



DISSERTATION | DOCTORAL THESIS

Titel | Title

Exploring diffusion mechanisms with machine learning

verfasst von | submitted by
Ing. Bc. Lukáš Kývala

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of
Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree
programme code as it appears on the
student record sheet:

UA 796 605 411

Dissertationsgebiet lt. Studienblatt | Field of
study as it appears on the student record
sheet:

Physik

Betreut von | Supervisor:

Univ.-Prof. Mag. Dr. Christoph Dellago

CONTENTS

ACKNOWLEDGMENTS	1
ABSTRACT	2
ZUSAMMENFASSUNG	3
LIST OF ABBREVIATIONS	5
1 INTRODUCTION	7
1.1 Molecular dynamics	8
1.2 Computational modeling methods	9
1.3 Machine learning potentials	10
1.3.1 Potential descriptors	11
1.3.2 Atomic interactions	13
1.3.3 Dataset preparation	14
1.3.4 Active learning with committee disagreement . .	15
1.3.5 Training	16
1.4 Two-Dimensional materials	18
1.4.1 Defects	19
1.5 Thesis summary	21
1.5.1 Chapter 2	21
1.5.2 Chapter 3	21
1.5.3 Chapter 4	24
1.5.4 Chapter 5	25
1.6 Final remarks and outlook	26
1.7 Publications and contributions	27
2 OPTIMIZING THE ARCHITECTURE OF BEHLER–PARRINELLO NEURAL NETWORK POTENTIALS	29
2.1 Introduction	29
2.2 Theory	30
2.2.1 High-dimension neural network potentials	30
2.2.2 Atom-centered symmetry functions	31
2.2.3 Weighted symmetry functions	32
2.3 Computational details	33
2.3.1 Dataset	33
2.3.2 Optimizer	33
2.3.3 Descriptors	34
2.4 Results and Discussion	35
2.4.1 QM9 Data Set	35
2.4.2 3BPA Data Set	40
2.5 Summary	41

3	DIFFUSION AND COALESCENCE OF PHOSPHORENE MONOVACANCIES STUDIED USING HIGH-DIMENSIONAL NEURAL NETWORK POTENTIALS	44
3.1	Introduction	44
3.2	Methods	46
3.2.1	First-principles calculations	46
3.2.2	Neural network potential	46
3.3	Results	49
3.3.1	Monovacancy	49
3.3.2	Monovacancy tracking	51
3.3.3	Kinetic properties	52
3.3.4	Divacancy and coalescence of monovacancies	55
3.4	Conclusions	58
4	ELECTRON-BEAM-INDUCED ADATOM-VACANCY COMPLEXES IN MONO- AND BILAYER PHOSPHORENE	59
4.1	Introduction	59
4.2	Results and Discussion	61
4.3	Conclusions	69
4.4	Methods	69
5	UNSUPERVISED IDENTIFICATION OF CRYSTAL DEFECTS FROM ATOMISTIC POTENTIAL DESCRIPTORS	72
5.1	Introduction	72
5.2	Results	74
5.2.1	Silicon	74
5.2.2	Hexagonal ice	76
5.3	Discussion	79
5.4	Methods	80
5.4.1	PaCMAP algorithm	80
5.4.2	Parametric PaCMAP	81
5.4.3	Datasets, atomistic potential descriptors, and PaCMAP hyperparameters	82
5.4.4	Software	83
A	SUPPLEMENTARY INFORMATION	85
A.1	Optimizing the architecture of Behler-Parrinello neural network potentials	85
A.2	Diffusion and coalescence of phosphorene monovacancies studied using high-dimensional neural network potentials	87
A.3	Unsupervised identification of crystal defects from atomistic potential descriptors	89
A.4	Phosphorene monovacancy tracking using PaCMAP	92
	BIBLIOGRAPHY	93

ACKNOWLEDGMENTS

The completion of my doctoral thesis would not have been possible without the support of my supervisor, colleagues, and family over the past four years. I am grateful to Christoph Dellago for giving me the opportunity to pursue a PhD in his research group, as well as for his unwavering support throughout my studies. I sincerely appreciate Andrea Angeletti, Cesare Franchini, Pablo Montero de Higes, Carsten Speckmann, and Jani Kotakoski for their invaluable collaboration, which has greatly enriched my academic experience. I would like to extend my gratitude to my colleagues in Christoph's research group, whose insightful feedback and comments have improved my work and broadened my perspective on various topics. I am also grateful to the Doctoral College Advanced Functional Materials (DCAFM) members for our enjoyable and productive retreats. In addition, I wish to acknowledge the administrative support provided by Claudia Rennhofer, Astrid Kellner-Windhager, Christiane Maria Losert-Valiente Kroon, and Julia Just, who have assisted me in various aspects of my doctoral studies^a.

^a I would also like to acknowledge the use of Grammarly, an advanced language tool, which has significantly contributed to improving the quality and clarity of the language in this thesis.

ABSTRACT

Defects significantly influence the properties of materials. Point defects are particularly relevant due to their potential positive impacts on applications such as the doping of semiconductors and their negative consequences, like hydrogen embrittlement of steel. However, point defects tend to diffuse, potentially altering the material’s properties over time. Therefore, it is essential to understand their diffusion behavior. Computational studies are well-suited for this task as they allow us to simulate materials at the atomic level. In this thesis, I employ machine learning potentials to accelerate molecular dynamics simulations and investigate diffusion mechanisms. Specifically, we investigated the diffusion and coalescence of monovacancies in phosphorene. This was facilitated by employing an active learning strategy that iteratively updated the neural network potential upon exploring the previously unknown potential energy surface. The same approach has been applied to study the stability and recombination of adatom-monovacancy complexes in phosphorene to support microscopic measurements. Before utilizing neural network potentials, I comprehensively analyzed the neural network’s hyperparameters to identify the optimal configuration. A key finding was that both insufficient and excessive numbers of fitting parameters can harm the accuracy of the neural network potential. The optimal number of parameters has been estimated to increase linearly with dataset size. This result is particularly relevant for active learning strategies that involve iteratively expanding the dataset. In the final stage of my research, I focused on the post-analysis of extended molecular dynamics simulations. Conducting extensive simulations is futile without extracting valuable information from them. In cases where simple order parameters are unavailable, such as tracking monovacancies, alternative methods must be employed. I have demonstrated the effectiveness of combining unsupervised algorithms with atomistic potential descriptors for distinguishing between various chemical environments with minimal computational overhead on multiple materials. In summary, this thesis uses neural network potentials to explore the diffusion mechanisms of point defects. It includes the optimal setup of hyperparameters, the construction of potentials using active learning, uncertainty control during simulations, and the post-analysis of diffusion events utilizing potential descriptors.

ZUSAMMENFASSUNG

Defekte haben einen erheblichen Einfluss auf die Eigenschaften von Materialien. Punktdefekte sind aufgrund ihrer potenziellen positiven Auswirkungen in Anwendungen wie der Dotierung von Halbleitern und ihrer negativen Folgen wie der Wasserstoffversprödung von Stahl besonders relevant. Punktdefekte neigen jedoch dazu zu diffundieren, wodurch sich Materialeigenschaften im Laufe der Zeit verändern können. Daher ist es wichtig ihr Diffusionsverhalten zu verstehen. Computergestützte Forschung ist für diese Aufgabe gut geeignet, da sie es uns ermöglicht, Materialien auf atomarer Ebene zu simulieren. In dieser Arbeit benutze ich Machine-Learning-Potentials, um Molekulardynamiksimulationen effizienter auszuführen und damit Diffusionsmechanismen zu untersuchen. Konkreter haben wir die Diffusion und Koaleszenz von Monovakanzen in Phosphoren untersucht. Dies wurde durch den Einsatz einer Active Learning Strategie ermöglicht, bei der unser sogenanntes Neural-Network-Potential bei der Erkundung einer zuvor unbekannten Potential-Energy-Surface iterativ aktualisiert wird. Dieselbe Methode wurde zur Untersuchung der Stabilität und Rekombination von Adatom-Monovakanz-Komplexen in Phosphoren angewandt, um mikroskopische Messungen zu unterstützen. Bevor ich diese Neural-Network-Potentials in Simulationen benutzte, wurden die respektiven Hyperparameter der neuronalen Netze umfassend analysiert, um die optimale Konfiguration zu ermitteln. Eine wichtige Erkenntnis hierbei war, dass sowohl eine zu geringe als auch eine zu hohe Anzahl von möglichen Parametern die Genauigkeit der Neural-Network-Potentials beeinträchtigen können. Die optimale Anzahl von Parametern nimmt schätzungsweise linear mit der Größe des Datensatzes zu. Dieses Ergebnis ist besonders relevant für Active Learning Strategien, die eine iterative Erweiterung des Datensatzes beinhalten. In der letzten Phase meiner Forschung konzentrierte ich mich auf die Nachanalyse von langzeitigen Molekulardynamiksimulationen. Die Durchführung umfangreicher Simulationen ist sinnlos wenn man nicht wertvolle Erkenntnisse aus ihnen gewinnen kann. In vielen Fällen, in denen einfache Ordnungsparameter nicht verfügbar sind, wie beispielsweise bei der Verfolgung von Monovakanzen, müssen alternative Methoden eingesetzt werden. Ich habe demonstriert, dass man mit einer Kombination aus unüberwachten Algorithmen und Deskriptoren von atomistischen Potenzialen mit minimalem Rechenaufwand zwischen verschiedenen chemischen Umgebungen in mehreren Materialien unterscheiden kann. Zusammenfassend geht es in dieser Arbeit um Neural-Network-Potentials zur Erforschung der Diffusionsmechanismen von Punktdefekten. Die Arbeit umfasst die Optimierung von Hyperparametern, die Konstruktion von Potenzialen durch Ac-

tive Learning, die Unsicherheitskontrolle während der Simulationen und die Nachanalyse von Diffusionsereignissen unter Verwendung von Potenzialdeskriptoren.

LIST OF ABBREVIATIONS

2D	two-dimensional
3BPA	3-(benzyloxy)pyridin-2-amine
3D	three-dimensional
AC	armchair
ACE	atomic cluster expansion
ACSF	atom-centered symmetry functions
AIMD	ab initio molecular dynamics
AL	active learning
aPIP	atomic permutationally invariant polynomials
BP	black phosphorus
ciNEB	climbing imaged nudged elastic band
CPU	central processing unit
DCAFM	doctoral college advanced functional materials
DFT	density functional theory
DV	divacancy
GAP	gaussian approximation potentials
GDML	gradient domain machine learning
HAADF	high angle annular dark field
HDNNP	high-dimensional neural network potential
HDNNP-MD	molecular dynamics driven by high-dimensional neural network potential
LAMMPS	large-scale atomic/molecular massively parallel simulator
MAE	mean absolute error
MD	molecular dynamics
ML	machine learning
MLP	machine learning potential
MPNN	message passing neural network
MS-EKF	multistream extended Kalman filter
MTP	moment tensor potentials

MV	monovacancy
NNP	neural network potential
PaCMAP	pairwise controlled manifold approximation projection
PBE	Perdew-Burke-Ernzerhof functional
PCA	principal component analysis
QBC	query by committee
ReLU	rectified linear unit
RMSE	root mean square error
SF	symmetry function
SH	simple hexagon
SNAP	spectral neighbor analysis potentials
SOAP	smooth overlap of atomic position
STEM	scanning transmission electron microscopy
STM	scanning tunnelling microscopy
SV	single vacancy
SVM	support vector machines
t-SNE	t-distributed stochastic neighbor embedding
TEM	transmission electron microscopy
TMD	transition metal dichalcogenides
UHV	ultra-high vacuum
UMAP	uniform manifold approximation and projection
VASP	vienna ab initio simulation package
VDSP	vienna doctoral school in physics
VSC	vienna scientific cluster
wACSF	weighted atom-centered symmetry functions
ZZ	zig-zag

INTRODUCTION

In 1929, Paul A. M. Dirac, in his seminal paper "Quantum Mechanics of Many-Electron Systems," announced¹:

"The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation."

As Dirac proposed, so it occurred. Various approximate methods of quantum mechanics (ab initio) with varying accuracy and computational complexity have been developed to describe the behavior of complex atomic systems, such as phase transitions, chemical reactions, or diffusion events^{2,3}. However, Dirac may not have anticipated that researchers would eventually seek to approximate these approximated methods likewise. The remarkable advances and achievements of machine learning in the 21st century have led researchers to explore its potential in chemistry and physics, resulting in the development of machine learning potential (MLP).

Supervised machine learning models aim to learn a function from a finite set of samples drawn from an underlying distribution. In the case of MLPs, they are primarily trained using ab initio reference data, which include both the energy of the system and the forces acting on individual atoms. If the training dataset is sufficiently large and well-sampled, the model's predictions are considered reliable within the region near the training space. The main advantage of MLPs is their significantly faster evaluation compared to ab initio calculations, which allows for the exploration of material properties in ways that were previously impractical. This capability includes studying system dynamics through molecular dynamics simulations and gaining insights into atomic-scale interactions and behavior over extended time scales. An illustrative example is the 1 μ s simulation of monovacancy diffusion in phosphorene achieved using the MLP presented in this thesis. This task would be beyond the reach of ab initio methods.

However, this increased level of approximation raises new challenges. Many open questions concern not only the machine learning model itself but also how to prepare a reference dataset and ensure the reliability of predictions. This thesis aims to provide answers to address these questions and many others.

1.1 Molecular dynamics

Molecular Dynamics (MD) is a computational method that enables detailed modeling at the atomic scale by simulating the motion of atoms and molecules. Typically, MD adopts a classical perspective, representing particles as point masses interacting through forces dependent on their mutual arrangement. While treating nuclei as classical non-quantum particles is a reasonable approximation in most situations, caution is required when simulating the movement of light atoms (such as He, H₂, and D₂) or when zero-point vibrations play a significant role, typically at low temperatures⁴.

MD is based on the premise that the nature of matter can be understood by examining the structure and motion of its constituent building blocks, with the dynamics emerging from solving the N-body problem. Due to the absence of an analytical solution for the classical N-body problem, numerical integration is necessary to update particle velocities v and positions r following Newton's second law⁵. A widely used method for numerical integration, which is also employed in this thesis, is the velocity Verlet algorithm⁴:

$$r(t + \Delta t) = r(t) + v(t)\Delta t + \frac{f(t)}{2m}\Delta t^2, \quad (1)$$

$$v(t + \Delta t) = v(t) + \frac{f(t + \Delta t) + f(t)}{2m}\Delta t, \quad (2)$$

where t is time, m is mass, and f is force.

By tracking the positions and momenta of individual particles through phase space (see Fig. 1), MD enables the calculation of various physical properties like temperature, pressure, and energy. The simulations can be run in a variety of phases, including gases, liquids, and solids, and can be used to study a wide range of phenomena, from the behavior of simple molecules like water and methane to the complex interactions between biomolecules like proteins and DNA⁴.

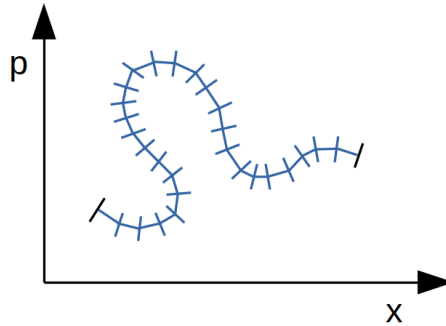


Figure 1: Molecular dynamics trajectory in phase space

Positions and velocities are updated through MD simulations (Eqs. 1 and 2), but determining the system energy and deriving forces for

a given configuration necessitate a distinct method. Historically, two primary approaches have been used: classical force fields and quantum mechanics. The third type, machine learning potentials (MLPs), has been increasingly adopted recently.

1.2 Computational modeling methods

Classical force fields employ empirical potentials to model interatomic and intermolecular interactions. This approach simplifies the quantum interactions arising from overlapping electron clouds into a system of masses connected by complex "springs." These potentials are not directly derived from quantum mechanical calculations. Rather, they are developed by adapting both structure and potential function to various sources, such as quantum mechanical calculations, experimental data (including thermodynamic and spectroscopic measurements), crystal structure analysis, measurements of transport properties, and collision studies using molecular beams, among others⁵. Originally proposed for liquid argon, the Lennard-Jones potential is the most well-known pairwise potential⁵:

$$u(r_{ij}) = \begin{cases} 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^6 \right], & \text{if } r_{ij} < r_c, \\ 0, & \text{if } r_{ij} \geq r_c. \end{cases} \quad (3)$$

For a pair of atoms i and j located at $\mathbf{r}_i$ and $\mathbf{r}_j$, the potential energy depends on the distance between them, $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. The parameter ϵ governs the strength of the interaction, while σ defines the length scale. The interaction is repulsive at short distances, becomes attractive as the atoms move apart, and is ultimately cut off at a separation of r_c . Classical force fields are mostly employed when the quantum-mechanical approach is computationally expensive (e.g., biomolecules) or when the system under study does not require a high level of accuracy or is relatively simple, such as in the case of argon⁶.

The quantum-mechanical approach (ab initio) describes the electronic structure of a material using wave functions resulting from Schrödinger's equation, which accounts for correlations between electrons and their interactions with nuclei. The quantum-mechanical approaches primarily rely on the Born-Oppenheimer approximation, which assumes that the wave functions of atomic nuclei and electrons can be treated separately, given that nuclei are much heavier than electrons. Consequently, the electronic wave function is calculated for a fixed configuration of nuclear coordinates. One widely adopted method within this framework is Density Functional Theory (DFT)⁷. DFT is founded on the two Hohenberg-Kohn theorems. The first theorem asserts that the total energy E of a system of interacting electrons in

an external potential (such as the Coulomb potential of nuclei) is a functional of the ground state electronic density ρ ⁸:

$$E = E[\rho(\mathbf{r})]. \quad (4)$$

The second theorem states that the true electronic ground state density is obtained by minimizing this energy functional⁸:

$$\left. \frac{\partial E[\rho]}{\partial \rho} \right|_{\rho_0} = 0. \quad (5)$$

However, the exact form of $E[\rho]$ is unknown, so the accuracy of DFT depends on the used approximations.

DFT-based methods, while less computationally intensive than other ab initio techniques⁹, still require significant computational resources for large-scale simulations or high-throughput screening applications. In stark contrast, classical force fields necessitate only a fraction of the computational power required by DFT, but this comes at the cost of limited accuracy. Historically, researchers faced a trade-off between computational complexity and accuracy. The emergence of MLPs has overcome this dichotomy¹⁰.

Trained on high-accuracy energies and forces obtained from DFT, MLPs deliver high accuracy at a fraction of the computational cost associated with ab initio methods¹¹. MLP has enabled accurate large-scale simulations that were previously unfeasible or excessively computationally expensive¹². Nevertheless, new challenges have emerged. The primary challenge is obtaining high-quality reference data for training these models.

1.3 Machine learning potentials

MLPs represent a third class of computational modeling approaches to predict interatomic interactions in various systems. MLPs are similar to classical force fields as they fit a potential function primarily using ab initio data. The key distinction lies in how the potential function is developed: in classical force fields, the structure and potential function are hand-crafted using expert knowledge, whereas in MLPs, the potential function is derived directly from the data during training. This flexibility allows MLPs to make much more accurate predictions within the range of the training data. However, this same flexibility can lead to potentially nonsensical results when applied to configurations vastly different from the training data.

The development of MLPs has been driven by advances in machine learning techniques, particularly those based on neural networks and kernel methods. These algorithms are well-suited to handle interatomic interaction's complexity and high-dimensional nature, which can be difficult to capture using traditional potential energy functions. There are several essential attributes that MLPs should possess¹³:

- Accuracy in prediction: The potential should achieve the highest possible agreement with reference data from quantum mechanical calculations or experimental results.
- Transferability to new data points: The ability to perform reliably across a diverse range of atomic configurations, particularly those not included in the potential's training.
- General applicability to various systems: The potential should have a functional form that does not include system-specific terms, ensuring broad applicability.
- Reactivity: The capability to accurately model the making and breaking of bonds.
- Computational efficiency and scalability: The potential should be efficient and scalable, enabling fast evaluation of potential energy even for large systems.
- Data efficiency: The potential should require the smallest possible number of reference data points to determine its parameters.

The following chapters explore the cornerstones of machine learning potentials, starting with potential descriptors, followed by the types of atomic interactions, dataset preparation, model construction with active learning and committee disagreement, and concluding with the training of MLPs.

1.3.1 Potential descriptors

The success of MLPs depends on transforming atomic configurations into input features for machine learning models. Using Cartesian coordinates is unsuitable due to the lack of invariances. Several properties should be considered to construct effective descriptors¹⁰:

- Translation invariance: the descriptor should remain unchanged regardless of any shifts in the coordinate system.
- Rotation invariance: the descriptor must remain unchanged despite any rotations of the coordinate system.
- Permutation invariance: the order of atomic indices should not affect the descriptor.
- Uniqueness: each descriptor, or at least a combination of descriptors, should be uniquely tied to a specific atomic environment.
- Compactness: descriptors should aim to condense as much information as possible and avoid redundancy.

- Computational efficiency: calculating the descriptor should be significantly less resource-intensive than the reference method.

The first descriptors meeting the listed criteria were Behler-Parrinello Atom-Centered Symmetry Functions (ACSF)¹¹. This approach transforms Cartesian coordinates into a set of hand-crafted symmetry functions to represent local chemical environments around a central atom in a specific cutoff. Symmetry functions are radial and angular types. The radial type captures the radial distribution of neighboring atoms around the central atom, and the angular type captures the angular distribution of neighboring atoms, considering the angles formed between pairs of neighbors and the central atom. A modified version called weighted atom-centered symmetry functions (wACSF)¹⁴ overcomes the undesirable scaling of ACSF with an increasing number of different elements in a material. An alternative approach that does not require hand-crafted descriptors like ACSF is the Smooth Overlap of Atomic Positions (SOAP)¹⁵. This method combines radial basis functions and spherical harmonics, which are then transformed into bispectrum coefficients to ensure symmetry invariance. Spectral Neighbor Analysis Potential (SNAP)¹⁶ differs from SOAP by assuming a linear relationship between atom energy and bispectrum components. Each atom’s local environment is represented through bispectrum components of local neighbor density, projected onto a hyperspherical harmonic basis in four dimensions. Moment Tensor Potential (MTP)¹⁷ uses moment tensors as an alternative to spherical harmonics for describing atomic environments, distinguishing it from methods like SOAP and SNAP that use spherical harmonic expansions. Finally, Atomic Cluster Expansion (ACE)¹⁸ provides a systematic framework for constructing higher-body-order complete polynomial basis functions. Higher-body-order terms are feasible due to the constant cost per basis function, which remains independent of the body order. Other atomic description-based approaches can be viewed as special cases of ACE, such as MTP, which is ACE formulated in Cartesian space¹⁹.

In addition to atomic description-based methods, researchers have also employed graph neural networks (GNNs). GNNs represent atomic configuration as graphs with atoms as nodes and apply message-passing operations to learn atomic environment representations. Pioneering works utilizing graph architecture were the DTNN²⁰, SchNet²¹, HIP-NN²², PhysNet²³, or DimeNet²⁴. However, these models did not improve upon atomic description-based approaches regarding computational efficiency and accuracy on public benchmarks²⁵. Furthermore, most of those GNN interatomic potentials use 2-body invariant messages, making them non-universal approximators²⁶. Recently, a new class of GNNs has been developed that utilizes equivariant internal features. Notable examples of these networks include Cormorant²⁷, Tensor Field Networks²⁸, EGNN²⁹, PaiNN³⁰, Equivariant Transformers³¹, SEGNN³², NewtonNet³³, and NequIP³⁴. These models have

outperformed atomic description-based approaches but face two significant limitations: the most accurate models rely on $L = 3$ spherical tensors as messages and require four to six message passing iterations, leading to relatively high computational costs and challenges in parallelizing across multiple GPUs²⁵. This bottleneck has been addressed by the MACE architecture, which requires only two message-passing iterations to achieve high accuracy, in contrast to the typical four or six iterations used by previous GNNs²⁵. This results in a potential that is both accurate and reasonably computationally efficient.

Although modern approaches like MACE offer higher accuracy than the original ACSF, ACSF remains in use, including in this thesis. This continued use is attributed to its extensive historical computational development, well-tested nature, simplicity, and computational efficiency, which allow it to be used on CPUs.

1.3.2 Atomic interactions

Body-order expansion models typically rely on nearsightedness, representing total energy E as a sum of atomic energies E_i :

$$E = \sum_i E_i(x_i). \quad (6)$$

Atomic energies are only dependent on the local chemical environment up to a cutoff radius¹¹ as interactions beyond this cutoff are assumed to be negligible.

However, long-range electrostatic and dispersion forces cannot always be neglected due to their significant impact on certain materials, e.g., ionic crystals³⁵. These interactions can be addressed by predicting atomic charges using local descriptors and subsequently calculating the long-range electrostatic energy using Ewald summation^{36,37}.

Relying solely on local chemical environments fails to capture long-range charge transfer and different charge states of a system. This limitation can be overcome by determining local environment-dependent electronegativities instead of charges. These electronegativities are then used for global charge equilibration, and the resulting charges incorporate charge transfer effects³⁸.

Incorporating long-range interactions is not always advisable, as it increases the complexity of the model. Before constructing MLPs, it is crucial to consider the relative importance of each interaction type. Excluding long-range interactions is a reasonable approach in scenarios where electrostatic interactions are negligible compared to short-range interactions. This was the case in my research, which focused exclusively on short-range interactions for studying defects in phosphorene.

1.3.3 Dataset preparation

Reference data are another crucial component of all ML models. Regardless of the sophistication of the architecture or feature extraction methods employed, poorly prepared reference data can hinder model performance. Fortunately, since our applications rely on ab initio generated data, noise is not typically a significant concern. If data corruption occurs during data generation, it is relatively straightforward to identify and remove problematic data points before training by implementing outlier detection measures or after training by analyzing parity plots for energies and forces. Notably, biased and misdistributed datasets pose a more substantial challenge.

The five primary methods of data collection, each suited to distinct purposes, are the following¹⁰:

- **Open-source material databases** provide a rich source of ab initio calculations. Utilizing open-source databases offers a highly effective approach for gathering relevant data for MLPs with minimum effort. Although these databases are extensive, they typically focus on pristine materials, limiting their applicability for studying diffusion mechanisms due to the lack of data on thermal fluctuations, defects, and transition states.
- **Random sampling** generates distorted structures by displacing atoms randomly or along normal modes. These displacements are drawn from uniform or Gaussian distributions, with the amplitude as a parameter. Atoms are typically displaced from a local minimum on the potential energy surface, but they can also be displaced from non-equilibrium configurations, such as those obtained during MD simulations. This method is efficient for exploring PES but is limited to regions close to starting configuration and is best combined with other sampling methods.
- **Ab Initio Molecular Dynamics (AIMD) sampling** employs MD driven primarily by DFT to sample the potential energy surface. AIMD's advantages lie in its ability to sample the phase region of interest. However, AIMD is typically limited by the picosecond sampling timescale, which is usually too short to capture all diffusion events. Consequently, this technique is suitable only for studying equilibrium or near-equilibrium properties, such as vibrational spectra or thermodynamic properties, or for generating an initial dataset for active learning sampling.
- **Active Learning (AL) sampling** starts with a small initial set of reference data to train a preliminary MLP. This potential is then employed in MD simulations, gathering additional configurations when model predictions become unreliable based on an uncertainty criterion. Once several additional configurations are

collected, the model is retrained with these new data points. This iterative process continues until the relevant potential energy surface is well-sampled. This strategy leverages AIMD to sample the desired phase space while removing restrictions imposed by computation complexity. As most timesteps involve MLP, ab initio calculations are required only for new configurations identified by uncertainty criterion.

- **Metadynamics sampling** biases the dynamics to visit unexplored regions on the potential energy surface. Although a powerful technique, it presents several challenges in setting up appropriate collective variables and parameters³⁹.

In conclusion, each data collection method has its applications. The open-source material databases provide a valuable data source, primarily for pristine materials. Random sampling effectively explores the potential energy surface but does not capture system dynamics, as the structures are generated without considering time evolution. Due to its short simulation time, AIMD sampling is restricted to equilibrium or near-equilibrium properties. AL and metadynamics sampling can be employed to sample the desired potential energy surface while reducing the need for ab initio calculations. However, they differ in complexity: active learning can be implemented without prior knowledge, while metadynamics requires some prior understanding.

1.3.4 Active learning with committee disagreement

AL is a subfield of machine learning based on the principle that the learning algorithm can actively choose the data from which it learns⁴⁰. AL systematically collects diverse data points to correct shortcomings in a machine learning model’s predictions. This process relies on accurately assessing the model’s uncertainty to guide the selection. A well-established practical strategy for AL with neural network potentials is called query by committee, where the estimate of uncertainty is the disagreement between a collection of models within an ensemble^{41,42}. Typically, there are five to ten neural networks in an ensemble, and these share the same architecture and hyperparameters but use a different initialization of the model parameters for training and different splits of the training/validation data. It is empirically observed that the variance of the ensemble predictions correlates with actual prediction error⁴³. In the query by committee strategy, the training set will be augmented with new ab initio data if the ensemble variance is large. AL strategy workflow is summarized in Fig. 2.

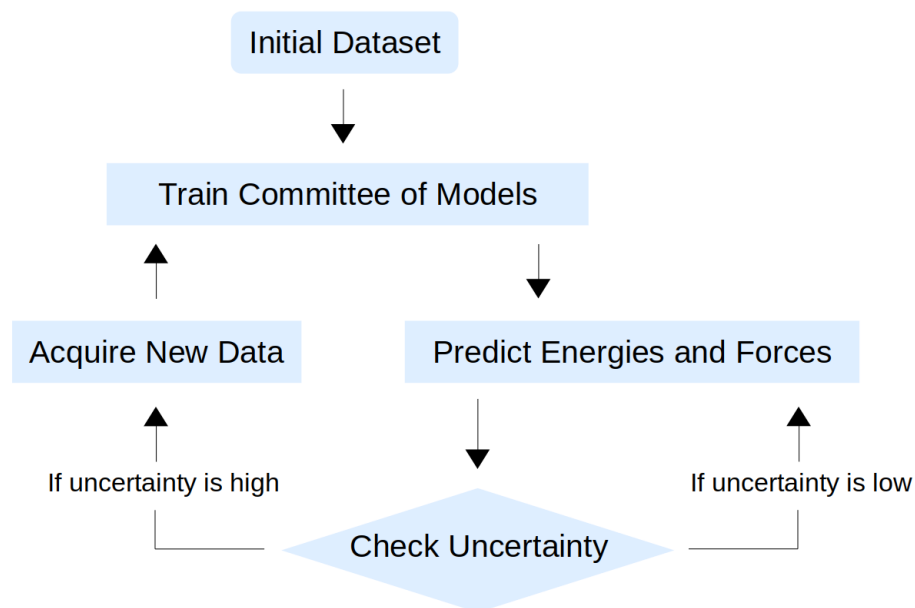


Figure 2: The flow diagram of active learning with query by committee

1.3.5 Training

Once the potential descriptors are established and reference data have been obtained, the next step is to train the model. First, basic terms related to training are reviewed⁴⁴, followed by an overview of the training process.

Basic terms

Loss function is a measurable way to evaluate the performance and accuracy of a machine learning model. The most common types of loss functions include mean square error (MSE) and root mean square error (RMSE), where $MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and $RMSE = \sqrt{MSE}$. Here, y_i represents the model's prediction, $\hat{y}_i$ is the reference value, and n is the number of samples.

Optimizers are algorithms that adjust learnable model parameters to minimize the loss function. Commonly used optimizers are based on gradient descent, including stochastic gradient descent⁴⁵ and ADAM⁴⁶. However, for small models, more advanced algorithms like L-BFGS⁴⁷ or extended Kalman filter⁴⁸ may be beneficial due to their superior performance compared to commonly used ADAM^{49,50}. These advanced algorithms are restricted to small models because their memory and computational requirements scale poorly with the number of learnable parameters.

Hyperparameters are settings that control the optimizer's behavior and are typically decided before training, such as batch size, the number of training epochs, and model architecture. The hyperparameters are often optimized by performing hyperparameter fine-tuning.

Batch size, a hyperparameter, defines the number of data points to work through before updating the model parameters. The trade-off between accuracy and speed can be understood in terms of the batch size. Large batch sizes result in faster training times but may lead to lower accuracy and overfitting. In contrast, small batches deliver better general performance due to the noise they add to the learning process, which prevents getting stuck in local loss minima.

Epoch is defined as the number of training iterations over the entire training dataset.

Learning curve plots the evolution of training and validation/test loss values as a function of the number of epochs or training set size. Fig. 3 illustrates an example of a learning curve for a number of epochs.

Neural network is an analytical function $y = f(\mathbf{x}; \theta)$ that maps an input $\mathbf{x}$ to an output y and learns the values of the parameters θ that result in the best function approximation. Neural networks are called networks because they are usually composed of multiple interconnected functions. For instance, we might have three functions $f^{(1)}$, $f^{(2)}$, and $f^{(3)}$ connected in a chain of layers such that $y = f^{(3)}(f^{(2)}(f^{(1)}(\mathbf{x})))$. Layers that act only as inputs to other layers without producing the final output y are called hidden layers. In previous example, $f^{(1)}$ and $f^{(2)}$ are hidden layers, while $f^{(3)}$ is the output layer. In a simple feedforward neural network, a layer f is $g(\sum_i^m w_i x_i + b)$, where w_i are learnable weights and b is a learnable bias. The non-learnable hyperparameters include m , the number of nodes in a layer, and g , the predetermined activation function. Activation functions are generally nonlinear for hidden layers and linear for the output layer.

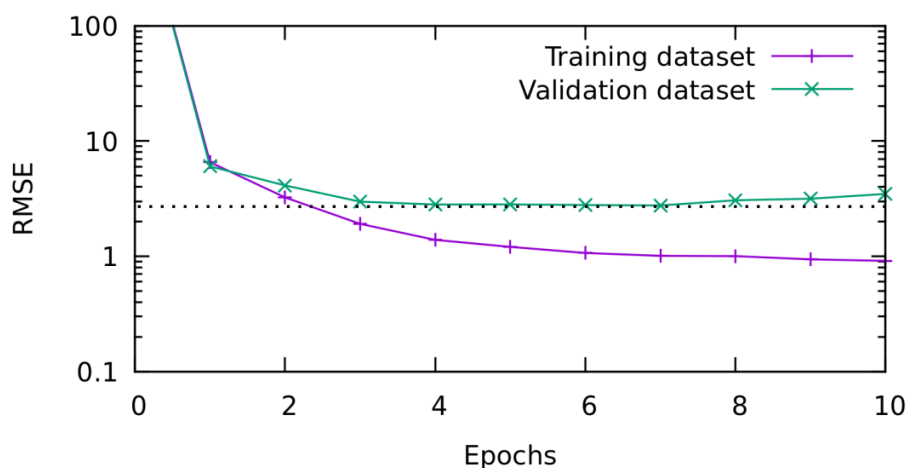


Figure 3: The learning curve shows how RMSE changes over the number of iterations (epochs) during training. Starting from the seventh epoch, the model's accuracy on the validation dataset begins to decrease. This learning curve resulted from training an MLP on the QM9 dataset, as described in Chapter 2.

Training process

The training process starts with data processing, which involves cleaning and normalizing the data. Data cleaning eliminates invalid data points and irrelevant outliers, whereas normalization adjusts the data to a standard scale for improved convergence during training. The primary goal of data processing is to ensure that all data used in the training are reliable and consistent.

A dataset is not entirely used for training. Instead, it is divided into training, validation, and testing sets. Most of the dataset (80-90%) comprises the training set, which is available to the optimizer to adjust model parameters to minimize the training loss function. In contrast, the validation dataset is not accessible to the optimizer but is evaluated after each epoch to detect the overfitting of training data. If the training loss decreases while the validation loss increases, the optimizer starts to memorize the training dataset instead of generalizing it. A common practice is to use early stopping, where the final model is selected based on the epoch that achieves the highest accuracy on both the training and validation datasets. In Fig. 3, this corresponds to epoch 7. The testing dataset is only evaluated once when the final model is selected, providing an unbiased metric of the model's ability to generalize.

