of the associated error or uncertainty. Ensemble UQ is routinely shown to be the most accurate UQ method. The main drawback of ensembling is the substantial computational overhead; for instance, using four independent models—largely considered a minimum—results in a 300% overhead for model training and evaluation. This high cost makes ensemble UQ infeasible for state-of-the-art GNN MLIPs, leading to the exploration of more efficient variants like shallow ensembles or local linear prediction rigidities (LLPR).

Heuristic methods define uncertainty by relating a model property to the uncertainty. A common property to use as a heuristic for model uncertainty is the distance to a training data point in either feature or latent space. The heuristic is computed along with the model outputs to predict the uncertainty of the model. The benefit of this approach is that it can often be applied after model training as a separate step allowing it to be used with pre-trained models. However, specifically in the case of distance based heuristics, the cost scales with the dataset size, decreasing its usefulness for uMLIPs.

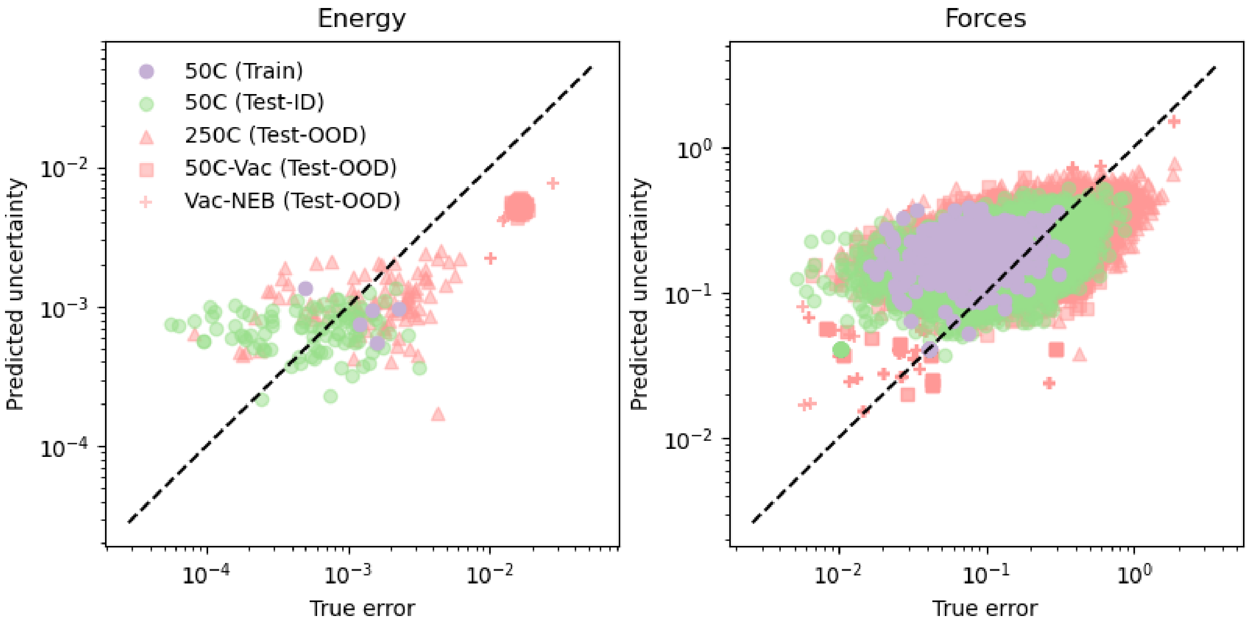

Figure 1: Demonstration of PHIN's uncertainty quantification for a model trained on lithium cobalt oxide ($LiCoO_2$) from a limited sample of 50^{o}C MD snapshots and then evaluated on in (50^{o}C MD) and out-of-distribution (250^{o}C MD, 50^{o}C MD with a vacancy, and an NEB simulation with a vacancy) datasets.

The key considerations for choosing which UQ method to use come down to application, where fine tuning uMLIPs using active learning is one of the most challenging applications. The large datasets required for training uMLIPs mean that it is critical that the method scales efficiently with dataset size and can use GNN architectures. This significant constraint means that even though Gaussian Process and deep ensembles are the gold standard, they are not seen as viable strategies because they make training foundation models prohibitively expensive. Furthermore, the large datasets of foundation models eliminate heuristic based distance methods as computing the heuristic becomes more expensive than the model evaluation.

PHIN has made significant strides to overcome this longstanding is and developed a proprietary UQ technique that blends ensemble and heuristic approaches to overcome the limitations of each. PHIN's proprietary UQ enables using a single model to compute the uncertainty. This significantly reduces the training and inference cost associated with ensemble approaches. In addition, PHIN blends a heuristic UQ method to identify the similarity between uncertain to sparsify identified points during active learning to minimize fine-tuning dataset sizes.

The performance of PHIN's single model UQ is seen in Figure 1. The figure shows the ability of PHIN's models to separately classify training, in-distribution, and out-of-distribution data. The fine-tuning dataset uses one of the most common cathode materials in lithium-ion batteries and investigates the accuracy of a model trained on crystalline $LiCoO_2$ at 50^oC and how it compares to in-distribution samples also taken at 50^oC and to out-of-distribution samples at 250^oC temperature, with a vacancy defect in the lithium sub-lattice simulated at 50^oC, and with a nudged elastic band (NEB) simulation that investigates the metastable configurations present during charging and discharging. The figure clearly shows distinct regions separating model performance on the training data and in- and out-of- distribution test data. Importantly, the uncertainty predicted by PHIN's UQ is highly correlated to the model error. This makes it possible to define thresholds based solely on the training data that can be used to systematically identify where a model is hallucinating. To further investigate the performance of PHIN's UQ, we next move to how it is integrated into an active learning pipeline for automating the process of fine-tuning uMLIPs.

Active Learning

Active learning enables automatically fine-tuning uMLIPs and combines UQ, model retraining, and automated data collection into a single system. PHIN's proprietary active learning algorithm is unique in how it embeds directly into simulation workflows. This eliminates the traditional two-step process of first generating a model and then simulating a material property. By integrating material simulations into the active learning algorithm, we can submit all simulations through active learning and allow active learning to govern fine-tuning.

PHIN's active learning algorithm for accurately simulating any material property is shown in Figure 2. The integrated active learning methodology makes the system very easy to integrate into traditional simulation workflows, where the active learning inputs are the initial structure and simulation protocol and the active learning output is the final material property calculated.

The performance of the active learning algorithm is dictated by the capacity and data efficiency of the MLIP architecture and performance of the UQ. The active learning performance is critical to being able to successfully scale active learning into production environments as the cost of data generation and model retraining need to be amortized over many simulation runs.

The improvement PHIN has made in its active learning algorithm can be seen in Figure 3. The baseline is comparable to other state-of-the-art active learning systems [cite], from which we reduced the time by 97% and the data generation cost by 99%. This particular benchmark focused on fine-tuning an MLIP to accurately model the melting of silicon from a model trained on crystalline data. This benchmark is challenging because it represents a significant out-of-distribution test, where the model is prone to hallucination. Moving from crystalline to melted, amorphous structures is key to the fine-tuning uMLIPs need, since most pretrained uMLIPs are built on (near-) crystalline data. Furthermore, the phase transition requires crossing an energy barrier to enter a new metastable state similar to chemical reactions. Being able to accurately capture changes to the structural order is therefore a key performance metric for active learning to capture well.

PHIN has been able to improve the data and time efficiency considerably. The time efficiency is critical to fine-tuning models quickly, while data-efficiency is necessary to reduce the cost of generating additional training data. For this benchmark, we are able to show that we are able to train a model that accurately simulates the melting of silicon in as little as 12 hours with 30 data points. Even though it is representative, the melting silicon benchmark is still relatively simple as there is only a single change to structural diversity. In reality, when modeling complex chemical reactions with rich structural diversity, it still takes weeks to fine-tune a model from a uMLIP. PHIN is pioneering system-specific fine-tuned models that are accurate for a particular system. This engineering limits the training cost associated with building the uMLIP.

While maintaining many fine-tuned models is a challenging infrastructure problem, it also provides an avenue to improve model performance considerably. Segmenting fine-tuned models to individual systems allows the ensemble to benefit from the theory behind mixture of expert models.

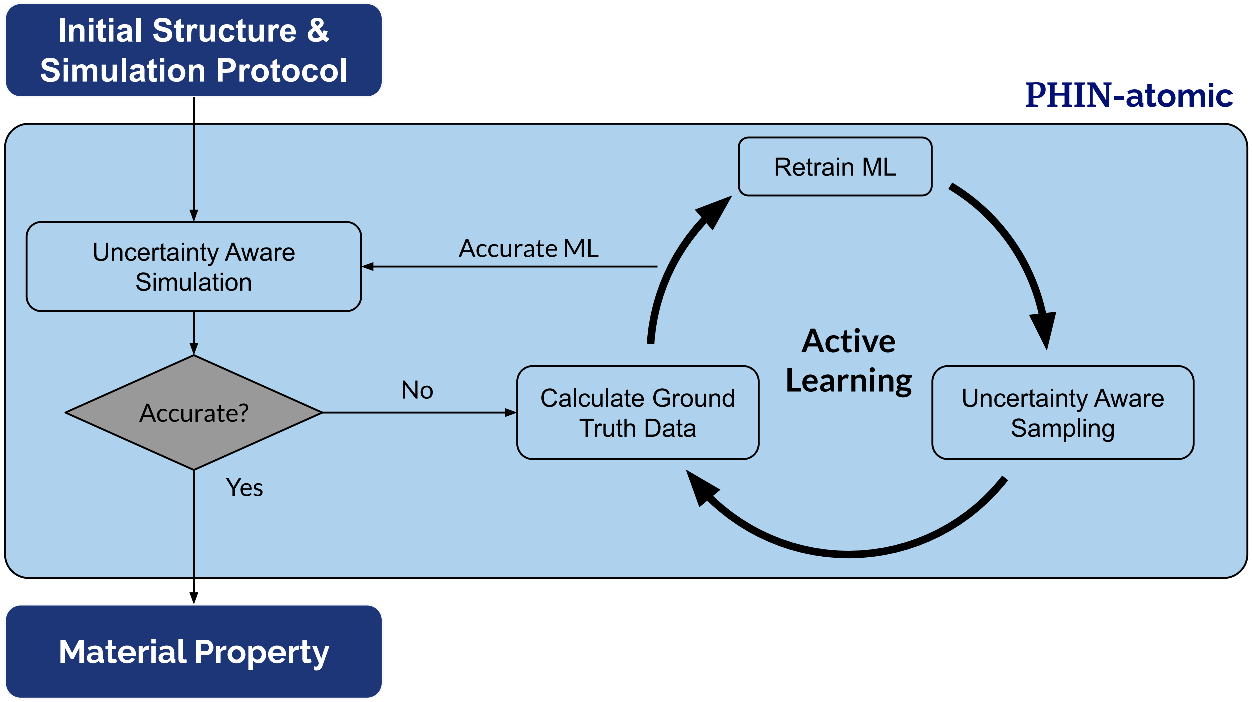
Initial Structure & Simulation Protocol
PHIN-atomic
Retrain ML
Uncertainty Aware Simulation
Accurate ML
Active Learning
Accurate?
No
Calculate Ground Truth Data
Uncertainty Aware Sampling
Yes
Material Property

Figure 2: Architecture of PHIN's active learning algorithm call PHIN-atomic. PHIN-atomic mimics traditional simulation workflows that start with an initial structure and simulation protocol and return a material property. PHIN-atomic starts by running uncertainty aware simulations using PHIN's single model UQ. It returns the simulation result if it is accurate. If not, PHIN-atomic fine-tunes the model using active learning by calculating new ground truth data, retraining the machine learning model, and performing local sampling. The fine-tuned model is finally used to run a simulation until it accurately simulates the desired process.

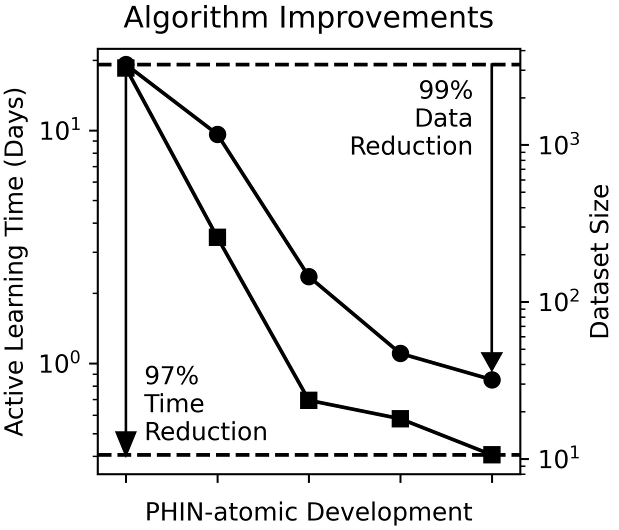

Figure 3: The iterative improvements to PHIN's machine learning model architecture and uncertainty predictions improved the active learning data efficiency by 99% and time efficiency by 97%.

MoE models enable increased parameter counts while maintaining inference cost by using a router to activate different regions of the network during inference. The router in this case is the active learning that turns on a fine-tuned model whose entire representation power can be used to improve performance on just its chemical system. Information from all other chemical systems are embedded in the starting weights for fine-tuning and are periodically updated when a new uMLIP is trained on all the data.

Our internal infrastructure currently maintains over 500 individual fine-tuned models representing different chemical compositions that have been queried through our web application.

Case Studies

The success of PHIN's active learning algorithm PHIN-atomic in accurately predicting the properties and behaviors of materials is shown for a number of systems from single elemental systems to surfaces, surface reactions, and phase changes across battery and semiconductors.