Data cleaning can only detect obvious errors and outliers but may not capture more subtle ones. Typical cases include badly converged ab initio calculations with incorrect energies and forces, although within ranges of other data points, or systematic errors resulting from altered hyperparameters in ab initio calculations. Corrupted data can be identified using parity plots, which visualize the comparison between a machine learning model's predictions and benchmark data. Corrupted data points do not typically align with the rest of the dataset in these plots.

1.4 Two-Dimensional materials

Two-dimensional (2D) materials are crystalline structures consisting of a single layer of atoms. The discovery of graphene by Geim and Novoselov in 2004 marked the beginning of the 2D material era⁵¹. Since then, these materials have attracted significant attention due to their unique physical and chemical properties, which arise from quantum size effects and differ greatly from those of their 3D bulk counterparts⁵². For instance, graphene exhibits unique properties such as giant carrier mobility⁵³, quantum Hall effect⁵⁴, large Young's modulus⁵⁵, and superior thermal conductivity⁵⁶. Besides graphene, various 2D materials with unique properties have been intensively studied, including black phosphorene⁵⁷, transition metal dichalcogenides/oxides^{58,59}, and transition metal carbides/nitrides/carbonitrides (MXenes)⁶⁰. The wide variety of available 2D materials provides various electronic prop-

erties, including metals, semimetals, insulators, and semiconductors with direct and indirect bandgaps, covering the electromagnetic spectrum from ultraviolet to infrared⁵².

2D materials can be used either as standalone sheets or stacked to form van der Waals heterostructures. While strong covalent bonds provide in-plane stability, relatively weak van der Waals forces are sufficient to hold the stack together. Careful engineering of stacking layers can yield properties and functionalities not found in standard materials⁶¹.

However, whether used as standalone layers or in heterostructures, 2D materials exhibit various defects that significantly affect their chemical, electronic, optical, magnetic, and mechanical properties. This has led to extensive research on understanding these defects, including tailoring them for specific applications⁶².

1.4.1 Defects

The second law of thermodynamics dictates the equilibrium defect concentrations in crystalline materials⁶². However, defect concentrations are usually higher in real materials due to their broad surfaces. A significant number of non-equilibrium defects are introduced during sample preparation. Additionally, irradiation with electrons or ions can create point defects through the ballistic ejection of atoms. Chemical reactions like oxidation can also lead to atom loss and further defects⁶³.

Defects in materials appear in various forms and are categorized by the extent of atomic disorder: point defects, line defects, planar defects, and volume defects. Point defects, such as vacancies or interstitial atoms, are zero-dimensional. Dislocations, on the other hand, are one-dimensional lines of defects. Grain boundaries or stacking faults extend in two dimensions, while inclusions and voids are three-dimensional. The reduced dimensionality of 2D materials limits the types of defects that can form. In 2D materials, point defects are similar to those in bulk crystals, but line defects play a distinct role as they separate domains with different crystal orientations. Indeed, three-dimensional defects do not exist in 2D materials⁶³.

Research of point defects is especially important in 2D materials due to their profound impact on material properties. On the negative side, they can reduce mechanical strength⁶⁴, act as scattering centers for electrons, leading to lower charge mobility⁶⁵, and interfere with lattice vibrations, resulting in decreased thermal conductivity⁶⁶. On the positive side, their reactivity can be leveraged for sensing applications⁶⁷, tuning electronic properties for electronic and optical uses⁶⁸, and even introducing magnetic properties⁶³.

A point defect involves a disordered arrangement of atoms near a lattice node, typically within a range of one to several lattice constants. Point defects are generally classified as intrinsic or extrinsic, depending

on their origin⁶². Intrinsic defects are deviations from the crystal's ideal structure or composition without introducing atoms or ions other than those already present⁶⁹. In contrast, extrinsic point defects result from the presence of different chemical elements in the crystal structure⁷⁰.

The simplest point defect is the **Stone-Wales defect**, which involves a 90° bond rotation without removing or adding atoms with fully saturated bonds⁶³.

Vacancies occur when one or more atoms are missing from a lattice site. Typically, vacancies with an even number of missing atoms are energetically favored, allowing for complete reconstruction and the saturation of dangling bonds. In contrast, vacancies with an odd number of missing atoms leave an open bond, making them less energetically stable^{62,71}.

Interstitial atoms arise when an atom occupies a normally unoccupied position within the crystal structure. In 2D materials, interstitials differ from those in bulk materials, as placing an atom in an in-plane position requires prohibitively high energy. Instead, additional atoms tend to use the third dimension and adhere to the surface⁶³. These adatoms, which may be either the same or a different element from the host material, interact with the lattice based on the strength of the bond. If the bond is weak, physisorption occurs via van der Waals interactions. If the bond is stronger, covalent bonding leads to chemisorption⁶³. Interstitials are often formed when an atom is displaced from its lattice site, leaving behind a vacancy.

Substitutional impurities occur when a lattice site is occupied by a different chemical element than the one it originally contained. For stable incorporation, the heteroatom must have a compatible size and electron configuration with the host lattice⁶³.

1.5 Thesis summary

This thesis focuses on understanding diffusion mechanisms in defective phosphorene using accelerated molecular dynamics driven by high-dimensional neural network potentials (HDNNPs) utilizing ACSFs¹¹. HDNNPs were constructed through active learning with queries by the committee. Chapter 2 focuses on the optimization of HDNNP hyperparameters. Chapters 3 and 4 are dedicated to dynamically investigating diffusive mechanisms in phosphorene, while Chapter 5 is dedicated to extracting information from long MD simulations using potential descriptors.

1.5.1 Chapter 2

Chapter 2 explores optimizing the architecture of HDNNPs. This chapter serves as a foundation for subsequent work, highlighting the importance of hyperparameters, such as the number of nodes, the number of hidden layers, the type of activation functions, and the size of the feature vector.

A key takeaway from this chapter is that optimal architecture relies on training dataset size. Although neural network architecture does not directly depend on training set size, I demonstrate that increasing the number of learnable parameters with larger datasets is necessary. Conversely, architecture that is too complex can be detrimental to small datasets. The findings are particularly crucial for active learning strategies, which often rely on fixed hyperparameters throughout a project. Such an approach may hinder the accuracy of potential, as demonstrated. Another finding is that unbounded symmetry functions outperform bounded ones in HDNNPs when working with small datasets. However, this advantage diminishes as the training set size increases. Additionally, I highlight the importance of correctly setting the number of descriptors describing local chemical environments. Too many or too few descriptors can lead to suboptimal results due to data sparsity and low information resolution.

1.5.2 Chapter 3

While Chapter 2 focuses on correctly setting up models, Chapter 3 uses HDNNPs to study diffusive events. This work forms the core of the thesis, as subsequent chapters are closely connected and build upon the findings presented here.

The primary focus of the thesis is on defective phosphorene, a two-dimensional monolayer of black phosphorus with unique electronic properties. Phosphorene is a promising semiconductor with a thickness-dependent electronic band gap (ranging from 2.2 eV in the monolayer to 0.3 eV in bulk)^{72,73}, high carrier mobility⁷⁴⁻⁷⁷, and in-plane an-

isotropy. When a monovacancy (MV) is introduced to phosphorene, the localized magnetic moment emerges from an unsaturated bond. Magnetic semiconductors are of significant interest in the field of spintronics⁷⁸. Despite theoretical predictions that MVs should be the most common defect after electron and ion irradiation^{79,80}, experimental studies have observed a predominance of divacancies and tetravacancies^{81–83}. Our study aims to explain this discrepancy.

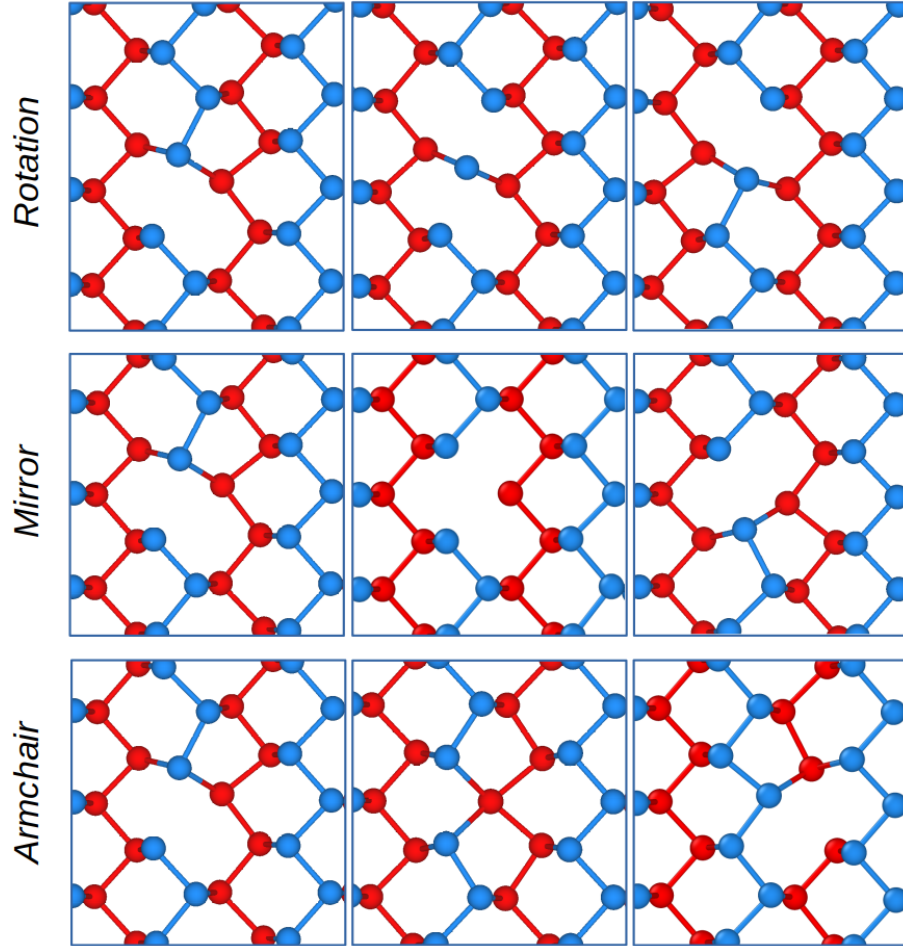


Figure 4: *Rotation* (top), *mirror* (middle), and *armchair* (bottom) hopping mechanisms in phosphorene as described by Cai et al.⁸⁴

This disagreement between theoretical and experimental studies can be explained in two ways: (1) by the rapid merging of MVs into more stable divacancies and (2) that MVs are too mobile and therefore not observable with electron microscopy, as suggested by previous research by Cai et al.⁸⁴.

Cai et al.⁸⁴ identified three distinct hopping mechanisms: *rotation*, *mirror*, and *armchair*, as shown in Fig. 4. Only the *rotation* mechanism has been validated by MD simulations. The authors calculated energy barriers for all three mechanisms, but the energy barriers alone are insufficient to determine the hopping rates ν , as the prefactor ν_s in the Arrhenius equation $\nu = \nu_s e^{-E_b/k_B T}$ is also needed.

The primary advantage of MD is its ability to calculate hopping rates ν directly from counting events over elapsed simulation time. For this task, we collaborated with Andrea Angeletti and Cesare Franchini, colleagues from Doctoral College Advanced Functional Materials (DCAFM). Work was divided such that I focused on HDDNPs and accelerated MD while Andrea generated DFT calculations using AIMD. We followed the active learning workflow presented in Fig. 2. During initial dataset generation, Andrea observed only the *rotation* mechanism using AIMD. Consequently, the other two mechanisms were not sampled in the first version of HDNNPs.

When the committee of HDNNPs approached the *mirror* transition during the MD simulation, the uncertainty became too high, leading to the termination of the simulation. However, it provided an ideal starting configuration for AIMD simulations, allowing us to observe the *mirror* transition with DFT for the first time. Having sampled the *mirror* transition with several AIMD simulations, HDNNP-MD simulations could also simulate this hopping mechanism correctly, and the search continued. After some time, the *armchair* mechanism was also found and validated by DFT. Notably, the active learning loop also included many configurations beyond those associated with hopping mechanisms.

With a well-established model, I conducted multiple MD simulations with varying temperatures totaling 1 μ s in cumulative simulation time. I observed only the three hopping mechanisms previously reported during this simulated period. Although our simulations reproduced the energy barriers, MD simulations revealed slightly different mechanisms for *mirror* and *armchair* hopping compared to previous work⁸⁴. Post-analysis further revealed that *rotation* hopping occurs on a picosecond timescale, whereas *mirror* and *armchair* hopping occurs on nanosecond timescales. It confirms monovacancies' strong itinerant nature and inability to be measured with current electron microscopy techniques.

To conduct post-analysis, I developed a new method to track the movement of MVs. MVs are not explicitly visible in atomic coordinates, and manual tracking was impractical due to the vast number of configurations. Instead, I utilized atomic energies as a proxy. The magnetic atom with missing bonds exhibits significantly higher atomic energy than its neighbors, being a reliable indicator of MV presence. Since each hopping event changes the proxy atom's index, MVs can be easily tracked by monitoring atomic energies.

The coalescence of monovacancies has also been explored through MD simulations accelerated by HDNNP. Although phosphorene is only a 2D structure, there are still numerous possibilities for MVs to approach each other. The complexity of the task makes a DFT study infeasible, which explains why no prior studies have investigated coalescence in phosphorene. Accelerated MD with HDNNPs allowed me to explore all relevant scenarios, identifying three distinct merging

mechanisms and metastable states arising from specific mutual MV orientations.

In summary, our findings indicate that MVs in phosphorene are unstable and cannot be employed at room temperatures. Their high mobility and low coalescence barrier make them susceptible to merging into more stable divacancies.

1.5.3 Chapter 4

Chapter 4 presents a second successful application of active learning combined with the query by committee technique, denoting its potential for exploring diffusive events. In Chapter 3, we demonstrated the mobility and tendency of MVs in phosphorene to coalesce. Experimentalists within DCAFM have successfully detected the stable MV-adatom complex using electron microscopy (Fig. 5). Our goal was to investigate whether the adatom can stabilize MV and understand the possible mechanisms underlying recombinations or hopping.

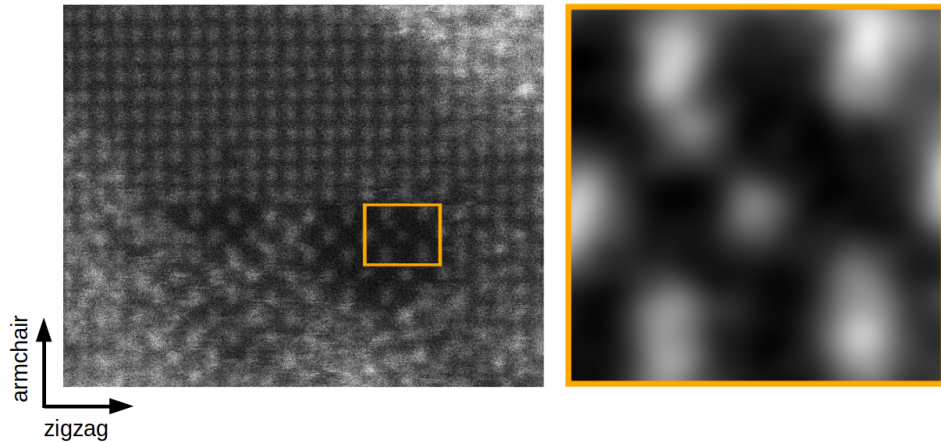


Figure 5: Adatom-vacancy complex in monolayer phosphorene. The right image is a zoomed-in view of the area indicated by the orange rectangle in the left figure.

We initiated multiple ab initio MD simulations at 300 K, applying equilibrium conditions and isotropic and anisotropic strains/stresses at 1 % and 3 %. These initial simulations revealed a hopping mechanism under anisotropic stress of 3 % in a zigzag direction. To identify additional hopping mechanisms and potential recombination processes, I trained HDNNP models and conducted multiple MD simulations powered by HDNNPs. This approach revealed the recombination process.

Our results indicate that recombination occurs in each tested strain and stress condition at levels of 1% and 3% within a few nanoseconds at 300 K, suggesting a low energy barrier. The discrepancy between experimental and computational studies may be attributed to an effect not considered in our simulations, such as the energy received by the monolayer during electron measurements or monovacancy bending. Figure

5 also shows that the MV-adatom complex is located near the bilayer interface, potentially leading to bond reorganization. Unfortunately, only one case of this stable MV-adatom complex has been observed, and we cannot rule out interactions with the bilayer. Therefore, the origin of stability for this complex remains unclear and requires further investigation.

1.5.4 Chapter 5

Running a lengthy simulation without the ability to extract valuable information is worthless. Thus, it is essential to identify the information of interest and choose methodologies to acquire it before starting a simulation. Extracting diffusion paths may be straightforward in certain scenarios, such as monitoring impurities in pristine structures. However, tracking monovacancies can be challenging due to the absence of a direct method. Although a novel approach was developed for tracking hopping using atomic energies in Chapter 3, fine-tuning and manual adjustments were necessary to determine the type of hopping mechanism, making it a labor-intensive process.

This motivated me to find a more efficient way to extract diffusion information that is automatic yet reasonable in terms of resources. It is impractical to run long MD simulations and recalculate order parameters for each structure afterward. Reusing pre-calculated MLP descriptors is an elegant approach that conserves computational resources, as these high-dimensional feature vectors encode valuable information about local chemical environments. However, working with high-dimensional vectors can be challenging. A superior approach is to perform dimensionality reduction first, reducing the vector to a lower dimension that can be visualized and classified, typically in two-dimensional space. The loss of information is not critical, as MLP descriptors are mostly strongly correlated.

However, there are many dimensionality reduction techniques, and one has to find the optimal solution for the problem. Due to simplicity, I first chose principal component analysis (PCA)⁸⁵. However, PCA is a linear approach, and as demonstrated in Chapter 5, it proves insufficient for achieving robust clustering. The nonlinear version of PCA is an autoencoder⁸⁶. After experimenting with autoencoders, they were deemed inappropriate for achieving the desired results. The stochastic training leads to unstable latent clustering due to the minimization of reconstruction errors rather than ordering latent variables. Additionally, autoencoders produce overlapping distributions as PCA, making it difficult to set clear boundaries between clusters.

For these reasons, I shifted my focus to graph-based methods for dimensionality reduction. They transform data into graphs and group nodes based on proximity, not preserving local densities and distances. Popular algorithms include the t-distributed stochastic neighbor embed-

ding (t-SNE)⁸⁷ and the uniform manifold approximation and projection (UMAP)⁸⁸. However, these algorithms are known to be sensitive to hyperparameters and the initialization process^{89–91}. This sensitivity can be particularly problematic in unsupervised analysis when the number of expected clusters is unknown in advance. A novel approach called pairwise controlled manifold approximation projection (PaCMAP)⁹² claims to address this lack of robustness.

I compared the PCA, UMAP, and PaCMAP algorithms to identify the most suitable approach for clustering local chemical environments using ACSFs. The tested systems included Si and H₂O phases, their point defects (monovacancies and interstitials), and surface/interface structures. The key finding was that PCA initialization is essential for stable clustering with both graph-based methods. Reported hyperparameter sensitivity was not confirmed, as the hyperparameters were fixed across all analyses to assess the robustness of the methods. Overall, the graph-based algorithms provided significantly better clustering results than PCA, which failed to yield valuable insights. UMAP outperformed PaCMAP in scenarios with significant class imbalances, such as the ice nucleus in liquid water (2,000 ice molecules vs. 72,000 water molecules). However, PaCMAP demonstrated superior cluster separation for water ice phases compared to UMAP, indicating better sensitivity in phase separation.

Lastly, I addressed a primary challenge with PaCMAP and graph-based dimensionality reduction algorithms: predicting new data points. I demonstrated that once clusters are formed and their interpretations are established, a simple neural network classifier can be trained and used to handle new data, effectively overcoming this limitation.

Since the manuscript focuses solely on silicon and water, Supplementary Information A.4 demonstrates the tracking of phosphorene monovacancies from Chapter 3.

1.6 Final remarks and outlook

The thesis has demonstrated the exploration and characterization of diffusion mechanisms using molecular dynamics accelerated by machine learning potentials. The potentials have been continuously improved through active learning, a robust approach applicable in scenarios where diffusion barriers are within the reach of molecular dynamics. The challenge with active learning lies in the frequent need to train potentials, which requires careful consideration of model complexity. High-dimensional neural network potentials are particularly well-suited for active learning strategy due to their simplicity, with the flexibility to switch to more sophisticated potentials as the training dataset comprehensively samples the relevant potential energy surface.

In cases where quantum mechanical calculations are costly, employing slightly elevated temperatures or larger timesteps can speed up

sampling. Integration errors introduced by longer timesteps are not a concern as long as the sampling of the relevant potential energy surface is preserved.

In closing, I reflect on machine learning potentials' current state and future. These potentials have undergone extensive development with numerous improvements over time. I believe that approaches such as MACE have reached a level of sophistication where research should be redirected from merely enhancing current models to focusing on broader applications. The emphasis should be placed on developing robust foundational models to make modeling more accessible. The first foundational model has already demonstrated significant promise for accelerating research⁹³. My experience also underscores the critical importance of uncertainty predictions to avoid significant errors. Furthermore, it is important to bear in mind that machine learning potentials are only as effective as the reference data they rely on. Improving the accuracy of reproducing reference data is futile if the reference data itself does not accurately represent the underlying physics.

1.7 Publications and contributions

The thesis contains four scientific publications at different stages of the peer review process. Chapters 2 and 3 have already been published, while Chapters 4 and 5 are under peer review. The list of publications and the corresponding individual contributions are as follows:

- **Chapter 2 and Appendix A.1 (published)**

Kývala, L., & Dellago, C. (2023). *Optimizing the architecture of Behler–Parrinello neural network potentials*. In *The Journal of Chemical Physics* (Vol. 159, Issue 9, pp. 094105). AIP Publishing. DOI: 10.1063/5.0167260

L. K. performed calculations, analysis, and wrote the initial draft.
C. D. provided consultation and finalized the article.

- **Chapter 3 and Appendix A.2 (published)**

Kývala, L., Angeletti, A., Franchini, C., & Dellago, C. (2023). *Diffusion and Coalescence of Phosphorene Monovacancies Studied Using High-Dimensional Neural Network Potentials*. In *The Journal of Physical Chemistry C* (Vol. 127, Issue 49, pp. 23743–23751). American Chemical Society (ACS). DOI: 10.1021/acs.jpcc.3c05713

L. K. developed MLPs, conducted and analyzed MD simulations accelerated by MLPs. A. A. performed DFT calculations, which included generating reference data, ciNEB analysis, and calculating

energy barriers and formation energies. L. K. and A. A. collaborated on active learning iterations and writing the initial draft. C. D. and C. F. provided consulting and finalizing the article.

- **Chapter 4 (submitted)**

Speckmann, C., Angeletti, A., Kývala, L., Lamprecht, D., Herterich, F., Mangler, C., Filipovic, L., Dellago, C., Franchini, C., and Kotakoski, J. (2024). *Electron-beam-induced adatom-vacancy-complexes in mono- and bilayer phosphorene*. Submitted to Physical Review X. The American Physical Society (APS). DOI: 10.48550/arXiv.2409.11102

C. S. designed the microscopy experiments, analyzed the data, designed the figures, and wrote the first draft of the manuscript. C. S. and D. L. prepared the samples and conducted the microscopy work. C. S., D. L., F. H., and C. M. were in charge of building and maintaining the experimental equipment. A. A. performed AIMD and structure relaxation. L. K. trained MLPs, conducted and analyzed MD simulations accelerated by MLPs. Consulting and finalizing the article were provided by J. K., C. D., and C. F.

- **Chapter 5 and Appendix A.3 (submitted)**

Kývala, L., Montero De Híjes, P., & Dellago, C. (2024). *Unsupervised identification of crystal defects from atomistic potential descriptors* Submitted to npj Computational Materials. Springer Nature. DOI: 10.48550/arXiv.2405.01320

L. K. designed the research and conducted clustering analyses. P. M. de H. prepared the water datasets. L. K. and P. M. de H. wrote the initial draft. C. D. provided consultation and finalized the article.

OPTIMIZING THE ARCHITECTURE OF BEHLER–PARRINELLO NEURAL NETWORK POTENTIALS

Kývala, L., & Dellago, C. (2023). *Optimizing the architecture of Behler–Parrinello neural network potentials*. In *The Journal of Chemical Physics* (Vol. 159, Issue 9, pp. 094105). AIP Publishing. DOI: 10.1063/5.0167260

The architecture of neural network potentials is typically optimized at the beginning of the training process and remains unchanged throughout. Here, we investigate the accuracy of Behler-Parrinello neural network potentials for varying training set sizes. Using the QM9 and 3BPA datasets, we show that adjusting the network architecture according to the training set size significantly improves accuracy. We demonstrate that both an insufficient and an excessive number of fitting parameters can have a detrimental impact on the accuracy of the neural network potential. Furthermore, we investigate the influences of descriptor complexity, neural network depth, and activation function on the model’s performance. We find that for the neural network potentials studied here, two hidden layers yield the best accuracy and that unbounded activation functions outperform bounded ones.

2.1 Introduction

Machine learning has emerged as a powerful tool in the field of atomistic simulations and is currently revolutionizing the way we predict and understand the properties of matter from the atomic scale. Since the first machine learning potential was published in 1995⁹⁴, the field has advanced significantly, and a wide range of methods are now available, including neural network potentials (NNP) of various types^{11,22,22–24,30,34,95–100,100–112}, Gaussian approximation potentials (GAP)¹⁵, Gaussian processes¹¹³, support vector machines (SVM)¹¹⁴, atomic permutationally invariant polynomials (aPIP)¹¹⁵, spectral neighbor analysis potentials (SNAP)^{16,116}, moment tensor potentials (MTP)¹⁷, gradient domain machine learning (GDML)¹¹⁷, and potentials based on the atomic cluster expansion (ACE)¹⁸.

In this work, we focus on neural network potentials, which are among the most established approaches¹¹⁸, and address the question of how to choose their architecture for optimum accuracy. Utilizing artificial neural networks to represent atomistic potentials has several advantages. First, neural networks are universal approximators¹¹⁹ and, therefore,

can capture complex, non-linear relationships between atomic positions and the corresponding energy and forces. Another notable advantage of neural networks is their ability to effectively learn from large data sets and, at the same time, offer rapid evaluation once they are trained. The evaluation time of neural network potentials scales with the architecture of the neural network itself rather than growing with the size of the data set, as is typically the case for most kernel-based methods¹²⁰.

Neural network potentials, however, also have certain limitations. Their complexity makes them challenging to understand theoretically, resulting in various approaches and topologies based primarily on empirical knowledge. Regardless of the approach, one common challenge in neural networks involves finding good hyperparameters, such as the number of features (descriptors) and the number of fit parameters (weights and biases).

The focus of this paper is to understand how the accuracy of Behler-Parrinello high-dimensional neural network potentials¹¹ depends on the number of descriptors used to characterize chemical environments and on the number of fitting parameters in the classical regime. We demonstrate that the number of parameters must be adjusted for each training set size and descriptor complexity to achieve optimal accuracy. In the following, we will briefly review the methodology of high-dimensional neural network potentials (HDNNPs) and discuss some computational details. Then, we will devise a strategy to find the optimum number of parameters for a given training set size and discuss the impact of neural network depth and activation functions.

2.2 Theory

2.2.1 High-dimension neural network potentials

HDNNPs were introduced by Behler and Parrinello in 2007¹¹ to model potential energy surfaces of condensed matter systems. One key idea of HDNNPs is that the potential energy is written as the sum of individual atomic contributions,

$$E = \sum_i E_i. \quad (7)$$

Each contribution E_i depends on the local chemical environment within a sphere centered at the respective atom. The environment is characterized by a set of descriptors called symmetry functions G . The symmetry functions (SFs) are constructed in a way that makes them invariant with respect to global translations and rotations, as well as permutations of atoms of the same species. These descriptors then serve as the input layer for a simple feedforward neural network specific to each chemical element. We refer to a topical review¹³ for a more detailed description of HDNNPs.

2.2.2 Atom-centered symmetry functions

HDNNPs utilize only radial and angular descriptors, although this representation has been shown to be incomplete in some cases¹²¹. Originally, Behler and Parrinello¹¹ introduced radial (2-body) symmetric functions as a sum of Gaussians with parameters η and μ :

$$G_i^{\text{rad}} = \sum_{j \neq i}^N e^{-\eta(r_{ij}-\mu)^2} f_c(r_{ij}), \quad (8)$$

where the summation runs over all neighbors j within a certain cutoff radius r_c , capturing interactions that quickly decrease with distance. The angular symmetry functions (3-body) also incorporate information about the angle θ_{ijk} spanned by two neighbors j and k and the central atom i :

$$\begin{aligned} G_i^{\text{ang}} = & 2^{1-\zeta} \sum_{j \neq i}^N \sum_{k \neq i, j}^N (1 + \lambda \cos \theta_{ijk})^\zeta \\ & \times e^{-\eta[(r_{ij}-\mu)^2 + (r_{ik}-\mu)^2 + (r_{jk}-\mu)^2]} \\ & \times f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}). \end{aligned} \quad (9)$$

Here, the parameter ζ controls the angular resolution, and λ is used to center the maxima of the cosine terms at 0 or π . To ensure a smooth decay to zero at the cutoff radius r_c , both radial and angular functions are modified by a smooth cutoff function f_c :

$$[h]f_c(r_{ij}) = \begin{cases} \frac{1}{2} \left[\cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1 \right], & \text{if } r_{ij} \leq r_c, \\ 0, & \text{if } r_{ij} > r_c. \end{cases} \quad (10)$$

Recently, Bircher et al.¹²² presented a class of atom-centered symmetry functions based on polynomials with compact support called polynomial symmetry functions:

$$G_{\text{rad}} = \sum_{j \neq i} f_P(r_{ij}, r_0, \Delta r), \quad (11)$$

and

$$G_{\text{ang}} = \sum_{j \neq i, k > j} f_P(r_{ij}, r_0, \Delta r) f_P(r_{ik}, r_0, \Delta r) f_P(\theta_{ijk}, \theta_0, \Delta \theta), \quad (12)$$

where the function $f_P(x, x_0, \Delta x)$ is defined as

$$f_P(x, x_0, \Delta x) = \begin{cases} f_P^2\left(\frac{x - x_0}{2\Delta x}\right), & \text{if } x_0 \leq x \leq x_0 + \Delta x, \\ f_P^2\left(\frac{x_0 - x}{2\Delta x}\right), & \text{if } x_0 - \Delta x \leq x \leq x_0, \\ 0, & \text{otherwise,} \end{cases} \quad (13)$$

with

$$f^{P^2}(x) = x^3(x(15 - 6x) - 10) + 1, \quad (14)$$

$$x_0 = \frac{1}{2}(x_{\max} + x_{\min}), \quad (15)$$

$$\Delta x = \frac{1}{2}(x_{\max} - x_{\min}). \quad (16)$$

Unlike the original Behler-Parinello symmetry functions¹¹, they offer greater flexibility in the choice of angular symmetry functions while requiring fewer floating point operations in their evaluation.

2.2.3 Weighted symmetry functions

Atom-centered symmetry functions (ACSFs) have been proven to work well for many systems¹²³, but poor scaling with the number of elements reduces their applicability to more complex systems. While the number of required radial symmetry functions scales linearly with the number of elements N_{elem} in the system, the number of required angular symmetry functions is proportional to $N_{\text{elem}}(N_{\text{elem}} + 1)/2$ to include all triplet combinations.

To overcome quadratic scaling with the number of elements, Gastegger et al.¹⁴ proposed a modification of the atom-centered symmetry functions. Instead of using a separate function for each combination of elements, the element information is encoded explicitly into the symmetry functions. Symmetry functions modified in this way are called weighted atom-centered symmetry functions (wACSFs) and take the form:

$$W_i^{\text{rad}} = \sum_{j \neq i}^N g(Z_j) e^{-\eta(r_{ij} - \mu)^2} f_c(r_{ij}), \quad (17)$$

for radial symmetry functions and

$$\begin{aligned} W_i^{\text{ang}} = & 2^{1-\zeta} \sum_{j \neq i}^N \sum_{k \neq i, j}^N h(Z_j, Z_k) (1 + \lambda \cos \theta_{ijk})^\zeta \\ & \times e^{-\eta[(r_{ij} - \mu)^2 + (r_{ik} - \mu)^2 + (r_{jk} - \mu)^2]} \\ & \times f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}), \end{aligned} \quad (18)$$

for angular symmetry functions. Here, Z_j is the atomic number of atom j . The weighting functions $g(Z_j)$ and $h(Z_j, Z_k)$ can then take various forms, but the simplest forms, $g(Z_j) = Z_j$ and $h(Z_j, Z_k) = Z_j Z_k$, seem to perform adequately well¹⁴. The polynomial symmetry functions can be modified similarly¹²².

With this modification, the number of symmetry functions per element remains the same as in the single-element system, and the total

number of symmetry functions increases linearly with the number of elements. However, in this approach, more information is compressed into each symmetry function, causing a loss of resolution in the description of local chemical environments, which can manifest itself with reduced accuracy with large training sets.

2.3 Computational details

2.3.1 Dataset

We optimized the architectures of HDNNPs for the QM9 and 3-(benzyloxy)pyridin-2-amine (3BPA) datasets^{124,125}. The QM9 dataset consists only of structures corresponding to energy minima, simplifying the task of predicting energies alone. Conversely, the 3BPA dataset consists of structures that are out of equilibrium, requiring learning both energies and forces.

The QM9 dataset is widely used as a machine learning potential benchmark¹²⁶. The dataset comprises 134k small organic molecules made up of H, C, O, N, and F atoms. The energies in the QM9 dataset are computed at the B3LYP/6-31G(2df,p) level of theory. Since 3,054 structures failed the consistency test¹²⁴, it is common practice to remove these structures from the dataset^{22,23,100}. From this pruned dataset containing 130,831 structures, we randomly selected 10,000 structures to serve as a validation set that controls early stopping¹⁰⁰. Structures that are not in the training or validation set are assigned to the test set.

The 3BPA dataset comprises 500 training structures obtained at room temperature. Out of these 500 structures, 50 were used as the validation set during the training process. The energies and forces for all structures were computed using density functional theory (DFT) with the ω B97X exchange-correlation functional and the 6-31G(d) basis set.

2.3.2 Optimizer

The n2p2 package¹²⁷ was used to obtain the neural network potential for the QM9 dataset. The training was carried out with multistream extended Kalman filter (MS-EKF) due to its higher performance for

Table 1: Standard deviations (SD) of the mean absolute error per atom (MAE/atom) calculated from five models differing by random seed for the parameters and with different train/test splits. The SDs are given in kcal/mol.

Training set size	1,012	3,172	10,044
SD (random seed)	0.59	0.16	0.02
SD (train/test split)	0.59	0.09	0.01

this type of task compared to gradient optimization algorithms^{48,50}. However, the nonlinear scaling with the number of parameters (weights and biases) and limited parallelization make this algorithm suitable only for rather small neural networks. We note that all results of this work were computed only on 12, 16, and 24 CPU threads.

Since the mean absolute error (MAE) is a common metric for the QM9 benchmark, here we apply the Kalman filter to this loss function. As for the 3BPA, the root mean square error (RMSE) has been minimized. The relative importance of each force in terms of energy updates has been set to 10.

Reported errors in this paper are a sample average over 5 models for cross-validation, 10 models for the QM9 data set, and 20 for the 3BPA dataset. The models differ in random seed, which affects the initialization of the weights and the mixing of training structures. A different approach with averaging over different train/test splits has also been reported²². However, we have found that the standard deviation estimated from different initial parameters is similar to that of the train/test split only for small training sets. As the number of structures in the training set increases, the standard deviation between model MAE due to different NN initialization outweighs the effect of the train/test split, as summarized in Table 1.

2.3.3 Descriptors

A previous study has demonstrated that the accuracy of a neural network potential can be affected by the number of descriptors employed¹⁴. In particular, it was shown that by transitioning from an unweighted model with 220 symmetry functions to a full-weighted model with 32 symmetry functions, the MAE was reduced by 25% for the QM9 dataset. To study the influence of descriptor complexity on the model accuracy, we introduced three types of models differing in complexity: full-weighted, angular-weighted, and unweighted models utilizing polynomial symmetry functions (refer to Supplementary information A.1 for detailed information). After pruning unrealized bonds, the complexity of our models can be described as follows: the full-weighted model contains 34 symmetry functions (7 weighted radial and 27 weighted angular), the angular-weighted model comprises 61 symmetry functions (34 unweighted radial and 27 weighted angular), and the unweighted model consists of 225 symmetry functions (18 unweighted radial and 207 unweighted angular) on average per element. The element-wise numbers of SFs are provided in Table 6.

We employed only the angular-weighted model to describe the chemical environment for the 3BPA data set. Since the 3BPA molecules do not contain fluorine, the average number of SFs is reduced to 54 (27 unweighted radial and 27 weighted angular).

Table 2: Comparison of Behler-Parinello ACSF and polynomial ACSF errors based on 5-fold cross-validation with a training set size of 8,000 and a validation set of 2,000.

Symmetry functions	MAE (kcal/mol)		
	Training	Validation	Testing
Gaussian: other work ¹⁴	1.18	1.84	1.83
Gaussian: this work	0.55	1.23	1.28
Polynomial	0.53	1.19	1.24

2.4 Results and Discussion

2.4.1 QM9 Data Set

First, we have reproduced 5-fold cross-validation with 8,000 training and 2,000 validation structures utilizing the original Gaussian descriptors from previous work¹⁴. The same 5-fold cross-validation has also been carried out for our polynomial set with full-weighted SFs. Table 2 compares 5-fold averaged MAE for both models. A significantly better accuracy has been achieved with Gaussian symmetry functions compared to the original work¹⁴. This accuracy improvement can be mainly attributed to adopting the multistream Kalman filter, which offers superior results compared to the previously used single-stream Kalman filter⁴⁸. The polynomial wACSFs slightly outperform the Gaussian wACSFs. Thus, we report another example showing that polynomial symmetry functions are an advantageous alternative to the original Gaussian SFs¹²².

Previous work¹⁴ mainly focused on introducing weighted ACSFs. For this purpose, only 10,000 structures were employed from the whole QM9 dataset. Here, we have determined the QM9 learning curves for our three models differing in complexity (see Descriptors section in Computational Details) with training sets ranging from 1,000 to 75,000 structures. The neural network architecture is the same as in the previous work¹⁴, i.e., there are two hidden layers with 10 and 50 nodes connected by a softplus activation function. The resulting BPNN learning curves are shown in Fig. 6.

The comparison of our three models shows that the model combining unweighted radial and weighted angular symmetry functions perform best. For a training set size of 10,000; the full-weighted model underperforms by 25%, while the unweighted model has twice the error. This indicates that the simplest model (full-weighted) is undertrained as the descriptors do not have sufficient resolution, and the most complex model (unweighted) is affected by the curse of dimensionality¹²⁸. Due to the significantly larger error, the model with 225 descriptors on average (unweighted model) is excluded from further analysis. Although the full-weighted and angular-weighted models perform well, a loss of

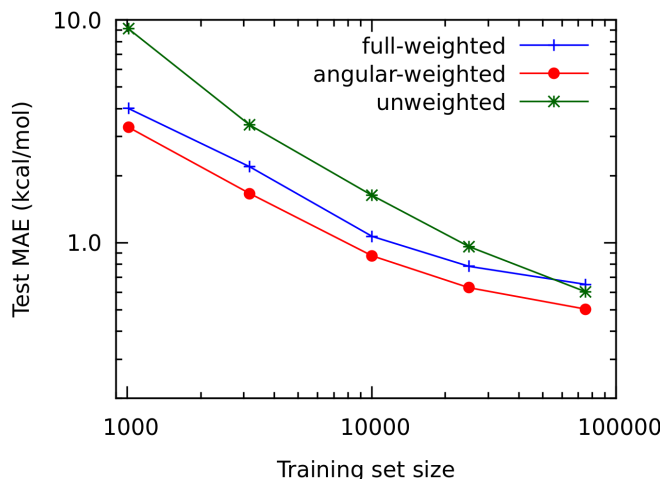


Figure 6: QM9 learning curves for our three BPNN models with a fixed architecture optimized for 10,000 structures.

linearity in the log-log plot is observed after 10,000 training structures for these two models.

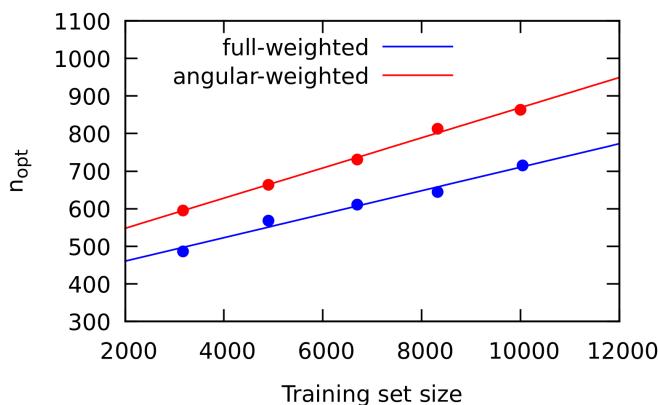


Figure 7: Optimal number of parameters n_{opt} as a function of training set size for the full-weighted and angular-weighted models. Linear functions are used to fit both models (solid lines).

The total error of a machine learning model originates from approximation, generalization, and optimization errors. The optimization origin of the error of our networks seems unlikely since the deviation from the power-law behavior (linearity in log-log plot) was not observed for the unweighted model. It also should not stem from the non-unique and biased distribution of the configuration space, as other methods do not experience this deviation^{22,23,100,129,130}. Hence, it seems likely that the errors are mostly approximation-based, i.e., the descriptors fail to sufficiently distinguish the different chemical environments, and/or the small number of parameters does not allow it to capture the finer features of the potential energy surface. Increasing the number of descriptors from the full-weighted to the angular-weighted model slightly reduces the underfitting but does not eliminate the slower

training pace. The learning curves shown in Fig. 6 were obtained using a neural network comprising two hidden layers connected with the softplus activation function. The hidden layers consisted of 10 and 50 nodes, as this particular architecture has been previously reported in the literature¹⁴. However, this neural network architecture has been optimized only for 8,000 training structures. As non-linearity in the log-log plot is observed from 10,000 training structures, an insufficient number of parameters might cause this behavior.

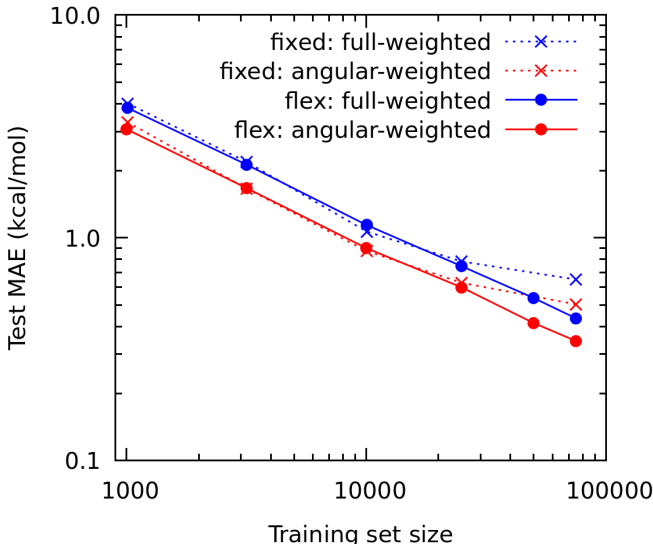


Figure 8: Comparison of fixed and flexible architectures for QM9 learning curves. The flexible architecture model dynamically adjusts the number of parameters according to the size of the training set. In contrast, the fixed model employs a predetermined architecture optimized specifically for 10,000 structures.

For training sets of sizes 1000, 3170, 4900, 6699, 8321, and 10000, we have investigated the dependence of the validation MAE on the number of parameters. The investigated neural network architectures consist of two hidden layers, where the number of nodes in both hidden layers is identical. For each training set size, we have trained 10 models starting with different random seeds for the parameters. The errors were then averaged and fit with the exponential function $f(x) = Ae^{-bx} + c$ (see Fig. 33). We define the optimal number of parameters n_{opt} using the fitted parameter b as $n_{\text{opt}} = 4.5/b$. With this definition, the deviation of the error from its asymptotic value c has decayed to 1% at n_{opt} . For both models, a linear dependence of n_{opt} on the training set size is observed, as shown in Fig. 7.

Next, we investigated whether extrapolating n_{opt} to larger training set sizes improves the learning curves for training sets of over 10,000 structures. The original number of parameters based on the 62-10-50-1 architecture corresponds to 1,231 for a hydrogen atom. For a training set size of 75,000 structures, we should have approximately 4,000 parameters according to our linear extrapolation. With this knowledge,

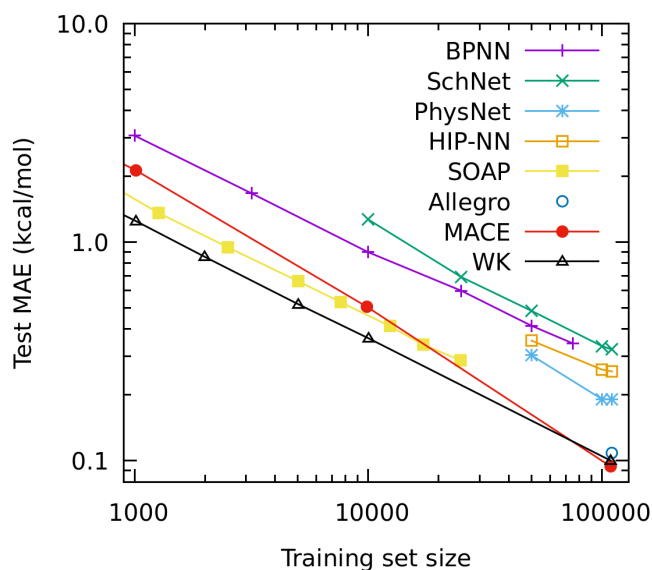


Figure 9: QM9 learning curves for our optimized angular-weighted BPNN model compared to other machine learning models^{22,23,100,112,131–133}.

we have retrained our models with predicted optimal architectures for each training set size. The resulting QM9 learning curves are depicted in Fig. 8. Linearity in the log-log plot is now also observed beyond a training set size of 10,000, indicating that the original neural network architecture does not provide sufficient complexity for larger training set sizes.

In Fig. 9, we present a comparison between our optimized angular-weighted BPNN with other machine learning models, i.e., with SchNet¹⁰⁰, PhysNet²³, HIP-NN²², SOAP¹³¹, Allegro¹¹², MACE¹³³, and Wigner-Kernels¹³². The BPNN model demonstrates comparable accuracy to SchNet¹⁰⁰ and HIP-NN²². However, recent parametric methods, such as MACE¹³³ or Allegro¹¹², achieve higher accuracy, albeit at the cost of complex architectures with significantly more parameters. For instance, the most accurate Allegro model encompasses 17.9 million parameters, exceeding the number of parameters of our architecture-optimized BPNN by more than three orders of magnitude.

As we have shown, it is beneficial to vary the number of fitting parameters with the size of the training set. However, if computation capacity is not a concern, could this approach be replaced by a fixed architecture where the optimal number of parameters for the largest training set is used for all training set sizes? Our largest training set with 75,000 structures needs 4,000 parameters (weights + biases). To verify this approach, we have gradually increased the number of parameters with a corresponding increase in training epochs for the smallest training dataset with 1,000 structures. Figure 10 shows how, at first, the validation MAE decreases with an increasing number of parameters. For more than 1,000 parameters, however, the validation MAE slowly starts to increase. Thus, if a model with 4,000 parameters

Table 3: MAE comparison of NNPs with varying depth across different training set sizes.

Training set size	MAE (kcal/mol)				
	1H	2H	3H	4H	5H
1,000	3.36	3.07	3.45	3.59	4.44
3,162	1.91	1.67	1.78	1.88	1.98
10,000	1.03	0.90	0.93	1.01	1.01
25,000	0.65	0.60	0.61	0.64	0.67
50,000	0.45	0.41	0.44	0.47	0.49
75,000	0.37	0.34	0.36	0.40	0.43

is employed instead of 600 parameters for a training set of size 1,000, the resulting model is 26% slower on CPU and also approximately 10% less accurate. Hence, both insufficient complexity and an overabundance of complexity degrade the accuracy of the potential energy prediction.

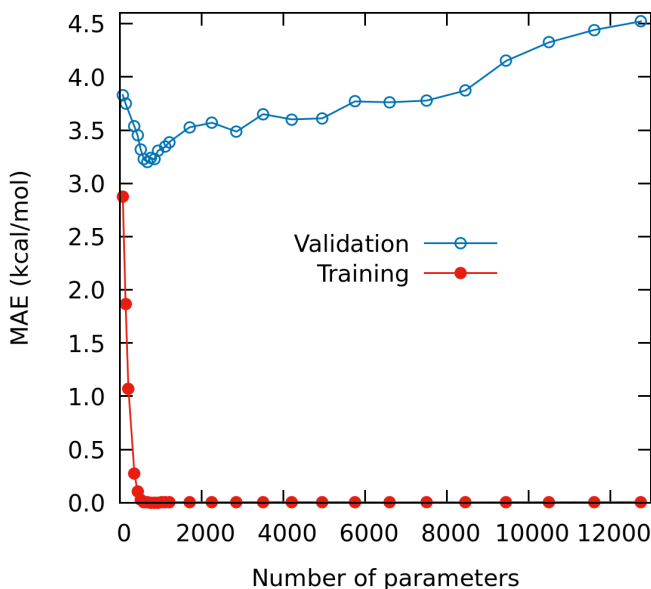


Figure 10: Validation and training MAE as a function of the number of parameters for the angular-weighted model with a training set size of 1000.

The depth of an HDNNP is another crucial hyperparameter to consider. We analyzed learning curves for neural networks with 1 to 5 hidden layers. We set the number of parameters for all models and training set sizes based on the optimal value determined for the 2 hidden layer model. We have found that increasing or decreasing the depth of the network does not result in improved accuracy, as listed in Tab. 3. The model with only one hidden layer exhibited an average MAE that was 11% higher compared to the model with 2 hidden layers. Similarly, the models with 3, 4, and 5 hidden layers displayed an MAE that was larger by 7%, 15%, and 26%, respectively. This observation

indicates a degradation of the accuracy with an increase in the depth of the neural network while maintaining the same number of parameters.

Nevertheless, it is crucial to acknowledge the inherent bias in this comparison, as optimizing the number of parameters was specifically tailored to the 2 hidden layer model. Consequently, it is conceivable that the optimal number of parameters is not exclusively dictated by the number of descriptors and the size of the training set, as discussed above, but also influenced by the depth of the neural network. Furthermore, the arrangement of parameters within the hidden layers may also impact the model’s performance. In this study, a uniform node distribution was employed, but it is plausible that alternative distributions could potentially yield more favorable outcomes.

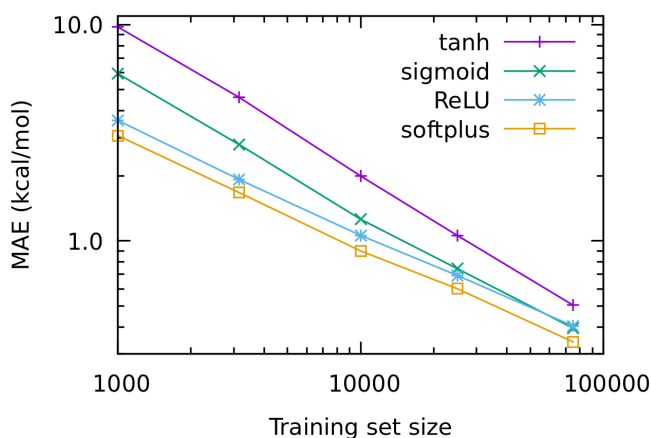


Figure 11: Comparison of activation functions for QM9 learning curves in models with two hidden layers.

For our model with two hidden layers, we also evaluated the performance of four activation functions: tanh, sigmoid, rectified linear unit (ReLU), and softplus. In Fig. 11, we observe that unbounded activation functions (ReLU and softplus) exhibit a significantly better generalization ability, particularly for smaller training set sizes. As the training set size increases, the difference between activation functions diminishes, and for a training set of 75,000 structures, the sigmoid function achieves the same level of accuracy as the ReLU function. Nevertheless, its MAE is still 16% higher than the softplus function, which prevails across all training set sizes.

2.4.2 3BPA Data Set

Given that the 3BPA molecule consists of 27 atoms, there are 81 corresponding force components for reference energy. We investigated the dependence of the validation RMSE on the training set size for datasets consisting of 100, 200, and 300 structures. The averaged RMSE values from 20 models, which were fitted using an exponential function, are depicted in Fig. 34. For each training set size, we determined the

optimal parameter value n_{opt} for energies and forces separately using the same approach as for the QM9 dataset. The resulting dependence is illustrated in Fig. 12.

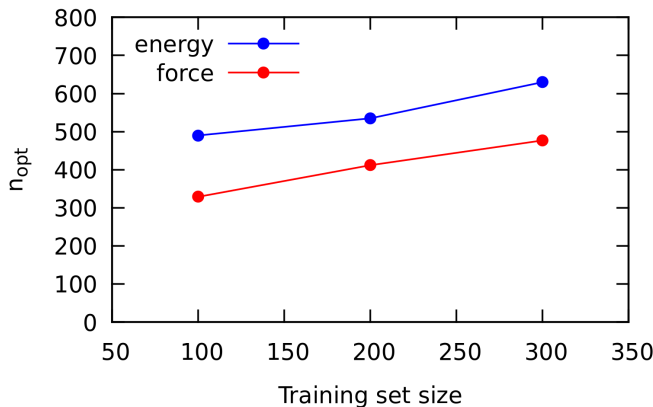


Figure 12: Optimal number of parameters n_{opt} as a function of training set size for energies and forces of the 3BPA dataset.

Similar to the findings for the QM9 dataset, we once again note a positive correlation between the optimum number of parameters and the size of the training set. However, the optimum number of parameters differs for energies and forces. As can be inferred from Fig. 12, about 150 more parameters are required for the optimal energy prediction for a given training set size even though 81 times more forces than energies are available. This suggests that a more complex architecture is required for energies than forces.

It is noteworthy that for the dataset consisting of a total of 24,600 reference data points (300 structures), achieving 99% accuracy within the classical regime only requires a modest number of 600 parameters.

2.5 Summary

In this paper, we have analyzed the learning curves of HDNNPs for the QM9 dataset using three models that employ varying numbers of descriptors, with average values of 34, 61, and 225 SFs. Among these models, the one with a combination of unweighted radial symmetry functions and weighted angular symmetry functions exhibits the highest accuracy. The full-weighted model demonstrates underfitting, likely due to insufficient information captured by the descriptors. On the other hand, the unweighted model tends to overfit within the investigated range of training set sizes.

Our calculations show that in order to achieve power-law behavior over the entire training set size range (from 1,000 to 75,000), the neural network architecture has to be adapted according to the training set size. For the angular-weighted model, we have empirically found a linear dependence of the optimal number of parameters n_{opt} on the training set size D , $n_{opt} = 0.045D + 526.224$. The correlation between

the number of parameters and the size of the training set is also evident in the 3BPA dataset. Interestingly, two distinct dependencies emerge for energies and forces, with the energies necessitating a larger number of parameters than the forces. Our results indicate that the evaluation time of optimized neural networks is not fully independent of the training set size. Hence, with each significant change in the training set size, e.g., during active learning, the optimal neural network architecture should be re-determined to fully use the available training data. Furthermore, we have demonstrated that an excess of parameters can be detrimental.

A comparison of learning curves for the QM9 data set shows that the accuracy achieved with our optimized angular-weighted BPNN is comparable to that of other, more complex machine learning models such as SchNet¹⁰⁰ and HIP-NN²². More recent methods such as MACE¹²⁵, Allegro¹¹² or Wigner Kernels¹³², however, achieve a substantially higher accuracy, but at the cost of greater complexity.

We also investigated the impact of varying the number of hidden layers for the QM9 dataset while maintaining a constant number of parameters. Our results revealed that utilizing two hidden layers yielded the lowest MAE, with the accuracy gradually decreasing as the number of layers increased. However, it is important to note that the number of parameters utilized for each training set size was optimized specifically for the two hidden layer models. This raises the question of whether the optimal number of parameters not only depends on the number of descriptors and training set size, as demonstrated, but on how the parameters are distributed over the layers. Finally, our findings indicate that unbounded activation functions, such as ReLU and softplus, outperform bounded activation functions (tanh and sigmoid) for small datasets, but the difference in performance gradually decreases as the dataset increases. Among the tested activation functions, the softplus function yielded the highest accuracy.

Acknowledgments

We acknowledge financial support by the Doctoral College Advanced Functional Materials – Hierarchical Design of Hybrid Systems DOC 85 doc.funds funded by the Austrian Science Fund (FWF) and the Vienna Doctoral School in Physics (VDSP). The computational results presented have been partly achieved using the Vienna Scientific Cluster (VSC).

Supplementary information

Supplementary information A.1 contains detailed information about the polynomial symmetry functions used in this work. In addition, the Supplementary information A.1 includes validation errors as a function

of the number of parameters for the QM9 and the 3BPA data sets (Fig. 33, 34, and 35).

DIFFUSION AND COALESCENCE OF PHOSPHORENE MONOVACANCIES STUDIED USING HIGH-DIMENSIONAL NEURAL NETWORK POTENTIALS

Kývala, L., Angeletti, A., Franchini, C., & Dellago, C. (2023). *Diffusion and Coalescence of Phosphorene Monovacancies Studied Using High-Dimensional Neural Network Potentials*. In *The Journal of Physical Chemistry C* (Vol. 127, Issue 49, pp. 23743–23751). American Chemical Society (ACS). DOI: 10.1021/acs.jpcc.3c05713

The properties of two-dimensional materials are strongly affected by defects that are often present in considerable numbers. In this study, we investigate the diffusion and coalescence of monovacancies in phosphorene using molecular dynamics simulations accelerated by high-dimensional neural network potentials (HDNNP). Trained and validated with reference data obtained with density functional theory (DFT), such surrogate models provide the accuracy of DFT at a much lower cost, enabling simulations on time scales by far exceeding those of first-principles molecular dynamics. Our micro-second long simulations reveal that monovacancies are highly mobile and move predominantly in the zigzag rather than armchair direction, consistent with the energy barriers of the underlying hopping mechanisms. In further simulations, we find that monovacancies merge into energetically more stable and less mobile divacancies, following different routes that may involve metastable intermediates.

3.1 Introduction

Two-dimensional (2D) materials are currently the subject of intense research because they show promise in a wide range of important applications, including electronics, valleytronics, catalysis, and biosensing¹³⁴. In particular, graphene has attracted much interest due to its large carrier mobility¹³⁵, but the lack of a band gap makes device applications difficult. Transition metal dichalcogenides (TMDs), including MoS₂ and WSe₂, feature band gaps within the 1.2–1.8 eV range¹³⁴. But as the valence and conduction bands of TMDs are made up of transition metal *d*-orbitals, their carrier mobility is a few orders of magnitude lower than that of graphene¹³⁶. In contrast, phosphorene, a 2D monolayer of black phosphorus, possesses a thickness-dependent electronic band gap (2.2 eV in the monolayer to 0.3 eV in bulk^{72,73}),

a high carrier mobility^{74–77} and in-plane anisotropy. Thanks to this unique combination of properties, phosphorene has been utilized in the fields of photo-catalysis^{137,138}, semiconductors^{139,140}, rechargeable batteries¹⁴¹, solar energy conversion¹⁴², and gas sensors¹⁴³.

Most properties and applications of materials are affected by the presence of defects, especially point defects¹⁴⁴. These imperfections give phosphorene its unique properties, such as hole-doping¹⁴⁵, local magnetic moments¹⁴⁵, improved activity for hydrogen evolution reactions¹⁴⁶, enhanced ion-transport for alkali-ion batteries¹⁴⁷, and unique photo-responses for neuromorphic computing^{148,149}. Defects, however, can also act in undesirable ways that lead to device failure. Therefore, understanding the formation and dynamics of defects is essential for designing stable and functional phosphorene-based devices.

The simplest point defect in phosphorene, *i.e.*, the monovacancy, is the subject of this work. The localized magnetic moment originating from an unsaturated bond makes the monovacancy particularly intriguing. Despite being predicted to be the most common defect in phosphorene after electron and ion irradiation^{79,80}, a predominance of divacancies and tetravacancies has been observed in experimental studies^{81–83}. Only recent work claims to have captured individual monovacancies⁸³. The difficulty of observation can be attributed to the highly itinerant nature of monovacancies⁸⁴, making it difficult to identify them on the time scales accessible experimentally. Another reason may be that monovacancies coalesce rapidly into more stable divacancies. However, the detailed mechanism of monovacancy coalescence has not yet been clarified.

In this work, we investigate the structure and kinetics of phosphorene monovacancies with extensive molecular dynamics (MD) simulations based on high-dimensional neural network potentials (HDNNP-MD)^{11,13}. In this machine learning approach, the potential energy of the system is represented using artificial neural networks trained on energies and forces calculated using density functional theory (DFT), leading to a considerable increase in efficiency. Our simulations with an aggregate length of more than 1 μs identify the mechanism and rates of the hopping processes underlying defect diffusion, confirm the mechanism postulated previously, and investigate with the nudged elastic band method⁸⁴. We then investigate how two monovacancies approach each other and coalesce into a divacancy, finding that this process can occur following three different mechanisms depending on the orientation of the monovacancies and their direction of approach. Under certain conditions, a long-lived monovacancy pair forms, in which the monovacancies are arranged so that all bonds are satisfied.

The remainder of the article is organized as follows. First, we outline the computational methods used in this work and provide technical details of the simulations. Subsequently, we present results obtained for the diffusion of a single vacancy, followed by a discussion of vacancy coalescence. Finally, we conclude our findings.

3.2 Methods

This section first discusses how the reference data are generated via first-principle calculations. We then describe the neural network model potential and explain how active learning was employed to generate additional reference data. Finally, we discuss the neural network parametrizations and describe the molecular dynamics setup.

3.2.1 First-principles calculations

All reference data used for the neural network training were generated with first-principle calculations performed using the Vienna Ab Initio Simulation Package (VASP)^{150,151}. The Perdew-Burke-Ernzerhof (PBE) functional¹⁵² was employed with an energy cutoff of 600 eV. The optimized values of the lattice parameters for the pristine phosphorene structure are found to be 3.30 Å and 4.62 Å along the zigzag and armchair directions, respectively. The criterion for the relaxation of the forces acting on each atom was set to 0.01 eV/Å. The supercell size chosen to simulate the system is $7 \times 5 \times 1$, containing 140 atoms in the pristine structure, with a vacuum region of 20 Å. The Brillouin zone was sampled using a $3 \times 3 \times 1$ Gamma-centered mesh. The *ab initio* molecular dynamics simulations were performed with the Langevin thermostat^{153,154} in the NPT ensemble^{155,156} applying zero external pressure and 300 K with a reduced energy cutoff of 250 eV. By allowing the relaxation of the two primitive lattice vectors along the zigzag and armchair direction, we inferred the average values at room temperature of the supercell containing the defective structure with one monovacancy. The obtained values employed in all calculations are 3.32 Å and 4.56 Å, respectively. The climbing imaged nudged elastic band (ciNEB)¹⁵⁷ calculations were performed by using a relaxation criterion of 0.02 eV/Å for the forces acting on each image. The formation energy of a defect is defined as $E_f = E_{\text{defective}} - N_p \times E_p$ where $E_{\text{defective}}$ is the energy of the structure with defect, N_p the number of atoms in the supercell and E_p the energy per phosphorus atom in the pristine configuration, $E_p = E_{\text{pristine}}/N_p$.

3.2.2 Neural network potential

In this work, we employ the HDNNP approach pioneered by Behler and Parrinello¹¹, in which the total potential energy of the systems is written as the sum of local, atom-centered contributions E_i ,

$$E = \sum_{i=1}^N E_i(\mathbf{G}), \quad (19)$$

where N denotes the number of atoms. Each contribution is encoded in an artificial neural network that depends on the local chemical en-

vironment described by a set of descriptors $\mathbf{G}$. These descriptors, also known as symmetry functions, are characteristic fingerprints of the chemical environment and are invariant with respect to translations, rotations, and permutations of atoms of the same species. We employ the so-called polynomial symmetry functions as descriptors, a family of atom-centered symmetry functions based on polynomials with compact support¹²². They provide more flexibility in the selection of angular symmetry functions while requiring fewer floating point operations for their evaluation than the original Behler-Parrinello symmetry functions¹¹. The forces acting on the atoms, necessary for the MD simulations, are given by the negative gradient of the potential energy, which is straightforward to calculate due to the simple analytical form of the neural network potential.

Active learning

To train our neural network potential, we use an active learning strategy called query by committee⁴¹, which has proven useful for creating training sets for machine learning potentials^{42,43,158,159}. First, a few short ab-initio MD simulations are performed, and configurations with energies and forces are collected from them. On this initial training set, several HDNNPs are trained with various training/validation splits and different initial conditions. The total energy $\bar{E}$ and the force components $\bar{F}_{k,\alpha}$ are then calculated as averages over the committee,

$$\bar{E} = \frac{1}{C} \sum_{c=1}^C E_c, \quad (20)$$

$$\bar{F}_{k,\alpha} = \frac{1}{C} \sum_{c=1}^C F_{k,\alpha,c}. \quad (21)$$

Here, C is the committee size, and E_c and $F_{k,\alpha,c}$ are the total energy and forces predicted by committee member c , respectively.

The committee disagreement is defined as the standard deviation of the committee predictions:

$$\sigma_E = \sqrt{\frac{1}{C} \sum_{c=1}^C (\bar{E} - E_c)^2}, \quad (22)$$

$$\sigma_{F_{k,\alpha}} = \sqrt{\frac{1}{C} \sum_{c=1}^C (\bar{F}_{k,\alpha} - F_{k,\alpha,c})^2}. \quad (23)$$

One expects that the predictions of different committee members are similar for configurations close to the training set and dissimilar elsewhere. Hence, the committee disagreement is a measure of uncertainty

of the neural network prediction⁴². The neural network-accelerated MD is stopped in the training phase when the committee disagreement exceeds a predetermined threshold. The simulation continues with DFT-MD, generating new configurations to update the training set. The neural networks are then retrained on this larger dataset, and the HDNNP-MD simulation is resumed and runs until the threshold is exceeded again. This procedure, repeated several times if necessary, ensures that the training set contains all configurations required to accurately reproduce not only the perfect lattice and stable defects but also short-lived intermediate configurations without leading to excessive training set sizes. The uncertainty estimation based on the committee approach was also used during all production runs to monitor the accuracy of the HDNNP for all configurations visited during our extensive MD simulations.

Neural network parameterizations

The HDNNPs used in this work were constructed and trained with the n2p2 package¹²⁷. Fifty polynomial symmetry functions (p2 type), comprising both radial and angular components, were used with a cutoff radius of $r_c = 7 \text{ \AA}$. The parameters of the symmetry functions are listed in Table 7 of the Supplementary information A.2. The atomic neural networks have a topology of 50-60-1 and employ the softplus function as the first activation function and the linear function for the output layer.

We have trained a committee of eight HDNNPs on a reference dataset of 5,342 structures containing 138, 139, and 140 atoms. Each of the eight committee members was initialized with different weights and biases, and a randomly chosen 10% of the dataset was used as a validation set. HDNNPs committee reproduces the energies and forces of the training set with a root mean squared error (RMSE) of 2.25 meV/atom and 45.5 meV/ $\text{\AA}$, respectively. Parity plots for energy and forces can be found in Fig. 36 of the Supplementary information A.2.

The dataset and the trained models are available on the public repository Zenodo¹⁶⁰.

Molecular Dynamics Simulation and Analysis

All HDNNP-MD simulations were performed using the large-scale atomic/molecular massively parallel simulator (LAMMPS)¹⁶¹. We simulated at three temperatures, namely 304 K, 330 K, and 359 K, to determine the temperature-dependent diffusion properties of the monovacancy. For each temperature, we carried out two independent MD runs. They add up to 0.4 μs of simulation time overall for 304 K and 330 K, and 0.3 μs for 359 K. Each simulation was first equilibrated for 1 ns in the canonical ensemble (NVT) with the Langevin thermostat, followed by a run in the microcanonical ensemble (NVE) for the

remaining simulation time ($0.2 \mu\text{s}$). We employed the velocity Verlet integrator with a time step of 1 fs. Only the NVE segments of the simulations were used to determine hopping rates. Diffusion coefficients were determined from the mean square displacement of the defects according to the Einstein formula using the MDanalysis code^{162,163}. A 9 ns/day simulation time has been realized utilizing 4 cores (AMD Ryzen 9 5950X CPU) for a supercell containing 139 atoms.

3.3 Results

Black phosphorene exhibits a non-planar honeycomb lattice comprising two sublattices. Phosphorus has a $3s^2 3p^3$ electronic configuration, and each atom forms three covalent bonds with its nearest neighbor, one of which lies on the different sublayer, as illustrated in Fig. 13.

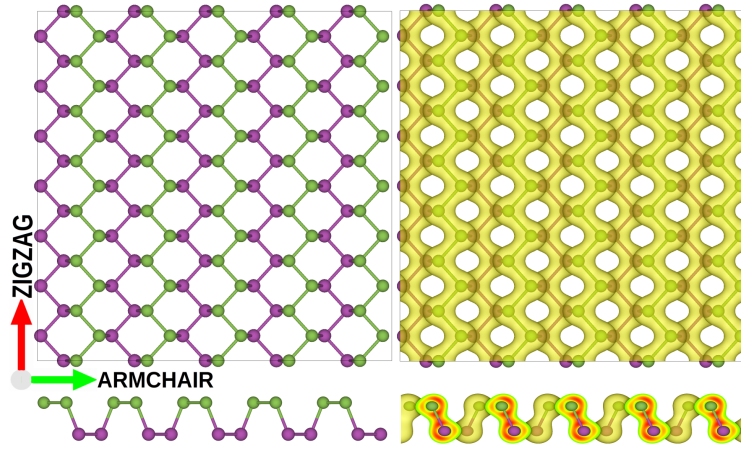


Figure 13: Left panel: Phosphorene supercell employed in our calculations, in top and side view. Atoms are colored in green and purple according to the sublattice to which they belong. Arrows indicate the directions for the zigzag and armchair. Right panel: Same structure as on the left with an isosurface of the total electronic charge distribution superimposed, again in top and side view. The colors indicate different values of the electronic charge density.

3.3.1 Monovacancy

The single vacancy defective structure SV(5|9), shown in Fig. 14, is obtained by removing one atom from the pristine structure and relaxing it. Here, the notation (5|9) refers to the two distorted 5- and 9-membered rings of atoms comprising the vacancy. According to our DFT calculations, the SV(5|9) defect has a formation energy of 1.65 eV and a total magnetic moment of $1 \mu_{B_m}$, in good agreement with previous results^{79,84,145,164–168}. The magnetism arises from a dangling bond atom, which forms only two bonds instead of three. Consequently, there is an unpaired electron, mainly localized in the vicinity of the broken bond (see Fig. 14). The density of states shows an occupied

peak above the Fermi energy near the bottom of the valence band. It corresponds to a defect state and represents the possibility of the atom with a broken bond to host another electron.

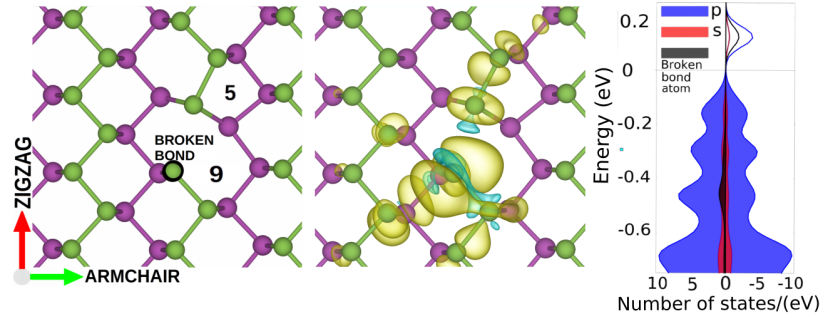


Figure 14: Left: Single vacancy SV(5|9) obtained by removing one atom from the top layer (green) and relaxing the energy. The black circle indicates the atom that takes part in only two bonds. Center: Difference between the total electronic charge densities of the spin-up (yellow) and spin-down components (cyan). Right: The electronic density of states, the positive and negative values representing the spin-up and spin-down components, respectively.