Battery Anode and Electrolyte Reactivity

Lithium-ion batteries have transformed the energy sector by providing affordable energy dense storage that is creating new product categories in mobility, defense, and infrastructure. A key aspect to solving lithium-ion safety and improving energy density is designing the solid-electrolyte interphase (SEI). The SEI prevents runaway degradation of the electrolyte and is formed through self-limiting reactions of the electrolyte with the anode. The simulation sizes and chemical complexity needed have made simulating the SEI computationally challenging for both DFT and classical interatomic potentials. Here we showcase the ability of PHIN-atomic to fine-tune MLIPs on on bulk lithium and a (110) lithium surface interacting with ethylene-carbonate. Using PHIN-atomic, we run atomic

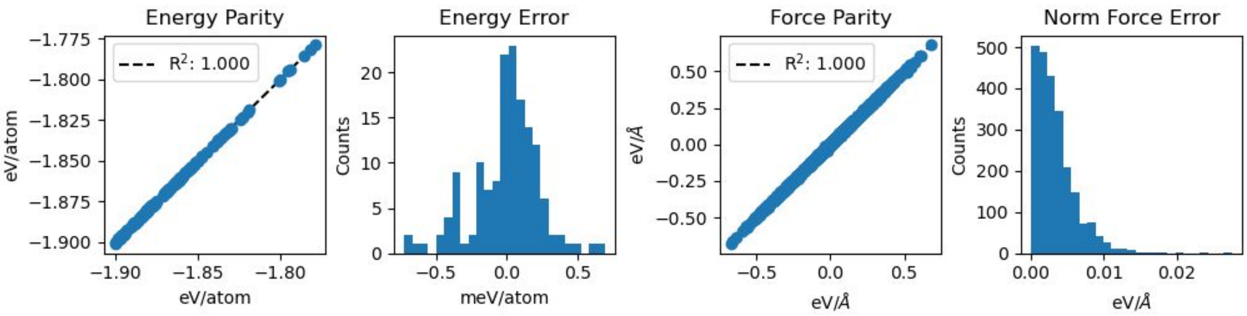

Figure 4: Accuracy comparison of density functional theory with machine learning interatomic potentials on the test set of an active learning run for pure lithium system.

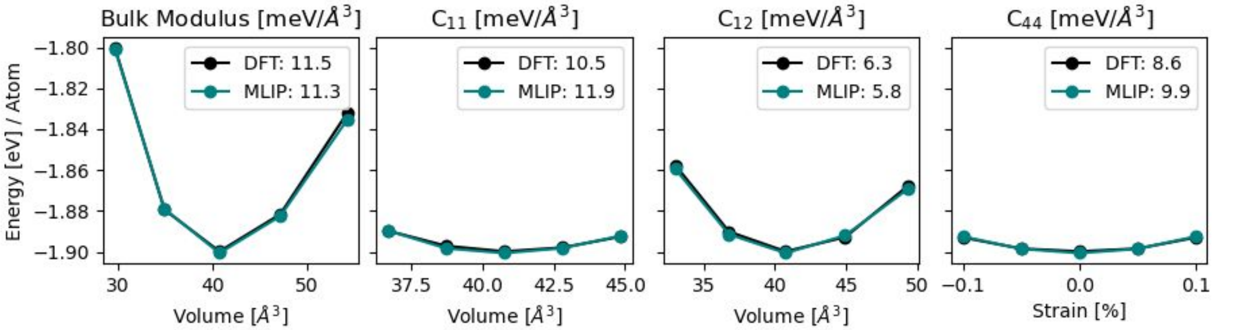

Figure 5: Comparison of energies and computed elastic constants for crystalline lithium structure for the bulk, C_{11}, C_{12}, and C_{44} elastic properties.

scale simulations in LAMMPS and ASE to simulate the mechanical properties of lithium and the fundamental chemical reactions that form the SEI.

Lithium Mechanical Properties

Before simulating the electro-chemical properties of the SEI, we ensure that PHIN-atomic can accurately predict the mechanical properties of bulk lithium. The MLIP accuracy is evaluated first on the properties the MLIP directly predicts, the energy and forces. Test datasets are curated by sampling the production simulations. The MLIP performance is reported through parity plots and error histograms on the test dataset. For the pure lithium system, the test mean absolute error on energies is 0.17 meV/atom while for forces is 3meV/Å leading to R^2>0.999 for both energy and forces.

Moving past the raw energies and forces, the fine-tuned MLIP accuracy is evaluated for the predicted mechanical properties of lithium. The energies for the bulk modulus, C_{11}, C_{12}, and C_{44} are plotted in Figure 5. The excellent agreement between the total energies computed from DFT and predicted by the machine learning model is expected due to the performance of the parity plots. Furthermore, the computed elastic moduli between the two methods are also in excellent agreement. This is particularly impressive for the slight energy differences present in the shear elastic modulus

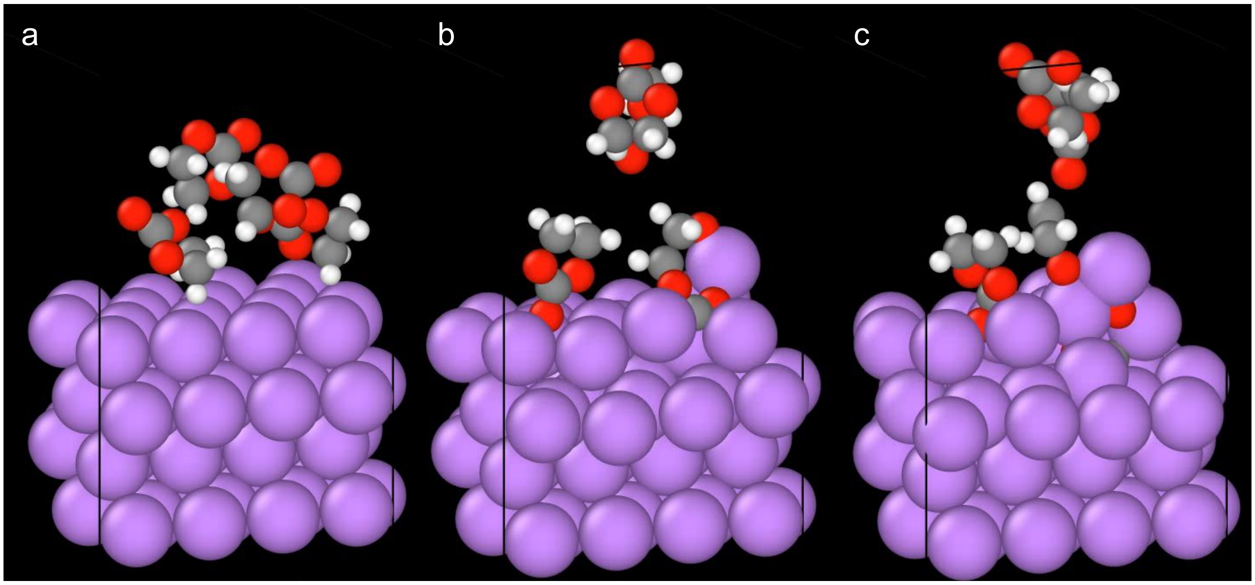

Figure 6: A DFT size unit cell of ethylene carbonate decomposing on a (110) lithium surface. (a) The initial state of the simulation. (b) The ring opening reaction mediated by a lithium atom. (c) The separation of ethylene and the precurors of the inorganic SEI layer.

C_{44}. The predicted C_{44} modulus is often incorrect and sometimes even negative because it is easy for MLIPs to predict random errors that have negative curvatures.

Surface Reactivity

With the performance of PHIN-atomic validated on lithium metal, we extend the fine-tuning from bulk lithium to a lithium surface interacting with an ethylene-carbonate solvent. Adding ethylene carbonate enables studying the fundamental mechanisms of SEI formation as it is predicted to be a key constituent in lithium-ion electrolytes for promoting stability. Here we use constant particle, volume, and temperature (NVT) simulations to predict the reactions that take place. We focus on two different simulations. The first is a smaller simulation that is run for 1ps. This simulation, which would have taken three months with DFT, can be run in ten minutes with a fine-tuned MLIP. The snapshots of the small NVT simulation in Figure 6. The snapshots clearly show the established reaction process for ethylene-carbonate decomposition [cite]. The molecules are (a) initialized above the surface before they (b) adsorb onto the surface and (c) decompose through the expected ring-opening mechanism to separate the carbonate from the ethylene.

With the trained MLIP, we can scale the simulation to larger simulations for collecting statistics on the decomposition reactions. In Figure 7, we showcase the size extensivity of the MLIP by running a 10ps NVT simulation of the same lithium ethylene-carbonate system. In this simulation, we see the same adsorption and ring opening reaction also observed in the smaller system and additionally observe the separation of the ethylene from the carbonate, where the carbonate forms lithium-carbonate (Li_2CO_3) and the ethylene offgases.

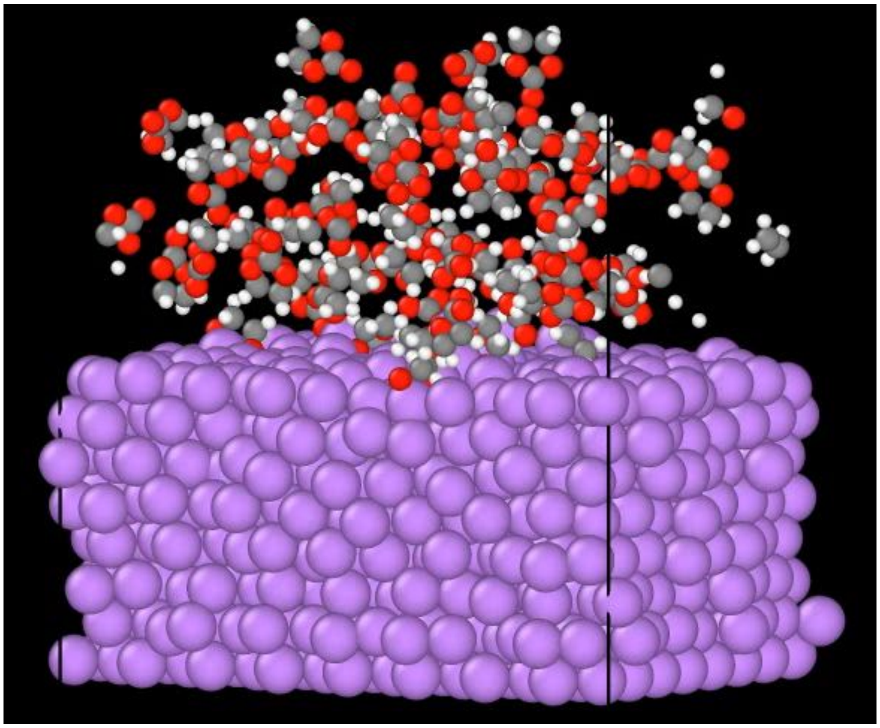

Figure 7: A snapshot from a simulation beyond the scope of DFT simulation sizes to show the size extensivity of MLIPs.

Silicon Case Study

Silicon has been a core technological material in the semiconductor industry for nearly a century and is becoming more critical in the energy storage market, making it a technologically relevant benchmark material with established material properties and digital models. To benchmark the performance of PHIN-atomic for fine-tuning on silicon, we simulate the elastic properties and melting temperature of diamond cubic crystalline silicon before moving to vacancy and surface energy calculations. We compare the accuracy of active learning to ground truth quantum mechanics data and experimental quantities as well as two alternative digital models: a Stillinger-Weber classical interatomic potential and the open-source uMLIP MACE. These alternative models provide insight into the relative performance of PHIN-atomic. The decades of development into the functional form of the Stillinger-Weber potential makes it an accurate method for digitally predicting silicon properties while the open-source uMLIP is an alternative machine learning method that does not use active learning.

Elastic Constants

Elastic properties are determined by applying a small deformation to a material and measuring the induced stress. Linear elastic constants of materials enable quickly determining how a material deforms under an applied strain, where the elastic response is the initial deformation characterized by a linear dependence between the stress and strain. The direction of the applied strain and measured stress give a tensor of all the elastic properties of a material.

For silicon, we measure the three independent elastic constants C_{11}, C_{12}, and C_{44} for each of the different digital models. Since silicon has been extensively researched for over half a century, the classical interatomic potentials are quite good and match experimental results well in Figure 8. The agreement between the classical and DFT and classical and experimental measurements are a good baseline for the accuracy a digital model should have to be useful in digital design. The error between PHIN's model and DFT/experiment for 0K/300K respectively is equal to or lower than the classical interatomic potential error for C_{11} and C_{44}, while the error for C_{12} is marginally higher. When we calculate the same elastic constants with the pretrained uMLIP MACE, the elastic constants it predicts are systematically lower than PHIN's. This is in alignment with previous work showing that the pretrained models have artificially low elastic constants [cite]. However, since the classical potential and PHIN MLIP overestimate C_{12}, the systematic softening of the pre-trained model means that it predicts C12 pretty well by chance.