Mechanisms of diffusion

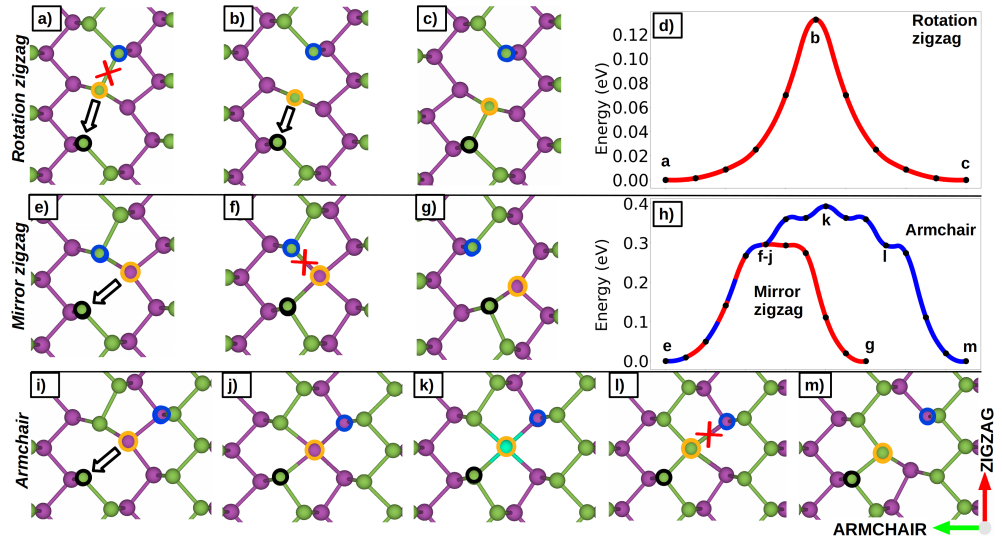


Figure 15: Mechanisms of diffusion of the monovacancy and corresponding ciNEB energy path. The black circle surrounds the atom with an initial unsaturated bond, whereas the yellow one indicates the atom that breaks a bond during the process, and the blue circle is where the vacancy ends the transition. a)-d) *rotation zigzag*. e)-h) *mirror zigzag*. i)-m) *armchair*. The *mirror* and *armchair* transitions undergo the same mechanism until the intermediate state f), where a four-folded atom is formed.

By performing DFT-MD calculations at 300 K, we have observed transitions of the monovacancy between different sites along the zigzag

direction. We call this transition a *rotation* since the final defect configuration can be obtained by applying a 180° rotation to the initial one. The mechanism for this transition is shown in Fig. 15(a)-(c). The black-circled atom has one unpaired electron and to saturate its p shell it attracts the yellow one. It leads to the breaking of the bond between the blue and yellow atoms. This hopping event can occur back and forth multiple times. At the transition state (b), there is competition between the black and the blue atoms, which exert an attractive force on the yellow one. The estimated energy barrier for this process is 0.13 eV, as inferred from the energy profile calculated with ciNEB and displayed in Fig. 15(d).

Two further types of vacancy transitions are possible toward different sites, which exhibit higher energy barriers compared to the *rotation*. Observing them in ab initio MD simulations would require prohibitively long simulation times. However, starting from the pre-adapted configuration obtained from the HDNNP-MD, allowed us to correctly model and reproduce these transitions with first-principle accuracy.

The so-called *mirror* transition, shown in Fig. 15(e)-(g), occurs along the zigzag direction as well, and the orientation of the final configuration can be obtained by applying a *mirror* operation to the initial one. The ciNEB estimated energy barrier is 0.3 eV, as indicated in Fig. 15(h). One phosphorus atom forms four bonds at the transition state, Fig 15(f). Starting from this configuration, a transition can occur following two distinct mechanisms: 15(e)-(g) we just discussed, and 15(i)-(m). In the latter, the four-folded atom can move along the direction perpendicular to the sheet to the other atomic sub-layer, Fig. 15(j). The associated energy barrier is estimated to be 0.1 eV. From here, the monovacancy can hop to two different sites along the armchair direction, Fig. 15(k) shows one possible transition state. Therefore, the overall process has an energy barrier of 0.4 eV, Fig. 15(l). This diffusion mechanism is referred to as *armchair* and is the only route for the monovacancy to switch from one sub-layer to another. Table 4 provides an overview and comparison of energy barriers with other works.

As a result, the combination of these three transition mechanisms allows the SV(5|9) to diffuse along every direction of the lattice across the whole monolayer.

3.3.2 Monovacancy tracking

Keeping track of the location of the monovacancy is necessary to assess its kinetic properties. It can be easily done for DFT-MD because of the local magnetic moment, which clearly distinguishes an atom with an unsaturated bond. For instance, Fig. 16 (top) shows two *rotation* hopping events between two atoms and the corresponding exchange of local magnetic moments.

Table 4: Energy barriers for monovacancy diffusion obtained by ciNEB and derived from HDNNP-MD (as explained in the section *Kinetic properties*), together with values reported in other works.

Energy barrier (eV)	<i>rotation</i>	<i>mirror</i>	<i>armchair</i>
NEB	0.13	0.30	0.39
HDNNP-MD	0.09	0.25	-
Hu ¹⁴⁵	-	-	0.40
Hu ¹⁶⁵	0.18	-	0.38
Cai ⁸⁴	0.18	0.30	0.40
Li ¹⁶⁶	-	-	0.40
Vierimaa ⁷⁹	0.09	0.25	-
Gaberle ¹⁶⁷	0.25	0.44	0.57
Rijan ¹⁶⁸	-	0.31	-

Within HDNNP-based simulations, the magnetic moment is inaccessible, and visual detection of monovacancy positions in long MD consisting of millions of timesteps is impractical. Here, we detect the position of the monovacancy exploiting a particular property of HDNNPs with minimal computational overhead. In HDNNPs, the total energy is calculated as the sum of local atomic energies (see Eq. 19). When a monovacancy is introduced into a material, its formation energy is spread around the monovacancy. Most of the formation energy is attributed to the atom with the unsaturated bond. By simply finding the atom with the highest atomic energy, we identify the position of the monovacancy in the material for any given configuration. A hopping process is then signaled by a change in the index of the atom with the highest energy, and the type of transition can be inferred from the relative positions of the new and old atoms.

Figure 16 compares the two techniques for the same MD trajectory. Atomic energy exchanges between atoms in the HDNNP-MD occur simultaneously as local magnetic moment swaps in the DFT-MD, validating our new approach.

3.3.3 Kinetic properties

Table 5: Hopping rates at 300 K and pre-exponential factors for MV diffusion derived from HDNNP-MP compared with previous work⁸⁴. All numbers are given in units of ns⁻¹.

	<i>rotation</i>		<i>mirror</i>		<i>armchair</i>	
	ν	ν_s	ν	ν_s	ν	ν_s
HDNNP	71	1800	0.035	180	0.013	-
Cai ⁸⁴	2.5	2200	0.021	2400	3.1×10^{-5}	160

Although energy barriers offer some insight into the kinetics of monovacancies, they are insufficient for the estimation of hopping rates ν due

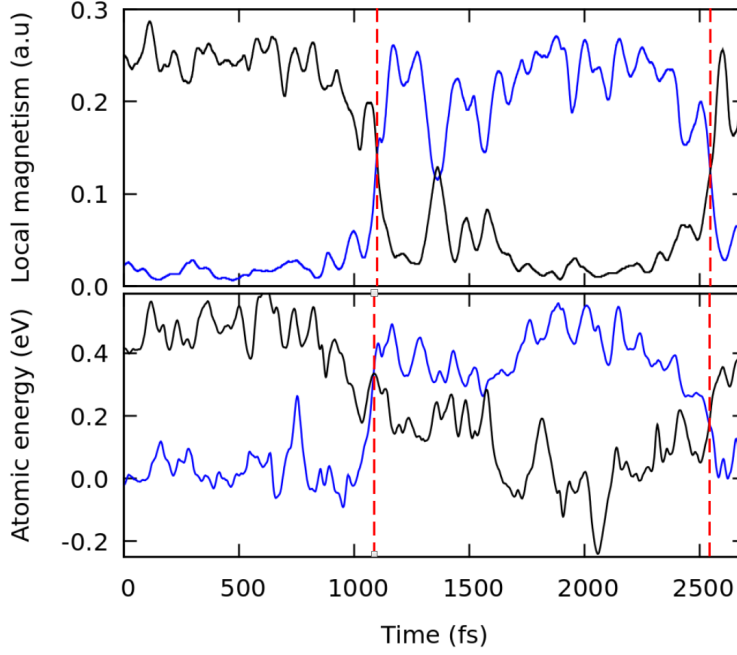


Figure 16: Monovacancy hopping can be detected by changes in either the local magnetic moments for DFT-MD (top) or the atomic energies for HDNNP-MD (bottom). The blue and black lines show the local magnetic moments and atomic energies for two atoms involved in the hopping of the defect. Both methods predict hopping at the same time on the same trajectory. The average atomic energy is subtracted from the atomic energies.

to the unknown pre-exponential factor ν_s , also called the characteristic frequency, in the Arrhenius equation:

$$\nu = \nu_s e^{-E_b/k_B T}. \quad (24)$$

The pre-exponential factor ν_s can be estimated in the harmonic approximation from the atomic vibrational frequencies in the energy minimum and the transition state associated with the motion of the MV, using the Vineyard formula¹⁶⁹. The hopping rate can also be obtained directly from an MD simulation without using the Arrhenius equation. If a given type of hopping mechanism is observed many times during an MD simulation, the hopping rate is simply given by the number of occurrences divided by the total simulation time. The hopping rates estimated directly from HDNNP-MD are shown in Fig. 17. The reported values at room temperature for zigzag *rotation*, zigzag *mirror*, and *armchair* hopping rates are 71 ns⁻¹, 0.035 ns⁻¹, and 0.013 ns⁻¹, respectively.

Although we also report highly itinerant atomic vacancies as previous work⁸⁴, a quantitative agreement in hopping rates is found only for the *mirror* transition. Our hopping rates are higher by three orders of magnitude for the *armchair* transition and one order of magnitude for the *rotation* transition. To reveal the cause of this discrepancy, we have factorized the hopping rates via the Arrhenius equation (Eq.

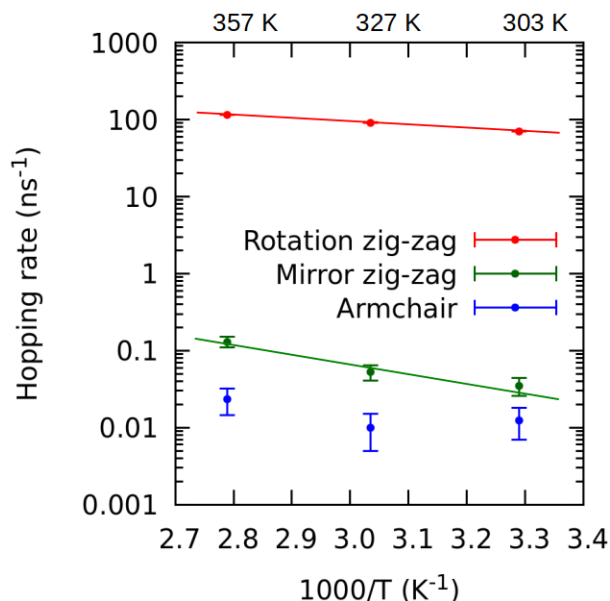


Figure 17: Arrhenius plot of the hopping rates for each of the three diffusion mechanisms at 304 K, 330 K, and 359 K, as determined by HDNNP-MD.

24) into the pre-exponential factor (Tab. 5) and the exponential factor depending on the energy barrier (Tab. 4). The prefactors agree very well for the *rotation* mechanism. The difference in hopping rates comes from the energy barrier, which is significantly smaller in our work (0.09 vs. 0.18 eV). The previously reported prefactor for the *mirror* mechanism is an order of magnitude larger than ours but also accompanied by a higher energy barrier (0.25 vs. 0.30 eV). These two effects compensate for each other, leading to similar hopping rates. Insufficient statistics prevent us from estimating the activation energy and factorizing the rate for the *armchair* transition, so the discrepancy cannot be explained in this case.

According to our results, the diffusion coefficient of a monovacancy in the in-plane direction is significantly greater than that of a divacancy and a tetravacancy. We determined the diffusion coefficient for the monovacancy to be 10^{-11} m²/s at 359 K. Experimental measurements report the diffusion coefficients for divacancy and tetravacancy to be 10^{-19} - 10^{-20} m²/s at 573 K⁸². This difference can be attributed to the size of the diffusion barrier, which is considerably lower for monovacancy, ranging from 0.1 to 0.4 eV, compared to divacancy and tetravacancy, which exhibit higher diffusion barriers in the range of 2.1 to 3.0 eV⁸².

A diffusion trajectory of the monovacancy at 328 K over a 200 ns simulation time is presented in Fig. 18. The monovacancy started at the origin and diffused through a combination of different mechanisms. Its maximum diffusion distance from the origin was 8 Å, but after 200 ns, it was only 3 Å. Within this timeframe, the monovacancy underwent 18×10^3 rotations, 11 mirror, and 3 armchair transitions. Only 5 rotations

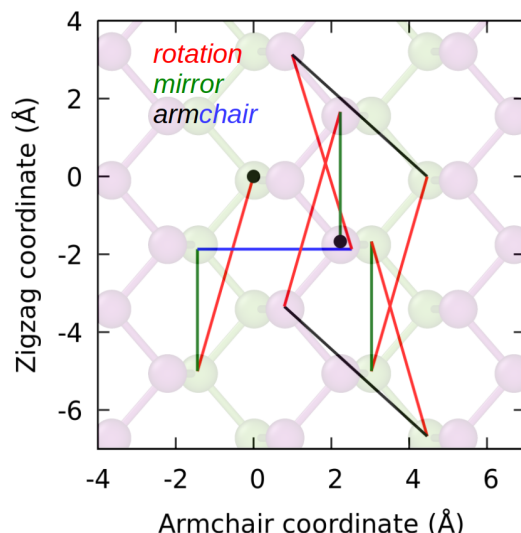


Figure 18: The trajectory of monovacancy diffusion during a period of 200 ns at 328 K, as simulated by HDNNP-MD. Each color corresponds to a specific mechanism type, with *rotation* represented by red, *mirror* by green, and the two types of *armchair* transition denoted by blue and black.

and 3 *mirror* events resulted in true displacements of the vacancy, as the remaining ones were negated by back hops. In contrast, all *armchair* transitions led to true displacements.

3.3.4 Divacancy and coalescence of monovacancies

A divacancy is created by removing two atoms from the pristine structure. Since all bonds are fully saturated in this case, there is no magnetism overall. The most stable divacancy has a formation energy of about 1.36 eV⁸⁴, which is significantly lower than the 3.3 eV formation energy of two separated SV(5|9). Such a significant energy difference and high monovacancy mobility imply that monovacancies tend to coalesce quickly. However, the coalescence of two monovacancies into one divacancy has not yet been documented.

Monovacancies can approach one another from different directions and with other configurations, leading to various types of mechanisms and divacancy conformations. Because of the complexity of the mechanism, MD is the appropriate method for studying vacancy coalescence. Since there is no prior knowledge of the coalescence of monovacancies, we introduce 50 initial arrangements differing in both distance and mutual orientation of the vacancies and let the system evolve with HDNNP-MD. In this process, many new structures were added to the reference dataset for training the HDNNP. Although we found many metastable states with formation energies of around 2.5 eV, our interest is in coalescence into stable divacancies with formation energies of 1.3–1.4 eV. Figure 19 shows the three initial arrangements of monovacancies and mechanisms of coalescence into the DV(5|8|5) with a

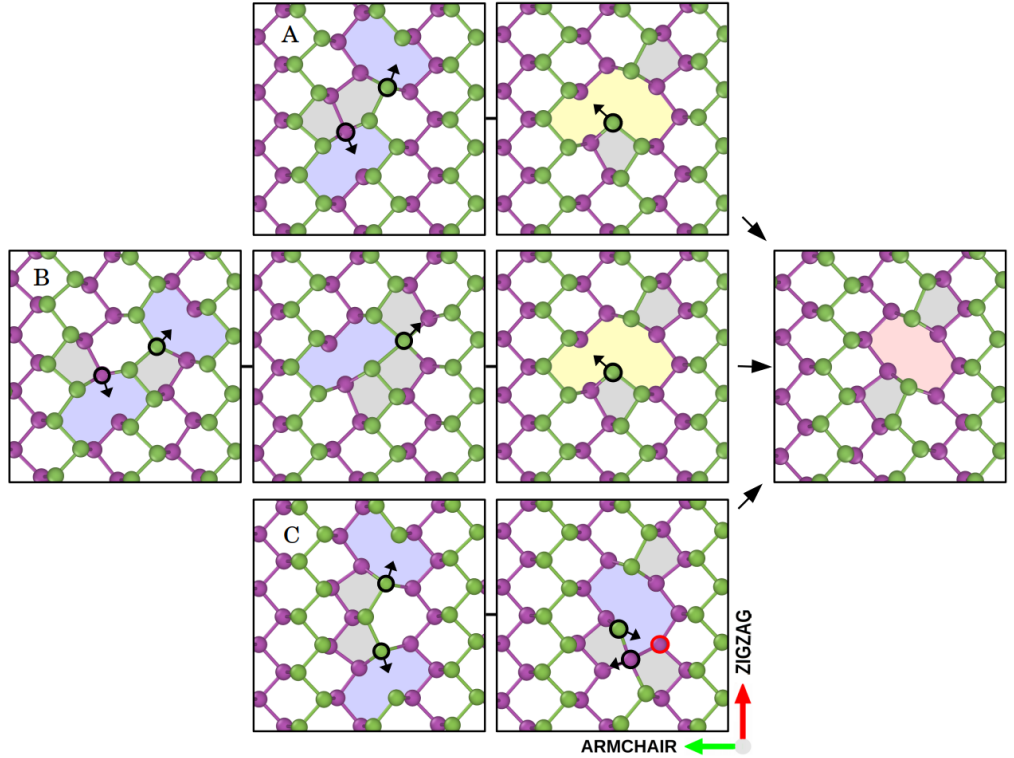


Figure 19: Divacancy formation mechanisms. Initial configurations of the two monovacancies that eventually coalesce to form the DV(5|8|5) are represented by the letters A, B, and C. Transitions A and B share the same last intermediate state before coalescence. The black arrows indicate the future direction of atomic movement. Colored areas indicate rings formed by more or less than 6 atoms.

1.4 eV formation energy. All three mechanisms have been identified by HDNNP and validated by DFT. The mutual orientation of monovacancies is identical except for the relative position in the armchair direction. This finding is crucial because diffusion is less likely to occur in this orientation than in a zigzag.

Transitions A and B share the last intermediate state before coalescence but go through different transition states in the initial stage. Mechanism A involves only two *rotation* transitions. In contrast, mechanism B needs the *armchair* transition in addition to the *rotation* transition. Two *rotations* are necessary also in the C transition. Based on the hopping sequence of the atoms, one of the black-labeled atoms forms four bonds, and the second one is left with two bonds. The divacancy is finally obtained when the four-bonded atom breaks the bond with the black atom, and the bond is formed with the other red atom.

The coalescence mechanism A is the fastest because only the *rotation* barrier, at 0.13 eV, needs to be crossed. Coalescence C initially only needs a *rotation*, but it takes a longer time than A because of the bond exchange of the black atoms with the red atom. Nevertheless, the process can be observed in a few ns. Since it requires an *armchair*

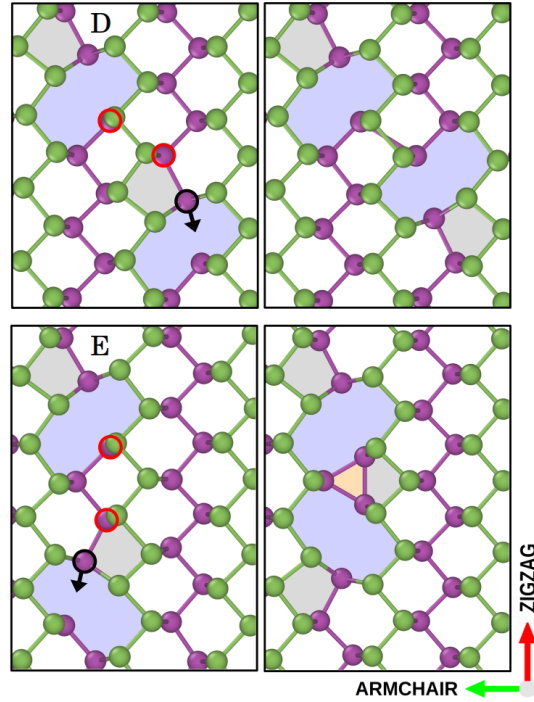


Figure 20: Metastable monovacancy pair formation mechanisms. MV pairs are formed if atoms with broken bonds (red circles) get close enough to make a bond. Mechanism E has also been reported in previous work¹⁷⁰.

transition with a 0.4 eV energy barrier, coalescence B proceeds the slowest of all.

The resulting DV(5|8|5) in Fig. 19 was experimentally observed as the most stable DV⁸² even though previous computational studies indicate that DV(5|7|7|5) should be more stable^{82,84}. Yao et al. attribute this to non-equilibrium conditions brought on by energetic e-beam irradiation. While the relative stability of DV(5|8|5) and DV(5|7|7|5) has not been resolved, the transition mechanisms between these two double vacancies and DV diffusion have been analyzed theoretically and experimentally⁸².

The initial orientation of atoms with dangling bonds plays an important role in the coalescence process. When these atoms are close to one another, they instantly form a bond that hinders further evolution into more stable divacancies. Two examples of this mechanism are shown in Fig. 20. Only *rotation* transitions of black atoms bring the two bonded red atoms together, which causes a bond to form right away. These states could be considered MV pairs rather than divacancies since they maintain MV shapes. Their formation energy, 2.3 eV, falls exactly between the stable divacancy (1.4 eV) and the two separated monovacancies (3.3 eV).

3.4 Conclusions

In this paper, we thoroughly analyzed the mobility of the phosphorene monovacancy. DFT-MD and HDNNP-MD have successfully reproduced all three previously identified diffusion mechanisms. The energy barriers calculated statically from DFT with the nudged elastic band method and dynamically from HDNNP-MD are in good agreement. MD simulations that are longer than a microsecond in total have not revealed any novel mechanisms. By assessing hopping rates directly from MD, we confirm the highly itinerant nature of the phosphorene monovacancy even at room temperature, preferentially along the zigzag direction.

We identified three mechanisms for the coalescence of two monovacancies into one divacancy, resulting in the stable DV(5|8|5) conformation with a formation energy of 1.4 eV. Coalescence can be observed on the ns timescale if MVs are close enough and ideally oriented. However, if the atoms with the dangling bond approach one another too early, a metastable MV pair rather than fully coalesced DV is produced, emphasizing the significance of the initial configuration. Monovacancies are predicted to be the most common defect after electron and ion irradiation. But since they quickly coalesce, their potential utilization poses a challenge, suggesting turning attention to more stable divacancies.

Acknowledgement

We acknowledge financial support by the Doctoral College Advanced Functional Materials – Hierarchical Design of Hybrid Systems DOC 85 doc.funds funded by the Austrian Science Fund (FWF) and the Vienna Doctoral School in Physics (VDSP) of the University of Vienna. The computational results presented have been partly achieved using the Vienna Scientific Cluster (VSC).

ELECTRON-BEAM-INDUCED ADATOM-VACANCY COMPLEXES IN MONO- AND BILAYER PHOSPHORENE

Speckmann, C., Angeletti, A., K yvala, L., Lamprecht, D., Herterich, F., Mangler, C., Filipovic, L., Dellago, C., Franchini, C., and Kotakoski, J. (2024). *Electron-beam-induced adatom-vacancy-complexes in mono- and bilayer phosphorene*. Submitted to Physical Review X. The American Physical Society (APS). DOI: 10.48550/arXiv.2409.11102

Phosphorene, a puckered two-dimensional allotrope of phosphorus, has sparked considerable interest in recent years due to its potential, especially for optoelectronic applications with its layer-number-dependant direct band gap and strongly bound excitons. However, detailed experimental characterization of its intrinsic defects and its defect creation characteristics under electron irradiation are scarce. Here, we report on the creation and stability of a variety of defect configurations under 60 kV electron irradiation in mono- and bilayer phosphorene, including the first experimental reports of stable adatom-vacancy complexes. Displacement cross section measurements in bilayer phosphorene yield a value of 7.7 ± 1.4 barn with an estimated lifetime of adatom-vacancy-complexes of 19.9 ± 0.7 s, while some are stable for up to 68 s under continuous electron irradiation. Surprisingly, ab initio-based simulations indicate that the complexes should readily recombine, even in structures strained by up to 3 %. The presented results will help to improve the understanding of the wide variety of defects in phosphorene, their creation, and their stability, which may enable new pathways for defect-engineered phosphorene devices.

4.1 Introduction

Semiconducting two-dimensional (2D) materials have been a research focus for some time as they provide the potential to create smaller and, more importantly, the thinnest possible devices in nanoelectronics^{171,172}. One of the most interesting materials is phosphorene, a single layer of black phosphorus (BP), a layered material where layers show a puckered, hexagonal structure with lattice parameters of 3.3 Å and 4.53 Å along the zig-zag (ZZ) and armchair (AC) directions, respectively¹⁷³. BP is also the thermodynamically most stable allotrope of phosphorus¹⁷⁴. Mono- or few-layer phosphorene has recently received increasing attention for its electronic and optoelectronic application

potential^{173,175–178}. This is due to its thickness-dependent direct band gap, which ranges from 1.73 eV in the monolayer to 0.35 eV in the bulk¹⁷⁹, mechanical flexibility¹⁸⁰, and its strongly bound excitons with binding energies as high as 0.9 eV⁷³. More importantly, due to its unique atomic structure, phosphorene shows interesting anisotropic properties, such as Raman mode sensitivity on polarization angle and sample orientation¹⁸¹, an anisotropic charge carrier effective mass^{74,182}, a negative Poisson's ratio¹⁸³, anisotropic bright excitons^{73,184}, as well as an anisotropic electrical and thermal conductance^{185,186} that could all be used for tailored applications. However, phosphorene is extremely sensitive to oxidation, leading to rapid degradation within hours or even minutes of being exposed to ambient conditions^{187,188}, complicating both the interpretation of measurements where samples were exposed to oxygen during the process as well as the potential of its applications.

It is well known that defects influence the properties of materials, which is why a thorough understanding of the creation and effects of these defects is needed. Density functional theory (DFT) calculations for the knock-on displacement thresholds in phosphorene under electron-irradiation indicate that beam damage would only occur with an elastically transferred kinetic energy of ≥ 6 eV for a pristine sample and ≥ 4.5 eV for a sample with defects⁷⁹; however, experimentally, this has not been studied yet. In addition, recent studies show that in semiconducting and insulating 2D materials, beam damage can occur below the knock-on threshold due to electronic excitations^{189–191}. This can also be expected to take place with phosphorene. In the case of vacancies, theory predicts monovacancies to induce a magnetic moment of $1 \mu_B$ ^{165,192} as well as an in-gap state^{81,192,193}, whereas divacancies are predicted to show neither^{81,192}. Defects have also been suggested to serve as nucleation sites for surface oxidation when exposed to oxygen or water vapor¹⁹⁴, while, in combination with strain, they can also lead to structural phase transitions serving as catalyzing sites for new phases¹⁹⁵. Experimental characterization of defects in mono- and few-layer phosphorene so far was mainly done using scanning tunneling microscopy (STM)^{196–202} and transmission electron microscopy (TEM)^{81,82,203,204}. A layer-by-layer thinning of multilayer phosphorene under 80 kV electron irradiation was reported for pristine samples^{203,204} and graphene/phosphorene heterostructures⁸¹ where for all cases, a (reconstructed) ZZ edge configuration during the thinning process was found to be preferred over an AC configuration. Furthermore, a detailed study of specific defect configurations in mono- and few-layer phosphorene at elevated temperatures of 523–673 K has reported the presence of divacancies, tetravacancies, and long defect lines, along with their most stable configurations, diffusion behavior, as well as possible transformation mechanisms⁸². Experimental observations of monovacancies in TEM are limited to a recent study claiming to have observed a monovacancy for the first time with this method⁸³,

whereas several studies were already published on the observations of monovacancies using STM^{196–199}. However, similar features in STM have also been attributed to substitutional tin impurities rather than phosphorus monovacancies^{200–202}.

In this study, we use scanning transmission electron microscopy (STEM) high-angle annular dark field (HAADF) imaging to record the effects of a 60 keV electron beam on a bilayer phosphorene sample, characterizing the created defects and their dynamics. In addition to previously reported vacancy configurations, we further identify a number of yet unreported defects. Notably, almost all vacancies show adjacent phosphorus interstitials or adatoms, which is the first observation of such defects in phosphorene. DFT was used to relax models of the observed structures, followed by STEM simulations to verify their correct identification. We estimate the displacement cross section of bilayer phosphorene under 60 keV electron irradiation to be 7.7 ± 1.4 barn with an estimated lifetime of adatom-vacancy-complexes under electron irradiation of 19.9 ± 0.7 s and a maximum observed lifetime of up to 68 s. Additionally, we give a rough estimate for the cross sections of the individual phosphorene sublayers towards and away from the electron source for the farther phosphorene layer of the bilayer to be 2.0 ± 0.3 barn and 10.4 ± 1.6 barn, respectively. For the other layer, much fewer defects were observed. Molecular dynamics simulations, accelerated by high-dimensional neural network potentials (HDNNP-MD), were employed to assess the stability of the adatom-vacancy-complexes both under equilibrium conditions, as well as under isotropic and anisotropic strain/stress of up to 3 %. However, all of these simulations resulted in recombining adatoms and defects far below the observed lifetimes.

4.2 Results and Discussion

The samples were mechanically exfoliated in an Ar atmosphere and brought into a vacuum without air exposure. For details on the sample preparation, we refer to the Methods section. A typical clean area of the sample can be seen in Fig. 21a, showing an AB stacked bilayer of phosphorene surrounded by (brighter) contamination. The exact composition of this type of contamination is unknown, but previous studies indicate a mixture of different hydrocarbons^{205–208} as well as polymer residue from the transfer process^{208,209}. The effect of electron irradiation on this structure can be seen in Fig. 21b, where defects were created in the area marked with the green square after ~ 8.5 s. The area within the green square is magnified in Fig. 21c, where a red arrow highlights a missing phosphorus atom and a cyan arrow a phosphorus interstitial or adatom. To verify the interpretation of this feature as an additional phosphorus atom, which is located next to a vacancy, which from now on will be called an adatom-vacancy-complex, we used DFT

first to relax a model of the observed structure followed by STEM image simulations. We find a good agreement between the simulated STEM image of the relaxed structure (Fig. 21d) and the experimental image. Details on the relaxation and the simulation process can be found in the Methods section. To our knowledge, this is the first experimental verification of an adatom-vacancy complex in phosphorene, which has previously only been predicted by theory⁷⁹. Previous (S)TEM studies on defects in phosphorene^{81,82,203,204} never mentioned the existence of phosphorus adatoms, which might be due to the lower electron acceleration voltage of 60 kV used here and/or because the sample was not heated during the experiments in this study and/or due to the low pressure in the objective area of the microscope ($\sim 10^{-10}$ mbar).

In Fig. 21e-h, starting with panel e, we show similar behavior for a small patch of monolayer phosphorene on the bottom of the image. This monolayer flake was created by prolonged electron irradiation, utilizing the previously reported layer-by-layer thinning^{203,204}. In most cases, this process created a pore before a monolayer became visible but occasionally resulted in a monolayer, allowing its imaging.

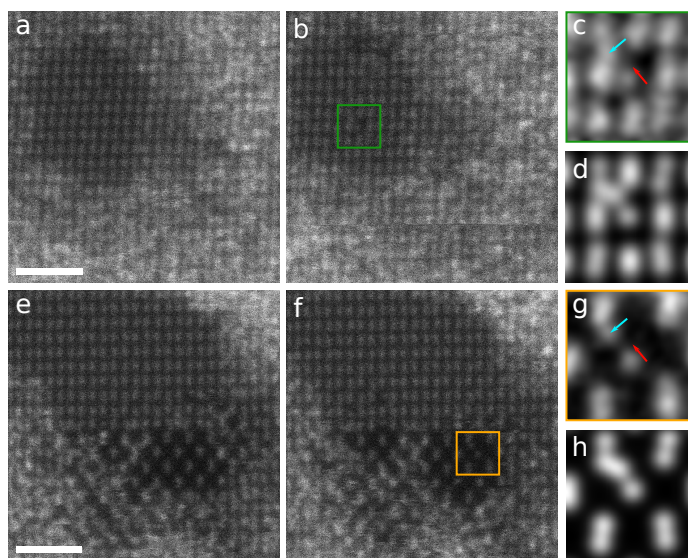


Figure 21: Creation of adatom-vacancy-complexes in bi- and monolayer phosphorene. Two STEM-HAADF image pairs of bilayer phosphorene (a-b) and a monolayer patch of phosphorene (e-f) recorded with $1024 \text{ px} \times 1024 \text{ px}$ and a frame time of $\sim 8.5 \text{ s}$. A zoomed-in image of the area marked with a green (orange) square in b (f) after applying a Gaussian blur of 6 px (4 px) is shown in c (g) where cyan and red arrows denote adatoms and vacancies, respectively. Simulated STEM-HAADF images of an adatom-vacancy complex in the bilayer (monolayer) phosphorene are shown in d (h). The scale bars are 1 nm.

In Fig. 21g, showing a magnified view of the area marked with an orange square in Fig. 21f, an adatom-vacancy-complex that was created via electron irradiation can be seen. Those were again verified using DFT and STEM image simulations, visible in Fig. 21h, where we can see

an excellent agreement between experiment and simulation. Given the low theoretical energy barrier for point defect diffusion in phosphorene (monovacancies: 0.09 eV⁷⁹ - 0.25 eV¹⁹³ along ZZ direction; 0.38 eV¹⁶⁵ - 0.44 eV¹⁹³ along AC direction; adatoms: 0.15 eV¹⁹³ - 0.23 eV⁷⁹ along ZZ direction; ≥ 1.19 eV⁷⁹ along AC direction) and therefore their expected extreme mobility at room temperature, it is surprising to find a phosphorus monovacancy and phosphorus adatom to be stable for the several milliseconds, required to record its feature in this image. This is especially intriguing, as only one experimental TEM study claims to have observed a monovacancy in phosphorene⁸³, with unfortunately somewhat ambiguous images.

To further estimate the probability of creating a phosphorus vacancy in the phosphorene bilayer, we identified all defects created during our measurements and correlated those with the electron doses applied to the sample between each appearance of a new defect. We then used Poissonian statistics as described in Ref.²¹⁰ to get a mean number of electrons needed for the creation of a single defect (see Fig. 22a). The bar plots, which required a higher electron dose to be created at each electron dose, are shown at each electron dose. According to Poisson statistics, this should show an exponential decline, as is the case here. The mean electron dose corresponding to the data for all defects in layer 1 yields a cross section of 7.7 ± 1.4 barn.