Melting Temperature

Predicting the transition from a solid to liquid is important for designing thermal stability, strength, and manufacturability. The melting temperature, however, is tricky to measure, both experimentally and digitally. In an experiment, the melting temperature is observed by either heating or cooling a material and watching the temperature change over time. The phase transition requires a certain amount of energy creating a plateau in the temperature versus time plot. However, to change phases a material often needs to be superheated or supercooled to overcome an activation energy needed to create the nuclei of the new phase. This is observed as a temperature overshooting the melting temperature before settling into the melting temperature. The activation barrier means that we cannot use the same procedure to digitally predict the melting temperature using an atomic scale

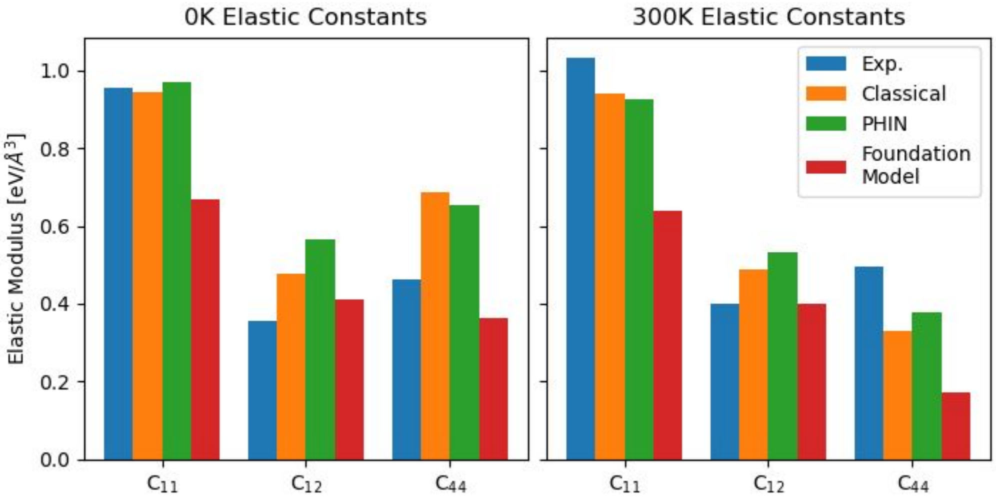

Figure 8: A comparison of the elastic constants of diamond silicon at 0K (left) and 300K (right) for DFT, Classical, PHIN, and Foundation Models.

simulation because of the time it would take to overcome the activation barrier. Instead, clever methods have been devised to sidestep the activation barrier and enable the accurate estimation of melting temperatures from atomic scale simulations. To overcome this, we use the SLUSHI method to initialize a supercell that is half solid and half liquid [cite]. During an NPT simulation, the solid-liquid interface can move, enabling the solid or liquid to grow. The heterogeneous nucleation has a much lower activation barrier enabling simulations of 100ps to be used for estimating the melting temperature. The trace of many simulations at different temperatures reveals the phase transition in an enthalpy versus temperature plot.

This method, although developed with density functional theory in mind, is still expensive to run with DFT. Therefore, we only compare the predicted melting temperature of PHIN against the pretrained model and a classical interatomic potential with the DFT prediction coming from the literature [DFT Si].

The enthalpy traces used by the SLUSHI method are reported in Figure 9 along with the computed melting temperatures. While the melting temperatures predicted with PHIN and classical interatomic potentials agree with DFT, both having roughly 5% error, the foundation model has more than a 25% (400K) error. The significant error of the pre-trained model is due to the dataset being biased towards low-temperature data. This is an example of a silent hallucination. While the model can be used to simulate melting, hallucinations in the force, energy, and stress predictions mean that its predictions are inaccurate. Without fine-tuning, the pretrained models don't give reliable estimates. Even worse, there is no indication from the model that it is inaccurate. The combination of the quiet failing and expert knowledge required to retrain these models make them inaccessible to large numbers of materials scientists.

To understand the silent hallucinations of the pre-trained foundation model more deeply, we investigate the accuracy of the model across the generated fine-tuning dataset. Because simulation errors accumulate during property predictions, the errors in MLIPs that lead to property predictions greater than 25% are actually quite small. The rigor with which energies, forces, and stresses must be calculated can be seen in Figure 10. Even though the R^2 values for foundation models (FMs)

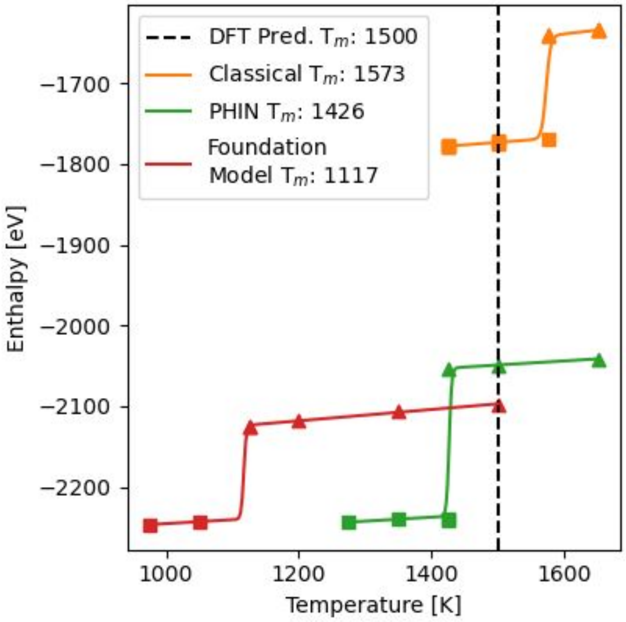

Figure 9: The melting temperature predictions predicted by classical interatomic potentials, PHIN's fine-tuned model, and an open source foundation model compared to a literature value for DFT.

are impressively high, even these small errors accumulate, leading to incorrect property predictions. Because of active learning, PHIN's models are able to eliminate even these small hallucinations yielding R^2 values greater than 0.99.

Vacancy Energy

So far, we have simulated the properties of periodic systems. We have shown that in these systems, foundation models were less accurate than fine-tuned models at predicting the elastic constants – a mechanical property – and the melting temperature – a thermodynamic property. While the previous properties were reasonably within the data distribution of foundation models, defected structures are decidedly outside of most open-source datasets. A key benefit of active learning is that we can easily extend training datasets to include these new structures and maintain high accuracy.

To show the difference of PHIN-atomic in predicting material performance for defected structures, we look at the vacancy energy and compare our performance to foundation models as well as DFT and classical interatomic potentials.

To calculate the vacancy energy in the dilute limit, we simulate the total energy of multiple structures with different vacancy concentrations and fit a linear equation for the total energy versus the number of atoms. The slope of this line is the cohesive energy and the intercept is the vacancy energy.