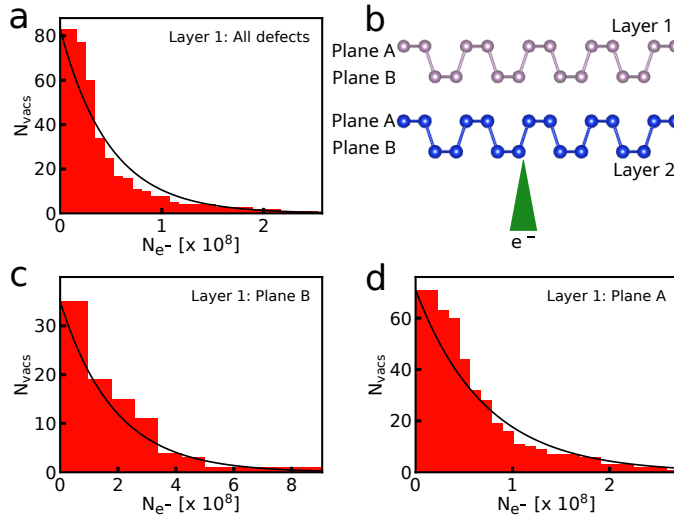


Figure 22: Statistics of the number of created defects in bilayer phosphorene. Panel b shows a schematic side view illustration of the phosphorene bilayer structure, indicating the different layers (colored in light purple and blue) and planes in the structure with respect to the incident electron beam. Panels a,c, and d exhibit the cumulative statistics for the number of electrons needed for the creation of a single defect, including all defects of layer 1 in panel a, only the defects in one plane (B) of layer 1 in panel c and those of the other plane (A) in panel d. Here, each bin indicates the number of defects created with an electron dose higher than the current one. Additionally, an exponential decline (shown in black) was fitted to the data (see text).

Note that most defects observed during these experiments were found in one of the two layers, indicating different cross sections for the individual layers. Following the previous argument by Vierimaa et al.⁷⁹, it is justified to assume that the layer hosting most of the defects is the one further away from the electron source (labeled layer 1 in Fig. 22b). To estimate the cross sections of the individual phosphorene sublayers (labeled planes A and B in Fig. 22b), we assigned the defects observed in the experiments to their distinct planes to test the hypothesis that one of the planes has a significantly higher cross section compared to the other. The results and statistical analysis are shown in Fig. 22c,d. Possible deviations from perfect Poissonian statistics could be explained by the thermal migration of some defects between the planes or due to the electron beam. The mean values obtained via these fits yield cross sections for the two planes of 2.0 ± 0.3 barn and 10.4 ± 1.6 barn for panels c and d, respectively.

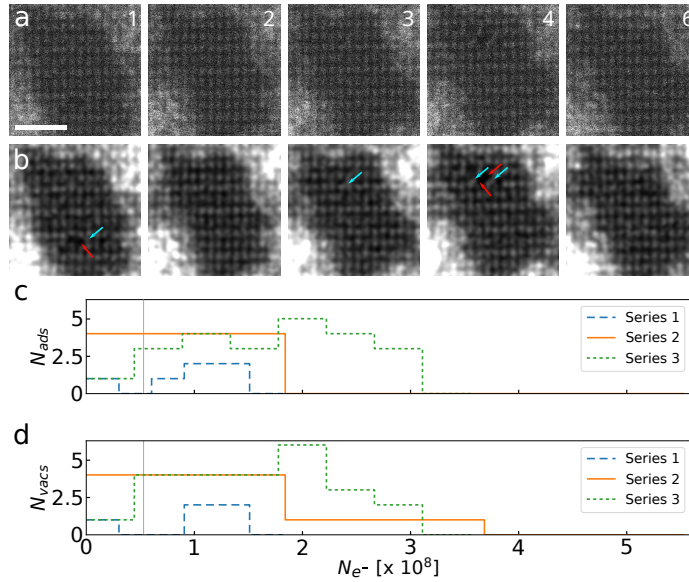


Figure 23: STEM-HAADF image series of bilayer phosphorene. Panel a shows five as-recorded images of an image series recorded with $512 \text{ px} \times 512 \text{ px}$ and a frame time of $\sim 2.2 \text{ s}$, where the frame number is indicated in the top right corner of each image. Frame 5 was omitted as it displayed the same structure as frame 4. A Gaussian blur (4 px) was applied to the images in panel b, and cyan and red arrows were added to indicate adatoms and vacancies, respectively. Panels c and d show the evolution of the number of adatoms and vacancies as a function of the total number of electrons impinging on the sample for three example image series. The vertical grey lines indicate the average number of electrons needed to create one defect in bilayer phosphorene. The dashed blue line labeled "Series 1" corresponds to the images shown in a and b. The scale bar is 1 nm.

It is possible to estimate the displacement energy threshold based on these cross sections by assuming a pure knock-on process using the McKinley-Feshbach-formalism²¹¹, while also considering the lattice vibrations^{212–215} using the Debye model with a Debye temperature

of 278.66 K for phosphorene²¹⁶. Those result in a threshold value of 4.79 ± 0.03 eV for the whole structure while the plane cross sections yield 5.01 ± 0.03 eV and 4.73 ± 0.03 eV for plane B and A, respectively. Those values are considerably lower than previously reported theoretical displacement threshold energies of 10.5 eV in pristine phosphorene or 7 eV in phosphorene with adatom-vacancy-complexes⁷⁹, which suggests that inelastic scattering effects can not be neglected for phosphorene, similar to MoS₂ and hBN^{189–191}.

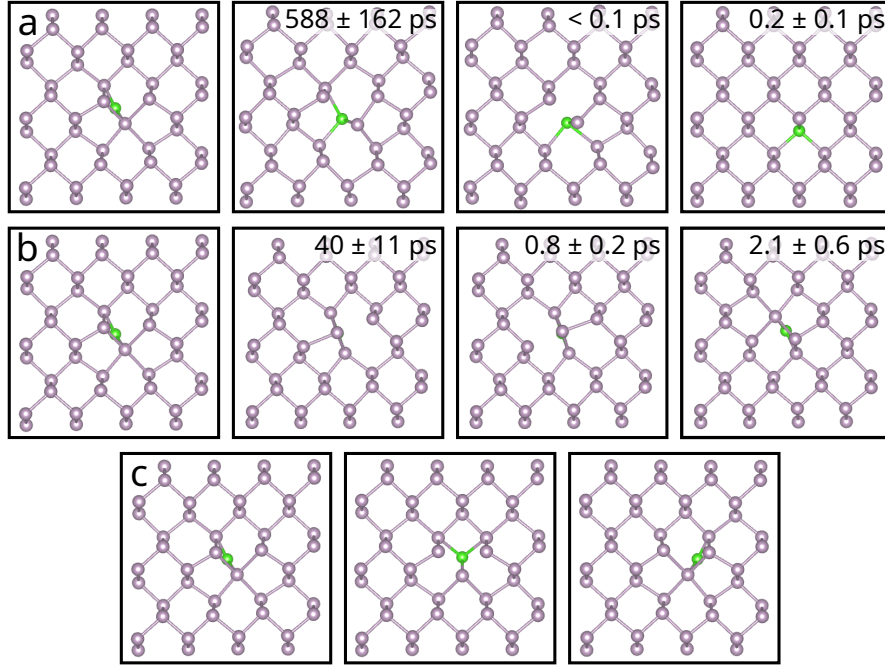


Figure 24: Transition mechanisms in monolayer phosphorene adatom-vacancy-complexes observed via HDNNP-MD and ci-NEB. A recombination process of a monovacancy and an adatom is shown in panel a, and two different types of hopping are depicted in panels b and c. The hopping in b shows an adatom-vacancy-complex hopping as a whole, while c shows how an adatom changes its orientation concerning the vacancy. All of these transitions start from the same adatom-vacancy-complex configuration. The model for this configuration is based on the experimental observation shown in Fig. 21g and was used to simulate the corresponding STEM image in Fig. 21h. Times in the top right corner of each frame in panels a and b indicate the time between the given and the previous frame as observed in HDNNP-MD (0 % strain, 300 K). Light purple atoms indicate phosphorus atoms in the phosphorene structure, while green atoms indicate the phosphorus adatom. HDNNP-MD was used to find the transition mechanisms shown in a and b, while ci-NEB calculations were used to identify the transition in c.

As for the adatom-vacancy-complexes, we not only observe their creation but also their recombination, as can be seen in the image series shown in Fig. 23a,b. Here, an adatom-vacancy-complex is present in the first frame, which recombines before the second frame is recorded, followed by the creation of two adjacent adatom-vacancy-complexes in the fourth frame, which are stable for a few seconds before recombining

to show a pristine layer in frame six again. Panels c and d show the number of adatoms and vacancies as a function of electron dose in three distinct image series. A clear trend towards the simultaneous creation and annihilation of adatoms and vacancies can be observed, although some additional vacancies and adatoms are also occasionally seen. The observation of more adatoms compared to the number of vacancies might be explained via adatom diffusion from outside the field of view of the analyzed images, while adatoms adjacent to vacancies almost always stay close to the defect. We observe lifetimes of these complexes under electron irradiation between 2.2 s and 68 s with a mean lifetime of 19.9 ± 0.7 s. As the distribution of these lifetimes appears Poissonian, we use the same technique to determine the mean lifetime for the number of electrons needed to create a single defect. Note that lifetimes below 2.2 s could not be measured as this is the time used to record a single frame.

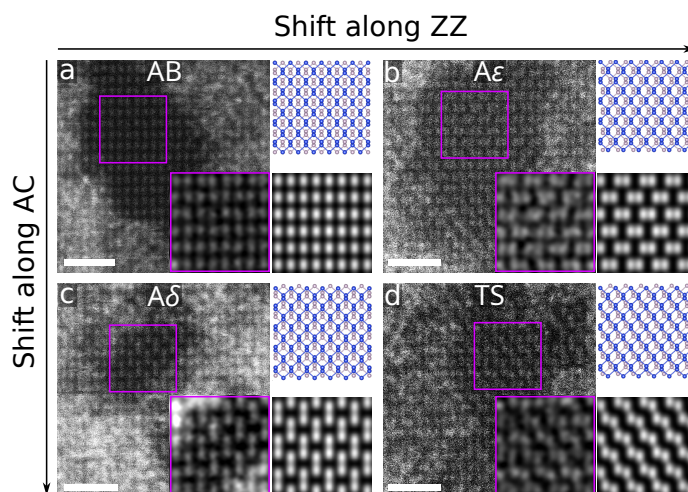


Figure 25: STEM-HAADF images of different BP stacking orders. Each panel highlights a different stacking order in bilayer phosphorene. The as-recorded image is visible on the left of each panel with a purple square indicating the zoomed-in area shown in the inset. A Gaussian blur was applied to the zoom-in. The top right of each panel features a structural model of the different stacking orders from the top view, and a STEM-HAADF simulation of this structure is shown on the bottom right. Atoms are colored in light purple and blue to indicate the two layers in the bilayer structure. Panels a and c were recorded with $1024 \text{ px} \times 1024 \text{ px}$ and a Gaussian blur of 6 px was applied in the inset while panels b and d were recorded with $512 \text{ px} \times 512 \text{ px}$ and a Gaussian blur of 4 px was applied in the inset. The scale bars are 1 nm.

We conducted HDNNP-MD simulations under various conditions to investigate the stability of the adatom-vacancy complex in monolayer phosphorene. Starting from a relaxed structure, similar to the experimental one shown in Fig. 21g, we tested different conditions, including equilibrium, and applied isotropic and anisotropic strains/stresses of 1% and 3%. In all cases, we consistently observed recombination, as shown in Fig. 24a, occurring on average within a timescale of less than

one nanosecond. Additionally, our simulations revealed two types of hopping mechanisms, one where the entire adatom-vacancy-complex is hopping and one where just the adatom is changing its orientation to the vacancy, as illustrated in Figs. 24b and c, respectively. Note that the hopping mechanism shown in Fig. 24b was already proposed by theory as a rotation along the ZZ direction for monovacancies in monolayer phosphorene, although no adatom was present for these simulations²¹⁷. Thus, the adatom-vacancy complex seems to be comparable to a vacancy without an additional adatom in terms of its migration. The transition shown in Fig. 24c is based on climbing image-nudged elastic band (ci-NEB) calculations only, revealing an energy barrier of 0.5 eV for this transition, which could not be verified by HDNNP-MD, as the adatom would always recombine with the vacancy via the transition shown in Fig. 24a before the adatom hopping could be observed. The validity of these mechanisms was confirmed through DFT calculations. Therefore, the observed stability of the complexes may be due to effects not accounted for in our simulation, such as out-of-equilibrium responses that can not be addressed at the DFT level. Possible explanations include a slight bending of the monolayer, an interaction with the bilayer interface, or a charge potential induced by STEM imaging.

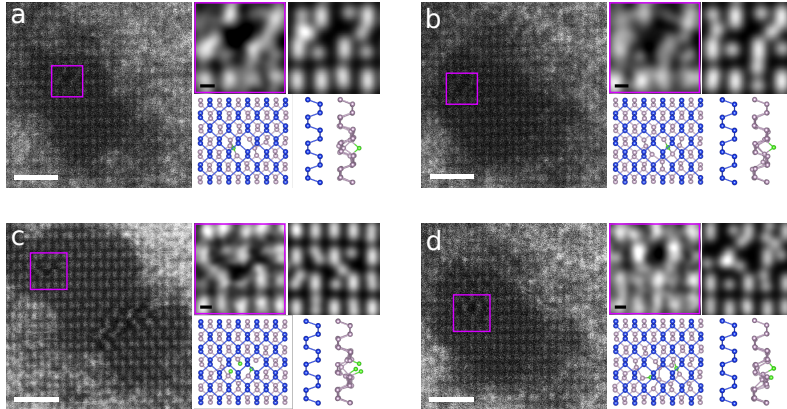


Figure 26: STEM-HAADF images of different adatom-vacancy configurations along the ZZ direction. Each panel shows a different defect configuration, including homoatomic adatoms in bilayer phosphorene. The as-recorded image is visible on the left of each panel, with a purple square indicating the area hosting a defect. A zoom-in of this area is shown after applying a Gaussian blur in the top right, framed with the same color as in the as-recorded image. The bottom right of each panel features a structural model of the defect configuration in both the top and side view, relaxed using DFT. The top right image without a frame is a STEM-HAADF simulation of this model. Atoms colored in light purple and blue indicate phosphorus atoms in the two different layers of the bilayer structure, and green atoms indicate phosphorus adatoms, respectively. Panels a, b, and d were recorded with 512 px $\times$ 512 px, and a Gaussian blur of 4 px was applied while panel c was recorded with 1024 px $\times$ 1024 px and a Gaussian blur of 6 px was applied. The scale bars are 1 nm in the as-recorded images and 1 Å in the zoomed-in images.

During measurements where the phosphorene sample was continuously irradiated with highly energetic electrons, we could also observe local shifts in the stacking order of the bilayer. As shown in Fig. 25, four distinct stacking orders could be identified. Panel a shows the known AB stacking order that was reported both for bulk BP²¹⁸ as well as thin layers¹⁸¹. The other stacking orders are characterized by a shift along the ZZ direction for the structure labeled here $A\epsilon$ (Fig. 25b), a shift along the AC direction for the $A\delta$ structure (Fig. 25c), and a shift along both directions for the TS structure (Fig. 25d). The $A\delta$ and TS stacking order have also been studied theoretically, showing their (meta-)stability as well as their influence on the electronic structure of the phosphorene bilayer²¹⁹, while we could not find any reports on the $A\epsilon$ structure in the literature. One noteworthy theoretical prediction for a transition from the AB into the TS or $A\delta$ stacking order is the induced transition from a direct to an indirect band gap, which is otherwise absent in BP. These findings mark the first experimental observation of these different stacking orders in bilayer phosphorene.

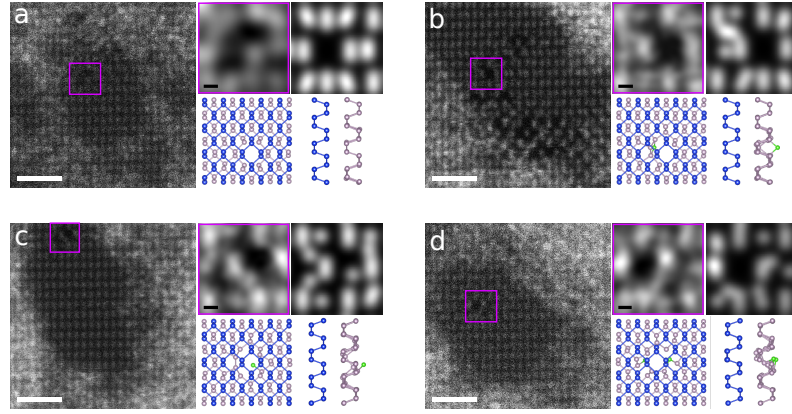


Figure 27: STEM-HAADF images of different adatom-vacancy configurations along the AC direction. Each panel shows a different defect configuration, including homoatomic adatoms in bilayer phosphorene. The as-recorded image is visible on the left of each panel, with a purple square indicating the area hosting a defect. A zoom-in of this area is shown after applying a Gaussian blur in the top right, framed with the same color as in the as-recorded image. The bottom right of each panel features a structural model of the defect configuration in both the top and side view, relaxed using DFT. The top right image without a frame is a STEM-HAADF simulation of this model. Atoms colored in light purple and blue indicate phosphorus atoms in the two different layers of the bilayer structure, and green atoms indicate phosphorus adatoms, respectively. Panels a, b, and c were recorded with $1024 \text{ px} \times 1024 \text{ px}$, and a Gaussian blur of 6 px was applied while panel d was recorded with $512 \text{ px} \times 512 \text{ px}$ and a Gaussian blur of 4 px was applied. The scale bars are 1 nm in the as-recorded images and 1 Å in the zoomed-in images.

Finally, we want to highlight the variety of defect configurations beyond monovacancies (see Fig. 21) observed in phosphorene bilayers. Fig. 26 shows a divacancy (panel a), trivacancy (panel b), and two different configurations of tetravacancies (panels c and d) along the ZZ

direction with a variable number of adatoms. Fig. 27 shows the same sequence of defects (divacancy in panel a, trivacancy in panel b, and two configurations of tetravacancies in panels c and d), this time orientated along the AC direction. Note that the structures in Fig. 26a and Fig. 26d have been previously reported without adatoms^{82,83}, while the other structures are yet unreported. Except for Fig. 27a, a reasonably good agreement between the simulated and the experimentally observed structures could be achieved, indicating the correct identification of these defects. The discrepancies for Fig. 27a might be explained by the occurrence of A δ stacking (see Fig. 25c) in the experimental image, which might influence the stability of the defect structure.

4.3 Conclusions

In conclusion, we characterized several unreported defect configurations in mono- and bilayer phosphorene, including stable adatom-vacancy complexes. The experimentally obtained displacement cross section of bilayer phosphorene under 60 kV electron irradiation was measured to be 7.7 ± 1.4 barn, with most defects being located in just one single layer. Within this layer, the two atomic planes also showed clearly distinct cross section values (10.4 ± 1.6 barn and 2.0 ± 0.3 barn, presumably for the planes away and towards the electron source, respectively). The average lifetime of adatom-vacancy-complexes was found to be 19.9 ± 0.7 s with a maximum observed lifetime of up to 68 s. Using DFT, the observation of different adatom-vacancy configurations was verified, while the extended lifetime of the adatom-vacancy-complexes could not be reproduced using HDNNP-MD, leaving the explanation for the stability of these structures as an open question that requires further investigation. These results further extend the understanding of the creation and stability of defects in phosphorene, which might lead to the creation of tailored defect-engineered phosphorene devices.

4.4 Methods

Sample preparation Phosphorene samples were produced inside an Ar glovebox with both O₂ and H₂O partial pressures below 0.5 ppm using the mechanical exfoliation technique introduced by Novoselov et al⁵¹. The bulk BP crystal was commercially bought from HQ Graphene. After exfoliation, thin flakes of BP were transferred onto a custom-made 200 μm thick silicon TEM grid with 3 $\times$ 3 windows covered by a 1 μm thick holey silicon nitride membrane using an all-dry viscoelastic stamping method²²⁰. Each window has a 5 $\times$ 5 matrix of 2 μm holes with a separation of 5 μm between holes. Samples were introduced to ultra-high vacuum (UHV) within 20 min after the transfer to the TEM grids was finished, and they were baked overnight at ca. 150°C before any measurements were conducted. Due to the unique design of

the CANVAS ultra-high vacuum system²²¹, connecting the glovebox directly to the loadlock of the system, the samples could be transferred into a vacuum (pressure $\sim 10^{-9}$ mbar) without exposure to ambient conditions, avoiding oxidation of the material to influence the experimental results. In between measurements, samples were kept in a vacuum to prevent degradation.

Scanning transmission electron microscopy The microscope used for this study is a Nion UltraSTEM 100 equipped with a cold field emission gun obtaining a probe size of ~ 1 Å. Images were recorded at room temperature using a HAADF detector with collection angles of 80 – 300 mrad at a constant electron acceleration voltage of 60 kV and a beam convergence semiangle of 30 mrad. Additionally, the number of impinging electrons on the sample was measured as explained in Ref.¹⁹⁰, resulting in a beam current of 16.5 ± 0.5 pA. The base pressure the samples were exposed to inside the microscope column was below 10^{-9} mbar at all times.

Computational methods First-principles calculations were performed using the Vienna Ab Initio Package^{150,151}, employing the same setup as in our previous study²¹⁷ with optimised lattice parameters for pristine phosphorene samples of 3.30 Å (along ZZ) and 4.62 Å (along AC) for the monolayer and 3.30 Å (along ZZ) and 4.50 Å (along AC) for the bilayer, respectively. We utilize the high-dimensional neural network potential (HDNNP)¹¹ for dynamic studies. This potential has been successfully employed for simulating the diffusion of monovacancies and coalescence of monovacancies in phosphorene in our previous work²¹⁷. Given that our current study deals with a similar problem, we employ the same set of descriptors, consisting of 50 symmetry functions, which characterize the local chemical environment of phosphorene up to a cutoff distance of 7 Å. The neural network architecture consists of two hidden layers, each with 40 nodes, utilizing the softplus activation function.

To discover the recombination mechanism, we employ an active learning strategy called query by committee⁴¹ and iteratively improve our neural network potential. This strategy has been proven effective in creating training sets for machine learning potentials^{42,43,158,159}, including those for phosphorene²¹⁷. The approach involves training several HDNNPs with different training/validation splits and initial conditions. During molecular dynamics (MD) simulation, we monitor the energy and force predictions of these committee members to identify configurations where predictions disagree. When the committee disagreement exceeds a predetermined threshold, we stop MD simulation accelerated by HDNNPs and continue it with ab initio MD, generating new configurations that are incorporated into the training set. The neural networks are then retrained on this enlarged dataset, and the MD simulation is resumed until the desired level of accuracy is achieved. The committee of 8 HDNNPs reproduces the energies and forces of the training set with a root mean squared error (RMSE)

of 0.7 meV/atom and 51.3 meV/Å, respectively. The reference data, comprising 4,904 configurations, and the trained models are publicly available on Zenodo²²².

The HDNNPs used in this work were constructed and trained with the n2p2 package¹²⁷. All HDNNP-MD simulations were conducted using the large-scale atomic/molecular massively parallel simulator (LAMMPS)¹⁶¹. We employed the microcanonical ensemble (NVE) and the velocity Verlet integrator with a time step of 1 fs.

STEM image simulations were performed using abTEM²²³, and structure models shown in the figures were created with VESTA²²⁴.

Data analysis Data was analyzed manually by identifying the different vacancy and adatom sites within each frame of each image series individually. Similar to Ref.¹⁹⁰, the uncertainty for this process will result in an increased uncertainty of the cross sections calculated with those data. This is taken into account by increasing the uncertainty obtained by fitting the distribution of defects to Poisson statistics by a factor of 2.5. Other uncertainties were deemed insignificant compared to this. To not skew the distribution of created defects by the simultaneously occurring recombination processes, image series were cut off when recombination was observed, while diffusion of defects in between frames was assumed when the number of defects in subsequent frames was identical but not their positions. To estimate the sublayer cross sections, defects in each image series were sorted according to their occurrence in the two different planes.

Acknowledgments

We acknowledge funding from the Austrian Science Fund (FWF) through the doctoral college Advanced Functional Materials–Hierarchical Design of Hybrid Systems (DOC 85) and via the projects 10.55776/DOC142 and 10.55776/P35318 as well as from the Vienna Doctoral School in Physics. The computational results presented have been partially achieved using the Vienna Scientific Cluster (VSC).

UNSUPERVISED IDENTIFICATION OF CRYSTAL DEFECTS FROM ATOMISTIC POTENTIAL DESCRIPTORS

Kývala, L., Montero De Hijes, P., & Dellago, C. (2024). *Unsupervised identification of crystal defects from atomistic potential descriptors* Submitted to npj Computational Materials. Springer Nature. DOI: 10.48550/arXiv.2405.01320

Identifying crystal defects is vital for unraveling the origins of many physical phenomena. Traditionally used order parameters are system-dependent and can be computationally expensive to calculate for long molecular dynamics simulations. Unsupervised algorithms offer an alternative independent of the studied system and can utilize pre-calculated atomistic potential descriptors from molecular dynamics simulations. We compare the performance of three such algorithms (PCA, UMAP, and PaCMAP) on silicon and water systems. Initially, we evaluate the algorithms for recognizing phases, including crystal polymorphs and the melt, followed by an extension of our analysis to identify interstitials, vacancies, and interfaces. While PCA is found unsuitable for effective classification, it has been shown to be a suitable initialization for UMAP and PaCMAP. Both UMAP and PaCMAP show promising results overall, with PaCMAP proving more robust in classification, except in cases of significant class imbalance, where UMAP performs better. Notably, both algorithms successfully identify nuclei in supercooled water, demonstrating their applicability to ice nucleation in water.

5.1 Introduction

Machine learning (ML) has transformed the computational modeling of atomic interactions, a cornerstone of molecular dynamics (MD) simulations. Current machine learning potentials (MLPs) allow us to reliably simulate large systems and long timescales with *ab initio* accuracy²²⁵. Beyond the fitting of potential energy surfaces, the development of ML tools has also benefited the analysis of atomistic simulations, providing valuable structural information. For instance, supervised learning approaches based on neural networks have been used to identify different phases in polymorphic systems^{226–232} as well as dynamical processes²³³. However, characterizing local atomic environments of crystal defects is a challenging task^{234–236}. Conveniently, the development of MLPs has come along with refined features describing local atomic

environments, such as the Smooth Overlap of Atomic Position (SOAP) descriptors²³⁷ or the atom-centered symmetry functions¹¹, which are applied in this work. Designing order parameters can be complex and computationally intensive, especially if no structural classification is available in advance. Therefore, unsupervised ML methods have emerged as a promising path towards discovering relevant structural features in complex systems²³⁸. Instead of trying to learn the local chemical environment from typically tens to hundreds of dimensions of atomic descriptors, it is common practice in unsupervised methods to initially reduce the dimensionality to two or three dimensions. Several methods for dimensionality reduction have been applied in the context of simulations^{239–244}. In particular, unsupervised classification tasks have been performed using i) topological graph order parameters combined with both principal component analysis (PCA)²⁴⁵ and diffusion maps^{246,247}, ii) Gaussian mixture models together with both PCA²⁴⁸ and neural-network-based autoencoders^{249,250}, iii) a band structure encoding along with the t-distributed stochastic neighbor embedding (t-SNE)⁸⁷, iv) simple three-body features²⁵¹ jointly with the uniform manifold approximation and projection (UMAP)⁸⁸, and v) SOAP combined with UMAP^{252–254} and the multidimensional scaling algorithm²⁵⁵.

To effectively analyze MD trajectories, it is crucial to employ algorithms that meet three fundamental criteria: (1) minimal computational overhead, which can be achieved by utilizing precomputed atomistic potential descriptors; (2) distinct separation of clusters for simple classification; and (3) robustness without variations in hyperparameters and initialization. Topological graph methods fail to satisfy the requirement of minimal computational overhead due to the additional calculations involved in generating graph-based order parameters²⁴⁵. Furthermore, metric-preserving algorithms such as PCA and autoencoders often produce distributions with overlapping clusters, lacking clear separation, as demonstrated in this work. Graph-based algorithms like t-SNE and UMAP are known to be sensitive to variations in hyperparameters and initialization procedures^{89–91}. The novel graph-based algorithm pairwise controlled manifold approximation projection (PaCMAP) has been reported to be more robust and superior compared to t-SNE and UMAP in toy problems⁹² and single-cell transcriptomic data analysis²⁵⁶.

Our study compares three unsupervised dimensionality reduction techniques for identifying crystal defects: PCA, UMAP, and PaCMAP. Our goal is to validate the reported superiority of PaCMAP over the established UMAP method. We maintain consistent hyperparameters across all studies to evaluate their robustness when cluster identities are unknown beforehand. The only exception is the initialization process, which has been shown to be crucial for achieving reliable results. We select two standard benchmark systems: Si and H₂O. First, we assess

the methods' ability to classify phases and then focus on locating point defects. For silicon, we also identify surface atoms, whereas for H₂O, we detect a hexagonal ice nucleus surrounded by supercooled water, an essential feature for studying ice nucleation^{257,258}. A detailed methodology for PaCMAP is provided, as this algorithm is less well-known than PCA and UMAP. Additionally, we discuss the nonparametric nature of PaCMAP and our approach to addressing it.

5.2 Results

5.2.1 Silicon

Phase classification

Silicon crystallizes in a diamond structure under standard conditions. When subjected to pressure, the diamond structure of silicon transforms into the β -Sn structure at approximately 10 GPa, the orthorhombic structure (space group Imma) at around 13 GPa, and the simple hexagonal structure at around 16 GPa²⁵⁹. High-pressure structures share the same unit cell, where atoms are positioned at $(0, 0, 0)$, $(0.5, 0.5, 0.5)$, $(0.5, 0, 0.5 + \nu)$, and $(0, 0.5, \nu)$. A ν value of 0.25 corresponds to the β -Sn structure, while the SH structure is characterized by $\nu = 0.5$. The Imma phase provides a continuous transition between these phases. Consequently, the β -Sn and SH phases closely resemble each other, and their differentiation may not always be trivial. It is confirmed by the PCA analysis in Fig. 28a, where these phases overlap, and a clear boundary cannot be unequivocally determined. Similarly, distinguishing the liquid phase from the solid is not achievable through PCA, as the liquid embedding overlaps with other solid phases, as already observed in water²⁶⁰.

Our preliminary experiments revealed that PaCMAP and UMAP highly depend on the initialization process. Specifically, using UMAP's default spectral initialization resulted in correct clustering (Fig. 28b) but also artificial cluster creation (Fig. 37a, in Supplementary information A.3) and misclassification of parts of the SH phase as β -Sn (Fig. 37b). These results were obtained with identical hyperparameters, only varying the random seed. Similar results are observed with PaCMAP random initialization. To address this instability, we switched to PCA initialization. Following this change, both algorithms successfully separated the liquid phase from the solid phases and clearly distinguished between β -Sn and SH phases, as shown in Figs. 28b and 28c. The misclassification rates for UMAP and PaCMAP were 0.10% and 0.08%, respectively.

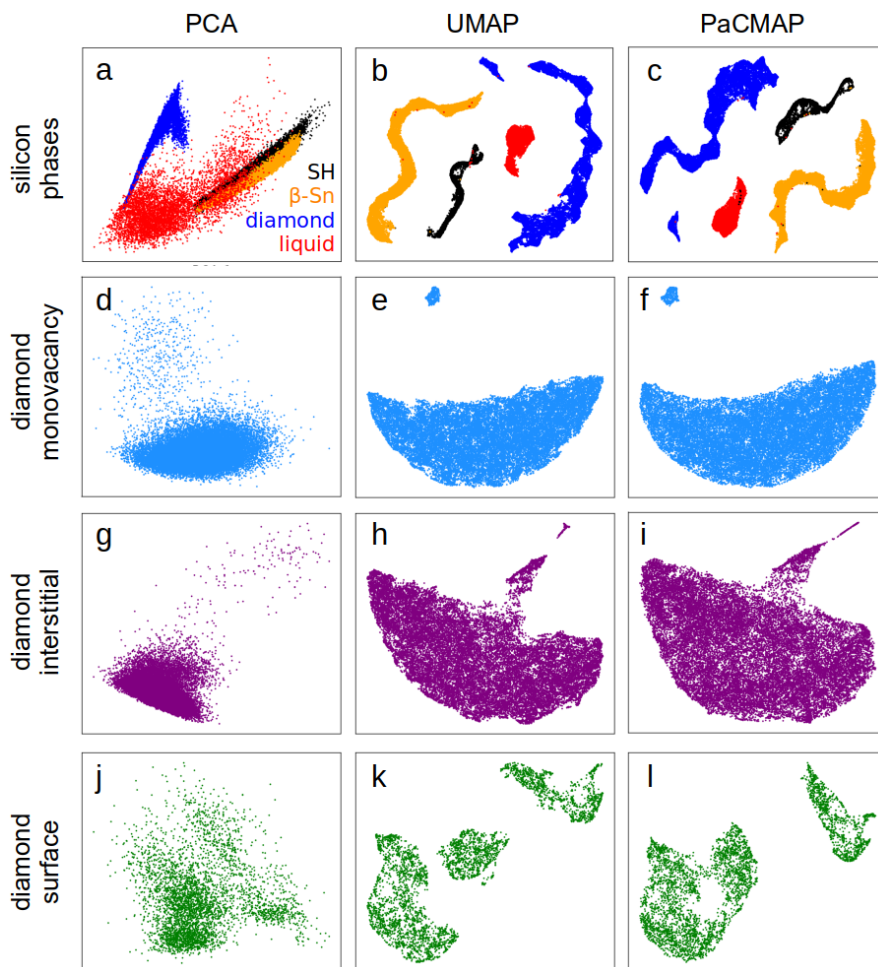


Figure 28: Comparison of clustering results using PCA, UMAP, and PaCMAP algorithms for silicon. Included are silicon phases (a-c), diamond monovacancy (d-f), diamond interstitial (g-i), and diamond surface (j-l).

Diamond point defects

The subsequent analysis aims to localize monovacancies and interstitials within silicon's diamond structure. Monovacancies can be identified by locating the four nearest neighbors that would have formed bonds with the missing atom if it had been present. The PCA plot in Fig. 28d shows two distributions without clear separation, whereas UMAP and PaCMAP clustering in Figs. 28e and 28f clearly distinguish atoms near monovacancies from those in other regions. Graph-based algorithms accurately identify the four nearest neighbors for 89% of structures, while the remaining cases exhibit variations (3, 5, and 6 neighbors) caused by local distortions introduced by interstitials. Nonetheless, the algorithms successfully locate the monovacancy in all tested cases.

Identifying interstitial atoms is straightforward when their identity remains constant throughout the simulation. However, additional analysis is required if an interstitial atom changes its identity over

time. While PCA provides no useful information (Fig. 28g), UMAP and PaCMAP categorize silicon systems with interstitial defects into two interconnected and one separated cluster, as shown in Figs. 28h and 28i. The largest cluster includes atoms far from the defect, while the middle interconnected cluster comprises atoms near the interstitial. The smallest clusters are composed of atoms not arranged on a diamond lattice. Figure 38 presents examples from the smallest cluster, showing local chemical environments with varying degrees of distortion.

Surface

We conclude our silicon analysis by examining (001) diamond surfaces using a relatively small dataset comprising 29 structures. The PCA analysis reveals only overlapping distributions (Fig. 28j). While UMAP divides the local environments into three clusters (Fig. 28k), PaCMAP identifies only two (Fig. 28l). Upon examining one of the surface structures depicted in Fig. 39, we find that the UMAP clusters primarily correspond to bulk atoms, surface atoms, and first-neighbors to surface atoms. PaCMAP misses the first neighbors to surface atoms.

5.2.2 Hexagonal ice

Phase classification

The second part of the paper is focused on the differentiation of ice-like and liquid-like water molecules in the context of defects and ice nucleation. It would be expected that classifying disordered and ordered phases would be easier than between similar crystal polymorphs, but this task was more challenging than the previous case of silicon. Mid-to long-range ordering is a critical feature for differentiation and is not primarily addressed by our potential descriptors. As shown in Figs. 29a-c, none of the tested methods distinguishes the water phase from the ice phase. This is due to the greater significance attributed to O-H bonds of water molecules compared to hydrogen bonds in used atomic potential descriptors. When the O-H bond length is fixed, as in the case of certain classical force fields, a complete separation between liquid and solid phases is possible. Thermal fluctuations of OH bonds can be reduced by applying a moving average to the molecular dynamics trajectory.