The data and fit for each method is shown in the left panel of figure 11 and the vacancy energies and the cohesive energies determined for each method are shown in the right two panels. We also include a model checkpoint from PHIN's active learning before fine tuning to the defect structures. The vacancy energy calculated from PHIN-atomic almost perfectly agrees with DFT for both the

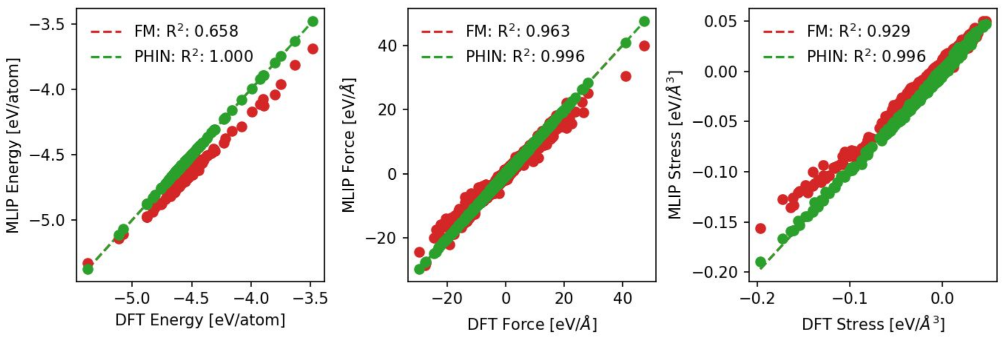

Figure 10: The energy, force, and stress parity plots for both PHIN's fine-tuned model and the foundation model tested on the fine-tuning dataset showing the exceptionally high accuracy needed for accurately predicting phase transformation temperatures.

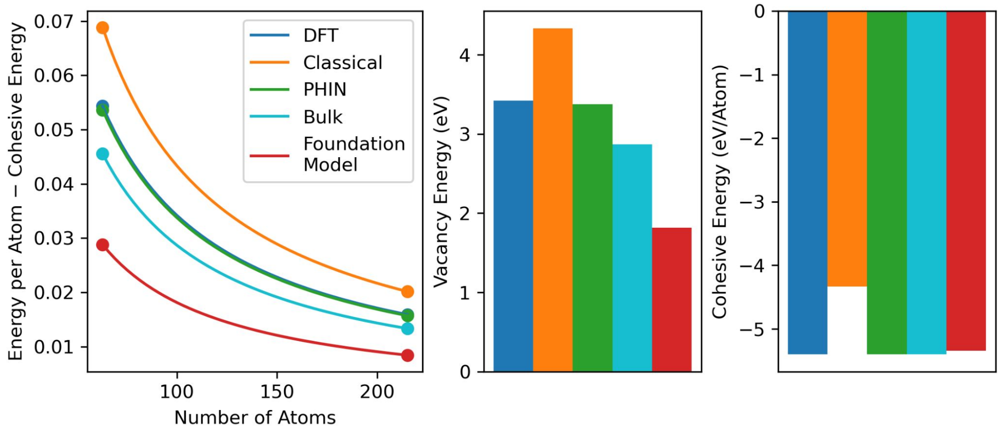

Figure 11: Vacancy energy of silicon computed at 0K from various atomic scale models. The raw energies are in the left panel with computed vacancy and cohesive energies in the center and right panels respectively from a linear model of the dilute vacancy energy.

raw data and the processed quantities, whereas the Bulk checkpoint model and the Foundation Model underestimate the vacancy energy. The Classical model over-predicts the DFT vacancy energy due to the classical potential being fit not just to DFT but also the experimental vacancy energy, which is greater than the DFT predicted value. The observation that the Bulk checkpoint is in better agreement to the Foundation model can most likely be attributed to the diversity of atomic environments the model was trained on for melted silicon, which is not present in the foundation model dataset.

PHIN-atomic's accuracy compared to DFT comes from ensuring it is interpolating on its training dataset. By showing we accurately reproduce DFT without any manual intervention, we establish confidence in the model's ability to accurately model properties to the same accuracy as DFT. When we transition to simulating the vacancy energy at room temperature, we can trust PHIN-atomic's predicted properties even though we cannot simulate it with DFT due to its high cost.

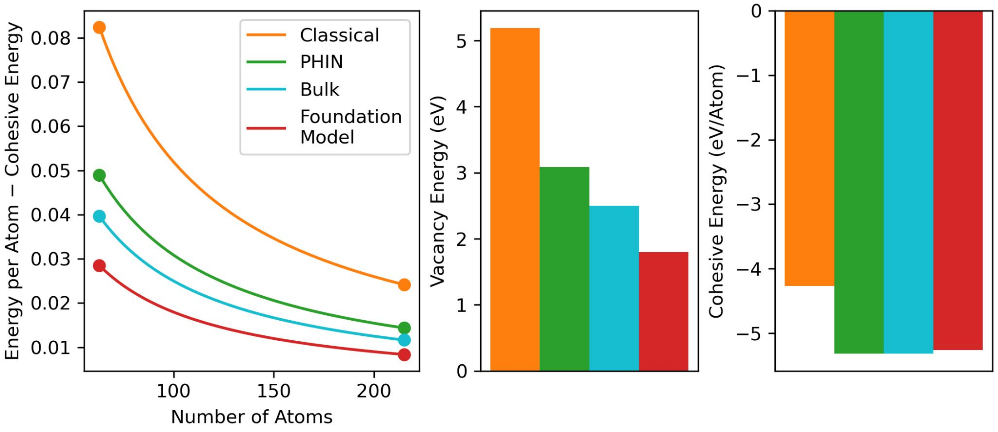

Figure 12: Vacancy energy of silicon computed at 300K from various atomic scale models. The raw energies are in the left panel with vacancy and cohesive energies in the center and right panels respectively, computed from a linear model of the dilute vacancy energy.

The diversity of predictions at 300K for the different models in Figure 12 highlights the traditional reluctance to use digital predictions to evaluate material properties. Without prior knowledge of model accuracy, given solely this figure, it is difficult to determine which model to trust. Low trust is one of the primary reasons preventing digital models from being adopted and after a scientist or company tries them once, they would probably not consider a digital model the second time. PHIN-atomic, however, builds trust with uncertainty quantification such that the model is always accurately predicting what the underlying value is, eliminating the issue of choosing an accurate model.

Surface Energy

Up to now, all the simulations have utilized periodic boundary conditions to approximate silicon infinitely repeating in space. In real-life, however, materials terminate at surfaces. The crystalline nature of materials results in a few preferential surface terminations with significantly lower energies. The lowest energy surface termination dictates the structure observed experimentally as well as how the material is faceted. For silicon, the diamond cubic crystal structure has two common surface terminations, defined by crystal lattice planes (100) and (111). These two surface terminations are shown in Figure 13. The two terminations are cleanly cut along the high symmetry crystalline directions, reducing the number of bonds that are broken.

As with the vacancy energy, we compare the energy of both of these surfaces calculated with PHIN-atomic to DFT, classical interatomic potentials, foundation models, and the bulk checkpoint fine-tuned for the periodic properties.