PaCMAP achieves complete water and ice phase separation at a moving average interval of 0.375 ps (Fig. 29f), with classification errors of only 1.14% for water and 0.68% for ice. As the moving average length increases, phase differentiation accuracy improves. For instance, a 0.75 ps moving average reduces errors to 0.34% for water and 0.14% for ice. In contrast, UMAP fails to achieve complete cluster separation at 0.375 ps (Fig. 29e) or even at 0.5 ps (Fig. 40). Only with a 0.75 ps moving average does UMAP entirely separate water and ice local

environments (Fig. 29h), albeit with a higher classification error for water (0.59%) compared to PaCMAP (0.34%). The classification error for ice remains similar, at 0.14%.

The increased classification error for water is primarily due to the probability of water molecules forming local ice-like structures. In contrast, ice molecules remain closely aligned with their crystal lattice structure.

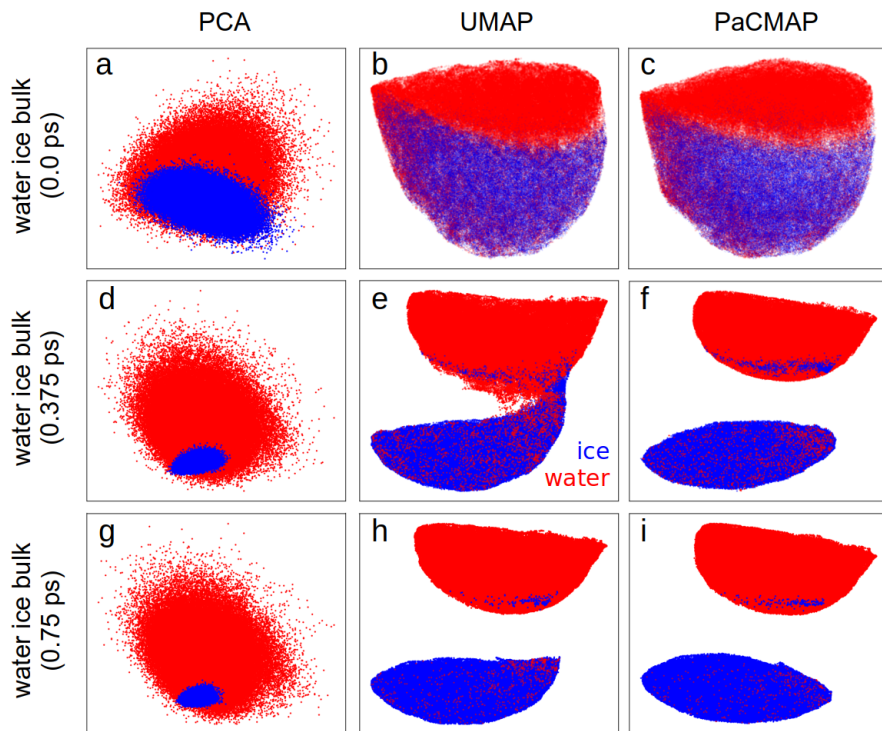


Figure 29: Comparison of clustering results using PCA, UMAP, and PaCMAP algorithms for bulk water and ice. Panels (a-c) show results without time averaging, (d-f) show results averaged over a 0.375 ps time interval, and (g-i) display results averaged over a 0.75 ps time interval.

Molecular defects

Time averaging has been shown to be essential in order to differentiate between water and ice phases using UMAP. Specifically, a moving interval of at least 0.75 ps is required to separate water and ice phases for UMAP effectively. Therefore, we employed this same interval in subsequent tests to identify molecular monovacancy, interstitial atoms, and ice nuclei in the liquid phase.

For localizing monovacancies in ice, UMAP proved the most effective for separating clusters (Figs. 30a-c), although complete separation is not achieved. The two emerging clusters correspond to distinct local chemical environments near monovacancy defects, specifically, two oxygen atoms missing a hydrogen donor and two oxygens missing a hydrogen acceptor per structure. These four oxygens near molecular

monovacancy enable straightforward tracking of the monovacancy location.

In the case of identifying interstitial atoms, even PCA demonstrates sufficient clustering (Fig. 30d). UMAP and PaCMAP not only identify interstitials in 99.3% of structures (top right clusters in Figs. 30e and 30f) but also their neighbor atoms (three clusters below interstitial cluster). Notably, the small cluster on the left represents a second neighbor to an interstitial atom that has been displaced from its crystal position due to interstitial diffusion.

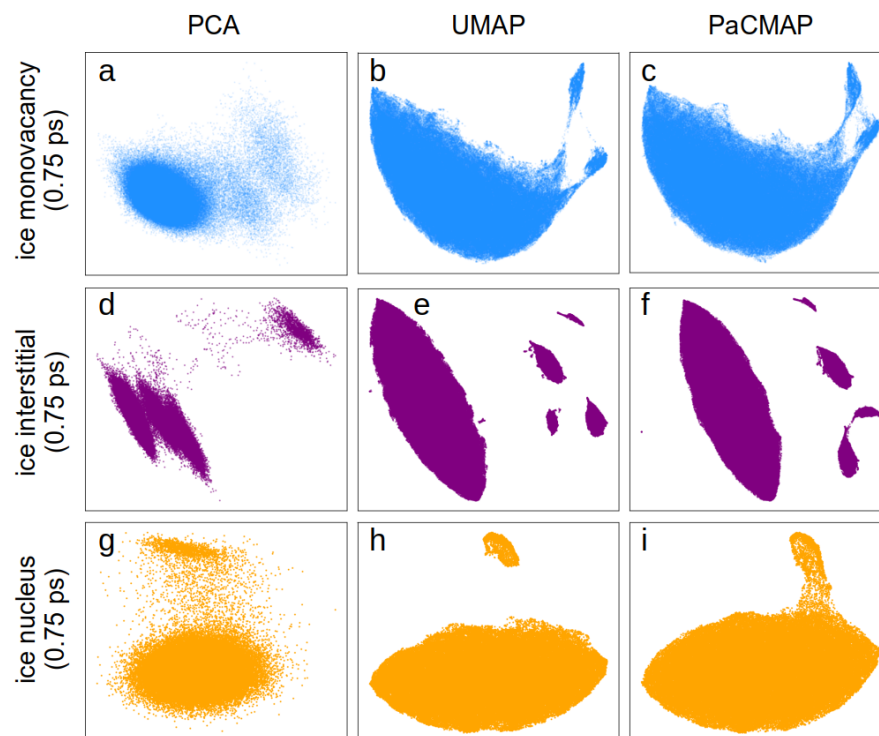


Figure 30: Comparison of clustering results using PCA, UMAP, and PaCMAP algorithms for special local environments in ice. Included cases are the ice monovacancy (a-c) and the ice interstitial (d-f) defects and the rounded ice nucleus in water (g-i), with potential descriptors averaged over a 0.75 ps time interval.

Ice nucleus in supercooled water

Our previous analysis compared the bulk structures of the water and ice phases. Identifying an ice nucleus within the liquid phase proved more challenging due to its distinct structure, influenced by interfacial stress. The interface is characterized by molecules with intermediate properties between liquid and solid, making it difficult to distinguish between them. To avoid finite-size effects in nucleation studies, the crystalline nucleus was significantly smaller than the surrounding liquid, resulting in an enormous class imbalance among molecules in the three phases (liquid, interface, and nucleus). Even using a

moving average of 0.75 ps, only UMAP could effectively differentiate between the liquid and ice nucleus, as illustrated in Figs. 30g-i. While PaCMAP showed a visible cluster forming in the upper right, it is not distinct. The structure contained approximately 75,000 molecules in the liquid phase and about 2,000 molecules in the solid phase, resulting in an enormous class imbalance. Augmenting the dataset with an additional 2,000 local environments of bulk ice structures enabled complete cluster separation even for PaCMAP (Fig. 41b). However, only partial overlap with the ice bulk environment is observed in the newly separated cluster. The non-overlapping segment corresponds to interface atoms, as indicated by the periodicity loss at the nucleus’s edge in Fig. 41c.

5.3 Discussion

We have demonstrated the applicability of atomistic potential descriptors for unsupervised identification of local atomic environments in silicon and water systems. By utilizing these potential descriptors, we can not only classify phases but also identify point defects and interfaces. We compared the clustering capabilities of two widely used methods, PCA and UMAP, with a novel approach, PaCMAP. Although PCA provided limited valuable information for phase or defect detection, its initialization served as an effective starting point for both UMAP and PaCMAP. Without PCA initialization, the clustering results of both methods were unstable.

Both UMAP and PaCMAP successfully separated silicon phases and identified diamond point defects. UMAP outperformed PaCMAP in distinguishing the first neighbors of the silicon surface, while PaCMAP only identified the surface atoms. Furthermore, UMAP could identify ice nuclei in water without additional data points, whereas PaCMAP required augmenting the dataset. This sensitivity to class imbalance suggests that PaCMAP may struggle with unbalanced datasets. In contrast, when class imbalance was not an issue, such as distinguishing between water and hexagonal ice, PaCMAP outperformed UMAP. PaCMAP required only a 0.375 ps moving average, whereas UMAP needed twice that time. Although hyperparameter tuning could potentially yield similar results for both methods, our focus was on unsupervised clustering, where the number of clusters is generally unknown a priori.

In conclusion, PaCMAP’s superiority over UMAP has not been confirmed. While PaCMAP excelled in separating the water and ice phases, it struggled with unbalanced datasets. Although UMAP clustering is sensitive to hyperparameter variations, it performs effectively with default settings, except for the initialization process, where the default spectral initialization produces unstable results. Both algorithms perform well overall, and the choice between them should be determined by the specifics of the system under study.

5.4 Methods

5.4.1 PaCMAP algorithm

The overarching objective of PaCMAP is to bring proximate data points within the high-dimensional space into closer proximity within the low-dimensional space, as well as more distant data points in the original space to greater distances within the low-dimensional space. Since it would be computationally demanding to consider all the distances between all atoms with each other during optimization, PaCMAP restricts the number of neighbors for each data point to a finite value. Neighbors are categorized into three types: near pairs (n_{NP}), mid-near pairs (n_{MN}), and further pairs (n_{FP}). Attractive forces are applied to near and mid-near pairs, whereas repulsive forces are exerted on further pairs. The selection of neighbors is a one-time process, and the neighbor pairs remain constant throughout the optimization phase. The whole algorithm unfolds in three steps:

Choosing neighbors

To determine near neighbors, an initial step involves the selection of a subset comprising the minimum of either $n_{NP} + 50$ or the total number of observations (N) nearest neighbors based on Euclidean distance. Subsequently, the nearest neighbors are identified according to the scaled distance metric $d_{ij}^2 = \frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{\sigma_i \sigma_j}$ between observation pairs (i, j), where $\mathbf{X}_i$ represents atomistic potential descriptors and σ_i is the average distance between observation i and its Euclidean nearest neighbors falling in the fourth to sixth positions. This scaling is implemented to accommodate potential variations in the magnitudes of neighborhoods across different regions of the feature space. In selecting mid-near pairs, six additional points are randomly sampled (uniformly), and the second nearest point among them is chosen. Lastly, further pairs are determined by randomly selecting n_{FP} additional points.

Initialize low dimension embedding $\mathbf{Y}$

Each datapoint i is associated with a 2D vector $\mathbf{y}_i$. As PaCMAP is a non-parametric algorithm without an underlying function that generates low-dimension embedding, values are generated by selecting the most important PCA components.

Minimize loss

With a predefined neighbor list for each datapoint and an initial position in the low-dimensional embedding provided by PCA, the opti-

mization process commences minimizing the loss function by varying $\mathbf{y}_i$ with the ADAM optimizer⁴⁶:

$$L = \sum_i^N \left(w_{NP} \sum_j^{n_{NP}} \frac{\tilde{d}_{ij}}{10 + \tilde{d}_{ij}} + w_{MN} \sum_k^{n_{MN}} \frac{\tilde{d}_{ik}}{10000 + \tilde{d}_{ik}} + w_{FP} \sum_l^{n_{FR}} \frac{1}{1 + \tilde{d}_{il}} \right) \quad (25)$$

where $\tilde{d}_{ab} = \|\mathbf{y}_a - \mathbf{y}_b\|^2 + 1$. The first two terms correspond to the attractive interactions between near and mid-near pairs. The third term represents a repulsive interaction that facilitates cluster separation. The pairs contribute to the loss function with weights determined by the coefficients w_{NP} , w_{MN} , and w_{FP} , collectively constituting the overall loss. These weights are dynamically updated throughout the algorithm as part of the optimization process. The dynamic update of weights follows a specific scheme during different iterations of the optimization process:

- Iterations 0 to 100: $w_{NP} = 2$,
 $w_{MN} = 1000 \left(1 - \frac{t-1}{100}\right) + 3 \left(\frac{t-1}{100}\right)$, $w_{FP} = 1$
- Iterations 101 to 200: $w_{NP} = 3$, $w_{MN} = 3$, $w_{FP} = 1$
- Iterations 201 to 450: $w_{NP} = 1$, $w_{MN} = 0$, $w_{FP} = 1$

The optimization process has three distinct phases designed to circumvent local optima. The initial phase focuses on global structure achieved through substantial weighting of mid-near pairs. As this phase progresses, weights on mid-near pairs gradually decrease, facilitating algorithmic focus shifting from global to local structures. The subsequent phase concentrates on enhancing local structure while retaining the global structure acquired in the first phase. Finally, the third phase prioritizes the refinement of local structure by reducing the weight of mid-near pairs to zero, accentuating the role of repulsive forces to separate cluster boundaries more distinctly. Fig. 42 provides an example of the optimization process for silicon phases.

5.4.2 Parametric PaCMAP

In this section, we address the most significant limitation of the PaCMAP algorithm: its nonparametric nature. It means that predicting new data points is not directly possible, as the algorithm does not construct a function $f(x) : R^d \rightarrow R^2$ that would map high-dimensional descriptors to a 2D latent space. This limitation can be a considerable constraint in certain real-world applications where the goal is to cluster training data and apply the learned mapping to new, simulated data points. A practical approach involves generating embedding/labels, which are then employed as output for supervised regression or classification tasks. Subsequently, a simple feedforward neural network can be easily

trained and rapidly produce predictions for new local environments. To illustrate this approach, we trained a neural network with three hidden layers, each comprising 50 nodes and employing ReLU activation functions for silicon and bulk liquid and ice phases.

In Fig. 31, we compare the neural network’s predictions with the test reference dataset (PaCMAP mapping), proving visually that the neural network has successfully learned the underlying mapping of PaCMAP. A similar comparison of bulk liquid and ice phases is provided in Fig. 43. However, classification is a more practical approach for real-world applications. Each cluster from the training data would be assigned a class, and a classifier would then predict the class of each cluster directly from the descriptors, eliminating the need for 2D embedding.

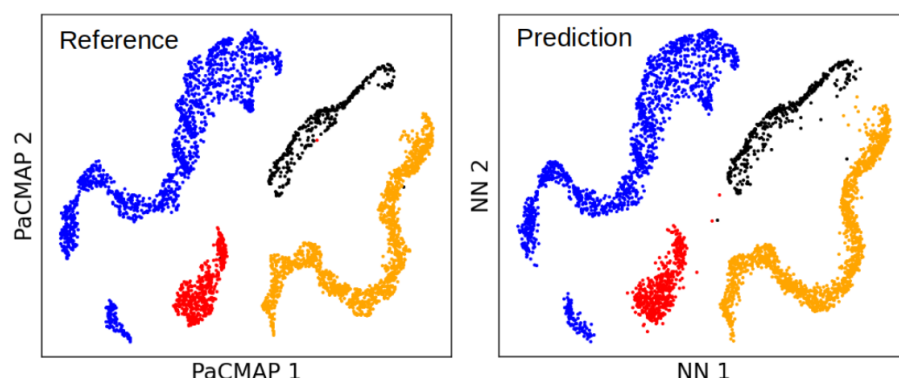


Figure 31: Parametric prediction of PaCMAP embedding. Comparison of reference test data (top) and neural network predictions (bottom) for distinguishing between silicon phases. The test dataset contains 4,000 local chemical environments.

5.4.3 Datasets, atomistic potential descriptors, and PaCMAP hyperparameters

We have selected two well-known systems for benchmarking unsupervised algorithms: silicon (Si) and water (H_2O). In the case of silicon, our focus is on identifying the crystal lattice, point defects, and the surface. Our emphasis for water molecules is on liquid and hexagonal ice distinction and molecular point defects.

The general-purpose interatomic potential for silicon encompasses 2,475 structures, which can be categorized into 23 different structure types²⁶¹. Not all structure types contain a sufficient number of structures for data-driven analysis. Therefore, we restrict our analysis to distinguishing between diamond, β -Sn, simple hexagon (SH), and liquid phase, but also localization of monovacancies, interstitial positions, and identifying surface atoms.

Water structures include bulk liquid, ice, and a spherical nucleus surrounded by supercooled water. All structures have been generated

by running MD simulations with large-scale atomic/molecular massively parallel simulator (LAMMPS)¹⁶¹ using a Behler-Parrinello neural network trained on ab initio data based on the RPBE-D3 zero damping density functional²⁶². Initial ice-Ih structures were generated with GenIce²⁶³ and then simulated in the anisotropic NpT ensemble at 250 K and 0 bar for 700 ps. Supercooled water presents prolonged relaxation, so producing equilibrated structures requires either very long simulations or the use of special techniques. Here, we have run a 32 replica parallel tempering simulation at constant 0 bar pressure during 8 ns covering from 211 K up to 335 K. Then, the last structure from the distribution of 250 K was selected and simulated for 2 ns more at constant temperature and pressure (250 K and 0 bar) for production-level data acquisition. To produce the nucleus system, we inserted a perfect spherical ice-Ih nucleus of about 2,000 water molecules into a supercooled water configuration, producing 78,856 water molecules. We pre-equilibrated the system using the TIP4P/Ice force field²⁶⁴ via the GROMACS package²⁶⁵. The temperature was set to 250 K, pressure to 0 bar, and the nucleus size barely changed during 500 ps. Then, we switched to the MLP and slightly heated the system towards 260 K for 10 ps. Finally, we have run 2.5 ns in the NpH ensemble for production-level structures at 0 bar. The nucleus remained relatively stable in size, and the average temperature quickly converged to ~ 255 K.

We employ Behler-Parinello descriptors (symmetry functions) to encode the local structure information due to their simplicity and widespread applicability¹¹. For silicon, we constructed a descriptor vector comprising 20 radial Gaussian symmetry functions and 15 angular polynomial symmetry functions¹²² (details are provided in Tables 8 and 9). These symmetry functions are designed generally without leveraging specific knowledge about the bonding¹⁴. A cutoff distance of 5 Å was set to prevent self-interaction in smaller supercells within the dataset. For water analysis, we chose well-tested symmetry functions commonly employed in molecular dynamics simulations accelerated by MLPs²⁶² with a cutoff of 6.35 Å. Only oxygen descriptors that include both oxygen-oxygen and oxygen-hydrogen interactions have been utilized in analyses.

Each data point corresponds to an individual atomic environment within the supercell without averaging over the cell or time unless otherwise specified.

Regarding the hyperparameters of the PaCMAP algorithm, we have fixed parameters for all presented cases. This entails $n_{NP} = 10$, $n_{MN} = 5$, and $n_{FP} = 25$ with PCA initialization.

5.4.4 Software

Production-level molecular dynamics simulations are performed with n2p2⁴⁸ – LAMMPS¹⁶¹. GROMACS²⁶⁵ is used for setting up the ice-

the nucleus system. Initial configurations of ice are generated with GenIce²⁶³. The symmetry functions are evaluated using the n2p2 package⁴⁸. The structures are rendered using OVITO²⁶⁶. Dimensionality reduction is performed using packages: scikit-learn for PCA²⁶⁷, UMAP²⁶⁸, and PaCMAP⁹². Neural networks trained on PaCMAP labels are constructed and interfered with PyTorch²⁶⁹.

Data availability

The atomic descriptors and corresponding structures are publicly available on Zenodo²⁷⁰.

Acknowledgments

We acknowledge financial support by the Doctoral College Advanced Functional Materials – Hierarchical Design of Hybrid Systems DOC 85 doc.funds and SFB-TACO 10.55776/F81 funded by the Austrian Science Fund (FWF), and by the Vienna Doctoral School in Physics (VDSP). The computational results presented have been partly achieved using the Vienna Scientific Cluster (VSC).

Additional information

Supplementary information A.3 includes complementary analyses and a list of the atomistic potential descriptors for silicon.

A.1 Optimizing the architecture of Behler-Parrinello neural network potentials

We have introduced a new set of polynomial symmetric functions. This set consists of 7 centered radial functions with linearly increasing width from 2 to 8 Å, as shown in Fig. 32. Such an arrangement is common, but the novelty of this set comes with the angular symmetry functions. Polynomial symmetry functions offer freedom in the choice of angle except for the continuity conditions at 0 and π . Since the QM9 dataset contains different kinds of molecules, specific angles cannot be singled out, and broad coverage is necessary. We have found that 4 angular symmetry functions cover the angular space well if we adjust them to mimic the first three harmonics (see Fig. 32). The radial part of the angular symmetry function is set identically to the radial symmetry functions (Eq. 12) to improve angular resolution. The ratio between the number of radial and angular symmetry functions is $R : 2R(R + 1)$ for the unweighted model and $4R$ for the weighted model, where R is the number of radial symmetry functions. The R is set at 7 for the full-weighted and angular-weighted models with a spacing of 1 Å. For the unweighted model, we had to reduce the R from 7 to 4 (2 Å spacing instead of 1 Å) to get alignment with the number of symmetry functions in the previously reported model¹⁴.

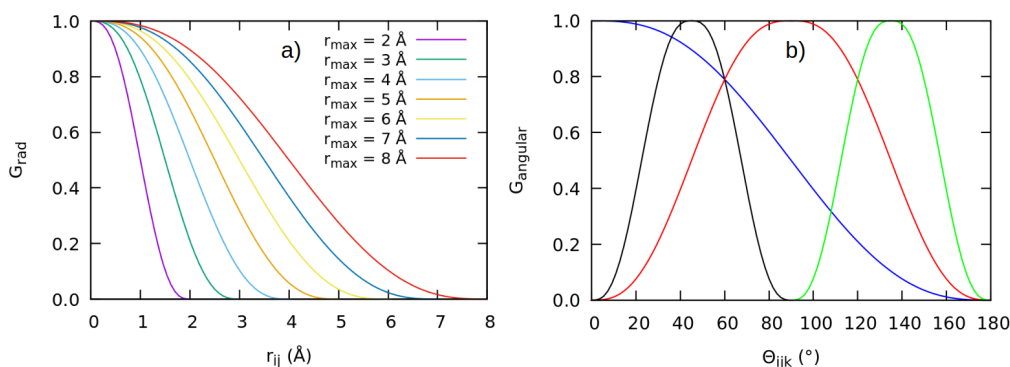


Figure 32: a) Radial polynomial symmetry functions for different r_{max} values. For each of the functions r_{min} is equal to $-r_{\text{max}}$. b) Angular parts of the angular polynomial symmetry functions. The parameters are $\theta_{\text{min}} = -180^\circ, \theta_{\text{max}} = 180^\circ$ (blue line), $\theta_{\text{min}} = 0^\circ, \theta_{\text{max}} = 180^\circ$ (red line), $\theta_{\text{min}} = 0^\circ, \theta_{\text{max}} = 90^\circ$ (black line), and $\theta_{\text{min}} = 90^\circ, \theta_{\text{max}} = 180^\circ$ (green line).

Dataset	Complexity	H	C	N	O	F
QM9	full-weighted	35	35	35	35	30
	angular-weighted	62	63	62	61	55
	unweighted	233	248	232	217	196
3BPA	angular-weighted	56	56	54	48	-

Table 6: Number of symmetry functions per element for different model complexities.

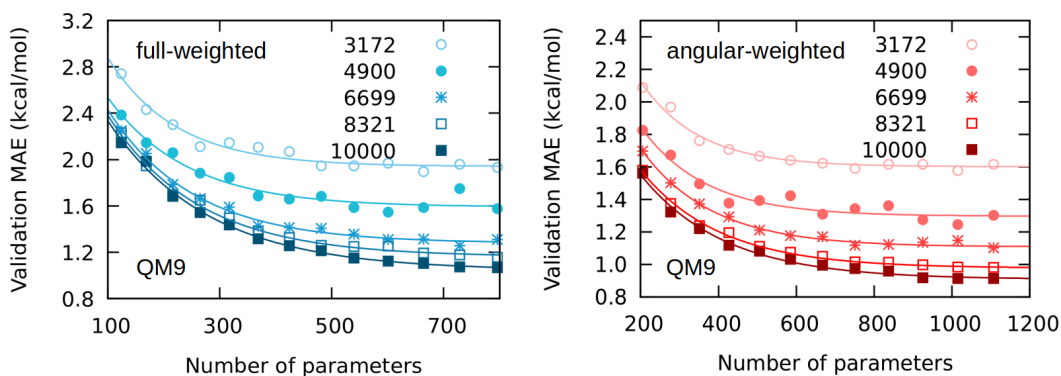


Figure 33: QM9: Validation MAE vs. number of parameters for full-weighted and angular-weighted for various training data sizes. The errors are averaged over 10 models trained starting with different seeds. The errors are fitted with an exponential function (solid lines) for each training set size.

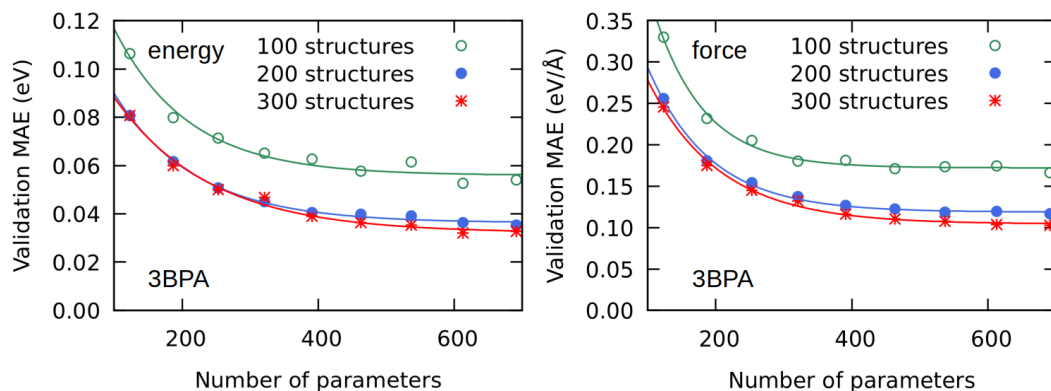


Figure 34: 3BPA: Validation MAE vs. number of parameters for energy and forces for various training data sizes. The errors are averaged over 20 models trained starting with different seeds. The errors are fitted with an exponential function (solid lines) for each training set size.

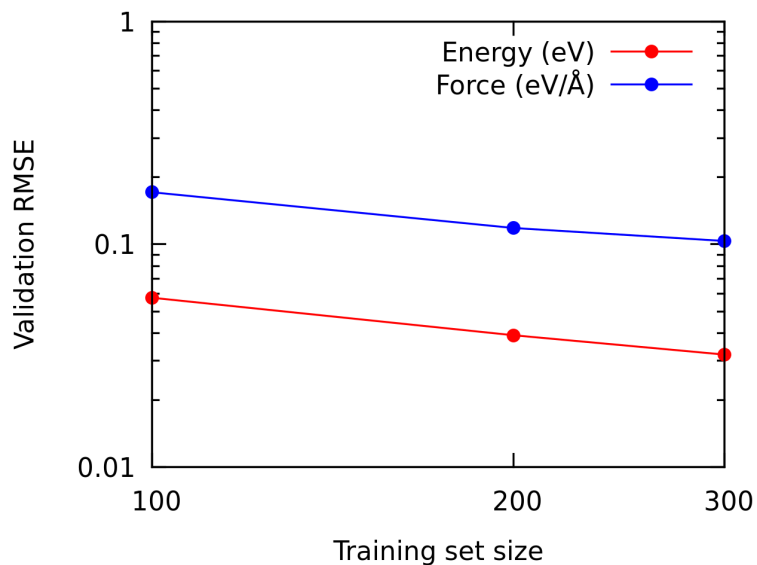


Figure 35: 3BPA learning curve with a flexible architecture.

A.2 Diffusion and coalescence of phosphorene monovacancies studied using high-dimensional neural network potentials

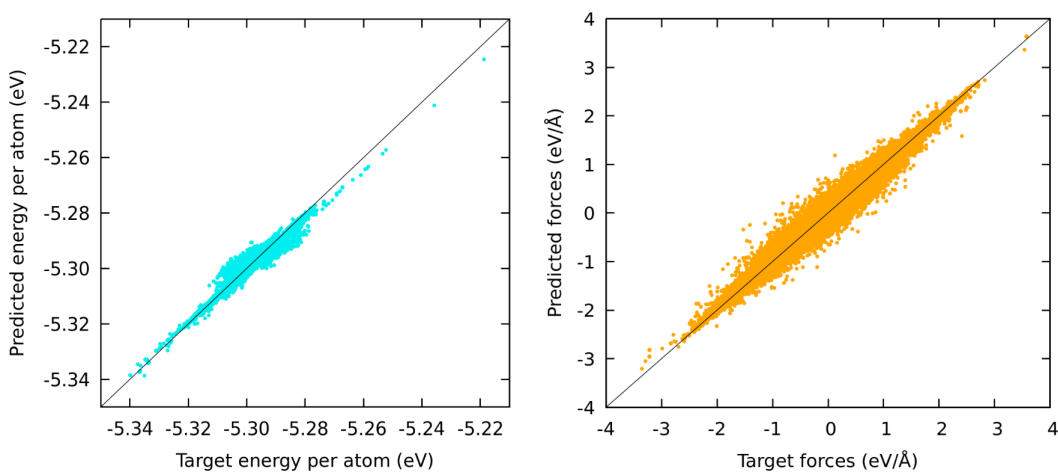


Figure 36: Parity plots comparing predicted values from an ensemble of HDNNPs with a size of 8 against target values obtained through DFT.

Table 7: Parameters of the radial and angular symmetry functions for phosphorene.

	r_{min} (Å)	r_{max} (Å)	θ_{min} (°)	θ_{max} (°)
G^{rad}	-7	7	-	-
	-6	6	-	-
	-5	5	-	-
	-4	4	-	-
	-3	3	-	-
	3	7	-	-
	2	6	-	-
	1	5	-	-
	0	4	-	-
	-1	3	-	-
G^{ang}	-7	7	-180	180
	-6	6	-180	180
	-5	5	-180	180
	-4	4	-180	180
	-3	3	-180	180
	3	7	-180	180
	2	6	-180	180
	1	5	-180	180
	0	4	-180	180
	-1	3	-180	180
	-7	7	0	180
	-6	6	0	180
	-5	5	0	180
	-4	4	0	180
	-3	3	0	180
	3	7	0	180
	2	6	0	180
	1	5	0	180
	0	4	0	180
	-1	3	0	180
	-7	7	0	90
	-6	6	0	90
	-5	5	0	90
	-4	4	0	90
	-3	3	0	90
	3	7	0	90
	2	6	0	90
	1	5	0	90
	0	4	0	90
	-1	3	0	90
	-7	7	90	180
	-6	6	90	180
	-5	5	90	180
	-4	4	90	180
	-3	3	90	180
	3	7	90	180
	2	6	90	180
	1	5	90	180
	0	4	90	180
	-1	3	90	180

A.3 Unsupervised identification of crystal defects from atomistic potential descriptors

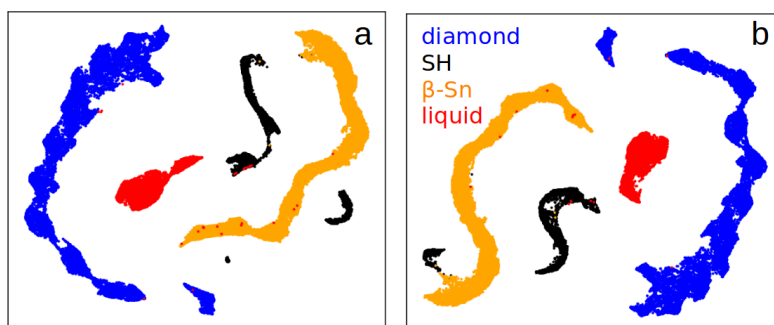


Figure 37: UMAP clustering for silicon phases with default spectral initialization. a) The SH phase (black) is incorrectly divided into three clusters. b) A portion of the SH phase (black) is merged with the β -Sn cluster (orange). Note that two blue clusters, also observed in PaCMAP, correctly represent the diamond phase but correspond to significantly different densities.

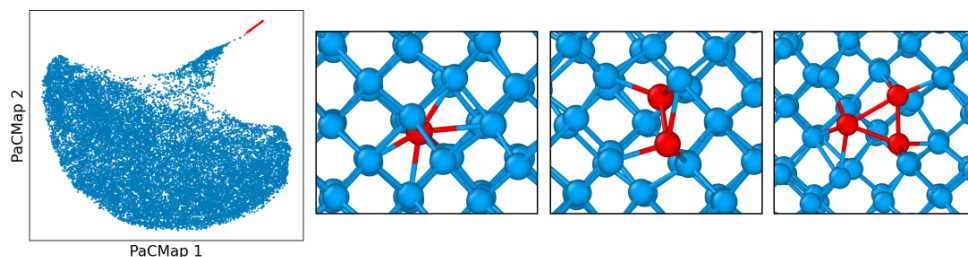


Figure 38: PaCMAP clustering of interstitials in the diamond silicon structure, showing three randomly chosen structures with atoms colored by their corresponding clusters.

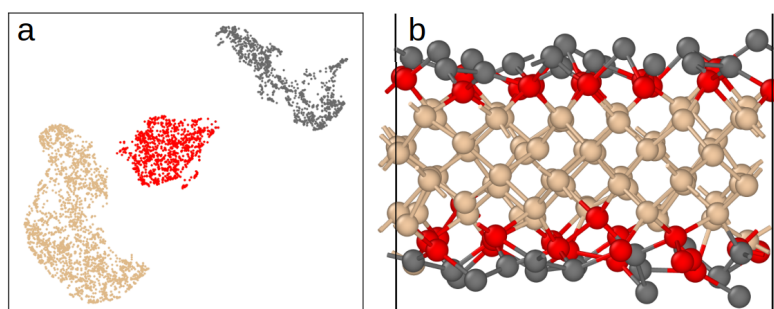


Figure 39: (a) UMAP clustering of silicon's diamond (001) surface. (b) Example of a silicon structure with atoms colored according to their UMAP clusters.

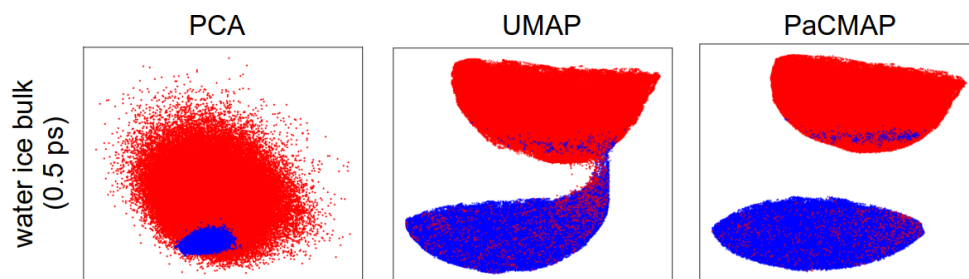


Figure 40: Comparison of clustering results using PCA, UMAP, and PaCMAP algorithms for water and ice bulk phases, with potential descriptors averaged over a 0.50 ps time interval.

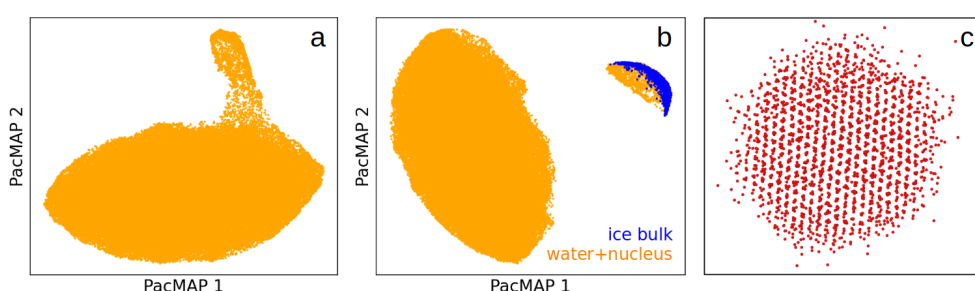


Figure 41: PaCMAP clustering of a water supercell: (a) with an ice nucleus, (b) with 2,000 additional ice bulk environments, and (c) highlighting the identified ice nucleus. Potential descriptors are averaged over a 0.75 ps time interval.

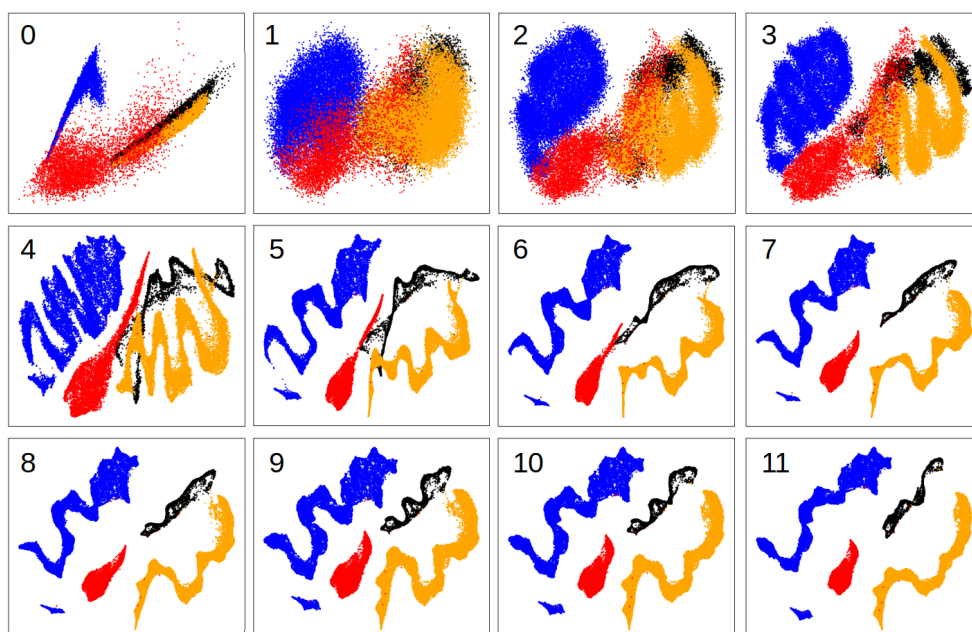


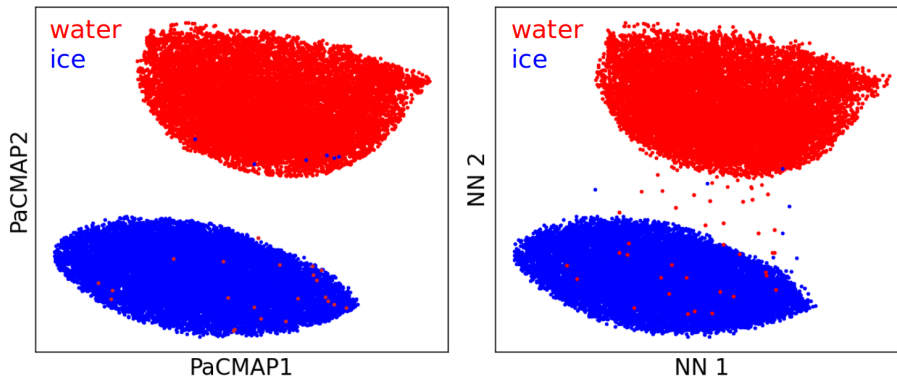
Figure 42: Visualizations of the low-dimensional embedding during the optimization process of PaCMAP on the silicon dataset. Iteration 0 shows the PCA initialization.

Table 8: Parameters of the radial symmetry functions for silicon.

G^{rad}	r_{min} (Å)	r_{max} (Å)	θ_{min} (°)
	0.0556	0.0	5.0
	0.0499	0.0	5.0
	0.0450	0.0	5.0
	0.0408	0.0	5.0
	0.0372	0.0	5.0
	0.0340	0.0	5.0
	0.0313	0.0	5.0
	0.0288	0.0	5.0
	0.0266	0.0	5.0
	0.0247	0.0	5.0
	18.0	3.000	5.0
	18.0	3.167	5.0
	18.0	3.333	5.0
	18.0	3.500	5.0
	18.0	3.667	5.0
	18.0	3.833	5.0
	18.0	4.000	5.0
	18.0	4.167	5.0
	18.0	4.333	5.0
	18.0	4.500	5.0

Table 9: Parameters of the angular symmetry functions for silicon.

G^{ang}	r_{min} (Å)	r_{max} (Å)	θ_{min} (°)	θ_{max} (°)
-3	3	3	-180	180
-3	3	3	0	360
-3	3	3	0	180
-3	3	3	0	120
-3	3	3	60	180
-4	4	4	-180	180
-4	4	4	0	360
-4	4	4	0	180
-4	4	4	0	120
-4	4	4	60	180
-5	5	5	-180	180
-5	5	5	0	360
-5	5	5	0	180
-5	5	5	0	120
-5	5	5	60	180

**Figure 43:** Comparison of reference test data (left) and neural network predictions (right) for distinguishing between water and ice. The test dataset comprises 30,000 molecules.

A.4 Phosphorene monovacancy tracking using PaCMAP

In Chapter 3, I employed atomic energies to track phosphorene monovacancies. Although identifying hopping events based on atomic energy differences was straightforward, determining the underlying mechanism required a closer examination of corresponding changes in Cartesian coordinates.

In contrast, unsupervised mechanisms such as PaCMAP do not necessitate manual effort. Different local chemical environments result in distinct PaCMAP clusters. In the case of phosphorene MV, clustering yielded five distinct clusters (Fig. 44, left). Each cluster's colors are depicted in Fig. 44, right, and correspond to specific local environments. Notably, PaCMAP identified four different local environments surrounding the MV: red, purple, blue, and green. Atomic energies alone can only detect the red cluster, which is insufficient for determining the hopping mechanism. However, tracking monovacancies with minimal effort becomes feasible with four distinct environments identified.

To illustrate this process, consider a scenario where an atom from the red cluster changes to gray while the green cluster persists, indicating a *rotation* mechanism. Alternatively, if an atom from the red cluster changes to green, a *mirror* transition occurs. If the red or green atom becomes purple, an *armchair* transition occurs. We can track monovacancies by monitoring how atoms change clusters solely based on potential descriptors without inspecting Cartesian coordinates.

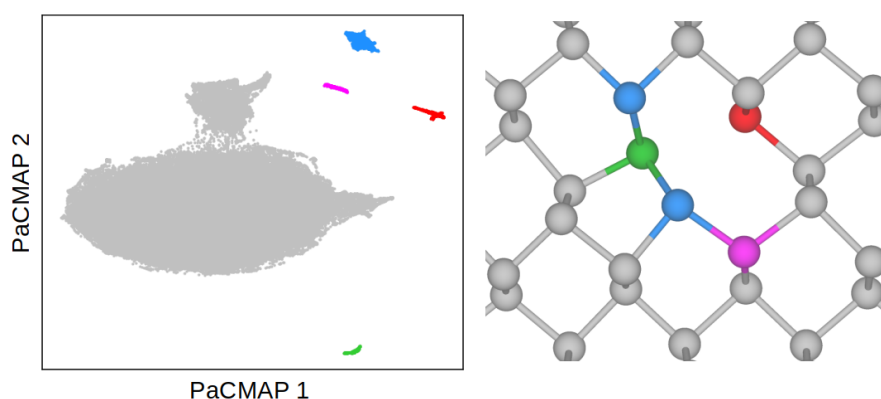


Figure 44: Left: PaCMAP clustering of defected phosphorene. Right: Phosphorene structure with atoms colored based on their respective clusters.

BIBLIOGRAPHY

- [1] P. DIRAC: Quantum mechanics of many-electron systems, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, **123**, 714–733 (1929).
- [2] R. M. MARTIN: *Electronic Structure: Basic Theory and Practical Methods*, Cambridge University Press (2004), ISBN: 9780511805769.
- [3] C. HONG, J. KIM, J. KIM, J. JUNG, S. JU, J. M. CHOI, S. HAN: Applications and training sets of machine learning potentials, *Science and Technology of Advanced Materials: Methods*, **3** (2023).
- [4] D. FRENKEL, B. SMIT: *Understanding molecular simulation*, Computational science series, Academic Press, San Diego, CA, 2 edition (2001), ISBN: 9780122673511.
- [5] D. C. RAPAPORT: *The Art of Molecular Dynamics Simulation*, Cambridge University Press, 2 edition (2004), ISBN: 9780511816581.
- [6] J. A. HARRISON, J. D. SCHALL, S. MASKEY, P. T. MIKULSKI, M. T. KNIPPENBERG, B. H. MORROW: Review of force fields and intermolecular potentials used in atomistic computational materials research, *Applied Physics Reviews*, **5** (2018).
- [7] W. KOHN, A. D. BECKE, R. G. PARR: Density functional theory of electronic structure, *The Journal of Physical Chemistry*, **100**, 12974–12980 (1996).
- [8] S. DAVID: *Planewaves, pseudopotentials, and the LAPW method*, New York: Springer (2006), ISBN: 9780387296845.
- [9] T. VAN MOURIK, M. BÜHL, M.-P. GAIGEOT: Density functional theory across chemistry, physics and biology, *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **372**, 20120488 (2014).
- [10] G. WANG, C. WANG, X. ZHANG, Z. LI, J. ZHOU, Z. SUN: Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations, *iScience*, **27**, 109673 (2024).
- [11] J. BEHLER, M. PARRINELLO: Generalized neural-network representation of high-dimensional potential-energy surfaces, *Physical Review Letters*, **98** (2007).
- [12] D. LU, H. WANG, M. CHEN, L. LIN, R. CAR, W. E, W. JIA, L. ZHANG: 86 pflops deep potential molecular dynamics simulation of 100 million atoms with ab initio accuracy, *Computer Physics Communications*, **259**, 107624 (2021).
- [13] J. BEHLER: Representing potential energy surfaces by high-dimensional neural network potentials, *Journal of Physics: Condensed Matter*, **26**, 183001 (2014).

- [14] M. GASTEGGER, L. SCHWIEDRZIK, M. BITTERMANN, F. BERZSENYI, P. MARQUETAND: wACSF—weighted atom-centered symmetry functions as descriptors in machine learning potentials, *The Journal of Chemical Physics*, **148**, 241709 (2018).
- [15] A. P. BARTÓK, M. C. PAYNE, R. KONDOR, G. CSÁNYI: Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons, *Physical Review Letters*, **104** (2010).
- [16] A. THOMPSON, L. SWILER, C. TROTT, S. FOILES, G. TUCKER: Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials, *Journal of Computational Physics*, **285**, 316 (2015).
- [17] A. V. SHAPEEV: Moment tensor potentials: A class of systematically improvable interatomic potentials, *Multiscale Modeling and Simulation*, **14**, 1153 (2016).
- [18] R. DRAUTZ: Atomic cluster expansion for accurate and transferable interatomic potentials, *Physical Review B*, **99** (2019).
- [19] B. CHENG: Cartesian atomic cluster expansion for machine learning interatomic potentials, *npj Computational Materials*, **10** (2024).
- [20] K. T. SCHÜTT, F. ARBABZADAH, S. CHMIELA, K. R. MÜLLER, A. TKATCHENKO: Quantum-chemical insights from deep tensor neural networks, *Nature Communications*, **8** (2017).
- [21] K. T. SCHÜTT, P.-J. KINDERMANS, H. E. SAUCEDA, S. CHMIELA, A. TKATCHENKO, K.-R. MÜLLER: Schnet: a continuous-filter convolutional neural network for modeling quantum interactions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, 992–1002, Curran Associates Inc., Red Hook, NY, USA (2017).
- [22] N. LUBBERS, J. S. SMITH, K. BARROS: Hierarchical modeling of molecular energies using a deep neural network, *The Journal of Chemical Physics*, **148** (2018).
- [23] O. T. UNKE, M. MEUWLY: PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges, *Journal of Chemical Theory and Computation*, **15**, 3678 (2019).
- [24] J. GASTEIGER, S. GIRI, J. T. MARGRAF, S. GÜNNEMANN: Fast and uncertainty-aware directional message passing for non-equilibrium molecules, *arXiv: 2011.14115* (2020).
- [25] I. BATATIA, D. P. KOVACS, G. SIMM, C. ORTNER, G. CSÁNYI: Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, in S. KOYEJO, S. MOHAMED, A. AGARWAL, D. BELGRAVE, K. CHO, A. OH (editors), *Advances in Neural Information Processing Systems*, volume 35, 11423–11436, Curran Associates, Inc. (2022).
- [26] S. N. POZDNYAKOV, M. CERIOTTI: Incompleteness of graph neural networks for points clouds in three dimensions, *Machine Learning: Science and Technology*, **3**, 045020 (2022).

- [27] B. ANDERSON, T. S. HY, R. KONDOR: Cormorant: Covariant molecular neural networks, in H. WALLACH, H. LAROCHELLE, A. BEYGEZIMER, F. D'ALCHÉ-BUC, E. FOX, R. GARNETT (editors), *Advances in Neural Information Processing Systems*, volume 32, Curran Associates, Inc. (2019).
- [28] N. THOMAS, T. E. SMIDT, S. M. KEARNES, L. YANG, L. LI, K. KOHLHOFF, P. F. RILEY: Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, *arXiv: 1802.08219* (2018).
- [29] V. G. SATORRAS, E. HOOGEBOOM, M. WELLING: E(n) equivariant graph neural networks, *arXiv: 2102.09844* (2022).
- [30] K. T. SCHÜTT, O. T. UNKE, M. GASTEGGER: Equivariant message passing for the prediction of tensorial properties and molecular spectra, *arXiv: 2102.03150* (2021).
- [31] P. THÖLKE, G. D. FABRITIIS: Torchmd-net: Equivariant transformers for neural network based molecular potentials, *arXiv: 2202.02541* (2022).
- [32] J. BRANDSTETTER, R. HESSELINK, E. VAN DER POL, E. J. BEKKERS, M. WELLING: Geometric and physical quantities improve e(3) equivariant message passing, *arXiv: 2110.02905* (2022).
- [33] M. HAGHIGHATLARI, J. LI, X. GUAN, O. ZHANG, A. DAS, C. J. STEIN, F. HEIDAR-ZADEH, M. LIU, M. HEAD-GORDON, L. BERTELS, H. HAO, I. LEVEN, T. HEAD-GORDON: Newtonnet: A newtonian message passing network for deep learning of interatomic potentials and forces, *arXiv: 2108.02913* (2021).
- [34] S. BATZNER, A. MUSAELIAN, L. SUN, M. GEIGER, J. P. MAILLOA, M. KORNBLUTH, N. MOLINARI, T. E. SMIDT, B. KOZINSKY: E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials, *Nature Communications*, **13** (2022).
- [35] E. R. SMITH: Electrostatic Energy in Ionic Crystals, *Proceedings of the Royal Society of London Series A*, **375**, 475 (1981).
- [36] N. ARTRITH, T. MORAWIETZ, J. BEHLER: High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide, *Physical Review B*, **83** (2011).
- [37] P. P. EWALD: Die berechnung optischer und elektrostatischer gitterpotentiale, *Annalen der Physik*, **369**, 253–287 (1921).
- [38] T. W. KO, J. A. FINKLER, S. GOEDECKER, J. BEHLER: A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer, *Nature Communications*, **12** (2021).
- [39] M. KULICHENKO, K. BARROS, N. LUBBERS, Y. W. LI, R. MESSERLY, S. TRETIK, J. S. SMITH, B. NEBGEN: Uncertainty-driven dynamics for active learning of interatomic potentials, *Nature Computational Science*, **3**, 230–239 (2023).
- [40] B. SETTLES: Active learning literature survey, *University of Wisconsin-Madison Department of Computer Sciences* (2009).

- [41] H. S. SEUNG, M. OPPER, H. SOMPOLINSKY: Query by committee, in *Proceedings of the fifth annual workshop on Computational learning theory*, COLT92, ACM (1992).
- [42] C. SCHRAN, K. BREZINA, O. MARSALEK: Committee neural network potentials control generalization errors and enable active learning, *The Journal of Chemical Physics*, **153**, 104105 (2020).
- [43] J. S. SMITH, B. NEBGEN, N. LUBBERS, O. ISAYEV, A. E. ROITBERG: Less is more: Sampling chemical space with active learning, *The Journal of Chemical Physics*, **148**, 241733 (2018).
- [44] I. GOODFELLOW, Y. BENGIO, A. COURVILLE: *Deep Learning*, MIT Press (2016), <http://www.deeplearningbook.org>.
- [45] H. ROBBINS, S. MONRO: A stochastic approximation method, *The Annals of Mathematical Statistics*, **22**, 400–407 (1951).
- [46] D. P. KINGMA, J. BA: Adam: A method for stochastic optimization, *arXiv: 1412.6980* (2017).
- [47] D. C. LIU, J. NOCEDAL: On the limited memory bfgs method for large scale optimization, *Mathematical Programming*, **45**, 503–528 (1989).
- [48] A. SINGRABER, T. MORAWIETZ, J. BEHLER, C. DELLAGO: Parallel multi-stream training of high-dimensional neural network potentials, *Journal of Chemical Theory and Computation*, **15**, 3075–3092 (2019).
- [49] K. LEE, D. YOO, W. JEONG, S. HAN: Simple-nn: An efficient package for training and executing neural-network interatomic potentials, *Computer Physics Communications*, **242**, 95–103 (2019).
- [50] Y. SHAO, F. M. DIETRICH, C. NETTELBLAD, C. ZHANG: Training algorithm matters for the performance of neural network potential: A case study of adam and the kalman filter optimizers, *The Journal of Chemical Physics*, **155**, 204108 (2021).
- [51] K. S. NOVOSELOV, A. K. GEIM, S. V. MOROZOV, D. JIANG, Y. ZHANG, S. V. DUBONOS, I. V. GRIGORIEVA, A. A. FIRSOV: Electric field effect in atomically thin carbon films, *Science*, **306**, 666–669 (2004).
- [52] M. ZENG, Y. XIAO, J. LIU, K. YANG, L. FU: Exploring two-dimensional materials toward the next-generation circuits: From monomer design to assembly control, *Chemical Reviews*, **118**, 6236–6296 (2018).
- [53] S. V. MOROZOV, K. S. NOVOSELOV, M. I. KATSNELSON, F. SCHEDIN, D. C. ELIAS, J. A. JASZCZAK, A. K. GEIM: Giant intrinsic carrier mobilities in graphene and its bilayer, *Physical Review Letters*, **100** (2008).
- [54] Y. ZHANG, Y.-W. TAN, H. L. STORMER, P. KIM: Experimental observation of the quantum hall effect and berry’s phase in graphene, *Nature*, **438**, 201–204 (2005).
- [55] C. LEE, X. WEI, J. W. KYSAR, J. HONE: Measurement of the elastic properties and intrinsic strength of monolayer graphene, *Science*, **321**, 385–388 (2008).

- [56] A. A. BALANDIN, S. GHOSH, W. BAO, I. CALIZO, D. TEWELDEBRHAN, F. MIAO, C. N. LAU: Superior thermal conductivity of single-layer graphene, *Nano Letters*, **8**, 902–907 (2008).
- [57] H. LIU, A. T. NEAL, Z. ZHU, Z. LUO, X. XU, D. TOMÁNEK, P. D. YE: Phosphorene: An unexplored 2d semiconductor with a high hole mobility, *ACS Nano*, **8**, 4033–4041 (2014).
- [58] S. JOSEPH, J. MOHAN, S. LAKSHMY, S. THOMAS, B. CHAKRABORTY, S. THOMAS, N. KALARIKKAL: A review of the synthesis, properties, and applications of 2d transition metal dichalcogenides and their heterostructures, *Materials Chemistry and Physics*, **297**, 127332 (2023).
- [59] K. ZHOU, G. SHANG, H. HSU, S. HAN, V. A. L. ROY, Y. ZHOU: Emerging 2d metal oxides: From synthesis to device integration, *Advanced Materials*, **35** (2023).
- [60] B. ANASORI, M. R. LUKATSKAYA, Y. GOGOTSI: 2d metal carbides and nitrides (mxenes) for energy storage, *Nature Reviews Materials*, **2** (2017).
- [61] A. K. GEIM, I. V. GRIGORIEVA: Van der waals heterostructures, *Nature*, **499**, 419–425 (2013).
- [62] F. LIU, Z. FAN: Defect engineering of two-dimensional materials for advanced energy conversion and storage, *Chemical Society Reviews*, **52**, 1723–1772 (2023).
- [63] F. BANHART, J. KOTAKOSKI, A. V. KRASHENINNIKOV: Structural defects in graphene, *ACS Nano*, **5**, 26–41 (2010).
- [64] J. H. KIM, J. H. JEONG, N. KIM, R. JOSHI, G.-H. LEE: Mechanical properties of two-dimensional materials and their applications, *Journal of Physics D: Applied Physics*, **52**, 083001 (2018).
- [65] I. DERETZIS, G. FIORI, G. IANNACONE, A. LA MAGNA: Effects due to backscattering and pseudogap features in graphene nanoribbons with single vacancies, *Physical Review B*, **81** (2010).
- [66] S. MAHENDRAN, J. CARRETE, A. ISACSSON, G. K. H. MADSEN, P. ERHART: Quantitative predictions of the thermal conductivity in transition metal dichalcogenides: Impact of point defects in mos2 and ws2 monolayers, *The Journal of Physical Chemistry C*, **128**, 1709–1716 (2024).
- [67] T. XIE, J. C. GROSSMAN: Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties, *Physical Review Letters*, **120** (2018).
- [68] M. BERNARDI, C. ATACA, M. PALUMMO, J. C. GROSSMAN: Optical and electronic properties of two-dimensional layered materials, *Nanophotonics*, **6**, 479–493 (2017).
- [69] N. BARSAN, K. SCHIERBAUM (editors): *Gas sensors based on conducting metal oxides*, Metal Oxides, Elsevier Science Publishing, Philadelphia, PA (2018), ISBN: 9780128112243.

- [70] C. MERCKLING: *Monolithic Integration of InGaAs on Si(001) Substrate for Logic Devices*, 71–114, Elsevier (2018), ISBN: 9780081020616.
- [71] J. KOTAKOSKI, A. V. KRASHENINNIKOV, K. NORDLUND: Energetics, structure, and long-range interaction of vacancy-type defects in carbon nanotubes: Atomistic simulations, *Phys. Rev. B*, **74**, 245420 (2006).
- [72] L. LIANG, J. WANG, W. LIN, B. G. SUMPTER, V. MEUNIER, M. PAN: Electronic bandgap and edge reconstruction in phosphorene materials, *Nano Letters*, **14**, 6400 (2014).
- [73] X. WANG, A. M. JONES, K. L. SEYLER, V. TRAN, Y. JIA, H. ZHAO, H. WANG, L. YANG, X. XU, F. XIA: Highly anisotropic and robust excitons in monolayer black phosphorus, *Nature Nanotechnology*, **10**, 517 (2015).
- [74] J. QIAO, X. KONG, Z.-X. HU, F. YANG, W. JI: High-mobility transport anisotropy and linear dichroism in few-layer black phosphorus, *Nature Communications*, **5**, 8693 (2014).
- [75] B. LIAO, J. ZHOU, B. QIU, M. S. DRESSELHAUS, G. CHEN: Ab initio study of electron-phonon interaction in phosphorene, *Physical Review B*, **91** (2015).
- [76] T. SOHIER, D. CAMPI, N. MARZARI, M. GIBERTINI: Mobility of two-dimensional materials from first principles in an accurate and automated framework, *Physical Review Materials*, **2** (2018).
- [77] G. GADDEMANE, W. G. VANDENBERGHE, M. L. V. DE PUT, S. CHEN, S. TIWARI, E. CHEN, M. V. FISCHETTI: Theoretical studies of electronic transport in monolayer and bilayer phosphorene: A critical overview, *Physical Review B*, **98** (2018).
- [78] H. KALITA, M. BHUSHAN, L. ROBINDRO SINGH: A comprehensive review on theoretical concepts, types and applications of magnetic semiconductors, *Materials Science and Engineering: B*, **288**, 116201 (2023).
- [79] V. VIERIMAA, A. V. KRASHENINNIKOV, H.-P. KOMSA: Phosphorene under electron beam: from monolayer to one-dimensional chains, *Nanoscale*, **8**, 7949 (2016).
- [80] S. GUPTA, P. PERIASAMY, B. NARAYANAN: Defect dynamics in two-dimensional black phosphorus under argon ion irradiation, *Nanoscale*, **13**, 8575–8590 (2021).
- [81] Y. LEE, S. LEE, J.-Y. YOON, J. CHEON, H. Y. JEONG, K. KIM: Fabrication and imaging of monolayer phosphorene with preferred edge configurations via graphene-assisted layer-by-layer thinning, *Nano Letters*, **20**, 559 (2019).
- [82] F. YAO, Y. CAI, Z. XIAO, G. ZHANG, R.-J. XIE, C. JIN: In situ transmission electron microscopy study of the formation and migration of vacancy defects in atomically thin black phosphorus, *2D Materials*, **8**, 025004 (2020).

- [83] A. R. BABOUKANI, S. M. AGHAEI, I. KHAKPOUR, V. DROZD, A. AASI, C. WANG: Defects investigation of bipolar exfoliated phosphorene nanosheets, *Surface Science*, **720**, 122052 (2022).
- [84] Y. CAI, Q. KE, G. ZHANG, B. I. YAKOBSON, Y.-W. ZHANG: Highly itinerant atomic vacancies in phosphorene, *Journal of the American Chemical Society*, **138**, 10199 (2016).
- [85] S. WOLD, K. ESBENSEN, P. GELADI: Principal component analysis, *Chemometrics and Intelligent Laboratory Systems*, **2**, 37–52 (1987).
- [86] M. A. KRAMER: Nonlinear principal component analysis using autoassociative neural networks, *AIChE Journal*, **37**, 233–243 (1991).
- [87] M. NUNEZ: Exploring materials band structure space with unsupervised machine learning, *Computational Materials Science*, **158**, 117 (2019).
- [88] L. MCINNES, J. HEALY, J. MELVILLE: Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv: 1802.03426* (2018).
- [89] M. WATTENBERG, F. VIÉGAS, I. JOHNSON: How to use t-sne effectively, *Distill* (2016).
- [90] S. ROY: *Trustworthy Dimensionality Reduction*, Master’s thesis, Indian Statistical Institute (2024).
- [91] E. BECHT, L. MCINNES, J. HEALY, C.-A. DUTERTRE, I. W. H. KWOK, L. G. NG, F. GINHOUX, E. W. NEWELL: Dimensionality reduction for visualizing single-cell data using umap, *Nature Biotechnology*, **37**, 38–44 (2018).
- [92] Y. WANG, H. HUANG, C. RUDIN, Y. SHAPOSHNIK: Understanding how dimension reduction tools work: An empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization, *Journal of Machine Learning Research*, **22**, 1 (2021).
- [93] I. BATATIA, P. BENNER, Y. CHIANG, A. M. ELENA, D. P. KOVÁCS, J. RIEBELSELL, X. R. ADVINCULA, M. ASTA, M. AVAYLON, W. J. BALDWIN, F. BERGER, N. BERNSTEIN, A. BHOWMIK, S. M. BLAU, V. CĂRARE, J. P. DARBY, S. DE, F. D. PIA, V. L. DERINGER, R. ELIJOŠIUS, Z. EL-MACHACHI, F. FALCIONI, E. FAKO, A. C. FERRARI, A. GENREITH-SCHRIEVER, J. GEORGE, R. E. A. GOODALL, C. P. GREY, P. GRIGOREV, S. HAN, W. HANDLEY, H. H. HEENEN, K. HERMANSSON, C. HOLM, J. JAAFAR, S. HOFMANN, K. S. JAKOB, H. JUNG, V. KAPIL, A. D. KAPLAN, N. KARIMITARI, J. R. KERMODE, N. KROUPA, J. KULLGREN, M. C. KUNER, D. KURLA, G. LIEPUONIUTE, J. T. MARGRAF, I.-B. MAGDĂU, A. MICHAELIDES, J. H. MOORE, A. A. NAIK, S. P. NIBLETT, S. W. NORWOOD, N. O’NEILL, C. ORTNER, K. A. PERSSON, K. REUTER, A. S. ROSEN, L. L. SCHAAF, C. SCHRAN, B. X. SHI, E. SIVONXAY, T. K. STENCZEL, V. SVAHN, C. SUTTON, T. D. SWINBURNE, J. TILLY, C. VAN DER OORD, E. VARGA-UMBRICH, T. VEGGE, M. VONDRÁK, Y. WANG, W. C. WITT, F. ZILLS, G. CSÁNYI: A foundation model for atomistic materials chemistry, *arXiv: 2401.00096* (2024).
- [94] T. B. BLANK, S. D. BROWN, A. W. CALHOUN, D. J. DOREN: Neural network models of potential energy surfaces, *The Journal of Chemical Physics*, **103**, 4129 (1995).

- [95] S. LORENZ, A. GROSS, M. SCHEFFLER: Representing high-dimensional potential-energy surfaces for reactions at surfaces by neural networks, *Chemical Physics Letters*, **395**, 210 (2004).
- [96] K. V. J. JOSE, N. ARTRITH, J. BEHLER: Construction of high-dimensional neural network potentials using environment-dependent atom pairs, *The Journal of Chemical Physics*, **136**, 194111 (2012).
- [97] B. JIANG, H. GUO: Permutation invariant polynomial neural network approach to fitting potential energy surfaces, *The Journal of Chemical Physics*, **139**, 054112 (2013).
- [98] K. YAO, J. E. HERR, S. N. BROWN, J. PARKHILL: Intrinsic bond energies from a bonds-in-molecules neural network, *The Journal of Physical Chemistry Letters*, **8**, 2689 (2017).
- [99] J. S. SMITH, O. ISAYEV, A. E. ROITBERG: ANI-1: an extensible neural network potential with DFT accuracy at force field computational cost, *Chemical Science*, **8**, 3192 (2017).
- [100] K. T. SCHÜTT, H. E. SAUCEDA, P.-J. KINDERMANS, A. TKATCHENKO, K.-R. MÜLLER: SchNet – a deep learning architecture for molecules and materials, *The Journal of Chemical Physics*, **148**, 241722 (2018).
- [101] K. YAO, J. E. HERR, D. W. TOTH, R. MCKINTYRE, J. PARKHILL: The TensorMol-0.1 model chemistry: a neural network augmented with long-range physics, *Chemical Science*, **9**, 2261 (2018).
- [102] L. ZHANG, J. HAN, H. WANG, R. CAR, W. E: Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics, *Physical Review Letters*, **120** (2018).
- [103] Y. ZHANG, C. HU, B. JIANG: Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation, *The Journal of Physical Chemistry Letters*, **10**, 4962 (2019).
- [104] R. ZUBATYUK, J. S. SMITH, J. LESZCZYNSKI, O. ISAYEV: Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network, *Science Advances*, **5** (2019).
- [105] Z. L. GLICK, D. P. METCALF, A. KOUTSOUKAS, S. A. SPRONK, D. L. CHENEY, C. D. SHERRILL: AP-net: An atomic-pairwise neural network for smooth and transferable interaction potentials, *The Journal of Chemical Physics*, **153**, 044112 (2020).
- [106] J. WANG, D. CAO, C. TANG, L. XU, Q. HE, B. YANG, X. CHEN, H. SUN, T. HOU: DeepAtomicCharge: a new graph convolutional network-based architecture for accurate prediction of atomic charges, *Briefings in Bioinformatics*, **22** (2020).
- [107] X. XIE, K. A. PERSSON, D. W. SMALL: Incorporating electronic information into machine learning potential energy surfaces via approaching the ground-state electronic energy as a function of atom-based electronic populations, *Journal of Chemical Theory and Computation*, **16**, 4256 (2020).

- [108] M. LIU, J. R. KITCHIN: SingleNN: Modified behler–parrinello neural network with shared weights for atomistic simulations with transferability, *The Journal of Physical Chemistry C*, **124**, 17811 (2020).
- [109] J. GASTEIGER, J. GROSS, S. GÜNNEMANN: Directional message passing for molecular graphs, in *International Conference on Learning Representations (ICLR)* (2020).
- [110] P. THÖLKE, G. D. FABRITIIS: Equivariant transformers for neural network based molecular potentials, in *International Conference on Learning Representations* (2022).
- [111] I. BATATIA, S. BATZNER, D. P. KOVÁCS, A. MUSAELIAN, G. N. C. SIMM, R. DRAUTZ, C. ORTNER, B. KOZINSKY, G. CSÁNYI: The design space of $e(3)$ -equivariant atom-centered interatomic potentials, *arXiv: 2205.06643* (2022).
- [112] A. MUSAELIAN, S. BATZNER, A. JOHANSSON, L. SUN, C. J. OWEN, M. KORNBLOTH, B. KOZINSKY: Learning local equivariant representations for large-scale atomistic dynamics, *Nature Communications*, **14** (2023).
- [113] J. VANDERMAUSE, S. B. TORRISI, S. BATZNER, Y. XIE, L. SUN, A. M. KOLPAK, B. KOZINSKY: On-the-fly active learning of interpretable bayesian force fields for atomistic rare events, *npj Computational Materials*, **6** (2020).
- [114] R. M. BALABIN, E. I. LOMAKINA: Support vector machine regression (LS-SVM)—an alternative to artificial neural networks (ANNs) for the analysis of quantum chemistry data?, *Physical Chemistry Chemical Physics*, **13**, 11710 (2011).
- [115] A. E. A. ALLEN, G. DUSSON, C. ORTNER, G. CSÁNYI: Atomic permutationally invariant polynomials for fitting molecular force fields, *Machine Learning: Science and Technology*, **2**, 025017 (2021).
- [116] M. A. WOOD, A. P. THOMPSON: Extending the accuracy of the SNAP interatomic potential form, *The Journal of Chemical Physics*, **148**, 241721 (2018).
- [117] S. CHMIELA, H. E. SAUCEDA, I. POLTAVSKY, K.-R. MÜLLER, A. TKATCHENKO: sGDML: Constructing accurate and data efficient molecular force fields using machine learning, *Computer Physics Communications*, **240**, 38 (2019).
- [118] E. KOCER, T. W. KO, J. BEHLER: Neural network potentials: A concise overview of methods, *Annual Review of Physical Chemistry*, **73**, 163 (2022).
- [119] K. HORNIK, M. STINCHCOMBE, H. WHITE: Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, 359 (1989).
- [120] X. TANG, X. WANG, D. JIA: Efficient large scale kernel ridge regression via ensemble SPSP approximation, in *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence - ICAI '18*, ACM Press (2018).