The 0K (111) surface energy is shown in Figure 14. The accuracy of fine-tuning models is shown clearly with PHIN-atomic and DFT being indistinguishable, where all the other models significantly underestimate the surface energy. This quality of agreement is only possible because PHIN-atomic ensures the accuracy of not just the final structure, but of all the structures throughout the relaxations. It has been shown that any uncertain structure throughout a relaxation will cause

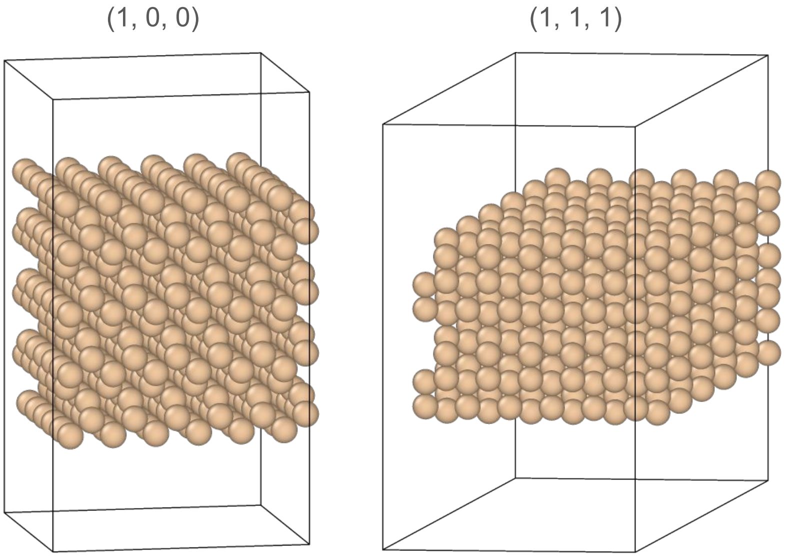

Figure 13: The two most common surface terminations of diamond silicon.

the final structure to be incorrect [cite].

Similar trends can be seen for the 0K surface energy of the (100) surface in 15. Again, DFT and PHIN-atomic are identical. The fitted curves in the left panel and the surface energy and volumetric energy in the right panels are indistinguishable. The Bulk checkpoint is the next best while the Foundation Model significantly underpredicts the surface energy. The surface energy for the classical potential is peculiarly zero. This may be because Classical potentials are restricted by a given functional form to describe an entire material's potential energy surface. It is quite likely the classical model does not have enough parameters to describe the (100) surface energy well. In particular, this classical potential contains only seven parameters that describe the potential energy surface. Tradeoffs are therefore necessary to determine which properties to fit well. Since the (100) surface is higher energy (and therefore less stable) than the (111) surface, it was most likely prioritized in the optimization.

With the accuracy established of PHIN's models at zero kelvin to match DFT values accurately, the values of the surface energies are computed at 300K Moving from 0K to 300K introduces kinetic energy to the silicon atoms. The kinetic energy allows the atoms to traverse the potential energy surface instead of remaining in the ground state configuration. The introduction of entropy into the system can have a large effect on the energetics of the system. The energies are now computed as free energies by taking an ensemble average from an NVT simulation. The ensemble average enables the thermodynamic weighting of the structures for capturing the diverse structures present in the simulation. The effect of temperature is immediately obvious when comparing the (100) predicted surface energies at 300K compared to 0K in Figure 16. All the surface energies decrease due to the configurational entropy of the surface reconstructions. Strikingly, even the classical interatomic potential becomes stable, while the surface energy predicted by PHIN-atomic decreases significantly.

The (111) surface energy (Figure 17) at 300K follows many of the same trends as the (100) surface. The configurational entropy decreases the surface energy considerably making it more stable. The surface energy predicted by PHIN-atomic decreases the most such that it is now larger

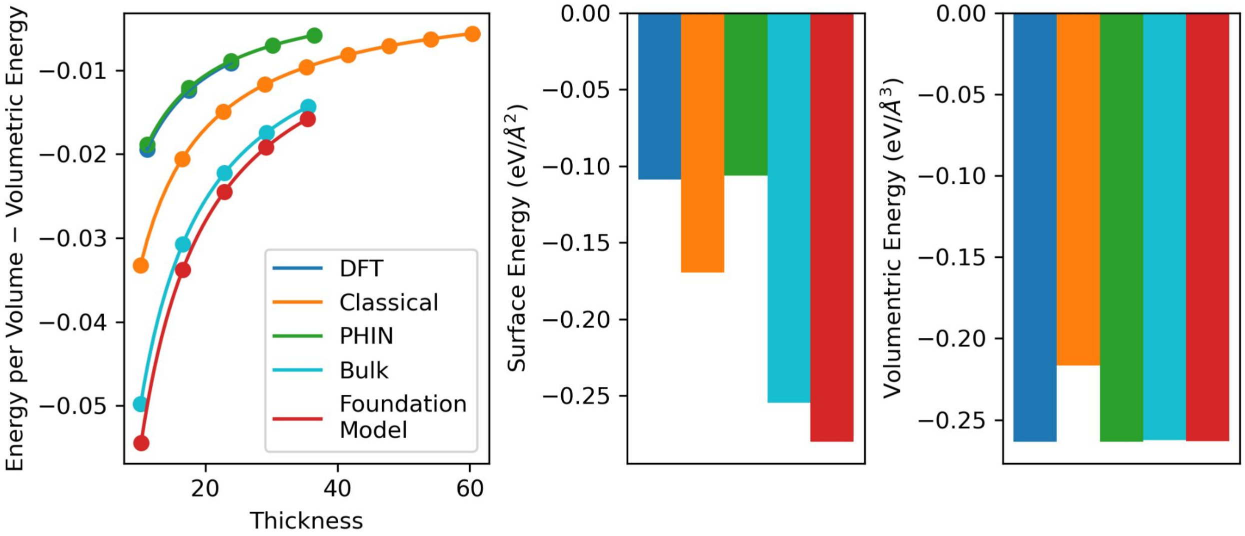

Figure 14: Zero kelvin surface energy of (111) diamond silicon. The left panel shows the difference of the total volumentric energy and the bulk volumetric energy. This data, when processed, yields the areal surface energy and volumentric energy in the center and right panes respectively.