- [121] S. N. POZDNYAKOV, M. J. WILLATT, A. P. BARTÓK, C. ORTNER, G. CSÁNYI, M. CERIOTTI: Incompleteness of atomic structure representations, *Physical Review Letters*, **125** (2020).
- [122] M. P. BIRCHER, A. SINGRABER, C. DELLAGO: Improved description of atomic environments using low-cost polynomial functions with compact support, *Machine Learning: Science and Technology*, **2**, 035026 (2021).
- [123] J. BEHLER: First principles neural network potentials for reactive simulations of large molecular and condensed systems, *Angewandte Chemie International Edition*, **56**, 12828 (2017).
- [124] R. RAMAKRISHNAN, P. O. DRAL, M. RUPP, O. A. VON LILIENFELD: Quantum chemistry structures and properties of 134 kilo molecules, *Scientific Data*, **1** (2014).
- [125] D. P. KOVÁCS, C. VAN DER OORD, J. KUCERA, A. E. A. ALLEN, D. J. COLE, C. ORTNER, G. CSÁNYI: Linear atomic cluster expansion force fields for organic molecules: Beyond RMSE, *Journal of Chemical Theory and Computation*, **17**, 7696 (2021).
- [126] K. T. SCHÜTT, S. CHMIELA, O. A. V. LILIENFELD, A. TKATCHENKO, K. TSUDA, K.-R. MÜLLER: *Machine Learning Meets Quantum Physics*, Springer Nature, Singapore (2020), ISBN: 9783030402457.
- [127] A. SINGRABER, J. BEHLER, C. DELLAGO: Library-based lammps implementation of high-dimensional neural network potentials, *Journal of Chemical Theory and Computation*, **15**, 1827 (2019).
- [128] L. CHEN: Curse of dimensionality, in *Encyclopedia of Database Systems*, 545–546, Springer US (2009).
- [129] A. P. BARTÓK, S. DE, C. POELKING, N. BERNSTEIN, J. R. KERMODE, G. CSÁNYI, M. CERIOTTI: Machine learning unifies the modeling of materials and molecules, *Science Advances*, **3** (2017).
- [130] F. A. FABER, A. S. CHRISTENSEN, B. HUANG, O. A. VON LILIENFELD: Alchemical and structural distribution based representation for universal quantum machine learning, *The Journal of Chemical Physics*, **148**, 241717 (2018).
- [131] M. J. WILLATT, F. MUSIL, M. CERIOTTI: Feature optimization for atomistic machine learning yields a data-driven construction of the periodic table of the elements, *Physical Chemistry Chemical Physics*, **20**, 29661 (2018).
- [132] F. BIGI, S. N. POZDNYAKOV, M. CERIOTTI: Wigner kernels: body-ordered equivariant machine learning without a basis, *arXiv: 2303.04124* (2023).
- [133] D. P. KOVÁCS, I. BATATIA, E. S. ARANY, G. CSÁNYI: Evaluation of the MACE force field architecture: From medicinal chemistry to materials science, *The Journal of Chemical Physics*, **159** (2023).
- [134] V. SHANMUGAM, R. A. MENSAH, K. BABU, S. GAWUSU, A. CHANDA, Y. TU, R. E. NEISIANY, M. FORSTH, G. SAS, O. DAS: A review of the synthesis, properties, and applications of 2d materials, *Particle and Particle Systems Characterization*, **39**, 2200031 (2022).

- [135] T. A. AMOLLO, G. T. MOLA, M. S. K. KIRUI, V. O. NYAMORI: Graphene for thermoelectric applications: Prospects and challenges, *Critical Reviews in Solid State and Materials Sciences*, **43**, 133 (2017).
- [136] A. RAWAT, N. JENA, D. DIMPLE, A. D. SARKAR: A comprehensive study on carrier mobility and artificial photosynthetic properties in group VI b transition metal dichalcogenide monolayers, *Journal of Materials Chemistry A*, **6**, 8693 (2018).
- [137] H. WANG, X. YANG, W. SHAO, S. CHEN, J. XIE, X. ZHANG, J. WANG, Y. XIE: Ultrathin black phosphorus nanosheets for efficient singlet oxygen generation, *Journal of the American Chemical Society*, **137**, 11376 (2015).
- [138] M. TAHIR, N. FATIMA, U. FATIMA, M. SAGIR: A review on the 2d black phosphorus materials for energy applications, *Inorganic Chemistry Communications*, **124**, 108242 (2021).
- [139] L. LI, Y. YU, G. J. YE, Q. GE, X. OU, H. WU, D. FENG, X. H. CHEN, Y. ZHANG: Black phosphorus field-effect transistors, *Nature Nanotechnology*, **9**, 372 (2014).
- [140] J. MIAO, L. ZHANG, C. WANG: Black phosphorus electronic and optoelectronic devices, *2D Materials*, **6**, 032003 (2019).
- [141] W. ZHENG, J. LEE, Z. GAO, Y. LI, S. LIN, S. P. LAU, L. Y. S. LEE: Laser-assisted ultrafast exfoliation of black phosphorus in liquid with tunable thickness for li-ion batteries, *Advanced Energy Materials*, **10** (2020).
- [142] J. DAI, X. C. ZENG: Bilayer phosphorene: Effect of stacking order on bandgap and its potential applications in thin-film solar cells, *The Journal of Physical Chemistry Letters*, **5**, 1289 (2014).
- [143] L. KOU, T. FRAUENHEIM, C. CHEN: Phosphorene as a superior gas sensor: Selective adsorption and distinct response, *The Journal of Physical Chemistry Letters*, **5**, 2675 (2014).
- [144] N. C. FREY, D. AKINWANDE, D. JARIWALA, V. B. SHENOY: Machine learning-enabled design of point defects in 2d materials for quantum and neuromorphic information processing, *ACS Nano*, **14**, 13406 (2020).
- [145] W. HU, J. YANG: Defects in phosphorene, *The Journal of Physical Chemistry C*, **119**, 20474 (2015).
- [146] Y. CAI, J. GAO, S. CHEN, Q. KE, G. ZHANG, Y.-W. ZHANG: Design of phosphorene for hydrogen evolution performance comparable to platinum, *Chemistry of Materials*, **31**, 8948 (2019).
- [147] S. M. ATASHZAR, S. JAVADIAN, H. GHARIBI, Z. REZAEI: Defective phosphorene as an anode material for high-performance li-, na-, and k-ion batteries: a first-principles study, *Nanoscale*, **12**, 20364 (2020).
- [148] T. AHMED, M. TAHIR, M. X. LOW, Y. REN, S. A. TAWFIK, E. L. H. MAYES, S. KURIAKOSE, S. NAWAZ, M. J. S. SPENCER, H. CHEN, M. BHASKARAN, S. SRIRAM, S. WALIA: Fully light-controlled memory and neuromorphic computation in layered black phosphorus, *Advanced Materials*, **33**, 2004207 (2020).

- [149] Y. REN, L. HU, J.-Y. MAO, J. YUAN, Y.-J. ZENG, S. RUAN, J.-Q. YANG, L. ZHOU, Y. ZHOU, S.-T. HAN: Phosphorene nano-heterostructure based memristors with broadband response synaptic plasticity, *Journal of Materials Chemistry C*, **6**, 9383 (2018).
- [150] G. KRESSE, J. FURTHMÜLLER: Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set, *Phys. Rev. B*, **54**, 11169 (1996).
- [151] G. KRESSE, D. JOUBERT: From ultrasoft pseudopotentials to the projector augmented-wave method, *Phys. Rev. B*, **59**, 1758 (1999).
- [152] J. P. PERDEW, K. BURKE, M. ERNZERHOF: Generalized gradient approximation made simple, *Phys. Rev. Lett.*, **77**, 3865 (1996).
- [153] W. G. HOOVER, A. J. C. LADD, B. MORAN: High-strain-rate plastic flow studied via nonequilibrium molecular dynamics, *Phys. Rev. Lett.*, **48**, 1818 (1982).
- [154] D. J. EVANS: Computer “experiment” for nonlinear thermodynamics of couette flow, *The Journal of Chemical Physics*, **78**, 3297 (1983).
- [155] M. PARRINELLO, A. RAHMAN: Crystal structure and pair potentials: A molecular-dynamics study, *Phys. Rev. Lett.*, **45**, 1196 (1980).
- [156] M. PARRINELLO, A. RAHMAN: Polymorphic transitions in single crystals: A new molecular dynamics method, *Journal of Applied Physics*, **52**, 7182 (1981).
- [157] G. HENKELMAN, B. P. UBERUAGA, H. JÓNSSON: A climbing image nudged elastic band method for finding saddle points and minimum energy paths, *The Journal of Chemical Physics*, **113**, 9901 (2000).
- [158] T. D. LOEFFLER, T. K. PATRA, H. CHAN, M. CHERUKARA, S. K. R. S. SANKARANARAYANAN: Active learning the potential energy landscape for water clusters from sparse training data, *The Journal of Physical Chemistry C*, **124**, 4907 (2020).
- [159] J. S. SMITH, B. NEBGEN, N. MATHEW, J. CHEN, N. LUBBERS, L. BURAKOVSKY, S. TRETIK, H. A. NAM, T. GERMANN, S. FENSIN, K. BARROS: Automated discovery of a robust interatomic potential for aluminum, *Nature Communications*, **12**, 1257 (2021).
- [160] L. KÝVALA, A. ANGELETTI, C. FRANCHINI, C. DELLAGO: Defected phosphorene dataset and trained hnp models [dataset], *Zenodo* (2023).
- [161] A. P. THOMPSON, H. M. AKTULGA, R. BERGER, D. S. BOLINTINEANU, W. M. BROWN, P. S. CROZIER, P. J. IN T VELD, A. KOHLMAYER, S. G. MOORE, T. D. NGUYEN, R. SHAN, M. J. STEVENS, J. TRANCHIDA, C. TROTT, S. J. PLIMPTON: LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales, *Computer Physics Communications*, **271**, 108171 (2022).
- [162] V. CALANDRINI, E. PELLEGRINI, P. CALLIGARI, K. HINSEN, G. KNELLER: nMoldyn - interfacing spectroscopic experiments, molecular dynamics simulations and models for time correlation functions, *École thématique de la Société Française de la Neutronique*, **12**, 201 (2011).

- [163] P. DE BUYL: tidynamics: A tiny package to compute the dynamics of stochastic and molecular simulations, *Journal of Open Source Software*, **3**, 877 (2018).
- [164] Y. GUO, J. ROBERTSON: Vacancy and doping states in monolayer and bulk black phosphorus, *Scientific Reports*, **5**, 14165 (2015).
- [165] T. HU, J. DONG: Geometric and electronic structures of mono- and di-vacancies in phosphorene, *Nanotechnology*, **26**, 065705 (2015).
- [166] X. LI, L. MA, D. WANG, X. C. ZENG, X. WU, J. YANG: Point defects in lines in single crystalline phosphorene: directional migration and tunable band gaps, *Nanoscale*, **8**, 17801 (2016).
- [167] J. GABERLE, A. L. SHLUGER: Structure and properties of intrinsic and extrinsic defects in black phosphorus, *Nanoscale*, **10**, 19536 (2018).
- [168] B. RIJAL, A. M. Z. TAN, C. FREYSOLDT, R. G. HENNIG: Charged vacancy defects in monolayer phosphorene, *Physical Review Materials*, **5**, 124004 (2021).
- [169] G. H. VINEYARD: Frequency factors and isotope effects in solid state rate processes, *Journal of Physics and Chemistry of Solids*, **3**, 121 (1957).
- [170] Y. CAI, S. CHEN, J. GAO, G. ZHANG, Y.-W. ZHANG: Evolution of intrinsic vacancies and prolonged lifetimes of vacancy clusters in black phosphorene, *Nanoscale*, **11**, 20987 (2019).
- [171] Z. LIN, C. WANG, Y. CHAI: Emerging Group-VI Elemental 2D Materials: Preparations, Properties, and Device Applications, *Small*, **16**, 2003319 (2020).
- [172] M. CHENG, J. YANG, X. LI, H. LI, R. DU, J. SHI, J. HE: Improving the device performances of two-dimensional semiconducting transition metal dichalcogenides: Three strategies, *Front. Phys.*, **17**, 63601 (2022).
- [173] L. KOU, C. CHEN, S. C. SMITH: Phosphorene: Fabrication, Properties, and Applications, *J. Phys. Chem. Lett.*, **6**, 2794 (2015).
- [174] C. E. HOUSECROFT, A. G. SHARPE: *Inorganic Chemistry (2nd Edition)*, Pearson Education, 2 edition (2004), ISBN: 9780130399137.
- [175] F. XIA, H. WANG, Y. JIA: Rediscovering black phosphorus as an anisotropic layered material for optoelectronics and electronics, *Nat. Commun.*, **5**, 4458 (2014).
- [176] X. LING, H. WANG, S. HUANG, F. XIA, M. S. DRESSELHAUS: The renaissance of black phosphorus, *PNAS*, **112**, 4523 (2015).
- [177] A. CARVALHO, M. WANG, X. ZHU, A. S. RODIN, H. SU, A. H. CASTRO NETO: Phosphorene: from theory to applications, *Nat. Rev. Mater.*, **1**, 1 (2016).
- [178] D. WAN, H. HUANG, Z. WANG, X. LIU, L. LIAO: Recent advances in long-term stable black phosphorus transistors, *Nanoscale*, **12**, 20089 (2020).

- [179] L. LI, J. KIM, C. JIN, G. J. YE, D. Y. QIU, F. H. DA JORNADA, Z. SHI, L. CHEN, Z. ZHANG, F. YANG, K. WATANABE, T. TANIGUCHI, W. REN, S. G. LOUIE, X. H. CHEN, Y. ZHANG, F. WANG: Direct observation of the layer-dependent electronic structure in phosphorene, *Nat. Nanotechnol.*, **12**, 21 (2017).
- [180] Q. WEI, X. PENG: Superior mechanical flexibility of phosphorene and few-layer black phosphorus, *Appl. Phys. Lett.*, **104**, 251915 (2014).
- [181] J. WU, N. MAO, L. XIE, H. XU, J. ZHANG: Identifying the Crystalline Orientation of Black Phosphorus Using Angle-Resolved Polarized Raman Spectroscopy, *Angew. Chem., Int. Ed.*, **54**, 2366 (2015).
- [182] T. LOW, A. S. RODIN, A. CARVALHO, Y. JIANG, H. WANG, F. XIA, A. H. CASTRO NETO: Tunable optical properties of multilayer black phosphorus thin films, *Phys. Rev. B*, **90**, 075434 (2014).
- [183] J.-W. JIANG, H. S. PARK: Negative poisson's ratio in single-layer black phosphorus, *Nat. Commun.*, **5**, 4727 (2014).
- [184] V. TRAN, R. SOKLASKI, Y. LIANG, L. YANG: Layer-controlled band gap and anisotropic excitons in few-layer black phosphorus, *Phys. Rev. B*, **89**, 235319 (2014).
- [185] R. FEI, A. FAGHANINIA, R. SOKLASKI, J.-A. YAN, C. LO, L. YANG: Enhanced Thermoelectric Efficiency via Orthogonal Electrical and Thermal Conductances in Phosphorene, *Nano Lett.*, **14**, 6393 (2014).
- [186] G. QIN, M. HU: Thermal Transport in Phosphorene, *Small*, **14**, 1702465 (2018).
- [187] J. PEI, X. GAI, J. YANG, X. WANG, Z. YU, D.-Y. CHOI, B. LUTHER-DAVIES, Y. LU: Producing air-stable monolayers of phosphorene and their defect engineering, *Nat. Commun.*, **7**, 10450 (2016).
- [188] S. KURIAKOSE, T. AHMED, S. BALENDHRAN, V. BANSAL, S. SRIRAM, M. BHASKARAN, S. WALIA: Black phosphorus: ambient degradation and strategies for protection, *2D Mater.*, **5**, 032001 (2018).
- [189] S. KRETSCHMER, T. LEHNERT, U. KAISER, A. V. KRASHENINNIKOV: Formation of Defects in Two-Dimensional MoS₂ in the Transmission Electron Microscope at Electron Energies below the Knock-on Threshold: The Role of Electronic Excitations, *Nano Lett.*, **20**, 2865 (2020).
- [190] C. SPECKMANN, J. LANG, J. MADSEN, M. R. A. MONAZAM, G. ZAGLER, G. T. LEUTHNER, N. McEVOY, C. MANGLER, T. SUSI, J. KOTAKOSKI: Combined electronic excitation and knock-on damage in monolayer MoS₂, *Phys. Rev. B*, **107**, 094112 (2023).
- [191] T. A. BUI, G. T. LEUTHNER, J. MADSEN, M. R. A. MONAZAM, A. I. CHIRITA, A. POSTL, C. MANGLER, J. KOTAKOSKI, T. SUSI: Creation of Single Vacancies in hBN with Electron Irradiation, *Small*, **19**, 2301926 (2023).
- [192] P. SRIVASTAVA, K. P. S. S. HEMBRAM, H. MIZUSEKI, K.-R. LEE, S. S. HAN, S. KIM: Tuning the Electronic and Magnetic Properties of Phosphorene by Vacancies and Adatoms, *J. Phys. Chem. C*, **119**, 6530 (2015).

- [193] J. GABERLE, A. L. SHLUGER: Structure and properties of intrinsic and extrinsic defects in black phosphorus, *Nanoscale*, **10**, 19536 (2018).
- [194] K. L. KUNTZ, R. A. WELLS, J. HU, T. YANG, B. DONG, H. GUO, A. H. WOOMER, D. L. DRUFFEL, A. ALABANZA, D. TOMÁNEK, S. C. WARREN: Control of Surface and Edge Oxidation on Phosphorene, *ACS Appl. Mater. Interfaces*, **9**, 9126 (2017).
- [195] Y. CHEN, X. SHI, M. LI, Y. LIU, H. XIAO, X. CHEN: Strain and defect engineering on phase transition of monolayer black phosphorene, *Phys. Chem. Chem. Phys.*, **20**, 21832 (2018).
- [196] C. D. ZHANG, J. C. LIAN, W. YI, Y. H. JIANG, L. W. LIU, H. HU, W. D. XIAO, S. X. DU, L. L. SUN, H. J. GAO: Surface Structures of Black Phosphorus Investigated with Scanning Tunneling Microscopy, *J. Phys. Chem. C*, **113**, 18823 (2009).
- [197] B. KIRALY, N. HAUPTMANN, A. N. RUDENKO, M. I. KATSNELSON, A. A. KHAJETOORIANS: Probing Single Vacancies in Black Phosphorus at the Atomic Level, *Nano Lett.*, **17**, 3607 (2017).
- [198] J. V. RIFFLE, C. FLYNN, B. ST. LAURENT, C. A. AYOTTE, C. A. CAPUTO, S. M. HOLLEN: Impact of vacancies on electronic properties of black phosphorus probed by STM, *J. Appl. Phys.*, **123**, 044301 (2018).
- [199] H. FANG, A. GALLARDO, D. DULAL, Z. QIU, J. SU, M. TELYCHKO, H. MAHALINGAM, P. LYU, Y. HAN, Y. ZHENG, Y. CAI, A. RODIN, P. JELÍNEK, J. LU: Electronic Self-Passivation of Single Vacancy in Black Phosphorus via Ionization, *Phys. Rev. Lett.*, **128**, 176801 (2022).
- [200] Z. QIU, H. FANG, A. CARVALHO, A. S. RODIN, Y. LIU, S. J. R. TAN, M. TELYCHKO, P. LV, J. SU, Y. WANG, A. H. CASTRO NETO, J. LU: Resolving the Spatial Structures of Bound Hole States in Black Phosphorus, *Nano Lett.*, **17**, 6935 (2017).
- [201] M. WENTINK, J. GABERLE, M. AGHAJANIAN, A. A. MOSTOFI, N. J. CURSON, J. LISCHNER, S. R. SCHOFIELD, A. L. SHLUGER, A. J. KENYON: Substitutional Tin Acceptor States in Black Phosphorus, *J. Phys. Chem. C*, **125**, 22883 (2021).
- [202] R. HARSH, S. MONDAL, D. SHARMA, M. BOUATOU, C. CHACON, M. ILYN, C. ROGERO, V. REPAIN, A. BELLEC, Y. GIRARD, S. ROUSSET, R. SANKAR, W. W. PAI, S. NARASIMHAN, J. LAGOUTE: Identification and Manipulation of Defects in Black Phosphorus, *J. Phys. Chem. Lett.*, **13** (2022).
- [203] Y. LEE, J.-Y. YOON, D. SCULLION, J. JANG, E. J. G. SANTOS, H. Y. JEONG, K. KIM: Atomic-scale imaging of few-layer black phosphorus and its reconstructed edge, *J. Phys. D: Appl. Phys.*, **50**, 084003 (2017).
- [204] S. LEE, Y. LEE, L. P. DING, K. LEE, F. DING, K. KIM: Atomically Sharp, Closed Bilayer Phosphorene Edges by Self-Passivation, *ACS Nano*, **16**, 12822 (2022).

- [205] Z. LI, Y. WANG, A. KOZBIAL, G. SHENOY, F. ZHOU, R. MCGINLEY, P. IRELAND, B. MORGANSTEIN, A. KUNKEL, S. P. SURWADE, L. LI, H. LIU: Effect of airborne contaminants on the wettability of supported graphene and graphite, *Nat. Mater.*, **12**, 925 (2013).
- [206] H. INANI, K. MUSTONEN, A. MARKEVICH, E.-X. DING, M. TRIPATHI, A. HUSSAIN, C. MANGLER, E. I. KAUPPINEN, T. SUSI, J. KOTAKOSKI: Silicon Substitution in Nanotubes and Graphene via Intermittent Vacancies, *J. Phys. Chem. C*, **123**, 13136 (2019).
- [207] A. PÁLINKÁS, G. KÁLVIN, P. VANCÓS, K. KANDRAI, M. SZENDRŐ, G. NÉMETH, M. NÉMETH, A. PEKKER, J. S. PAP, P. PETRIK, K. KAMARÁS, L. TAPASZTÓ, P. NEMES-INCZE: The composition and structure of the ubiquitous hydrocarbon contamination on van der Waals materials, *Nat. Commun.*, **13**, 6770 (2022).
- [208] R. TILMANN, C. BARTLAM, O. HARTWIG, B. TYWONIUK, N. DOMINIK, C. P. CULLEN, L. PETERS, T. STIMPEL-LINDNER, N. McEVOY, G. S. DUESBERG: Identification of Ubiquitously Present Polymeric Adlayers on 2D Transition Metal Dichalcogenides, *ACS Nano*, **17**, 10617 (2023).
- [209] J. J. SCHWARTZ, H.-J. CHUANG, M. R. ROSENBERGER, S. V. SIVARAM, K. M. MCCREARY, B. T. JONKER, A. CENTRONE: Chemical Identification of Interlayer Contaminants within van der Waals Heterostructures, *ACS Appl. Mater. Interfaces*, **11**, 25578 (2019).
- [210] T. SUSI, J. KOTAKOSKI, D. KEPAPTSOGLU, C. MANGLER, T. C. LOVEJOY, O. L. KRIVANEK, R. ZAN, U. BANGERT, P. AYALA, J. C. MEYER, Q. RAMASSE: Silicon–Carbon Bond Inversions Driven by 60-keV Electrons in Graphene, *Phys. Rev. Lett.*, **113**, 115501 (2014).
- [211] W. A. MCKINLEY, H. FESHBACH: The Coulomb Scattering of Relativistic Electrons by Nuclei, *Phys. Rev.*, **74**, 1759 (1948).
- [212] J. C. MEYER, F. EDER, S. KURASCH, V. SKAKALOVA, J. KOTAKOSKI, H. J. PARK, S. ROTH, A. CHUVILIN, S. EYHUSEN, G. BENNER, A. V. KRASHENINNIKOV, U. KAISER: Accurate Measurement of Electron Beam Induced Displacement Cross Sections for Single-Layer Graphene, *Phys. Rev. Lett.*, **108**, 196102 (2012).
- [213] J. C. MEYER, F. EDER, S. KURASCH, V. SKAKALOVA, J. KOTAKOSKI, H. J. PARK, S. ROTH, A. CHUVILIN, S. EYHUSEN, G. BENNER, A. V. KRASHENINNIKOV, U. KAISER: Erratum: Accurate Measurement of Electron Beam Induced Displacement Cross Sections for Single-Layer Graphene [Phys. Rev. Lett. 108, 196102 (2012)], *Phys. Rev. Lett.*, **110**, 239902 (2013).
- [214] T. SUSI, C. HOFER, G. ARGENTERO, G. T. LEUTHNER, T. J. PENNYCOOK, C. MANGLER, J. C. MEYER, J. KOTAKOSKI: Isotope analysis in the transmission electron microscope, *Nat. Commun.*, **7**, 13040 (2016).
- [215] A. CHIRITA, A. MARKEVICH, M. TRIPATHI, N. A. PIKE, M. J. VERSTRAETE, J. KOTAKOSKI, T. SUSI: Three-dimensional ab initio description of vibration-assisted electron knock-on displacements in graphene, *Phys. Rev. B*, **105**, 235419 (2022).

- [216] G. QIN, Q.-B. YAN, Z. QIN, S.-Y. YUE, M. HU, G. SU: Anisotropic intrinsic lattice thermal conductivity of phosphorene from first principles, *Phys. Chem. Chem. Phys.*, **17**, 4854 (2015).
- [217] L. KÝVALA, A. ANGELETTI, C. FRANCHINI, C. DELLAGO: Diffusion and coalescence of phosphorene monovacancies studied using high-dimensional neural network potentials, *J. Phys. Chem. C*, **127**, 23743–23751 (2023).
- [218] R. HULTGREN, N. S. GINGRICH, B. E. WARREN: The Atomic Distribution in Red and Black Phosphorus and the Crystal Structure of Black Phosphorus, *J. Chem. Phys.*, **3**, 351 (1935).
- [219] A. ALHASSAN, M. YU: Insight into the stacking effect on shifted patterns of bilayer phosphorene: a comprehensive first-principles study, *Nanotechnology*, **35**, 155701 (2024).
- [220] A. CASTELLANOS-GOMEZ, M. BUSCEMA, R. MOLENAAR, V. SINGH, L. JANSSEN, H. S. J. v. D. ZANT, G. A. STEELE: Deterministic transfer of two-dimensional materials by all-dry viscoelastic stamping, *2D Mater.*, **1**, 011002 (2014).
- [221] C. MANGLER, J. MEYER, A. MITTELBERGER, K. MUSTONEN, T. SUSI, J. KOTAKOSKI: A Materials Scientist’s CANVAS: A System for Controlled Alteration of Nanomaterials in Vacuum Down to the Atomic Scale, *Microsc. Microanal.*, **28**, 2940 (2022).
- [222] L. KÝVALA, A. ANGELETTI: Dataset of adatom-monovacancy complexes in phosphorene and trained hdnnp models, *Zenodo* (2024).
- [223] J. MADSEN, T. SUSI: abTEM: Transmission Electron Microscopy from First Principles, *Open Res. Europe*, **1**, 13015 (2021).
- [224] K. MOMMA, F. IZUMI: VESTA 3 for three-dimensional visualization of crystal, volumetric and morphology data, *J. Appl. Crystallogr.*, **44**, 1272 (2011).
- [225] J. BEHLER, G. CSÁNYI: Machine learning potentials for extended systems: a perspective, *The European Physical Journal B*, **94**, 1 (2021).
- [226] P. GEIGER, C. DELLAGO: Neural networks for local structure detection in polymorphic systems, *The Journal of chemical physics*, **139** (2013).
- [227] M. FULFORD, M. SALVALAGLIO, C. MOLTENI: Deepice: A deep neural network approach to identify ice and water molecules, *Journal of Chemical Information and Modeling*, **59**, 2141 (2019).
- [228] M. DIJKSTRA, E. LUIJTEN: From predictive modelling to machine learning and reverse engineering of colloidal self-assembly, *Nature materials*, **20**, 762 (2021).
- [229] R. S. DEFEVER, C. TARGONSKI, S. W. HALL, M. C. SMITH, S. SARUPRIA: A generalized deep learning approach for local structure identification in molecular simulations, *Chemical science*, **10**, 7503 (2019).

- [230] S. ISHIAI, K. ENDO, K. YASUOKA: Graph neural networks classify molecular geometry and design novel order parameters of crystal and liquid, *The Journal of Chemical Physics*, **159** (2023).
- [231] S. ISHIAI, K. ENDO, P. E. BRUMBY, A. K. SUM, K. YASUOKA: Novel approach for designing order parameters of clathrate hydrate structures by graph neural network, *The Journal of Chemical Physics*, **160** (2024).
- [232] D. KUROSHIMA, M. KILGOUR, M. E. TUCKERMAN, J. ROGAL: Machine learning classification of local environments in molecular crystals, *Journal of Chemical Theory and Computation*, **20**, 6197–6206 (2024).
- [233] J. HUANG, G. HUANG, S. LI: A machine learning model to classify dynamic processes in liquid water, *ChemPhysChem*, **23**, e202100599 (2022).
- [234] I. MOSQUERA-LOIS, S. R. KAVANAGH, A. WALSH, D. O. SCANLON: Identifying the ground state structures of point defects in solids, *npj Computational Materials*, **9**, 25 (2023).
- [235] J. WANG, T. LI, Z. FU, B. FU, C. GOU: Identification of self-interstitial atoms and vacancies in crystalline materials in atomistic simulation, *Computer Physics Communications*, **299**, 109138 (2024).
- [236] A. GORFER, R. ABART, C. DELLAGO: Structure and thermodynamics of defects in na-feldspar from a neural network potential, *Physical Review Materials*, **8**, 073602 (2024).
- [237] A. P. BARTÓK, R. KONDOR, G. CSÁNYI: On representing chemical environments, *Physical Review B*, **87**, 184115 (2013).
- [238] M. CERIOTTI: Unsupervised machine learning in atomistic simulations, between predictions and understanding, *The Journal of chemical physics*, **150** (2019).
- [239] S. FRICKENHAUS, S. KANNAN, M. ZACHARIAS: Efficient evaluation of sampling quality of molecular dynamics simulations by clustering of dihedral torsion angles and sammon mapping, *Journal of computational chemistry*, **30**, 479 (2009).
- [240] M. L. TEODORO, G. N. PHILLIPS JR, L. E. KAVRAKI: Understanding protein flexibility through dimensionality reduction, *Journal of Computational Biology*, **10**, 617 (2003).
- [241] H. BABA, R. URANO, T. NAGAI, S. OKAZAKI: Prediction of self-diffusion coefficients of chemically diverse pure liquids by all-atom molecular dynamics simulations, *Journal of Computational Chemistry*, **43**, 1892 (2022).
- [242] V. SPIWOK, B. KRÁLOVÁ: Metadynamics in the conformational space nonlinearly dimensionally reduced by isomap, *The Journal of Chemical Physics*, **135** (2011).

- [243] N. G. SGOURAKIS, M. MERCED-SERRANO, C. BOUTSIDIS, P. DRINEAS, Z. DU, C. WANG, A. E. GARCIA: Atomic-level characterization of the ensemble of the $\alpha\beta$ (1–42) monomer in water using unbiased molecular dynamics simulations and spectral algorithms, *Journal of molecular biology*, **405**, 570 (2011).
- [244] A. GLIELMO, B. E. HUSIC, A. RODRIGUEZ, C. CLEMENTI, F. NOÉ, A. LAIO: Unsupervised learning methods for molecular simulation data, *Chemical Reviews*, **121**, 9722 (2021).
- [245] J. CHAPMAN, N. GOLDMAN, B. C. WOOD: Efficient and universal characterization of atomic structures through a topological graph order parameter, *npj Computational Materials*, **8**, 37 (2022).
- [246] W. F. REINHART, A. W. LONG, M. P. HOWARD, A. L. FERGUSON, A. Z. PANAGIOTOPOULOS: Machine learning for autonomous crystal structure identification, *Soft Matter*, **13**, 4733 (2017).
- [247] W. F. REINHART, A. Z. PANAGIOTOPOULOS: Automated crystal characterization with a fast neighborhood graph analysis method, *Soft matter*, **14**, 6083 (2018).
- [248] M. SPELLINGS, S. C. GLOTZER: Machine learning for crystal identification and discovery, *AIChE Journal*, **64**, 2198 (2018).
- [249] E. BOATTINI, M. DIJKSTRA, L. FILION: Unsupervised learning for local structure detection in colloidal systems, *The Journal of chemical physics*, **151** (2019).
- [250] X. ZHANG, Y. YAO, H. LI, A. PYTHON, K. MOCHIZUKI: Fast crystal growth of ice vii owing to the decoupling of translational and rotational ordering, *Communications Physics*, **6**, 164 (2023).
- [251] W. F. REINHART: Unsupervised learning of atomic environments from simple features, *Computational Materials Science*, **196**, 110511 (2021).
- [252] A. OFFEI-DANSO, A. HASSANALI, A. RODRIGUEZ: High-dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning, *Journal of Chemical Theory and Computation*, **18**, 3136 (2022).
- [253] S. DI PINO, E. D. DONKOR, V. M. SÁNCHEZ, A. RODRIGUEZ, G. CASSONE, D. SCHERLIS, A. HASSANALI: Zundeig: The structure of the proton in liquid water from unsupervised learning, *The Journal of Physical Chemistry B*, **127**, 9822 (2023).
- [254] E. D. DONKOR, A. OFFEI-DANSO, A. RODRIGUEZ, F. SCIORTINO, A. HASSANALI: Beyond local structures in critical supercooled water through unsupervised learning, *The Journal of Physical Chemistry Letters*, **15**, 3996 (2024).
- [255] J. B. KRUSKAL: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika*, **29**, 1 (1964).
- [256] H. HUANG, Y. WANG, C. RUDIN, E. P. BROWNE: Towards a comprehensive evaluation of dimension reduction methods for transcriptomic data visualization, *Communications Biology*, **5** (2022).

- [257] P. M. PIAGGI, J. WEIS, A. Z. PANAGIOTOPOULOS, P. G. DEBENEDETTI, R. CAR: Homogeneous ice nucleation in an ab initio machine-learning model of water, *Proceedings of the National Academy of Sciences*, **119**, e2207294119 (2022).
- [258] P. MONTERO DE HIJES, J. R. ESPINOSA, C. VEGA, C. DELLAGO: Minimum in the pressure dependence of the interfacial free energy between ice Ih and water, *The Journal of Chemical Physics*, **158** (2023).
- [259] G. A. VORONIN, C. PANTEA, T. W. ZERDA, L. WANG, Y. ZHAO: In situ x-ray diffraction study of silicon at pressures up to 15.5 gpa and temperatures up to 1073 K, *Physical Review B*, **68** (2003).
- [260] B. MONSERRAT, J. G. BRANDENBURG, E. A. ENGEL, B. CHENG: Liquid water contains the building blocks of diverse ice phases, *Nature communications*, **11**, 5757 (2020).
- [261] A. P. BARTÓK, J. KERMODE, N. BERNSTEIN, G. CSÁNYI: Machine learning a general-purpose interatomic potential for silicon, *Physical Review X*, **8** (2018).
- [262] T. MORAWIETZ, A. SINGRABER, C. DELLAGO, J. BEHLER: How van der Waals interactions determine the unique properties of water, *Proceedings of the National Academy of Sciences*, **113**, 8368–8373 (2016).
- [263] M. MATSUMOTO, T. YAGASAKI, H. TANAKA: Genice: Hydrogen-disordered ice generator, *Journal of Computational Chemistry*, **39**, 61 (2017).
- [264] J. ABASCAL, E. SANZ, R. GARCÍA FERNÁNDEZ, C. VEGA: A potential model for the study of ices and amorphous water: Tip4p/ice, *The Journal of chemical physics*, **122** (2005).
- [265] M. J. ABRAHAM, T. MURTOLO, R. SCHULZ, S. PÁLL, J. C. SMITH, B. HESS, E. LINDAHL: Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers, *SoftwareX*, **1**, 19 (2015).
- [266] A. STUKOWSKI: Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool, *Modelling and Simulation in Materials Science and Engineering*, **18** (2010).
- [267] F. PEDREGOSA, G. VAROQUAUX, A. GRAMFORT, V. MICHEL, B. THIRION, O. GRISEL, M. BLONDEL, P. PRETTENHOFER, R. WEISS, V. DUBOURG, J. VANDERPLAS, A. PASSOS, D. COURNAPEAU, M. BRUCHER, M. PERROT, E. DUCHESNAY: Scikit-learn: Machine learning in Python, *Journal of Machine Learning Research*, **12**, 2825 (2011).
- [268] L. MCINNIS, J. HEALY, J. MELVILLE: Umap: Uniform manifold approximation and projection for dimension reduction, *arXiv: 1802.03426* (2018).
- [269] A. PASZKE, S. GROSS, F. MASSA, A. LERER, J. BRADBURY, G. CHANAN, T. KILLEEN, Z. LIN, N. GIMELSHEIN, L. ANTIGA, A. DESMAISON, A. KOPF, E. YANG, Z. DEVITO, M. RAISON, A. TEJANI, S. CHILAMKURTHY, B. STEINER,

- L. FANG, J. BAI, S. CHINTALA: Pytorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems* 32, 8024–8035, Curran Associates, Inc. (2019).
- [270] L. KÝVALA, P. MONTERO DE HIJES: Atomistic potential descriptors and structures of defected water and silicon systems [dataset], *Zenodo* (2024).