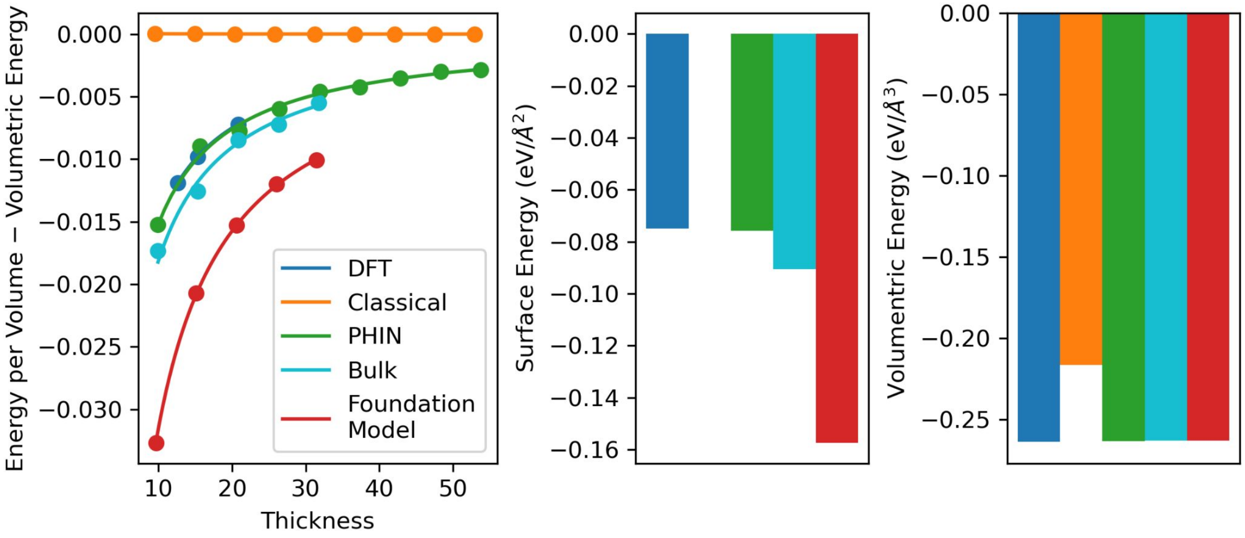

Figure 15: Zero kelvin surface energy of (100) diamond silicon. The left panel shows the difference of the total volumentric energy and the bulk volumetric energy. This data, when processed, yields the areal surface energy and volumentric energy in the center and right panes respectively.

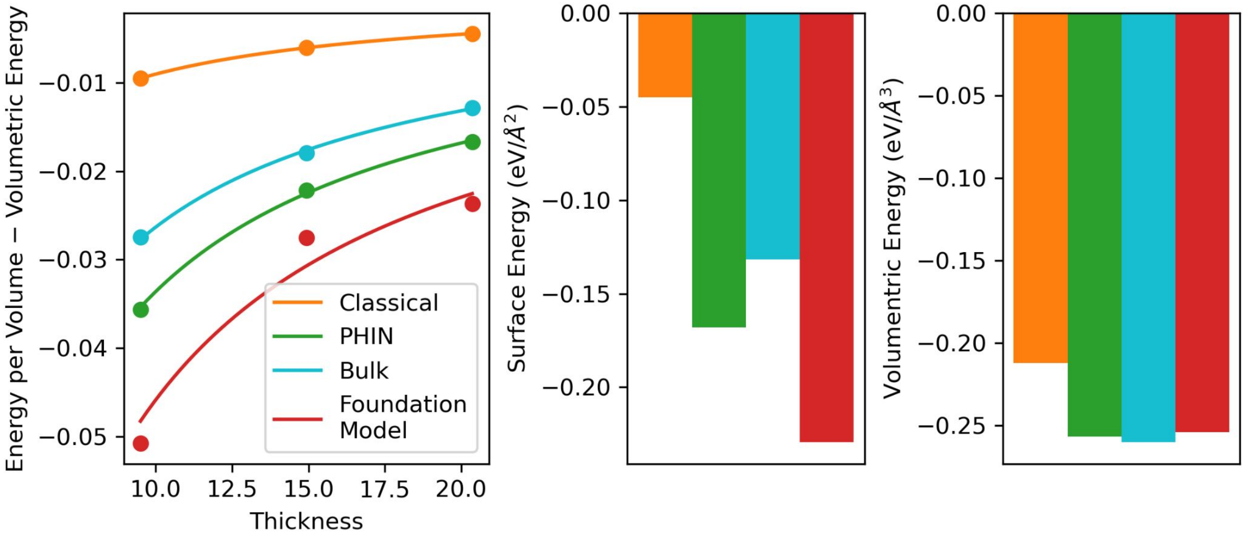

Figure 16: 300K surface energy of (100) diamond silicon. The left panel shows the difference of the total volumentric energy and the bulk volumetric energy. This data, when processed, yields the areal surface energy and volumentric energy in the center and right panes respectively.

than the surface energy predicted by the classical interatomic potential. Importantly, the values predicted by PHIN-atomic match the experimental observation that the (111) surface is more stable.

Of note is the deviation of the foundation model data points in red from the linear model for the surface energy in the left pane of Figure 16. The deviation points to hallucinations present in the model predicting different physics with different thicknesses. This deviation for just the pretrained model in addition to the baseline differences between the model surface energy predictions between the two surfaces and between the two temperatures highlight the importance of trusting the accuracy of an interatomic potential. Since DFT cannot be used to compute non-zero temperature properties due to the excessive cost of the models. PHIN's trustable machine learning interatomic potentials enable us to accurately model both surfaces across temperatures at DFT accuracy.

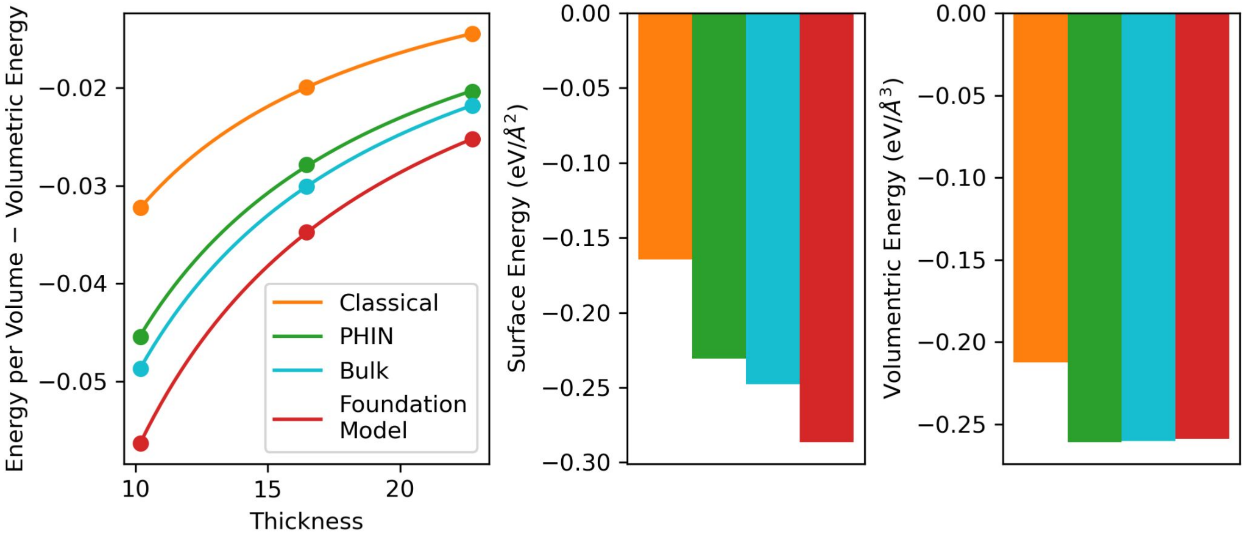

Figure 17: 300K surface energy of (111) diamond silicon. The left panel shows the difference of the total volumentric energy and the bulk volumetric energy. This data, when processed, yields the areal surface energy and volumentric energy in the center and right panes respectively.